Hakutulosten esittäminen

Pekka Tapio Aalto

Helsinki 16.11.2001

Käyttöliittymätutkimus -seminaari
Raportti

HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen laitos

Hakutulosten esittäminen

Pekka Tapio Aalto

Käyttöliittymätutkimus -seminaari
Tietojenkäsittelytieteen laitos
Helsingin yliopisto
16.11.2001, 13 sivua

Internetin käytön laajentumisen myötä on kehitetty hakujärjestelmiä, joilla voidaan etsiä annetuilla hakusanoilla www-sivuja. Hakukoneen löytämät sivut palautetaan tyypillisesti listana, jossa eri aiheita käsittelevät sivut ovat usein sekoittuneet keskenään. Tässä raportissa keskitytään siihen, miten käyttäjälle voidaan tarjota riittävästi tietoa, jonka avulla hän pystyy nopeammin päättelemään, sisältääkö tulosdokumentti käyttäjän etsimää tietoa. Ensin esitellään menetelmiä, joilla tulosdokumentit voidaan ryhmitellä. Tämän jälkeen esitellään järjestelmiä, joissa on pyritty ratkaisemaan hakutulosten esittämisen ongelmia.

Aiheluokat (Computing Reviews 1998): H.3.3, H.5.2, H.5.3

Avainsanat: hakutulos, esittäminen, ryhmittely

Sisältö

1 Johdanto 1

2 Hakutulosten ryhmittely 1
2.1 Ryhmittely rakenteellisen tiedon avulla 2
2.2 Ryvästys 2
2.3 Luokittelu 3

3 Hakutulosten esittäminen 3
3.1 Tekstimuotoiset menetelmät 4
3.1.1 Cha-Cha 4
3.1.2 SWISH 5
3.1.3 Grouper 6
3.2 Graafiset menetelmät 8
3.2.1 NIRVE 8
3.2.2 Tehostetut kuvakkeet 9

4 Yhteenveto 11

Lähteet 12

1 Johdanto

Internetin käytön laajentumisen myötä on tiedon hakeminen muuttunut entistäkin vaikeammaksi. Tiedon löytämiseksi on kehitetty hakujärjestelmiä, joilla voidaan etsiä annetuilla hakusanoilla sivuja, joissa hakusana esiintyy. Suurin osa näistä hakujärjestelmistä palauttaa järjestetyn listan tulosdokumenteista. Käyttäjä joutuu usein käymään lävitse pitkän listan hakutuloksia, joissa eri aiheita käsittelevät www-sivut ovat sekoittuneet keskenään. Haettavan tiedon löytäminen voi muodostua ongelmalliseksi tilanteissa, joissa annetulla hakusanalla on useita eri merkityksiä. Sama ongelma ilmenee myös yleisesti tekstikannoissa, joissa haku tehdään suuresta tietomäärästä.

Tämän ongelman ratkaisemiseksi on hakukoneita pyritty kehittämään monella tavalla. Hakukoneiden tehokkuutta pyritään parantamaan ensinnäkin palauttamalla korkeampilaatuisia hakutuloksia. Toisaalta käyttäjälle pyritään tarjoamaan riittävästi tietoa, jonka avulla hän pystyy nopeammin päättelemään sisältääkö tulosdokumentti käyttäjän etsimää tietoa [WFR01]. Tässä raportissa keskitytään lähinnä jälkimmäiseen tapaan, jossa tutkitaan, mitkä seikat auttavat käyttäjää hakutulosten arvioinnissa.

Tulosdokumenttien ryhmitteleminen on yksi lähtökohta hakutulosten esittelemiseksi tehokkaasti. Luvussa 2 esitellään lyhyesti kolme erilaista menetelmää, joilla voidaan ryhmitellä hakukoneiden tuloksia ryhmiin. Ryhmittely tapahtuu rakenteellisen tiedon avulla, ryvästäen tai luokitellen Luvussa 3 on esimerkkejä järjestelmistä, joiden tavoitteena on ollut ratkaista hakutulosten esittämiseen liittyviä ongelmia. Esiteltävät järjestelmät on jaettu tekstimuotoisiin ja graafisiin järjestelmiin. Tekstimuotoisia järjestelmiä ovat Cha-Cha, SWISH ja Grouper. Graafisista järjestelmistä esitellään NIRVE ja tehostettujen kuvakkeiden järjestelmä. Luvussa 4 on yhteenveto.

