3I1-R-9-2

構造的関連性学習を用いた論文特許検索

Cross-Domain Academic Search using Structural Correspondence Learning

森純一郎
Junichiro Mori

東京大学総括プロジェクト機構「プラチナ社会」総括寄付講座
Presidential Endowed Chair for "Platinum Society"

We propose a method to automatically associate documents from different domains such as scientific paper and patent. The proposed method enables cross-domain academic search on the scientific data. Borrowing ideas from multi-task learning and structural correspondence learning, our approach automatically identifies correspondences among the words from different domains using a small number of so-called concepts. Our method models the correlation between the concepts and all other words by training linear classifiers on the documents from different domains.

1. Introduction

The vast amount of human knowledge has been accumulated in the form of academic papers. The Web of Science, one of academic citation indexes, currently is indexing tens of millions of academic publications and the number of its records is rapidly growing. With the recent advances in Web technologies, we can easily access the large amount of scientific data. For example, the Web of Science provides an intuitive Web interface that enables a user to search for academic publications from multiple databases. Furthermore, Microsoft Academic Search provides many ways to explore not only academic publications but also other information such as authors, organizations, and keywords. It also recently provides a set of APIs to allow a user to develop own tools on top of the data.

Given the large amount of scientific data from a variety of information domains that is easily available from online, one of key questions is how to associate the information from different domains. From the viewpoint of information search, it is a task of associating a source (e.g., scientific paper) from one domain to a target (e.g., patent) from another domain or the other way round. There exist several potential linkages between different domains such as paper–patent, paper–firm, and paper–website, which can be useful to search for related information on the vast amount of scientific data. For example given a scientific paper about a particular topic, one might like to find related patents or firms/products on the topic. The simple way of associating documents from different domains is to do thesaurus-based mapping among the domains or to model domain-independent concept using comparable corpora. However, vocabulary of one domain (e.g., scientific paper) is often different from vocabulary of another domain (e.g., patent), which cause sparse-overlapping regions of the feature space when mapping documents from two different domains. One way to overcome this mapping problem is to find a cross-domain representation for documents in different domains, which enables extend the representation of a document by transferring the knowledge between domains. Intuitively, such a cross-domain representation can be considered as a concept space that underlies different domains.

In this project, we propose a method to automatically associate documents from different domains such as scientific paper, patent, and firm/product of scientific data. The proposed method enables cross-domain academic search on the scientific data. Borrowing ideas from the field of multi-task learning and structural correspondence learning in the field of natural language processing, our approach automatically identifies correspondences among the words from different domains using a small number of so-called concepts. A concept is a commonly used keyword from one domain and another domain, which holds a relevant semantics to respective domains. We plan to develop a cross-domain academic search engine on top of the proposed method that enable search for related information from different domains of scientific data.

Our contributions of the project are twofold: First, a general method for cross-domain academic search using structural corresponding learning. Second, a cross-domain academic search engine that enables search for different domains between academic paper and patent and between academic paper and firm and between academic paper and website.

Cross-domain academic search improves the process of retrieving scientific data that is distributed over a variety of online information sources. For example, a user might use an academic paper of his or her interest as an input. Then, a cross-domain academic search engine will return related patents, firms, and websites from different information sources. Because our method can be easily incorporated with a search engine, major academic search engines such as Microsoft Academic Search, CiNII (NII), and J-Global (JST) will enjoy a function of cross-domain academic search that can be provided to users.

2. Related Research

In this project, we propose a method to automatically associate documents from different domains such as scientific paper, patent, and firm of scientific data. Underlying ideas are based on structural correspondence learning in the field of natural language processing [Blitzer 06]. Structural correspondence learning is a the-

連絡先: 森純一郎，東京大学総括プロジェクト機構「プラチナ社会」総括寄付講座，東京都文京区本郷 7-3-1 伊藤国際学術研究センター，03-5841-1596，jmori@platinum.u-tokyo.ac.jp

ory to automatically induce correspondences among features from different domains. Several applications of structural correspondence learning have been recently studied in the field of Natural Language Processing and Information Retrieval [Wang et al. 11] [Prettenhofer and Stein 10]. However, to the best of our knowledge, the identification and utilization of the theory of structural correspondence learning to cross-domain academic search on the vast amount of scientific data has not been studied before.

