Innehållsförteckning

1 Inledning 1

1.1 Bakgrund 1
1.2 Uppdragsgivare 2
1.3 Problem 2
1.4 Uppdrag 2
1.5 Arbetssätt 2
1.5.1 Förarbete 3
1.5.2 Insamling och analys av data 3
1.5.3 Utvärdering 4
1.6 Målgrupp 4

2 Utvidgad bakgrund 5

2.1 Elektroniska postsystem 5
2.1.1 Användarfunktioner i elektroniska postsystem 5
2.1.2 Programvaror 6
2.2 Examensarbetet av Naoufel Ben-Cheikh och Magnus Zakrisson 7
2.3 Statistiska urvalsmetoder 8
2.3.1 Obundet slumpmässigt urval 8
2.3.2 Chi-två 8

3 Artikelinsamling och analys 10

3.1 Val av typer 10
3.2 Insamling och kategorisering av artiklar 11
3.3 Metod för urval av ord 13
3.4 Utveckling av datorprogram 14
3.5 Programkörning 15
3.6 Urval av ord 15

4 Statistikanalys 18

4.1 Chi-två program 18
4.2 Tolkning av chi-två värdet 18

5 Utvärdering av resultat 19

5.1 Toppen på listan 19
5.2 Innehållsord 19
5.3 Innehållsfattiga ord 19
5.4 Formord 20

6 Slutsatser 22

Litteraturförteckning 23

Ordlista 24

1 Inledning

Elektroniska meddelande- och postsystem har blivit en viktig del i datakommunikationsvärlden. Ett sådant system är e-post (alternativ benämning: elektronisk post), där data och meddelanden kan överföras från en dator till en annan via telefonlinjer, mikrovågor, kommunikationssatelliter eller annan telekommunikationsutrustning [Office system, 1993]. Dessa system har tack vare deras effektivitet snabbt fått stor spridning.

För att kunna göra gruppkommunikationssystem ännu mer effektiva pågår forskning inom området med hjälp av bland annat forskningsprojektet IntFilter som drivs av SICS i samarbete med andra parter [jmf kap 1.2].

1.1 Bakgrund

E-posten som blir allt vanligare ute i affärsvärlden har mycket stora fördelar. Speciellt när en enskild användare via höghastighetsmodem kan utnyttja systemet dygnet runt. I professionellt bruk kan en användare kommunicera med hundratals samtalspartner och utväxla en stor mängd brev per dag. De e-postsystem som finns idag är tyvärr asymmetriska, dvs det är mycket lättare för en avsändare att skicka ett brev än det är för en mottagare att läsa brevet. När avsändaren har skrivit ett meddelande är kostnaden (i tid och ansträngning) att skicka det till flera mottagare mycket lägre än att skicka det till en mottagare. Speciellt stort blir problemen när kommunikationen går mellan flera användare, dvs vid gruppkommunikation.

Man kan naturligtvis föreställa sig vilka problem som uppstår för mottagaren av brevet. De vanligaste är:

- att brevlådan översvämmas av mängder med brev som kommer in och sättet att systematiskt sköta svarshanteringen blir mycket svårt.

- att sökningen efter relevant information bland mottagna brev blir nästan omöjlig.

De flesta e-postsystemen innehåller funktioner med vars hjälp samtliga ovanstående problem kan hanteras manuellt. I teorin kan en systematisk användning av dessa funktioner bemästra flertalet av de problem som normalt kan uppstå. I praktiken lider de flesta användare av brist på kompetens respektive tid, vilket ofta omöjliggör effektiva arbetsrutiner.

En radikal lösning på detta problem som samtidigt skulle öka användarvänligheten vore att sortera meddelandena i hanterliga mindre klasser på ett sådant sätt att det lätt går att välja meddelanden att läsa, genom att utveckla ett interaktivt intelligent filtersystem som skall hjälpa användaren att välja de

relevanta meddelandena i informationsströmmen [Ben-Cheikh och Zakrisson, DSV vt-94].

1.2 Uppdragsgivare

SICS (Swedish Institute of Computer Science) är ett forskningsinstitut som bedriver avancerad forskning inom flera områden i datavetenskap och har stark position inom bland annat språkteknologi, programmeringsmetodik, användaranpassade system och kunskapsbaserade system [SICS verksamhetsberättelse, 1993].

Inom SICS finns språkteknologigruppen som forskar i området språkteknologi (jmf eng Natural Language Processing). Området omfattar teknologi för användning av mänskligt språk för kommunikation mellan användaren och datorbaserade system samt datorstödd hantering av texter. Typiska tillämpningar är översättningsstöd, informationssökning, språkbaserade och talstyrda gränssnitt.

I samarbete med Data- och Systemvetenskapliga institutionen (DSV) driver språkteknologigruppen IntFilter-projektet. Det är finansierat av NUTEK (Närings- och teknikutvecklingsverket) och syftar till att bygga ett intelligent filtersystem för elektroniska meddelandehanteringssystem. Systemet kommer att behöva kunskap om användaren, dokument och kommunikation mellan användaren och systemet. Inom projektet kommer man att undersöka processen vid sortering och val av relevant information, som görs av användaren [IntFilter, 1992].

1.3 Problem

Ett av problemen i IntFilter-projektet är hur man med enkla metoder skall kunna särskilja olika sorters typer av texter.

l.4 Uppdrag

Språkteknologigruppen på SICS, som arbetar med IntFilter-projektet, har givit oss i uppdrag att göra en förstudie för att undersöka om en enkel metod som ordstatistik skulle kunna användas för text- och informationskategorisering efter texttyp.

1.5 Arbetssätt

Ett tidigare examensarbete [Ben-Cheikh och Zakrisson, DSV vt-94] har gjorts med liknande ämne [jmf kap 2.2] och vi kompletterar deras resultat med ytterligare

en undersökning baserad på det ovannämnda examensarbetet men på andra språkstatistiska undersökningar.

Texterna som kommer att behandlas är artiklar på engelska hämtade från Usenet News [jmf kap 2.1.2]. Efter en introduktion och analys av uppdraget har vi fastställt följande tre steg:

- Förarbete.
- Insamling och analys av data.
- Utvärdering av resultat.

1.5.1 Förarbete

Denna fas syftade till att vi skulle få nödvändig bakgrundskunskap om projektet IntFilter och om statistik. Detta gjorde vi genom arbetsmöten med vår handledare, Jussi Karlgren (projektmedlem), och med hjälp av litteratur inom området. Bland litteraturen finns examensarbetet *Genrekategorisering av text för filtrering av elektroniska meddelanden* [Ben-Cheikh och Zakrisson, DSV vt-94] som används som utgångspunkt för uppdraget [jmf kap 2.2].

1.5.2 Insamling och analys av data

En stor del av vårt arbete har gått åt till att samla in de texter, klassificerade artiklar, som kommer att användas för att få ut ordstatistik. Vi har valt att undersöka en delmängd av ord för att utveckla en metod. Utvärderingen sker enligt väletablerade statistiska principer. Vi har utfört följande arbetssteg:

1. Insamling av 1 000 artiklar, med den statistiska metoden obundet slumpmässigt urval [jmf kap 2.3.1], fördelade på 26 olika konferenser (newsgroups) från Usenet News.

2. Manuell kategorisering av artiklarna efter typer [jmf kap 3.1] enligt Kevin Crowston, Thomas W. Malone och Felix Lin [Crowston, Malone och Lin, 1988].