2 Hakutulosten ryhmittely

Hakutulosten ryhmittely on erittäin tarpeellista, jos käyttäjä on antanut vain muutamia hakusanoja. Tällaiset kyselyt palauttavat yleensä suuren määrän heterogeenisiä dokumentteja eri aihealueilta. Tämän kaltaisessa tilanteessa hakutulosten pitäisi opastaa

käyttäjää sopivan lähtökohdan löytämisessä [Che99]. Seuraavaksi esitellään kolme erilaista menetelmää, joilla hakutulokset voidaan jaotella ryhmiin.

2.1 Ryhmittely rakenteellisen tiedon avulla

Tässä menetelmässä dokumentit ryhmitellään niiden *rakenteellisen tiedon* (structural information) perusteella. Ryhmittelyssä käytetty rakenteellinen tieto on jokaisesta dokumentista saatavissa olevaa tietoa. Tämä tieto voi olla esimerkiksi dokumentin sijainti sivustolla tai dokumenttiin viittaavien linkkien määrä [ChD00].

DynaCat-järjestelmä käytti UMLS:n (Unified Medical Language System) lääketieteellisen hakuteoksen terminologiaa hakutulosten ryhmittelemiseksi [ChD00, DCC01]. Tässä raportissa esitelty Cha-Cha –järjestelmä ryhmitteli dokumentit lyhimmän juuritasolta dokumenttiin kulkevan polun perusteella. [ChD00, Che99].

Manuaalisesti ylläpidettävät järjestelmät ovat hyödyllisiä, mutta ne ovat vaikeita ylläpitää. Automaattisesti ylläpidettävät järjestelmät puolestaan johtavat usein heterogeenisiin luokittelukriteereihin ja voivat olla vaikeita ymmärtää [ChD00].

2.2 Ryvästys

Hakutulokset voidaan *ryvästää* (clustering), jolloin dokumenteista muodostetaan ryhmiä dokumenttien keskinäisen samankaltaisuuden perusteella. Ryvästyksessä dokumenteista pyritään muodostamaan saman aihealueen asioita käsitteleviä dokumenttiryhmiä. Ryvästystä voidaan tarkentaa valitsemalla jokin rypäs tarkemman tutkimuksen kohteeksi. Valitusta ryppäästä muodostetaan uusi ryvästys, joka esitetään käyttäjälle [ChD00].

Ryvästyksen ongelmana on hakutulosten tosiaikaiseen ryvästämiseen tarvittava aikavaatimus, sillä joillakin algoritmeilla aikavaatimus kasvaa dokumenttien suhteen eksponentiaalisesti. Toinen ongelma on ryvästyksellä muodostettujen ryhmien nimeäminen, sillä ryppäät nimetään tyypillisesti niissä usein esiintyvillä ilmaisuilla. Erityisen ongelmallista on käsityksen muodostaminen ryppään dokumenttien sisällöistä sen otsikon perusteella [ChD00]. Useimmat ryvästysalgoritmit käsittelevät dokumentteja järjestämättömänä joukkona sanoja, jolloin menetetään arvokasta tietoa [ZaE99].

Ryvästyksestä esitellään esimerkkeinä Zamirin ja Etzionin kehittämä Grouper-järjestelmä, jossa hakutulokset ryhmitellään suffiksipuun ryvästyksellä sekä NIRVE-järjestelmä, jossa käsitteiden pohjalta muodostettuja ryppäitä voidaan tutkia kolmiulotteisesti.

2.3 Luokittelu

Luokittelussa (classification) haun tulosdokumentit ryhmitellään Yahoo-hakemiston kaltaisiin aiheluokkiin. Aiheluokat muodostuvat tyypillisesti joukosta yleisiä, ylemmän tason luokkia sekä näitä tarkentavista alemman tason luokista. Tulosdokumenttien automaattisessa luokittelussa käytetään hyväksi tilastollisia menetelmiä dokumenttien ryhmittelemisessä. Luokittelu tapahtuu mallilla, jota opetetaan ensin joukolla luokiteltuja dokumentteja. Dokumentit saadaan tyypillisesti www-hakemistoista. Opetuksen jälkeen mallia käytetään uusien dokumenttien luokittelemiseksi. Mallin palauttamaa luokittelua käytetään hyväksi tulosten esittämisessä käyttäjälle [ChD00].

Luokittelun vahvuus on sen käyttämä yhdenmukainen luokittelu, jolla voidaan auttaa käyttäjää nopeasti paikallistamaan tarvitsemansa tiedon. Ongelmallista luokittelussa on riittävän tarkkojen luokkarajojen määrittely. Tämän lisäksi hakujen tulokset eivät välttämättä sijoitu mihinkään luokkaan riittävän hyvin [ChD00].

Luokittelevasta ryhmittelystä esitellään esimerkkinä Chenin ja Dumaisin kehittelemä SWISH-hakujärjestelmä, jolla hakukoneen palauttamat tulosdokumentit luokitellaan suoritushetkellä ja luokitukset esitellään käyttäjälle.

