A Method of Ontology-aided Expertise Matching for Facilitating Knowledge Exchange

Eduard Babkin, Nikolay Karpov, Tatiana Babkina

National Research University Higher School of Economics,
25/12 Bolshaja Pecherskaja Ulitsa, Nizhny Novgorod 603155, Russia,
{eababkin, nkarpov, tbabkina}@hse.ru

Abstract. The paper proposes a new method for facilitating knowledge exchange by seeking relevant university experts for commenting actual information events arising in the open environment of a modern economical cluster. This method is based on a new mathematical model of ontology concepts matching. We propose to use in the formal core of our method a new modification of Latent Dirichlet allocation. The method and the mathematical model of ontology matching were validated in the form of a software-based solution: the newly designed decision support system titled EXPERTIZE. The system regularly monitors different text sources in the Internet, performs document analysis and provides university employees with critical information about relevant events according a developed matching algorithm. In the proposed solution we made several contributions to the advances of knowledge processing, including: new modifications of topic modeling method suitable for application in expert finding tasks, integration of new algorithms and existing ontology services to show feasibility of the solution.

Keywords: expert finding, natural language processing, topic modeling.

1 Introduction

Emerging and successful growing of new forms of inter-organizational cooperation known as regional, innovation or university clusters [1] in national economies became a significant phenomenon of the modern world-wide socio-economical system. Sustainable exchange of expertise and professional knowledge between stakeholders of innovation clusters plays an important role in knowledge-based economics [2]. For this task an university undoubtedly should be a catalyst which provides expert evaluation and opinions. Critical problems and major strategic choices should be commented, discussed and exposed for multiple stakeholders including industry mass-media and society.

Until now there is no big success of tight integration of university community within the framework of emerging innovative clusters. Informational links are developed by *ad hoc* manner, major activities are implemented inside the stable university-based structures like incubators and business parks. Communication with business experts and mass media shows that in modern turbulent information environments it is the paradigm of information and knowledge exchange which should be modernized. The modernized paradigm of information and knowledge exchange should facilitate reactive or even proactive behavior of university community in response to critical emerging economic or social phenomena in the open environment of innovative cluster-based economy of knowledge.

Traditional analytical methods which provide modern university community with current information about important discussion topics and critical issues lack of comprehensiveness and become too slow. In nowadays practice of universities the best solutions primarily include manual analysis of mass media and internet resources and further slow distribution of information about relevant public events through the inefficient hierarchical organizational structure (from the schools, faculties towards department and employees).

We believe that advanced methods of automated and automatic knowledge management belong to critical scientific foundations of modernization the paradigm of information and knowledge exchange. A specifically designed combination of automated text processing and ontology-based knowledge engineering may improve quality of information analysis and reduce university's response time.

There are many interesting systems which approaches are close to our knowledge exchange idea. The one of it is Media Information Logistics project (Media-ILOG) which is concerns the domain of mass media too. The goal of the Media-ILOG [3], was to improve information flow inside a local newspaper JonkopingsPosten.

In our research we limited the scope of the aforementioned global problem to the key issue of real-time matching between relevant university experts and actual information events arising in the open environment of the economical innovation cluster. We offer a solution of that issue in the form of new automated method of experts finding for facilitating knowledge exchange between the university and heterogeneous community of the innovation cluster.

In contrast to Media-ILOG which is used semantic matching approach proposed by Billig et al. [4] The core of our method is a modification of Latent Dirichlet allocation. [5] It is algorithmically implemented in the newly designed decision support system titled EXPERTIZE. The system regularly monitors different text sources in the Internet, performs document analysis and provide university employees with critical information about relevant events according the specific relevance matching algorithm.

The high level design structure of EXPERTIZE software system includes several principal components. They are Crawler, Data Modeler, Data Store, GUI and Matcher. We match an input document not only with a single expert from our dataset, but with a scientific areas of interest, which is a category of the formal ontology. Each category is represented as a probability distribution of latent topics, so we match distribution of latent topics in the query document with the category using the maximum-likelihood estimation.

In the result of software implementation EXPERTIZE software system has been implemented as a software service. Now it is in an operating state, and regularly collects data from the several information resources available in Internet: library of HSE[9] and Elibrary[10]. Open systems interfaces allow EXPERTIZE get real-time access to the areas of domain interest of the employees of HSE from the InfoPort service [6].

