

# Multidisciplinary Expertise Retrieval with an Application in R&D Management

Choochart Haruechaiyasak      Alisa Kongthon

Human Language Technology Laboratory (HLT)  
National Electronics and Computer Technology Center (NECTEC)  
Thailand Science Park, Klong Luang, Pathumthani 12120, Thailand  
choochart.haruechaiyasak, alisa.kongthon@nectec.or.th

## ABSTRACT

Previous works in Expertise Retrieval (ER) mainly focused on finding people with a specific knowledge within an organization. In this paper, we propose a new challenging task called Multidisciplinary Expertise Retrieval (MULTI-ER). We define the MULTI-ER as a process of finding a group of expert candidates whose combined expertise is required to solve a multidisciplinary R&D problem. The MULTI-ER is different from the ordinary ER in two following ways. Firstly, the problem considered in the MULTI-ER is of a larger scale, and thus, requires multidisciplinary expertise from more than one person. Secondly, the scope of expert finding is not only limited within an organization, but could extend to cover people from different organizations and institutions around the world. As an illustration, a case study on the research subject of Emerging Infectious Diseases (EIDs) is used for a discussion in the context of the proposed MULTI-ER framework.

## Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

## General Terms

Design, Management

## Keywords

Expertise retrieval, expertise identification, expert finding, expert profiling, expertise modeling, R&D management

## 1. INTRODUCTION

Most publicly available search engines sometimes return a long list of search results which do not exactly match the user's query. These search results are Uniform Resource Locators (URLs) which point to some specific Web pages. Therefore, such search engines only perform *document* retrieval task, rather than the actual *Information* Retrieval (IR) as many users expect. One important reason is due to the unstructured and open-domain characteristics of the Web contents. To allow the IR technique become more practical, some domain-specific IR tasks have been proposed.

One of these tasks is *Expertise Retrieval* (ER) which has recently gained increasing attention among researchers in the IR community.

Previous works in the ER can be broadly classified into two groups: *expert finding* and *expert profiling* [3]. The expert finding task aims to identify a list of people who carry some certain knowledge specified by the input query. Typical approach applied for the expert finding is based on the construction of some IR models around expert candidates and topics [2, 7, 8, 9]. The expert profiling, on the other hand, focuses on identifying the area of expertise associated with a given person [4, 6]. To construct an expert profile, two types of information which can be used to describe an expert are *topical* and *social* information. The topical information represents domain and degree of knowledge in which an expert possesses. The social information measures an association aspect among experts such as research project collaboration, publication co-authoring and program committee assignment.

To support the evaluation of the ER task, some related corpora have been proposed during the past few years. The first publicly available corpus is the TREC 2005 Enterprise Track [5]. The corpus consists of various contents, such as Web pages, emails, and source codes, collected by crawling on the World Wide Web Consortium (W3C) Web site. The assigned expert search task is to identify a list of W3C people who are experts for each given topical query. The main drawback of the TREC corpus is the topical queries were directly drawn from the working groups. By using this knowledge, the models which are constructed on documents related the working groups would obviously yield good performance. This makes the TREC corpus less realistic. A more recent corpus is the CSIRO Enterprise Research Collection (CERC) which represents some real-world search activity within an enterprise [1]. The highlight of the CSIRO corpus is the use of internal staffs called *science communicators* to create some topics and perform the judgment.

The previous ER task, however, only focused on finding people with a specific expertise to solve a small-scale problem at the intra-organizational level. In this paper, we propose a new challenging task called *Multidisciplinary Expertise Retrieval* (MULTI-ER). The MULTI-ER is a process which identifies and forms a group of expert candidates whose combined expertise is required to solve a multidisciplinary R&D problem. For example, organizing a research forum to discuss on the global warming issue would require different experts with various knowledge such as scientists

in various fields, sociologists and policy makers.

Compared to the previous ER task, the MULTI-ER has two following differences.

- **Problem scale:** Typical ER task focuses on small and specific problems, e.g., finding programmers who are experts on Java network programming. On the other hand, the MULTI-ER considers problems which are much larger and broader, e.g., global warming, emerging infectious diseases, alternative energy resources and international terrorism. To solve these problems, multiple experts with multidisciplinary expertise are required.
- **Expert scope:** Problems considered for the ordinary ER occur within an organization, and thus, require only the internal employees. The MULTI-ER, on the other hand, considers larger-scale problems which could be interorganizational, national or even global level. Therefore, experts from many different organizations and institutions with various knowledge and expertise may be required to successfully solve the problems.