3. Metod för urval av ord, dvs ta fram en bra metod för att göra urval av ord. Detta steg kommer att ske i en iterativ testfas tillsammans med steg 4 och 5.

4. Utveckling av ett datorprogram som använder sig av metoden från steg 3 och räknar fram, av alla artiklar, antalet gånger och den procentuella andelen orden har förekommit.

5. Programkörning.

6. Urval av ord, dvs vilka ord vi skall använda i statistikanalysen.

7. Statistikanalyser med chi-två (chi-square) metoden [jmf kap 2.3.2].

Dessa arbetssteg beskrivs mer ingående i kap 3 och 4.

1.5.3 Utvärdering av resultat

Vi skall tillsammans med vår handledare utvärdera de resultat vi får från statistikanalyserna. Resultaten visar om det går att kategorisera texter med enkla metoder. Dessa metoder kan komma att användas och vidareutvecklas i IntFilter-projektet.

1.6 Målgrupp

För att en läsare utan svårigheter skall kunna följa denna uppsats krävs det att man har en del statistiska kunskaper. Examensarbetet behandlar ett praktiskt datavetenskapligt problem med statistiska metoder, men kunskap i lingvistik och språkvetenskap är inte nödvändiga.

På sid 24 i uppsatsen finns en ordlista där det finns förklaringar av en del ord som finns med i uppsatsen. Alla dessa ord är markerade med en asterisk (*) första gången de förekommer i uppsatsen.

2 Utvidgad bakgrund

Detta kapitel kommer att ge läsaren förberedande kunskap i olika användbara begrepp och undersökningsmetoder, som kommer att användas i vårt uppdrag. Syftet är att ge en bra överblick över de områden som behandlas av IntFilter-projektet och examensarbetet.

2.1 Elektroniska postsystem

En stor del av den kommunikation mellan människor som idag görs med de vanligaste tekniska kommunikationsmedierna, såsom brev och telefonsamtal, blir oerhört mycket effektivare om den istället går via elektroniska postsystem [jmf Palme, 1989].

System för elektronisk post används redan idag av miljontals människor runt om i världen. Om några decennier rör det sig kanske om hundratals miljoner människor. Eftersom de flesta elektroniska postsystem redan idag är hopkopplade i nät, kan användare skicka meddelanden över hela världen.

Det finns tre tillämpningar som motiverar ett nätverks existens: skrivar- och fildelning samt elektronisk post. De två första var orsaken till att de lokala nätverken skapades överhuvudtaget, medan e-post från början kom in som en intressant men inte helt nödvändig finess. För den som börjat använda e-post blir det dock snabbt lika nödvändigt som fil- och skrivardelning [Nätverk & Kommunikation, Okt 1994].

Kommunikation sker normalt asynkront, vilket innebär att avsändare och mottagare läser och skriver sina brev oberoende av varandra vid olika tidpunkter. En del system tillåter simultan kommunikation, så att en dialog kan ske i realtid. Den enda skillnaden mellan detta och ett vanligt telefonsamtal är som bekant att dialogen i e-postsystem sker digitalt och inte analogt. De flesta föredrar därför telefon i sådana fall då simultan kommunikation är nödvändig [jmf Palme, 1989].

Ett aldrig så avancerat och välkonfiguerat e-postsystem är i det närmaste värdelöst om det inte används. Det är därför nödvändigt att få igång användarna om systemet skall nå framgång. Nyttan med e-postsystem uppstår först när organisationen ändrat sitt beteende och blivit effektivare. Elektronisk post är ett relativt nytt medium, som ändrar människors och organisationers sätt att tänka. Att helt eller delvis ersätta andra kommunikationssätt måste vara målet med e-postsystem.

2.1.1 Användarfunktioner i elektroniska postsystem

Personliga brev är den mest grundläggande funktionen vid elektronisk post. Användarna startar det elektroniska brevsystemet på sin dator och skriver sina meddelanden, varefter de ger kommandot för att sända iväg dem.
Mottagaren startar regelbundet det elektroniska brevsystemet och frågar om det finns några nyinkomna brev. När mottagarna har läst ett brev, får de möjlighet att vidta olika åtgärder med brevet, såsom svara på brevet, spara det eller ta bort det ur datorns minne [jmf Palme, 1989].

2.1.2 Programvaror

Det finns tre nät för elektronisk post med stor spridning i Sverige, Internet*, MEMO* (eller Postnet) och X.400*.

Internet-mail är en stor standard för meddelandehantering som de flesta e-postsystem har kopplingar till och som miljoner användare runt om i världen är anslutna till. För svenska universitet och högskolor administreras Internet genom Sunet*, som har sin bas på KTH i Stockholm.

MEMO, som är det mest använda företagsinterna meddelandesystemet i Sverige, skapades ursprungligen av VolvoData och används idag av många andra företag och myndigheter.

Vid University of British Columbia i Kanada utvecklades 1984 ett av världens första meddelandesystem baserade på X.400-standarden. Det fick namnet EAN och har snabbt fått en stor spridning inom den akademiska världen. EAN marknadsförs kommersiellt under namnet Envoy 400 av företaget Sydney Development i Kanada [Nätverk & Kommunikation, Sept 1994].

Konferenssystem

Konferenssystem är ett hjälpmedel för gruppkommunikation, dvs när fler än två personer kommunicerar med varandra. Dessa används framför allt för erfarenhetsutbyte mellan experter på olika håll, inom företag eller mellan olika företag.

De mest kända konferenssystemen [jmf Palme, 1989] är:

- EIES*-familjen, som utmärker sig för en uppsjö av nya idéer och experimenteran-de.

- Kom-familjen, omfattande KOM*, PortaCOM* och SuperKOM*. Dessa utmärker sig för att ha god integrering mellan hanteringen av personligt adresserade brev och konferensinlägg.

- Usenet News, ett världsomspännande distribuerat nät av hopkopplade konferenssystem.

- Participate, ett amerikanskt system i vilket varje konversation behandlas som ett separat möte.

Usenet News

Usenet News är ett konferenssystem på Internet som har miljontals läsare och eftersom varje individ har tillgång till alla konferenser är det samtidigt mycket populärt. Diskussionerna är organiserade i konferenser som finns för en stor mängd olika ämnen, allt från medicin (sci.med) till skvaller (talk.rumours). Vem som helst kan starta en ny konferens och ämnesindelningen grenar sedan ut sig som ett träd. Om man har fastnat med ett programmeringsproblem, t ex i Prolog*, kan man lösa detta genom att skicka ett brev till comp.lang.prolog (dvs konferensen computer, undergruppen language och under-undergruppen prolog) med en beskrivning av problemet. Inom några dagar får man förhoppningsvis tillbaka olika svar från olika personer som kanske har haft erfarenhet av liknande problem.

Usenet News distribueras över hela världen och det finns både globala och lokala konferenser.

2.2 Examensarbetet av Naoufel Ben-Cheikh och Magnus Zakrisson

Examensarbetet *Genrekategorisering av text för filtrering av elektroniska meddelanden* hade som syfte att utreda möjligheterna att utveckla system, som genom induktiv inlärning lär sig att kategorisera texter. Detta gjordes med insamling av totalt 102 träningsartiklar[1] och senare också 30 testartiklar[2] från 10 olika konferenser i Usenet News. De kriterier för kategorisering som Ben-Cheikh och Zakrisson använde var bland annat, *artikelns längd, antal rader, antal ord per mening, antal tecken, antal citerade rader* och *utropstecken*, dvs ytliga textstrukturkriterier. Vi skall däremot använda ordstatistiska kriterier.

