Automated Subject Classification of Textual Documents in the Context of Web-Based Hierarchical Browsing

Golub, Koraljka

2007

Automated Subject Classification of Textual Documents in the Context of Web-Based Hierarchical Browsing

Automated Subject Classification of Textual Documents in the Context of Web-Based Hierarchical Browsing

Ph.D. Thesis

Koraljka Golub

LUND UNIVERSITY

Department of Electrical and Information Technology,
Faculty of Engineering

ISBN: 91-7167-042-4
ISRN: LUTEDX/TEIT-07/1038-SE

Printed in E-huset, Lund, Sweden

To my sister, parents and own relief

Contents

Abstract xiii

Preface xv

Acknowledgments xvii

Summary **1**

1 Introduction 1

2 Background 3
- 2.1 Terminology 3
- 2.2 Approaches to automated classification (II) 3
- 2.3 Evaluation challenge 5
- 2.4 Hierarchical subject browsing 7

3 Methodology 8
- 3.1 User studies and performance measures 8
- 3.2 Document collections 10
- 3.3 Engineering Information thesaurus and classification scheme 11
- 3.4 The string-matching classification algorithm 12

4 Results 13
- 4.1 Usage of web-based browsing (I, VIII) 13
- 4.2 Suitability of DDC and Ei for hierarchical web-based browsing (I, VIII) 14
- 4.3 Improving the string-matching algorithm (III, IV, V, VI) 15
- 4.4 The string-matching algorithm on an abstracts collection (VI) 20
- 4.5 String-matching algorithm on a web-page collection (VIII) 22

5 Concluding remarks 22

References 23

I. Users Browsing Behaviour in a DDC-Based Web Service: A Log Analysis **29**

1 Introduction 31

2 Background 33

3 Methodology 34
3.1 Cleaning the log files 34
3.2 Defining sessions 35
3.3 Defining activity types 35
3.4 Creating datasets for studying information-seeking behaviour 36

4 Results 36
4.1 Global usage 36
4.2 Information-seeking activities 37
4.3 DDC usage 43

5 Conclusion 50
5.1 Future work 52

Acknowledgments 53

References 53

II. Automated Subject Classification of Textual Web Documents **57**

1 Introduction 59

2 Approaches to automated classification 62
2.1 Text categorization 62
2.2 Document clustering 67
2.3 Document classification 72
2.4 Mixed approach 77

3 Discussion 78
3.1 Features of automated classification approaches 78
3.2 Evaluation 79
3.3 Application for subject browsing 80

Acknowledgments 81

References 81

III. Automated Subject Classification of Textual Web Pages, Based on a Controlled Vocabulary: Challenges and Recommendations 93

1 Introduction 95

2 Related work 97

3 Approach 99
3.1 Algorithm 99
3.2 Term list 101
3.3 Document collection 102

4 Methodology 103

5 Problems identified 103
5.1 Class not found at all 103
5.2 Class found but below threshold 106
5.2.1 Example demonstrating the problem 107
5.2.2 Recommendations 107
5.3 Wrong automatically assigned class 108
5.3.1 Recommendations 110
5.4 Automatically assigned class that is not really wrong 111
5.4.1 Recommendations 111

6 Concluding remarks 111

Acknowledgments 112

References 112

IV. Importance of HTML Structural Elements and Metadata in Automated Subject Classification 117

1 Introduction 119

2 Background 120
2.1 Related work 120
2.2 Evaluation challenge 121
2.3 Description of the algorithm 121

3 Methodology 122
3.1 Document collection 122
3.2 Methods for evaluation and deriving significance indicators 123

4 Significance indicators 125
4.1 General 125
4.2 Precision and recall 125
4.3 Semantic distance 127
4.4 Deriving significance indicators 127

4.5 Evaluation 129

5 Conclusion 130

Acknowledgments 131

References 131

V. The Role of Different Thesaurus Terms and Captions in Automated Subject Classification 135

1 Introduction 137

2 Methodology 138
2.1 String-matching algorithm 138
2.2 Document collection 139
2.3 Term lists 140
2.4 Stop-word list and stemming 141
2.5 Evaluation methodology 141

3 Experimental results 142
3.1 Averaged results for all the classes 142
3.2 Individual classes 144

4 Concluding remarks 146

Acknowledgments 146

References 147

VI. Automated Classification of Textual Documents Based on a Controlled Vocabulary in Engineering 151

1 Introduction 153

2 Methodology 156
2.1 String-matching algorithm 156
2.2 Document collection 160
2.3 Evaluation methodology 161

3 Improving the algorithm 164
3.1 Term weights 165
3.2 Cut-offs 169
3.3 Enriching the term list with new terms 170

4 Results 172
4.1 Weights and cut-offs 172
4.2 Enhancing the term list with new terms 178

4.3 Terms analysis and shortened term lists 179

5 Conclusion 181

Acknowledgments 182

References 182

VII. Comparing and Combining Two Approaches to Automated Subject Classification of Text 187

1 Introduction 189

2 Methodology 191
2.1 String-matching algorithm 191
2.2 Linear support vector machines (SVM) algorithm 195
2.3 Document collection 196
2.4 Evaluation 197
2.5 Term lists (features) 198

3 Results 199
3.1 Comparison in original settings 199
3.2 Combining term lists (features) 200

4 Conclusion 202

Acknowledgments 202

References 203

VIII. Automated Subject Classification of Engineering Web Pages in Hierarchical Browsing: A User Study 207

1 Introduction 209

2 Background 210
2.1 Classification algorithm 210
2.2 Engineering Information classification scheme 214

3 Methodology 214
3.1 Web page collection 214
3.2 User study design 214

4 Results 221
4.1 Browsing 221
4.2 Automatically assigned classes 229

5 Conclusions 237

Acknowledgments 240

References 240

Appendix 1. Hierarchical tree for the class 9 (Engineering, General) of the Engineering Information classification scheme. 243

Appendix 2. Pre-experiment questionnaire. 247

Appendix 3. Post-experiment questionnaire. 249

Appendix 4. Hierarchical view of an ideal sequence of browsing steps for each task. 251

Appendix 5. Tasks rotation scheme. 253

Appendix 6. Instructions sheet. 255

Appendix 7. Main user study screen. 259

Abstract

With the exponential growth of the World Wide Web, automated subject classification has become a major research issue. Organizing web pages into a hierarchical structure for subject browsing has been gaining more recognition as an important tool in information-seeking processes.

The most frequent approach to automated classification is machine learning. It, however, requires training documents and performs well on new documents only if they are similar enough to the former. In the thesis, a string-matching algorithm based on a controlled vocabulary was explored. It does not require training documents, but instead reuses the intellectual work invested into creating the controlled vocabulary. Terms from the Engineering Information thesaurus and classification scheme were matched against text of documents to be classified. Plain string-matching was enhanced in several ways, including term weighting with cut-offs, exclusion of certain terms, and enrichment of the controlled vocabulary with automatically extracted terms. The final results were comparable to those of state-of-the-art machine-learning algorithms, especially for particular classes. Concerning web pages, it was indicated that all the structural information and metadata available in web pages should be used in order to achieve the best automated classification results; however, the exact way of combining them proved not to be very important.

In the context of browsing, the biggest difference between three approaches to automated classification (machine learning, information retrieval, library science) is whether they use controlled

vocabularies. It has been claimed that well-structured, high-quality classification schemes, such as those used predominantly in library science approaches, could serve as good browsing structures. In the thesis it was shown that Dewey Decimal Classification and Engineering Information classification scheme are suitable for the task. Moreover, a log analysis of a large web-based service using Dewey Decimal Classification demonstrated that browsing is used to a much larger degree than searching.

The final conclusion is that an appropriate controlled vocabulary, with a large number of entry vocabulary designating classes, could be utilised in automated classification. If the same controlled vocabulary has an appropriate hierarchical structure, it could at the same time provide a good browsing structure to the automatically classified collection of documents.

Preface

The thesis is a compilation of papers listed below. The papers are re-printed with minor changes in formatting, grammar, style and obvious errors corrected. References to the papers are made throughout the Summary, using upper-case Roman numbers associated with the papers (**I**, **II**, **III**, **IV**, **V**, **VI**, **VII**, and **VIII**).

I. Koch, T., Golub, K. and Ardö, A. (2006), "Users browsing behaviour in a DDC-based web service: a log analysis", *Cataloging & Classification Quarterly,* Vol. 42 No. 3/4, pp. 163-186.

II. Golub, K. (2006), "Automated subject classification of textual web documents", *Journal of Documentation,* Vol. 62 No. 3, pp. 350-371.

III. Golub, K. (2006), "Automated subject classification of textual web pages, based on a controlled vocabulary: challenges and recommendations", *New Review of Hypermedia and Multimedia*, Vol. 12 No. 1, pp. 11-27.

IV. Golub, K., and Ardö, A. (2005), "Importance of HTML structural elements and metadata in automated subject classification", *Proceedings of the 9th European Conference on Research and Advanced Technology for Digital Libraries, Vienna, Austria, 18-23 September*, pp. 368-378.

V. Golub, K. (2006), "The role of different thesaurus terms in automated subject classification of text", *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, Hong Kong, China, 18-22 December,* pp. 961-965.

VI. Golub, K., Hamon, T., and Ardö, A. (2007), "Automated classification of textual documents based on a controlled vocabulary in engineering", submitted to *Knowledge Organization*.

VII. Golub, K., Ardö, A., Mladenić, D., and Grobelnik, M. (2006), "Comparing and combining two approaches to automated subject classification of text", *Proceedings of 10th European Conference on Research and Advanced Technology for Digital Libraries, Alicante, Spain, 17-22 September*, pp. 467-470.

VIII. Golub, K. (2007), "Automated subject classification of engineering web pages in hierarchical browsing: a user study", to be submitted.

Acknowledgments

Hats off to Anders Ardö! He has been the best supervisor in so many ways, impossible to list every one of them. Above all I appreciated that he made himself available for discussions at any time; and always with a smile. As his first Ph.D. student, I can only hope he will not give up mentoring because I took up so much of his energy.

Traugott Koch, the man of strongest opinions, has been one of the most outspoken critics of my work. Deepest thanks for his will to stay around even when he moved from Sweden and for helping with my work during his leisure time.

Many thanks to numerous colleagues for valuable feedback on earlier versions of my work: Tatjana Aparac Jelušić, all teachers at the Nordic Research School in Library and Information Science (NORSLIS, http://www.norslis.net/), in particular Pia Borlund, Wolfgang Glänzel, Birger Hjørland, Birger Larsen, Kalervo Järvelin, Marianne Lykke Nielsen, Nils Pharo, Jesper Schneider, and Irene Wormell. I thank NORSLIS as a whole for providing a wonderful network of researchers, many of whom became great colleagues and friends and helped me go through with the studies.

I would also like to thank the opponent, Douglas Tudhope, as well as members of the Ph.D. committee, Lennart Björneborn, Marianne Lykke Nielsen, Birgitta Olander and Ingeborg Sölvberg, for taking time out from their busy schedules to review this thesis.

This research was funded by the Swedish Agency for Innovation Systems (P22504-1 A), the European Union Sixth Frameworks Programme on Information Society Technologies (Alvis, IST-1-002068-STP) and DELOS Network of Excellence on Digital Libraries.

I thank all those who contributed to this thesis with informal reviews, suggestions, and proofreading, in particular Johan Eklund, Ingo Frommholz, Repke de Vries, Jessica Lindholm, Boris Badurina, Sanjica Faletar Tanacković, Martina Dragija Ivanović, Liv Fugl; as well as friends outside the field: Jonas, Goran, Kristoffer, Bruce, Andreas and Håkan.

I also wish to thank very much everyone from the Department of Electrical and Information Technology and the Department of Cultural Sciences for all their help and support, as well as for creating a wonderful working milieu. Thank you Marcus, Maja, Martin, Håkan, Thomas, Fredrik, Sasha, Dima, Sara, Gunilla and Olof. I remain especially indebted to Suleyman.

Finally, I would not have ended up in Sweden in the first place if it were not for my Croatian teachers, especially Tatjana Aparac Jelušić, Aleksandra Horvat, and Jadranka Lasić-Lazić, who sparked my interest and opened up the world of information science to me. The same goes for all other Croatian colleagues and friends of mine who made science so much fun. LIDA (Libraries in the Digital Age) conferences were another crucial stepping stone; special thanks to Emil Levine!

I remain indebted to all of them as well as to so many other colleagues, friends and relatives spread around the world, for their wonderful support and standing by my side in spite of the distance, and even when I was a real pain in the behind. Thank you Mum, Dad, Višnja, Sanjica, Martina, Karmenka, Marko, Andrea, Jasna, Suzi, Nikola, Iva, Ivana, Tatjana, Tihana, Darko, Ampi, Nevena, Gordana, Irena and all the others!

In Lund, 1 October 2007

Koraljka

Summary

1 Introduction

Automated subject classification of textual documents has been a challenging research issue for several decades now, major motivation being the high cost of human-based subject classification. The interest has especially grown in the 1990s when the number of digital documents started to increase exponentially. Because of the vast amount of available documents it was recognized that established objectives of bibliographic systems could be left behind (Svenonius 2000, 20-21), and that automated means could be a solution to preserve them (30). Automated subject classification of textual documents today finds its use in a wide variety of applications, such as hierarchical organization of documents for browsing, e-mail filtering, focused crawling and many others (see Sebastiani 2002, 6-9).

The dominant approach to automated subject classification is the one based on machine learning (Sebastiani 2002). While reported results can be rather good, evaluation is most often conducted in laboratory-like conditions, i.e., outside the application context. Moreover, machine learning requires a collection of human-classified documents from which an algorithm learns; such collections are in many subject areas and document types unavailable.

Organizing web pages into a hierarchical structure for subject browsing is gaining more recognition as an important tool in information-seeking processes. However, in comparison to searching little research has been conducted on browsing, while

automated subject classification in the context of browsing has been hardly studied at all.

The aim of the thesis was to study approaches to automated subject classification in the application context of hierarchical browsing. A main focus was to explore a string-matching approach to automated subject classification which does not require pre-classified documents but instead utilizes the intellectual work that was put into building a good controlled vocabulary. Advantages and challenges of automatically classifying *web pages* were examined in particular. Automated subject classification was examined not only by comparison to pre-assigned classes, but also through users' judgements on the correct placement of documents while browsing. Browsing behaviour was also studied in two different situations: a large human-classified web page collection, and a collection of automatically classified web pages.

Research questions of the thesis were many. Major ones include the following:

1) How is hierarchical browsing in a large web service used (if at all)?
2) Are established classification schemes such as Dewey Decimal Classification (OCLC Dewey Services 2007) and Engineering Information (Milstead 1995) suitable for hierarchical web-based browsing?
3) Which approaches to automated subject classification exist, and what are their advantages and disadvantages, especially in relation to hierarchical web-based browsing?
4) What are the challenges of applying a string-matching classification algorithm on a collection of pre-classified web pages?
5) How can automated subject classification performance of the string-matching algorithm be improved?
6) Which automated subject classification performance can the string-matching algorithm yield when measured on a collection of pre-classified paper abstracts?
7) Which automated subject classification performance can the string-matching algorithm yield when measured on a collection of harvested web pages and evaluated by end-users?

The Summary is structured as follows: in the second section (Background) general information such as definitions and research challenges are given; third section (Methodology) describes the main classification algorithm, document collections and performance measures; in fourth section (Results) major results are presented and discussed; and, finally, concluding remarks and implications for further research are given (Concluding remarks).

2 Background

2.1 Terminology

Classification is, to the purpose of this thesis, defined as "…the multistage process of deciding on a property or characteristic of interest, distinguishing things or objects that possess that property from those which lack it, and grouping things or objects that have the property or characteristic in common into a class. Other essential aspects of classification are establishing relationships among classes and making distinctions within classes to arrive at subclasses and finer divisions" (Chan 1994, 259).

Automated subject classification (in further text: automated classification) denotes machine-based organization of related information objects into topically related groups. In this process human intellectual processes are replaced by, for example, statistical and computational linguistics techniques. In the related literature automated classification can also be referred to as *automated indexing* (Moens 2000; Lancaster 2003). Terms *automatic* and *automated* are both used. Here the term *automated* is chosen because it more directly implies that the process is machine-based.

In difference to searching, *browsing* in general relies on recognition of patterns (e.g. sequences of words) rather than recall from memory of search terms (Large et al. 1999, 179). *Hierarchical browsing* in this thesis refers to using a hierarchical tree structure in which information resources are organized by topic.

2.2 Approaches to automated classification (II)

In my opinion and as discussed in **II**, one can distinguish between three major approaches to automated classification, viewed in the context of browsing: text categorization (machine learning), document clustering (information retrieval), and document

classification (library science) (research question **3**). In their broadest sense these terms could be considered synonymous, which is probably one of the reasons why they are interchangeably used in the literature. In this thesis terms *text categorization* and *document clustering* are chosen because they tend to be the prevalent terms in the literature of the corresponding research communities, while *document classification* is used to consistently distinguish between the three approaches.

In document clustering, both clusters (classes) into which documents are classified and, to a limited degree, relationships between them are automatically produced. Labelling the clusters is a major research problem, with relationships between them, such as those of equivalence, related-term and hierarchical relationships, being even more difficult to automatically derive (Svenonius 2000, 168). In addition, "[a]utomatically-derived structures often result in heterogeneous criteria for category membership and can be difficult to understand" (Chen and Dumais 2000, 146). Also, clusters' labels and relationships between them change as new documents are added to the collection; unstable class names and relationships are in information retrieval systems user-unfriendly, especially when used for subject browsing.

Text categorization (machine learning) is the most widespread approach to automated classification of text. Here characteristics of subject classes, into which documents are to be classified, are learnt from documents with human-assigned classes. However, human-classified documents are often unavailable in many subject areas, for different document types or for different user groups. If one would judge by the standard Reuters Corpus Volume 1 collection (Lewis et al. 2004), some 8,000 training and testing documents would be needed per class. A related problem is that text categorization algorithms perform well on new documents only if they are similar enough to the training documents. The issue of document collections was pointed out by Yang (1999) who showed how similar versions of one and the same document collection had a strong impact on performance.

Traditionally, research in text categorization seems to be focused on improving algorithm performance, and experiments are conducted under laboratory-like conditions. Studies in which web pages were categorized into hierarchical structures for browsing generally do not involve well-developed classification schemes, but

home-grown structures such as search engines directories that are not well structured and/or maintained. In addition, often only few categories with one or two hierarchical levels are used in experiments, each consequently containing an "unbrowsable" number of documents.

In document classification, matching is conducted between a controlled vocabulary and text of documents to be classified. A major advantage of this approach is that it does not require training documents, while still maintaining a pre-defined structure. If using a well-developed classification scheme, it will also be suitable for subject browsing in information retrieval systems. This would be less the case with automatically created classes and structures of document clustering or home-grown directories not created in compliance with professional principles and standards. Apart from improved information retrieval, another motivation to apply controlled vocabularies in automated classification is to reuse the intellectual effort that has gone into creating such a controlled vocabulary (see also Svenonius 1997).

2.3 Evaluation challenge

According to ISO standard on methods for examining documents, determining their subjects, and selecting index terms (International Organization for Standardization 1985), human-based subject indexing is a process involving three steps: 1) determining subject content of a document, 2) conceptual analysis to decide which aspects of the content should be represented, and 3) translation of those concepts or aspects into a controlled vocabulary. These steps, in particular the second one, are based on a specific library's policy in respect to its document collections and user groups. Thus, when evaluating automatically assigned classes against the human-assigned ones, it is important to know the collection indexing policies.

Another problem to consider when evaluating automated classification is the fact that certain subjects are erroneously assigned. When indexing, people make errors such as those related to exhaustivity policy (too many or too few subjects become assigned), specificity of indexing (which usually means that the assigned subject is not the most specific one available), they may

omit important subjects, or assign an obviously incorrect subject (Lancaster 2003, 86-87).

For document collections used in the thesis it was not possible to obtain indexing policies. Because of this and erroneous classes, without a thorough qualitative analysis of automatically assigned classes one cannot be sure whether, for example, the classes assigned by the algorithm, but not human-assigned, are actually wrong, or if they were left out by mistake or because of the indexing policy.

In addition, it has been reported that different people, whether users or professional subject indexers, would assign different subjects to the same document. Studies on inter- and intra-indexer consistency report generally low indexer consistency (Olson and Boll 2001, 99-101). Markey (1984) reviewed 57 indexer consistency studies and reported that consistency levels ranged from 4% to 84%, with only 18 studies showing over 50% consistency. There are two main factors that seem to affect it:

1) Higher exhaustivity and specificity of subject indexing both lead to lower consistency, i.e., indexers choose the same first term or class notation for the major subject of the document, but the consistency decreases as they choose more subjects;
2) The bigger the vocabulary, or, the more choices the indexers have, the less likely they will choose the same terms or class notations (Olson and Boll 2001, 99-101).

Both factors were present in experiments of the thesis.

Today evaluation in automated classification experiments is mostly conducted under controlled conditions, ignoring the above-discussed issues. As Sebastiani (2002, 32) puts it, "…the evaluation of document classifiers is typically conducted experimentally, rather than analytically. The reason is that… we would need a formal specification of the problem that the system is trying to solve (e.g., with respect to what correctness and completeness are defined), and the central notion… that of membership of a document in a category is, due to its subjective character, inherently nonformalizable."

Because methodology for such experiments has yet to be developed, as well as limited resources, in all but one studies of the thesis the common approach to evaluation was followed, i.e., the

assumption was that human-assigned classes in document collections were correct, and automatically assigned classes were compared against them.

2.4 Hierarchical subject browsing

While it has been reported that users prefer searching to browsing (Lazonder 2003; Nielsen 1997), browsing has been claimed to have a number of advantages. It is an intuitive activity which is cognitively easier than searching and helps clarify an information problem (Large et al. 1999, 192). It is especially useful when users are not looking for a specific information resource, when they lack experience in performing searching, and when they are not familiar with the subject or its structure and terminology (Koch and Zettergren 1999).

Examples of web-based services offering subject browsing include quality-controlled subject gateways such as Intute (Intute Consortium 2006a), or those provided by commercial search engines such as Yahoo! Directory (Yahoo! 2007) or Google Directory (Google 2007). However, subject browsing generally does not seem to be very well supported in information services on the World Wide Web (in further text: Web). For example, in his study on browsing strategies and implications for design of web search engines, Xie (1999) reports that existing browsing features of search engines are insufficient to users. One of the possible reasons for this underdevelopment could be that people to a large extent believe that browsing is less useful. Even within the Renardus project, an initial belief about potential user requirements was that end-users preferred searching to browsing (Tuominen et al. 2000). After the browsing interface was built, it was shown that browsing was much favoured (**I**). Large et al. (1999, 180) claim that today we acknowledge that users can often express their information need only in very general terms and that the need can be met only by incorporating both browsing and searching capabilities in information retrieval systems.

Controlled vocabularies (classification schemes, thesauri, subject heading systems) have traditionally been used in libraries, and in indexing and abstracting services, some since the 19th century. With the Web, new versions of vocabularies emerged within the computer science and the Semantic Web communities:

ontologies and search-engine directories of web pages. All these vocabularies have distinct characteristics and are consequently better suited for some applications than others. For example, subject heading systems normally do not have detailed hierarchies of terms, while classification schemes consist of hierarchically structured groups of classes. In classification schemes similar documents are grouped together into classes and relationships between the classes are established; thus they are better suited for subject browsing than other controlled vocabularies (Vizine-Goetz 1996; Koch and Zettergren 1999). This is partly confirmed by the fact that they have been used by several web-based services, especially those providing information resources for academic users, such as BUBL Information Service (2005), INFOMINE (2007) etc.

Different classification schemes have different characteristics; for subject browsing the following are important: the bigger the collection, the more depth should the hierarchy contain; hierarchically flat schemes are not effective for browsing; classes should contain more than just one or two documents (Schwartz 2001, 48). Search-engine directories and other home-grown schemes on the Web, "…even those with well-developed terminological policies such as Yahoo… suffer from a lack of understanding of principles of classification design and development. The larger the collection grows, the more confusing and overwhelming a poorly designed hierarchy becomes…" (76). For these reasons it was decided to use the following two classification schemes: Dewey Decimal Classification (OCLC Dewey Services 2007), which has been used (and updated) in libraries for more than a century now; and, the Engineering Information classification scheme (Milstead 1995), which has been used and maintained in the Compendex database (Engineering Information 2007).

3 Methodology

3.1 User studies and performance measures

3.1.1 User studies

In order to study browsing behaviour based on classification schemes, two methods were applied.

First, log analysis of a large web-based service providing integrated searching and browsing access to quality-controlled web resources classified into Dewey Decimal Classification (**I**). Log analysis was chosen because users do not need to be directly involved in the study, user behaviour is captured in natural conditions, and every activity inside the service is tracked. This study encompassed 16 months of usage. Own software for log analysis has been developed, since existing software packages did not support all the needed tasks.

Second, a user study was conducted in which evaluation of browsing was combined with evaluation of classification correctness (**VIII**). The study was based on web pages which had been automatically crawled and classified into the Engineering Information classification scheme. It involved 40 engineering students or experts who were, given four tasks, supposed to find the most appropriate class in the classification scheme and evaluate whether the top ranked web pages were on the topic of their task. The data were collected through questionnaires, logging users' browsing steps, correctness assessments, and, in six cases, observation.

3.1.2 Performance measures

Performance of automated classification was evaluated using two main methods.

First, comparison to human-assigned classes, employing standard evaluation measures, precision, recall and F1 (Sebastiani 2002, 40-41) (**IV**, **V**, **VI**, **VII**):

Precision = correctly automatically assigned classes / all automatically assigned classes

Recall = correctly automatically assigned classes / all human-assigned classes

F1 = 2*Precision*Recall / (Precision + Recall)

The Ei classification scheme has a solid hierarchical structure, thus allowing for a rather credible test on partial overlap. The topical relatedness of classes is expressed in numbers representing the classes (class notation): the more initial digits any two classes have

in common, the more related they are. For example, 933.1.2 for *Crystal Growth* is closely related to 933.1 for *Crystalline Solids*, both of which belong to 933 for *Solid State Physics*, and finally to 93 for *Engineering Physics*. Each digit represents one hierarchical level: class 933.1.2 is at the fifth hierarchical level, 933.1 at the fourth etc. Thus, comparing two classes at only first few digits (later referred to as partial matching) instead of all the five also makes sense.

In addition, the average number of classes assigned to each document was studied. In the context of hierarchical browsing based on a classification scheme, having too many classes assigned to a document would place one document in too many different places, which would create the opposite effect of the original purpose of a classification scheme, that of grouping similar documents together. Several other factors, such as the number of documents that are classified, whether main concept is discovered were also taken into consideration (**VI**).

Second, a user study was conducted in which evaluation of browsing was combined with evaluation of classification correctness, as described above (**VIII**).

3.2 Document collections

Browsing based on classification schemes was studied on two collections of web pages. One was human-classified using Dewey Decimal Classification; it contained over 80,000 web pages (**I**). The other comprised about 19,000 web pages that were automatically crawled and classified into the Engineering Information classification scheme (**VIII**).

The easiest way of studying how to achieve improvements in automated classification is through a collection of documents that were human-classified. The collection used in **IV** contained web pages which were human-classified into classes of the Engineering Information classification scheme. Since it comprised only about 1,000 documents and bigger collections of web pages did not exist, further experiments for improvements of the classification algorithm were conducted on a collection of some 35,000 paper bibliographic records (also comprising abstracts) from the Compendex database, also human-classified into the Engineering Information

classification scheme (**V**, **VI**). Performance comparison of the string-matching and machine-learning algorithms was conducted on a similar set from Compendex, comprising about 24,000 bibliographic records (**VII**).

3.3 Engineering Information thesaurus and classification scheme

Engineering Information thesaurus and classification scheme (in further text: Ei controlled vocabulary), in the field of engineering, consists of two parts: a thesaurus of engineering terms, and a hierarchical classification scheme of engineering topics (Milstead 1995). In information retrieval systems, these two controlled vocabulary types have each traditionally had distinct functions: the thesaurus has been used to describe a document with as many controlled terms as possible, while the classification scheme has been used to group similar documents together to the purpose of allowing systematic browsing.

The Ei classification scheme is hierarchical and consists of six main classes divided into 38 finer classes which are further subdivided into 182 classes. These are subdivided even further, resulting in some 800 individual classes in a five-level hierarchy. In **IV**, most of these, around 750, were automatically assigned.

In studies **V**, **VI**, and **VIII**, 92 classes were used. They all belonged to class 900, *Engineering, General*. The reason for choosing this group of classes was that it covers both natural sciences such as physics and mathematics, and social sciences fields such as engineering profession and management. The literature of the latter tends to contain more polysemic words than the former, and as such presents a more complex challenge for automated classification. For study **VII**, six classes were selected, the ones which had most documents in the document collection.

A major advantage of the Ei controlled vocabulary for automated classification is that thesaurus descriptors are mapped to classes of the classification scheme. These mappings have been created manually (intellectually) and are an integral part of the thesaurus. Compared with captions[1] alone, mapped thesaurus terms provide a rich additional vocabulary for every class: instead of

[1] A caption is a class notation expressed in words, e.g., in the Ei classification scheme “Electric and Electronic Instruments” is the caption for class “942.1”.

having only one term per class (there is only one caption per class), there are on average 88 terms per class (for the 92 classes used in **V**, **VI**, and **VIII**). In addition, Ei comprises a large portion of composite terms (3,474 in the total of 4,411 distinct terms for the 92 classes); as such, it provides a rich and precise vocabulary with the potential to reduce the risks of false hits.

3.4 The string-matching classification algorithm

This section describes the string-matching classification algorithm used in **III, IV, V, VI, VII** and **VIII**. The algorithm classifies documents into classes of the Ei classification scheme, in order to provide a browsing interface to the document collection. It searches for terms from the Ei controlled vocabulary, in text of documents to be classified.

Based on thesaurus terms and captions, a term list was created that served as an input to the algorithm. The list was formed as an array of triplets (for a formal description, see Ardö 2007):

Weight: **Term** (single word, Boolean term or phrase) = **Class**

It contained class captions, thesaurus terms (Term)[2], classes to which the terms and captions map or denote (Class), and weight indicating how appropriate the term is for the class to which it maps or which it designates (Weight). *Single-word terms* were terms consisting of one word. *Boolean terms* were phrases consisting of two or more words which were "translated" into Boolean expressions; thus, all the words had to be present but in any order or in any distance from each other.

The algorithm searches for strings from a given term list in the document to be classified and if the **Term** is found, the **Class(es)** assigned to that string in the term list are assigned to the document. One class can be designated by many terms, and each time a term is found, the corresponding **Weight** is added to the score for the class. The scores for each class are summed up and classes with scores above a certain cut-off value (heuristically defined) are selected as the final ones for the document being classified.

[2] Geographical names in the thesaurus part, all mapping to class 950, were excluded on the grounds that they were not engineering-specific.

4 Results

4.1 Usage of web-based browsing (I, VIII)

With the purpose of determining whether hierarchical web-based browsing is being used and, if so, how (research question **1**), a study of a large web-based service was conducted (**I**). The service chosen was Renardus[3], which offered integrated searching and browsing access to about 80,000 quality web pages from major European subject gateways. Both browsing and searching options were elaborately developed. The main navigation feature was browsing based on a well-established classification scheme, Dewey Decimal Classification (DDC) (OCLC Dewey Services 2007). Browsing-support features were also provided: the graphical fish-eye display, search entry to browsing pages, and merging resources' descriptions from all related collections. Simple and advanced searching were available as well, the latter allowing several combinations of search terms and search fields, and options to limit searches in different ways.

In contrast to some other research reporting that searching is used more than browsing (Lazonder 2003; Nielsen 1997) and, perhaps, common belief that that is the case, this study clearly indicated that browsing as an information-seeking activity is highly used, given proper conditions. About 80% of all activities in Renardus were browsing activities; only 5% were the searching ones. The DDC directory-style browsing was the single clearly dominating activity in Renardus (60%). Two-thirds of it was done in unbroken sequences, some of them surprisingly long; while the majority limited themselves to 10 such steps, long unbroken sequences of up to 86 steps were found. The browsing support features, especially graphical fish-eye display and search entry to browsing pages, were also heavily used: they made up 13% of all activities.

One factor contributing to the dominance of browsing was that the majority of users (71%) were referred from search engines directly to browsing pages in Renardus. In addition, layout of the

[3] Renardus was developed and maintained until 2002. Parts of it can be still seen at http://renardus.it.lth.se/.

home page "invited" browsing, so also users starting at the home page (22%) predominantly used the browsing part of the service.

Users coming to the service at one of its browsing pages showed different navigation behaviour than users starting at the homepage. The latter performed almost twice as many activities per session, used searching pages five times as often, and visited other pages three times as often. They were a minority but used the service elaborately, in the way system designers had imagined and intended.

Transitions between different types of activities were rare; if they took place, it was mostly between different browsing activities. Switching between browsing and searching occurred in 7% of the sessions, far less than expected and hoped for.

In the last study (**VIII)** it was also indicated that browsing was well accepted. Most suggestions to improve browsing that arose from **VIII** had been already implemented in Renardus, in the form of browsing-support features. In conclusion, the study lead to the hypothesis that browsing and its support features are perceived popular and useful in services like Renardus.

4.2 Suitability of DDC and Ei for hierarchical web-based browsing (I, VIII)

In order to determine whether established classification schemes are suitable for hierarchical web-based browsing (research question **2**), two different classification schemes were examined: a general-subject one, DDC (**I**), and a subject-specific one, Engineering Information (Ei) (Milstead 1995) (**VIII**).

Log analysis of DDC usage (**I**) provided several insights into the suitability of its structure and vocabulary. The findings from the log analysis can, however, only help create hypotheses and need to be complemented by investigative sessions with users. As reported above, DDC browsing was the most dominant activity in Renardus, which is one indication of its suitability. Analysis of a sample of 100 search queries submitted to a search engine showed that most matched terms in DDC captions. Most successful queries used two to three query terms. When analysing browsing and jumps between different parts of the DDC directory, they occurred in less than half of the sessions with unbroken DDC-directory browsing. In those sessions on average less than two jumps were carried out, which is not too high. Also, the overall mean probability of moving between

DDC main classes in a session is small (3%). These results imply that DDC is suitable for browsing.

The **VIII** study indicated that the Ei classification scheme is also generally well suited for browsing. The majority of participants found the right class, they reported that it was quite easy finding it and were quite certain they found the right class. Still, participants' comments indicated some inadequacy of the classification scheme. The need for several improvements of classification schemes was implied:

- Follow consistent division principles when building classification structures;
- Modify captions so that they better reflect concepts they represent; and,
- Allow for a larger entry vocabulary, which would directly help both finding the ideal class fast and improve recall in automated classification.

4.3 Improving the string-matching algorithm (III, IV, V, VI)

While the first two research questions were related to browsing, the five remaining ones are on automated classification. Different approaches to automated classification (research question **3**) were discussed in 2.2. The string-matching algorithm (questions **4, 5, 6,** and **7**) is dealt with from this section onward.

With the purpose of determining problems and improvements of the string-matching classification algorithm (research questions **4** and **5**), four studies were conducted (**III, IV, V,** and **VI**).

4.3.1 Challenges and recommendations (III)

In order to identify challenges of applying a string-matching classification algorithm on a collection of web pages (research question **4**), an analysis of 70 misclassified ones was conducted (**III**). Four major types of problems were identified:

1) Class not found at all;
2) Class found but below a pre-defined cut-off value;
3) Wrong automatically assigned class; and,
4) Correct automatically assigned class which was not human-assigned.

The following reasons behind these were recognized and ways to deal with them proposed:

1) Classes were not found when the term list was missing the right terms designating the classes. In certain cases this was due only to a simple form variation or different ordering of a term's constituent elements.

The latter problem had been in previous experiments dealt with stemming, but at the expense of decreased precision. A partial solution could be to manually introduce regular expressions[4] to the term list. An automated solution would be to apply computational linguistics methods to create new variations of the existing terms. Cases in which term forms are entirely different from the ones found in text could be dealt with by enriching the term list with synonyms for concepts already covered by the thesaurus as well as new ones.

In a later study (**VI**), automated multi-word morpho-syntactic analysis and synonym acquisition were used to introduce new variations of terms and synonyms for existing concepts. In (**V**), the term list was enriched with more term types from the Ei thesaurus than there were in the original term list.

2) Certain classes found by the algorithm were not assigned as final ones because their scores did not reach the cut-off value (see last paragraph in 3.4). Different weighting schemes and cut-offs should be experimented with. These were experimented with in later studies (**IV**, **VI**).

3) Wrong automatically assigned classes seemed to have been caused by three different problems. First, terms found on web pages were homonyms or distant synonyms to concepts designated by the same terms on the term list. Second, terms from the term list found on web pages represented an *instance* of the concept designated by the term and was not *about* such an instance. Third, mappings between a thesaurus term and a class were too distant.

Suggestions proposed included the following:

[4] A regular expression uses certain syntax rules to match a set of strings; a wildcard character is an example.

- Adding context to single and ambiguous terms by, e.g., enriching them with their broader terms;
- Introducing synonyms;
- Creating a stop-list of homonyms always yielding incorrect classes; and
- Classifying hyperlinked web pages and comparing each other's classes.

In **VI,** significant improvements were achieved by excluding those terms from the term list that had been previously shown to mainly find wrong classes, due to the three problems described above.

4) For different reasons, certain classes that were assigned automatically should or could have been human-assigned. Due to such omissions, automatically assigned classes in research studies should be also evaluated for accuracy by subject experts. Such a user study was conducted and reported in **VIII**.

4.3.2 HTML structural elements and metadata (IV)

The aim of the **IV** study was to determine the importance of distinguishing between different parts of a web page in automated classification. The hypothesis was that best results are achieved when different weights (significance indicators) are assigned to classes, based on where the terms designating the classes are found on a web page. Four web-page elements were studied: title, headings, internal metadata, and main text. The document collection consisted of some 1,000 web pages in engineering, to which Ei classes have been human-assigned.

Potential weights were obtained using several different methods: precision and recall based on both total and partial overlap, semantic distance and multiple regression. The derived weights were tested against the baseline, in which all the four elements were given equal weight. The results showed that looking for Ei terms in all the four elements was necessary since not all of them occur on every web page. However, the exact way of combining the weights for terms found in those elements, turned out not to be highly important: the best combination of weights was 3% better than the baseline. In this

formula big significance is given to classes that were assigned based on the title: the score of one class is the sum of score of that class found based on title and multiplied by 86, the score from metadata multiplied by 6, the score from headings multiplied by 5, and the score from main text.

4.3.3 Improvements achieved on paper abstracts (V, VI)

The **V** study explored to what degree different types of terms in the Ei thesaurus and classification scheme influence automated classification performance. Preferred terms, their synonyms, broader, narrower, related terms, and captions were examined in combination with a stemmer and a stop-word list. The document collection comprised some 35,000 scientific paper abstracts from the Compendex database. A subset of the Ei thesaurus and classification scheme was used, containing 92 classes from the area of General Engineering.

The results showed that preferred terms perform best, and captions worst. Stemming in most cases improved performance, while the stop-word list did not have a significant impact. The majority of classes were found when using all types of terms, and stemming: recall was 73%. The remaining 27% of classes were not found because terms designating the classes on the term list did not exist in the text of documents being classified. The sole number of terms designating a class did not seem to be related to classification performance for that class.

The study implied that all types of terms should be included on a term list in order to achieve best recall. Higher weights could be given to preferred terms, captions and synonyms, as they yield highest precision.

In **V** neither weights nor cut-offs were experimented with; instead, all the classes that were found for a document were assigned to it. In the context of hierarchical browsing based on a classification scheme, having too many classes assigned to a document would place one document to many different places, which would create the opposite effect of the original purpose of a classification scheme, that of grouping similar documents together.

The aim of the **VI** study was to further improve the classification algorithm, especially to:

1) Achieve precision levels similar to recall gained in **V,** by applying different weights and cut-offs; and
2) Increase levels of recall to more than those achieved in **V,** by natural language processing methods.

In order to systematically vary different parameters, 14 weighting schemes evolved, combining weights for different term and class types, Ei term types, number of words a term contains and number of times each of the words occur in other terms. A stop-word list and stemming were also tested.

Also, the effect of several different cut-offs was investigated: 1) the score of classes to be selected as final had to have a minimum percentage of the sum of all the classes' scores (different values for the minimum percentage were tested); 2) the rule that if there were no class with the required score, the one with the highest score would be assigned; and, 3) based on the indexing principle of specificity, score propagation was introduced, where scores for classes at deeper hierarchical levels were increased by scores of classes at upper hierarchical levels.

In order to further improve recall, the basic term list was enriched with new terms. These terms were extracted using multi-word morpho-syntactic analysis and synonym acquisition, based on the original preferred and synonymous terms, those two term types giving best precision. Extracted synonyms were verified by a subject expert.

The **VI** study showed that the string-matching algorithm could be enhanced in a number of ways, including the following:

1) Weights: adding different weights to the term list based on whether a term is single, phrase or Boolean, which type of class it maps to, and Ei term type, improves precision and relevance order of assigned classes, the latter being important for browsing;
2) Cut-offs: selecting as final classes those above a certain cut-off level improves precision and F1;
3) Enhancing the term list with new terms based on morpho-syntactic analysis and synonyms acquisition improves recall;
4) Excluding terms that in most cases gave wrong classes yields best performance in terms of F1, where the improvement is due to increased precision levels.

4.4 The string-matching algorithm on an abstracts collection (VI)

Finally, performance of the string-matching algorithm on an abstracts collection can be reported (research question **6**). At third and second hierarchical levels mean F1 reached up to 60%. For all hierarchical levels, best achieved mean F1 was 38%, when only those terms that found classes correct in the majority of cases were included on the term list. In this case 65% of documents were classified. Best recall was 76%, when the basic term list was enriched with new terms, and precision 99% when only those terms that only gave correct classes were included. When using the original term list, without any term exclusion, precision of individual classes was up to 98%.

These results are comparable to machine-learning algorithms (see Sebastiani 2002), which are considered to be the best ones but require training documents and are collection-dependent. Another benefit of classifying documents into classes of well-developed classification schemes is that they are suitable for subject browsing, unlike automatically-developed controlled vocabularies or home-grown directories often used in document clustering and text categorization (see section 2.2).

The study also showed that different versions of the algorithm could be implemented so that it best suits the application of the automatically classified document collection. If high recall were required, such as, for example, in focused crawling, cut-offs would not be used. Or, if providing directory-style browsing interface to a collection of automatically classified web pages, the pages could be ranked by relevance based on scores. In such a directory, one would want to limit the number of web pages per class, e.g., assign only the class with highest probability that it is correct, as it is done in the Thunderstone's web site catalog (Thunderstone 2007). Since for 14 classes at top three hierarchical levels mean F1 is almost twice as good as for the complete matching, this classification approach would suit better those information systems in which fewer hierarchical levels are needed, like the Intute subject gateway on engineering (Intute Consortium 2006b).

4.4.1 Comparison to a machine-learning algorithm (VII)

In this exploratory study the string-matching algorithm was compared to a machine-learning one, the latter being support vector machine (SVM). Document collection consisted of a subset of about 24,000 Compendex paper abstracts, classified into six different classes, two of them from class 900.

SVM on average outperformed the string-matching algorithm. The first hypothesis, that SVM would yield better recall and string-matching better precision, was confirmed only on one of the classes. The second hypothesis was that classification performance could be improved by confederating the two algorithms. Terms (features) used by one algorithm were combined with the other algorithm's terms in five different ways. The results showed that SVM performed best in its original setting, while recall and F1 of string-matching improved when using the SVM terms.

Since this study had been conducted before further improvements were introduced to the string-matching algorithm (**VI**), performance based on those improvements could be reported only now. It was shown that performance for two classes from class 900, when using the full term list with 8,099 terms, was better than when only preferred terms, synonyms and captions were used (**V**). When using a shortened term list containing 1,308 terms that on a similar document collection always yielded correct classes, with the 5% cut-off, precision for both classes was 1.0. These results are better than SVM in any setting. In the same setting, recall is, however, less than 0.1. When using the same shortened list with 1,308 terms, applying stemming and no cut-offs, best recall achieved for class 903.3 is 0.61. This recall is the same as in the first experiment (**VII**), but in the second (**VI**) precision is higher, so F1 is higher too (0.46). Values of F1 are still lower than when the term list was enriched with *tf-idf* centroid terms produced as part of the SVM algorithm.

While SVM used in the study outperforms the string-matching approach in recall and F1, it should be remembered that when it comes to real-life information systems such as digital libraries, pre-classified document collections (especially of web pages) are rarely available. String-matching algorithms could in such cases be the feasible solution.

4.5 String-matching algorithm on a web-page collection (VIII)

In order to determine performance of the string-matching algorithm on a collection of harvested web pages as evaluated by end-users (research question **7**), the **VIII** study was conducted. It involved 40 engineering subject experts and 4 tasks, with some 19,000 web pages automatically crawled and classified into the Ei classification scheme. Each task was to find information on a given topic by browsing the Ei classification scheme. Once the most appropriate class was reached, top ranked web pages were to be evaluated if they were on the topic of the task.

Top ranked web pages in each of the four classes were on average deemed partly correct. A major problem with determining whether a web page is in the right class or not is that there were large differences among participants in their judgements – a number of web pages were evaluated as correct, partly correct and incorrect by different participants. A major reason is probably the issue of "aboutness" and related subjectivity in deciding which topic a document is dealing with.

As with browsing, evaluations between the four tasks differed. This is in compliance with previous results of the algorithm's performance, based on a pre-classified collection of paper abstracts, where it was shown that certain classes have better performance than others (**VI**). Worst results in both studies were gained for class 901.1.1 (Societies and institutions), which can be attributed to the fact that only one term exists for this class on the term list. Also, most terms designating the other three classes are rather field-specific and less ambiguous than the one term designating class 901.1.1 (*societies @and institutions*).

5 Concluding remarks

In the thesis it was shown that hierarchical web-based browsing was much used and that well-developed classification schemes such as DDC and Ei were suitable for the task. In the context of browsing, three main approaches to automated classification were recognized. In order to provide good browsing structures as results of their complex and much-researched automated classification algorithms, text categorization and document clustering approaches would need to employ suitable controlled vocabularies. Improvements of the

string-matching approach were achieved in several ways and evaluation implied that results are comparable to those of state-of-the-art machine-learning algorithms could be achieved, especially for certain classes and applications.

While subject browsing was shown to be useful in a large web-based service, it needs to be further investigated to determine to what degree it is suitable for different users' tasks. Some controlled vocabularies are being adjusted for new roles in the online environment and adjustments have been proposed in the literature and indicate in the thesis; still, more research is needed on what controlled vocabularies need to be like to support browsing, as well as on their suitability for automated classification. Also, while Ei proved to be suitable to a certain degree, which characteristics of controlled vocabularies are beneficial for automated classification needs to be further studied.

Owing to recognized evaluation issues, it is difficult to estimate to what degree automated classification tools of today are applicable in operative information systems. The subjectivity of deriving the correct interpretation of a document's subject matter, much discussed in the literature, has been indicated (**VI**) and shown (**VIII**) in the thesis. Evaluation results depend on several factors, such as document collection, application context, or user tasks. It is believed that evaluation methodology of automated classification where all the different factors would be included, perhaps through a triangulation of standard collection-based evaluation and user studies, should be a major further research question.

References

Ardö, A. (2007), "Crawler internal operation", available at: http://combine.it.lth.se/documentation/DocMain/node6.html (accessed 5 September 2007).

BUBL Information Service (2005), Centre for Digital Library Research, Strathclyde University, Glasgow, available at: http://bubl.ac.uk/ (accessed 31 August 2007).

Chan, L.M. (1994), *Cataloging and classification: an introduction*, 2nd ed., McGraw-Hill, New York.

Chen, H., and Dumais, S.T. (2000), "Bringing order to the Web: automatically categorizing search results", *Proceedings of the ACM International Conference on Human Factors in Computing Systems, Den Haag*, pp. 145-152.

Engineering Information (2007), "Compendex", Engineering Information, Elsevier, available at: http://www.ei.org/databases/compendex.html (accessed 30 August 2007).

Google (2007), *Google Directory*, Google, available at: http://directory.google.com/ (accessed 30 August 2007).

INFOMINE: scholarly Internet resource collections (2007), Library of the University of California, available at: http://infomine.ucr.edu/ (accessed 30 August 2007).

International Organization for Standardization (1985), *Documentation – Methods for examining documents, determining their subjects, and selecting index terms*: *ISO 5963,* Geneva, International Organization for Standardization.

Intute Consortium (2006a), *Intute*, available at: http://www.intute.ac.uk/ (accessed 30 August 2007).

