Finding Topic-centric Identified Experts based on Full Text Analysis

Hanmin Jung, Mikyoung Lee, In-Su Kang, Seung-Woo Lee, Won-Kyung Sung

Information Service Research Lab., KISTI, Korea
jhm@kisti.re.kr

Abstract. This paper shows a method for finding topic-centric experts from open access metadata and full text documents. Topic-centric information including experts is served on OntoFrame, which is a Semantic Web-based academic research information service supporting R&D activities. URI scheme-based OntoFrame provides three entity pages: topic, person, and event. 'Persons by Topic' in topic page lists up topic-centric identified experts. SPARQL query is used to retrieve them from RDF triple store through backward chaining. We gathered CiteSeer open access metadata and full text documents with the amount of about 110,000 papers. Using about 160,000 abundant topics, OntoFrame now serves topic-centric identified experts and relevant information acquired by full text analysis.

1 Introduction

Finding experts is useful in such cases: seeking for consultants, collaborators, and speakers. It also provides a source of information to supplement or complement academic sources including metadata [7], thus, receives increased attention in recent years. However, identification resolution is not considered significantly even though this research topic mainly deals with persons. Many studies concentrate only on string-based person names [1] [2] [5] [6]. Semantic Web can be one of competent solutions for managing identified experts through underlying URI scheme. Another consideration is to guarantee reliability on the results of the task. Deep analysis based on full text documents is needed in that topically-classified documents in high precision ensure finding the right persons for each topic. On the basis of these considerations, we propose an experts-finding method based on identity resolution and full text analysis, and further extract topic-centric information such as 'Topic Trends' and 'Institutions by Topic'. Chapter 2 indicates several previous studies. Chapter 3 explains how to acquire topic-centric information based on a Semantic Web Framework.

2 Related Studies

The sources for finding experts are various: documents, programs, e-mails, databases, citations, communities and so on. Finding expertise information from e-mails with four simple binary association methods was proposed by [1]. [5] investigated the

expertise of users and experts by combining information retrieval techniques. However, such e-mails and communities are insufficient to extract the right experts for a specific topic because they give clues about only relationship and context.
An experts-finding study based on full text documents related with persons and on a set of terms in them was introduced [2]. It extracts similar experts by measuring similarity between term vectors. However, it is not able to indicate which topics are related with experts, but only provides a bundle of persons as the results. ExpertFinder [6] recommends persons with a lot of documents for a given topic. A keyword phrase is used to retrieve relevant documents, but the results are unsatisfactory because reasonable candidates are not listed within the top three or four candidates in most cases. Its slow response time and incorrect relationship between persons and documents are also problems. Another interesting study, performed by [8], introduced three innovative points: document authority in terms of their PageRanks, co-occurrence model, and multiple levels of associations between experts and query terms. It finds variants in experts' names for identity recognition, but failed to identify different persons with the same name uniquely.