[9] publications.hse.ru

[10] elibrary.ru

A set of practical use cases show that EXPERTIZE properly matches the actual information about discussion topics and information events.

The article has the following structure. After the introduction Section 2 contains related works overview in the information modeling and semantic matching to experts' domains. In Section 3 we observe essentials and formal foundations of our method. Main design decisions and functionality of EXPERTIZE software system are described in Section 4. Section 5 provides the readers with case study of application of that system in a real life information environment. Section 6 concludes the work, giving comparison results of our method and other known approaches and defining open research questions for further investigation.

2 Overview of Relevant Formal Methods for Expert Finding

As soon as our task is to match ontology concepts of expertise with plain text of news it is strongly related to the common expertise retrieval task. The past decade has appeared tremendous interest in expertise retrieval as an emerging subdiscipline. From 2005 the Enterprise Track at the Text REtrieval Conference (TREC) provided a common platform for researchers to empirically assess methods and techniques devised for expert finding [7]. The TREC Enterprise test collections are based on public facing web pages of large knowledge-intensive organizations, such as the World Wide Web Consortium (W3C) and the Commonwealth Scientific and Industrial Research Organisation (CSIRO).

Balog et al. 2012 [8] highlights state of the art models and algorithms relevant to this field. They classified expert finding approaches as follows:

- profile-based model;
- document-based model;
- hybrid model.

A profile-based model for expert finding using information retrieval proposed in Balog and de Rijke [9]. A candidate's skill is represented as a score over documents that are relevant given a knowledge area. The relevance of a document is estimated using standard generative language model techniques.

In the other approach, the method of document-based expert finding does not create a profile for each expert. It uses documents to match candidates to queries. The idea is to first find documents that are relevant to the topic and then locate the experts associated with these documents. The document models are also referred to as query-dependent approaches. Later, Fang and Zhai [10] presented a general probabilistic model for expert finding and showed how the document-based model can be adapted in this schema.

Balog et al. [8] applied this approach to a language model–based framework for expert finding. They also used the profilebased approach in their system and showed that the document-based approach performs better than the profile-based model. Serdyukov and Hiemestra [11] proposed a hybrid model for expert finding which combines both profile- and document-based approaches.

Semantic analysis of texts for expert finding with required competencies proposed by Fomichov on the basis of Formal Concept Analysis [12]. The approach allows to build and compare semantic representations of expert profile using the theory of K-representation and a model of linguistic database.

A topic modeling approach for expert finding proposed by Balog et al. [13]. Instead of modeling candidate profiles or documents, they built a model for each input query and used this model to calculate the probability of candidates given queries. Their approach is similar to the document likelihood method, which is used in language model–based information retrieval. Based on their results, this model underperforms the profile- and document based approaches. The main reason of its poor performance is the sparsity of the models built from the queries. Their definition of topic, however, is different from ours. The term topic in their work refers to query words that users use to search for experts, whereas in the present work we use the term topic as a set of concepts that are extracted from a collection using a topic modeling algorithm. There are multiple known methods for topic modeling of document which are Latent Semantic Analysis (LSA) [14] Latent Dirichlet allocation (LDA) [5] et al.

The topic modeling approach is based on the assumption that words in a document are independent of one another (bag of words) and of their order in the text. Similarly, documents in a corpus **D** are independent of one another and unordered. Distribution of words $\mathbf{W}$ is determined by the set of latent topics $\mathbf{Z}$. Each topic has its own word distribution (phi) and each document has distribution over topics (theta).

Traditional topic-based information retrieval approach is exploited by Wei and Croft, 2006 [15]. The extracted topics are used for information retrieval; whereas the to-be-retrieved documents are used in the retrieval step, i.e., the distribution of topics over words (phi) is used for estimating $P(\mathbf{Q}/\mathbf{Z})$, where $\mathbf{Q}$ – is a set of word in query. The distribution of documents over topics (theta) is used for estimating $P(\mathbf{Z}/\mathbf{D})$.

