Hyperteksti tieteellisessä julkaisutoiminnassa

Mika Tähtinen

Helsinki 10.12.2004
HELSINGIN YLIOPISTO
Hypermediajärjestelmät-seminaari
Tietojenkäsittelytieteen laitos

HELSINGIN YLIOPISTO – HELSINGFORS UNIVERSITET – UNIVERSITY OF HELSINKI

Tiedekunta/Osasto – Fakultet/Sektion – Faculty/Section	Laitos – Institution – Department	
Matemaattis-luonnontieteellinen	Tietojenkäsittelytieteen laitos	
Tekijä – Författare – Author		
Mika Tähtinen		
Työn nimi – Arbetets titel – Title		
Hyperteksti tieteellisessä julkaisutoiminnassa		
Oppiaine – Läroämne – Subject		
Tietojenkäsittelytiede		
Työn laji – Arbetets art – Level	Aika – Datum – Month and year	Sivumäärä – Sidoantal – Number of pages
Seminaariesitelmä	10.12.2004	10 sivua + 0 liitesivua

Tiivistelmä – Referat – Abstract

Tieteen kehittyessä tieteellisten julkaisujen määrä kasvaa ja samalla dokumenttien löytäminen ja arvioiminen vaikeutuu. Eksaktien tieteiden kehityksen myötä myös tieteellisistä dokumenteista on tullut hyvin määriteltyjä, joten niiden automaattinen käsittely ja yhdistäminen hypertekstiin on luontevaa. ResearhIndex on osoittautunut automaationsa ansiosta toimivaksi ratkaisuksi ja sen CiteSeer-prototyyppi on noussut suureen suosioon erityisesti tietojenkäsittelytieteessä.

Avainsanat – Nyckelord – Keywords

ResearchIndex, CiteSeer

Säilytyspaikka – Förvaringställe – Where deposited

Muita tietoja – Övriga uppgifter – Additional information

Hypermediajärjestelmät–seminaari, syksy 2004

Sisältö

1 Johdanto

Tieteellisten julkaisujen suuri määrä on jo vuosisatoja tuottanut ongelmia tieteenharjoittajille tiedonhaussa. Samalla kun tiede kehittyy, syntyy uusia dokumentteja ja tutkimustuloksia. Uusien artikkelien julkaisemisessa paperimuotoon on aina pieni viive, jonka takia tieteenharjoittajilla ei aina ole uusinta tietoa käytettävissään. Toisaalta tiedejulkaisuja on niin paljon, että kaikkien lukeminen on yksittäiselle tutkijalle mahdotonta.

Frédérique Harmsze arvioi vuonna 2000, että tieteenharjoittajat lukevat noin 40% alansa relevantista kirjallisuudesta. Tiedejulkaisujen määrä kaksinkertaistuu noin 10-15 vuodessa ja yhdestä julkaisusta löytyvien dokumenttien määrä ja pituus kasvaa [Har00]. Dokumenttien suuri määrä on saanut aikaan sen, että harva tiedekirjasto pystyy nykyään säilyttämään kaikkia tiedejulkaisuja paperimuodossa.

Eksaktien tieteiden kehityksen myötä myös tieteellisistä dokumenteista on tullut eksakteja ja formaalisti muotoiltuja. Tieteelliset dokumentit tarjoavat siis paljon mahdollisuuksia hypertekstin käyttöön. Nykyään kaikki julkaistu teksti tuotetaan tietokoneita käyttäen, joten elektroninen tallennus ja levitys on looginen ja helppo vaihtoehto paperijulkaisujen rinnalle.

Internet tarjoaa tieteenharjoittajille keinot tutkimustulosten julkaisuun ja nopeaan kommunikointiin. Lisäksi internetin kautta voidaan helposti vaihtaa monenlaista tutkimusdataa. Kuten kaikessa internetin sisällössä ongelmana ei kuitenkaan ole tiedon saanti vaan sen ylitarjonta. Internetiin voi kuka tahansa julkaista omia tutkimustuloksiaan ja dokumenttejaan, mutta korrektien julkaisujen löytäminen voi olla vaikeaa. Tarvitaan siis toimivia ratkaisuja tieteellisten dokumenttien julkaisuun ja arviointiin.

