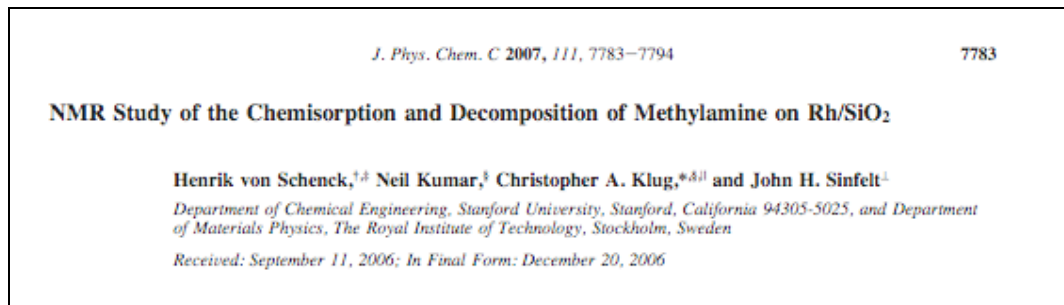


Förslag till examensarbete från Enheten för vetenskaplig information och lärande

Ulf Kronman, 2008-02-06

Vem har skrivit vad – och var?



Allteftersom intresset för utvärdering av forskning ökar hamnar forskarnas artiklar och citeringarna till dessa allt mer i fokus. Studier av vem som publicerat vad och hur mycket det citerats används allt oftare för medelstilldelning, tjänstetillsättningar och bedömning av olika forskningssatsningar, både lokalt på universiteten och centralt från statsmakternas sida.

Hur får man då reda på vem som publicerat vad? Man kan förstås fråga forskarna efter listor på sina artiklar, men de listorna kräver en hel del administrativt arbete att ta fram och ger ingen information om hur mycket en artikel är citerad. Vill man inte störa forskarna och dessutom få citeringsinformation kan man använda de kommersiella systemen Web of Science och Scopus. I dessa system finns en stor del av världens vetenskapliga publicering förtecknad, dessutom med information om hur mycket artiklarna är citerade.

Ett grundläggande problem med informationen om artiklarna i de kommersiella systemen är att den registreras i den textform databasleverantörerna får den från förlagen. Identifieringen av författarna består oftast av en lista med "accentstādade" efternamn följt av en eller ett par initialer. Adressinformationen består av en lista med adresser till olika universitet och institutioner, så som forskarna har skrivit sina adresser i artikeln. Ingen rättning eller normalisering av författarnamn och adresser görs. Ofta finns heller ingen koppling mellan författaruppgifterna och adressuppgifterna.

På Enheten för vetenskaplig information och lärande (KTH-biblioteket) pågår just nu ett projekt för att ladda ned artikelinformation om KTHs vetenskapliga publicering från systemet Scopus och med hjälp av automatisk matchning försöka identifiera vilka forskare på KTH som är författare till artiklarna och vid vilken eller vilka skolor forskarna har varit aktiva när de skrev artikeln. Matchningen bygger på en kombination av tolkning av adresser och tolkning av författarnamn i kombination med uppslag mot ett organisationsregister och ett personalregister.

Matchningen fungerar för närvarande i runt 80 % av fallen, men skulle sannolikt kunna förbättras avsevärt med textanalytiska och statistiska angreppssätt. Vi söker därför nu efter en KTH-student med inriktning mot datorbehandling av talat och skrivet språk, som är intresserad av att ägna sitt examensarbete åt denna problemställning.

Den tekniska miljön för systemet är: Ubuntu Linux server, PostgreSQL databas, programspråk Python och ramutvecklingsverktyg Django.

För mer information; kontakta Ulf Kronman, tfn: 790 7447, e-post: ukronman@kth.se