En av de förbättringsmöjligheter som tagits upp i uppsatsen *Genrekategorisering av text för filtrering av elektroniska meddelanden* är bland annat ökning av antalet träningsartiklar till 1 000 st.

De slutsatser författarna kom fram till är att nya okända artiklar kan klassificeras med cirka 65% säkerhet. Författarna tror även att genom att åtgärda de förbättringsmöjligheter som nämnts i uppsatsen [vt-94] kommer att öka säkerheten betydligt mer, till ca 90% säkerhet.

[1] Artiklar som används för att träna kategoriseringsalgoritmer [jmf Ben-Cheikh, Zakrisson, DSV vt-94].
[2] De artiklar som datorn själv skall kunna klassificera [jmf Ben-Cheikh, Zakrisson, DSV vt-94].

2.3 Statistiska urvalsmetoder

Statistisk analys involverar ofta ett försök att generalisera från data som finns tillgängliga. Det finns flera olika urvals- och analysmetoder som en statistiker kan använda sig av.

Vid planering av en statistisk undersökning är det viktigt att ha klart för sig vilken urvalsmetod som skall användas. Det finns flera olika metoder som används av statistiker över hela världen. Vissa är komplicerade och svåra att förstå medan andra inte är det.

Urvalsmetoderna kan delas upp i två kategorier, slumpmässiga och icke slumpmässiga. Det går en grundläggande skiljelinje mellan dem. De slumpmässiga (dvs obundet slumpmässigt urval, systematiskt urval, stratifierat urval eller gruppurval) karakteriseras av att varje urvalsenhet skall ha en känd sannolikhet som är större än noll för att bli utvald [Dahmström, 1994].

2.3.1 Obundet slumpmässigt urval

Den teoretiskt enklaste slumpmässiga urvalsmetoden är obundet slumpmässigt urval. Urvalet görs obundet i den mening att inga restriktioner läggs på att enheter med speciella egenskaper skall komma med i urvalet i vissa förutbestämda proportioner [Dahmström, 1994].

Antag att populationen består av N st enheter, i vårt fall alla artiklar i Usenet News, som samtidigt är elementen i undersökningen. Obundet slumpmässigt urval innebär att varje element har samma sannolikhet att bli utvald och att varje tänkbart urval om n element har samma sannolikhet. Det finns två varianter på denna metod. Dragning av element med återlämning, som i praktiken är ovanlig, och dragning av element utan återlämning, som vi har använt. I vårt fall har det inneburit, med dragning utan återlämning, att när en artikel har valts har den tagits bort ur mängden av de resterande artiklarna utan möjlighet att bli omvald.

Om man skall göra ett urval från en liten population kan man i praktiken skriva upp alla namnen på elementen och välja ut det sökta antalet genom att blanda och dra. Om populationen är stor kan man välja ut antalet element som behövs från en slumptalstabell [Dahmström, 1994].

2.3.2 Chi-två

Chi-två är en metod för att undersöka matriser för eventuella regelbundenheter utan att behöva sätta sig in i vad matrisen representerar. Chi-två ger ett enkelt skalärt mått på sannolikheten för att raderna och kolumnerna i en matris är relaterade på ett systematiskt sätt. Däremot ger inte chi-två någon fingervisning alls om vad relationen är eller vad den betyder.

Grundidén är att jämföra de observerade data i matrisen med förväntade värden givet att rad- och kolumnsummorna bibehålls som de är.

Som exempel ser vi här data från en liten korpus (textmassa) av talad och skriven engelsk text. Vi vill undersöka om pronomina *anybody* och *anyone* har systematiska skillnader som har att göra med språkets karaktär.

Ord	I skriven text	I talad text	Summa
anyone	57	17	**74**
anybody	31	32	**63**
Summa	**88**	**49**	**137**

Figur 2.1 Observerad matris.

Ord	I skriven text	I talad text	Summa
anyone	48	26	**74**
anybody	40	23	**63**
Summa	**88**	**49**	**137**

Figur 2.2 Förväntad matris.

Ord	I skriven text	I talad text
anyone	9	-9
anybody	-9	9

Figur 2.3 Observerad - förväntad matris.

Sannolikhet[3]	Chi-två
0,1000	2,7055
0,0500	3,8414
0,0250	5,0238
0,0100	6,6349
0,0050	7,8794
0,0001	10,8280

Figur 2.4 Chi-två tabell.

Första steget är att vi bygger en matris med de observerade värdena [jmf figur 2.1]. Sedan bygger vi en liknande matris där cellerna innehåller förväntade värden, givet rad- och kolumnsummorna. Om vi kallar de förväntade värdena e(i,j), radsumman r(i), kolumnsumman k(i) och totalsumman S (S=137 i exemplet) så blir:

$$e(i,j) = S\frac{r(i)}{S}\frac{k(i)}{S}$$

Vi får alltså medelvärdet av alla slumpmässiga fördelningar [jmf figur 2.2]. Vi subtraherar nu den observerade matrisen med den förväntade och får en tabell med differenserna [jmf figur 2.3].

Vi kvadrerar varje tal för att få bort minustecknen och dividerar sedan talen med det förväntade värdet. Summan av dessa tal är chi-två talet för den observerade matrisen:

$$Chi-två = \frac{81}{48}+\frac{81}{26}+\frac{81}{40}+\frac{81}{23}=10{,}35$$

Detta tal jämförs med en tabell som talar om hur stor sannolikheten är att matrisen innehåller regelbundenheter eller inte. Exempel på tabellen för 2x2-matriser ser vi på figur 2.4. (det finns naturligtvis en tabell för varje rimlig storlek på matriser eftersom chi-två talet är beroende av matrisstorleken).

Vi jämför vårt tal med talen i tabellen. Om det tal vi har fått inte finns i tabellen, som i vårt fall här, tar vi det minsta närmaste värdet, i vårt fall 7,8794. Som tolkning av värdet, kan man säga att sannolikheten att vår observerade matris är slumpmässigt fördelad, är mycket liten, dvs ungefär 0,005. Det betyder, om man ska tolka materialet, att *anyone* är mer skriftspråkligt än *anybody*.

[3]Dvs sannolikheten att matrisen är slumpmässigt fördelad.

3 Artikelinsamling och analys

Detta kapitel inleder vi med att beskriva vilka typer (alternativa benämningar: kategorier, klasser) vi använt och hur vi gick tillväga för att samla in de data som vi har behandlat. Vi berättar även hur vi kom fram till en metod för att göra urval av ord som ska användas i statistikanalysen. I slutet av kapitlet tar vi upp hur själva programkörningen gick till, vad programmet gör och hur vi behandlat resultaten av programkörningen.

3.1 Val av typer

Crowston, Malone och Lin [1988] har arbetat med ett projekt där de har använt sig av en objektorienterad programmeringsmodell som har gjort det möjligt att karakterisera information som utväxlas och behandlas mellan människor. Projektet skulle undersöka vilken relation som finns mellan informationsteknologi och organisationsstruktur. Vi kommer inte att fördjupa oss i projektet eller i modellen, eftersom det inte är relevant för vår uppsats.

Baserade på programmeringsmodellen och analysen av informationen, tog de fram fem typer av meddelanden, nämligen, *Policy*, *Problem*, *RequestForComments* (*RFC*), *Comment* och *andra klassificerade typer* (*Conference-System-Related*, *Announcements*, *Personal)*. Enligt författarna kan de flesta meddelanden klassificeras i dessa typer.

