

Tekstikaevandamine: infoeraldus

Jüri Reimand
Arvutiteaduse instituut, Tartu Ülikool
thcmob@ut.ee

Andmekaevandamise uurimisseminar MTAT.03.169.
Arvutiteaduse instituut, Tartu Ülikool
Detsember 2003, lk. 161–177

Kokkuvõte

Tekstikaevandamiseks nimetatakse kombineeritud automaatset protsessi, mis analüüsib struktureerimata loomuliku keele teksti, leidmaks mittetriviaalset informatsiooni ja teadmisi. Käesoleva töö eesmärgiks on tutvustada informatsiooni eraldamist kui tekstikaevandamise tehnoloogiat. Näitena vaatleme MedLine abstraktide baasist valk-valk interaktsioonide eraldamise süsteemi.

1 Sissejuhatus

Vaatamata arvutustehnika suurenevale osakaalule on tavalise teadustöö lõpp-produkt - artikkel, väitekiri või raport - ikkagi tekstikujul. Digitaalselt kättesaadavate teadustööde hulk on viimaste kümnendite jooksul eksponentsiaalselt kasvanud. Üha enam pabermeedia sisetust jõuab internetiväljaannete kaudu digitaal-

meediasse. Lisaks on kättesaadavad spetsiaalsed teadusartiklite andmebaasid. Näiteks biomeditsiiniliste abstraktide andmebaas MedLine (<http://www.ncbi.nlm.nih.gov/PubMed/>) Sisaldab üle 11 miljoni artikli abstrakti (2001 a.). Nendest 2,8 miljonit sisaldavad sõnu 'gene' või 'protein', kusjuures 2/3 tulemustest on avaldatud viimase kümnendi jooksul. Spetsiifilisem päring, 'epidermal growth factor receptor', annab üle 10 000 vastuse.

Infoajastul on andmete liikumise pudelikaelaks saanud inimene ise. Valdav enamus kättesaadavast infost on ebakvaliteetne, vananenud või irrelevantne. Vajaliku info ignoreerimise ja kaotatud võimaluste osakaal aina suureneb. Sadade lehekülgede teksti regulaarne analüüsimine ja nõelte heintest eraldamine käib inimesele üle jõu. Tekkinud on vajadus suurte tekstihulkade töötlemise ja haldamise vahendite järele.

Informatsiooni ülekülluse probleem on viinud oluliste arenguteni andmekaevandamise (ik

Data Mining, DM) valdkonnas. Andmekaevandamine on tehnika, mis otsib seaduspärasusi struktureeritud andmete kogumites. Tekstima-terjali puhul on tegemist peidetud grammatilise struktuuriga andmetega. Seega on suur osa digitaalsest informatsioonist traditsioonilistele andmekaevanduse meetoditele kättesaamatu.

Tekstikaevandamiseks (ik *Text Mining, TM*) nimetatakse kombineeritud automaatset protsessi, mis analüüsib struktureerimata loomuliku keele teksti, leidmaks mittetriviaalset informatsiooni ja teadmisi (Feldman 2003).

Tekstikaevandamine hõlmab tehnikaid, mille eesmärgiks on loomuliku keele tekstidest struktureeritud andmete leidmine ning andmekaevandamiseks sobivale kujule viimine. Tekstikaevandamise üheks eesmärgiks on tekstikogude haldus ja eeltötlus ning tekstides sisalduva informatsiooni struktureerimine. Struktureerimiseks kasutatakse loomuliku keele töötuse abil tekstide rühmitamist ning terminite ja informatsiooni eraldamist. Samuti tegeleb tekstikaevandamine töötuse vahe- ning lõpptulemuste säilitamisega ning analüüsiga, kasutades üldiseid andmekaevandamise meetodeid, nagu jao- tuse analüüs, klasterdamine, otsustuspuud, asotsiatsioonireeglid jms.

Andmekaevandamisest tuntud probleemid laienevad ka tekstikaevandamise valdkonda. Vaadeldavad andmekogud on tüüpiliselt suured ja mitmemõõtmelised, suur osa andmetest on müra, esineb ülesobitust, leitud mustrid pole alati arusaadavad. Kuid tekstikaevanduses tekivad ka spetsiifilisemad probleemid. Tekstiline info on mõeldud eelkõige inimeste poolt töötlemiseks ning selle struktuur on keeruline ja halvasti defineeritav. Tekst on alati suurel määral mitmetimõistetav, seda nii leksilisel kui

semantilisel tasemel. Lisaks peab tekstikaevandamine tihti tegelema mitmekeelsusega.

Käesoleva töö eesmärgiks on tutvustada tekstikaevandamisega seotud tehnoloogiaid. Põhjalikumalt peatume informatsiooni eraldamisel kui andmekaevandamisele kõige olulisemal tehnikal. Samuti vaatleme valk-valk interaktsioonide eraldamist MedLine abstraktide baasist.

2 Infootsing

Infootsing (Van Rijsbergen 1979) (ik *Information Retrieval, IR*) on üks vanimaid tekstikaevandamise meetodeid, mille vastu on seoses Interneti kasvuga taas huvi tekkinud. Infootsing seisneb mingist dokumentide kogust kasutaja päringule vastavate dokumentide alamhulga filtreerimises. Päring formuleeritakse tavaliselt loogilise avaldisena. Tagastatud dokumendid võivad olla tähtsuse järgi järjestatud. Samuti on kasutuses dokumentides otsingusõnede esiletõstmine. Kiiremad infootsingusüsteemid suudavad teha mõnesekundisi päringuid üle gigabaidiste tekstihulkade (Blaschke, Hirschman, & Valencia).

Infootsing põhineb paljuski informatsiooni- teoorial, tõenäosusteoorial ja statistikal. Lihtsaimad infootsingusüsteemid vaatlevad tekste kui sõnade kogumeid. Paremad rakendused võimaldavad ka terminite kaalumist, otsingut fraasi järgi, lähendatud otsingut (ik *Proximity Search*) või otsingut tesauruse abil. Filtreerimisel kasutatakse peamiselt otsingusõnade statistilisi omadusi, näiteks nende esinemissagedust pealkirjades ja tekstis. Siit tulenebki infootsingusüsteemide peamine puudus. Infootsing ei pööra tähelepanu vaadeldava teksti sisule. Süsteemide

vigade allikaks on sisendi sõnavaraline varieeruvus, näiteks sünonüümide ('*valk*', '*proteiin*') ja homonüümide ('*pank*') kasutus, erinevad grammatilised vormid, näiteks '*tuba*' ja '*toad*'. Samuti põhjustab vigu välisviitamine, s.t. olukord, kus samale objektile viidatakse lauses erineval viisil asesõnade, lühendite või määrsõnade abil. Kaasaegsetes rakendustes on infootsing enamasti kombineeritud mõne muu tekstikaevanduse meetodiga.

