Combining Topic Model and Co-author Network for KAKEN and DBLP linking

Duy-Hoang Tran[1] , Hideaki Takeda[2], Minh-Triet Tran[1]

[1] Faculty of Information Technology,
University of Science Ho Chi Minh City, Vietnam
{tdhoang, tmtriet}@fit.hcmus.edu.vn
[2] National Institute of Informatics,
takeda@nii.ac.jp

Abstract. The Web of Data is based on two simple ideas: use the RDF data model to public structured data on the Web and use RDF links to interlink data from different data sources. In this paper, we describe our experience in linking KAKEN, a database provides the latest information of researcher projects in Japan, and the DBLP Computer Science Bibliography. Using these links one can navigate from a computer scientist in KAKEN to his publications in the DBLP database. The problem of linking KAKE researchers and DBLP authors based on their name is having a poor result. We proposed combining LDA based topic model and co-author network approach to improve accuracy of linking.

Keywords: Web of Data, LDA Model, Co-author Network, Connected Triple.

1 Introduction

The Web of Data is constantly growing over the last three years and has started to span data sources from a wide range of domains such as geographic information, people, companies, music, life-science data, books, and scientific publications. With the constant growth of scientific publications, bibliographic data sources become widespread. Linked datasets such as CiteSeer, ACM or DBLP are often consulted to find publications in a given domain or identify people working in an area of interest. KAKEN is database of Grants-in-Aid for Scientific Research contains the "Project Selected" documents and the research report summaries. The Grants-in-Aid for Scientific Research is granted whole field of science, and this database provides the latest information of the research projects in Japan exhaustively. KAKEN RDF allows asking sophisticated queries against datasets derived from KAKEN to other datasets on the Web. In this paper we will deal with the problem of linking researchers in KAKEN and authors in DBLP.

Our challenge is the lack of associated properties of the entities in two data sources that should be link. We carry out analysis and evaluation of approaches which related to the data linking problem and propose using SILK framework [] to discover links between KAKEN researchers and DBLP authors based on their names. But the entity names are often ambiguous. For example, the name “Hiroshi Suzuki” refers to 27

researchers in KAKEN. Also, the name “Hiroshi Suzuki” can be written “H. Suzuki” in DBLP. We list two types of ambiguity: (1) different researchers share the same name and (2) one author has different name aliases. We propose topic-based similarity measure and co-author network to improve the reliability of links. The main idea behind our solution is that if a researcher and an author are the same person, his papers and projects must be related to the same topic and they have some social relationships with the same person. We collect paper titles as a topic feature for an author and project titles as a topic feature for a researcher. We calculate the topic-based similarity of two features by LDA based topic model. Also, the researchers have co-member relationships with other member in the same project and the authors have co-author relationships with other author in the same paper. Using both co-member and co-author relationships we define a relationship-based network and use Connected Triple similarity to determine reliability of a link. The main results of this paper are (1) constructing a topic based similarity measure based on LDA model, (2) constructing a relationship based similarity measure based on co-author network approach, (3) building a system combining two measures to determine the reliability of a link, (4) accuracy of linking increase from 41,02% to 86,04%.

This paper is structured as follows: Section 2 given an overview of data sources KAKEN and DBLP. Section 3 describes the system architecture with two major modules: LDA based similarity measure and Connected Triple similarity measure. Section 4 reports the results of implementation and review related work in Section 5.

2 Related work

LinkedMDB [3] provides a demonstration of connecting several major existing movie web resources. Because the data sources are about movie, LinkedMDB chooses movie titles as feature to discover *owl:sameAs* links. The problem is matching only the titles may not be sufficient due to different representations of the same title. They use proper string similarity function and specific record matching techniques to achieve high accuracy. However, it is not easy to apply this idea to our problem because person name is more ambiguous than film title. SILK [4] use a declarative language for specifying which types of RDF links between data sources should be discovered as well as which conditions entities must fulfill in order to be linked. Depending on which data sources are linked, SILK has different thresholds (“accept” and “verify”) for identifying similarity heuristics and qualifying the amounts of discovered links. This approach, however, only focuses on links of pairs of data sources: there is no guarantee that the information extracted from two data sources will be enough to find suitable entities in remains data sources. [5] present an algorithm to detect hidden *owl:sameAs* links or hidden relations in data sets. The main idea behind this solution is to extract useful features by applying supervised learning on frequent graphs. Then, using these extracted features to discover entities in data sources. This approach is not appropriate for our problem because it needs the existing links between data sources to discover the other hidden links.

3 Data sources

KAKEN is the database which is established and provided by the National Institute of Informatics (NII) with support of the Ministry of Education, Culture, Sports, Science and Technology (MEXT) and the Japan Society for the Promotion of Science (JSPS). It is a part of "GeNii, the National Institute of Informatics academic content portal", established by NII and provides the latest information of the research projects in Japan. The database has more than 180,000 researchers and 2.7 millions projects in November 2010. DBLP (Digital Bibliography & Library Project) is a computer science bibliography website hosted at Universität Trier, in Germany. It was originally a database and logic programming bibliography site, and has existed at least since the 1980s. DBLP listed more than 1.3 million articles on computer science in January 2010. Journals tracked on this site include VLDB, a journal for very large databases, the IEEE Transactions and the ACM Transactions. Conference proceedings papers are also tracked. It is mirrored at five sites across the Internet. DBLP (L3S) is an effort to extract structured information from DBLP and to make this information available on the Web. DBLP (L3S) allows you to ask sophisticated queries against DBLP, and to link other data sets on the Web to DBLP data. This is a database published with D2R Server. It can be accessed using: (1) your plain old web browser, (2) Semantic Web browsers, (3) SPARQL clients.