Another topic-based model is proposed by Momtazi and Naumann [16]. This model outperforms the state-of-the-art profile- and document-based models. To-be-retrieved documents are not used in the retrieval step. Instead, we only use these documents for training LDA, i.e., to be-retrieved documents are used as a corpus to extract topics in an off-line process. Then, in the retrieval step, we only use the distribution of topics over words (phi) for estimating both $P(\mathbf{Q}/\mathbf{Z})$ and $P(e/\mathbf{Z})$ where e – is an expert label.

In a paper [17] the researchers show how to use a topic-based model with scientific ontology, where each document labeled with a category in scientific classification taxonomy **C**. They represent each category *c* as a conditional probabilistic distribution $P(\mathbf{Z}/c)$ which denotes the probability of category *c* being labeled with topic *z*. By utilizing LDA, $P(\mathbf{Z}/c)$ is a $|\mathbf{Z}|$-dimension vector of topic distribution. The main requirement for this approach is to estimate the probability $P(z_k/c)$, which cannot be obtained directly from LDA. However, according to the Bayes formula authors calculate $P(z_k/c)$ by

$$P(z_k/c) = \frac{P(c/z_k)P(z_k)}{\sum_k P(c,z_k)} \tag{1}$$

where $P(c/z_k)$ and $P(z_k)$ can be obtained from LDA. As soon as $\sum_k P(c,z_k)$ is constant for different *c* and $P(z_k)$ is uniform distribution we have

$$P(\mathbf{Z}/c) \propto P(c/\mathbf{Z}) \tag{2}$$

On the basis of explored papers the best way to solve our task is to match between relevant university experts and actual information events using topic-based model, which is proposed by Momtazi and Naumann [16]. Thus, we should implement the model for papers in Russian language concluded in our enterprise dataset. With the help of approach described in [17] this topic-based approach can be applied to use with scientific ontology. To show feasibility of the solution, we archive an integration of new algorithms and existing ontology services.

3 The Essentials and Formal Foundations of the Method Proposed

In our previously designed InfoPort system [6] for solving the expert finding problem we proposed to translate a user-specified query to a corresponding SPARQL query which is evaluated against a specific set of RDF repositories. The query result consisted of a relevant category of scientific classification taxonomies and keywords. The search algorithm of InfoPort system retrieved all persons who labeled with this query.

In the current research our new system EXPERTIZE works automatically: it gets news event as a query and matches it to the most relevant scientist, who can provide expert evaluation and opinions about it. In other word we arrange experts in order to relevance to the event.

On the one hand news events are represented as news in natural language format, thus we have ability to extract semantic information from the text. On the other hand each expert has texts in the form of written papers or records of spoken interviews and tutorials. This material contains rich semantic information about personal interests and abilities.

There are some formal models suitable for implementation of context analysis such as a Distributional Semantic Model (DSM) [18][19] and Latent Semantic Analysis [20][14] and Latent Dirichlet Allocation [5].

In our project we use an extension of Latent Dirichlet Allocation which is a generative formal model that uses latent groups to explain results of observations – data similarity in particular. For instance, if the observations are words in the documents, one can posit that each document is a combination of a small number of topics and that each word in the document is connected with one of the topics. Latent Dirichlet Allocation (LDA) is one of topic-modeling methods and was first introduced by its authors as a graphical model for topic detection.

In our approach by training the LDA model, we form the statistical portrait of its author. A person writing a text has a set of topics in their mind, and each document has a certain distribution of these topics. The author first selects the topic to write on; within this topic, there is a distribution of words that may occur in any document that contains this topic. The next word in the text is generated within the distribution. Then the same procedure is repeated. On each iteration, the author either selects a new topic or continues to use the previous one, and generates the next word within the active topic [5].

The first step of our method for expert finding is a training the model on a collection of texts. We get an estimate of two discrete distribution functions. The following is distribution of probabilities of words $\mathbf{W}$ in topics $\mathbf{Z}$:

$$P(w_i / z_k);\ i \in \overline{1,|\mathbf{W}|}, k \in \overline{1,|\mathbf{Z}|} \tag{3}$$

Distribution of probabilities of topics $\mathbf{Z}$ in documents $\mathbf{D}$:

$$P(z_k / d_n);\ k \in \overline{1,|\mathbf{Z}|}, n \in \overline{1,|\mathbf{D}|} \tag{4}$$

Semantic representation of query news document d_0 can be also calculated using built LDA model. It is distribution of probabilities of topics $\mathbf{Z}$ in documents d_0 – $P(z_k / d_0);\ k \in \overline{1,|K|}$.