Intute Consortium (2006b), *Intute: Science, engineering and technology – engineering*, available at: http://www.intute.ac.uk/sciences/engineering/ (accessed 30 August 2007).

Koch, T., and Zettergren, A.-S. (1999) "Provide browsing in subject gateways using classification schemes", *EU Project DESIRE II*, available at http://www.mpdl.mpg.de/staff/tkoch/publ/class.html.

Lancaster, F.W. (2003), *Indexing and abstracting in theory and practice,* 3rd ed, Facet, London.

Large, A., Tedd, L., and Hartley, R. (1999), *Information seeking in the online age,* K. G. Saur, London etc.

Lazonder, A.W. (2003), "Principles for designing web searching instruction", *Education and Information Technologies* 8 (June 2003), pp. 179-193.

Lewis, D.D., Yang, Y., Rose, T., and Li, F. (2004), "RCV1: a new benchmark collection for text categorization research", *The Journal of Machine Learning Research,* 5, pp. 361-397.

Markey, K. (1984), "Interindexer consistency tests: a literature review and report of a test of consistency in indexing visual materials", *Library & Information Science Research,* 6, pp. 155-77.

Milstead, J, ed. (1995), *Ei thesaurus,* 2nd ed., Engineering Information Inc., Hoboken, NJ.

Moens, M.-F. (2000), *Automatic indexing and abstracting of document texts,* Kluwer, Boston.

Nielsen, J. (1997), "Search and you may find", *Jakob Nielsen's Alertbox,* July 15, available at: http://www.useit.com/alertbox/9707b.html (accessed 30 August 2007).

OCLC Dewey Services (2007), "Dewey Decimal Classification System", available at: http://www.oclc.org/dewey/about/default.htm (accessed 31 August 2007).

Olson, H.A., and Boll, J.J. (2001), *Subject analysis in online catalogs*, 2nd ed., Libraries Unlimited, Englewood, CO.

Schwartz, C. (2001), *Sorting out the Web: approaches to subject access,* Ablex, Westport, CT.

Sebastiani, F. (2002), "Machine learning in automated text categorization", *ACM Computing Surveys*, Vol. 34 No. 1, pp. 1-47.

Svenonius, E. (1997), "Definitional approaches in the design of classification and thesauri and their implications for retrieval and for automatic classification", *Proceedings of the Sixth International Study Conference on Classification Research,* pp. 12-16.

Svenonius, E. (2000), *The intellectual foundations of information organization,* MIT Press, Cambridge, MA.

Thunderstone (2007), "About the Thunderstone web site catalog", available at: http://search.thunderstone.com/texis/websearch/about.html, (accessed 31 August 2007).

Tuominen, K., Kanner, J., Miettinen, M., and Heery, R. (2000), "User requirements for the broker system: Renardus project deliverable D1.2", available at: http://renardus.it.lth.se/about_us/project_ deliverables.html (accessed 31 August 2007).

Vizine-Goetz, D. (1996), "Using library classification schemes for Internet resources", *OCLC Internet Cataloging Project Colloquium,* available at: http://webdoc.sub.gwdg.de/ebook/aw/oclc/man/colloq/v-g.htm (accessed 31 August 2007).

Xie, H. (1999), "Web browsing: current and desired capabilities", *20th Annual National Online Meeting, 18-20 May, New York, US,* pp. 523-37.

Yahoo! (2007), *Yahoo! Directory,* available at: http://dir.yahoo.com/, (accessed 30 August 2007).

Yang, Y. (1999), "An evaluation of statistical approaches to text categorization", *Journal of Information Retrieval*, Vol. 1 No. 1/2, pp. 67-88.

Paper I

PAPER I.
Users Browsing Behaviour in a DDC-Based Web Service: A Log Analysis

Abstract

This study explores the navigation behaviour of all users of a large web service, Renardus, using web log analysis. Renardus provides integrated searching and browsing access to quality-controlled web resources from major individual subject gateway services. The main navigation feature is subject browsing through the Dewey Decimal Classification (DDC) based on mapping of classes of resources from the distributed gateways to the DDC structure. Among the more surprising results are the hugely dominant share of browsing activities, the good use of browsing support features like the graphical fish-eye overviews, rather long and varied navigation sequences, as well as extensive hierarchical directory-style browsing through the large DDC system.

Koch, T., Golub, K. and Ardö, A, (2006), "Users browsing behaviour in a DDC-based web service: a log analysis", *Cataloging & Classification Quarterly,* Vol. 42 No. 3/4, pp. 163-186.

1 Introduction

As many research communities are increasingly concerned with issues of interaction design, one of the current foci in information science is on user behaviour in seeking information on the World Wide Web. A frequently applied methodology for studying this behaviour is log analysis. This approach has several advantages: users do not need to be directly involved in the study, a picture of user behaviour is captured in non-invasive conditions, and every activity inside the system can be tracked.

User log studies mainly use the average analytical approaches of existing software packages for statistical reporting. Such software provides limited knowledge of user behaviour (Hochheiser and Shneiderman 2001), since it only produces comparatively general insights into aspects of information services, such as number of users per month or the mostly followed hyperlink, and thus tells little about specific navigation behaviour.

A variety of aspects of user information-seeking behaviour using log analysis have been studied previously, in digital libraries (Jones et al. 2000), web search engines (Silverstein et al. 1999; Ozmutlu et al. 2004; Beitzel et al. 2004), and other web-based information services. Browsing behaviour has not been studied that much.

The common belief seems to be that users prefer searching to browsing: Lazonder (2003, 181) claims "…students strongly prefer searching to browsing". Nielsen (1997) states the following: "Our usability studies show that more than half of all users are search-dominant, about a fifth of the users are link-dominant, and the rest exhibit mixed behaviour. The search-dominant users will usually go straight for the search button when they enter a website: they are not interested in looking around the site; they are task-focused and want to find specific information as fast as possible. In contrast, the link-dominant users prefer to follow the links around a site: even when they want to find specific information, they will initially try to get to it by following promising links from the home page. Only when they become hopelessly lost will link-dominant users admit defeat and use a search command. Mixed-behaviour users switch between search and link-following, depending on what seems most promising to them at any given time but do not have an inherent preference".

These observations have implications for building searching-oriented user interfaces. However, those results could be dependent on a number of issues that might have not yet been recognized. One such issue is, for example, the role of the web page layout in "favouring" either of the two strategies. Xie (1999) conducted a study on browsing strategies and implications for design of web search engines. The study reports that existing browsing features of search engines are insufficient to users. Even within the Renardus project, an initial belief about potential user requirements was that end-users preferred searching to browsing (Tuominen 2000, 23). After the browsing interface had been built, it showed that browsing was much favoured.

The overall purpose of our project was to gain insights into real users' navigation and especially browsing behaviour in a large service on the Web. This knowledge could be used to improve such services, in our case the Renardus service (Renardus 2002a) which offers a large DDC browsing structure. Renardus is a distributed web-based service which provides integrated searching and browsing access to quality controlled web resources from major individual subject gateway services across Europe. (The Renardus project was funded by the European Union's Information Society Technologies 5th Framework Programme until 2002).

The research aimed at studying the following topics: the unsupervised usage behaviour of all Renardus users, complementing the initial Renardus user enquiry; detailed usage patterns (quantitative/qualitative, paths through the system); the balance between browsing, searching and mixed activities; typical sequences of user activities and transition probabilities in a session, especially in traversing the hierarchical DDC browsing structure; the degree of usage of the browsing support features; and typical entry points, referring sites, points of failure and exit points. Because of the high cost of full usability lab studies, we also wanted to explore whether a thorough log analysis could provide valuable insights and working hypotheses as the basis for good usage and usability studies at a reasonable cost.

The paper provides short background information about Renardus (2 Background); the methodology applied in this study is described in section three (3 Methodology); the analysis, hypotheses and results regarding the general usage of Renardus, the browsing behaviour and the usage of the DDC are presented in the fourth

section (4 Results). A summary of the results and some ideas for further investigation conclude the paper (5 Conclusion).

2 Background

Renardus (Renardus 2002a) exploits the success of subject gateways, where subject experts select quality resources for their users, usually within the academic and research communities. This approach has been shown to provide a high quality and valued service, but encounters problems with the ever-increasing number of resources available on the Internet. Renardus is based on a distributed model where major subject gateway services across Europe can be searched and browsed together through a single interface provided by the Renardus broker. The Renardus partner gateways cover over 80,000 predominantly digital, web-based resources from within most areas of academic interest, mainly written in English.

The Renardus service allows searching several subject gateways simultaneously. What is searched are "catalogue records" (metadata) of quality controlled web resources, not the actual resources. There are two ways to search the service, either through a simple search box that is available on the Renardus "Home page" or through the "Advanced search" page allowing combination of terms and search fields and providing options to limit searches in a number of different ways. A pop-up window of a list of words alphabetically close to the entered word (for title, DDC, subject and document type) supports the search term selection.

Apart from searching, Renardus offers subject browsing in a hierarchical directory-style (see, for example, Technology | Agriculture | Animal husbandry (Renardus 2002c)). It is based on intellectual mapping of classification systems used by the distributed gateway services to the DDC. There are also several browsing-support features. The graphical fish-eye display presents the classification hierarchy as an overview of all available categories that surround the category one started from, normally one level above and two levels below in the hierarchy. This allows users to speed up the browsing and get an immediate overview of the relevant Renardus browsing pages for a subject. The feature "Search entry into the browsing pages" offers a shortcut to categories in the browsing tree where the search term occurs. The lower half of the

browsing pages, as a result of the classification mapping, offers the links to the "Related Collections" of the chosen subject. In case users do not want to jump to the parts of the gateways offering related collections, an option of Merging the resource-descriptions from all related collections is available.

For a more detailed description of Renardus, see, for example, Koch et al. (2003). All related publications are given at the web page "Renardus project archive and associated research and development" (Renardus 2002b).

3 Methodology

Before Renardus was finally released and the EU project concluded in 2002, an end user evaluation of the Renardus pilot subject gateway (Alhainen et al. 2002) was carried out during fall 2001 which led to some service improvements. The results and shortcomings of this initial user study stimulated us to try the full study of Renardus user logs which is presented in this paper.

Log analysis was chosen because it costs considerably less than full usability lab studies and has the advantage that it is an unobtrusive means of capturing unsupervised usage. This thorough log analysis required several steps which are described below: cleaning of the log files, defining of user sessions, categorization into activity types and the creation of datasets and structures to allow the creation of statistics and the testing of hypotheses.

3.1 Cleaning the log files

The log files used spanned 16 months between summer 2002 and late fall 2003. They first had to be cleaned from entries created by search engine robots, crackers (users performing unauthorized activities), local administration, images, etc. The largest group of removed entries, almost half of all log entries, was that containing images and style sheets (1,107,378). Further, 516,269 entries were removed because they originated from more than 650 identified robots, and an additional 12,647 entries because they were from crackers. Various other entries not relating to real usage of Renardus for information seeking, e.g., 17,586 redirections, about 9,000 local administrative activities, error codes and HTTP head entries, had to be removed.

Thus, in the first step, the total number of 2,299,642 log entries was reduced to 631,711 entries. From this dataset, only some general Renardus usage statistics was derived. For the analysis of real user behaviour in Renardus, several further steps and separate datasets were required.

3.2 Defining sessions

After cleaning the log, all entries were grouped into user sessions. A session was heuristically defined as containing all entries coming from the same IP-address and a time gap of less than one hour from the prior entry from the same IP-address.

3.3 Defining activity types

Each log entry was classified into one of eleven different main activities offered by Renardus. These activities were then used to characterize user behaviour, via a typology of usages and sequences of activities.

Browsing activities:

- "Gen. Browse", hierarchical directory-style browsing of the DDC (e.g., see Renardus 2002c);
- "Graph. Browse", graphical fisheye presentation of the classification hierarchy (e.g., see Graphical browsing page: Renardus 2002d);
- "Text Browse", text version of the graphical fisheye presentation;
- "Search Browse", search entry into the browsing structure;
- "Merge Browse", merging of results from individual subject gateways;
- "Browse", DDC top level browsing page on the home page.

Searching activities:

- "Simple Search" with "showsimpsearch" for result display;
- "Adv. Search", advanced search with "showadvsearch" for result display and "scan" for scanning certain data indices.

Other activities:

- "Home Page"; "Help"; "Other" other informational pages, including project documentation.

3.4 Creating datasets for studying information-seeking behaviour

To try to make sure that we studied only human behaviour in Renardus, we removed, in a further step, another 82,490 entries judged as probable machine activities. This determination was based on heuristic criteria, for example, all sessions containing only one entry; sessions shorter than two seconds.

Most of the analysis in this paper regarding human activities in Renardus is based on a dataset containing 464,757 entries grouped into 73,434 user sessions. Only in a few calculations (especially in the section "Browsing sessions") did we use a further subset of this dataset. The different datasets were stored in a relational database and SQL was used to query them to create statistical tables and to test various hypotheses against the log file data.

4 Results

4.1 Global usage

Renardus was accessed from 99,605 unique machines (IP-numbers) during the 16 month period studied. With 351 unique top-level domains or countries identified (a considerable part of the IP-numbers could not be identified), it is apparent that Renardus has a truly global audience. IP-numbers from the USA topped the list with about 30%, other .net and .com domains followed with 8-10%. Renardus Project partner countries were led by Finland with 5%. Canada, Australia, the Philippines, Italy and India were other countries exceeding 1% of the IP-numbers.

The user sessions are of considerable length: 33% are longer than 2 minutes and 10% are longer than 10 minutes. The time users might have been exploring participating gateways after leaving Renardus is not included.

The figures indicate that more than 851 different hosts referred users to Renardus. As much as 56% of all referred sessions came from various Google servers and 24% from Yahoo!

Renardus seemed to be able to attract and keep many "faithful" users during the first 16 months after release. Thirteen percent of all unique user machines were returning to the service, which is a comparatively good value.

4.2 Information-seeking activities

4.2.1 Main activities, transitions

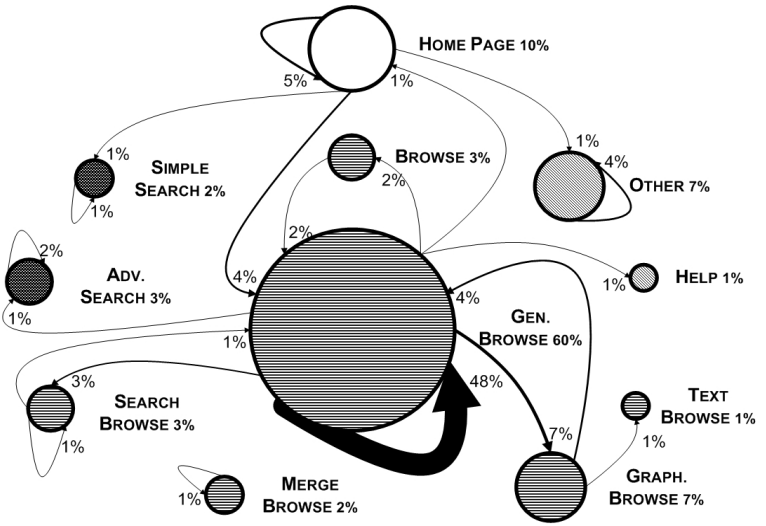


Figure 1. Main Renardus features, indicating their share in all activities, and major transitions between the activities.

Figure 1 illustrates the share of each activity and transition in the following ways: the share of each of the main activities is indicated by the circle size; and the share of the major transitions between different activities is indicated by arrow size. Only values above 1% are displayed. It shows that 60% of all Renardus activities are directory-style browsing using the DDC structure (Gen. Browse; for the abbreviations used here and throughout the paper, see the description under 3.3 Defining activity types). Forty-eight percent of all transitions in Renardus are steps from one such topical page/DDC class to another.

The four special browsing support features are comparatively well used. As many as 45% of the sessions dominated by browsing use two or more different types of browsing activities. As many as 14% use three to five different types (see Browsing sessions below).

Use of the graphical DDC browsing overview (Graph. Browse) is the second most frequent activity in Renardus (7%), after the directory-style browsing. The transition from the dominant directory browsing in the DDC structure to a graphical display is clearly the largest single transition in Renardus, after subsequent directory browsing steps.

Related to Gen. Browse, in 11% of the cases, directory-style browsing has been followed by the usage of the graphical overview (see Figure 2). For further reasoning about these findings see below.

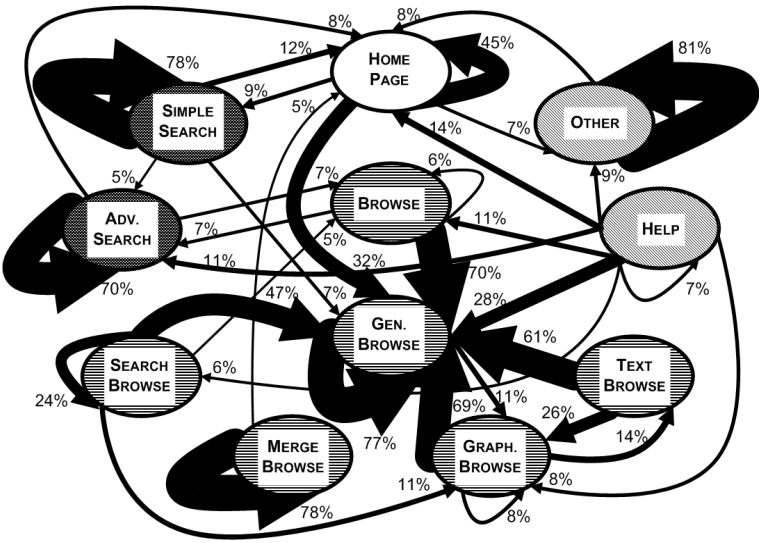


Figure 2. Transition probabilities (more than 5%-transitions only).

Figure 2 illustrates another important finding. Users tend to stay in the same feature and group of activities, whether it is a single activity like Gen. Browse or a group like browsing, searching or looking for background information, despite the provision of a full navigation bar on each page of the Renardus service. In particular, the transitions between browsing and searching activities are less frequent than expected and hoped for. Figure 2 demonstrates this by

displaying the main transitions from each feature to other features of the service (the percentages displayed close to the arrows relate to the feature they originate from). For example, 77% of all transitions from one Gen. Browse activity are directed to another Gen. Browse activity and 11% to Graph. Browse.

As the early user study in 2001 showed (Alhainen et al. 2002, table 18), the Renardus pilot service was mostly considered very easy or easy to navigate already, although a fifth of the respondents found navigating through the different parts of the service difficult or very difficult. We conclude that advanced online services need to provide some kind of search strategy support. They need to be designed for receiving the user where he/she first enters the system and to assist with user navigation through the whole system with more than a ubiquitous navigation bar (which is offered by Renardus on all pages).

4.2.2 General navigation sequences

Many users engage in several different activities during their session: about 46% in one activity, 20% in two, 16% in three different activities. About 18% of the user sessions have between 4 and 11 different activities.

Many users employ a surprisingly rich variety of navigation and browsing sequences and often alternate between many different features. For example, one session has the following sequence (the first number indicates the number of immediate repetitions of the same feature, the second gives the length of this activity in seconds).

> home 3 3; genbrowse 4 31; graphbrowse 1 1; genbrowse 3 17; home 1 1 ; browse 1 1; genbrowse 2 3; searchbrowse 1 1; genbrowse 7 152; searchbrowse 1 1; genbrowse 4 24; help 1 1; home 1 1; genbrowse 4 29; graphbrowse 1 1; searchbrowse 1 1; genbrowse 1 1; searchbrowse 1 1; genbrowse 1 1; browse 1 1; genbrowse 3 2; graphbrowse 1 1; genbrowse 2 2; home 2 2

When we look at the most frequent sequences of activity types (immediate repetition of the same type is not counted), we find 4,810 different sequences. The top ten sequences are presented in Table 1. The most frequent sequences, apart from mergebrowse and showsimpsearch, are (in and) between browsing activities.

Table 1. Most frequent sequences of activity types.

type of activity	sessions	%
(repetitions of) genbrowse	30,606	41.70%
other	7,403	10.10%
genbrowse-graphbrowse-genbrowse	3,860	5.30%
genbrowse-graphbrowse	3,590	4.90%
genbrowse-searchbrowse	2,812	3.30%
(repetitions of) mergebrowse	2,391	3.30%
(repetitions of) showsimpsearch	1,705	2.30%
genbrowse-browse-genbrowse	1,635	2.20%
genbrowse-searchbrowse-genbrowse	1,236	1.70%
genbrowse-browse	1,035	1.40%
all less frequent sequences		23.80%

When we look at a more detailed table of sequences including immediate repetitions of the same activity (not reproduced here), the dominance of browsing and the very high number of variations in navigation is well illustrated. In 73,434 user sessions we find as many as 16,377 different sequences; however, the top 10 most frequent sequences (with more than 1,000 instances each) cover 41.7% of all sessions. In the top 6, and numbers 9-11 among the 11 most frequent sequences, the user exclusively repeats the same activity. Only numbers 7 and 8 involve a switch between different activities (from genbrowse to graphbrowse and from genbrowse to searchbrowse). In the five most frequent cases genbrowse is the repeated activity. The sequences where only the same activity type is repeated cover about 50% of all sessions. This further underlines our earlier finding that a surprisingly large part of the users stay in the same (group of) activities.

4.2.3 Browsing vs. searching

The levels of usage of the main Renardus features are highly uneven (see Figure 1). The most surprising finding is the clear dominance of browsing activities, about 80%. Depending how "dominance of browsing" is defined: 76% of all activities are browsing; 80.5% of all sessions are dominated by browsing. Searching has a much lower share, between 3 and 6%.

This is a highly unusual ratio compared to other published evaluations and common beliefs (see Introduction). A possible reason is that most of the browsing pages are indexed by search

engines. Seventy-one percent of the users reached browsing pages directly via search engines and start their Renardus navigation at a browsing page. Taken together with the clear tendency to stay in the same (group of) features, these facts "favour" browsing. Additionally, the layout of the home page invites browsing by putting the browsing structure on top of the search box. Still, among users starting at the home page, 57% browse and only 12.5% search (only 22% of all users enter Renardus at the home page/the "front door" of the service, however).

In spite of the dominance of browsing and the tendency to stay in the same group of activities, we see a certain amount of switching between browsing and searching during the same session. In as few as 7.3 % of all sessions users switch between a browse and a search activity, out of which 4.5% of sessions have one switch, 1.9% have two, 0.4% have three, and 0.5% have more than three switches.

The largest number of switches per session is 20. Out of 27 different kinds of switches between browsing and searching, 7 start with a search. Switching from browsing to searching is much more frequent than the opposite. Users at the search pages need to be pointed to the benefits of browsing.

4.2.4 Browsing sessions

For the calculations in this section we use a subset of our usual dataset, containing 378,267 entries in 58,954 user sessions, defined by a share of more than 50% browsing activities: sessions where "browsing is dominant".

The shares of sessions with a certain number of different activities are almost the same as for all Renardus sessions (see the beginning of General navigation sequences). So, even sessions with dominant browsing show as much variety in activities as most other sessions.

Many browsing sessions use more than one type of browsing activity, including the browsing support features: Graph. Browse, Text Browse and Merge Browse. As many as 45% of the sessions dominated by browsing show two or more browsing activities and 14% three to five different types of browsing. We find up to 95 individual browse activities per session, with gracefully degrading numbers from two activities and down.

4.2.5 Two different groups of users

Because of the big influence of referrers like search engines, 71% of the human user sessions start at browsing pages pointed to by referrers, whereas 22% start at the homepage (16,300 out of 73,434 sessions). This quantitatively surprising result stimulated us to check if these two "groups" of users show significantly different navigation behaviour. Sessions starting at home have almost twice as many entries per session than sessions starting elsewhere (10 vs. 5.8 entries per session; 35.8% of all entries). Thus, home starters carry out many more activities per session than the other user group.

Users jumping into the middle of the Renardus service are carrying out browsing activities in 87% of all cases and only 2.7% searching activities (Table 2).

Table 2. Types of activities for the two different groups of users.

type of activity	starting at home		starting elsewhere	
	entries	%	entries	%
browsing	94,215	56.6	259,471	87
searching	20,831	12.5	8,099	2.7
other	51,139	30.9	30,684	10.3
total	166,503		298,254	

Users starting at the Renardus Home page/"frontdoor" show a level of browsing of almost 57%, and 12.5 % searching. Three times as often they visit other pages and five times as often search pages compared to the other group. These are probably the users who go deliberately to Renardus, whereas a large part of users starting elsewhere, most often in the browsing pages, end up there "ignorantly" after a search in a search engine. The latter overwhelmingly stay in the browsing activities.

People starting elsewhere have a much higher percentage of browsing among their activities. Home starters, however, do considerably more browsing activities compared to their share of all sessions: 53.2% of the sessions show more than 11 browsing activities and 36.8% more than 30 browsing activities.

Figure 3 shows that the home starters clearly dominate the sessions with many browsing activities. A more detailed analysis shows that they are active in browsing activities to a higher and increasing degree starting with 8 browsing activities, compared with

their share in all sessions (22%). Quite the opposite is true for users starting elsewhere. They are overrepresented up to the level of nine browsing activities with an ever-decreasing tendency.

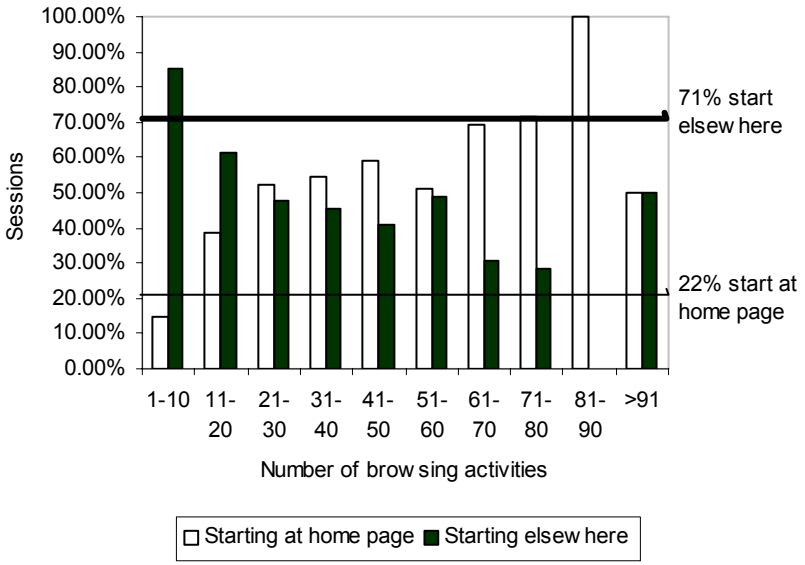


Figure 3. Browsing activities of the two groups of users.

Home starters also exceed their share when it comes to the number of different activity types, all types are counted (in browsing sessions) except when there are three different activities. From five different activities and higher, they have more than twice their share and dominate clearly.

When it comes to the number of different browsing types (in browsing sessions), home starters exceed their share when carrying out between three and five different browsing types.

4.3 DDC usage

4.3.1 DDC analysis

Analysis of the popularity of DDC sections and classes and the navigation behaviour of users in the DDC structure allow good insights into the distribution of topical interests and the suitability of

the DDC system and vocabulary. The findings from the log analysis can, however, only help create hypotheses and need to be complemented by investigative sessions with the users. The most frequently used parts of the DDC hierarchy at the top hierarchical level are given in Table 3.

Table 3. Most frequently used parts of the DDC hierarchy at the top hierarchical level.

entries	DDC	class
50,784	3	Social sciences
46,209	5	Science
30,955	6	Technology
26,015	2	Religion
22,081	7	Arts & recreation
17,994	8	Literature
16,828	9	History & geography
16,527	0	Computers, information & general reference
13,839	4	Language
13,428	1	Philosophy & psychology

All DDC classes show generally good usage levels (users jumping to one class and not continuing browsing are not counted). Compared to what one would expect in a global internet setting, Religion ranks surprisingly high and Computers etc. unexpectedly low (see Table 3). Here the vocabulary used in the DDC captions could play a role, e.g., many computing-related terms used in Internet searching do not directly occur in the captions.

On the second hierarchical level, surprisingly large topical areas are Christian denominations (DDC 28), German & related literatures (83), Social problems (36) and Earth Sciences (55; see Figure 4).

Unexpectedly frequent visits to individual topics like 552.1 Igneous rocks (the sixth most visited individual page with 2,436 directory browsing activities) could be due to the fact that little information might be found about such a concept in the search engines or to the fact that other sites made prominent links to this topic page in Renardus.

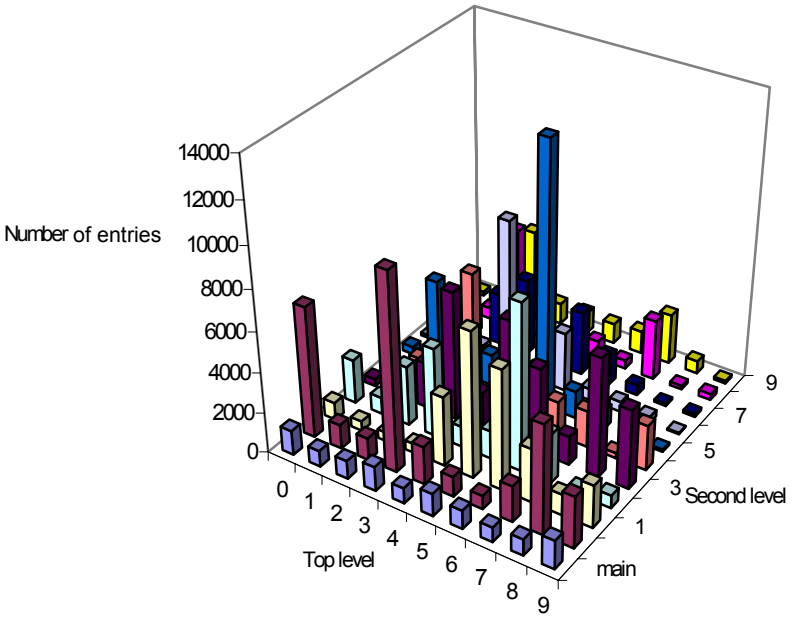


Figure 4. Most frequently used parts of the DDC hierarchy at the second hierarchical level.

4.3.2 Directory style of browsing in the DDC hierarchy

The directory-style of browsing in the DDC-based browsing structure is clearly the dominant activity in Renardus (about 60%). Sixty-seven percent of all browsing activities are DDC directory browsing (254,660 out of 378,264 entries in browsing sessions). Two-thirds of the latter (167,628) appear in unbroken sequences. In these cases, not even browse support features are used between directory browsing steps. While the clear majority of users limit themselves to 10 or fewer steps (for distribution see the Figure 5), we found surprisingly long unbroken browsing sequences of up to 86 steps in the DDC directory trees.

These are very unexpected results. People looking for information on the Web are often said to use as few clicks as necessary, switching frequently to other services and activities, having a very short attention spans. Browsing the DDC hierarchies in a directory style of steps at such quantity and lengths is one of the most significant results of this log study.

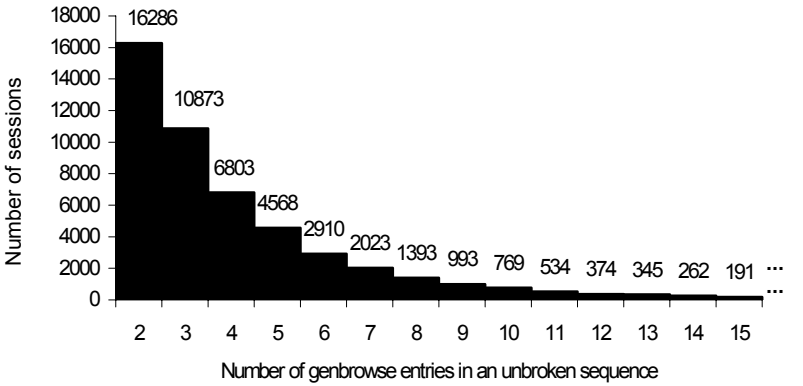


Figure 5. Number of genbrowse activities in sessions (up to 15).

4.3.3 Jumping in the DDC hierarchy

Since the DDC browsing area in the Renardus user interface displays the higher levels in the hierarchy, in addition to the "parent" and the "child" classes, we can find out to what degree users jump levels in the DDC hierarchy during unbroken directory browsing sequences.

Two of the support features, the graphical overviews and the "search entry to browsing pages", were designed to relieve users from the "pain" of having to jump around in the hierarchy. Jumping one step up and another step down in the directory-style display is probably faster and easier than using the support features; moving farther away would possibly have been easier using the support features.

The following sequence is an example of a session featuring jumps within unbroken directory browsing:

> start 62-; go to 624; go to 624.1; jump to 62-; go to 625; go to 625.1; go to 625; go to 62-; go to 627; jump to 628; go to 628.1

In sessions featuring unbroken directory browsing, 20.2% of all steps are jumps. Jumps occur in 40.8% of these sessions. In the sessions with jumps, on average 1.7 jumps are carried out. This is a decent number of cases but not excessively high. Many users make use of the support features, especially the graphical overviews,

instead of jumping in the directory. This finding indicates, at least, that the necessity to jump in the hierarchy is not putting off users.

As seen from Figure 6, the probability for a user in one session to browse in several main DDC classes increases with the length of the session. This might seem natural but it also implies that the longer the session, the shorter time spent within one main DDC class before moving to another. Each point in the figure is based on several sessions that together contain more than 2,000 browsing entries. Due to the heavy dominance of shorter sessions, the overall mean probability of moving between DDC main classes in a session is 3%.

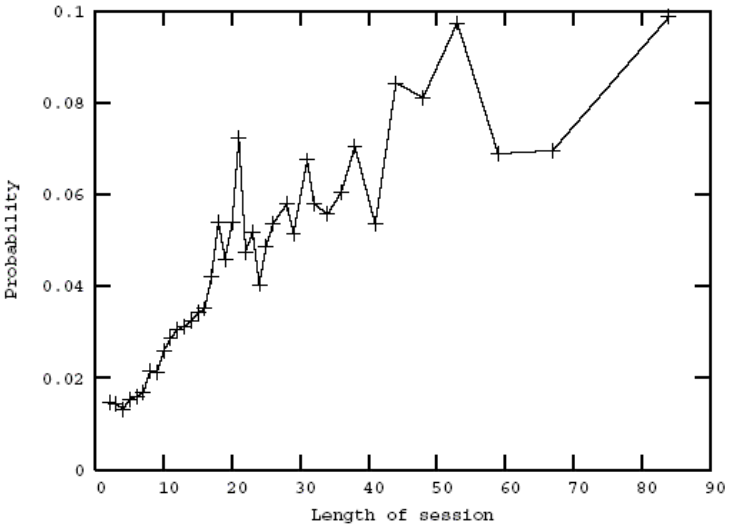


Figure 6. Probability of moving between DDC main classes.

Figure 7 shows a few individual sessions plotted with the number of browsing steps versus the visited DDC classes. For example, all classes within the "1—" branch of DDC are displayed between 100 and 200 on the vertical axis in such a way that the hierarchy is preserved, e.g., the closer two classes are in the hierarchy the closer they are plotted in the figure. Thus a horizontal line indicates that the user stays within a narrow area of DDC while vertical parts indicate jumps between different areas of DDC. The letter "G" indicates that the graphical overview was used while an

"S" indicates that the search entry to the browsing structure was used at the indicated points in the sequence.

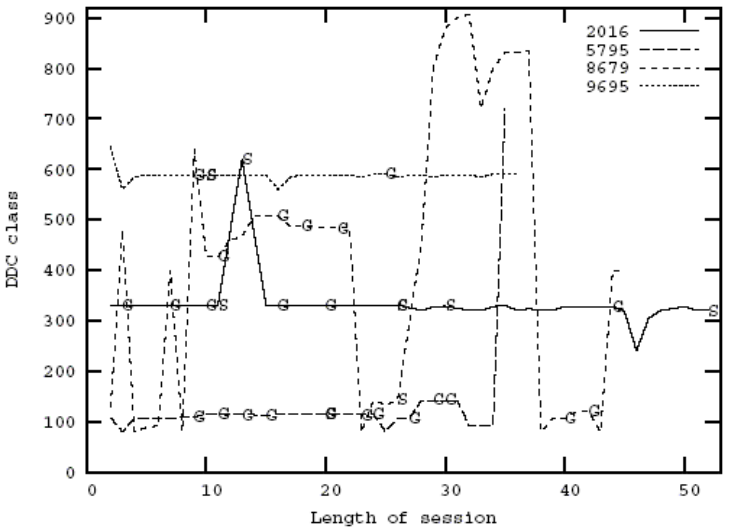


Figure 7. DDC browsing behaviour per session.

4.3.4 Keywords and browsing

We wanted to find out whether the user managed to come close to his/her topic of interest when browsing DDC pages in Renardus. In order to obtain an indication of that, we compared the keywords entered by a given user into the search engine (Google) respectively entered into Renardus Search with the browsing pages visited subsequently by the same user.

The following list of examples shows keywords entered into Google and the => Renardus DDC class the user selected from the search result:

ancient continents => History of ancient world; of specific continents, countries, localities; of extraterrestrial worlds
perspective drawing => Drawing & decorative arts
"statistics of south america" => General statistics of specific continents, countries, localities in modern world
writing systems and etymology => Standard language--description and analysis
kinds of sedimentary rocks => Specific kinds of rocks

The sample studied showed very good hits in the Renardus DDC pages. Most queries matched terms in the DDC caption (which is also used as the title of the page), about 13% of the cases had partial hits there and partial matches in other class and directory "titles" mentioned on the same page (parent, child DDC classes; names of mapped directories from cooperating subject gateways). Most successful questions used 2-3 query terms; only 3% used one term.

It seems that a good initial hit is required to invite users to continue browsing in Renardus (the data here is derived from sessions containing more than one browsing activity).

The result says more about the search engine's ranking algorithm (Google in the case of our sample) than about the Renardus pages and the suitability of the DDC captions. Part of the reason for the share of the good hits we see is the fact that only DDC pages from Renardus where the user's search terms hits several words in the title or top half of the DDC page in Renardus have a chance to appear in the top of Google search results. Only such pages have a chance to be selected and visited by the user.

When checking queries and hits in Other Renardus pages (background and project information), we found great results, too: most hits seemed relevant and we could not find many wrong hits on topical questions.

The following examples show search terms entered into Renardus Search (Q:) and DDC classes/pages used during browsing. Each case was gathered from the entire session and was limited to sessions starting with Search and continuing with Browse activities. In the examples, queries and DDC captions are separated by a semicolon.

> Q=chopin; vieuxtemps;
> DDC=Arts & recreation; Music; Composers and traditions of music;

> Q=paperin+valmistus; paperin+valmistus; papermaking; paper+technology;
> DDC=Technology; Engineering; Engineering of railroads and roads; Engineering of railroads and roads; Engineering of railroads and roads; Railroads; Railroads; Astronautical engineering; Technology; Engineering; Engineering and allied operations; Engineering mechanics and materials; Science; Chemistry; Chemistry; Organic chemistry; Technology; Chemical engineering; Chemical engineering and related technologies; Biotechnology; Biotechnology; Pulp and paper technology; Genetic engineering; Electrical engineering; lighting; superconductivity; magnetic engineering; applied optics; paraphotic technology; electronics; communications engineering; computers; Electrical engineering; lighting; superconductivity; magnetic

engineering; applied optics; paraphotic technology; electronics; communications engineering; computers; Electronics; Special topics; Optoelectronics; Pulp and paper technology; Conversion of pulp into paper, and specific types of paper and paper products; General topics; Properties, tests, quality controls;

The results of the evaluations of our sample remind us that users frequently follow more than one topic of interest during one session in an information system. In our sample 70% of all users seemed to pursue one topic in a session, 23% two topics, 2% three topics and 5% seemed to browse around without specific question. In some cases, topics looked for in Renardus Search are not pursued when browsing, in other cases, a new topic (most often one) is investigated after the switch to browsing.

5 Conclusion

The main purpose of this study was to explore the navigation behaviour of all users of a large web service, Renardus, using web log analysis, in order to improve the user interface and, especially, the browsing features of the system. In addition, we aimed at gaining some more general insights into users browsing and navigation in large classification structures, the benefits from system support and the problems and failures that occurred.

Our study indicates that a thorough log analysis can indeed provide a deeper understanding of user behaviour and service performance. Being an unobtrusive means of capturing unsupervised usage and offering a complete and detailed picture of user activities, log analysis can reveal quantitatively comprehensive, sometimes unexpected results, far beyond plain statistics.

In contrast to common beliefs, our study clearly indicates that browsing as an information-seeking activity is highly used, given proper conditions. About 80% of all activities in Renardus are browsing activities. A contributing reason to that dominance is the fact that a very high percentage (71%) of the users are referred from search engines or other linking sites directly to a browsing page in Renardus. The layout of the home page “invites” browsing, which certainly contributes to the fact that even users starting at the home page predominantly use the browsing part of the service.

Our study leads to a hypothesis which deserves further research: browsing is perceived as useful and dominates navigation in services similar to Renardus and under proper conditions.

The good use of the browsing support features, especially graphical overview and search entry to browsing pages, suggests that it would be worthwhile to further develop such support.

Since most visitors jump into the middle of the service, there might be a need to redesign the browsing pages so they would better serve as full-fledged starting points for comprehensive Renardus exploration. The ubiquitous navigation bar seems not sufficiently inviting. In making such changes, it would also be important to better understand the details of site indexing and ranking algorithms in search engines.

The study of navigation sequences shows that users employ a rich variety of navigation and browsing sequences, including rather long and highly elaborate paths through the system. Nevertheless, quantitatively dominating is, to a quite surprising degree, the tendency to stay in the same group of activities or individual activity, whether browsing, searching or background information. This finding points us to the importance of providing "search strategy" support to the users at the page where their actions take place.

From the behaviour as documented in the log files we could identify two clearly different groups of users: people starting at the home page/frontdoor of the service (22%), and the majority of the users starting elsewhere. There are dramatic differences in their activity in the service. People starting at the home page show almost twice as many activities per session, and use the non-browsing features three to five times as often. Their share of the browsing activities is smaller, but they primarily engage in the long sequences of browsing activities (8 and longer) and employ more types of browsing activities and more types of other activities in a session. The home page starters are seemingly a minority but represent high quality of usage of the service in a way the system designers have imagined and intended.

The DDC directory browsing is the single clearly dominating activity in Renardus (60%). Two thirds of it is done in unbroken directory browsing sequences. We see a surprising average and total length of such browsing sequences, opposing the common belief of the short attention span of users of online services.

Thus, we acquire the surprising hypothesis that sequential, directory style of hierarchical (classification) browsing is found to

be popular and useful in large services like Renardus, especially when there is graphical support.

Comparisons between search terms used and topics browsed indicated a very good chance to obtain relevant results from Renardus browsing when more than one search term was used. People using Renardus Search were capable of finding browsing pages corresponding to their queries. The system invited users to pursue more than one topic during a session.

5.1 Future work

Our findings indicate that log analysis has a clear potential as a method for studying information behaviour and the proper design of information systems. A lot could be gained from future work to investigate questions such as:

- To what degree does the actual design of the system influence user behaviour, especially with regard to the difference in usage levels of browsing versus searching activities?
- Can we identify additional specific usage and browsing patterns and different behaviours of specific user groups?
- What is the influence of the use of end-user adapted and multilingual DDC captions on browsing behaviour?
- How can we provide search strategy support and further improve the support for systematic browsing of large subject structures?
- What is the importance of the details of site indexing in search engines for the discovery of and navigation in large browsing systems?
- How can pages be redesigned so that they better serve as full-fledged starting points?

For more important results and improvements one would need to go beyond the log analysis and:

- Evaluate user behaviour in supervised sessions/usability lab;
- Evaluate the accuracy and success of Renardus to help answering user questions;

- Use local URLs to identify what pages outside Renardus users explore as a result of Renardus navigation (links to participating subject gateways).

Acknowledgments

The Swedish Agency for Innovation Systems provided the main funding for this research. This work was partially funded by European Union (EU) under project ALVIS – Super-peer Semantic Search Engine (EU 6. FP, IST-1-002068-STP). This work was partially funded by DELOS – Network of Excellence on Digital Libraries (EU 6. FP IST, G038-507618).

References

Alhainen, T., Eerola, T., Heikkinen, R., and Lainkari, J. (2002), "User evaluation report: Renardus project deliverable D5.2", available at: http://renardus.it.lth.se/about_us//project_deliverables.html, (accessed 31 August 2007).

Beitzel, S.M., Jensen, E.C., Chowdhury, A., Grossman, D., and Frieder, O. (2004), "Hourly analysis of a very large topically categorized web query log", *Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval, Sheffield, United Kingdom*, pp. 321-328.

Hochheiser H., and Shneiderman, B. (2001), "Using interactive visualizations of WWW log data to characterize access patterns and inform site design", *Journal of the American Society for Information Science and Technology* Vol. 52 No. 4, pp. 331–343.

Jones, S., Cunningham, S.J., McNab, R. and Boddie, S. (2000), "A transaction log analysis of a digital library", *International Journal on Digital Libraries* Vol. 3 No. 2, pp. 152-169.

Koch, T., Neuroth, H., and Day, M. (2003), "Renardus: cross-browsing European subject gateways via a common classification system (DDC)", *Proceedings of the IFLA Satellite Meeting sponsored by the IFLA Section on Classification and Indexing and the IFLA Section on Information Technology, 14-16 August 2001, Dublin, OH, USA*, pp. 25-33.

Lazonder, A.W. (2003), "Principles for designing web searching instruction", *Education and Information Technologies* 8 (June 2003), pp. 179-193.

Nielsen, J. (1997), "Search and you may find", *Jakob Nielsen's Alertbox,* July 15, available at: http://www.useit.com/alertbox/9707b.html (accessed 30 August 2007).

Ozmutlu, S., Spink, A., and Ozmutlu, H.C. (2004), "A day in the life of web searching: an exploratory study", *Journal of Information Processing and Management* Vol. 40 No. 2, pp. 319-345.

Renardus (2002a), "Renardus home page" (2002), available at: http://renardus.it.lth.se/, (accessed 31 August 2007) [only a demonstrator].

Renardus (2002b), "Renardus project archive and associated research and development", available at: http://renardus.it.lth.se/about_us/project_archive.html (accessed 31 August 2007).

Renardus (2002c), "Technology: Agriculture: page", available at: http://renardus.it.lth.se/cgibin/genDDCbrowseSQL.pl?ID=10191&node=AAZNG (accessed 31 August 2007).

Renardus (2002d), "Technology ...: Mining for specific materials", available at: http://renardus.it.lth.se/cgi-bin/imageDDCbrowseSQL.pl?node=ABDPH&ID=10193&pmat=N&pnavnode=Y&pgraph=matcirc (accessed 31 August 2007).

Silverstein, C., Marais, H., Henzinger, M., and Moricz, M. (1999), "Analysis of a very large web search engine query log", *SIGIR Forum,* Vol. 33 No. 1, pp. 6-12, available at: http://doi.acm.org/10.1145/331403.331405 (accessed 31 August 2007).

Tuominen, K., Kanner, J., Miettinen, M., and Heery, R. (2000), "User requirements for the broker system: Renardus project deliverable D1.2", available at: http://renardus.it.lth.se/about_us/project_deliverables.html (accessed 31 August 2007).

Xie, H. (1999), "Web browsing: current and desired capabilities", *20th Annual National Online Meeting, 18-20 May, New York, US,* pp. 523-37.

Paper II

PAPER II.
Automated Subject Classification of Textual Web Documents

Abstract

The purpose of this literature review was to provide an integrated perspective to similarities and differences between approaches to automated classification in different research communities (machine learning, information retrieval and library science), and point to problems with the approaches and automated classification as such. Three main approaches were identified: text categorization, document clustering and a third one referred to as document classification. Major similarities and differences between the three approaches were identified: document pre-processing and utilization of Web-specific document characteristics is common to all the approaches; major differences are in applied algorithms, employment or not of the vector space model and of controlled vocabularies. Problems of automated classification are recognized.

Golub, K. (2006), "Automated subject classification of textual web documents", *Journal of Documentation,* Vol. 62 No. 3, pp. 350-371.

1 Introduction

Classification is, to the purpose of this paper, defined as "…the multistage process of deciding on a property or characteristic of interest, distinguishing things or objects that possess that property from those which lack it, and grouping things or objects that have the property or characteristic in common into a class. Other essential aspects of classification are establishing relationships among classes and making distinctions within classes to arrive at subclasses and finer divisions" (Chan 1994, 259). Automated subject classification (in further text: automated classification) denotes machine-based organization of related information objects into topically related groups. In this process human intellectual processes are replaced by, for example, statistical and computational linguistics techniques. In the literature on automated classification, the terms automatic and automated are both used. Here the term automated is chosen because it more directly implies that the process is machine-based.