3 Acquiring Topic-Centric Information

3.1 OntoFrame: an Academic Research Information Service

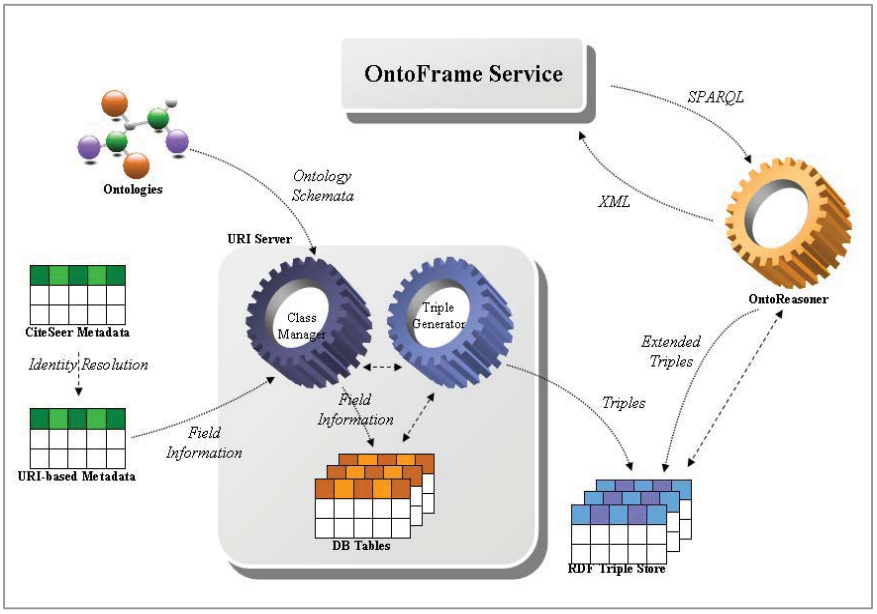

Fig. 1. OntoFrame Architecture

OntoFrame is a Semantic Web-based service which provides academic research information for supporting R&D activities [3]. Its two main components are URI server and OntoReasoner (inference engine). The latter interacts with user interfaces through receiving SPARQL queries and returning XML results. We introduce SPARQL rather than inflexible SQL because it is easy to construct queries with only knowledge on ontology schema. OntoReasoner also expands knowledge in ways of forward-chaining inference. The URI server has several functions: ontology schema parsing and loading, DB schema creation, ontology instance loading, and RDF triple generation as shown in figure 1. When a new instance is inserted into the server, triple generator makes triples for the instance. The triples are then stored in RDF triple store, and further would be referred by OntoReasoner.
OntoFrame distinguishes from other academic research information services such as CiteSeer (http://citeseer.ist.psu.edu/) and Google Scholar (http://scholar.google.com/) because it provides information acquired by inference beyond metadata. 'Persons by Topic', 'Topic Trends', and 'Social Network' are representative information served by OntoFrame.

3.2 Data Gathering and Refining

The Open Archives Initiative (OAI, http://www.openarchives.org/) develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content. CiteSeer (http://citeseer.ist.psu.edu/oai.html) also supports OAI, and thus allows downloading its own open access metadata which includes title, authors, publication year and so on.
Identity resolution is an obligatory task for transforming string-based data to semantic data [4]. Various forms of institution names in the metadata are mapped to a set of normalized institution names[1], e.g. "U. Kassel" and "University of Kassel." We also identify different persons with the same name. There are a few metadata fields available for distinguishing authors such as affiliation, e-mail, and co-authors. It is possible to determine whether two authors with the same name are different or not using their affiliations and e-mails. However, affiliation and e-mail fields are not obligatory in many cases including CiteSeer metadata. Co-authorship information plays an important role in resolving identity problems because co-author field is usually filled up in metadata, and further many authors maintain co-authorship relation regardless of affiliation change. We consider two authors with the same name as the identical person when they share the identical co-author(s), otherwise they remain as different persons. 'sameAs' relation would compensate the short coverage of this method based on co-authorship. All of their information, including papers and topics, will be merged as one when we connect two authors with 'sameAs' relation later.
After identity resolution, we assign URI for each entity; for example, paper "A Bayesian Multiple Models Combination Method for Time Series Prediction" with 'http://www.kisti.re.kr/isrl/ResearchRefOntology#ART_00000000000000458673', topic "markov model" with 'http://www.kisti.re.kr/isrl/ResearchRefOntology#TOP_00000000000000046687' and person

[1] currently, about 14,000

"V. Petridis" with 'http://www.kisti.re.kr/isrl/ResearchRefOntology#PER_00000000000000128292'.

