# Term Weighting in Expert Search Task: Analyzing Communication Patterns[*]

Anna Kravchenko and Dmitry Romanov

Higher School of Economics,
Research and Educational Center of Information Management Technologies
Kirpichnaya ul. 33/4, 105679 Moscow, Russia

**Abstract.** The goal of the expert search task is finding knowledgeable persons within the enterprise. In this paper we focus on its distinctions from the other information retrieval tasks. We review the existing approaches and propose a new term weighting scheme which is based on analysis of communication patterns between people.
The effectiveness of the proposed approach is evaluated on a collection of e-mails from an organization of approximately 1500 people. Results show that it is possible to take into account communication structure in the process of term weighting, effectively combining communication-based and document-based approaches to expert finding.

**Keywords:** term weighting, expert finding, expert search task, graph analysis, referral systems

## 1 Introduction

The problem of locating desired information or source of knowlege has been faced by almost everyone, and sometimes finding a right person may be even more valuable than finding the right documents. In large enterprises such as companies and goverment agencies it may become a task of major importance.

For example, according to Yimam-Seid&Kobsa [14] people may search/tend to seek for an expert as a source of information for the following reasons:

- Access to non-documented information. Not all information in organizations can possibly be explicitly documented or information may be deliberately not made publicly available for economic, social and political reasons.
- Specification need. Userw may be unable to formulate a plan to solve a problem, or to pick a direction of research and resorts to seeking experts.

[*] This work was conducted with financial support from the Government of the Russian Federation (Russian Ministry of Science and Education) under contract 13.G25.31.0096 on "Creating high-tech production of cross-platform systems for processing unstructured information based on open source software to improve management innovation in companies in modern Russia".

- Leveraging on others expertise (group efficiency). E.g. finding a piece of information that a relevant expert would know/find with less effort than the seeker or filtering reliable information from the huge mass of information available.
- Interpretation need. Userw may need help understanding the information even if he/she manages to find it from documents.
- Socialization need. Users may prefer the human dimension that is involved in asking an expert as opposed to interacting with documents and computers.

It may be also important to identify users or groups of users who are more knowlegeable than the others. In a big enterprise it may help discover new ideas, the most valuable employees and sources of innovations.

An expert search system is designed to help with these tasks. In contrast with classical information retrieval where documents are retrieved, given a query, the system has to return a ranked list of person names in response.

Most approaches to expert finding tend to copy standard information retrieval methods, such as Vector Space Model [15], representing user profiles as documents. We argue that it may not be the best way, since the need to search for experts mostly emerges in business environments (like corporate mail, for example), which implies a completely different data organization. Authors are usually connected to each other and data mostly consists of messages, not isolated documents.

Existing methods can be divided into communication-based, that take into account the communication patterns between users, and content-based, which focus strictly on the documents content. An essential notion in content-based methods is term significance. It is used to estimate the generality or specificity of a term and evaluate how important is a word to a document, collection or corpus. The process of calculating term significance is called term weighting and the most common approach to term weighting is the tf-idf measure and it's variations [8].

The problem of existing term weighting schemes as well as other content-based methods is that they don't take into account communication patterns between users. They may be combined with communication-based approaches sequentially, but there is no existing method that allows communication structure to affect the term's weight. Developing such method could be a significant contribution to the domain.

In this paper we analyse the specifics of the task, review the existing approaches to expert finding and term weighting and propose a new weighting scheme developed specifically for expertise evaluation.

## 2 Task specifics

As it has been mentioned earlier, the expert search task has different properties from the standart information retrieval task. The standard IR task was developed for analysing web pages and text collections such as online libraries. An expert search task needs to deal with forums, social networks, blogs and corporate e-mails.

We consider the following properties of the expert search task the most important:

- In business environment size of the available corpora is limited by the company lifespan, therefore number of documents is relatively small. Same counts for social networks.
- Corpora may contain information that normal text collections don't, like authorship or communication patterns.
- Communications between authors are not the same as hyperlinks. The possibility to relate every document to its author creates a hierarchical structure, which is absent in the standard IR task.

Naturally, different data organization demands alternative approaches to its processing, that take into account additional features and information they provide. Existing but rare approaches to standart IR such as time marks may also prove useful. Further investigation of the impact of some of this information to extracting data from business corpora is required.

