Slutrapport för projektet

VISAT

av

Johan Hagman
Leif Grönqvist
Sven David Bergström
Jens Allwood

Projektledare: Jens Allwood
Projektfinansiär: FRN

ImmigrantInstitutet, Borås

Inst f Lingvistik, SSKKII, Göteborgs Universitet

1 Introduktion

Hösten ’95 skissade en liten forskargrupp inom SSKKII på ett projekt där tanken var att kombinera några intressanta *information retrieval* -tekniker i en och samma sökmotor som man sedan skulle erbjuda allmänheten att använda och utvärdera. Projektet fick namnet ‘VISAT’, inspirerat av frasen •Visualiserad InformationsSökning med Automatiska Tesauri•, Teknikerna som avsågs är utvecklade inom lingvistik och till viss del datalogi.

En av de lingvistiska tekniker man tänkte på var *automatisk expansion av sökfrågan*, vilket kan göras på åtminstone två sätt: *morfologiskt*, där man genererar flera möjliga böjningsvarianter till de(t) ord man uppgett; och *samförekomstrelaterat*, där ord som statistiskt sett (baserat på stora textmaterial som tidigare studerats) ofta brukar uppträda tillsammans med de(t) ord man skrivit in i sin sökfråga också läggs till för att utöka chanserna att hitta dokument som kan vara av intresse — fastän de inte innehåller just det ord man först tänkte på.

En annan möjlighet som projektets initiativtagare själva hade saknat vid websökningar var att kunna göra sökningen samtidigt till flera sökmotorer, där de som är mest uppdaterade och omfattande förstås är mest intressanta att parallellkoppla. Exakt vilka detta är förändras från år till år, ja t o m från månad till månad, efterhand som konkurrerande •stormotorer• byggs ut och moderniseras. Till denna finess tänkte man sig att lägga möjligheten att samtidigt kunna skicka denna sökfråga till en databas över bibliografiskt material. Forskargruppen bestämde att testa denna idé på Immigrant-Institutets material och associerade därmed denna institution till projektet. Immigrant-Institutet ligger i Borås.

2 Första fasen: montering av web-servrar och web-sajtar

Våren ’96 fick vi besked om finansiellt stöd från FRN att förverkliga vår projektidé, med start hösten samma år. Det bestämdes att man skulle börja med en inventering av Immigrant-Institutets bibliografiska material och att anpassa detta för att göra det tillgängligt på webben. I samband med detta, hösten ’96, installerades servrar för den kommande sökmotorn och för databashanteringen. Detta arbete pågick i perioder under ett års tid.

Cirklarna med siffror i vänsterkanten av nedanstående text hänför till skissen i Bilaga 1 och de olika modulerna där. Presentationen i stycke 2 följer i stort den ordning i vilken momenten har verkställts. Ordningen av delarna i stycke 3 däremot återspeglar själva sökmotorns algoritm, eller, om man vill, informationsflödet vid användningen av prototypen.

2.1 Web-servrar

Vi har installerat och konfigurerat en dator som server och en som arbetsstation för projektet. Båda är lokaliserade till SSKKII, Göteborgs Universitet. Förutom standardprogramvara har även en webserver installerats på serverdatorn.

Vi har även installerat erforderligt operativsystem och server-mjukvara på Immigrant-Institutet och förutom installation av standardprogramvara samt konfiguration av en webserver där, har dess personal fått undervisning i NETSCAPE COMMUNICATOR och HTML för bl a löpande websidekonstruktion.

2.2 Behovsanalys

① En behovs- och verksamhetsanalys av Immigrant-Institutet har gjorts för att se på vilket sätt dess bibliografiska material och övrig information bäst kan göras tillgänglig på webben. Resultatet av denna analys föreslog bl a inscanning av text, strukturering och design av websidor, samt nya uppdateringsrutiner av det webbaserade materialet.