3.3 Topic Extraction

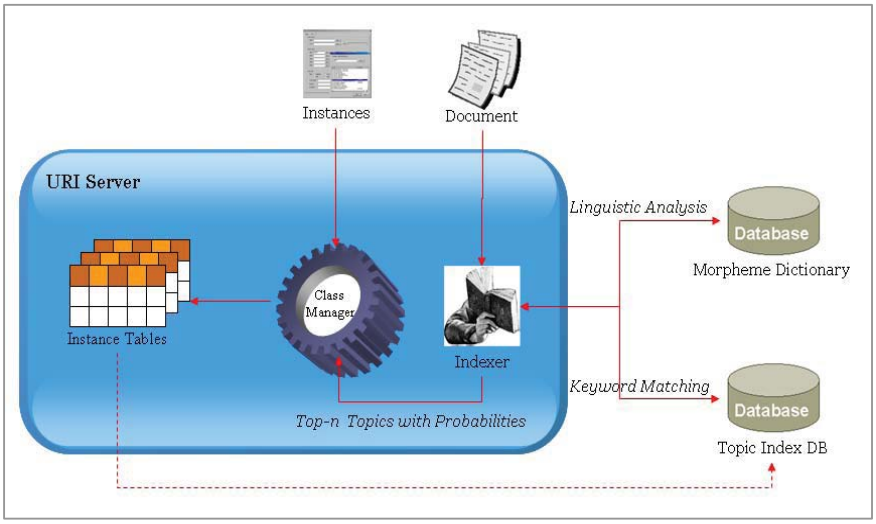

Fig. 2. Workflow of Topic Extraction based on Full Text Documents

Extracting topics from papers is the most basic task to acquire topic-centric experts. As full text documents as well as metadata of CiteSeer are available, we use the documents. Extracted topics are assigned to each paper. The followings explain the stages of the extraction as shown in figure 2; First, indexer extracts index terms from a given document. Second, the terms are matched with topic keywords in topic index DB[2]. Third, successfully matched terms are ranked by the following algorithms, and then we select *top-n* (currently, five) topics for the input document.

(1) Index term list: The *k*th document $D_k = \{t_{k1},...,t_{km}\}$ have *m* index terms. t_{ki} indicates the *i*th index term in the document.

(2) Topic keyword list: Topic keyword list $S = \{s_1,...,s_p\}$ has *p* keywords.

(3) TF (Term Frequency) of index term: $tf_{D_k}(t)$ is the term frequency of index term *t* in document D_k.

[2] Topic keyword and topic are the same in this study. Successfully matched index terms are also a subset of topic keywords because the terms are always a member of topic keywords in topic index DB.

(4) TF of the index term matched with topic keyword: $tf_{D_k^S}(t)$ is the term frequency of the index term *t* found in topic keyword DB. The frequency originates from $tf_{D_k}(t)$.

Topic weighting formula: $r(t) = \frac{tf_{D_k^s}(t)}{\sum_{t' \in Top5} tf_{D_k^s}(t')}$ for *top-5* ranked topics

3.4 Finding Experts

Many factors can be considered for finding experts: the number of papers, impact factor of sources, the degree of citations, hub persons in social network and so on. Currently, we take into account only the number of papers for several reasons. A great portion of source field in CiteSeer open access metadata has no information. Citation information also may be incomplete when compared with CiteSeer service page. We also do not consider social network because prosperous co-authorship with other persons does not always guarantee specialty on a topic.
Acquiring topic-centric experts on OntoFrame requires querying to RDF triple store based on DBMS. 'Persons by Topic' is retrieved directly from the database through SPARQL query (shown as follows) and automatic SPARQL-to-SQL conversion. The query searches papers (?accomplishment) of which topic area is *topicTerm*, and then retrieves authors (?person) of the papers. Figure 3 shows backward chaining flow starting from *topicTerm*.

```
SELECT ?person ?perRep ?perEngName ?perKorName ?institution ?instEng-
Name ?instKorName
WHERE
{
        ?topicArea isrl:hasTopicTermOfAccomplishment topicTerm .
        ?accomplishment isrl:hasTopicAreaOfAccomplishment ?topicArea .
        ?accomplishment isrl:createdByPerson ?person .

        OPTIONAL {?perRep isrl:standForSameAsGroupOf ?person . }

        OPTIONAL { ?person isrl:engNameOfPerson ?perEngName . }
        OPTIONAL { ?person isrl:korNameOfPerson ?perKorName . }

        ?person isrl:hasInstitutionOfPerson ?institution .

        OPTIONAL { ?institution isrl:engNameOfInstitution ?instEngName . }
        OPTIONAL { ?institution isrl:korNameOfInstitution ?instKorName . }
}
ORDER BY ?person
```

'createdByPerson' is one of derived properties induced by user-defined inference rules. It reduces the distance of backward path to find 'Persons by Topic' in ways that go through directly to 'Person' rather than without passing through 'CreatorInfo' (the dotted line in figure 3). After retrieving persons, OntoReasoner performs post-processing for ranking them by descending order of the number of their own papers.

