Informaatiotutkimuksen päivät 2010
21. - 22. lokakuuta, Tampere

ABSTRAKTI

Työkalu ontologioiden editointiin ja ontologiapohjaiseen tiedonhakuun

Feza Baskaya
Department of Information Studies and Interactive Media
University of Tampere, Finland
Feza.Baskaya@uta.fi

Jaana Kekäläinen
Department of Information Studies and Interactive Media
University of Tampere, Finland
Jaana.Kekalainen@uta.fi

Kalervo Järvelin
Department of Information Studies and Interactive Media
University of Tampere, Finland
Kalervo.Jarvelin@uta.fi

TIIVISTELMÄ

Perinteisten tiedonhakujärjestelmien käyttäjien ongelmia ovat sanaston yhteensopimattomuus ja sumeat hakutavoitteet. Tämä johtuu siitä, että hakukäsitteet voidaan ilmaista teksteissä niin monin tavoin. Sanaston yhteensopimattomuutta voidaan lievittää laajennetulla haulla ja asiakirjojen merkitsemisellä sanastojen, tunnisteiden tai ontologioiden avulla. Toisaalta käyttäjille olisi hyötyä ontologiapohjaisesta annotoinnista, toisaalta käyttäjien tulee analysoida hakutuloksia ja muotoilla hakunsa uudelleen päästäkseen eroon epämääräisistä hakutavoitteista. Tässä artikkelissa ehdotamme ontologioita hakuun ja annotointiin ja kuvailemme työkalun, joka tukee ontologioiden muokkausta online-tilassa ja ontologiapohjaista monikielistä dokumenttien hakua ja analyysia. Se tarjoaa keinon tiedon etsimiseen useista lähteistä (samanaikaisesti) ja dokumenttien analyysiin, mm. ontologiapohjaisen tiivistämisen, annotoinnin ja klusteroinnin avulla.

Avainsanat
Ontologiat, tiivistäminen, luokittelu, käyttöliittymä

1. JOHDANTO

Moderneissa verkkoympäristöissä on miljardeja haettavissa olevia dokumentteja. Tekstidokumentit kirjoitetaan ja niitä haetaan monilla eri kielillä; ei-tekstimuotoiset dokumentit (kuten kuvat tai videot) voidaan hakea vain annotaatioiden avulla. Koska dokumenttien tuotanto ja julkaisu on hajautettua, ei ole yleisesti sovittuja sääntöjä niiden ulkoasusta tai sisällön esittämisestä. Näin ollen käyttäjä kohtaa tutut ongelmat, eli sanaston yhteensopimattomuuden, mikä tarkoittaa vaikeutta arvata niiden dokumenttien tarkka tekstisisältö, joita hän haluaa saada käyttöönsä (access), ja sumeita hakutavoitteita. Jälkimmäinen ongelma liittyy tiedonhakuun, koska tiedontarpeet ovat dynaamisia, kun

käyttäjät oppivat enemmän ongelmistaan. On myös ilmeistä, että iedontarpeiden ilmaisemiseen yhdellä tai useammalla (vieraalla) liittyy ongelmia.
Tässä artikkelissa ehdotamme ontologioita haku- ja analyysivälineiksi. Niiden avulla voimme hakea ja analysoida dokumenttikokoelmia, joiden dokumenteissa saattaa olla, muttei välttämättä ole, annotointeja. Pyrimme korjaamaan tiedonhakijan kohtaamat ongelmat seuraavasti: (1) Tiedon hakija käyttää omaa ontologiaansa tai olemassa olevaa ontologiaa, joka on räätälöity hänen tarpeisiinsa. (2) dokumenteissa ja kyselyissä käytettyjen ilmaisujen epäjohdonmukaisutta hallitaan ottamalla olennaisia ilmaisuja synonyymeina (ja käännöksinä) hakuontologiaan (Järvelin 2001). (3) Kaikki dokumentit, joissa on tekstiä ja mahdollisesti myös annotointeja, ovat haettavissa. Pyrimme tukemaan käyttäjiä, jotka eivät osaa kunnolla ilmaista tarpeitaan. Lisäksi olemme yhdistäneet useita työkaluja, jotka tukevat tiedon tutkimista hakukäyttöliittymässä: online-ontologiaeditori, online-ontologiapohjainen annotointi ja ontologiapohjainen kirjanmerkkien hallinta. Haku- ja tutkimustyökalun, WebExplorerin, perusidea kuvaillaan lähteessä (Baskaya 2009) Tässä tutkimuksessa kehittelemme työkalua lisäämällä ominaisuuksia online-ontologianmuokkaukseen, annotointiin, ja ontologiapohjaiseen tiivistämiseen.

