EXPERT FINDING IN DOMAINS WITH UNCLEAR TOPICS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY

BY

GONCA HÜLYA SELÇUK DOĞAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF INFORMATION SYSTEMS

FEBRUARY 2012

EXPERT FINDING IN DOMAINS WITH UNCLEAR TOPICS

Submitted by **Gonca Hülya Selçuk Doğan** in partial fulfillment of the requirements for the degree of **Master of Science in Information Systems, Middle East Technical University** by,

Prof. Dr. Nazife Baykal
Director, **Informatics Institute** _____

Prof. Dr. Yasemin Yardımcı Çetin
Head of Department, **Information Systems** _____

Assist. Prof. Dr. Tuğba Temizel Taşkaya
Supervisor, **Information Systems, METU** _____

Prof. Dr. Adnan Yazıcı
Co-Supervisor, **Computer Engineering, METU** _____

Examining Committee Members:

Assoc. Prof. Dr. Pınar Şenkul
Computer Engineering, METU _____

Assist. Prof. Dr. Tuğba Temizel Taşkaya
Information Systems, METU _____

Prof. Dr. Adnan Yazıcı
Computer Engineering, METU _____

Assist. Prof. Dr. Erhan Eren
Information Systems, METU _____

Assoc. Prof. Dr. Sevgi Özkan
Information Systems, METU _____

Date: 02.03.2012

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name and Surname : Gonca Hülya SELÇUK DOĞAN
Signature :

ABSTRACT

EXPERT FINDING IN DOMAINS WITH UNCLEAR TOPICS

SELÇUK DOĞAN, Gonca Hülya
M.S., Department of Information Systems
Supervisor: Assist. Prof. Dr. Tuğba TAŞKAYA TEMİZEL
Co-Supervisor: Prof. Dr. Adnan YAZICI

February 2012, 151 pages

Expert finding is an Information Retrieval (IR) task that is used to find the needed experts. To find the needed experts is a noticeable problem in many commercial, educational or governmental organizations. It is highly crucial to find the appropriate experts, when seeking referees for a paper submitted to a conference or when looking for a consultant for a software project. It is also important to find the similar experts in case of the absence or the inability of the selected expert. Traditional expert finding methods are modeled based on three components which are a supporting document collection, a list of candidate experts and a set of pre-defined topics. In reality, most of the time pre-defined topics are not available. In this study, we propose an expert finding system which generates a semantic layer between domains and experts using Latent Dirichlet Allocation (LDA). A traditional

expert finding method (voting approach) is used in order to match the domains and the experts as the baseline method. In case similar experts are needed, the system recommends experts matching the qualities of the selected experts. The proposed model is applied to a semi-synthetic data set to prove the concept and it performs better than the baseline method. The proposed model is also applied to the projects of the Technology and Innovation Funding Programs Directorate (TEYDEB) of The Scientific and Technological Research Council of Turkey (TÜBİTAK) as a case study. The experimental results show that our model is satisfiable compared to the baseline method. In our experiments, we use a new ground truth set which is generated based on the choices of three raters by using the Kappa statistics.

Keywords: Expert finding, Similar experts, Voting, Topic generation, Kappa statistics

ÖZ

KONULARIN BELİRSİZ OLDUĞU ALANLARDA UZMAN BULMA

SELÇUK DOĞAN, Gonca Hülya
Yüksek Lisans, Bilişim Sistemleri Bölümü
Tez Yöneticisi: Yard. Doç. Dr. Tuğba TAŞKAYA TEMİZEL
Ortak Tez Yöneticisi: Prof. Dr. Adnan YAZICI

Şubat 2012, 151 sayfa

Uzman bulma, ihtiyaç duyulan uzmanları bulmak için kullanılan bir bilgi elde etme (Information Retrieval - IR) yöntemidir. Birçok ticaret, eğitim veya kamu kuruluşu için ihtiyaç duyulan uzmanın bulunması dikkate değer bir problemdir. Bir konferansa gönderilmiş olan makalenin değerlendirilmesi için hakem aranırken ya da bir yazılım projesi için danışman aranırken uygun uzmanı bulmak son derece önemlidir. Seçilmiş olan uzmana erişilememesi ya da uzmanın müsait olmaması durumunda benzerlerinin bulunması da önemlidir. Geleneksel uzman bulma yöntemleri; destekleyici belge kümesi, uzman aday listesi ve ön tanımlı konular olmak üzere üç bileşen temel alınarak modellenmiştir. Gerçekte çoğu zaman ön tanımlı konular bulunmamaktadır. Bu çalışmada, Latent Dirichlet Allocation (LDA) kullanılarak alanlar ve uzmanlar arasında anlamsal bir katman oluşturan bir uzman bulma sistemi önerilmektedir. Alanları ve uzmanları eşleştirmek için geleneksel bir uzman bulma

yöntemi (oylama yöntemi), temel yöntem olarak kullanılmaktadır. Benzer uzmanlara ihtiyaç duyulduğunda, sistem seçilmiş uzmanların niteliklerini eşleştirerek uzmanlar önerir. Önerilen yöntem, kavram ispatı için yarı sentetik bir veri kümesine uygulanmıştır ve temel yönteme göre daha iyi performans göstermiştir. Önerilen yöntem aynı zamanda örnek olay olarak Türkiye Bilimsel ve Teknolojik Araştırma Kurumu (TÜBİTAK) Teknoloji ve Yenilik Destek Programları Başkanlığı (TEYDEB) projelerine uygulanmıştır. Modelimiz temel yöntem ile karşılaştırıldığında deney sonuçları modelimizin memnun edici olduğunu göstermiştir. Deneylerimizde üç değerlendiricinin tercihlerini temel alan ve Kappa istatistiği kullanılarak oluşturulmuş yeni bir asıl veri kümesi kullanılmıştır.

Anahtar Kelimeler: Uzman bulma, Benzer uzmanlar, Oylama yöntemi, Konu çıkarsama, Kappa istatistiği

To My Family

ACKNOWLEDGMENTS

I am deepfuly grateful to my supervisor Asst. Prof. Dr. Tuğba Taşkaya Temizel and my co-supervisor Prof. Dr. Adnan Yazıcı, who have guided me throughout my research with their invaluable suggestions and criticisms, and encouraged me in this long path. I would like to address my special thanks to Asst. Prof Dr. Erhan Eren, Assoc. Prof. Dr. Sevgi Özkan and Assoc. Prof. Dr. Pınar Şenkul, for their valuable comments and offerings.

I am also grateful to TÜBİTAK for the permission of using the TEYDEB and ARBİS data in this thesis study. I owe my deepest gratitude to Dr. Kıvanç Dinçer, Vice President of TÜBİTAK, for his invaluable contribution on the permission.

I would like to address my thanks to my managers and colleagues at TÜBİTAK BİLGEM UEKAE / Software and Data Engineering Department (G222), especially to the members of İKİS, KAYS, and PİKAY projects, and the members of the Quality Management Department.

I would like to thank Mesut Çeviker for his support on the last steps of my study.

I am deeply thankful to my husband, Serdar. He has made available his support in a number of ways. He supported me with his valuable ideas as a colleague, and more importantly I am grateful for his patience and hearty support.

Finally, I am also deeply thankful to my family, especially my mother Suzan, and my dear sister Sevinç, for their everlasting love, support and encouragement in every day of my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ARBİS	:	Researchers Knowledge Base (Araştırmacı Bilgi Sistemi)
CCS	:	Computing Classification System
IR	:	Information Retrieval
LDA	:	Latent Dirichlet Allocation
LSI	:	Latent Semantic Indexing
MCMC	:	Markov Chain Monte Carlo
PLSA	:	Probabilistic Latent Semantic Analysis
PLSI	:	Probabilistic Latent Semantic Indexing
POC	:	Proof of Concept
RD	:	Research Area Keywords
SME	:	Small and Medium Size Enterprise
SVD	:	Singular Value Decomposition
TC	:	List of Technology Codes
TEYDEB	:	Technology and Innovation Funding Programs Directorate (Teknoloji ve Yenilik Destek Programları Başkanlığı)
TREC	:	Text Retrieval Conference
TÜBİTAK	:	The Scientific and Technological Research Council of Turkey (Türkiye Bilimsel ve Teknolojik Araştırma Kurumu)

CHAPTER 1

INTRODUCTION

World Wide Web is currently used heavily to find information about people as well as any other information. We search for friends, colleagues, and sometimes experts with some specific skills. For instance, when seeking reviewers for papers submitted to a conference or when looking for a consultant for a software project, it becomes critical to find appropriate experts. Therefore "expert finding" recently has become an important task.

The Information Retrieval (IR) systems that meet the "expertise need" are called expert search (expert finding) systems, which can meet the "expertise need" in two ways. The first one is "expertise identification" ("Who are the experts on topic X?") and the second one is "expertise selection" ("What does expert Y know?") [19]. Since the definition of "expert search task" is related to "expertise identification" subject in the Text Retrieval Conference (TREC) in 2005, studies in this area have gained momentum. The traditional expert finding systems with language models usually use two approaches as, "query-dependent" and "query-independent". Query-dependent models rank the associated experts after finding relevant documents for a query. Query-independent models rank candidates according to textual representations of candidates. "Query-dependent" and "query-independent" approaches have also been modeled by using generative probabilistic models and language models [5], [20]. Another traditional expert finding method is introduced by MacDonald & Ounis [18], in which expert finding is handled as a "voting process". Textual representations of candidates are utilized as implicit votes for determining the indirect similarity between topics and experts. Candidate profiles are used in order to calculate the similarity between the documents and topics.

Therefore, relevant topics and experts can be matched. Expert finding as a voting approach reduces dependency on the data set, which generates a more flexible model.

The common point in these studies mentioned above is the assumption that the topics of the corpus are pre-defined. But topics may not present or may not describe the corpus clearly in real-world systems. On the other hand, a topic may be described using different keywords by two different people. Using different words for the same meaning, different experts can be obtained as relevant. In this case, the traditional methods are not satisfiable. In addition finding the most relevant experts related with the domains (projects, papers, subjects, etc.) is not sufficient in some cases as;

- The experts are not interested in the matched domains any more.
- If time is important, the experts' agenda is not available to deal with the selected domain.

Considering the above problems some new questions arise which are the primary motivation for our study. These questions are; "How can we assign experts to documents which do not have any specific topic or have topics but not sufficiently clear or explanatory?" and "Who are the similar experts to the matched ones?". In the literature, many researchers introduced some models to try to find answers for these questions. For example, in the study of Zhang, Tang, Liu, & Li [34], a semantic level is generated as latent topics for finding experts using Probabilistic Latent Semantic Analysis (PLSA) with language models. Language models may not be applied to all of the corpora because of some assumptions and restrictions. In most of the former expert finding studies, it is assumed that an evidence of candidate experts as a name or an email address is present in the supporting documents. This assumption can be run over using the voting approach in expert finding. The other model is done by Balog & de Rijke [4]. Here, the main point of the argument is to be able to suggest alternative experts in case of the unavailability of the previously recommended expert. Another model uses a hidden semantic layer [17] which aims to associate a group of expert with a large scale

multidisciplinary R&D project. The topics and the members of the hidden semantic layer, are generated from the R&D problem using Latent Dirichlet Allocation (LDA).

In this study we propose a new model for expert finding which consists of the following steps:

- Composes the expert finding in a semantic level using the voting approach,
- Proposes an expert finding model based on Latent Dirichlet Allocation (LDA) with Gibbs Sampling,
- Recommends similar experts to the matched ones with domains.

With this proposed model we introduce a new expert finding system. More specifically the contributions of this study are;

- Proposing an expert finding system that generates a semantic level between domains and experts using LDA. The proposed model extracts explanatory and clear topics from domains. Contrary to typical expert finding systems, more than one viewpoint is used, which are generated and pre-defined topics.
- As a result of the proposed model, a ranked list of relevant experts is retrieved. After finding the most relevant experts for domains, the proposed model with the finding similar experts task targets to find the similars of the most relevant experts.
- Our model is applied to a semi-synthetic data set to prove the concept. We get better performance than the baseline method.
- Our proposed model is applied to a real life problem to match experts with the funded projects of Technology and Innovation Funding Programs Directorate (TEYDEB). The experimental results show that our model is satisfiable for this real life application. In our experiments, we use a new ground truth set which is generated based on the choices of three raters by using the Kappa statistics.

In this thesis, we propose an expert search system that combines a baseline approach that retrieves relevant experts by finding similar experts' to those relevant

ones. This thesis is organized as follows. We first review the background works of expert finding and topic modeling in Chapter 2. The baseline expert search approach is presented in Chapter 3. Chapter 4 presents the proposed model and Chapter 5 presents the studies to prove our concept. Experimental design and evaluation are presented in Chapter 6. Chapter 7 provides further experiments and discussions with expert search by finding similar experts. Chapter 8 concludes the thesis and provides possible future research directions based on the thesis work.

CHAPTER 2

BACKGROUND

This chapter puts forth the previous work done in the literature on the expert finding and on the related fields.

2.1. Expert Finding

Expert finding is an approach of Information Retrieval (IR) which meets the "expertise need". Various aspects of expert finding, including expertise identification, "Who are experts on topic X?" and expertise selection, "Who does expert Y know?" is studied by McDonald & Ackerman [19].

Smirnova & Balog [22] tackled expert finding task from different viewpoints:

- Expertise retrieval, which takes a mostly system-centered approach,
- Expertise seeking, which studies related human aspects.

2.1.1. Expertise Retrieval

In expertise retrieval, expert finding system methods which are mostly system-centered approaches are used. From the perspective of the expertise retrieval, expert finding is focused on identifying good topical matches between the expertise need and supporting document collections.

In 2005, an "expert search task" is defined by Text Retrieval Conference (TREC) related to "expertise identification" subject. Furthermore, in each year between 2005 and 2008, an "expert search task" is given by TREC Enterprise track [10], [23], [3], [7]. Each defined task is accelerated the studies about expert finding. The

expert search tasks, defined for Enterprise track, include the following three components [12]:

- A supporting document collection,
- A list of expert candidates (or a task to find expert candidates using their email addresses),
- A set of topics.

As a result of the "expert search task", a ranked list of the expert candidates for a given topic is demanded. The key challenge of the tasks is eliciting the association between a person and an expertise area based on the supporting document collection.

For retrieving and ranking experts on a given topic or user query in the Enterprise track, various methods such as probabilistic or language models [33], [5], [20], [12], graph-based approaches [11] and voting models [18] are used.

Probabilistic and language models may be either query-dependent (also called as document-based) or query-independent (also called as candidate-based) [5], [20]. Query-dependent models rank the associated experts after finding the relevant documents for a given query. Query-independent models rank the candidates according to the textual representations of the candidates also known as the profiles of the candidates. Balog, Azzopardi, & Rijke [5] applied these models to the 2005 edition of the TREC Enterprise track. All evaluation measures of the query-dependent model are higher and the response time of the query-dependent model is also more reasonable than the query-independent model.

A similar approach to language models was proposed in the study of Cao, Liu, Bao, & Li [8] which was referred as two-stage language model. The proposed two-stage language model consists of two parts, relevance model and co-occurrence model. The relevance model shows the relevancy between a document and a query, and the co-occurrence model shows the relevancy between a person and a query. The proposed model can be regarded as a method to develop a query-dependent model.

Petkova & Croft [20] combined the query-independent and the query-dependent models in a model which is referred as the hierarchical model, to provide flexibility in gathering information. The advantages of the hierarchical model are defined as follows,

- While query-independent models concatenate the texts of different document formats explicitly, in the hierarchical model, concatenation is done combining probability distributions.
- Similar to the query-dependent models, the hierarchical model also gathers the information in document collections. But unlikely, the hierarchical model deals with only a subset of the document collection rather than the entire collection.

The hierarchical model is applied to the TREC 2005 Enterprise track. As a result, the model effectively composed evidence for expertise.

In the study of MacDonald & Ounis [18], expert finding task is handled as a "voting process". The textual representations of the candidates (candidate profiles) are evaluated as implicit votes for determining the indirect similarity between topics and the experts. The candidate profiles are constructed based on the similarity between the documents and topics. So the relevant topics and experts can be matched. The voting approach is applied to the TREC 2005 Enterprise track. Eleven data fusion techniques are used in the experiments. The results of the experiments are compared to the median run of all participants of TREC 2005 (MAP 0.1402). According to the comparison, most of the data fusion techniques have increased performance over the median run. As well as the increased performance, another advantage of the voting approach is that it can be easily applied to enterprise data sets without any specific setting.

In expert finding, mostly a relevance score is calculated using the score of relevancy between the query and the different supporting documents of the candidate experts. In the study of Zhang, Tang, Liu, & Li [34], two models were suggested for the calculation of the relevance score:

- Composite model: The model expects that all the terms in a query should occur in each support document in the same order as they are arranged in the query.
- Hybrid model: This model is more flexible than the composite model but this model also expects all the terms in a query should occur in the supporting documents. The order of the terms is trivial.

In most of the studies hybrid model is used while calculating the relevancy scores. Although the hybrid model is more flexible than the composite model, also it has some restrictions. Although the hybrid model does not consider the order of the terms, the model looks for the term itself. The synonyms of the terms or the terms with similar meanings are not covered. A mixture model is proposed in this study which covers the semantic issues as well as the evidences gathered with text mining which is described in detail in the Section 2.3.

2.1.2. Expertise Seeking

Expertise seeking models are interested in how people choose an expert. Several studies have identified the factors that affect decisions of people [21], [16]. Woudstra & Hooff [31] studied on factors related to the quality and accessibility. The factors identified are listed in Table 2-1.

In the study of Hoffman, Balog, Bogers, & de Rijke [16], some of the contextual factors are modeled into an expert search system which also recommends similar experts. The modeled factors are topic of knowledge, organizational structure, media experience, reliability, up-to-dateness, and contacts. For the experiments, a questionnaire and the UvT Expert Collection which is introduced in the study of Balog, Bogers, Azzopardi, de Rijke & van den Bosch [6] is used. The factors, organizational structure, position, media experience, and contacts are identified as significant after the experiments.

Table 2-1: Factors found by the study of Woustra & Hooff

Factor	**Description**
Quality-related factors	
Topic of knowledge	the match between the knowledge of an expert and a given task
Perspective	the expected perspective of the expert, e.g., due to academic background
Reliability	the validity, credibility, or soundness of the expert's knowledge based on the expert's competence
Up-to-dateness	how recent the expert's knowledge is
Accessibility-related factors	
Physical proximity	how close or far away the expert is located
Availability	the time and effort involved in contacting the expert
Approachability	how comfortable the participant feels about approaching the expert
Cognitive effort	the cognitive effort involved in understanding and communicating with the expert and processing the obtained information
Saves time	how much time the participant saves when contacting this expert
Other Factors	
Familiarity	whether and how well the participant knows the expert
Contacts	the relevance of the expert's contacts

Additionally, Smirnova & Balog [22] have focused on the factors that influence people's choice that are time to contact an expert, and the knowledge value gained after. For the experimental evaluation, the UvT Expert Collection is used. In the study, as a baseline method query-dependent model which is defined by Balog, Azzopardi, & Rijke [5] is used. They propose a user-oriented method which models the social network distance between the user and an expert according to organizational hierarchy, geographical location and collaboration. The reported experiments of the study demonstrate considerable improvements over the baseline method.

2.2. Finding Similar Experts

Constructing a similar expert list is another information retrieval task studied in the literature [4]. Here, the main point of the argument is to be able to suggest alternative experts in case of the unavailability of the previously recommended expert. Similarities between experts are calculated taking into account the following aspects:

- Through their collaborations,
- Through documents they are associated with,
- Through discriminative terms they are associated with, the terms with the highest tf-idf values for each document,
- Through vectors of weighted terms where the discriminative terms are weighted using tf-idf values.

The best performance is gained with the similarities between the vectors of weighted terms. In the study, describing an expert with a vector of weighted terms that is extracted from the supporting documents of the expert is identified as the most effective way.

Finding similar expert task of Balog et al. [6] is a content-based approach as an expertise retrieval method. Furthermore, in the study Hoffman et al. [16] the task finding similar experts is improved by including the contextual factors to their model which is an expertise seeking approach. Content-based finding similar experts approach is compared to the improved method in which contextual factors are included. In most of the evaluation measures, the improved method shows higher performance.

2.3. Expert Finding through Topic Modeling

Traditional methods used for "expertise retrieval" usually take into consideration the relevancy between the queries and the supporting documents of the candidate experts according to the occurrences of the query terms in the supporting documents. Traditional methods lack the ability of determining the semantic knowledge.

According to Fang & ChengXiang [12], "a set of topics" is a requirement for an expert finding task. However, in some cases of expert finding the topics are predefined or the predefined topics do not describe the corpus clearly. For instance, the predefined topics cannot handle polysemy (words with multiple meanings) and synonymy (multiple words with similar meanings). In a corpus related with both Music and Sports areas, the word "play" could be related with both of the areas. In such cases, topics can be generated from the corpus.

Topics can be extracted from a corpus using various algorithms as Latent Semantic Indexing (LSI), Probabilistic Latent Semantic Indexing (pLSI), and Latent Dirichlet Allocation (LDA).

Kongthon et al. [17] describe LSI, pLSI and LDA as follows:

> "LSI uses singular value decomposition (SVD) to reduce high-dimensional term-by-document matrix to a lower dimensional representation called latent semantic space. LSI cannot capture some aspects of polysemy and synonymy because SVD is actually designed for normally-distributed data. pLSI approach which models each word in a document as a sample from a mixture model, where the mixture components are multinomial random variables that can be viewed as representations of topics. However, the pLSI model encounters overfitting problem because the number of parameters grows linearly with the number of documents. LDA is then introduced to correct such problem. The basic idea behind LDA is that documents are represented as random mixtures over latent topics, where each topic is represented by a probability distribution over words."

LDA is used in many studies [15], [30]. In the study of Griffiths & Steyvers [15], a generative model based on LDA is used. They defined the problem as considering the posterior distribution over the assignments of words w to topics z, $P(w|z)$. Because, it requires computing a probability distribution over a large discrete state space as in Bayesian statistics and statistical physics. This problem is addressed by using Monte Carlo procedure that requires little memory, and competitive in speed and performance with existing algorithms. They used Markov chain Monte Carlo

while sampling from the target distribution. In Markov chain Monte Carlo, a chain is constructed to converge to the target distribution, and samples are then taken from that Markov chain [15].

In Griffiths & Steyvers [15] and Steyvers & Griffiths [24], a Markov Chain Monte Carlo (MCMC) algorithm, Gibbs Sampling is presented for extracting topics. Gibbs Sampling (Alternating Conditional Sampling) is a specific form of MCMC. The next state of Gibbs Sampling algorithm is estimated from sampling all variables conditioned on current values of other variables.

In the study of Griffiths & Steyvers [15], Gibbs sampling algorithm is compared with variational Bayes and expectation propagation. For the comparison, a dataset consisted of a set of 2000 images; each containing 25 pixels in a 5x5 grid is used. All of the 3 algorithms are run using the same initial conditions for 4 times. Variational Bayes and expectation propagation algorithms are run until convergence, and Gibbs sampling algorithm is run for 1000 iterations. During the runs, the number of floating point operations is tracked for calculating the perplexity. Perplexity is a standard measure for evaluating the performance of the statistical models of natural language which indicates the uncertainty in predicting a single word. As a result, although all 3 algorithms captured the underlying topics, Gibbs sampling algorithm is executed more rapidly than either variational Bayes or expectation propagation.

A mixture model, an expert finding approach with topic modeling, is proposed by Zhang, Tang, Liu, & Li [34] which determines the semantic knowledge. The mixture model is represented as the hidden semantic layer between the terms and the supporting documents using Probabilistic Latent Semantic Analysis (PLSA). Using the mixture model, Zhang et al. [34] do not model the queries and support documents directly, but each hidden theme layer is associated with the queries and the supporting documents. The mixture model is evaluated in a real-world system, ArnetMiner. The mixture model is compared with the traditional language models. Results of the experiments showed that the proposed method is performed better

than the traditional language models. The study of Zhang et al. [34] is the formalization of the expert finding problem in a semantic-level using PLSA.

A hidden semantic layer is used in the study of Kongthon, Haruechaiyasak, & Thaiprayoon [17]. The aim of this study is associating a group of expert with a large scale multidisciplinary R&D project. The topics and the members of the hidden semantic layer, are generated from the R&D problem using LDA. On the other hand, the profiles of experts are generated using the knowledge areas, the social information including CV, personal homepage, research papers and the social associations between experts. Finally, the relevancy between the generated profiles and the topics are identified. The proposed method is presented through a case study of Emerging Infectious Diseases R&D problem. The problem is firstly analyzed with various keywords using the Compendex database to gather research publications. From the related research publications, latent topics are generated using LDA. Related experts with the generated topics are illustrated to the users' of the system. According to the study of Griffiths & Steyvers [15], LDA with the inference method Gibbs sampling has better performance than the PLSA which is used in the firstly formalized expert finding method in semantic-level.

CHAPTER 3

BASELINE APPROACH: EXPERT SEARCH AS A VOTING PROBLEM

This chapter explains the baseline approach of our proposed system. The study of MacDonald & Ounis [18] is used as the baseline approach. In this chapter the method, calculations and experiments are presented. Equations, figures and tables are cited from the study of MacDonald & Ounis [18].

3.1. Overview of "Expert Search as a Voting Problem"

When the knowledge level of an organization increases, finding experienced people in a specific area and bringing out the overlapping areas of interest among these people become difficult.

When we need information about an issue, besides searching the related documents, we usually need to consult people who have knowledge on that issue. Because, by using this way, we can access the information easier and faster than searching it in the related documents. This saves time and the time is the one of most valuable things in business life today. Yimam-seid & Kobsa have identified possible scenarios when people may seek an expert as a source of information to complement other sources as documents, and databases [32]. These are:

- *Access to non-documented information:* All the information in organizations cannot be completely documented.
- *Specification need:* Problems that require the information may not be defined specifically.

- *Leveraging on other's expertise (group efficiency):* Finding the appropriate person can solve the problem with less effort.
- *Interpretation need:* Deriving implications or interpretations of the information can be easier.
- *Socialization need:* Rather than interacting with documents, users prefer the human dimension.

According to the scenarios listed above, an expert search system will be of benefit. The expert search system provides a list of related candidates based on user queries. In order to return the related candidate list;

- Candidate experts list,
- Textual evidences for creating candidate profiles and mapping the user queries with subjects

should be obtained.

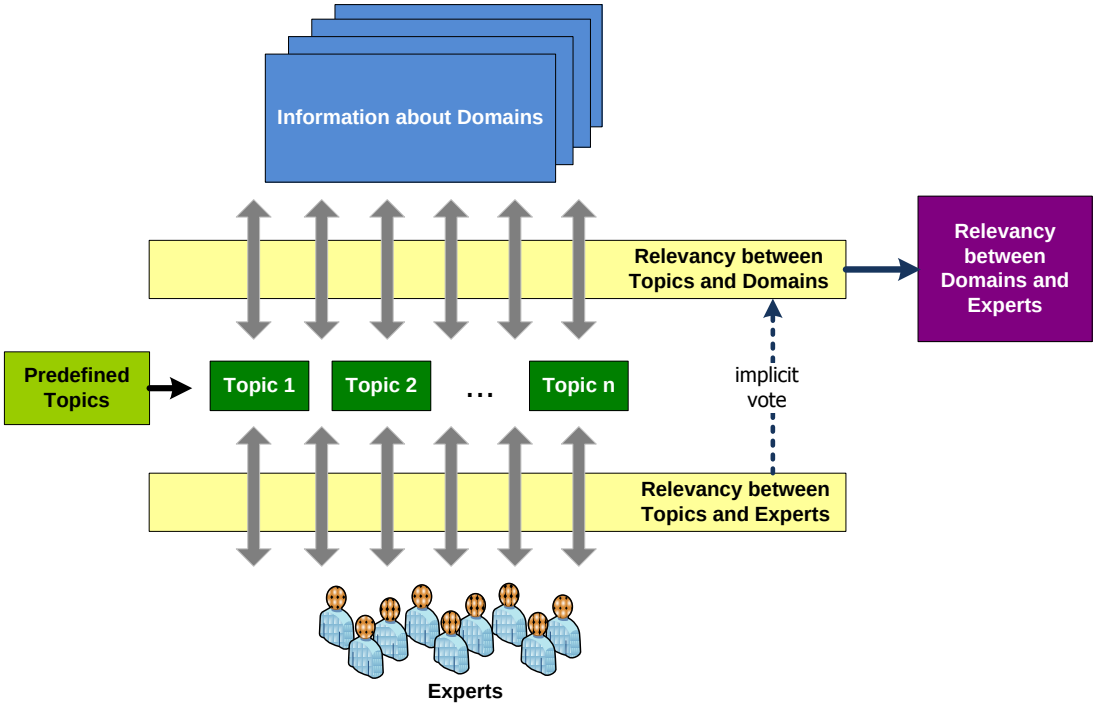


Figure 3-1: Overview of the baseline approach

In the studies of MacDonald & Ounis [18], the ranking of documents according to the query results and locating this document in a candidate profile for any retrieved

document, are evaluated as an implicit vote. In this study, various data fusion techniques have been used to aggregate the document ranks with votes in the profiles. The overview of the method is shown in the Figure 3-1.

The reasons for choosing this study as baseline are listed below;

- Set of documents retrieved for a query and set of documents belonging to the profile of a candidate are combined by data fusion techniques in the baseline approach. With the data fusion techniques we can compose the set of documents and the set of scientists retrieved for a query.
- Data fusion techniques can be used for combining multiple sources. As well as topics, other textual representations can be integrated with data fusion techniques.

In the baseline approach, expert search is considered as a voting process. As a result of an expert search query, a rank of retrieved documents is obtained. The profiles of candidates consist of relevant documents with the candidate. As a voting process, documents in a candidate's profile are considered as an implicit vote to the rank of retrieved documents for a query.

The problem is defined as "How are the votes for each candidate incorporated to produce the final ranking of experts?". Data fusion techniques are utilized to combine the votes for the candidates.

Two main classes of data fusion techniques are used in the study:

- Combining rankings using the ranks of the retrieved documents.
- Combining rankings using the scores of the retrieved documents.

The candidate profiles consist of the related documents in the corpus. The relationship between the candidates and the documents is established by searching the names and e-mails of candidates in the documents. The user queries are executed within the documents located in the corpus and a ranked relevant document list is retrieved at the end of a query. For instance, as it can be seen in the Figure 3-2; if the name or e-mail of a candidate C_1 exists in the documents

"D_a, D_d, D_e", the profile of this candidate will be created as $profile(C_1): \{D_a, D_d, D_e\}$. Then if we consider the document list retrieved by Query1 as $R(Q_1): \{D_a, D_b, D_c, D_d\}$ and if we give 1 vote to every document in a candidate profile; the results in the $R(Q_1)$ list would get 2 votes for the candidate C_1.

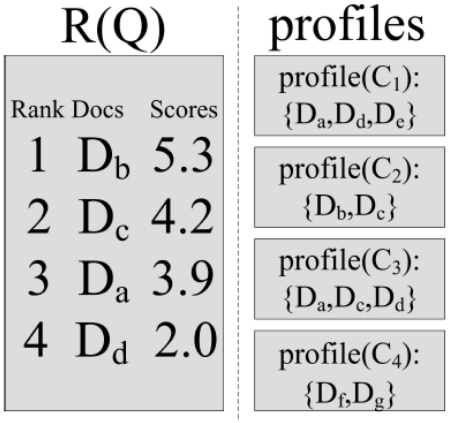


Figure 3-2: A simple example from expert search

In order to find the score $score\_cand(C, Q)$ corresponding to a query, the voting operation is applied in three different ways as listed below;

- The number of retrieved documents voting for each candidate
- The scores of the retrieved documents voting for each candidate
- The ranks of the retrieved documents voting for each candidate

The data fusion techniques used in referred study are given in Figure 3-3.

For implementing the application, TREC 2005 Enterprise track data set is used. This data set is indexed using the Terrier library. The stopwords are removed and a weak stemming algorithm that applies only the first two steps of Porter's stemming algorithm is used.

Name	Relevance score of candidate is:
Votes	‖D(C,Q)‖
RR	sum of inverse of ranks of docs in $D(C,Q)$
BordaFuse	sum of ($\|R(Q)\|$ - ranks of docs in $D(C,Q)$)
CombMED	median of scores of docs in $D(C,Q)$
CombMIN	minimum of scores of docs in $D(C,Q)$
CombMAX	maximum of scores of docs in $D(C,Q)$
CombSUM	sum of scores of docs in $D(C,Q)$
CombANZ	CombSUM $\div \|D(C,Q)\|$
CombMNZ	$\|D(C,Q)\| \times$ CombSUM
expCombSUM	sum of exp of scores of docs in $D(C,Q)$
expCombANZ	expCombSUM $\div \|D(C,Q)\|$
expCombMNZ	$\|D(C,Q)\| \times$ expCombSUM

Figure 3-3: Summary of expert search data fusion techniques

At the end of the study the results in the Figure 3-4 are obtained. When the results are compared to the median run of all participants of TREC 2005 (MAP 0.1402), the most data fusion techniques show higher performance than this median run. The rank based techniques, RR and BordaFuse, give the best results in three weighting models. When we look at the score-based techniques, we see that the results are different in terms of performance. CombMNZ and CombSUM techniques and their exponential variants are discussed as the strongest in the score-based techniques.

	BM25			PL2			DLH13		
Fusion	MAP	ΔMAP	P@10	MAP	ΔMAP	P@10	MAP	ΔMAP	P@10
Votes	0.1691	(+21%)	0.3180	0.1661	(+18%)	0.3100	0.1650	(+17%)	0.3080
RR	0.1940$^{>}$	(+39%)	**0.3560**	0.1758	(+25%)	0.3120	0.1849$^{>}$	(+32%)	**0.3500**
BordaFuse	0.1774	(+27%)	0.3360	0.1691	(+21%)	0.3160	0.1738$^{>}$	(+24%)	0.3280
CombANZ	0.0316$^{\ll}$	(-77%)	0.0380	0.0344$^{\ll}$	(-75%)	0.0420	0.0313$^{\ll}$	(-78%)	0.0240
CombMED	0.1055$^{<}$	(-25%)	0.1900	0.1022$^{\ll}$	(-27%)	0.1720	0.1089$^{\ll}$	(-22%)	0.1880
CombMIN	0.0654$^{\ll}$	(-53%)	0.1380	0.0637$^{\ll}$	(-55%)	0.1380	0.0728$^{\ll}$	(-48%)	0.1500
CombMAX	0.1756	(+25%)	0.3120	0.1630	(+16%)	0.2960	0.1632	(+16%)	0.3080
CombSUM	0.1769	(+26%)	0.3280	0.1736	(+26%)	**0.3240**	0.1743$^{>}$	(+24%)	0.3180
CombMNZ	0.1747	(+25%)	0.3280	0.1733	(+25%)	0.3220	0.1715	(+22%)	0.3220
expCombANZ	0.0333$^{\ll}$	(-76%)	0.0340	0.0300$^{\ll}$	(-79%)	0.0380	0.0333$^{\ll}$	(-76%)	0.0420
expCombSUM	**0.1980**$^{>}$	(+41%)	0.3420	0.1757$^{>}$	(+25%)	0.3120	0.1792$^{>}$	(+28%)	0.3380
expCombMNZ	0.1970$^{>}$	(+40%)	0.3420	**0.1816**$^{>}$	(+30%)	0.3220	**0.1873**$^{\gg}$	(+34%)	0.3440

Figure 3-4: Performance of the 11 data fusion techniques for expert search

3.2. Experimental Settings

While evaluating our baseline approach, we tried to employ the same conditions as in the study of the baseline approach. So we used the same data set, libraries and measures for the experiments. Differently, we only used a restricted set of data fusion techniques. By analyzing the results of the baseline approach, we chose two different data fusion techniques with the highest relevance scores, which are "Reciprocal Rank" based on ranks and "ExpCombMNZ" based on scores. Although, we chose two techniques, we also decided to utilize "CombMNZ" technique to evaluate the results with "ExpCombMNZ".

Data Set

TREC 2005 Enterprise track data set is used for the experiments. This test collection consists of 331,037 documents collected from the World Wide Web Consortium (W3C) website in 2005 [10].

Type	Scope	Size (GB)	Docs	avdocsize (KB)
Email	lists	1.855	198,394	9.8
Code	dev	2.578	62,509	43.2
Web	www	1.043	45,975	23.8
Wiki web	esw	0.181	19,605	9.7
Misc	*other*	0.047	3,538	14.1
Web	people	0.003	1,016	3.6
	all	5.7	331,037	18.1

Figure 3-5: Details of W3C Corpus

The W3C test collection includes 1,092 candidate experts. 50 topics that were published for the expert search task of TREC 2005 Enterprise track are used in the experiments.

Indexing and Retrieving

Data set is indexed using Terrier [29] which is developed by the School of Computing Science, University of Glasgow. During indexing, each document is defined as content, title and incoming text of hyperlinks. For stop words removal process, the default stop words list of Terrier is used. A weak stemming algorithm, that performs only the first two steps of Porter's stemming algorithm, is used to increase the precision values.

Data Fusion Techniques

Reciprocal Rank (RR) is adapted to the expert search task. In RR technique, the combined ranking is determined by the sum of reciprocal rank of the documents that are both present in the result set of a query and the profile of a candidate. Adapting from Reciprocal Rank, the score of a candidate's expertise is defined as:

$$score\_cand_{RR}(C,Q) = \sum_{d \in R(Q) \cap profile(C)} \frac{1}{rank_d} \qquad \text{(Equation 3-1)}$$

where $rank_d$ is the rank of document d in the retrieved result set of query $R(Q)$.