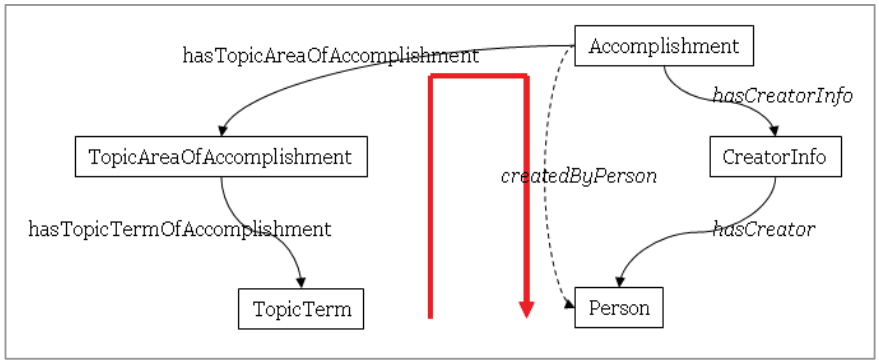

Fig. 3. Backward Chaining Path for Finding 'Persons by Topic' (Experts for a Topic)

3.5 Topic-centric Information

OntoFrame provides several entity-centric pages such as topic, person, and event. Each entity page consists of a stack of information related with a specific entity. For example, topic page serves 'Search Results', 'Topic Trends', 'Also Try', 'Persons by Topic', 'Institutions by Topic', 'Papers by Topic', and 'Researcher Group (Social Network)' as shown in figure 4. 'Topic Trends' shows relevant topics by year. We define the relevance as the topics extracted from the same paper. 'Institutions by Topic' for dominant institutions is similar to 'Persons by Topic'. 'Papers by Topic' shows papers classified semantically into a topic.

4 Conclusions

We gathered 114,337 papers (2000 ~ 2006) from CiteSeer open access metadata. They include 161,853 persons and 17,093 institutions. 160,568 topic keywords[3] were extracted from titles and abstracts. Average consuming time for extracting maximum 5 topics from a paper is about 1.6 seconds. Within three seconds are enough to generate an entity page including 'Persons by Topic' on OntoFrame[4].

[3] Simple and compound nouns were extracted automatically and filtered manually by human dictionary constructors.

[4] The whole system will appear in Poster/Demo Track of ISWC2007.

