Web-sivuston käyttäjien ryhmittely

Seminaariraportti

Sami Linnanvuo

Tietämyksen muodostamisen seminaari
Helsingin yliopisto
Tietojenkäsittelytieteen laitos

Sisältö

1 JOHDANTO 3

2 KÄYTTÄJÄRYHMIEN MUODOSTAMINEN 3

3 KLUSTEROINTIMENETELMIEN ESITTELY 4

3.1 AUTOCLASS 4
3.2 ITSEORGANISOITUVAT KARTAT 5
3.2 CLUSTER MINING 5

4 SYÖTEAINEISTO 7

5 TULOKSET 8

5.1 MALLIEN KOKO JA LUKUMÄÄRÄ 8
5.2 EROTTELEVUUS JA KATTAVUUS 9
5.3 LÖYDETYT KÄYTTÄJÄRYHMÄT 10

6 YHTEENVETO 11

LÄHTEET 12

1 Johdanto

Webin käyttäjien ryhmittelyllä eli klusteroinnilla tarkoitetaan implisiittisten kayttäjäryhmien löytämistä käyttöhistorian perusteella. Usein sivuston käytössä on löydettävissä käyttäytymismalleja, jotka kuvaavat tietyn käyttäjäryhmän käyttäytymistä sivustossa. Käyttäjäryhmien tunnistamisesta on hyötyä esimerkiksi sivuston personoinnissa, jolloin sivuston sisältö voidaan esittää eri käyttäjäryhmille juuri heidän kiinnostusprofiilinsa mukaisesti. Löydettyjä käyttäjäryhmiä voidaan hyödyntää myös markkinoinnissa tai sivuston rakenteen analysoinnissa, jolloin sivuston rakenne ja navigointi pyritään sovittamaan käyttäjäryhmien tarpeiden mukaiseksi.

Lähestymistapoja klusterointiin on useita ja eri sovelluskohteet vaativat eri klusterointimenetelmiä. Tässä raportissä esitellään kolme klusterointialgoritmia, jotka soveltuvat käyttäjien ryhmittelemiseen. Käyttäjäryhmien muodostamista lähestytään kahdella eri tavalla: Sivupyyntöjä analysoitaessa etsitään sivujoukkoja, jotka kuvaavat käyttäjäryhmien kiinnostuksia. Sivusiirtymien analysoinnissa tutkitaan käyttäjien navigaatiomalleja sivustossa. Raportti perustuu Paliouras et al artikkeliin *Clustering the users of large web sites into communities* [Pal00].

2 Käyttäjäryhmien muodostaminen

Klusterointimenetelmää valittaessa on otettava huomioon kohdealueen vaatimukset. Webin käyttäjien ryhmittelyssä tärkeimmät kriteerit klusterointimenetelmälle on skaalautuvuus suurille datamäärille ja päällekkäisten klustereiden salliminen. Suuressa sivustossa voi olla satoja tuhansia käyttäjiä ja käytöstä kertynyt historiadata on helposti useita gigatavuja. Monet yleiset klusterointimenetelmät muodostavat ainoastaan erillisiä klustereita, mikä on turha rajoite webin käyttäjien ryhmittelyssä: yhden käyttäjän tulisi voida kuulua useaan ryhmään.

Käyttäjäryhmien löytämisen lisäksi tavoitteena on löytää käyttäytymismallit, jotka kuvaavat käyttäjiä. Mallien löytämiseksi tutkitaan löydettyjen ryhmien kuvaavia

ominaisuuksia. Käyttäjäryhmät kuvataan samoilla määreillä kuin data, josta ne on muodostettu. Jos data sisältää vain tietoja sivuista, joissa käyttäjät ovat käyneet, silloin myös käyttäjäryhmät kuvataan sivuina.