Policy

Till typen Policy hör meddelanden som förklarar en verksamhets policy eller är svar på frågor som handlar om ett systems funktioner. I denna typ ingår inte meddelanden som ger kommentarer angående någon policy eller är en beskrivning av problem med policyn. Exempel på meddelande som kan ingå i *Policy*:

- För att logga in sig i systemet X, skriv LOGIN <användare> ...
- Plan B kommer inte att uppdatera poster i databasen ...
- Använd inte systemet på fredagar eftersom ...

Comment

Till typen Comment hör meddelanden som ger kommentarer eller åsikter om något. Det kan vara svar till bland annat *Policy* eller *RFC*, men får däremot inte vara ett meddelande som förklarar en verksamhets policy eller är ett svar på en fråga som handlar om ett systems funktioner. Exempel på meddelande som ingår i *Comment*:

- För att förklara min åsikt ...
- Systemet skulle kunna göras bättre om ...
- Som svar på Marias fråga, skulle jag hålla med om att ...

RFC

Till typen RFC hör meddelanden som ber om kommentarer om något. Hit hör inte meddelanden som ber om hjälp angående implementeringar för ett system. Exempel på meddelande som ingår i *RFC*:

- Skulle ni kunna skicka mig en artikel med följande information ...
- Kan någon tala om för mig hur jag skulle kunna ...

Problem

Till typen Problem hör meddelanden från en person som ber om hjälp om något. Hit räknas inte meddelanden som ber om kommentarer om befintlig policy i en verksamhet. Exempel på meddelande som ingår i *Problem*:

- Vad kommer att hända med oss under dessa regler ...?
- Jag kommer inte förbi en bugg som ...

Andra klassificerade meddelanden

- **Conference-System-Related**: Meddelanden som vänder sig till en grupp eller som drar uppmärksamheten till en ny grupp.

- **Announcements**: Meddelanden som handlar om sökta arbeten, hus m.fl.

- **Personal**: Meddelanden med utseendet *Gott nytt år Johan*.

Vi har rationaliserat bort typen *andra klassificerade typer* och använder *Conference-System-Related, Announcement* och *Personal* som fristående typer på samma nivå som *Policy, Problem, RequestForComments* och *Comment*. Vi har döpt om typen Conference-System-Related till Conferences.

3.2 Insamling och kategorisering av artiklar

Vi inledde med att välja ut 25 st konferenser med obundet slumpmässigt urval [jmf kap 2.3.1], ur en lista över konferenser i Usenet News. Från varje konferens valde vi 40 st artiklar (även här med obundet slumpmässigt urval) som skulle utgöra de 1 000 artiklar vi skall behandla. Efter det att vi läst igenom alla artiklar, kategoriserade vi dem manuellt efter de typer som redovisats ovan [jmf kap 3.1]. Det visade sig dock, efter kategoriseringen, att antalet artiklar inte kom upp till de 1 000 som beräknat. Detta berodde på att vi var tvungna att ta bort några artiklar

som inte var på engelska, eftersom vi har beslutat att utföra undersökningen på engelska artiklar. För att ersätta de borttagna artiklarna var vi tvungna att lägga till ytterligare en konferens, varifrån vi inkluderade de artiklar som fattades. Det visade sig att 21 st artiklar inte gick att klassificera enligt de 7 typer vi redovisat i kap 3.1. Procentuellt är det inte en sådan stor andel (2,1%), därför tar vi inte hänsyn till dessa artiklar i vår undersökning. I tabellen nedan [jmf figur 3.1] ser vi fördelningen av artiklarna på de 26 konferenserna och 7 typerna, även de oklassificerade artiklarna finns med i tabellen (se Unclass).

Newsgroups	**CATEGORIES**								**Number of Articles**
	Announce	Comment	Confere	Unclass	Personal	Policy	Problem	RFC	
alt.activism.death-penalty	0	35	0	0	0	0	0	5	40
alt.artcom	4	4	15	3	0	0	7	5	38
alt.horror	1	26	1	0	1	1	3	12	45
alt.activism	0	29	2	0	0	10	0	1	42
alt.alien.visitors	0	26	0	2	9	2	0	2	41
alt.backrubs	5	13	1	2	8	3	1	2	35
alt.astrology	1	14	0	3	5	2	5	6	36
alt.architecture.alternative	2	7	1	3	3	4	5	7	32
bionet.general	6	10	0	0	3	3	7	11	40
comp.archives	0	9	3	1	0	28	0	0	41
comp.graphics.visualization	3	2	3	0	2	10	9	8	37
comp.infosystems	0	12	6	0	1	8	3	10	40
comp.lang.prolog	1	13	7	0	1	5	5	8	40
comp.lang.smalltalk	4	8	0	0	1	5	14	6	38
comp.multi-media	1	14	5	0	0	0	18	11	49
comp.protocols.ppp	1	10	0	0	0	0	21	7	39
comp.theory	1	6	15	0	2	5	3	9	41
comp.windows.news	0	21	3	1	0	9	2	2	38
comp.virus	0	12	0	5	1	9	8	4	39
comp.sys.amiga.cd32	0	6	0	0	0	0	1	0	7
eunet.jokes	0	34	0	0	0	0	0	0	34
gnu.emacs.vm.info	0	10	0	0	3	9	13	5	40
misc.books.technical	10	11	2	0	3	11	2	7	46
misc.creativity	0	23	0	1	3	3	2	2	34
rec.video.desktop	1	5	0	0	0	0	11	29	46
talk.rumours	1	38	0	0	0	1	2	0	42
Number of Articles	42	398	64	21	46	128	142	159	1000

Figur 3.1 Tabell över fördelning av artiklar bearbetade på konferenser och typer

Namnen på de flesta konferenser är förkortade som vi kan se på figur 3.1. Vad en del av dessa förkortningar står för har det varit svårt att få uppgift om. Det finns en artikel med en enradsbeskrivning över alla konferenser i Usenet News, men denna har varit omöjlig att hitta eftersom man måste veta exakt var man ska söka. Några förkortningar har vi dock lyckats ta reda på vad de står för, som bland annat *alt = alternative*, *comp = computer*, *sys = system*, *lang = language*.

Artikel före redigering:

Article 1688 of gnu.emacs.vm.info:
x-gateway: relay2.UU.NET from info-vm to gnu.emacs.vm.info; Wed, 30 Mar 1994 22:37:23 EST
Date: Wed, 30 Mar 1994 22:29:04 -0500
From: kyle@crystal.WonderWorks.com (Kyle Jones)
Message-ID: <9403310329.AA19870@crystal.WonderWorks.com>
Subject: Headers & Folders
References: <ROLFL.94Mar29142721@ask.uio.no>
Newsgroups: gnu.emacs.vm.info
Path: news.kth.se!sunic!pipex!uunet!wendy-fate.uu.net!info-vm
Sender: info-vm-request@uunet.uu.net
Lines: 7

Rolf Marvin Bxe Lindgren writes:
> How can I specify which headers to display accoring to which >folder I am visiting?

Set vm-visible-headers and vm-invisible-header-regexp in your vm-mode-hook. Test for various folder names, matching against buffer-file-name.

Artikel efter redigering:

Set vm-visible-headers and vm-invisible-header-regexp in your vm-mode-hook. Test for various folder names, matching against buffer-file-name.

Figur 3.2 Exempel på artikel före och efter redigering.

