University of Twente at the TREC 2007 Enterprise Track: Modeling relevance propagation for the expert search task

Pavel Serdyukov, Henning Rode, Djoerd Hiemstra
Database Group, University of Twente
PO Box 217, 7500 AE
Enschede, The Netherlands
{serdyukovpv, rodeh, hiemstra}@cs.utwente.nl

ABSTRACT

This paper describes several approaches which we used for the expert search task of the TREC 2007 Enterprise track. We studied several methods of relevance propagation from documents to related candidate experts. Instead of one-step propagation from documents to directly related candidates, used by many systems in the previous years, we do not limit the relevance flow and disseminate it further through mutual documents-candidates connections. We model relevance propagation using random walk principles, or in formal terms, discrete Markov processes. We experiment with infinite and finite number of propagation steps. We also demonstrate how additional information, namely hyperlinks among documents, organizational structure of the enterprise and relevance feedback may be utilized by the presented techniques.

1. INTRODUCTION

This is the third year of TREC 2007 Enterprise Track and the first year the University of Twente (Database group) participates in it. We submitted results only for the expert search task, the task of finding knowledgeable persons in the organization in response to the user query.

Most popular approaches to expert finding basically consist of two stages. They calculate some score, or the probability of a document to be relevant to the query, and then represent a candidate's expertise score as a sum of scores of related documents. For example, one of these methods, described in works by Balog et al. [1], follows the language modeling principle of IR and considers the expertise level of the candidate expert e to be calculated as:

$$Expertise(e) = \sum_{D \in R} P(D|Q)P(e|D) \qquad (1)$$

where if $P(D)$ is distributed unform, then $P(D|Q) \approx P(Q|D)$, what is the probability of the document D to generate the query Q. This probability is the measure of document relevance according to LM-based IR. $P(e|D)$ is the probability of association between the candidate and the document, what may depend on various factors: on the part of the document where a candidate mentioned, on our confidence that the mentioning of some person matches a specific candidate etc.

If we look at the Formula 1 we may notice that it mathematically describes a probabilistic process, in which a user selects a document among the ranked ones with some probability, reads the document, finds all candidate experts mentioned in it and makes an inquiry to one of them with some probability. The selection of a document depends on its rank, or probabilistic score, and the following selection of a candidate depends on the level of its benefit/responsibility to the content of the document. We may look at this process as at one-step relevance probability propagation from documents to related candidate experts.

However, in reality its not likely that reading only one document and consulting only one person is enough to completely satisfy a personal information need in the enterprise. We may imagine that a user is willing to question several people and hence to find more contacts in the ranked documents by reading more of them. We may also find natural that a user goes over the ranked documents, not only coming back to them again and again, but following hyperlinks among documents. The discovery of new experts may be possible not through documents only, but also with the help of candidates the user is in contact with already: for example, they can send the user to their colleagues in the same department who expectedly possess similar expertise.

In our methods described in the next section we try to overcome the limitations of the one-step relevance propagation. We show how we model the process of collecting the expertise in the enterprise representing it as a multi-step or a non-stop process of consulting with both kinds of possible knowledge sources: documents and people. The rest of the paper contains our experimental results and the conclusions.

2. RELEVANCE PROPAGATION AS A RANDOM WALK

Our main goal in this work was to propagate the relevance coming from documents not only to immediate candidate neighbors, but further, through consequent connections in document-candidates graph, toward candidates and documents connected to starting nodes indirectly. We consider that the process of finding a relevant expert can not be modeled only as a one-step move from the ranked documents to the related persons, as described in Section 1. We present two kinds of random walk models: with finite and infinite number of steps. We consider that the probability of being in the specific candidate's state at the end of the walk reflects the personal expertise of this candidate. Obviously, if a candidate is very central (i.e. is mentioned in many documents, which are in turn contain references to many candidates and so on) then the probability to end up with her will be higher.

However, in both methods we assume that during the walk a user always has some strategy which allows him to find ex-

perts relevant exactly to his query, but not just authoritative in the organization. This means that a user would always prefer not to move too far away from the ranked documents and/or to return to them regularly, taking a document from the ranked list again and starting a new walk. Now we describe our two random walk models implementing this strategy in two different ways.