Infootsingusüsteemide hindamiseks kasutatakse peamiselt kahte väärtust. Katvus (ik *recall*) hindab süsteemi poolt leitud andmete ja kogu otsitava andmehulga suhet. Täpsus (ik (*precision*)) hindab tagastatud andmete õigsust. Veel on võetud kasutusse täpsuse *P* ja katvuse *R* kombinatsioon, nn *F*-väärtus, kus β tähistab täpsuse kaalu (Van Rijsbergen 1979).

$$F = \frac{(\beta^2 + 1) * P * R}{\beta^2 * P + R}$$

Analoogseid mõõte kasutatakse ka teiste tekstikaevandamistehnikate hindamiseks.

3 Tekstide rühmitamine

Tekstide rühmitamine (ik *Text Categorisation*) on tegevus, mille käigus märgendatakse loomuliku keele tekstid teemade järgi eelnevalt defineeritud rühmadesse (Feldman 2003). Rühmitamine aitab vähendada tekstide hulka, võimaldades vaatluseks valida ainult teemaga haakuvaid tekste. Tekstide rühmitamises domineerib kaks lähenemist: Teadmuspõhine meetod ja masinõppe meetod.

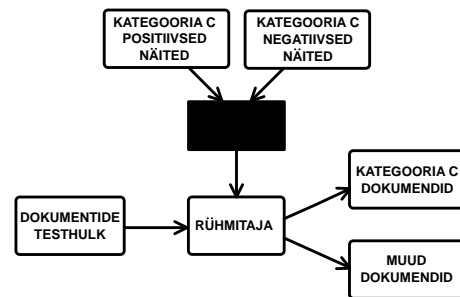
Teadmuspõhises rühmitamises koostab teadmusinsener koostöös antud valdkonna spetsia-

listiga rühmitamise reeglid. Näiteks Reuterile uudiste analüüsiks ehitatud CONSTRUE süsteemis on reeglid konjunktsioonide disjunktsiooni-dena.

```
IF ((wheat & farm) OR
(wheat & commodity) OR
(bushels & export) OR
(wheat & tonnes) OR
(wheat & winter & ¬soft ))
THEN Wheat
ELSE ¬Wheat
```

Hayesi väitel on süsteem väikeste dokumendihulkade (723 dokumenti) korral saavutanud 90% täpsuse (Feldman 2003). Peamiseks probleemiks on teadmuse hankimise protsessi kulukus. Samuti ei ole teadmuspõhised süsteemid kuigi paindlikud. Teadmusbasis tuleb reegli lisamise või muutmise korral üle kontrollida ka ülejäänud reeglite hulk ning eemaldada kattuvad ja vastuolud.

Masinõppe meetodil tekste rühmitava süsteemi peamiseks komponendiks on kogumik märgendatud ja eeldefineeritud kategooriatesse jagatud treeningdokumente. Testkogumik sisaldab iga vaadeldava kategooria kohta nii positiivseid kui negatiivseid näiteid.



Joonis 1: Masinõppepõhine tekstide rühmitaja

Teostatav rühmitamine võib olla kas '*tugev*'

või 'pehme'. Tugeva (automaatse) rühmitamise korral määratakse tekstile iga rühma kohta vastavusse tõeväärtus, 'true', kui tekst kuulub vaadeldavasse rühma, 'false' vastasel juhul.

Pehme rühmitamine seisneb rühma staatuse väärtuse (ik *Categorization Status Value, CSV*) arvutamises. CSV võib omandada reaalarvulisi väärtusi [0..1] ning see näitab, millisel määral võib vaadeldavat teksti pidada antud rühma kuuluvaks. Pehme klassifitseerimine on paindlikum ja veakindlam kui tugev, selle abil on võimalik tekstide sorteerimine tähtsuse järgi.

Tekstide klassifitseerimiseks kasutatakse andmekaevandamisest tuntud võtteid, nagu tõenäosuslikke ja näitepõhiseid klassifitseerijaid, proportsionaalset reeglite õppimist, klasterdamist ja tugivektormasinaid.

Tekstide rühmitamine on andmekaevanduse vaatenurgast liiga üldine protsess. Defineeritud kategooriatele saab vastavusse seada paar-kolm vaadeldava teksti kõige olulisemat teemat, mis on küll abiks töödeldava info mahu vähendamise seisukohast, kuid ei ole piisav informatsiooni eraldamiseks.

4 Informatsiooni eraldamine

Informatsiooni eraldamine (ik *Information Extraction, IE*) (Gaizauskas & Wilks 1998; Cowie & Lehnert 1996; Feldman *et al.* 2002) kombineerib teksti struktureerimiseks loomuliku keele töötluse vahendeid, leksilisi ressursse ja semantilisi piiranguid. Informatsiooni eraldamise protsessi sisendandmeteks on loomuliku keele tekstide kogum, väljundina tagastatakse hulk vajaliku informatsiooniga täidetud mallvorme (ik *templates*).

Kui infootsingu eesmärgiks on oluliste dokumentide eristamine mitteolulistest, siis info eraldamise peamine ülesanne on antud dokumentidest müra eemaldamine ja olulise info struktureerimine. Need kaks tehnoloogiat on seega teineteist täiendavad ning annavad võimaluse uute võimsate vahendite loomiseks.

Võrreldes lihtsamate sõnapõhiste analüüsi vahenditega tekstide rühmitamises tagab infoeraldus konkreetsemad ja täpsemad andmed andmekaevanduse meetodite jaoks. Eraldatud kontseptsioonid ja suhted on kindlamalt seotud vaadeldava dokumendi valdkonnaga. Eri-nevalt tekstide rühmitamisest on infoeralduses oluliste seoste märgendite arv piiramata. Kui rühmitamise tulemusel märgendatakse vaadeldav mõneleheküljeline tekst vastavalt peamisele teemale 1-5 kategooriasse, siis infoeraldusprotsess lisab vastavalt teksti sisule 20-50 märgendit, mis annab andmekaevanduse meetodite rakendamiseks palju parema aluse (Feldman 2003).

Järgnevas toome näite infoeraldussüsteemi sisendist ja väljundist (Cardie 1997).