Kuvaavien ominaisuuksien valinta perustuu ajatukseen, että käyttäjäryhmää kuvaava ominaisuus on erityinen sille käyttäjäryhmälle, jos sen yleisyys ryhmän sisällä on merkittävästi suurempi kuin ryhmän ulkopuolella. Jotta ryhmiä kuvaavat mallit olisivat hyödyllisiä, niitä ei saa olla liikaa ja niiden oltava tarpeeksi yksinkertaisia. Kaikki löydetyt mallit eivät välttämättä ole mielenkiintoisia. Eri ryhmiä kuvaavien mallien on oltava mahdollisimman *erittteleviä*. Jos eri ryhmien mallit ovat keskenään hyvin samanlaisia, ei ryhmien muodostamisessa ole järkeä. Jos etsitään malleja, jotka ovat pieniä ja hyvin erittteleviä, päädytään helposti malleisin, jotka kattavat vain hyvin pienen osan käyttäjistä. Tämän ongelman välttämiseksi otetaan käyttöön myös ominaisuus *kattavuus*, joka kertoo kuinka suuren osan aineistosta malli kattaa. Valitut neljä ominaisuutta: mallien määrä, koko, erittelevyys ja kattavuus ovat riippumattomia tässä tutkimuksessa käytettyjen klusterointialgoritmien suorituskyvylle, joten niitä voidaan käyttää valittujen klusterointimenetelmien arvioinnissa.

3 Klusterointimenetelmien esittely

Tässä kappaleessa esitellään kolme klusterointimenetelmää webin käyttäjien ryhmittelemiseksi. *AutoClass* on teollisuudessa paljon käytetty Bayes-menetelmää käyttävä klusterointimenetelmä. *Itseorganisoituvat kartat* (self-organizing maps) muodostaa samanlaisia klustereita kuin AutoClass, mutta vaatii syötteenä datan lisäksi muodostettavien klustereiden lukumäärän. Kolmas, graafimenetelmään perustuva *cluster mining,* on varsin uusi menetelmä ja se esitellään muita tarkemmin.

3.1 AutoClass

AutoClass on käsitepohjainen klusterointimenetelmä (conceptual clustering) täydennettynä Bayes-luokittelijalla parhaimman klusterin paikantamiseksi. AutoClass

poikkeaa muista klusterointimenetelmistä siinä, että se ei määritä klustereille alkioita, vaan ainoastaan todennäköisyyden jolla alkio kuuluu kuhunkin klusteriin. Määrittäessään mallia klusterille, AutoClass käyttää käsitettä *vaikutus* (influence), joka kertoo kuinka tärkeä kukin attribuutti on klusterille. Klusterin malliksi valitaan attribuutit, joilla on suurin vaikutus kyseisessä klusterissa.

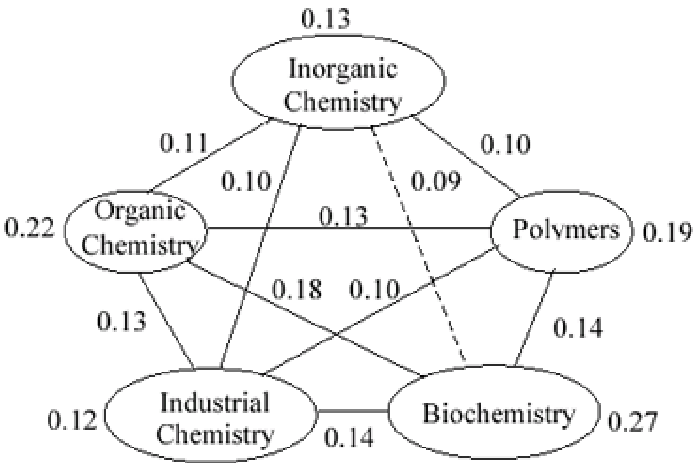
3.2 Itseorganisoituvat kartat

Itseorganisoituvat kartat (self-organizing maps, SOM), on yksi käytetyimmistä menetelmistä klusteroinnissa. SOM perustuu *k-means*-tyyppiseen menetelmään, jossa iteratiivisesti määritetään prototyyppivektorit kullekin klusterille eli kartan yksikölle. Kuten neuroverkkojen neuronit, myös SOM-kartan yksiköt kilpailevat syötealkioista ja mukautuvat jokaisen voittamansa alkion mukaan siten, että voittanut yksikkö muuttaa painojaan uuden alkion suuntaan. Samalla viereiset yksiköt muuttavat painojaan alkiosta poispäin, korostaen yksiköiden välistä rajaa.

Jokainen verkon yksikkö kuvaa yhtä klusteria ja klusterin malli saadaan, samoin kuin AutoClass-menetelmässä, valitsemalla attribuutit, joilla on suurin vaikutus kyseisessä klusterissa. Klustereiden lukumäärä on määritettävä kartalle etukäteen, mikä on SOM-menetelmän heikkous.