Vi gjorde några modifieringar på artiklarna som handlade om lösningar på problem eller liknande. Alla dessa artiklar börjar med upprepningen av själva problemet som på varje rad börjar med tecknet > [jmf figur 3.2].

För att ordstatistiken vi skulle få fram skulle bli trovärdig var vi tvungna att radera de rader som beskrev problemet. Vi raderade även de onödiga rader som systemet sätter in [jmf figur 3.2].

3.3 Metod för urval av ord

Ett av de problem som vi har stött på under vårt arbete är hur vi ska välja de ord som skall undersökas med chi-två metoden.

I början var tanken att vi skulle, med hjälp av vårt datorprogram, få ut ord som förekom minst ett procentuellt antal gånger, t ex 0,05% och därefter välja ord manuellt. Det visade sig, efter ett antal försök, att det var svårt att välja det rätta procenttalet. Om procenttalet var 0,10% fick man en alldeles för liten lista. Om man däremot ändrade procenttalet till 0,05% blev listan alldeles för stor, över 200 sidor. Säkerheten att man fick med alla ord av signifikant karaktär var inte heller så stor. Vi blev tvungna att prova en ny metod. Den gick ut på att filtrera de texter som skulle undersökas. Det finns två välkända metoder för att filtrera texter:

- Filtrering enligt en lista med onödiga ord. Man har en lista med onödiga ord som jämförs med texten man vill filtrera. Som resultat får man alla ord som inte finns med i listan.

- Filtrering enligt en lista med signifikanta ord. Man har en lista med signifikanta ord som jämförs med texten man vill filtrera. Som resultat får man alla ord som finns med i listan.

Eftersom det är arbetsamt att filtrera bort enligt en lista med onödiga ord (listan blir mycket stor) valde vi att filtrera enligt en lista med signifikanta ord. Programmet ska läsa in en lista med ord och sedan skriva ut antalet och den procentuella andelen dessa ord förekommer i texten. Problemet som nu uppstod var att välja vilka ord som skulle ingå i denna lista. Vi försökte ta fram en lista manuellt genom att läsa igenom olika texter och välja de ord vi ansåg vara intressanta. Vi tog fram en lista på omkring 70 ord och testkörde programmet men graden av tillförlitlighet blev inte så hög.

Till slut kom vi fram till en tredje filtreringsmetod som kanske låter lite komplicerad och tidskrävande, men som visade sig vara bäst i vårt fall. Eftersom vi skall undersöka hur våra texter skiljer sig beroende på typ, med ordstatistik, är skillnaden som finns mellan förekomsten av olika ord relevant. Vi vill endast ha de ord som är skevt fördelade, dvs förekommer ofta i en typ och kanske inte alls i en annan typ. Därför har vi beslutat att endast välja de ord som har en viss skillnad mellan typerna. För att ett ord skall bli valt skall skillnaden mellan den högsta och den lägsta andelen som ordet förekommer i typerna vara större eller lika med den fastställda skillnaden.

För att förtydliga detta följer här ett enkelt exempel. Säg att vi har tre olika typer, typ1, typ2 och typ3. Vi skall i detta exempel endast välja de ord som har en minimal skillnad på 0,02% (dvs den fastställda skillnaden är 0,02%). I typ1 förekommer ordet *data* 0,10%, i typ2 0,12% och i typ3 0,09%. Vi beräknar skillnaden mellan den största = 0,12 och den minsta = 0,09 andelen som ordet förekommer i texterna. Skillnaden är 0,03% (alltså större än den fastställda skillnaden som enligt ovan var 0,02%) och ordet väljs för att ingå i undersökningen.

Det kommer att ske ett tvåstegs-urval av ord. Det första steget sker i samband med programkörningen. Vi har ännu inte fastställt den minimala skillnaden men den kommer att fastställas efter några testkörningar av programmet. Det andra steget sker direkt efter programkörningen, då vi kommer att göra ett manuellt urval av de ord som klassificerar sig som skeva, eftersom vi behöver en matris med rimlig storlek.

3.4 Utveckling av datorprogram

Vi har utvecklat ett datorprogram i C++* som tar emot som parametrar ett antal textfiler[4] (max 7 st plus utskriftsfilen) och ett nummer som den fastställda minsta skillnaden mellan förekomsterna av orden. Programmet bygger upp ett binärt träd* med alla ord som förekommer och räknar ut hur många gånger varje ord förekommer i varje textfil. Sedan söks trädet igenom och alla ord vars skillnad är mindre än den fastställda skillnaden filtreras bort enligt filtreringsmetoden ovan [jmf kap 3.3].

Eftersom det inte var fastställt ifall vi skulle göra statistikanalysen på ett redan befintligt statistikprogram eller skriva ett eget program, i antingen C++ eller Prolog som beräknar chi-två talet lade vi in en extra finess i detta program så att matrisen även kan skrivas ut i Prologform.

Som resultat får man ut en matris som innehåller endast de utvalda orden, antalet gånger och den procentuella andelen orden har förekommit i varje textfil sorterade efter skevhet [jmf figur 3.3].

[4]En textfil innehåller artiklar av samma typ.

Minsta skillnad = 0,09000

Words	11com.txt		11pol.txt	
A	6	1.4963	52	3.5471
AND	4	0.9975	31	2.1146
ANNOUNCE	4	0.9975	0	0.0000
ANNOUNCEMENTS	5	1.2469	0	0.0000
COMPUTER	6	1.4963	3	0.2046
DATA	1	0.2494	17	1.1596
FOR	11	2.7431	12	0.8186

Figur 3.3 Exempel på utskrift av programmet. Minsta skillnad = 0,09%. Filerna som jämförs är 11com.txt[5] och 11pol.txt[6]. Orden, antalet ordförekomster och procentandelen skrivs ut.

3.5 Programkörning

Vi inledde med att försöka fastställa den minsta skillnaden genom att göra olika testkörningar. Vi provade med 0,05% som minsta skillnad men det resulterade i en alldeles för stor lista, om man däremot ökade minsta skillnaden till 0,10% blev listan alldeles för liten och risken att vi inte fått med viktiga ord var stor. Efter ytterligare testkörningar kunde vi fastställa minsta skillnaden till 0,09%.

Vi slog ihop alla texter som hör till en viss typ oavsett konferens [jmf figur 3.1, sammanslagning kolumnvis], så att vi fick sju olika textfiler. Programkörningen skedde med dessa filer och som resultat fick vi en lista på 6 sidor med alla skeva ord.

3.6 Urval av ord

Efter en genomgång av listan som producerades av programkörningen bestämde vi oss för att göra 4 olika sorters urval. Inom varje urval valde vi ut mellan 10 och 20 ord (vi valde att inte ta mer än 20 ord för att inte matrisen skulle bli för stor).

- Toppen på listan: de 10 mest skeva orden valdes ut som ett första urval eftersom de borde ge starkast resultat [jmf figur 3.4].

[5]Textfilen 11com.txt innehåller alla artiklar som ingår i konferens 11 och är av typen comment.
[6]Textfilen 11pol.txt innehåller alla artiklar som ingår i konferens 11 och är av typen policy.