## 3 Existing Approaches

### 3.1 Content-based schemes

The most direct approach to expert finding is adapting Vector Space Model for the task. VSM, one of the most successful solutions to IR, has proved to be highly effective for many types of documents. VSM represents queries and documents as vectors in an n-dimensional information space. Then it systematically compares each document vector with the query vector to find the documents nearest to the query. To adapt VSM for referral systems, each person is represented with a vector based on a "document" consisting of all messages they sent to the user. A person vector contains weights for each term in the query based on how frequently the term was used by the sender in his or her email messages. An example of this approach has been described in [15].

Let $P_i$ be the person vector for colleague $p_i$. This vector contains the computed weights for all terms in the sender's email that also occurs in the query vector. Adaptation of TF-IDF metrics and variations are the most common for calculating the weights.

$$w_i k = \frac{tf_{ik} \log(N/n_k)}{\sqrt{\sum_{t=1}^{n} [(tf_{ik}) \log(N/n_k]^2}} \ . \tag{1}$$

Here, $N$ = number of colleagues of the user, $n_k$ = number of colleagues who have used term $k$, $tf_{ik}$ = term frequency; number of times colleague $p_i$ has used term $k$.

It can be separated into the term frequency factor (tf) and the inverse document frequency factor (idf, collection factor).

The first gives a measure of the importance of the term t within the particular context, document or user profile. The simplest case is the occurrence count of a term in a document. The more a term is encountered in a certain context, the more it contributes to the meaning of the context

Inverse document frequency factor diminishes the weight of terms that occur very frequently in the collection and increases the weight of terms that occur rarely. It is a measure of the general importance of the term (obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient).

Variations of this scheme include term relevance weight (i.e. probabilistic term inverse frequency) and inverse term frequency, a good overview can be found in [8].

The calculation of the query vector, $Q_j$ , can be taken directly from information retrieval methods with computed weights for each term in the query vector $< w_{j1}, w_{j2}, ...w_{jn} >$. Several different query vectors has been proposed, for example, by Salton and Buckley [12]:

$$w_{ik} = [0.5 + \frac{0.5tf_{ik}}{tf_{max}}] \log N/n_k \ . \tag{2}$$

Here $tf_{max}$ = maximum frequency of all terms in query $j$.

After person vectors have been calculated for each sender, the similarity between the query vector and a person vector can be calculated as the cosine of the angle between those vectors. Using this sort of similarity measure, the agent will return to the user a list of potential experts ranked in decreasing order of computed similarity values upon a query. The term weights can be normalized, so the similarity is between 0 and 1. It is convenient to think of this value as the likelihood or confidence that a person will be able to answer the query of the user.

### 3.2 Communication-based schemes

The ranking scheme and it's variations described above don't take into account link structure or communication patterns. A rather more popular approach for expert finding are schemes that focus specifically on those properties, which has proven to be more effective [2]. An overview of graph-based algorithms is given in [4].

One of the most famous is PageRank. In this approach the obtained ranking is the principal eigenvector of the adjacency matrix of the digraph in which edges correspond to links (messages) between users. This approach turns out to be virtually idenctical to the well-known PageRank algorithm for ranking web pages, the only difference being that in PageRank low-weight edges are added between all nodes in both directions.

The simplest is the successor model, where messages are viewed as directed edges pointed from greater to lesser expertise. All the people 'downstream' from a given node are considered of lesser expertise. This one measure of expertise is simply the count of such people (nodes) Another is positional power function (PPF), where the ranks$(ri|i = 1..n)$ satisfy the following system of equations:

$$r_i = \sum_{j \, \in \, S_i} \frac{1}{n}(r_j + 1) \ . \tag{3}$$

An adaptation of HITS algorithm, a precursor to PageRank was also proposed in [2].

Neither of those pays attention to term weighting, focusing only on communication patterns.

## 4 Proposed approach

The goal of our work was is to investigate the impact of communication pattern analysis to expertise evaluation in business corpora and the possibility of using this information during term weighting.

In case of expert finding the main task of term weighting is to divide generic terms from professional vocabulary. Selecting rare terms doesn't always work, as well as the TF-IDF scheme, since rare terms may be generic as well. However, significance of the term may correlate with the patterns of communicaion between people who use it. Generic terms tend to come up randomly in documents of any domain created by authors of all kinds of occupation and level of professional expertise, while professional terminology is mostly used between experts and people asking for a consultation, and experts are normally few and tend to interact with each other. Those people should form a tight cluster, while generic term users would be scattered around the enterprise with relatively small amount of connections.

The main idea of our approach is to calculate the number of clusters that people form for every term in the corpora and to compare it to the number of clusters that would be formed if the links between users were distributed randomly.