3.2 Cluster mining

Cluster mining-menetelmässä syöteaineistosta muodostetaan verkko, jonka solmuina ovat attribuutit ja jonka solmujen väliset kaaret kuvaavat attribuuttien välisiä esiintymiä syöteaineistossa. Esimerkiksi jos käyttäjä on käynyt sivuilla ”Organic Chemistry” ja ”Polymers”, lisätään kaari kyseisten solmujen välille. Solmujen ja kaarien painot lasketaan attribuuttien frekvenssien ja attribuuttien välisten yhteyksien perusteella. Kaaren paino normalisoidaan jakamalla se kaaren päissä olevista solmuista sillä, jolla on suurempi frekvenssi. Normalisoinnilla vähennetään sellaisten attribuuttien vaikutusta, jotka esiintyvät usein koko syöteaineistossa. Jotta verkon kaaria saataisiin vähennettyä, otetaan käyttöön *kynnysarvo* (connectivity threshold), jota pienemmän arvon omaavat kaaret poistetaan.



Kuva 1. Normalisoitu graafi.

Kun verkko on muodostettu, siitä voidaan muodostaa klustereita etsimällä aliverkkoja graafialgoritmilla. Sopivia menetelmiä ovat [PeE99]

- *Yhtenäisten komponenttien* (connected components) etsintä. Etsitään puusta aliverkot, joissa jokaisen solmuparin välissä on kaari. Yhtenäisten komponenttien etsintä voidaan suorittaa lineaarisessa ajassa.

- *Maksimaalisten klikkien* (maximal cliques) etsintä. Klikki on aliverkko, jossa jokaisen solmuparin välissä on kaari. Maksimaalinen klikki ei ole minkään toisen klikin alijoukko. Koska maksimaalisen klikin etsintä vie pahimmassa tapauksessa eksponentiaalisen ajan, etsittävien klikkien koolle on syytä asettaa yläraja.

Toisin kuin AutoClass -ja SOM-menetelmissä, klusterille ei tarvitse erikseen muodostaa mallia, vaan jokainen verkosta saatava klusteri kuvaa suoraan käyttäjien tyypillisiä ominaisuuksia, koska verkon solmut kuvaavat käyttäjäryhmän tyypillisiä attribuutteja.

4 Syöteaineisto

Tutkimuksen aineistona käytettiin palvelimen käyttölokia sivustosta ”Information Retrieval in Chemistry” (http://macedonia.chem.demokritos.gr), joka sisältää useita tuhansia sivuja ja suuren määrän aktiivisia käyttäjiä. Yksittäisiä hakuja oli kaiken kaikkiaan 137 150. Aineiston esiprosessoinnissa jaoteltiin ensin yksittäiset haut sessioihin ja sen jälkeen muodostettiin aineistosta attribuuttivektorit. Sessiot muodostettiin seuraavasti:

1. Järjestä loki päivämäärän ja IP-osoitteen mukaan.

2. Määritä aikaikkuna, jonka sisällä samasta IP-osoitteesta tulevat haut luetaan samaan sessioon kuuluvaksi.

3. Ryhmittele sivupyynnöt sessioksi etsimällä aikaikkunan sisällä olevat samasta IP-osoitteesta tulevat pyynnöt.

Sopivan aikaikkunan määrittämiseksi generoitiin frekvenssijakauma sivusiirtyminä minuuttien mukaan, jonka perusteella voitiin todeta, että siirtyminen sivulta toiselle yli tunnin viiveellä oli hyvin harvinaista, joten session aikaikkunaksi määritettiin yksi tunti. Näin laskettuna sessioita muodostui kaiken kaikkiaan 11 893 kappaletta.

Attribuuttivektoreiden muodostamiseksi kokeiltiin kahta eri lähestymistapaa. Ensimmäisessä vektorit muodostettiin siten, että jokainen attribuutti vastasi yhtä sivuston sivua. Toisessa lähestymistavassa attribuutit kuvasivat sivusiirtymiä session aikana. Kummassakin tavassa vektorit sisälsivät totuusarvon attribuutille sen mukaan, sisältyikö attribuutti sessioon. Attribuuttien vähentämiseksi joukosta poistettiin sivut, joiden esiintymisfrekvenssi aineistossa oli vähemmän kuin 30 ja sivusiirtymät, joiden frekvenssi oli vähemmän kuin 20. Lisäksi poistettiin sivusiirtymät samojen sivujen välillä. Karsinnan jälkeen vektoreihin jäi 229 sivua ja 251 sivusiirtymää.