'4 Apr Dallas - Early last evening, a tornado swept through an area northwest of Dallas, causing extensive damage. Witnesses confirm that the twister occurred without warning at approximately 7:15 p.m. and destroyed two mobile homes. The Texaco Station, at 102 Main Street, Farmers Branch, TX, was also severely damaged, but no injuries were reported. Total property damages are estimated to be \$350,000.'

```
event = tornado
date = 4/3/97
time = 19:15
location = Farmers Branch
```

```
"northwest of Dallas"  
TX : USA  
Damage = "mobile homes" (2)  
"Texaco Station" (1)  
Estimated losses = $350,000  
Injuries = none
```

Infoeraldussüsteemi peamiseks omapäraks on selle seotus vaadeldavate tekstide sisulise valdkonnaga, nn. süsteemi domeeniga. Domeenipõhisus tuleneb otseselt asjaolust, et info eraldamise protsess tegeleb tekstide semantilise ja leksilise analüüsiga, mille käigus võetakse tekstist vastavalt kontekstile objektid ning nendevahelised seosed. Lisaks on oluline sisendandmete sarnane peidetud (grammatiline) struktuur. Tüüpiliseks näiteks on terrorirünnakuid puudutavad meediaartiklid või molekulaarbioloogia teaduslike artiklite abstraktid.

Loomuliku keele töötuse seisukohast pakub infoeraldus mitmeid väljakutseid. Infoeralduse ülesanded on selgelt piiritletud ja tegelevad reaalelu tekstidega, esitades keerukaid loomuliku keele probleeme. Infoeraldussüsteemide jõudlus on võrreldav sama valdkonna inimeksperimentide tööga.

Esimesed infoeralduse alased arengud toimusid juba 1960-tel aastatel. Infoeralduse areng kiirenes märgatavalt 1980-tel aastatel, kui hakkasid toimuma USA kaitseministeeriumi DARPA projekti raames rahastatavad MUC (Message Understanding Conference) konverentsid ja võistlused hindamaks loodud süsteemide headust.

Infoeraldussüsteemi headuse hindamiseks kasutatakse samuti täpsuse ja katvuse mõõte. Katvus näitab infoeralduse kontekstis seda, kui suur osa olulisest tekstist on kaetud tagastatud mall-

vormidega. Täpsus hindab, milline osa tagastatud vormidest kajastab olulist teksti. Lisaks süsteemi täpsusele ja katvusele saab hinnata mallide täitmise õigsust, mallide alatäituvust (ik *undergeneration*, *UND*) ja ületäituvust (ik *overgeneration*, *OVG*) ning vigase info osakaalu tagastatud mallides, nn substitutsiooni (ik *substitution*, *SUB*). Tänapäevaste süsteemide täpsus varieerub olenevalt meetodikast ja valdkonnast 40-90% ning jääb keskmiselt ca 70% juurde. Nõutav täpsus on suuresti konkreetsest valdkonnast, kuid andmebaaside loomiseks kasutatavates süsteemides põhjustab 30% viga liiga palju müra. Cowie ja Lehnert'i arvates on rahuldavaks ja inimanalüütikutega võrreldavaks tasemeks 90% täpsus (Cowie & Lehnert 1996).

4.1 Üldise infoeraldussüsteemi ehitus

Infoeraldussüsteemid on reeglina kompleksed, koosnevad paljudest komponentidest ning nende ühtne defineerimine pole lihtne. Üks võimalikest üldistest informatsiooni eraldamise protsessi definitsioonidest on antud Hobbsi poolt IV MUC konverentsil osalenud süsteemide kohta. Infoeraldussüsteemiks nimetatakse moodulite ja muundajate kaskaadi, mis igal sammul lisavad struktuuri ja kaotavad (loodetavasti) ebavajalikku informatsiooni, rakendades manuaalselt defineeritud või automaatselt genereeritud reeglistikku (Hobbs 1993). Järgnevas on toodud Hobbs'i üldise infoeraldussüsteemi moodulid (Hobbs 1993).

1. Teksti tükeldaja (ik *text zoner*), milles jagatakse tekst segmentideks. Tüüpiliselt eraldatakse eraldatakse juba struktureeritud in-

fo (tabelid, loetelud), pealkirjad ja alapealkirjad. Pikemate tekstide korral jagatakse tekst ka lõikude järgi mingisse struktuuri.

2. Eeltötleja, mis jagab segmendid väikesemateks üksusteks. Esimesena jagatakse tekst kirjavahemärke jälgides lauseteks, kusjuures tähelepanu tuleb pöörata näiteks lühenditele ('etc. '), isikunimedele ('Alan J. Smith') ning numbrilistele andmetele ('17.5').

Järgmiseks sammuks on lausete tükeldamine sõnadeks ja fraasideks ning sõnaliikide märgendamine (ik *Part of Speech Tagging, POS*), mille käigus seatakse igale sõnale vastavusse tema liiki tähistav märgend. Samuti toimub siin pärisnimede tuvastamine ning klassifitseerimine.

Sõnaliikide märgendamiseks on olemas erineva granulaarsuse ja keerukusega märgendikogusid, näiteks 56-elementiline Penn Treebank (www.cis.upenn.edu/treebank). Märgendikogu keerukus mõjutab oluliselt süsteemi jõudlust. Järgnevas näites tähistab /N nimisõna, /V tegusõna, /A määrsõna, /AD määravat artiklit, /P eessõna, /T aega, /TA ajaatribuuti, /NO arvu, /AJ omadussõna ning /PM lauseeraldusmärki.

'Witnesses confirm that the twister occurred without warning at approximately 7:15 p.m.'

Witnesses/N confirm/V that/A the/AD twister/N occurred/V without/A warning/N at/P approximately/A 7:15/T p.m./TA and destroyed/V two/NO mobile/AJ homes/N../PM

Sõnaliikide märgendamine võib olla kombineeritud morfoloogilise analüüsiga, mille käigus eraldatakse sõnatüvest /R sufiksid /SF ja prefiksid /PF ning määratakse grammatiline arv ja ajavorm.

'Witnesses' => witness/R + es/SF

3. Filter, mis eemaldab lausete hulgast ebaolulised, vähendamaks töödeldava info mahtu. Lause olulisuse mõõtmiseks kasutatakse peamiselt kahte lähenemist. Üheks võimaluseks on lausetest teatavate võtmefraaside otsimist. Vaadeldava lause võib lugeda ebaoluliseks, kui temas ei leidu ühtegi võtmefraasi. Võtmefraaside hulga loomine võib toimuda manuaalselt või automaatselt. Alternatiivina hinnatakse lause olulisust lauses esinevate sõnade statistilise mudeli põhjal. Uuringud näitavad, et teksti lõikes väga sageli või harva esinevad terminid ei ole olulise tähtsusega (Cardie 1997).
4. Eel-süntaksianalüsaator, mille ülesandeks on leksiliste üksuste (sõnade) hulgast struktuuride (sõnaühendite, fraaside) leidmine.

