1. Diacritice
2. Paginile de inceput: titlu, autor, cuprins, declaratii
3. Aliniere justify
4. Punctuatie
5. Sistemele participante pe modelul sugerat (algoritm, resurse…) – pag. 2
6. Bibliografie pe modelul de la http://thor.info.uaic.ro/~adiftene/publicatii.html

CriES **Cr**oss-lingual **E**xpert **S**earch – CLEF 2010

Workshop-ul CriES face parte din conferinta din 2010 a CLEF Initiative (Conference and Labs of the Evaluation Forum). Proiectul adreseaza problema cautarii de experti pentru mai multe limbi in contextul social media. Temele principale sunt: metode de gasire a expertilor multilinguali, analiza social media in legatura cu cautarea expertilor, selectarea datelor si evaluarea rezultatelor cautarii expertilor.

Provocarea initiala (CriES Pilot Challenge) propusa de organizatorii de la CLEF consta in urmatoarea problema: participantilor le sunt oferite 2 colecții de date: intrebari (1) si intrebari/raspunsuri (2).

(1) Primul set de date consta in 60 intrebari din diferite domenii, impartite egal pentru 4 limbi: 15 in Engleza, 15 in Spaniola, 15 in Franceza, 15 in Germana.

(2)Al doilea set consta intr-un subset al unei colectii de date pusa la dispozitie de catre Yahoo! Answers ce contine perechi de intrebari si serii de raspunsuri la acele intrebari oferite de diferiti utilizatori in cele 4 limbi de mai sus. Pentru fiecare intrebare este marcat cel mai bun raspuns, impreuna cu ID-ul autorului sau, si ID-ul autorului intrebarii. In plus este dat un DiGraf unde nodurile reprezinta ID-urile utilizatorilor. Muchiile indica legaturile dintre intrebarile puse si raspunsurile date de acestia( cine cui a raspuns).

Scopul este indentificarea expertilor care ar putea raspunde la topicurile (intrebarile) din primul set pe baza intrebarile si raspunsurile deja existente in al doilea. Pentru fiecare topic trebuie produsa o lista ordonata de experti (ID-urile userilor) care pot raspunde (au expertiza in acel domeniu).

Evaluarea de catre organizatori se face folosind result pooling cu masuri standard de evaluare: Precision at cutoff level 10 (P@10), Mean Reciprocal Rank (MRR). Aceste rezultate sunt comparate cu un BaseLine calculate in prealabil folosind BM25 si Z-score, precum si intre ele.

http://www.google.ro/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&ved=0CE4QFjAB&url=http%3A%2F%2Fciteseerx.ist.psu.edu%2Fviewdoc%2Fdownload%3Fdoi%3D10.1.1.90.943%26rep%3Drep1%26type%3Dpdf&ei=qmLET5C8H8Sn4gTFoPDhCQ&usg=AFQjCNG52Yk8vCOfK10qV1C8NjVQhGU0Nw

Solutiile aplicate de catre participanti impreuna cu rezultatele si concluziile lor sunt prezentate in paginile urmatoare:

Algoritm

Resurse

Rezultate

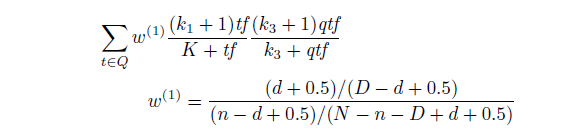
Avantaje

Dezavantaje

Baseline

Acesta este un model standard pus la dispozitie de catre organizatori.

Se folosesc indexe specific limbajelor pentru profilele textelor expertilor. Acestea reprezinta toate raspunsuri pentru fiecare expert intr-o anumita limba. Topicurile sunt traduse folosind Google Translate in fiecare limba pentru a obtine rezultate specific acestora. Acestea abordare s-a facut pe baza analizei distributiei raspunsurilor expertilor in diferite limbi facuta anterior. Este folosit modelul BM25 pentru gasirea informatiilor si Z-Score pentru normalizare iar scorurile finale pentru fiecare expert pentru un topic specific sunt obtinute prin agregare.

Scorul BM25 pentru un document si o cerere/intrebare Q is definit astfel:

Unde *Q* este intrebarea ce contine termini t, iar *w*(1) este relevanta RSJ (Robertson/ Sparck-Jones) a termenului *t* in *Q*.