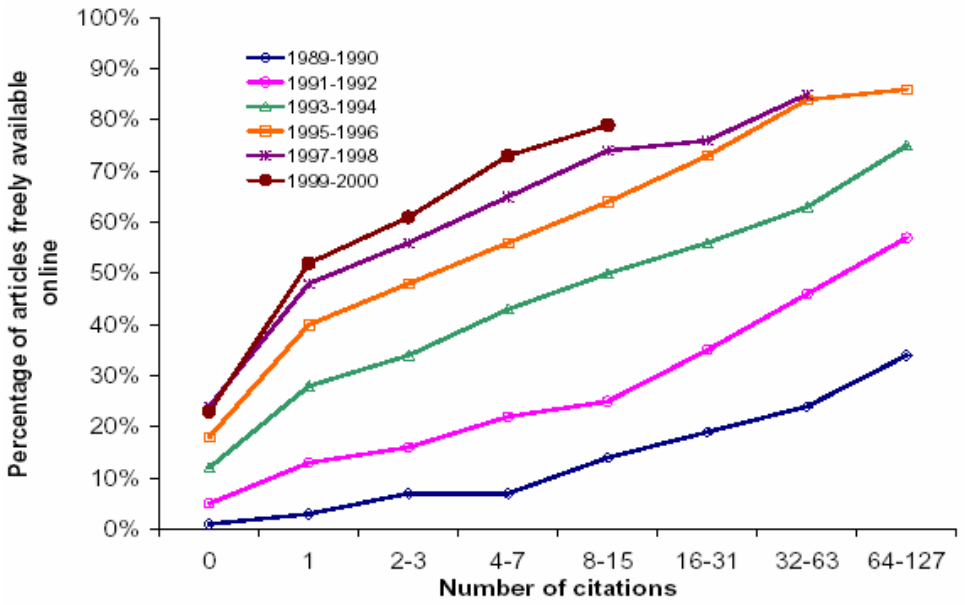
Tässä työssä esitellään tieteellisten dokumenttien julkaisemiseen suunniteltu ResearchIndex-järjestelmä. ResearchIndex on muutaman vuoden olemassaolonsa aikana vakiinnuttanut paikkansa tiedeyhteisössä syrjäyttäen edeltäjänsä suosiollaan ja tehokkuudellaan. Hypertekstin hyväksikäyttöön tieteellisessä julkaisutoiminnassa on ehdotettu muitakin ratkaisuja, kuten Harmszen esittämä modulaarinen rakenne [Har00]. ResearchIndexin etu

muihin ratkaisuihin nähden on kuitenkin sen automaattisuus, joka ei vaadi dokumenttien kirjoittajilta mitään erityisiä toimenpiteitä ja merkintätapoja.

2 ResearchIndex

ResearchIndex-järjestelmä on esittelynsä jälkeen tullut tieteenharjoittajille tutuksi CiteSeer-prototyyppinsä[1] avulla. Järjestelmän toimintaperiaatteena on indeksoida ja esittää tieteellisiä dokumentteja niiden tekemien viittauksien perusteella.

Artikkelissaan ”Online or invisible?”[2] [Law01] Steve Lawrence, yksi ResearchIndex-järjestelmän kehittäjistä, esittää näkemyksiään ja tutkimustuloksiaan siitä, miten tieteellisten dokumenttien saatavuus internetissä suhtautuu dokumenttiin tehtyjen viittauksien määrään. Tutkimuksen tuloksena on kuvassa 1 esitetty kuvaaja.