In the second step, the extracted topics are used to calculate the probability of query document d_0 given candidates E and categories C. Both E and C represented as a word in the LDA model. Thus, P(Q/E) and P(Q/C) is calculated based on the topics that are distributed over candidate names (E) or scientific domain topics (C).

$$P(d_0 / E) = \sum_{z \in Z} P(d_0 / z, E) P(z / E) \tag{5}$$

$$P(d_0 / C) = \sum_{z \in Z} P(d_0 / z, C) P(z / C) \tag{6}$$

By assuming conditional independence between d_0 and E, C and the document d_0 as equiprobable with other documents we have

$$P(d_0 / z, E) = P(d_0 / z, C) = P(d_0 / z) = \frac{P(z / d_0) P(d_0)}{P(z)} \propto P(z / d_0) \tag{7}$$

Using (2) and (7) from (5) and (6) we get following simple formulas

$$P(d_0 / E) \propto \sum_{z \in Z} P(z / d_0) P(E / z) \tag{8}$$

$$P(d_0 / C) \propto \sum_{z \in Z} P(z / d_0) P(C / z) \tag{9}$$

We rank the categories C from scientific classification taxonomy according to the maximum-likelihood estimation. Most probable categories are chosen and associated with expert.

$$c_{\max} = \arg\max_{c \in C} \left(P(d_0 / c) \right) \tag{10}$$

We perform the same approach to rank experts E from a set of employees of the company.

4. Software Design of EXPERTIZE

The described method for experts finding was practically implemented during design and implementation of the system for matching between relevant university experts and actual information events arising in the open environment of the economical cluster. Such system was called EXPERTIZE. The following services are distinguished in the high-level design of that system (Fig.2):

1. Web Crawler;
2. Data Modeler;

3. Data Store;
4. Graphical User Interface (GUI);
5. Matcher.

EXPETIZE actively uses our InfoPort Service [2]. That semantic service provides in the form of formal ontology factual information about more than three hundred employees of Higher School of Economics (HSE NRU)[11] branch at Nizhny Novgorod. The InfoPort data is represented as RDF triples. Triples include hierarchical information as it originally is in the source. The first level is an alphabetical ordered list of group of scientist, second is a scientist with his personal interests and papers, and third is papers with its features.

The components of the EXPETIZE system can be classified as Online and Offline services. Both are interacted with InfoPort via native REST interface. Offline ones work within monthly period to update information regularly. Online services work on demand, when user activates it by web interface.

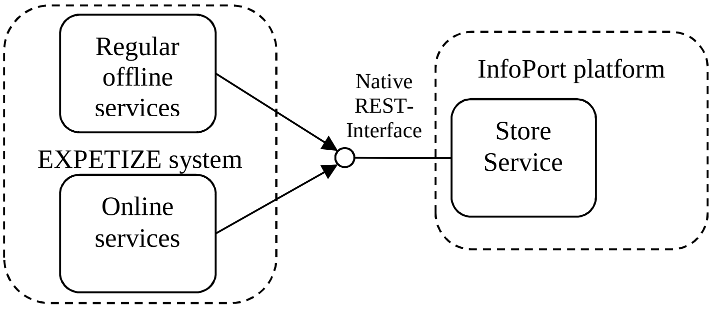


Fig. 1. Interaction of EXPETIZE services with InfoPort platform.

Offline processing begins with crawler Service by scheduler. It makes a request via REST-interface to the InfoPort Store Service to take a list of papers' URI (Uniform Resource Identifiers). As soon as each paper is available online the Crawler gets it by URI and extracts paper's features from page using XML parser. Paper's features include: authors, title, abstract, free keywords, scientific categories of ontology. This information is collected to the Data Store with the help of MySQL[12] base as a Temporal raw data. Implementation of Crawler Service uses Python[13] programming language and Lxml[14] library for HTML processing.

Preprocessing in the Data Modeler service includes the following steps:

- get temporal raw data;
- tokenize the text;
- lemmatize the tokens;
- index the words using the dictionary of lemmas;
- filter out the words that are too frequent (stop words) or too rare (used only once);
- index authors and scientific categories;
- form bag of words using lemmas, authors and categories;

[11] http://www.hse.ru/en/
[12] http://www.mysql.com
[13] http://www.python.org
[14] http://lxml.de

- build LDA model with a given number of topics *K*.