2.3 Www-material på Immigrant-Institutets hemsidor

④ En första version av orienterande hemsidor konstruerades åt Immigrant-Institutet (se bilagan 2 som delvis återger innehållet på institutets hemsida på http://www.immi.se/) med följande huvudsakliga punkter:

- Immigration, Emigration
- Flyktingar
- Rasism och mänskliga rättigheter
- Flerspråkig nyhetslänk
- Länder i världen
- Institutets elektroniska bibliotek
- Frågor och svar om migration

Dessutom finns nu web-sidor med följande innehåll:

- Public Holidays Worldwide
- Art Gallery
- Virtual Tour (en rundvandring i ImmigrantInstitutets lokaler)
- Books to Sale: (Internet-annonsering)
- Meeting, presentation of the Director
- Clickable Map of Sweden
- Virtual Library
- Welcome Greetings (prototyp med ljud)

Ett antal Internet-baserade diskussionsgrupper har etablerats på nätet om immigrant-relaterade ämnen och därtill har även ett litet system för •inter–cultural communication in virtual worlds on the www• implementerats, på ett experimentellt stadium.

2.4 Inscanning av bibliografiska data

→ En scanner har införskaffats och erforderlig mjukvara installerats. Scannern har använts främst till att göra tryckt material tillgängligt genom Immigrant-Institutets webbsidor men även en del bildmaterial har gjorts •maskinläsbart•.

Det textdatamaterial vi för närvarande har består av:

- Lista på böcker, c:a 2.000 stycken, där de i nuläget sökbara databas-fälten är c:a 30 till antalet och uppger det som biblioteks-databaser brukar uppge, dvs författare, titel, år, antal sidor, typ/ämne, språk samt var boken finns, mm.
- Ett 500-tal websidor med information, omfattande c:a 250.000 ord.
- Lista på c:a 700 tidskrifter samt information om dessas placering.

3 Andra fasen: tillverkning av gränssnitt och sökmotor

Under andra fasen av projektet har vi implementerat själva sökmotorn med dess faciliteter. Resultatet är en prototyp som vi lagt upp på adressen **http://www.ling.gu.se/visat/** . I detta stycke beskrivs dess egenskaper och funktion.

3.1 Web-gränssnittet

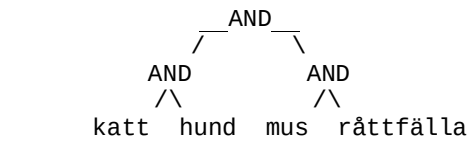
{ Vi har designat ett webgränssnitt, eller •interface• som möjliggör sökning både av standardtyp och enligt mer elaborerade metoder. Användaren har full frihet att koppla in och ur de finesser som erbjuds utöver de faciliteter som redan finns vid sökmotorer som t ex AltaVista och HotBot.

De extra faciliteter vi erbjuder är:

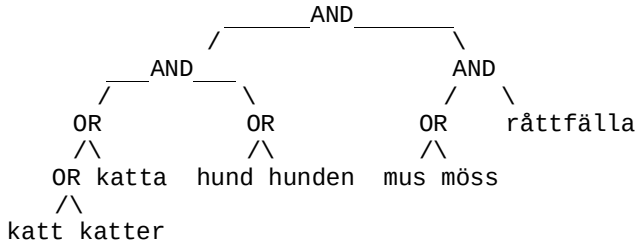
- Möjlighet för användaren att uppge huruvida speciella språk önskas. Användaren ska m a o kunna ange vilka språk som han/hon kan tillgodogöra sig info på.
- Domänfiltrering, dvs att man kan (de-)aktivera webdomäner beroende på intresset vid en viss sökning. T ex kanske man bara vill ha information från japanska webbservrar och aktiverar därmed endast källor med adressuffixet •.jp•.

