**Identificare Experţi**

**Rotariu Ştefan**

**Facultatea de Informatica**

**Universitatea Alexandru Ioan Cuza**

**Iasi**

**Cuprins**

CriES Challenge…………………………………………………………………………………………………….3

Baseline…………………………………………………………………………………………………………….….4

Multilingual Expert Search using LOD as Interlingual Representation………………..…5

HITS and Misses: Combining BM25 with Hits Algorithm.................…………………..….8

Identify Experts from a Domain of Interest…………………………………………………………11

Evaluare………………………………………………………………………………………………………….….13

Bibliografie………………………………………………………………………………………………….……..15

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

Workshop-ul CriES face parte din conferinţa din 2010 a CLEF Initiative (Conference and Labs of the Evaluation Forum). Proiectul adresează problema căutării de experţi pentru mai multe limbi în contextul social media. Temele principale sunt: metode de găsire a experţilor multilinguali, analiza social media în legătură cu căutarea experţilor, selectarea datelor şi evaluarea rezultatelor căutării experţilor.

Provocarea iniţială (**CriES Pilot Challenge**) propusă de organizatorii de la CLEF constă în următoarea problemă: participanţilor le sunt oferite 2 colecții de date: întrebări (1) şi întrebări/răspunsuri (2).

(1) Prima colecţie de date constă în 60 întrebări din diferite domenii, împărţite egal pentru 4 limbi: 15 în Engleză, 15 în Spaniolă, 15 în Franceză, 15 în Germană.

(2)A doua colecţie constă într-un subset al unei colecţii de date pusă la dispoziţie de către Yahoo! Answers ce conţine perechi de întrebări şi serii de răspunsuri la acele întrebări oferite de diferiţi utilizatori în cele 4 limbi de mai sus. Pentru fiecare întrebare este marcat cel mai bun răspuns, împreună cu ID-ul autorului său, şi ID-ul autorului întrebării. În plus, este dat un DiGraf unde nodurile reprezintă ID-urile utilizatorilor. Muchiile indică legăturile dintre întrebările puse şi răspunsurile date de aceştia(cine cui a răspuns).

Scopul este indentificarea experţilor care ar putea răspunde la topicurile (întrebările) din prima colecţie pe baza întrebările şi răspunsurile deja existente în a doua. Pentru fiecare topic trebuie produsă o listă ordonată de experţi (ID-urile userilor) care pot răspunde(au expertiză în acel domeniu).

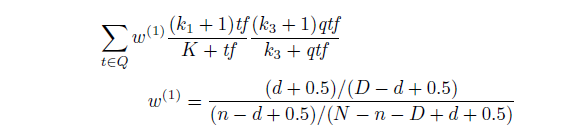
Evaluarea de către organizatori se face folosind result pooling cu măsuri standard de evaluare: Precision at cutoff level 10 (P@10), Mean Reciprocal Rank (MRR). Aceste rezultate sunt comparate cu un BaseLine calculate în prealabil folosind BM25 şi Z-score, precum şi între ele.

Soluţiile aplicate de către participanţi împreună cu rezultatele şi concluziile lor sunt prezentate în paginile următoare:

Baseline

Acesta este un model standard pus la dispoziţie de către organizatori.

Se folosesc indexuri specifice limbajelor pentru profilele textelor experţilor. Acestea reprezintă toate răspunsuri date de fiecare expert într-o anumită limbă. Topicurile sunt traduse folosind Google Translate în fiecare limbă pentru a obţine rezultate specifice acestora. Acestea abordare s-a făcut pe baza analizei distribuţiei răspunsurilor experţilor în diferite limbi făcută anterior. Este folosit modelul BM25 pentru găsirea informaţiilor şi Z-Score pentru normalizare iar scorurile finale ale fiecărui expert pentru un topic specific sunt obţinute prin agregare.