Kuva 1: Artikkeleihin tehtyjen viittausten määrä suhteessa todennäköisyyteen, että kyseinen artikkeli on saatavilla internetistä [Law00]. Tutkimuksessa käytettiin 119 924 dokumenttia CiteSeer-järjestelmästä. Tutkituista viittauksista on poistettu kirjoittajan omiin töihinsä tekemät viittaukset.

---

[1] CiteSeer.IST Scientific Literature Digital Library http://citeseer.ist.psu.edu/, 1.12.2004.

[2] Artikkelin lopullinen nimi Nature-lehdessä on ”Free online availability substantially increases a paper’s impact”.

Kuvasta 1 nähdään, että dokumentin vapaa saatavuus internetissä nostaa dokumenttiin tehtyjä viittauksia. Vastaavasti dokumentit, joihin viitataan paljon, ovat todennäköisemmin saatavilla internetistä kuin dokumentit, joihin on vähän viittauksia. Dokumentin laatu ei ole siis ainoa tekijä, joka vaikuttaa dokumenttiin tehtyjen viittausten määrään. ResearchIndexin käyttämä arvosteluperuste on kuitenkin nykyaikana pätevä esimerkiksi tietojenkäsittelytieteen julkaisuissa, koska suurin osa alan merkittävistä julkaisuista löytyy internetistä.

2.1 Tieteellisten julkaisujen löytäminen

Ensimmäinen askel tieteellisten julkaisujen indeksoimisessa on dokumenttien löytäminen. ResearchIndexissä myös tämä vaihe on osittain automatisoitu. Järjestelmä etsii tieteellisiä dokumentteja internetistä yleisillä hakukoneilla. Hakusanoina käytetään tyypillisiä tieteellisiin julkaisuihin viittaavia avainsanoja, kuten ”PostScript”, ”PDF”, ”technical report”, ”conference”, ”proceedings”, ”publications” ja ”papers” [LGB99a, LGB99b].

ResearchIndex valvoo myös tutkijoiden käyttämiä sähköpostilistoja ja uutisryhmiä sekä käyttää suoria yhteyksiä julkaisujen kustantajien tietojärjestelmiin. Yksittäisillä tutkijoilla on myös mahdollisuus itse ilmoittaa järjestelmälle uusista julkaisuista. Järjestelmässä on indeksoitu myös maksullisten julkaisusivustojen artikkeleita.

2.2 Dokumenttien indeksointi

Kun indeksoitavat dokumentit on paikallistettu, alkaa niiden käsittely. PDF ja PostScript –muotoiset tiedostot muunnetaan tekstitiedostoiksi. ResearchIndexiä esittelevässä artikkelissaan [LGB99a] Lawrence et al. toteaa muunnoksen tapahtuvan New Zealand Digital Library –projektin tuottamalla PreScript[3] –ohjelmalla. CiteSeer-prototyyppiä tarkemmin käsittelevässä artikkelissa [LGB99b] kuitenkin viitataan Digital Virtual Paper –projektin pstotext[4]-ohjelmaan.

---

[3] New Zealand Digital Library, PreScript, http://www.nzdl.org/html/prescript.html, 1.12.2004.

[4] Virtual Paper, pstotext http://www.research.compaq.com/SRC/virtualpaper/pstotext.html, 1.12.2004.

Käsittelyn aluksi tarkistetaan, että käsiteltävä dokumentti on tieteellinen dokumentti etsimällä viittaus-osiota. Viittausosio voidaan tunnistaa joko suoraan osion otsikosta tai tunnistamalla itse viittauslista. Viittausosion käsittelystä kerrotaan tarkemmin seuraavassa aliluvussa.

Järjestelmään lisättäville dokumenteille luodaan täysi teksti-indeksointi. ResearchIndexin teksti-indeksi on suunniteltu jatkuvaa päivitystä varten, koska järjestelmä pyritään pitämään mahdollisimman hyvin ajan tasalla. Dokumenttien sanoista pidetään yllä tyypillistä hajautustaulua. Hajautustaulun solujen sisältönä on käänteisindeksit kokoelmiin niistä dokumenteista, joissa kyseisen solun esittämä sana esiintyy. Näihin kokoelmiin on lisäksi talletettu paikat, missä kyseinen sana esiintyy kussakin dokumentissa. Mikäli mahdollista, dokumenteista talletetaan myös erillisiä osia, kuten tiivistelmä, avainsanat ja kyseiseen dokumenttiin läheisesti liittyvät dokumentit (related work).