In CombMNZ, the combined score is determined by multiplying the number of documents from the profile of a candidate that are in the result set of the query, $R(Q)$, with the sum of these documents scores.

$$score\_cand_{CombMNZ}(C,Q) = \|R(Q) \cap profile(C)\| \sum_{d \in R(Q) \cap profile(C)} score_d \qquad \text{(Equation 3-2)}$$

ExpCombMNZ is a variant of CombMNZ technique. The score of documents are transformed by applying the exponential function (e^{score}). Applying the exponential

function, the distance between the scores of the high-scored and low-scored documents is amplified.

$$score\_cand_{expCombMNZ}(C,Q) = \|R(Q) \cap profile(C)\| \sum_{d \in R(Q) \cap profile(C)} e^{score_d} \quad \text{(Equation 3-3)}$$

Evaluation Measures

Mean Average Precision (MAP) and Precision@10 (P@10) measures are used to evaluate the retrieval performance. MAP is used to assess the overall quality of the ranking and P@10 is used to assess the accuracy of the top-ranked candidates retrieved by the system.

3.3. Experimental Results

By implementing the baseline approach, we get the results which are listed in Table 3-1.

Table 3-1: Experimental Results

	BM25		**PL2**		**DLH13**	
Fusion Technique	**MAP**	**P@10**	**MAP**	**P@10**	**MAP**	**P@10**
RR	0.1345	0.2000	0.1249	0.1552	0.1300	0.2104
CombMNZ	0.1228	0.1760	0.1202	0.2063	0.1245	0.2021
expCombMNZ	0.1442	0.2049	0.1381	0.2292	0.1460	0.2396

When we compare the results given in the paper of MacDonald & Ounis with the results we obtained, we have seen that the results are distributed as expected. However the results we obtained are 30% less than the results given in paper. the possible reasons for lower MAP and P@10 values are listed below:

- In the study of MacDonald & Ounis [18], since no API settings are described in the paper, we cannot be sure that we have the same experimental setup as they have
- The candidate profiles may be set using a different way. For the candidate profiles, we search the name and email alias of the candidates in the documents.
    - The email addresses can be designed basically as follows, "<local part>@<domain name>.<alias>". In the Enterprise track, because of the restrictions of Terrier we search for only the "<local part>" of the email addresses.
    - Terrier also does not index the characters like a dot, so if the local part contains a dot, the email address cannot be retrieved.

Although our results are 30% less than the results obtained by MacDonald & Ounis [18], the data fusion technique, expCombMNZ, performed better than the median run of all participants of TREC 2005 (MAP 0.1402).

CHAPTER 4

EXPERT FINDING IN DOMAINS WITH UNCLEAR TOPICS

4.1. Overview of Proposed System

As mentioned in the Chapter 2, we already explained that most of the expert finding systems are required to have the following items [12]:

- A supporting document collection,
- A list of expert candidates (generally names or email addresses),
- A set of topics.

A document collection with well-defined set of categories is not available at all times. Even some collections have topics; these topics may be insufficient to represent the entire collection. So we limited our problem as follows:

How can we assign experts to the documents which do not have any topics or have topics but not sufficiently clear or explanatory?

We propose an expert finding system with the following capabilities:

- Covering the basic requirement expert finding as matching projects, papers or documents with experts.
- Implementation of an expert finding in a semantic level.
- Proposal of an expert finding model based on Latent Dirichlet Allocation (LDA) with Gibbs Sampling.

In this thesis, the items to be matched with the candidate experts are referred as "domains". A domain could be a document, a project, or a paper to be matched with the candidate experts.

In the study of Zhang et al. [34], a hidden 'semantic' theme layer between a topic and the document collection of experts is modeled. In our problem, the hidden 'semantic' theme layer is modeled between the domains and the candidate experts. Figure 4-1 shows the overview of the proposed model, *Matching Experts and Domains with Unclear Topics*. The numbers, *(1)*, *(2)*, *(3)* and *(4)* shows the execution sequence of the processes.

In our model:

- "Document collections" are the supporting explanatory documents, keywords, and other information about experts or domains.
- "Corpus" is used to describe all of the document collections which are related with domains or experts.
- The topics extracted from the document collection of the domains are defined as "generated topics".
- Topics which are already defined for the corpus are defined as "pre-defined topics".
- "Topic" is used to describe all of the topics, generated topics and pre-defines topics.

The process "(1) Extracting Topics" is the first process of the model which generates the hidden semantic theme layer between domains and experts. The hidden semantic theme layer contains one or more topic sets, and a topic set may contain one or more topics.

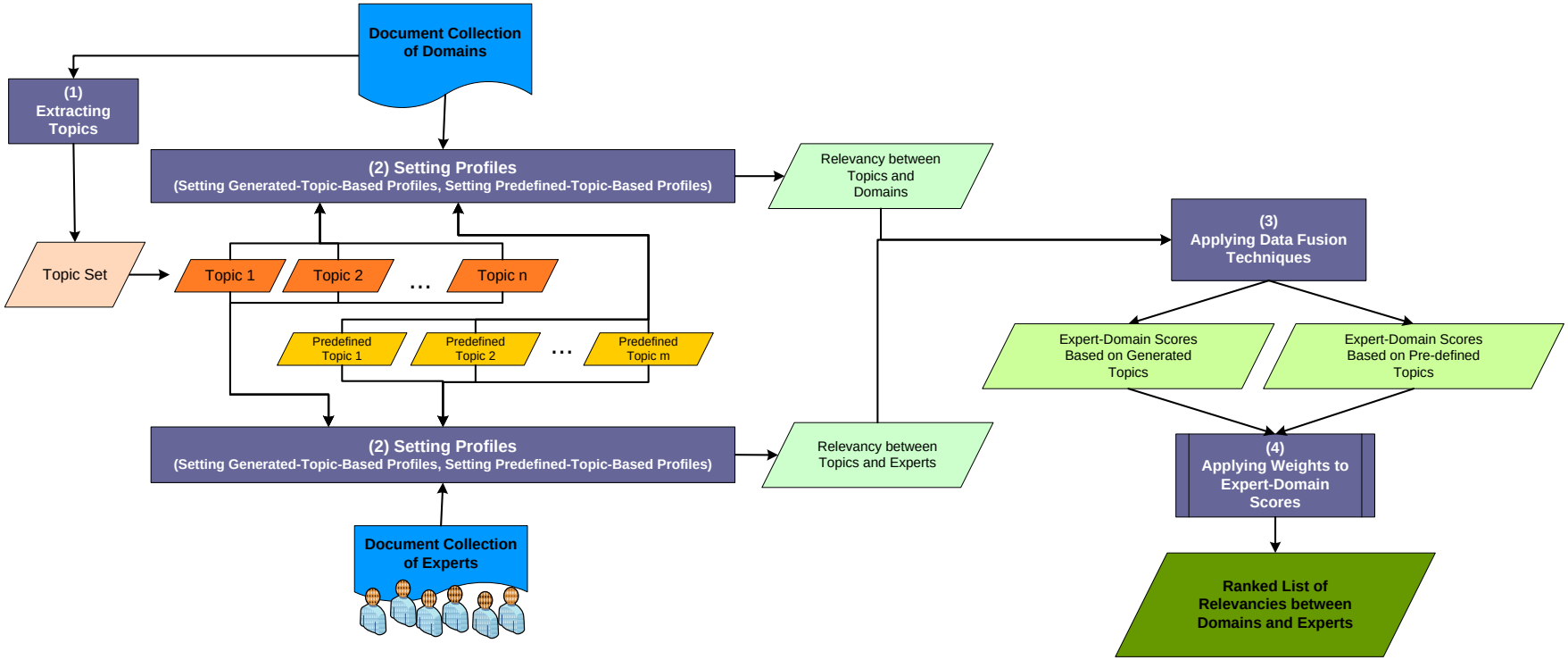


Figure 4-1: Overview of the proposed model

The generated topics, t_g and the predefined topics, t_p are the inputs of the process "(2) Setting Profiles". The process is performed for these two types of the topics. "Setting Generated-Topic-Based Profiles" and "Setting Predefined-Topic-Based Profiles" are the sub-process of the process. Setting profiles process is also performed for the domains and the experts in parallel. The outputs of the parallel processes are the profiles of the domains (relevancies between topics and domains), and the profiles of the experts (relevancies between topics and experts). The outputs of the processes are as follows:

- The profile of a domain which is a set of the predefined topics related to a domain, is defined as $D(t_p)$.
- The profile of an expert which is a set of the predefined topics related to an expert, is defined as $E(t_p)$.
- The profile of a domain which is a set of the generated topics related to a domain, is defined as $D(t_g)$.
- The profile of an expert which is a set of the predefined topics related to an expert, is defined as $E(t_g)$.

The profiles of the domains and the experts are generated using text mining methods. The profiles contain the related topics with the experts or the domains, and a relevancy score. For generating the profiles, we search for every term of the topics in the document collections of the domains and the experts. For instance, if none of the terms of a topic is found in the document collection of an expert, we cannot define a relevancy between the expert and the topic. So the relevancy score between the topic and the expert is set to zero (0). On the other hand, if all of the words of a topic is found in the document collection of an expert, the relevancy score is set to a greater point. The generated profiles constitute the inputs of the "Applying Data Fusion Techniques" process.

During the process "(3) Applying Data Fusion Techniques", the profiles of experts and domains are composed using data fusion techniques. There are two outputs of the process:

- Expert-Domain scores based on the generated topics, resulted as the $score_{DataFusionTechnique}(E, D, t \in t_g)$. $D(t_g)$ and $E(t_g)$ are the inputs for the calculation.
- Expert-Domain scores based on the predefined topics, resulted as the $score_{DataFusionTechnique}(E, D\ t \in t_p)$. $D(t_p)$ and $E(t_p)$ are the inputs for the calculation.

The process "(4) Applying Weights" performs different weights to find out the impact of the generated topics and predefined topics on the relevancy scores. The weights are applied, and final scores of expert-domain matching are calculated as $score_{DataFusionTechnique}(E, D)$. Finally, a ranked list of the relevancy scores between domains and experts are generated. For each data fusion technique, a different ranked list is generated.

4.1.1. Extracting Topics

In our model, there is an assumption as predefined topics are not present in the corpus or the defined topics are poor or unclear to summarize the corpus. So we generate topics from the document collection of the domains.

A topic has a unique number and a set of keywords, and a topic set consists of many topics. The topics are generated by applying the *Gibbs Sampling Algorithm* on the document collection of a domain.

LDA is used for extracting topics and Gibbs sampling algorithm is used for inference [15]. LDA provides a statistical approach to document clustering based on words that appear in a document [9]. For the implementation of LDA and Gibbs sampling, we have used LingPipe API [2].

In Gibbs sampling algorithm, the next state is reached by sequentially sampling all variables from their distribution when conditioned on the current values of all other variables and the data [15]. The next state is calculated by using the probability of the word w_i under topic j, and the probability of topic j in document d_i. Applying

this algorithm, the full conditional distribution described by Griffiths & Steyvers [15] is given in the Equation 4-1, which is used to assign words to topics. The first ratio expresses the probability of the word w_i under topic j, and the second ratio expresses the probability of topic j in document d_i.

$$P(z_i = j \,|z_{-i}, w) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha} \qquad \text{(Equation 4-1)}$$

- z_i is a vector of assignments. z_i includes the words w which are assigned to topic j.
- z_{-i} is the vector of assignments which is assigned before z_i.
- n is the total number of word instances.
- $n_{-i}^{(\cdot)}$ is a count that does not include the current assignment of z_i.
- α and β are hyperparameters which are used to form a good compromise between the number of topics per document and the number of words per topic.
- T is the number of topics.

The z_i variables are initialized to values in $\{1, 2, \ldots, T\}$ for determining the initial state of the Markov chain. The chain is run for a number of iterations, each time finding a new state by sampling each z_i from the distribution specified by Equation 4-1. After enough iteration for the chain to approach the target distribution, the current values of the z_i variables are recorded [15]. In our model, the recorded values of the z_i variables are formed the topics in a topic set.

4.1.2. Setting Profiles

The process "Setting Profiles" contains the following two sub-processes:

- Setting predefined-topic-based profiles.
- Setting extracted-topic-based profiles.

For each domain and expert, we create profiles by using the textual evidences generated $D(t)$, and $E(t)$ from the topic extraction method. As each domain and

expert already has associated keywords defined in the system, we incorporate these keywords to the domain and expert profiles respectively. The relevance between experts and domains are then computed using the pre-defined keywords and the textual evidences with data fusion techniques.

The inputs of the "Setting Profiles" process are document collections of the experts and the domains and the topics whether predefined or extracted. The output of the process is a profile. For instance, an expert's profile consists;

- A word of a topic which is found in the document collection of the expert,
- The score of the word in the expert's document collection,
- The rank of the word which is gathered after sorting all the words in the profile in order according to the scores.

Profiles are generated matching the keywords of topics with the document collections of the domains and the experts using a weight model. Weight model is used for assigning the scores to the matched documents in a document collection. The maximum number of keywords matched gets the highest scores. We set the ranks of the domains or the experts with respect to a topic by sorting the scores. Item has the highest score get the highest rank, rank 1. The highest score varies depending on weight models, data fusion techniques, and the corpus.

4.1.2.1. Setting Predefined-Topic-Based profiles

The predefined topics include keywords which are defined using common forms. The examples of predefined topics can be;

- The keywords of conference papers which are set by authors,
- The technology codes of a project which are set by owners of a project,
- The research area keywords of a scientist which are set by himself.

For instance, in TÜBİTAK (The Scientific and Technological Research Council of Turkey), TEYDEB (Technology and Innovation Funding Programs Directorate) has several funding programs which are described in the Section 5. Corporations apply to these funding programs with their projects to be funded. Technology codes defined by TÜBİTAK are assigned to the projects by the applicants. In some cases,

technology codes cannot represent the projects clearly. For instance, a project is related with "Geographic Information Systems – Coğrafi Bilgi Sistemleri" while the technology codes of the project are as follows:

- Computer Science and Technology – Bilgisayar Bilimleri ve Teknolojisi,
- Software Engineering – Yazılım Mühendisliği,
- Computer Graphics – Bilgisayarda Grafik.

If there are predefined topics in a corpus, while setting profiles we have to give a standard score/rank for each associated topic. For example consider a corpus consisting of peer-reviewed conference papers each of which is associated with a set of keywords provided by the authors. While choosing the keywords, the authors often try to find the most relevant ones according to their papers content. However, these keywords often turn out to be very generic and are not able to represent the finer details of the paper. Consequently, such pre-determined keywords by the authors should be associated with the highest ranks i.e. rank 1 or 100/100 points of score but we also need supplementary keywords produced by topic extraction methods which should have a score less than 100 or a rank lower than the pre-determined keywords' rank as shown in the Figure 4-2.

Profiles of Domains for Predefined Topic 1 D(tp1)

Rank	Project	Score
1	P1	100
1	P2	100
1	P4	100

Profiles of Experts for Predefined Topic 1 E(tp1)

Rank	Candidate	Score
1	S1	100
1	S5	100

Figure 4-2: Sample domain and expert profiles based on the predefined topics

4.1.2.2. Setting Extracted-Topic-Based Profiles

The "voting" process is used for generating the profiles of the domains and the experts. For instance, if a domain and an expert are related with the same topic, then the scores/ranks of this topic in the profiles are assigned according to their relevancy degrees. Sample domain and expert profiles corresponding to a topic can be seen in Figure 4-3 below.

Profiles of Domains for Generated Topic 1 D(tg1)

Rank	Domain	Score
1	D2	5.7
2	D1	4.3
3	D4	1.2

Profiles of Experts for Generated Topic 1 E(tg1)

Rank	Expert	Score
1	E1	2.3
2	E5	0.7

Figure 4-3: Sample domain and expert profiles based on the generated topics

4.1.3. Applying Data Fusion Techniques

In the proposed model, the score/rank of experts with respect to a topic is taken into consideration. As well as the domains, the profiles of the experts also affect the relevancy between an expert and a domain.

For instance, consider the supporting documents of two different experts which contain "Veri Madenciliği, Yapay Sinir Ağları" and "Veri Madenciliği, Karar Destek ve İş Zekası Sistemleri" phrases. A topic can be extracted from this corpus containing the keywords, "analizi, ambarı, destek, iş, karar, madenciliği, olap, veri, veriler, zekası". In this situation, if we do not consider the score/rank of experts with respect to a topic, the relation of experts with topics cannot be differentiated. An example can be seen in Table 4-1 below. 2 keywords of the topic match for "E-1"

and 6 keywords of the topic are match for "E-2". In this case, the relevancy between the topic and "E-2" is more powerful. But if we do not consider the score/rank of experts' profiles, the relevancies between "E-1" and the topic; "E-2" and the topic will be considered equally.

Table 4-1: Example of Expert – Topic Profile Relations

Experts	**Sample Phrases Related with the Experts**	**Topic Keywords**	**Relevant**	**Matched Keyword Count**
E-1	Veri Madenciliği, Yapay Sinir Ağları	analizi, ambarı, destek, iş, karar, madenciliği, olap, veri, veriler, zekası	True	2
E-2	Veri Madenciliği, Karar Destek ve İş Zekası Sistemleri	analizi, ambarı, destek, iş, karar, madenciliği, olap, veri, veriler, zekası	True	6

Considering the impact of the scores of expert profiles, two different weights are used with the generated-topic-based profiles. These weights are as below:

- Only the "score" and "rank" values of domain profiles are used which is defined as, $D(t_g)$.
- "Score" or "rank" values of both domain and expert profiles are used as equally weighted, $\left(D\left(t_g\right) \times 0.5\right) + \left(E\left(t_g\right) \times 0.5\right)$.

The data fusion techniques, "Reciprocal Rank", "CombMNZ", and "expCombMNZ" are employed in the proposed model. The "Reciprocal Rank" score is calculated using the formula below:

$$score_{RR}(E, D, t \in t_g, t_p) = \left(\sum_{t \in D(t) \cap E(t)} \frac{1}{rank\_d_t} \times w_d\right) + \left(\sum_{t \in D(t) \cap E(t)} \frac{1}{rank\_e_t} \times (1 - w_d)\right) \quad \text{(Equation 4-2)}$$

where, w_d is the weight of the domain profile, $1 - w_d$ is the weight of the expert profile. $rank\_d_t$ shows the rank of the topic according to the score in the domain profile and $rank\_e_t$ shows the rank of the topic according to the score in the expert profile.

The scores of profiles are used while applying the "CombMNZ". The number of the expert profiles and the domain profiles which overlap on a given topic t, is also handled as $\|D(t) \cap E(t)\|$. The weighted scores of the expert profiles and the domain profiles are utilized in the calculation of the final score similar to the Reciprocal Rank. The calculation of "CombMNZ" is given in the Equation 4-3.

$$score_{CM}(E, D, t \in t_{g,}\, t_p) = (\|D(t) \cap E(t)\| \sum_{t \in D(t) \cap E(t)} score\_d_t \times w_d) + (\|D(t) \cap E(t)\| \sum_{t \in D(t) \cap E(t)} score\_e_t \times (1 - w_d)) \quad \text{(Equation 4-3)}$$

- $score\_d_t$ is the score of a domain with respect to a topic.
- $score\_e_t$ is the score of an expert with respect to a topic.

Similarly with the "CombMNZ", while applying "expCombMNZ" the scores of profiles are used. The Equation 4-4 is used for calculating "expCombMNZ".

$$score_{EC}(E, D, t \in t_g, t_p) = (\|D(t) \cap E(t)\| \sum_{t \in D(t) \cap E(t)} e^{score\_d_t} \times w_d) + (\|D(t) \cap E(t)\| \sum_{t \in D(t) \cap E(t)} e^{score\_e_t} \times (1 - w_d)) \quad \text{(Equation 4-4)}$$

- $\|D(t) \cap E(t)\|$, is the count of the expert profiles and the domain profiles which overlap on a given topic t.
- $e^{score\_d_t}$ is the exponential functional of the score of a domain with respect to a topic.
- $e^{score\_e_t}$ is the exponential functional of the score of an expert with respect to a topic.

After applying the data fusion techniques, the combined relevancy scores between the experts and the domains are found. Figure 4-4 shows an overview of the function, applying data fusion techniques. A list is shown as a sample. In the proposed model, different expert-domain relevancy scores are calculated for each data fusion technique.

4.1.4. Applying Weights to Expert-Domain Scores

The outputs of the "Setting Profiles" process are,

- The relevancy scores of experts and domains based on the generated topics, $score_{DataFusionTechnique}(E, D, t \in t_g)$.
- The relevancy scores of experts and domains based on the pre-defined topics, $score_{DataFusionTechnique}(E, D\ t \in t_p)$.

We apply different weights to these relevancy scores, to derive one relevancy score and to state the impact of generated topics to the proposed model.

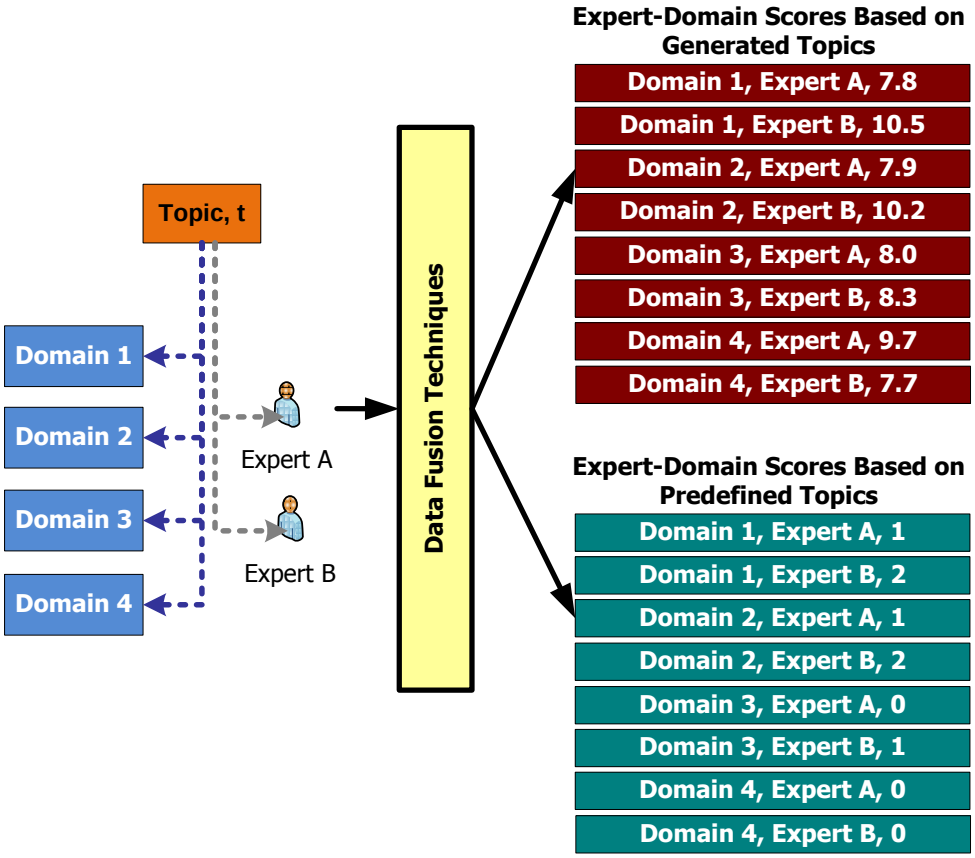


Figure 4-4: Overview of applying data fusion techniques

4.2. The Proposed Model with Finding Similar Experts

With expert finding systems, matching experts with relevant domains is aimed. For instance, the matched experts do not correspond to the domains. This could be occur in many cases, some of the cases are listed below:

- The experts are not interested in the matched domains any more.
- If time is important, the experts' agenda is not available to deal with the selected domain.

A new question is added to our proposed model:

Who are the similar experts of the matched experts?

We extend the proposed model based on this question. The proposed model gives the ranked lists of relevancy scores between the domains and the experts. So for finding similar experts task, we choose the top experts and as an output we find the

similar experts. The structure of the proposed model with finding similar experts is shown in the Figure 4-5.

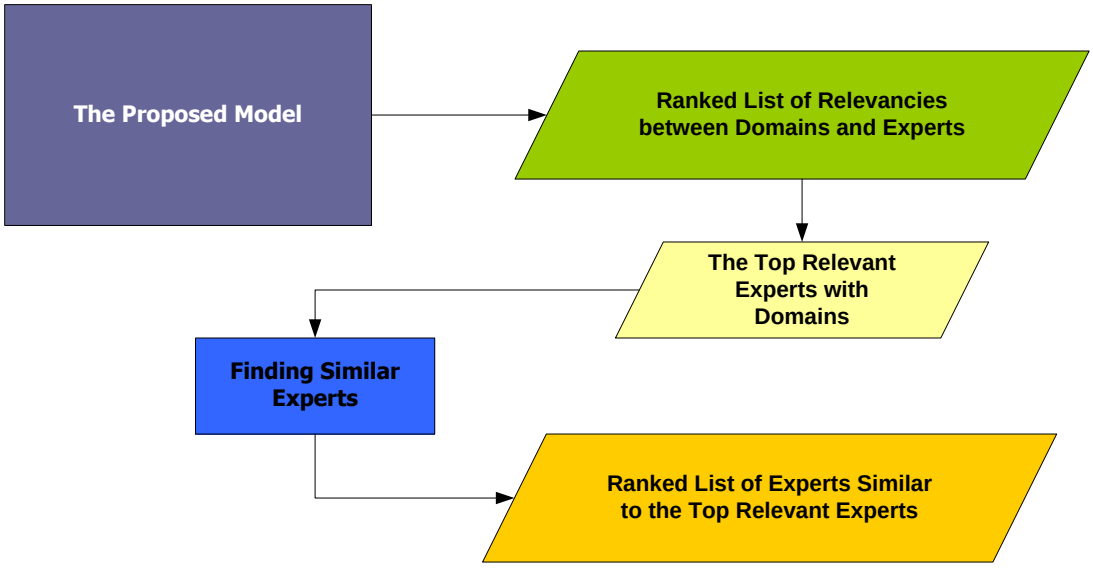


Figure 4-5: The structure of the proposed model with finding similar experts

During the "Finding Similar Experts" process, the similarities between structured representations of the experts could be calculated. The structured representations are defined in common ways. A structured representation can be;

- The relevant documents in a profile of an expert,
- The top discriminative terms of an expert,
- Some descriptive qualifications as "Java Experience with the values of – Beginner, Good, Very Good", "Work Experience defined by years".

The cosine similarity is used to determine the similarity between the lists of two experts' terms using the Equation 4-5.

$$sim_x(e, e') = \cos\big(l_x(e), l_x(e')\big) = \frac{l_x(e) \cdot l_x(e')}{\|l_x(e)\| \, \|l_x(e')\|} \qquad \text{(Equation 4-5)}$$

where, $l_x(e)$ and $l_x(e')$ values are representing the list of terms associated with e and e'experts.

If there are more than one list of terms described an expert then weights can be given to the different similarities. The calculation of similarities for two different lists is given in the Equation 4-6.

$$similarity(e, e') = (sim_{X1}(e, e') \times w) + (sim_{X2}(e, e') \times (1 - w)) \quad \text{(Equation 4-6)}$$

where, w is the weight for the similarity scores.

CHAPTER 5

CASE STUDY: TÜBİTAK – TEYDEB

Technology and Innovation Funding Programs Directorate (TEYDEB) is one of the funding programs directorates in The Scientific and Technological Research Council of Turkey (TÜBİTAK). TEYDEB provides five funding programs which are grouped in three subjects and listed as below [26]:

- Industrial R&D Funding Programs – 1501, 1509
- SME (Small and Medium Size Enterprise) Funding Programs – 1505, 1507
- Project Brokerage Events Funding Program - 1503

Organizations can apply to these funding programs with their projects to be funded. Projects can have various areas of focus and research. The applicant organizations should describe their projects in detail online via PRODİS [27].

Although application documents contain more information, we use a restricted set of information which is allowed for use in our study. In the data set, information about projects is listed below:

- Project Id: For every project, a unique number is generated by the system automatically. We have used project id for associating projects with the other information.
- Name: Name of projects.
- Keywords: Applicants can define keywords of projects using at most 100 characters.
- Summary: Summary is the short description of a project. Summary can contain at most 1000 characters.

- Expertise Areas: Expertise areas contain the information about technological and scientific research fields of projects. Expertise areas can be defined up to 500 characters.
- Summary of Innovation: If exist, the innovative parts of projects can be defined.
- Summary of Purpose: The purpose of projects can be defined.
- Technology Codes: The scientific and technological scenes of TÜBİTAK are defined as a structured list. Applicants have to choose at least one technology code and at most three technology codes from this list which are related to research areas of projects.

On the other hand, TÜBİTAK have a scientist knowledge base called "Araştırmacı Bilgi Sistemi" (ARBİS). ARBİS contains background information about researchers, academicians, professionals;

- Who are working or studying in Turkey.
- Who are Turkish citizens and working or studying abroad.

In this study, "Scientist" is used for all researchers, professionals who are registered to ARBİS. Scientists from ARBİS are chosen to evaluate project proposals and monitor accepted projects on behalf of TÜBİTAK. Projects are assigned to the scientists according to their expertise and research areas.

Although scientists have much more information in ARBİS, we use a restricted set of information which is allowed for use in our study. In the data set, information of scientists is listed below:

- Scientist Id: For every scientist who is registered to ARBİS, a unique number is generated by the system automatically. We use scientist id for differentiating the scientists and associating with the other information.
- Research Areas: Scientists are able to define keywords of their research areas as free text in Turkish. 1024 characters can be defined for the research areas. Research areas are not defined by scientists mandatorily.

- Technology Codes: The scientific and technological scenes of TÜBİTAK are defined in a structured list. Scientists may choose one or more technology codes from this list which are related to their research areas.

A sample scientist with dummy data is defined in Table 5-1.

Table 5-1: A sample scientist with dummy data

Scientist Id	1
Research Areas	Veri Madenciliği Yapay Sinir ağları Veritabanı Sistemleri
Technology Codes	999 – Computer Science and Technology 998 – Database Technologies

Currently, TEYDEB specialists use a simple straightforward database querying system to find experts from ARBİS. When they assign evaluators to the projects, they;

- Search World Wide Web to find up-to-date information about the scientists registered to ARBİS according to the project subjects.
- Try to match the research areas of scientists with the project's technological areas.
- Look for the previous performances of scientists if they are previously assigned to the projects. For the previous performance of a scientist,
    - Submission of periodical reports on time
    - Quality of periodical reports

  are the criteria which are considered in.
- Look for the references given by other scientists when looking at "socialization need" perspective defined in the study of Yimam-seid and Kobsa [32]. A scientist can be recommended by another scientist for a specific topic.

When matching the projects with the scientists, choosing the "right" scientist depends on the research level of the specialist or the level of acknowledgement

about that scientist. After the scientists are assigned to the projects for evaluation, the scientists can reject the evaluation requests according to their availability, ethical issues or other personal reasons. In such cases the requirements;

- Matching the "right" scientists with the project,
- Suggesting the similar scientists if the assigned scientist has an excuse,

stand out. Fulfilling these requirements, we apply our proposed model to the TEYDEB data set (For ethical clearance, please refer to APPENDIX A). The proposed model and the finding similar experts task present scientific solutions for TEYDEB. The core competencies of this case study are listed below;

- Finding the appropriate scientists for a given project proposal.
- Finding similar scientists to the top appropriate scientists.

By implementing the proposed model to TEYDEB and ARBİS, the scientists and projects are automatically matched and the similar profiles of the matched scientists are automatically obtained, ranked and presented.

Data Set

TEYDEB and ARBİS data sets contain confidential data. We get permission from TÜBİTAK to use TEYDEB and ARBİS data sets in our thesis study (Please refer to APPENDIX A).

The technology codes of TÜBİTAK are used to describe the projects and the scientists. Although the technology codes definitions have three levels of hierarchy, they are not detailed enough to describe an area of interest of a project or a scientist exactly. Some of the technology codes which are related with "Computer Science" are given in the Table 5-2 [28]. For instance, the two projects with the same technology code "Sayısal Algoritmalar" can contain very different information in the fields of project which are keywords, summary, expertise areas, summary of innovation, summary of purpose.

Table 5-2: Sample technology codes

Code	Level 1	Level 2	Level 3
330000	Teknolojik Bilimler (Mühendislikler)		
331400		Bilgisayar Bilimleri ve Teknolojisi	
331401			Bilişimsel Model
331402			Karmaşıklık Kuramı
331403			Biçimsel (Formal) Diller
331411			Sayısal Algoritmalar
331412			Benzetim(Simülasyon) ve Modelleme

While applying the methods,

- The technology codes are used as the predefined topics. The technology codes are related with both of the scientists and the projects.
- The scientists are used as the experts.
- The projects are used as the domains.

5.1. The Baseline Method

The baseline method is applied to TEYDEB and ARBİS data sets as explained in the Section 3. The overview of the baseline method applied to the TEYDEB and ARBİS data sets is given in the Figure 5-1.

The baseline method is applied using only the technology codes which are both defined to describe the projects and the scientists. During the process "Applying Data Fusion Techniques", the scores or the ranks of the scientist profiles are not used. The scientist profiles are used as an implicit vote to the projects' profiles.

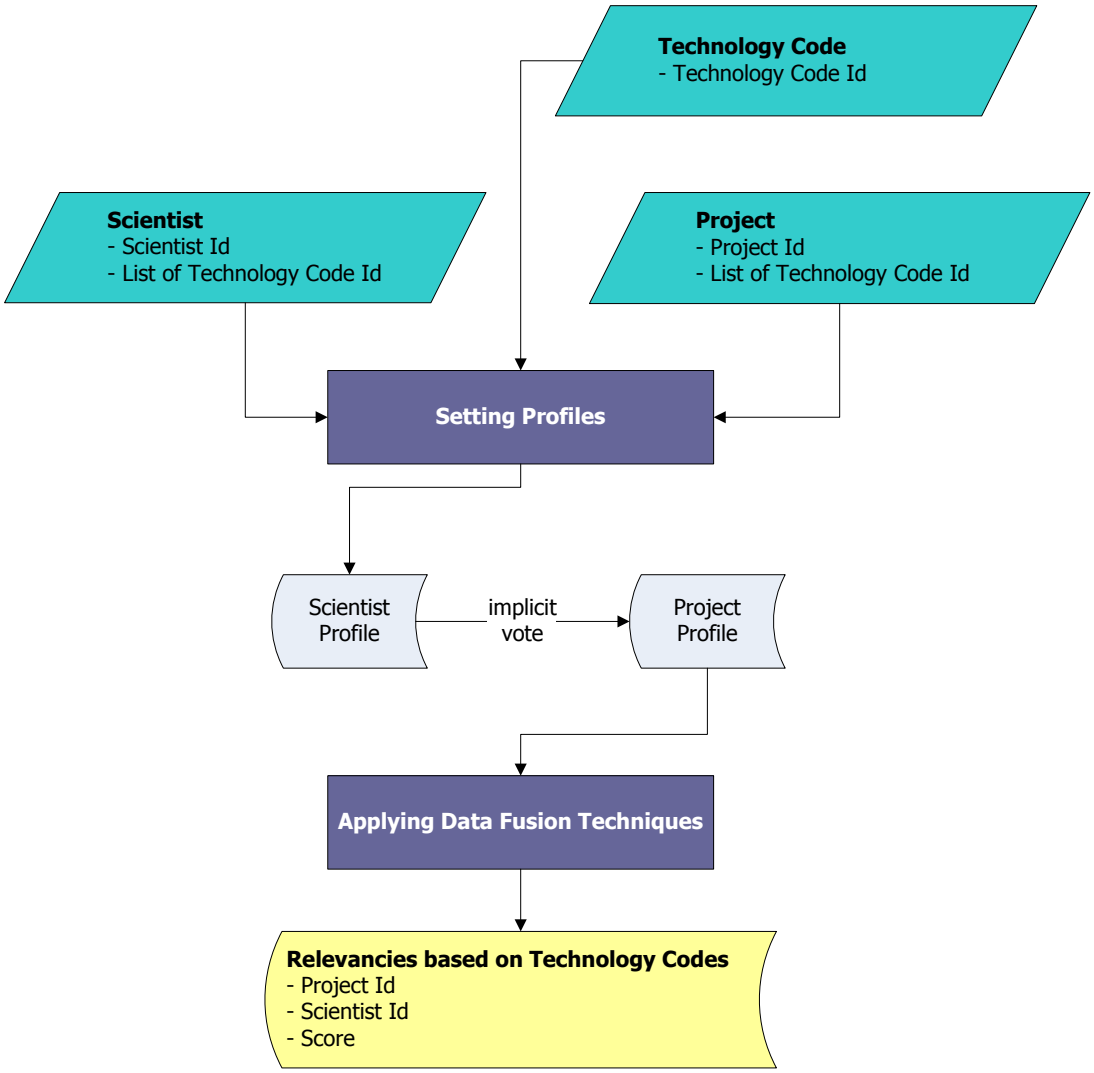


Figure 5-1: Overview of baseline method applied to TEYDEB

5.2. The Proposed Model and Finding Similar Experts Task

The structure of the proposed model applied to TEYDEB and ARBİS data sets is given in the Figure 5-2. The structure also contains the "Finding Similar Experts" task.