Words	announce	comment	conference	personal	policy	probl	RFC	Sum
I	0,3555	1,4381	0,2472	2,0737	0,5474	3,3063	2,3051	10,2733
AND	2,4095	1,8275	3,1358	1,5442	1,9790	1,7494	1,6878	14,3332
THE	3,3576	4,9246	3,8685	3,7062	4,7529	3,5951	3,6645	27,8694
YOU	0,7189	0,7977	0,3483	1,5222	0,8832	0,3680	0,5845	5,2228
TO	1,8170	2,5616	1,7021	1,9854	2,2625	2,8478	2,3117	15,4881
IT	0,3318	1,1882	0,2126	1,3016	0,7469	0,9908	0,7487	5,5206
ARTICLE	0,3397	0,5690	0,1025	1,1692	0,1548	0,7983	1,0902	4,2237
UNIVERSITY	0,2370	0,0513	1,1126	0,1544	0,0752	0,0849	0,1248	1,8402
SUBJECT	0,3950	0,4954	0,1493	1,1471	0,2377	0,8322	1,1033	4,3600
OF	3,0021	2,7000	3,2338	2,9561	2,4664	2,2646	3,1063	19,7293
Sum	12,9641	16,5534	14,1127	17,5601	14,106	16,8374	16,7269	108,8606

Figur 3.4 Tabell över observerad procentuell fördelning av de 10 mest skeva orden beroende på typer.

- Tydliga innehållsord: ord med mycket betydelseinnehåll, som borde förväntas vara skeva [jmf figur 3.5].

Words	announce	comment	conference	personal	policy	probl	RFC	Sum
UNIVERSITY	0,2370	0,0513	1,1126	0,1544	0,0752	0,0849	0,1248	1,8402
VIDEO	0,0316	0,0424	0,0935	0,0000	0,0185	0,3057	0,6305	1,1222
CONFERENCE	0,0474	0,0078	0,3935	0,0000	0,0131	0,0057	0,0131	0,4806
MULTIMEDIA	0,0553	0,0190	0,3226	0,0000	0,0109	0,1529	0,1445	0,7052
MAIL	0,3160	0,1116	0,1583	0,0221	0,2203	0,1755	0,2102	1,2140
PROGRAMMING	0,2923	0,0402	0,0543	0,0000	0,0414	0,0226	0,0328	0,4836
CHAIR	0,0316	0,0067	0,2834	0,0000	0,0011	0,0000	0,0000	0,3228
APPLICATION	0,2528	0,0089	0,0709	0,0000	0,0283	0,0453	0,0263	0,4325
SYSTEMS	0,0869	0,0234	0,2578	0,0221	0,0698	0,1019	0,0854	0,6473
LANGUAGE	0,2291	0,0223	0,0437	0,0000	0,0327	0,0057	0,0657	0,3992
SMALLTALK	0,1659	0,0223	0,0000	0,0221	0,0076	0,1585	0,2233	0,5997
CONSTITUTION	0,2212	0,0123	0,0000	0,0000	0,0033	0,0000	0,0000	0,2368
SOFTWARE	0,1580	0,0312	0,0618	0,0000	0,0992	0,2208	0,0854	0,6564
CHICAGO	0,0237	0,0056	0,0151	0,2206	0,0065	0,0000	0,0000	0,2715
HALL	0,1975	0,0033	0,0362	0,0000	0,0153	0,0000	0,0000	0,2523
HOTEL	0,0000	0,0022	0,1975	0,0000	0,0000	0,0000	0,0000	0,1997
VIRUS	0,0079	0,0781	0,0015	0,0221	0,0840	0,1755	0,0657	0,4348
ARCHITECTURE	0,0316	0,0134	0,1568	0,0662	0,0076	0,0453	0,0854	0,4063
UFO	0,0000	0,0234	0,0000	0,1324	0,0000	0,0000	0,0000	0,1558
Sum	2,5438	0,5566	3,3213	0,662	0,834	1,7211	1,8785	11,5173

Figur 3.5 Tabell över observerad procentuell fördelning av innehållsord beroende på typer.

- Innehållsfattiga ord: ord vars betydelse inte är specifikt relaterad till ämnesområdet. Dessa ord valdes för att undersöka om det finns innehållsord som varierar mellan typerna och där variationen inte går att förklara med enkla innehållsvariationer [jmf figur 3.6].

Words	announce	comment	conference	personal	policy	probl	RFC	Sum
ONE	0,1343	0,2990	0,1658	0,3530	0,2192	0,2208	0,1707	1,5628
THINK	0,0079	0,1495	0,0075	0,1765	0,0469	0,0566	0,0919	0,5368
FIND	0,0395	0,0591	0,0166	0,1324	0,0829	0,1812	0,1051	0,6168
OUT	0,0553	0,2198	0,0573	0,1544	0,1221	0,1868	0,2036	0,9993
USING	0,1343	0,0591	0,0693	0,0882	0,1003	0,2095	0,0854	0,7461
EVEN	0,0316	0,0926	0,0075	0,1544	0,0676	0,0623	0,0722	0,4882
ONLY	0,0869	0,1718	0,0739	0,2206	0,1123	0,1189	0,1182	0,9026
SUCH	0,1659	0,0803	0,0362	0,0221	0,0992	0,0793	0,1445	0,6275
MORE	0,1580	0,2131	0,0799	0,1544	0,1407	0,1359	0,1248	1,0068
SEND	0,1817	0,0591	0,0889	0,0662	0,0665	0,0453	0,1051	0,6128
MUST	0,0632	0,0491	0,1357	0,0000	0,0425	0,0226	0,0328	0,3459
LOOK	0,0553	0,0524	0,0106	0,0662	0,1450	0,0396	0,0591	0,4282
Sum	1,1139	1,5049	0,7492	1,5884	1,2452	1,3588	1,3134	8,8738

Figur 3.6 Tabell över observerad procentuell fördelning av innehållsfattiga ord beroende på typer.

- Formord: ord som vid första åtanke inte borde vara skeva. Dessa ord valdes eftersom det är förvånande att de finns med i listan överhuvudtaget. Man skulle kunna förvänta sig att de skulle vara jämt fördelade [jmf figur 3.7].

Words	announce	comment	conference	personal	policy	probl	RFC	Sum
I	0,3555	1,4381	0,2472	2,0737	0,5474	3,3063	2,3051	10,2733
AND	2,4095	1,8275	3,1358	1,5442	1,9790	1,7494	1,6878	14,3332
THE	3,3576	4,9246	3,8685	3,7062	4,7529	3,5951	3,6645	27,8694
YOU	0,7189	0,7977	0,3483	1,5222	0,8832	0,3680	0,5845	5,2228
A	1,4299	2,0272	1,5573	1,9413	1,8733	2,3382	2,0293	13,1965
IS	0,8611	1,3689	0,8608	1,1913	1,3979	1,3418	1,2281	8,2499
FOR	1,1850	0,8992	1,2317	0,8604	1,0435	1,1833	1,3660	7,7691
MY	0,0395	0,2388	0,0271	0,3750	0,1265	0,5265	0,4203	1,7537
HAVE	0,2054	0,5210	0,1930	0,5956	0,3609	0,6907	0,5451	3,1117
ME	0,1422	0,2343	0,0256	0,3309	0,1069	0,3906	0,5188	1,7493
WAS	0,0948	0,5132	0,0452	0,4853	0,2170	0,1415	0,2102	1,7072
BUT	0,1422	0,4184	0,0950	0,4412	0,2453	0,5605	0,4531	2,3557
IN	1,5721	1,6211	1,3900	1,2133	1,5320	1,2738	1,3791	9,9814
YOUR	0,1027	0,2243	0,2156	0,5074	0,2770	0,1302	0,1248	1,5820
AM	0,1027	0,1093	0,1206	0,1103	0,0251	0,3906	0,3546	1,2132
Sum	12,7191	17,1636	13,3617	16,8983	15,3679	17,9865	16,8713	110,3684

Figur 3.7 Tabell över observerad procentuell fördelning av formorden beroende på typer.