Automated classification has been a challenging research issue for several decades now. Major motivation has been the high cost of manual classification. Interest has grown rapidly since 1997, when search engines could not do with just text retrieval techniques, because the number of available documents grew exponentially. Owing to the ever-increasing number of documents, there is a danger that recognized objectives of bibliographic systems (Svenonius 2000, 20-21) would be left behind; automated means could be a solution to preserve them (30). Automated classification of text finds its use in a wide variety of applications, such as: organizing documents into subject categories for topical browsing, including grouping search results by subject; topical harvesting; personalized routing of news articles; filtering of unwanted content for Internet browsers; and many others (Sebastiani 2002; and Jain et al. 1999).

In the narrower focus of this paper is automated classification of textual web documents into subject categories for browsing. Web documents have specific characteristics such as hyperlinks and anchors, metadata, and structural information, all of which could serve as complementary features to improve automated classification. On the other hand, they are rather heterogeneous; many of them contain little text, metadata provided are sparse and

can be misused, structural tags can also be misused, and titles can be general ("home page", "untitled document"). Browsing in this paper refers to seeking for documents via a hierarchical structure of subject classes into which the documents had been classified. Research has shown that people find browsing useful in a number of information-seeking situations, such as: when not looking for a specific item, when one is inexperienced in searching (Koch and Zettergren 1999), or unfamiliar with the subject in question and its terminology or structure (Schwartz 2001, 76).

In the literature, terms such as classification, categorization and clustering are used to represent different approaches. In their broadest sense these terms could be considered synonymous, which is probably one of the reasons why they are interchangeably used in the literature, even within the same research communities. For example, Hartigan (1996, 2) says: "The term cluster analysis is used most commonly to describe the work in this book, but I much prefer the term classification..." Or: "…classification or categorization is the task of assigning objects from a universe to two or more classes or categories" (Manning and Schütze 1999, 575).

In this paper terms text categorization and document clustering are chosen because they tend to be the prevalent terms in the literature of the corresponding communities. Document classification and mixed approach are used in order to consistently distinguish between the four approaches. Descriptions of the approaches are given below:

1) *Text categorization*. It is a machine-learning approach, in which also information retrieval methods are applied. It consists of three main parts: categorizing a number of documents to pre-defined categories, learning the characteristics of those documents, and categorizing new documents. In the machine-learning terminology, text categorization is known as supervised learning, since the process is "supervised" by learning categories' characteristics from manually categorized documents.
2) *Document clustering*. It is an information-retrieval approach. Unlike text categorization, it does not involve pre-defined categories or training documents and is thus called unsupervised. In this approach the clusters and, to a limited degree, relationships between clusters are derived automatically

from the documents to be clustered, and the documents are subsequently assigned to those clusters.

3) *Document classification.* In this paper it stands for a library science approach. It involves an intellectually created controlled vocabulary (such as classification schemes), into classes of which documents are classified. Controlled vocabularies have been developed and used in libraries and in indexing and abstracting services, some since the end of the 19th century.
4) *Mixed approach.* Sometimes methods from text categorization or document clustering are used together with controlled vocabularies. In the paper such an approach is referred to as a mixed approach.

To the author's knowledge no review paper on automated text classification attempted to discuss more than one community's approach. Individual approaches of text categorization, (document) clustering and document classification have been analysed by Sebastiani (2002), Jain et al. (1999), and Toth (2002), respectively.

This paper deals with all the approaches, from an integrated perspective. It is not aimed at detailed descriptions of approaches, since they are given in the above-mentioned reviews. Nor does it attempt to be comprehensive and all-inclusive. It aims to point to similarities or differences as well as problems with the existing approaches. In what aspects and to what degree are today's approaches to automated classification comparable? To what degree can the process of subject classification really be automated, with the tools available today? What are the remaining challenges? These are the questions touched upon in the paper.

The paper is laid out as follows: explorations of individual approaches as to their special features (description, differences, evaluation), application and employment of characteristics of web pages are given in the second section (approaches to automated classification), followed by a discussion (third section).

2 Approaches to automated classification

2.1 Text categorization

2.1.1 Special features

2.1.1.1 Description of features

Text categorization is a machine-learning approach, which has also adopted some features from information retrieval. The process of text categorization consists of three main parts:

1) The first part involves manual categorization of a number of documents to pre-defined categories. Each document is represented by a vector of terms. (The vector space model comes from information retrieval). These documents are called training documents because, based on those documents, characteristics of categories they belong to are learnt.
2) By learning the characteristics of training documents, for each category a program called classifier is constructed. After the classifiers have been created, and before automated categorization of new documents takes place, classifiers are tested with a set of test documents, which were not used in the first step.
3) The third part consists of applying the classifier to new documents.

In the literature, text categorization is known as supervised learning, since the process is "supervised" by learning from manually pre-categorized documents. As opposed to text categorization, clustering is known as an unsupervised approach, because it does not involve manually pre-clustered documents to learn from. Nonetheless, due to the fact that manual pre-categorization is rather expensive, semi-supervised approaches, which diminish the need for a large number of training documents, have also been implemented (Blum and Mitchell 1998; Liere and Tadepalli 1998; McCallum et al. 2000).

2.1.1.2 Differences within the approach

A major difference among text categorization approaches is in how classifiers are built. They can be based on Bayesian probabilistic learning, decision tree learning, artificial neural networks, genetic

algorithms or instance-based learning – for explanation of those, see, for example, Mitchell (1997). There have also been attempts of classifier committees (or metaclassifiers), in which results of a number of different classifiers are combined to decide on a category (e.g., Liere and Tadepalli 1998). One also needs to mention that not all algorithms used in text categorization are based on machine learning. For example, Rocchio (1971) is actually an information retrieval classifier, and WORD (Yang 1999) is a non-learning algorithm, invented to enable comparison of learning classifiers' categorization accuracy. Comparisons of learning algorithms can be found in Schütze et al. 1995, Li and Jain (1998), Yang (1999), or Sebastiani (2002).

Another difference within the text categorization approach is in the document pre-processing and indexing part, where documents are represented as vectors of term weights. Computing the term weights can be based on a variety of heuristic principles. Different terms can be extracted for vector representation (single words, phrases, stemmed words etc.), also based on different principles; characteristics of web documents, such as mark-up for emphasized terms and links to other documents, are often experimented with (Gövert et al. 1999). The number of terms per document needs to be reduced not only for indexing the document with most representative terms, but also for computing reasons. This is called dimensionality reduction of the term space. Dimensionality reduction methods could include removal of non-informative terms (not only stop words); also, taking only parts of the web document, its snippet or summary (Mladenic and Grobelnik 2003), has been explored. For an example of a complex document representation approach, a word clustering one, see Bekkerman et al. (2003); for another example, based on latent semantic analysis, see Cai and Hofmann (2003).

Several researches have explored how hierarchical structure of categories into which documents are to be categorized could influence the categorization performance. Koller and Sahami (1997) used a Bayesian classifier at each node of the classification hierarchy and employed a feature selection method to find a set of discriminating features (i.e., words) for each node. They showed that, in comparison to a flat approach, using hierarchical structure could improve classification performance. Similar improvements

were reported by McCallum et al. (1998), Dumais and Chen (2000), and Ruiz and Srinivasan (1999).

2.1.1.3 Evaluation methods

Various measures are used to evaluate different aspects of text categorization performance (Yang 1999). Effectiveness, the degree to which correct categorization decisions have been made, is often evaluated using performance measures from information retrieval, such as precision (correct positives/predicted positives) and recall (correct positives/actual positives). Efficiency can also be evaluated, in terms of computing time spent on different parts of the process. There are other evaluation measures, and new are being developed such as those that take into account degrees to which a document was wrongly categorized (Dumais et al. 2001; Sun et al. 2001). For more on evaluation measures in text categorization, see Sebastiani (2002, 32-39). Evaluation in text categorization normally does not involve subject experts or users.

Yang (1999) claims that the most serious problem in text categorization evaluations is the lack of standard document collections and shows how different versions of the same collection have a strong impact on the performance, and other versions do not. Some of the document collections used by the text categorization community are: Reuters-21578 (2004), which contains newswire stories classified under categories related to economics; OHSUMED (Hersh 1994), containing abstracts from medical journals categorized under Medical Subject Headings (MeSH); the U.S. Patent database in which patents are categorized into the U.S. Patent Classification System; 20 Newsgroups DataSet (1998), containing about 20,000 postings to 20 different Usenet newsgroups. For web documents there is WebKB (WebKB 2001), Cora (McCallum et al. 1999), and samples from directories of web documents such as Yahoo! (2005). All these collections have a different number of categories and hierarchical levels. There seems to be a tendency to conduct experiments on a relatively small number of categories with few hierarchical levels, which is usually not suitable for subject browsing tasks.

2.1.1.4 Characteristics of web pages

A number of issues related to categorization of textual web documents have been dealt with in the literature. Hypertext-specific

characteristics such as hyperlinks, HTML tags and metadata have all been explored.

Yang et al. (2002) defined five hypertext regularities of web document collections, which need to be recognized in order to chose an appropriate text categorization approach:

1) No hypertext regularity; in which case standard classifiers for text are used;
2) Encyclopaedia regularity, when documents with a certain category label only link to documents with the same category label, in which case the text of each document could be augmented with the text of its neighbours;
3) Co-referencing regularity, when neighbouring documents have a common topic; in which case the text of each document can be augmented with the text of its neighbours, but text from the neighbours should be marked (e.g., prefixed with a tag);
4) Preclassified regularity, when a single document contains hyperlinks to documents with the same topic, in which case it is sufficient to represent each page with names of the pages it links with; and,
5) Metadata regularity, when there are either external sources of metadata for the documents on the Web, in which case we extract the metadata and look for features that relate documents being categorized, or metadata are contained within the META, ALT and TITLE tags.

Several other papers discuss characteristics of document collections to be categorized. Chakrabarti et al. (1998b) showed that including documents that cite, or are cited by the document being categorized, as if they were local terms, performed worse than when those documents were not considered. They achieved improved results applying a more complex approach with refining the class distribution of the document being classified, in which both the local text of a document and the distribution of the estimated classes of other documents in its neighbourhood, were used. Slattery and Craven (2000) showed how discovering regularities, such as words occurring on target pages and on other pages related by hyperlinks, in both training and test document sets could improve categorization accuracy. Fisher and Everson (2003) found out that link information could be useful if the document collection had a sufficiently high

link density and links were of sufficiently high quality. They introduced a frequency-based method for selecting the most useful citations from a document collection.

Blum and Mitchell (1998) compared two approaches, one based on full-text, and the other based on anchor words, and found out that anchor words alone were slightly less powerful than the full-text alone, and that the combination of the two was best. Glover et al. (2002) reported that the text in citing documents close to the citation often has greater discriminative and descriptive power than the text in the target document. Similarly, Attardi et al. (1999) used information from the context where a URL that refers to that document appears and got encouraging results. Fürnkranz (1999) included words that occurred in nearby headings and in the same paragraph as anchor-text, which yielded better results than using the full-text alone. In his later study, Fürnkranz (2002) used portions of texts from all pages that point to the target page: the anchor text, the headings that structurally precede it, the text of the paragraph in which it occurs, and a set of linguistic phrases that capture syntactic role of the anchor text in this paragraph. Headings and anchor text seemed to be most useful. In regards to metadata, Ghani et al. (2001) reported that metadata could be very useful for improving classification accuracy.

2.1.1.5 Application

Text categorization is the most frequently used approach to automated classification. While a large portion of research is aimed at improving algorithm performance, it has been applied in operative information systems, such as Cora (McCallum et al. 2000), NorthernLight (Dumais et al. 2002, 69-70), and the Thunderstone web site catalog (Thunderstone 2005). However, detailed information about approaches used in commercial directories is mostly not available, due to their proprietary nature (Pierre 2001, 9). There are other examples of applying machine-learning techniques to web pages and categorizing them into browsable structures. Mladenic (1998) and Labrou and Finin (1999) used the Yahoo! directory (Yahoo! 2005). Pierre (2001) categorized web pages into industry categories, although he used only top-level categories of the North American Industrial Classification System.

Apart from organizing web pages into categories, text categorization has been applied for categorizing search engine

results (Chen and Dumais 2000; Sahami et al. 1998). It also finds its application in document filtering, word sense disambiguation, speech categorization, multimedia document categorization, language identification, text genre identification, and automated essay grading (Sebastiani 2002, 5).

2.1.1.6 Summary

Text categorization is a machine-learning approach, with the vector-space model and evaluation measures borrowed from information retrieval. Characteristics of pre-defined categories are learnt from manually categorized documents. Within text categorization, differences occur in several aspects: algorithms, methods applied to represent documents as vectors of term weights, evaluation measures and document collections used.

The potential added value of web document characteristics, which have been compared and experimented with, are, for example, anchor words, headings words, text near the URL for the target document, inclusion of linked document's text as being local. When deciding which methods to use, one needs to determine which characteristics are common to the documents to be categorized; for example, augmenting the document to be classified with the text of its neighbours will yield good results only if the source and the neighbours are related enough.

Text categorization is the most widespread approach to automated classification, with a lot of experiments being conducted under controlled conditions. There seems to be a tendency to use a small number of categories with few hierarchical levels, which is usually not suitable for subject browsing tasks. Several examples of its application in operative information systems exist.

2.2 Document clustering

2.2.1 Special features

2.2.1.1 Description of features

Document clustering is an information retrieval approach. As opposed to text categorization, it does not involve manually pre-categorized documents to learn from, and is thus known as an unsupervised approach.

The process of document clustering involves two main steps:

1) Documents to be clustered are represented by vectors, which are then compared to each other using similarity measures. Like in text categorization, different principles can be applied at this stage to derive vectors (which words or terms to use, how to extract them, which weights to assign based on what, etc.). Also, different similarity measures can be used, the most frequent one probably being the cosine measure.
2) In the following step, documents are grouped into clusters using clustering algorithms. Two different types of clusters can be constructed: partitional (or flat), and hierarchical.

Partitional algorithms determine all clusters at once. A usual example is *k*-means, in which first a *k* number of clusters is randomly generated; when new documents are assigned to the nearest centroid (centre of a cluster), centroids for clusters need to be re-computed.

In hierarchical clustering, a hierarchy of clusters is built. Often agglomerative algorithms are used: first, each document is viewed as an individual cluster; then, the algorithm finds the most similar pair of clusters and merges them. Similarity between documents can be calculated in a number of ways. For example, it can be defined as the maximum similarity between any two individuals, one from each of the two groups (single-linkage), as the minimum similarity (complete-linkage), or as the average similarity (group-average linkage).

For a review of different clustering algorithms, see Jain et al. (1999), Rasmussen (1992), and Fasulo (1999).

Another approach to document clustering is self-organizing maps (SOMs). SOMs are a data visualisation technique, based on unsupervised artificial neural networks, that transform high-dimensional data into (usually) two-dimensional representation of clusters. For a detailed overview of SOMs, see Kohonen (2001). There are several research examples of visualization for browsing using SOMs (Heuser et al. 1998; Poincot et al. 1998; Rauber and Merkl 1999; Goren-Bar et al. 2000; Schweighofer et al. 2001; Yang et al. 2003; Dittenbach et al. 2004).

2.2.1.2 Differences within the approach

A major difference within the document clustering community is in algorithms (see above). While previous research showed that

agglomerative algorithms performed better than partitional ones, some studies indicate the opposite. Steinbach et al. (2000) compared agglomerative hierarchical clustering and *K*-means clustering and showed that *K*-means is at least as good as agglomerative hierarchical clustering. Zhao and Karyps (2002) evaluated different partitional and agglomerative approaches and showed that partitional algorithms always lead to better clustering solutions than agglomerative algorithms. In addition, they presented a new type of clustering algorithms called constrained agglomerative algorithms that combined the features of both partitional and agglomerative algorithms. This solution gave better results than agglomerative or partitional algorithms alone. For a comparison of hierarchical clustering algorithms, and added value of some linguistics features, see Hatzivassiloglou et al. (2000). Different enhancements to algorithms have been proposed (see, for example, Liu et al. 2002; Mandhani et al. 2003; Slonim et al. 2003).

Since in document clustering (including SOMs) clusters and their labels are produced automatically, deriving the labels is a major research challenge. In an early example of automatically derived clusters (Garfield et al. 1975), which were based on citation patterns, labels were assigned manually. Today a common heuristic principle is to extract between five and ten of the most frequent terms in the centroid vector, then to drop stop-words and perform stemming, and choose the term which is most frequent in all documents of the cluster. A more complex approach to labelling is given by Glover et al. (2003). They used an algorithm to predict "parent, self, and child terms"; self terms were assigned as clusters' labels, while parent and children terms were used to correctly position clusters in the cluster collection.

Another problem in document clustering is how to deal with large document collections. According to Jain et al. (1999, 316), only the *K*-means algorithm and SOMs have been tested on large collections. An example of an approach dealing with them was presented by Haveliwala et al. (2000), who developed a technique they managed to apply to 20 million URLs.

2.2.1.3 Evaluation methods

Similarly to text categorization, there are many evaluation measures (e.g., precision and recall), and evaluation normally does not include subject experts or users.

Document collections often used are fetched from Text REtrieval Conferences (TREC 2004). In the development stage is the INEX initiative project (INitiative for the Evaluation of XML Retrieval 2004), within which a large document collection of XML documents, over 12,000 articles from IEEE publications from the period of 1995-2002, would be provided.

2.2.1.4 Characteristics of web pages

A number of researchers have explored the potential of hyperlinks in the document clustering process. Weiss et al. (1996) were assigning higher similarities to documents that have ancestors and descendants in common. Their preliminary results also illustrated that combining term and link information yields improved results. Wang and Kitsuregawa (2002) experimented with best ways of combining terms from web pages with words from in-link pages (pointing to the web page) and out-link pages (leading from the web page), and achieved improved results.

Other web-specific characteristics have been explored. Information about users' traversals in the category structure has been experimented with (Chen et al. 2002), as well as usage logs (Su et al. 2001). The hypothesis behind this approach is that the relevancy information is objectively reflected by the usage logs; for example, it is assumed that frequent visits by the same person to two seemingly unrelated documents indicate that they are closely related.

2.2.1.5 Application

Clustering is the unsupervised classification of objects, based on patterns (observations, data items, feature vectors) into groups or clusters (Jain et al. 1999, 264). It has been addressed in various disciplines for many different applications (Jain et al. 1999); in information retrieval, documents are the ones that are grouped or clustered (hence the term document clustering).

Traditionally, document clustering has been applied to improve document retrieval (for a review, see Willet 1988; for an example, see Tombros and Rijsbergen 2001). In this paper the emphasis is on automated generation of hierarchical clusters structure and subsequent assignment of documents to those clusters for browsing.

An early attempt to cluster a document collection into clusters for the purpose of browsing was Scatter/Gather (Cutting et al. 1992). Scatter/Gather would partition the collection into clusters of related

documents, present summaries of the clusters to the user for selection, and when the user would select a cluster, the narrower clusters were presented; when the narrowest cluster would be reached, documents were enumerated. Another approach is presented by Merchkour et al. (1998). First the source collection (an authoritative collection representative in the domain of interest of the users) would be clustered for the user to browse it, with the purpose of helping him/her with defining the query. Then the query would be submitted via a web search engine to the target collection, which is the World Wide Web. The results would be clustered into the same categories as in the source collection. Kim and Chan (2003) attempted to build a personalized hierarchy for an individual user, from a set of web pages the user visited, by clustering words from those pages. Other research has been conducted in automated construction of vocabularies for browsing (Chakrabarti et al. 1998a; Wacholder et al. 2001).

Another application of automated generation of hierarchical category structure and subsequent assignment of documents to those categories is organization of web search engine results (Vivisimo 2004; MetaCrawler web search 2005; Zamir et al. 1997; Zamir and Etzioni, 1998; Palmer et al. 2001; Wang and Kitsuregawa 2002).

2.2.1.6 Summary

Like in text categorization, in document clustering documents are first represented as vectors of term weights. Then they are compared for similarity, and grouped into partitional or hierarchical clusters using different algorithms. Characteristics of web documents similar to those from text categorization approach have been explored.

In evaluation, precision, recall and other measures are used, while end-users and subject experts are normally left out.

Unlike text categorization, document clustering does not require either training documents, or pre-existing categories into which the documents are to be grouped. The categories are created when groups are formed – thus, both the names of the groups and relationships between them are automatically derived. The derivation of names and relationships is the most challenging issue in document clustering.

Document clustering was traditionally used to improve information retrieval. Today it is better suited for clustering search-engine results than for organizing a collection of documents for

browsing, because automatically derived cluster labels and relationships between the clusters are incorrect or inconsistent. Also, clusters change as new documents are added to the collection – such instability of browsing structure is not user-friendly either.

2.3 Document classification

2.3.1 Special features

2.3.1.1 Description of features

Document classification is a library science approach. The tradition of automating the process of subject determination of a document and assigning it to a term from a controlled vocabulary partly has its roots in machine-aided indexing (MAI). MAI has been used to suggest controlled vocabulary terms to be assigned to a document.

The automated part of this approach differs from the previous two in that it is generally not based on either supervised or unsupervised learning. Documents and classes are not necessarily represented by vectors. In document classification, the algorithm typically compares terms extracted from the text to be classified, to terms from the controlled vocabulary. At the same time, this approach does share similarities with text categorization and document clustering: the pre-processing of documents to be classified includes stop-words removal; stemming can be conducted; words or phrases from the text of documents to be classified are extracted and weights are assigned to them based on different heuristics. Web-page characteristics have also been explored, although to a lesser degree.

The most important part of this approach is controlled vocabularies, most of which have been created and maintained for use in libraries and indexing and abstracting services, some of them for more than a century. These vocabularies have devices to "control" polysemy, synonymy, and homonymy of the natural language. They can have systematic hierarchies of concepts, and a variety of relationships defined between the concepts. There are different types of controlled vocabularies, such as classification schemes, thesauri and subject heading systems. With the World Wide Web, new types of vocabularies emerged within the computer science and the Semantic Web communities: ontologies and search-engine directories of web pages. All these vocabularies have distinct characteristics and are consequently better suited for some

classification tasks and applications than others (Koch and Day 1997; Koch and Zettergren 1999; see also Vizine-Goetz 1996). For example, subject heading systems normally do not have detailed hierarchies of terms (exception: Medical Subject Headings), while classification schemes consist of hierarchically structured groups of classes. The latter are better suited for subject browsing. Also, different classification schemes have different characteristics of hierarchical levels. For subject browsing the following are important: the bigger the collection, the more depth should the hierarchy contain; classes should contain more than just one or two documents (Schwartz 2001, 48). On the other hand, subject heading systems and thesauri have traditionally been developed for subject indexing to describe topics of the document as specifically as possible. Since both classification schemes and subject headings or thesauri provide users with different aspects of subject information and different searching functions, their combined usage has been part of practice in indexing and abstracting services. Ontologies are usually designed for very specific subject areas and provide rich relationships between terms. Search-engine directories and other home-grown schemes on the Web, "…even those with well-developed terminological policies such as Yahoo… suffer from a lack of understanding of principles of classification design and development. The larger the collection grows, the more confusing and overwhelming a poorly designed hierarchy becomes…" (Schwartz 2001, 76).

Although well-structured and developed, existing controlled vocabularies need to be improved for the new roles in the electronic environment. Adjustments should include:

1) Improved currency and capability for accommodating new terminology;
2) Flexibility and expandability – including possibilities for decomposing faceted notation for retrieval purposes;
3) Intelligibility, intuitiveness, and transparency – it should be easy to use, responsive to individual learning styles, able to adjust to the interests of users, and allow for custom views;
4) Universality – the scheme should be applicable for different types of collections and communities and should be able to be integrated with other subject languages; and

5) Authoritativeness – there should be a method of reaching consensus on terminology, structure, revision, and so on, but that consensus should include user communities (Schwartz 2001, 77-78).

Some of the controlled vocabularies are already being adjusted, such as: AGROVOC, the agricultural thesaurus (Soergel et al. 2004), WebDewey, which is the Dewey Decimal Classification adapted for the electronic environment (OCLC Dewey Services 2005), and California Environmental Resources thesaurus (CERES 2003).

2.3.1.2 Differences within the approach

The differences occur in document pre-processing, which includes word or phrase extraction, stemming etc., heuristic principles (such as weighting based on where the term/word occurs or occurrence frequency), linguistic methods, and controlled vocabulary applied.

The first major project aimed at automated classification of web pages based on a controlled vocabulary was the Nordic WAIS/World Wide Web Project (1995), which took place at Lund University Library and National Technological Library of Denmark (Ardö et al. 1994; Koch 1994). In this project automated classification of the World Wide Web and Wide Area Information Server (WAIS) databases using Universal Decimal Classification (UDC) was experimented with. A WAIS subject tree was built based on two top levels of UDC, i.e., 51 classes. The process involved the following steps: words from different parts of database descriptions were extracted, and weighted based on which part of the description they belonged to; by comparing the extracted words with UDC's vocabulary a ranked list of suggested classifications was generated. The project started in 1993, and ended in 1996, when WAIS databases came out of fashion.

GERHARD is a robot-generated web index of web documents in Germany (GERHARD 1999, 1998; Möller et al. 1999). It is based on a multilingual version of UDC in English, German and French, adapted by the Swiss Federal Institute of Technology Zurich (Eidgenössische Technische Hochschule Zürich – ETHZ). GERHARD's approach included advanced linguistic analysis: from captions, stop words were removed, each word was morphologically analysed and reduced to stem; from web pages stop words were also removed and prefixes were cut off. After the linguistic analysis,

phrases were extracted from the web pages and matched against the captions. The resulting set of UDC notations was ranked and weighted statistically, according to frequencies and document structure.

Online Computer Library Center's (OCLC) project Scorpion (Scorpion 2004) built tools for automated subject recognition, using DDC. The main idea was to treat a document to be indexed as a query against the DDC knowledge base. The results of the "search" were treated as subjects of the document. Larson (1992) used this idea earlier, for books. In Scorpion, clustering was also used, for refining the result set and for further grouping of documents falling in the same DDC class (Subramanian and Shafer 1998). The System for Manipulating And Retrieving Text (SMART) weighting scheme was used, in which term weights were calculated based on several parameters: the number of times that the term occurred in a record; how important the term was to the entire collection based on the number of records in which it occurred; and, the normalization value, which is the cosine normalization that computes the angle between vector representations of a record and a query. Different combinations of these elements have been experimented with.

Another OCLC project, WordSmith (Godby and Reighart 1998), was to develop software to extract significant noun phrases from a document. The idea behind it was that the precision of automated classification could be improved if the input to the classifier were represented as a list of the most significant noun phrases, instead as the complete text of the raw document. However, it showed that there were no significant differences.

OCLC currently works on releasing Faceted Application of Subject Terminology (FAST 2004), based on the Library of Congress Subject Headings (LCSH), which are modified into a post-coordinated faceted vocabulary. The eight facets to be implemented are: Topical, Geographic (Place), Personal Name, Corporate Name, Form (Type, Genre), Chronological (Time, Period), Title and Meeting Place. FAST could also serve as a knowledge base for automated classification, like the DDC database did in Scorpion (FAST 2003).

Wolverhampton Web Library (WWLib) is a manually maintained library catalogue of British Web resources, within which experiments on automating its processes were conducted (Wallis and Burden 1995; Jenkins et al. 1998). Original classifier from 1995

was based on comparing text from each document to DDC captions. In 1998 each classmark in the DDC captions file was enriched with additional keywords and synonyms. Keywords extracted from the document were weighted on the basis of their position in the document. The classifier began by matching documents against class representatives of top ten DDC classes and then proceeded down through the hierarchy to those subclasses that had a significant measure of similarity (Dice's coefficient) with the document.

"All" Engineering (EELS 2003) is a robot-generated web index of about 300,000 web documents, developed within DESIRE (DESIRE project 1999; DESIRE 2000), as an experimental module of the manually created subject gateway Engineering Electronic Library (2003; Koch and Ardö 2000). Engineering Index (Ei) thesaurus was used; in this thesaurus, terms are enriched with their mappings to Ei classes. Both Ei captions and thesaurus terms were matched against the extracted title, metadata, headings and plain text of a full-text document from the World Wide Web. Weighting was based on term complexity and type of classification, location and frequency. Each pair of term-class codes was assigned a weight depending on the type of term (Boolean, phrase, single word), and the type of class code (main code, the class to be used for the term, or optional code, the class to be used under certain circumstances); a match of a Boolean expression or a phrase was made more discriminating than a match of a single word; a main code was made more important than an optional code. Having experimented with different approaches for stemming and stop-word removal, the best results were gained when an expanded stop-word list was used, and stemming was not applied. The DESIRE project proved the importance of applying a good controlled vocabulary in achieving the classification accuracy: 60% of documents were correctly classified, using only a very simple algorithm based on a limited set of heuristics and simple weighting.

Another robot-generated Web index, Engine-e (Engine-e 2004), used a slightly modified automated classification approach to the one developed in "All" Engineering (Lindholm et al. 2003). Engine-e provided subject browsing of engineering documents based on Ei terms, with six broader categories as starting points.

The project Bilingual Automatic Parallel Indexing and Classification (BINDEX 2001; Nübel at al. 2002) was aimed at indexing and classifying abstracts from engineering in English and

German, using the English INSPEC thesaurus and classification, FIZ Technik's bilingual thesaurus, "Engineering and Management", and the Classification Scheme, "Fachordnung Technik 1997". They performed morpho-syntactic analysis of a document, which consisted of identification of single and multiple-word terms, tagging and lemmatization, and homograph resolution. The extracted keywords were checked against the INSPEC thesaurus and the German part of "Engineering and Management", and classification codes were derived. Keywords which were not in the thesaurus were assigned as free indexing terms.

2.3.1.3 Evaluation methods

Measures such as precision and recall have been used. This approach differs from the other two approaches in that evaluation of document classification tends to also involve subject experts or intended users (Koch and Ardö 2000), which is in line with traditional library science evaluations.

Examples of document collections that have been used are harvested web documents (GERHARD, "All" Engineering), and bibliographic records of Internet resources (Scorpion).

2.3.1.4 Summary

Document classification is a library science approach. It differs from text categorization and document clustering in that well-developed controlled vocabularies are employed, whereas vector space model and algorithms based on vector calculations are generally not used. Instead, selected terms from documents to be classified are compared against terms in the chosen controlled vocabulary, whereby often computational linguistic techniques are employed.

In evaluation, performance measures from information retrieval are used, and, unlike the other two approaches, subject experts or users tend to be involved.

In the focus of research are mainly publicly available operative information systems that provide browsing access to their document collections.

2.4 Mixed approach

Mixed approach is the term used here to refer to a machine-learning or an information-retrieval approach, in which also controlled

vocabularies that have been traditionally used in libraries and indexing and abstracting services are used. There do not seem to be many examples of this approach. Frank and Paynter (2004) applied machine-learning techniques to assign Library of Congress Classification (LCC) notations to resources that already have an LCSH term assigned. Their solution has been applied to INFOMINE (subject gateway for scholarly resources, http://infomine.ucr.edu/), where it is used to support hierarchical browsing. There are also cases in which search engine results were grouped into pre-existing subject categories for browsing (Pratt 1997).

Other mixed approaches are also possible, such as the one applied in the Scorpion project (see section 2.3.1.2). The emergence of this approach demonstrates the potentials for utilizing ideas and methods from another community's approach.

3 Discussion

3.1 Features of automated classification approaches

Several problems with automated classification in general have been identified in the literature. As Svenonius (2000, 46-49) claims, automating subject determination belongs to logical positivism – a subject is considered to be a string occurring above a certain frequency, is not a stop word and is in a given location, such as a title. In clustering algorithms, inferences are made such as "if document A is on subject X, then if document B is sufficiently similar to document A (above a certain threshold), then document B is on that subject". It is assumed that concepts have names, which is common in science, but is not always the case in humanities and social sciences. Automated classification in certain domains has been entirely unexplored, due to lack of suitable document collections or good-quality controlled vocabularies. Another critique given is the lack of theoretical justifications for vector manipulations, such as the cosine measure that is used to obtain vector similarities (Salton 1991, 975).

In regards to similarities and differences between the approaches, document pre-processing (e.g., selection of terms) is common to all the approaches. Various web page characteristics have also been explored by all the three communities, although mostly within the text categorization approach. Major differences

between the three approaches are in applied algorithms, employment or not of the vector-space model and of controlled vocabularies, especially as to how well-suited they are for subject browsing (see 3.3 Application for subject browsing). Since there are similarities between approaches, the hypothesis is that idea exchange and co-operation between the three communities would be beneficial. The hypothesis does seem to be supported by the emergence of the mixed approach. They could all benefit from at least looking into each other's approaches to document pre-processing and indexing, and exchanging ideas about properties of web pages and how they could be used. However, there seems to be little co-operation or idea exchange among them. This is also supported by the fact that, to the author's knowledge, no review paper on automated classification attempted to discuss more than one community's approach. A recent bibliometric study (Golub and Larsen 2005) shows that the three communities are quite clearly mutually independent when looking at citation patterns; and that document clustering and text categorization are closer to each other, while the document classification community is almost entirely isolated. Further research is needed to determine why direct and indirect links are lacking between the document classification and the other two communities, in spite of emergence of the mixed approach.

3.2 Evaluation

The problem of deriving the correct interpretation of a document's subject matter has been much discussed in the library science literature (while much less so in machine learning and information retrieval communities). It has been reported that different people, whether users or subject indexers, would assign different subject terms or classes to the same document. Studies on inter-indexer and intra-indexer consistency report generally low indexer consistency. There are two main factors that seem to affect it:

1) Higher specificity and higher exhaustivity both lead to lower consistency (indexers choose the same first term for the major subject of the document, but the consistency will decrease as they choose more terms); and
2) The bigger the vocabulary, or, the more choices the indexers have, the less likely will they choose the same terms.

The document collection's purpose is another important factor in deciding which classes or terms are to be chosen or made more prominent (Olson and Boll, 99-101).

Having the above in mind, performance measures need to be questioned and evaluation has to be dealt with in the broader contexts of users and their tasks. Subject experts or intended end-users have been mostly excluded from evaluation in text categorization and document clustering approaches, while the document classification approach tends to involve them to a larger degree, corresponding to the tradition of evaluating other library services.

Owing to poor evaluation, it is difficult to estimate to what degree the automated classification tools of today are really applicable in operative information systems and for which tasks.

3.3 Application for subject browsing

Research in text categorization seems to be mainly in improving categorization performance, and experiments are conducted under controlled conditions. Research in which web pages have been categorized into hierarchical structures for browsing generally does not involve well-developed classification schemes, but home-grown structures such as directories of search engines that are not structured and maintained well enough.

In document clustering, clusters' labels and relationships between the clusters are automatically produced. Labelling of the clusters is a major research problem, with relationships between the categories, such as those of equivalence, related-term and hierarchical relationships, being even more difficult to automatically derive (Svenonius 2000, 168). "Automatically-derived structures often result in heterogeneous criteria for category membership and can be difficult to understand" (Chen and Dumais 2000). Also, clusters change as new documents are added to the collection. Unstable category names in Web services and digital libraries, for example, are not user-friendly. Koch and Zettergren (1999) suggest that document clustering is better suited for organizing web search engine results.

Document classification approach employs well-developed classification schemes, which are suitable for subject browsing. However, future research should include improving controlled

vocabularies for browsing in the electronic environment, as well as making them more suitable for automated classification.

Acknowledgments

Many thanks to Traugott Koch, Anders Ardö, Tatjana Aparac Jelušić, Johan Eklund, Ingo Frommholz, Repke de Vries and the Journal of Documentation reviewers for providing valuable feedback on earlier versions of the paper.

References

20 Newsgroups DataSet (1998), *The 4 Universities Data Set*, available at: http://www-2.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/news20.html (accessed 22 December 2004).

Ardö, A. et al. (1994), ”Improving resource discovery and retrieval on the Internet: the Nordic WAIS/World Wide Web project summary report”, *NORDINFO Nytt*, Vol. 17 No. 4, pp. 13-28.

Attardi, G., Gullì, A., and Sebastiani, F. (1999), “Automatic web page categorization by link and context analysis”, *Proceedings of the European Symposium on Telematics, Hypermedia and Artificial Intelligence*, pp. 105-119.

Bekkerman, R. et al. (2003), “Distributional word clusters vs. words for text categorization”, *Journal of Machine Learning Research*, Vol. 3, pp. 1183-1208.

BINDEX (2001), “HLT Project Factsheet: BINDEX”, HLTCentral, available at: http://www.hltcentral.org/projects/print.php?acronym=BINDEX (accessed 22 December 2004).

Blum, A., and Mitchell, T. (1998), “Combining labeled and unlabeled data with co-training”, *Proceedings of the Workshop on Computational Learning Theory*, pp. 99-100.

Cai, L., and Hofmann, T. (2003), “Text categorization by boosting automatically extracted concepts”, *Proceedings of the 26th ACM International Conference on Research and Development in Information Retrieval,* pp. 182-189.

CERES (2003), “CERES thesaurus effort”, CERES The California Environmental Resources Evaluation System, available at: http://ceres.ca.gov/thesaurus/ (accessed 22 December 2004).

Chakrabarti, S. et al. (1998a), “Automatic resource compilation by analyzing hyperlink structure and associated text”, *Proceedings of the Seventh International Conference on World Wide Web 7, Brisbane, Australia,* pp. 65-74.

Chakrabarti, S., Dom, B., and Indyk, P. (1998b), ”Scalable feature selection, classification and signature generation for organizing large

text databases into hierarchical topic taxonomies", *Journal of Very Large Data Bases*, Vol. 7 No. 3, pp. 163-178.

Chan, L.M. (1994), *Cataloging and classification: an introduction*, 2nd ed., McGraw-Hill, New York.

Chen, H., and Dumais, S.T. (2000), "Bringing order to the Web: automatically categorizing search results", *Proceedings of the ACM International Conference on Human Factors in Computing Systems, Den Haag*, pp. 145-152.

Chen, M., LaPaugh, A., and Singh, J.P. (2002), "Categorizing information objects from user access patterns", *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, *4-9 November*, pp. 365-372.

Cutting, D. et al. (1992), "Scatter/Gather: a cluster-based approach to browsing large document collections", *Proceedings of the 15th Annual International ACM/SIGIR Conference, Copenhagen,* pp. 318-329.

DESIRE (2000), "DESIRE: Development of a European Service for Information on Research and Education", DESIRE, available at: http://www.desire.org/ (accessed 22 December 2004).

DESIRE Project (1999), Lunds Universitets Bibliotek, available at: http://www.lub.lu.se/desire (accessed 22 December 2004).

Dittenbach, M., Berger, H., and Merkl, D. (2004), "Improving domain ontologies by mining semantics from text", *Proceedings of the First Asian-Pacific Conference on Conceptual Modeling, Dunedin, New Zealand,* Vol. 31, pp. 91-100.

Dumais, S.T., and Chen, H. (2000), "Hierarchical classification of web content", *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, July 24-28, 2000, Athens, Greece, pp. 256-263.

Dumais, S.T., Lewis, D.D., and Sebastiani, F. (2002), "Report on the Workshop on Operational Text Classification Systems (OTC-02)", *ACM SIGIR Forum,* Vol. 35 No. 2, pp. 8-11.

EELS (2003), *'All' Engineering resources on the Internet: a companion service to EELS*, EELS, Engineering E-Library, Sweden, available at: http://eels.lub.lu.se/ae/ (accessed 22 December 2004).

Engine-e (2004), Lund University Libraries, available at: http://engine-e.lub.lu.se/ (accessed 22 December 2004).

Engineering Electronic Library (2003), Lund University Libraries, available at: http://eels.lub.lu.se/ (accessed 22 December 2004).

FAST (2003), "FAST as a knowledge base for automated classification", OCLC projects, available at: http://www.oclc.org/research/projects/fastac/ (accessed 7 August 2005).

FAST (2004), "FAST: faceted application of subject terminology", OCLC projects, available at: http://www.oclc.org/research/projects/fast/ (accessed 22 December 2004).

Fasulo, D. (1999), "An analysis of recent work on clustering algorithms: technical report", University of Washington, available at: http://citeseer.ist.psu.edu/208269.html (accessed 31 August 2004).

Fisher, M., and Everson R. (2003), "When are links useful? Experiments in text classification", *Proceedings of the 25th European Conference on Information Retrieval, Pisa, IT,* pp. 41-56.

Frank, E., and Paynter, G.W. (2004), "Predicting Library of Congress Classifications From Library of Congress Subject Headings", *Journal of the American Society for Information Science and Technology*, Vol. 55 No. 3, pp. 214-227.

Fürnkranz, J. (1999), "Exploiting Structural Information for Text Classification on the WWW", *Proceedings of the 3rd Symposium on Intelligent Data Analysis*, pp. 487-497.

Fürnkranz, J. (2002), "Hyperlink ensembles: a case study in hypertext classification", *Information Fusion*, Vol.3 No. 4, pp. 299-312.

Garfield, E., Malin, M.V., and Small, H. (1975), "A System for Automatic Classification of Scientific Literature", Reprinted from *Journal of the Indian Institute of Science* Vol. 57 No. 2, pp. 61-74. (Reprinted in: *Essays of an Information Scientist*, Vol. 2, pp. 356-365).

GERHARD (1998), *GERHARD: German harvest automated retrieval and directory,* available at: http://www.gerhard.de/ (accessed 22 December 2004).

GERHARD (1999), "GERHARD - navigating the Web with the Universal Decimal Classification System", available at: http://www.gerhard.de/info/dokumente/vortraege/ecdl99/html/index.htm (accessed 22 December 2004).

Ghani, R., Slattery, S., and Yang, Y. (2001), "Hypertext Categorization using Hyperlink Patterns and Metadata", *Proceedings of the 18th International Conference on Machine Learning*, pp. 178-185.

Glover, E.J. et al. (2002), "Using Web structure for classifying and describing web pages", *Proceedings of the Eleventh International Conference on World Wide Web, Honolulu, Hawaii, USA,* pp. 562-569.

Glover, E.J. et al. (2003), "Inferring hierarchical descriptions", *Proceedings of the Eleventh International Conference on Information and Knowledge Management, November 4-9, 2002,* pp. 507-514.

Godby, J., and Reighart, R. (1998) "The WordSmith indexing system", (*OCLC Digital Archive*), available at: http://digitalarchive.oclc.org/da/ViewObject.jsp?fileid=0000003487:000000090408&reqid=33836 (accessed 22 December 2004).

Golub, K. and Larsen, B. (2005), "Different Approaches to Automated Classification: Is There an Exchange of Ideas?", *Proceedings of the 10th International Conference of the International Society for*

Scientometrics and Informetrics, Stockholm, Sweden, 24-28 July, Vol. 1. pp. 270-274.

Goren-Bar, D. et al. (2000), "Supervised learning for automatic classification of documents using self-organizing maps", *Proceedings of the First DELOS Network of Excellence Workshop on Information Seeking, Searching and Querying in Digital Libraries, Zűrich, Switzerland, 11-12 December*. Available at: http://citeseer.ist.psu.edu/456294.html (accessed 31 August 2007)

Gövert, N., Lalmas, M., and Fuhr, N. (1999), "A probabilistic description-oriented approach for categorising web documents", *Proceedings of the Eighth International Conference on Information and Knowledge Management*, pp. 475-482.

Hartigan, J.A. (1996), "Introduction", *Clustering and classification*, World Scientific, Singapore, pp. 3-5.

Hatzivassiloglou, V., Gravano, L., and Maganti, A. (2000), "An investigation of linguistic features and clustering algorithms for topical document clustering", *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Athens, Greece,* pp. 224-231.

Haveliwala, T.H., Gionis, A., and Indyk, P. (2000), "Scalable techniques for clustering the Web", *Third International Workshop on the Web and Databases, May*, pp. 129-134.

Hersh, W.R. (1994), "OHSUMED: An interactive retrieval evaluation and new large test collection for research", *Proceedings of the 17th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 192-201.

Heuser, U., Babanine, A., and Rosenstiel, W. (1998), "HTML documents classification using (non-linear) principal component analysis and self-organizing maps", *Proceedings of the Fourth International Conference on Neural Networks and their Applications, 11-13 March, Marseilles, France,* pp. 291-295

INitiative for the Evaluation of XML Retrieval (2004), DELOS Network of Excellence for Digital Libraries, available at: http://inex.is.informatik.uni-duisburg.de/ (accessed 22 December 2004).

Jain, A.K., Murty, M.N., and Flynn, P.J. (1999), "Data clustering: a review", *ACM Computing Surveys*, Vol. 31 No. 3, pp. 264-323.

Jenkins, C. et al. (1998), "Automatic classification of web resources using java and Dewey Decimal Classification", *Computer Networks & ISDN Systems*, Vol. 30, pp. 646-648.

Kim, H.R., and Chan, P.K. (2003), "Learning implicit user interest hierarchy for context in personalization", *Proceedings of the International Conference on Intelligent User Interfaces*, pp. 101-108.

Koch, T. (1994), "Experiments with automatic classification of WAIS databases and indexing of WWW", *Internet World & Document Delivery World International 94, London, May*, pp. 112-115.

Koch, T., and Ardö, A. (2000), "Automatic classification", *DESIRE II D3.6a, Overview of results,* available at: http://www.lub.lu.se/desire/DESIRE36a-overview.html (accessed 22 December 2004).

Koch, T., and Day, M. (1997), "The role of classification schemes in Internet resource description and discovery", *EU Project DESIRE, Deliverable D3.2.3*, available at: http://www.lub.lu.se/desire/radar/reports/D3.2.3/ (accessed 22 December 2004).

Koch, T., and Zettergren, A.-S. (1999) "Provide browsing in subject gateways using classification schemes", *EU Project DESIRE II*, available at: http://www.lub.lu.se/desire/handbook/class.html (accessed 22 December 2004).

Kohonen, T. (2001), *Self-Organizing Maps*, 3rd ed., Springer-Verlag, Berlin.

Koller, D., and Sahami, M. (1997), "Hierarchically classifying documents using very few words", *Proceedings of the 14th International Conference on Machine Learning*, pp. 170-178.

Labrou, Y., and Finin, T. (1999), "Yahoo! as an ontology: using Yahoo! categories to describe documents", *Proceedings of the 8th ACM International Conference on Information and Knowledge Management*, pp. 180-187.

Larson, R.R. (1992), "Experiments in automatic Library of Congress Classification", *Journal of the American Society for Information Science*, Vol. 43 No. 2, pp. 130-148.

Li, Y.H., and Jain, A.K. (1998), "Classification of text documents", *The Computer Journal*, Vol. 41 No. 8, pp. 537-546.

Liere, R., and Tadepalli, P. (1998), "Active learning with committees: preliminary results in comparing winnow and perceptron in text categorization", *Proceedings of the 1st Conference on Automated Learning and Discovery*, pp. 591-596.

Lindholm, J., Schönthal, T., and Jansson, K. (2003), "Experiences of Harvesting Web Resources in Engineering using Automatic Classification", *Ariadne* No. 37, available at: http://www.ariadne.ac.uk/issue37/lindholm/.

Liu, X. et al. (2002), "Document clustering with cluster refinement and model selection capabilities", *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland,* pp. 191-198.

McCallum, A. et al. (1998), "Improving text classification by shrinkage in a hierarchy of classes", *ICML-98, 15th International Conference on Machine Learning,* pp. 359-367.

McCallum et al. (1999), "Building domain-specific search engines with machine learning techniques", *AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace*.

McCallum, A. et al. (2000), "Automating the construction of Internet portals with machine learning", *Information Retrieval Journal*, Vol. 3, pp. 127-163.

Mandhani, B., Joshi, S. and Kummamuru K. (2003), "A matrix density based algorithm to hierarchically co-cluster documents and words", *Proceedings of the Twelfth International Conference on World Wide Web, Budapest, Hungary,* pp. 511-518.

Manning, C. and Schütze, H. (1999), *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, MA.

Merchkour, M., Harper, D.J. and Muresan, G. (1998), "The WebCluster project: using clustering for mediating access to the World Wide Web", *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia,* pp. 357-358.

MetaCrawler Web Search (2005), available at: http://metacrawler.com (accessed 5 August 2005).

Mitchell, T. (1997), *Machine Learning*. McGraw Hill, New York, NY.

Mladenic, D. (1998), "Turning Yahoo into an automatic web-page classifier", *Proceedings of the 13th European Conference on Artificial Intelligence,* pp. 473-474.

Mladenic, D. and Grobelnik, M. (2003), "Feature selection on hierarchy of web documents", *Decision Support Systems*, Vol. 35 No. 1, pp. 45-87.