The proposed method was analyzed on a collection of e-mails from an organization of approximately 1500 people consisting of 1270000 messages.

## 5 Algorithm realisation

For every term, information about the number of occurrences, number of people using the term and messages id was included in the index. For every message senders and receivers were known.

For every $n = 3..50$ all terms that were used by n people was selected. Then for every term the number of clusters $k$ its users were forming was computed, and finally for every pair $(n, k)$ the number of terms was measured.

Then a random distribution of messages has been modelled.

For every $n = 3..50$ 200 terms were randomly selected from the table, then, according to number of term users $n_1$ and number of term encounters $n_2$, $n_1$

people were randomly selected from the list of users and liked ("sending messages") to $n_2$ any other users on the list. Uniform distribution was used for the probability of creating link with another user. As a result an alternative index was created, then numbers of all $(n, k)$ pairs from this index were computed according to the algorithm described above.

## 6 Results

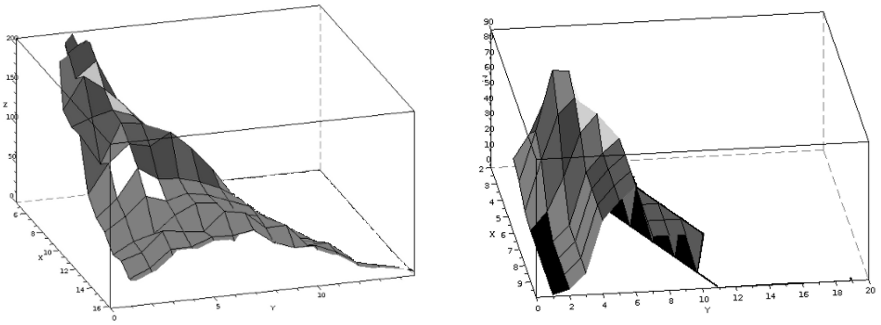

**Fig. 1.** Distribution pattern for the factual communication structure (on the left) and the randomly generated one (on the right).

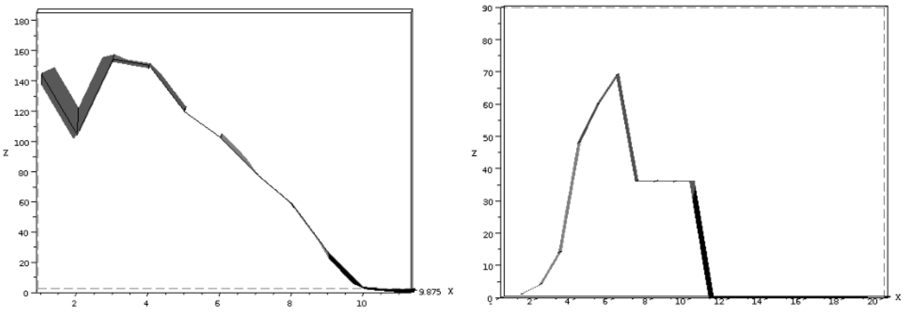

**Fig. 2.** Distribution for n=10. The graph on the right shows randomly generated communications

Fig. 1 and Fig. 2 depict the results of the experiment.

As it can be seen, random distribution differs prominently from the factual one.

The random distribution pictures show that the number of terms, users of which form a single cluster, is vanishingly small. It is highly improbable for such terms to occur by chance.

At the same time, pictures of the factual distribution contain a much higher number of "single cluster" terms, and the distribution graph itself is flatter. Therefore it can be concluded that those terms do not appear at random and may share some common properties.

It also appeared that probability of term being professional instead of generic corresponds with the number of clusters. Terms with smaller k values tend to be more significant.

Taking all these things into account, we find it reasonable to assume that number of clusters can be used as an important indicator of term significance.

This leads to a following weighting scheme:

$$w_{ij} = \frac{p}{p_r} \ . \tag{4}$$

$p = \frac{N_{nk}}{N_n}$, $p_r = \frac{N_{nk,r}}{N_{n,r}}$, where $p$ is the probability of n users forming $k$ clusters for factual distributuion, $p_r$ the probability of it for random distributuion, $N_{nk}$ and $N_{nk,r}$ are the number of terms, users of which form $k$ clusters, in factual and random distribution accordingly, $N_n$ and $N_{n,r}$ are the total number of terms that are used by n people.

If $w_{ij} > 1$ term is considered a professional one and if $w_{ij} < 1$ the term is considered general.