..destroyed/V two/NO mobile/AJ homes/N../PM

..destroyed/V two/NO (mobile homes)/N../PM

5. Süntaksianalüsaator, mille sisendiks on üks lause kui sõnade ja fraaside jada ja väljundiks võimalikult terviklik lausepuu. Puu tippudeks on sõnad või fraasid ning tipud on ühendatud süntaktiliste seostega. Teksti parsimist käsitleme järgmises sektsioonis.

6. Fragmentide kombineerija, milles toimub parsimise tulemusel tekkinud lausepuude hulga kombineerimine tekstipuuks. Otsese lähenemisena teisendatakse iga lause loogiliseks avaldiseks, avaldised liidetakse konjunktsioonidega kokku ning eemaldatakse vastuolud ja kattuvused. Alternatiivne meetod vaatleb iga lauset eraldi ning püüab järgmist lauset sobitada olemasoleva struktuuri lünkadesse.
7. Semantiline tõlgendaja, mis genereerib parsitud tekstipuust semantilise struktuuri, ühendades argumendid (sõnad, fraasid) ning neile vastavad predikaadid (täenduslikud seosed). See ning järgnevad komponendid on teadmuspõhised ning enamal määral süsteemi domeenist sõltuvad.
8. Leksiline ühestaja (ik *disambiguation*), mis muudab semantilise struktuuri üldised või mitmetimõistetavad predikaadid ning objektid, näiteks homonüümid 'palk - palga, palgi', konkreetseteks ja ühesteks. Moodulis valitakse võimalike kandidaattähenduste hulgast parim võimalik kandidaatide valimiseks leiavad rakendust elektronkuul sõnaraamatud, leksikonid jms. Teiseks variandiks on domeenipõhine treeningkorpus, milles on näidatud vastava valdkonna jaoks sobivad tähendused. Uuemaks suunaks on meetodite kombinatsioon.
9. Välisviidete lahendamine, milles tuvatatakse samade objektide erinevad kirjeldused. Viidete ühendamise tulemusel muundatakse parsitud puustruktuur graafiks. Välisviidete lahendamine on infoeralduse suuremaks väljakutseks ning see tuleb

vaatluse alla ühes järgnevas sektsioonis.

10. Mallvormide generaator, mis saab sisendiks loodud semantilise struktuuri ning tagastab täidetud mallvormi. Viimane etapp seisneb sisuliselt sisendstruktuuri iteratiivses lihtsustamises kuni malli täitmiseks vajaliku väljundi leidmiseni.

Loomulikult ei kasuta kõik süsteemid täpselt sellist struktuuri. Samuti võivad esineda erinevused protsesside järjekorras. Sageli on 6. ja 7. samm vastupidises järjekorras (Gaizauskas & Wilks 1998).

4.2 Pinnapealne ja sügav lähenemine

Infoeralduse rakendustes toimuvad arengud on viinud diskussioonini loomuliku keele analüüsi teemadel. Laias laastus võib seisukohad jagada kaheks - sügavat ja teoreetilist ning pinnapealset (ik *shallow*) ja rakendustele orienteeritud lähenemist.

Infoeraldussüsteemi arhitektuuri vaadeldes saab parsimise osa jagada kaheks, lause taseme komponendiks ja teksti taseme komponendiks.

Lause taseme komponent töötleb üksikuid sisendi lauseid, tegeledes lause sõnadeks ja fraasideks jagamisega, sõnaliikide märgendamisega ning lause loogilisele kujule viimisega.

Tugevamaid süntaktilise analüüsi mehhanismid rakendavad loomuliku keele lause struktuuri täielikuks analüüsiks kontekstivabu grammatikaid, mis on koostatud võimalikult üldise ülesandepüstituse jaoks ning on loomult keerukad. Niisugused lauseanalüsaatorid töötavad reeglina

eksponentsiaalse ajaga ning kipuvad hätta jääma lausetega, mille pikkus ületab 20-30 sõna (Cowie & Lehnert 1996).

Pinnapealses lähenemises on kasutuses lõplikud mustriotsijad (ik *pattern matchers*), mis otsivad sisendi grammatilise struktuuri osaliseks analüüsimiseks domeenipõhiseid leksiliselt käivitatavaid mudeleid. Näiteks saab järgnevate lihtsate heuristikute abil tuvastada teatud juhtude jaoks grammatilise arvu ja aja.

'..witnesses confirmed..'

REEGEL:

IF noun+'es'

plural(noun);

REEGEL:

IF verb+'ed' AND¬'have'

pastPresent(verb);

Tervikteksti tasemel töötav komponent võtab sisendiks esimese komponendi poolt väljastatud hulga loogilisi lauseid ning integreerib selle semantiliste reeglite abil tervet teksti hõlmavasse struktuuri, mis on aluseks lõpliku vormi täitmisel.

Teksti semantiline analüüs võib samuti olla üldisemal või domeenispetsiifilisel kujul ning tulemuseks anda enam või vähem üldise formalismi. Tugeva parsimise korral kasutatakse üldiste tähenduslike reeglite abil lausetel põhinevate teoreemide tõestamist, et leida sisendile Occami habemenoa printsiibile vastavat vähima kuluga selgitust. Vastupidist, minimaalses lähenemises teostatakse semantilist analüüsi vaid nende lausete kohta, mis paistavad sisaldavat malli täitmiseks olulist infot. Lausete võrku ühendamiseks on sisse toodud *ad hoc* domeenispetsiifilised heuristikud.

'A twister destroyed two mobile houses.'

REEGEL:

target = object(destroyed);

Kahe erineva lähenemise võrdlemiseks võib tuua järgneva näite (Gaizauskas & Wilks 1998). MUC-3 jaoks arvutuslingvistilise teooria vaimus valminud SRI TACITUS kasutas lausete parsimiseks täielikku lingvistilist analüüsi ning tervikteksti semantiliseks analüüsiks teoreemide tõestamist. MUC-4 jaoks valminud TACITUSE järglane FASTUS kasutas edukalt lõplike muundurite kaskaadi ning lausete osaliseks analüüsiks valdkonnaspetsiifilist reeglistikku ja kitsast heuristikat. Tulemused on märgatavalt erinevad: TACITUS'el kulus 100 teksti töötlemiseks 36 tundi, FASTUS läbis 100 teksti 12 minutiga. Võrdluseks, mallitöötlusprogramm mõõtis keskmise inimanalüütiku tööajaks 20 tundi (Cowie & Lehnert 1996). Kuigi MUC-3 ja MUC-4 ülesandepüstituste erinevuse tõttu ei ole süsteemide veakindlus otseselt võrreldav, tuleb mainida, et FASTUS andis 16% võrra rohkem vigu, seda peamiselt vähenenud katvuse tõttu. Autorite sõnul ei olnud esimene lähenemine mitte vale, vaid ülesande jaoks ebasobiv. TACITUS'e eesmärgiks oli teksti mõistmine, FASTUS keskendus aga info eraldamisele, milleks teksti täielik mõistmine ei ole otseselt vajalik.