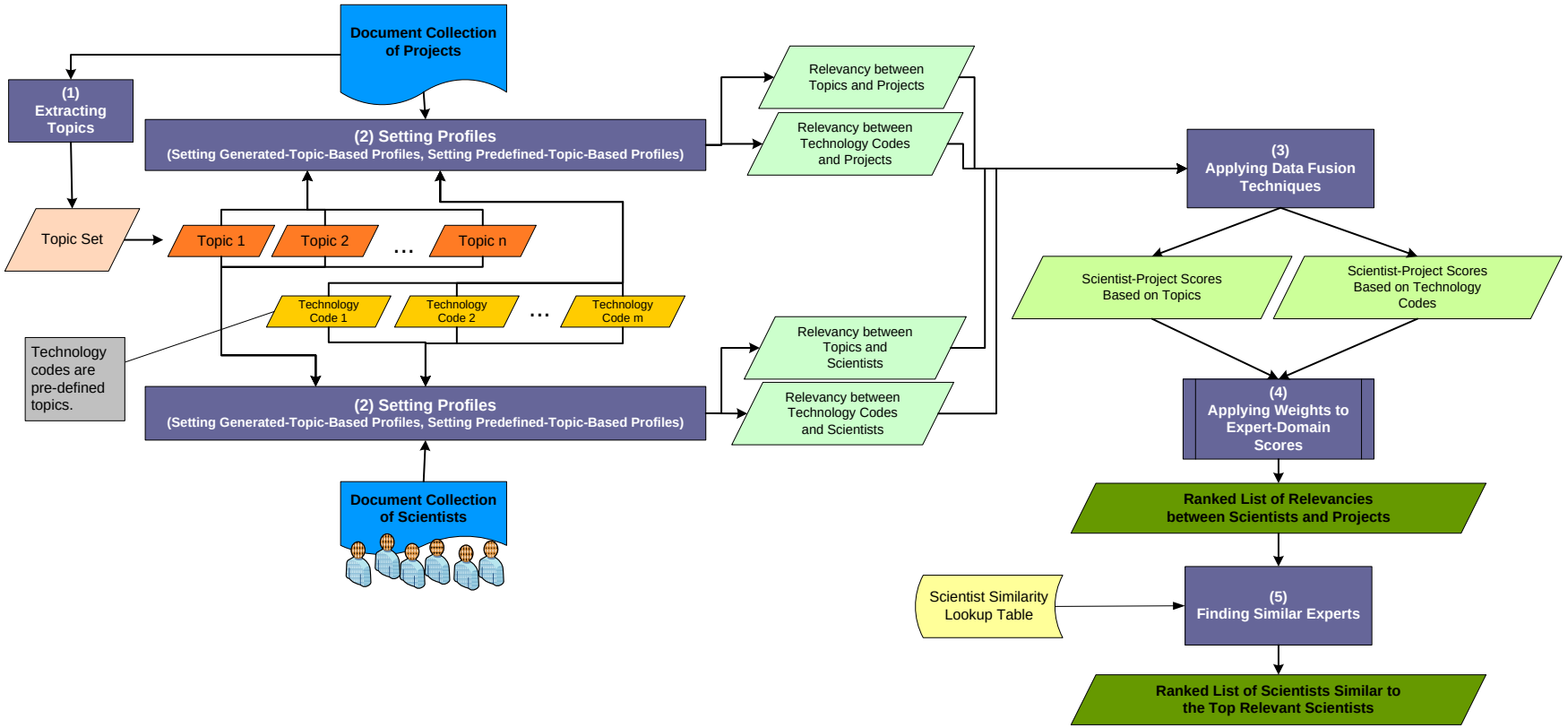


Figure 5-2: The structure of proposed model applied to TEYDEB

As explained in the Sections 4.1 and 4.2, the processes,

- Extracting topics,
- Setting profiles,
- Applying data fusion techniques,
- Applying weights to expert-domain scores,
- Finding similar experts

are implemented.

While extracting topics, the following projects' fields are utilized;

- name,
- keywords,
- summary,
- expertise areas,
- summary of innovation,
- summary of purpose

So we use all of the information about projects excluding the technology codes to extract the topics because all of the information can contain valuable keywords about projects.

During the process of setting profiles, the profiles based on the technology codes and the profiles based on the topics are calculated. Because the technology codes are predefined topics, we set their scores to the top score which can be retrieved in TEYDEB-ARBİS data sets using DLH13 weight model and we set their ranks to the rank 1. The following formula given in the Equation 5-1 is used to calculate the top score of the weight model DLH13 for TEYDEB-ARBİS data sets [18].

$$score(d,Q) = \sum_{t \in Q} \frac{qtw}{tf + 0.5} \cdot \left(log_2 \left(\frac{tf \,.\, avg\_l}{l} \cdot \frac{N}{F} \right) + 0.5 \, log_2 \left(2\pi tf \left(1 - \frac{tf}{l}\right) \right) \right) \quad \text{(Equation 5-1)}$$

- d is the document.
- Q is the query.

- qtw is given by qtf/qtf_{max} , where qtf is the query term frequency and qtf_{max} is the maximum query term frequency.
- tf is the term frequency term frequency of the term t in document d.
- avg_l is the average document length in the collection.
- l is the document length.
- N is the number of documents in the collection.
- F is the frequency of the query term in the collection.

For the process of finding similar experts, the similarities between experts are calculated formerly and stored in a lookup table as "Scientist Similarity Lookup Table". While calculating the similarities between scientists, "Research Areas" and "Technology Codes" are utilized. Different weights are given to the similarities of the "Research Areas" and "Technology Codes". Technology codes are structured variables where the research areas are free text variables so scientists can define any keyword that they are related with. We use the research areas as term vectors for calculating the cosine similarity. For instance, if two scientists have research areas defined as "işaret işleme, görüntü işleme, video işleme, haberleşme" and "işaret işleme, örüntü tanıma", we constitute the term vectors of the scientists using the unique terms in the research areas as follows:

Table 5-3: Sample term vectors

Scientist	**Sample Term Vectors**
Scientist 1	{işaret, işleme, görüntü, video, haberleşme}
Scientist 2	{işaret, işleme, örüntü, tanıma}

The following equation is used for calculating the similarities between scientists:

$$similarity(s, s') = (sim_{TC}(s, s') \times w) + (sim_{RA}(s, s') \times (1 - w)) \quad \text{(Equation 5-2)}$$

where,

- $sim_{TC}(s, s')$ is the similarity between scientists s and s' corresponding to the technology codes of the scientists,
- $sim_{RA}(s, s')$ is the similarity between scientists s and s' corresponding to the term vectors that are constituted from the research areas of the scientists,
- w is the weight of the similarity corresponding to the technology codes of two scientists which is used to clarify the effect of the technology codes and the research areas on the similarities.

CHAPTER 6

PROOF OF CONCEPT

6.1. Proof of Concept: Topic Extraction Approach

A semantic layer could be generated by extracting topics from in a corpus. The generated semantic layer can handle the different words with same meanings or the words with multiple meanings. Various algorithms can be used to generate topics from a corpus. In our proposed model, we utilize LDA and Gibbs sampling for inference. We have applied Gibbs sampling algorithm on the abstracts of four articles about expert finding which are Balog & Rijke [4], Balog, Azzopardi, & Rijke [5], MacDonald & Ounis [18], Hoffman, Balog, Bogers, & Rijke [16] to prove the concept.

Table 6-1: Summary information about papers

Number	Reference	Keywords Defined by Authors	Category
1	Balog, Azzopardi, & Rijke, 2006	Expert Finding, Enterprise Search	C1
2	Balog & Rijke, 2007	Expert Finding, Similar Experts, Expert Representation	C1
3	MacDonald & Ounis, 2006	Voting, Expert Finding, Expertise Modelling, Expert Search, Information Retrieval, Ranking, Data Fusion	C2
4	Hoffman, Balog, Bogers, & Rijke, 2010	-	C3

As given in the Table 6-1, the papers which are used for proof of concept are categorized into 3 categories intuitively. The "Category" column represents the category of the paper (For more information about the papers, please refer to APPENDIX B).

We manually assigned the categories to the papers. The first paper (*number 1*) is related with one of the traditional methods in expert finding. The second paper (*number 2*) is a bit different from the first paper. Although the second paper's subject is *finding similar experts*, the abstract of the paper does not contain very much of the exact keywords. So the first and the second papers are categorized together. The third paper (*number 3*) is also related with one of the traditional methods in expert finding, but the abstract of this paper contains exact keywords like *voting*, and *data fusion techniques*. So the third paper is separated from the first two paper. The fourth (*number 4*) and the last paper do not contain keywords set by the authors. The fourth paper is mostly about contextual factors in expert finding, it also contains keywords about *finding similar experts* but these keywords are not repeated very much. So the fourth paper is categorized separately from the other papers.

We implement three runs in our experiments. The parameters of the experiments are given in the Table 6-2. The number of topics indicates the number of the generated topics from the data set. The number of abstracts indicates number of the abstracts which are used in the topic generation process.

Table 6-2: The parameters of the experiments

Run	**Number of Topics**	**Number of Abstracts**
Extraction#1	2	3
Extraction#2	2	4
Extraction#3	3	4

In the Extraction#1, we extract two topics from the first three abstracts with the numbers 1, 2, and 3. The topics are generated by the algorithm, and each paper's

relevance with each topic is calculated and found as shown in Table 6-3. The distribution of the topics is similar to the categories of the papers.

Table 6-3: The generated two topics and their associated keywords based on the three abstracts

Topic	**Keywords**	**Relevant Papers**
Topic 0	experts, performance, finding, given, trec, topic, second	1, 2
Topic 1	expert, voting, techniques, models, using, search, query	3

In the Extraction#2, we add the fourth abstract (*number 4*) to the data set. Two topics are generated from the abstracts. The generated topics are given in Table 6-4.

Table 6-4: The generated two topics and their associated keywords based on the four abstracts

Topic	**Keywords**	**Relevant Papers**
Topic 0	expert, models, finding, experts, expertise, factors, based	4
Topic 1	voting, techniques, performance, using, given, query, system	1, 2, 3

The two topics are generated from the four papers, but according to the categorization these four abstracts should be grouped into three topics. With generated two topics, two of the abstracts which are *number 1* and *number 2* are related with different topics. The abstract of papers which are *number 3* and *number 4* are related with right topics. The generated two topics do not sufficiently describe the four abstracts.

Finally, in the Extraction#2, we generate three topics from these four abstracts. In the manual categorization, there are three different categories. As a result of this

experiment, the distribution of keywords and the relations of abstracts with the generated topics are similar to the manual categorization. The papers, *number 1* and *number 2* are matched with the same topic *Topic 2*. The papers, *number 3* and *number 4* are matched with different topics from each other and also from the papers *number 1* and *number 2*. The generated topics and the associated papers are given in Table 6-5.

Table 6-5: The generated three topics and their associated keywords based on the four abstracts

Topic	**Keywords**	**Relevant Papers**
Topic 0	voting, techniques, search, using, query, problem, approach	3
Topic 1	expert, models, finding, expertise, factors, based, content	4
Topic 2	performance, experts, given, system, example, people, topic	1, 2

With this POC study, we indicate that Gibbs sampling algorithm can be used for generating topics in our proposed system. For different data sets, the acceptable number of topics varies. While the acceptable number of topics is 3 for the 4 abstracts, 20 or 100 topics can be significant for other data sets.

6.2. Proof of Concept: Expert Finding in Domains with Unclear Topics

We apply the proposed model to a semi-synthetic data set to demonstrate the success of the model. The study of the semi-synthetic data set is referred as proof of concept (POC) of the proposed model, expert finding in domains with unclear topics.

Papers are used to create the semi-synthetic data set for POC. Three papers about five different subjects are downloaded from ACM, using the Computing Classification System (CCS) [1]. While choosing the five different subjects, we pay attention to

choose subjects from different breakdowns. The chosen subjects and their whole breakdown structure are given in Table 6-6.

Table 6-6: Semi-synthetic data set for POC

Subject	**ACM Computing Classification System (CCS)**
S1	D. Software D.2 Software Engineering D.2.4 Software/Program Verification Subjects: Assertion checkers
S2	H. Information Systems H.2 Database Management H.2.8 Database applications Subjects: Spatial databases and GIS
S3	I. Computing Methodologies I.2 Artificial Intelligence I.2.7 Natural Language Processing Subjects: Machine translation
S4	C. Computer Systems Organization C.2 Computer-Communication Networks C.2.2 Network Protocols Subjects: Protocol architecture (OSI model)
S5	K. Computing Milieux K.3 Computers And Education K.3.1 Computer Uses in Education Subjects: Distance learning

The three papers related with each subject are chosen randomly from ACM. There is one restriction, the chosen paper must contain keywords that are set by the authors of the paper.

We consider the titles and the abstracts of the papers as document collections of the domains. The keywords of the papers are assumed as an expert. We handle the defined keywords as the keywords of an expert. For instance, the keywords of a paper are "cross-language, machine translation, machine-readable dictionary, two-phase". We define one expert from these keywords. In this study, "Expert" is referred to the keywords of each paper. As a ground truth of the data set, we

assume that a paper should be related with only one expert who has the keywords of the paper. Every paper in the data set has only one right expert.

Totally, the data set contains 15 articles (For more information about articles, please refer to APPENDIX C) and 15 experts.

The experiment parameters of the baseline and the proposed methods are listed below and can be seen from the Figure 6-1 and Figure 6-2:

- Pre-defined topics are obtained from the breakdown structure of ACM.
- Generated topics are extracted from the document collection of papers.
- Weight model, DLH13 is used.
- Data Fusion Techniques as Reciprocal Rank, CombMNZ, and expCOMBMNZ.
- Weight of the domain profile is a weight which defines whether the scores of the expert profiles are used or not. The values can be 0.5 (the scores of paper and expert profiles are equally weighted), 1 (the scores of paper profiles are used only).
- Stemming is used, describes whether a stemming algorithm is used for the corpus or not.

The semi-synthetic data set is indexed with using the Terrier library. The stopwords are removed. The experiments are repeated with Porter's stemming algorithm and without it. The weight model "DLH13" is used only, because the performance of the weight model is not required for the experiments. So we choose the weight model which gets the highest scores in the experiments in the Section 3.3 Experimental Results. These experimental settings are same for the applications of the baseline method and the proposed model.

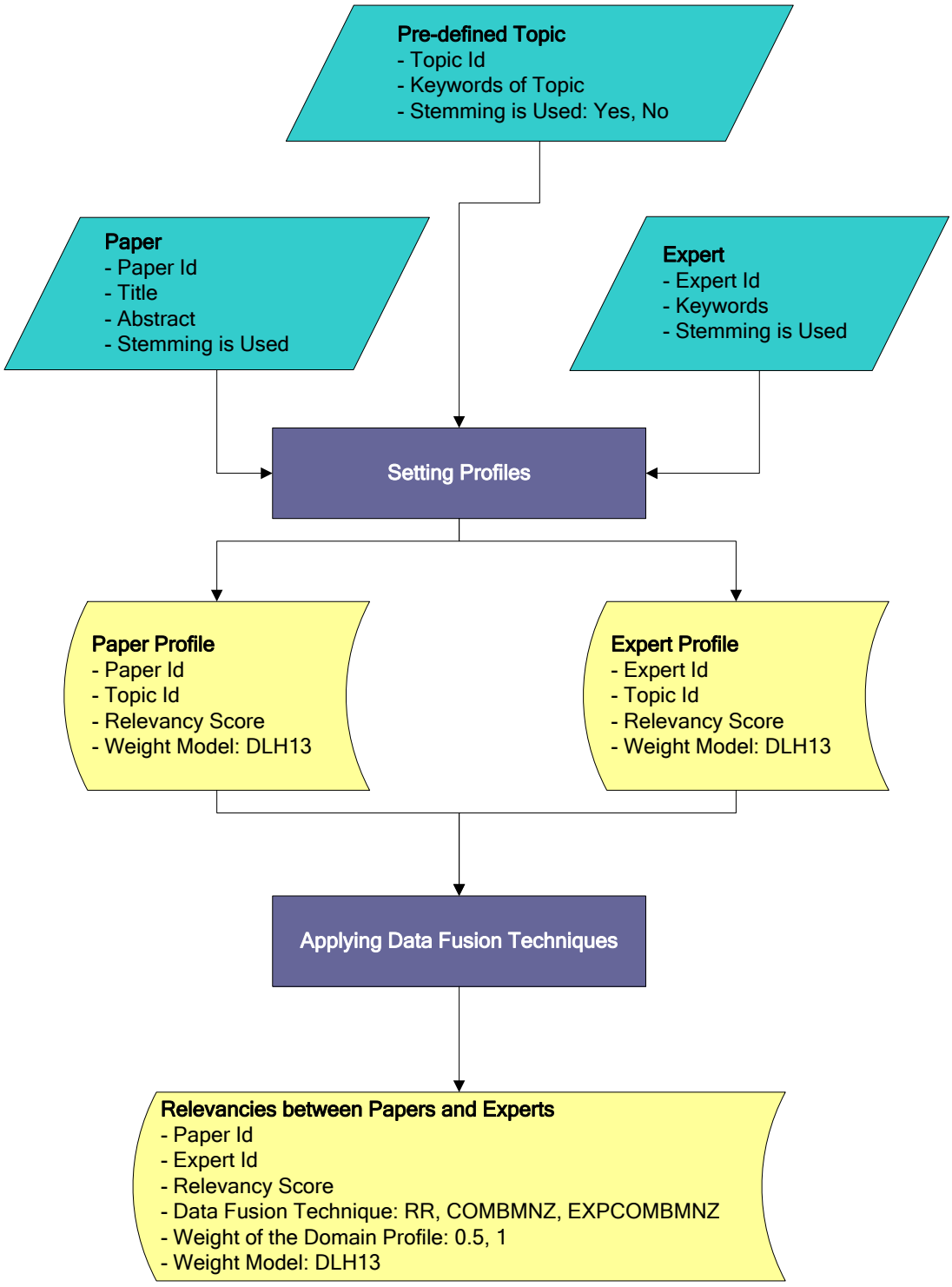


Figure 6-1: Structure of the baseline method application on the semi-synthetic data

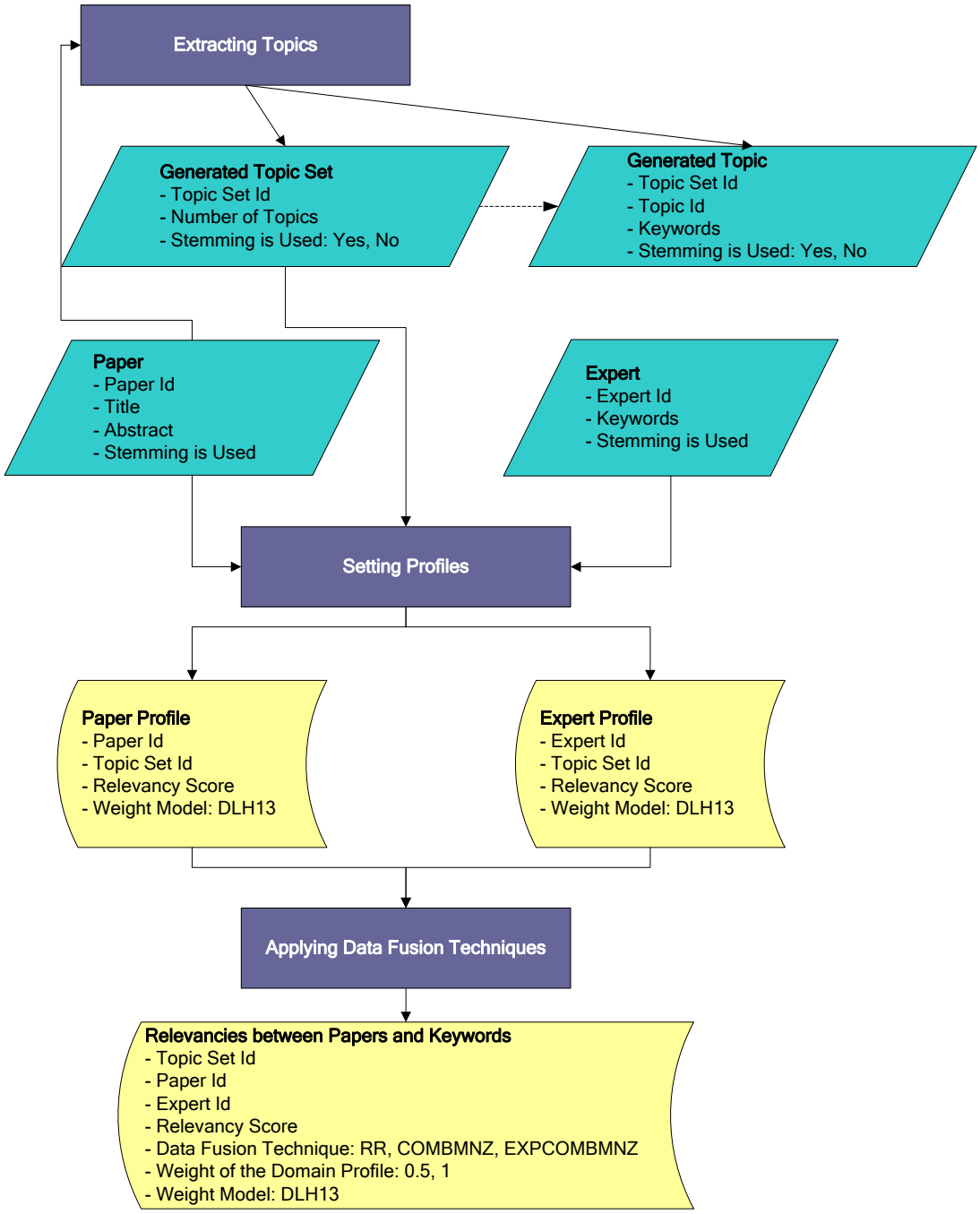


Figure 6-2: Structure of the proposed model application on the semi-synthetic data

For evaluating the results of the experiments we use the traditional IR measures. The measures are listed below:

- Precision
- Recall
- Precision@n (P@n)

- Recall@n (R@n)

where n can be {1, 5, 10, 15}.

Experiments

As a first step, we apply the baseline method to the data set. As predefined topics, we use the ACM Computing Classification System. Five different subjects are assumed as topics. Every unique word of the breakdown structure is defined as a keyword of the topic. The five predefined topics are given in Table 6-7.

Table 6-7: List of pre-defined topics

Subjects	**Topics**	**ACM Computing Classification System (CCS)**
S1	T1	software, engineering, program, verification, assertion, checkers
S2	T2	information, systems, database, management, applications, spatial, databases, gis
S3	T3	computing, methodologies, artificial, intelligence, natural, language, processing, machine, translation
S4	T4	computer, systems, organization, communication, networks, network, protocols, protocol, architecture, osi, model
S5	T5	computing, milieux, computers, and, education, computer, uses, in, education, distance, learning

The top results of applying the baseline method to the semi-synthetic data set is given in Table 6-8 (For all of the results, please refer to APPENDIX E). The top results are very similar to each other. The results can be compared:

- With all the data fusion techniques applied, we get the top results with different values of the parameters.
- The highest precision and P@1 values are obtained with using a stemming algorithm.
- In the semi-synthetic data set, using the scores of the expert profiles (weight of the domain profile = 0.5) do not improve the results.

Table 6-8: Top results of the baseline method

Topic	Data Fusion Technique	Weight of the Domain Profile	Stemming is Used	Precision	Recall	P@1	P@5
T3	CombMNZ, expCombMNZ	0.5, 1	No	0.2	0.8	0.2	0.16
T3	RR	1	No	0.2	0.8	0.2	0.16
T1	RR, expCombMNZ	0.5, 1	Yes	0.25	0.5	0.5	0.1

The proposed model is applied to the semi-synthetic data set, utilizing only the generated topics. The topic set containing different number of topics are generated. Using various topic sets, we can consider the impact of the number of topics to the proposed model. Topic sets containing 5, 10, 15, 20, 30 topics are generated. Generated topic sets are listed in the Table 6-9 (For generated topics, please refer to APPENDIX D). These topic sets are generated for both cases, using the Porter's stemming algorithm and not using the stemming algorithm.

Table 6-9: List of the generated topic sets

Topic Set	Number of Topics
1	5
2	10
3	15
4	20
5	30

The top results of the proposed model of the semi-synthetic data set is given in Table 6-10 (For all of the results, please refer to APPENDIX E). For all of the results, the precision values are smaller than the baseline method; on the other hand recall values increase.

While stemming is not used the highest P@1 value is obtained from the topic set "5" which contains 30 topics. In the experiment, as the data fusion technique

Reciprocal Rank is used. 13 of the 15 papers are matched with the right keywords in the first order where the value of P@1 is 0.8667.

The highest P@1 value is obtained from the topic set "4" which contains 20 topics, using the Porter's stemming algorithm. In the experiment, as the data fusion technique expCombMNZ is used. 14 of the 15 papers are matched with the right keywords in the first order where the value of P@1 is 0.9333.

P@5 value is 0.2 for all of the top results. The first five experts that matched with the papers contain the right expert for all of the top results.

Table 6-10: Top results of the proposed model

Topic Set	**Data Fusion Technique**	**Weight of the Domain Profile**	**Stemming is Used**	**Precision**	**Recall**	**P@1 / R@1**	**P@5**
5	COMBMNZ	0.5	No	0.0986	1	0.7333	0.2
5	COMBMNZ	1	No	0.0986	1	0.6667	0.2
5	EXPCOMBMNZ	0.5, 1	No	0.0986	1	0.8	0.2
5	RR	0.5	No	0.0986	1	0.8667	0.2
5	RR	1	No	0.0986	1	0.8	0.2
4	COMBMNZ	0.5, 1	Yes	0.0806	1	0.6667	0.2
4	EXPCOMBMNZ	0.5	Yes	0.0806	1	0.8	0.2
4	EXPCOMBMNZ	1	Yes	0.0806	1	0.9333	0.2
4	RR	0.5, 1	Yes	0.0806	1	0.6667	0.2
5	COMBMNZ	0.5, 1	Yes	0.0801	1	0.8667	0.2
5	EXPCOMBMNZ	0.5	Yes	0.0801	1	0.8	0.2
5	EXPCOMBMNZ	1	Yes	0.0801	1	0.7333	0.2
5	RR	0.5	Yes	0.0801	1	0.8	0.2
5	RR	1	Yes	0.0801	1	0.8667	0.2

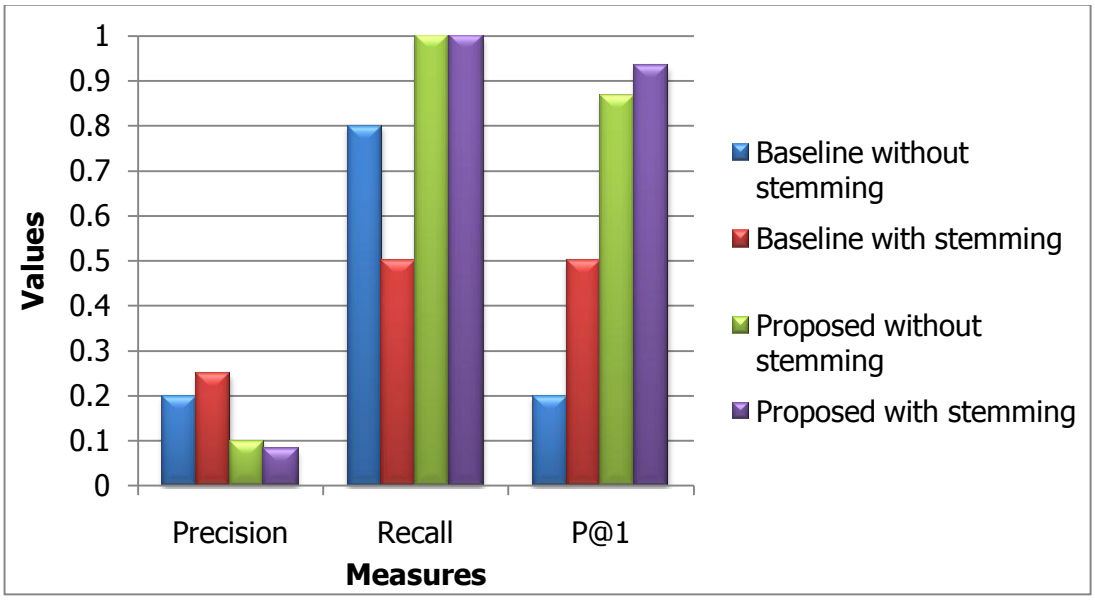


Figure 6-3: Comparison of the baseline method and the proposed model

The comparison of the results of the baseline and the proposed model are given in the Figure 6-3. The precision values have decreased because the retrieved experts for a paper have increased. The recall values have increased because all of the right experts are retrieved using the generated topics in the proposed model. In the baseline method, the half of the papers could be matched to the right experts in the first rank. On the other hand, in the proposed model nearly all of the right experts are matched to the papers in the first rank.

CHAPTER 7

PERFORMANCE EVALUATION

7.1. Experimental Environment

The frameworks, programming languages, APIs and databases used for developing the proposed system and their purpose of use are listed in the table below.

Table 7-1: Experimental Environment

Source (Framework/Programming Language/API/Database)	**Source Name - Version**	**Purpose of Use**
Programming Language	Java 1.6	Implementation
Framework	Eclipse 1.2.2	Implementation
API	Terrier 2.2.1	Indexing and retrieval from the data set
API	LingPipe 4.1.0	Extracting topics with Gibbs Sampling Algorithm
Database	Oracle 10g Express Edition	Storing calculated values (Scores, ranks, etc.)

7.2. Data Set

The data set used in the experiments contains a subset of the projects funded by TEYDEB and the scientists in the ARBİS knowledge base. The ARBİS and TEYDEB data sets are referred as corpus in this section. The projects related with "Information Technologies" and the scientists who selected "Computer Science and Technology" for their technology codes are chosen.

As a preprocessing phase, a project is removed from the corpus when;

- The project does not have any assigned evaluators,
- The assigned evaluators of the project do not have any research areas or technology codes.

A scientist is removed from the corpus when this scientist does not have any research area keywords or technology codes. After preprocessing, there remain approximately 500 projects and 1200 scientists in the corpus. For the experiments, 80% of the TEYDEB data set is used for training and the remaining 20% of projects is used for testing. The training and test sets are constructed randomly.

The assignment of the scientists to the projects which are done by TEYDEB experts are used as ground truth. The number of evaluators assigned to each project in the training set and test set are given in the Table 7-2 and Table 7-3 respectively.

Table 7-2: The percentages of the projects that are in the training set, in respect of the number of the evaluators assigned to the projects

Number of Evaluators Assigned to Projects in the Training Set	**Percentage of Projects with the Number of Evaluators in the Training Set**
1	53%
2	38%
3	8%
4	1%

Table 7-3: The percentages of the projects that are in the test set, in respect of the number of the evaluators assigned to the projects

Number of Evaluators Assigned to Projects in the Test Set	**Percentage of Projects with the Number of Evaluators in the Test Set**
1	53%
2	42%
3	5%

Although there are some English keywords and abbreviations, the corpus is mostly defined in Turkish. The corpus is indexed using Terrier [14], [29]. Before indexing, the stopwords are removed. The Turkish stopword list published by The Natural Language Process Group of Fatih University [25] is used. The stopword list is extended by adding words which are suggested as stopwords in the corpus as "vb, göre, ilgili" (For the entire stopwords list, please refer to APPENDIX F).

For setting the profiles of the projects and the scientists, the topics are searched and results are retrieved using Terrier. While indexing and retrieving data by Terrier, the following parameters are used with non-standard values:

- Collection class to be indexed is defined as "*UTFCollection*".
- Matching retrieval size is not limited and set to zero (0).
- The stopwords file is changed with the Turkish stopwords file.
- Before indexing, the information of projects and scientists are generated as XML documents.
- XML tag values that documents contain are defined to be processed while indexing.
- Before retrieval, a topic file which will be searched in the corpus is given as an input to the Terrier API.

Same as the POC works, DLH13 statistical document weight model is used during the retrieval. For DLH13, the default parameters of Terrier are used.

7.3. Topic Extraction Method

The topic sets are extracted from the TEYDEB data set which contains information of projects. For training of LDA, Gibbs sampling algorithm is used which is developed by LingPipe [2].

The input parameters of LingPipe API, used for Gibbs sampling algorithm are listed below [9]:

- Corpus:
    - Text to be sampled to extract topics.

- All information about the projects is used excluding only the project number.

- Minimum count of a word:
  - Minimum instances of a word in the corpus which can be used. For instance, if minimum count of a word set to "2" then a word used only once is pruned out from the corpus.
  - Minimum count of a word is set to zero. So we include all of the words while generating topics.
- Number of topics:
  - Number of topics which will be extracted from the corpus.
  - 30, 50, 100, 200, 219 topics are extracted from the corpus. Also the number 219 is used for the extraction because there are 219 unique technology codes in the corpus. The impact of number of the technology codes is examined.
- Document-topic prior:
  - Smoothing term. Each document can be modeled from many different topics. When the document-topic prior is lower, the algorithm is encouraged to model a document using fewer topics. So the words in a document are distributed to fewer topics.
  - As suggested in the study of Carpenter & Baldwin [9], $1/number\ of\ topics$ is used as the document-topic prior.
- Topic-word prior:
  - Smoothing term. Each topic can be modeled using words of the corpus. When the topic-word prior is higher, the probabilities of each word is encouraged to be more balanced. Topic-word prior moves the distribution of words in topics closer to the uniform distribution.
  - As suggested in the study of Carpenter & Baldwin [9], $1/number\ of\ words$ is used as topic-word prior.
- Burn-in:
  - Number of samples thrown away during the burn-in phase.
  - 2000 samples are generated in the burn-in phase.
- Sample Lag:
  - Period between samples after burn-in phase.

    - For the experiments, no lag is defined.
- Number of samples:
    - Number of samples that will be taken.
    - 5000 samples are generated.
- Random:
    - Random number generator required for the sampler.
    - Java random number generator *java.util.Random* class is used. The different seeds are not statistically significant so we set the explicit seed to a prime number "157101".
    - Student's t-test is applied using different seeds to show the value of the seed does not affect the generated topics. The seed value is not statistically significant. The details of the t-test are given in the section 7.4.

The input parameters of the LingPipe API to get extracted topics are listed below:

- Words per topic:
    - Define the number of words that a topic contains.
    - Are set to 10 are in the experiments for each topic.

7.4. Choosing the Seed Parameter: Student's T-test

Using fifteen different seeds, topic sets are generated from the TEYDEB data set. Experiments with the same parameters except the seed values and the topic numbers are repeated. In the experiments of the t-test, the experimental settings used are as follows:

- *Data fusion technique:* Reciprocal Rank, CombMNZ.
- *Topic numbers:* 30, 50, 100, 200, 219.
- *Stemming:* 15 topics are generated using the Porter's stemming algorithm and 15 topics are generated not using it.
- *Weight of the domain profile:* The scores of the project and the scientist profiles are used equally, so 0.5 is used for the weight.

- *Project-scientist similarity weight:* The weights for the relevancy score based on topics and the relevancy score based on the technology codes are used equally as 0.5.
- *Seed:* 15 different prime numbers are used for generating topics.

We apply a two-tailed t-test because we expect that different seeds do not affect the results. In the experiments, topics are generated from the same data set with different seeds, so the type of the t-test is paired or dependent. The recall values are used as the inputs of the t-test.

T-tests Using Reciprocal Rank

In the Figure 7-1, the recall values of the experiments are given in which Reciprocal Rank is used as the data fusion technique and stemming algorithm is not used for generating the topics. Figure 7-2 shows the recall values of the experiments are given in which Reciprocal Rank is used as the data fusion technique and the Porter's stemming algorithm is used for generating the topics. The highest recall values are gained in the experiments with the number of topics, 200 and 219. The standard deviation of the recall values is lower with the number of topics, 200 and 219 than the others. The mean and the standard deviation values based on number of topics are given in tables, Table 7-4 and Table 7-5.