The MULTI-ER has a potential application in R&D management. Typical tasks in R&D management are, for example, organizing a forum or a meeting to discuss on a certain problem issue and forming a research workgroup to collaborate on a given project. These R&D management tasks are usually performed manually with the following steps. Firstly, the assigned problem is analyzed to identify all related topics. The next step involves searching and obtaining a list of potential candidates who are considered experts in each of the related topics. The last step is the mapping between each related topic and the candidates based on their profiles. All of the above processes require management staffs and managers who are fully trained and highly experienced.

Although the proposed MULTI-ER framework cannot fully replace humans in performing the R&D management tasks, it could provide a decision support function to improve the overall efficiency and effectiveness. Based on the MULTI-ER framework, a system could be implemented to assist and guide users in performing the tasks step by step. Many techniques in the fields such as IR, NLP and machine learning could be applied to make the process more automatic and efficient. For example, given a R&D problem as a query, the system could perform the IR and text mining tasks to automatically retrieve and extract all related keywords associated with the problem. These related keywords could be organized or clustered into a set of topics which is then verified by the users.

In next section, we present the proposed MULTI-ER framework and give a comparative discussion to the ordinary ER. In Section 3, a case study on the research subject of Emerging Infectious Diseases (EIDs) is used for a discussion in the context of the proposed MULTI-ER framework. The conclusion is given in the last section.

## 2. THE MULTI-ER FRAMEWORK

To support the MULTI-ER task, we propose a framework which contains different components to handle all related processes as illustrated in Figure 1. The proposed framework consists of three main components which can be explained in details as follows.

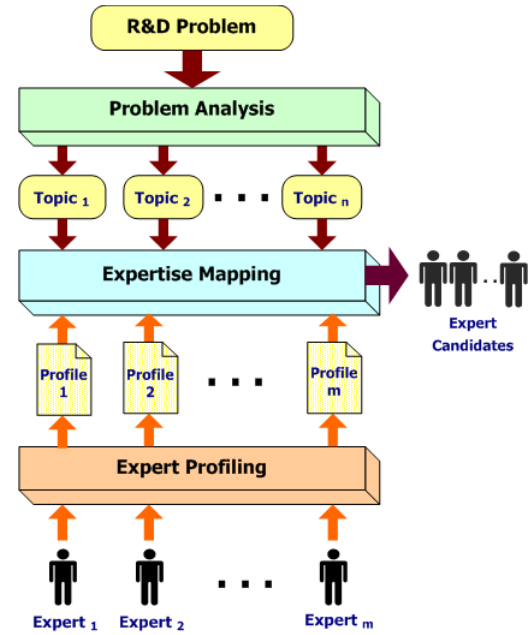


Figure 1: The Proposed MULTI-ER Framework.

### 2.1 Problem Analysis

The main function of the problem analysis is to analyze a given R&D problem and identify a set of  $n$  related topics. To support this process, the IR and text mining techniques could be applied. Search engines can be used to find some relevant documents on the problem. Text mining could then be applied to extract key terms related to the problem. The extracted terms could be clustered to form a set of topics. In the ordinary ER task, the problem analysis does not exist since the problem is very specific and equivalent to a small topic.

One important issue concerning the problem analysis is the information resource for supporting the process. To ensure the maximum topical coverage on a given problem, many resources should be included. Some potential resources include publication and patent databases. Most of these well-organized databases also provide some hierarchical concepts or categories which could be used to form the required set of topics. Many related IR techniques, such as query expansion and citation analysis, could also be applied to help find the relevant topics. The problem analysis is, however, difficult to evaluate since the success on solving a given problem depends on many factors besides the topical coverage. However, a group of initial experts could be asked to verify whether the set of topics meet the requirement to solve a given problem.

### 2.2 Expert Profiling

The expert profiling is a task which has previously been explored in the context of ER. The main goal of the expert profiling is to identify the expertise associated with an expert. The output from the expert profiling is a set of  $m$  profiles describing each of the  $m$  experts. Previous approaches in constructing expert profiles used two types of information, *topical* and *social*, to describe an expert. The topical information represents the knowledge area of an expert. The social information measures the social association

**Table 1: Comparison between the ordinary ER and the proposed MULTI-ER**

Factors	ER	MULTI-ER
Problem scale	Small and specific	Large and multidisciplinary
Expert scope	Within a single organization	Across multiple organizations
Data resources	Intranet and internal DBs	The Internet and outside DBs
Organizational level	Intra-organizational, e.g., W3C and CSIRO	Interorganizational, e.g., UN, WHO and FAO

between an expert to others.