3 Hakutulosten esittäminen

Hakukoneiden palauttamat tuloslistat voidaan esittää usealla eri tavalla. Ensin tutustutaan kolmeen tekstimuotoiseen menetelmään, joissa hakutulokset esitetään tekstin muodossa. Tämän jälkeen perehdytään kahteen graafiseen menetelmään, joissa hakutulokset on esitetty graafisesti.

3.1 Tekstimuotoiset menetelmät

3.1.1 Cha-Cha

Cha-Cha on Kalifornian yliopistossa kehitetty hakujärjestelmä, jonka tavoitteena on organisoida intranetistä tehtyjen hakujen tulokset tavalla, jolla käyttäjä saa kuvan tulosdokumenttien suhteesta sivustoon. Tulosdokumentit ryhmitellään niiden sivustolla olevan sijainnin perusteella käyttämällä hyväksi sivun rakenteellista tietoa. Cha-Chan tavoitteena on auttaa käyttäjiä ymmärtämään paremmin hakutuloksia sekä niiden välisiä suhteita, ja auttaa haettavan tiedon löytämisessä sekä muodostamaan käsityksen sivuston yleisrakenteesta [Che99].

Cha-Cha rakentuu kahdesta komponentista: tiedon indeksoijasta ja kyselyn käsittelijästä. Tiedon indeksoija käy lävitse intranetin sivustoja etsien kaikki ne sivut, joihin löytyy polku sivuston päätasolta. Tässä vaiheessa sivuista etsitään ja tallennetaan niiden lyhimmät polut päätason sivuille. Polun lisäksi sivusta tallennetaan mm. sen otsikko, sivun koko sekä viimeisin muokkauspäivä. Sivujen indeksitiedot tallennetaan Kalifornian yliopistossa kehitettyyn Cheshire II –hakujärjestelmään, jota käytetään Cha-Chan hakukoneena [Che99].

Cha-Chan kyselyn käsittelijä vastaanottaa käyttäjän tekemät kyselyt ja palauttaa käsiteltyjen kyselyiden tulokset. Käyttäjän antamista hakusanoista generoidaan kysely, joka välitetään Cheshire II -hakukoneelle. Kyselyä tehtäessä käyttäjällä on mahdollisuus valita näytetäänkö tulosdokumentit listamuotoisesti (list view) vai jäsennettynä näkymänä (outline view). Jos käyttäjä valitsee listamuotoisen näkymän, niin hänelle muodostetaan perinteinen listamuotoinen tulossivu. Jos käyttäjä sen sijaan valitsee jäsennetyn näkymän, niin käsittelijä hakee tulosdokumenttien metatiedot ja muodostaa hierarkian tulos-dokumenttien lyhimpien polkujen perusteella. Muodostetusta hierarkiasta generoidaan käyttäjälle jäsennetty näkymä. Kuvassa 1 on Cha-Chan palauttama, jäsennettynä näkymänä generoitu tulossivu, kun järjestelmällä haettiin earthquake-sanan esiintymiä sivustosta [Che99].

Jäsennetyn näkymän ylimmän tason hierarkia koostuu yleensä palvelimista, joissa tulosdokumentit sijaitsevat. Käyttäjällä on mahdollisuus saada tulosdokumentista

yhteenveto, johon on koottu dokumentin otsikon lisäksi muutama lause dokumenteista joissa hakusana esiintyy [Che99].

Tutkijat tekivät pienen käyttötutkimuksen sekä keräsivät kyselyllä tietoa Cha-Chan käyttäjiltä. Tulokset osoittivat, että osanottajat kokivat jäsennetyn näkymän helpoksi käyttää. Yli puolet kyselyyn vastanneista ilmoittivat, että he pystyivät löytämään jäsennetyn näkymän avulla tietoa, jota eivät olisi muulla tavalla löytäneet [Che99].