Möller, G. et al. (1999), "Automatic classification of the WWW using the Universal Decimal Classification", *Proceedings of the 23rd International Online Information Meeting. London, 7-9 Dec*, pp. 231-238.

Nordic WAIS/World Wide Web Project (1995), Lund University Libraries, available at: http://www.lub.lu.se/W4/ (accessed 22 December 2004).

Nübel, R. et al. (2002) "Bilingual indexing for information retrieval with AUTINDEX", *Third International Conference on Language Resources and Evaluation, 29th, 30th & 31st May, Las Palmas de Gran Canaria (Spain),* pp. 1136-1149.

OCLC Dewey Services (2005), "About DDC: research: a vital part of ongoing development", available at: http://www.oclc.org/dewey/about/research/ (accessed 8 August 2005).

Olson, H.A., and Boll, J.J. (2001), *Subject analysis in online catalogs*, 2nd ed., Libraries Unlimited, Englewood, CO.

Palmer, C.R. et al. (2001), "Demonstration of hierarchical document clustering of digital library retrieval results", *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries, Roanoke, Virginia,* pp. 451.

Pierre, J.M. (2001), "On the automated classification of web sites", *Linköping Electronic Articles in Computer and Information Science*, Vol. 6 No. 001.

Poincot, P., Lesteven, P.S., and Murtagh, F. (1998), "A spatial user interface to the astronomical literature", *Astronomy & Astrophysics*, 2 May, pp. 183-191.

Pratt, W. (1997), "Dynamic organization of search results using the UMLS", *American Medical Informatics Association Fall Symposium*, pp. 480-484.

Rasmussen, E. (1992), "Clustering algorithms", *Information retrieval: Data structures and algorithms*, Prentice Hall, Engelwood Cliffs, NJ, pp. 419-442.

Rauber, A., Merkl, D. (1999), "SOMLib: A digital library system based on neural networks", *Proceedings of the Fourth ACM Conference on Digital Libraries, Berkeley, California, United States,* pp. 240-241.

Reuters-21578 (2004), available at: http://www.daviddlewis.com/ resources/testcollections/reuters21578/, (accessed 3 August 2005).

Rocchio, J.J. (1971), "Relevance feedback in information retrieval", *The SMART Retrieval System: Experiments in Automatic Document Processing*, Prentice-Hall, Englewood Cliffs, NJ, pp. 313-323.

Ruiz, M.E. and Srinivasan, P. (1999), "Hierarchical neural networks for text categorization", *Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval,* pp. 281-282.

Sahami, M., Yusufali, M., and Baldonado, M.Q. (1998), "SONIA: a service for organizing networked information autonomously", *3rd ACM Conference on Digital Libraries, Pittsburgh,* pp. 200-209.

Salton, G. (1991), "Developments in automatic text retrieval", *Science*, Vol. 253, pp. 974-979.

Schütze, H., Hull, D.A., and Pedersen, J.O. (1995), "A comparison of classifiers and document representations for the routing problem", *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle,* pp. 229-237.

Schwartz, C. (2001), *Sorting out the Web: approaches to subject access,* Ablex, Westport, CT.

Schweighofer, E., Rauber, A. and Dittenbach, M. (2001), "Automatic text representation, classification and labeling in European law", *ICAIL 2001*, pp. 78-87.

Scorpion (2004), OCLC software, available at: http://www.oclc.org/ research/software/scorpion/default.htm (accessed 22 December 2004).

Sebastiani, F. (2002), "Machine learning in automated text categorization", *ACM Computing Surveys*, Vol. 34 No. 1, pp. 1-47.

Slattery, S., and Craven, M. (2000), "Discovering test set regularities in relational domains", *Proceedings of the 17th International Conference on Machine Learning*, pp. 895-902.

Slonim, N., Friedman, N., and Tishby, N. (2003), "Unsupervised document classification using sequential information maximization", *Proceedings of the 25th ACM International Conference on Research and Development of Information Retireval, Tampere, Finland,* pp. 129-136.

Soergel, D. et al. (2004), "Reengineering thesauri for new applications: the AGROVOC example", *Journal of Digital Information*, Vol. 4 No. 4, Article no. 257, available at: http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Soergel/.

Steinbach, M., Karypis, G., and Kumar, V. (2000), "A comparison of document clustering techniques", *KDD Workshop on Text Mining, Boston, MA, 20-23 August*, available at http://www.cs.cmu.edu/~dunja/KDDpapers/Steinbach_IR.pdf.

Su, Z. et al. (2001), "Correlation-based document clustering using web logs", *Proceedings of the 34th Annual Hawaii International Conference on System Sciences*, 3-6 January, Vol. 5, pp. 5022.

Subramanian, S., and Shafer, K.E. (1998) "Clustering", OCLC Publications, available at: http://digitalarchive.oclc.org/da/ViewObject.jsp?objid=0000003409, (accessed 22 December 2004).

Sun, A., Lim, E.-P., and Ng, W.-K. (2001), "Hierarchical Text Classification and Evaluation", *ICDM 2001, IEEE Int. Conf. on Data Mining*, pp. 521-528.

Svenonius, E. (2000), *The intellectual foundations of information organization,* MIT Press, Cambridge, MA.

Thunderstone (2005), *Thunderstone's Web Site Catalog*, available at: http://search.thunderstone.com/texis/websearch (accessed 4 August 2005).

Tombros, A., and van Rijsbergen, C.J. (2001), "Query-sensitive similarity measures for the calculation of interdocument relationships", *Proceedings of the Tenth International Conference on Information and Knowledge Management, Atlanta, Georgia, USA,* pp. 17-24.

Toth E. (2002), "Innovative solutions in automatic classification: a brief summary", *Libri*, Vol. 25 No. 1, pp. 48-53.

TREC (2004), *TREC: Text REtrieval Conference*, National Institute of Standards and Technology, available at: http://trec.nist.gov/ (accessed 22 December 2004).

Vivisimo (2004), *Clusty the clustering engine*, available at: http://www.clusty.com (accessed 22 December 2004).

Vizine-Goetz, D. (1996), "Using library classification schemes for internet resources", *OCLC Internet Cataloging Project Colloquium*,

available at: http://webdoc.sub.gwdg.de/ebook/aw/oclc/man/colloq/v-g.htm (accessed 4 April 2006).

Wacholder, N., Evans, D.K., and Klavans, J.L. (2001), "Automatic identification and organization of index terms for interactive browsing", *Proceedings of the ACM-IEEE Joint Conference on Digital Libraries, Roanoke, Virginia, June*, pp. 128-134.

Wallis, J., and Burden, P. (1995), "Towards a classification-based approach to resource discovery on the Web", University of Wolverhampton, available at: http://www.scit.wlv.ac.uk/wwlib/position.html (accessed 22 December 2004).

Wang, Y., and Kitsuregawa, M. (2002), "Evaluating Contents-Link Coupled Web Page Clustering for Web Search Results", *Proceedings of the Eleventh International Conference on Information and Knowledge Management, McLean, Virginia, USA,* pp. 499-506.

WebKB (2001), *CMU World Wide knowledge base (Web -> KB) project*, available at: http://www-2.cs.cmu.edu/~webkb/, (accessed 22 December 2004).

Weiss, R., et al. (1996), "HyPursuit: a hierarchical network search engine that exploits content-link hypertext clustering", *Proceedings of the Seventh ACM Conference on Hypertext, Washington, DC, March*, pp. 180-193.

Willet, P. (1988), "Recent trends in hierarchic document clustering: a critical review", *Information Processing and Management*, Vol. 24 No. 5, pp. 577-597.

Yahoo! (2005), *Yahoo! Directory,* available at: http://dir.yahoo.com/, (accessed 8 August 2005).

Yang, C., Chen H., and Hong, K. (2003), "Visualization of large category map for Internet browsing", *Decision Support Systems (DSS)*, Vol. 35 No. 1, pp. 89-102.

Yang, Y. (1999), "An evaluation of statistical approaches to text categorization", *Journal of Information Retrieval*, Vol. 1 No. 1/2, pp. 67-88.

Yang, Y., Slattery, S., and Ghani, R. (2002), "A study of approaches to hypertext categorization", *Journal of Intelligent Information Systems*, Vol. 8 Nos 2-3, pp. 219-241.

Zamir, O., and Etzioni, O. (1998), "Web document clustering: a feasibility demonstration", ACM SIGIR'98, Australia, pp. 46-54.

Zamir, O. et al. (1997), "Fast and intuitive clustering of web documents", *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, pp. 287-290.

Zhao, Y., and Karypis, G. (2002), "Evaluation of hierarchical clustering algorithms for document dataset", *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, *McLean, Virginia,* pp. 515-524.

Paper III

PAPER III.
Automated Subject Classification of Textual Web Pages, Based on a Controlled Vocabulary: Challenges and Recommendations

Abstract

The primary objective of this study was to identify and address problems of applying a controlled vocabulary in automated subject classification of textual web pages, in the area of engineering. Web pages have special characteristics such as structural information, but are at the same time rather heterogeneous. The classification approach used comprises string-to-string matching between words in a term list extracted form the Ei (Engineering Information) thesaurus and classification scheme, and words in the text to be classified. Based on a sample of 70 web pages, a number of problems with the term list are identified. Reasons for those problems are discussed and improvements proposed. Methods for implementing the improvements are also specified, suggesting further research.

Golub, K. (2006), "Automated subject classification of textual web pages, based on a controlled vocabulary: challenges and recommendations", *New Review of Hypermedia and Multimedia*, Vol. 12 No. 1, pp. 11-27.

1 Introduction

Classification is, to the purpose of this paper, defined as "…the multistage process of deciding on a property or characteristic of interest, distinguishing things or objects that possess that property from those which lack it, and grouping things or objects that have the property or characteristic in common into a class. Other essential aspects of classification are establishing relationships among classes and making distinctions within classes to arrive at subclasses and finer divisions" (Chan 1994, 259). Automated subject classification (in further text: automated classification) denotes machine-based organization of related information objects into topically related groups. In this process human intellectual processes are replaced by, for example, statistical and computational linguistics techniques. Automated classification is also referred to as automated indexing (Moens 2000; Lancaster 2003). In the literature on automated classification and indexing, the terms automatic and automated are both used. Here the term automated is chosen because it more directly implies that the process is machine-based.

Automated classification has been a challenging research issue for several decades now. A major motivation has been high costs of manual subject classification, in terms of time and human resources. The interest has rapidly grown with the advancement of the World Wide Web, on which the number of available documents has been growing exponentially. Due to the ever-increasing number of documents, there is a danger that recognized objectives of bibliographic systems (finding, collocating, choice, acquisition, navigation) (Svenonius 2000, 20-21) would become left behind; automated means could be a solution to preserve them (30). Apart from bibliographic systems, automated classification finds its use in a wide variety of applications, such as grouping search results by subject; topical harvesting; personalized routing of news articles; filtering of unwanted content for Internet browsers; and many others (Jain et al. 1999; Sebastiani 2002).

According to Golub (2006), one can distinguish between three major approaches to automated classification, the largest being text categorization (coming from machine-learning community), followed by document clustering (information-retrieval community), and document classification, coming from library science

community. While the first two approaches use complex algorithms, they by tradition hardly utilize controlled vocabularies. The library science community research focuses less on algorithms and more on operational systems using controlled vocabularies. The latter approach is more or less based on string-to-string matching of controlled vocabulary terms and text in documents to be classified. Usually weighting schemes are applied with the purpose of indicating degrees to which a term from a document to be classified is significant for the document's topicality. Controlled vocabularies (such as classification schemes, thesauri, subject heading systems) have been traditionally used in libraries, and in indexing and abstracting services, some since the 19th century. They have the devices to control polysemy, synonymy, and homonymy of the natural language, and as such could serve as good-quality structures for subject searching and browsing. Another motivation to apply this approach is to reuse the intellectual effort that has gone into creating such a controlled vocabulary. For further details on the advantages of using pre-existing controlled vocabularies as well as on different approaches to automated classification and indexing see Moens (2000), Browne (2003a, 2003b), and Golub (2006).

String-to-string matching has been explored in linguistics, and controlled vocabularies have been used in automated subject indexing. However, controlled vocabularies largely differ from one another as to their suitability for the task of automated classification or indexing, especially since they have been traditionally designed for other tasks. To the author's knowledge, Engineering Information thesaurus and classification scheme (Milstead 1995) has not been explored in this specific respect by others. In addition, the documents that have been mostly dealt with in these two areas were more traditional document forms, such as research papers, news articles etc., and not web pages. Web pages have specific characteristics such as hyperlinks and anchors, metadata, and structural information, all of which could serve as complementary features to improve automated classification. On the other hand, they are rather heterogeneous; many of them contain little text, metadata provided are sparse and can be misleading, structural tags can be misused, and titles can be without any information significant of the content (e.g., "home page", "untitled document").

This paper is aimed at determining the problems of using controlled vocabularies in automated classification of textual web

pages in the field of engineering, using a string-to-string matching approach based on the Ei thesaurus and classification scheme. The study is based on a sample of 70 web pages and is of a qualitative character.

The paper is laid out as follows: background information with related research and recognized problems are given in the following section; the classification approach used is described in detail in third section; the methodology is given in section 4; in the last two sections, the problems are identified and discussed, followed by concluding remarks and recommendations for further research.

2 Related work

A number of projects and studies have been conducted for the purpose of classifying web pages, using classification schemes. In the Nordic WAIS/World Wide Web Project (Ardö et al. 1994), World Wide Web documents and WAIS (Wide Area Information Server) databases were being automatically classified, using Universal Decimal Classification (UDC). A WAIS subject tree was built based on two top levels of UDC, i.e., 51 classes. The process involved the following steps: words from different parts of database descriptions were extracted; by comparing the extracted words with UDC's vocabulary a list of suggested classifications was generated; the words were weighted based on which part of the description they belonged to. They report that 10% of the total 660 databases do not have any classification at all, which is mostly because there are no significant keywords in the descriptions, and suggest that they would extend their UDC vocabulary, which would help solve this problem.

A later project called GERHARD (German Harvest Automated Retrieval and Directory) was aimed at creating a robot-generated web index of web documents in Germany (Möller et al. 1999). The web index was based on a multilingual version of UDC in English, German and French. GERHARD's approach involved advanced linguistic analysis: 1) processing captions[1], which included stop-words removal, morphological analysis of each word and its reduction to the stem; 2) removing prefixes and stop-words from web pages. The words and phrases were then extracted and matched

[1] A caption is a class notation expressed in words; e.g., in UDC, 'telescopes' is the caption for class '520.2'.

against the captions. The resulting set of UDC notations was ranked and weighted statistically according to frequencies and to the structure of the document.

A major web page classification project was OCLC's (Online Computer Library Center) Scorpion, within which tools were built for automated subject recognition, using Dewey Decimal Classification (DDC). The basic idea was to treat a document to be indexed as a query against a DDC knowledge base. The results of the "search" were treated as subjects of the document. Scorpion also used clustering, to refine the result set and to further group documents falling in the same DDC class (Subramanian and Shafer 1998). Another OCLC project, WordSmith (WordSmith Project), provided support for automated classification. The software developed used a variety of computational linguistics methods to extract significant noun phrases from a document. The idea behind it was that the precision of automated classification could be improved if the input to the classifier were represented as a list of the most significant noun phrases, instead of the complete text of the raw document. However, it showed that there were no significant differences.

Other examples of automated classification of web pages using controlled vocabularies include Wolverhampton Web Library (WWWLib), a manually maintained library catalogue of British web resources, within which experiments with automated classification were conducted (Wallis and Burden 1995; Jenkins et al. 1998). The earlier study is based on matching of words from the document against words in DDC and reports 34% correctly classified documents. The latter study takes advantage of DDC's hierarchy: words are extracted from the document and assigned weights; they are first matched at the top level, proceeding further down the hierarchy until a significant match is found with a leaf node. Also, in a study by Prabowo et al. (2002) ontologies were built with which to classify web pages. The ontologies were based on DDC and Library of Congress Classification (LCC). In the process of web page classification, a feed-forward neural network was used to assist the classifier in measuring the similarity between the web page and a class representative. Term weighting was based on position on a web page, and the number of term occurrences. They claim that their approach results in improved classification accuracy, but also point

to the problem of ontology incompleteness in the process of automated classification.

Related work also includes automated classification of documents other than web pages, using controlled vocabularies such as MeSH (Medical Subject Headings) (Indexing Initiative 2006; Roberts and Souter 2000) or INSPEC thesaurus (Plaunt and Norgard 1997; BINDEX 2001). Lexis-Nexis has developed an approach called SmartIndexing®, based on a controlled list of terms and their profiles, created by indexers, while assignment of those terms is done automatically (Tenopir 1999). In a study similar to ours, but based on medical documents and applying International Code of Disease (ICD) classification scheme (Ribeiro-Neto et al. 2001), five different cases of the classification algorithm failures were discovered: no class found, due to the fact that the ICD alphabetical index is incomplete; a class was found but did not correspond to the manually assigned class because the specialist did not assign an appropriate class; names for narrower concepts did not exist in the alphabetical index; the presence of a human expert is required since specialist knowledge is needed to deduce the class; and, the class assigned by the algorithm was wrong.

3 Approach

3.1 Algorithm

The study is based on an automated classification approach (Koch and Ardö 2000) that has been developed within the DESIRE project (DESIRE 2000) to produce 'All' Engineering (EELS 2003), an experimental module of the manually created subject gateway Engineering Electronic Library (EELS) (Engineering Electronic Library 2003) (no longer maintained).

The algorithm classifies web pages into classes of the Ei classification scheme. Mappings exist between the Ei classes and Ei thesaurus' descriptors; both the captions of classes and the descriptors are matched against extracted title, headings, metadata, and main text of a web page. A list of suggested classes and corresponding relevance scores (S) is produced using the following algorithm:

$$S = \sum_{locs} \left(\sum_{terms} \left(freq[loc_j][term_i] * weight[term_i] * weight[loc_j] \right) \right) .$$

Term weight (*weight[term$_i$])* is taken from the term list (see 3.2). Location weight (*weight[loc$_j$]*) is defined for locations like title, metadata, HTML headings, and main text. The applied formula was 86*scoreTitle, 5*scoreHeadings, 6*scoreMetadata, 1* scoreText, as determined in Golub and Ardö 2005. Frequency (*freq[loc$_j$][term$_i$]*) is the number of times *term$_i$* occurs in the text of location *loc$_j$*.

Only classes with scores above a pre-defined cut-off value are selected as the classes for the document: best results are achieved when the final classes selected are those with scores that contain at least five percent of the sum of all the scores assigned in total, or, if such a class does not exist, the class with the top score is selected. According to the policies for the collection, on the average three classes per document are automatically assigned.

Having experimented with different approaches for stemming and stop-word removal, best results were gained when an expanded stop-word list was used, and when stemming was not applied – stemming was shown to improve recall at the expense of precision (Koch and Ardö 2000, chapter 5).

Precision, recall and F1 measure were used as standard evaluation measures (Moens 2001, 104-105). Precision is the ratio of correct automatic assignments divided by the total number of correct automatic assignments. Recall is the ratio of correct automatic assignments divided by the total number of automatic assignments. The F1 measure is a combination of precision and recall:

F1 (precision, recall) =
2 * precision * recall / (precision + recall) .

By comparing automatically assigned classes to manually assigned ones at all the five levels of specificity (Ei has five hierarchical levels), the F1 measure was 0.26, whereas if comparison was done by reducing all the classes to the first two hierarchical levels, F1 was 0.59 (Golub and Ardö 2005). Also, an additional evaluation was performed, in which a subject expert evaluated both the automatically and manually assigned classes of a random sample of 109 web pages. Based on this type of evaluation, the automated approach gained the F1 of 0.66.

3.2 Term list

The term list used in our approach is based on the Ei thesaurus and classification scheme (Milstead 1995). Here is an extract from the classification scheme:

4 Civil Engineering
…
44 Water and Waterworks Engineering
441 Dams and Reservoirs
…
445 Water Treatment
445.1 Water Treatment Techniques
445.1.1 Potable Water Treatment Techniques
…

Thesaurus contains the following types of terms and relationships: descriptors and their synonyms, related terms, broader and narrower terms, and scope notes. Thesaurus descriptors are mapped to classification codes. The classification codes are organized into 6 categories which are divided into 38 subjects, which are further subdivided into 182 specific subject areas. These are yet further subdivided, resulting in over 800 individual classes in a five-level hierarchy. There are some 20,000 terms in the thesaurus, with an average of 11 terms assigned per class.

The term list contains class captions, descriptors and synonyms, and their mappings to Ei classes. It is organized as a list of triplets: term (single word, Boolean term, phrase), class to which it maps, and weight. Boolean terms consist of words that must all be present but in any order or in any distance from each other. The Boolean terms are not explicitly part of the Ei thesaurus, so they had to be created in a pre-processing step. They are considered to be those terms from the Ei thesaurus, which contain the following strings: *and*, *vs.* (short for *versus*), *,* (comma), *;* (semi-colon, separating different concepts in class captions), *(* and *)* (parentheses, indicating the context of a homonym), *:* (colon, indicating a more specific description of the previous term in a class captions), and -- (double dash, indicating *heading--subheading* relationship).

Concerning weighting, a main class is made more important than an optional class: in the Ei thesaurus, main class is the class to use for the term, while optional class is to be used under certain circumstances. Phrases are assigned the highest weights (40 for main class, 20 for optional class), since they normally are most

discriminatory. Boolean AND-expressions are the next best (15 for main class, 10 for optional class). Single words can be too general and/or have several meanings or uses that make them less specific and should thus be assigned a small weight. In the first run of the experiments, they were assigned the same weights as Boolean entries (15 for main class, 10 for optional class).

Upper-case words from the Ei are left in upper case in the term list, assuming that they are acronyms. All other words containing at least one lower-case letter are converted into lower case.

Here is an excerpt from the Ei thesaurus, based on which the excerpt from the term list (further below) was created:

```
TM Active solar buildings
MC 402
MC 643.1
MC 657.1

TM Catalyst activity
UF Catalysts--Activity
MC 803
MC 804
OC 802.2
```

TM stands for the descriptor, and UF for synonym; MC represents the main class, and OC an optional class. Below is an excerpt from the term list, as based on those two examples:

```
40: active solar buildings=657.1, 643.1, 402,
15: activity @and catalysts=803, 804,
10: activity @and catalysts=802.2,
40: catalyst activity=803, 804,
20: catalyst activity=802.2,
```

3.3 Document collection

The document collection used in the study comprises a sample of 1,000 web pages from the EELS subject gateway (Engineering Electronic Library 2003). EELS web pages have been selected and classified by librarians for end users of the gateway. Using the described algorithm (see 3.1), each class is automatically assigned a score indicating the degree of certainty that it is the correct one. Every document also has manually assigned Ei classes, against which the automatically assigned classes were compared.

4 Methodology

To the purpose of identifying problems of the described approach to automated classification, a qualitative analysis was applied to selected 70 web pages. These were web pages from the document collection (see 3.3), which had all the automatically assigned classes wrong, at least in comparison to the (pre-existing) intellectually assigned classes.

Each web page in the sample was thoroughly examined. The entire text of the web page was read in detail, including title, headings, metadata and the rest of the web page's content. *All* automatically assigned classes were looked at and compared against the intellectually assigned ones, even those below the cut-off value, which were not selected as *the* classes for the document (see 3.1).

The nature of problem for each automatically assigned class that was not intellectually assigned was specified as being one of the following:

- Class not found at all (112 instances in the sample);
- Class found but below threshold (71 instances in the sample);
- Wrong automatically assigned class (148 instances in the sample); and,
- Automatically assigned class that is not really wrong (36 instances in the sample).

For each class it was determined why it was automatically assigned, by looking at terms found on a web page and comparing them against terms in the term list designating corresponding classes.

5 Problems identified

5.1 Class not found at all

In the sample, two major types of disagreement have been identified:

1) Intellectually assigned and automatically assigned classes are of equal length, i.e., they are at the same hierarchical level of specificity: first, second, or third. Several cases were found where classes overlapped in the first three digits, i.e., they were

both at the third hierarchical level (e.g., 921.5 vs. 921.3), and several more overlapped in the first two digits (e.g., 402 vs. 409) or in the first digit only (e.g., 901.4 vs. 912.3). Also, in several instances, intellectually and automatically assigned classes did not overlap even in the first digit (e.g., 901.3 and 723.3).

2) Intellectually assigned and automatically assigned classes are not of equal length, i.e., they are at different levels of specificity. In other words, automatically assigned classes are either broader than manually assigned ones (e.g., 723 vs. 723.5), or narrower (e.g., 723.1.1 vs. 72). Also, in several instances, intellectually and automatically assigned classes did not overlap not even in the first digit (e.g., 655.1 and 731).

These classes were not found because the words in the term list designating the classes were not found in the text of the web page to be classified. The same problem has been discovered in automated classification of medical documents (Ribeiro-Neto et al. 2001, 398).

5.1.1 Recommendations

An ideal approach would be to re-design the Ei thesaurus and classification scheme by adding more synonyms, as well as introducing new concepts. Synonyms could be introduced based on an intellectually produced synonym list such as WordNet (WordNet), although the right sense for each term would also need to be selected manually. Another alternative would be to automatically extract additional terms for each class, from documents known to belong to the class.

To provide for different word forms and different ordering of words in a term, regular expressions could be used, although they need to be manually introduced to the term list. An automated alternative would be to apply natural language processing tools and methods.

5.1.2 Example demonstrating the problem

URL: http://www.iiasa.ac.at/ (as downloaded at the time of the classification process)
Words from title: IIASA home page
Words from metadata: /
Words from headings: international institute for applied systems analysis
Words from the main text: IIASA LOGO welcome to the international institute for applied systems analysis IIASA is a non-governmental research organization located in austria international teams of experts from various disciplines conduct scientific

studies on environmental economic technological and social issues in the context of human dimensions of global change the institute is sponsored by 17 national member organizations in north america europe and asia for more details home - what's new - general info - research publications - options - world - web map - search international institute for applied systems analysis. A-2361 laxenburg austria phone 43-2236-807-0 fax 43-2236-71313 web support for optimal viewing of this site use at least netscape navigator 3 and adobe acrobat reader 3

One of the manually assigned classes: 901.4, which stands for 'Impact of Technology on Society'

The automatically assigned class that is closest to the above manually assigned class: 912.3, which stands for 'Operations Research' and belongs to 912 class for 'Industrial Engineering and Management'

Term list for 901.4:

20: behavioral science computing=911.2, 912.2, 901.4, 461.4,
20: data processing @and social @and behavioral sciences applications=911.2, 912.2, 901.4, 461.4,
20: economic @and social effects=911.2, 901.4,
20: economic effects=911.2, 901.4,
40: engineering @and social aspects=901.4,
40: impact of technology on society=901.4,
20: public risks=901.4,
20: risk studies @and public risks=901.4,
20: robots @and industrial @and socioeconomic aspects=911.2, 901.4,
40: social aspects=901.4,
20: social effects=911.2, 901.4,
20: social sciences computing=911.2, 912.2, 901.4, 461.4,
20: social sciences=911.2, 912.2, 901.4, 461.4,
20: socioeconomic effects=911.2, 901.4,
40: sociological aspects=901.4,
20: sociological effects=911.2, 901.4,
20: technological forecasting=901.4,
20: technology @and economic @and sociological effects=911.2, 901.4,
20: technology transfer=911.2, 901.4,

Term list for 912.3:

10: PERT=912.3,
20: bioengineering @and systems science=731.1, 461.1, 912.3,
20: complex systems=731.1, 461.1, 912.3,
20: composite systems=731.1, 461.1, 912.3,
20: cost effectiveness=912.2, 912.3,
20: data processing @and PERT=912.3,
20: interconnected systems=731.1, 461.1, 912.3,
20: large scale systems=731.1, 461.1, 912.3,
40: operations research=912.3,
20: program evaluation @and review technique=912.3,
20: resource allocation=912.3,
40: system analysis=912.3,
40: system design=912.3,
20: system science=731.1, 461.1, 912.3,
20: system theory=461.1, 912.3, 731.4,
40: systems analysis=912.3,
40: systems design=912.3,
20: systems science @and cybernetics @and large scale systems=731.1, 461.1, 912.3,
20: systems science @and cybernetics @and system theory=461.1, 912.3, 731.4,
20: systems science @and cybernetics=731.1, 461.1, 912.3,

20: systems science=731.1, 461.1, 912.3,
20: systems theory=461.1, 912.3, 731.4,
Results of automated classification: Based on the term list, class 912.3 was assigned since one of the terms designating it, 'systems analysis', was found in the headings and main text of the web page. The matching instances are marked by grey rectangles. The 912.3 was automatically assigned a score of 280, according to the formula given in 3.1:

$$Score_{912.3} = 1_{frequency} * 5_{headings} * 40_{termtype} + 2_{frequency} * 1_{plaintext} * 40_{termtype} = 280$$

5.1.3 Examples of suggested improvements

Enriching the term list for 901.4 with synonyms, such as the following ones, would be beneficial:

environmental issues
economic issues
technological issues
technological change
social issues
systems analysis

In a similar text, one could find different word forms, such as noun, verb and adjectival forms, singular or plural etc. For example, the following could be introduced to the 901.4 term list:

social science
technology @ forecasting
sociology @ effects
sociology @ effect

Also, introducing different ordering should be provided for, e.g., for *impact of technology on society* we could have:

technology @ impact @ society
technology @ impact
technology @ society

5.2 Class found but below threshold

The main reason for not assigning correct classes that were discovered automatically has to do with weighting and cut-off values. This is because only classes with scores above a pre-defined cut-off value are selected as *the* classes for the document: the final classes selected are those with scores that contain at least 5% of the sum of all the scores assigned in total, or, if such a class does not exist, the class with the top score is selected (see 3.1). Another reason could be that the classification algorithm is made to always pick the most specific class as the final one, which is in accordance with the given policy for intellectual classification.

5.2.1 Example demonstrating the problem

URL: http://www.luth.se/depts/mt/hallf/index2.html
(as downloaded at the time of the classification process)
Words from title: solid mechanics hållfasthetslära luleå univ of technology Sweden
Words from metadata: /
Words from headings: /
Words from the main text (excerpt): staff research areas current projects equipment what is solid mechanis under graduate education… material is called solid rather than fluid if it can also fracture mechanics computational computer numercal simulations upport a substantial shearing force over the time scale of some natural process or technological application of interest shearing forces are directed parallel rather than perpendicular to the material surface on which they act the force per unit of area is called shear stress for example consider a vertical metal rod that is fixed to a support at its upper end…
Two of the manually assigned classes: 421, which stands for 'Strength of Building Materials; Mechanical Properties', and 422, which stands for 'Strength of Building Materials; Test Equipment and Methods'
Automatically derived classes selected as the classes for the document: 931.1, which stands for 'Mechanics', 901.2, which stands for 'Education', 901.3, which stands for 'Engineering Research' and 901 for 'Engineering Profession'. These selected classes were the ones that had a score containing at least 5% of the sum of all the scores assigned in total; as given below, the sum of all the scores of all the automatically assigned classes was 11775.
All the automatically derived classes for the document: 38 different classes were automatically derived and ranked (score is in the brackets):
931.1 (3795), 901.2 (1935), 901.3 (1845), 901 (1815), 421 (525), 933.2 (150), 481.1.2 (120), 933.1 (120), 804.2 (105), 481.1 (105), 933 (90), 604.1 (90), 641.1 (90), 657.2 (75), 931.3 (60), 535.1 (60), 804 (60), 931.2 (60), 818.1 (60), 741.1 (60), 412 (45), 444 (45), 422 (45), 657 (45), 408.1 (45), 802.3 (45), 414 (45), 545.3 (30), 483.1 (30), 461.2 (30), 812.3 (30), 531.1 (15), 903.2 (15), 801.4 (15), 932.2 (15), 482.2 (15), 505 (15), 631 (15).

5.2.2 Recommendations

Experiment with different heuristics for weights and cut-offs. For example, weights for terms could be determined based on document vs. collection frequencies, or based on how many documents in which the term occurs are actually relevant vs. the number of documents where it occurs but is not relevant for the document (Salton and Buckley 1988). Statistical methods such as multiple regression could also be used, to derive appropriate weights automatically (see Golub and Ardö 2005, 372).

A different solution could be to include all the automatically found classes, and assign them the automatically derived weights. Weights indicating term importance for a certain document have also been attributed by human indexers performing intellectual (manual) indexing (Moens 2000, 58).

5.3 Wrong automatically assigned class

Based on the sample, four different sub-problems have been identified:

1) Words recognized as homonyms or distant synonyms, e.g.,:

- *association* - a web page on *international association of drilling contractors* is wrongly classified as belonging to class *chemical reactions*, because the word *association* in the term list is mapped to that class;
- *paper* – a web page of the *journal of the electrochemical society* is wrongly classified as belonging to class *pulp and paper* because of the word *paper*, which is on the web page found in the context of research papers published in the journal, but in the term list it is mapped to the class *pulp and paper*;
- *architecture* and *facilities* – a web page on computers is wrongly classified as belonging to class *buildings and towers* because the word *architecture* referring to computer architecture, is found on the web page but in the term list is mapped to that class; a web page on *labs and facilities* is also wrongly classified as *buildings and towers* because the word *facilities* in the term list mapped to that class;
- *information technology* – a web page on computer information centre is wrongly classified as belonging to class *information science* because the term *information technology* found on the web page is in the term list mapped to that class;
- *systems analysis* – a web page about an institute studying technological impact on society is wrongly classified as belonging to class *control systems* in control engineering, because the term *systems analysis* is mapped to that class;
- *safety* – a web page about the world wide web virtual library safety is wrongly classified as belonging to class *accidents and accident prevention* in engineering, because the word *safety* in the term list is mapped to that class;
- *hardware* – a web page containing string *bibliographies on software/hardware engineering and formal methods* is wrongly classified as belonging to class *small tools and hardware* because the word *hardware* is mapped to that class;

2) Word found on a web page is there because it is an instance of what it represents, and it is not about such instances, e.g.,:

   - a web page containing bibliographies or allowing access to databases on computer science is classified as *information science* or as *database systems*, instead of being classified as *computer science*;
   - a web page on SQL standard or a web page on a classification system is wrongly classified as *codes and standards* in engineering;
   - a web page that is an information service for artificial intelligence is classified as *information services*;
   - a web page on online tutorials and e-learning programs for technical fields is wrongly classified as a web page on *education*;

3) Too distant term-class mappings, including cases when one term in the term list is mapped to several different classes, e.g.,:

   - word *bibliographies* is mapped to *information science*;
   - terms *policy*, *technology*, and *public policy* are all mapped to *engineering profession*;
   - word *air* is mapped to *chemical products generally*;
   - word *textbooks* is mapped to *education* and *information dissemination*;
   - term *social sciences* is mapped to *human engineering*, *industrial economics*, *management* in industrial engineering, and *impact of technology on society*;
   - word *cryptography* is mapped to *telephone and other line communications,* to *electronic equipment, radar, radio and television*, to *electro-optical communication*, and *computer software, data handling and applications*;

4) Words mentioned on the web page have little to do with the web page's "aboutness", e.g., an institution's web page is wrongly classified as *facsimile systems and technology*, because among their contact information, there is also a fax number, and the word *fax* is mapped to that class.

5.3.1 Recommendations

Word-sense disambiguation in context is needed. Homonym and polysem resolution could be improved by introducing the rule in the algorithm that, before making a final decision whether to assign a certain class, it checks if similar classes (e.g., within the same top-level class) have also been assigned to that document. For example, if class 811.1 for *Pulp and Paper* is assigned because the word *paper* is found (in the term list: *15: paper=811.1*), the algorithm would check if other classes starting with *811* have also been automatically assigned:

```
811 :: Cellulose, Paper and Wood Products
811.1.1 :: Papermaking Processes
811.1.2 :: Papermaking Equipment
811.2 :: Wood and Wood Products
811.3 :: Cellulose and Derivatives
```

Another approach would be to classify also sub-pages of the web page being classified to confirm the classification of the main page. For example, if class 811.1 for *Pulp and Paper* is assigned to a web page, the algorithm could check if the web pages to which this page links are assigned other classes starting with *811*. This could at the same time solve the problem of web pages containing hardly any text.

It is also possible to provide context in the term list itself, by applying Boolean operators to create new, context-enriched terms, e.g., by adding a broader term or class caption to single words and homonyms. For example, term item *15: association=802.2*, could be replaced by combining the single word *association* with words from classes at level 800, for *Chemical Engineering*:

```
40: association @and chemical engineering=505.1
```

It might help if single words are given lower weight as well. A different approach would be to enrich the context by introducing synonyms of the correct senses from WordNet (WordNet), which, again, needs to be done manually. Yet another way to proceed would be to determine in which context in the document collection homonyms mostly occur; if a homonym most often occurs in a non-Ei context, put it on the stop list, or use a large negative weight in the term list. Also individual rules could be introduced, such as, the

class *facsimile systems and technology* could be assigned when there are not any numbers following the word *fax*. Ideally, there would also be an expert going through the list and removing those entries that are too distant in meaning from the class they are to designate.

Another problem is that the classification algorithm assigns a class based on a Boolean term, no matter how far from each other elements of the Boolean term are on a web page. A proximity measure could be introduced in the algorithm, defining that, for example, two words connected by a Boolean AND should not have more than seven words between them.

5.4 Automatically assigned class that is not really wrong

In the sample, a number of classes were found that were not really wrong, but were not intellectually assigned. For example, a web page on chemistry and biochemistry is automatically assigned a class for *biology*; or, a web page on *mechanical engineering, plant and power* and *electric transmission and distribution* is assigned *nuclear reactors*, which is not really wrong because *nuclear power* is also what it is about. This could have to do with the subject indexing policy, such as exhaustivity (see Moens 2000, 72).

5.4.1 Recommendations

Further research is needed to determine to what degree and in which contexts classes that are automatically assigned are to be discarded as incorrect or could actually be useful to end-users. Methodology for such user-based studies needs to be developed.

6 Concluding remarks

In the study, several different problems of string-to-string matching approach to automated classification were identified and discussed. The focus of the study was a collection of web pages in the field of engineering. Web pages present a special challenge: because of their heterogeneity the same principle (e.g., words from headings are more important than main text) is not applicable to all the web pages of a collection. For example, utilizing information from headings on all web pages might not give improved results, since headings are sometimes used simply instead of using bold or a larger font size.

The matching was based on a term list derived from the Ei thesaurus and classification scheme, which contains a number of different types of terms and relationships, and thesaurus descriptors are mapped to classification codes. As a result of these elaborate relationships, on average 11 terms are assigned per class. This allows for relatively good classification results by using only the simple string-to-string matching.

However, a number of weaknesses of the described approach were identified, and ways to deal with those were proposed for further research. These include enriching the term list with synonyms and different word forms, adjusting the term weights and cut-off values and word-sense disambiguation. In our further research the plan is to implement automated methods. On the other hand, the suggested manual methods (e.g., adding synonyms) would, at the same time, improve Ei's original function, that of enhancing retrieval. Having this purpose in mind, manually enriching a controlled vocabulary for automated classification or indexing would not necessarily create additional costs.

Acknowledgments

The author would like to thank Anders Ardö, reviewers and editors whose suggestions helped improve the paper.

The research was funded by ALVIS, an EU Sixth Framework Programme, Information Society Technologies (IST-1-002068-STP).

References

Ardö, A. et al. (1994), "Improving resource discovery and retrieval on the Internet: the Nordic WAIS/World Wide Web project summary report", *NORDINFO Nytt*, Vol. 17 No. 4, pp. 13-28.

BINDEX (2001), "HLT Project Factsheet: BINDEX", HLTCentral, available at: http://www.hltcentral.org/projects/print.php?acronym=BINDEX (accessed 18 April 2006).

Browne, G. (2003a), "Automatic categorisation: Part 1: principles of classification", *Online Currents*, Vol. 18 No. 1, pp. 17-22.

Browne, G. (2003b), "Automatic categorisation: Part 2: technology", *Online Currents*, Vol. 18 No. 2, pp. 7-11.

Chan, L.M. (1994), *Cataloging and classification: an introduction*, 2nd ed., McGraw-Hill, New York.

DESIRE (2000), "DESIRE: Development of a European Service for Information on Research and Education", DESIRE, available at: http://www.desire.org/ (accessed 22 December 2004).

EELS (2003), 'All' Engineering resources on the Internet: a companion service to EELS, EELS, Engineering E-Library, Sweden, available at: http://eels.lub.lu.se/ae/ (accessed 13 January 2006).

Engineering Electronic Library (2003), Lund University Libraries, available at: http://eels.lub.lu.se/ (accessed 13 January 2006).

Golub, K. (2006), "Automated subject classification of textual Web documents" *Journal of Documentation*, Vol. 62 No. 3, pp. 350-371.

Golub, K., and Ardö, A. (2005), "Importance of HTML structural elements and metadata in automated subject classification", *Proceedings of the 9th European Conference on Research and Advanced Technology for Digital Libraries, Vienna, Austria, 18-23 September*, pp. 368-378.

"Indexing Initiative" (2006), National Library of Medicine, available at: http://ii.nlm.nih.gov/ (accessed 13 January 2006).

Jain, A.K., Murty, M.N., and Flynn, P.J. (1999), "Data clustering: a review", *ACM Computing Surveys*, Vol. 31 No. 3, pp. 264-323.

Jenkins, C. et al. (1998), "Automatic classification of web resources using java and Dewey Decimal Classification", *Computer Networks & ISDN Systems*, Vol. 30, pp. 646-648.

Koch, T., and Ardö, A. (2000), "Automatic classification of full-text HTML-documents from one specific subject area", *DESIRE II D3.6a* available at: http://www.lub.lu.se/desire/DESIRE36a-WP2.html (accessed 13 January 2006).

Lancaster, F.W. (2003), *Indexing and abstracting in theory and practice,* 3rd ed, Facet, London.

Milstead, J, ed. 1995. *Ei thesaurus*. 2nd ed. Hoboken, NJ: Engineering Information Inc.

Moens, M.-F. (2000), *Automatic indexing and abstracting of document texts,* Kluwer, Boston.

Möller, G. et al. (1999), Automatic classification of the WWW using the Universal Decimal Classification, in *Proceedings of the 23rd International Online Information Meeting*, pp. 231-238.

Plaunt, C., and Norgard, B.A. (1998), "An association-based method for automatic indexing with controlled vocabulary", *Journal of the American Society for Information Science* Vol. 49 No. 10, pp. 887-902.

Prabowo, R. et al. (2002), "Ontology-based automatic classification for the Web pages: design, implementation and evaluation", *Proceedings of the 3rd International Conference on Web Information Systems Engineering*, pp. 182-191.

Ribeiro-Neto, B, Laender, A.H.F., and de Lima, L.R.S (2001), "An experimental study in automatically categorizing medical documents",

Journal of the American Society for Information Science and Technology, Vol. 52 No. 5, pp. 391-401.

Roberts, D., and Souter, C. (2000), "The automation of controlled vocabulary subject indexing of medical journal articles" *Aslib Proceedings*, Vol. 52 No. 10, pp. 384-401.

Salton, G., and Buckley, C. 1988, "Term weighting approaches in automatic text retrieval", *Information Processing and Management*, Vol. 24 No. 5, pp. 513-523. Reprinted in: K. Sparck Jones, P. Willett (Eds.), (1997), *Readings in information retrieval,* Morgan Kaufman, San Francisco, pp. 323-328.

Sebastiani, F. (2002), "Machine learning in automated text categorization", *ACM Computing Surveys*, Vol. 34 No. 1, pp. 1-47.

Subramanian, S., and Shafer, K.E. (1998), "Clustering", available at: http://digitalarchive.oclc.org/da/ViewObject.jsp?objid=0000003409 (accessed 13 January 2006).

Svenonius, E. (2000), *The intellectual foundations of information organization,* MIT Press, Cambridge, MA.

Tenopir, C. (1999), "Human or automated, indexing is important", *Library Journal*, Vol. 124 No. 18, pp. 34-38.

Wallis, J., and Burden, P. (1995), "Towards a classification-based approach to resource discovery on the Web", University of Wolverhampton, available at: http://www.scit.wlv.ac.uk/wwlib/position.html (accessed 13 January 2006).

WordNet, available at: http://wordnet.princeton.edu/ (accessed 13 January 2006).

"WordSmith Project", OCLC, available at: http://www.oclc.org/dewey/about/research/ (accessed 13 January 2006).

Paper IV

PAPER IV.
Importance of HTML Structural Elements and Metadata in Automated Subject Classification

Abstract

The aim of the study was to determine how significance indicators assigned to different web page elements (internal metadata, title, headings, and main text) influence automated classification. The document collection that was used comprised some 1,000 web pages in engineering, to which Engineering Information classes had been manually assigned. The significance indicators were derived using several different methods: (total and partial) precision and recall, semantic distance and multiple regression. It was shown that for best results all the elements have to be included in the classification process. The exact way of combining the significance indicators turned out not to be overly important: using the F1 measure, the best combination of significance indicators yielded no more than 3% higher performance results than the baseline.

Golub, K., and Ardö, A. (2005), "Importance of HTML structural elements and metadata in automated subject classification", *Proceedings of the 9th European Conference on Research and Advanced Technology for Digital Libraries, Vienna, Austria, 18-23 September*, pp. 368-378.

1 Introduction

Automated subject classification has been a challenging research issue for several decades now, a major motivation being high costs of manual classification. The interest rapidly grew around 1997, when search engines could not do with just full-text retrieval techniques, because the number of available documents grew exponentially. Due to the ever-increasing number of documents, there is also a danger that recognized objectives of bibliographic systems (finding, collocating, choice, acquisition, navigation) (Svenonius 2000, 20-21) would be left behind; automated means could be a solution to preserve them (30). Automated subject classification of text finds its use in a wide variety of applications, such as: organizing documents into subject categories for topical browsing, which includes grouping search results by subject; topical harvesting; and, many others (see Sebastiani 2002; Jain et al. 1999).

A frequent approach to web-page classification has been a bag-of-words representation of a document, in which all parts of a web page are considered to be of equal significance. However, unlike other text documents, web pages have certain characteristics, such as internal metadata, structural information, hyperlinks and anchors, which could serve as potential indicators of subject content. For example, words from title could be more indicative of a page's content than headings. The degree to which different web page elements are indicative of its content is in this paper referred to as significance indicator.

With the overall purpose of improving our classification algorithm (see section 2.3), the aim was to determine the importance of distinguishing between different parts of a web page. Significance of four elements was studied: title, headings, metadata, and main text.

The paper is structured as follows: in the second section a literature review is given, evaluation issues are discussed and the algorithm used is described (2 Background); in the third section document collection as well as methodology for deriving significance indicators are described (3 Methodology); deriving and testing the significance indicators is presented in section 4 (4 Significance indicators). The paper ends with conclusions and further research (5 Conclusion).

2 Background

2.1 Related work

A number of issues related to automated classification of documents and significance of their different parts have been explored in the literature. Kolcz et al. (2001) studied news stories features and found out that initial parts of a story (headline and first two paragraphs) give best results, reflecting the fact that news stories are written so as to capture readers' attention. Pierre (2001) gained best results in targeted spidering when using contents of keywords and description metatags as the source of text features, while body text decreased classification accuracy. Ghani et al. (2001) also showed that metadata can be very useful for improving classification accuracy. Blum and Mitchell (1998) compared two approaches, one based on full-text, and one based on anchor words pointing to the target pages, and showed that anchor words alone were slightly less powerful than the full-text alone, and that the combination of the two was best. Glover et al. (2002) claimed that text in citing documents close to the citation often had greater discriminative and descriptive power than text in target documents. Similarly, Attardi et al. (1999) also used information from the context where a URL that refers to that document appears and had encouraging results. Fürnkranz (2002) used portions of texts from all pages that point to the target page: the anchor text, the headings that structurally precede it, the text of the paragraph in which it occurs, and a set of (automatically extracted) linguistic phrases that capture syntactic role of the anchor text in the paragraph; headings and anchor text proved to be most useful.

On the other hand, Ghani et al. (2001) claim that including words from linked neighbourhoods should be done carefully since the neighbourhoods could be rather "noisy". Different document collections contain web pages of various characteristics. If certain characteristics are common to the majority of web pages in the collection, an appropriate approach taking advantage of those could be applied, but if the web pages are very heterogeneous, it is difficult to take advantage of any of the web-specific characteristics (Yang et al. 2002; Fisher and Everson 2003; Slattery and Craven 2000).

2.2 Evaluation challenge

The problem of deriving the correct interpretation of a document's subject matter has been much discussed in the library science and related literature. It has been reported that different people, whether users or subject indexers, would assign different subject terms or classes to the same document. Studies on inter-indexer and intra-indexer consistency report generally low indexer consistency. There are two main factors that seem to affect it: 1) higher exhaustivity and specificity of subject indexing both lead to lower consistency (indexers choose the same first term for the major subject of the document, but the consistency decreases as they choose more classes or terms); 2) the bigger the vocabulary, or, the more choices the indexers have, the less likely they will choose the same classes or terms (Olson and Boll 2001, 99-101).