4.3 Süntaksireeglite automaatne genereerimine

Esimeste edukate infoeraldussüsteemide töö olnes suurel määral ekspertide poolt defineeritud keerukast reeglite kogumist. Infoeralduse uuemad arengud kulgevad teoreetiliselt lingvistikalt empiirilistest meetodite suunas (Cardie 1997). Sellest on ka tingitud püüe tuletada teks-

tistruktuure ja lingvistilist üldistust süsteemide poolt töödeldavatest tekstidest või kasutada reeglite tuletamiseks masinõppe meetodeid.

Statistiliste meetodite kasutamine on osutunud edukaks infoeralduse protsessi leksilise analüüsiga seotud ülesannete lahendamisel. Näiteks on sõnaliikide tüübi määramisel tihti abiks vaadeldava sõna esinemisjaotuse omadused mõnest suurest tekstihulgast. Tihti kombineeritakse statistilisi võtteid elektrooniliste ühe- või kahekeelsete elektrooniliste sõnaraamatute kasutamiseks. On tekkinud ka loominguuline uurimissuund, mis tegeleb leksiliste ressursside automaatse leidmise ja uutele valdkondadele kohandamisega.

Masinõppe meetodeid kasutavad korpusepõhised keeleõppe algoritmid (Cardie 1997), mida saab edukalt rakendada infoeralduse esimestes etappides. Niisugustel algoritmidel põhinevad komponendid omandavad teatud keeletöötuse protsessi uurides hulka vaadeldava protsessi läbiviimise näiteid. Seega oleneb algoritmi töö treeningandmete (korpuse) olemasolust ja headusest. Algoritm täidab oma ülesannet edukalt seni, kuni vaadeldava sisendinfo kirjastiil ja vorm sarnanevad läbitud korpusele. Vastasel juhul tuleb koostada uus korpus ning komponent täielikult ümber treenida. Teoreetiliselt on võimalikud ka üldised ning teistele domeenidele porditavad õppealgoritmid, kuid senised 'üldised' sõnakogud on jäänud üsna väikeseks (Cardie 1997).

4.4 Teadmuse kasutamine

Teadmuseks (ik *knowledge*) nimetatakse informatsiooni tähendust omavas kontekstis, mis on viidud programmi poolt mõistetavale struktuur-

sele kujule ning mida programm saab kasutada ülesannete lahendamisel.

Eelnevas kirjeldatud loomuliku keele analüüsis esinenud viited aksioomidele, domeenipõhistele reeglitele, *ad hoc* heuristikale, jms viitavad asjaolule, et infoeraldussüsteemides on kasutuses teadmuse esitus ja töötlemine. Teaduspõhiste reeglite rakendamine on möödapäasmatu semantiliselt analüüsis, kuna see eeldab süsteemilt tekstiga seonduva "reaalse maailma" taustinfo tundmist.

Teadmuse rakendamises infoeraldussüsteemides tuleb tähelepanu pöörata nn. olulise hinna eeldusele (ik *Significant Price Assumption*), mis väidab, et iga teadmusühiku (reegli) loomine toimub kasutajale teatud kuluga ning see kulu on piisavalt oluline tekitamiseks teadmuse töötlemise pudelikaelu (ik *knowledge-engineering bottlenecks*)(Cowie & Lehnert 1996). Nimelt on teadmusbaasi loomine mahukas ülesanne, mis eeldab teadmusinseneri ning vaadeldava valdkonna spetsialisti koostööd.

Infoeraldussüsteemide arengus on üheks oluliseks uurimissuunaks teadmusega seotud pudelikaelade vältimine. Teadmuse korral on peamiseks ressursse nõudvaks tegevuseks just hankimine ja süsteemile vastuvõetavale kujule viimine. Ühest küljest tuleks teadmusega seotud kulu minimeerida, teisalt aga on teatud tasemel teadmuse töötlus paratamatult vajalik.

Eelnevas viidatud edukas TACITUS süsteem kasutas lausete analüüsiks paljuski valdkonnaspetsiifilist reeglistikku ja kitsast heuristikat, nn pinnapealset (ik *shallow*) teadmust. Pinnapealse teadmuse reeglits sobib näiteks eelnevas toodud tornaado sihtmärgi tuletamise reegel.

Pinnapealse teadmuse korral saab hanki-

da minimaalsete kuludega suurt teadmushulka. Massachussetsi ülikooli UMass tööühma uurimused näitavad, et vastavate vahendite (näit. AutoSlog) abil on võimalik uue valdkonna teadmusbaas luua vähem kui 10 inimtunniga. Vähesese keeletöötuse kogemusega tudengite poolt loodud baasid annavad võrreldavaid tulemusi ekspertide omadega (Cowie & Lehnert 1996).

Pinnapealse teadmuse kriitikaks võib tuua asjaolu, et kitsalt spetsiifiline teadmus on harva laiendatav ning selle edukas teisendamine mõne muu ülesandepüstituse või domeeni jaoks ei ole tõenäoline. Samas tuleks teadmuse hankimise kulud viia nii madalale, et teadmusbaasi loomine ei moodustaks infoeraldussüsteemi loomises olulist osa. Sellisel juhul võib baasi vaadelda kui spetsiaalselt ühekordseks kasutamiseks loodud väärtust. Viimatimainitud asjaolu on oluline just infoeraldussüsteemide porditavuse seisukohast.

4.5 Välisviidete lahendamine

Infoeraldussüsteemide üheks suuremaks väljakutseks on välisviidete lahendamine (ik *anaphora/coreference resolution*) ehk objektide nimede ja erinevate kirjelduste ning vastavate viidete ühendamine (Feldman 2003). Nimetatud protsess on tähtis seetõttu, et vaadeldava teksti jaoks olulisi objekte viidatakse korduvalt erinevate lausete ja isegi lõikude raames ning identseid kirjeldusi ei pruugi esineda. Viidete lahendamine on tundlik teksti struktuuri ja semantika suhtes, samuti sõltub see eelnevate etappide (teksti tükeldamine, sõnaliikide märgendamine, fraaside tuvastamine) tulemusest.

'Early last evening, a tornado swept through an area northwest of Dallas, causing extensive damage. Witnesses confirm that the twis-

ter occurred without warning at approximately 7:15 p.m.'