3.1.2 SWISH

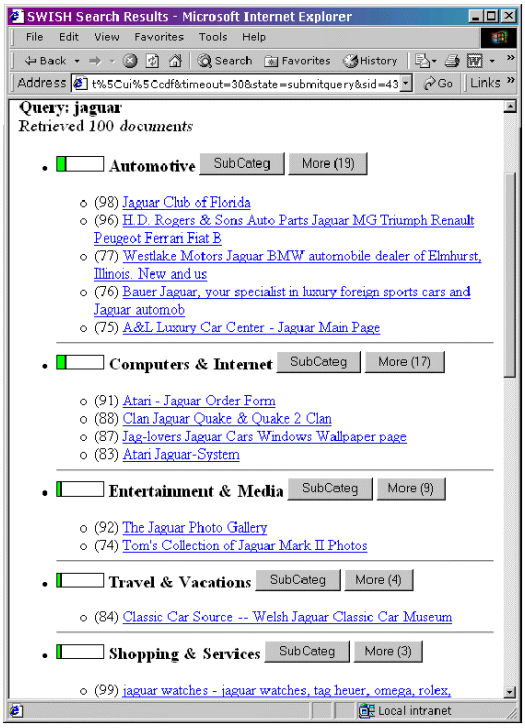
SWISH on Kalifornian yliopiston ja Microsoftin tutkimusosaston yhdessä kehittelemä hakujärjestelmä, joka luokittelee hakutulokset hierarkisiin luokkiin. SWISHissä käytetään automaattista hakutulosten luokittelua. Luokitetuilla www-sivuilla opetetaan ensin tilastollista luokittelumallia, jonka jälkeen mallia voidaan käyttää hakutulosten luokittelemiseksi automaattisesti [ChD00].

Alustusosassa mallin opettamiseen käytettiin LookSmartin hakemistoa. Artikkelin kirjoitushetkellä LookSmartissa oli 13 ylemmän tason luokkaa ja 150 alemman tason luokkaa, yhteensä luokkia oli yli 17000. Kokonaisuudessaan luokkien opettamiseen kului muutama tunti [ChD00].

Järjestelmän käyttöliittymä välittää hakusanat käyttäjän valitsemalle hakukoneelle ja luokittelee hakukoneen palauttamat tulosdokumentit. Jokainen palautettu tulosdokumentti luokitellaan yhteen tai useampaan luokkaan käyttämällä opetettua mallia. Lopulta luokitelluista dokumenteista generoidaan tulossivu. Tulossivulle otetaan ainoastaan ne ylimmän tason luokat, joille tulosdokumentit luokittuivat. Ensimmäiselle tulossivulle otetaan ainoastaan pieni osa tulosjoukon dokumenteista. Valinta tapahtuu tutkijoiden toteuttamalla valikoivalla heuristiikalla, jolla pyritään löytämään käyttäjän kannalta mielenkiintoisimmat tulosdokumentit. Luokitus ja sivulle tulevat dokumentit tulostetaan lukumäärällisesti laskevassa järjestyksessä. Kuvassa 2 on esimerkki SWISHin generoimasta tulossivusta, jossa on luokiteltu jaguar-sanalle suoritetun haun tulosdokumentit [ChD00].

Tutkijat suorittivat laajemman käyttötestin, jossa testattiin seitsemää eri variaatiota. Variaatioissa tulossivun luokitukset oli esitetty eri menetelmillä. Käyttötutkimukset

Kuva 1. Cha-Cha hakujärjestelmän palauttama jäsennetty näkymä.



Kuva 2. SWISH hakujärjestelmän luokittelemat tulosdokumentit.

osoittivat, että luokitellut tulokset olivat tehokkaampia kuin listamuotoiset tulokset jopa tilanteessa, jossa luokka kerrottiin jokaisen tulosdokumentin yhteydessä. Parhaimpaan tulokseen päästiin kun sekä luokkien nimet että yksittäisten sivujen otsikot oli annettu [DCC01].

Tutkijat huomasivat, että dokumentti luokiteltiin noin 70% varmuudella oikeaan luokkaan verrattuna siihen, miten ihminen olisi sen luokitellut. Ristiriitaisuudet johtuivat enimmäkseen siitä, että oikean luokan lisäksi oli määritelty vaihtoehtoisia luokkia tai yhtään sopivaa luokkaa ei oltu määritelty. Tämän lisäksi tutkimukset osoittivat, että käyttäjät löysivät haettavan tiedon 50% nopeammin listamuotoiseen tulossivuun verrattuna [ChD00].

3.1.3 Grouper

Grouper on Washingtonin yliopistossa kehitetty hakutulosten ryhmittelijä, joka on toteutettu HuskySearch-metahakupalvelun yhteyteen. HuskySearch suorittaa käyttäjän määrittelemän kyselyn usealla yleisellä hakukoneella ja kokoaa osakyselyiden

tulosdokumentit yhteen. Grouper ryvästää tulosdokumentit ryppäisiin ja nimeää ryppään lauseilla, joiden perusteella tulosdokumentti ryvästetään. [ZaE99].

Grouper käyttää tutkijoiden kehittelemää *suffiksipuun ryvästystä* (suffix tree clustering, STC), joka on nopea ja toimii lineaarisessa ajassa. Ryvästys tapahtuu täydentyvästi, jolloin tulosdokumenttien ryvästys voidaan aloittaa heti kun ensimmäiset hakutulokset on saatu hakukoneelta. Näiden ominaisuuksien lisäksi suffiksipuun ryvästys tuottaa johdonmukaisia ryppäitä, sillä dokumentit ryvästetään yksittäisten sanojen lisäksi myös useamman sanan lauseilla [ZaE99].