2. WebExplorer -- ontologiaan perustuva työkalu

Ontologian rakentaminen on vaativa tehtävä, joten ontologioiden uudelleen käyttäminen olisi kannattavaa. Saatavilla olevat ontologiat eivät kuitenkaan välttämättä ole tiedonhakuun sopivia, jos ne eivät täytä seuraavia vaatimuksia:

- henkilökohtaisuus: ontologiat ovat henkilökohtaisia tai tarpeeksi läheisiä, niin että hakija tuntee ne.
- Pienimuotoisuus: ontologiat ovat tarpeeksi pieniä, jotta hakija voi hallita niiden sisällön.
- Kartoitus: ontologiat kartoittavat hakijoiden käsitteille synonyymit ja niihin liittyvät avainsanat
- Monikielisyys: ontologiat kartoittavat hakijoiden käsitteiden avainsanat halutuilla hakukielillä
- Muokattavuus: muutokset tiedonhakijoiden kiinnostuksen kohteissa (käsitteissä) voidaan helposti koodata ontologioihin tarvitsematta tarkistaa asiakirjan annotointeja.

Ontologiaeditori (Search Ontology Editor, ShOE) tukee seuraavia WebExplorerin ominaisuuksia:

- ShOE tukee useita formaatteja, kuten XML:ää, joten ontologiat on helppo muuntaa siihen ja käyttää ShOEssa.
- Online-muokkauksen avulla voidaan personoida ontologia valitsemalla olemassa

olevan ontologian kiinnostavat osat ja lisäämällä omia käsitteitä ja synonyymeja.

- Muokattu ontologia voidaan helposti jakaa.
- Ontologia on saatavilla hakua, klusterointia, luokittelua ja tiivistämistä varten.

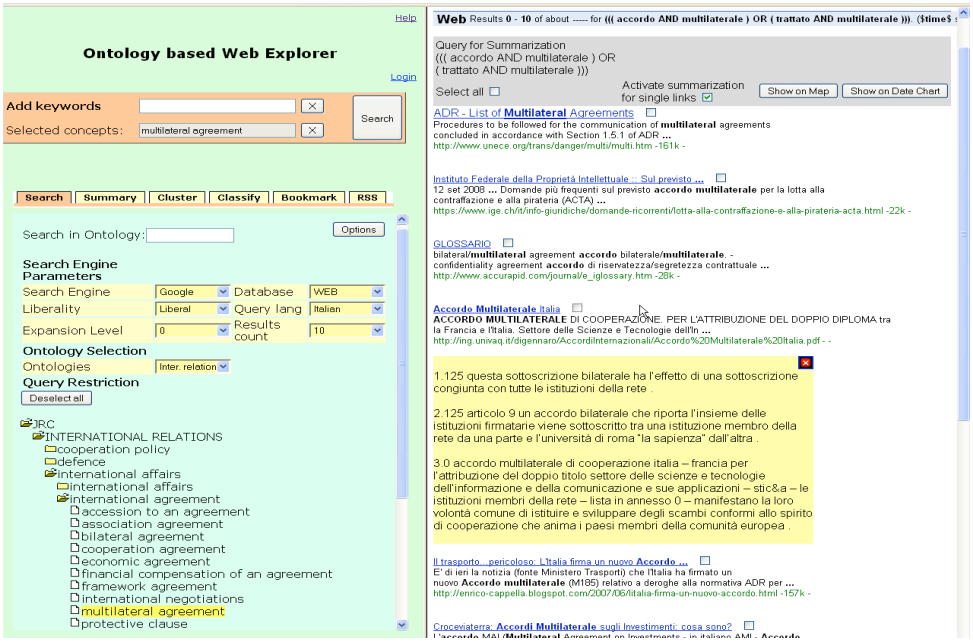
Lisätoimintoja tarvitaan tiedon analysointia varten:

- Avainsanahaku: tiedonhakija voi yhdistää haussa ontologian ulkopuolisia sanoja ontologiakäsitteisiin.
- Tiivistelmät: WebExplorer tarjoaa kahden tyyppisiä tiivistelmiä: a) alkuperäiseen hakuun perustuvia tiivistelmiä ja b) tiivistelmä, jossa lähtökohtana ovat ontologian käsitteet.
- Annotointi: haetut dokumentit voidaan merkitä valittuja ontologioita vasten, myös kielirajojen ylitse, dokumentissa olevien kiinnostavien käsitteiden identifioimiseksi.
- Klusterointi: haettujen asiakirjojen nopea klusterointi ja klusterin merkintäkäsitteiden (cluster label terms) esittäminen tärkeiden uusien käsitteiden identifioimiseksi.
- Kirjanmerkit: tiedonhakija voi säilyttää linkkejä ontologiassa myöhempää konsultaatiota ja analyysia varten.
- Tasainen toimintojen integrointi: niin tekstisanoja, klusterinimiä kuin identifioituja käsitteitä voidaan käyttää joustavasti seuraavassa haku-iteraatiossa ja kehittyvän tiedontarpeen analyysissa.