Table 7-4: Mean and standard deviation of the recall values (Reciprocal Rank, without using stemming)

Number of Topics	30 Topics	50 Topics	100 Topics	200 Topics	219 Topics
Mean	0.9729	0.9805	0.9889	0.9931	0.9931
Standard Deviation	0.0068	0.0054	0.0037	0.0018	0.0024

Table 7-5: Mean and standard deviation of the recall values (Reciprocal Rank, using stemming)

Number of Topics	30 Topics	50 Topics	100 Topics	200 Topics	219 Topics
Mean	0.9823	0.9867	0.9920	0.9950	0.9960
Standard Deviation	0.0045	0.0026	0.0025	0.0023	0.0015

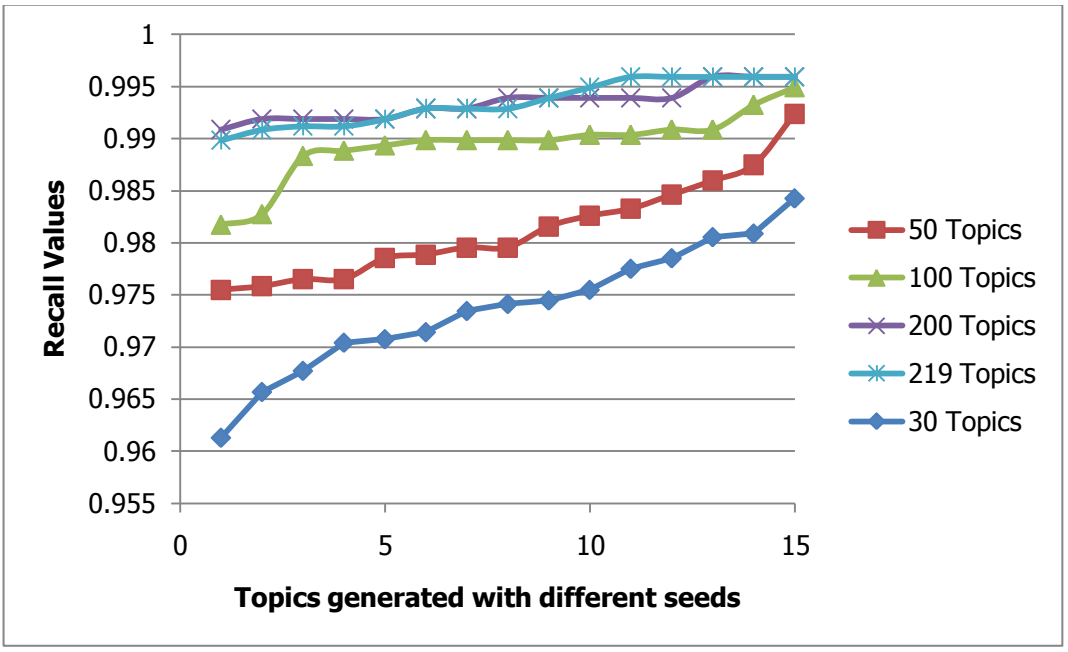


Figure 7-1: Recall values of the t-test experiments using Reciprocal Rank and not using stemming

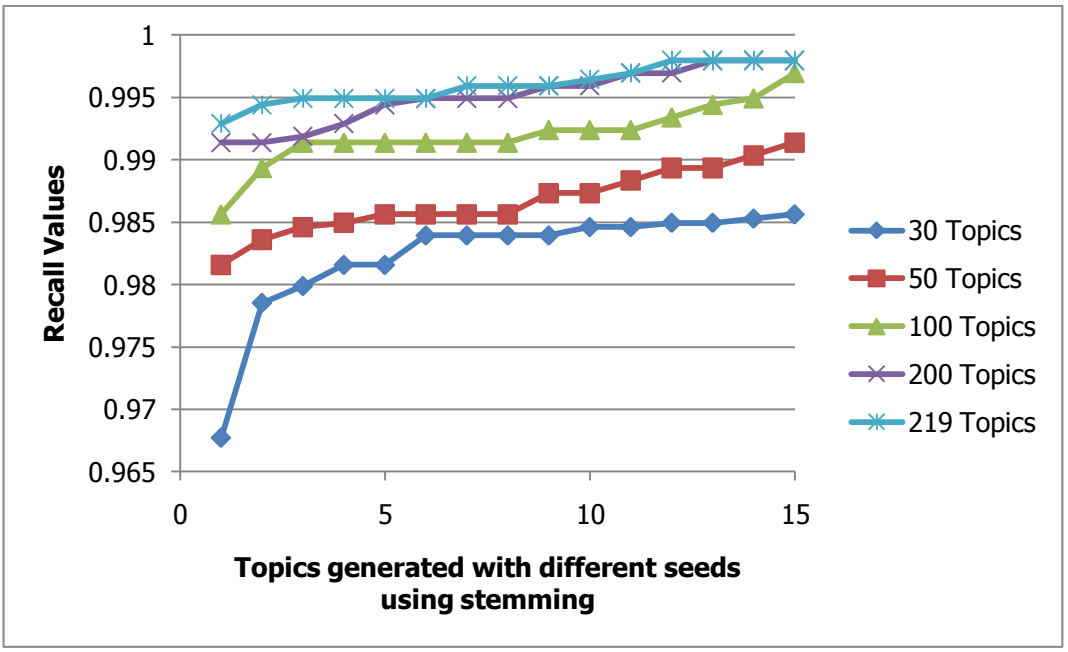


Figure 7-2: Recall values of the t-test experiments using Reciprocal Rank and stemming

The results of the t-test using Reciprocal Rank are given in the Table 7-6 and Table 7-7. P-values are gained similarly for both using the Porter's stemming algorithm and not using it. According to p-values, most of the experiments indicate that different seeds are statistically significant ($p < 0.05$) against different topics. An exceptional case is occurred in the tests between 200 topics and 219 topics. 200 topics and 219 topics are not statistically significant from each other.

Table 7-6: P-values of the t-tests (Reciprocal Rank, without using stemming)

	P-Values
30 Topics - 50 Topics	0.0007
30 Topics - 100 Topics	0.0017×10^{-4}
30 Topics - 200 Topics	0.0009×10^{-6}
30 Topics - 219 Topics	0.0053×10^{-6}
50 Topics - 100 Topics	0.0057×10^{-2}
50 Topics - 200 Topics	0.0014×10^{-4}
50 Topics - 219 Topics	0.0014×10^{-4}
100 Topics - 200 Topics	0.0002
100 Topics - 219 Topics	0.0026
200 Topics - 219 Topics	0.9869

Table 7-7: P-values of the t-tests (Reciprocal Rank, using stemming)

	P-Values
30 Topics - 50 Topics	0.0083
30 Topics - 100 Topics	0.0011×10^{-2}
30 Topics - 200 Topics	0.0056×10^{-4}
30 Topics - 219 Topics	0.0027×10^{-5}
50 Topics - 100 Topics	0.0064×10^{-3}
50 Topics - 200 Topics	0.0098×10^{-5}
50 Topics - 219 Topics	0.0098×10^{-5}
100 Topics - 200 Topics	0.0036×10^{-1}
100 Topics - 219 Topics	0.0057×10^{-2}
200 Topics - 219 Topics	0.1519

T-test Using CombMNZ

In the Figure 7-3, the recall values of the experiments are given in which CombMNZ is used as the data fusion technique and stemming algorithm is not used for generating the topics. Figure 7-4 shows the recall values of the experiments are given in which CombMNZ is used as the data fusion technique and the Porter's stemming algorithm is used for generating the topics. The highest recall values are gained in the experiments with the number of topics, 200 and 219. The standard deviation of the recall values is lower with the number of topics, 200 and 219 than the others. The mean and the standard deviation values based on number of topics are given in tables, Table 7-8 and Table 7-9.

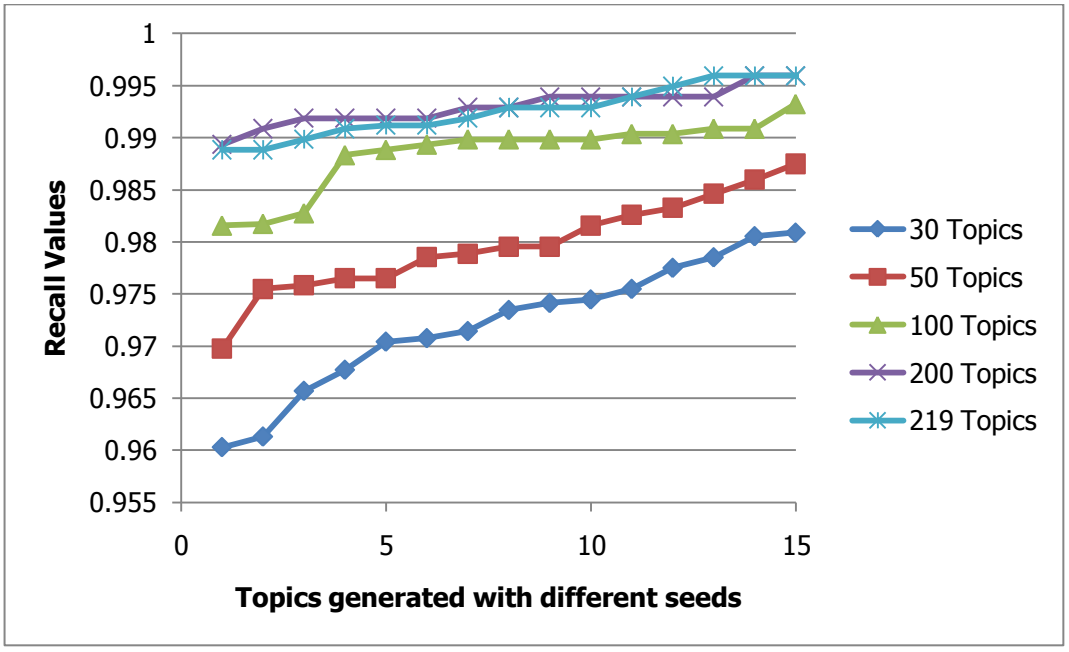


Figure 7-3: Recall values of the t-test experiments using CombMNZ and without stemming

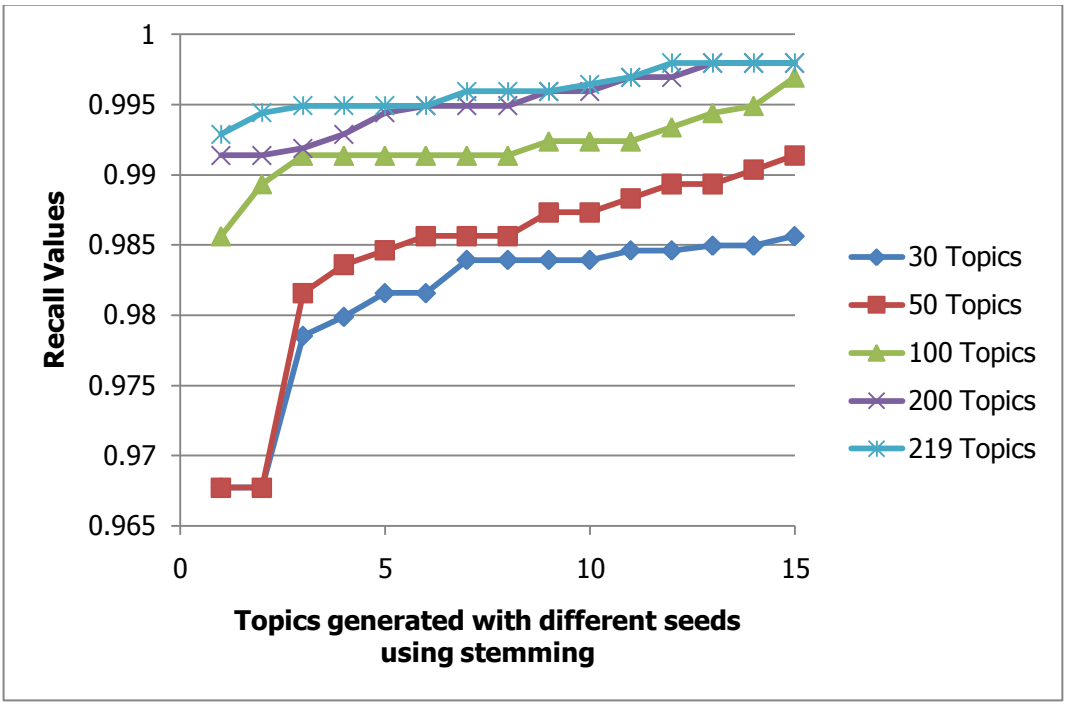


Figure 7-4: Recall values of the t-test experiments using CombMNZ and using stemming

Table 7-8: Mean and standard deviation of the recall values (CombMNZ, without using stemming)

Number of Topics	30 Topics	50 Topics	100 Topics	200 Topics	219 Topics
Mean	0.9729	0.9805	0.9889	0.9931	0.9927
Standard Deviation	0.0068	0.0054	0.0037	0.0018	0.0025

Table 7-9: Mean and standard deviation of the recall values (CombMNZ, using stemming)

Number of Topics	30 Topics	50 Topics	100 Topics	200 Topics	219 Topics
Mean	0.9811	0.9843	0.9920	0.9950	0.9960
Standard Deviation	0.0058	0.0072	0.0025	0.0023	0.0015

The results of the t-test using CombMNZ are given in the Table 7-10 and Table 7-11. P-values are gained similarly for both using the Porter's stemming algorithm and not using it. According to p-values, most of the experiments indicate that different seeds are statistically significant ($p < 0.05$) against different topics. In the

case stemming is not used, some exceptional cases are occurred in the tests between 219 topics and 30, 100, 200 topics. 219 topics are not statistically significant from 30, 100 and 200 topics each other. While stemming is used, most of the p-values are greater than 0.05. Except the t-tests between 100 and 200 topics, and 100 and 219 topics the topics are not statistically significant from each other.

Table 7-10: P-values of the t-tests (CombMNZ, without using stemming)

	P-Values
30 Topics - 50 Topics	0.0007
30 Topics - 100 Topics	0.0017×10^{-4}
30 Topics - 200 Topics	0.0092×10^{-7}
30 Topics - 219 Topics	0.5961
50 Topics - 100 Topics	0.0057×10^{-2}
50 Topics - 200 Topics	0.0014×10^{-4}
50 Topics - 219 Topics	0.0014×10^{-4}
100 Topics - 200 Topics	0.0024×10^{-1}
100 Topics - 219 Topics	0.3803
200 Topics - 219 Topics	0.3333

Table 7-11: P-values of the t-tests (CombMNZ, using stemming)

	P-Values
30 Topics - 50 Topics	0.6129
30 Topics - 100 Topics	0.2508
30 Topics - 200 Topics	0.2292
30 Topics - 219 Topics	0.2217
50 Topics - 100 Topics	0.1454
50 Topics - 200 Topics	0.1359
50 Topics - 219 Topics	0.1359
100 Topics - 200 Topics	0.0003
100 Topics - 219 Topics	0.0057×10^{-2}
200 Topics - 219 Topics	0.1519

7.5. Evaluation Measures

For evaluating the results of experiments, the traditional evaluation measures in IR are used which are; *precision*, the accuracy of the suggested candidates' expertise; and *recall*, the number of candidates with the relevant expertise retrieved. To assess the accuracy of top-ranked candidates retrieved by the system, Precision@n (P@n) and Recall@n (R@n) measures are calculated.

P@n and R@n are calculated as below;

$$P@n = \frac{|\{relevant\ scientists\ at\ n\} \cap \{retrieved\ scientists\}|}{|\{retrieved\ scientists\ at\ n\}|} \quad \text{(Equation 7-1)}$$

$$R@n = \frac{|\{relevant\ scientists\ at\ n\} \cap \{retrieved\ scientists\}|}{|\{relevant\ scientists\ at\ n\}|} \quad \text{(Equation 7-2)}$$

where n is the cut-off rank for retrieved and relevant scientists. In the equation, only the top n relevant and retrieved scientists are considered. For instance, for a project with only one assigned evaluator, P@5, P@10 and P@15 measures will be calculated as;

$$P@5 = \frac{1}{5} = 0.2 \quad , \quad P@10 = \frac{1}{10} = 0.1 \quad , \quad P@15 = \frac{1}{15} = 0.0667$$

Because the project has only one assigned evaluator, the intersection of relevant and retrieved scientists can only be 1. As the maximum values, 0.2 for P@5, 0.1 for P@10 and 0.0667 for P@15 can be gained. The maximum P@5, P@10 and P@15 values for all projects grouped by the assigned evaluators are given in the Table 7-12.

Table 7-12: The maximum precision values of projects in the training set, grouped by number of assigned evaluators

Number of Evaluators Assigned to Projects in the Training Set	Percentage of Projects with the Number of Evaluators in the Training Set	Maximum P@5	Maximum P@10	Maximum P@15
1	53%	0.2	0.1	0.0667
2	38%	0.4	0.2	0.1334
3	8%	0.6	0.3	0.2
4	1%	0.8	0.4	0.2667

We can calculate the maximum precision values as given below;

$$P@n = \frac{\sum_{i=\{1,2,3,4\}} \|P_i\| \times Max_i(P@n)}{\|P\|} \qquad \text{(Equation 7-3)}$$

where;

- i is the number of evaluators assigned to the projects in the training set.
- $\|P_i\|$, is the number of projects with i number of assigned evaluators.
- $Max_i(P@n)$, is the maximum $P@n$ value for projects with i number of assigned evaluators.
- $\|P\|$, is the total number of projects in the training set.

Finally, the maximum precision values for all of the projects in the training set are calculated using the Equation 7-3 and given in the Table 7-13.

Table 7-13: The maximum precision values for training data set

Maximum P@5	0.3132
Maximum P@10	0.1566
Maximum P@15	0.1044

According to Table 7-13, even if all the projects and scientists are matched as they matched in the ground truth set; the maximum precision values can be 0.3132 for P@5, 0.1566 for P@10 and 0.1044 for P@15.

With the same method and equations, the maximum precision values are calculated for the test set. The distribution of projects in the test set and the number of evaluators that are assigned to the projects are given in the Table 7-14. The maximum precision values for all of the projects in the test set are calculated using the Equation 7-3 and given in the Table 7-15.

Table 7-14: The maximum precision values of projects in the test set, grouped by number of assigned evaluators

Number of Evaluators Assigned to Projects in the Test Set	Percentage of Projects with the Number of Evaluators in the Test Set	Maximum P@5	Maximum P@10	Maximum P@15
1	53%	0.2	0.1	0.0667
2	42%	0.4	0.2	0.1334
3	5%	0.6	0.3	0.2

Table 7-15: The maximum precision values for test data set

Maximum P@5	0.304
Maximum P@10	0.1520
Maximum P@15	0.1014

7.6. Evaluation Parameters

The parameters used for topic extraction are listed below:

- *Number of topics:* A topic set consists of topics. This parameter defines the number of topics that a topic set will contain.
- *Stemming:* While generating the topic sets, the topic sets are generated using stemming or not.

We have extracted several topic sets from the corpus with different values of number of topics and stemming parameters. The projects in the training set are related with 219 different technology codes. So we have defined a range of different

topic sizes, such as 30, 50, 100, 200 and 219, to find the most appropriate number of topics. The features of the topic sets are listed in Table 7-16.

Table 7-16: Characteristics of the generated topic sets

Topic Set	**Number of Topics**	**Stemming**
1	30	No
2	50	No
3	100	No
4	200	No
5	219	No
6	30	Yes
7	50	Yes
8	100	Yes
9	200	Yes
10	219	Yes

The other parameters used in the experiments are listed below:

- *Data fusion techniques:* Used for combining scores of retrieved documents and scientist profiles. "RR", "CombMNZ" and "expCombMNZ" data fusion techniques are used.
- *Weight of the domain profile:* In the proposed model the weight of the domain profile gets two different values: 0.5 and 1. The scores of the domain and the expert profiles are used equally, or the scores of the domain profiles are used only.
- *Project-scientist similarity weight:* For finding the similarity between projects and scientists, two different similarity scores, topic based and technology code based, are calculated. To assess the effect of these similarities to the model, different weights are applied. The weights used in experiments are listed below:

Table 7-17: Weights of project-scientist relevancy scores

Code of weights	**Weight of relevancy score based on topics**	**Weight of relevancy score based on technology codes**
TOPIC03_TECH07	0.3	0.7
TOPIC05_TECH05	0.5	0.5
TOPIC07_TECH03	0.7	0.3

- *Scientist-scientist similarity weight:* The similarity between two scientists is calculated based on the research areas and the technology codes. To assess the effect of these similarities, different weights are used during experiments. The weights used are listed below:

Table 7-18: Weights of scientist-scientist similarity scores

Code of weights	**Weight of similarity score based on research areas**	**Weight of similarity score based on technology codes**
RD03_TC07	0.3	0.7
RD05_TC05	0.5	0.5
RD07_TC03	0.7	0.3

The maximum relevancy score is the score that can be given to a topic related to a project or a scientist. The relevancy scores between the topics and the projects/scientists are calculated based on the relevancy degree between a topic and a project/scientist. In the proposed model, the topics are generated and then the relations are calculated. In the generation phase, we could not know or assign the relevancy scores to the topics. But the technology codes are chosen by the scientists for themselves, or by the applicants of the projects for their projects. Because the technology codes are assigned based on the declarations of the scientists or the applicants, they are thought as the top relevant scientists or applicants. While setting profiles based on the technology codes, the maximum relevancy score is calculated using the Equation 5-1. The maximum relevancy scores for CombMNZ and expCombMNZ data fusion techniques are given in the Table 7-19.

Table 7-19: Maximum relevancy scores of the technology codes

Data Fusion Technique	**Maximum Relevancy Score**
CombMNZ	1700
expCombMNZ	3×10^{20}

7.7. Evaluation Strategies

7.7.1. Baseline Method

The baseline method discussed in the Section 5.1, is implemented for the TEYDEB and ARBİS data sets. The scientists are matched with the projects according to the technology codes. The P@n, and R@n values where n can be {5, 10, 15, 20} are given in the Table 7-20. All of the P@n and R@n values are close to 0.

Table 7-20: The results of the experiments of the baseline method

P@5	0.0009	**R@5**	0.0029
P@10	0.0015	**R@10**	0.01
P@ 15	0.0023	**R@15**	0.0209
P@ 20	0.0019	**R@20**	0.0249

7.7.2. Original TEYDEB Data Set

The motivation of the experiments done with the TEYDEB data set is to solve a real-life problem with our proposed model which is proved with a semi-synthetic data set.

According to the t-tests for choosing the optimum number of topics for the experiments, the highest recall values are obtained in the experiments with 200 and 219 topics. The results of the experiments with the original TEYDEB data set are given for only these two values of the number of topics (For all of the results, please refer to APPENDIX G).

The top results obtained from the experiments of the training set are given in the tables Table 7-21, Table 7-22, and Table 7-23. The top results are obtained by a score-based data fusion technique, expCombMNZ. Project-scientist similarity weight do not differ the results. The scores of both the domain (project) and the scientist profiles are used. Although, P@n and R@n values are higher than the values of the baseline method, they are also very close to 0.

Table 7-21: The top P@5 and R@5 results of the experiments of the training set

Number of Topics	Stemming is Used	Weight Model	Data Fusion Technique	Project-Scientist Similarity Weight	Weight of the Domain Profile	P@5	R@5
200	No	DLH13	EXPCOMBMNZ	TOPIC03_TECH07, TOPIC05_TECH05, TOPIC07_TECH03	0.5	0.0089	0.0240
200	Yes	DLH13	EXPCOMBMNZ	TOPIC03_TECH07, TOPIC05_TECH05, TOPIC07_TECH03	0.5	0.0085	0.0267

Table 7-22: The top P@10 and R@10 results of the experiments of the training set

Number of Topics	Stemming is Used	Weight Model	Data Fusion Technique	Project-Scientist Similarity Weight	Weight of the Domain Profile	P@10	R@10
200	No	DLH13	EXPCOMBMNZ	TOPIC03_TECH07, TOPIC05_TECH05, TOPIC07_TECH03	0.5	0.0053	0.0287
200	Yes	DLH13	EXPCOMBMNZ	TOPIC03_TECH07, TOPIC05_TECH05, TOPIC07_TECH03	0.5	0.0063	0.0353

Table 7-23: The top P@15 and R@15 results of the experiments of the training set

Number of Topics	Stemming is Used	Weight Model	Data Fusion Technique	Project-Scientist Similarity Weight	Weight of the Domain Profile	P@15	R@15
200	No	DLH13	EXPCOMBMNZ	TOPIC03_TECH07, TOPIC05_TECH05, TOPIC07_TECH03	0.5	0.0054	0.0467
200	Yes	DLH13	EXPCOMBMNZ	TOPIC03_TECH07, TOPIC05_TECH05, TOPIC07_TECH03	0.5	0.0062	0.0519

The evaluation measures are compared in the Figure 7-5. Although, all of the values are close to 0, they are increased by the proposed model.

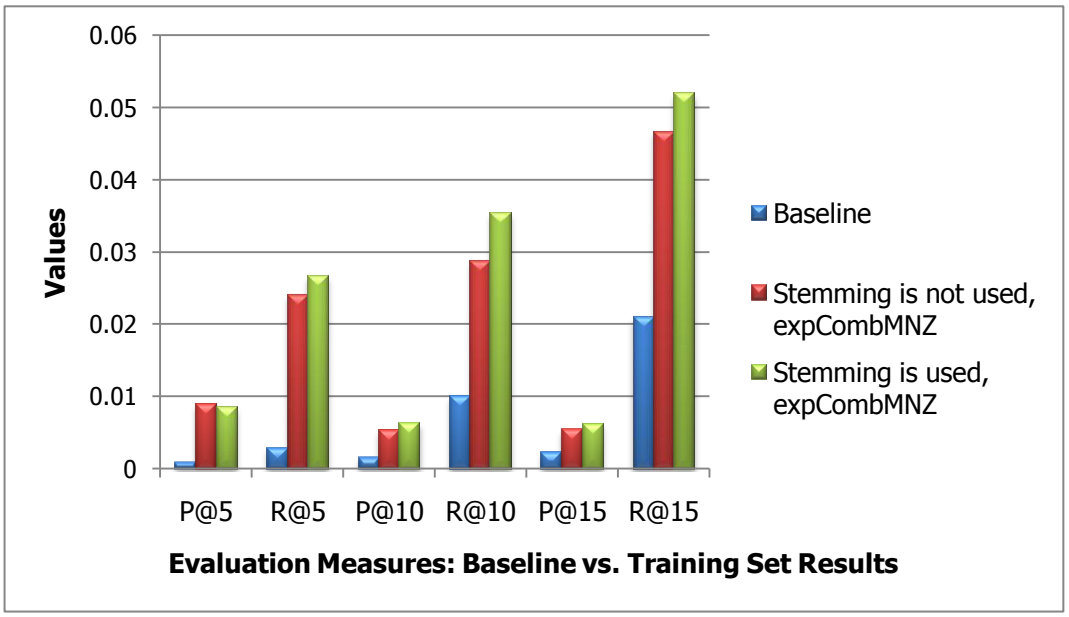


Figure 7-5: Comparison of the evaluation measures between the baseline method and the proposed model on the training set

The top results obtained from the experiments of the test set are given in the tables, Table 7-25, and Table 7-26. The top results are obtained by a score-based data fusion technique, expCombMNZ and a rank-based data fusion technique, Reciprocal Rank. Project-scientist similarity weight do not differ the results as in the training set experiments. In the top results, when Reciprocal Rank is used, the ranks of the domain profiles are used only. But with the expCombMNZ, the scores of the both the domain (project) and the scientist profiles are used. Although, P@n and R@n values are higher than the values of the baseline method, they are also very close to 0.

Table 7-24: The top P@5 and R@5 results of the experiments of the test set

Number of Topics	Stemming is Used	Data Fusion Technique	Project-Scientist Similarity Weight	Weight of the Domain Profile	P@5	R@5
219	Yes	RR	TOPIC03_TECH07, TOPIC05_TECH05, TOPIC07_TECH03	1	0.0065	0.0149
200	No	EXPCOMBMNZ	TOPIC03_TECH07, TOPIC05_TECH05, TOPIC07_TECH03	0.5	0.0033	0.0122

Table 7-25: The top P@10 and R@10 results of the experiments of the test set

Number of Topics	Stemming is Used	Data Fusion Technique	Project-Scientist Similarity Weight	Weight of the Domain Profile	P@10	R@10
219	Yes	RR	TOPIC03_TECH07, TOPIC05_TECH05, TOPIC07_TECH03	1	0.0049	0.0271
200	No	EXPCOMBMNZ	TOPIC03_TECH07, TOPIC05_TECH05, TOPIC07_TECH03	0.5	0.0024	0.0163

Table 7-26: The top P@15 and R@15 results of the experiments of the test set

Number of Topics	Stemming is Used	Data Fusion Technique	Project-Scientist Similarity Weight	Weight of the Domain Profile	P@15	R@15
219	Yes	RR	TOPIC03_TECH07, TOPIC05_TECH05, TOPIC07_TECH03	1	0.0043	0.0352
200	No	EXPCOMBMNZ	TOPIC03_TECH07, TOPIC05_TECH05, TOPIC07_TECH03	0.5	0.0022	0.0203

The evaluation measures are compared in the Figure 7-6. Except the values P@15 and R@15, all of the values are increased by the proposed model. As in the experiments of the training set, the values obtained by the test set are also very close to 0.

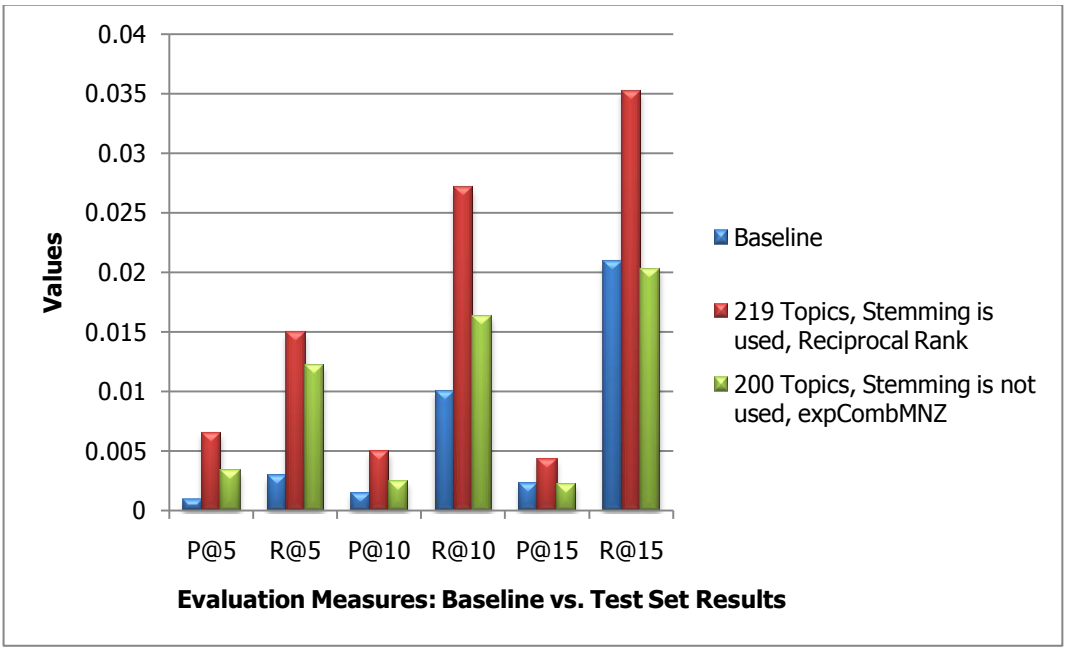


Figure 7-6: Comparison of the evaluation measures between the baseline method and the proposed model on the test set

7.7.3. Subsampled TEYDEB Data Set

The results of the proposed model are increased with respect to baseline method, but the measures cannot be acceptable for an expert finding system, for an information retrieval system. The reason of the low measures is thought as that of the ground truth set. While the proposed model deals with the research areas of the scientists, TEYDEB specialists do not only consider the keywords of the research areas. The points that TEYDEB specialists consider while matching scientists and projects can be listed as follows:

- Research areas of the scientists should be appropriate for the projects to be matched. But a scientist with a lot of research area may not be matched to any projects. The knowledge level of a scientist with a lot of research area may be considered inadequate for evaluating a project.
- Previous evaluation performance of a scientist in other TEYDEB projects is satisfactory.
- Scientists who have evaluated similar projects previously are taken into account.

- Scientists who are working for a university in the same city or the same region with the applicant organization may be primarily preferred.
- Scientists who have related research areas with the projects and who are not evaluated any TEYDEB project previously are preferred to extend the evaluator set.

The ground truth set could not be valid for evaluating the proposed model according to the criteria of TEYDEB specialists. As a result, three TEYDEB specialists rate the scientist sets for the selected projects for generating a valid ground truth set.

We choose thirty five projects in which evaluation of the proposed model is unsuccessful. Most of the selected projects could not be matched to any scientist which are in the ground truth set. For every project, a different scientist set is generated. Every scientist set is consisted of the research area keywords of,

- The first twenty scientists which are matched by the proposed model,
- The scientists which are matched by the TEYDEB specialists in the ground truth set,
- Twenty scientists which are chosen randomly.

The raters, TEYDEB specialists, choose the most related five scientists with every project. After the ratings, Kappa statistics is applied to show if there is an agreement or not. For applying Kappa statistics, Cohen's Kappa for more than two annotators with multiple classes developed by Jeroen Geertzen is used [13].

The Kappa-value (K) may range from -1 to 1, where K=-1, shows a total disagreement; K=1 shows a total agreement. All of the K values in the results are greater than 0; K value range from 0.11 to 0.84 in this experiment. According to the results, the projects have K-values lower than 0.5 are excluded from the set. K-values for the two projects (K=0.48, 0.49), are very close to 0.5, so they are included to the set.

Table 7-27: Agreement scores of the Kappa statistics for the chosen projects

Agreement Scores				
Variable	**Evaluator1+Evaluator2**	**Evaluator1+Evaluator3**	**Evaluator2+Evaluator3**	**Average Kappa Values**
Project1	PA=0.80 PE=0.16 K=0.76	PA=0.80 PE=0.16 K=0.76	PA=1.00 PE=0.20 K=1.00	K=0.84 / 15 pairs
Project3	PA=0.80 PE=0.16 K=0.76	PA=0.60 PE=0.12 K=0.55	PA=0.60 PE=0.12 K=0.55	K=0.62 / 15 pairs
Project4	PA=0.80 PE=0.16 K=0.76	PA=0.60 PE=0.12 K=0.55	PA=0.60 PE=0.12 K=0.55	K=0.62 / 15 pairs
Project7	PA=0.80 PE=0.16 K=0.76	PA=0.80 PE=0.16 K=0.76	PA=0.60 PE=0.16 K=0.52	K=0.68 / 15 pairs
Project9	PA=1.00 PE=0.20 K=1.00	PA=0.40 PE=0.08 K=0.35	PA=0.40 PE=0.08 K=0.35	K=0.57 / 15 pairs
Project12	PA=0.60 PE=0.12 K=0.55	PA=0.40 PE=0.08 K=0.35	PA=0.80 PE=0.16 K=0.76	K=0.55 / 15 pairs
Project15	PA=1.00 PE=0.20 K=1.00	PA=0.80 PE=0.16 K=0.76	PA=0.80 PE=0.16 K=0.76	K=0.84 / 15 pairs
Project16	PA=0.80 PE=0.16 K=0.76	PA=0.40 PE=0.08 K=0.35	PA=0.40 PE=0.08 K=0.35	K=0.49 / 15 pairs
Project17	PA=0.60 PE=0.12 K=0.55	PA=0.60 PE=0.12 K=0.55	PA=0.80 PE=0.16 K=0.76	K=0.62 / 15 pairs
Project18	PA=0.80 PE=0.16 K=0.76	PA=0.80 PE=0.16 K=0.76	PA=0.80 PE=0.16 K=0.76	K=0.76 / 15 pairs
Project20	PA=0.80 PE=0.16 K=0.76	PA=0.40 PE=0.08 K=0.35	PA=0.60 PE=0.12 K=0.55	K=0.55 / 15 pairs
Project22	PA=0.80 PE=0.16 K=0.76	PA=0.60 PE=0.12 K=0.55	PA=0.40 PE=0.08 K=0.35	K=0.55 / 15 pairs
Project23	PA=0.60 PE=0.12 K=0.55	PA=0.60 PE=0.12 K=0.55	PA=1.00 PE=0.20 K=1.00	K=0.70 / 15 pairs
Project25	PA=0.80 PE=0.16 K=0.76	PA=0.60 PE=0.12 K=0.55	PA=0.80 PE=0.16 K=0.76	K=0.69 / 15 pairs
Project29	PA=0.80 PE=0.16 K=0.76	PA=0.60 PE=0.12 K=0.55	PA=0.60 PE=0.12 K=0.55	K=0.62 / 15 pairs
Project32	PA=0.60 PE=0.12 K=0.55	PA=0.60 PE=0.12 K=0.55	PA=1.00 PE=0.20 K=1.00	K=0.70 / 15 pairs
Project33	PA=0.40 PE=0.08 K=0.35	PA=0.60 PE=0.12 K=0.55	PA=0.60 PE=0.12 K=0.55	K=0.48 / 15 pairs
Project34	PA=0.80 PE=0.16 K=0.76	PA=1.00 PE=0.20 K=1.00	PA=0.80 PE=0.16 K=0.76	K=0.84 / 15 pairs
Project35	PA=1.00 PE=0.20 K=1.00	PA=0.60 PE=0.12 K=0.55	PA=0.60 PE=0.12 K=0.55	K=0.70 / 15 pairs

There are 19 projects on which agreement is provided. The three raters agreed on 54% of the projects approximately. Agreement scores of the chosen projects are given in the Table 7-27 (For all of the agreement scores, please refer to APPENDIX H). In the results,

- PA shows the observed proportion of agreement,
- PE shows the proportion of agreement expected by chance,
- K shows the Kappa-value.

For the chosen 19 projects, the chosen scientist sets are generated from the choices of the raters. The scientists, at least two raters agreed on, are included to the set of related scientists to generate a new research-area-based ground truth set.

Table 7-28: Comparison of the number of scientists assigned to the projects

Project	**Number of scientists - agreed on by the raters**	**Number of scientists - formerly assigned**	**Number of scientists - both assigned formerly and agreed on by the raters**
Project1	5	2	0
Project3	4	1	0
Project4	4	1	0
Project7	6	3	3
Project9	5	2	2
Project12	5	1	1
Project15	5	2	2
Project16	4	3	0
Project17	4	2	1
Project18	4	3	2
Project20	5	1	0
Project22	5	2	0
Project23	5	1	1
Project25	5	3	0
Project29	4	1	1
Project32	5	1	0
Project33	4	1	1
Project34	5	1	1
Project35	5	1	0

In Table 7-28, the number of scientists assigned to the projects is given. For the chosen 19 projects,

- Number of scientists who are chosen by the raters,
- Number of scientists who are formerly assigned to the projects by TEYDEB specialists

are given. Also the number of scientists who are overlapped in both of the sets that are chosen by raters and assigned by TEYDEB specialists is given in Table 7-28. In 9 of the 19 projects, different scientists are chosen by raters. In 8 of the 19 projects, the chosen scientists by raters include all of the scientists formerly assigned by TEYDEB specialists.

We repeat the evaluations of experiments with the research-area-based ground truth set. The top results of the experiments done by 200 and 219 topics are given in the tables, Table 7-29, Table 7-30, Table 7-31, Table 7-32 for different P@n and R@n values (For all of the results, please refer to APPENDIX I).

The top results represent taht the P@n and R@n values are satisfiable in the experiments with the subsampled data set than the experiments with the original TEYDEB data set.