Hakusana syötetään Grouperille ja valitaan, kuinka monta tulosdokumenttia kultakin hakupalvelimelta palautetaan. Grouper suorittaa kyselyt eri hakukoneissa ja aloittaa tulosdokumenttien ryvästyksen. Kun Grouper on vastaanottanut ja ryvästänyt kaikkien hakukoneiden tulokset, voidaan tulossivu generoida. Ryvästyksen tulokset tulostetaan taulukkoon, jossa jokaisesta ryppäästä kerrotaan tulosdokumenttien määrä, ryppään kuvaavat lauseet sekä kolmen esimerkkidokumentin otsikko. Kuvaavat lauseet esiintyvät useimmissa ryppään dokumenteissa. Kuvassa 3 on esimerkki Grouperin generoimasta tulossivusta, jossa israel-sanalla suoritetun haun tulokset on ryvästetty [ZaE99].

Tutkiessaan Grouperin käyttökelpoisuutta tutkijat käyttivät apuna Grouperin ja

Query: israel
Documents: 272, Clusters: 15, Average Cluster Size: 15.1 documents

Cluster	Size	Shared Phrases and Sample Document Titles
1 View Results Refine Query Based On This Cluster	16	Society and Culture (56%), Faiths and Practices (56%), Judaism (69%), Spirituality (56%); Religion (56%) , organizations (43%) Ahavat Israel - The Amazing Jewish Website! Israel and Judaism Judaica Collection
2 View Results Refine Query Based On This Cluster	15	Ministry of Foreign Affairs (33%), Ministry (87%) Publications and Data of the BANK OF ISRAEL Consulate General of Israel to the Mid-Atlantic Region The Friends of Israel Gospel Ministry
3 View Results Refine Query Based On This Cluster	11	Israel Tourism (36%), Comprehensive Israel (36%), Tourism (64%) Interactive Israel tourism guide - Jerusalem Ambassade d'Israel Travel to Israel Opportunites
4 View Results Refine Query Based On This Cluster	7	Middle East (57%), History (57%); WAR (42%) , Region (42%) , Complete (42%) , Listing (42%) , country (42%) Israel at Fifty: Our Introduction to The Six Day War Machal - Volunteers in the Israel's War of Independence HISTORY: The State of Israel
5 View Results Refine Query Based On This Cluster	22	Economy (68%), Companies (55%), Travel (55%) Israel Hotel Association Israel Association of Electronics Industries Focus Capital Group - Israel

Kuva 3. Grouperin tulossivu, jonka tulokset on ryvästetty israel-sanan tulosdokumenteista.

HuskySearchin käyttölokeja. Lokeja vertailemalla tutkijat huomasivat, että käyttäjät vierailivat useammin Grouperin generoiman ryvästetyn tulossivun dokumenteissa kuin HuskySearchin listamuotoisen tulossivun dokumenteissa [ZaE99].

3.2 Graafiset menetelmät

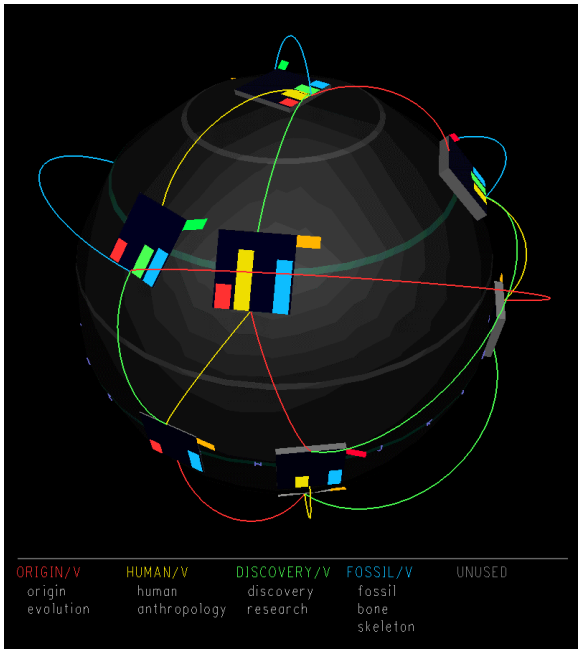
3.2.1 NIRVE

National Institute of Standards and Techology (NIST) –järjestön visualisoinnin ja virtuaalitodellisuuden ryhmä on kehittänyt prototyypin, jolla käyttäjä pystyy muodostamaan yleiskuvan hakutuloksista. Sebrechtsin ja kumppaneiden kehittelemä *NIRVE* (NIST Information Retrieval Visualization Engine) on järjestelmä, jolla pystyy tutkimaan hakutuloksia sekä kaksi- ja kolmiulotteisesti että tekstimuotoisena [Seb99].