Kun klusterointimenetelmiä sovellettiin aineistoon, ensimmäisessä lähestymistavassa saatiin tulokseksi joukko sivuja, jotka ovat suosittuja käyttäjäryhmän sisällä.

Tuloksista voidaan suoraan nähdä käyttäjien kiinnostukset. Klusteroimalla sivusiirtymiä saatiin käyttäjät ryhmiteltyä navigoinnillisten mallien mukaan, jotka kuvaavat tyypillisiä polkuja, joita käyttäjät sivustossa etenevät.

5 Tulokset

Kappaleessa 3 esiteltyjä klusterointimenetelmiä sovellettiin syöteaineistoon. Tuloksia arvioitiin neljällä kriteerillä: syntyneiden mallien määrällä, mallien keskimääräisellä koolla, erottelevuudella ja kattavuudella.

5.1 Mallien koko ja lukumäärä

AutoClass muodosti aineistosta automaattisesti 14 klusteria sivuvektoreille ja 4 klusteria sivusiirtymille. Analysoitaessa klustereiden malleja havaittiin, että sekä sivuvektoreilla että sivusiirtymävektoreilla on suuri määrä attribuutteja, joilla on pieni vaikutus. Näin ollen kynnysarvolla on suuri vaikutus mallien muodostumisessa. Kynnysarvolla 0,2 mallit sisälsivät keskimäärin vain hieman yli 10 attribuuttia, mikä vastaa 5% attribuuttien kokonaismäärästä.

SOM puolestaan vaatii klustereiden lukumäärän määrittämistä etukäteen. Muodestettavien klustereiden lukumääräksi asetettiin 16. Samoin kuin AutoClassin tapauksessa, myös SOM muodosti malleja, joissa attribuuttien määrä oli vähäinen. Kynnysarvolla 0,2, attribuutteja oli keskimäärin alle 5.

Cluster mining -algoritmissa graafin kytkösten kynnysarvolla on vaikutusta sekä muodostuvien klustereiden määrään että mallien kokoon. Jotta malleja olisi muodostunut kohtuullinen määrä (alle 40), kynnysarvo täytyi nostaa yli 0,2, mutta toisaalta yli 0,1 kynnysarvot muodostivat niin vähän attribuutteja, että voidaan puhua paremminkin sivupareista kuin kuin sivujoukoista. Kynnysarvolla 0,2 muodostui 76 mallia joissa attribuutteja oli keskimäärin vähän yli kaksi.

5.2 Erottelevuus ja kattavuus

Klusteroinnin hyödyllisyyteen vaikuttaa suoraan kattavuus, eli mallien tulisi kattaa riittävän suuri joukko käyttäjiä. Toisaalta mallien tulisi olla mahdollisimman erottelevia eli klustereiden tulisi olla mahdollisimman etäällä toisistaan. Kattavuus on yleensä riippuvainen mallien lukumäärästä ja koosta siten, että kattavuus on pieni silloin kun sekä mallien lukumäärä että koko ovat pieniä. Erottelevuutta ja kattavuutta arvioitiin tekemällä mittauksia eri vaikutus- ja kynnysarvoilla. Käyristä nähdään, että kaikilla testatuilla klusterointimenetelmillä on taipuvaisuutta muodostaa klustereita, jotka ovat hyvin erottelevia, mutta huonoja kattavuudeltaan. AutoClass suoriutuu parhaiten molemmissa testeissä.

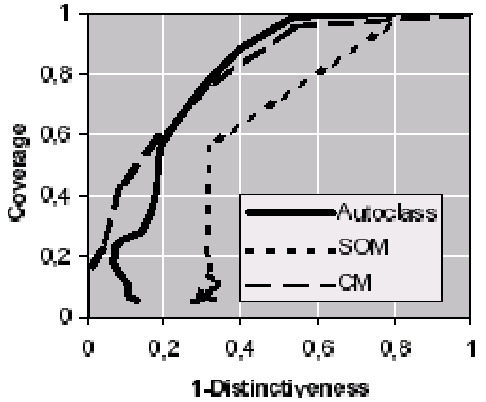


Figure 2. Trade-off curves for the page-based representation.