Scorul BM25 pentru un document şi o cerere/întrebare Q îs definit astfel:

Unde Q este întrebarea ce conţine termini t, iar w(1) este relevanţa RSJ (Robertson/ Sparck-Jones) a termenului t în Q.

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

-N este numărul de documente în colecţie, D este numărul de documente presupus sau ştiut a fi relevant pentru topicul curent.

-n este frecvenţa documentelor pentru termen, d este numărul de documente relevante ce conţin termenul.

-tf este frecvenţa termenului într-un document, qtf este frecvenţa termenului în domeniu.

-K = k1( (1 – b) + b \* doclen/avg\_doclen ), unde doclen şi avg\_doclen sunt lungimea documentului, respectiv lungimea media a documentelor.

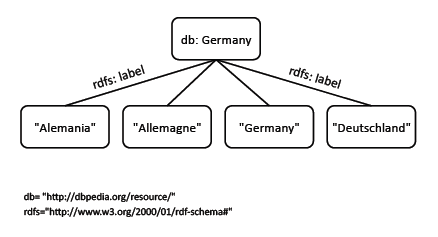
Multilingual Expert Search using Linked Open Data as Interlingual Representation

Daniel M. Herzig and Hristina Taneva Karlsruhe Institute of Technology

***Resurse***

Acest model rezolva problema mai multor limbaje folosind **Linked Open Data**. În majoritatea modelelor de IR (Information Retrieval, adică găsirea/recuperarea informaţiilor) documentele sunt văzute şi reprezentate ca Bag-of-Words (mulţime de cuvinte). În acest model acestea sunt tratate ca Bag-of-Resources, unde elementele conţinute sunt tratate ca resurse LOD(Linked Open Data). Acestea sunt structuri de date publice, documentate, cu legături între ele pe web ce folosesc o semantică web standard, în principiu RDF (Resource Description Framework).

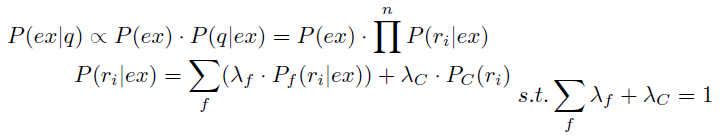
Conceptele sunt identificate prin URI-uri (Uniform Resource Identifiers) care sunt procesate de maşini. Deci nu trebuie să fie lizibile, deşi de multe ori au denumiri în engleză. La acestea pot fi ataşate etichete care, în acest caz, sunt cuvintele/expresiile în cele 4 limbi ce identifică conceptul respectiv. De exemplu pentru ‘Germania’ avem URI-ul “db:Germany” şi etichetele “Alemania”, “Allemagne”, “Germany”, “Deutschland”.



Se folosesc structurile din DBpedia, o colecţie populară de date bazat pe legăturile interlinguale din Wikipedia pentru etichete. Se foloseşte Wikipedia Miner Toolkit pentru extragerea informaţiilor (resurselor) din documentele propuse. Se identifică candidaţi din text şi se elimină ambiguităţi, apoi se compară cu textul înconjurător şi structura Wikipedia.

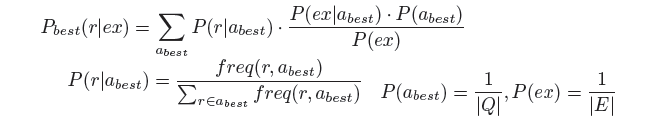
***Algoritm***

Căutarea experţilor se realizează folosind modele mixte. Potenţialii experţi se ierarhizează după probabilitatea P(ex|q) că un expert ex din E(useri) să răspundă la o întrebare q din Q (topicuri), unde q={r1,…,rn} (resursele ce formează q).