In this study we start from the assumption that manual classes in our document collection are correct, and compare results of automated classification against them. The classification system used in the study is Engineering Information (Ei), which is rather big (around 800 classes) and deep (five hierarchical levels), allowing many different choices. Without a thorough qualitative analysis of automatically assigned classes we cannot be sure if the classes assigned by the algorithm, which were not manually assigned, are actually wrong.

2.3 Description of the algorithm

This study is based on an automated classification approach (Ardö and Koch 1999) that has been developed within the DESIRE project (DESIRE 2000) to produce "All" Engineering (EELS 2003), an experimental module of the manually created subject gateway Engineering Electronic Library (EELS) (no longer maintained).

The algorithm classifies web pages into classes of the Ei classification scheme. Mappings exist between the Ei classes and Ei thesaurus descriptors; both the captions of classes and the descriptors are matched against extracted title, headings, metadata, and main text of a web page. Each time a match is found, the document is assigned the corresponding class, which is awarded a relevance score, based on which term is matched (single word, phrase, Boolean), the type of class matched (main or optional) (*weight[term]*), and the part of the web page in which the match is

found (*weight[loc]*). A match of a phrase (a number of words in exact order) or a Boolean expression (all terms must be present but in any order) is made more discriminating than a match of a single word; a main class is made more important than an optional class (in the Ei thesaurus, main class (code) is the class to use for the term, while optional class (code) is to be used under certain circumstances). A list of suggested classes and corresponding relevance scores (S) is produced using the following algorithm:

$$S = \sum_{locs} \left(\sum_{terms} \left(freq[loc_j][term_i] * weight[term_i] * weight[loc_j] \right) \right) . \quad (1)$$

Only classes with scores above a pre-defined cut-off value (see section 4.5) are selected as the classes for the document. Having experimented with different approaches for stemming and stop-word removal, the best results were gained when an expanded stop-word list was used, and stemming was not applied. For more information on the algorithm, see Ardö and Koch (1999) and Koch and Ardö (2000).

3 Methodology

3.1 Document collection

The document collection used in the study comprises a selection of web pages from the EELS subject gateway (Engineering Electronic Library 2003). EELS web pages have been selected and classified by librarians for end users of the gateway.

For the study, only pages in English were kept, the reason being that Ei captions and descriptors are in English. Also, some other pages were removed because they contained very little or no text. (The problem of pages containing hardly any text could be dealt with in the future, by propagating the class obtained for their subordinate pages.) The final document collection consisted of 1,003 web pages in the field of engineering.

The data were organized in a relational database. Each document in the database was assigned Ei classes derived from the following elements:

- Title;
- Headings;
- Metadata; and,
- Page's main text (Text).

Each class was automatically assigned a score indicating the degree of certainty that it is the correct one. Every document also had manually assigned Ei classes (Manual), against which the automatically assigned classes were compared.

3.2 Methods for evaluation and deriving significance indicators

Various measures have been used to evaluate different aspects of automated classification performance (Yang 1999). Effectiveness, the degree to which correct classification decisions have been made, is often evaluated using performance measures from information retrieval, such as precision and recall, and F1 measure being the harmonic mean of the two. Solutions have also been proposed to measure partial overlap, i.e., the degree of agreement between correct and automatically assigned classes (see, for example, Ceci and Malerba 2003).

In this study, three methods have been used for evaluating and deriving the significance of different web-page elements:

1) Total and partial precision, recall, and F1 measures (using macroaveraging);
2) Semantic distance; and,
3) Multiple regression.

1) The Ei classification scheme has a solid hierarchical structure, allowing for a rather credible test on partial overlap. Three different levels of overlap were tested: total overlap; partial overlap of the first three digits, e.g., *932.1* and *932.2* are considered the same; and, partial overlap of the first two digits, e.g., *932* and *933* are considered the same. Partial overlap of the first four digits has not been conducted because there were few classes of five-digit length in the document collection.

2) In the literature, different similarity measures have been used for hypermedia navigation and retrieval (see, for example, Tudhope and Taylor 1997). Semantic distance, a numerical value representing the difference in meaning between two concepts or terms, is one of them. There are different ways in which to calculate it. For example, the measure of clicking distance in a directory-browsing tree can be used. We used the hierarchical structure of the Ei classification scheme as the means of obtaining the following (rather arbitrary) measures of semantic distance between any two classes:

- 4, when the classes differ already in the first digit (e.g., 601 vs. 901);
- 2, when the classes differ already in the second digit (e.g., 932 vs. 901);
- 1, when the classes differ in the third digit (e.g., 674.1 vs. 673.1); and
- 0.5, when the classes differ in the fourth digit (e.g., 674.1 vs. 674.2).

Those values reflect how the hierarchical system is structured; e.g., we say that class 6 and class 7 are more distant from each other than classes 63 and 64, which are in turn more distant in meaning than 635.1 and 635.2.

Calculations were conducted using the average distance between manually and automatically assigned classes. For each document, average distances were calculated for each of the four elements, and then the values were averaged for all the documents. When there was more than one manually assigned class per document, the semantic distance was measured between an automatically assigned class and that manually assigned class which was most similar to the automatically assigned one.

3) Multiple regression was used in a rather simplified way: scores assigned based on individual elements of a web page were taken as independent variables, while the final score represented the dependent variable. The dependent variable was set to either 1000 or 0, corresponding to a correct or an incorrect class respectively.

4 Significance indicators

4.1 General

In Table 1 basic classification characteristics and tendencies of our document collection are given. All the documents (1,003) have at least one, and no more than six manually assigned classes, the majority having up to three classes. Manual assignment of classes was based on collection-specific classification rules.

Concerning automatically assigned classes based on different parts of a page, not all the pages have classes based on all of them. Classes based on text are assigned to the majority of documents, while those based on metadata to the least number of documents. Based on only title, headings, or metadata, less than 50% of the documents would get classified at all. On average, per every document there are 2 manually assigned classes, 2 classes based on title, 4 based on headings, 9 based on metadata, and 18 classes based on text.

In the whole collection there are 753 different classes assigned, either manually or automatically. The largest variety comes from the group of classes assigned based on text (675), which is more than twice as many as manually assigned (305).

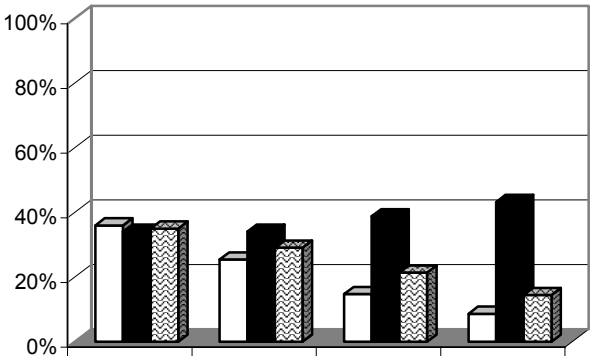
Table 1. Distribution of classes in the document collection. First data row shows how many documents have been classified, second row how many classes have been assigned in the whole of the document collection, and the last row how many different individual classes, out of some 800 possible, have been assigned.

	Manual	Title	Headings	Metadata	Text
Number of classified doc.	1003	411	391	260	964
In the data collection	1943	827	1504	2227	17089
Different classes	305	174	329	406	675

4.2 Precision and recall

Figure 1 shows the degree of automated classification accuracy when words are taken solely from the four different parts of the web page. While title tends to yield best precision, which is 27% more than the worst element (text), text gives the best recall, but only 9%

more than the worst element (title). Precision and recall are averaged using the F1 measure, according to which title performs the best (35%), closely followed by headings (29%), metadata (21%) and text (15%).

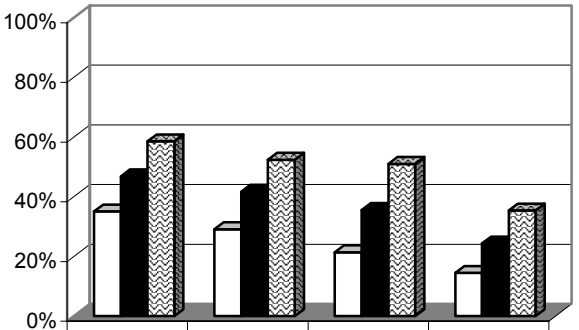


	Title	Headings	Metadata	Text
Precision	36%	25%	15%	9%
Recall	34%	34%	39%	43%
F	35%	29%	21%	14%

Figure 1. Precision, recall and F1 measure

4.2.1 Partial precision and recall

When testing the algorithm performance for partial overlap (Figure 2), precision and recall for all parts of a web page give much better results (title in 2-digit overlap achieves 59%). The ratio between their performance for both two- and three-digit overlap is the same as for total overlap: title performs the best, followed by headings, metadata and text.



	Title	Headings	Metadata	Text
F complete overlap	35%	29%	21%	14%
F 3-digit overlap	47%	42%	36%	24%
F 2-digit overlap	59%	52%	51%	35%

Figure 2. F1-measure values for total overlap, three- and two-digit overlap.

4.3 Semantic distance

Using the semantic distance method, the calculations (Table 2) show that automatically assigned classes are on the average wrong in the third and second digits. Just like precision and recall results for partial overlap (see section 4.2), best results (smallest semantic distances) are achieved by title, followed by headings, metadata and text.

Table 2. Semantic distance.

	Title	Headings	Metadata	Text
Mean distance	1.3	1.7	1.8	2.2

4.4 Deriving significance indicators

As we saw in section 4.1, not every document has all the four elements containing sufficient terms for automated classification. Thus, in order to get documents classified, we need to use a combination of them. How to best combine them has been experimented with in this section, by applying results gained in

evaluation using the F1 measure, semantic distance, and multiple regression.

The symbols used in formulae of this section are:

- S – final score for the automatically assigned class;
- STi – score for the automatically assigned class based on words in Title;
- SH – score for the automatically assigned class based on words in Headings;
- SM – score for the automatically assigned class based on words in Metadata; and,
- STe – score for the automatically assigned class based on words in Text.

The baseline, in which all the elements have equal significance, is represented with the following formula:

$$S = STi + SH + SM + STe . \quad (2)$$

Based on evaluation results, the following co-efficients, representing significance indicators, have been derived (the co-efficients were normalized by reducing the smallest co-efficient to one and by rounding others to integer values):

A. Based on total overlap and F1 measure values:

$$S = 3*STi + 2*SH + 2*SM + STe . \quad (3)$$

These co-efficients have been derived by simply taking the F1 measure values of each of the algorithms (Figure 1). Similar co-efficients have also been derived using partial overlap (respectively): 2, 2, 2, 1 in three-digit overlap, and 2, 1, 1, 1 in two-digit overlap.

B. Based on multiple regression, with scores not normalized for the number of words contained in title, headings, metadata, and text:

$$S = 86*STi + 5*SH + 6*SM + STe . \quad (4)$$

C. Based on multiple regression, with scores normalized for the number of words contained in title, headings, metadata, and text:

$$S = STi + SH + SM + 5*STe\ . \qquad (5)$$

D. On the basis of semantic distance results, the best significance indicator performs less than twice as well as the worst one, so all co-efficients are almost equal, as in (2).

4.5 Evaluation

4.5.1 Defining a cut-off

As described in section 2.3, each document is assigned a number of suggested classes and corresponding relevance scores. Only a few classes with best scores, those above a certain cut-off value, are finally selected as the classes representing the document.

Different cut-offs, that would give best precision and recall results, were experimented with. Also, the number of documents that would be assigned at least one class, and the number of classes that would be assigned per document, were taken into consideration. Best results were achieved when the final classes selected were those with scores that contained at least 5% of all the scores assigned to all the classes, or, if such a class had not existed, the class with the top score was selected. In this case, F1 was 27%, there were about 4,000 classes assigned as final, and all documents were classified. This is the cut-off we used in the study.

4.5.2 Results

As seen from Table 3, the evaluation showed that different significance indicators make hardly any difference in terms of classification algorithm performance. Co-efficients in (3) and (5) are similar to the ones in the baseline (2), and, compared to the baseline (2), which performs 23% in F1, normalized multiple regression (5) performs worse by 1%, while the formula based on F1 measure (3) performs the same. The best result was achieved using non-normalized multiple regression (4), which performs by 3% better than the baseline. This formula gives big significance indicator to classes that were assigned based on the title.

Table 3. Results of applying different co-efficients as significance indicators.

	Baseline (2)	F1 (3)	Regression (4)	Regression N. (5)
Precision	16%	17%	21%	16%
Recall	39%	39%	35%	38%
F1	23%	23%	26%	22%
Number of pages	1003	1003	1003	1003
Number of classes	5174	5063	4073	5147

5 Conclusion

The aim of this study was to determine the significance of different parts of a web page for automated classification: title, headings, metadata, and main text. The significance indicators were derived using several different methods: (total and partial) precision and recall, semantic distance, and multiple regression. The study showed that using *all* the structural elements and metadata is necessary since not all of them occur on every page. However, the exact way of combining the significance indicators turned out not to be highly important: the best combination of significance indicators is only 3% better than the baseline.

Reasons for such results need to be further investigated. One could guess that this is due to the fact that the web pages in our document collection were rather heterogeneous; on the other hand, they were selected by librarians for end users of an operational service, and as such they might indicate what such web-page collections are like. Apart from heterogeneity, the problem could be that metadata were abused, and that certain tags were misused (e.g., instead of using appropriate tags for making text bold, one used a headings tag, which has the same effect on the screen).

Concerning evaluation of automated classification in general, further research is needed to determine the true value of the classification results. To that purpose information specialists and users could be involved, to compare their judgments as to which classes are correctly assigned. Also, in order to put the evaluation of classification into a broader context, a user study based on different information-seeking tasks would be valuable.

Other related issues of further interest include:

- Determining significance of other elements, such as anchor text, location at the beginning of the document versus location at the end, etc.;
- Comparing the results with new versions of the web pages in the collection, e.g., maybe the quality of titles improves with time, and structural tags or metadata are less misused etc.; and,
- Experimenting with other web page collections.

Acknowledgments

The research was funded by ALVIS, an EU Sixth Framework Programme, Information Society Technologies (IST-1-002068-STP), and The Swedish Agency for Innovation Systems (P22504-1 A).

References

Ardö, A., and Koch, T. (1999), "Automatic classification applied to the full-text Internet documents in a robot-generated subject index", *Proceedings of the 23rd International Online Information Meeting, London*, pp. 239-246.

Attardi, G., Gullì, A., and Sebastiani, F. (1999), "Automatic web page categorization by link and context analysis", *Proceedings of the European Symposium on Telematics, Hypermedia and Artificial Intelligence*, pp. 105-119.

Blum, A., and Mitchell, T. (1998), "Combining labeled and unlabeled data with co-training", *COLT: Proceedings of the Workshop on Computational Learning Theory*, pp. 99-100.

Ceci, M., and Malerba, D. (2003), "Hierarchical classification of HTML documents with WebClassII", *ECIR*, pp. 57-72.

DESIRE (2000), "DESIRE: Development of a European Service for Information on Research and Education", DESIRE, available at: http://www.desire.org/ (accessed 22 December 2004).

EELS (2003), *'All' Engineering resources on the Internet: a companion service to EELS*, EELS, Engineering E-Library, Sweden, available at: http://eels.lub.lu.se/ae/ (accessed 22 December 2004).

Engineering Electronic Library (2003), Lund University Libraries, available at: http://eels.lub.lu.se/ (accessed 22 December 2004).

Fisher, M., and Everson R. (2003), "When are links useful? Experiments in text classification", *Proceedings of the 25th European Conference on Information Retrieval, Pisa, IT,* pp. 41-56.

Fürnkranz, J. (2002), "Hyperlink ensembles: a case study in hypertext classification", *Information Fusion*, Vol.3 No. 4, pp. 299-312.

Ghani, R., Slattery, S., and Yang, Y. (2001), "Hypertext Categorization using Hyperlink Patterns and Metadata", *Proceedings of the 18th International Conference on Machine Learning*, pp. 178-185.

Glover, E.J. et al. (2002), "Using Web structure for classifying and describing web pages", *Proceedings of the Eleventh International Conference on World Wide Web Honolulu, Hawaii, USA,* pp. 562-569.

Jain, A.K., Murty, M.N., and Flynn, P.J. (1999), "Data clustering: a review", *ACM Computing Surveys*, Vol. 31 No. 3, pp. 264-323.

Koch, T., and Ardö, A. (2000), "Automatic classification of full-text HTML-documents from one specific subject area", *DESIRE II D3.6a* available at: http://www.lub.lu.se/desire/DESIRE36a-WP2.html (accessed 13 January 2006).

Kolcz, A., Prabakarmurthi, V., Kalita, J., and Alspector, J. (2001), "Summarization as feature selection for text categorization", *Proceedings of the Tenth International Information and Knowledge Management,* pp. 365-370.

Olson, H.A., and Boll, J.J. (2001), *Subject analysis in online catalogs*, 2nd ed., Libraries Unlimited, Englewood, CO.

Pierre, J.M. (2001), "On the automated classification of web sites", *Linköping Electronic Articles in Computer and Information Science*, Vol. 6 No. 001.

Sebastiani, F. (2002), "Machine learning in automated text categorization", *ACM Computing Surveys*, Vol. 34 No. 1, pp. 1-47.

Slattery, S., and Craven, M. (2000), "Discovering test set regularities in relational domains", *Proceedings of the 17th International Conference on Machine Learning*, pp. 895-902.

Svenonius, E. (2000), *The intellectual foundations of information organization,* MIT Press, Cambridge, MA.

Tudhope, D., Taylor C. (1997), "Navigation via similarity: automatic linking based on semantic closeness", *Information Processing and Management*, Vol. 33 No. 2, pp. 233-242.

Yang, Y. (1999), "An evaluation of statistical approaches to text categorization", *Journal of Information Retrieval*, Vol. 1 No. 1/2, pp. 67-88.

Yang, Y., Slattery, S., and Ghani, R. (2002), "A study of approaches to hypertext categorization", *Journal of Intelligent Information Systems*, Vol. 8 Nos 2-3, pp. 219-24.

Paper V

PAPER V.
The Role of Different Thesaurus Terms and Captions in Automated Subject Classification

Abstract

The paper aims to explore to what degree different types of terms in the Engineering Information (Ei) thesaurus and classification scheme influence automated subject classification performance. Preferred terms, their synonyms, broader, narrower, related terms, and captions are examined in combination with a stemmer and a stop-word list. The algorithm comprises string-to-string matching between words in the documents to be classified and words in term lists derived from the Ei thesaurus and classification scheme. The document collection for evaluation consists of some 35,000 scientific paper abstracts from the Compendex database. A subset of the Ei thesaurus and classification scheme is used, comprising 92 classes at up to five hierarchical levels from General Engineering. The results show that preferred terms perform best, whereas captions perform worst. Stemming in most cases shows to improve performance, whereas the stop-word list does not have a significant impact.

Golub, K. (2006), "The role of different thesaurus terms in automated subject classification of text", *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence, Hong Kong, China, 18-22 December,* pp. 961-965.

1 Introduction

Automated subject classification (in further text: automated classification) denotes machine-based organization of information objects into topically related groups. Automated classification has been a challenging research issue for several decades now. The importance of controlled vocabularies such as thesauri in automated classification has been recognized in recent research (Koch and Ardö 2000; Bang et al. 2006; Medelyan and Witten 2006; Garcés et al. 2006).

Vocabulary control in thesauri is achieved in several ways, out of which the following are beneficial for automated classification:

- The terms are usually noun phrases, which are content words;
- The meaning of the term is restricted to that most effective for the purposes of a particular thesaurus, which is indicated by the addition of scope notes and definitions, providing additional context for automated classification;
- Three main types of relationships are displayed in a thesaurus: 1) equivalence (synonyms, lexical variants); 2) hierarchical (generic, whole-part or instance relationships); 3) associative (terms that are closely related conceptually but not hierarchically and are not members of an equivalence set). In automated classification, equivalence terms allow for discovering concepts and not just words expressing them. Hierarchies provide additional context for determining the correct sense of a term, and so do associative relationships.

The purpose of the paper is to explore to what degree different types of terms in the Ei (Engineering Information) thesaurus and classification scheme (Milstead 1995) influence classification performance. Preferred terms, their synonyms, related, broader, narrower terms and captions are examined in combination with a stemmer and a stop-word list. The study would imply which terms with which weights to use in classification.

2 Methodology

2.1 String-matching algorithm

The algorithm searches for terms from the Ei thesaurus and classification scheme in documents to be classified. In order to do this, a term list is created, containing class captions, different thesauri terms and classes which the terms and captions denote. The list consists of triplets: term (single word, Boolean term or phrase), class which the term designates or maps to, and weight. Boolean terms consist of words that must all be present but in any order or in any distance from each other. Boolean terms in this form were not explicitly part of Ei, but were created to our purpose. They were considered to be those terms which in Ei contained the following strings: *and*, *vs.* (short for *versus*), *,* (comma), *;* (semi-colon, separating different concepts in class captions), *(* and *)* (parentheses, indicating the context of a homonym), *:* (colon, indicating a more specific description of the previous term in a class captions), and -- (double dash, indicating *heading--subheading* relationship). Upper-case words from the Ei thesaurus and classification scheme are left in upper case in the term list, assuming that they are acronyms. All other words containing at least one lower-case letter are converted into lower case. Geographical names are excluded on the grounds that they are not being engineering-specific in any sense.

The following is an excerpt from the Ei thesaurus and classification scheme, based on which the excerpt from the term list (further below) was created:

From the classification scheme (captions):
931.2 Physical Properties of Gases, Liquids and Solids
…
942.1 Electric and Electronic Instruments
…
943.2 Mechanical Variables Measurements

From the thesaurus:
TM Amperometric sensors
UF Sensors--Amperometric measurements
MC 942.1
…
TM Angle measurement
UF Angular measurement
UF Mechanical variables measurement--Angles
BT Spatial variables measurement
RT Micrometers

MC 943.2
…
TM Anisotropy
NT Magnetic anisotropy
MC 931.2

TM stands for the preferred term, UF for synonym, BT for broader term, RT for related term, NT for narrower term; MC represents the main class. Below is an excerpt from the All term list (see 2.3.), as based on the above examples:

1: electric @and electronic instruments =942.1,
1: mechanical variables measurements =943.2,
1: physical properties of gases @and liquids @and solids =931.2,
1: amperometric sensors =942.1,
1: sensors @and amperometric measurements =942.1,
1: angle measurement =943.2,
1: angular measurement =943.2,
1: mechanical variables measurement @and angles =943.2,
1: spatial variables measurement =943.2,
1: micrometers =943.2,
1: anisotropy =931.2,
1: magnetic anisotropy =913.2

The algorithm looks for strings from a given term list in the document to be classified and if the string (e.g., *magnetic anisotropy* from the above list) is found, the class(es) assigned to that string in the term list (*913.2* in our example) are assigned to the document. One class can be designated by many terms, and each time the class is found, the corresponding weight (*1* in our example) is assigned to the class. The scores for each class are summed up and classes with scores above a certain cut-off (heuristically defined) can be selected as the final ones for that document. In this particular study the weight is always *1* and all the classes are assigned as the final ones. Both weights and cut-offs will be dealt with in further research, based on the results of this study.

2.2 Document collection

Document collection consists of a subset of 35,166 paper titles and abstracts from the Compendex database (Engineering Information 2006), classified in the 92 selected Ei classes (see 2.3. Term lists). On average, 2.2 classes per document have been intellectually assigned (by people who are experts in the subject and in indexing).

Compendex is a commercial database so the subset cannot be made available to others. However, the authors can provide records' identification numbers on request.

2.3 Term lists

The Ei classification scheme is organized into six categories which are divided into 38 subjects, which are further subdivided into 182 specific subject areas. These are further subdivided, resulting in some 800 individual classes in a five-level hierarchy.

For this study one of the six main classes was selected, together with all its subclasses: class 900 – Engineering, General. The reason for choosing this class is that it contains both natural sciences such as engineering physics, and social sciences such as engineering profession and engineering management. The latter tend to use more polysemic words than the former, and as such present a more complex challenge for automated classification.

Within the 900 main class, there are 99 subclasses, but since for seven of them the number of documents in Compendex was few (less than 100), it was decided to exclude those seven classes from the study altogether.

The table below (Table 1) shows how many different types of terms there are in the 92 classes (Total), and the average number of terms per class (Avg./class).

Table 1. The number of different types of terms.

	All	BT	Ca	NT	PT	RT	ST
Total	8099	932	92	1423	1691	4378	1739
Avg./class	88	10	1	15	18	48	19

For the study, seven different term lists were created, each containing the following types of terms:

1) **All:** captions, preferred terms, synonyms, related, narrower and broader terms; 8099 entries.
2) **Broader (BT):** broader terms; 932 entries.
3) **Captions (Ca):** captions[1]; 92 entries.

[1] "Caption" stands for the term translating the class notation (e.g., class notation "900" has caption "Engineering, General").

4) **Narrower (NT):** narrower terms; 1423 entries.
5) **Preferred (PT):** preferred terms; 1691 entries.
6) **Related (RT):** related terms; 4378 entries.
7) **Synonyms (ST):** non-preferred terms; 1739 entries.

2.4 Stop-word list and stemming

Terms and captions in the Ei thesaurus and classification scheme can also contain words which are frequently used in many contexts and as such are not very indicative of any document's topicality (e.g., word *general* in the Ei class caption *Engineering, General*). The stop-word list used contained 429 words, and was taken from Onix text retrieval toolkit (Onix text retrieval toolkit). For stemming, the Porter Stemming Algorithm was used (Porter 1980). The stop-word list was applied to the term lists, and stemming to both the term lists and documents.

2.5 Evaluation methodology

Assuming that intellectually assigned classes in the document collection are correct, evaluation in this study is based on comparison of automatically derived classes against the intellectually assigned ones. The subset of the Ei thesaurus and classification scheme used in the experiment comprises 92 classes at five hierarchical levels. These 92 classes are all related to each other – often there is only a small topical difference between them. The topical relatedness is expressed in numbers representing the classes – the more initial digits any two classes have in common, the more related they are. Thus, comparing the classes at only the first few digits instead of all the five (each representing one hierarchical level), would also make sense. Still, the evaluation in this study is conducted based on all the five different levels, i.e., an automatically assigned class is considered correct only when it is exactly the same as an intellectually assigned class for the same document.

Apart from the standard micro-averaged and macro-averaged precision, recall and F1 measures (Sebastiani 2002, 33), the results are compared based on the number of documents that were assigned at least one class (Clas. doc. in Tables 2, 3 and 4), and the average number of classes assigned to each document (Avg. nbr. clas. in Tables 2, 3 and 4). There are about 2.2 classes intellectually

assigned per document, and the aim of automated classification is to achieve similar.

3 Experimental results

3.1 Averaged results for all the classes

Table 2. No stop-word list, no stemming.

	All	BT	Ca	NT	PT	RT	ST
Clas. doc. %	**96.8**	85.9	16.6	56.3	74.2	94.6	39.4
Avg. clas. nbr.	16.1	6.8	0.2	1.0	**1.7**	11.2	0.7
Macroa. P	0.11	0.11	0.43	0.29	**0.48**	0.12	0.37
Macroa. R	**0.54**	0.24	0.05	0.07	0.22	0.37	0.11
Microa. P	0.07	0.08	**0.43**	0.21	0.30	0.08	0.33
Microa. R	**0.54**	0.26	0.04	0.10	0.23	0.41	0.10
Macroa. F1	0.15	0.10	0.06	0.08	**0.22**	0.13	0.12
Microa. F1	0.13	0.12	0.07	0.13	**0.26**	0.13	0.15
Avg. F1s	0.14	0.11	0.07	0.10	**0.24**	0.13	0.14

As seen from Tables 2, 3 and 4, the best performance measured as mean F1s (Avg. F1s) has the Preferred term list, and the worst one the Captions list. As seen from Table 3, and by comparing the mean F1s of Tables 2 and 3, stemming showed to be beneficial in four out of the seven different term lists: Captions, Narrower, Preferred, and Synonyms. In All and Related lists the F1 performance became worse due to too lowered precision. Table 4 shows impact of the stop-word list, and in comparison to Table 2, the mean of F1s improved for Narrower and Preferred terms. For other terms the stop-word list did not do much of a difference since the thesaurus contains only content words, as seen from the last two rows in Table 4. The last row (Stop-w. %) shows the percentage of stop-words in all the terms on a list. In all lists apart from the shortest list (Captions), less than 10% are stop-words. Captions list has only 92 terms (Table 1) but the terms in this list are mostly longer than terms in other lists. This is due to the fact that captions' original function in a classification scheme is to describe well what the corresponding class notation stands for, and not to be a distinct term.

Table 3. No stop-word list, stemming.

	All	BT	Ca	NT	PT	RT	ST
Clas. doc. %	**99.4**	97.2	28.6	87.3	95.6	99.1	71.3
Avg. nbr. clas.	28.3	12.8	0.4	**2.6**	4.2	19.9	1.6
Macroa. P	0.09	0.09	**0.42**	0.27	0.40	0.10	0.33
Macroa. R	**0.72**	0.38	0.07	0.14	0.36	0.54	0.16
Microa. P	0.06	0.06	**0.36**	0.15	0.20	0.07	0.22
Microa. R	**0.73**	0.38	0.06	0.19	0.38	0.59	0.16
Macroa. F1	0.13	0.10	0.08	0.11	**0.27**	0.13	0.15
Microa. F1	0.10	0.11	0.10	0.17	**0.26**	0.12	0.18
Avg. F1s	0.11	0.11	0.09	0.14	**0.26**	0.12	0.17

Table 4. Stop-word list, no stemming.

	All	BT	Ca	NT	PT	RT	ST
Clas. doc. %	**97.8**	86.5	16.5	70.0	81.1	95.6	44.6
Avg. nbr. clas.	17.5	7.1	0.2	1.4	**2.1**	12.2	0.9
Macroa. P	0.11	0.11	0.42	0.30	**0.47**	0.12	0.36
Macroa. R	**0.56**	0.24	0.05	0.08	0.23	0.38	0.12
Microa. P	0.07	0.08	**0.42**	0.22	0.29	0.08	0.27
Microa. R	**0.59**	0.26	0.04	0.14	0.28	0.42	0.11
Macroa. F1	0.15	0.10	0.06	0.09	**0.22**	0.13	0.13
Microa. F1	0.13	0.12	0.07	0.17	**0.28**	0.13	0.16
Avg. F1s	0.14	0.11	0.07	0.13	**0.25**	0.13	0.14
Stop-w. nbr.	473	39	13	131	156	259	101
Stop-w. %	5.8	4.2	14.1	9.2	9.2	5.9	1.1

Concerning the number of classes per document that get automatically assigned, when using Captions less than one class is assigned on average even when stemming is applied; Narrower and Synonyms improve with stemming, close to our aim of 2.2 classes that have been intellectually assigned. The most appropriate number of classes are assigned when Preferred terms are used with stop-words. Based on All, Broader and Related lists, too many classes get assigned, but that could be dealt with in the future by introducing cut-offs (see last paragraph of 2.1.).

The results are similar for the number of documents that become classified: in all the three tables the lowest number of documents are classified using Captions (less than 30% even when stemming is applied), then using Synonyms and Narrower terms in

their best case (71% and 87% respectively). When looking at Table 1, such results could be expected for Captions since only one term designates a class. On the other hand, Preferred and Synonyms lists have similar number of terms, but Preferred performs almost twice as good when stemming is not applied. This reflects the fact that preferred terms in the Ei thesaurus occur more frequently in the documents than their synonyms.

By comparing results in Table 5 with numbers of terms in each term list (Table 1), we can see that only a certain number of terms are found in the documents being classified, ranging from 29% for All (when no stop-words and no stemming are used), to 74% for Captions (when stemming is used).

Table 5. Number of found terms from term lists.

	All	BT	Ca	NT	PT	RT	ST
No stop-words, no stemming	2348	387	62	651	823	1453	553
No stop-words, stemming	2730	413	68	789	960	1632	701
Stop-words, no stemming	2359	385	62	657	823	1455	558

3.2 Individual classes

As seen earlier from Table 1, each class is on average designated by 88 terms, ranging from 1 to 756 terms per class. The majority of terms are related terms, followed by synonyms and preferred terms.

Table 6 lists top performing classes using the All term list, no stemming and no stop-word list (their F1, either micro-averaged or macro-averaged, is above 0.30). The table also shows the number of each type of terms per class. We can see that the sole number of terms designating a class does not seem to be proportional to the performance. Moreover, these best-performing classes do not have a similar distribution of types of terms designating them, i.e., the percentage of certain term types does not seem to be directly related to performance.

Table 6. Top performing classes and number of terms.

Class	All	BT	Ca	NT	PT	RT	ST	F1
941.1	24	4	1	1	3	12	3	0.32
933.1.1	135	13	1	7	21	54	39	0.32
931.3	510	42	1	52	93	234	88	0.33
921.5	58	6	1	8	9	25	9	0.33
932.2.1	50	3	1	4	9	16	17	0.35
944.4	11	1	1	0	1	8	0	0.38
903	28	1	1	3	5	17	1	0.38
933.3	5	0	1	0	1	2	1	0.39
903.3	54	7	1	4	9	17	16	0.42
913.4.3	1	0	1	0	0	0	0	0.54
Total	876	77	10	79	151	385	174	n/a
Avg./class	88	8	1	8	15	39	17	n/a

Table 7. Worst-performing classes and number of terms.

Class	All	BT	Ca	NT	PT	RT	ST	F1
911.5	13	2	1	0	2	6	2	0.02
941.4	88	10	1	9	18	38	12	0.02
913	31	5	1	1	6	18	0	0.02
913.3.1	25	1	1	0	2	19	2	0.02
941	126	12	1	18	18	55	22	0.03
922	34	4	1	3	6	14	6	0.03
912.1	185	9	1	42	19	96	18	0.03
933	185	19	1	38	24	86	17	0.03
943.3	360	43	1	41	73	141	61	0.03
911.3	17	2	1	0	2	8	4	0.03
Total	1064	107	10	152	170	481	144	n/a
Avg./class	106	11	1	15	17	48	14	n/a

Table 7 lists worst-performing classes using the All term list, no stemming and no stop-word list (their F1, either micro-averaged or macro-averaged, is 0.03 or less). As it is the case with the best-performing classes, the worst-performing classes do not have a similar number of classes designating them, neither do they have a similar distribution of types of terms designating them.

Table 8 compares performance of the same worst-performing classes as Table 7, in regards to involving or not the stop-word list and stemming. The differences are very small, but stemming has a negative effect in 8 out of 10 cases, whereas stop-word list improves in two cases and worsens in two others.

Table 8. F1 of worst-performing classes in relation to stemming and stop-words.

	F1		
Class	No stemming, no stop-word list	Stemming, no stop-word list	No stemming, stop-word list
911.5	0.017	0.011	0.017
941.4	0.022	0.022	0.022
913	0.022	0.021	0.023
913.3.1	0.023	0.016	0.030
941	0.025	0.021	0.020
922	0.027	0.019	0.027
912.1	0.027	0.020	0.027
933	0.028	0.021	0.025
943.3	0.029	0.029	0.029
911.3	0.030	0.028	0.030

4 Concluding remarks

The majority of classes is found when using the All term list and stemming: micro-averaged recall is 73% (Table 3). The remaining 27% of classes were not found because the words in the term list designating the classes did not exist in the text of the documents to be classified.

In the study, no weighting or cut-offs were applied, but will be experimented with in the future. This study implies that all types of terms should be used for a term list (All) in order to achieve best recall, but that higher weights could be given to preferred terms, captions and synonyms, as the latter yield highest precision. Stemming seems useful for achieving higher recall, and could be balanced by introducing weights for stemmed terms. Stop-word list could be applied to captions, narrower and preferred terms.

Acknowledgments

This work was supported by the IST Programme of the European Community under ALVIS (IST-1-002068-STP). The author thanks Anders Ardö whose detailed comments helped improve the paper.

References

Bang, S.L., Yang, J.D., and Yang, H.J. (2006) "Hierarchical document categorization with k-NN and concept-based thesauri", *Information Processing and Management*, Vol. 42, pp. 387-406.

Engineering Information (2006), "Compendex", Engineering Information, Elsevier, available at: http://www.ei.org/databases/compendex.html (accessed 30 June 2006).

Garcés, P.J., Olivas, J.A., and Romero, F.P. (2006), "Concept-matching IR systems versus word-matching information retrieval systems: considering fuzzy interrelations for indexing Web pages", *Journal of the American Society for Information Science and Technology,* Vol. 57 No. 4, pp. 564–576.

Koch, T., and Ardö, A. (2000), "Automatic classification", *DESIRE II D3.6a, Overview of results),* available at: http://www.lub.lu.se/desire/ DESIRE36a-overview.html (accessed 22 December 2004).

Medelyan, O., and Witten, I. (2006), "Thesaurus based automatic keyphrase indexing", *Proceedings of Joint Conference on Digital Libraries*, pp. 296-297.

Milstead, J, ed. (1995), *Ei thesaurus,* 2nd ed., Engineering Information Inc., Hoboken, NJ.

"Onix text retrieval toolkit: Stop word list 1", available at: http://www.lextek.com/manuals/onix/stopwords1.html (accessed 30 June 2006).

Porter, M.F. (1980), "An algorithm for suffix stripping", *Program,* Vol. 14 No. 3, pp. 130-137.

Sebastiani, F. (2002), "Machine learning in automated text categorization", *ACM Computing Surveys*, Vol. 34 No. 1, pp. 1-47.

Paper VI

PAPER VI.
Automated Classification of Textual Documents Based on a Controlled Vocabulary in Engineering

Abstract

We explored a string-matching algorithm based on a controlled vocabulary, which does not require training documents – instead it reuses the intellectual work put into creating the controlled vocabulary. Terms from the Engineering Information thesaurus and classification scheme were matched against title and abstract of engineering papers from the Compendex database. Simple string-matching was enhanced by several methods such as term weighting schemes and cut-offs, exclusion of certain terms, and enrichment of the controlled vocabulary with automatically extracted terms. The best results are 76% recall when the controlled vocabulary is enriched with new terms, and 79% precision when certain terms are excluded. Precision of individual classes is up to 98%. These results are comparable to state-of-the-art machine-learning algorithms.

Golub, K., Hamon, T., and Ardö, A. (2007), "Automated classification of textual documents based on a controlled vocabulary in engineering", submitted to *Knowledge Organization*.

1 Introduction

Subject classification is organization of objects into topically related groups and establishing relationships between them. In automated subject classification (in further text: automated classification) human intellectual processes are replaced by, for example, statistical and computational linguistics techniques. Automated classification of textual documents has been a challenging research issue for several decades. Its relevance is rapidly growing with the advancement of the World Wide Web. Due to high costs of human-based subject classification and the ever-increasing number of documents, there is a danger that recognized objectives of bibliographic systems (Svenonius 2000, 20-21) would be left behind; automated means could provide a solution to preserve them (30).

Automated classification of text has many different applications (see Sebastiani 2002 and Jain et al. 1999); in this paper, the application context is that of information retrieval. In information retrieval systems, e.g., library catalogues or indexing and abstracting services, improved precision and recall are achieved by controlled vocabularies, such as classification schemes and thesauri. The specific aim of the classification algorithm is to provide a hierarchical browsing interface to a document collection, through a classification scheme.

In our opinion, one can distinguish between three major approaches to automated classification: text categorization, document clustering, and document classification (Golub 2006a).

In document clustering, both subject clusters or classes into which documents are classified and, to a limited degree, relationships between them are automatically produced. Labeling the clusters is a major research problem, with relationships between them, such as those of equivalence, related-term and hierarchical relationships, being even more difficult to automatically derive (Svenonius 2000, 168). In addition, "[a]utomatically-derived structures often result in heterogeneous criteria for category membership and can be difficult to understand" (Chen and Dumais 2000, 146). Also, clusters' labels and relationships between them change as new documents are added to the collection; unstable class

names and relationships are in information retrieval systems user-unfriendly, especially when used for subject browsing.

Text categorization (machine learning) is the most widespread approach to automated classification of text. Here characteristics of subject classes, into which documents are to be classified, are learnt from documents with human-assigned classes. However, human-classified documents are often unavailable in many subject areas, for different document types or for different user groups. If one would judge by the standard Reuters Corpus Volume 1 collection (RCV1) (Lewis et al. 2004), some 8,000 training and testing documents would be needed per class. A related problem is that the algorithm performs well on new documents only if they are similar enough to the training documents. The issue of document collections was also pointed out by Yang (1999) who showed how certain versions of one and the same document collection had a strong impact on performance.

In document classification, matching is conducted between a controlled vocabulary and text of documents to be classified. A major advantage of this approach is that it does not require training documents. If using a well-developed classification scheme, it will also be suitable for subject browsing in information retrieval systems. This would be less the case with automatically-developed classes and structures of document clustering or home-grown directories not created in compliance with professional principles and standards. Apart from improved information retrieval, another motivation to apply controlled vocabularies in automated classification is to reuse the intellectual effort that has gone into creating such a controlled vocabulary (see also Svenonius 1997).

The importance of controlled vocabularies such as thesauri in automated classification has been recognized in recent research. Bang et al. (2006) used a thesaurus to improve performance of a k-NN classifier and managed to improve precision by 14%, without degrading recall. Medelyan and Witten (2006) showed how information from a subject-specific thesaurus improved performance of keyphrase extraction by more than 1.5 times in F1, precision, and recall.

The overall purpose of this experiment is to gain insights into what degree a good controlled vocabulary such as Engineering Information thesaurus and classification scheme (Milstead 1995) (in

further text: Ei controlled vocabulary) could be used in automated classification of text, using string-matching. Vocabulary control in thesauri is achieved in several ways (Aitchinson et al. 2000). We believe that the following could be beneficial in the process of automated classification:

- Terms in thesauri are usually noun phrases, which are content words;
- Three main types of relationships are displayed in a thesaurus:
  1) equivalence (e.g., synonyms, lexical variants);
  2) hierarchical (e.g., generic, whole-part, instance relationships);
  3) associative (terms that are closely related conceptually but not hierarchically and are not members of an equivalence set).

  In automated classification, equivalence terms could allow for discovering concepts and not just terms expressing the concepts. Hierarchies could provide additional context for determining the correct meaning of a term; and so could associative relationships;
- When a term has more than one meaning in the thesaurus, each meaning is indicated by the addition of scope notes and definitions, providing additional context for automated classification.

In a previous paper (Golub 2006b) it was explored to what degree different types of Ei thesaurus terms and Ei classification captions influence performance of automated classification. In short, the algorithm searched for terms from the Ei controlled vocabulary in engineering documents to be classified (see 2.1). The majority of classes were found when using all the types of terms: preferred terms, their synonyms, related, broader, narrower terms and captions, in combination with a stemmer: recall was 73%. The remaining 27% of classes were not found because the words in the term list designating the classes did not exist in the text of the documents to be classified. No weighting or cut-offs were applied in the experiment. Apart from showing that all those types of terms should be used for a term list in order to achieve best recall, it was also indicated that higher weights could be given to preferred terms (from the thesaurus), captions (from the classification scheme) and

synonyms (from the thesaurus), as those three types of terms yielded highest precision.

The aim of this experiment is to improve the classification algorithm based on string-matching between the Ei controlled vocabulary and engineering documents to be classified. We especially wanted to do the following:

1) Achieve precision levels similar to recall achieved in a previous experiment (Golub 2006b, 964) by applying different weights and cut-offs.
2) Increase levels of recall to more than those achieved in the previous experiment by adding new terms extracted using natural language processing methods such as multi-word morpho-syntactic analysis and synonym extraction.

The paper is structured as follows: the next section, (2 Methodology) describes the applied string-matching classification algorithm, document collection and the evaluation methodology. The third section (3 Improving the string-matching algorithm) describes methods for enhancement of the string-matching algorithm, including the enrichment with automatically extracted terms. In fourth section, (4 Results) analyzes and discusses the results. Major conclusions and implications for further research are presented in the fifth section (5 Conclusion).

2 Methodology

2.1 String-matching algorithm

This section describes the classification algorithm used in the experiment. It is based on searching for terms from the Ei controlled vocabulary, in the field of engineering, in text of documents to be classified (also in the field of engineering). The Ei controlled vocabulary consists of two parts: a thesaurus of engineering terms, and a hierarchical classification scheme of engineering topics. These two controlled vocabulary types have each traditionally had distinct functions: the thesaurus has been used to describe a document with as many controlled terms as possible, while the classification scheme has been used to group similar documents together to the purpose of shelving them and allowing systematic browsing. The

aim of the algorithm was to classify documents into classes of the Ei classification scheme in order to provide a browsing interface to the document collection. A major advantage of Ei is that thesaurus descriptors are mapped to classes of the classification scheme. These mappings have been made manually (intellectually) and are an integral part of the thesaurus. Compared with captions[1] alone, mapped thesaurus terms provide a rich additional vocabulary for every class: instead of having only one term per class (there is only one caption per class), in our experiment there were on average 88 terms per class.

Pre-processing steps of Ei included normalizing upper- and lower-case words. Upper-case words were left in upper case in the term list, assuming that they were acronyms; all other words containing at least one lower-case letter were converted into lower case. The first major step in designing the algorithm was to extract terms from Ei into what we call a term list. It contained class captions, thesaurus terms (Term), classes to which the terms and captions map or denote (Class), and weight indicating how appropriate the term is for the class to which it maps or which it designates (Weight). Geographical names, all mapping to class 95, were excluded on the grounds that they are not engineering-specific. The term list was formed as an array of triplets:

Weight: **Term** (single word, Boolean term or phrase) = **Class**

Single-word terms were terms consisting of one word. *Boolean terms* were terms consisting of two or more words that must all be present but in any order or in any distance from each other. Boolean terms in this form were not explicitly part of Ei, but were created to our purpose. They were considered to be those terms which in Ei contained the following strings: *and*, *vs.* (short for *versus*), *,* (comma), *;* (semi-colon, separating different concepts in class captions), *(* and *)* (parentheses, indicating the context of a homonym), *:* (colon, indicating a more specific description of the previous term in a class captions), and -- (double dash, indicating *heading--subheading* relationship). These strings we replaced with *@and* which indicated the Boolean relation in the term. All other

[1] A caption is a class notation expressed in words, e.g., in the Ei classification scheme "Electric and Electronic Instruments" is the caption for class "942.1".

terms consisting of two or more words were treated as *phrases*, i.e., strings that need to be present in the document in the exact same order and form as in the term. Ei comprises a large portion of composite terms (3,474 in the total of 4,411 distinct terms in our experiment); as such, Ei provides a rich and precise vocabulary with the potential to reduce the risks of false hits.

The following are two excerpts from the Ei classification scheme and thesaurus, based on which the excerpt from the term list (further below) is created:

From the classification scheme:
931.2 Physical Properties of Gases, Liquids and Solids
...
942.1 Electric and Electronic Instruments
...
943.2 Mechanical Variables Measurements

From the thesaurus:
TM Amperometric sensors
UF Sensors--Amperometric measurements
MC 942.1
...
TM Angle measurement
UF Angular measurement
UF Mechanical variables measurement--Angles
BT Spatial variables measurement
RT Micrometers
MC 943.2
...
TM Anisotropy
NT Magnetic anisotropy
MC 931.2

All the different thesaurus terms as well as captions were added to the term list. Despite the fact that choosing all types of thesaurus terms might lead to precision losses, we decided to do just that in order to achieve maximum recall, as shown in a previous paper (Golub 2006b). In the thesaurus, TM stands for the preferred term, UF ("Used For") for an equivalent term, BT for broader term, RT for related term, NT for narrower term; MC represents the main class; sometimes there is also OC, which stands for optional class, valid only in certain cases. Main and optional classes are classes from the Ei classification scheme that have been made manually (intellectually) and are an integral part of the thesaurus. Based on the above excerpts, the following term list would be created:

1: physical properties of gases @and liquids @and solids = 931.2,
1: electric @and electronic instruments = 942.1,

1: mechanical variables measurements = 943.2,
1: amperometric sensors = 942.1,
1: sensors @and amperometric measurements = 942.1,
1: angle measurement = 943.2,
1: angular measurement = 943.2,
1: mechanical variables measurement @and angles = 943.2,
1: spatial variables measurement = 943.2,
1: micrometers = 943.2,
1: anisotropy = 931.2,
1: magnetic anisotropy = 931.2,

The number at the beginning of each triplet is weight estimating the probability that the term of the triplet designates the class; in this example it is set to 1 as a baseline, and experiments with different weights are discussed later on.