Table 7-29: The results of the experiments evaluated with subsampled data set, P@2, R@2

Number of Topics	**Stemming is Used**	**Data Fusion Technique**	**Project-Scientist Similarity Weight**	**Weight of the Domain Profile**	**P@2**	**R@2**
200	No	EXPCOMBMNZ	TOPIC03_TECH07, TOPIC05_TECH05, TOPIC07_TECH03	1	0.2368	0.2368
219	Yes	EXPCOMBMNZ	TOPIC03_TECH07, TOPIC05_TECH05, TOPIC07_TECH03	0.5	0.1842	0.1842

Table 7-30: The results of the experiments evaluated with subsampled data set, P@3, R@3

Number of Topics	Stemming is Used	Data Fusion Technique	Project-Scientist Similarity Weight	Weight of the Domain Profile	P@3	R@3
200	No	EXPCOMBMNZ	TOPIC03_TECH07, TOPIC05_TECH05, TOPIC07_TECH03	1	0.2105	0.2105
219	Yes	EXPCOMBMNZ	TOPIC03_TECH07, TOPIC05_TECH05, TOPIC07_TECH03	0.5	0.1404	0.1404

Table 7-31: The results of the experiments evaluated with subsampled data set, P@4, R@4

Number of Topics	Stemming is Used	Data Fusion Technique	Project-Scientist Similarity Weight	Weight of the Domain Profile	P@4	R@4
200	No	EXPCOMBMNZ	TOPIC03_TECH07, TOPIC05_TECH05, TOPIC07_TECH03	1	0.1842	0.1842
219	Yes	EXPCOMBMNZ	TOPIC03_TECH07, TOPIC05_TECH05, TOPIC07_TECH03	0.5	0.1316	0.1316

Table 7-32: The results of the experiments evaluated with subsampled data set, P@5, R@5

Number of Topics	Stemming is Used	Data Fusion Technique	Project-Scientist Similarity Weight	Weight of the Domain Profile	P@5	R@5
200	No	EXPCOMBMNZ	TOPIC03_TECH07, TOPIC05_TECH05, TOPIC07_TECH03	1	0.1789	0.1921
219	Yes	EXPCOMBMNZ	TOPIC03_TECH07, TOPIC05_TECH05, TOPIC07_TECH03	0.5	0.1053	0.1105

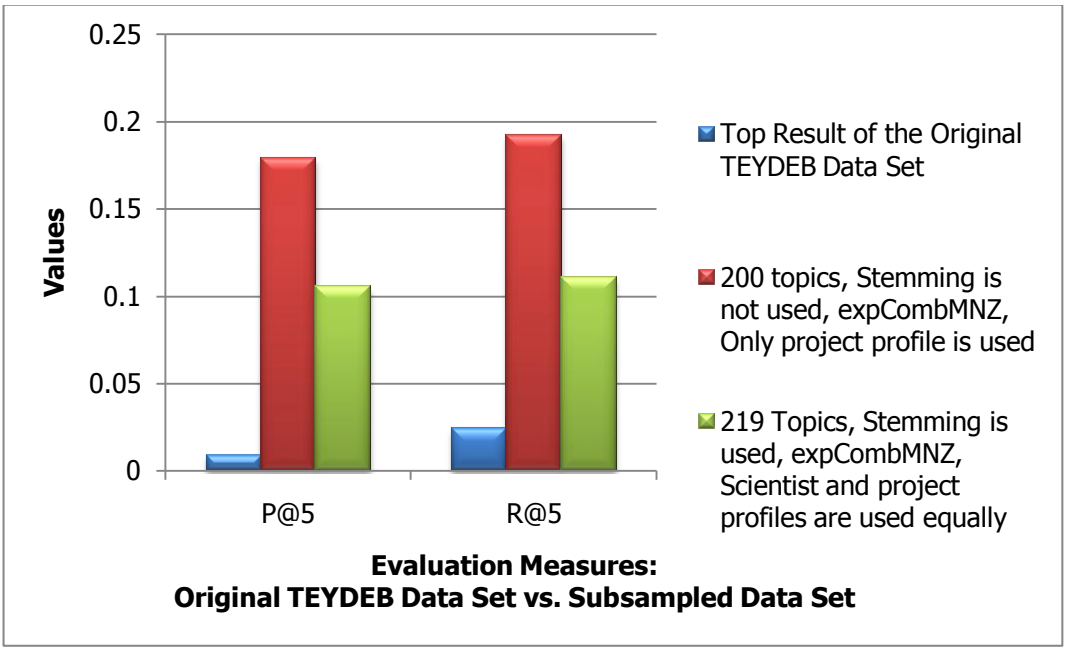


Figure 7-7: Comparison of the results – Original TEYDEB data set vs. subsampled data set

In Figure 7-7, P@5 and R@5 values are compared for the baseline method and the proposed model. Both of the top results are gained with the data fusion technique expCombMNZ. With the new ground truth set, precision and recall values are increased. Higher precision and recall values are obtained in the experiment in which;

- Number of topics, 200,
- A score based data fusion technique, expCombMNZ is used,
- Stemming is not used,
- Only domain (project) profile scores are used.

In most of the cases, the project-scientist similarity weight does not affect the evaluation measures.

7.7.4. Finding Similar Experts on Subsampled TEYDEB Data Set

"Finding Similar Experts" task is executed for the topics which are generated without using a stemming algorithm. The data fusion technique, expCombMNZ is used for the experiments because of the highest performance. The project-scientist

similarity weight do not affect the results in the experiments with the subsampled data set, so "TOPIC05_TECH05" is used for the experiments of finding similar experts. Same as all of the experiments, DLH13 is used for the weight model. The top results of the experiments are given in the Table 7-33. The number of selected experts shows the number of the experts which are selected to be matched to the similar experts. As the number of similar experts, two different values as 5 and 10 are used. But all of the top results are obtained when 10 similar experts are chosen.

Table 7-33: Top results of the finding similar experts process

Number of Topics	**Stemming is Used**	**Scientist-Scientist Similarity Weight**	**Weight of the Domain Profile**	**Number of Similar Experts**	**P@2**	**R@2**	**P@3**	**R@3**
200	No	RD03_TC07	1	10	0.0263	0.0263	0.0175	0.0175
200	No	RD07_TC03	1	10	0.0263	0.0263	0.0175	0.0175
200	Yes	RD03_TC07	1	10	0.0263	0.0263	0.0175	0.0175
200	Yes	RD05_TC05	1	10	0.0263	0.0263	0.0175	0.0175
200	Yes	RD07_TC03	1	10	0.0263	0.0263	0.0175	0.0175

All of the measures have decreased about 10% against the experiments of the proposed model with the subsampled TEYDEB data set. In most of the experiments except the top resulted ones, the P@n and R@n are 0.

While the similarities between experts are calculated by different three weights for the scientist-scientist similarity, the different weights have no effect on the results.

CHAPTER 8

RESULTS AND DISCUSSION

The proposed model is implemented for two different corpora, for the semi-synthetic data set and for the TEYDEB and ARBİS data sets.

In the experiments of the semi-synthetic data set, 93% of the papers (domains) are matched to the right keywords (experts). The semi-synthetic data set is in English. Although we expected that using Porter's stemming algorithm increases the evaluation measures, using Porter's stemming algorithm or not using it do not differ the results as expected. When Porter's stemming algorithm is used, the performance of matching the papers to the keywords is decreased from 93% to 86%. Although the decrease of the performance, both of the results are very successful to prove our concept.

The proposed model is also implemented for a real-life problem in two different conditions:

- For the original TEYDEB data set: The scientists assigned to the projects by TEYDEB specialists are used as the ground truth set.
- For the subsampled TEYDEB data set: Appropriate scientists are rated by three TEYDEB specialists for the chosen 35 projects. If there is an agreement on the rated scientists for a project then the project and the rated scientists are used for the ground truth set.

The results of the experiments with the original TEYDEB data set are not acceptable. The reason of the unsuccessful results is the different matching strategies of the proposed model and TEYDEB specialists. The scientists assigned to the projects by TEYDEB specialists are used as the ground truth set. TEYDEB specialists consider various criteria for matching the scientists to the projects, so the ground truth set which is conducted by TEYDEB specialists is not appropriate for the evaluation. A research-area-based ground truth set is generated using the rates of the three TEYDEB specialists. With the new research-area-based ground truth set, the proposed model performs better than the baseline method. Using the subsampled TEYDEB data set, 57% of the projects are assigned to the right scientists. Although the performance of the proposed model on subsampled TEYDEB data set is acceptable, it is not successful as the POC work.

The TEYDEB data set which is used as a case study contains projects about information systems. Although the domains of the projects are similar, most of the projects have different scopes from each other. For instance, a project for which the proposed model could not match the right scientists is about automation of manufacturing raw material. This project contains keywords as "üretim", "montaj". But these keywords are not used often in the TEYDEB data set. So the probability of these keywords used in the topics is very low. For instance, although the keyword, "üretim" is a frequently-used keyword in this project, it does not appear in any of the topics. As a result, this project could be matched to the scientists over more general keywords as "yazılım", "otomasyon", "bilgi".

Although the choices of the three raters based on keywords for the Kappa statistics, in some choices the information about a project do not contain the keywords of the chosen scientist. A project about "elektronik süreç yönetimi" is matched to a scientist who has 9 keywords and only a keyword "yönetim bilişim sistemleri" matches with the terms of the project information. The project and the keyword are related on the terms "yönetim" and "sistem", which are not specific for the project. The proposed model set profiles and relations between the projects and the scientists based on the terms. In this case, a relation based on two terms is

evaluated as a weak relation by the proposed model and the proposed model is not successful.

On the other hand, the proposed model matched the projects to the scientists successfully. A project with a main scope "gömülü sistemler", contains the terms "iletişim protokolleri", "gömülü sistem", "şifreleme", "haberleşme" frequently. These terms are specific to the domains and could be present in the research area keywords of the scientists who are studying or working in this domain. This project is matched to the right scientists successfully by the proposed model.

Additionally, some of the evaluation parameters do not affect the results of the experiments. In most of the experiments, the project-scientist similarity weight does not affect the evaluation measures. The project-scientist similarity weight is used to evaluate the effects of the technology codes and the generated topics on the model. Although this weight do not differ the results mostly, we can say that the technology codes cannot describe the projects or the scientists sufficiently. Because, the baseline method is based on the technology codes and the results of the experiments are very low.

In all of the experiments on the finding similar experts process, the scientist-scientist similarity weight does not also affect the measures. Evaluation strategy of the finding similar experts task could be different. The raters only choose five scientists from a set. Using the finding similar experts task, we extend the set of matched scientists, but the set of the scientists chosen by raters do not extend. As a result, all of the measures are decreased 10% as expected.

Most of the successful results are obtained by the score based data fusion technique, expCombMNZ. While calculating the scores of the terms;

- The frequency of the term,
- The document frequency of the term,
- The term frequency in the document and in the corpus,

are used. These measures do not affect the rank of a term.

Language of the corpus is also important for the performance of the system. Depending on the domain, the keywords of a domain may vary. More successful results are obtained in the English corpus. This can be caused from several reasons. For the information systems domain, most of the keywords in English do not have a strict translation in Turkish. Also the corpus used for the POC in English is defined in a structured way. If the research area keywords of the scientists and the information about projects are defined more clearly, the performance of the proposed model could increase.

CHAPTER 9

CONCLUSIONS AND FUTURE WORK

9.1. Conclusion

In this study, an expert finding system is proposed for the domains with unclear topics. A voting approach is used as a baseline method. Baseline method is improved by generating topics from the domains. The achievement of the proposed model is proved by a semi-synthetic data set. The proposed model is also applied to TEYDEB and ARBİS data sets as a case study. The proposed model performs better than the baseline method for both proof of concept and case study.

A hidden semantic level is built between the domains and the experts. The hidden semantic level is formed by extracting explanatory and clear topics from the domains. LDA and Gibbs sampling are used which are competitive with other topic generation methods. To the best of our knowledge, there is no previous study that uses LDA and Gibbs sampling algorithm with the voting approach for expert finding.

Unlikely to the traditional expert finding systems more than one viewpoint is used which are pre-defined and generated topics.

The proposed model is extended by the finding similar experts task. After finding the most relevant experts for the domains, the finding similar experts task targets to recommend the experts with similar qualifications.

Our proposed model is applied to a real life problem. TEYDEB and ARBİS data sets are used to match the projects with the relevant scientists based on textual evidences.

TEYDEB and ARBİS data sets which are the data sets for the case study cannot be used directly for the proposed model. Although some scientists registered to the ARBİS do not have any information, they may be assigned to the projects. ARBİS and TEYDEB data sets are preprocessed before applying the proposed model. On the other hand, the assignments between the projects and the scientist are not appropriate for the performance evaluation of our study. TEYDEB specialists consider much more criteria than the matching of the research area keywords. Finally, we generate a new ground truth set for the performance evaluation.

9.2. Future Work

In this study, the proposed expert finding system uses textual evidences for matching the domains and the experts. Textual evidences are extracted from the supporting documents of the domains and the experts. In reality, we may consider factors which are not textual sources. For instance, up-to-dateness in a topic, accessibility of the experts may be also important. A major consideration for the future may be extending the proposed model using contextual factors [16].

Another extension point may be extending the generated topics with a dictionary. Extending topics may increase the number of the matched experts. But the set of matched experts may consist of more specialized set of experts could be formed.

REFERENCES

[1] ACM, Inc., Association for Computing Machinery. (2012). *ACM Computing Classification System*. Retrieved February 4, 2012, from ACM Digital Library: http://dl.acm.org/ccs.cfm?CFID=64838794&CFTOKEN=63313451

[2] Alias-i. (2011). LingPipe Home. Retrieved August 29, 2011, from Alias-i: http://alias-i.com/lingpipe/

[3] Bailey, P., Vries, A. P., Craswell, N., & Soboroff, I. (2007). Overview of the TREC 2007 Enterprise Track. Proceedings of the Fourteenth Text REtrieval Conference (TREC 2007). National Institute of Standards and Technology (NIST).

[4] Balog, K., & Rijke, M. d. (2007). Finding Similar Experts. Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2007) (pp. 821-822). Amsterdam: ACM.

[5] Balog, K., Azzopardi, L., & Rijke, M. d. (2006). Formal Models for Expert Finding in Enterprise Corpora. Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2006) (pp. 43-50). Seattle, Washington: ACM.

[6] Balog, K., Bogers, T., Azzopardi, L., de Rijke, M., & van den Bosch, A. (2007). Broad expertise retrieval in sparse data environments. Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2007) (pp. 551-558). Amsterdam: ACM.

[7] Balog, K., Thomas, P., Craswell, N., Soboroff, I., Bailey, P., & Vries, A. P. (2009). Overview of the TREC 2008 Enterprise Track. Proceedings of the Seventeenth Text REtrieval Conference (TREC 2008). NIST.

[8] Cao, Y., Liu, J., Bao, S., & Li, H. (2005). Research on Expert Search at Enterprise Track of TREC 2005. Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005). National Institute of Standards and Technology (NIST).

[9] Carpenter, B., & Baldwin, B. (June 2011). Text Analysis with LingPipe 4. LingPipe Inc.

[10] Craswell, N., Vries, A. P., & Soboroff, I. (2005). Overview of the TREC-2005 Enterprise Track. Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005). Gaithersburg, Maryland: National Institute of Standards and Technology (NIST).

[11] Das, S., Mitra, P., & Giles, C. L. (2011). Ranking Authors in Digital Libraries. Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries (JCDL 2011) (pp. 251-254). New York: ACM.

[12] Fang, H., & ChengXiang, Z. (2007). Probabilistic Models for Expert Finding. Proceedings of the 29th European conference on IR research (ECIR 2007) (pp. 418-430). Rome: Springer-Verlag.

[13] Geertzen, J. (2010). Cohen's Kappa for more than two annotators. Retrieved February 2, 2012, from Jeroen Geertzen: http://cosmion.net/jeroen/software/kappao/

[14] Goswami, P. (n.d.). Batch (TREC) Indexing and Retrieval using Terrier 2.2.1. Retrieved August 29, 2011, from www.isical.ac.in/~mtc0907/TerrierGuide.pdf

[15] Griffiths, T. L., & Steyvers, M. (2004). Finding Scientific Topics. Proceedings of the National Academy of Sciences of the United States of America , 101, 5228-5235.

[16] Hoffman, K., Balog, K., Bogers, T., & Rijke, M. d. (2010). Contextual Factors for Finding Similar Experts. Journal of American Society for Information Science and Technology , 61, 994-1014.

[17] Kongthon, A., Haruechaiyasak, C., & Thaiprayoon, S. (2009). Expert Identification for Multidisciplinary R&D Project Collaboration. Management of Engineering & Technology, 2009. PICMET 2009. Portland International Conference on, (pp. 1474-1480). Portland, OR.

[18] MacDonald, C., & Ounis, I. (2006). Voting for Candidates: Adapting Data Fusion Techniques for an Expert Search Task. Proceedings of the 15th ACM international conference on Information and knowledge management (CIKM 2006) (pp. 387-396). Arlington, Virginia: ACM.

[19] McDonald, D. W., & Ackerman, M. S. (2000). Expertise recommender: a flexible recommendation system and architecture. Proceedings of the 2000 ACM conference on Computer supported cooperative work (CSCW 2000) (pp. 231-240). Philadelphia, Pennsylvania: ACM.

[20] Petkova, D., & Croft, W. C. (2006). Hierarchical Language Models for Expert Finding in Enterprise Corpora. Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2006) (pp. 599-608). IEEE Computer Society.

[21] Shami, N. S., Ehrlich, K., & Millen, D. R. (2008). Pick me!: Link Selection in Expertise Search Results. Conference on Human Factors in Computing Systems (CHI 2008) (pp. 1089-1092). Florence: ACM.

[22] Smirnova, E., & Balog, K. (2011). A User-Oriented Model for Expert Finding. Proceedings of the 33rd European conference on Advances in information retrieval (ECIR 2011) (pp. 580-592). Dublin: Springer-Verlag.

[23] Soboroff, I., & Craswell, N. (2006). Overview of the TREC 2006 Enterprise Track. Proceedings of the Fifteenth Text REtrieval Conference (TREC 2006). National Institute of Standards and Technology (NIST).

[24] Steyvers, M., & Griffiths, T. (2007). Probabilistic Topic Models. In Latent Semantic Analysis: A Road to Meaning. Laurence Erlbaum.

[25] The Natural Languge Processing Group at Fatih University. (2012). Turkish Stop Word List 1.1. Retrieved August 29, 2011, from The Natural Languge Processing Group: http://nlp.ceng.fatih.edu.tr/?p=101

[26] Türkiye Bilimsel ve Teknolojik Araştırma Kurumu (TÜBİTAK). (2012). TÜBİTAK - TEYDEB. Retrieved August 29, 2011, from Türkiye Bilimsel ve Teknolojik Araştırma Kurumu: www.teydeb.tubitak.gov.tr

[27] Türkiye Bilimsel ve Teknolojik Araştırma Kurumu (TÜBİTAK). (2008). TÜBİTAK TEYDEB Proje Değerlendirme ve İzleme Sistemi - PRODİS. Retrieved August 29, 2011, from Türkiye Bilimsel ve Teknolojik Araştırma Kurumu (TÜBİTAK): https://eteydeb.tubitak.gov.tr

[28] Türkiye Bilimsel ve Teknolojik Araştırma Kurumu. (2012). TEYDEB Teknoloji Kodları. Retrieved February 1, 2012, from Türkiye Bilimsel ve Teknolojik Araştırma Kurumu: http://www.tubitak.gov.tr/sid/481/pid/478/cid/3761/index.htm

[29] University of Glasgow, School of Computing Science, Terrier Team. (2011). Terrier Information Retrieval Platform Homepage. Retrieved August 29, 2011, from Terrier: http://terrier.org/

[30] Wei, W., Barnaghi, P., & Bargiela, A. (2010). Probabilistic Topic Models for Learning Terminological Ontologies. IEEE Transactions on Knowledge and Data Engineering , 22, 1028-1040.

[31] Woudstra, L., & Hooff, B. v. (2008). Inside the Source Selection Process: Selection Criteria for Human Information Sources. Information Processing and Management , 44, 1267-1278.

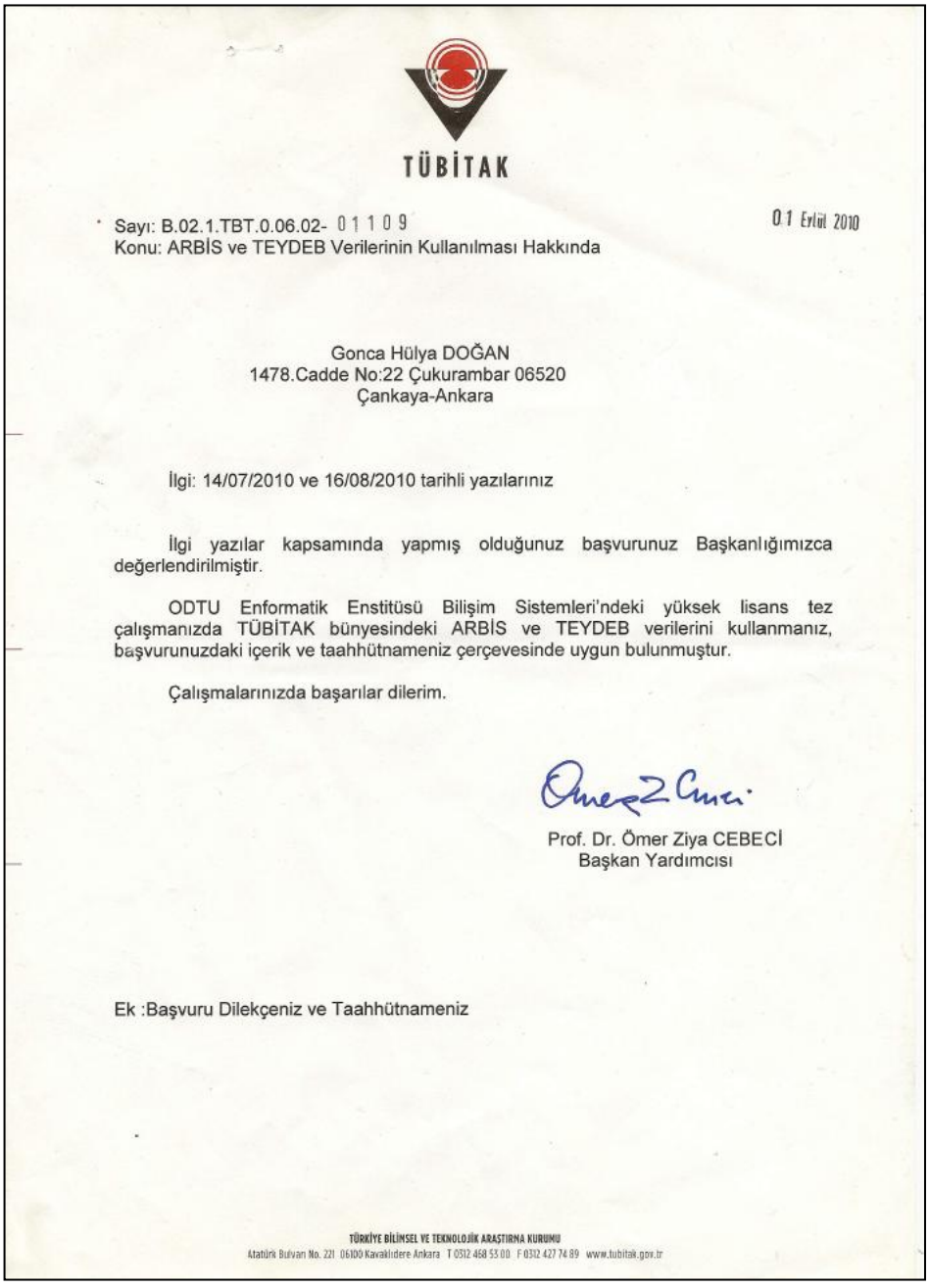
[32] Yimam-seid, D., & Kobsa, A. (2000). Expert Finding Systems for Organizations: Problem and Domain Analysis and the DEMOIR Approach. Beyond Knowledge Management: Management Expertise Workshop (ECSCW 1999). 13, pp. 276-283. MIT Press.

[33] Zhai, C., & Lafferty, J. (2001). A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval. Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2001) (pp. 334-342). New Orleans, Louisiana: ACM.

[34] Zhang, J., Tang, J., Liu, L., & Li, J. (2008). A Mixture Model for Expert Finding. Proceedings of the 12th Pacific-Asia conference on Advances in knowledge discovery and data mining (PAKDD 2008) (pp. 466-478). Osaka: Springer-Verlag.

APPENDICES

APPENDIX A – Ethics Clearance



TÜBİTAK

Sayı: B.02.1.TBT.0.06.02- 01109
Konu: ARBİS ve TEYDEB Verilerinin Kullanılması Hakkında

01 Eylül 2010

Gonca Hülya DOĞAN
1478.Cadde No:22 Çukurambar 06520
Çankaya-Ankara

İlgi: 14/07/2010 ve 16/08/2010 tarihli yazılarınız

İlgi yazılar kapsamında yapmış olduğunuz başvurunuz Başkanlığımızca değerlendirilmiştir.

ODTU Enformatik Enstitüsü Bilişim Sistemleri'ndeki yüksek lisans tez çalışmanızda TÜBİTAK bünyesindeki ARBİS ve TEYDEB verilerini kullanmanız, başvurunuzdaki içerik ve taahhütnameniz çerçevesinde uygun bulunmuştur.

Çalışmalarınızda başarılar dilerim.

Prof. Dr. Ömer Ziya CEBECİ
Başkan Yardımcısı

Ek :Başvuru Dilekçeniz ve Taahhütnameniz

TÜRKİYE BİLİMSEL VE TEKNOLOJİK ARAŞTIRMA KURUMU
Atatürk Bulvarı No. 221 06100 Kavaklıdere Ankara T 0312 468 53 00 F 0312 427 74 89 www.tubitak.gov.tr

Table B-1: The data set used for the POC of Gibbs sampling

Title	**Abstract**
Finding Similar Experts	The task of finding people who are experts on a topic has recently received increased attention. We introduce a different expert finding task for which a small number of example experts is given (instead of a natural language query), and the system's task is to return similar experts. We define, compare, and evaluate a number of ways of representing experts, and investigate how the size of the initial example set affects performance. We show that more finegrained representations of candidates result in higher performance, and larger sample sets as input lead to improved precision.
Formal Models for Expert Finding in Enterprise Corpora	Searching an organization's document repositories for experts provides a cost effective solution for the task of expert finding. We present two general strategies to expert searching given a document collection which are formalized using generative probabilistic models. The first of these directly models an expert's knowledge based on the documents that they are associated with, whilst the second locates documents on topic, and then finds the associated expert. Forming reliable associations is crucial to the performance of expert finding systems. Consequently, in our evaluation we compare the different approaches, exploring a variety of associations along with other operational parameters (such as topicality). Using the TREC Enterprise corpora, we show that the second strategy consistently outperforms the first. A comparison against other unsupervised techniques, reveals that our second model delivers excellent performance.

Table B-1 (continued)

Contextual Factors for Finding Similar Experts	Expertise-seeking research studies how people search for expertise and choose whom to contact in the context of a specific task. An important outcome are models that identify factors that influence expert finding. Expertise retrieval addresses the same problem, expert finding, but from a system-centered perspective.The main focus has been on developing content-based algorithms similar to document search. These algorithms identify matching experts primarily on the basis of the textual content of documents with which experts are associated. Other factors, such as the ones identified by expertise-seeking models, are rarely taken into account. In this article, we extend content-based expert-finding approaches with contextual factors that have been found to influence human expert finding. We focus on a task of science communicators in a knowledge-intensive environment, the task of finding similar experts, given an example expert. Our approach combines expertise-seeking and retrieval research. First, we conduct a user study to identify contextual factors that may play a role in the studied task and environment. Then, we design expert retrieval models to capture these factors. We combine these with content-based retrieval models and evaluate them in a retrieval experiment. Our main finding is that while content-based features are the most important, human participants also take contextual factors into account, such as media experience and organizational structure. We develop two principled ways of modeling the identified factors and integrate them with content-based retrieval models. Our experiments show that models combining content-based and contextual factors can significantly outperform existing content-based models.

Table B-1 (continued)

Voting for Candidates: Adapting Data Fusion Techniques for an Expert Search Task	In an expert search task, the users' need is to identify people who have relevant expertise to a topic of interest. An expert search system predicts and ranks the expertise of a set of candidate persons with respect to the users' query. In this paper, we propose a novel approach for predicting and ranking candidate expertise with respect to a query. We see the problem of ranking experts as a voting problem, which we model by adapting eleven data fusion techniques. We investigate the effectiveness of the voting approach and the associated data fusion techniques across a range of document weighting models, in the context of the TREC 2005 Enterprise track. The evaluation results show that the voting paradigm is very effective, without using any collection specific heuristics. Moreover, we show that improving the quality of the underlying document representation can significantly improve the retrieval performance of the data fusion techniques on an expert search task. In particular, we demonstrate that applying field-based weighting models improves the ranking of candidates. Finally, we demonstrate that the relative performance of the adapted data fusion techniques for the proposed approach is stable regardless of the used weighting models.

APPENDIX C – POC of Expert Finding in Domains with Unclear Topics: Semi-Synthetic Data Set

Table C-1: Semi-synthetic data set

Predefined Topic By ACM	**Title**	**Abstract**	**Authors**
D. Software D.2 Software Engineering D.2.4 Software/Program Verification Subjects: Assertion checkers	A Step-Wise Refinement Approach for Enhancing e-Voting Acceptance	The successful transformation of e-Government from a nice idea into a successful reality had been hindered by a variety of factors ranging from bureaucratic and legislative inertia to the inability of countries to achieve a sufficient IT penetration in their societies. Nowadays, the fall in IT prices, the development of innovative IT solutions and the rise in IT literacy in a number of countries has, at least tackled the latter issue. However, people still are not as enthusiastic, as it was envisaged by technocrats and politicians, in using IT solutions to pass from e-Government to e-Governance, a notable example of which is e-Voting. In this paper we argue that efforts to introduce complex e-Government and e-Participation applications should be gradual and develop solutions hand-in-hand with in-field trials that increase (also gradually) in complexity and people inclusiveness, so as to handle the various forms of social inertia successfully. We present our experience in the e-Voting domain and suggest that a similar approach in eVoting (and other demanding e-Government/e-Participation applications) could fare better to success than introducing to people a system that suddenly appears and claims to be the "perfect", all-in-one, solution.	Christos Manolopoulos, Dimitris Sofotassios, Polyxeni Nakou, Yannis Stamatiou, Anastasia Panagiotaki, Paul Spirakis

Table C-1 (continued)

Predefined Topic By ACM	**Title**	**Abstract**	**Authors**
H. Information Systems H.2 Database Management H.2.8 Database applications Subjects: Spatial databases and GIS	An Enhanced Indoor Pedestrian Model Supporting Spatial DBMSs	Two-dimensional geographic information systems (GISs) are mature technology and applications such as car navigation systems are commonplace. As indoor positioning techniques are developing, indoor 3D models are attracting increasing attention. However, modeling and implementing indoor 3D models applicable to real-time, client-server environments such as 2D GIS is a challenge and no working applications have yet been reported. As part of a multi-stage project that aims to build 3D indoor applications running in real-time, we are currently developing a fire evacuation system. Although not definitely required at this stage, we used a spatial DBMS as the input data instead of CAD files; the process of building floor plans and stairs is shown here. In developing the simulation model, we improved the existing 'floor field' model such that it can accommodate the visibility factor. While the previous floor field model does not capture the visibility effect, we revised the algorithm so it can give different walking speeds to pedestrians based on the level of visibility to the exits from where the pedestrians are located. We show the process of building the proposed 3D model and test the simulation system using a campus building.	Suyeong Kwak, Hyunwoo Nam, Chulmin Jun

Table C-1 (continued)

Predefined Topic By ACM	**Title**	**Abstract**	**Authors**
H. Information Systems H.2 Database Management H.2.8 Database applications Subjects: Spatial databases and GIS	Valid Scope Computation for Location-Dependent Spatial Query in Mobile Broadcast Environments	Wireless data broadcast is an efficient and scalable means to provide information access for a large population of clients in mobile environments. With Location-Based Services (LBSs) deployed upon a broadcast channel, mobile clients can collect data from the channel to answer their location-dependent spatial queries (LDSQs). Since the results of LDSQs would become invalid when mobile client moves to new locations, the knowledge of valid scopes for LDSQ results is necessary to assist clients to determine if their previous LDSQ results can be reused after they moved. This effectively improves query response time and client energy consumption. In this paper, we devise efficient algorithms to determine valid scopes for various LDSQs including range, window and nearest neighbor queries along with LDSQ processing over a broadcast channel. We conduct an extensive set of experiments to evaluate the performance of our proposed algorithms. While the proposed valid scope algorithm incurs only little extra processing overhead, unnecessary LDSQ reevaluation is significantly eliminated, thus providing faster query response and saving client energy.	Ken C. K. Lee, Josh Schiffman, Baihua Zheng, Wang-Chien Lee

Table C-1 (continued)

Predefined Topic By ACM	Title	Abstract	Authors
H. Information Systems H.2 Database Management H.2.8 Database applications Subjects: Spatial databases and GIS	Time Geography Inverted: Recognizing Intentions in Space and Time	Mobile intention recognition is the problem of inferring a mobile user's intentions from her behavior in geographic space. Such behavior is constrained in space and time. Current approaches, however, have difficulties to handle temporal constraints. We therefore propose using the framework of time geography to formalize and visualize both spatial and temporal constraints for the mobile intention recognition problem. A new rule language is introduced which allows for modeling intentions with spatial and temporal constraints. A location-based game application demonstrates that interpreting a user's spatio-temporal behavior sequence in terms of intentions reduces ambiguity compared to mobile intention recognition without temporal constraints.	Peter Kiefer, Martin Raubal, Christoph Schlieder
D. Software D.2 Software Engineering D.2.4 Software/Program Verification Subjects: Assertion checkers	Reasoning about Comprehensions with First-Order SMT Solvers	This paper presents a technique for translating common comprehension expressions (sum, count, product, min, and max) into verification conditions that can be tackled by two off-the-shelf first-order SMT solvers. Since a first-order SMT solver does not directly support the bound variables that occur in comprehension expressions, the challenge is to provide a sound axiomatisation that is strong enough to prove interesting programs and, furthermore, that can be used automatically by the SMT solver. The technique has been implemented in the Spec# program verifier. The paper also reports on the experience of using Spec# to verify several challenging programming examples drawn from a textbook by Dijkstra and Feijen.	K. Rustan M. Leino, Rosemary Monahan

Table C-1 (continued)

Predefined Topic By ACM	**Title**	**Abstract**	**Authors**
I. Computing Methodologies I.2 Artificial Intelligence I.2.7 Natural Language Processing Subjects: Machine translation	Effective Arabic-English Cross-Language Information Retrieval via Machine-Readable Dictionaries and Machine Translation	In Cross-Language Information Retrieval (CLIR), queries in one language retrieve relevant documents in other languages. Machine-Readable Dictionary (MRD) and Machine Translation (MT) are important resources for query translation in CLIR. We investigate MT and MRD to Arabic-English CLIR. The translation ambiguity associated with these resources is the key problem. We present three methods of query translation using a bilingual dictionary for Arabic-English CLIR. First, we present the Every-Match (EM) method. This method yields ambiguous translations since many extraneous terms are added to the original query. To disambiguate the query translation, we present the First-Match (FM) method that considers the first match in the dictionary as the candidate term. Finally, we present the Two-Phase (TP) method. We show that good retrieval effectiveness can be achieved without complex resources using the Two-Phase method for Arabic-English CLIR. We also empirically evaluate the effectiveness of the MT-based method using short, medium, and long queries from TREC. The effects of the query length on the quality of the MT-based CLIR are investigated.	Mohammed Aljlayl, Ophir Frieder

Table C-1 (continued)

Predefined Topic By ACM	**Title**	**Abstract**	**Authors**
I. Computing Methodologies I.2 Artificial Intelligence I.2.7 Natural Language Processing Subjects: Machine translation	Discriminative Phrase-Based Models for Arabic Machine Translation	A design for an Arabic-to-English translation system is presented. The core of the system implements a standard phrase-based statistical machine translation architecture, but it is extended by incorporating a local discriminative phrase selection model to address the semantic ambiguity of Arabic. Local classifiers are trained using linguistic information and context to translate a phrase, and this significantly increases the accuracy in phrase selection with respect to the most frequent translation traditionally considered. These classifiers are integrated into the translation system so that the global task gets benefits from the discriminative learning. As a result, we obtain significant improvements in the full translation task at the lexical, syntactic, and semantic levels as measured by an heterogeneous set of automatic evaluation metrics.	Cristina Espana-Bonet, Jesus Gimenez, Lluis Marquez
I. Computing Methodologies I.2 Artificial Intelligence I.2.7 Natural Language Processing Subjects: Machine translation	Statistical Machine Translation	Statistical machine translation (SMT) treats the translation of natural language as a machine learning problem. By examining many samples of human-produced translation, SMT algorithms automatically learn how to translate. SMT has made tremendous strides in less than two decades, and new ideas are constantly introduced. This survey presents a tutorial overview of the state of the art. We describe the context of the current research and then move to a formal problem description and an overview of the main subproblems: translation modeling, parameter estimation, and decoding. Along the way, we present a taxonomy of some different approaches within these areas. We conclude with an overview of evaluation and a discussion of future directions.	Adam Lopez

Table C-1 (continued)

Predefined Topic By ACM	Title	Abstract	Authors
C. Computer Systems Organization C.2 Computer-Communication Networks C.2.2 Network Protocols Subjects: Protocol architecture (OSI model)	Frequency-Aware Rate Adaptation and MAC Protocols	There has been burgeoning interest in wireless technologies that can use wider frequency spectrum. Technology advances, such as 802.11n and ultra-wideband (UWB), are pushing toward wider frequency bands. The analog-to-digital TV transition has made 100- 250 MHz of digital whitespace bandwidth available for unlicensed access. Also, recent work on WiFi networks has advocated discarding the notion of channelization and allowing all nodes to access the wide 802.11 spectrum in order to improve load balancing. This shift towards wider bands presents an opportunity to exploit frequency diversity. Specifically, frequencies that are far from each other in the spectrum have significantly different SNRs, and good frequencies differ across sender-receiver pairs. This paper presents FARA, a combined frequency-aware rate adaptation and MAC protocol. FARA makes three departures from conventional wireless network design: First, it presents a scheme to robustly compute per-frequency SNRs using normal data transmissions. Second, instead of using one bit rate per link, it enables a sender to adapt the bitrate independently across frequencies based on these per-frequency SNRs. Third, in contrast to traditional frequency-oblivious MAC protocols, it introduces a MAC protocol that allocates to a sender-receiver pair the frequencies that work best for that pair. We have implemented FARA in FPGA on a wideband 802.11-compatible radio platform. Our experiments reveal that FARA provides a 3.1× throughput improvement in comparison to frequency-oblivious systems that occupy the same spectrum.	Hariharan Rahul, Farinaz Edalat, Dina Katabi, Charles Sodini