Se presupune că P(q)=P(ex)=1 din teorema lui Bayes. P(ri|ex) e aproximată ca suma proporţională de f parametrii şi informaţia pe intreg corpusul C. λC este fixat la 0.1 iar λf este format din λbest şi λall , parametrii corespunzători modelelor:

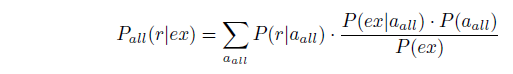
**Best-Answer Model**

Preia întrebarea q şi răspunsul cel mai bun a\* impreună ca abest şi leagă abest de expertul celei mai bune întrebari. Freq(r,a) este frecvenţa resursei r în textul a. P(ex|abest) este 1 dacă ex este autorul abest / 0 altfel.



**All-other-Answers Model**

Preia toate răspunsurile în afară de cel mai bun ca aall şi leagă de expertul celei mai bune întrebări. Se presupune că expertul care a dat cel mai bun răspuns a discutat şi celelalte răspunsuri, eventual corectat.



Parametrii pentru cel mai bun rezultat au fost calculate evaluând rezultatele în funcţie de MAP crescând λbest de la 0(nu se ia în calcul cel mai bun răspuns) la 0.9(se ia în calcul doar cel mai bun răspuns) în trepte de 0.05. Cea mai bună performanţă a fost observată cand λbest =0.66 ,deci λall =0.34.

***Avantaje***

Avantajul la acest model este aparenta lipsă a ambiguităţii între limbaje precum şi exploatarea conexiunilor între diferitele resurse existente pe Web ce pot oferi alte informaţii ajutătoare – dar acest lucru nu este folosit încă în proiect.

***Dezavantaje***

În unele întrebari nu au fost gasite resurse din cauza limbajului informal sau prescurtat. În run1 şi 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

***Resurse***

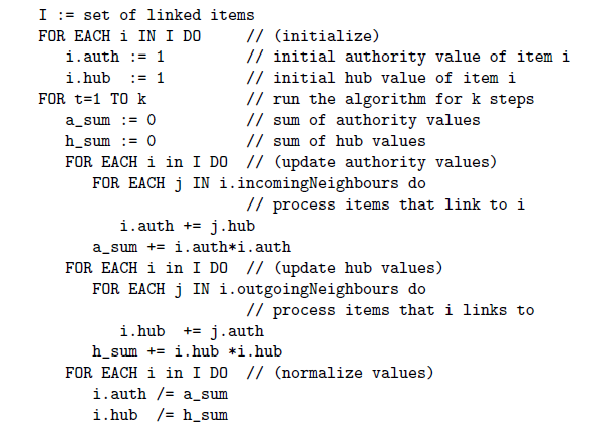
Acest model utilizează metode similare cu cele folosite de către organizatori, găsind informaţiile despre experţi prin BM25, similar cu sisteme existente de FAQ search (Frequently Asked Questions) şi ordonând rezultatele prin algoritmul HITS.

**Algoritmul HITS**

Acesta, împreună cu PageRank este folosit în evaluarea paginilor web. Spre deosebire de PageRank ce produce un scor static independent de cerere (întrebare), HITS produce 2 valori la execuţia cererii pe rezultatele iniţiale: autoriatatea(valoarea conţinutului paginii web) şi hub-ul(valoarea link-urilor către alte pagini).

În acest context autorităţile reprezintă experţii unui domeniu(expertiză) fiind calculate ca suma hub-urilor, iar hub-ul reprezintă persoanele interesate în acesta fiind calculate ca suma valorilor autorităţilor. Acestea se calculează pornind de la valorea 1 şi iterand procesul de adunare de K ori.

***Algoritm***



După preprocesarea făcută prin Lucene toolkit şi metode standard de tokenizare colecţia CriES de întrebări-răspunsuri este împărţită în 2 tipuri de documente.

Documentele de Rapunsuri (A) conţin ID-ul expertului împreună cu răspunsul său şi categoria întrebării. Documentele de Întrebări (Q) conţin ID-ul expertului împreună cu întrebarea la care răspunde şi toate categoriile din document. Aşadar se vor face 2 tipuri de căutări: bazate pe răspunsurile, respectiv întrebările asemănătoare cu un topic. În aceste căutări se foloseşte algoritmul de evaluare şi extragere BM25 pentri Lucene descris în Baseline(pag 2). Apoi rezultatele se ierarhizează şi reordonează folosind algoritmul HITS. Intrarea pentru acesta este graficul oferit împreună cu documentele.