The algorithm searches for strings from a given term list in the document to be classified and if the string (e.g., *magnetic anisotropy* from the above list) is found, the class(es) assigned to that string in the term list (*931.2* in our example) are assigned to the document. One class can be designated by many terms, and each time a term is found, the corresponding weight (1 in our example) is added to a score for the class. The scores for each class are summed up and classes with scores above a certain cut-off (heuristically defined, discussed later on) are selected as the final ones for the document being classified.

The Ei classification scheme is hierarchical and consists of six main classes divided into 38 finer classes which are further subdivided into 182 classes. These are subdivided even further, resulting in some 800 individual classes in a five-level hierarchy. For this experiment one of the six main classes was selected, together with all its subclasses: class 9, *Engineering, General*. The reason for choosing this class was that it covers both natural sciences such as physics and mathematics, and social sciences fields such as engineering profession and management. The literature of the latter tends to contain more polysemic words than the former, and as such presents a more complex challenge for automated classification. Within the 9 class, there are 99 subclasses. However, for seven of them the number of documents in a database based on which the document collection was created (see 2.2 Document collection) were few, less than 100. Thus those seven classes were excluded from the experiment altogether. These were: 9 (Engineering, General), 902 (Engineering Graphics; Engineering Standards; Patents), 91 (Engineering Management), 914 (Safety

Engineering), 92 (Engineering Mathematics), 93 (Engineering Physics), and 94 (Instruments and Measurement). Of the remaining 92 classes, the distribution at the five different hierarchical levels is as follows: at the fifth hierarchical level 11 classes, at the fourth 67, at the third 14, and at the second hierarchical level 5.

2.2 Document collection

The document collection comprised 35,166 bibliographic records[2] from the Compendex database (Engineering Information 2006). The records were selected by simply retrieving the top 100 or more of them upon entering the class notation. A minimum of 100 records per class were downloaded at several different points in time during the years of 2005 and 2006.

For each record there was at least one of the 92 selected classes that were human-assigned (see 2.1). A subset of this collection was created to include only those records where main class[3] was class 9; this subset contained 19237 documents.

From each bibliographic record (in further text: document) the following elements were extracted: an identification number, title, abstract and human-assigned classes (*Ei classification codes*). Thesaurus descriptors (in Compendex called *Ei controlled terms*) were not extracted since the purpose of this experiment was to compare automatically assigned classes (and not descriptors) against the human-assigned ones. Below is an example of one document:

Identification number: 03337590709
Title: The concept of relevance in IR
Abstract: This article introduces the concept of relevance as viewed and applied in the context of IR evaluation, by presenting an overview of the multidimensional and dynamic nature of the concept. The literature on relevance reveals how the relevance concept, especially in regard to the multidimensionality of relevance, is many faceted, and does not just refer to the various relevance criteria users may apply in the process of judging relevance of retrieved information objects. From our point of view, the multidimensionality of relevance explains why some will argue that no consensus has been reached on the relevance concept. Thus, the objective of this article is to present an overview of the many different views and ways by which the concept of relevance is used - leading to a consistent and compatible understanding of the concept. In addition, special attention is paid to the type of situational relevance. Many researchers perceive situational relevance as the most realistic type of user relevance, and therefore situational relevance is discussed with reference to its

[2] Compendex being a commercial database, the document collection cannot be made available to others, but the authors are willing to provide documents' identification numbers on request.
[3] The first one listed in the *Ei classification codes* field of the record.

potential dynamic nature, and as a requirement for interactive information retrieval (IIR) evaluation.
Ei classification codes: 903.3 Information Retrieval & Use, 723.5 Computer Applications, 921 Applied Mathematics

Automated classification was based on title and abstract, and automatically assigned classes were compared against human-assigned ones (Ei classification codes in the example). On average, 2.2 classes per document were human-assigned, ranging from 10 to 1.

2.3 Evaluation methodology

2.3.1 Evaluation Challenge

According to ISO standard on methods for examining documents, determining their subjects, and selecting index terms (International Organization for Standardization 1985), human-based subject indexing is a process involving three steps: 1) determining subject content of a document, 2) conceptual analysis to decide which aspects of the content should be represented, and 3) translation of those concepts or aspects into a controlled vocabulary. These steps, in particular the second one, are based on a specific library's policy in respect to its document collections and user groups. Thus, when evaluating automatically assigned classes against the human-assigned ones, it is important to know the human-based indexing policies. Unfortunately, we were unable to obtain indexing policies applied in the Compendex database. What we could derive from the document collection was the number of human-assigned classes per document, which were used in evaluation. However, without a thorough qualitative analysis of automatically assigned classes one cannot be sure whether, for example, the classes assigned by the algorithm, but not human-assigned, are actually wrong, or if they were left out by mistake or because of the indexing policy. A further issue is that we did not know whether the articles had been human-classified based on their full-text or/and abstracts; we had, however, only abstracts.

Another problem to consider when evaluating automated classification is the fact that certain subjects are erroneously assigned. When indexing, people make errors such as those related to exhaustivity policy (too many or too few terms become assigned), specificity of indexing (which usually means that people do not

assign the most specific term), they may omit important terms, or assign an obviously incorrect term (Lancaster 2003, 86-87). In addition, it has been reported that different people, whether users or professional subject indexers, would assign different subject terms or classes to the same document. Studies on inter- and intra-indexer consistency report generally low indexer consistency (Olson and Boll 2001, 99-101). Markey (1984) reviewed 57 indexer consistency studies and reported that consistency levels range from 4% to 84%, with only 18 studies showing over 50% consistency. There are two main factors that seem to affect it:

1) Higher exhaustivity and specificity of subject indexing both lead to lower consistency, i.e., indexers choose the same first term for the major subject of the document, but the consistency decreases as they choose more classes or terms;
2) The bigger the vocabulary, or, the more choices the indexers have, the less likely will they choose the same classes or terms (Olson and Boll 2001, 99-101).

Both of these two factors were present in our experiment:

1) High exhaustivity: on average, 2.2 classes per document had been human-assigned, ranging from 10 to 1;
2) Ei controlled vocabulary is rather big (we chose 92 classes) and deep (five hierarchical levels), allowing many different choices.

Today evaluation in automated classification experiments is mostly conducted under controlled conditions, ignoring the above-discussed issues. As Sebastiani (2002, 32) puts it, "…the evaluation of document classifiers is typically conducted experimentally, rather than analytically. The reason is that… we would need a formal specification of the problem that the system is trying to solve (e.g., with respect to what correctness and completeness are defined), and the central notion… that of membership of a document in a category is, due to its subjective character, inherently nonformalizable." Because of the fact that methodology for such experiments has yet to be developed, as well as limited resources, we followed the common approach to evaluation and started from the assumption that human-assigned classes in the document collection were correct, and compared automatically assigned classes against them.

2.3.2 Evaluation measures

The subset of the Ei controlled vocabulary we used comprised 92 classes that are all topically related to each other. The topical relatedness is expressed in numbers representing the classes: the more initial digits any two classes have in common, the more related they are. For example, 933.1.2 for *Crystal Growth* is closely related to 933.1 for *Crystalline Solids*, both of which belong to 933 for *Solid State Physics*, and finally to 93 for *Engineering Physics*. Each digit represents one hierarchical level: class 933.1.2 is at the fifth hierarchical level, 933.1 at the fourth etc. Thus, comparing two classes at only first few digits (later referred to as partial matching) instead of all the five also makes sense. Still, unless specifically noted, the evaluation in this experiment was conducted based on all the five different levels (later referred to as complete matching), i.e., an automatically assigned class was considered correct only if all its digits were the same as a human-assigned class for the same document.

Evaluation measures used were the standard microaveraged and macroaveraged precision, recall and F1 (Sebastiani 2002, 40-41), for both complete and partial matching:

$$\text{Precision} = \frac{\text{correctly automatically assigned classes}}{\text{all automatically assigned classes}}$$

$$\text{Recall} = \frac{\text{correctly automatically assigned classes}}{\text{all human-assigned classes}}$$

$$F1 = 2*\text{Precision}*\text{Recall} / (\text{Precision} + \text{Recall})$$

In macroaveraging the results are first calculated for each class, and then summed and divided by the number of classes. In microaveraging the results for each part of every equation are summed up first (e.g., all correctly automatically assigned classes are added together, all automatically assigned classes are added together), and then the "aggregated" values are used in one equation. Equations for macroaveraged and microaveraged precision are given below:

$$\text{Precision}_{\text{macroaveraged}} = \frac{\text{sum of precision values for each class}}{\text{number of all classes}}$$

$Precision_{microaveraged}$ = sum of correct automated assignments for each class / sum of all automated assignments for each class

In microaveraging more value is given to classes that have a lot of instances of automatically assigned classes and the majority of them are correct, while in macroaveraging the same weight is given to each class, no matter if there are many or few automatically assigned instances of it. The differences between macroaveraged and microaveraged values can be large, but whether one is better than the other has not been agreed upon (Sebastiani 2002, 41-42). Thus, in this experiment, it is the mean macroaveraged and microaveraged F1 that is mostly used.

In order to examine different aspects of the automated classification performance, several other factors were also taken into consideration:

- Whether the (human-assigned) main class is found;
- The number of documents that got automatically assigned at least one class;
- Whether the class with highest score was the same as the human-assigned main class;
- The distribution of automatically versus human-assigned classes; and,
- The average number of classes assigned to each document. There were 2.2 human-assigned classes per document, and our aim was to achieve similar. In the context of hierarchical browsing based on a classification scheme, having too many classes assigned to a document would place one document to too many different places, which would create the opposite effect of the original purpose of a classification scheme, that of grouping similar documents together.

3 Improving the algorithm

The major aim of the experiment was to improve the algorithm that was previously experimented with in Golub 2006b, where highest (microaveraged) recall was 73% when all types of terms were included in the term list. In that experiment neither weights nor cut-

offs were experimented with, so all the classes that were found for a document were assigned to it. Here we wanted to achieve as high as possible precision levels by use of term weighting and class cut-offs. In order to also allow for better recall, the basic term list was enriched with new terms extracted from documents in the Compendex database, using multi-word morpho-syntactic analysis and synonym acquisition.

3.1 Term weights

The aim of this part of the experiment was to achieve as high as possible precision levels by use of weighting and cut-offs. As shown in Golub 2006b, all types of terms need to be used in the term list for maximum recall. Thus, all the different types of terms and their mappings to classes were merged into the final term list. This resulted in a number of duplicate cases which were dealt with in the following manner:

- If one term mapping to the same class was a caption, a preferred term, and a synonym at the same time, the highest preference was, based on their performance (see Table 4), given to captions, followed by preferred terms, followed by synonyms, while others were removed from the list;
- If one term mapping to both optional class (OC) and main class (MC) was a caption, a preferred term, and a synonym at the same time, the highest preference was, based on their performance (see Table 2), given to captions, followed by preferred terms, followed by synonyms, while others were removed from the list;
- If one thesaurus term of the same type mapped to both optional class (OC) and main class (MC), the one that mapped to the optional class was removed (based on their performance, see Table 2).

The final term list consisted of 8099 terms, out of which 92 were captions (all mapped to main class (MC)), 668 were broader terms, 729 narrower, 1653 preferred, 3224 related, and 1733 were synonym terms. This big number of terms that have been human-mapped to classes indicates potential usefulness of such a controlled

vocabulary in a string-matching algorithm for automated classification.

In order to systematically vary different parameters, the following 14 weighting schemes evolved:

1) **w1**: All terms in the term list were given the same weight, 1. This term list served as a baseline.

2) **w134**: Different term types were given different weights: single-word terms 1, phrases 3, and Boolean terms 4.

   These weights were heuristically derived in a separate experiment (Table 1). Three different term lists were created, each containing only single-word terms, phrases or Boolean terms. Weight 1 was assigned to all of them. The documents were classified using these three terms lists and their performance was compared for precision.

   Table 1. Single, phrase and Boolean term lists and their performance as a basis for weights.

	Single	Phrase	Boolean
Avg. precision (%)	8	26	33
Derived weight	1	3	4

   Avg. precision (%) is mean microaveraged and macroaveraged precision. Derived weights were based on dividing precision values (Avg. precision) by the lowest precision value (in this case 8).

3) **w12**: Terms mapping to a main class (MC) were given weight 2, and those mapping to an optional class (OC) were given weight 1.

   These weights were heuristically derived in a separate experiment (Table 2). Two different term lists were created, one containing only those terms that map to a main class, and another one containing only those terms that map to an optional class. Weight 1 was assigned to all of them. The documents were classified using these two terms lists and their performance was compared for precision.

Table 2. Main code and optional code term lists and their performance as a basis for weights.

	MC	OC
Avg. precision (%)	13	6
Derived weight	2	1

Avg. precision (%) is mean microaveraged and macroaveraged precision. Derived weights were based on dividing precision values (Avg. precision) by the lowest precision value (in this case 6).

4) **w134_12**: This list was a combination of the two preceding lists. Weights for term type 1, 3, and 4 for single, phrase or Boolean term were multiplied by the weight for the type of class to which the term mapped – 1 or 2 for optional or main class.

5) **wOrig**: As used in the original term weighting scheme when the string-matching algorithm based on Ei was first applied (Koch and Ardö 2000). These weights were intuitively derived. They combined types of terms depending if it were a single-word term, Boolean or phrase, and whether the assigned class was main (MC) or optional (OC).

Table 3. Weights in the original algorithm.

	Phrase	Boolean	Single
OC	4	2	1
MC	8	3	2

6) **w1234**: With weights for different Ei term type as experimented with in Golub 2006b (captions are from the classification scheme, all others are thesaurus terms).

Table 4. Different types of thesaurus terms captions and their performance as a basis for weights.

	Broader	Captions	Narrower	Preferred	Related	Synonyms
Avg. precision (%)	10	43	25	39	10	35
Derived weight	1	4	2	4	1	3

7) **w134_1234**: This list was a combination of two previous lists, w134 and w1234. Weights for term type 1, 3, and 4 for single, phrase or Boolean term were multiplied by the weight for the type of Ei term as given in Table 4.

8) **w134_12_1234**: This list was a combination of two previous lists, w134_12 and w1234. Weights for term type 1, 3, and 4 for single, phrase or Boolean term were multiplied by the weight for the type of class to which the term mapped – 1 or 2 for optional or main class, and by the weight for the type of Ei term as given in Table 4.

9) **wTf10**: In this list weights were based on the number of words the term consisted of, and of the number of times each of its words occurred in other terms (cf. *tf-idf*, term frequency – inverse document frequency, Salton and McGill 1983, 63, 205). If *f* were the frequency with which a word *w* from the term *t* occurred in other terms, term *t* consisting of *n* words, then the weight *weight* of that term was calculated as follows:

$$\text{weight}_t = \log(n) \cdot (1/f_{w1} + 1/f_{w2} + \ldots + 1/f_{wn})$$

Logarithm was applied in order to reduce the impact of parameter *n*, i.e., to avoid getting overly high weights for terms consisting of several sparse words. In order to get integers as weights, the weights were multiplied by 10, rounded and increased by 1 to avoid zeros.

10) **wTf10Boolean**: As in wTf10, with all the phrases modified into Boolean terms. This list was created in order to study the influence of phrases and Boolean terms on precision and recall.

11) **wTf10Phrases**: As in wTf10, with all the Boolean terms modified into phrases. This list was created in order to study the influence of phrases and Boolean terms on precision and recall.

12) **wTf10_12**: As in wTf10, with those weights multiplied by the weight for the type of class to which the term maps – 1 or 2 for optional or main class. The multiplication was done before the rounding.

13) **wTf10_1234**: As in wTf10, with those weights multiplied by the weight for the type of relationship (Table 4). The multiplication was done before the rounding.

14) **wTf10_12_1234**: As in wTf10_12, with those weights multiplied by the weight for the type of relationship (Table 4). The multiplication was done before the rounding.

3.1.1 Stop-word list and stemming

Although the terms and captions in the Ei controlled vocabulary are usually noun phrases which are good content words, they can also contain words which are frequently used in many contexts and as such are not very indicative of any document's topicality (e.g., word *general* in the Ei class caption *Engineering, General*). Thus, a stop-word list was used. It contained 429 such words, and was taken from Onix text retrieval toolkit (Onix text retrieval toolkit). For stemming, the Porter's algorithm (Porter 1980) was used. The stop-word list was applied to the term lists, and stemming to the term lists as well as documents.

3.2 Cut-offs

In a previous experiment (Golub 2006b) cut-offs were not used – instead, all the classes that were found for a document were assigned to it. In the context of hierarchical browsing based on a classification scheme, having too many classes assigned to a document would place one document to many different places, which would create the opposite effect of the original purpose of a classification scheme (grouping similar documents together). In the document collection, there were 2.2 human-assigned classes per document, and the aim of automated classification was to achieve similar. The effect of several different cut-offs was investigated:

1) All automatically derived classes are assigned as final ones (no cut-off).
2) In order to assign a certain class as final, the score of that class had to have a minimum percentage of the sum of all the classes' scores. Different values for the minimum percentage were

tested: 1, 5, 10, 15 and 20, as well as some others (see section 4 Results).

3) The second type of cut-off in combination with the rule that if there were no class with the required score, the one with the highest score would be assigned.
4) In order to follow the subject classification principle of always assigning the most specific class possible, the principle of score propagation was introduced. The principle was implemented so that the scores for classes at deeper hierarchical levels were a sum of their own score together with scores of classes at upper hierarchical levels if such were assigned.

3.3 Enriching the term list with new terms

In the previous experiment (Golub 2006b), highest achieved recall was 73% (microaveraged), when all types of terms were included in the term list. In order to further improve recall, the basic term list was enriched with new terms. These terms were extracted from bibliographic records of the Compendex database, using multi-word morpho-syntactic analysis and synonym acquisition, based on the existing preferred and synonymous terms (as they gave best precision results).

Multi-word morpho-syntactic analysis was conducted using a parser FASTER (Jacquemin 1996) which analyses raw technical texts and, based on built-in meta-rules, detects morpho-syntactic variants. The parser exploits morphological (derivational and inflectional) information as given by the database CELEX (Baayen et al. 1995). Morphological analysis was used to identify derivational variants, such as:

effect of gravity: gravitational effect
architectural design: design of the proposed architecture
supersonic flow: subsonic flow
structural analysis: analysis of the structure

Syntactical analysis was used to:

a) insert word inside a term, such as:

flow measurement: flow discharge measurements
distribution of good: distribution of the finished goods
construction equipment: construction related equipment
intelligent distributed control: intelligent control

b) permute components of a term, such as:

control of the inventory: inventory control
flow control: control of flow
development of a flexible software: software development

c) add a coordinated component to a term, such as:

project schedule and management: project management
control system: control and navigation system

Synonyms were acquired through a rule-based system SynoTerm (Hamon and Nazarenko 2001) which infers synonymy relations between complex terms by employing semantic information extracted from lexical resources. First the documents were preprocessed and tagged with part-of-speech information and lemmatized. Then terms were identified through the YaTeA term extractor (Aubin and Hamon 2006). The semantic information provided by the database WordNet (Fellbaum 1998; WordNet) was used as a bootstrap to acquire synonym terms of the basic terms. The synonymy of the complex candidate terms was assumed to be compositional, i.e., two terms were considered synonymous if their components were identical or synonymous (e.g., building components: construction components, building components: construction elements).

Although verification by a subject expert is desirable for all automatically derived terms, due to limited resources only the extracted synonyms were verified. Checking the synonyms is also most important since computing those leads to a bigger semantic shift than morphological and syntactical operations do. The verification was conducted by a subject expert, a fifth-year student of engineering physics. Suggested synonym terms were displayed in the user interface of SynoTerm. The verification was not strict: derived terms were kept if they were semantically related to the basic term. Thus, hyperonym (generic/specific) or meronym (part/whole) terms were also accepted as synonyms. The expert spent 10 hours validating the derived terms. Of the 292 automatically acquired synonyms, 168 (57.5%) were validated and used in the experiment.

4 Results

4.1 Weights and cut-offs

Based on each of the 14 term lists, the classification algorithm was run on the document collection of 35,166 documents (see 2.2). As described earlier (2.3.2), several aspects were evaluated and different evaluation measures were used; thus, for each term list, the following types of results were obtained:

1) **min 1**: if no classes were assigned because their final scores were below the pre-defined cut-off value (described in 3.2), the class with the highest score was assigned;
2) **cut-off**: the applied cut-off value;
3) **min 1 correct**: number of documents that were assigned at least one correct class;
4) **min 1 auto**: number of documents that were assigned at least one class;
5) **avg auto/doc**: average number of classes that were assigned per document, based on documents that were assigned at least one class;
6) **macroa P**: macroaveraged precision;
7) **macroa R**: macroaveraged recall;
8) **macroa F1**: macroaveraged F1;
9) **microa P**: microaveraged precision;
10) **microa R**: microaveraged recall;
11) **microa F1**: microaveraged F1;
12) **mean F1s**: arithmetic mean of macroaveraged and microaveraged F1 values.

Table 5 shows results for list w134_12_1234 which has combined weights for term type (single, phrase or Boolean), type of class, and Ei term type. In order to provide an example of how results for every other term list were analyzed, we discuss results for this list in detail.

Best recall is achieved when no cut-off is applied, 0.54, but in that case on average 17 classes are assigned per document, which is too many in comparison to 2.2 that are human-assigned. This setting is appropriate in applications such as focused crawling where documents are ranked based on weights. When the most appropriate

number of classes for our purpose is assigned (2.63), recall is 0.22. Best precision is gained when cut-off value is highest (20): 0.37 macroaveraged, 0.28 microaveraged. In that setting the average number of classes assigned per document is 1.5. Best mean macroaveraged and microaveraged F1 is 0.22, when cut-offs are 10 or 15.

Best precision results are gained when cut-off is highest, best recall when there is no cut-off. More than twice as many documents are assigned correct classes when no cut-off is used. All these results suggest that weights are not very appropriate. Still, when looking at the F1 values, in comparison to the baseline (first column), an improvement of six percent is achieved when using the w134_12_1234 term list.

Table 5. Results for term list w134_12_1234.

min 1	no						yes				
cut-off	0	1	5	10	15	20	1	5	10	15	20
min 1 correct	**24036**	21403	17403	14339	12320	10278	21403	17403	14425	12774	11606
min 1 auto	34053	34053	34050	33270	30433	26587	34053	34053	34053	34053	34053
avg auto/doc	16.65	9.77	5.02	2.69	1.91	1.47	9.46	4.86	**2.55**	1.65	1.11
macroa P	0.11	0.14	0.18	0.25	0.32	**0.37**	0.14	0.18	0.25	0.31	0.35
macroa R	**0.54**	0.42	0.29	0.21	0.17	0.14	0.42	0.29	0.22	0.18	0.15
macroa F1	0.19	0.21	0.22	**0.23**	0.22	0.20	0.21	0.22	**0.23**	**0.23**	0.21
microa P	0.07	0.10	0.13	0.19	0.24	**0.28**	0.10	0.13	0.19	0.23	0.27
microa R	**0.54**	0.43	0.30	0.22	0.18	0.14	0.43	0.30	0.22	0.18	0.16
microa F1	0.13	0.16	0.19	0.20	0.20	0.19	0.16	0.18	0.20	**0.21**	0.20
mean F1s	0.16	0.19	0.20	**0.22**	0.21	0.19	0.19	0.20	**0.22**	**0.22**	0.21

The same experiment was run on all the other term lists. When looking at mean F1 values, the differences between the term lists are not larger than four percent. Performance of the different lists measured in precision and recall is also similar. Three lists that perform best in terms of mean F1 are w1234, w134_1234 and w134_12_1234 – all of them based on weights for different Ei term types. The biggest number of documents with correct classes is found with the wTf10Boolean list in which all phrases were converted into Boolean terms.

When using cut-offs, two sets of experiments were conducted: one with assigning at least the class with highest score, and the other following the threshold calculation only. Because the former results in more documents with assigned correct classes, in further

experiments the rule to assign at least the class with highest score is applied.

4.1.1 Stop-words removal and stemming

Next, the influence of stop-words removal and stemming was tested (as described in 3.1.1). For this experiment three lists that performed best in the previous one were chosen: w1234, w134_1234 and w134_12_1234. Every list was run against stop-words removed, stemming, and both the stop-words removed and stemming, each in combination with different cut-off values: 5, 10 and 15.

Improvements when using either stemming or stop-words removal or both are achieved in majority of cases up to two percent. There is also a slight increase in the number of correctly found classes without finding more incorrect classes. The differences between the three term lists measured in mean F1 are minor – one or two percent. The best term list is w134_12_1234 used in combination with stemming and stop-words removal and cut-off 10 – best mean F1 is 0.24. For this list more cut-offs were experimented with for better results; the value of 9 proved to perform best but better only on a third decimal digit than that of 10.

4.1.2 Individual classes, partial matching, distribution of classes

We further wanted to investigate performance at the level of individual classes, partial matching as well as how automatically assigned classes are distributed in comparison to human-assigned ones. We used the best-performing w134_12_1234 term list and setting (applying stemming and stop-words removal, cut-off 9).

It was shown that certain classes perform much better than the average. Performance of different classes varies quite a lot. For example, top three performing classes as measured in precision are different from top three classes for recall or F1:

- Top three in precision:
  - Cellular Manufacturing (913.4.3), precision 0.98;
  - Electronic Structure of Solids (933.3), precision 0.97; and,
  - Information Retrieval and Use (903.3), precision 0.82.
- Top three in recall:
  - Amorphous Solids (933.2), recall 0.61;
  - Crystal Growth (933.1.2), recall 0.52; and,

  - Manufacturing (913.4), recall 0.50.
- Top three in F1:
  - Crystal Growth (933.1.2), F1 0.45;
  - Amorphous Solids (933.2), F1 0.44; and,
  - Optical Variables Measurement (941.1), F1 0.40.

As expected, the algorithm performs better when evaluation is based on partial matching between automatically and human-assigned classes. As seen from Table 6, at the second hierarchical level F1 is up to 0.66 and at third 0.59. At the second hierarchical level the best F1 is achieved by classes *Engineering mathematics* (represented by notation 92) and *General engineering* (90), both of which have by far the smallest number of terms designating them (*terms*), while other three classes have many more terms and similar performance measured in mean F1. At the third hierarchical level, the class that performs best of all is 921 *Applied Mathematics*, while the worst one is 943 *Mechanical and Miscellaneous Instruments*. In conclusion, for the 14 classes at top three hierarchical levels mean F1 is almost twice as good as for the complete matching, which implies that our classification approach would suit better those information systems in which fewer hierarchical levels are needed, like the Intute subject gateway on engineering (Intute Consortium 2006).

Table 6. Results for partial matching at the second and third hierarchical levels, and number of terms per each class.

	General			*Management*				*Maths*		*Physics*			*Instruments*			
	90			*91*				*92*		*93*			*94*			
F1	0.65			0.5				**0.66**		0.51			0.49			
terms	679			1922				848		2902			1748			
	901	*902*	*903*	*911*	*912*	*913*	*914*	*921*	*922*	*931*	*932*	*933*	*941*	*942*	*943*	*944*
F1	0.35	0.27	**0.53**	0.32	**0.36**	0.26	0.29	**0.59**	0.33	0.44	0.33	**0.48**	0.28	0.36	0.2	**0.44**
terms	275	241	163	237	596	393	696	628	220	1648	801	453	422	373	604	349

The variations in performance between individual classes for both complete and partial matching are quite big, but at this stage it is difficult to say why. Further research is needed to explore what the factors contributing to performance are.

Using the same best setting achieved so far, the algorithm was also evaluated for distribution of automatically assigned classes in comparison to that of the human-assigned ones. The comparison was based on how often two classes get assigned together when using the algorithm in comparison to when they get human-assigned. Figure 1 shows the frequency distribution of assigned class pairs. The x-coordinate presents human-assigned class pairs ordered by descending frequency. One point represents one class pair: e.g., the pair of classes 912.2 and 903 occurs most frequently in human-based classification (48 times, as marked on the y-coordinate) and is represented by point 1 on the x-coordinate; point 500 on the x-coordinate represents the 913.5 and 911 pair that occurs 3 times, as marked on the y-coordinate. Thus, the smoothest line (Human-assigned) represents the human-assigned classes. The minimum of 2538 pairs of classes that both the algorithm and people have produced are shown.

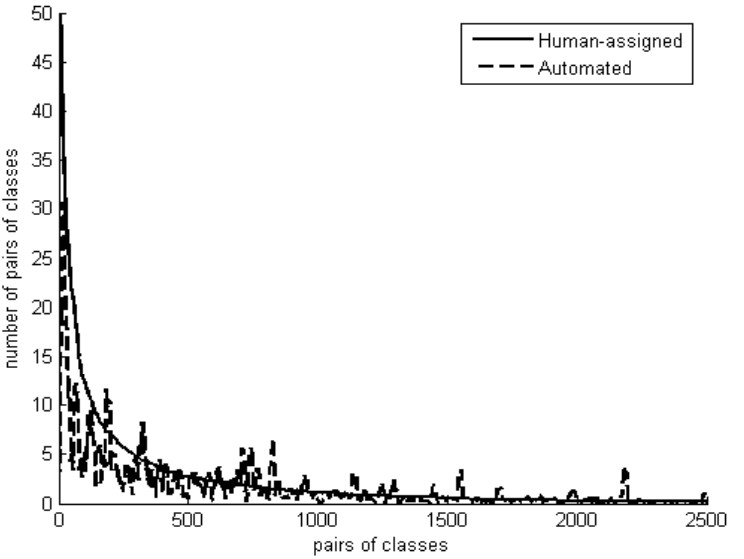


Figure 1. Frequency distribution of assigned pairs of classes (2538 pairs).

A correlation of 0.38 exists between the human-assigned classes and automatically assigned classes (Automated). However, for the 100 most frequent pairs, the correlation drops to 0.21. In the top 10 most frequent pairs of classes, there is no overlap at all. In conclusion, the distribution of human-assigned and automatically assigned classes is more correlated when looking at all pairs of classes occurring together, but less so for more frequently occurring pairs.

4.1.3 Score propagation and main classes

A relevant subject classification principle is to always assign the most specific class available. This principle provided us with a basis for score propagation, in which scores of classes at narrower (more specific) hierarchical levels were increased by scores assigned to their broader classes (later referred to as "propagated down"). In another run, this was slightly varied, so that the broader classes from which scores were propagated to their narrower classes were removed ("propagated down, broader removed").

These types of score propagation were tested on the best performing term list and setting (w134_12_1234 with stemming and stop-words removal). In complete matching, "propagated down" performs best. However, it is slightly worse than when not using score propagation at all. In partial matching, both "propagated down" and "propagated down, broader removed" perform slightly better than the original on the first two or three hierarchical levels, and slightly worse on the fourth and fifth ones. These not-so-good results with score propagation can be partially explained by the fact that the term list contained both broader and narrower terms, which was done in order to achieve best recall (Golub 2006b).

We further analyzed the degree to which the one most important concept of every document is found by the algorithm. To this purpose, a subset of (19,153) documents was used which had the human-assigned main class in class 9 (there is one main class per document). In complete matching 78% of main classes are found when no cut-offs are applied. When cut-offs are applied, 22% of main classes are found. In partial matching, more main classes are found at the second and third hierarchical levels when using both types of score propagation, up to 59% and 38% respectively. Thus,

score propagation could be used in services for which fewer hierarchical levels are needed (e.g., Intute Consortium 2006).

4.2 Enhancing the term list with new terms

In the previous experiment (Golub 2006b), highest achieved recall was 73% (microaveraged), when all types of terms were included in the term list. In order to further improve recall, the basic term list was enriched with new terms. These terms were extracted from bibliographic records of the Compendex database, using multi-word morpho-syntactic analysis and synonym acquisition, based on the existing preferred and synonymous terms (as they gave best precision results). The number of terms added to the term list was as follows:

1) Based on multi-word morpho-syntactic analysis:

   - derivation: 705, out of which 93 adjective to noun, 78 noun to adjective, and 534 noun to verb;
   - permutation: 1373;
   - coordination: 483;
   - insertion: 742; and
   - preposition change: 69.

2) Based on semantic variation (synonymy): 292 automatically extracted, out of which 168 were verified as correct by the subject expert.

In order to examine the influence of different types of extracted terms, nine different term lists were created and the classification was based on each of them. It was shown that the number of terms is not proportional to performance, e.g., permutation-based extraction comprises 1373 terms, and, when stemming is applied, has performance as measured in mean F1 of 0.02, whereas coordination comprises 403 terms, with performance of 0.07. These two cases can be explained by the fact that permutation also implies variation based on insertion and preposition change (e.g., engineering for commercial window systems: system engineering) which leads to bigger semantic shift than the identification of term variant based on the coordination. By combining all the extracted terms into one term

list, the mean F1 is 0.14 when stemming is applied, and microaveraged recall is 0.11, which would imply that enriching the original Ei-based term list with these newly extracted terms should improve recall. In comparison to results gained in Golub 2006b, where microaveraged recall with stemming is 0.73, here the best recall, also microaveraged and with stemming, is 0.76.

The next step was to assign appropriate weights to the newly extracted terms (Table 7). We used the w134_12_1234 term list, earlier shown to perform best. The result as measured in mean F1 is the same as in the original, 0.24 (cut-off 10, stemming applied but not stop-word removal). The difference is that recall and the number of correctly assigned classes increases by 3%, but precision decreases. Thus, depending on the final application, terms extracted in this way could be added to the term list or not.

Table 7. Performance of the w1 term list enriched with all automatically extracted terms.

	all combined			
stemming	no	yes	no	yes
stop-words out	no	no	yes	yes
min 1 correct	24479	29639	26039	30466
min 1 auto	34086	34966	34425	34987
avg auto/doc	16.79	28.61	18.06	29.68
macroa P	0.11	0.09	0.11	0.09
macroa R	0.54	0.71	0.55	0.72
macroa F1	0.19	0.16	0.18	0.15
microa P	0.07	0.06	0.07	0.06
microa R	0.55	0.73	0.59	**0.76**
macroa F1	0.13	0.11	0.13	0.10
mean F1	0.16	0.13	0.16	0.13

4.3 Terms analysis and shortened term lists

In the original term list there were 4,411 distinct terms. In the document collection, 53% of them were found. The average length of the terms found was between one and two words, while the longer ones were less frequently found.

Of the terms found in the collection, based on 16% of them correct classes were always found, while based on 43% of them

incorrect classes were always found. For a sample of documents containing terms that were shown to always yield incorrect results, we had a male subject expert confirm whether the documents were in the wrong class according to his opinion. For 10 always-incorrect terms with most frequent occurrences, the subject expert looked at 30 randomly selected abstracts containing those terms. Based on his judgments, it was shown that 24 out of those 30 documents were indeed incorrectly classified, but there were also 6 which he deemed to be correct. This is another indication of how problematic it is to evaluate subject classification in general, and automated subject classification in particular. Perhaps one way would be to have a number of subject experts agree on all the possible subjects and classes for every document in a test collection for automated classification; another way could be to evaluate automated classification in context, by end-users.

Based on the term analysis, three new term lists were extracted from the original one, and tested for performance:

1) Containing only those terms that found classes which were always correct (1,308 terms). When cut-off is between 5 and 10, macroaveraged precision reaches 0.89, and microaveraged 0.99, when neither stemming nor stop-words removal are applied. Stemming does not really improve general performance because recall increases only little, by 0.03, while precision decreases by 0.2. However, when using only those 1,308 terms, only 5% of documents are classified. The best mean F1, 0.15, is achieved when stemming and the stop-word removal are used.
2) Containing those terms that found classes which were correct in more instances than they were incorrect (1,924 terms). This list yields best mean F1, 0.38. This value is achieved when stemming is used but no stop-words are removed. There are 65% of documents that are classified, with the average number of classes 1.7. When stemming is not used, precision levels are 0.75 for microaveraged, and 0.79 for macroaveraged.
3) Containing all terms excluding those that found classes which were always incorrect (4,751 terms). The mean F1 is 0.25, when cut-off is 10 and both stop-words removal and stemming are used. The slight improvement in comparison to the original list is due to increase in precision.

5 Conclusion

In comparison to previous results (Golub 2006b) the experiment showed that the string-matching classification algorithm could be enhanced in the following ways:

1) Weights: adding different weights to the term list based on whether a term is single, phrase or Boolean, which type of class it maps to, and Ei term types, improves precision, mean F1, and relevance order of assigned classes, the latter being important for browsing;
2) Cut-offs: selecting as final classes those above a certain cut-off level improves precision and F1. Assigning at least the class with highest score improves the number of documents that are classified, and the number of documents that are correctly classified;
3) Converting all phrases into Boolean terms increases the number of documents with correct classes;
4) Stemming, stop-words removal or the two in combination improve F1;
5) Score propagation improves finding the main class at the top three and two hierarchical levels;
6) Enhancing the term list with new terms based on morpho-syntactic analysis and synonyms acquisition improves recall;
7) Excluding terms that in most cases gave wrong classes yields best performance in terms of F1, where the improvement is due to higher precision levels; and
8) Best precision levels are achieved when only those terms that always gave correct classes are used.

The best achieved recall is 76%, when the basic term list is enriched with new terms, and precision 79%, when only those terms previously shown to yield correct classes in the majority of documents are used. Performance of individual classes, measured in precision, is up to 98%. At third and second hierarchical levels mean F1 reaches up to 60%.

These results are comparable to machine-learning algorithms (see, for example, Sebastiani 2002), which are considered to be the best ones but require training documents and are collection-dependent. Another benefit of classifying documents into classes of

well-developed classification schemes is that they are suitable for subject browsing, unlike automatically-developed controlled vocabularies or home-grown directories often used in document clustering and text categorization (Golub 2006a).

The experiment has also shown that different versions of the algorithm could be implemented so that it best suits the application of the automatically classified document collection. If the application requires high recall, such as, for example, in focused crawling, cut-offs would not be used. Or, if one provides directory-style browsing interface to a collection of automatically classified web pages, web pages could be ranked by relevance based on weights. In such a directory, one might want to limit the number of web pages per class, e.g., assign only the class with highest probability that it is correct, as it is done in the Thunderstone's web site catalog (Thunderstone 2005).

Most appropriate weights have still to be discovered. Future research should also involve testing automated classification in the context of an application and by end users, because of the problem of "aboutness". The applicability of the string-matching approach mostly depends on the controlled vocabulary itself. While Ei proved to be suitable, which characteristics of controlled vocabularies are beneficial for automated classification needs to be further studied.

Acknowledgments

Many thanks to Traugott Koch for providing us with detailed comments on the manuscript, which helped improve the paper considerably. The authors also wish to thank two subject experts who helped in evaluation. This work was supported by the IST Programme of the European Community under ALVIS (IST-002068-STP).

References

Aitchinson, J., Gilchirst, A., Bawden, D. (2000), *Thesaurus construction and use: a practical manual,* 4th ed., Aslib, London.

Aubin, S., and Hamon, T. (2006), "Improving term extraction with terminological resources", *Proceedings of the 5th International Conference on NLP, FinTAL,* pp. 380-387.

Baayen, R.H., Piepenbrock, R., and Gulikers, L. (1995), *The CELEX lexical database*, release 2, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA. [CD-ROM].

Bang, S.L., Yang, J.D., and Yang, H.J. (2006) "Hierarchical document categorization with k-NN and concept-based thesauri", *Information Processing and Management*, Vol. 42, pp. 387-406.

Chen, H., and Dumais, S.T. (2000), "Bringing order to the Web: automatically categorizing search results", *Proceedings of ACM International Conference on Human Factors in Computing Systems, Den Haag*, pp. 145-152.

Engineering Information (2006), "Compendex", Engineering Information, Elsevier, available at: http://www.ei.org/databases/compendex.html (accessed 30 June 2006).
engineering/ (accessed 30 August 2007).

Fellbaum, C. (1998), *WordNet: an electronic lexical database,* MIT Press, Cambridge, MA.

Golub, K. (2006a), "Automated subject classification of textual Web documents", *Journal of Documentation,* Vol. 62 No. 3, pp. 350-371.

Golub, K. (2006b), "The role of different thesauri terms in automated subject classification of text", *Proceedings of the International Conference on Web Intelligence, Hong Kong,* pp. 961-965.

Hamon, T., Nazarenko, A. (2001), "Detection of synonymy links between terms: experiment and results", *Recent Advances in Computational Terminology,* pp. 185-208.

International Organization for Standardization (1985), *Documentation – Methods for examining documents, determining their subjects, and selecting index terms*: *ISO 5963,* Geneva, International Organization for Standardization.

Intute Consortium (2006), *Intute: Science, engineering and technology – engineering*, available at: http://www.intute.ac.uk/sciences/ (accessed 30 August 2007).

Jacquemin, C. (1996), "A symbolic and surgical acquisition of terms through variation", *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing,* pp. 425-438.

Jain, A.K., Murty, M.N., and Flynn, P.J. (1999), "Data clustering: a review", *ACM Computing Surveys*, Vol. 31 No. 3, pp. 264-323.

Koch, T., and Ardö, A. (2000), "Automatic classification", *DESIRE II D3.6a, Overview of results),* available at: http://www.lub.lu.se/desire/ DESIRE36a-overview.html (accessed 22 December 2004).

Lancaster, F.W. (2003), *Indexing and abstracting in theory and practice,* 3rd ed, Facet, London.

Lewis, D.D., Yang, Y., Rose, T., and Li, F. (2004), "RCV1: a new benchmark collection for text categorization research", *The Journal of Machine Learning Research,* 5, pp. 361-397.

Markey, K. (1984), "Interindexer consistency tests: a literature review and report of a test of consistency in indexing visual materials", *Library & Information Science Research,* 6, pp. 155-77.

Medelyan, O., and Witten, I. (2006), "Thesaurus based automatic keyphrase indexing", *Proceedings of Joint Conference on Digital Libraries*, pp. 296-297.

Milstead, J, ed. (1995), *Ei thesaurus,* 2nd ed., Engineering Information Inc., Hoboken, NJ.

Olson, H.A., and Boll, J.J. (2001), *Subject analysis in online catalogs*, 2nd ed., Libraries Unlimited, Englewood, CO.

"Onix text retrieval toolkit: Stop word list 1", available at: http://www.lextek.com/manuals/onix/stopwords1.html (accessed 30 June 2006).

Porter, M.F. (1980), "An algorithm for suffix stripping", *Program,* Vol. 14 No. 3, pp. 130-137.

Salton, G., and McGill, M.J. (1983), *Introduction to modern information retrieval,* McGraw-Hill, Auckland.

Sebastiani, F. (2002), "Machine learning in automated text categorization", *ACM Computing Surveys*, Vol. 34 No. 1, pp. 1-47.

Svenonius, E. (1997), "Definitional approaches in the design of classification and thesauri and their implications for retrieval and for automatic classification", *Proceedings of the Sixth International Study Conference on Classification Research,* pp. 12-16.

Svenonius, E. (2000), *The intellectual foundations of information organization,* MIT Press, Cambridge, MA.

Thunderstone (2005), *Thunderstone's Web Site Catalog*, available at: http://search.thunderstone.com/texis/websearch (accessed 4 August 2005).

WordNet, "WordNet Search", available at: http://wordnet.princeton.edu/perl/webwn (accessed 13 July 2007).

Yang, Y. (1999), "An evaluation of statistical approaches to text categorization", *Journal of Information Retrieval*, Vol. 1 No. 1/2, pp. 67-88.

Paper VII

PAPER VII.
Comparing and Combining Two Approaches to Automated Subject Classification of Text

Abstract

In this exploratory study a machine-learning (SVM), and a string-matching algorithm were compared, as to their performance and (dis)advantages. Document collection consisted of a subset of Compendex bibliographic records classified into six different classes. It was shown that SVM on average outperforms the string-matching approach. Our hypothesis that SVM would yield better recall and string-matching better precision was confirmed only on one of the classes. Our second hypothesis was that classification performance could be improved by confederating the two algorithms. Terms (features) used by one algorithm were combined with the other algorithm's terms in five different ways. The results have shown that SVM performs best in its original setting, while recall and F1 of string-matching improve when using the SVM terms. Further research should include a larger document collection, recognition of conditions in which one algorithm performs better than the other, and possibly combining the two algorithms by using predictions.

Expanded version of:

Golub, K., Ardö, A., Mladenić, D., and Grobelnik, M. (2006), "Comparing and combining two approaches to automated subject classification of text", *Proceedings of 10th European Conference on Research and Advanced Technology for Digital Libraries, Alicante, Spain, 17-22 September*, pp. 467-470.

1 Introduction

Subject classification is organization of objects into topically related groups and establishing relationships between them. In automated subject classification (in further text: automated classification) intellectual process of manual text classification is modelled by, for example, statistical and computational linguistics techniques. Automated classification of textual documents has been addressed by research community for several decades now, the interest rapidly growing with the advancement of the World Wide Web. Due to high costs of human-based subject classification and the ever-increasing number of documents there is a danger that recognized objectives of bibliographic systems (Svenonius 2000, 20-21) would be left behind; automated means could be a solution to preserve them (30). Automated classification of text has many different applications (Sebastiani 2002); in this paper, the application context is that of information retrieval.

The most widespread approach to automated classification of text is machine learning. In machine learning, characteristics of subject classes into which documents are classified are modelled (learnt) from human-based document classification. However, large-enough collections of human-classified documents are often unavailable in many subject areas, or for different user groups and document types. If one would judge by the standard Reuters Corpus Volume 1 collection (RCV1) (Lewis et al. 2004), some 8,000 training and testing documents per class would be needed. A related problem is that the algorithm performs well on new documents only if they are similar enough to the training documents. The issue of document collections was also pointed out by Yang (1999) who showed how certain versions of one and the same document collection had a strong impact on performance.

A less used approach is the string-matching one, in which text of documents to be classified is compared against terms designating classes of a classification scheme into which the documents are to be classified. Usually weighting schemes are applied to indicate how well a term designates a class and the degree to which a term from a document to be classified is significant for the document's topicality. A major advantage of this approach is that it does not require training documents. Another motivation to use this approach

is to reuse the intellectual work put into creation and maintenance of well-developed controlled vocabularies.

Comparisons of different algorithms have been conducted by Yang (1999). She compared a string-matching algorithm called WORD (72), based on a vector-space model, to machine-learning algorithms, and showed that the latter outperformed the former. Her approach is different from our string-matching approach in that she used vector-based representations and comparisons. Moreover, the categories and document collection in her study are entirely different from ours (see 2.1 and 2.3) – she used Reuters articles, categorized into Reuters-made subject structure.

While studies combining different algorithms for automated classification with controlled vocabularies are few (see Golub and Larsen 2005), indications of potential benefits exist. Bang et al. (2006) used a thesaurus to improve performance of the k-NN classifier and managed to improve precision by 14%, without degrading recall. Medelyan and Witten (2006) showed how information from a thesaurus improved performance of keyphrase extraction by more than 1.5 times in F1, precision, and recall.

The purpose of our exploratory study was to gain indications as to potential advantages of combining machinelearning with string-matching algorithms. Both algorithms were to classify documents into classes of the chosen classification scheme in the field of engineering. Performance of a machine learning algorithm against a string-matching one was compared, including different derivation of sets of terms: controlled vocabulary terms (extracted from the used thesaurus and classification scheme), automatically extracted terms based on a centroid *tf-idf* vector of each class (term frequency – inverse document frequency (Salton and McGill 1983, 63, 205)), and automatically extracted terms that distinguish one class from the others based on support vector machines.

The paper is structured as follows: in the next section (section 2) methodology is described, including the two algorithms, document collection, and experimental setting; in section 3 results are presented and discussed; the paper ends with conclusions and indications for further research (section 4).

2 Methodology

2.1 String-matching algorithm

The string-matching algorithm is based on finding terms from the Ei (Engineering Information) thesaurus and classification scheme (Milstead 1995), in the field of engineering, in text of documents to be classified (also in the field of engineering). The Ei thesaurus and classification scheme (in further text: Ei) consist of two parts: a thesaurus of engineering terms, and a hierarchical classification scheme of engineering topics. These two controlled vocabulary types have traditionally each had distinct functions: the thesaurus has been used to describe a document with as many controlled terms as possible, while the classification scheme has been used to group similar documents together to the purpose of shelving them and allowing systematic browsing. A major advantage of Ei is that thesaurus descriptors are mapped to classes of the classification scheme. These mappings have been made manually (intellectually) and are an integral part of the thesaurus. Compared to captions[1] alone, mapped thesaurus terms provide a rich additional vocabulary for every class: instead of having only one term per class (there is only one caption per class), in our study there were on average 150 terms per class.