Hobbsi poolt tehtud uuringud (Feldman 2003) näitasid, et lihtsaimaks osutub pärisnimede ja erinevate sünonüümide lahendamine (näit. *'Soome laht'* kui geograafiline üksus, *'GM'* kui General Motorsi akronüüm). Selleks otstarbeks kasutatakse sageli leksikone (ik *gazetteer*), näiteks Gaizauskase poolt arendatud LASie süsteemis. Mõnevõrra raskemaks loetakse asesõnaliste viidete lahendamist (*'meie, see, teised'*). Kõige keerukamaks ja vigadele altimaks osutub määravate (ik *definitive*) viidete lahendamine (*'John Smith'* kui *'uus kandidaat'*).

Ühe võimaliku lähenemisena pakub Hobbs välja teadmuspõhise heuristilise algoritmi, kus iga esineva viite korral vaadeldakse olenevalt tüübist mingit hulka eelnevaid viiteid (Feldman 2003). Pärisnimede korral vaadeldakse terve teksti viiteid, määrsõnade korral kõiki antud lõigus varem esinenud viiteid ning määravate viidete korral antud lõigus ning eelnevas lõigus esinenud viiteid. Ainsaks erandiks on konstruktsioon *'see X'* (ik *the X*), kus X on organisatsioon/firma vms. Siis laiendatakse otsingu skooopi tervele tekstile. Vaadeldavast kandidaatide hulgast eemaldatakse keeleliselt ebasobivad (ainsus/mitmus, sugu, objekti tüüp). Seejärel järjestatakse viitekandidaadid, eelistades neid, mis esinesid

1. vaadeldavas lauses eespool,
2. eelnevas lauses eespool,
3. ülejäänud lausetes tagapool, alates teksti algusest.

Meetodi testimisel selgus, et 82% juhtudest

pani algoritmi viitele vastavusse õige objekti (Feldman 2003).

5 Bioloogilise info eraldamine: valk-valk interaktsioonid

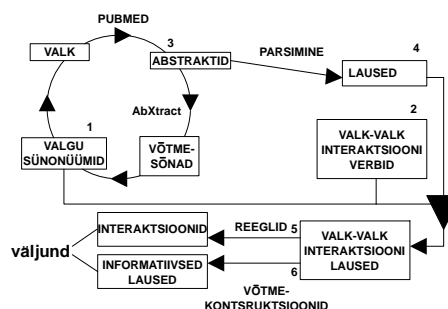
Bioinformaatika edu põhineb paljuski asjaolul, et suurt osa eksperimentaalandmetest, näiteks molekulaarbioloogia sekventse, kogutakse juba mõnda aega struktureeritud kujul mitmesugustesse andmebaasidesse lihtsustamaks edasist eksperimenteerimist. Niisuguste andmebaaside korral saab uurida makromolekulaarse info sidumise võimalusi vähemstruktureeritud (tekstilise) informatsiooniga. Järgnevas kirjeldame süsteemi, mille ülesandeks on valk-valk interaktsioonide eraldamine teadusartiklite abstraktide andmebaasist MedLine (Blaschke *et al.* 1999; ?).

5.1 Süsteemi ehitus

Süsteem saab sisendiks kasutaja poolt valitud valgu nime. Esimese sammuna otsitakse MedLine abstraktide baasist vaadeldava valkude seotud abstraktide hulk. Seejärel viiakse AbXtract süsteemi abil läbi abstraktide analüüs, mis tagastab hulga abstraktidest leitud võtmesõnu, mille hulgast valib kasutaja manuaalselt vaadeldava valgu sünonüümid ja sellega seonduvad valgud. Nendest valitakse abstraktidest edasiseks eraldamiseks mingi alamhulk. Vajadusel võib teha MedLine baasis sünonüümide järgi uue otsingu. Valitud abstraktide laused parsitakse. Järgmisena teatavate reeglite abil viiakse läbi

analüüs ning jagatakse laused kaheks: illustreerivad laused ning valk-valk interaktsioone sisaldavad laused. Analüüsis otsitakse lausetest vaadeldava valgu nimetust või sünonüümi ja mõne eeldefineeritud võtmesõna olemasolu. Tulemusena tagastatakse kasutajale parsitud lausetest koostatud graaf, mille tippudeks on valgud ning servadeks interaktsioonid. Kokku koosneb süsteem alamülesandest:

1. Valkude nimede eraldamine
2. Võtmesõnade eraldamine
3. Tekstikorpuse kogumine
4. Lausete parsimine
5. Lausete analüüs
6. Võtmekonstruktsioonide eraldamine



Joonis 2: valk-valk interaktsioonide eraldamine

Järgnevas vaatleme iga komponenti eraldi.

1. Valkude nimede tuvastamine on seotud välisviidete lahendamise probleemiga, mis on infoeraldussüsteemide üheks suuremaks raskuseks. Käesolevas süsteemis otsitakse

abstraktidest küll vaadeldava valgu ja temaga seotud valkude viiteid, kuid edasiseks tööks eeldatakse kasutaja sekkumist, kelle ülesandeks on leitud terminite hulgast vaadeldava valgu korrektsete sünonüümid ja seotud valkude valimine. Süsteemi järgmistes realisatsioonides on lubatud rakendada spetsiifilisemaid meetodeid, näiteks morfoloogilist analüüsi.

2. Võtmesõnade hulga koostamine toimub käesolevas versioonis manuaalselt. Nendeks on 14 valk-valk interaktsioonidega seotud sõnatüve.

- acetylat-e (-ed, -es, -ion)
- activat-e (-ed, -es, -ion)
- associated with
- bind (-ing, -s, -s to, /bound)
- destabiliz-e (-ed, -es, -ation)
- inhibit (-ed, -s, -ion)
- interact (-ed, -ing, -s, -ion)
- is conjugated to
- modulat-e (-ed, -es, -ion)
- phosphorylat-e (-ed, -es, -ion)
- regulat-e (-ed, -es, -ion)
- stabiliz-e (-ed, -es, -ion)
- suppress (-ed, -es, -ion)
- target

Võtmesõnade manuaalne defineerimine võimaldab vältida keeruka semantilise analüüsi läbimist. Seoses teadustöö abstrakti suhteliselt selge grammatilise

struktuuri ja mahupiiranguga on tõenäoline, et oluline osa interaktsioone on kirjeldatud eelnevate predikaatide abil. Võtmesõnade hulga genereerimiseks võib kasutada ka molekulaarbioloogia terminoloogia kogumikke, näiteks Julian (www.mblab.gla.ac.uk/julian), või ingliskeelset elektronsõnaraamatut, näiteks WordNet (www.cogsci.princeton.edu/wn).

3. Tekstikorpuse koostamiseks kasutatakse päringut valgu nime järgi Medline abstraktide andmebaasi. Tulemuste parandamiseks võidakse teha iteratiivselt kaks või enam päringut, milles sisendiks antakse eelmisest päringust saadud valkude sünonüümid mingi alamhulk.