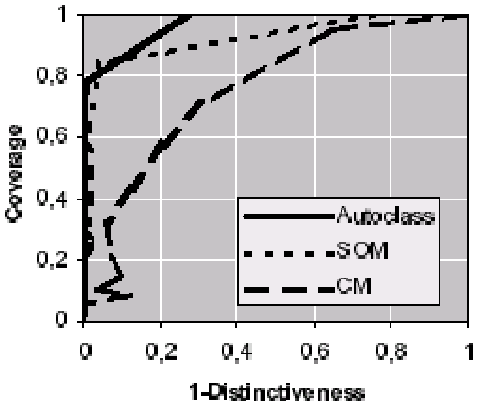


Figure 3. Trade-off curves for the transition-based representation.

Kuva 3. Erottelevuuden ja kattavuuden suhteet käyrinä

5.3 Löydetyt käyttäjäryhmät

Koska klustereiden halutaan olevan mahdollisimman kattavia, korkean kynnysarvon tuottamat klusterit hylätään. Toisaalta liian matala kynnysarvo kasvattaa huomattavasti klustereiden kokoa ja vähentää niiden erottelevuutta. Alla olevassa kuvassa on listattu valitut klusterit. Tulosten perusteella AutoClass näyttää muodostavan parhaita käyttäjäryhmiä ja SOM pärjää lähes yhtä hyvin sivusiirtymien klusteroinnissa (taulukon alaosa).

METHOD (REPRESENTATION)	THRESHOLD	MODELS	SIZE	COVERAGE	DIST/NESS
AUTOCLASS (P)	0.20	13	15.45	0.66	0.78
SOM (P)	0.05	16	11.75	0.56	0.68
CM (P)	0.20	76	2.18	0.59	0.81
AUTOCLASS (T)	0.05	4	49.25	0.78	1.00
SOM (T)	0.05	16	13.75	0.84	0.96
CM (T)	0.15	83	2.18	0.43	0.79

Kuva 3. Tulokset valituilla kynnysarvoilla

Sivuston ylläpitäjälle käyttäjäryhmät kertovat miten sivuston aiheiden kiinnostavuus on jakautunut eri käyttäjien kesken ja millaisia reittejä käyttäjät liikkuvat sivustossa. Jos löydetyt käyttäjäryhmät poikkeavat paljon odotetusta, se voi olla merkki sivuston rakenteen huonosta suunnittelusta. Alla olevassa taulukossa on kuvattu löydettyihin malleihin liittyvät mallit. Sivumallit on kuvattu listana yksittäisiä sivuja puolipisteellä eroteltuna. Sivusiirtymämallit on kuvattu nuolilla yhdistettynä sivuina.

METHOD (REP-RESENTATION)	COMMUNITY MODEL
AUTOCLASS (P)	Atmospheric-Chem Databases; Pharmaceutical-Chem Databases; Medicinal-Chem Databases; Chemistry Overview Databases; Hellas; Internet Overview.
SOM (P)	Computational Chemistry Journals; Chemistry Overview Databases; Thermochemistry.
CM (P)	Engineering; Environmental Sciences; Crystallography; Other Topics.
AUTOCLASS (T)	Chemistry ->related -> Internet; demok -> akno.
SOM (T)	Chemistry -> Internet -> WWW; Chemistry -> Biochemistry.
CM (T)	Internet -> Institute; Akno -> stats -> stats_all -> awards.

Kuva 4. Löydetyt käyttäjäryhmät

6 Yhteenveto

Tässä raportissa esiteltiin kolme klusterointimenetelmää, joilla avulla voidaan löytää mielenkiintoista tietoa web-sivuston käyttäjistä. Käyttäjistä muodostettiin käyttäjäryhmiä sivulatausten ja sivusiirtymien perusteella. Eri menetelmiä vertailtiin eri kynnysarvoilla ja syntyneet mallit arvioitiin. Testien perusteella kolmesta menetelmästä AutoClass suorioutui parhaiten tässä testissä, mutta menetelmiä tulisi testata myös muissa sivustoissa ennen kuin menetelmien paremmuutta voi arvioida pidemmälle.

Lähteet

Pal00 Paliouras, G., Papatheodorou, C., Karkaletsis, V., Spyropoulos, C. Clustering the users of the large web sites into communities. *Proceedings of the 23rd ACM SIGIR Conference on R&D in IR (SIGIR)*, 128-135, Athens, Greece, 2000.

PeE99 Perkowitz, M., Etzioni, O. Towards adaptive web sites: Conceptual framework and case study. *Proceedings of Eight WWW Conference*, 1999.