Pre-processing steps of Ei included normalizing upper- and lower-case words. Upper-case words were left in upper case in the term list, assuming that they were acronyms; all other words containing at least one lower-case letter were converted into lower case. The first major step in designing the algorithm was to extract terms from Ei into what we call a term list. It contained class captions, thesaurus preferred terms, their synonyms (Term), classes to which the terms and captions map or denote (Class), and weight indicating how appropriate the term is for the class to which it maps or which it designates (Weight). Geographical names, all mapping to class 95, were excluded on the grounds that they are not engineering-specific. The term list was formed as an array of triplets:

[1] A caption is a class notation expressed in words, e.g., in The Ei classification scheme ‘Electric and Electronic Instruments’ is the caption for class ‘942.1’.

Weight: **Term** (single word, Boolean term or phrase) = **Class**

Single-word terms were terms consisting of one word. *Boolean terms* were terms consisting of two or more words that must all be present but in any order or in any distance from each other. Boolean terms in this form were not explicitly part of Ei, but were created to our purpose. They were considered to be those terms which in Ei contained the following strings: *and*, *vs.* (short for *versus*), *,* (comma), *;* (semi-colon, separating different concepts in class captions), *(* and *)* (parentheses, indicating the context of a homonym), *:* (colon, indicating a more specific description of the previous term in a class captions), and -- (double dash, indicating *heading--subheading* relationship). These strings we replaced with *@and* which indicated the Boolean relation in the term. All other terms consisting of two or more words were treated as *phrases*, i.e., strings that need to be present in the document in the exact same order and form as in the term. Ei comprises a large portion of composite terms (771 out in the total of 899 in our study); as such, Ei provides a rich and precise vocabulary with the potential to reduce the risks of false hits.

The following are two excerpts from the Ei classification scheme and thesaurus, based on which the excerpt from the term list (further below) would be created:

From the classification scheme:
...
402 Buildings and Towers
...
722.3 Data Communication
(Equipment and Techniques)
...
903.3 Information Retrieval and Use
...

From the thesaurus:
TM Architecture
NT Architectural design
RT Buildings
MC 402
...
TM Wide area networks
UF Computer networks--Wide area networks
UF WAN
BT Computer networks
OC 722.3
...
TM Geographic information systems
UF GIS
BT Nonbibliographic retrieval systems
RT Database systems

RT Mapping
MC 903.3

In the thesaurus, TM ("TerM") stands for the preferred term, UF ("Used For") for synonym, BT for broader term, RT for related term, NT for narrower term; MC represents the main class, and OC optional class to be used when appropriate for the specific document being indexed (Milstead 1995, ix). MC and OC are classes from the Ei classification scheme that have been made manually (intellectually) and are an integral part of the thesaurus. Captions, preferred terms and their synonyms were added to the term list. Based on the above excerpts, the following term list would be created:

15: buildings @and towers = 402
40: data communication @and equipment @and techniques = 722.3
40: information retrieval @and use = 903.3
15: architecture = 402
20: wide area networks = 722.3
20: computer networks @and wide area networks = 722.3
10: WAN = 722.3
40: geographic information systems = 903.3
15: GIS = 903.3

The number at the beginning of each triplet is weight indicating the probability that the term correctly designates the class. The weights were derived heuristically:

- A main class is made more important than an optional class, since the rule in the Ei thesaurus is that the optional class is to be used only under certain circumstances.
- Phrases are assigned highest weights (40 for main class, 20 for optional class), since they are normally most discriminatory.
- Boolean terms are the next best (15 for main class, 10 for optional class).
- Single words were assigned the same weights as Boolean entries (15 for main class, 10 for optional class). However, experiments have shown that they can be too general and/or have several meanings or uses that make them less specific; thus, in further experiments they should be assigned a lower weight.

The algorithm looks for strings from a given term list in a document to be classified and if the string (e.g., *wide area networks* from the above list) is found, the class(es) assigned to that string in the term

list (*722.3* in our example) is/are assigned to the document. One class can be designated by many terms, and each time a term is found, the corresponding weight (*20* in our example) is added to the class. The scores for each class are summed up and classes with scores above a certain cut-off (heuristically defined) can be selected as the final ones for that document. Unless specifically noted, no cut-offs were used in this study for string-matching.

The Ei classification scheme consists of six main classes divided into 38 more specific classes which are further subdivided into 182 classes. These are subdivided even further, resulting in some 800 individual classes in a five-level hierarchy. For this exploratory study only six classes were selected: 723.1.1 (Computer Programming Languages), 723.4 (Artificial Intelligence), 903.3 (Information Retrieval and Use), 722.3 (Data Communication Equipment and Techniques), 402 (Buildings and Towers), and 903 (Information Science). These classes were chosen because we had a document collection for them (2.3) from previous experiments.

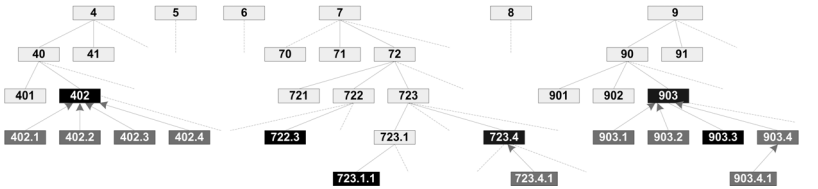


Figure 1. An extract from the Ei classification scheme and its hierarchical structure. Classes used in the study were the following:

- Class **402** (Buildings and Towers) with included subclasses 402.1 (Industrial and Agricultural Buildings), 402.2 (Public Buildings), 402.3 (Residences), 402.4 (Towers).
- Classes **722.3** (Data Communication Equipment and Techniques).
- Class **723.1.1** (Computer Programming Languages).
- Class **723.4** (Artificial Intelligence) with subclass 723.4.1 (Expert Systems).
- Class **903** (Information Science) with subclasses 903.1 (Information Sources and Analysis), 903.2 (Information Dissemination), 903.4.1 (Libraries), and 903.4 (Information Services).
- Class **903.3** (Information Retrieval and Use).

Figure 1 shows an excerpt from the Ei classification scheme and its hierarchical structure. The hierarchical level is reflected in the number of digits, starting from one digit at the top level, and ending with five digits at the deepest hierarchical level. In Fig. 1 the six classes selected for the study are the ones in black rectangles. In the term list, for the six classes also captions and thesauri terms from their subclasses were added (dark-grey rectangles with arrows pointing at their superclasses).

2.2 Linear support vector machines (SVM) algorithm

The second algorithm we experimented with was the support vector machine (SVM) (Cortes and Vapnik 1995), as it is currently considered the state-of-the-art algorithm in text classification. This algorithm trains a linear classifier of the form $\text{sgn}(\mathbf{w}^T \mathbf{x} + b)$. Here, $\mathbf{x}$ is a training example, $\mathbf{w}$ is the "normal" of the hyperplane separating examples of two different classes, i.e., a vector that is perpendicular to the plane and defines its orientation. The constant b defines the position of the hyperplane in space. Learning is posed as an optimization problem with the goal of maximizing the *margin*, i.e., the distance between the separating hyperplane $\mathbf{w}^T\mathbf{x}+b=0$ and the nearest training vectors. To choose $\mathbf{w}$ and b, SVM minimizes the criterion function

$$f(\mathbf{w}, b) := 1/2 \, \|\mathbf{w}\|^2 + C \, \Sigma_i \, \xi_i$$

subject to the constraints $\forall i: \; y_i(\mathbf{w}^T \mathbf{x}_i - b) \geq 1 - \xi_i$ and $\xi_i \geq 0$. Thus the learner has to look for a trade-off between maximizing the width of the margin (which is inversely proportional to $\|\mathbf{w}\|$) and avoiding classification errors on the training documents (represented by the slack variables ξ_i). The parameter C defines the relative importance of these two objectives. After learning, the classifier predicts the class labels of test documents using the formula $\textit{prediction}(\mathbf{x}) := \text{sgn}(\mathbf{w}^T \mathbf{x} - b)$. Thus, the decision process involves the term $\mathbf{w}^T \mathbf{x}$ that can be seen as the score of the document $\mathbf{x}$ and the threshold b above which the documents are assigned a positive label.

We used binary linear SVM, the implementation from TextGarden (Grobelnik and Mladenic 2007). Text of documents to be classified was preprocessed by removing stop-words and representing each document, using the standard bag-of-words approach containing individual words. Representation of each document was further enriched by frequent phrases, which were

considered to be those consisting of up to five consecutive words (as described in Mladenic and Grobelnik 2003) and occurring at least four times in the document collection. The binary classification model was automatically constructed for each of the six classes, taking training documents of the class as positive and training documents of other classes as negative. Each document from the document collection was classified by each of the six models. Finally, each document was assigned those classes that were above the threshold of zero. The classification model was trained on one part of the document collection, leaving the other part to be classified using the standard statistical method called cross validation. In other words, the document collection was randomly divided into several disjoint parts (in our case 10) of approximately equal size. Then 10 classification models were generated, each taking one of the 10 parts as testing and the remaining 9 parts as training documents. We report average performance (precision, recall) over the 10 testing.

2.3 Document collection

The document collection comprised 24,106 bibliographic records[2] (in further text: documents) from the Compendex database (Engineering Information 2006). For each of the three classes 723.1.1, 723.4 and 903.3 there were 4,400 documents, which is the download maximum allowed by the Compendex database provider. For other three classes all the documents existing at the time of download were taken: 4,283 documents for class 402, 3,823 for class 903, and 2,800 documents for class 722.3. For each document there was at least one of the six classes that were human-assigned (1.1 on average, ranging from four in one document to one in the majority of the documents).

From each document the following elements were extracted: an identification number, title, abstract and human-assigned classes (in Compendex called *Ei classification codes*). Thesauri descriptors (in Compendex called *Ei controlled terms*) were not extracted since the purpose of the experiment was to compare automatically

[2] Compendex being a commercial database, the document collection cannot be made available to others, but the authors are willing to provide documents' identification numbers on request.

assigned classes (and not descriptors) against the human-assigned ones. Automated classification was based on title and abstract. E.g.,:

Identification number: 03337590709
Title: The concept of relevance in IR
Abstract: This article introduces the concept of relevance as viewed and applied in the context of IR evaluation, by presenting an overview of the multidimensional and dynamic nature of the concept. The literature on relevance reveals how the relevance concept, especially in regard to the multidimensionality of relevance, is many faceted, and does not just refer to the various relevance criteria users may apply in the process of judging relevance of retrieved information objects. From our point of view, the multidimensionality of relevance explains why some will argue that no consensus has been reached on the relevance concept. Thus, the objective of this article is to present an overview of the many different views and ways by which the concept of relevance is used - leading to a consistent and compatible understanding of the concept. In addition, special attention is paid to the type of situational relevance. Many researchers perceive situational relevance as the most realistic type of user relevance, and therefore situational relevance is discussed with reference to its potential dynamic nature, and as a requirement for interactive information retrieval (IIR) evaluation.
Ei classification codes: 903.3 Information Retrieval & Use, 723.5 Computer Applications, 921 Applied Mathematics

2.4 Evaluation

The topic of "aboutness" has been much discussed in the literature, and inter- and intra-indexer consistency levels have been reported to be generally low (Olson and Boll 2001, 99-101). However, in this study we assumed that human-assigned classes in the document collection were correct, and compared results of automated classification against them. Evaluation measures used were the standard precision, recall and F1 measures (Sebastiani 2002, 40-41):

$$\text{Precision} = \frac{\text{correctly automatically assigned classes}}{\text{all automatically assigned classes}}$$

$$\text{Recall} = \frac{\text{correctly automatically assigned classes}}{\text{all human-assigned classes}}$$

$$F1 = 2*\text{Precision}*\text{Recall} / (\text{Precision} + \text{Recall})$$

When results were averaged for all classes, it was macroaveraging that was used. In macroaveraging, the results are first calculated for each class, and then they are summed and divided by the number of classes. Unlike microaveraging, macroaveraging shows the ability of

a classifier to behave well also on categories with low generality (33).

2.5 Term lists (features)

Apart from terms in original settings of the two algorithms, terms used by one algorithm were combined with the other algorithm's terms in five different ways. In experiments with string-matching algorithm, the following two term lists were additionally created:

1) The one comprising what we named *descriptive* terms. These were terms extracted from the documents belonging to the six classes. For every class highest ranked terms from their centroid *tf-idf* vectors (term frequency – inverse document frequency, Salton and McGill 1983, 63, 205) were taken. The number of terms per class was the same as the number of terms in the original term list (Table 1).
2) The one comprising what we named *distinctive* terms. These were terms extracted through the process of training the SVM algorithm on a binary classification problem for each class, i.e., terms that distinguished one class form the others. For every class highest ranked terms were taken (as proposed in Mladenic et al. 2004), so that the number of terms per class was the same as the number of terms in the original term list.

The weights assigned to the descriptive and distinctive term lists were based on the original weights derived by the SVM algorithm. They were multiplied by 100 and rounded, so the final values ranged between 1 and 38, similar to values in the original term list of the string-matching algorithm.

The linear SVM in the original setting was trained with no term selection except the stop-word removal (see 2.2). Additionally, three experiments were conducted using term selection, taking:

1) *Controlled* terms. Only those terms which were present in the original term list used by the string-matching algorithm, i.e., with terms taken from the Ei thesaurus and classification scheme.
2) *Descriptive* terms. We model each class by the centroid of all its documents (summing all the document-vectors). For every class

with the highest *tf-idf* values from their centroid *tf-idf* vectors of the class documents were taken.

3) *Distinctive* terms. We model each class with linear SVM model (obtained to enable binary classification of the class against all the other classes). For every class with the highest weight from its linear SVM model were taken.

For descriptive and distinctive terms we selected the highest ranked terms only (in our case 1,000) to avoid overfitting the data by using too many terms having low weight and consequently low value as descriptive or distinctive terms.

3 Results

3.1 Comparison in original settings

Our first hypothesis was that, as the string-matching algorithm is based on human-derived terms, it would yield higher precision, and that the machine-learning one with its automatically constructed model based on training documents would yield higher recall. As seen from Table 1, this hypothesis was confirmed only on one of the six classes: 903.3. On average, SVM outperforms the string-matching algorithm in both precision and recall.

Table 1. Performance of algorithms in their original settings. "Nbr of terms" stands for the number of original terms per class in the string-matching algorithm.

	String-matching				SVM		
Class	Nbr of terms	Precision	Recall	F1	Precision	Recall	F1
402	423	0.49	0.58	0.53	**0.91**	**0.93**	**0.92**
722.3	292	0.26	0.12	0.16	**0.79**	**0.76**	**0.78**
723.1.1	137	0.32	0.34	0.33	**0.79**	**0.73**	**0.76**
723.4	61	0.39	0.37	0.38	**0.81**	**0.65**	**0.72**
903	58	0.61	0.28	0.38	**0.74**	**0.71**	**0.73**
903.3	26	**0.97**	0.32	0.48	0.78	**0.73**	**0.76**
Average		0.51	0.33	0.40	**0.80**	**0.75**	**0.78**

In string-matching, best achieved F1 is for class 402, which could be attributed to the fact that it has the highest number of terms designating it in the term list (423). Class 903.3, on the other hand,

has only 26 different terms, but for this class string-matching outperforms SVM in precision (0.97 vs. 0.79). Since SVM performs best in class 402 as well, another assumption could be that this class is in a subject field where texts are most easily classified and have fewer ambiguities. Reasons for differences in performance of the classes need to be further investigated.

3.2 Combining term lists (features)

Our second hypothesis was that confederating the two algorithms via combining their sets of terms would result in performance better than either of the two algorithms in their original settings. This hypothesis was not confirmed. As seen from Tables 2 and 3, and in comparison to original settings (given in Table 1), best results are still achieved by SVM in its original setting.

Table 2. Performance of SVM using different term lists.

	Controlled			Descriptive			Distinctive		
Class	Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall	F1
402	0.68	0.85	0.75	**0.87**	0.89	**0.88**	0.67	**0.91**	0.77
722.3	0.48	0.64	0.55	**0.76**	0.73	**0.75**	0.62	**0.77**	0.68
723.1.1	0.59	0.57	0.58	**0.79**	0.72	**0.75**	0.61	**0.77**	0.68
723.4	0.67	0.38	0.49	**0.8**	0.63	**0.70**	0.66	**0.66**	0.66
903	0.41	0.36	0.39	**0.72**	**0.66**	**0.68**	0.59	0.64	0.62
903.3	0.60	0.48	0.53	**0.78**	0.72	**0.75**	0.67	**0.76**	0.71
Average	0.57	0.55	0.55	**0.79**	0.72	**0.75**	0.64	**0.75**	0.69

SVM performance results for each of the classes, based on the three different term lists (2.5) are given in Table 2. SVM using the original setting (Table 1) has best performance for all classes, closely followed by SVM with only descriptive terms. Compared to the original setting, using distinctive and controlled terms further decreases the SVM performance. These results confirm previous findings in automated classification studies (Brank et al. 2002) that applying feature (term) selection to reduce the feature space before training linear SVM usually decreases the performance. Using controlled terms in combination with some other classification algorithm that is known to benefit from feature selection might be worth further testing. One can also see that in five out of six cases descriptive terms yield better precision, and distinctive better recall.

This is somewhat counterintuitive to the fact that descriptive terms capture vocabulary of each class independently of the other classes while distinctive terms associate each class to terms that best distinguish the class from the other classes in hand.

Our hypothesis that using the controlled terms for SVM feature selection would increase its performance was not confirmed. We attribute that to a large overlap between the controlled terms and documents' terms that enable SVM to find the right terms for a good quality model. We expect that the advantage of using controlled vocabulary with SVM would be evident on such document collections where documents inside classes would have very diverse vocabulary (e.g., many synonyms, highly inflected languages); in that case simple statistics over words would be of limited use and some natural language processing might be needed, such as lemmatization or incorporating some kind of semantics. One would expect that human-generated control vocabulary would partially compensate for the difficulty. The overlap between the controlled terms and terms selected by SVM from documents would probably be rather small.

Table 3. Performance of string-matching using different terms lists.

	Descriptive			Distinctive		
Class	Precision	Recall	F1	Precision	Recall	F1
402	**0.52**	0.89	**0.66**	0.35	**0.92**	0.50
722.3	**0.46**	**0.69**	**0.55**	0.31	0.60	0.41
723.1.1	**0.43**	0.75	**0.55**	0.25	**0.79**	0.38
723.4	**0.58**	**0.68**	**0.62**	0.36	0.65	0.47
903	**0.53**	0.63	**0.58**	0.36	**0.68**	0.47
903.3	**0.47**	**0.69**	**0.56**	0.42	0.59	0.49
Average	**0.50**	**0.73**	**0.59**	0.34	0.70	0.46

In the string-matching experiment based on SVM-derived term lists (2.5), improved performance was achieved when cut-offs were applied. The best cut-off, when still enough classes per documents would be assigned (1.1 is in human-assigned classes in the document collection), was when classes selected as the final ones were those with scores that contained at least 15% of the sum of all the scores assigned to all the classes per document.

When comparing performance in the original setting with Table 3, one can see that using descriptive terms improves performance of

the string-matching algorithm, especially when measured in recall and F1. Precision is increased in four classes, and decreased in two. Still, by far the best precision for class 903.3 is achieved in string-matching algorithm's original setting. As it is the case with SVM (Table 2), descriptive terms seem to perform better than distinctive ones, apart from recall for three classes.

4 Conclusion

From this exploratory study, based on a sample of six classes in one subject area and in one very homogeneous document collection, we can see that machine learning approach using linear SVM algorithm outperforms the string-matching approach in F1, recall and precision on all but one class, where the string-matching algorithm achieves higher precision. It was also shown that SVM performs best in its original settings. On the other hand, string-matching is improved by using (mostly) descriptive terms in all classes but one for which best precision, better than with SVM, is achieved in string-matching's original setting.

While SVM used in the study mostly outperforms the string-matching approach, one should remember that when it comes to real-life information systems such as digital libraries, pre-classified document collections (e.g., especially of web pages) are rarely available. String-matching algorithms could in such cases be the feasible solution.

Further research should include more classes and a larger document collection, recognition of conditions in which one algorithm performs better than the other, and possibly combining the two algorithms by using their predictions. String-matching could be further experimented with to see if improvements could be achieved.

Acknowledgments

This work was supported by the Slovenian Research Agency and the IST Programme of the European Community under ALVIS (IST-1-002068-STP).

References

Bang, S.L., Yang, J.D., and Yang, H.J. (2006) "Hierarchical document categorization with k-NN and concept-based thesauri", *Information Processing and Management*, Vol. 42, pp. 387-406.

Brank, J., Grobelnik, M., Milic-Frayling, N., and Mladenic, D. (2002), "Interaction of feature selection methods and linear classification models", *Proceedings of the ICML-2002 Workshop on Text Learning,* pp. 12-17.

Cortes, C., and Vapnik, V.N. (1995), "Support-vector networks", *Machine Learning,* Vol. 20 No. 3, pp. 273-297.

Engineering Information (2006), "Compendex", Engineering Information, Elsevier, available at: http://www.ei.org/databases/compendex.html (accessed 30 June 2006).

Golub, K. and Larsen, B. (2005), "Different Approaches to Automated Classification: Is There an Exchange of Ideas?", *Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics, Stockholm, Sweden, 24-28 July,* Vol. 1. pp. 270-274.

Grobelnik, M., and Mladenic, D. (2007), *Text Mining Recipes,* Springer, Berlin, Heidelberg, New York. Accompanying software available at: http://www.textmining.net (accessed 3 September 2007).

Lewis, D.D., Yang, Y., Rose, T., and Li, F. (2004), "RCV1: a new benchmark collection for text categorization research", *The Journal of Machine Learning Research,* 5, pp. 361-397.

Medelyan, O., and Witten, I. (2006), "Thesaurus based automatic keyphrase indexing", *Proceedings of Joint Conference on Digital Libraries*, pp. 296-297.

Milstead, J, ed. (1995), *Ei thesaurus,* 2nd ed., Engineering Information Inc., Hoboken, NJ.

Mladenic, D., and Grobelnik, M. (2003), "Feature selection on hierarchy of web documents", *Journal of Decision Support Systems,* Vol. 35, pp. 45-87.

Mladenic, D., Brank, J., Grobelnik, M., Milic-Frayling, N. (2004), "Feature selection using linear classifier weights: interaction with classification models", *Proceedings of the Twenty-Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, July 25th-29th,* pp. 234-241.

Olson, H.A., and Boll, J.J. (2001), *Subject analysis in online catalogs*, 2nd ed., Libraries Unlimited, Englewood, CO.

Salton, G., and McGill, M.J. (1983), *Introduction to modern information retrieval,* McGraw-Hill, Auckland.

Sebastiani, F. (2002), "Machine learning in automated text categorization", *ACM Computing Surveys*, Vol. 34 No. 1, pp. 1-47.

Svenonius, E. (2000), *The intellectual foundations of information organization,* MIT Press, Cambridge, MA.

Yang, Y. (1999), ”An evaluation of statistical approaches to text categorization”, *Journal of Information Retrieval*, Vol. 1 No. 1/2, pp. 67-88.

Paper VIII

PAPER VIII. Automated Subject Classification of Engineering Web Pages in Hierarchical Browsing: A User Study

Abstract

A user study was conducted to investigate performance of an automated classification algorithm on a harvested collection of engineering web pages, in the context of hierarchical browsing. Terms from the Engineering Information (Ei) thesaurus and classification scheme were matched against text of about 19,000 web pages. The study involved 40 engineering students or experts who were, given four tasks of different types, to find the most appropriate class in the Ei classification scheme and evaluate whether the top ranked web pages in that class were on the topic of their task. The data were mostly collected through questionnaires, logging and correctness assessments. Based on the study it was shown that the Ei classification scheme is suited for browsing. Concerning automated classification performance, the top ranked web pages in each of the four classes were on average deemed to be partly correct. As in previous research, it was shown that some classes are better automatically classified than others. Success of browsing showed to be dependent on classification accuracy.

Golub, K. (2007), "Automated subject classification of engineering web pages in hierarchical browsing: a user study", to be submitted.

1 Introduction

With the exponential growth of the World Wide Web, automated subject classification of web pages has become a major research issue. Compared to other types of text documents such as traditional research papers, with web pages there is a special challenge. They tend to be rather heterogeneous: very short with hardly any text, or very long, with titles that are often general ("Home page") or non-existent ("Untitled document"), and metadata that are inconsistent or misused. User-based evaluation of automated classification has been called for but rarely conducted (see Golub 2006a).

Organizing web pages into a hierarchical structure for subject browsing has been gaining more recognition as an important tool in information-seeking processes. Usefulness of classification schemes for browsing web resources has been reported but rarely researched (Koch et al. 2006; Vizine-Goetz 1996; Koch and Zettergren 1999; see also Soergel 2004). In this study the aim was to investigate performance of an automated classification algorithm on a collection of engineering web pages, in the context of hierarchical browsing. This has, to the knowledge of the author, not been conducted before.

One could distinguish between several different approaches to automated classification (Golub 2006a; Sebastiani 2002; Moens 2000; Jain et al. 1999). In this study a string-matching algorithm is applied, which does not require pre-classified training documents, often unavailable, especially for web pages. The algorithm searches for strings from the Engineering Information (Ei) thesaurus and classification scheme (Milstead 1995) in text of web pages to be classified. When a string is found, the class which it designates is assigned to the web page. Performance results for the algorithm on a collection of paper abstracts were reported as comparable to state-of-the-art algorithms in machine learning, especially for certain classes (Golub et al. 2007).

Different classification schemes have different characteristics; for subject browsing the following are important: hierarchical structure; the bigger the collection, the more depth should the hierarchy contain; and, classes should contain more than just one or two documents (Schwartz 2001, 48). Search-engine directories and other home-grown schemes on the Web, "…even those with well-

developed terminological policies such as Yahoo… suffer from a lack of understanding of principles of classification design and development. The larger the collection grows, the more confusing and overwhelming a poorly designed hierarchy becomes…" (76). For these reasons it was decided to use the Ei classification scheme, which has been used and maintained in the Compendex database (Engineering Information 2006).

The purpose of the study was to investigate 1) how suitable the Ei classification scheme is for browsing, and 2) how the classification algorithm performs on a harvested collection of web pages. The relationship between the two is also analysed. Two major research questions were:

1) Are users able to navigate the Ei classification structure?
2) How well are the web pages classified, as judged by the users?

The paper is structured as follows: in the following section, the classification algorithm and the classification scheme are described (2 Background); in the third section the study design is presented (3 Methodology); in the fourth section (4 Results) the results are analyzed and discussed; and, final remarks are given in the last section (5 Conclusions).

2 Background

2.1 Classification algorithm

This section describes the classification algorithm used in the study. The purpose was to classify a collection of web pages into classes of the Ei classification scheme in order to provide a browsing interface to the collection. The algorithm searches for terms from the Ei thesaurus and classification scheme (in further text: Ei controlled vocabulary), in the field of engineering, in text of web pages to be classified. The Ei controlled vocabulary consists of two parts: a thesaurus of engineering terms, and a hierarchical classification scheme of engineering topics. These two controlled vocabulary types have traditionally each had distinct functions: the thesaurus has been used to describe a document with a number of controlled terms and thus allow as many access points as possible, while the

classification scheme has been used to group similar documents together to the purpose of shelving them and allowing systematic browsing.

A major advantage of the Ei controlled vocabulary for automated classification is that thesaurus descriptors are mapped to classes of the classification scheme. These mappings have been made manually (intellectually) and are an integral part of the thesaurus. Compared with captions[1] alone, mapped thesaurus terms provide a rich additional vocabulary for every class: instead of having only one term per class (there is only one caption per class), in the study there were on average 14 terms per class.

Pre-processing steps of Ei included normalizing upper- and lower-case words. Upper-case words were left in upper case in the term list, assuming that they were acronyms; all other words containing at least one lower-case letter were converted into lower case. The first major step in designing the algorithm was to extract terms from Ei into a term list. It contained class captions, thesaurus terms, classes to which the terms and captions map or denote, and weight indicating how appropriate the term is for the class to which it maps or which it designates. (Geographical names were excluded on the grounds that they were not engineering-specific.) The term list was formed as an array of triplets:

Weight: **Term** (single word, Boolean term or phrase) = **Class**

Single-word terms were terms consisting of one word. *Boolean terms* were terms consisting of two or more words that must all be present but in any order or in any distance from each other. Boolean terms in this form were not explicitly part of Ei, but were created to our purpose. They were considered to be those terms which in Ei contained the following strings: *and*, *vs.* (short for *versus*), *,* (comma), *;* (semi-colon, separating different concepts in class captions), *(* and *)* (parentheses, indicating the context of a homonym), *:* (colon, indicating a more specific description of the previous term in a class captions), and -- (double dash, indicating *heading--subheading* relationship). These strings we replaced with

[1] A caption is a class notation expressed in words, e.g., in the Ei classification scheme "Electric and Electronic Instruments" is the caption for class "942.1".

@and which indicated the Boolean relation in the term. All other terms consisting of two or more words were treated as *phrases*, i.e., strings that need to be present in the document in the exact same order and form as in the term.

The following are two excerpts from the Ei classification scheme and thesaurus, based on which the excerpt from the term list (further below) is created:

From the classification scheme:
931.2 Physical Properties of Gases, Liquids and Solids
...
942.1 Electric and Electronic Instruments
...
943.2 Mechanical Variables Measurements

From the thesaurus:
TM Amperometric sensors
UF Sensors--Amperometric measurements
MC 942.1
...
TM Angle measurement
UF Angular measurement
UF Mechanical variables measurement--Angles
BT Spatial variables measurement
RT Micrometers
MC 943.2
...
TM Anisotropy
NT Magnetic anisotropy
MC 931.2

All the different thesaurus terms as well as captions were added to the term list. While choosing all the types of thesaurus terms might lead to precision losses, the decision was to use them all to achieve maximum recall, as shown in a previous paper (Golub 2006b). In the thesaurus, TM stands for the preferred term, UF ("Used For") for an equivalent term, BT for broader term, RT for related term, NT for narrower term. MC represents the main class; sometimes there is also OC, which stands for optional class, valid only in certain cases. Main and optional classes are classes from the Ei classification scheme that have been manually mapped to thesaurus terms and are an integral part of the thesaurus. Based on the above excerpts, the following term list would be created:

1: physical properties of gases @and liquids @and solids = 931.2,
1: electric @and electronic instruments = 942.1,
1: mechanical variables measurements = 943.2,
1: amperometric sensors = 942.1,

1: sensors @and amperometric measurements = 942.1,
1: angle measurement = 943.2,
1: angular measurement = 943.2,
1: mechanical variables measurement @and angles = 943.2,
1: spatial variables measurement = 943.2,
1: micrometers = 943.2,
1: anisotropy = 931.2,
1: magnetic anisotropy = 931.2,

The number at the beginning of each triplet is weight estimating the probability that the term of the triplet designates the class; in this example it is set to 1.

The algorithm looks for strings from a given term list in a web page to be classified and if the string (e.g., *magnetic anisotropy* from the above list) is found, the class(es) designating that string in the term list (*931.2* in the example) is(are) assigned to the web page. One class can be designated by many terms, and each time the class is found, the corresponding weight (*1* in the example) is added to a score for the class. The scores for each class are summed up and classes with scores above a certain cut-off (heuristically defined) can be selected as the final ones for the web page being classified.

Best weights and cut-offs determined in a previous experiment (Golub et al. 2007) were used. Weights 1, 3, and 4 for single, phrase or Boolean term were multiplied by the weight for the type of class to which the term mapped, 1 or 2 for optional or main class, and by the weight for the type of Ei term (broader 1, narrower 2, preferred 4, related 1, synonyms 3 and captions 4). Also, only those terms that had always found correct classes (1,308 terms) were included. Stemming and stop-words removal were not applied. In order to assign a certain class as final, the score of that class had to have at least 10% of the sum of all the classes' scores (for each web page); in case not a single class with high enough score existed, the one with the highest score was assigned.

As shown in another experiment (Golub and Ardö 2005), text coming from all the four parts of a web page (title, headings, main text, metadata) should be included in the process of automated classification. The same weights that had been shown to perform best in that experiment were used in the study: the scores of classes found in each part were multiplied with the following weights:

$$\text{Score(class)} = 86\cdot\text{Score(Title)} + 5\cdot\text{Score(Headings)} + 6\cdot\text{Score(Metadata)} + \text{Score(Main Text)}.$$

2.2 Engineering Information classification scheme

The Ei classification scheme is hierarchical and consists of six main classes divided into 38 finer classes which are further subdivided into 182 classes. These are subdivided even further, resulting in some 800 individual classes in a five-level hierarchy. For this study one of the six main classes was selected, together with its subclasses: class 9, *Engineering, General*. The reason for choosing this class was that it covers both natural sciences such as physics and mathematics, and social science fields such as engineering profession and management. The literature of the latter tends to contain more polysemic words than the former, and as such presents a more complex challenge for automated classification. Within the 9 class, there are 99 subclasses; their distribution at the five different hierarchical levels is as follows: 11 classes at the fifth hierarchical level, 67 at the fourth, 16 at the third, and 5 at the second one (see Appendix 1).

3 Methodology

3.1 Web page collection

The collection was automatically created, i.e., web pages were collected using the Combine focused crawler (Ardö 2007). The topic list used by the crawler was the same as the one used later on by the classification algorithm – containing 1,308 terms (see 2, see also Golub et al. 2007). No stemming and no stop words removal were applied. The collection contained 18,895 web pages, crawled in the period between 10 and 15 May 2007. There were 518 seed web pages, taken from the Intute subject gateway, the topic of Engineering General (Intute Consortium 2006).

Once the pages were crawled, they were classified as described in section 2. On average, 1.5 classes were assigned per web page. No web pages were classified into the top two hierarchical levels, because of the classification principle to assign the most specific class available.

3.2 User study design

The purpose of the study was to investigate how suitable the Ei classification scheme is for browsing, and how the classification

algorithm performs on a harvested collection of web pages. The relationship between the two is also analysed. Two major research questions were:

1) Are users able to navigate the Ei classification structure?
2) How well are the web pages classified, as judged by the users?

3.2.1 Experimental setting

The setting of the study was inspired by the INEX interactive track (Larsen et al. 2006). It comprised the following steps, with the last three ones repeated for each of four search tasks (described in section 3.2.3):

- Invitation to participate. Participants were recruited through personal contact, advertising the study on mailing lists and billboards. Each participant was awarded two cinema tickets.
- Participation consent form. Each participant was asked to sign a participation form which gave information about the study and the participant's role in it.
- Written instructions: a two-page instructions about the information retrieval system and evaluation criteria (Appendix 6).
- Pre-study questionnaire on participants' background and their previous searching and browsing experience (Appendix 2).
- Search task description.
- Search session, in which every move was logged. In addition, six participants were asked to "think-aloud" (Lewis and Rieman 1994, chapter 5), which was video-taped.
- Post-task questionnaire on participants' certainty of their decisions and general satisfaction (Appendix 3).

Each participant was first given the participation consent form and then written instructions on how to conduct the searching. After optional testing of the system and the obligatory pre-study questionnaire, the first search task was described and the first searching session began. In the searching session, the participant had to find the class where he/she thought most web pages on the topic of the search task should be. Every click in the browsing tree was logged using a home-made software. Once the class was found, the

participant was to evaluate whether web pages in that class were on the topic of the task. At least top 10 web pages (ranked by descending scores described in 2.) had to be evaluated in a row; at most 40 were offered per screen. The maximum number of web pages evaluated per class was 40, by three participants. Once the participant decided to be finished with the task, he/she filled-in a post-task questionnaire.

The language of the study was English. The vast majority of web pages were also in English. The study took place in the period between 21 May and 5 June 2007. It was conducted at one of the Department's computer rooms, one participant per computer. Six of the participants were video-taped in the researcher's office. The researcher was always present and available for help or clarifications. Each session was predicted to last one hour, as stated in the invitation to participate and the participation consent form.

3.2.2 User interface

Since, to the author's knowledge, an operative information system employing the full-scale Ei classification scheme as a browsing structure did not exist, a simple home-made interface was created (Figure 1; Appendix 7). It consisted of two parts: in the upper part of the screen, a clickable hierarchical browsing tree of the Ei classes was provided, and in the lower part web pages classified in the class clicked on were listed, in a descending relevance order based on classification scores. In the upper part of the screen also most important instructions were given; detailed instructions were provided on a separate sheet of paper (Appendix 6). For each web page there was an automatically extracted title, automatically extracted sentences ("Summary"), hyperlink to the original web page and a small evaluation form. In the evaluation form the participant was asked to judge whether the web page is on the topic of a given task. Four options were available: "correct", "partly correct", "incorrect" and "impossible for me to say".

The retrieval system and the questionnaires were pilot-tested by two additional participants. Based on their input, several minor changes on the user interface were implemented.

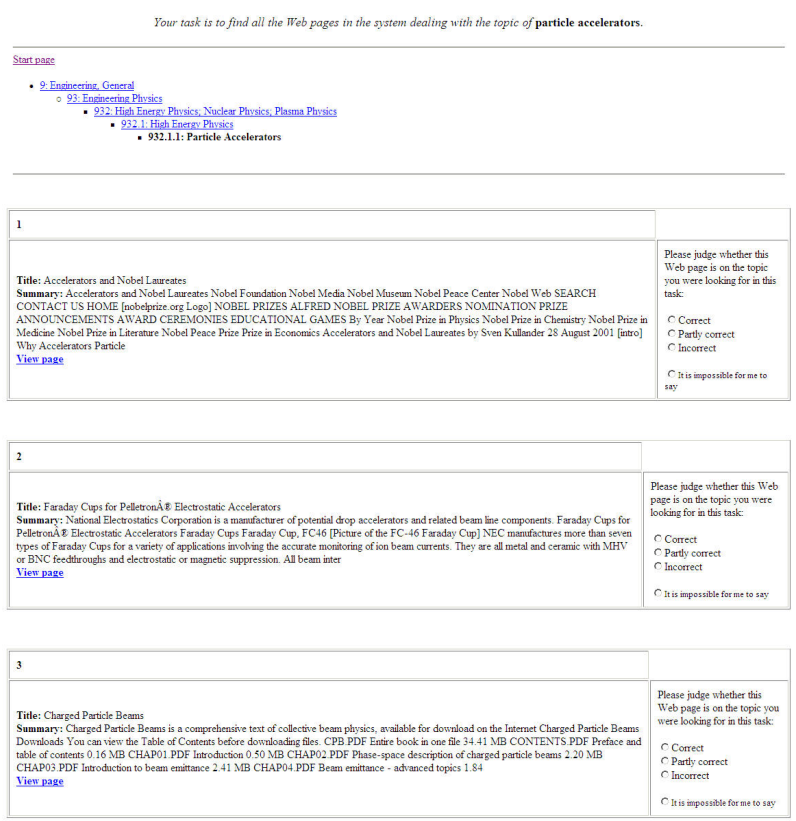


Figure 1. Screen shots of main parts of the user interface: the task definition, browsing tree and top-ranked web pages with evaluation forms.

3.2.3 Search tasks

Each participant was given four search tasks. Similar number of tasks was used in related studies (e.g. Nielsen 2004; Larsen et al. 2006). The tasks were presented in one of 24 rotated sequences, as recommended by Borlund (2003) (Appendix 5).

In each task the participant was to find web pages on an assigned topic (cf. Ingwersen and Järvelin 2005, 73). The tasks were formulated so that they would suit both the browsing and classification parts of the study. Exact formulations of the four tasks were as follows:

1) "*Your task is to find all the web pages in the system dealing with the topic of* **particle accelerators**."

2) *"Your task is to find all the web pages in the system dealing with the topic of* **magnetic instruments**."

3) "*Your task is to find all the web pages in the system dealing with the topic of* **differentiation and integration."**

4) *"Your task is to find all the web pages in the system dealing with the topic of* **professional organizations in the field of engineering**."

In order to investigate browsing and classification accuracy in a variety of contexts, each search task was from a different subject field and of a different character. Two tasks were in the basic sciences as applied in engineering, and two in general engineering; two were at fifth, and two at the fourth hierarchical level; in two topic names were the same as class captions and two were entirely different:

- **Task 1** was in the field of physics, on the topic of *particle accelerators*. The class was at the fifth hierarchical level (932.1.1), the class caption was the same as the topic name, and words from the topic name did not exist anywhere in captions of higher level classes.
- **Task 2** was in the field of instruments and measurements, on the topic of *magnetic instruments*. The class was at the fourth hierarchical level (942.3), the class caption was the same as the topic name. One word (*instruments*) from the topic name was part of captions at second (94 *Instruments and measurement*) and third (942 *Electric and Electronic Measuring Instruments*) hierarchical levels.
- **Task 3** was in the field of mathematics, on the topic of *differentiation and integration*. The class was at the fourth hierarchical level (921.2), the class caption (*Calculus*) was entirely different from the topic name, and words from the topic name did not exist anywhere in the higher level class captions.
- **Task 4** was in the field of engineering profession, on the topic of *professional organizations in the field of engineering*. The

class was at the fifth hierarchical level (901.1.1), the class caption (*Societies and Institutions*) was entirely different from the topic name, and one word (*professional*) from the topic name existed in its noun form at second (901 *Engineering Profession*) and third hierarchical levels (901.1 *Engineering Professional Aspects*).

3.2.4 Data collection and analysis methods

In order to analyse browsing (research question 1), quantitative data was collected, by logging the browsing steps and determining if correct classes were found. Participants' browsing steps and selected classes were compared against an "ideal" browsing path and an "ideal" class for each search topic. These were pre-determined by the researcher, by simply looking at the whole browsing tree and identifying what the best matching class for each search topic would be, and what the shortest route to the class would be. For each task, the shortest possible ("ideal") browsing path and class were known and the participants' steps were compared against them.

For the classification part of the study (research question 2), quantitative data were collected through correctness assessments. For each web page listed under a class selected as the most appropriate for the given task, a form was offered with four options from which to choose: "correct", "partly correct", "incorrect" and "impossible for me to say". According to the written instructions, these were to be chosen in the following cases:

- Correct – if the web page is about the topic;
- Partly correct – if the web page can be considered to be on the topic, but is mixed with other topics; and,
- Incorrect – if the web page has absolutely no relation to the topic.

The participants were asked to avoid the option "Impossible for me to say" and use it only in cases when content is not available, e.g., if a Web server is down. Apart from indicating different level of topic coverage (cf. weighted indexing in Lancaster 2003, 187-188), the three options are also analogous to related experiments were relevance of documents is assessed and three options, "relevant", "partly relevant" and "irrelevant" are made available (e.g. Nielsen 2004).

And then, we looked at evaluations at both ideal classes for each tasks and others selected by users as *the* class for a task.

For both parts of the study, a post-task questionnaire on participants' certainty of their decisions and general satisfaction was to be filled-in after each task. In addition, clarifying, qualitative data were collected for six participants by observation based on the "think-aloud" protocol. After completing their search tasks, all the participants were asked if they had any comments, which were also recorded and analyzed. At this stage, comments were received from 12 participants. Through post-task questionnaires, 31 comments were collected, submitted by 16 participants.

Based on post-task questionnaires, relation between browsing and classification accuracy was investigated. For each question, 159 answers were collected. Correlations were calculated between every pair of questions that could indicate an influence of classification accuracy on browsing and vice versa. As the answers were of ordinal variable type, Spearman's rank correlation was used (Vaughan 2001, 140-143). The calculations were conducted in Matlab, where the original data were given as input to a formula for direct calculation of Spearman's *rho* values.

Information on participants' background and their previous searching and browsing experience were collected in a pre-study questionnaire (Appendix 2).

3.2.5 Participants

There were 40 participants, students and researchers in the field of engineering. The participants were selected randomly: they were the first 40 people who agreed to take part in the study.

Information on participants' background and their previous searching and browsing experience was collected through the pre-study questionnaire (Appendix 2). All the participants had very good English skills and a lot of online searching experience. The majority (87.5%) were between 20 and 30 years old, male (88%), taking or having completed their Master's degree in the field of computer engineering (86%), with a very good or excellent knowledge of English (85%). All of them had at least four years of online searching experience and 90% of them claimed they generally found what they were looking for on the World Wide Web. They used search engines once or twice a week (3.8 on a scale from 1 to 5

where 1 stands for "Never", 2 for "Once of twice a year", 3 for "Once or twice a month", 4 for "Once or twice a week", and 5 for "One or more times a day"), in significant contrast to professional information services: they used library catalogues and Lund University's service providing free access to commercial databases once or twice a year (1.8 the former, 1.6 the latter), and engineering-specific database Compendex (also freely available for Lund University's students and researchers), hardly ever (1.1). On average they used hierarchical directory-style browsing of e.g., search engines or other information databases once or twice a month (2.5).

4 Results

4.1 Browsing

How good the Ei classification scheme is for browsing was measured by counting the number of participants finding the pre-defined ideal class in the four tasks as well as the number of classes visited before reaching the final class. Comments received from the participants were also analyzed. Influence of automated classification accuracy on browsing decisions is examined in section 4.3.

4.1.1 Analysis based on browsing steps

As described earlier (section 2.4.1), the participants were instructed to first find the class they considered most appropriate for the task at hand, and then evaluate whether web pages listed under it were on the topic of the task. However, in several cases web pages in more than one class were evaluated.

On average for all the four tasks, the majority (29 out of 40 participants) found the right class. Approximately two other classes per task were considered correct by at least two participants.

In Table 1 responses from the post-task questionnaire related to browsing are presented. The results are reported in separate columns for those who found the right class ("right class found") and for those who did not ("right class not found"). On a scale from 1 to 3, where 1 stands for "not at all", 2 for "somewhat" and 3 for "very", participants who found the right class reported on average for all the four tasks that it was easy (2.3) to find the right class ("easycat") and that they were rather certain they found it ("certaincat") (2.6).

Those who did not find the right class were less sure they found it (1.7) and for them it was less easy to find an appropriate class (1.9). Both groups reported that they were somewhat familiar with the topics (2.1).

Table 1. Results from post-task questionnaires related to browsing.

	right class found			right class not found		
task	easycat	certaincat	familiar	easycat	certaincat	familiar
1	2.2	2.9	1.8	1.8	2.0	1.9
2	2.7	3.0	1.9	2.0	1.6	2.2
3	2.4	2.5	2.7	1.7	1.4	2.2
4	1.9	2.0	2.0	1.9	1.7	1.9
average	2.3	2.6	2.1	1.9	1.7	2.1

For two of the tasks the ideal browsing path takes four steps and for other two five steps (see Appendix 4). On average the participants who found the right class took 15 steps; all participants took 16 steps. Browsing in each task is discussed separately below.

Task 1: *particle accelerators* (932.1.1)

In Task 1 the shortest possible number of steps was five (including the step in which one decided that he/she reached the right class). There were six participants who followed this shortest path. The majority (21) took up to 15 steps to come to the ideal class. On average participants who found the ideal class took 16 steps; all participants took 19 steps. An example of a 15-step sequence taken by one participant is given below:

93 → 932 → 932.2 → 932 → 93 → 931 → 931.3 → 93 → 9 → 94 → 9 → 93 → 932 → 932.1 → 932.1.1

As seen from Table 2, the first browsing step the 70% majority took was correct – 93 *Engineering Physics*. Eight participants chose 94 *Instruments and Measurement* and four 90 *Engineering, General*. Of those who took the correct first step, the majority chose the correct second step (82.1%). The weakest point was the third ideal step, choosing a specific class within the broad area of 932 *High Energy Physics; Nuclear Physics; Plasma Physics* – only 43.5% made the right decision. Most of them were not sure if *particle accelerators* belonged to 932.1 *High Energy Physics*, 932.2 *Nuclear Physics* or

932.3 *Plasma Physics*. Since the three classes at the fourth hierarchical level seem quite clear, each representing one of the three concepts from their broader class 932, this could be attributed to the fact that the participants were less than somewhat familiar with the topic of the task (1.8 and 1.9, see Table 1).

Table 2. Ideal browsing steps for Task 1. Percentage in "step taken by" is calculated in relation to the number of participants taking the preceding ideal step.

	ideal sequence	step taken by
step 1	93 Engineering Physics	70.0%
step 2	932 High Energy Physics; Nuclear Physics; Plasma Physics	82.1%
step 3	932.1 High Energy Physics	43.5%
step 4	932.1.1 Particle Accelerators	80.0%
step 5	confirmed	87.5%

There were 31 participants who found the right class. Six other classes were deemed correct by at least one participant. Classes chosen by at least two participants were the following:

- 932 *High Energy Physics; Nuclear Physics; Plasma Physics*, which is a correct but not the most specific class in the hierarchy. Considering a broader class as correct, especially when relevant resources were discovered, is a defendable error;
- 931.3 *Atomic and Nuclear Physics*, which is wrong. This class's broader class 931 *Applied Physics Generally* is also wrong, but the first step class (93 *Applied Physics Generally*) is correct;
- 932.2.1 *Fission and Fusion Reactions*, which is wrong. This class's broader class 932.2 *Nuclear Physics* is also wrong, but the first and second step classes are correct.