WebExplorerin käyttöliittymä on suunniteltu toimimaan tehokkaasti ja perustuu modulaariseen arkkitehtuuriin, jotta sitä voidaan käyttää monipuolisesti tiedonhakuun internetissä ja jotta se tukee erilaisia ontologioita. Käyttöliittymä koostuu kahdesta pääpaneelista. Vasen paneeli on WebExplorerin toimintoja varten ja oikea tulosten esittämistä varten (Kuva 1). Vasemmassa paneelissa on seuraavat välilehdet vastaaville toiminnoille: Haku, Tiivistelmä, Klusteri, Luokittelu, Kirjanmerkit ja RSS. Luokittelu-välilehteä voidaan käyttää asiakirjojen annotointiin. WebExplorer aloittaa oletustoimintona Haku-välilehdellä. Ontologiaeditori on toteutettu uuteen ikkunaan. Muokattuaan ontologiaa käyttäjä voi välittömästi käyttää sitä tiedon hakuun ja analyysiin.

WebExplorerin tyypillinen käyttöskenaario on aktiivisten, tarpeen mukaan valittavissa olevien ontologioiden määrittämä vertikaalinen haku verkkodomainissa. Tutkimustoimintojen avulla käyttäjä voi tarkentaa/kehittää tiedontarpeitaan ja merkitä löydetyt asiakirjat luokittelemalla ne vasten valitsemaansa aktiivista ontologiaa. Samanaikaisesti käyttäjä voi saada uusia ajatuksia ontologioiden rakenteesta ja sisällöstä

ja sitten muokata niitä. Tämä mahdollistaa ontologioiden personoinnin ja sen myötä personoidun annotoinnin. Erityisominaisuutena on, että haku voidaan tehdä yhden ontologian (tai vapaan tekstin) kautta ja luokittelu/annotointi toisen kautta. Merkityt asiakirjat voidaan "varastoida" kirjanmerkkeihin annotointien kanssa tai säilyttää ulkoisesti.



Kuva 1. WebExplorerin käyttöliittymä englanti - italia - kieltenvälisessä tiedonhaussa yhden dokumentin tiivistelmän kanssa.

3. KESKUSTELUA JA PÄÄTELMÄT

Olemme esitelleet WebExplorerin, internet-hakutyökalun, jossa on useita toimintoja tiedonhakijoiden omien avainsanojen avulla. Haku on mahdollista internetissä ja tietyissä kokoelmissa, ja RSS-syötteitä voidaan vastaanottaa. Erilaisia ontologioita voidaan käyttää olettamatta dokumenttien annotointia, ja uusia voidaan helposti luoda, muokata online-tilassa ja jakaa. Mahdollisuus muokata olemassa olevaa ontologiaa auttaa myös ontologioiden uudelleenkäyttöä ja yleisten ontologioiden personointia. WebExplorer tarjoaa dokumentin analyysitoiminnon, joka luo ontologiapohjaisen online-tiivistelmän yksittäisistä dokumenteista ja dokumenttiryhmistä. Tiivistelmän näkökulmaa voidaan joustavasti muuttaa. Analyysia voidaan jatkaa klusteroimalla dokumentteja ja identifioimalla klusteroinnin painopistekäsitteitä, joita voidaan käyttää seuraavalla hakukierroksella haun muuttamiseen. Myös ontologiapohjainen kieltenvälinen luokittelu ja kirjanmerkkien hallinta ovat saatavilla. Luokittelua ja klusterointia voidaan käyttää automaattiseen syöttöön haun uudelleen muotoiluun. Nämä toiminnot tukevat tehokkaasti dokumenttin analysointia ja käyttäjän oppimista epäselvien, sekavien hakutavoitteiden tapauksessa.

4. Viitteet

[1] Baskaya, F et al. (2009). WebExplorer: A Tool for Ontology-based Information Exploration. IADIS WWW/Internet 2009, Vol II,pp. 223-229.

[2] Järvelin, K., Kekäläinen, J. and Niemi, T. (2001). ExpansionTool: Concept-Based Query Expansion and construction. *Information Retrieval,* Vol. 4, No. 3/4, pp. 231-255.