Table C-1 (continued)

Predefined Topic By ACM	Title	Abstract	Authors
C. Computer Systems Organization C.2 Computer-Communication Networks C.2.2 Network Protocols Subjects: Protocol architecture (OSI model)	Self-Certified Sybil-Free Pseudonyms	Accurate and trusted identifiers are a centerpiece for any security architecture. Protecting against Sybil attacks in a privacy-friendly manner is a non-trivial problem in wireless infrastructureless networks, such as mobile ad hoc networks. In this paper, we introduce self-certified Sybil-free pseudonyms as a means to provide privacy-friendly Sybil-freeness without requiring continuous online availability of a trusted third party. These pseudonyms are self-certified and computed by the users themselves from their cryptographic longterm identities. Contrary to identity certificates, we preserve location privacy and improve protection against some notorious attacks on anonymous communication systems.	Leonardo A. Martucci, Markulf Kohlweiss, Christer Andersson, Andriy Panchenko
	Optimized Ant Based Routing Protocol for MANET	The basic routing problem in MANET deals with methods to transport a packet across a network from source node to destination node. In this paper, we introduce a new ant based routing protocol to optimize the route discovery and maximize the efficiency of routing in terms of packet delivery ratio (PDR) using the blocking expanding ring search (Blocking-ERS), third party route reply, local route repair and n-hop local ring techniques. These techniques control the overhead and minimize the end-to-end delay with improvement of PDR. The Optimized-Ant routing protocol is based on ad hoc on-demand distance vector (AODV) and inspired by the ant-colony optimization (ACO) used to solve complex optimization problems and utilizes a collection of mobile agents as "ants" to perform optimal routing activities. Exhaustive simulations are carried out and it is observed that, Optimized-Ant performs better than AODV.	Ashima Rout, Srinivas Sethi, Debajyoti Mishra

Table C-1 (continued)

Predefined Topic By ACM	**Title**	**Abstract**	**Authors**
K. Computing Milieux K.3 Computers And Education K.3.1 Computer Uses in Education Subjects: Distance learning	Group Formation in Computer-Supported Collaborative Learning	Group formation in CSCL environments is either done manually with little support from the system, or the system needs an elaborated model of the learning domain in order to select potential peer learners and to form learning groups in a pedagogically sound way. Our research objectives include the integration of collaborative learning into the learning environment so that knowledge about the collaboration context can be used to support collaboration, including group formation without the need for a detailed model of the learning domain. In this paper we describe how so-called Intended Points of Cooperation (IPoCs) can be integrated into a (web-based) course. The course author defines at which points in the course a collaborative activity should occur and specifies the cooperative activity, i.e., type and size of the learning group, the collaboration type, and additional material for each activity. We explain how the system can utilize the knowledge about the collaboration context in order to form appropriate learning groups. Finally, we illustrate our approach with examples from the project "L³: Lifelong learning as a utility", a German federally funded project which serves as a use case.	Martin Wessner, Hans-Rüdiger Pfister

Table C-1 (continued)

Predefined Topic By ACM	Title	Abstract	Authors
K. Computing Milieux K.3 Computers And Education K.3.1 Computer Uses in Education Subjects: Distance learning	Experiences Teaching Operating Systems Using Virtual Platforms and Linux	Operating system courses teach students much more when they provide hands-on kernel-level project experience with a real operating system. However, enabling a large class of students to do kernel development can be difficult. To address this problem, we created a virtual kernel development environment in which operating systems can be developed, debugged, and rebooted in a shared computer facility without affecting other users. Using virtual machines and remote display technology, our virtual kernel development laboratory enables even distance learning students at remote locations to participate in kernel development projects with on-campus students. We have successfully deployed and used our virtual kernel development environment together with the open-source Linux kernel to provide kernel-level project experiences for over nine hundred students in the introductory operating system course at Columbia University.	Jason Nieh, Chris Vaill
D. Software D.2 Software Engineering D.2.4 Software/Program Verification Subjects: Assertion checkers	Safe Composition of Product Lines	Programs of a software product line can be synthesized by composing modules that implement features. Besides high-level domain constraints that govern the compatibility of features, there are also low-level implementation constraints: a feature module can reference elements that are defined in other feature modules. Safe composition is the guarantee that all programs in a product line are type safe: i.e., absent of references to undefined elements (such as classes, methods, and variables). We show how safe composition properties can be verified for AHEAD product lines using feature models and SAT solvers.	Sahil Thaker, Don Batory, David Kitchin, William Cook

Table C-1 (continued)

Predefined Topic By ACM	Title	Abstract	Authors
K. Computing Milieux K.3 Computers And Education K.3.1 Computer Uses in Education Subjects: Distance learning	Reflecting on Online Learning Designs Using Observed Behavior	Educators, as designers of resources, experiences, and environments for learning, make judgments and assumptions about learners and how design choices will affect them. While some uncertainties can be resolved through the design process, others must be addressed experientially, through action (implementation or enactment) punctuated by reflection. Online learning designs, since they are often motivated by broad, asynchronous accessibility, offer both unique challenges and opportunities for design reflection. The challenges tend to concern greater diversity among larger learner populations, and therefore a need to account for greater potential variance in learner experiences. The opportunities arise from the nature of the medium, where use can be passively observed through interactions between learners and the learning environment. In this paper, we address the use of observed behavior as a lens for design reflection on a large corpus of online learning resources focusing on cybersecurity for adult learners.	Larry Howard, Julie Johnson, Carin Neitzel

APPENDIX D – The Topic Sets Generated for the POC: Matching Experts and Domains with Unclear Topics

Table D-1: Generated topics without using any stemming algorithm

TOPIC SET	TOPIC ID	WORDS OF TOPICS
1	T000	arabic,based,clir,machine,present,translation,using,query,method,first
1	T001	applications,kernel,development,indoor,model,students,virtual,systems,system,project
1	T002	behavior,constraints,location,time,temporal,spatial,problem,new,mobile,environments
1	T003	ant,fara,presents,protocol,frequency,frequencies,routing,sybil,wireless,spectrum
1	T004	design,domain,group,learners,order,smt,safe,product,paper,learning
2	T000	based,observed,paper,problem,using,resources,phrase,online,design,behavior
2	T001	development,students,systems,virtual,using,technology,system,operating,kernel,level
2	T002	arabic,english,clir,language,method,queries,translation,query,present,machine
2	T003	applications,developing,indoor,models,visibility,voting,type,model,domain,building
2	T004	ant,route,techniques,sybil,self,routing,pseudonyms,certified,privacy,optimized
2	T005	composition,constraints,experience,government,product,safe,solvers,programs,people,feature
2	T006	collaboration,environments,group,learning,system,project,learners,experiences,environment,course
2	T007	fara,wireless,wider,spectrum,sender,rate,frequency,mac,frequencies,protocol
2	T008	broadcast,client,location,new,temporal,valid,time,spatial,mobile,ldsq
2	T009	first,paper,smt,support,spec,solver,presents,overview,made,order
3	T000	broadcast,ldsqs,proposed,valid,process,ldsq,client,clients,knowledge,environments

Table D-1 (continued)

TOPIC SET	TOPIC ID	WORDS OF TOPICS
3	T001	behavior,challenges,design,local,online,reflection,observed,learners,learner,designs
3	T002	based,experiences,level,problem,using,wireless,systems,paper,models,location
3	T003	applications,building,developing,shown,stage,visibility,simulation,model,indoor,field
3	T004	clir,language,match,query,retrieval,translation,resources,queries,mt,method
3	T005	fara,frequencies,frequency,sender,spectrum,wider,snrs,rate,protocols,mac
3	T006	challenge,development,experience,students,tackled,virtual,successfully,real,operating,kernel
3	T007	attacks,certified,friendly,pseudonyms,sybil,third,self,privacy,party,networks
3	T008	activity,collaboration,context,group,project,system,learning,environment,domain,course
3	T009	arabic,discriminative,english,present,statistical,translation,semantic,phrase,overview,machine
3	T010	behavior,constraints,intentions,space,temporal,time,spatial,problem,new,mobile
3	T011	composition,comprehension,feature,programs,solver,solvers,safe,product,modules,lines
3	T012	ant,aodv,manet,protocol,routing,techniques,route,pdr,optimized,node
3	T013	approach,countries,government,people,successful,voting,solutions,participation,inertia,hand
3	T014	access,across,first,presents,spec,support,smt,paper,order,made
4	T000	algorithms,data,location,new,significantly,time,spatial,proposed,mobile,client
4	T001	classifiers,evaluation,machine,semantic,task,translation,translate,statistical,phrase,first
4	T002	comprehension,solver,voting,using,tackled,support,solvers,paper,domain,experience
4	T003	behavior,design,challenges,designs,learner,observed,reflection,online,learners,diversity
4	T004	ambiguity,based,context,local,using,problem,present,language,complex,arabic

Table D-1 (continued)

TOPIC SET	TOPIC ID	WORDS OF TOPICS
4	T005	clir,retrieval,resources,query,queries,mt,method,dictionary,match,english
4	T006	fara,frequencies,frequency,protocols,sender,spectrum,wider,snrs,rate,mac
4	T007	countries,government,inertia,solutions,people,participation,literacy,hand,factors,enthusiastic
4	T008	automatically,implemented,made,first,order,presents,spec,wide,smt,overview
4	T009	computer,discriminative,form,project,remote,technology,system,real,model,distance
4	T010	broadcast,ldsq,valid,results,response,processing,ldsqs,dependent,channel,clients
4	T011	ant,node,aodv,optimized,pdr,routing,techniques,source,route,party
4	T012	attacks,free,friendly,pseudonyms,trusted,sybil,self,privacy,freeness,certified
4	T013	applications,visibility,stage,simulation,process,models,indoor,building,field,developing
4	T014	development,display,introductory,level,operating,successfully,virtual,students,linux,kernel
4	T015	approaches,intention,introduced,user,temporal,space,recognition,intentions,constraints,behavior
4	T016	address,environment,experiences,course,formation,knowledge,learning,need,large,including
4	T017	access,across,manet,paper,systems,wireless,third,protocol,networks,experiments
4	T018	composition,lines,variables,safe,programs,product,modules,level,elements,feature
4	T019	activity,collaborative,collaboration,elaborated,group,order,type,sound,groups,examples
5	T000	aware,notion,pair,protocols,rate,presents,oblivious,frequency,fara,frequencies
5	T001	behavior,resources,reflection,online,greater,learner,learners,observed,medium,designs
5	T002	course,group,need,potential,using,remote,paper,learning,experiences,environment
5	T003	challenge,process,unique,visibility,simulation,pedestrians,developing,floor,indoor,increasing

Table D-1 (continued)

TOPIC SET	TOPIC ID	WORDS OF TOPICS
5	T004	clir,effectiveness,first,method,readable,translation,retrieval,mt,match,dictionary
5	T005	ad,hoc,network,protocol,third,node,introduce,end,ant,aodv
5	T006	broadcast,results,valid,response,dependent,determine,knowledge,ldsqs,deployed,clients
5	T007	attacks,certified,freeness,privacy,self,trusted,sybil,pseudonyms,friendly,free
5	T008	classifiers,local,selection,task,trained,semantic,phrase,levels,discriminative,frequent
5	T009	based,translate,way,statistical,finally,learning,points,research,integrated,context
5	T010	effectively,energy,incurs,moved,queries,saving,query,quality,ldsq,evaluate
5	T011	automatically,paper,smt,technique,verification,solver,presents,overview,comprehension,expressions
5	T012	based,terms,using,show,complex,methods,models,optimized,far,basic
5	T013	conditions,first,implemented,occur,sound,support,spec,order,interesting,groups
5	T014	applications,innovative,notable,societies,step,social,nice,improved,effect,field
5	T015	algorithms,problem,techniques,new,channel,location,means,mobile,language,approaches
5	T016	access,digital,mac,significantly,spectrum,wider,unlicensed,snrs,sender,experiments
5	T017	address,diversity,large,proposed,technology,locations,environments,design,client,data
5	T018	algorithm,stage,time,spatial,geographic,modeling,previous,processing,introduced,current
5	T019	blocking,carried,manet,packet,ring,routing,route,pdr,optimization,local
5	T020	composition,lines,product,safe,variables,programs,modules,line,elements,feature
5	T021	attracting,shown,system,real,campus,level,model,project,including,building
5	T022	adaptation,architecture,distance,party,systems,work,users,source,enables,bands

Table D-1 (continued)

TOPIC SET	TOPIC ID	WORDS OF TOPICS
5	T023	activity,domain,examples,objectives,type,formation,elaborated,collaborative,author,collaboration
5	T024	behavior,temporal,user,space,geography,intentions,inverted,recognition,intention,constraints
5	T025	columbia,development,kernel,operating,shared,virtual,students,platforms,linux,difficult
5	T026	across,networks,paper,wideband,wireless,receiver,overhead,made,advocated,improve
5	T027	approach,transformation,voting,tackled,factors,people,rebooted,successfully,governance,enthusiastic
5	T028	countries,evoting,government,inertia,participation,successful,solutions,literacy,hand,fare
5	T029	ambiguity,heterogeneous,machine,taxonomy,translation,present,incorporating,evaluation,arabic,english

Table D-2: Generated topics using the Porter's stemming algorithm

TOPIC SET	TOPIC ID	WORDS OF TOPICS
1	T000	base,client,constraint,introduc,locat,spatial,time,mobil,ldsq,intent
1	T001	applic,develop,approach,cours,govern,level,system,project,model,kernel
1	T002	ant,fara,improv,spectrum,rout,protocol,optim,network,mac,frequenc
1	T003	collabor,design,implement,learn,experi,environ,learner,paper,smt,order
1	T004	arab,clir,effect,first,machin,translat,queri,present,method,languag
2	T000	algorithm,ldsq,mobil,queri,time,spatial,process,locat,improv,client
2	T001	applic,govern,indoor,vote,solut,peopl,increas,floor,approach,build
2	T002	address,cours,collabor,design,learn,need,reflect,observ,learner,group

Table D-2 (continued)

TOPIC SET	TOPIC ID	WORDS OF TOPICS
2	T003	comprehens,featur,modul,product,safe,solver,program,order,line,domain
2	T004	develop,oper,virtual,system,student,project,level,environ,kernel,experi
2	T005	behavior,certifi,constraint,privaci,recognit,sybil,tempor,space,pseudonym,intent
2	T006	ant,network,node,problem,term,rout,optim,new,introduc,complex
2	T007	across,wider,spectrum,sender,protocol,present,fara,frequenc,adapt,mac
2	T008	activ,base,first,model,smt,techniqu,support,paper,implement,challeng
2	T009	arab,machin,phrase,translat,queri,present,method,languag,clir,english
3	T000	algorithm,queri,scope,valid,result,ldsq,broadcast,channel,data,client
3	T001	across,base,effici,node,third,wireless,protocol,network,introduc,improv
3	T002	ambigu,arab,base,method,queri,retriev,mt,english,effect,clir
3	T003	adapt,fara,frequenc,sender,spectrum,wider,snr,rate,pair,mac
3	T004	address,behavior,context,model,reflect,resourc,need,learner,learn,integr
3	T005	constraint,intent,locat,propos,tempor,time,spatial,problem,new,mobil
3	T006	complex,countri,govern,peopl,success,vote,solut,inertia,hand,gradual
3	T007	approach,exampl,form,smt,support,techniqu,spec,order,introduc,increas
3	T008	activ,collabor,cours,includ,type,util,knowledg,group,format,domain
3	T009	challeng,comput,design,observ,paper,process,onlin,implement,experi,environ
3	T010	develop,enabl,kernel,student,technolog,virtual,system,project,oper,level
3	T011	applic,build,field,model,stage,visibl,pedestrian,indoor,geograph,floor

Table D-2 (continued)

TOPIC SET	TOPIC ID	WORDS OF TOPICS
3	T012	ad,ant,certifi,pseudonym,self,sybil,rout,privaci,optim,local
3	T013	composit,comprehens,featur,program,solver,verifi,safe,product,modul,line
3	T014	automat,discrimin,evalu,phrase,statist,translat,present,overview,machin,first
4	T000	ambigu,discrimin,english,machin,phrase,statist,problem,overview,evalu,arab
4	T001	certifi,free,ident,protect,self,trust,sybil,pseudonym,privaci,friendli
4	T002	address,learn,select,potenti,need,natur,model,integr,context,cooper
4	T003	comprehens,implement,cours,level,program,spec,techniqu,support,solver,order
4	T004	develop,kernel,particip,real,virtual,system,student,project,oper,distanc
4	T005	challeng,technolog,process,paper,larg,experi,environ,comput,enabl,data
4	T006	clir,dictionari,effect,match,mt,resourc,retriev,queri,method,languag
4	T007	base,network,protocol,wireless,third,system,significantli,node,local,improv
4	T008	access,behavior,design,affect,divers,observ,opportun,reflect,onlin,learner
4	T009	ad,approach,demand,parti,subproblem,transform,term,qualiti,introduc,complex
4	T010	broadcast,ldsq,valid,set,scope,result,queri,effici,channel,client
4	T011	algorithm,locat,applic,mobil,problem,spatial,user,time,propos,new
4	T012	ant,block,optim,perform,techniqu,rout,ring,packet,manet,aodv
4	T013	activ,type,knowledg,includ,group,format,form,approach,domain,collabor
4	T014	adapt,fara,frequenc,pair,sender,spectrum,wider,snr,rate,mac
4	T015	allow,constraint,invert,tempor,space,recognit,receiv,intent,bitrat,behavior

Table D-2 (continued)

TOPIC SET	TOPIC ID	WORDS OF TOPICS
4	T016	build,floor,geograph,field,increas,model,stage,visibl,pedestrian,indoor
4	T017	automat,final,idea,made,semant,translat,smt,present,interest,first
4	T018	composit,modul,verifi,variabl,safe,refer,product,line,element,featur
4	T019	countri,gradual,govern,inertia,number,solut,vote,success,peopl,literaci
5	T000	campu,hand,real,successfulli,system,show,level,enabl,develop,distanc
5	T001	certifi,trust,sybil,self,friendli,ident,privaci,pseudonym,protect,free
5	T002	ambigu,evalu,machin,present,translat,smt,overview,languag,base,automat
5	T003	comprehens,spec,variabl,verifi,textbook,solver,express,order,smt,program
5	T004	across,introduc,network,parti,ratio,tradit,third,pdr,node,bit
5	T005	complex,least,synthes,unnecessari,vote,transform,rang,inabl,countri,idea
5	T006	associ,readabl,resourc,queri,cross,match,mrd,mt,dictionari,clir
5	T007	adapt,comparison,fpga,mac,pair,uwb,scheme,opportun,frequenc,enabl
5	T008	ad,english,local,phase,retriev,method,investig,effect,arab,complex
5	T009	demand,techniqu,util,tackl,exampl,paper,research,sound,form,domain
5	T010	awar,band,fara,protocol,snr,wider,spectrum,sender,oblivi,digit
5	T011	envisag,literaci,peopl,success,trial,solut,particip,inertia,govern,gradual
5	T012	broadcast,spatial,time,propos,geograph,mobil,overhead,previou,locat,environ
5	T013	creat,kernel,machin,project,student,virtual,teach,remot,oper,linux
5	T014	algorithm,move,paper,wireless,work,process,natur,implement,channel,data

Table D-2 (continued)

TOPIC SET	TOPIC ID	WORDS OF TOPICS
5	T015	activ,manet,point,includ,cooper,explain,format,group,cours,collabor
5	T016	ant,aodv,end,packet,protocol,rout,ring,perform,optim,block
5	T017	applic,field,increas,shown,visibl,model,floor,factor,attract,enhanc
5	T018	base,system,technolog,sourc,comput,set,significantli,simul,improv,collect
5	T019	context,defin,knowledg,project,semant,type,support,select,model,integr
5	T020	composit,line,product,safe,softwar,refer,modul,implement,element,featur
5	T021	behavior,space,tempor,sequenc,constraint,intent,invert,recognit,current,constrain
5	T022	client,depend,energi,queri,result,valid,scope,respons,ldsq,effici
5	T023	approach,introduc,problem,term,user,secur,new,hoc,commun,handl
5	T024	centerpiec,rate,receiv,present,document,final,first,interest,enthusiast,condit
5	T025	build,client,gi,multi,report,stair,stage,pedestrian,indoor,dbm
5	T026	access,deploi,larg,refin,tend,normal,experi,debug,affect,cours
5	T027	address,reflect,resourc,potenti,enact,interact,learn,need,environ,author
5	T028	classifi,discrimin,made,phrase,standard,task,statist,reveal,occupi,frequent
5	T029	behavior,greater,observ,opportun,popul,onlin,learner,divers,challeng,design

Table E-1: Results of the baseline method without using a stemming algorithm

Topic Set	Data Fusion Technique	Weight of the Domain Profile	Precision	Recall	P@1 / R@1	P@5	R@5	P@10	R@10	P@15	R@15
1	COMBMNZ	0.5	0	0	0	0	0	0	0	0	0
1	COMBMNZ	1	0	0	0	0	0	0	0	0	0
1	EXPCOMBMNZ	0.5	0	0	0	0	0	0	0	0	0
1	EXPCOMBMNZ	1	0	0	0	0	0	0	0	0	0
1	RR	0.5	0	0	0	0	0	0	0	0	0
1	RR	1	0	0	0	0	0	0	0	0	0
2	COMBMNZ	0.5	0.1429	0.5714	0	0.1143	0.5714	0.0571	0.5714	0.0381	0.5714
2	COMBMNZ	1	0.1429	0.5714	0.1429	0.1143	0.5714	0.0571	0.5714	0.0381	0.5714
2	EXPCOMBMNZ	0.5	0.1429	0.5714	0.1429	0.1143	0.5714	0.0571	0.5714	0.0381	0.5714
2	EXPCOMBMNZ	1	0.1429	0.5714	0.1429	0.1143	0.5714	0.0571	0.5714	0.0381	0.5714
2	RR	0.5	0.1429	0.5714	0.1429	0.1143	0.5714	0.0571	0.5714	0.0381	0.5714
2	RR	1	0.1429	0.5714	0.1429	0.1143	0.5714	0.0571	0.5714	0.0381	0.5714
3	COMBMNZ	0.5	0.2	0.8	0.2	0.16	0.8	0.08	0.8	0.0533	0.8
3	COMBMNZ	1	0.2	0.8	0.2	0.16	0.8	0.08	0.8	0.0533	0.8
3	EXPCOMBMNZ	0.5	0.2	0.8	0.2	0.16	0.8	0.08	0.8	0.0533	0.8
3	EXPCOMBMNZ	1	0.2	0.8	0.2	0.16	0.8	0.08	0.8	0.0533	0.8
3	RR	0.5	0.2	0.8	0	0.16	0.8	0.08	0.8	0.0533	0.8
3	RR	1	0.2	0.8	0.2	0.16	0.8	0.08	0.8	0.0533	0.8
4	COMBMNZ	0.5	0.0952	0.5714	0.1429	0.1143	0.5714	0.0571	0.5714	0.0381	0.5714
4	COMBMNZ	1	0.0952	0.5714	0.1429	0.0857	0.4286	0.0571	0.5714	0.0381	0.5714
4	EXPCOMBMNZ	0.5	0.0952	0.5714	0	0.0857	0.4286	0.0571	0.5714	0.0381	0.5714
4	EXPCOMBMNZ	1	0.0952	0.5714	0.1429	0.0857	0.4286	0.0571	0.5714	0.0381	0.5714

Table E-1 (continued)

Topic Set	Data Fusion Technique	Weight of the Domain Profile	Precision	Recall	P@1 / R@1	P@5	R@5	P@10	R@10	P@15	R@15
4	RR	0.5	0.0952	0.5714	0	0.1143	0.5714	0.0571	0.5714	0.0381	0.5714
4	RR	1	0.0952	0.5714	0.1429	0.0857	0.4286	0.0571	0.5714	0.0381	0.5714
5	COMBMNZ	0.5	0.1333	0.6667	0.3333	0.1333	0.6667	0.0667	0.6667	0.0444	0.6667
5	COMBMNZ	1	0.1333	0.6667	0.1667	0.1333	0.6667	0.0667	0.6667	0.0444	0.6667
5	EXPCOMBMNZ	0.5	0.1333	0.6667	0	0.1333	0.6667	0.0667	0.6667	0.0444	0.6667
5	EXPCOMBMNZ	1	0.1333	0.6667	0.1667	0.1333	0.6667	0.0667	0.6667	0.0444	0.6667
5	RR	0.5	0.1333	0.6667	0.1667	0.1333	0.6667	0.0667	0.6667	0.0444	0.6667
5	RR	1	0.1333	0.6667	0.1667	0.1333	0.6667	0.0667	0.6667	0.0444	0.6667

Table E-2: Results of the baseline method using the Porter's stemming algorithm

Topic Set	Data Fusion Technique	Weight of the Domain Profile	Precision	Recall	P@1 / R@1	P@5	R@5	P@10	R@10	P@15	R@15
1	COMBMNZ	0.5	0.25	0.5	0	0.1	0.5	0.05	0.5	0.0333	0.5
1	COMBMNZ	1	0.25	0.5	0	0.1	0.5	0.05	0.5	0.0333	0.5
1	EXPCOMBMNZ	0.5	0.25	0.5	0.5	0.1	0.5	0.05	0.5	0.0333	0.5
1	EXPCOMBMNZ	1	0.25	0.5	0	0.1	0.5	0.05	0.5	0.0333	0.5
1	RR	0.5	0.25	0.5	0.5	0.1	0.5	0.05	0.5	0.0333	0.5
1	RR	1	0.25	0.5	0	0.1	0.5	0.05	0.5	0.0333	0.5
2	COMBMNZ	0.5	0.1111	0.4444	0	0.0889	0.4444	0.0444	0.4444	0.0296	0.4444
2	COMBMNZ	1	0.1111	0.4444	0.111111	0.0889	0.4444	0.0444	0.4444	0.0296	0.4444
2	EXPCOMBMNZ	0.5	0.1111	0.4444	0	0.0889	0.4444	0.0444	0.4444	0.0296	0.4444
2	EXPCOMBMNZ	1	0.1111	0.4444	0.1111	0.0889	0.4444	0.0444	0.4444	0.0296	0.4444
2	RR	0.5	0.1111	0.4444	0.1111	0.0889	0.4444	0.0444	0.4444	0.0296	0.4444

Table E-2 (continued)

Topic Set	Data Fusion Technique	Weight of the Domain Profile	Precision	Recall	P@1 / R@1	P@5	R@5	P@10	R@10	P@15	R@15
2	RR	1	0.1111	0.4444	0.1111	0.0889	0.4444	0.0444	0.4444	0.0296	0.4444
3	COMBMNZ	0.5	0.0714	0.5	0.0833	0.0833	0.4167	0.05	0.5	0.0333	0.5
3	COMBMNZ	1	0.0714	0.5	0.0833	0.0667	0.3333	0.05	0.5	0.0333	0.5
3	EXPCOMBMNZ	0.5	0.0714	0.5	0.0833	0.05	0.25	0.05	0.5	0.0333	0.5
3	EXPCOMBMNZ	1	0.0714	0.5	0.0833	0.0667	0.3333	0.05	0.5	0.0333	0.5
3	RR	0.5	0.0714	0.5	0.0833	0.0833	0.4167	0.05	0.5	0.0333	0.5
3	RR	1	0.0714	0.5	0.0833	0.0667	0.3333	0.05	0.5	0.0333	0.5
4	COMBMNZ	0.5	0.0833	0.5	0.0833	0.0833	0.4167	0.05	0.5	0.0333	0.5
4	COMBMNZ	1	0.0833	0.5	0.0833	0.0833	0.4167	0.05	0.5	0.0333	0.5
4	EXPCOMBMNZ	0.5	0.0833	0.5	0.0833	0.0667	0.3333	0.05	0.5	0.0333	0.5
4	EXPCOMBMNZ	1	0.0833	0.5	0.0833	0.0833	0.4167	0.05	0.5	0.0333	0.5
4	RR	0.5	0.0833	0.5	0	0.0667	0.3333	0.05	0.5	0.0333	0.5
4	RR	1	0.0833	0.5	0.0833	0.0833	0.4167	0.05	0.5	0.0333	0.5
5	COMBMNZ	0.5	0.0667	0.3333	0.2	0.0667	0.3333	0.0333	0.3333	0.0222	0.3333
5	COMBMNZ	1	0.0667	0.3333	0.0667	0.0667	0.3333	0.0333	0.3333	0.0222	0.3333
5	EXPCOMBMNZ	0.5	0.0667	0.3333	0.0667	0.0667	0.3333	0.0333	0.3333	0.0222	0.3333
5	EXPCOMBMNZ	1	0.0667	0.3333	0.0667	0.0667	0.3333	0.0333	0.3333	0.0222	0.3333
5	RR	0.5	0.0667	0.3333	0.1333	0.0667	0.3333	0.0333	0.3333	0.0222	0.3333
5	RR	1	0.0667	0.3333	0.0667	0.0667	0.3333	0.0333	0.3333	0.0222	0.3333

Table E-3: Results of the proposed model without using a stemming algorithm

Topic Set	Data Fusion Technique	Weight of the Domain Profile	Precision	Recall	P@1 / R@1	P@5	R@5	P@10	R@10	P@15	R@15
1	COMBMNZ	0.5	0.0907	0.9333	0.2	0.1733	0.8667	0.0933	0.9333	0.0622	0.9333
1	COMBMNZ	1	0.0907	0.9333	0.1333	0.1333	0.6667	0.0933	0.9333	0.0622	0.9333
1	EXPCOMBMNZ	0.5	0.0907	0.9333	0.2	0.1867	0.9333	0.0933	0.9333	0.0622	0.9333
1	EXPCOMBMNZ	1	0.0907	0.9333	0.0667	0.16	0.8	0.0933	0.9333	0.0622	0.9333
1	RR	0.5	0.0907	0.9333	0.2	0.1867	0.9333	0.0933	0.9333	0.0622	0.9333
1	RR	1	0.0907	0.9333	0.2	0.1467	0.7333	0.0933	0.9333	0.0622	0.9333
2	COMBMNZ	0.5	0.1211	1	0.4	0.1867	0.9333	0.1	1	0.0667	1
2	COMBMNZ	1	0.1211	1	0.3333	0.1867	0.9333	0.1	1	0.0667	1
2	EXPCOMBMNZ	0.5	0.1211	1	0.4	0.1867	0.9333	0.1	1	0.0667	1
2	EXPCOMBMNZ	1	0.1211	1	0.4667	0.2	1	0.1	1	0.0667	1
2	RR	0.5	0.1211	1	0.4	0.2	1	0.1	1	0.0667	1
2	RR	1	0.1211	1	0.4	0.1867	0.9333	0.1	1	0.0667	1
3	COMBMNZ	0.5	0.1078	1	0.6667	0.2	1	0.1	1	0.0667	1
3	COMBMNZ	1	0.1078	1	0.4	0.2	1	0.1	1	0.0667	1
3	EXPCOMBMNZ	0.5	0.1078	1	0.7333	0.2	1	0.1	1	0.0667	1
3	EXPCOMBMNZ	1	0.1078	1	0.6667	0.2	1	0.1	1	0.0667	1
3	RR	0.5	0.1078	1	0.6667	0.2	1	0.1	1	0.0667	1
3	RR	1	0.1078	1	0.6	0.2	1	0.1	1	0.0667	1
4	COMBMNZ	0.5	0.0826	1	0.6	0.2	1	0.1	1	0.0667	1
4	COMBMNZ	1	0.0826	1	0.6	0.2	1	0.1	1	0.0667	1
4	EXPCOMBMNZ	0.5	0.0826	1	0.6667	0.2	1	0.1	1	0.0667	1
4	EXPCOMBMNZ	1	0.0826	1	0.6667	0.2	1	0.1	1	0.0667	1
4	RR	0.5	0.0826	1	0.8	0.2	1	0.1	1	0.0667	1
4	RR	1	0.0826	1	0.8	0.2	1	0.1	1	0.0667	1
5	COMBMNZ	0.5	0.0986	1	0.7333	0.2	1	0.1	1	0.0667	1

Table E-3 (continued)

Topic Set	Data Fusion Technique	Weight of the Domain Profile	Precision	Recall	P@1 / R@1	P@5	R@5	P@10	R@10	P@15	R@15
5	COMBMNZ	1	0.0986	1	0.6667	0.2	1	0.1	1	0.0667	1
5	EXPCOMBMNZ	0.5	0.0986	1	0.8	0.2	1	0.1	1	0.0667	1
5	EXPCOMBMNZ	1	0.0986	1	0.8	0.2	1	0.1	1	0.0667	1
5	RR	0.5	0.0986	1	0.8667	0.2	1	0.1	1	0.0667	1
5	RR	1	0.0986	1	0.8	0.2	1	0.1	1	0.0667	1

Table E-4: Results of the proposed model using the Porter's stemming algorithm

Topic Set	Data Fusion Technique	Weight of the Domain Profile	Precision	Recall	P@1 / R@1	P@5	R@5	P@10	R@10	P@15	R@15
1	COMBMNZ	0.5	0.0886	0.8	0.1333	0.1333	0.6667	0.08	0.8	0.0533	0.8
1	COMBMNZ	1	0.0886	0.8	0.1333	0.1333	0.6667	0.08	0.8	0.0533	0.8
1	EXPCOMBMNZ	0.5	0.0886	0.8	0.1333	0.1467	0.7333	0.08	0.8	0.0533	0.8
1	EXPCOMBMNZ	1	0.0886	0.8	0.2	0.16	0.8	0.08	0.8	0.0533	0.8
1	RR	0.5	0.0886	0.8	0.2667	0.16	0.8	0.08	0.8	0.0533	0.8
1	RR	1	0.0886	0.8	0.1333	0.12	0.6	0.08	0.8	0.0533	0.8
2	COMBMNZ	0.5	0.0879	1	0.4667	0.2	1	0.1	1	0.0667	1
2	COMBMNZ	1	0.0879	1	0.2667	0.1867	0.9333	0.1	1	0.0667	1
2	EXPCOMBMNZ	0.5	0.0879	1	0.4	0.2	1	0.1	1	0.0667	1
2	EXPCOMBMNZ	1	0.0879	1	0.2667	0.2	1	0.1	1	0.0667	1
2	RR	0.5	0.0879	1	0.4	0.2	1	0.1	1	0.0667	1
2	RR	1	0.0879	1	0.3333	0.1867	0.9333	0.1	1	0.0667	1
3	COMBMNZ	0.5	0.0901	1	0.6	0.1867	0.9333	0.1	1	0.0667	1

Table E-4 (continued)

Topic Set	Data Fusion Technique	Weight of the Domain Profile	Precision	Recall	P@1 / R@1	P@5	R@5	P@10	R@10	P@15	R@15
3	COMBMNZ	1	0.0901	1	0.4667	0.1733	0.8667	0.1	1	0.0667	1
3	EXPCOMBMNZ	0.5	0.0901	1	0.6667	0.1867	0.9333	0.1	1	0.0667	1
3	EXPCOMBMNZ	1	0.0901	1	0.6	0.2	1	0.1	1	0.0667	1
3	RR	0.5	0.0901	1	0.6	0.1867	0.9333	0.1	1	0.0667	1
3	RR	1	0.0901	1	0.5333	0.1733	0.8667	0.1	1	0.0667	1
4	COMBMNZ	0.5	0.0806	1	0.6667	0.2	1	0.1	1	0.0667	1
4	COMBMNZ	1	0.0806	1	0.6667	0.2	1	0.1	1	0.0667	1
4	EXPCOMBMNZ	0.5	0.0806	1	0.8	0.2	1	0.1	1	0.0667	1
4	EXPCOMBMNZ	1	0.0806	1	0.9333	0.2	1	0.1	1	0.0667	1
4	RR	0.5	0.0806	1	0.6667	0.2	1	0.1	1	0.0667	1
4	RR	1	0.0806	1	0.6667	0.2	1	0.1	1	0.0667	1
5	COMBMNZ	0.5	0.0801	1	0.8667	0.2	1	0.1	1	0.0667	1
5	COMBMNZ	1	0.0801	1	0.8667	0.2	1	0.1	1	0.0667	1
5	EXPCOMBMNZ	0.5	0.0801	1	0.8	0.2	1	0.1	1	0.0667	1
5	EXPCOMBMNZ	1	0.0801	1	0.7333	0.2	1	0.1	1	0.0667	1
5	RR	0.5	0.0801	1	0.8	0.2	1	0.1	1	0.0667	1
5	RR	1	0.0801	1	0.8667	0.2	1	0.1	1	0.0667	1

APPENDIX F – Turkish Stopword List

Table F-1: Turkish Stopword List

a	çok	i	onlara	tamam
acaba	çünkü	için	onlardan	tüm
altı	d	içinde	onların	tümü
ama	da	iki	onların	u
ancak	daha	ile	onu	ü
artık	de	ise	onun	üç
asla	değil	işte	orada	üzere
aslında	demek	j	oysa	v
az	diğer	k	oysaki	var
b	diğeri	kaç	ö	ve
bana	diğerleri	kadar	öbürü	veya
bazen	diye	kendi	ön	veyahut
bazı	dokuz	kendine	önce	y
bazıları	dolayı	kendini	ötürü	ya
bazısı	dört	ki	öyle	ya da
belki	e	kim	p	yani
ben	elbette	kime	r	yedi
beni	en	kimi	rağmen	yerine
benim	f	kimin	s	yine
beş	fakat	kimisi	sana	yoksa
bile	falan	l	sekiz	z
bir	felan	m	sen	zaten
birçoğu	filan	madem	senden	zira
birçok	g	mı	seni	q
birçokları	gene	mı	senin	w
biri	gibi	mi	siz	x
birisi	ğ	mu	sizden	olarak
birkaç	h	mu	size	kullanılacaktır
birkaçı	hâlâ	mü	sizi	konularında