***Avantaje***

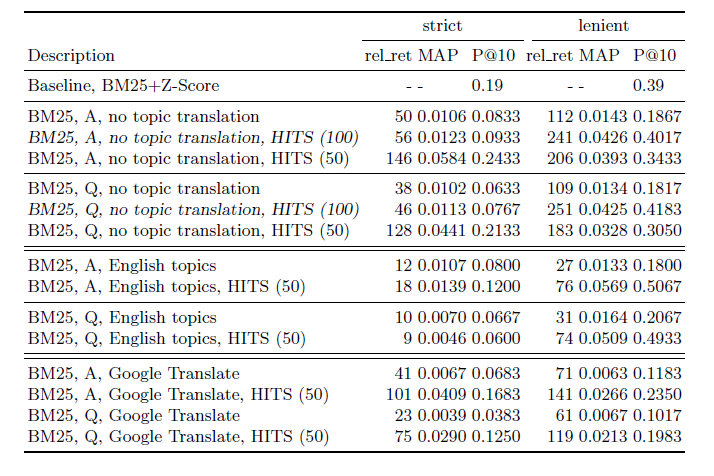
Se iau în calcul mai mulţi factori, inclusiv legăturile dintre utilizatorii care pun întrebări şi cei ce răspund(se pot considera utilizatorii ce pun multe întrebări care au răspunsuri bune ca surse valoroase de informaţii, chiar dacă nu sunt consideraţi experţi).

***Dezavantaje***

Metoda nu pare să îmbunătăţească cu mult rezultatele iar în unele cazuri acestea sunt chiar sub Baseline.

***Rezultate***

S-au efectuat mai multe serii de experimente: pe cele 2 tipuri de documente (A) şi (Q), aplicând algoritmul HITS pe primele 100, respectiv 50 de rezultate. Iniţial nu s-au tradus documentele, dar, după analiza distribuţiei limbajelor în documente (marea majoritate sunt în engleză), s-au făcut rulări cu Google Translate şi doar pentru topicurile în engleză.

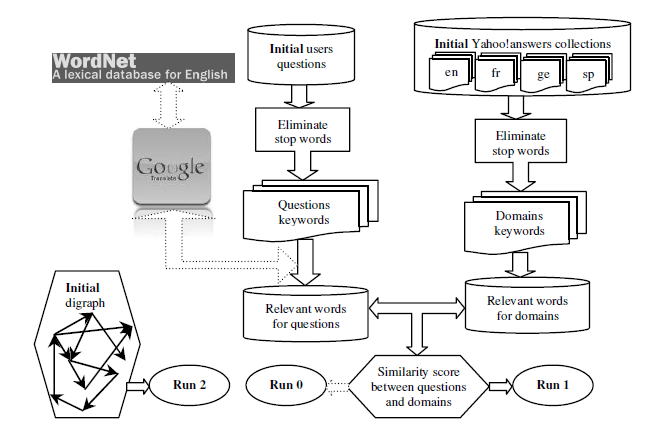


**Identify Experts from a Domain of Interest**

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

***Resurse***

Această abordare propune un system format din module care se ocupă cu diferite tipuri de procesări. Acestea includ eliminarea cuvintelor neimportante, obţinerea de sinonime pentru cuvinde în engleză, traducere, analiza/compararea textelor, analiza grafului.



***Algoritm***

Procesarea iniţială constă în următoarele acţiuni:

Pentru fiecare domeniu de interes se extrag informaţiile relevante prin eliminarea cuvintelor nefolositoare pentru fiecare limba (Ex: “the”, “and” pentru engleză, “la”, ”et” pentru franceză) din <title> şi <description> şi adăugând ce rămâne la un nou tag <keywords>.