At present time, there are several methods for building LDA models, that is, methods of searching for parameters of all distribution functions in the model. All of the methods are iteration-based and are similar in structure to the Expectation Maximization (EM) algorithm. They are:

- Online Variational Bayes algorithm [21];
- Gibbs Sampling [22];
- Expectation Propagation [23].

Among these algorithms, we use the Online Variational Bayes algorithm as it is the most precise one [21]. It is well implemented in the Gensim[15] toolkit. resource-intensive algorithm.

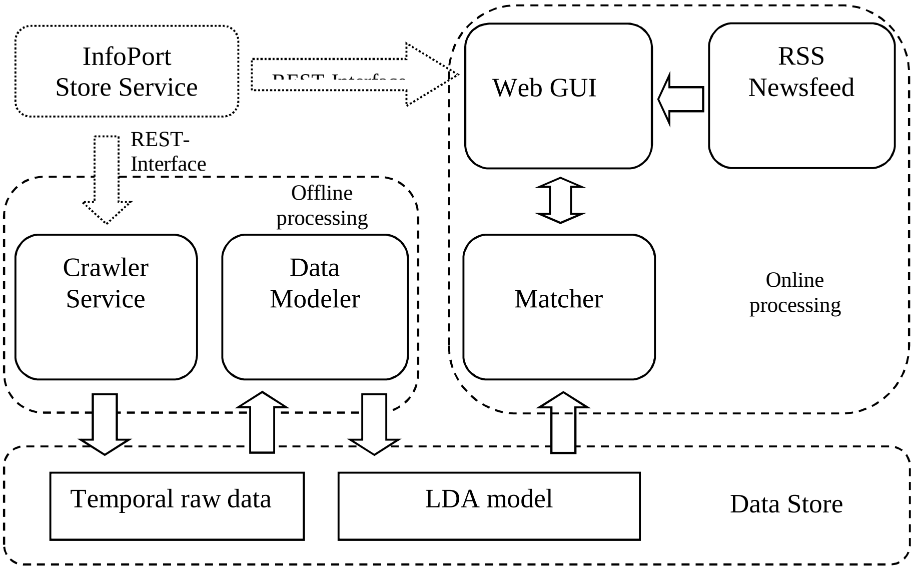


Fig. 2. Principle design of the EXPERTIZE system.

Online processing performs on demand of user by opening Web GUI. Web interface activates RSS Newsfeed, which gets and displays 10 last news from the RSS feed and an empty textbox. User can choose one of its 10 news or paste the text to the textbox manually. When user specify input query the GUI transfer it to the Matcher. In turn, this component performs online semantic search. A semantic representation of event is matched with semantic representations of scientific categories and experts by applying the formula (8) and (9) and selecting top 5 of the units. So, the Mather component returns 5 URIs to the GUI.

To provide user friendly output of the finding result GUI component makes a request to the Infoport Service. It gets features of the selected units: full name, expert's photo URL, expert's department.

[15] http://radimrehurek.com/gensim

The crawled collection includes 4132 units but only 1492 papers are in the Russian language. So, we decide to extend collection with the help of eLibrary[16] scientific database. This is a biggest scientific database in the Russian language. We extracted a part of this base connected only with Information Technologies field. It includes 9127 papers not older than 2011 year.

5 Case study

Evaluation of our proposed method and the EXPERTIZE system was performed empirically. We choose experts from our pool. This pool includes more than three hundreds of professors and researchers of the HSE NRU branch at Nizhny Novgorod[17]. According to an experts field of the study we chose news, which one can be comment by expert and put it to EXPERTIZE. If this expert appears in the list, proposed by the system, we mark such attempt as a successful match.

Let's take a case study. There is an expert Sidorov Dmitry V. whose profile includes a set of scientific domain topics, which he is interested in. There are:

- w_1 - innovation projects,
- w_2 - venture investments,
- w_3 - innovative potential estimation
- etc.

Each scientific domain topics coded as one word and we have pre-created table which is distribution of probabilities of words $\mathbf{W}$ in latent topics $\mathbf{Z}$: $P(w_i / z_k);\ i \in \overline{1,|\mathbf{W}|}, k \in \overline{1,|\mathbf{Z}|}$. It usually has small number of elements higher than zero.