-Parametrii *k1*, *k3* si *b* sunt default *k1* = 1.2, *k3* = 7 si *b* = 0.75.

-*N* este numarul de documente in colectie, *D* este numarul de documente presupus sau stiut a fi relevant pentru topicul curent.

-*n* este frecventa documentelor pentru termen, *d* este numarul de documente relevante ce contin termenul.

-*tf* este frecventa termenului intr-un document, *qtf* este frecventa termenului in domeniu.

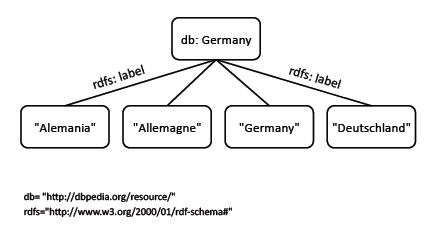
-*K* = *k1( (1 – b) + b \* doclen/avg\_doclen )*, unde *doclen* si *avg\_doclen* sunt lungimea documentului si lungimea media a documentelor.Multilingual Expert Search using Linked Open Data as Interlingual Representation

Daniel M. Herzig and Hristina Taneva Karlsruhe Institute of Technology

Acest model rezolva problema mai multor limbaje folosind Linked Open Data. In majoritatea modelelor de IR (Information Retrieval, adica gasirea/recuperarea informatiilor) documentele sunt vazute si reprezentate ca Bag-of-Words (multime de cuvinte). In acest model ele sunt tratate ca Bag-of-Resources, unde elementele continute sunt tratate ca resurse LOD(Linked Open Data). Acestea sunt structuri de date publice, documentate, cu legaturi intre ele pe web ce folosesc o semantica web standard, in principiu RDF (Resource Description Framework).

**Resurse folosite**

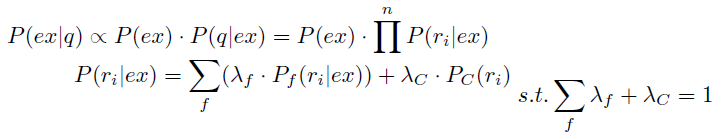
Conceptele sunt identificate prin URI-uri (Uniform Resource Identifiers) care sunt procesate de masini. Deci nu trebuie sa fie lizibile, desi de multe ori au denumiri in engleza.La acestea pot fi atasate etichete care, in acest caz, sunt cuvintele/expresiile in cele 4 limbi ce identifica conceptual respectiv. De exemplu pentru ‘Germania’ avem URI-ul “db:Germany” si etichetele “Alemania”, “Allemagne”, “Germany”, “Deutschland”.



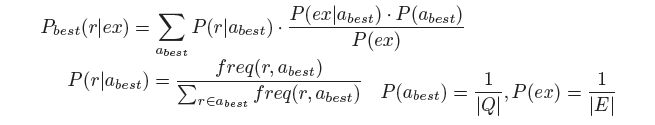
Se folosesc structurile din DBpedia, un set popular de date bazat pe legaturile interlinguale din Wikipedia pentru etichete. Se foloseste Wikipedia Miner Toolkit pentru extragerea informatiilor (resurselor) din documentele propuse. Se identifica candidati din text si elimina ambiguitati, apoi se compara cu textul inconjurator si structura Wikipedia.

**Avantaje**

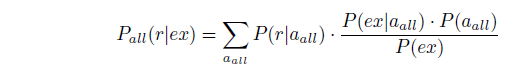
La acest model este aparenta lipsa a ambiguitatii intre limbaje precum si exploatarea conexiunilor intre diferite resurse existente pe Web ce pot oferi alte informatii ajutatoare – dar acest lucru nu este folosit inca in proiect.

Cautarea expertilor se realizeaza folosind modele mixte. Potentialii experti se ierarhizeaza dupa probabilitatea P(ex|q) ca un expert ex din E(useri) sa raspunda la o intrebare q din Q (topicuri), unde q={r1,…,rn} (resursele ce formeaza q).Se presupune ca P(q)=P(ex)=1 din teorema lui Bayes. P(ri|ex) e aproximata ca suma proportionala de f parametrii si informatia pe intreg corpusul C. λC este fixat la 0.1 iar λf este format din λbest si λall , parametrii corespunzatori modelelor:

**Best-Answer Model**

Preia intrebarea q si raspunsul cel mai bun a\* impreuna ca abest si leaga abest de expertul celei mai bune intrebari. Freq(r,a) este frecventa resursei r in textul a. P(ex|abest) este 1 daca ex e autorul abest / 0 altfel.