There are some probabilities in formula 1 which we also use in our work. The probability of query to be generated by the document language model is calculated as:

$$P(Q|D) = \prod_{q \in Q} P(q|D), \tag{2}$$

$$P(q|D) = (1 - \lambda_G)\frac{tf(q,D)}{|D|} + \lambda_G \frac{\sum_D tf(q,D)}{\sum_D |D|} \tag{3}$$

where $tf(q,D)$ is a term frequency of q in the document D, $|D|$ is the document length and λ_G is a Jelinek-Mercer smoothing parameter - the probability of a term to be generated from the global language model. In all our experiments it is set to 0.5, what is standard in retrieval tasks.

In our work, we also exploit the probabilities of selection a document given a person and of selection a person given a document:

$$P(D|e) = \frac{a(e,D)}{\sum_D a(e,D)}, P(e|D) = \frac{a(e,D)}{\sum_e a(e,D)}, \tag{4}$$

where $a(e,D)$ is the non-normalized association score between the candidate e and the document D.

2.1 K-step random walk

In this approach, we imagine that a user after getting the list of ranked documents with the list of related candidates attached:

- selects one document with a probability proportional to its probability of relevance,
- makes K-steps of two kinds: if a user is in the document node, then one of related candidates is selected, or if a user is in the candidate node, then one of documents related to this candidate is selected.

If we consider this process as a walk over a bipartite graph with documents and candidates layers of nodes, than this becomes a process of moving to a node from an opposite layer at each step, starting from some node in a document layer.

We consider this walk as finite, so we believe in this case that at some point a user is tired or satisfied with some candidate and stops their search process. So, we iteratively calculate the probability that a random surfer will end up in a certain candidate starting the walk from the one of ranked documents, using the following formulas:

$$P_0(D) = \frac{P(Q|D)}{\sum_{D \in R} P(Q|D)}, P_0(e) = 0, \tag{5}$$

$$P_k(D) = \sum_{e \to D} P(D|e) P_{k-1}(e), \tag{6}$$

$$P_k(e) = \sum_{D \to e} P(e|D) P_{k-1}(D) \tag{7}$$

The probability of starting this walk from a specific document is proportional to its probability to be relevant.

2.2 Infinite random walk

In our second approach, we assume that the walk process has an infinite number of steps. Stationary probabilities of ending up in the candidate nodes are considered estimators of their expertise. However, the stationary distribution of a discrete Markov process which we use in our modeling does not depend on the initial distribution over states. In other words, the more steps the user does, the less important that the walk was started from certain document nodes and hence their relevance has much less influence on the selection of candidates. In the case of a non-stop walk along the paths with infinite number of nodes, the relevance appearing from ranked documents is propagated so often that it just gets spread equally over all nodes in the graph at some step in the future.

In order to retain the importance for a candidate to stay in proximity to relevant documents, we introduce the possibility to return regularly to the documents from any node of the graph and start the walk through mutual documents-candidates links again. We consider that the probability of jumping to the specific document equals its probability to be relevant, what makes candidates which are situated closer to them to be visited more often during a normal walk. The following formulas are used for iterations until convergence:

$$P_k(D) = \lambda \frac{P(Q|D)}{\sum_{D \in R} P(Q|D)} + (1 - \lambda) \sum_{e \to D} P(D|e) P_{k-1}(e), \tag{8}$$

$$P_k(e) = \sum_{D \to e} P(e|D) P_{k-1}(D) \tag{9}$$

λ is the probability that at any step the user decides to make a jump and not to follow outgoing links anymore. The described Markov process is stochastic and irreducible (since each candidate and document is reachable due to introduced jumps to documents from any node) and hence has the stationary distribution.

2.3 Adding organizational and document links

Our approaches are graph-based in nature. Usually, for this family of algorithms, the introduction of the new information into the analysis often comes to adding new discovered links among analyzed entities. This often helps to model their mutual relations and directions of influence better. The scenario of searching for expertise in the enterprise may include not only the moving from relevant documents to the candidates found in them, but also along document-document and candidate-candidate connections. In the case of the CSIRO collection, we can extract both of these types. Most CSIRO documents are made for web publication and hence often refer to each other. Candidates in turn often belong to the specific department in the CSIRO institute, whose name is usually the part of their email address. For example, the email address Alan.Smith@ento.csiro.au shows that Alan Smith works at the CSIRO Entomology research department. We inter-link all candidates in the same department and also take into account the hyperlinks among documents. After adding these new transitions to our documents-candidates graph, we have the following formulas for the Infinite Random Walk iterations:

$$P_k(D) = \lambda \frac{P(Q|D)}{\sum_{D \in R} P(Q|D)} + (1-\lambda)((1-\mu_D) \sum_{e \to D} P(D|e) P_{k-1}(e) +$$

$$+ \mu_D \sum_{D' \to D} P(D|D') P_{k-1}(D')), \quad (10)$$

$$P_k(e) = (1-\mu_e) \sum_{D \to e} P(e|D) P_{k-1}(D) + \mu_e \sum_{e' \to e} P(e|e') P_{k-1}(e'), \quad (11)$$

where μ_D is the probability of following document-document connections being in the document node, μ_e is the probability of following candidate-candidate connections being in the candidate node. The new transition probabilities are calculated as:

$$P(D|D') = 1/N_{D'}, P(e|e') = 1/N_{e'}, \quad (12)$$

where $N_{D'}$ is the number of outgoing document links from the document D' and $N_{e'}$ is the number of outgoing candidate links from the candidate e'.

2.4 Adding relevance feedback

The useful feature of the relevance propagation approaches is that user feedback easily fits their framework. While in the absence of feedback information the relevance appears solely from the documents content, in the case of feedback it comes into the system also directly from user and shared among positively judged documents. Considering that, we utilized the list of relevant documents for a query provided by TREC in a simple way. We found the document among the ranked ones with the highest score and gave twice higher score to the feedback documents assuming that this relevance probability will be propagated to the adjacent candidates and further.

3. RELATED WORK

The presented approaches belong to the family of document-centric expert finding methods. They consider that the best estimator for the candidates expertise is the aggregated relevance of related documents [1, 9] or of surrounding text windows around candidate mentionings in these documents [3, 8]. Some document-centric methods already utilized the social network built using links among persons extracted from top documents. Campbell et al. [2] proposed the use of HITS algorithm [6] which performed better than just ranking by candidate's in-degree (related documents number). However, Chen et al. [4] found that a document-centric approach is still better than HITS based only ranking.

Markov chains are widely used in IR, mostly as variations of Google's Pagerank [7]. Some of them use strategies for the propagation of document relevance similar to ours. Jeh and Widom presented the Personalized Pagerank algorithm where random surfer jumped more likely to the documents which user initially preferred [5]. The approach by Richardson and Domingos makes a random, but "intelligent" surfer not only to follow hyperlinks, but also to move always in the direction of more relevant documents [10]. The propagation of item preference among similar users is modeled with discrete K-step Markov process for collaborative recommendation by Song et al. [11].

	MAP	MRR
qorwkstep	0.1441	0.2250
qorw	0.1463	0.2378
qorwnewlinks	0.1481	0.2478
feedbackrun	0.2371	0.3517

Table 1: TREC 2007 Expert search official results, candidates recognition with email-addresses only, 1500 documents retrieved

4. RESULTS

The CSIRO collection was indexed using Snowball stemmer at the text parsing stage. This year participants were not given a list of candidates, but just a structural template of their email addresses helping to recognize them in the documents. For the purpose of finding candidate experts, we extracted all emails from the collection with *csiro.au* domain and *firstname.lastname*-like first part. We also made an automatic match of emails with the same first part, but with different subdomains to one candidate identifier, found *csiro.au* email address without subdomains and the same *firstname.lastname* part. For example: Alan.Smith@cmis.csiro.au, Alan.Smith@ento.csiro.au → Alan.Smith@csiro.au. If the email address without subdomains did not exist in the collection for the specific person, it was made up. We also had a list of email addresses to be banned which were not personal, but organizational addresses (e.g. publishing.photos@csiro.au). Documents were retrieved using language model-based ranking function (see Formula 2). For the analysis we took only top-K ranked documents.