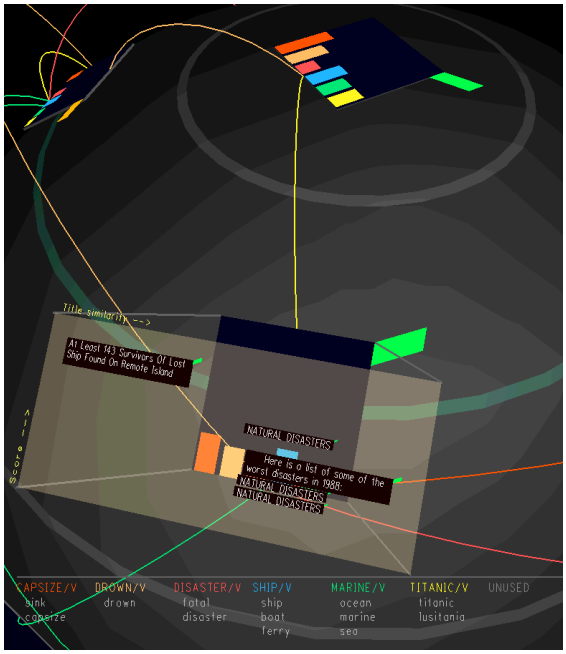
NIRVEssä hakutuloksen dokumentit ryvästetään *käsitteiden* (concept) perusteella. Käsite on joukko hakusanoja, jotka ovat merkitykseltään lähellä toisiaan. Ryppäät muodostuvat käsitteistä siten, että ryppääseen tulee ainoastaan ne dokumentit, joista löytyy kyseisen ryppään käsitteet. Kuvassa 4 on muodostetut ryppäät sijoitettu tasoittain pallon pinnalle. Ylimmässä tasossa sijaitsee rypäs, jonka dokumentit sisältävät kaikki käsitteet. Seuraavista tason ryppäistä puuttuu vuorollaan jokin ylemmän tason käsitteistä. Ryppäät kytketään ylemmän tason ryppääseen puuttuvan käsitteen tunnusvärin mukaisella yhdysviivalla. Jokaisesta ryppäästä on pallon pinnalla pylväskuvaaja, joka kertoo kuinka paljon kyseinen rypäs käsittelee kutakin käsitettä. Ryppään korkeus pallon pinnasta ilmaisee ryppäässä olevien dokumenttien määrän [Seb99].

Kuvassa 5 on yksittäisen ryppään tiedot suurennettu tutkittavaksi. Ryppään dokumentit sijoitetaan avautuvaan suorakaiteeseen siten, että samankaltaiset otsikot sijoittuvat suunnilleen samalle pystysarakkeelle. Dokumentti sijoittuu hakutuloksessa sitä paremmin mitä ylempänä dokumentin otsikko sijaitsee suorakaiteessa [Seb99].

NIRVEn kaksiulotteinen malli vastaa toiminnaltaan kolmiulotteista mallia. Kaksiulotteisessa mallissa pallo on litistetty kaksiulotteiseen muotoon ja ryppäät on sijoitettu samalla tavalla kuin pallon pinnalla. Tekstipohjainen malli pyrkii kertomaan tekstinä ryppäistä kaiken sen, mitä kaksi- ja kolmiulotteisissa malleissa ilmaistaan grafiikan avulla [Seb99].



Kuva 4. NIRVEn hakutulosnäkymä, jossa ryppäät on sijoitettu pallon pinnalle.



Kuva 5. Yksittäisen ryppään tiedot on suurennettu tarkasteltavaksi NIRVEn hakutulosnäkymässä.

Tutkijat tekivät käyttötestin, jonka tarkoituksena oli verrata keskenään tekstipohjaista, kaksi- ja kolmiulotteista mallia. Testi osoitti, että kolmiulotteisen mallin käyttönopeus oli harjoituksen jälkeen verrattavissa kaksiulotteisen ja tekstipohjaisen mallin käyttönopeuteen. Lisäksi ilmeni, että käyttäjät joilla oli enemmän kokemusta tietokoneen käytöstä pystyivät käyttämään kolmiulotteista mallia nopeammin kuin kaksiulotteista mallia. Testissä ilmeni myös, että asioiden ja toimintojen värikoodaus oli erittäin käyttökelpoinen ja tehokas menetelmä [Seb99].