4 Statistikanalys

Vid statistikanalysen använde vi oss av antalet förekomster orden hade och inte den procentuella andelen. Förklaringen är att chi-två förutsätter att man vill studera skillnaden mellan förekomster, inte mellan proportioner av förekomster. Till exempel kan en kategori vara mycket liten och då kan något ord slumpmässigt få en stor procentuell andel av förekomsterna. I själva verket är faktumet att det förekommer t ex fyra gånger istället för sex, som man förväntat sig, det mest intressanta.

För att visa att man inte kan transformera matriser hur som helst så tar vi exemplet på sidan 9. Om vi skulle multiplicera alla tal i cellerna med hundra skulle chi-två värdet också bli hundra gånger större (1 035) och det vill vi inte eftersom detta resulterar i en missvisande tolkning av värdet.

4.1 Chi-två program

För beräkningen av chi-två talet har vi utvecklat ett datorprogram i Prolog. Programmet läser in den observerade matrisen och beräknar den förväntade matrisen med hänsyn till antal ordförekomster. Därefter beräknar den chi-två värdet för matrisen.

4.2 Tolkning av chi-två värdet

Topp	6313,02
Innehållsord	5784,82
Innehållsfattiga	1437,41
Formord	5380,10

Figur 4.1 Chi-två värde för olika urval.

De beräkningar vi har fått fram [jmf figur 4.1] visar ett högt chi-två värde i alla urval. Det stärker våra misstankar om att den metod vi använde för att göra urval av skeva ord var riktig. Dessutom visar de höga chi-två talen att våra ord verkligen är skeva och att de inte är slumpmässigt fördelade. Det som vi reagerade mot i början var just dessa höga chi-två tal, eftersom de tabeller över chi-två fördelningar vi har, inte innehåller så höga chi-två tal. Efter genomgångar av resultaten med vår handledare kom vi fram till att detta bara bevisar att sannolikheten att våra matriser är slumpmässigt fördelade är mycket liten, mindre än 0,005 [jmf figur 4.2].

Frihetsgrader[7]	**0,900**	**0,100**	**0,050**	**0,025**	**0,010**	**0,005**
070	55,3290	85,5271	90,5312	95,0231	100,425	104,251
080	65,2778	96,5782	101,879	106,629	112,329	116,321
090	73,2912	107,565	113,145	118,136	124,116	128,299
100	82,3581	118,498	124,342	129,261	135,807	140,169

Figur 4.2 Ett utdrag av en chi-två tabell ur Aczel [1993]. Chi-två talen finns under sannolikheten .

[7]Frihetsgrader som betyder storleken på matrisen, dvs antal celler.

5 Utvärdering av resultat

Alla fyra urvalen av ord visar mycket starka resultat över icke slumpmässiga förekomster. I detta kapitel går vi igenom dem urval för urval och analyserar om det finns trovärdiga förklaringar till varför vi fick sådana höga chi-två värden.

5.1 Toppen på listan

Resultatet i toppen på listan är inte så förvånande. Egentligen bekräftar den bara att vårt urvalssätt med att ta de ord som hade mer än 0,09% i differens är korrekt. Dessa ord med störst differens är mycket riktigt skevt fördelade.

5.2 Innehållsord

Att innehållsord är skevt fördelade är naturligtvis inte helt förvånande. Texterna i de olika typerna handlar delvis om olika saker. Ordet *chair* (ordförande) förekommer flest gånger i texter om konferenser. *University* är till exempel överrepresenterad i typen *conferences*. Orden *video* och *multimedia* har samma distribution och förekommer flest gånger i *RFC* och *problem* (detta tyder på att den nya teknologin leder till problem för användare). Ett annat ord som förekommer ofta i typen *problem* är *help* och det är inte helt överraskande. *Announcement* har ju att göra med jobbannonser, där förekommer ord som är vanliga i sådana sammanhang, som till exempel *language*, *programming* och *mail*. Exempel med satser som innehåller innehållsord ur vårt material, ser vi i figur 5.1.

Satser ur Announcement:

If you are a strong C programmer, familiar with graphics programming (or are willing to learn, like I), creative, and willing to spend a lot of spare time on this during the summer, please e-mail me.

Medical Center Information Systems is seeking a Sr. Analyst Programmer to develop PC/workstation applications using Smalltalk.

Satser ur Conferences:

XX is the Cecilia Fitler Moore Professor and Chair of Computer and Information Science at the University of Pennsylvania and has been on that faculty since 1974.

Satser ur Problem:

Please help!
I need a software-only decompression algorithm, with a few restrictions:

Figur 5.1 Satser som innehåller skeva innehållsord.

5.3 Innehållsfattiga ord

Innehållsfattiga ord är på många sätt intressantare än de innehållsrika orden. För att kunna använda innehållsord som kriterium för att skilja kategorier av texter åt, så vill man inte nödvändigtvis tänka på innehållet. Dessa ord förekommer oavsett innehåll i texterna.

Satser ur problem

I would like to find a ppp client for OS/2.

Satser ur policy

To look up e-mail addresses and other information about our users.

You'll have to look at the creation date of the file - the file name will be always one and the same.

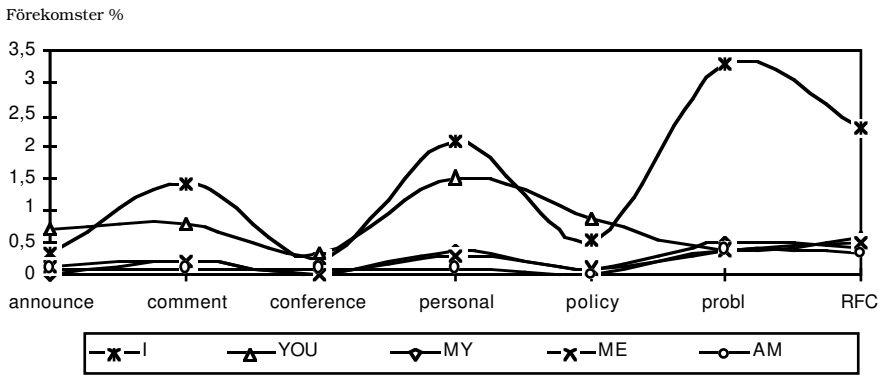
Figur 5.2 Satser som innehåller innehållsfattiga ord.

Even är mycket vanlig i personliga texter jämfört med andra ord. Ordet tillhör ett mer personligt språk. *Find* och *out* har samma distribution och förekommer mest i *problem* och *personal*. Ordet *look* förekommer flest antal gånger i typen *policy*, vilket inte är så konstigt. Exempel med satser som innehåller innehållsfattiga ord, ur vårt material, ser vi i figur 5.2.

5.4 Formord

Formord är ord som *jag, du, den, det, är, för* och liknande, dvs ord som har ingen eller liten betydelse, men som fogar ihop de mer betydelsebärande orden till en mening. Dessa ord verkar uppvisa en skev fördelning i våra texter. Det kanske kan tyckas överraskande eftersom formord borde förekomma likformigt över alla sorters språk (texter).

Vid närmare eftertanke är det så att språkets karaktär skiftar från stilsort till stilsort. Därigenom skiftar också meningsstrukturen och med den också formorden som används.