Constructing expert profiles for the MULTI-ER task is different from the previous approaches proposed for the ER task. In the MULTI-ER, the problem scale is larger and the expert scope are broader than those in the ER. Therefore, the set of terms and concepts used to describe the area of expertise is extensively increased. The hierarchical concepts provided with the databases, such as the Library of Congress Classification (LCC) and the International Patent Classification (IPC), could be effectively applied to form the area of expertise.

Another important difference is the supporting information resource used for extracting the expertise. For the previous ER task, the expert scope is limited within an organization. Therefore, the information used to support the profile construction are, for example, Web pages, personal homepages, emails, blogs and Web board messages, which are available on the organization’s intranet. For the proposed MULTI-ER, the information needed for building a profile must be obtained from outside of the organization. Web search engines are perhaps the main tool for gathering the information related to each expert. In addition to providing the topical information, the databases such as publications and patent records are also useful in identifying the social aspect a given expert. The social information can be analyzed and extracted from, for example, co-authoring and co-citation information.

### 2.3 Expertise Mapping

Once the expert profiles are available and all relevant topics to the problem are identified, the expertise mapping performs the matching between the profiles and the topics. The relationship between an expert’s profile and a set of topics are one-to-many, i.e., a person could have more than one area of expertise which can be mapped into multiple topics. The output from this process is a group of candidates whose combined expertise are needed to solve the given problem.

The main design issue in the expertise mapping is the efficient ranking scheme to select an appropriate set of candidate experts to successfully solve a problem. Making the decision to form a group of candidates is not straightforward as it depends on many factors. Some important factors are experience levels of the experts and the success level of the previously performed works.

Table 1 summarizes the differences between the proposed MULTI-ER and the ordinary ER. The comparing factors are problem scale, expert scope, data resources and organization level. As mentioned in the introduction section, two main differences between the MULTI-ER and the ordinary ER are the problem scale and the expert scope. Another differences are the supporting data resources used to construct the models and the organizational level which corresponds to the problems.

Due to the larger problem scale and the broader expert scope, the MULTI-ER requires the access to external DBs and the Internet. To construct an evaluation corpus for the MULTI-ER, the potential resources could be obtained from some international organizations who deals with some multidisciplinary problem issues such as the United Nations (UN), the World Health Organization (WHO) and the Food and Agriculture Organization (FAO).

### 3. A CASE STUDY OF EMERGING INFECTIOUS DISEASES (EIDS)

To better understand the proposed MULTI-ER framework, we present a discussion through a case study of R&D management in Emerging Infectious Diseases (EIDs). The research subject in EIDs has currently received a lot of attentions due to the periodically reports of avian influenza outbreak. This case study aims to explore the possibility of using converging technologies which can cross discipline and contribute to the prevention and management of EIDs that are (and could become) widespread in the APEC (Asia-Pacific Economic Cooperation) region <sup>1</sup>. The resulting technology roadmaps will be recommended to related APEC groups, member economies, and industry for further implementing, especially to develop the technologies. The EIDs problem requires multidisciplinary expertise, ranging from medical science, biotechnology, nanotechnology, material sciences, to information and communication technology. Figure 2 illustrates examples of converging technologies that can help prevent and manage EIDs.

The EIDs case study can be applied in the MULTI-ER context as follows:

- **Problem analysis:** We started analyzing the problem by gathering related information from different sources such as reports from WHO and FAO and research publications from related journals and conferences. Then we did a series of focused interview with an initial group of experts to obtain various kind of information such as the current problems and issues with EIDs and what products and services are needed in order to combat EIDs. We also applied text mining approach to help identify research topics and problems. This method could help discover topics that our initial group of experts might not be familiar with. Using various methods to analyze the problem along with the verification from the group of experts, five research topics related to combating EIDs were identified as:

1. Bioterrorism and Surveillance System,
2. Earth and Climate Observation,
3. Disease Detection,

<sup>1</sup>EID: Roadmapping Converging Technologies for Combat Emerging Infectious Diseases, <http://www.apecforesight.org>

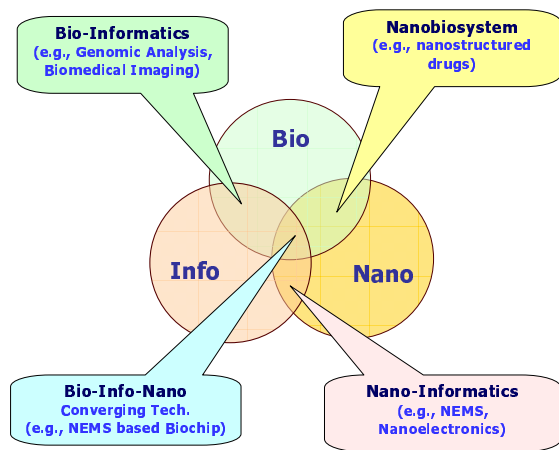


Figure 2: Multidisciplinary expertise for EIDs.