3.2.2 Tehostetut kuvakkeet

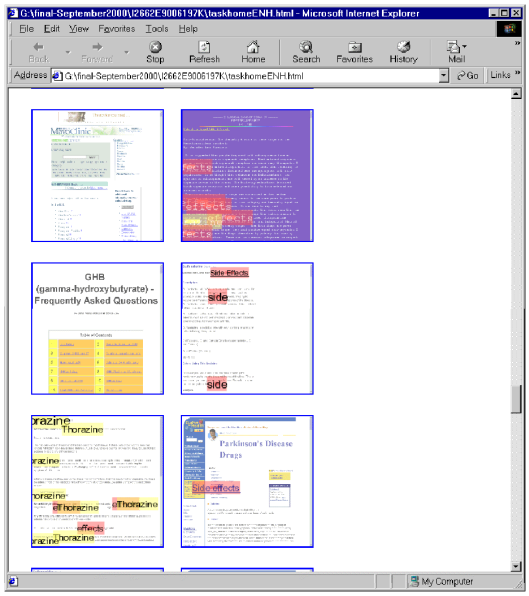
Woodruff ja kumppanit esittelevät artikkelissaan [Woo01] menetelmän, jolla hakutulokset esitellään tekstuaalisesti *tehostettuina kuvakkeina* (enchanced thumbnails). Tehostetuissa kuvakkeissa yhdistyvät pelkkien kuvakkeiden ja tekstimuotoisen yhteenvedon hyödyt. Tehostetuissa kuvakkeissa tulosdokumentista luodaan pienennetty kuvake, josta ilmenee tulosdokumentin tyyli ja sommittelu [Woo01].

Tehostetut kuvakkeet muodostettiin kolmessa vaiheessa. Ensimmäisessä vaiheessa alkuperäisen tulosdokumentin elementtien ulkoasua muokattiin esikäsittelijällä. Esimerkiksi tulosdokumentin otsikkotekstien tekstikokoa kasvatettiin siten, että otsikko oli

luettavissa muodostetusta kuvakkeesta. Toisessa vaiheessa kuvake luotiin käsittelijällä. Viimeisessä vaiheessa kuvakkeelle tehtiin värikäsittely, jolla kuvakkeen kaikkia värejä vaalennettiin selvästi. Tämän jälkeen kuvakkeeseen lisättiin hakusanat siten, että ne tulevat kuvakkeessa selvästi esille. Kuvassa 6 on esimerkki järjestelmän palauttamasta tulossivusta, jossa tulosdokumentit esitetään tehostettuina kuvakkeina [Woo01].

Tutkijat testasivat tehostettujen kuvakkeiden käyttöä käyttötestillä, johon osallistui 18 henkilöä. Käyttötestiä varten tutkijat hakivat jokaista testikysymystä kohden sata ensimmäistä Google-hakukoneen palauttamaa sivua. Haetut sivut tallennettiin paikallisesti ja sivuista luotiin sekä pelkät että tehostetut kuvakkeet, jotka myös tallennettiin paikallisesti [Woo01].

Tutkimus osoitti odotetusti, että eri menetelmien tehokkuus riippui hyvin paljon haettavan tiedon aihealueesta. Testi kuitenkin osoitti, että tehostettujen kuvakkeiden käyttö oli keskimäärin tehokkaampaa kuin pelkkien kuvakkeiden tai tekstiyhteenvetojen käyttö. Useimmat tutkimushenkilöt havaitsivat, että tehostettujen kuvakkeiden käyttö oli intuitiivisempaa kuin pelkkien kuvakkeiden tai tekstiyhteenvetojen käyttö. Niiden käyttämiseen tarvitsi myös vähemmän työtä. Puolet tutkimushenkilöistä valitsi tehostetut kuvakkeet parhaimmaksi yhteenvetomenetelmäksi, muista osanottajista useimmat



Kuva 6. Esimerkki Woodruffin ja kumppaneiden järjestelmästä, jossa tulosdokumentit esitetään tehostettuina kuvakkeina.

ajattelivat tehostettujen kuvakkeiden sopivan tietyn tyyppisiin tilanteisiin [Woo01].

Tehostettujen kuvakkeiden toteuttaminen kaupalliseen hakukoneeseen oli tutkijoiden mielestä mielenkiintoinen ongelma. Toteuttaminen tuo esille monia huomioon otettavia seikkoja, kuten haun tuloskuvakkeiden lataamisesta syntyvä kaistan tarpeen sekä kuvakkeiden generoimiseen tarvittava ajantarpeen [Woo01].

