An adaptive ontology-based approach to identify correlation between publications

Huayu Wu
School of Computing
National University of Singapore
wuhuayu@comp.nus.edu.sg

Hideaki Takeda
National Institute of Informatics
Tokyo, Japan
takeda@nii.ac.jp

Masahiro Hamasaki
National Institute of Advanced Industrial Science and Technology (AIST)
Tokyo, Japan
hamasaki@ni.aist.go.jp

Tok Wang Ling
School of Computing
National University of Singapore
lingtw@comp.nus.edu.sg

Liang Xu
School of Computing
National University of Singapore
xuliang@comp.nus.edu.sg

ABSTRACT

In this paper, we propose an adaptive ontology-based approach for related paper identification, to meet most researchers' practical needs. By searching ontology, we can return a diverse set of papers that are explicitly and implicitly related to an input paper. Moreover, our approach does not rely on known ontology. Instead, we build and update ontology for a collection with any domain of interest. Being independent from known ontology, our approach is much more adaptive for different domains.

Categories and Subject Descriptors

H.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Information Search and Retrieval

General Terms

Measurement

Keywords

correlation of publications, ontology-based, adaptive approach

1. INTRODUCTION

Searching adequate related papers from online resources is needed by most researchers. Two typical ways to find related papers in most online collections are based on citation tracing and keyword tracing. Citation tracing retrieves related papers based on citations (references) of an input paper [6, 4]. However, citation normally lacks preciseness and completeness. A paper may cite another paper to which it is related on a minor part; and we can never expect a paper cites all related papers in the collection.

Keyword search in most online libraries and web search engines normally returns a meaningful set of related papers. A main problem is that only papers containing exactly the same keywords are returned. Papers with similar concepts but different author-defined terminology, or papers with related techniques to the input technical terms cannot be returned by keyword search. Another keyword tracing based approach is paper clustering [7]. However, a paper may be related to papers in different clusters, from different aspects. Putting a paper into a unique cluster will break down its relevance to other clusters.

In this paper, we propose an adaptive ontology-based approach for related paper identification. Ontology often provides very instructive information, which makes data processing more accurate. Due to the high cost of ontology engineering, ontology is normally built based on purposes and domains. To aid an online collection to return related papers of an input paper, ontology seems not helpful because the collection can be an online library or a publisher's website with papers across multiple domains. To incorporate useful background information into related paper identification, we propose an adaptive ontology construction method which builds ontology for a collection based on its current papers and updates the ontology as papers increase. After that for each given paper, we find all relevant papers by searching the ontology, and return a set of ranked results.

2. ADAPTIVE ONTOLOGY

We build a weighted network of terms as an ontology for any paper collection, regardless how many domains it covers. The ontology is adaptive because it is constructed based on the current papers in the collection, and updated as the collection expands. Because most published papers are accurate in concept description, the adaptive ontology based on those papers has a high accuracy, though we do not manually revise it with expertise.

2.1 Data sources and term base

As mentioned early, for each collection we use the papers it contains to construct ontology. Because processing the full text of every paper is time consuming, we only use paper abstracts as data sources to build ontology. Normally the core concepts of each paper are all highlighted in its abstract.

Conceptually an ontology is modeled as a graph, in which each node represents a technical *term*. A term can be a single word or a phrase to describe a technical concept, method,

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0632-4/11/03.

theorem and so on. We obtain 124,222 author-provided terms from CiNii [1], which is a general-purposed database for academic papers that provides metadata of more than 12 million papers, as an initial term base. This term base can be extended by incorporating new terms in particular online paper collections, which are provided by paper authors or mined from papers in the corresponding collection [5].

2.2 Term relationship finding

After obtaining the paper abstracts and a term base for a paper collection, we try to find useful terms and the relationship between them to construct ontology. In our first study, we ignore the types of relationships, and only consider a unique *related* relationship between every two terms. As a follow-up project, we will use NLP techniques to further analyze the text to classify different types of relationship, to improve the quality of our system.

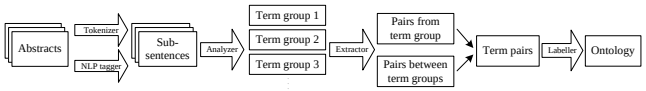

Figure 1: Ontology construction process

The general process is shown in Fig. 1. Generally we first tokenize the abstracts into sub-sentences. For each sentence in an abstract, if there is no comma we simply consider it as a sub-sentence. If a sentence contains two or more parts delimited by comma, we check each part to see whether it contains a verb to connect two nouns, under the help of an NLP tagger [3]. If a part contains such a verb, we consider it as a separate sub-sentence; otherwise, we do not tokenize it from other sub-sentences. Then we identify all term pairs within each sub-sentence. As long as two terms appear within the same sub-sentence, we consider they are related. More explanations can be found in our report [2].