4. Disease Diagnosis,
5. Disease Identification

- **Expert profiling:** After we identified the problems and what products and services are needed in order to combat EIDs, we started identifying technologists and researchers who have expertise in such areas both locally and internationally by looking at their CVs and research papers. We also looked at their research social activities, e.g., publication co-authoring, research project collaboration, research societies they belong to.
- **Expertise mapping:** To map the identified topics with the profile of experts, we conducted three roundtable meetings among the program committee and our target experts to form a consensus.

The above example shows that the process of identifying experts for a large multidisciplinary research project depends significantly on the human experts. This could lead to some disadvantages including incomprehensive topical coverage and biased selection of expert candidates. Thus, the proposed MULTI-ER framework could provide decision support function to assist in making the overall processes more efficient and effective.

#### 4. CONCLUSIONS AND OPEN DISCUSSION ISSUES

We proposed and gave detailed discussion on a framework for a new task called Multidisciplinary Expertise Retrieval (MULTI-ER). Two main differences between the existing ER and the proposed MULTI-ER are the scope of experts and scale of problems to be solved. The MULTI-ER focuses on much larger-scale problems which could be further segmented into smaller topics. These topics are varied and thus require different experts beyond the scope of a single organization.

The MUTLI-ER introduces many new challenging research issues which can be organized into two groups: research on related techniques and development of an evaluation corpus. Some questions which must be considered on each issue are listed as follows.

- **Problem analysis:** What types of information should be considered? How to make sure that all relevant topics are included to solve the problem successfully?
- **Expert profiling:** What types of information are needed to describe the expertise of a person? How to model an expert's experience for profile construction?
- **Expertise mapping:** How to rank the expertise scores on a given topic? Should the social information be weighted more than the topical information?
- **Corpus construction:** Which organization should be considered? How to obtain the information resources to build a real-world corpus to evaluate the framework? How many topic queries should be included?
- **Performance metrics:** What types of performance measures are suitable for evaluating the MULTI-ER framework?
- **User involvement:** The experience and feedback from the users are very useful to develop a successful framework. How to introduce the framework to the people in R&D management? Which process is considered the most important for the users?

We believe that the proposed MULTI-ER is a potential and practical extension to the ordinary ER. The MULTI-ER opens up many interesting research issues which need to be discussed further.

#### 5. REFERENCES

- [1] P. Bailey, N. Craswell, I. Soboroff and A. P. de Vries. "The CSIRO enterprise search test collection," *ACM SIGIR Forum*, 41(2), pp. 42–45, 2007.
- [2] K. Balog, L. Azzopardi and M. de Rijke "Formal models for expert finding in enterprise corpora," In *Proc. of SIGIR 2006*, pp. 43–50, 2006.
- [3] K. Balog, T. Bogers, L. Azzopardi, M. de Rijke and A. van den Bosch. "Broad expertise retrieval in sparse data environments," In *Proc. of SIGIR 2007*, pp. 551–558, 2007.
- [4] K. Balog and M. de Rijke. "Determining expert profiles (with an application to expert finding)," In *Proc. of IJCAI 2007*, pp. 2657–2662, 2007.
- [5] N. Craswell, A. de Vries, and I. Soboroff. "Overview of the TREC-2005 Enterprise Track," In *The Fourteenth Text REtrieval Conf. Proc. (TREC 2005)*, 2006.
- [6] C. Macdonald and I. Ounis. "Voting for candidates: adapting data fusion techniques for an expert search task," In *Proc. of CIKM 2006*, pp. 387–396, 2006.
- [7] D. Mimno and A. McCallum. "Expertise modeling for matching papers with reviewers," In *Proc. of KDD 2007*, pp. 500–509, 2007.
- [8] M. Steyvers, P. Smyth, M. Rosen-Zvi and T. Griffiths. "Probabilistic author-topic models for information discovery," In *Proc. of KDD 2004*, 2004.
- [9] J. Zhang, M. S. Ackerman and L. Adamic. "Expertise networks in online communities: structure and algorithms," In *Proc. of WWW 2007*, pp. 221–230, 2007.