4 Yhteenveto

Hakukoneella suoritettavan haun tulokset voidaan esittää listamuotoisesti tai esittämällä dokumentit ryhmiteltyinä joukkoina. Perinteinen, listamuotoinen menetelmä on tyypillisesti pitkä lista hakukoneen palauttamia tulosdokumentteja, jossa eri aiheita käsittelevät www-sivut ovat usein sekoittuneet keskenään. Listamuotoisen tuloslistan ongelmia voidaan välttää ryhmittelemällä hakukoneen palauttamat tulosdokumentit. Tulosdokumentit voidaan ryhmitellä kolmella tavalla.

Ensimmäinen tapa on dokumenttien ryhmittely rakenteellisen tiedon perusteella. Rakenteellinen tieto on jokaisesta dokumentista saatavissa olevaa tietoa. Tämä tieto voi olla esimerkiksi dokumentin sijainti sivustolla tai dokumenttiin viittaavien linkkien määrä. Esimerkkinä tämän tavan järjestelmästä esiteltiin Cha-Cha.

Toinen tapa on ryvästys. Ryvästysalgoritmit pyrkivät ryhmittelemään dokumentteja niiden keskinäisen samankaltaisuuden perusteella. Ryvästyksen ongelmia ovat tyypillisesti se, että ryppäiden muodostamiseen kuluu paljon aikaa, sekä muodostettujen ryppäiden vaikea kuvaaminen. Ryvästystä käyttäviä järjestelmiä esiteltiin Grouper ja NIRVE.

Kolmas vaihtoehto on tulosdokumenttien luokittelu. Luokittelu tapahtuu mallilla, jota opetetaan ensin joukolla luokiteltuja dokumentteja. Opetuksen jälkeen mallia käytetään uusien dokumenttien luokittelemiseksi. Luokittelua käyttävänä järjestelmänä esiteltiin SWISH.

Vaihtoehtona perinteiselle listamuotoiselle menetelmälle esiteltiin tehostetut kuvakkeet, jossa tulosdokumentit esitetään niitä mallintavina kuvakkeina. Hakutulosten esittäminen kuvakkeina on tehokasta, sillä ihminen pystyy hahmottamaan kuvan suunnilleen samassa ajassa kuin pystyy lukemaan yhden sanan [Woo01].

Graafisen tiedon käsitteleminen on yleisesti huomattavasti tehokkaampaa kuin tekstimuotoisen tiedon käsitteleminen. NIRVEn kolmiulotteinen malli hyödyntää erittäin tehokkaasti tietojen esittämistä visuaalisessa muodossa. NIRVEn kaltaisen järjestelmän rajoittavaksi tekijäksi muodostuu sen asettamat laitteistovaatimukset.

Toisaalta SWISHin kaltainen tulosdokumenttien luokittelu ja esittäminen saattaa olla www-sivujen näkökulmasta tehokkain tapa hakutulosten esittämiseksi. Tehokkaaksi menetelmän tekee sen kyky luokitella tulosdokumentit selväkielisiin, hierarkisiin luokkiin. Dokumenttien luokittelutarkkuus on menetelmän heikoin kohta. Kun tulosdokumentit pystytään luokittelemaan riittävän tarkasti niin tällä menetelmällä on monia etuja muihin menetelmiin verrattuna.

Lähteet

ChD00 Chen, H., Dumais, S., Bringing Order to the Web: Automatically Categorizing Search Results. Proc. of the CHI 2000 conference on Human factors in computing systems 2000, Hague, Netherlands, April 1-6, 2000, 145-152.

Che99 Chen, M. et al., Cha-Cha: A System for Organizing Intranet Search Results. Proc. of the 2nd USENIX Symposium on Internet Technologies & Systems (USITS), Boulder, Colorado, USA, October 11-14, 1999.

DCC01 Dumais, S., Cutrell, E., Chen, H., Optimizing Search by Showing Results in Context. Proc. of the SIGCHI conference on Human factors in computing systems 2001, Seattle, Washington, USA, 31 March-5 April, 2001, 277-283.

Seb99 Sebrechts, M. et al., Visualization of Search Results: A Comparative Evaluation of Text, 2D, and 3D Interfaces. Proc. of the SIGIR conference on Research and Development of Information Retrieval 1999, Berkeley, California, USA, August 15-19, 1999, 3-10.

Woo01 Woodruff, A. et al., Using Thumbnails to Search the Web. Proc. of the SIGCHI conference on Human factors in computing systems 2001, Seattle, Washington, USA, 31 March-5 April, 2001, 198-205.

ZaE99 Zamir, O., Etzioni, O., Grouper: A Dynamic Clustering Interface to Web Search Results. Proc. of the Eighth International World Wide Web Conference (WWW8), Toronto, Canada, May 11-14, 1999.