Monista tieteellisistä artikkeleista löytyy useita eri versioita. Esimerkiksi samasta dokumentista voi löytyä tekninen julkaisu, konferenssijulkaisu ja jotain tiedelehteä varten tehty lyhennetty versio. ResearchIndex tunnistaa ja yhdistää saman artikkelin eri versiot tutkimalla lauseiden esiintymisiä saman kirjoittajan dokumenteissa.

2.3 Viitteiden indeksointi

ResearchIndex käyttää ACI-järjestelmää (Autonomous Citation Indexing) dokumenttien viittausten indeksointiin. Jokaisesta käsiteltävästä dokumentista käsitellään sen tekemät viittaukset ja linkataan käsiteltävä dokumentti viitattuihin dokumentteihin. Viittausindeksien avulla tieteellisiä dokumentteja voidaan selata aikaulottuvuudessa. Dokumentin tekemät viittaukset osoittavat aina taaksepäin ajassa jo aiemmin kirjoitettuihin dokumentteihin. Eteenpäin ajassa taas liikutaan dokumenttiin myöhemmin tehtyjen viittausten kautta.

ACI-järjestelmän vahva puoli on sen automaattisuus. Aiemmat viittausindeksointijärjestelmät vaativat työtä artikkelien kirjoittajilta tai indeksien ylläpitäjiltä [LGB99a]. ResearchIndex indeksoi dokumentteja automaattisesti vaatimatta minkään tietyn standardin tai merkkauksen noudattamista. Dokumentissa täytyy kuitenkin olla rakennetta sen verran, että viittauksien jäsentäminen on ylipäätänsä mahdollista. Automaattinen viittausindek-

sointi parantaa myös järjestelmän tehokkuutta siten, että jo järjestelmässä olevia dokumentteja voidaan helposti indeksoida uudestaan, kun niihin tehdään muutoksia. Näin saadaan aikaan paremmin ajan tasalla oleva viittausindeksi.

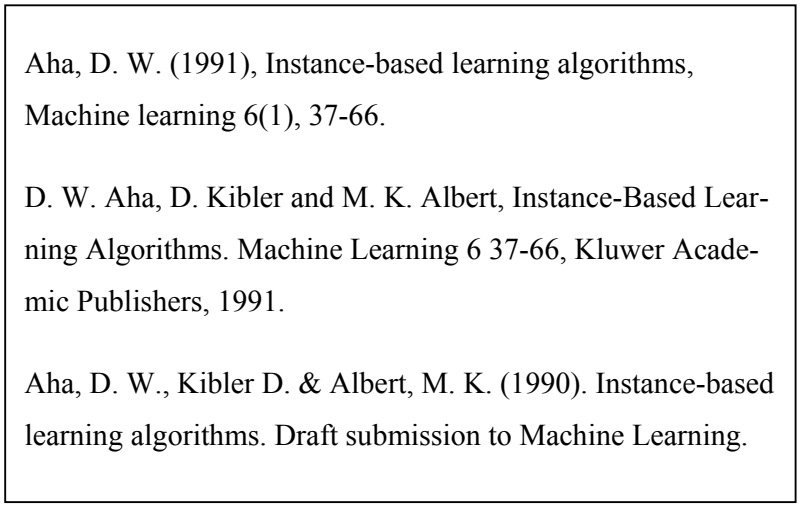
Käsiteltävän dokumentin viittauslista erotellaan yksittäisiksi viittauksiksi viittauksien tunnisteiden, viittauksien välien ja sisennyksien perusteella. Kuvassa 2 esitetyn esimerkin viittauksissa kukin viittaus alkaa viittaustunnisteella, itse viittauksen sisältö on sisennetty ja viittauksien välissä on kappaleenvaihto.