We set a threshold to filter out the rare term pairs. Then we construct a graph with nodes of all terms found and edges of relatedness between each pair of nodes. Furthermore, each edge is assigned a label to indicate its importance. We assume the frequency of the co-occurrence of two terms is directly proportional to the importance of this pair, and the frequency of each term is inversely proportional to the importance of a term pair. The idea is similar to the TF*IDF measure in IR. The importance of the relationship between two terms A and B is defined as:

$$importance(A,\ B) = \frac{fre(A,B)}{avg(fre(A),fre(B))}$$

where *fre(A, B)* is the frequency of the co-occurrence of A and B, *avg(fre(A), fre(B))* is the average of the frequency of A and the frequency of B, in all sub-sentences.

Once more papers are added to the collection, regardless of whether they are on the existing topics or introduce new topics, the ontology simply expands to include new nodes or new edges, and updates the occurrences of relevant nodes and edges.

3. RELEVANT PAPER IDENTIFICATION

3.1 Inverted index

We represent each paper by a set of terms. The terms are either provided by the authors or extracted from the abstract or the full paper, as we do in extending the term base in Section 2.1. Then we build an inverted list for each term, to index all the papers containing this term.

3.2 Finding relevant papers

For an input paper P with a set of terms, we aim to find all papers in the collection that are related to P, ranked by relatedness. The general idea is for each term in P, we find all terms linked to it in the ontology (including itself). Then using the inverted index, we can get a set of papers containing any of these terms, which are considered related to P. Then we rank these papers.

We compute a score for each related paper based on the number of its terms linked to the input paper terms in the ontology, the importance of each term pair and the percentage of all its terms that are related to the input paper terms. The score of a candidate related paper p_i with m terms, w.r.t. an input paper p with n terms is:

$$score(p_i) = \frac{m'}{m}\sum_{m}^{j=1}\sum_{n}^{k=1} importance(t_{i_j},\ t_k)$$

where m' is the number of terms in p_i that are identical or linked to some terms in p in the ontology, t_{i_j} is the j-th term in p_i, and t_k is the k-th term in p.

Generally we find all pairs of terms from the input paper and each related paper separately, and sum up the *weights* of these pairs. The weight of a pair of terms depends on the positional relationship between them in the ontology: (1) the two terms correspond to the same node (they are identical term), then the weight between them is 1; (2) the two terms are linked by an edge, then the weight is the importance of the corresponding edge (defined in Section 2.2) in the ontology and (3) the two terms are neither identical nor linked in the ontology, then the weight is 0 (or computed by the path information, e.g. [7], in our ongoing work). Last, for each related paper we normalize the summed score to avoid a paper with too many terms getting a higher score.

4. EXPERIMENTS

We use 1345 proceeding papers from the latest four WWW conferences and the latest four SIGMOD conferences as a paper collection, and construct an ontology with 2005 terms (nodes) and 10904 relationships (edges). We choose three papers from each conference and find all related papers of each of them, with ranking. The results show that for each paper, our approach can effectively return a list of related papers, and the ranking is reasonable from a researcher's aspect. The detailed experiments can be found in [2].

5. REFERENCES

[1] http://ci.nii.ac.jp

[2] http://www.comp.nus.edu.sg/∼wuhuayu/fulltext.pdf

[3] http://www-tsujii.is.s.u-tokyo.ac.jp/∼tsuruoka/postagger

[4] M. M. Kessler. Bibliographic coupling between scientific papers. In American Document, 14: 10-25, 1963.

[5] A. Parameswaran, H. Garcia-Molina, and A. Rajaraman. Towards the Web of Concepts: Extracting Concepts from Large Datasets. Technical Report 2009. Stanford InfoLab.

[6] H. Small. Co-citation in the scientific literature: a new measure of the relationship between two documents. In Journal of the American Society for Information Science, 24: 265-269, 1973.

[7] X. Zhang, L, Jing, X. Hu, M. Ng, and X. Zhou. A comparative study of ontology based term similarity measures on PubMed document clustering. In DASFAA, pp. 115-126, 2007.