Table.1. Example of probabilities distribution of words $\mathbf{W}$ in latent topics $P(w_1 / z_k); k \in \overline{1,|\mathbf{Z}|}$

	z_1	z_2	…	z_{58}	…	z_{200}
w_1	0	0.04		0.1		0

We find news with title «Yandex company pays for big data»[18], which he can be able to comment as an expert. This news is about investment of Russian IT giant to an Israeli startup company. As each other documents in the collection it can be found probabilities distribution of latent topics z in document. This news goes as an input to the Matcher component where it converts to the probability distribution over latent topics (4) using the pre-built LDA model. The number of topics we set equal to 200.

Table.2. Example of probabilities distribution of topics $\mathbf{Z}$ in documents d_0 – $P(z_k / d_0);\ k \in \overline{1,|K|}$.

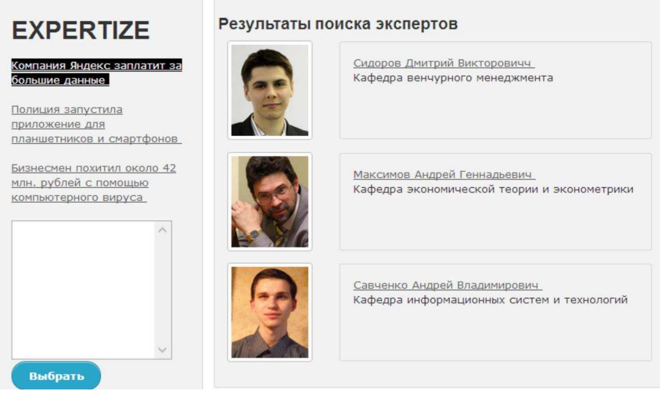
	z_1	z_2	…	z_{58}	…	z_{200}
d_0	0	0.21		0.058		0.034

[16] http://elibrary.ru

[17] http://nnov.hse.ru/en/

[18] http://www.kommersant.ru/doc/2469831

Next, using formula (10) the algorithm chooses each categories ***c*** from scientific classification taxonomy and finds $P(d_0 / C)$. Top 5 of experts who has maximum $P(d_0 / C)$ is shown in the system. A result is presented in Fig. 3.



Компания Яндекс заплатит за большие данные

Полиция запустила приложение для планшетников и смартфонов

Бизнесмен похитил около 42 млн. рублей с помощью компьютерного вируса

Результаты поиска экспертов

Сидоров Дмитрий Викторовичч
Кафедра венчурного менеджмента

Максимов Андрей Геннадьевич
Кафедра экономической теории и эконометрики

Савченко Андрей Владимирович
Кафедра информационных систем и технологий

Fig. 3. Graphical user interface of the EXPERTIZE system.

As soon as our target expert Sidorov Dmitry V. is presented in the output we mark this trial as a successful one. From 100 trials we get 43 successful matches.

Table 3. Experimental results with different model of ontology matching.

Document-based	0.31
Candidate-based	0.22
Topic-based	0.43

We choose the Topic-base approach because in comparison with other approaches (Document-based and Candidate-based) this one gets the best results.

6 Conclusion

In this article we presented a new approach to support rapid exchange of knowledge in innovation clusters based on reactive experts finding. The proposed method of expert finding uses open Internet resources and existing ontological services like InfoPort [6] to get access to the approved skills of potential experts.

During our research we developed a new formal method based on Latent Dirichlet allocation, which includes a software-based solution for matching between relevant university experts and actual information events arising in the open environment of the economical cluster. This solution allows performing real-time matching between Internet news and areas of interest of university employees with further quick notification about possible participation of relevant employees in interviews, informational programs and discussions. In the proposed solution we made several contributions to the advances of knowledge processing, including: new modifications of topic modeling method suitable for application in expert finding tasks, integration of new algorithms and existing ontology services to show feasibility of the solution**.**

A software design of decision support system EXPERTIZE was developed for practical application of the method proposed. The first use cases of the EXPERTIZE system show their relevance and ability to solve the task specified.