Law01	Lawrence, S., Online or Invisible? Nature, Volume 411, Number 6837, p. 521, 2001. http://www.neci.nec.com/~lawrence/papers/online-nature01/, 1.12.2004.
LBG99a	Lawrence, S., Bollacker, K., Giles, C.L., Digital Libraries and Autonomous Citation Indexing. IEEE Computer, Volume 32, Number 6, pages 67-71, 1999. http://citeseer.nj.nec.com/lawrence99digital.html, 1.12.2004.

Kuva 2: Esimerkki viittauslistan jäsentämisestä. Rivi alkaa viittaustunnisteella, viittauksen sisältö on sisennetty ja viittaukset on eroteltu toisistaan kappaleenvaihdolla.

Yksittäinen viittaus jäsennellään heuristiikoiden avulla eri kenttiin, kuten viittaustunniste, dokumentin nimi, kirjoittaja, julkaisuvuosi ja sivunumerot. Tässä käytetään metodia, jossa invariantit selvitetään ensin. Ensimmäiseksi haetaan siis viittauksen kentät, joilla on yleensä yhdenmuotoinen syntaksi. Esimerkiksi viittaustunnisteet ovat yleensä rivin alussa ja kirjoittajien nimet ennen artikkelin nimeä. Viittaustunnisteet myös muotoutuvat usein kirjoittajien nimien alkuosista. Apuna kenttien tunnistamisessa käytetään tietokantaa tutkijoiden ja dokumenttien nimistä. Kun invariantit on tunnistettu, voidaan saatua tietoa käyttää hyväksi muita kenttiä tutkittaessa. Lisäksi käyttämällä säännöllisiä lausekkeita voidaan tunnistaa myös erilaiset viittauskäytännöt, esimerkiksi onko viittaustunnisteeseen otettu kaikkien kirjoittajien vai vain yhden kirjoittajan nimen alkukirjaimet.

Samaan dokumenttiin tehtyjen viittauksien muodot vaihtelevat paljon. Kuvassa 3 on esimerkki koneoppimisjulkaisuista löytyneistä viittauksista yhteen artikkeliin. ResearchIndex pystyy yhdistämään kaikki esimerkissä annetut viittaukset samaan artikkeliin.



Kuva 3: Esimerkki samaan artikkeliin tehtyjen lainauksien variaatioista. [LGB99a].