Structural correspondence learning was originally proposed for domain adaptation. Domain adaptation refers to the problem of adapting a statistical classifier trained on data from one source domain(s) to a different target domain. Cross-domain academic search is closely related to an unsupervised domain adaptation problem by automatically inducing correspondences between documents from different domains. In our case concepts correspond to the generalized features across domains. Our method models the correlation between the concepts and all other features by training linear classifiers on the unlabeled data from different domains. This model is used to induce correspondences among features from the different domains and to learn a shared representation that is meaningful across different domains.

3. Method

The proposed method for cross-domain academic search is illustrated in Figure. 1 and it consists of following main steps.

- Collect scientific data including academic papers, patents, firms, and Wikipedia. We plan to use Microsoft Academic Search APIs to collect the information about academic papers. In particular we focus on the field of computer science for efficiency reasons. We also plan to collect scientific data including academic papers, patents, and firms using CiNII APIs and J-Global APIs. For this, we plan to closely cooperate with Prof. Hideaki Takeda at NII and Dr. Takayuki Saito at JST.

- Extract a concept across different domains. Given the large amount of scientific data, we next extract commonly used keywords, so-called concepts, across different domains. We use the following heuristics to extract a set of concepts: First, we select a subset vocabulary from one domain, which contains frequent keywords. Second, for each keyword in the vocabulary we find its same (or similar) keyword in the vocabulary from another domain.

- Identify correspondences between different domains using concepts. We model the correlation between each concept and all other words. For this purpose, we train linear classifiers that predict whether or not concept occur in a document, based on the other words. A training set is created for each concept. The training set contains documents from the domains that concepts are extracted. Thus, the classifiers can be considered as cross-domain classifiers. We then reduce the dimension of a parameter matrix of the linear classifiers to further identify correlations across concepts. In other words, we obtain common substructures among the linear classifiers that can be used as a mapping function to associate the original representation of a document to the concept space with its cross-domain representation.

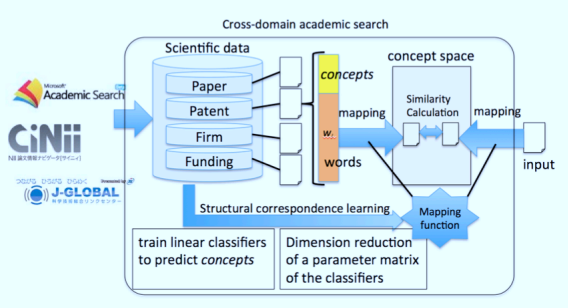

図 1: Proposed Method

- Using the mapping function, we extend the representation of each document to the cross-domain representation. Then we compute a similarity between the documents from different domains in the extended concept space. Based on the similarity, we return the ranked list of documents from respective domains that are relevant to a given document. We plan to investigate different similarity functions to rank the result depending on a domain.

To evaluate the proposed method, we build a bench dataset for cross-domain academic search. We use standard evaluation measures in information retrieval such as average precision (AP) and mean reciprocal rank (MRR) to evaluate the accuracy of the proposed method. We also plan to make the search engine available online and conduct a user study.

4. Conclusions

We propose a method to automatically associate documents from different domains such as scientific paper and patent. The proposed method enables cross-domain academic search on the scientific data. Borrowing ideas from multi-task learning and structural correspondence learning, our approach automatically identifies correspondences among the words from different domains using a small number of so-called concepts. Our method models the correlation between the concepts and all other words by training linear classifiers on the documents from different domains.

参考文献

[Blitzer 06] J. Blitzer, R. McDonald, and F. Pereira: Domain adaptation with structural correspondence learning, Proceedings of EMLNLP06, pp.120–128, 2006.

[Prettenhofer 10] P. Prettenhofer and B. Stein: Cross-language text classification using structural correspondence learning, Proceedings of ACL10, pp.1118–1127, 2010.

[Wang et al. 11] H. Wang, H. Huang, F. Nie, and C. Ding: Cross-language web page classification via dual knowledge transfer using nonnegative matrix tri-factorization, Proceedings of SIGIR11, pp.993–942, 2011.