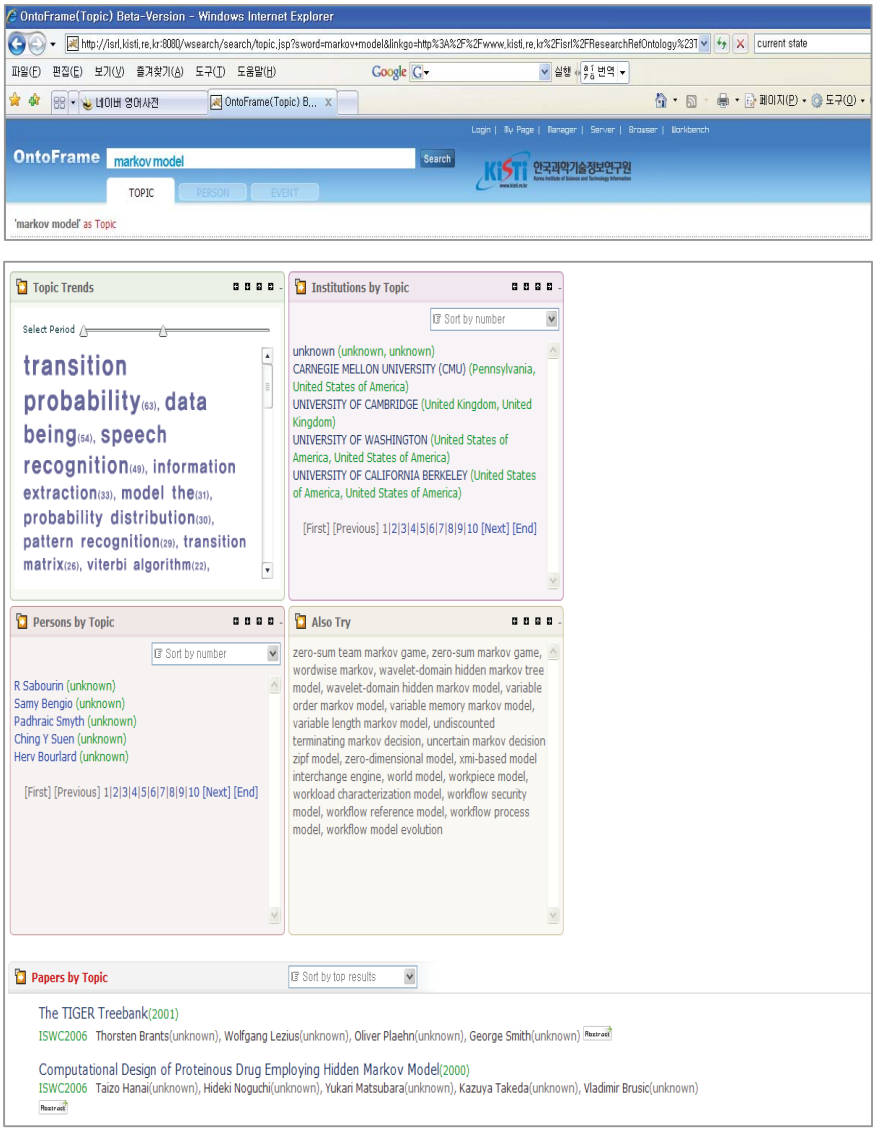

Fig. 4. Example of Topic Page for 'markov model' ('Persons by Topic' shows ranked experts.)

This paper showed a method for finding topic-centric identified experts from CiteSeer open access metadata and full text documents. Topic extraction based on full text analysis enables to construct topically-classified papers, and inference makes propagation to persons and institutions. SPARQL query retrieves URI-based 'Persons by Topic' from RDF triple store. Our future work includes introducing usability test to

evaluate the performance of topic extraction and experts-finding in comparative ways with Google Scholar and CiteSeer.

References

1. Balog, K. and Rijke, M.: Finding Experts and Their Details in E-mail Corpora. *In Proceedings of the 15th International Conference on World Wide Web* (2006)
2. Balog, K. and Rijke, M.: Finding Similar Experts. *In Proceedings of the 30th Annual International ACM SIGIR Conference* (2007)
3. Jung, H., Lee, M., Sung, W., and Park, D.: *Semantic Web-Based Services for Supporting Voluntary Collaboration among Researchers Using an Information Dissemination Platform*. In Journal of Data Science Journal 6(1) (2007)
4. Jung, H. and Sung, W.: Construction of Semantic Web-based Knowledge Using Text Processing. *In Proceedings of the 4th International Conference on Information Technology : New Generations* (2007)
5. Liu, X., Croft, W., and Koll, M.: Finding Experts in Community-Based Question-Answering Services. *In Proceedings of the 14th ACM International Conference on Information and Knowledge Management* (2005)
6. Mattox, D., Maybury, M., and Morey, D.: Enterprise Expert and Knowledge Discovery. *In Proceedings of the 8th International Conference on Human-Computer Interaction* (1999)
7. Yimam, D.: *Expert Finding Systems for Organizations: Domain Analysis and the DEMOIR Approach*. Beyond Knowledge Management: Sharing Expertise. MIT Press (2000)
8. Zhu, J., Song, D., Rüger, S., Eisenstadt, M., and Motta, E.: The Open University at TREC 2006 Enterprise Track Expert Search Task. *In Proceedings of the 15th Text REtrieval Conference* (2006)