4. Süsteem kasutab lausete parsimiseks pinnapealset tehnikat ning vaadeldava valdkonna jaoks efektiivset reeglite hulka. Antud realisatsioon ei tegele kahe lingvistiliselt keeruka juhuga. Esiteks ei vaadelda lauseid, mis kirjeldavad tugevat negatiivset seost. Teiseks võib analüüsist välja jääda informatsioon, mis sisaldub peidetud kujul mitmes lauses.

5. Teksti analüüsiks eraldatakse esmalt grammatiliste eraldajate (',', ',', ';') kaudu fragmendid või laused, mida hakatakse eraldi vaatlema. Eelistatud on fragmendid, mis sisaldavad vähemalt kahte nimetust. Antud versioonis on realiseeritud üldkuju '*valkA - interaktsioon - valkB*', mille leidmiseks vaadeldakse lihtsate heuristiliste meetmete abil sõnade asetust ja sagedust. Keerukamate kujude '*valkA - valkB - interaktsioon*'

ja 'interaktsioon - valkA - valkB' analüüs on realiseerimata.

5.2 Valkude interaktsioonide võrgu konstrueerimine

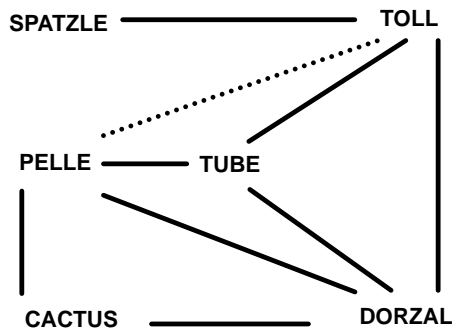
Järgnevas vaatleme süsteemi tööd lihtsa interaktsioonide võrgu koostamisel, milleks on äädikakärbsse (*Drosophila melanogaster*) pellenimelise valguga seotud valkude võrk. Analüüsitud tekstikorpus koosneb 6728 abstraktist. Võrgus esinevate valkude hulka valiti käsitsi 6 valku.

- pelle - valgukinaas
- dorsal - transkriptsiooni faktor
- toll - transmembraani retseptor
- tube - membraaniga seotud tundmatu funktsioon
- spatze - rakuväline ligand tollile
- cactus - dorsal'i inhibiitor

Asjakohase kirjanduse manuaalsel uurimisel tuvastati 9 pelle'ga seotud interaktsiooni. Süsteem leidis nendest korrektselt 8 interaktsiooni. Leitud interaktsioonide võrk on kujutatud joonisel.

Joonisel kujutatud seoseid saab kirjeldada järgnevalt:

1. dorsal ja cactus moodustavad tsütoplasmilise kompleksi.
2. spazle saabub raku pinnale ning seondub retseptoriga toll.



Joonis 3: *Drosophila* pelle interaktsioonide võrk

3. pelle saabub raku pinnale ning seondub/fosforüleerub membraaniga seotud tube' külge.
4. aktiveeritud toll indutseerib tube ja pelle kohaliku seondumise plasma membraanile.
5. pelle ja tube seonduvad dorsal'iga (mis on seotud cactus'ega) ning moodustavad neljast komponendist koosneva kompleksi.
6. pelle fosforüleerib cactus'e.
7. cactus'e fosforüleerumine põhjustab cactus'e vabastamise ja degradeerumise, mille järel dorsal translokeerub tuuma ning suunab geeniekspressiooni.

Punktiiriga tähistatud seos pelle ja toll'i vahel on ilmselt vale. Seos on eraldatud näiteks lausest '*Tube and pelle then transduce the signal from activated toll to a complex of dorsal and cactus.*' Siinkohal viitab võtmesõna '*activated*' toll'i seisundile, mitte pelle ja toll'i vahelisele interaktsioonile.

Järgnevalt on toodud automaatselt tuvastatud interaktsioonid ning nende esinemise arv.

1. dorsal - binds - tube (5)
2. dorsal - regulate - toll (5)
3. pelle - activate(d)- dorsal (10)
4. pelle - activated - toll (5)
5. pelle - interact - dorsal (6)
6. pelle - regulates - dorsal (5)
7. pelle - regulates - tube (5)
8. spatzle - activate - toll (8)
9. spatzle - activates - toll (6)
10. toll - phosphorylates - tube (5)
11. toll - regulated - dorsal (6)
12. tube - activate - dorsal (5)
13. tube - activates - pelle (4)
14. tube - interact - dorsal (10)
15. tube - interact - pelle (4)

Süsteemi poolt kasutatud spazle ja toll vahe-
list interaktsiooni (kujul 'valkA -interaktsioon -
valkB') kirjeldavad laused.

1. MED 97070052: interaktsioon: activate;
valgud: spatzle; toll;
'This process is thought to restrict the ac-
tion of three follicle cell gene functions,
encoded by windbeutel, nudal, and, pipe, to
the ventral follicle cells, where they lead to
the localized activation of a serine prote-
ase cascade required to produce the active
spatzle ligand to **activate** the **toll** receptor.'

2. MED 94170368: interaktsioon: activates;
valgud: spatzle; toll;
'Spatzle acts immediately upstream of the
membrane protein toll in the genetic pat-
hway, suggesting that **spatzle** could encode
the ventrally localized ligand that **activates**
the receptor activity of **toll**.'
3. MED 98362749: interaktsioon: activate;
cleaved; valgud: spatzle; toll;
'Proteolytically **cleaved spatzle** could the-
refore dimerize and **activate** the **toll** recep-
tor by inducing receptor dimerization.'

Laused, mis on mõnel keerukamal üldkujul
ning mis jäävad antud süsteemi poolt vaatluse
alt välja:

1. MED 96033803: Interaktsioon: activated;
valgud: spatzle; toll;
'The ligand for the **toll** receptor is thought
to be **spatzle**, a secreted protein that is **ac-
tivated** by proteolytic cleavage.'
2. MED 98175880: Interaktsioon: inhibited;
valgud: spatzle; toll; cactus;
'Here we demonstrate a dorsalizing ac-
tivity for the heterologous easter, **spatzle**
and **toll** proteins in uv-ventralized xeno-
pus embryos, which is **inhibited** by a co-
injected dominant **cactus** variant.'
3. MED 96422863: Interaktsioon: regulates;
valgud: spatzle; toll; dorsal;
'After fertilization, the initial asymmetry
of the egg chamber is used by the **spatzle**
toll pathway to generate within the emb-
ryo anuclear gradient of the transcription
factor **dorsal**, which **regulates** the regional
expression of a set of zygotic genes.'