Table F-1 (continued)

birşey	hangi	mü	sizin	olan
birşeyi	hangisi	n	son	yapılacaktır
biz	hani	nasıl	sonra	ayrıca
bize	hatta	ne	ş	yapılması
bizi	hem	ne kadar	şayet	kapsamında
bizim	henüz	ne zaman	şey	çalışmaktadır
böyle	hep	neden	şeyden	almıştır
böylece	hepsi	nedir	şeye	sahiptir
bu	hepsine	nerde	şeyi	üzerinde
buna	hepsini	nerede	şeyler	sayesinde
bunda	her	nereden	şimdi	farklı
bundan	her biri	nereye	şöyle	oluşan
bunu	herkes	nesi	şu	oluşacaktır
bunun	herkese	neyse	şuna	sağlayacaktır
burada	herkesi	niçin	şunda	tarafından
bütün	hiç	niye	şundan	aşağıda
c	hiç kimse	o	şunlar	edilebilir
ç	hiçbiri	on	şunu	olduğu
çoğu	hiçbirine	ona	şunun	geliştirilecektir
çoğuna	hiçbirini	ondan	t	vb
çoğunu	ı	onlar	tabi	

APPENDIX G – The Results of Case Study: The Proposed Model

Table G-1: Results of the proposed model in training set

Number of Topics	Stemming is Used	Data Fusion Technique	Project-Scientist Similarity Weight	Weight of the Domain Profile	P@5	R@5	P@10	R@10	P@15	R@15
30	No	COMBMNZ	TOPIC03_TECH07	0.5	0.0081	0.0220	0.0053	0.0277	0.0051	0.0397
30	No	COMBMNZ	TOPIC03_TECH07	1	0.0073	0.0179	0.0055	0.0292	0.0053	0.0428
30	No	COMBMNZ	TOPIC05_TECH05	0.5	0.0077	0.0213	0.0053	0.0277	0.0051	0.0397
30	No	COMBMNZ	TOPIC05_TECH05	1	0.0073	0.0179	0.0055	0.0292	0.0053	0.0428
30	No	COMBMNZ	TOPIC07_TECH03	0.5	0.0061	0.0166	0.0055	0.0287	0.0049	0.0382
30	No	COMBMNZ	TOPIC07_TECH03	1	0.0073	0.0179	0.0055	0.0292	0.0053	0.0428
30	No	EXPCOMBMNZ	TOPIC03_TECH07	0.5	0.0069	0.0196	0.0047	0.0257	0.0051	0.0436
30	No	EXPCOMBMNZ	TOPIC03_TECH07	1	0.0065	0.0156	0.0053	0.0308	0.0051	0.0436
30	No	EXPCOMBMNZ	TOPIC05_TECH05	0.5	0.0069	0.0196	0.0049	0.0267	0.0051	0.0436
30	No	EXPCOMBMNZ	TOPIC05_TECH05	1	0.0065	0.0156	0.0053	0.0308	0.0051	0.0436
30	No	EXPCOMBMNZ	TOPIC07_TECH03	0.5	0.0069	0.0196	0.0047	0.0257	0.0051	0.0436
30	No	EXPCOMBMNZ	TOPIC07_TECH03	1	0.0065	0.0156	0.0053	0.0308	0.0051	0.0436
30	No	RR	TOPIC03_TECH07	0.5	0.0061	0.0152	0.0045	0.0240	0.0047	0.0402
30	No	RR	TOPIC03_TECH07	1	0.0065	0.0167	0.0045	0.0235	0.0045	0.0367
30	No	RR	TOPIC05_TECH05	0.5	0.0069	0.0172	0.0053	0.0281	0.0050	0.0423
30	No	RR	TOPIC05_TECH05	1	0.0037	0.0098	0.0043	0.0242	0.0037	0.0313
30	No	RR	TOPIC07_TECH03	0.5	0.0053	0.0132	0.0049	0.0294	0.0046	0.0396
30	No	RR	TOPIC07_TECH03	1	0.0008	0.0020	0.0022	0.0122	0.0034	0.0314
50	No	COMBMNZ	TOPIC03_TECH07	0.5	0.0061	0.0135	0.0061	0.0314	0.0058	0.0458
50	No	COMBMNZ	TOPIC03_TECH07	1	0.0065	0.0156	0.0065	0.0345	0.0061	0.0497
50	No	COMBMNZ	TOPIC05_TECH05	0.5	0.0065	0.0142	0.0059	0.0321	0.0061	0.0480

Table G-1 (continued)

Number of Topics	Stemming is Used	Data Fusion Technique	Project-Scientist Similarity Weight	Weight of the Domain Profile	P@5	R@5	P@10	R@10	P@15	R@15
50	No	COMBMNZ	TOPIC05_TECH05	1	0.0065	0.0156	0.0065	0.0345	0.0061	0.0497
50	No	COMBMNZ	TOPIC07_TECH03	0.5	0.0057	0.0128	0.0063	0.0348	0.0058	0.0475
50	No	COMBMNZ	TOPIC07_TECH03	1	0.0065	0.0156	0.0065	0.0345	0.0061	0.0497
50	No	EXPCOMBMNZ	TOPIC03_TECH07	0.5	0.0081	0.0193	0.0067	0.0358	0.0059	0.0470
50	No	EXPCOMBMNZ	TOPIC03_TECH07	1	0.0093	0.0240	0.0075	0.0409	0.0066	0.0548
50	No	EXPCOMBMNZ	TOPIC05_TECH05	0.5	0.0081	0.0193	0.0067	0.0358	0.0059	0.0470
50	No	EXPCOMBMNZ	TOPIC05_TECH05	1	0.0093	0.0240	0.0075	0.0409	0.0066	0.0548
50	No	EXPCOMBMNZ	TOPIC07_TECH03	0.5	0.0081	0.0193	0.0067	0.0358	0.0059	0.0470
50	No	EXPCOMBMNZ	TOPIC07_TECH03	1	0.0093	0.0240	0.0075	0.0409	0.0066	0.0548
50	No	RR	TOPIC03_TECH07	0.5	0.0077	0.0223	0.0055	0.0313	0.0062	0.0524
50	No	RR	TOPIC03_TECH07	1	0.0061	0.0169	0.0053	0.0321	0.0054	0.0483
50	No	RR	TOPIC05_TECH05	0.5	0.0073	0.0216	0.0059	0.0318	0.0055	0.0463
50	No	RR	TOPIC05_TECH05	1	0.0012	0.0041	0.0034	0.0220	0.0041	0.0365
50	No	RR	TOPIC07_TECH03	0.5	0.0069	0.0206	0.0053	0.0308	0.0055	0.0468
50	No	RR	TOPIC07_TECH03	1	0.0008	0.0020	0.0016	0.0098	0.0024	0.0216
100	No	COMBMNZ	TOPIC03_TECH07	0.5	0.0065	0.0179	0.0051	0.0274	0.0051	0.0426
100	No	COMBMNZ	TOPIC03_TECH07	1	0.0069	0.0189	0.0053	0.0294	0.0053	0.0436
100	No	COMBMNZ	TOPIC05_TECH05	0.5	0.0069	0.0199	0.0055	0.0304	0.0050	0.0402
100	No	COMBMNZ	TOPIC05_TECH05	1	0.0069	0.0189	0.0053	0.0294	0.0053	0.0436
100	No	COMBMNZ	TOPIC07_TECH03	0.5	0.0061	0.0169	0.0055	0.0318	0.0051	0.0423
100	No	COMBMNZ	TOPIC07_TECH03	1	0.0069	0.0189	0.0053	0.0294	0.0053	0.0436
100	No	EXPCOMBMNZ	TOPIC03_TECH07	0.5	0.0069	0.0196	0.0061	0.0352	0.0055	0.0480
100	No	EXPCOMBMNZ	TOPIC03_TECH07	1	0.0065	0.0176	0.0053	0.0297	0.0053	0.0456
100	No	EXPCOMBMNZ	TOPIC05_TECH05	0.5	0.0069	0.0196	0.0061	0.0352	0.0055	0.0480

Table G-1 (continued)

Number of Topics	Stemming is Used	Data Fusion Technique	Project-Scientist Similarity Weight	Weight of the Domain Profile	P@5	R@5	P@10	R@10	P@15	R@15
100	No	EXPCOMBMNZ	TOPIC05_TECH05	1	0.0065	0.0176	0.0053	0.0297	0.0053	0.0456
100	No	EXPCOMBMNZ	TOPIC07_TECH03	0.5	0.0069	0.0196	0.0061	0.0352	0.0055	0.0480
100	No	EXPCOMBMNZ	TOPIC07_TECH03	1	0.0065	0.0176	0.0053	0.0297	0.0053	0.0456
100	No	RR	TOPIC03_TECH07	0.5	0.0065	0.0189	0.0065	0.0396	0.0061	0.0532
100	No	RR	TOPIC03_TECH07	1	0.0045	0.0142	0.0041	0.0237	0.0039	0.0355
100	No	RR	TOPIC05_TECH05	0.5	0.0069	0.0203	0.0055	0.0338	0.0049	0.0439
100	No	RR	TOPIC05_TECH05	1	0.0028	0.0112	0.0028	0.0199	0.0032	0.0294
100	No	RR	TOPIC07_TECH03	0.5	0.0045	0.0139	0.0049	0.0301	0.0039	0.0368
100	No	RR	TOPIC07_TECH03	1	0.0016	0.0051	0.0026	0.0189	0.0024	0.0247
200	No	COMBMNZ	TOPIC03_TECH07	0.5	0.0057	0.0159	0.0039	0.0213	0.0042	0.0338
200	No	COMBMNZ	TOPIC03_TECH07	1	0.0053	0.0139	0.0051	0.0257	0.0053	0.0423
200	No	COMBMNZ	TOPIC05_TECH05	0.5	0.0041	0.0128	0.0030	0.0172	0.0028	0.0227
200	No	COMBMNZ	TOPIC05_TECH05	1	0.0053	0.0139	0.0051	0.0257	0.0053	0.0423
200	No	COMBMNZ	TOPIC07_TECH03	0.5	0.0028	0.0088	0.0026	0.0152	0.0023	0.0189
200	No	COMBMNZ	TOPIC07_TECH03	1	0.0053	0.0139	0.0047	0.0240	0.0049	0.0385
200	No	EXPCOMBMNZ	TOPIC03_TECH07	0.5	0.0089	0.0240	0.0053	0.0287	0.0054	0.0467
200	No	EXPCOMBMNZ	TOPIC03_TECH07	1	0.0085	0.0233	0.0055	0.0297	0.0061	0.0548
200	No	EXPCOMBMNZ	TOPIC05_TECH05	0.5	0.0089	0.0240	0.0053	0.0287	0.0054	0.0467
200	No	EXPCOMBMNZ	TOPIC05_TECH05	1	0.0085	0.0233	0.0053	0.0287	0.0061	0.0548
200	No	EXPCOMBMNZ	TOPIC07_TECH03	0.5	0.0089	0.0240	0.0053	0.0287	0.0054	0.0467
200	No	EXPCOMBMNZ	TOPIC07_TECH03	1	0.0085	0.0233	0.0053	0.0287	0.0059	0.0538
200	No	RR	TOPIC03_TECH07	0.5	0.0069	0.0206	0.0045	0.0254	0.0042	0.0365
200	No	RR	TOPIC03_TECH07	1	0.0012	0.0051	0.0028	0.0213	0.0031	0.0308
200	No	RR	TOPIC05_TECH05	0.5	0.0061	0.0176	0.0049	0.0277	0.0039	0.0348

Table G-1 (continued)

Number of Topics	Stemming is Used	Data Fusion Technique	Project-Scientist Similarity Weight	Weight of the Domain Profile	P@5	R@5	P@10	R@10	P@15	R@15
200	No	RR	TOPIC05_TECH05	1	0.0008	0.0030	0.0012	0.0101	0.0016	0.0189
200	No	RR	TOPIC07_TECH03	0.5	0.0053	0.0152	0.0047	0.0274	0.0038	0.0335
200	No	RR	TOPIC07_TECH03	1	0.0008	0.0030	0.0012	0.0091	0.0011	0.0118
219	No	COMBMNZ	TOPIC03_TECH07	0.5	0.0045	0.0118	0.0043	0.0223	0.0041	0.0348
219	No	COMBMNZ	TOPIC03_TECH07	1	0.0065	0.0162	0.0053	0.0277	0.0061	0.0510
219	No	COMBMNZ	TOPIC05_TECH05	0.5	0.0032	0.0108	0.0034	0.0203	0.0038	0.0321
219	No	COMBMNZ	TOPIC05_TECH05	1	0.0069	0.0169	0.0053	0.0277	0.0061	0.0510
219	No	COMBMNZ	TOPIC07_TECH03	0.5	0.0028	0.0098	0.0043	0.0237	0.0035	0.0291
219	No	COMBMNZ	TOPIC07_TECH03	1	0.0069	0.0169	0.0051	0.0270	0.0055	0.0477
219	No	EXPCOMBMNZ	TOPIC03_TECH07	0.5	0.0069	0.0189	0.0053	0.0270	0.0050	0.0412
219	No	EXPCOMBMNZ	TOPIC03_TECH07	1	0.0069	0.0179	0.0051	0.0267	0.0049	0.0399
219	No	EXPCOMBMNZ	TOPIC05_TECH05	0.5	0.0069	0.0189	0.0053	0.0270	0.0050	0.0412
219	No	EXPCOMBMNZ	TOPIC05_TECH05	1	0.0069	0.0179	0.0051	0.0267	0.0046	0.0382
219	No	EXPCOMBMNZ	TOPIC07_TECH03	0.5	0.0069	0.0189	0.0053	0.0270	0.0049	0.0402
219	No	EXPCOMBMNZ	TOPIC07_TECH03	1	0.0069	0.0179	0.0051	0.0267	0.0046	0.0382
219	No	RR	TOPIC03_TECH07	0.5	0.0057	0.0172	0.0049	0.0287	0.0042	0.0399
219	No	RR	TOPIC03_TECH07	1	0.0012	0.0041	0.0026	0.0176	0.0030	0.0270
219	No	RR	TOPIC05_TECH05	0.5	0.0049	0.0145	0.0043	0.0270	0.0043	0.0392
219	No	RR	TOPIC05_TECH05	1	0.0008	0.0020	0.0018	0.0115	0.0020	0.0203
219	No	RR	TOPIC07_TECH03	0.5	0.0049	0.0145	0.0041	0.0254	0.0038	0.0321
219	No	RR	TOPIC07_TECH03	1	0	0	0.0020	0.0135	0.0018	0.0162
30	Yes	COMBMNZ	TOPIC03_TECH07	0.5	0.0045	0.0105	0.0051	0.0270	0.0043	0.0348
30	Yes	COMBMNZ	TOPIC05_TECH05	0.5	0.0045	0.0105	0.0053	0.0291	0.0045	0.0368
30	Yes	COMBMNZ	TOPIC07_TECH03	0.5	0.0037	0.0088	0.0041	0.0237	0.0038	0.0311

Table G-1 (continued)

Number of Topics	Stemming is Used	Data Fusion Technique	Project-Scientist Similarity Weight	Weight of the Domain Profile	P@5	R@5	P@10	R@10	P@15	R@15
30	Yes	COMBMNZ	TOPIC03_TECH07	1	0.0041	0.0098	0.0053	0.0277	0.0042	0.0338
30	Yes	COMBMNZ	TOPIC05_TECH05	1	0.0041	0.0098	0.0053	0.0277	0.0042	0.0338
30	Yes	COMBMNZ	TOPIC07_TECH03	1	0.0041	0.0098	0.0053	0.0277	0.0042	0.0338
30	Yes	EXPCOMBMNZ	TOPIC03_TECH07	0.5	0.0053	0.0149	0.0047	0.0254	0.0050	0.0412
30	Yes	EXPCOMBMNZ	TOPIC05_TECH05	0.5	0.0053	0.0149	0.0047	0.0254	0.0050	0.0412
30	Yes	EXPCOMBMNZ	TOPIC07_TECH03	0.5	0.0053	0.0149	0.0047	0.0254	0.0050	0.0412
30	Yes	EXPCOMBMNZ	TOPIC03_TECH07	1	0.0053	0.0139	0.0057	0.0335	0.0053	0.0463
30	Yes	EXPCOMBMNZ	TOPIC05_TECH05	1	0.0053	0.0139	0.0057	0.0335	0.0053	0.0463
30	Yes	EXPCOMBMNZ	TOPIC07_TECH03	1	0.0053	0.0139	0.0057	0.0335	0.0053	0.0463
30	Yes	RR	TOPIC03_TECH07	0.5	0.0061	0.0189	0.0047	0.0274	0.0047	0.0416
30	Yes	RR	TOPIC05_TECH05	0.5	0.0065	0.0199	0.0049	0.0284	0.0050	0.0439
30	Yes	RR	TOPIC07_TECH03	0.5	0.0053	0.0176	0.0053	0.0325	0.0049	0.0439
30	Yes	RR	TOPIC03_TECH07	1	0.0065	0.0189	0.0045	0.0270	0.0042	0.0368
30	Yes	RR	TOPIC05_TECH05	1	0.0041	0.0149	0.0034	0.0220	0.0032	0.0308
30	Yes	RR	TOPIC07_TECH03	1	0.0028	0.0112	0.0030	0.0210	0.0038	0.0392
50	Yes	COMBMNZ	TOPIC03_TECH07	0.5	0.0069	0.0172	0.0053	0.0270	0.0046	0.0382
50	Yes	COMBMNZ	TOPIC05_TECH05	0.5	0.0053	0.0139	0.0047	0.0243	0.0043	0.0345
50	Yes	COMBMNZ	TOPIC07_TECH03	0.5	0.0028	0.0078	0.0028	0.0176	0.0039	0.0348
50	Yes	COMBMNZ	TOPIC03_TECH07	1	0.0065	0.0162	0.0057	0.0301	0.0053	0.0436
50	Yes	COMBMNZ	TOPIC05_TECH05	1	0.0065	0.0162	0.0057	0.0301	0.0053	0.0436
50	Yes	COMBMNZ	TOPIC07_TECH03	1	0.0065	0.0162	0.0057	0.0301	0.0053	0.0436
50	Yes	EXPCOMBMNZ	TOPIC03_TECH07	0.5	0.0069	0.0169	0.0057	0.0301	0.0057	0.0477
50	Yes	EXPCOMBMNZ	TOPIC05_TECH05	0.5	0.0069	0.0169	0.0057	0.0301	0.0057	0.0477
50	Yes	EXPCOMBMNZ	TOPIC07_TECH03	0.5	0.0069	0.0169	0.0057	0.0301	0.0057	0.0477

Table G-1 (continued)

Number of Topics	Stemming is Used	Data Fusion Technique	Project-Scientist Similarity Weight	Weight of the Domain Profile	P@5	R@5	P@10	R@10	P@15	R@15
50	Yes	EXPCOMBMNZ	TOPIC03_TECH07	1	0.0077	0.0203	0.0069	0.0402	0.0062	0.0548
50	Yes	EXPCOMBMNZ	TOPIC05_TECH05	1	0.0077	0.0203	0.0069	0.0402	0.0062	0.0548
50	Yes	EXPCOMBMNZ	TOPIC07_TECH03	1	0.0077	0.0203	0.0069	0.0402	0.0062	0.0548
50	Yes	RR	TOPIC03_TECH07	0.5	0.0053	0.0125	0.0053	0.0267	0.0054	0.0473
50	Yes	RR	TOPIC05_TECH05	0.5	0.0045	0.0108	0.0051	0.0267	0.0049	0.0423
50	Yes	RR	TOPIC07_TECH03	0.5	0.0037	0.0118	0.0039	0.0210	0.0035	0.0308
50	Yes	RR	TOPIC03_TECH07	1	0.0077	0.0221	0.0057	0.0330	0.0049	0.0441
50	Yes	RR	TOPIC05_TECH05	1	0.0028	0.0096	0.0030	0.0215	0.0041	0.0384
50	Yes	RR	TOPIC07_TECH03	1	0.0012	0.0061	0.0034	0.0225	0.0028	0.0272
100	Yes	COMBMNZ	TOPIC03_TECH07	0.5	0.0065	0.0172	0.0057	0.0291	0.0053	0.0450
100	Yes	COMBMNZ	TOPIC05_TECH05	0.5	0.0057	0.0152	0.0057	0.0321	0.0054	0.0477
100	Yes	COMBMNZ	TOPIC07_TECH03	0.5	0.0045	0.0139	0.0049	0.0311	0.0049	0.0439
100	Yes	COMBMNZ	TOPIC03_TECH07	1	0.0081	0.0189	0.0055	0.0281	0.0049	0.0409
100	Yes	COMBMNZ	TOPIC05_TECH05	1	0.0081	0.0189	0.0055	0.0281	0.0049	0.0409
100	Yes	COMBMNZ	TOPIC07_TECH03	1	0.0081	0.0189	0.0057	0.0301	0.0049	0.0409
100	Yes	EXPCOMBMNZ	TOPIC03_TECH07	0.5	0.0077	0.0193	0.0059	0.0311	0.0049	0.0389
100	Yes	EXPCOMBMNZ	TOPIC05_TECH05	0.5	0.0081	0.0203	0.0059	0.0311	0.0049	0.0389
100	Yes	EXPCOMBMNZ	TOPIC07_TECH03	0.5	0.0081	0.0203	0.0059	0.0311	0.0049	0.0389
100	Yes	EXPCOMBMNZ	TOPIC03_TECH07	1	0.0077	0.0203	0.0049	0.0250	0.0050	0.0419
100	Yes	EXPCOMBMNZ	TOPIC05_TECH05	1	0.0077	0.0203	0.0049	0.0250	0.0050	0.0419
100	Yes	EXPCOMBMNZ	TOPIC07_TECH03	1	0.0077	0.0203	0.0049	0.0250	0.0050	0.0419
100	Yes	RR	TOPIC03_TECH07	0.5	0.0057	0.0149	0.0043	0.0243	0.0047	0.0382
100	Yes	RR	TOPIC05_TECH05	0.5	0.0053	0.0139	0.0039	0.0216	0.0045	0.0352
100	Yes	RR	TOPIC07_TECH03	0.5	0.0049	0.0152	0.0041	0.0250	0.0045	0.0406

Table G-1 (continued)

Number of Topics	Stemming is Used	Data Fusion Technique	Project-Scientist Similarity Weight	Weight of the Domain Profile	P@5	R@5	P@10	R@10	P@15	R@15
100	Yes	RR	TOPIC03_TECH07	1	0.0024	0.0078	0.0024	0.0128	0.0028	0.0247
100	Yes	RR	TOPIC05_TECH05	1	0	0	0.0014	0.0098	0.0020	0.0196
100	Yes	RR	TOPIC07_TECH03	1	0	0	0.0010	0.0068	0.0018	0.0179
200	Yes	COMBMNZ	TOPIC03_TECH07	0.5	0.0024	0.0068	0.0037	0.0206	0.0038	0.0348
200	Yes	COMBMNZ	TOPIC05_TECH05	0.5	0.0016	0.0047	0.0022	0.0132	0.0034	0.0325
200	Yes	COMBMNZ	TOPIC07_TECH03	0.5	0	0	0.0022	0.0142	0.0034	0.0325
200	Yes	COMBMNZ	TOPIC03_TECH07	1	0.0073	0.0196	0.0053	0.0291	0.0054	0.0456
200	Yes	COMBMNZ	TOPIC05_TECH05	1	0.0073	0.0196	0.0051	0.0284	0.0050	0.0429
200	Yes	COMBMNZ	TOPIC07_TECH03	1	0.0057	0.0159	0.0039	0.0206	0.0042	0.0365
200	Yes	EXPCOMBMNZ	TOPIC03_TECH07	0.5	0.0085	0.0267	0.0063	0.0353	0.0062	0.0519
200	Yes	EXPCOMBMNZ	TOPIC05_TECH05	0.5	0.0085	0.0267	0.0063	0.0353	0.0062	0.0519
200	Yes	EXPCOMBMNZ	TOPIC07_TECH03	0.5	0.0085	0.0267	0.0063	0.0353	0.0062	0.0519
200	Yes	EXPCOMBMNZ	TOPIC03_TECH07	1	0.0073	0.0216	0.0057	0.0313	0.0059	0.0505
200	Yes	EXPCOMBMNZ	TOPIC05_TECH05	1	0.0077	0.0237	0.0059	0.0333	0.0058	0.0495
200	Yes	EXPCOMBMNZ	TOPIC07_TECH03	1	0.0077	0.0237	0.0059	0.0333	0.0058	0.0495
200	Yes	RR	TOPIC03_TECH07	0.5	0.0081	0.0213	0.0067	0.0402	0.0058	0.0500
200	Yes	RR	TOPIC05_TECH05	0.5	0.0053	0.0169	0.0051	0.0308	0.0046	0.0423
200	Yes	RR	TOPIC07_TECH03	0.5	0.0032	0.0088	0.0045	0.0270	0.0041	0.0365
200	Yes	RR	TOPIC03_TECH07	1	0.0016	0.0071	0.0020	0.0132	0.0030	0.0291
200	Yes	RR	TOPIC05_TECH05	1	0.0004	0.0020	0.0010	0.0071	0.0015	0.0149
200	Yes	RR	TOPIC07_TECH03	1	0	0	0.0002	0.0020	0.0018	0.0156
219	Yes	COMBMNZ	TOPIC03_TECH07	0.5	0.0041	0.0125	0.0041	0.0270	0.0034	0.0311
219	Yes	COMBMNZ	TOPIC05_TECH05	0.5	0.0024	0.0088	0.0037	0.0250	0.0031	0.0291
219	Yes	COMBMNZ	TOPIC07_TECH03	0.5	0.0020	0.0078	0.0034	0.0243	0.0032	0.0311

Table G-1 (continued)

Number of Topics	Stemming is Used	Data Fusion Technique	Project-Scientist Similarity Weight	Weight of the Domain Profile	P@5	R@5	P@10	R@10	P@15	R@15
219	Yes	COMBMNZ	TOPIC03_TECH07	1	0.0065	0.0159	0.0047	0.0250	0.0051	0.0446
219	Yes	COMBMNZ	TOPIC05_TECH05	1	0.0065	0.0159	0.0045	0.0230	0.0047	0.0409
219	Yes	COMBMNZ	TOPIC07_TECH03	1	0.0061	0.0152	0.0037	0.0189	0.0041	0.0348
219	Yes	EXPCOMBMNZ	TOPIC03_TECH07	0.5	0.0073	0.0179	0.0053	0.0291	0.0053	0.0487
219	Yes	EXPCOMBMNZ	TOPIC05_TECH05	0.5	0.0081	0.0220	0.0053	0.0291	0.0053	0.0487
219	Yes	EXPCOMBMNZ	TOPIC07_TECH03	0.5	0.0081	0.0220	0.0053	0.0291	0.0053	0.0487
219	Yes	EXPCOMBMNZ	TOPIC03_TECH07	1	0.0073	0.0199	0.0055	0.0311	0.0055	0.0527
219	Yes	EXPCOMBMNZ	TOPIC05_TECH05	1	0.0073	0.0199	0.0055	0.0311	0.0055	0.0527
219	Yes	EXPCOMBMNZ	TOPIC07_TECH03	1	0.0073	0.0199	0.0055	0.0311	0.0055	0.0527
219	Yes	RR	TOPIC03_TECH07	0.5	0.0065	0.0203	0.0055	0.0358	0.0054	0.0510
219	Yes	RR	TOPIC05_TECH05	0.5	0.0061	0.0193	0.0049	0.0301	0.0049	0.0456
219	Yes	RR	TOPIC07_TECH03	0.5	0.0057	0.0172	0.0049	0.0311	0.0043	0.0399
219	Yes	RR	TOPIC03_TECH07	1	0.0004	0.0020	0.0016	0.0105	0.0018	0.0166
219	Yes	RR	TOPIC05_TECH05	1	0	0	0.0004	0.0017	0.0011	0.0118
219	Yes	RR	TOPIC07_TECH03	1	0	0	0.0004	0.0017	0.0005	0.0047

Table G-2: Results of the proposed model in training set

Number of Topics	Stemming is Used	Data Fusion Technique	Project-Scientist Similarity Weight	Weight of the Domain Profile	P@5	R@5	P@10	R@10	P@15	R@15
30	No	COMBMNZ	TOPIC03_TECH07	0.5	0.0016	0.0027	0.0024	0.0108	0.0049	0.0515
30	No	COMBMNZ	TOPIC03_TECH07	1	0.0016	0.0027	0.0033	0.0149	0.0033	0.0271

Table G-2 (continued)

Number of Topics	Stemming is Used	Data Fusion Technique	Project-Scientist Similarity Weight	Weight of the Domain Profile	P@5	R@5	P@10	R@10	P@15	R@15
30	No	COMBMNZ	TOPIC05_TECH05	0.5	0.0016	0.0027	0.0024	0.0108	0.0049	0.0515
30	No	COMBMNZ	TOPIC05_TECH05	1	0.0016	0.0027	0.0033	0.0149	0.0033	0.0271
30	No	COMBMNZ	TOPIC07_TECH03	0.5	0.0016	0.0027	0.0024	0.0108	0.0049	0.0515
30	No	COMBMNZ	TOPIC07_TECH03	1	0.0016	0.0027	0.0033	0.0149	0.0033	0.0271
30	No	EXPCOMBMNZ	TOPIC03_TECH07	0.5	0	0	0.0016	0.0068	0.0022	0.0136
30	No	EXPCOMBMNZ	TOPIC03_TECH07	1	0.0049	0.0108	0.0049	0.0271	0.0043	0.0352
30	No	EXPCOMBMNZ	TOPIC05_TECH05	0.5	0	0	0.0016	0.0068	0.0022	0.0136
30	No	EXPCOMBMNZ	TOPIC05_TECH05	1	0.0049	0.0108	0.0049	0.0271	0.0043	0.0352
30	No	EXPCOMBMNZ	TOPIC07_TECH03	0.5	0	0	0.0016	0.0068	0.0022	0.0136
30	No	EXPCOMBMNZ	TOPIC07_TECH03	1	0.0049	0.0108	0.0049	0.0271	0.0043	0.0352
30	No	RR	TOPIC03_TECH07	0.5	0.0016	0.0027	0.0008	0.0027	0.0027	0.0271
30	No	RR	TOPIC03_TECH07	1	0.0049	0.0122	0.0049	0.0230	0.0049	0.0379
30	No	RR	TOPIC05_TECH05	0.5	0.0016	0.0027	0.0008	0.0027	0.0027	0.0271
30	No	RR	TOPIC05_TECH05	1	0.0049	0.0122	0.0049	0.0230	0.0049	0.0379
30	No	RR	TOPIC07_TECH03	0.5	0.0016	0.0027	0.0008	0.0027	0.0027	0.0271
30	No	RR	TOPIC07_TECH03	1	0.0049	0.0122	0.0049	0.0230	0.0049	0.0379
50	No	COMBMNZ	TOPIC03_TECH07	0.5	0.0081	0.0325	0.0049	0.0407	0.0043	0.0474
50	No	COMBMNZ	TOPIC03_TECH07	1	0.0049	0.0203	0.0049	0.0366	0.0054	0.0610
50	No	COMBMNZ	TOPIC05_TECH05	0.5	0.0081	0.0325	0.0049	0.0407	0.0043	0.0474
50	No	COMBMNZ	TOPIC05_TECH05	1	0.0049	0.0203	0.0049	0.0366	0.0054	0.0610
50	No	COMBMNZ	TOPIC07_TECH03	0.5	0.0081	0.0325	0.0049	0.0407	0.0043	0.0474
50	No	COMBMNZ	TOPIC07_TECH03	1	0.0049	0.0203	0.0049	0.0366	0.0054	0.0610
50	No	EXPCOMBMNZ	TOPIC03_TECH07	0.5	0.0016	0.0041	0.0008	0.0041	0.0022	0.0136
50	No	EXPCOMBMNZ	TOPIC03_TECH07	1	0.0033	0.0122	0.0024	0.0149	0.0027	0.0312

Table G-2 (continued)

Number of Topics	Stemming is Used	Data Fusion Technique	Project-Scientist Similarity Weight	Weight of the Domain Profile	P@5	R@5	P@10	R@10	P@15	R@15
50	No	EXPCOMBMNZ	TOPIC05_TECH05	0.5	0.0016	0.0041	0.0008	0.0041	0.0022	0.0136
50	No	EXPCOMBMNZ	TOPIC05_TECH05	1	0.0033	0.0122	0.0024	0.0149	0.0027	0.0312
50	No	EXPCOMBMNZ	TOPIC07_TECH03	0.5	0.0016	0.0041	0.0008	0.0041	0.0022	0.0136
50	No	EXPCOMBMNZ	TOPIC07_TECH03	1	0.0033	0.0122	0.0024	0.0149	0.0027	0.0312
50	No	RR	TOPIC03_TECH07	0.5	0.0016	0.0027	0.0008	0.0027	0.0011	0.0068
50	No	RR	TOPIC03_TECH07	1	0.0016	0.0081	0.0016	0.0122	0.0022	0.0203
50	No	RR	TOPIC05_TECH05	0.5	0.0016	0.0027	0.0008	0.0027	0.0011	0.0068
50	No	RR	TOPIC05_TECH05	1	0.0016	0.0081	0.0016	0.0122	0.0022	0.0203
50	No	RR	TOPIC07_TECH03	0.5	0.0016	0.0027	0.0008	0.0027	0.0011	0.0068
50	No	RR	TOPIC07_TECH03	1	0.0016	0.0081	0.0016	0.0122	0.0022	0.0203
100	No	COMBMNZ	TOPIC03_TECH07	0.5	0.0049	0.0163	0.0057	0.0407	0.0054	0.0556
100	No	COMBMNZ	TOPIC03_TECH07	1	0.0049	0.0163	0.0057	0.0407	0.0049	0.0474
100	No	COMBMNZ	TOPIC05_TECH05	0.5	0.0049	0.0163	0.0057	0.0407	0.0054	0.0556
100	No	COMBMNZ	TOPIC05_TECH05	1	0.0049	0.0163	0.0057	0.0407	0.0049	0.0474
100	No	COMBMNZ	TOPIC07_TECH03	0.5	0.0049	0.0163	0.0057	0.0407	0.0054	0.0556
100	No	COMBMNZ	TOPIC07_TECH03	1	0.0049	0.0163	0.0057	0.0407	0.0049	0.0474
100	No	EXPCOMBMNZ	TOPIC03_TECH07	0.5	0.0049	0.0108	0.0024	0.0108	0.0022	0.0149
100	No	EXPCOMBMNZ	TOPIC03_TECH07	1	0.0065	0.0203	0.0057	0.0366	0.0049	0.0447
100	No	EXPCOMBMNZ	TOPIC05_TECH05	0.5	0.0049	0.0108	0.0024	0.0108	0.0022	0.0149
100	No	EXPCOMBMNZ	TOPIC05_TECH05	1	0.0065	0.0203	0.0057	0.0366	0.0049	0.0447
100	No	EXPCOMBMNZ	TOPIC07_TECH03	0.5	0.0049	0.0108	0.0024	0.0108	0.0022	0.0149
100	No	EXPCOMBMNZ	TOPIC07_TECH03	1	0.0065	0.0203	0.0057	0.0366	0.0049	0.0447
100	No	RR	TOPIC03_TECH07	0.5	0	0	0.0024	0.0095	0.0016	0.0095
100	No	RR	TOPIC03_TECH07	1	0.0049	0.0122	0.0057	0.0366	0.0054	0.0515

Table G-2 (continued)

Number of Topics	Stemming is Used	Data Fusion Technique	Project-Scientist Similarity Weight	Weight of the Domain Profile	P@5	R@5	P@10	R@10	P@15	R@15
100	No	RR	TOPIC05_TECH05	0.5	0	0	0.0024	0.0095	0.0016	0.0095
100	No	RR	TOPIC05_TECH05	1	0.0049	0.0122	0.0057	0.0366	0.0054	0.0515
100	No	RR	TOPIC07_TECH03	0.5	0	0	0.0024	0.0095	0.0016	0.0095
100	No	RR	TOPIC07_TECH03	1	0.0049	0.0122	0.0057	0.0366	0.0054	0.0515
200	No	COMBMNZ	TOPIC03_TECH07	0.5	0.0016	0.0041	0.0024	0.0108	0.0033	0.0257
200	No	COMBMNZ	TOPIC03_TECH07	1	0	0	0.0024	0.0122	0.0038	0.0312
200	No	COMBMNZ	TOPIC05_TECH05	0.5	0.0016	0.0041	0.0024	0.0108	0.0033	0.0257
200	No	COMBMNZ	TOPIC05_TECH05	1	0	0	0.0024	0.0122	0.0038	0.0312
200	No	COMBMNZ	TOPIC07_TECH03	0.5	0.0016	0.0041	0.0024	0.0108	0.0033	0.0257
200	No	COMBMNZ	TOPIC07_TECH03	1	0	0	0.0024	0.0122	0.0038	0.0312
200	No	EXPCOMBMNZ	TOPIC03_TECH07	0.5	0.0033	0.0122	0.0024	0.0163	0.0022	0.0203
200	No	EXPCOMBMNZ	TOPIC03_TECH07	1	0.0033	0.0081	0.0016	0.0081	0.0016	0.0108
200	No	EXPCOMBMNZ	TOPIC05_TECH05	0.5	0.0033	0.0122	0.0024	0.0163	0.0022	0.0203
200	No	EXPCOMBMNZ	TOPIC05_TECH05	1	0.0033	0.0081	0.0016	0.0081	0.0016	0.0108
200	No	EXPCOMBMNZ	TOPIC07_TECH03	0.5	0.0033	0.0122	0.0024	0.0163	0.0022	0.0203
200	No	EXPCOMBMNZ	TOPIC07_TECH03	1	0.0033	0.0081	0.0016	0.0081	0.0016	0.0108
200	No	RR	TOPIC03_TECH07	0.5	0	0	0.0016	0.0068	0.0011	0.0068
200	No	RR	TOPIC03_TECH07	1	0.0033	0.0081	0.0016	0.0081	0.0022	0.0163
200	No	RR	TOPIC05_TECH05	0.5	0	0	0.0016	0.0068	0.0011	0.0068
200	No	RR	TOPIC05_TECH05	1	0.0033	0.0081	0.0016	0.0081	0.0022	0.0163
200	No	RR	TOPIC07_TECH03	0.5	0	0	0.0016	0.0068	0.0011	0.0068
200	No	RR	TOPIC07_TECH03	1	0.0033	0.0081	0.0016	0.0081	0.0022	0.0163
219	No	COMBMNZ	TOPIC03_TECH07	0.5	0.0033	0.0122	0.0016	0.0122	0.0022	0.0190
219	No	COMBMNZ	TOPIC03_TECH07	1	0.0033	0.0122	0.0016	0.0122	0.0022	0.0203