Cuvintele cheie se traduc în engleză cu Google Translate şi, folosind WordNet, se obţine o listă de sinonime pentru acestea. Sinonimele se traduc apoi în limbile originale şi se adăugă la un nou tag <synonyms>.

Pentru a creste viteză fişierul XML original(toate domeniile) se divide în funcţie de categorii în 204 fişiere mai mici pentru o procesare mai uşoară.

S-au executat 3 rulări pentru a testa diferitele abordări:

**Run 0**: Pentru fiecare topic se determină categoria şi fişierul corespunzător acesteia (din cele 204 calculate anterior). Se calculează un scor de similaritate între topic şi perechile de întrebare/răspunsuri: pentru fiecare cuvânt din topic ce aparţine tag-ului <keywords> scorul creşte cu 2 iar pentru <synonyms> cu 1.

Se adună scorurile obţinute pentru fiecare user ce a răspuns des. Se consideră experţi primii 10 ordonaţi descrescător după scorul obţinut.

**Run1**: Se executa aceleaşi calcule de la Run 1 fără a lua în considerare tag-ul <synonyms>.

**Run2**: În această abordare se foloseşte doar digraful oferit de Yahoo! . Se consideră pentru fiecare user numărul de răspunsuri date într-un anumit domeniu ca fiind numărul de muchii cu întrebări în domeniul respective (folosind tag-ul <category> din fişier-ul cu întrebări). Experţii se ordonează descrescător după acest scor.

***Avantaje***

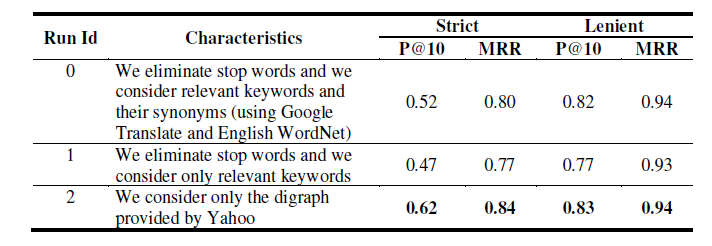
Algoritmul şi modul de procesare a informaţiilor este uşor de înţeles şi rezultatele sunt îmbunătăţite considerabil. În plus se iau în calcul factori specifice fiecărei limbi(sinonimele).

***Dezavantaje***

Este necesară parcurgerea datelor de mai multe ori şi compararea multor termeni, lucru ce poate rezulta în timp de calcul mare.

***Rezultate***

Rezultatele sunt mai bune când se iau în calcul şi sinonimele dar cele mai bune rezultate, surprinzător, sunt în Run2 când se ia în calcul doar graful pentru identificarea experţilor.



Evaluare

S-a folosit result pooling pentru evaluarea rezultatelor găsite de către grupurile participante. Pentru fiecare rulare primii 10 experţi au fost consideraţi şi evaluaţi în funcţie de profilele experţilor. Evaluatorii au primit topic-urile şi profilele complete at expertitor formate din toate răspunsurile postate, şi au considerat doar datele acestea (toate cunoştinţele experţilor sunt presupuse a fi prezente în răspunsuri). Pe baza acestora s-au asignat uple topic-expert următoarelor clase de relevanţă:

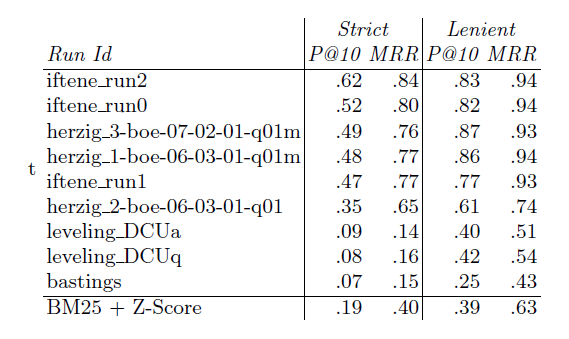
2 Expert care probabil poate să răspundă.