Eelneva näite põhjal võib järeldada, et süsteem suudab vabalt leida lihtsamaid ja sagedi esinevaid interaktsioone. Tulemusena võivad keerukad või harvemini mainitud interaktsioonid paista vähemolulistena või jääda koguni vaatluse alt välja. Siit lähtub järgmine probleem: kuidas eristada üldiseid interaktsioone tulemustest, mis on saavutatud eksperimentaalsetes tingimustes. Samuti võib puudulik informatsioon, näiteks mõni vaatluse alt välja jäänud valk, põhjustada interaktsioonide eeldustest mittekorrektsete järelduste tegemist.

Antud süsteemi ehituse juures tuleb tähele panna asjaolu, et töö on rangelt piiratud süsteemi domeenipõhise teadmusega. Süsteemi lihtsad statistilised algoritmid tulevad toime abstraktide piiratud inglise keele kasutusega, lühilausete ja spetsiaalse molekulaarbioloogia terminoloogiaga. Samas on kasutuses kvantitatiivne lähenemine - ühe kasuliku (ja keeruka) lause mõistmise asemel hinnatakse võimalikult paljude erinevate interaktsioonide tuvastamist. Järelikult kasvab süsteemi jõudlus ja täpsus märgatavalt, kui kaasata suurem tekstikorpuse. Teiseks on abiks tekstikorpuse kvaliteedi parandamine, näiteks kaasata otsingusse ainult otsestelt *Drosophila* vaadeldavate valkude funktsioonidega seotud abstraktid FlyBase (flybase.bio.indiana.edu) andmebaasist.

6 Kokkuvõte

Digitaalselt kättesaadava informatsiooni hulk on viimase kümnendi jooksul hüppeliselt kasvanud. Andmekogudest vajaliku ja korrektse informatsiooni leidmine on aina keerulisem ning info leviku pudelikaelaks on inimene ise. Infor-

matsiooni ülekülluse probleem on viinud oluliste arenguteni andmekaevandamise valdkonnas.

Tekstikaevandamine on andmekaevandamise haru, mille eesmärgiks on struktureerimata andmetest ehk loomulikus keeles tekstist mittetruuaalse info leidmine ning selle viimine andmekaevandamiseks sobivale kujule. Tekstikaevandamise meetoditest vaadeldakse infootsingut, tekstide rühmitamist ning infoeraldust.

Käesolevas töös vaatleme põhjalikumalt informatsiooni eralduse tehnoloogiat ning näitena valk-valk interaktsioonide eraldamise süsteemi. Infoeraldus on protsess, mille sisendiks on loomuliku keele tekst ning väljundiks tekstis sisaldunud andmetega täidetud mallvorm. Infoeraldus on andmekaevandamise seisukohast olulisim tekstikaevandamise haru, kuna selle abil on võimalik loomuliku keele keerukast struktuurist eraldada vajalik informatsioon ning viia see edasiseks andmekaevandamiseks sobivale kujule. Infoeraldussüsteemi väljund sisaldab andmebaaside loomiseks ja kaevandamiseks piisavalt spetsiifilist infot.

Kaasaegsed infoeraldussüsteemid kasutavad aina rohkem valdkonnaspetsiifilisi ja heuristilisi vahendeid. Kogemused näitavad, et tekstist info eraldamiseks ei pea süsteem tingimata teksti sisu loomuliku keele meetodite abil parsides mõistma, vaid tihtipeale piisab lihtsamatest meetoditest. Meie arvates on infoeralduse tehnoloogia olulisteks arengusuundadeks kasvav automaatsus ning vähenevad süsteemi loomise kulud. Nii on võimalik luua lihtsalt porditavaid süsteeme, tagada suuremat veakindlust ning rakendada paremini juba olemasolevaid lingvistilisi ressursse.

7 Tulevikusuunad

Kaasaegsete infoeraldussüsteemide peamiseks eesmärgiks on tekstist autori poolt kodeeritud sõnumi eraldamine ja struktureerimine. Kahtlemata huvitavamaks väljakutseks on tegelik tekstikaevandamine eesmärgiga leida suurest tekstide hulgast väiteid kombineerides täiesti uut materjali, mida ei saa järeltada ühestki üksikust tekstist. Näiteks saab ravimatute haiguste sümptomeid, ravimeid ja tulemusi käsitlevatest abstraktidest saadud info alusel automaatselt püstitada huvitavaid senipüstitamata hüpoteese.

Teiseks väga oluliseks suunaks on üldise infoeraldussüsteemi kesta loomine, mida oleks võimalik ilma ekspertide abita uue domeeni või keele jaoks ette valmistada. Seniste süsteemide loomine eeldab lisaks valdkonnaekspertide keeleeksperti ja teadmusinseneri osalust ning keerukat välisviidete ja mitmetimõistetavuse lahendamise algoritme. Uuemad arengud lubavad treenitavaid süsteeme ning leksilise ja semantilise reeglistiku automaatset omandamist kombinatsiooniliselt teiste tekstikaevandamise meetoditega.

Infoeraldussüsteemide suuremaks väljakutseks on kindlasti veel inimeksperdiga võrreldes parem täpsus. Tänapäevaste süsteemide täpsus varieerub olenevalt meetodikast ja valdkonnast 40-90% ning jääb keskmiselt ca 70% juurde. Cowie ja Lehrert'i arvates (Cowie & Lehnert 1996) on rahuldavaks tasemeks 90% täpsus. Selliste parameetrite saavutamiseks ei ole näha ühtegi maagilist lahendust, kuid arvata on, et grammatiliste ja leksiliste ressursside kasvades paraneb ka infoeraldussüsteemide üldine heaolu.

Viited

- Blaschke, C.; Andrade, M. A.; Ouzounis, C.; and Valencia, A. 1999. Automatic extraction of biological information from scientific text: Protein-protein interactions. 60–67.
- Blaschke, C.; Hirschman, L.; and Valencia, A. Information extraction in molecular biology.
- Cardie, C. 1997. Empirical methods in information extraction. *AI Magazine* 18(4):65–80.
- Cowie, J., and Lehnert, W. 1996. Information extraction. *Communications of the ACM* 39(1):80–91.
- Feldman, R.; Regev, Y.; Finkelstein-Landau, M.; Hurvitz, E.; and Koganl, B. 2002. Mining biomedical literature using information extraction.
- Feldman, R. 2003. Mining the biomedical literature using semantic analysis and natural language processing techniques, a link analysis approach.
- Gaizauskas, R., and Wilks, Y. 1998. Information extraction: Beyond document retrieval. *Journal of Documentation* 54(1):70–105.
- Hobbs, J. R. 1993. The generic information extraction system. 87–91.
- Van Rijsbergen, C. J. 1979. *Information Retrieval, 2nd edition*. Dept. of Computer Science, University of Glasgow.