3.2 Automatiska expansioner av sökfrågan

★ Användarens fråga kan vara av tre typer: •alla ord• (implicit konjunktiv), •åtminstone något ord• (implicit disjunktiv) eller •boolsk fråga• (explicit). Innan en fråga kan behandlas måste den normaliseras till boolsk form. Implicita frågor omvandlas genom att alla ordmellanrum ersätts med 'AND' resp. 'OR' Internt används en trädform för frågor så att frågan `(alla orden) katt hund mus råttfälla` ser ut så här:



Genom att slå upp orden _katt, hund, mus_ och _råttfälla_ i automatgenererade lexikon kan vi expandera trädet till nedanstående vilket motsvarar `(katt OR katter OR katta) AND (hund OR hunden) AND (mus OR möss) AND råttfälla`:



När man letar efter något i en databas eller på internet är det ibland svårt att komma på vilka sökord som kommer att ge bra resultat. Om orden man skriver in är för vanliga så får man ofta tusentals träffar och om orden är ovanliga så blir det kanske inga träffar alls. Eftersom vi i det här projektet har koncentrerat oss på att förbättra sökmöjligheterna i Immigrant-Institutets databaser, som består av relativt lite information jämfört med hela Internet, så är det störst risk att användaren får för få träffar. Därför erbjuder VISAT's sökmotor användaren att låta göra en automatisk expansion av sökfrågan. Med detta menar vi att till de sökord som användaren har uppgett kommer synonymer och andra på något sätt associerade ord att läggas till för att på så sätt bredda sökfrågan och öka chanserna att hitta relevanta websidor och andra data på www.

Ett slags förteckning av ord och deras synonymer och andra på ett eller annat sätt (ofta men inte nödvändigtvis hierarkiskt) •semantiskt associerade• ord och uttryck kallas •tesaurus•. Vi har gjort en inventering av maskinläsbara tesauri för eventuell integrering i sökverktyget. Här avser vi i första hand svenska tesauri eftersom huvuddelen av texterna är på svenska. Den enda svenska vi har fullständig tillgång till är Brings tesaurus från 1930. Den är ganska stor och innehåller totalt 76034 ordtyper. Trots detta finns bara 3,2% av orden i Immigrant-Institutets texter med i denna tesaurus, vilket gör Brings tesaurus är nästan oanvändbar i det här fallet. Anledningen till detta är att en del av texterna inte är på svenska men även att denna tesaurus är en allmän (och bitvis något gammaldags) sådan medan språket i Immigrant-Institutets databas är mycket domänspecifikt, vilket för övrigt gäller de flesta textsamlingar.

En annan lovande idé som vi tittat på är att koppla in två- eller flerspråkliga lexika för att göra det möjligt för icke-svenskspråkiga användare att specificera sina sökvillkor och sökord (som därefter översätts till andra språk före sökningen på www och i Immigrant-Institutets egna databaser). Även här får man problem med domänspecifika ord. Även mycket stora lexikon fångar endast några få procent av orden i Immigrant-Institutets texter. Som illustation kan nämnas att en stor samling (12 miljoner ord varav 450.823 unika) av tidningstexter, romaner, mm täcker upp endast 34% av orden i Immigrant-Institutets texter och då är detta material mycket större än de flesta existerande lexika.

Ett ytterligare problem är att bra flerspråkiga lexika i datorläsbar form är dyra, speciellt om man vill använda dem så fritt som i ett sökverktyg. De experiment vi gjort med flerspråkiga lexika visar att det är värt att satsa på men med de små gratislexikon vi har tillgång till nu får man ingen nämnvärd förbättring vid sökning i en flerspråkig databas.

Vi har använt två huvudmetoder för att expandera sökfrågan:

- *Stemming* handlar om att utgående från en grammatik över ett språks vanliga ändelser, böjnings- och stavningsregler göra en automatisk, pseudo-morfologisk parsning.
- *Samförekomstanalys* innebär att man tar fram ord som oftare än förväntat förekommer tillsammans med förhoppningen att de skall vara relaterade.