1 Expert pentru care este posibil să răspundă.

0 Expert care probabil nu poate să răspundă.

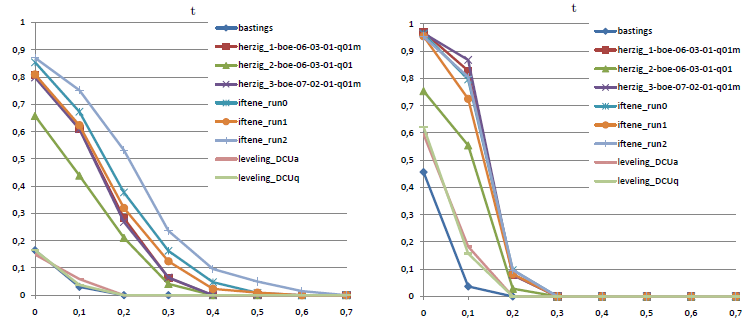
**Strict assessment** ia în considerare doar clasa 2 iar **Lenient assessment** ia în considerare clasele 2 şi 1.

**Rezultate participantilor**



După cum se vede, s-au obţinut valori sub baseline în unele abordări dar şi valori mult mai bune, mai ales pentru evaluarea Strictă.

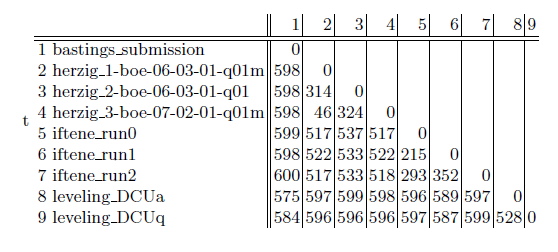
**Precision/Recall Curve based on interpolated Recall**



**Strict assessment Lenient assessment**

**Suprapunerea experţilor** între rulări este prezentată mai jos. Comparând 2 rulări numerele prezentate reprezintă câţi experţi pentru fiecare topic NU au fost găsiţi de ambele metode (câţi nu coincid). S-ar observant că aceasta suprapunere este foarte mică. Chiar şi pentru cele mai bune rulări a 2 grupuri diferite (iftene run2, herzig 3-boe-07-02-01-q01m) coincid doar **14%** deşi valorile P@10 şi MRR sunt apropiate.

**Matrice de disimilaritate**



**Bibliografie**

* Iftene, A., Luca, B., Cărăușu, G., Merchez, M. 2010. Identify Experts from a Domain of Interest. *Notebook Paper for the CLEF 2010 LABs Workshop*, ISBN 978-88-904810-0-0, ISSN 2038-496322-23, 22-23 September, Padua, Italy.
* Herzig, M.,Daniel, Hristina, T. 2010. Multilingual Expert Search using Linked Open Data as Interlingual Representation. *Notebook Paper for the CLEF 2010 LABs Workshop,* September, Padua, Italy.
* Leveling, J., Gareth, J., F. 2010. Hits and Misses: Combining BM25 with HITS for Expert Search. *Notebook Paper for the CLEF 2010 LABs Workshop,* September, Padua, Italy.
* Sorg, P., Cimiano, P., Schultz, A., Sizov, S. 2010. Overview of the Cross-lingual Expert Search (CriES) Pilor Challenge. 2010. In: Working Notes of the CLEF 2010 Lab Sessions, 20-23 September, Padua, Italy.
* Sorg, P., Cimiano, P. Enriching the Crosslingual Link Structure of Wikipedia – A Classification-Based Approach. AAAI2008, Institute AIFB, University of Karlsruhe, D-76128 Karlsruhe, Germany.
* Manning, C., Raghavan, P., Hinrich, S. Introduction to Information Retreival. 2009. Cambridge University Press. Cambridge, England.
* Sorg, P., Cimiano, P. 2011. Finding the right expert: Discriminative models for expert retrieval. Institute AIFB, University of Karlsruhe, Germany.