In the experiments whose results we submitted to TREC, we retrieved 1500 documents since for the previous TREC testbeds, the number of retrieved documents necessary for the maximum performance varied from 1000 to 8000. We set the jumping probability λ to 0.01, μ_e and μ_D were both set to 0.1, the association scores $a(e, D)$ used in formulas 4 were set to 1.0. The number of steps in K-Step Random Walk was 5. In total, we submitted 4 runs:

- **qorwkstep**: K-Step Random Walk algorithm,
- **qorw**: Infinite Random Walk algorithm,
- **qorwnewlinks**: Infinite Random Walk algorithm with inter-document and organizational links added ($\mu_e, \mu_D > 0$),
- **feedbackrun**: Infinite Random Walk algorithm with the introduced feedback information.

Further we present the official results and the results of our successive experiments for the two main measures: the Mean Average Precision (MAP) and the Mean Reciprocal Rank (MRR).

We see in Table 1 that all algorithms used to produce the official runs showed similar performance, but the infinite random walk with and without additional links was slightly better. Adding new links did not improve MAP significantly, but MRR got more visible improvement.

However, after the submission and receiving results and relevance judgments from TREC, we continued our experiments. At first, we made an important observation: many candidates mentionings in the corpus are not accompanied with email addresses. If we take the *firstname.lastname* part

of all extracted emails and additionally try to detect candidates in documents using these names (with a space instead of a dot in the middle), we find much more occurrences of the candidates in the collection. This improvement of candidates recognition tremendously influenced the overall performance (see Table 2).

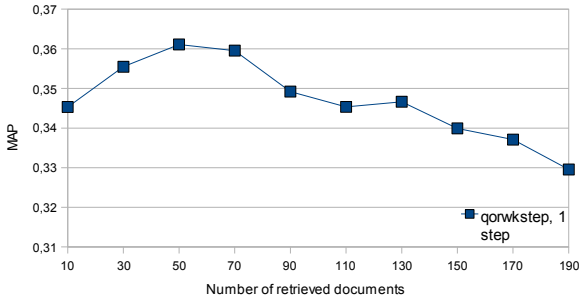

Figure 1: Experiments with different numbers of retrieved documents and one-step relevance propagation

We also discovered that the best number of documents retrieved is much smaller than that we used in our TREC submission and that was optimal for the previous collection. We demonstrate the performance of the baseline method described in Section 1 (which is also the K-Step random walk with 1 step) for the different numbers of retrieved documents in Figure 1. Besides the average increase of the performance due to using full names in candidates recognition, we also see that it is better to retrieve just 50 documents. Such a decrease in performance with retrieval of each next document can be explained by several reasons for the CSIRO collection. At first, this year there are only few experts per query and therefore they are probably so authoritative in the organization that all appear in the top most relevant documents. At second, since the list of candidates is not predefined, we get more and more candidates competing to be ranked higher while retrieving more documents. This means that it becomes much harder to distinguish among them, especially because the expertise of the ones found in many lower-ranked documents is apparently supported by just the amount and not by the quality of expertise evidence.

We repeated all experiments that we made for the TREC submission for all tested methods with 50 retrieved documents and with the additional candidates recognition using full names. We see in Table 2 that all methods appeared to be better than the baseline method, especially ones based on infinite random walk. We also see that the K-Step Random Walk only slightly outperforms the baseline method and adding new links also gives not so much improvement to the Infinite Random Walk method. However, these methods are dependent on more parameters and hence are more sensitive to their proper setup. So, we show further that with the tuned settings, these methods become much more powerful.

We continue by experimenting with different numbers of steps for K-step Random Walk. As we see in Figure 1, the K-Step Random Walk algorithm reaches its maximum performance ($MAP = 0.382$) after making 15 steps. Moreover, after making about 50 steps it starts to show a very stable behavior and does not change its performance ($MAP \approx 0.370$) while a walk gets longer.

	MAP	MRR
qorwkstep, k=1	0.3610	0.5088
qorwkstep, k=5	0.3636	0.5088
qorw	0.3790	0.5328
qorwnewlinks	0.3820	0.5322

Table 2: TREC 2007 Expert search unofficial results with additional candidates recognition by full names and 50 retrieved documents

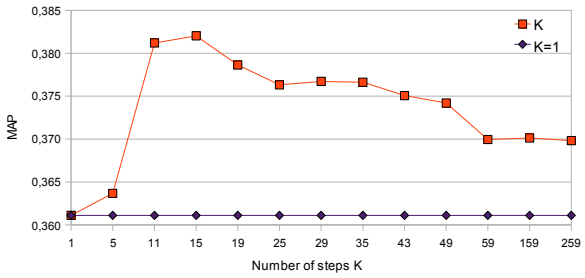

Figure 2: Experiments with different K for K-Step Random Walk

In the follow-up experiments shown in Figure 2 we test different values for μ_e and μ_D. We discover that adding links among documents does not influence the performance of the Infinite Random Walk positively. However, introducing links among candidates shows small, but noticeable improvement with a proper setup of μ_e.