A simpler, though less accurate scheme is

$$w_{ij} = \frac{k}{k_{mostprobable}} \ . \tag{5}$$

Here if the number of users is $n$ and they form $k$ clusters, $k_{mostprobable}$ is the value that has the highest number of $(n, k)$ pairs in random distribution.

Those schemes can be used by itself or in a combination with TF-IDF as an additional factor.

Precise evaluation is currently unachievable due to the small amount of annotated data available - there are no corpora focusing on the term level. The direction of our future work is building a full expert search system based on this scheme. It will allow to evaluate the approach using standart expert search corpora, annotated on text level, such as, for example, the TREC 2005 task.

## 7 Conclusion

In this paper we have shown that it is possible to make analysis of communication patterns a part of a term weighting scheme, and have suggested an example of such scheme. We have also focused on the distinctions between expert finding and standard IR task.

The proposed approach allows to take into account the specifics of the expert search task. It also allows to adapt the existing approaches (such as vector space models, for example) for expert finding more effectively.

For further investigation, we are focusing on proper evaluation of the scheme, combining it with other term weighting schemes. and improving it by exploring the impact of other properties like graph density.

## References

1. Balog, K., De Rijke, M.: Determining Expert Profiles (with an Application to Expert Finding). In IJCAI07: Proc. 20th Intern. Joint Conf. on Artificial Intelligence, pp. 2657–2662 (2007)
2. Campbell, C.S., Maglio, P.P., Cozzi, A., Dom, B.: Expertise Identification Using Email Communications. In Proceedings of the Twelfth International Conference on Information and Knowledge Management, pp. 528–531 (2003)
3. Chen, J., Geyer, W., Dugan, C., Muller, M., Guy, I.: Make New Friends, but Keep the Old: Recommending People on Social Networking Sites. In Proceedings of the 27th International Conference on Human Factors in Computing Systems, pp. 201–210 (2009)
4. Dom B., Eiron I., Cozzi A., Zhang Y.: Graph-Based Ranking Algorithms for E-mail Expertise Analysis. In proceedings of 8th AGM SIGMOD Workshop on Research issues in Data Mining and Knowledge Discovery (2003)
5. Feldman, R., M. Fresko, Y. Kinar, Y. Lindell, O. Liphstat, M. Rajman, Y. Schler, and O. Zamir: Text Mining at the Term Level. Principles of Data Mining and Knowledge Discovery: pp. 6573. (1998)
6. Golbeck, J., and J. Hendler: Reputation Network Analysis for Email Filtering. In Proceedings of the First Conference on Email and Anti-Spam, 44:5458. (2004)
7. Kolari, P., Finin, T., Lyons, K., Yesha, Y.:Expert Search Using Internal Corporate Blogs. In Workshop on Future Challenges in Expertise Retrieval, SIGIR 2008, pp. 2–5. (2008)
8. Lan, M., Sung, S.Y., Low, H.B., Tan. C.L.: A Comparative Study on Term Weighting Schemes for Text Categorization. (2005)
9. Macdonald, C., Ounis, I.:Voting for Candidates: Adapting Data Fusion Techniques for an Expert Search Task. In Proceedings of the 15th ACM International Conference on Information and Knowledge Management, pp. 387–396. (2006)
10. Mesnage, CS, Carman, M.: Piloted Search and Recommendation with Social Tag Cloud-based Navigation. In 1st Workshop On Music Recommendation And Discovery (WOMRAD), ACM RecSys, 2010, Barcelona, Spain. (2010)
11. Petkova, D., Croft, W.B.: Hierarchical Language Models for Expert Finding in Enterprise Corpora. In Tools with Artificial Intelligence, 2006. ICTAI06. 18th IEEE International Conference On, pp. 599608. (2006)
12. Salton. G., Buckley C.: Term Weighting Approaches in Automatic Text Retrieval Gerard Salton. Proceedings. 2005 IEEE International Joint Conference (2005)
13. Serdyukov, P., Rode, H., Hiemstra, D.: Modeling Relevance Propagation for the Expert Search Task. Technical report, TREC 2007 Enterprise Track (2007)Dom
14. Yimam, S., Kobsa, A.: Expert Finding Systems for Organizations: Problem and Domain Analysis and the DEMOIR Approach. (2003)
15. Yu, B., Singh, M.P.: A Multiagent Referral System for Expertise Location. In Working Notes of the AAAI Workshop on Intelligent Information Systems, pp. 66–69 (1999)