Figur 5.3 Graf över hur orden *I, me, you, my, me* och *am* skiftar mellan de olika typerna.

Till exempel visar personliga pronomen upp stora skiftningar mellan kategorierna. *I, me, my*, och *am* förekommer relativt sett ofta i kategorierna *comment, personal, problem, RFC* och relativt sett sällan i kategorierna *conferences* och *policy* [jmf figur 5.3].

Det är inte så överraskande om man tänker efter. I frågor om hjälp eller kommentarer hänvisar man ofta till sig själv.

Satser ur Problem:

I am struggling with a seemingly simple question and would appreciate some help.

The problem is that I have to add some products again
which I subtracted more than once, and doing everything right gets messy (and I'm afraid again exponential ;-( )

Satser med you ur Policy:

You can pick up a copy by anonymous FTP from site "ai.uga.edu", directory "ai.prolog.standard".

If you discover what you believe to be a virus on your Macintosh system, please report it to the vendor/author of your anti-virus software package for analysis.

Figur 5.4 Satser som innehåller skeva formord.

Ordet *you* har inte riktigt samma fördelning. Det förekommer ofta i *personal* precis som *I, me, my,* och *a m* men till skillnad från dem ofta i kategorin *policy*. Det är inte heller så konstigt. Exempel med satser som innehåller skeva formord, ur vårt material, ser vi i figur 5.4.

6 Slutsatser

Vårt uppdrag var att göra en förstudie för att undersöka om en enkel metod som ordstatistik skulle kunna användas för text- och informationskategorisering efter texttyp. Vi undersökte detta genom att se ifall vissa ordförekomster varierar mellan olika sorters typer av texter. För att åstadkomma detta använde vi oss av ordstatistik och en enkel egenutvecklad metod för att hitta de skeva orden [jmf kap 3.3]. Vi använde oss av en statistisk test (chi-två) för att undersöka om ordförekomsterna är slumpmässigt fördelade i olika typer av texter.

Att ordstatistik är en framkomlig väg för att kategorisera text och information efter typ bekräftades av att de chi-två värden vi fick ut ur statistikanalysen var mycket höga. Som summering av statistikanalysen kan vi säga att sannolikheten att ordförekomsterna är slumpmässigt fördelade, i olika typer av texter, är endast 0,005 [jmf kap 4.2]. Det finns övertygande förklaringar till varför ord varierar systematiskt. Några av dessa förklaringar har vi gått igenom i kapitel 5.

Under arbetets gång upptäckte vi hur vissa ord hör ihop med andra, dvs orden följer ett visst mönster [jmf figur 5.3]. Vilket kan vara mycket användbart vid utveckling av ett intelligent filtersystem. Resultaten visar att texter *om* olika saker inte bara har olika *innehåll* utan även olika *form*. Detta går att se med automatisk analys av texterna. Denna studie har visat var någonstans skillnaderna finns. Texterna i typen *RFC* har till exempel fler ord *multimedia* i sig än texter ur typen *policy*. Det kan till exempel bero på att området multimedia är så nytt att många har frågor om ämnet.

IntFilter-projektet skall nu ta resultaten från denna förstudie, i kombination med Ben-Cheikh och Zakrissons slutsatser och utreda exakt hur dessa skall tillämpas.

För att förbättra resultaten ytterligare kan man vid manuell kategorisering av texter låta fler personer läsa igenom alla texterna och kategorisera dem var för sig. Därefter kan man sammanställa alla kategoriseringar och enbart ta med i undersökningen de texter som hamnade i samma kategori hos varje person. På detta sätt kan man vara säker på att kategoriseringen blir mer objektiv.

Vidare rekommenderar vi att texterna man väljer kommer från olika källor och inte enbart från Usenet News, som i vårt fall. Man kan, om man vill, öka antalet texter i undersökningen, men vi tror inte att det nödvändigtvis skulle ge bättre resultat.

Något vi rekommenderar starkt är att man genomför ännu en undersökning med mer specifika typer och inte så generella, som i vårt fall.

Litteraturförteckning

Aczel, 1993 Amir D. Aczel, *Complete Business Statistics*, Second Edition, USA, Bentley College, 1993.

Ben-Cheikh och Zakrisson, vt-94 Naoufel Ben-Cheikh, Magnus Zakrisson, *Genrekategorisering av text för filtrering av elektroniska meddelanden*, Examensarbete, DSV, Vt-94.

Crowston, Malone och Lin, 1988 Kevin Crowston, Thomas W. Malone och Felix Lin, *Human Computer* Interaction: *Cognitive Science and Organisational Design*: A Case Study of Computer Conferencing, Volume 3, Number 1, Artikel 3, Massachusettes Institute of Technology (MIT), 1987-88.

Dahmström, 1994 Karin Dahmström, *Från Datainsamling till Rapport*, Studentlitteratur, Lund, 1994.

IntFilter, 1992 IntFilter, *NUTEK-ansökan*, 1992.

Karlgren och Cutting, 1994 Jussi Karlgren, Douglas Cutting, *Recognizing Text Genres with Simple Metrics Using Discrimant Analysis*, Skickad till Coling 94, Kyoto, Japan. Kan också fås genom DSV: IntFilterrapport nr 3.

Nätverk & kommunikation, Sept 1994 Nätverk & kommunikation, *Årgång 4, nummer 5*, September 1994.

Nätverk & kommunikation, Okt 1994 Nätverk & kommunikation, *Årgång 4, nummer 6*, Oktober 1994.

Office System, 1993 Office System, *Microsoft(R)*, Encarta, 1993.

Palme, 1989 Jakob Palme, *Elektronisk post*, Studentlitteratur, Lund, 1989.

SICS verksamhetsberättelse, 1993 SICS verksamhetsberättelse, 1992/1993.

Walters Lexikon, 1988 Walters Lexikon, *Computerlexikon*, Stockholm, 1988.

Ordlista

Binärt träd	Typ av trädstrukturerad index. Tillämpas mest då mycket sökning sker i datorns interna minne efter relevant information.
C++	C++ är ett objektorienterad programmeringsspråk.
EIES	EIES står för Electronic Information Exchange System. Det är ett datorstött konferenssystem, utvecklat av Murray Turrof vid New Jersey Institute of Technology.
Internet	Ett större nät, bestående av flera hopkopplade nät. Beteckningen används ofta för det hopkopplade nät, bestående av de olika akademiska näten i USA, vilka baseras på TCP/IP.
KOM	Datorstött konferenssystem, som utvecklats av FOA (Försvarets forskningsanstalt) 1978.
MEMO	System för elektronisk post för IBM-datorer, som utvecklats av VolvoData och säljs av Verimation AB. MEMO är känt för att vara effektiv och väl anpassad till IBM-operativsystem.
PortaCOM	Europeiska Gemenskapen (EG) beslöt 1979 att utveckla ett nytt konferenssystem, baserat på KOM. Det systemet kallas nu för PortaCOM.
Prolog	Prolog är ett logikprogrammeringsspråk.
Sunet	Det svenska akademiska forskningsnätet.
SuperKOM	En starkt förbättrad version av KOM, utvecklat av QZ UniversitetsData AB.
X.400-standarden	X.400 är CCITT's (Comité Consultatif International Télégraphique et Téléphonique) namn på den för framtiden viktigaste standarden för elektronisk post. ISO kallar samma standard för MOTIS. CCITT och ISO är två internationella standardiseringsorganisationer.