Stemming

Olika ordformer, ändelser och felstavningar ställer till en hel del problem vid informationssökning. Därför har vi utvecklat program som går igenom alla ord och försöker identifiera par av ord som troligen är olika varianter eller former av samma ord. Vi har använt regler av typen:

- Ordslut som `-a`, `-an`, `-ans`, `-as`, `-ornas`, `-orna`, `-ors` och `-or` betraktas som lika, för att fånga böjningar.
- Ekvivalensregler som ‘`oe`’ = ‘`ö`’; ‘`ph`’ = ‘`f`’; ‘`ae`’ = ‘æ’ = ‘`ä`’ fångar stavningsvarianter.

Ett mer generellt sätt är att leta efter ordpar med en stor gemensam del och på så sätt hitta att exempelvis `avdragsgilla` och `avdragsreglerna` är relaterade.

Ännu mer generellt är att beräkna något som kallas *editeringsavstånd*. Detta är ett avståndsmått mellan två ord som beräknas genom att titta på hur många operationer (insättning, borttagning, ersättning) som behövs för att gå från det ena ordet till det andra där enheterna för operationerna kan vara bokstäver eller vokal/konsonant-kluster. Vi gör båda varianterna och relaterar editeringsavståndet till ordens längd för att kunna ta fram de 15.000 bästa för varje

uppdelning. `Axelsson` och `Axelson` har exempelvis editeringsavståndet 1 med avseende på bokstäver.

Fördelen med dessa två mer generella metoder är att de fångar en hel del böjningar och felstavningar för alla språk, faktiskt även en del stavningsvarianter mellan språk. Därför fungerar detta ganska bra trots att vi bara skrivit grammatikregler för svenska.

De olika metoderna för stemming resulterar i en slags ekvivalenslista där ord som ser •tillräckligt• lika ut klumpas ihop. Genom att variera de olika tröskelvärdena ovan så kan man välja hur många ord det skall vara i listan, men ju fler ordpar som accepteras desto fler kommer att hamna där felaktigt. Här följer några exempel från listan:

```
emigrasjon, emigratio, emigration, emigrationen, emigra-
tionens, emigrationer, emigrations, migracion, migraciones,
migrasjon, migration, migrationen, migrations, migratoria,
migratorio, migratorios migratory

reading, heading, headings, leading ,readings

ungdom, ungdomar, ungdomarna, ungdomarnas, ungdomars,
ungdomen, ungdomens, ungdommens, ungdoms
```

Som synes fungerar det ganska bra för det mesta och i vår tillämpning kan en viss felprocent accepteras. Det första exemplet visar att övergång mellan språk kan fås, det andra att det kan bli fel ibland, och det tredje en ren felstavning som fångas. Med de värden vi valt så får vi 12.029 ekvivalensklasser med totalt 38.088 unika ord utifrån grundmaterialet som innehåller 86.234 unika ord.

Samförekomstanalys

Stemming-processen bygger på att ord som ser ungefär likadana ut har med varandra att göra. Vid samförekomstanalys utgår man istället från att de ord som förekommer tillsammans oftare än förväntat, rent statistiskt, är relaterade. Den här sortens analys blir bättre ju större textmaterial man har att analysera. Vi har använt sammanlagt en dryg miljon ord och eftersom dessa texter ligger inom samma domän så blir resultatet ganska bra. Även här måste man sätta en tröskel någonstans för att få en förhållandevis stor andel korrekta ordpar. Vi har valt att räkna antalet samförekomster med högst 25 ords avstånd för varje ordpar, och sedan relatera det till ordens frekvens enligt:

$$s^2 * (1-1/f_1) * (1-1/f_2) \ / \ (f1*f_2)$$

... där s är antalet samförekomster och f_1, f_2 är ordens frekvens.

Sedan har vi klippt vid ett empiriskt fastställt värde. Syftet med att använda den här formeln är att få fram par av ord som förekommer tillsammans mycket oftare än förväntat, samtidigt som man vill sortera bort ord med allt för låg frekvens.