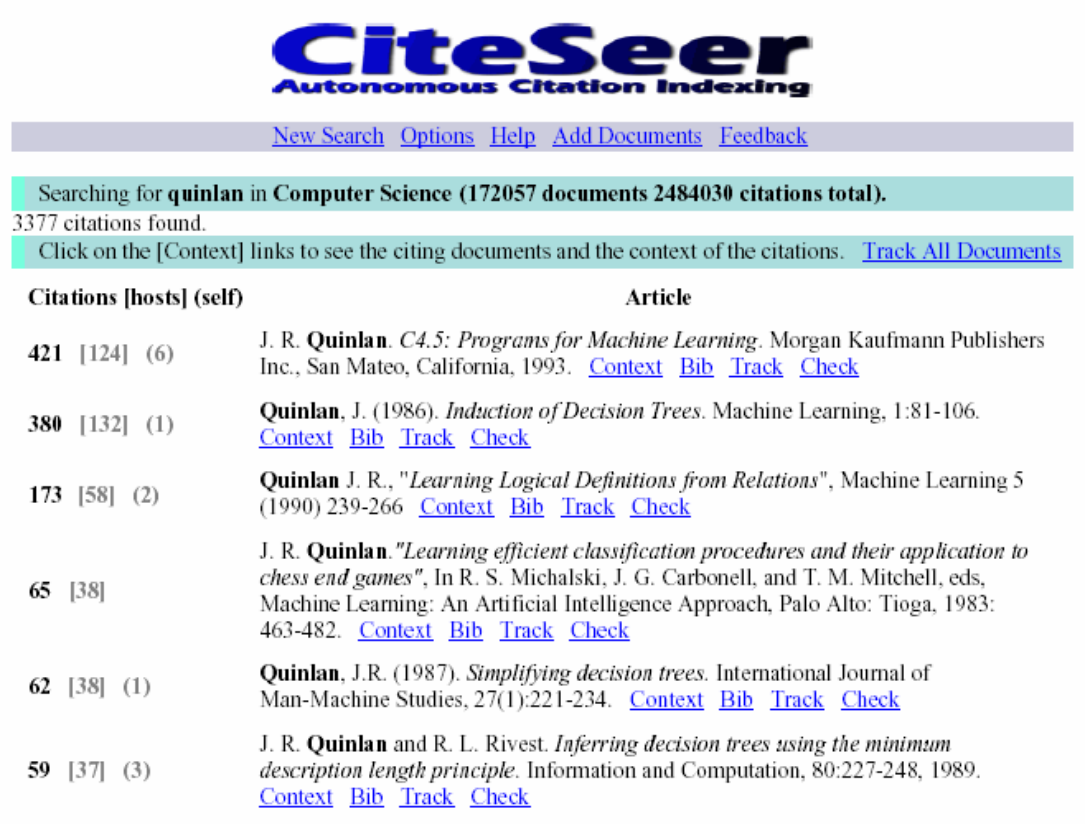
Kun viittaus tiettyyn dokumenttiin on käsitelty, ResearchIndex hakee tekstistä kontekste-ja missä viittauksia on tehty. Säännöllisten lausekkeiden avulla tekstistä poimitaan viittauksien asiayhteydet ja ne talletetaan viittauksien mukana järjestelmän tietokantaan.

2.4 Dokumenttien selaaminen

ResearchIndex sisältää oman hakukoneensa, jonka avulla voidaan hakea dokumentteja sekä niiden nimien ja kirjoittajien että viittauksien perusteella. Hakutulokset järjestetään artikkeleihin tehtyjen viittauksien määrän mukaan. Haun tehtyään käyttäjä voi jatkaa artikkeleiden selaamista viittauslinkkien mukaan kyseisen artikkelin tekemien viittausten tai kyseiseen artikkeliin tehtyjen viittauksien suuntaan.

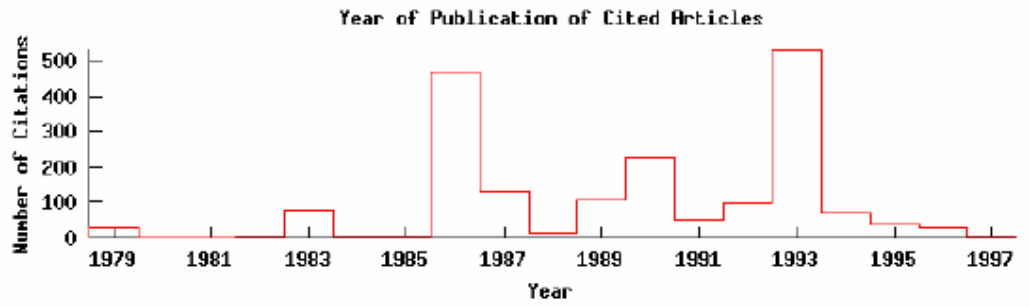
Kuvassa 4 on CiteSeeristä saatu hakutulos hakusanalle ”quinlan”. Kunkin artikkelin edessä on siihen tehtyjen viittauksien määrä. Hakasuluissa ilmaistu ”hosts”-arvo kertoo kuinka monelta eri web-sivustolta löydetyistä dokumenteista on viitattu kyseiseen artikkeliin. Lisäksi suluissa oleva ”self”-arvo ilmaisee artikkeliin kohdistuvat niin sanotut

itseviittaukset, jotka ovat artikkelin kirjoittajan muissa dokumenteissa olevia viittauksia, jotka eivät kuulu viittauksien kokonaismäärään.