Table G-2 (continued)

Number of Topics	Stemming is Used	Data Fusion Technique	Project-Scientist Similarity Weight	Weight of the Domain Profile	P@5	R@5	P@10	R@10	P@15	R@15
219	No	COMBMNZ	TOPIC05_TECH05	0.5	0.0033	0.0122	0.0016	0.0122	0.0022	0.0190
219	No	COMBMNZ	TOPIC05_TECH05	1	0.0033	0.0122	0.0016	0.0122	0.0022	0.0203
219	No	COMBMNZ	TOPIC07_TECH03	0.5	0.0033	0.0122	0.0016	0.0122	0.0022	0.0190
219	No	COMBMNZ	TOPIC07_TECH03	1	0.0033	0.0122	0.0016	0.0122	0.0022	0.0203
219	No	EXPCOMBMNZ	TOPIC03_TECH07	0.5	0.0033	0.0068	0.0024	0.0095	0.0016	0.0095
219	No	EXPCOMBMNZ	TOPIC03_TECH07	1	0	0	0.0008	0.0027	0.0016	0.0108
219	No	EXPCOMBMNZ	TOPIC05_TECH05	0.5	0.0033	0.0068	0.0024	0.0095	0.0016	0.0095
219	No	EXPCOMBMNZ	TOPIC05_TECH05	1	0	0	0.0008	0.0027	0.0016	0.0108
219	No	EXPCOMBMNZ	TOPIC07_TECH03	0.5	0.0033	0.0068	0.0024	0.0095	0.0016	0.0095
219	No	EXPCOMBMNZ	TOPIC07_TECH03	1	0	0	0.0008	0.0027	0.0016	0.0108
219	No	RR	TOPIC03_TECH07	0.5	0	0	0.0024	0.0108	0.0016	0.0108
219	No	RR	TOPIC03_TECH07	1	0.0033	0.0081	0.0024	0.0108	0.0016	0.0108
219	No	RR	TOPIC05_TECH05	0.5	0	0	0.0024	0.0108	0.0016	0.0108
219	No	RR	TOPIC05_TECH05	1	0.0033	0.0081	0.0024	0.0108	0.0016	0.0108
219	No	RR	TOPIC07_TECH03	0.5	0	0	0.0024	0.0108	0.0016	0.0108
219	No	RR	TOPIC07_TECH03	1	0.0033	0.0081	0.0024	0.0108	0.0016	0.0108
30	Yes	COMBMNZ	TOPIC03_TECH07	0.5	0.0033	0.0122	0.0041	0.0217	0.0033	0.0257
30	Yes	COMBMNZ	TOPIC03_TECH07	1	0.0016	0.0041	0.0041	0.0230	0.0043	0.0339
30	Yes	COMBMNZ	TOPIC05_TECH05	0.5	0.0033	0.0122	0.0041	0.0217	0.0033	0.0257
30	Yes	COMBMNZ	TOPIC05_TECH05	1	0.0016	0.0041	0.0041	0.0230	0.0043	0.0339
30	Yes	COMBMNZ	TOPIC07_TECH03	0.5	0.0033	0.0122	0.0041	0.0217	0.0033	0.0257
30	Yes	COMBMNZ	TOPIC07_TECH03	1	0.0016	0.0041	0.0041	0.0230	0.0043	0.0339
30	Yes	EXPCOMBMNZ	TOPIC03_TECH07	0.5	0.0016	0.0027	0.0024	0.0095	0.0033	0.0217
30	Yes	EXPCOMBMNZ	TOPIC03_TECH07	1	0.0033	0.0068	0.0049	0.0257	0.0043	0.0339

Table G-2 (continued)

Number of Topics	Stemming is Used	Data Fusion Technique	Project-Scientist Similarity Weight	Weight of the Domain Profile	P@5	R@5	P@10	R@10	P@15	R@15
30	Yes	EXPCOMBMNZ	TOPIC05_TECH05	0.5	0.0016	0.0027	0.0024	0.0095	0.0033	0.0217
30	Yes	EXPCOMBMNZ	TOPIC05_TECH05	1	0.0033	0.0068	0.0049	0.0257	0.0043	0.0339
30	Yes	EXPCOMBMNZ	TOPIC07_TECH03	0.5	0.0016	0.0027	0.0024	0.0095	0.0033	0.0217
30	Yes	EXPCOMBMNZ	TOPIC07_TECH03	1	0.0033	0.0068	0.0049	0.0257	0.0043	0.0339
30	Yes	RR	TOPIC03_TECH07	0.5	0.0033	0.0068	0.0057	0.0312	0.0060	0.0556
30	Yes	RR	TOPIC03_TECH07	1	0.0016	0.0041	0.0016	0.0081	0.0027	0.0190
30	Yes	RR	TOPIC05_TECH05	0.5	0.0033	0.0068	0.0057	0.0312	0.0060	0.0556
30	Yes	RR	TOPIC05_TECH05	1	0.0016	0.0041	0.0016	0.0081	0.0027	0.0190
30	Yes	RR	TOPIC07_TECH03	0.5	0.0033	0.0068	0.0057	0.0312	0.0060	0.0556
30	Yes	RR	TOPIC07_TECH03	1	0.0016	0.0041	0.0016	0.0081	0.0027	0.0190
50	Yes	COMBMNZ	TOPIC03_TECH07	0.5	0.0016	0.0027	0.0016	0.0068	0.0016	0.0108
50	Yes	COMBMNZ	TOPIC03_TECH07	1	0.0016	0.0041	0.0049	0.0230	0.0038	0.0312
50	Yes	COMBMNZ	TOPIC05_TECH05	0.5	0.0016	0.0027	0.0016	0.0068	0.0016	0.0108
50	Yes	COMBMNZ	TOPIC05_TECH05	1	0.0016	0.0041	0.0049	0.0230	0.0038	0.0312
50	Yes	COMBMNZ	TOPIC07_TECH03	0.5	0.0016	0.0027	0.0016	0.0068	0.0016	0.0108
50	Yes	COMBMNZ	TOPIC07_TECH03	1	0.0016	0.0041	0.0049	0.0230	0.0038	0.0312
50	Yes	EXPCOMBMNZ	TOPIC03_TECH07	0.5	0	0	0.0008	0.0041	0.0033	0.0271
50	Yes	EXPCOMBMNZ	TOPIC03_TECH07	1	0.0049	0.0122	0.0049	0.0271	0.0043	0.0434
50	Yes	EXPCOMBMNZ	TOPIC05_TECH05	0.5	0	0	0.0008	0.0041	0.0033	0.0271
50	Yes	EXPCOMBMNZ	TOPIC05_TECH05	1	0.0049	0.0122	0.0049	0.0271	0.0043	0.0434
50	Yes	EXPCOMBMNZ	TOPIC07_TECH03	0.5	0	0	0.0008	0.0041	0.0033	0.0271
50	Yes	EXPCOMBMNZ	TOPIC07_TECH03	1	0.0049	0.0122	0.0049	0.0271	0.0043	0.0434
50	Yes	RR	TOPIC03_TECH07	0.5	0.0016	0.0027	0.0041	0.0230	0.0038	0.0352
50	Yes	RR	TOPIC03_TECH07	1	0.0033	0.0081	0.0041	0.0190	0.0038	0.0312

Table G-2 (continued)

Number of Topics	Stemming is Used	Data Fusion Technique	Project-Scientist Similarity Weight	Weight of the Domain Profile	P@5	R@5	P@10	R@10	P@15	R@15
50	Yes	RR	TOPIC05_TECH05	0.5	0.0016	0.0027	0.0041	0.0230	0.0038	0.0352
50	Yes	RR	TOPIC05_TECH05	1	0.0033	0.0081	0.0041	0.0190	0.0038	0.0312
50	Yes	RR	TOPIC07_TECH03	0.5	0.0016	0.0027	0.0041	0.0230	0.0038	0.0352
50	Yes	RR	TOPIC07_TECH03	1	0.0033	0.0081	0.0041	0.0190	0.0038	0.0312
100	Yes	COMBMNZ	TOPIC03_TECH07	0.5	0.0016	0.0041	0.0016	0.0122	0.0016	0.0203
100	Yes	COMBMNZ	TOPIC03_TECH07	1	0.0016	0.0041	0.0016	0.0081	0.0027	0.0285
100	Yes	COMBMNZ	TOPIC05_TECH05	0.5	0.0016	0.0041	0.0016	0.0122	0.0016	0.0203
100	Yes	COMBMNZ	TOPIC05_TECH05	1	0.0016	0.0041	0.0016	0.0081	0.0027	0.0285
100	Yes	COMBMNZ	TOPIC07_TECH03	0.5	0.0016	0.0041	0.0016	0.0122	0.0016	0.0203
100	Yes	COMBMNZ	TOPIC07_TECH03	1	0.0016	0.0041	0.0016	0.0081	0.0027	0.0285
100	Yes	EXPCOMBMNZ	TOPIC03_TECH07	0.5	0.0033	0.0122	0.0033	0.0244	0.0038	0.0407
100	Yes	EXPCOMBMNZ	TOPIC03_TECH07	1	0	0	0.0008	0.0041	0.0016	0.0163
100	Yes	EXPCOMBMNZ	TOPIC05_TECH05	0.5	0.0033	0.0122	0.0033	0.0244	0.0038	0.0407
100	Yes	EXPCOMBMNZ	TOPIC05_TECH05	1	0	0	0.0008	0.0041	0.0016	0.0163
100	Yes	EXPCOMBMNZ	TOPIC07_TECH03	0.5	0.0033	0.0122	0.0033	0.0244	0.0038	0.0407
100	Yes	EXPCOMBMNZ	TOPIC07_TECH03	1	0	0	0.0008	0.0041	0.0016	0.0163
100	Yes	RR	TOPIC03_TECH07	0.5	0	0	0.0016	0.0068	0.0011	0.0068
100	Yes	RR	TOPIC03_TECH07	1	0.0016	0.0027	0.0049	0.0298	0.0038	0.0379
100	Yes	RR	TOPIC05_TECH05	0.5	0	0	0.0016	0.0068	0.0011	0.0068
100	Yes	RR	TOPIC05_TECH05	1	0.0016	0.0027	0.0049	0.0298	0.0038	0.0379
100	Yes	RR	TOPIC07_TECH03	0.5	0	0	0.0016	0.0068	0.0011	0.0068
100	Yes	RR	TOPIC07_TECH03	1	0.0016	0.0027	0.0049	0.0298	0.0038	0.0379
200	Yes	COMBMNZ	TOPIC03_TECH07	0.5	0	0	0.0008	0.0081	0.0016	0.0203
200	Yes	COMBMNZ	TOPIC03_TECH07	1	0	0	0.0008	0.0081	0.0022	0.0244

Table G-2 (continued)

Number of Topics	Stemming is Used	Data Fusion Technique	Project-Scientist Similarity Weight	Weight of the Domain Profile	P@5	R@5	P@10	R@10	P@15	R@15
200	Yes	COMBMNZ	TOPIC05_TECH05	0.5	0	0	0.0008	0.0081	0.0016	0.0203
200	Yes	COMBMNZ	TOPIC05_TECH05	1	0	0	0.0008	0.0081	0.0022	0.0244
200	Yes	COMBMNZ	TOPIC07_TECH03	0.5	0	0	0.0008	0.0081	0.0016	0.0203
200	Yes	COMBMNZ	TOPIC07_TECH03	1	0	0	0.0008	0.0081	0.0022	0.0244
200	Yes	EXPCOMBMNZ	TOPIC03_TECH07	0.5	0	0	0.0016	0.0081	0.0016	0.0163
200	Yes	EXPCOMBMNZ	TOPIC03_TECH07	1	0.0033	0.0081	0.0033	0.0163	0.0033	0.0325
200	Yes	EXPCOMBMNZ	TOPIC05_TECH05	0.5	0	0	0.0016	0.0081	0.0016	0.0163
200	Yes	EXPCOMBMNZ	TOPIC05_TECH05	1	0.0033	0.0081	0.0033	0.0163	0.0033	0.0325
200	Yes	EXPCOMBMNZ	TOPIC07_TECH03	0.5	0	0	0.0016	0.0081	0.0016	0.0163
200	Yes	EXPCOMBMNZ	TOPIC07_TECH03	1	0.0033	0.0081	0.0033	0.0163	0.0033	0.0325
200	Yes	RR	TOPIC03_TECH07	0.5	0	0	0.0008	0.0027	0.0016	0.0108
200	Yes	RR	TOPIC03_TECH07	1	0.0016	0.0081	0.0024	0.0163	0.0027	0.0230
200	Yes	RR	TOPIC05_TECH05	0.5	0	0	0.0008	0.0027	0.0016	0.0108
200	Yes	RR	TOPIC05_TECH05	1	0.0016	0.0081	0.0024	0.0163	0.0027	0.0230
200	Yes	RR	TOPIC07_TECH03	0.5	0	0	0.0008	0.0027	0.0016	0.0108
200	Yes	RR	TOPIC07_TECH03	1	0.0016	0.0081	0.0024	0.0163	0.0027	0.0230
219	Yes	COMBMNZ	TOPIC03_TECH07	0.5	0	0	0.0016	0.0122	0.0027	0.0271
219	Yes	COMBMNZ	TOPIC03_TECH07	1	0	0	0.0024	0.0163	0.0022	0.0244
219	Yes	COMBMNZ	TOPIC05_TECH05	0.5	0	0	0.0016	0.0122	0.0027	0.0271
219	Yes	COMBMNZ	TOPIC05_TECH05	1	0	0	0.0024	0.0163	0.0022	0.0244
219	Yes	COMBMNZ	TOPIC07_TECH03	0.5	0	0	0.0016	0.0122	0.0027	0.0271
219	Yes	COMBMNZ	TOPIC07_TECH03	1	0	0	0.0024	0.0163	0.0022	0.0244
219	Yes	EXPCOMBMNZ	TOPIC03_TECH07	0.5	0	0	0.0008	0.0041	0.0005	0.0041
219	Yes	EXPCOMBMNZ	TOPIC03_TECH07	1	0.0016	0.0041	0.0024	0.0163	0.0033	0.0325

Table G-2 (continued)

Number of Topics	Stemming is Used	Data Fusion Technique	Project-Scientist Similarity Weight	Weight of the Domain Profile	P@5	R@5	P@10	R@10	P@15	R@15
219	Yes	EXPCOMBMNZ	TOPIC05_TECH05	0.5	0	0	0.0008	0.0041	0.0005	0.0041
219	Yes	EXPCOMBMNZ	TOPIC05_TECH05	1	0.0016	0.0041	0.0024	0.0163	0.0033	0.0325
219	Yes	EXPCOMBMNZ	TOPIC07_TECH03	0.5	0	0	0.0008	0.0041	0.0005	0.0041
219	Yes	EXPCOMBMNZ	TOPIC07_TECH03	1	0.0016	0.0041	0.0024	0.0163	0.0033	0.0325
219	Yes	RR	TOPIC03_TECH07	0.5	0	0	0.0016	0.0068	0.0011	0.0068
219	Yes	RR	TOPIC03_TECH07	1	0.0065	0.0149	0.0049	0.0271	0.0043	0.0352
219	Yes	RR	TOPIC05_TECH05	0.5	0	0	0.0016	0.0068	0.0011	0.0068
219	Yes	RR	TOPIC05_TECH05	1	0.0065	0.0149	0.0049	0.0271	0.0043	0.0352
219	Yes	RR	TOPIC07_TECH03	0.5	0	0	0.0016	0.0068	0.0011	0.0068
219	Yes	RR	TOPIC07_TECH03	1	0.0065	0.0149	0.0049	0.0271	0.0043	0.0352

APPENDIX H – Agreement Scores of the Kappa Statistics

Table H-1: Agreement Scores

Agreement Scores				
Variable	**Evaluator1+Evaluator2**	**Evaluator1+Evaluator3**	**Evaluator2+Evaluator3**	**Average Kappa Values**
Project1	PA=0.80 PE=0.16 K=0.76	PA=0.80 PE=0.16 K=0.76	PA=1.00 PE=0.20 K=1.00	K=0.84 / 15 pairs
Project2	PA=0.60 PE=0.12 K=0.55	PA=0.40 PE=0.08 K=0.35	PA=0.40 PE=0.08 K=0.35	K=0.41 / 15 pairs
Project3	PA=0.80 PE=0.16 K=0.76	PA=0.60 PE=0.12 K=0.55	PA=0.60 PE=0.12 K=0.55	K=0.62 / 15 pairs
Project4	PA=0.80 PE=0.16 K=0.76	PA=0.60 PE=0.12 K=0.55	PA=0.60 PE=0.12 K=0.55	K=0.62 / 15 pairs
Project5	PA=0.60 PE=0.12 K=0.55	PA=0.40 PE=0.08 K=0.35	PA=0.20 PE=0.04 K=0.17	K=0.35 / 15 pairs
Project6	PA=0.40 PE=0.08 K=0.35	PA=0.20 PE=0.04 K=0.17	PA=0.20 PE=0.04 K=0.17	K=0.23 / 15 pairs
Project7	PA=0.80 PE=0.16 K=0.76	PA=0.80 PE=0.16 K=0.76	PA=0.60 PE=0.16 K=0.52	K=0.68 / 15 pairs
Project8	PA=0.80 PE=0.16 K=0.76	PA=0.20 PE=0.08 K=0.13	PA=0.40 PE=0.08 K=0.35	K=0.41 / 15 pairs
Project9	PA=1.00 PE=0.20 K=1.00	PA=0.40 PE=0.08 K=0.35	PA=0.40 PE=0.08 K=0.35	K=0.57 / 15 pairs
Project10	PA=0.60 PE=0.12 K=0.55	PA=0.20 PE=0.16 K=0.05	PA=0.20 PE=0.08 K=0.13	K=0.24 / 15 pairs
Project11	PA=0.40 PE=0.08 K=0.35	PA=0.20 PE=0.04 K=0.17	PA=0.20 PE=0.04 K=0.17	K=0.23 / 15 pairs
Project12	PA=0.60 PE=0.12 K=0.55	PA=0.40 PE=0.08 K=0.35	PA=0.80 PE=0.16 K=0.76	K=0.55 / 15 pairs
Project13	PA=0.60 PE=0.12 K=0.55	PA=0.20 PE=0.04 K=0.17	PA=0.40 PE=0.08 K=0.35	K=0.35 / 15 pairs
Project14	PA=0.20 PE=0.04 K=0.17	PA=0.20 PE=0.04 K=0.17	PA=0.00 PE=0.00 K=0.00	K=0.11 / 15 pairs
Project15	PA=1.00 PE=0.20 K=1.00	PA=0.80 PE=0.16 K=0.76	PA=0.80 PE=0.16 K=0.76	K=0.84 / 15 pairs
Project16	PA=0.80 PE=0.16 K=0.76	PA=0.40 PE=0.08 K=0.35	PA=0.40 PE=0.08 K=0.35	K=0.49 / 15 pairs
Project17	PA=0.60 PE=0.12 K=0.55	PA=0.60 PE=0.12 K=0.55	PA=0.80 PE=0.16 K=0.76	K=0.62 / 15 pairs
Project18	PA=0.80 PE=0.16 K=0.76	PA=0.80 PE=0.16 K=0.76	PA=0.80 PE=0.16 K=0.76	K=0.76 / 15 pairs
Project19	PA=0.60 PE=0.12 K=0.55	PA=0.00 PE=0.00 K=0.00	PA=0.20 PE=0.04 K=0.17	K=0.24 / 15 pairs

Table H-1 (continued)

Agreement Scores				
Variable	**Evaluator1+Evaluator2**	**Evaluator1+Evaluator3**	**Evaluator2+Evaluator3**	**Average Kappa Values**
Project20	PA=0.80 PE=0.16 K=0.76	PA=0.40 PE=0.08 K=0.35	PA=0.60 PE=0.12 K=0.55	K=0.55 / 15 pairs
Project21	PA=0.40 PE=0.08 K=0.35	PA=0.60 PE=0.12 K=0.55	PA=0.40 PE=0.08 K=0.35	K=0.41 / 15 pairs
Project22	PA=0.80 PE=0.16 K=0.76	PA=0.60 PE=0.12 K=0.55	PA=0.40 PE=0.08 K=0.35	K=0.55 / 15 pairs
Project23	PA=0.60 PE=0.12 K=0.55	PA=0.60 PE=0.12 K=0.55	PA=1.00 PE=0.20 K=1.00	K=0.70 / 15 pairs
Project24	PA=0.60 PE=0.12 K=0.55	PA=0.40 PE=0.08 K=0.35	PA=0.20 PE=0.04 K=0.17	K=0.35 / 15 pairs
Project25	PA=0.80 PE=0.16 K=0.76	PA=0.60 PE=0.12 K=0.55	PA=0.80 PE=0.16 K=0.76	K=0.69 / 15 pairs
Project26	PA=0.40 PE=0.08 K=0.35	PA=0.60 PE=0.12 K=0.55	PA=0.40 PE=0.08 K=0.35	K=0.41 / 15 pairs
Project27	PA=0.40 PE=0.08 K=0.35	PA=0.00 PE=0.00 K=0.00	PA=0.00 PE=0.00 K=0.00	K=0.12 / 15 pairs
Project28	PA=0.60 PE=0.12 K=0.55	PA=0.20 PE=0.04 K=0.17	PA=0.00 PE=0.00 K=0.00	K=0.24 / 15 pairs
Project29	PA=0.80 PE=0.16 K=0.76	PA=0.60 PE=0.12 K=0.55	PA=0.60 PE=0.12 K=0.55	K=0.62 / 15 pairs
Project30	PA=0.20 PE=0.04 K=0.17	PA=0.60 PE=0.12 K=0.55	PA=0.20 PE=0.04 K=0.17	K=0.29 / 15 pairs
Project31	PA=0.60 PE=0.12 K=0.55	PA=0.20 PE=0.04 K=0.17	PA=0.40 PE=0.08 K=0.35	K=0.35 / 15 pairs
Project32	PA=0.60 PE=0.12 K=0.55	PA=0.60 PE=0.12 K=0.55	PA=1.00 PE=0.20 K=1.00	K=0.70 / 15 pairs
Project33	PA=0.40 PE=0.08 K=0.35	PA=0.60 PE=0.12 K=0.55	PA=0.60 PE=0.12 K=0.55	K=0.48 / 15 pairs
Project34	PA=0.80 PE=0.16 K=0.76	PA=1.00 PE=0.20 K=1.00	PA=0.80 PE=0.16 K=0.76	K=0.84 / 15 pairs
Project35	PA=1.00 PE=0.20 K=1.00	PA=0.60 PE=0.12 K=0.55	PA=0.60 PE=0.12 K=0.55	K=0.70 / 15 pairs

APPENDIX I – The Results of Case Study: The Proposed Model using Kappa Statistics for the Ground Truth Set

Table I-1: The results of proposed model using kappa statistics for the ground truth set

Number of Topics	Stemming is Used	Data Fusion Technique	Project-Scientist Similarity Weight	Weight of the Domain Profile	P@2	R@2	P@3	R@3	P@4	R@4	P@5	R@5
219	No	COMBMNZ	TOPIC03_TECH07	0.5	0.1316	0.1316	0.1228	0.1228	0.1053	0.1053	0.1263	0.1342
219	No	EXPCOMBMNZ	TOPIC03_TECH07	0.5	0.1316	0.1316	0.1579	0.1579	0.1447	0.1447	0.1789	0.2
219	No	COMBMNZ	TOPIC05_TECH05	0.5	0.1316	0.1316	0.1228	0.1228	0.1053	0.1053	0.1263	0.1342
219	No	EXPCOMBMNZ	TOPIC05_TECH05	0.5	0.1316	0.1316	0.1579	0.1579	0.1447	0.1447	0.1789	0.2
219	No	EXPCOMBMNZ	TOPIC07_TECH03	0.5	0.1316	0.1316	0.1579	0.1579	0.1447	0.1447	0.1789	0.2
219	No	EXPCOMBMNZ	TOPIC03_TECH07	1	0.1053	0.1053	0.1579	0.1579	0.1842	0.1842	0.1789	0.1974
219	No	EXPCOMBMNZ	TOPIC05_TECH05	1	0.1053	0.1053	0.1579	0.1579	0.1842	0.1842	0.1789	0.1974
219	No	COMBMNZ	TOPIC07_TECH03	0.5	0.1053	0.1053	0.1053	0.1053	0.0921	0.0921	0.1158	0.1211
219	No	EXPCOMBMNZ	TOPIC07_TECH03	1	0.1053	0.1053	0.1579	0.1579	0.1842	0.1842	0.1789	0.1974
219	No	RR	TOPIC03_TECH07	0.5	0.0526	0.0526	0.0351	0.0351	0.0263	0.0263	0.0211	0.0237
219	No	COMBMNZ	TOPIC05_TECH05	1	0.0526	0.0526	0.0526	0.0526	0.0395	0.0395	0.0316	0.0342
219	No	RR	TOPIC05_TECH05	0.5	0.0526	0.0526	0.0351	0.0351	0.0263	0.0263	0.0211	0.0237
219	No	RR	TOPIC07_TECH03	0.5	0.0526	0.0526	0.0351	0.0351	0.0263	0.0263	0.0211	0.0237
219	No	COMBMNZ	TOPIC03_TECH07	1	0.0263	0.0263	0.0351	0.0351	0.0395	0.0395	0.0316	0.0316
219	No	RR	TOPIC03_TECH07	1	0.0263	0.0263	0.0175	0.0175	0.0395	0.0395	0.0316	0.0368
219	No	RR	TOPIC05_TECH05	1	0.0263	0.0263	0.0175	0.0175	0.0263	0.0263	0.0211	0.0237
219	No	RR	TOPIC07_TECH03	1	0.0263	0.0263	0.0175	0.0175	0.0263	0.0263	0.0211	0.0237
219	No	COMBMNZ	TOPIC07_TECH03	1	0	0	0.0175	0.0175	0.0263	0.0263	0.0211	0.0237
219	Yes	EXPCOMBMNZ	TOPIC03_TECH07	0.5	0.1842	0.1842	0.1404	0.1404	0.1316	0.1316	0.1053	0.1105
219	Yes	EXPCOMBMNZ	TOPIC05_TECH05	0.5	0.1842	0.1842	0.1404	0.1404	0.1316	0.1316	0.1053	0.1105

Table I-1 (continued)

Number of Topics	Stemming is Used	Data Fusion Technique	Project-Scientist Similarity Weight	Weight of the Domain Profile	P@2	R@2	P@3	R@3	P@4	R@4	P@5	R@5
219	Yes	EXPCOMBMNZ	TOPIC07_TECH03	0.5	0.1842	0.1842	0.1404	0.1404	0.1316	0.1316	0.1053	0.1105
219	Yes	EXPCOMBMNZ	TOPIC03_TECH07	1	0.1579	0.1579	0.1579	0.1579	0.1184	0.1184	0.1368	0.1474
219	Yes	EXPCOMBMNZ	TOPIC05_TECH05	1	0.1579	0.1579	0.1579	0.1579	0.1184	0.1184	0.1368	0.1474
219	Yes	EXPCOMBMNZ	TOPIC07_TECH03	1	0.1579	0.1579	0.1579	0.1579	0.1184	0.1184	0.1368	0.1474
219	Yes	COMBMNZ	TOPIC03_TECH07	0.5	0.1316	0.1316	0.1053	0.1053	0.0921	0.0921	0.1158	0.1211
219	Yes	COMBMNZ	TOPIC05_TECH05	0.5	0.0789	0.0789	0.1053	0.1053	0.0921	0.0921	0.0947	0.0974
219	Yes	COMBMNZ	TOPIC07_TECH03	0.5	0.0526	0.0526	0.0351	0.0351	0.0526	0.0526	0.0632	0.0658
219	Yes	RR	TOPIC03_TECH07	0.5	0.0263	0.0263	0.0175	0.0175	0.0658	0.0658	0.0737	0.0842
219	Yes	COMBMNZ	TOPIC03_TECH07	1	0	0	0.0175	0.0175	0.0132	0.0132	0.0211	0.0211
219	Yes	RR	TOPIC03_TECH07	1	0	0	0.0175	0.0175	0.0263	0.0263	0.0421	0.0447
219	Yes	COMBMNZ	TOPIC05_TECH05	1	0	0	0	0	0	0	0.0105	0.0105
219	Yes	RR	TOPIC05_TECH05	1	0	0	0.0175	0.0175	0.0132	0.0132	0.0316	0.0342
219	Yes	RR	TOPIC05_TECH05	0.5	0	0	0	0	0.0263	0.0263	0.0421	0.05
219	Yes	COMBMNZ	TOPIC07_TECH03	1	0	0	0	0	0	0	0.0105	0.0105
219	Yes	RR	TOPIC07_TECH03	1	0	0	0.0175	0.0175	0.0132	0.0132	0.0316	0.0342
219	Yes	RR	TOPIC07_TECH03	0.5	0	0	0	0	0.0263	0.0263	0.0316	0.0395
200	No	EXPCOMBMNZ	TOPIC03_TECH07	1	0.2368	0.2368	0.2105	0.2105	0.1842	0.1842	0.1789	0.1921
200	No	EXPCOMBMNZ	TOPIC05_TECH05	1	0.2368	0.2368	0.2105	0.2105	0.1842	0.1842	0.1789	0.1921
200	No	EXPCOMBMNZ	TOPIC07_TECH03	1	0.2368	0.2368	0.2105	0.2105	0.1842	0.1842	0.1789	0.1921
200	No	EXPCOMBMNZ	TOPIC03_TECH07	0.5	0.1316	0.1316	0.1404	0.1404	0.1447	0.1447	0.1684	0.1816
200	No	EXPCOMBMNZ	TOPIC05_TECH05	0.5	0.1316	0.1316	0.1404	0.1404	0.1447	0.1447	0.1684	0.1816
200	No	EXPCOMBMNZ	TOPIC07_TECH03	0.5	0.1316	0.1316	0.1404	0.1404	0.1447	0.1447	0.1684	0.1816
200	No	COMBMNZ	TOPIC03_TECH07	0.5	0.1053	0.1053	0.1228	0.1228	0.0921	0.0921	0.0842	0.0895
200	No	COMBMNZ	TOPIC05_TECH05	0.5	0.1053	0.1053	0.1228	0.1228	0.0921	0.0921	0.0842	0.0895

Table I-1 (continued)

Number of Topics	Stemming is Used	Data Fusion Technique	Project-Scientist Similarity Weight	Weight of the Domain Profile	P@2	R@2	P@3	R@3	P@4	R@4	P@5	R@5
200	No	COMBMNZ	TOPIC07_TECH03	0.5	0.0789	0.0789	0.1228	0.1228	0.0921	0.0921	0.0842	0.0895
200	No	COMBMNZ	TOPIC03_TECH07	1	0.0526	0.0526	0.0526	0.0526	0.0395	0.0395	0.0526	0.0579
200	No	RR	TOPIC03_TECH07	1	0.0263	0.0263	0.0175	0.0175	0.0132	0.0132	0.0316	0.0368
200	No	RR	TOPIC03_TECH07	0.5	0.0263	0.0263	0.0351	0.0351	0.0526	0.0526	0.0421	0.0526
200	No	RR	TOPIC05_TECH05	1	0.0263	0.0263	0.0175	0.0175	0.0132	0.0132	0.0316	0.0368
200	No	RR	TOPIC05_TECH05	0.5	0.0263	0.0263	0.0351	0.0351	0.0395	0.0395	0.0316	0.0395
200	No	RR	TOPIC07_TECH03	1	0.0263	0.0263	0.0175	0.0175	0.0132	0.0132	0.0316	0.0368
200	No	RR	TOPIC07_TECH03	0.5	0.0263	0.0263	0.0351	0.0351	0.0395	0.0395	0.0316	0.0395
200	No	COMBMNZ	TOPIC05_TECH05	1	0	0	0.0175	0.0175	0.0132	0.0132	0.0105	0.0105
200	No	COMBMNZ	TOPIC07_TECH03	1	0	0	0	0	0	0	0	0
200	Yes	EXPCOMBMNZ	TOPIC03_TECH07	1	0.1053	0.1053	0.1053	0.1053	0.0921	0.0921	0.0947	0.1
200	Yes	EXPCOMBMNZ	TOPIC03_TECH07	0.5	0.1053	0.1053	0.1053	0.1053	0.1053	0.1053	0.0947	0.1
200	Yes	EXPCOMBMNZ	TOPIC05_TECH05	1	0.1053	0.1053	0.1053	0.1053	0.0921	0.0921	0.0947	0.1
200	Yes	EXPCOMBMNZ	TOPIC05_TECH05	0.5	0.1053	0.1053	0.1053	0.1053	0.1053	0.1053	0.0947	0.1
200	Yes	EXPCOMBMNZ	TOPIC07_TECH03	1	0.1053	0.1053	0.1053	0.1053	0.0921	0.0921	0.0947	0.1
200	Yes	EXPCOMBMNZ	TOPIC07_TECH03	0.5	0.1053	0.1053	0.1053	0.1053	0.1053	0.1053	0.0947	0.1
200	Yes	COMBMNZ	TOPIC03_TECH07	0.5	0.0789	0.0789	0.0877	0.0877	0.0921	0.0921	0.1158	0.1237
200	Yes	COMBMNZ	TOPIC05_TECH05	0.5	0.0789	0.0789	0.0877	0.0877	0.0789	0.0789	0.1053	0.1105
200	Yes	RR	TOPIC03_TECH07	1	0.0263	0.0263	0.0175	0.0175	0.0132	0.0132	0.0105	0.0105
200	Yes	RR	TOPIC03_TECH07	0.5	0.0263	0.0263	0.0175	0.0175	0.0132	0.0132	0.0105	0.0105
200	Yes	RR	TOPIC05_TECH05	1	0.0263	0.0263	0.0175	0.0175	0.0132	0.0132	0.0105	0.0105
200	Yes	RR	TOPIC05_TECH05	0.5	0.0263	0.0263	0.0175	0.0175	0.0132	0.0132	0.0105	0.0105
200	Yes	COMBMNZ	TOPIC07_TECH03	0.5	0.0263	0.0263	0.0526	0.0526	0.0526	0.0526	0.0632	0.0658
200	Yes	RR	TOPIC07_TECH03	1	0.0263	0.0263	0.0175	0.0175	0.0132	0.0132	0.0105	0.0105

Table I-1 (continued)

Number of Topics	Stemming is Used	Data Fusion Technique	Project-Scientist Similarity Weight	Weight of the Domain Profile	P@2	R@2	P@3	R@3	P@4	R@4	P@5	R@5
200	Yes	RR	TOPIC07_TECH03	0.5	0.0263	0.0263	0.0175	0.0175	0.0132	0.0132	0.0105	0.0105
200	Yes	COMBMNZ	TOPIC03_TECH07	1	0	0	0	0	0.0132	0.0132	0.0105	0.0105
200	Yes	COMBMNZ	TOPIC05_TECH05	1	0	0	0	0	0	0	0	0
200	Yes	COMBMNZ	TOPIC07_TECH03	1	0	0	0	0	0	0	0	0

METU LIBRARY

TEZ FOTOKOPİ İZİN FORMU

ENSTİTÜ

Fen Bilimleri Enstitüsü ☐

Sosyal Bilimler Enstitüsü ☐

Uygulamalı Matematik Enstitüsü ☐

Enformatik Enstitüsü ☒

Deniz Bilimleri Enstitüsü ☐

YAZARIN

Soyadı : (Selçuk) Doğan
Adı : Gonca Hülya
Bölümü : Bilişim Sistemleri

TEZİN ADI (İngilizce) : Expert Finding in Domains with Unclear Topics

TEZİN TÜRÜ : Yüksek Lisans ☒ Doktora ☐

1. Tezimin tamamı dünya çapında erişime açılsın ve kaynak gösterilmek şartıyla tezimin bir kısmı veya tamamının fotokopisi alınsın. ☒
2. Tezimin tamamı yalnızca Orta Doğu Teknik Üniversitesi kullancılarının erişimine açılsın. (Bu seçenekle tezinizin fotokopisi ya da elektronik kopyası Kütüphane aracılığı ile ODTÜ dışına dağıtılmayacaktır.) ☐
3. Tezim bir (1) yıl süreyle erişime kapalı olsun. (Bu seçenekle tezinizin fotokopisi ya da elektronik kopyası Kütüphane aracılığı ile ODTÜ dışına dağıtılmayacaktır.) ☐

Yazarın imzası ..... Tarih .....