EXAMENSARBETE Entity-based Search
A take on the intelligent book
STUDENT Erik Zander
HANDLEDARE Pierre Nugues (LTH), Markus Klang (LTH)
EXAMINATOR Jacek Malec (LTH)

Konceptbaserad sök

POPULÄRVETENSKAPLIG SAMMANFATTNING **Erik Zander**

Tänk dig att vi ska söka upp *Paris*, svaret kommer direkt - stad i Frankrike. Dock är det personen *Paris* vi söker. Vi prövar igen, *Paris personen*, vi får nu träff på den mytologiska personen *Paris* men vi sökte den fiktiva personen *Tom Paris* från Star Trek. Det är uppenbart att det är svårt att veta vad vi söker efter, för datorer har det varit omöjligt. Omöjligt är dock ett relativt begrepp när det gäller teknik. Detta projekt har jobbat med att ta aktuella forskningsresultat om språkteknik och skapa en sökmotorprototyp med hjälp av dessa.

Vår prototyp är en sökmotor där en användare kan söka bland 3500 böcker. Det vanliga fallet är att en sökning startar med en textsträng, sökningen letar bara efter att bokstäverna skall vara rätt. När sen resultatet presenteras, visas de koncept som sökningen har hittat i anslutning till sökordet. Ett koncept kan vara en person, ett land, en stad, osv. Om vi återgår till vår sökning om *Paris* kan till exempel sökmotorn presentera *Paris (staden)*, *Paris (grekisk mytologi)*, *Paris (Sci-fi karaktär)*. Vi ser att om vi får dessa alternativ presenterade kan vi som användare snabbare komma in på det vi letade efter. Det vi inte är intresserade av filtrerar vi bort.

Hur går då detta till? För att svara på det ger vi först en överblick vad som krävs för att det skall vara möjligt. Vi behöver veta vilka koncept en bok innehåller; vi behöver spara den informationen någonstans; vi vill kunna länka ut till online resurser som Wikipedia, alternativt till grafvarianter av Wikipedia så som Wikidata eller DBpedia.

För att veta vad en bok innehåller tar vi hjälp av ett API (Langforia) som är skrivet av Markus Klang på LTH. Till API:et skickar vi varje paragraf i boken och får tillbaka vilka koncept den inne-håller. Vi sparar denna information i samma databas som vi har böckerna i. Mer precist sparar vi informationen i varje digital bok genom att vi berikar paragrafen.

När vi får informationen från Langforia får vi även ett så kallat Qid. Ett Qid är en identifierare på Wikidata.org och med den kan vi hitta mer information om ett specifikt koncept. Den komplementerande informationen kan till exempel vara när en person dog eller hur stor befolkning ett land har. I vår applikation visar vi denna information i en informationsruta till höger.

Vad skall det vara bra för? Genom att veta vilka koncept som texterna innehåller kan vi både ge en precisare och mer berikad upplevelse. Det krävs dock att applikationen arbetar med texterna innan och begränsningar i hur det fungerar med maskininlärning gör att vi inte kan jobba med alla typer av texter samtidigt. Bäst blir det därför om vi har ett material som vi vet ungefär vad det är men som vi behöver kunna söka bättre i. Har vi ett sådant material kan vi erbjuda en mycket bra upplevelse till användaren.