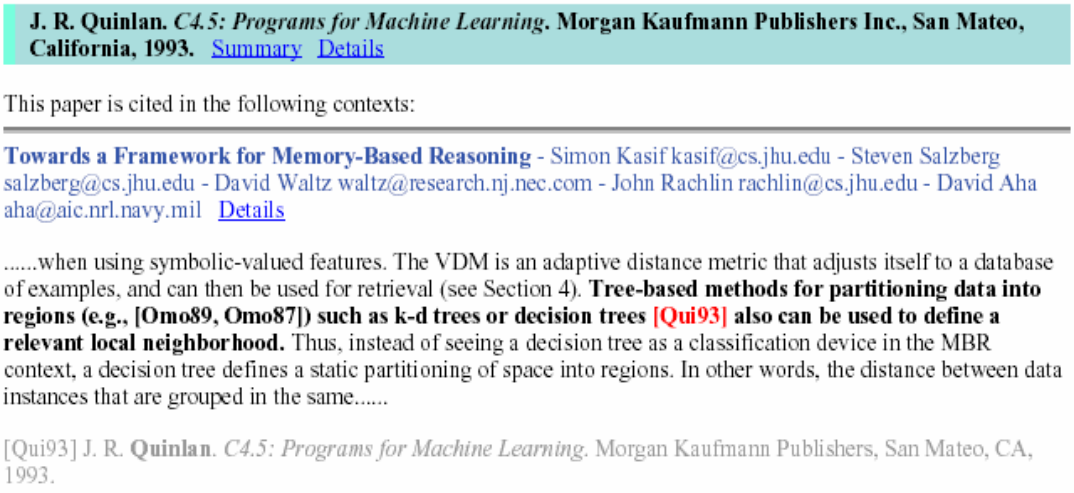
Kuva 4: CiteSeeristä saatu hakutulos hakusanalla ”quinlan” [LGB99a].

ResearchIndex esittää hakujen yhteydessä myös viittaavien artikkelien julkaisuvuodet kuvaajana. Kuvaajasta voidaan päätellä tietyn artikkelin ajantasaisuus; mikäli artikkeliin ei ole tehty viittauksia kyselyn ajankohtaa edeltävinä vuosina, on artikkelin tieto mahdollisesti voinut vanhentua. Kuvassa 5 on esimerkki kuvan 4 hakutuloksen yhteydessä esitetystä kuvaajasta. Kuvaajasta voidaan päätellä, että J. R. Quinlan on julkaissut merkittäviä artikkeleita vuosien 1986 ja 1993 tienoilla, koska näiden vuosien jälkeen tehtyjen viittauksien määrä on korkea.



Kuva 5: Hakutuloksen artikkeleiden julkaisuvuosien jakauma [LGB99a].

Viitteiden yhteyksien selvittämistä helpotetaan näyttämällä viitteiden yhteydessä kontekstit, joissa viitteitä on tehty. Kuvassa 6 on J. R. Quinlanin kirjaan ”C4.5 Programs for Machine Learning” tehty viittaus. Kunkin viittaavan dokumentin tietojen yhteydessä on ote dokumentista, jossa viittausta on käytetty sekä viittauksen muoto kyseisessä dokumentissa. Näiden tietojen avulla käyttäjä löytää helposti muita kyseessä olevan artikkelin aiheeseen liittyviä artikkeleita sekä voi muodostaa niistä kontekstien perusteella helposti mielikuvan.