Using topic-based model proposed by Momtazi and Naumann [16] we have achieved about 0.43 amount of mean average precision (MAP) on our own queries. The same approach on TREC 2005 and 2006 queries, shows 0.248 and 0.471 amount of MAP respectively [16]. So, precision of EXPERTIZE system is not much less than achieved on TREC 2006. The estimation of recall and f-measure in our EXPERTIZE system less interesting because in general user doesn't need a full set of various experts. One or two most relative experts usually enough for facilitating knowledge exchange.

As soon as we perform expert matching with scientific categories we can apply cross-language expertise retrieval by applying multi-language scientific ontology. It would be our prospective work.

References

1. Asheim, B., Cooke, P., Martin, R.: Clusters and regional development: Critical reflections and explorations. Econ. Geogr. 84, 109–112 (2008).
2. Klimova, N., Litvintseva, M.: Universities Innovation Clusters: Approaches for National Competitiveness Paradigm, (2011).
3. Sandkuhl, K., Öhgren, A., Smirnov, A., Shilov, N., Kashevnik, A.: Ontology construction in practice: Experiences and recommendations from industrial cases. 9th International Conference on Enterprise Information Systems, 12-16, June 2007, Funchal, Madeira–Portugal (2007).
4. Billig, A., Blomqvist, E., Lin, F.: Semantic matching based on enterprise ontologies. On the Move to Meaningful Internet Systems 2007: CoopIS, DOA, ODBASE, GADA, and IS. pp. 1161–1168. Springer (2007).
5. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation, (2003).
6. Babkin, E., Karpov, N., Kozyrev, O.: Towards Creating an Evolvable Semantic Platform for Formation of Research Teams. In: Kobyliński, A. and Sobczak, A. (eds.) Perspectives in Business Informatics Research. pp. 200–213. Springer Berlin Heidelberg (2013).
7. Craswell, N., de Vries, A.P., Soboroff, I.: Overview of the TREC 2005 Enterprise Track. Trec. pp. 199–205 (2005).
8. Balog, K., Fang, Y., de Rijke, M., Serdyukov, P., Si, L., others: Expertise Retrieval, (2012).
9. Balog, K., De Rijke, M.: Determining Expert Profiles (With an Application to Expert Finding). IJCAI. pp. 2657–2662 (2007).
10. Fang, H., Zhai, C.: Probabilistic models for expert finding. Advances in Information Retrieval. pp. 418–430. Springer (2007).
11. Serdyukov, P., Hiemstra, D.: Modeling documents as mixtures of persons for expert finding. Advances in Information Retrieval. pp. 309–320. Springer (2008).

12. Fomichov, V.: Semantics-Oriented Natural Language Processing: Mathematical Models and Algorithms. Springer (2009).
13. Balog, K., Bogers, T., Azzopardi, L., De Rijke, M., Van Den Bosch, A.: Broad expertise retrieval in sparse data environments. Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 551–558. ACM (2007).
14. Dumais, S.T.: Latent semantic analysis. Annu. Rev. Inf. Sci. Technol. 38, 188–230 (2004).
15. Wei, X., Croft, W.B.: LDA-based document models for ad-hoc retrieval. Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 178–185. ACM (2006).
16. Momtazi, S., Naumann, F.: Topic modeling for expert finding using latent Dirichlet allocation, (2013).
17. Zhu, H., Cao, H., Xiong, H., Chen, E., Tian, J.: Towards expert finding by leveraging relevant categories in authority ranking. Proceedings of the 20th ACM international conference on Information and knowledge management. pp. 2221–2224. ACM (2011).
18. Baroni, M., Lenci, A.: Distributional memory: A general framework for corpus-based semantics, (2010).
19. Turney, P.D.: Similarity of semantic relations, (2006).
20. Landauer, T.K., Foltz, P.W., Laham, D.: An introduction to latent semantic analysis. Discourse Process. 25, 259–284 (1998).
21. David M. Blei, Matthew D. Hoffman: Online Learning for Latent Dirichlet Allocation, http://books.nips.cc/papers/files/nips23/NIPS2010_1291.pdf, (2010).
22. Thomas L. Griffiths, Mark Steyvers: Finding scientific topics, (2004).
23. Thomas Minka, John Lafferty: Expectation-propagation for the generative aspect model. Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence. pp. 352–359. Morgan Kaufmann Publishers Inc. (2002).