Den resulterade listan innehåller 16.668 huvudord med i genomsnitt 5,8 ord som tycks vara relaterade till respektive huvudord. Sammanlagt 34.598 unika ord finns med i den resulterade lexikonfilen. Några exempel:

arbetssökande	annonsering, arbetserbjudanden, förbigå, otillbörlig, otillbörligt, särbehandla
enkelrum	anmäler, dubbelrum, hotell, kr, natt, per, plaza, pris, priset
immigration	emigration, ohdic, imm
lindgren	arnia, astrid, ilon, ksi, nasza, rabén, sjögren, wikland

Sammanfattningsvis kan sägas att det totalt i lexikonfilerna (stemming + samförekomstanalys) finns 58.237 olika ord representerade, vilket motsvarar 67% av det totala antalet unika ord i Immigrant-Institutets texter.

3.3 Parallellt utskick av sökfrågan till flera sökmotorer

3.3.1 Val av sökmotorer att skicka frågan till

När vi skulle välja vilka av Internets sökmotorer som vi skulle använda i projektet beaktade vi följande kriterier:

- De måste klara boolska frågor. Detta är en direkt följd av att vi vill kunna utnyttja automatisk språkbehandling. I en sökmotor som inte kan hantera boolska frågor kan vi bara ställa frågor som `(något av orden) hund katt` och `(alla orden) hund katt`. Om vi vill söka efter dokument som innehåller orden _hund_ och _katt_ eller andra former av orden så blir frågan i traditionell notation: `(hund OR hunden OR hundar) AND (katt OR katta OR katter OR kattor)`.
- De måste täcka i princip hela nätet. En hel del sökmotorer söker i mindre än hälften av de dokument som finns på Internet, t.ex. för att de bara avser att täcka vissa länder eller ämnesområden, eller för att de är dåligt skötta. Ett av målen med projektet var att användaren skulle få välja begränsningar själv snarare än vara begränsad av sökmotorn.
- De måste vara aktuella. Alltför många sökmotorer uppdaterar sina index så sällan att de i sina söksvar hänvisar till dokument som inte längre finns eller missar nya dokument som ännu inte registrerats.

Vi valde AltaVista och HotBot för att dessa uppfyller kraven ovan och dessutom är allmänt väl ansedda. AltaVista var den första sökmotorn som hade ambitionen att indexera hela webben, och har utvecklats stadigt sedan dess. HotBot har vunnit ett flertal utmärkelser för sin förmåga att lämna relevanta svar på frågor. På senare tid har tyvärr både och tvingats ge efter på de två sista punkterna ovan p g a webens oerhört snabba expansion.

Förutom de krav vi ställde har AltaVista och HotBot ytterligare några finesser: AltaVista talar om vilket språk dokumentet är skrivet på, vilket återspeglas i svarssammanställningen. HotBot poängsätter de refererade dokumenten i relation till det första dokumentet. Tyvärr har vi ännu inte lyckats använda den in-

formationen på ett konstruktivt sätt; för närvarande ser vi det som att en sådan (sporadisk) uppgift i sammanställningen snarare skulle leda till förvirring.

3.3.2 Själva utskicket av sökfrågan

När användaren har slipat färdigt sin fråga och väljer att söka vidtar ett intensivt arbete på serversidan. Det CGI-program som användaren anropar startar ett program för varje sökmotor som användaren valt (AltaVista, HotBot, och/eller Immigrant-Institutets dito). Underprogrammen skickar frågorna till sina respektive sökmotorer, väntar på svaren och lagrar dem i en databas.

3.4 Presentation av sökresultatet från flera sökmotorer

♠ När alla sökmotorer har lämnar svar går huvudprogrammet igenom de inkomna svaren och sorterar dem enligt följande:

- Dokument som refereras av *alla* befrågade sökmotorer kommer först.
- Dokument som refereras av *alla utom en* av de befrågade sökmotorerna kommer därnäst, osv.
- Övriga dokument kommer växelvis från de olika sökmotorerna och presenteras i den ordning som de hade i sökmotorernas svar. Det finns tyvärr inget standardiserat poängsystem som kan användas.