Thus, all the participants chose the correct second hierarchical level, 93 *Engineering Physics*. The reason why they chose different classes within physics could be attributed to the fact that they were less familiar with the topic. The participants who found the right class were very certain that they found it (2.9). This could be partly explained by the fact that the class caption was the same as the topic name. Participants who did not find the right class were only

somewhat sure (2.0); for them also finding the class was less easy (1.8 on the scale) than for those who found it (2.2). Both groups were a bit less than somewhat familiar with the topic from before (1.9 for those who did not find it and 1.8 for those who did).

Task 2: *magnetic instruments* (942.3)

In Task 2 the shortest possible number of steps was four (including the step in which one decided that he/she reached the right class). There were 18 participants who followed this shortest path. The majority (24 of 35) took up to six steps to find the ideal class; on average eight steps were taken. An example of a six-step sequence taken by one participant is given below:

94 → 943 → 943.3 → 94 → 942 → 942.3

As seen from Table 3, more than half of the participants (60%) took the correct first step; 13 participants chose 93 *Engineering Physics* and 3 chose 90 *Engineering, General*. Of those who took the correct first step, the majority chose the correct second step; those who did not, chose 943 *Mechanical and Miscellaneous Measuring Instruments* which could be, because of the word *miscellaneous*, considered justifiable. The majority chose the correct third step and also confirmed the class to be final.

Table 3. Ideal browsing steps for Task 2.

	ideal sequence	step taken by
step 1	94 Instruments and Measurement	60.0%
step 2	942 Electric and Electronic Measuring Instruments	79.2%
step 3	942.3 Magnetic Instruments	94.7%
step 4	confirmed	94.4%

There were 35 participants who found the right class. Eight other classes were deemed correct by at least one participant. The class chosen by at least two was 931.1 *Mechanics*. This class is incorrect, although its higher levels 931 *Applied Physics Generally* and 93 *Engineering Physics* could be considered correct to some degree. The participants who found the right class were very certain that they found it (3.0 on the scale, as seen from Table 1). This could be partly explained by the fact that the class caption was the same as

the topic name. Participants who did not find the right class were far less sure (1.6). For them also finding the class was less easy (2.0) than for those who found it (2.7). Both groups were somewhat familiar with the topic from before (2.2 for those who did not find it, and 1.9 for those who did).

Task 3: *differentiation and integration* (921.2)

In Task 3 the shortest possible number of steps was four (including the step in which one decided that he/she reached the right class). Eight participants followed this shortest path. The majority (22) took up to 14 steps to come to the ideal class; on average, 15 steps were taken by those was found the ideal class, and 18 by all. An example of a 14-step sequence taken by one participant is given below:

92 → 921 → 921.2 → 921 → 921.3 → 921 → 921.5 → 921 → 92 → 922 → 922.2 → 92 → 921 → 921.2

Table 4. Ideal browsing steps for Task 3.

	ideal sequence	step taken by
step 1	92 Engineering Mathematics	85.0%
step 2	921 Applied Mathematics	91.2%
step 3	921.1 Calculus	51.6%
step 4	confirmed	37.5%

As seen from Table 4, the first browsing step the majority took was correct (85%) – 92 *Engineering Mathematics*. Of the remaining six, two went to 90 *Engineering General*, two to 91 *Engineering Management* and two to 94 *Instruments and measurement*. Of those who took the correct first step, the majority chose the correct second step (91.2%). A weak point was the third step, choosing a specific class within the broad area of 921 *Applied Mathematics*: only a tight majority picked the right class (51.6%). The weakest point was coming to the right class – only 37.5% decided it was the ideal class, while others went one level up, to class 921 *Applied Mathematics*. The author believes that the reason for the latter two could be attributed to the fact that the class caption was entirely different from the topic name. Also, the participants not finding the ideal class were not very familiar with the topic (2.2 on the scale, see Table 1). Another reason could be that the participants were mainly Swedish and did not take mathematics courses in English: the

English word *calculus* shares little more than etymology with the Swedish word *kalkyl* in common usage (the corresponding Swedish word is *analys*).

There were 28 participants who found the right class. Nine other classes were deemed correct by at least one participant. Classes chosen by at least two of them were the following:

- 921.6 *Numerical Methods* which is wrong, but its broader class 921 *Applied Mathematics* is correct; and
- 921.4 *Combinatorial Mathematics* which is wrong, but also has its broader class 921 *Applied Mathematics* correct.

Thus, all participants have chosen correct second and third hierarchical levels (92 *Engineering Mathematics* and 921 *Applied Mathematics*). The author believes that the reason why some of them did not chose the ideal class at the fourth hierarchical level could be the same as above for the fourth browsing step. The participants who found the right class were between somewhat and very certain they found the right class (2.5 on the scale, as seen from Table 1). The certainty level is high, although a bit lower than in the first two tasks; this could be partly explained by the fact that the class caption was different from the topic name. Participants who did not find the right class were rather unsure (1.4 on the scale). For them finding the class was also less easy (1.7) than for those who found it (2.4). The group who found the ideal class was quite familiar with the topic from before (2.7), while the group who did not find the ideal class was somewhat familiar (2.2).

Task 4: ***professional organizations in the field of engineering*** **(901.1.1)**

In Task 4 the shortest possible number of steps was five (including the step in which one decided that he/she reached the right class). There were five participants who followed this shortest path. On average 19 steps were taken by those who found the ideal class, and also by all. An example of a 19-step sequence taken by one participant is given below:

9 → 91 → 912 → 912.2 → 9 → 90 → 901 → 901.1 → 901.1.1 → 901.1 → 90 → 901 → 9 → 91 → 913 → 9 → 90 → 901 → 901.1 → 901.1.1

This example also demonstrates how the participant found the right class but continued looking at other classes until he/she made the decision he/she was most sure of. Reasons why this was the case need further investigation.

As seen from Table 5, the first ideal browsing step was taken by half of the participants; the second half chose class 91 *Engineering Management*, which can be, because of the nature of the topic, considered to be partly correct. All those who took the ideal first step, chose the correct second step. While the majority also chose the correct third and fourth step, only half of those who came to the right final class realized it was the right one. The reason for this could be that the class caption was entirely different from the topic name.

Table 5. Ideal browsing steps for Task 4.

	ideal sequence	step taken by
step 1	90 Engineering, General	50.0%
step 2	901 Engineering Profession	100.0%
step 3	901.1 Engineering Professional Aspects	70.0%
step 4	901.1.1 Societies and Institutions	71.4%
step 5	confirmed	50.0%

There were 20 participants who found the right class. Nine other classes were deemed correct by at least one participant. Classes chosen by at least two participants were the following:

- 912.2 *Management*, which could be considered correct, although there is a class that describes the topic better;
- 901.1 *Engineering Professional Aspects*, which is correct but not the most specific class that can be found in the classification scheme. Considering a broader class as correct, especially when relevant resources were discovered, is a defendable error;
- 901.3 *Engineering Research*, which could be considered correct, although there is a class that better describes the content; and
- 912.1 *Industrial Engineering*, which is also somewhat related to the topic of the task.

Reasons for choosing these different classes could be attributed to the fact that the topic of this task can belong to more than one strict

class. Participants who found the right class were somewhat sure they found the ideal class (2.0 on the scale), less than for the other three tasks. This could be explained by the nature of the topic, the fact that the class caption was entirely different from the topic name, and also by the fact that the top ranked web pages in this class were evaluated to be between partly correct and incorrect, worse than in any other task. Comments by participants confirmed that this task was more ambiguous than others, e.g., "This one seems like a broad topic, it could include anything that works with engineering professionally, quite a lot" or "All companies are also at least semi-professional so I feel I can pretty much go to any category and still find things."

Participants who did not find the right class were less sure they reached it (1.7). Both groups reported that finding the right class was somewhat easy (1.9), but on average it was more difficult than in any other task. The familiarity with the topic was reported to be medium by both groups (2.0 where correct and 1.9 where incorrect).

4.1.2 Analysis based on comments

Since comments were not obligatory, this section could provide only indications. Several participants said that they preferred searching to browsing, and some of them provided reasons such as the following ones:

- "I prefer searching because here you always need to go up and down";
- "I think you need to know something about the topic before you start using the tree; the hierarchy helps if you know a little bit, but if you have no clue…"

These issues could be dealt with by enhancing the hierarchical interface by, for example, adding a search box for words in class captions with synonym search (easily provided for Ei since the classes are mapped to thesaurus terms), and returning the hierarchical tree expanded around the class in which caption the search term is found. If a term searched for is contained in, say, two different contexts, returning the two hierarchical trees would serve as a disambiguation device and help the user chose the exact meaning he/she is looking for. The suggestion to allow for searching for words from class captions was provided by quite a few

participants as well. Another suggestion they made was to provide some kind of a support for explaining classes, e.g., describing each caption with what it contains or adding an expand box with what is below the class.

Several participants expressed that they liked browsing. One said that he/she had never tried browsing before, but by the time he/she arrived to the last task, was happy with it and the experience had been pleasing. Another said he/she didn't really like the hierarchy, but believed it could be useful once used to it.

One participant complained that for two classes it was not obvious that they would be where they were; however, the fact that he/she did find them shows that it is possible to find one's way through the Ei structure although every individual would probably structure subjects differently. Two participants commented that class 94 *Instruments and Measurement* represents an application area, while others are scientific, and that there are overlaps in subjects between them – e.g., *instruments* could be part of physics as well. Overlaps in topic representations in the classification scheme were reported as an issue by others as well, but could be dealt with a search entry into a synonym list of class captions (already part of the Ei thesaurus). Moreover, overlaps in topics exist in disciplines themselves and a good classification scheme should reflect such overlaps. Another suggestion was to exclude the words *general* and *engineering* from captions of second-level classes as he/she perceived them as redundant. Several wanted to see more specific subclasses; the reason could be that there were too many web pages in one class, more than 40 in most of them.

Most of these suggestions had been already implemented in another web-based service where browsing and browsing support features were shown to be heavily used, more than searching (Koch et al. 2006).

4.2 Automatically assigned classes

In the second part of the study, correctness of automatically assigned classes was analyzed. The most common approach is by comparing automatically assigned classes against human-assigned ones. For web pages there are few collections with human-assigned classes. To the author's knowledge, there is one maintained collection using Ei classes, Intute subject gateway on engineering (Intute Consortium 2006). However, web pages in this collection are classified mainly

into top hierarchical levels, or six classes altogether: 900, 910, 920, 930, 940, 901.2. Since the aim was to study how the algorithm performs also at third, fourth and fifth hierarchical levels, this collection did not suffice. Also, the number of web pages in the Intute branch on general engineering is about 1,500, too small for classification evaluation. Moreover, the problem of documents' "aboutness" has been much discussed in the literature and the need for evaluating automated classification by end users has been proposed but seldom conducted (see Golub 2006a).

4.2.1 Automatically assigned classes against human-assigned ones

Since web pages were automatically crawled and classified, no pre-existing human-assigned classes were available. Of the 518 web pages that were classified both by the Intute subject gateway (Intute Consortium 2006) and by the algorithm, 320 of them were put in the same class as in the Intute subject gateway. Because of the small sample and because Intute has most web pages only at the top two hierarchical levels, further comparison was not conducted.

4.2.2 Automatically assigned classes as judged by the user study participants

Table 6 shows the number of web pages evaluated in different tasks. In ideal classes there were on average 40 different web pages evaluated per task with 10 different participants evaluating each web page. There were in total 36 evaluations or 23 different web pages that were deemed as "impossible for me to say" by at least one participant. Four web pages were deemed as "impossible for me to say" by two or more participants. One of them was "under construction", two had very little text, and one was extremely long (178 pages if printed). Some of them were also not in English. Because of the small number of "impossible to say" decisions, they were not counted in tables following this one.

Table 6. Number of evaluations and evaluated web pages.

	task 1	task 2	task 3	task 4
Number of evaluations	620	644	593	617
Number of different Web pages evaluated	170	154	172	176
Number of evaluations in the ideal class	392	497	417	297
Number of different web pages in the ideal class evaluated	40	31	40	40

Table 7 shows for each task the ideal class and averaged evaluations for 1) all the evaluated web pages, and 2) top 10 ranked web pages that were at the same time most frequently evaluated. The scale used in this part of the study was from 1 to 3, where 1 stands for "correct", 2 for "partly correct" and 3 for "incorrect". The differences between all and top 10 evaluated web pages do not seem to be significant, but are a little better for the top 10 pages, which is in accord with the fact that the higher ranked ones should be more correct. The fact that the differences are not very significant implies that the algorithm performs equally well for resources listed further down on the page. Neither were there significant differences when comparing evaluations made by participants who were very certain that their judgements were accurate ("only when certain") against the average of all participants ("different certainty").

Table 7. Correctness of automatically assigned ideal classes.

		different certainty		only when certain		
task	correct class	for all evaluated web pages	for top 10	for all evaluated web pages	for top 10	average
1	932.1.1	1.9	1.8	1.8	1.8	1.8
2	942.3	2.0	1.8	2.1	1.9	2.0
3	921.2	2.0	2.0	2.0	2.0	2.0
4	901.1.1	2.6	2.5	2.6	2.5	2.6
average	n/a	2.1	2.0	2.1	2.1	2.1

As with browsing, the performance of the algorithm as judged by the participants differs between the tasks. Based on 1,603 evaluations of 151 different web pages in ideal classes (last two rows in Table 6), the top ranked web pages in each of the four classes were on average deemed partly correct (2.1 on the scale). Best classification was achieved for Task 1 (1.8), and the worst one for Task 4 (2.6). This was also confirmed by comments given by several participants who said that web pages in Task 1 seemed quite good but that they were confused with web pages in Task 4.

Table 8. Correctness of automatically assigned other classes.

task	other classes	number of participants choosing the class	number of evaluated web pages	for all evaluated Web pages (different certainty degrees)
1	932	5	104	2.3
	931.3	5	54	2.6
	932.2.1	3	30	2.7
2	931.1	3	44	2.2
3	921.6	4	50	2.8
	921.4	2	23	2.4
4	912.2	8	119	**2.3**
	901.1	7	108	**1.9**
	901.3	4	48	**2.1**
	912.1	2	11	2.2

While the majority chose the ideal class, it was also important to see how the participants who chose other classes as correct ones evaluated web pages put in those classes. Table 8 reports evaluations of web pages for non-ideal classes that were chosen by at least two participants. (In the instructions it was clearly written that one should evaluate correctness of web pages in relation to the task at hand, and not the caption of the chosen class. However, think-aloud sessions have shown that the latter could also have been the case.)

Table 9. Results from post-task questionnaires related to correctness of classes and general experience.

	right class found			right class not found		
task	easyclass	certainclass	experience	easyclass	certainclass	experience
1	2.3	2.4	2.4	1.9	2.3	1.7
2	2.3	2.4	2.4	2.0	2.0	1.6
3	2.3	2.4	2.3	1.6	1.7	1.4
4	2.4	2.3	2.0	2.0	2.0	1.8
average	2.3	2.4	2.3	1.9	2.0	1.6

In Tasks 1, 2 and 3 topics of web pages in non-ideal classes were deemed more wrong than in the ideal class. In Task 4, however, evaluations of web pages in the non-ideal classes were deemed more correct than in the ideal class, the best one being its broader class, 901.1. This could be partly explained by the fact that

this topic could be considered to fit in several classes (cf. Task 4 in 3.1.1), and as such is harder to automatically classify.

Table 9 shows results from post-task questionnaires related to correctness of classes as well as general experience ("experience"). The results are reported separately for those who found the right class ("right class found") and for those who did not ("right class not found"). On a scale from 1 to 3, where 1 stands for "not at all", 2 for "somewhat" and 3 for "very", participants who found the right class reported on average for all the four tasks that it was easy (2.3) to decide whether the web pages in the selected class were on the topic of the task ("easyclass") and that they were certain of their evaluation indicating whether web pages in the class were on the topic of the task ("certainclass") (2.4). On a scale from 1 to 3 where 1 stands for "frustrating", 2 for "neutral" and 3 for "pleasing", this group deemed the whole experience to be between neutral and pleasing (2.3). Those who did not find the right said that it was less than somewhat easy to decide whether web pages in the selected class were on the topic of the task (1.9) and were less certain of their evaluation indicating whether web pages in the class were on the topic of the task (2.0). This group deemed the whole experience to be somewhere between frustrating and neutral (1.6).

The fact that evaluations between the four tasks differed is in compliance with previous results of the algorithm's performance, based on a pre-classified collection of research abstracts, where it was shown that certain classes have better performance than others (Golub et al. 2007). Table 10 presents precision and recall for the four classes, but measured on the collection of paper abstracts. While precision is almost total for all the four classes, recall is weakest for class 901.1.1 (Task 4), which can be attributed to the fact that only one term exists for this class on the term list. Also, most terms designating the other three classes are rather field-specific and thus less ambiguous than the term designating class 901.1.1.

Table 10. Performance for the four classes on the collection of paper abstracts (described in Golub et al. 2007), with the same parameters as in this study.

class	term	instances found	precision	recall
932.1.1	accelerators @and betatron	1	0.97	0.08
	electron sources	1		
	particle beams	3		
	accelerators @and electrostatic	1		
	accelerators @and magnets	4		
	accelerator magnets	1		
	accelerators @and targets	4		
	electron accelerators	2		
	accelerators @and synchrotron	3		
	synchrotron x-ray radiation	2		
	storage rings	11		
	particle beam dynamics	2		
	synchrotron ultraviolet radiation	1		
	linear accelerators	2		
942.3	gradiometers @and magnetic	7	1	0.12
	compasses	1		
	magnetometers	35		
	fluxmeters	2		
921.2	kinetic theory of gases	1	1	0.03
	differential relational calculus	1		
	differentiation @and calculus	1		
	integral equations	12		
	maxwell's equations	9		
901.1.1	societies @and institutions	1	1	0.001

4.2.1 The problem of "aboutness"

The challenge of documents' aboutness has been much discussed in the literature. In practice this problem is reflected in low inter- and intra-indexer consistency (see Olson and Boll 2001, 99-101). Markey (1984) reviewed 57 indexer consistency studies and reported that consistency levels range from 4% to 84%, with only 18 studies showing over 50% consistency. In this study there were cases when one web page was at the same time evaluated as correct, incorrect and partly correct. For top 10 pages in all the 4 tasks, 26 pages were at least once evaluated as correct, incorrect and partly

correct; 11 pages were evaluated with two different values; and, only 3 pages were evaluated with the same value. Based on "think-aloud" sessions and participants' comments, reasons behind their decisions and such big discrepancies between evaluations were analyzed:

- There were three cases implying that some participants used only summaries, in spite of the clearly provided instruction "In order to evaluate whether the web page is about the topic given in the task, please open and look at the page ('View page') because it would be incorrect to judge only based on the Title and Summary". An example of a comment implying they used only summaries was "…this task was a little bit hard to extract information from web sites because the researching subject wasn't described in web sites' summaries…"
- Most differences between how people judged one and the same web page occurred due to mixing web page's topic with its genre. There are many web pages offering not just factual information on a topic, but also (or instead) describe a related software, provide a search engine, or sell products like magnetic instruments. While according to instructions, "incorrect" was supposed to be chosen only when the web page had absolutely no relation to the topic, some judged commercial web pages as incorrect. E.g., a participant leaving a comment "a lot of commercial websites and almost no didactic website" on average evaluated web pages in Task 3 as 2.4 (between partly correct and incorrect), while the average for that task was 2.0 (partly correct).
- Task interpretation. Some evaluated a web page based on whether it had anything to do with the topic at hand, as instructed, while others based it on their own task interpretation and introduced criteria such as usefulness and quality. E.g., a participant saying "this page could be useful if we know something but for a beginner, no" judged that web page as incorrect. Or, a web page providing only a definition on compasses was judged by one participant as partly correct because "it provides only superficial information".

These findings illustrate the problem of aboutness in general and at the same time make the results of the study as to classification accuracy somewhat dubious.

4.3 Browsing and classification accuracy

Finding one's way through the browsing tree tends to be related to whether web pages are classified in appropriate classes. This is supported by participants' comments:

- "I am not sure I found the right category for professional organizations". Actually this participant did find the right class, but was unsure of that because web pages in the class were by that participant evaluated as mostly incorrect (2.6 on average). This topic belongs to Task 4 where web pages were by all deemed to be least correct.
- "Most of them are incorrect so I think I chose the wrong category". This participant really did not find the right class, so in this case web pages were correctly indicating that he/she should look for another class.
- Another participant said: "It is easier with search bar, this is quite frustrating; lot's of useless web pages". In this case the right class was not found, but the comment indicates how the correctness of automated classification influenced the participant's preference for searching.

Based on each participant's post-task questionnaire answers for every task (159 per question) the following significant correlations were recognized with probability above 95% (Sheskin 2000, Table A18):

- Certainty that the right class was found and certainty of one's evaluation whether web pages in the found class were on the topic (Spearman correlation coefficient is 0.33).
- Easiness to find the right class and easiness to decide whether web pages in the found class were on the topic: the correlation is (Spearman correlation coefficient is 0.31).

- Certainty that the right class was found and easiness to decide whether web pages in the found class were on the topic (Spearman correlation coefficient is 0.35).

5 Conclusions

The study was to investigate performance of an automated classification algorithm on a collection of engineering web pages, in the context of hierarchical browsing. Two major research questions were whether users were able to navigate the Ei classification structure and how accurately were web pages in a certain class classified. The study involved 4 tasks and 40 participants. The participants had a very good knowledge of English, at least four years of online searching experience, frequent usage of search engines, once or twice a month they used hierarchical browsing and were generally finding the desired information.

The study showed that the Ei classification scheme is generally well suited for browsing. The majority of participants found the right class, they reported that it was quite easy finding it and were quite certain they found the right class. Also, those who found ideal classes deemed the whole user study experience between neutral and pleasing. This was the case in spite of the fact that the participants on average made 15 steps to reach the ideal class, while the shortest browsing path would take only 5. However, the number of browsing steps needs to be put in relation to at least two other factors: examples have shown that some participants have systematically looked at a number of other classes to make sure they are reached the most appropriate one; and, the hierarchical tree showed only the path to the last class clicked on, preventing distant jumps. Other possible factors could be inadequacy of the classification scheme and participants' unfamiliarity with it. Inadequacy of the classification scheme was indicated as to the following: captions contain redundant words like *engineering*; division of subject areas is not very logical: basic sciences, mathematics and physics, as applied in engineering, are at the same hierarchical level as *Instruments and Measurement* and *Engineering, General*; and, class *Engineering, General* contains a mixture of topics such as engineering profession, graphics and libraries. Exact reasons behind taking longer paths than required need to be further studied.

Majority of the participants selected correct second and third hierarchical level classes in all the four tasks. More wrong classes were chosen at the fourth hierarchal level, which could be attributed to: 1) participants' unfamiliarity with the subject at the required specific level; and, 2) class captions being different from topic names. The latter is confirmed by the fact that approving the right class arrived at as correct was problematic in the two tasks in which class captions were entirely different from topic names. For those two tasks also lower certainty levels were reported. The lowest certainty level was obtained for one of those two tasks in which web pages were judged as more incorrect than in other tasks, another possible contributing factor. Also, the classification scheme could have for some reason been inappropriate. Exact reasons need to be further investigated.

Top ranked web pages in each of the four classes were on average deemed partly correct. A major problem with determining whether a web page is in the right class or not is that there were large differences among participants in their judgements – a number of web pages were evaluated as correct, partly correct and incorrect by different participants. A major reason is probably the reported problem of aboutness" and related subjectivity in deciding which topic a document is dealing with. Other factors were also recognized in the study:

- Some participants, despite the instructions, based their evaluations only on summaries, instead of full-text web pages;
- Others interpreted tasks more narrowly and evaluated web pages based also on other criteria such as quality and usefulness;
- It was hard to evaluate web pages' topicality when there was hardly any or very much text.

As with browsing, evaluations between the four tasks differed. This is in compliance with previous results of the algorithm's performance, based on a pre-classified collection of research abstracts, where it was shown that certain classes have better performance than others (Golub et al. 2007).

Although the classification of web pages was on average judged as only partly correct, and while there is evidence that correct placement of web pages and browsing success are related, the majority of the participants were able to navigate the Ei

classification structure well and deemed the whole experience to be between neutral and pleasing.

Several improvements for browsing have been identified, most of which had been already implemented, and heavily used, in a browsing-based Web service called Renardus (Koch et al. 2006):

- Describing class captions and/or listing their subclasses from start;
- Allowing for searching for words from class captions with synonym search (easily provided for Ei since the classes are mapped to thesauri terms);
- When searching for class captions, returning the hierarchical tree expanded around the class in which caption the search term is found; and
- Because automatically-produced summaries could be misleading, avoid presenting them until their quality is sufficiently improved.

The need for several improvements of classification schemes was indicated:

- Follow consistent division principles when building classification structures;
- Modify captions so that they better reflect concepts they represent;
- Allow for a larger entry vocabulary, which would directly help both finding the ideal class fast and improve recall in automated classification.

Further research should include determining other reasons behind browsing failures in the Ei classification scheme. It should also deal with problems of disparate evaluations of one and the same web page. This could mean applying a better user study methodology that would help participants make more homogeneous decisions. Another way would be to harvest web pages that are more uniform in terms of quality or genre (see, for example, Custard and Sumner 2005; Nicholson 2003), or to design more narrowly specified search tasks for a certain purpose.

Acknowledgments

Many thanks to Birger Larsen who provided suggestions on the user study design, based on the interactive track at INEX2005. Thanks also to Anders Ardö and Traugott Koch whose comments on earlier versions helped significantly improve the paper. This work was supported by the IST Programme of the European Community under ALVIS (IST-002068-STP).

References

Ardö, A. (2007) "Focused crawler: Combine system homepage", available at: http://combine.it.lth.se/ (accessed 3 September 2007).

Borlund, P. (2003), "The IIR evaluation model: a framework for evaluation of interactive information retrieval systems", Information Research, Vol. 8 No. 3, paper no. 152, available at: http://informationr.net/ir/8-3/paper152.html (accessed 5 September 2007).

Custard, M. and Sumner, T. (2005), "Using machine learning to support quality judgments", D-Lib Magazine, October Vol. 11 No. 10, available at: http://www.dlib.org/dlib/october05/custard/10custard.html (accessed 3 September 2007).

Engineering Information (2006), "Compendex", Engineering Information, Elsevier, available at: http://www.ei.org/databases/compendex.html (accessed 30 June 2006).

Golub, K. (2006a), "Automated subject classification of textual Web documents", Journal of Documentation, Vol. 62 No. 3, pp. 350-371.

Golub, K. (2006b), "The role of different thesauri terms in automated subject classification of text", *Proceedings of the International Conference on Web Intelligence, Hong Kong,* pp. 961-965.

Golub, K., and Ardö, A. (2005), "Importance of HTML structural elements and metadata in automated subject classification", *Proceedings of the 9th European Conference on Research and Advanced Technology for Digital Libraries, Vienna, Austria, 18-23 September,* pp. 368-378.

Golub, K., Hamon, T., and Ardö, A. (2007), "Automated classification of textual documents based on a controlled vocabulary in engineering", *submitted to Knowledge Organization*.

Ingwersen, P., and Järvelin, K. (2005), The turn: integration of information seeking and retrieval in context, Springer, Dordrecht.

Intute Consortium (2006), Intute: Science, engineering and technology – engineering general, available at: http://www.intute.ac.uk/sciences/cgi-bin/browse.pl?id=25682 (accessed 30 August 2007).

Jain, A.K., Murty, M.N., and Flynn, P.J. (1999), "Data clustering: a review", ACM Computing Surveys, Vol. 31 No. 3, pp. 264-323.

Koch, T., and Zettergren, A.-S. (1999) "Provide browsing in subject gateways using classification schemes", *EU Project DESIRE II*, available at http://www.mpdl.mpg.de/staff/tkoch/publ/class.html.

Koch, T., Golub, K., and Ardö, A. (2006), "Users browsing behaviour in a DDC-based Web service: a log analysis", Cataloging & Classification Quarterly, Vol. 42 No. 3/4, pp. 163-186.

Lancaster, F.W. (2003), *Indexing and abstracting in theory and practice,* 3rd ed, Facet, London.

Larsen, B., Malik, S. and Tombros, A. (2006), "The interactive track at INEX2005", Advances in XML Information Retrieval and Evaluation, 4th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2005, Dagstuhl Castle, Germany, November 28-30, 2005, Revised Selected Papers, pp. 398-410.

Lewis, C. and Rieman, J. (1994), "Task-centered user interface design: a practical introduction", available at: http://www.hcibib.org/tcuid/ (accessed 3 September 2007).

Markey, K. (1984), "Interindexer consistency tests: a literature review and report of a test of consistency in indexing visual materials", Library & Information Science Research, 6, pp. 155-77.

Milstead, J, ed. (1995), Ei thesaurus, 2nd ed., Engineering Information Inc., Hoboken, NJ.

Moens, M.-F. (2000), Automatic indexing and abstracting of document texts, Kluwer, Boston.

Nicholson, S. (2003), "Bibliomining for automated collection development in a digital library setting: using data mining to discover Web-based scholarly research works", Journal of the American Society for Information Science and Technology Vol. 54 No. 12, pp. 1081-1090.

Nielsen, M.L. 2004, "Task-based evaluation of associative thesaurus in real-life environment", *Proceedings of the ASIST 2004 Annual Meeting, Providence, Rhode Island, November 13-18*, pp. 437-447.

Olson, H.A., and Boll, J.J. (2001), Subject analysis in online catalogs, 2nd ed., Libraries Unlimited, Englewood, CO.

Schwartz, C. (2001), Sorting out the Web: approaches to subject access, Ablex, Westport, CT.

Sheskin, D.J. (2000), Handbook of parametric and nonparametric statistical procedures, 2nd ed., Chapman & Hall, Boca Raton etc.

Sebastiani, F. (2002), "Machine learning in automated text categorization", ACM Computing Surveys, Vol. 34 No. 1, pp. 1-47.

Soergel, D. et al. (2004), "Reengineering thesauri for new applications: the AGROVOC example", Journal of Digital Information, Vol. 4 No. 4, Article no. 257, available at: http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Soergel/.

Vaughan, L. (2001), Statistical methods for the information professional: a practical, painless approach to understanding, using, and interpreting statistics, Information Today, Inc., Medford, NJ.

Vizine-Goetz, D. (1996), "Using library classification schemes for internet resources", OCLC Internet Cataloging Project Colloquium, available at: http://webdoc.sub.gwdg.de/ebook/aw/oclc/man/colloq/v-g.htm (accessed 4 April 2006).

Appendix 1. Hierarchical tree for the class 9 (Engineering, General) of the Engineering Information classification scheme.

All but the stricken-through classes were used in the study.

90: Engineering, General
- 901: Engineering Profession
  - 901.1: Engineering Professional Aspects
    - 901.1.1: Societies and Institutions
  - 901.2: Education
  - 901.3: Engineering Research
  - 901.4: Impact of Technology on Society
- ~~902: Engineering Graphics; Engineering Standards; Patents~~
  - ~~902.1: Engineering Graphics~~
  - ~~902.2: Codes and Standards~~
  - ~~902.3: Legal Aspects~~
- 903: Information Science
  - 903.1: Information Sources and Analysis
  - 903.2: Information Dissemination
  - 903.3: Information Retrieval and Use
  - 903.4: Information Services
    - 903.4.1: Libraries

91: Engineering Management
- 911: Cost and Value Engineering; Industrial Economics
  - 911.1: Cost Accounting
  - 911.2: Industrial Economics
  - 911.3: Inventory Control
  - 911.4: Marketing
  - 911.5: Value Engineering
- 912: Industrial Engineering and Management
  - 912.1: Industrial Engineering
  - 912.2: Management
  - 912.3: Operations Research
  - 912.4: Personnel
- 913: Production Planning and Control; Manufacturing
  - 913.1: Production Engineering
  - 913.2: Production Control
  - 913.3: Quality Assurance and Control
    - 913.3.1: Inspection
  - 913.4: Manufacturing
    - 913.4.1: Flexible Manufacturing Systems
    - 913.4.2: Computer Aided Manufacturing
    - 913.4.3: Cellular Manufacturing
  - 913.5: Maintenance
  - 913.6: Concurrent Engineering
- ~~914: Safety Engineering~~
  - ~~914.1: Accidents and Accident Prevention~~

~~914.2: Fires and Fire Protection~~
~~914.3: Industrial Hygiene~~
~~914.3.1: Occupational Diseases~~

92: Engineering Mathematics
- 921: Applied Mathematics
  - 921.1: Algebra
  - 921.2: Calculus
  - 921.3: Mathematical Transformations
  - 921.4: Combinatorial Mathematics (Includes Graph Theory, Set Theory)
  - 921.5: Optimization Techniques
  - 921.6: Numerical Methods
- 922: Statistical Methods
  - 922.1: Probability Theory
  - 922.2: Mathematical Statistics

93: Engineering Physics
- 931: Applied Physics Generally
  - 931.1: Mechanics
  - 931.2: Physical Properties of Gases, Liquids and Solids
  - 931.3: Atomic and Molecular Physics
  - 931.4: Quantum Theory
  - 931.5: Gravitation and Relativity
- 932: High Energy Physics; Nuclear Physics; Plasma Physics
  - 932.1: High Energy Physics
    - 932.1.1: Particle Accelerators
  - 932.2: Nuclear Physics
    - 932.2.1: Fission and Fusion Reactions
  - 932.3: Plasma Physics
- 933: Solid State Physics
  - 933.1: Crystalline Solids
    - 933.1.1: Crystal Lattice
    - 933.1.2: Crystal Growth
  - 933.2: Amorphous Solids
  - 933.3: Electronic Structure of Solids

94: Instruments and Measurement
- 941: Acoustical and Optical Measuring Instruments
  - 941.1: Acoustical Instruments
  - 941.2: Acoustic Variables Measurements
  - 941.3: Optical Instruments
  - 941.4: Optical Variables Measurements

942: Electric and Electronic Measuring Instruments
- 942.1: Electric and Electronic Instruments
- 942.2: Electric Variables Measurements
- 942.3: Magnetic Instruments
- 942.4: Magnetic Variables Measurements

943: Mechanical and Miscellaneous Measuring Instruments
- 943.1: Mechanical Instruments
- 943.2: Mechanical Variables Measurements
- 943.3: Special Purpose Instruments

944: Moisture, Pressure and Temperature, and Radiation Measuring Instruments
- 944.1: Moisture Measuring Instruments
- 944.2: Moisture Measurements
- 944.3: Pressure Measuring Instruments
- 944.4: Pressure Measurements
- 944.5: Temperature Measuring Instruments
- 944.6: Temperature Measurements
- 944.7: Radiation Measuring Instruments
- 944.8: Radiation Measurements

Appendix 2. Pre-experiment questionnaire.

This appendix contains the questionnaire used to collect demographic information about the test persons, their educational background and information-seeking experience.

Before-experiment questionnaire

1. Your age:
○ 20 or younger ○ 21 - 25 ○ 26 - 30 ○ 31 - 35 ○ 35 or older

2. Your gender:
○ Male ○ Female

3. How would you rate your knowledge of English? Please indicate on a scale from 1 to 5, where 1 means very bad and 5 very good.
Poor Medium Excellent
○ 1 ○ 2 ○ 3 ○ 4 ○ 5

4. What is the major field of your education for Master's degree?
Choose

If you chose "Other", please state your major field in the box below:

5. How many years have you approximately been doing searching for information on the World Wide Web?
○ less than 1 ○ 1 - 3 ○ 4 - 6 ○ 7 or more

6. When looking for information, how often do you search through Web search engines?
○ Never ○ Once or twice a year ○ Once or twice a month ○ Once or twice a week ○ One or more times a day

7. When looking for information, how often do you search through ELIN?
○ Never ○ Once or twice a year ○ Once or twice a month ○ Once or twice a week ○ One or more times a day

8. When looking for information, how often do you search through library catalogs (LIBRIS, Lovisa)?
○ Never ○ Once or twice a year ○ Once or twice a month ○ Once or twice a week ○ One or more times a day

9. When looking for information, how often do you search through Engineering Village databases (Compendex, Inspec)?
○ Never ○ Once or twice a year ○ Once or twice a month ○ Once or twice a week ○ One or more times a day

10. When looking for information, how often do you use hierarchical directory-style browsing of e.g. search engines or other information databases (see examples below)?
○ Never ○ Once or twice a year ○ Once or twice a month ○ Once or twice a week ○ One or more times a day

11. When looking for information, I generally find what I am looking for on the World Wide Web. Please choose the statement closest to your experience.
○ I strongly disagree ○ I disagree ○ Not sure ○ I agree ○ I strongly agree

Submit your answers

Appendix 3. Post-experiment questionnaire.

This appendix contains the questionnaire used to collect information about the test persons' thoughts on how they performed in the task.

Post-task questionnaire

1. How easy was it to find the right category in the hierarchical structure for your task?
○ Not at all ○ Somewhat ○ Very

2. How certain are you that you found the right category?
○ Not at all ○ Somewhat ○ Very

3. Before looking at the Web pages, how familiar were you with the topic?
○ Not at all ○ Somewhat ○ Very

4. How easy was it to decide whether the Web pages in the category were on the topic of the task?
○ Not at all ○ Somewhat ○ Very

5. How certain are you of your evaluation indicating whether the Web pages in the category were on the topic of the task?
○ Not at all ○ Somewhat ○ Very

6. How would you rate this experience?
○ Frustrating ○ Neutral ○ Pleasing

7. Is there anything else that you think would be important for people analyzing your answers to know? If so, please write them in the box below.

Submit your answers

Appendix 4. Hierarchical view of an ideal sequence of browsing steps for each task.

Task 1:

9: Engineering, General
93: Engineering Physics
932: High Energy Physics; Nuclear Physics; Plasma Physics
932.1: High Energy Physics
932.1.1: Particle Accelerators

Task 2:

9: Engineering, General
94: Instruments and Measurement
942: Electric and Electronic Measuring Instruments
942.3: Magnetic Instruments

Task 3:

9: Engineering, General
92: Engineering Mathematics
921: Applied Mathematics
921.2: Calculus

Task 4:

9: Engineering, General
90: Engineering, General
901: Engineering Profession
901.1: Engineering Professional Aspects
901.1.1: Societies and Institutions

Appendix 5. Tasks rotation scheme.

Each participant was given a different identification number (id). Based on the id, the tasks were automatically rotated, as given in this appendix.

	tasks			
id	first	second	third	fourth
1	1	2	3	4
2	1	2	4	3
3	1	3	2	4
4	1	3	4	2
5	1	4	2	3
6	1	4	3	2
7	2	1	3	4
8	2	1	4	3
9	2	3	1	4
10	2	3	4	1
11	2	4	1	3
12	2	4	3	1
13	3	1	2	4
14	3	1	4	2
15	3	2	1	4
16	3	2	4	1
17	3	4	1	2
18	3	4	2	1
19	4	1	2	3
20	4	1	3	2
21	4	2	1	3
22	4	2	3	1
23	4	3	1	2
24	4	3	2	1
25	1	2	3	4
26	1	2	4	3
27	1	3	2	4
28	1	3	4	2
29	1	4	2	3
30	1	4	3	2
31	2	1	3	4
32	2	1	4	3
33	2	3	1	4
34	2	3	4	1
35	2	4	1	3
36	2	4	3	1
37	3	1	2	4
38	3	1	4	2
39	3	2	1	4
40	3	2	4	1

Appendix 6. Instructions sheet.

DETAILED INSTRUCTIONS

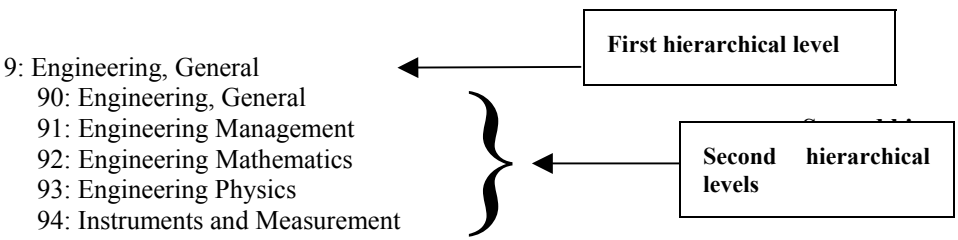
Computer login: **userstudy**
Password: **********

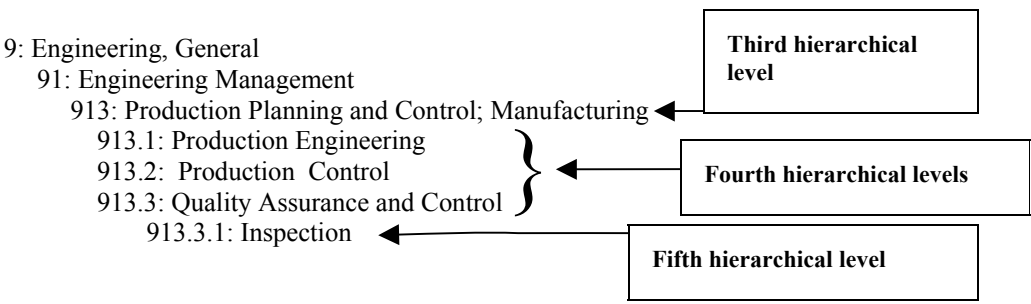
Open the Internet browser and type the following URL: http://juvel.it.lth.se/login.php.

On the initial screen, once you have tested the system and are ready to start with the experiment, you will need to enter your **name and surname**, as well as an **ID number** which we gave you. Then click on the **"Start the study"** button.

After the initial questionnaire ("**Before-experiment questionnaire"**), you will get to the description of your first task. When you have understood your task, click on the button "**Start experiment based on the first task**."

You will come to the searching system. The searching system does not offer a search box where you could enter your search term. Instead, it provides you with a hierarchical structure of topics. Through this structure you can browse by clicking on the different categories to arrive at their more specific sub-categories. Below is an example of different branches in this hierarchical structure:





This hierarchical structure is one branch of a bigger structure called "Engineering Index" classification scheme which has been used for classifying documents in the field of engineering. The numbers to the left of the names of categories represent the hierarchical level, e.g., four numbers mean that the category is at the fourth hierarchical level.

This hierarchical structure is shown in the top half of the screen, whereas in the lower part of the screen we have listed Web pages that belong into the last class you clicked on. At first and second levels, as well as at some third-level categories, there will be no Web pages, because they were classified into more specific levels.

Once you find the right category for your task, you will be asked to evaluate whether the Web pages listed as belonging to that category have been put in the right place or not. You will use the form that is given next to the description (Title, Summary, Link) of each Web page. In order to evaluate whether the Web page is about the topic given in the task, please open and look at the page ('**View page**') because it would be incorrect to judge only based on the Title and Summary.

When you open the page by clicking on '**View page**', what you will get is an online version. The fact that the page maybe does not look nice etc., should not influence your decision whether it is in the right category. It is only the text of the Web page based on which you should make the decision. Also, it is only one page you will view, not the whole Web site, and your judgement should be thus based only the content of that one Web page. If it happens that there is an error and you can't see the Web page ("Page not found" or so), please mark "Impossible for me to say" and move on to the next Web page.

In the form choose one of the three options:

1) **Correct** – if the Web page is about the topic

2) **Partly correct** – if the Web page **can be** considered to be on the topic, but is mixed with other topics
3) **Incorrect** – if the Web page has **absolutely no relation** to the topic.

Please avoid choosing the option 'Impossible for me to say.' Instead, try to figure out the page's content and choose one of the top three options.

Please evaluate **at least top 10** Web pages in each task. Please evaluate them in the order they are presented, **do not jump over** any of them. Please do not count cases where you selected "It is impossible for me to say" among the 10 evaluated Web pages.

After you submit the evaluation, you will get to the screen where you can choose whether you still want to evaluate more Web pages in the current task, or if you want to go on to another task. The screen will look like this:

You have now submitted evaluation for the last XY records you have evaluated. To continue evaluating other Web pages in this task, please click on the button "Go back to evaluation."

Go back to evaluation

If you have evaluated the minimum of 10 Web pages in this task, you may proceed to the post-task questionnaire for this task, and then to the second task, by clicking on the button below.

Finish the task and go to the post-task questionnaire

After clicking "Finish the task and go to the post-task questionnaire", you will come to the screen with several questions on the task you've just finished. After filling in that questionnaire and submitting your answers, you will get to a new screen describing the following task.

OBS!

During the study you should not talk to other people participating in the study or look at each other's screens.
We also need to ask you to turn your cell phone off.

Appendix 7. Main user study screen.

Upper part of the screen contains task description, hierarchical tree and most important instructions. Lower part of screen comprises web pages and evaluative forms with each of them. Up to 40 web pages with forms were listed on the same page.

Your task is to find all the Web pages in the system dealing with the topic of **particle accelerators**.

Start page

- 9: Engineering, General
  - 93: Engineering Physics
    - 932: High Energy Physics; Nuclear Physics; Plasma Physics
      - 932.1: High Energy Physics
        - **932.1.1: Particle Accelerators**

Please read the "Detailed instructions" document we gave you.

Basic steps include the following:
1) Browse the hierarchical tree of categories in order to come to the right place for your topic.
2) If you believe you have come to the right place for your topic, please evaluate whether the pages in that category (if any found in the category, they are listed below these instructions) are on that topic.

Important:

- In order to evaluate whether the Web page is about the topic given in the task, please open the Web page (**'View page'**) as it would be obvioulsy wrong to judge only based on the Title and Summary.
- Button "Send evaluation" needs to be clicked for every Web page.
- Please evaluate at least top 10 Web pages before you move on to the following task.
- Please do not skip any Web pages.
- Please avoid choosing the option 'Impossible for me to say.' Instead, try to figure out the page's content and choose one of the top three options.

1

Title: Accelerators and Nobel Laureates
Summary: Accelerators and Nobel Laureates Nobel Foundation Nobel Media Nobel Museum Nobel Peace Center Nobel Web SEARCH CONTACT US HOME [nobelprize.org Logo] NOBEL PRIZES ALFRED NOBEL PRIZE AWARDERS NOMINATION PRIZE ANNOUNCEMENTS AWARD CEREMONIES EDUCATIONAL GAMES By Year Nobel Prize in Physics Nobel Prize in Chemistry Nobel Prize in Medicine Nobel Prize in Literature Nobel Peace Prize Prize in Economics Accelerators and Nobel Laureates by Sven Kullander 28 August 2001 [intro] Why Accelerators Particle
View page

Please judge whether this Web page is on the topic you were looking for in this task:
- Correct
- Partly correct
- Incorrect
- It is impossible for me to say

2

Title: Faraday Cups for PelletronÂ® Electrostatic Accelerators
Summary: National Electrostatics Corporation is a manufacturer of potential drop accelerators and related beam line components. Faraday Cups for PelletronÂ® Electrostatic Accelerators Faraday Cups Faraday Cup, FC46 [Picture of the FC-46 Faraday Cup] NEC manufactures more than seven types of Faraday Cups for a variety of applications involving the accurate monitoring of ion beam currents. They are all metal and ceramic with MHV or BNC feedthroughs and electrostatic or magnetic suppression. All beam inter
View page

Please judge whether this Web page is on the topic you were looking for in this task:
- Correct
- Partly correct
- Incorrect
- It is impossible for me to say

3

Title: Charged Particle Beams
Summary: Charged Particle Beams is a comprehensive text of collective beam physics, available for download on the Internet Charged Particle Beams Downloads You can view the Table of Contents before downloading files. CPB.PDF Entire book in one file 34.41 MB CONTENTS.PDF Preface and table of contents 0.16 MB CHAP01.PDF Introduction 0.50 MB CHAP02.PDF Phase-space description of charged particle beams 2.20 MB CHAP03.PDF Introduction to beam emittance 2.41 MB CHAP04.PDF Beam emittance - advanced topics 1.84
View page

Please judge whether this Web page is on the topic you were looking for in this task:
- Correct
- Partly correct
- Incorrect
- It is impossible for me to say

4

Title: Partner Institutions
Summary: Partner Institutions Partner*> [EUROFEL] European FEL Design Study Partner Task Management Calendar Documents Messages Recruitment Links Contacts [] Home > Partner [] Partner Website Hosted by DESY Partner Institutions [] Berliner Elektronenspeicherring-Gesellschaft fÃ¼r Synchrotronstrahlung mbH, Berlin, Germany Designed and operates German's only 3rd generation synchrotron light source. Currently BESSY is developing the plans for a 2nd generation CW FEL based on a 2.3 GeV superconducting driv
View page

Please judge whether this Web page is on the topic you were looking for in this task:
- Correct
- Partly correct
- Incorrect
- It is impossible for me to say