**All-other-Answers Model**

Preia toate raspunsurile in afara de cel mai bun ca aall si leaga de expertul celei mai bune intrebari. Se presupune ca expertul care a dat cel mai bun raspuns a discutat si celelalte raspunsuri, eventual corectat.

Parametrii pentru cel mai bun rezultat au fost calculate evaluand rezultatele in functie de MAP crescand λbest de la 0(nu se ia in calcul cel mai bun raspuns) la 0.9(se ia in calcul doar cel mai bun raspuns) in trepte de 0.05. Cea mai buna performanta a fost observata cand λbest =0.66 ,deci λall =0.34.

In unele intrebari nu au fost gasite resurse din cauza limbajului informal sau prescurtat. In run1 si run3 dintre cele 3 trimise la workshop aceste resurse au fost completate manual.

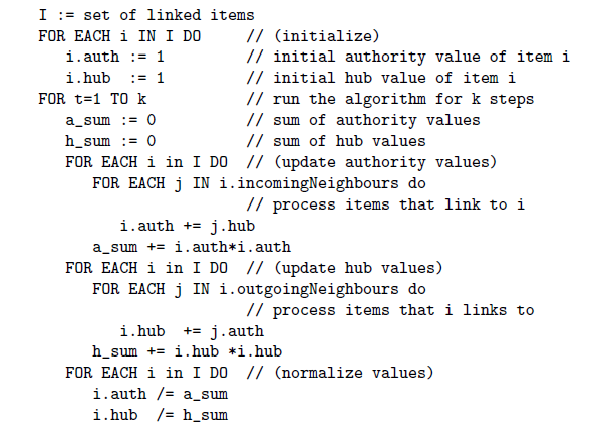
HITS and Misses: Combining BM25 with HITS

for Expert Search

Johannes Leveling and Gareth J. F. Jones Dublin City University

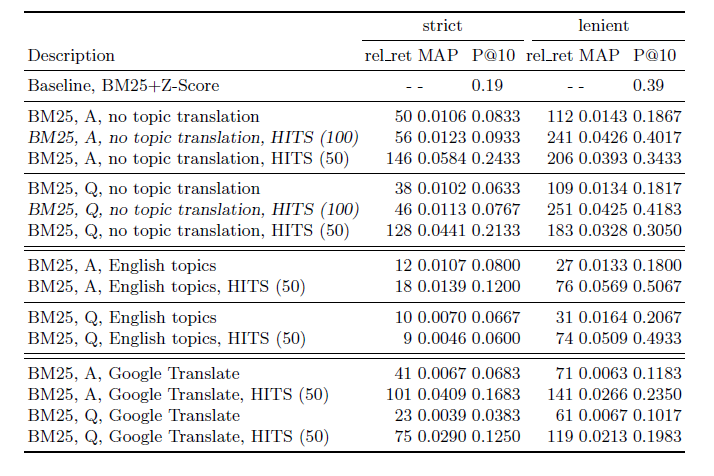
Acest model utilizeaza metode similare cu cele folosite de catre organizatori, gasind informatiile despre experti prin BM25, similar cu sisteme existente de FAQ search (Frequently Asked Questions) si ordonand rezultatele prin algoritmul HITS.

**Algoritmul HITS**

Acesta, impreuna cu PageRank este folosit in evaluarea paginilor web. Spre deosebire de PageRank ce produce un scor static independent de cerere (intrebare), HITS produce 2 valori la executia cererii pe rezultatele initiale: autoriatatea(valoarea continutului paginii web) si hub-ul(valoarea link-urilor catre alte pagini). In acest context autoritatile reprezinta expertii unui domeniu(expertiza) fiind calculate ca suma hub-urilor, iar hub-ul reprezinta persoanele interesate in acesta fiind calculate ca suma valorilor autoritatilor. Acestea se calculeaza pornind de la valorea 1 si iterand procesul de adunare de K ori.