4 Summering och möjliga vidareutvecklingar

4.1 Resultaten av projektet

En webserver har installerats på immigrantinstitutet och en uppsättning orienterande websidor har ställts samman. Dessutom har personalen instruerats i hur sidorna skall uppdateras och byggas ut. En fungerande prototyp av en sökmotor, specialdesignad för Immigrantinstitutet har utvecklats. Den har en hel del finesser som man inte brukar hitta men även standardkomponenter:

- Sökmotorn har ett flexibelt gränssnitt. Enkel ordsökning kan utföras utan kunskaper om avancerad informationssökning.
- Förutom att söka bland de lokala dokumenten kan sökfrågan skickas vidare till två av de större sökmotorerna i världen: AltaVista och HotBot.
- Sökfrågan kan utformas som ett godtyckligt logiskt uttryck av sökord.
- Sökfrågan kan breddas genom att dess ord expanderas till flera variater m h a:
  - samförekomstanalys - ord som oftare än förväntat förekommer tillsammans i texter antas ha med varandra att göra.
  - ortografisk likhet - om endast ett fåtal bokstäver skiljer mellan två ord eller om skillnaden uppfyller böjningsregler eller stavningsvariantsregler antas de vara relaterade.
  - efter expansionen kan sökfrågan fortfarande editeras.
- Sökningen kan inskränkas till dokument från vissa årtal (endast implementerat för HotBot och AltaVista).
- Vissa internetdomäner kan väljas eller väljas bort (endast implementerat för HotBot och AltaVista eftersom alla ImmigrantInstitutets WWW-sidor ligger inom domänen Sverige).
- Dokument på vissa språk kan väljas eller väljas bort (ej implementerat i nuläget).

- Internt används databasen MYSQL och WWW-indexeringssystemet HT://DIG samt programmeringsspråket PERL.

4.2 Framtida vidareutvecklingar

Det finns mycket intressant arbete kvar att göra inom området då det inom VISAT endast funnits utrymme för prototypisk testning i liten skala. Några av de angrepps-punkter vi ser är:

- Det vore mycket intressant att utvidga sökdomänen från Immigrantinstitutet till exempelvis Sverige. Detta skulle kräva en hel del datorkraft och lagringsutrymme samt en genomtänkt infrastruktur för insamling och bearbetning av data.

- För att få språkfiltrering att fungera tillfredställande krävs utveckling av en egen språkidentifieringsmodul som är specialiserad på de språk som är intressanta i den aktuella sökdomänen.

- Sökfrågeexpansionen kan förbättras ytterligare. En intressant utvidgning vore att använda existerande flerspråkiga lexika för att ytterligare förbättra möjligheten att finna dokument på andra språk än det man använt i sökfrågan.

- En metod som kallas LSI (Latent Semantic Indexing) som bygger på att man räknar fram vektorer i en n-dimensionell rymd som svarar mot ord eller dokument. När samtliga dokument indexerats med LSI kan de dokument som matchar en sökfråga placeras ut på skärmen så att de som - med avseende på de ingående orden - tycks handla om samma sak hamnar nära varandra. Dessutom möjliggör LSIn frågor av typen: •ge mig de fem dokument i sökdomänen som ligger närmast dessa fem• eller •det dokument som ligger längst ifrån dessa•. Man skulle även kunna ange •negat-iva• sökord: •ge mig dokument som handlar om chips men inte om datorer•.

- Rangordningen av sökresultatet måste förbättras. Ett problem i vår prototyp är att Alta Vistas rangordning är ganska dålig och nästan omöjlig att jämföra med HotBots och den lokala sökmotorns resultat. En väl fungerande rangordning eliminerar också behovet av att begränsa antalet träffar - de man får först är ändå de bästa.

De lingvistiska metoderna som använts för att bredda sökfrågan är till stor hjälp vid informationssökning, men till en fullt användbar sökmotor anpassad för en större domän, som t ex hela Sverige, återstår mycket arbete. Med hittills gjorda erfarenheter och de idéer som uppstått och klarnat under detta projekts gång känner vi oss inspirerade att fortsätta detta •bygge•.