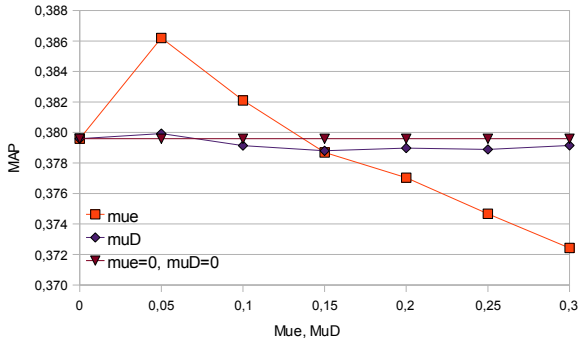

Figure 3: Experiments with values for μ_e and μ_D

The ability to efficiently use the feedback information is very important for a modern IR system. We present a performance of all methods with the same parameter set which we used for TREC submission and with incorporated document relevance feedback in Table 3.

We see that since the task became much easier with the presence of feedback, it is harder to distinguish between methods. However, all of them are slightly better than the baseline and especially the performance of the K-Step Random Walk algorithm and the Infinite Random Walk with additional links.

	MAP	MRR
qorwkstep, k=1	0.4401	0.5720
qorwkstep, k=5	0.4528	0.5840
qorw	0.4409	0.5782
qorwnewlinks	0.4472	0.5838

Table 3: TREC 2007 Expert search unofficial results with candidates recognition by full names and 50 retrieved documents + using relevance feedback

5. CONCLUSIONS

We described a number of approaches for relevance propagation in the documents-candidates network for the expert search task. We used two kinds of discrete Markov processes (random walks) for the modeling. They both appeared to be better than a classic approach with one-step relevance propagation. We also showed how to utilize the additional information: hyperlinks among documents, organizational structure of the enterprise and the user feedback.

In the future, it would be interesting to continue managing relevance flow among documents and candidates. It is probably possible to obtain some candidates priors and hence let a user make random jumps not only to documents. Also, we may allow user not only to jump, but also to walk in the direction of more relevant documents or candidates with a higher prior.

6. REFERENCES

[1] K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 43–50, 2006.

[2] C. S. Campbell, P. P. Maglio, A. Cozzi, and B. Dom. Expertise identification using email communications. In *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*, pages 528–531, New York, NY, USA, 2003. ACM Press.

[3] Y. Cao, J. Liu, S. Bao, and H. Li. Research on expert search at enterprise track of trec 2005. In *Proceedings of 14th Text Retrieval Conference (TREC 2005)*, 2005.

[4] J. X. S. T. Haiqiang Chen, Huawei Shen and X. Cheng. Social Network Structure behind the Mailing Lists: ICT-IIIS at TREC 2006 Expert Finding Track. In *Proceeddings of the 15th Text REtrieval Conference (TREC 2006)*, 2006.

[5] G. Jeh and J. Widom. Simrank: a measure of structural-context similarity. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 538–543, New York, NY, USA, 2002. ACM Press.

[6] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999.

[7] P. Lawrence, B. Sergey, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford University, 1998.

[8] W. Lu, S. Robertson, A. Macfarlane, and H. Zhao. Window-based Enterprise Expert Search. In *Proceeddings of the 15th Text REtrieval Conference (TREC 2006)*, 2006.

[9] C. Macdonald and I. Ounis. Voting for candidates: adapting data fusion techniques for an expert search task. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 387–396, New York, NY, USA, 2006. ACM Press.

[10] M. Richardson and P. Domingos. The intelligent surfer: Probabilistic combination of link and content information in pagerank. In *NIPS*, 2001.

[11] X. Song, B. L. Tseng, C.-Y. Lin, and M.-T. Sun. Personalized recommendation driven by information flow. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 509–516, New York, NY, USA, 2006. ACM Press.