Dupa preprocesarea facuta prin Lucene toolkit si metode standard de tokenizare setul CriES de intrebari-raspunsuri este impartit in 2 tipuri de documente. Documentele de Rapunsuri (A) contin ID-ul expertului impreuna cu raspunsul sau si categoria intrebarii. Documentele de Intrebari (Q) contin ID-ul expertului impreuna cu intrebarea la care raspunde si toate categoriile din document. Asadar se vor face 2 tipuri de cautari: bazate pe raspunsurile, respectiv intrebarile asemanatoare cu un topic. In aceste cautari se foloseste algoritmul de evaluare si extragere BM25 pentri Lucene descris in Baseline (pag 2). Apoi rezultatele se ierarhizeaza si reordoneaza folosind algoritmul HITS. Intrarea pentru acesta este graficul oferit impreuna cu documentele.

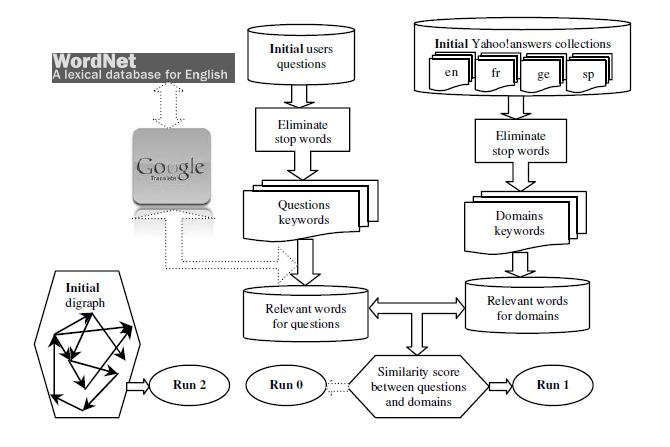
S-au efectuat mai multe serii de experimente: pe cele 2 tipuri de documente (A) si (Q), aplicand algoritmul HITS pe primele 100, respectiv 50 de rezultate. Initial nu s-au tradus documentele dar, dupa analiza distributiei limbajelor in documente (marea majoritate sunt in engleza), s-au facut rulari cu Google Translate si doar pentru topicurile in engleza.



**Identify Experts from a Domain of Interest**

Adrian Iftene, Bogdan Luca, Georgiana Cărăușu, Madălina Merchez Faculty of Computer Science, “Alexandru Ioan Cuza” University, Iasi

Aceasta abordare propune un system format din module care se ocupa cu diferite tipuri de procesari. Acestea includ eliminarea cuvintelor neimportante, obtinerea de sinonime pentru cuvinde in engleza, traducere, analiza/compararea textelor, analiza grafului.



Procesarea initiala consta in urmatoarele actiuni:

Pentru fiecare domeniu de interes se extrag informatiile relevante prin eliminarea cuvintelor nefolositoare pentru fiecare limba (Ex: “the”, “and” pentru engleza, “la”,”et” pentru franceza) din <title> si <description> si adaugand ce ramane la un nou tag <keywords>.

Cuvintele cheie se traduc in engleza cu Google Translate si, folosind WordNet se obtine o lista de sinonime pentru acestea. Sinonimele se traduc apoi in limbile originale si se adauga la un nou tag <synonyms>.

Pentru a creste viteza fisierul XML original(toate domeniile) se divide in functie de categorii in 204 fisiere mai mici pentru o procesare mai usoara.

S-au executat 3 rulari pentru a testa diferitele abordari:

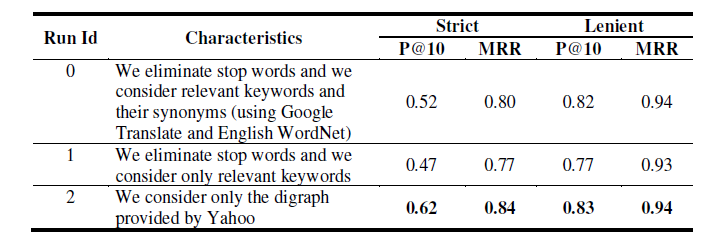
Run 0: Pentru fiecare topic se determina categoria si fisierul corespunzator acesteia (din cele 204 calculate anterior). Se calculeaza un scor de similaritate intre topic si perechile de intrebare/raspunsuri: pentru fiecare cuvant din topic ce apartine tag-ului <keywords> scorul creste cu 2 iar pentru <synonyms> cu 1.