J. R. Quinlan. ***C4.5: Programs for Machine Learning*****. Morgan Kaufmann Publishers Inc., San Mateo, California, 1993.** Summary Details

This paper is cited in the following contexts:

Towards a Framework for Memory-Based Reasoning - Simon Kasif kasif@cs.jhu.edu - Steven Salzberg salzberg@cs.jhu.edu - David Waltz waltz@research.nj.nec.com - John Rachlin rachlin@cs.jhu.edu - David Aha aha@aic.nrl.navy.mil Details

.....when using symbolic-valued features. The VDM is an adaptive distance metric that adjusts itself to a database of examples, and can then be used for retrieval (see Section 4). **Tree-based methods for partitioning data into regions (e.g., [Omo89, Omo87]) such as k-d trees or decision trees [Qui93] also can be used to define a relevant local neighborhood.** Thus, instead of seeing a decision tree as a classification device in the MBR context, a decision tree defines a static partitioning of space into regions. In other words, the distance between data instances that are grouped in the same.....

[Qui93] J. R. **Quinlan**. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA, 1993.

Kuva 6: Viittaushaun tuloksissa näytetään viittaavien artikkelien tietojen lisäksi konteksti, missä viittaus on tehty [LGB99a].

Täyden teksti-indeksoinnin ansiosta järjestelmään indeksoituihin dokumentteihin voidaan kohdistaa myös sisältöhakuja. ResearchIndex tukee Boolean-hakuja, lausehakuja (phrase search) sekä sanojen lähekkäisyyteen perustuvia hakuja (proximity search). Poiketen normaaleista hakukoneista ResearchIndex ei hylkää yleisiä sanoja kuten ”the”, koska tieteellisiä julkaisuja haettaessa halutaan usein hakea artikkelin tarkalla nimellä tai esi-

merkiksi kirjoittajan nimikirjaimien perusteella. Sisältöhakuja tehdessä ResearchIndex antaa hakutuloksena hakuun sopivien dokumenttien otsikot sekä kontekstit, joissa hakusanat esiintyivät.

3 Yhteenveto

ResearchIndex on CiteSeer-prototyypillään osoittautunut muutaman vuoden olemassaolonsa aikana toimivaksi järjestelmäksi ja on vakiinnuttanut paikkansa erityisesti tietojenkäsittelytieteen alalla. Tieteellisen julkaisun löytyminen CiteSeeristä viitteineen on nykyään jo jonkinlainen statusarvo artikkeleille.

Viittausindeksointiin perustuva järjestelmä helpottaa tiedeyhteisön kommunikointia antamalla huomiota artikkeleihin tehtyihin korjauksiin ja artikkeleihin kohdistuvaan kritiikkiin. Lisäksi viittauksien perusteella voidaan analysoida tieteen tutkimustrendejä ja tunnistaa tieteen uusien alojen syntymistä ja kehittymistä.

Muitakin hypertekstiin pohjaavia järjestelmiä on ehdotettu ja kehitetty sekä ennen CiteSeerin käyttöönottoa että sen jälkeen. Näyttää kuitenkin siltä, että tehokkaasti automatisoitu ja suuren suosion saanut ResearchIndex-järjestelmä tulee tulevaisuudessakin pitämään paikkansa tiedemaailmassa.

Lähteet

Har00 Harmsze, F., A Modular Structure for Scientific Articles in an Electronic Environment. Ph. D. thesis, University of Amsterdam, 2000. http://www.science.uva.nl/projects/commphys/papers/thesisfh/pdf/pdfversion.html, 1.12.2004.

Law01 Lawrence, S., Online or Invisible? Nature, Volume 411, Number 6837, p. 521, 2001. http://www.neci.nec.com/~lawrence/papers/online-nature01/, 1.12.2004.

LBG99a Lawrence, S., Bollacker, K., Giles, C.L., Digital Libraries and Autonomous Citation Indexing. IEEE Computer, Volume 32, Number 6, pages 67-71, 1999. http://citeseer.nj.nec.com/lawrence99digital.html, 1.12.2004.

LBG99b Lawrence, S., Bollacker, K., Giles, C.L., Indexing and Retrieval of Scientific Literature. In Eighth International Conference on Information and Knowledge Management, CIKM 99, pages 139-146, 1999. http://citeseer.ist.psu.edu/lawrence99indexing.html, 1.12.2004.