Se aduna scorurile obtinute pentru fiecare user ce a raspuns des. Se considera experti primii 10 ordonati descrescator dupa scorul obtinut.

Run1: Se executa aceleasi calcule de la Run 1 fara a lua in considerare tag-ul <synonyms>.

Run2: In aceasta abordare se foloseste doar digraful oferit de Yahoo! . Se considera pentru fiecare user numarul de raspunsuri date intr-un anumit domeniu ca fiind numarul de muchii cu intrebari in domeniul respective (folosind tag-ul <category> din fisier-ul cu intrebari). Expertii se ordoneaza descrescator dupa acest scor.

Rezultatele sunt mai bune cand se iau in calcul si sinonimele dar cele mai bune rezultate, surprinzator, sunt in Run2 cand se ia in calcul doar graful pentru identificarea expertilor.



Evaluare

S-a folosit result pooling pentru evaluarea rezultatelor gasite de catre grupurile participante. Pentru fiecare rulare primii 10 experti ai fost considerati si evaluati in functie de profilele expertilor. Evaluatorii au primit topic-urile si profilele complete at expertitor formate din toate raspunsurile postate, si au considerat doar datele acestea (toate cunostintele expertilor sunt prepuse a fi prezente in raspunsuri). Pe baza acestora s-au asignat uple topic-expert urmatoarelor clase de relevanta:

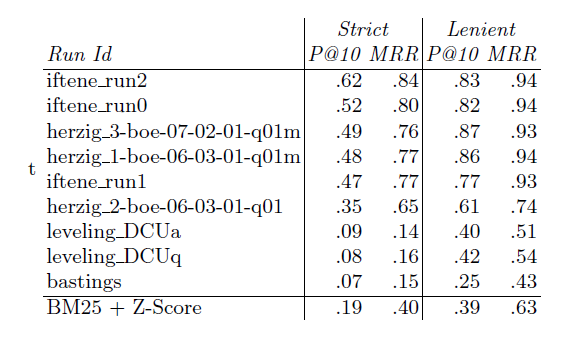
2 Expert care probabil poate sa raspunda.

1 Expert pentru care este posibil sa raspunda.

0 Expert care probabil nu poate sa raspunda.

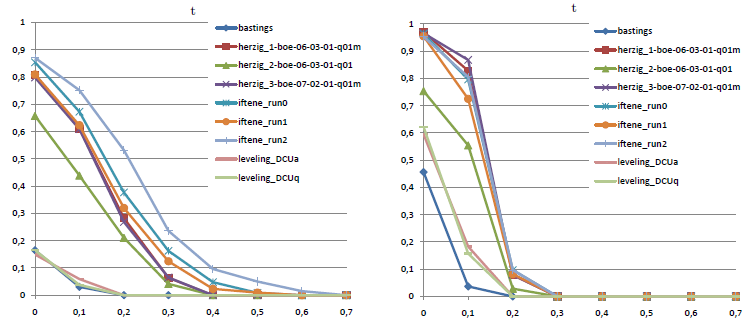
*Strict assessment* ia in considerare doar clasa 2 iar *Lenient assessment* ia in considerare clasele 2 si 1.

**Rezultate participantilor**



Dupa cum se vede, s-au obtinut valori sub baseline in unele abordari dar si valori mult mai bune, mai ales pentru evaluarea Stricta.

**Precision/Recall Curve based on interpolated Recall**



**Strict assessment Lenient assessment**

Suprapunerea expertilor intre rulari este prezentata mai jos. Comparand 2 rulari numerele prezentate reprezinta cati experti pentru fiecare topic Nu au fost gasiti de ambele metode (cati Nu coincid). S-ar observant ca aceasta suprapunere este foarte mica. Chiar si pentru cele mai bune rulari a 2 grupuri diferite (iftene run2, herzig 3-boe-07-02-01-q01m) coincid doar **14%** desi valorile P@10 si MRR sunt apropiate.

**Matrice de disimilaritate**