4 System architecture

The KAKEN and DBLP linking system have four main components. SILK framework is use to discover link between two data sources based on personal name. Then combining LDA based similarity and Connected Triple similarity to compute the similarity between two entities. Final, deciding a valid link if the similarity score above a threshold θ.

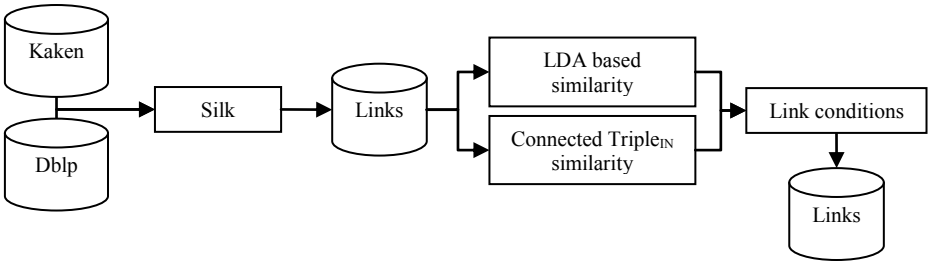

Fig. 1. KAKEN and DBLP linking system architecture

4.1 LDA based similarity

The main idea is that each author is associated with a multinomial distribution over topics and each topic is associated with a multinomial distribution over words. An author with multiple documents is modeled as a distribution over topics. This is a generative model for document collections, the topic model, that simultaneously models the content of documents and the interests of authors. This generative model represents each document with a mixture of topics, as in state-of-the-art approaches like Latent Dirichlet Allocation (Blei et al., 2003), and extends these approaches to author modeling by allowing the mixture weights for different topics to be determined by the authors of the document. By learning the parameters of the model, we obtain the set of topics that appear in a corpus and their relevance to different documents, as well as identifying which topics are used by which authors. The algorithm has three steps:

Step 1: finding hidden topic by using Latent Dirichlet Allocation, training data is list of document, each document is list of paper title in a conference.

Step 2: extract the entity topic feature: an author is featured as the list of paper titles; a researcher is featured as the list of project titles. Appling LDA topic model for these features we have vectors that each dimension corresponds to probability that an author or researcher is related to a hidden topic.

Step 3: using cosine similarity to compare similarity between two vector, from which determine that the researcher and the author is the same person

4.1 Connected Triple$_{IN}$ similarity

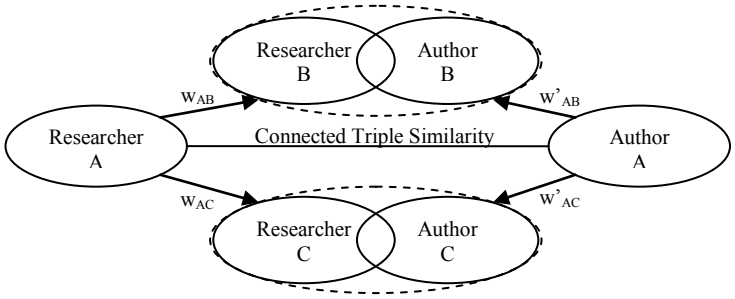

Fig. 2. Combining co-member and co-author network

Assume that all researcher and author that have the same name is a same person. The researchers that are member of the same project will have co-member relationships. The author will have the co-author relationships with other author that have the same paper. Using both co-member and co-author relationships we have a network G in which authors/researchers are represented as vertices V , and relationships builds the edges E. Then for each link researcher and author, we calculate the Connected Triples Similarity of two entities. A Connected Triple $\wedge = \{V_\wedge, E_\wedge\}$ can be described as a sub graph of G consisting of three vertices with $V_\wedge = \{A_1, A_2, A_3\} \subset V$ and $E_\wedge = \{e_{A1,A2}, e_{A1,A3}\} \in E$, $\{e_{A2,A3}\} \notin E$. The edges in the co-author network will be weighted

according to Liu et al. [3]. With V = $\{v_1,...,v_n\}$ as the set of n authors, m the amount of publications A = $\{a_1,...,a_k,...a_m\}$ and f (a_k) the amount of authors of publications a_k the weight between two authors v_i and v_j for publications a_k is calculated by:

$$g(i,j,k) = \frac{1}{f(a_k)-1} \tag{1}$$

There by the weight between two authors for one publication is smaller the more authors collaborated on this publication. Considering the amount of publications two authors i and j collaborated on together, an edge between these authors is calculated with (2) which leads to higher weights the more publications the two authors share.

$$C_{ij} = \sum_{k=1}^{m} g(i,j,k) \tag{2}$$

Applying a normalization the weight between two authors i and j considering the amount of co-authors and publications is calculated by (3) leading to a directed co-author graph.

$$w_{ij} = \frac{C_{ij}}{\sum_{r=1}^{n} C_{ir}} \tag{3}$$

The similarity of two authors using Connected Triples can consequently be either calculated on incoming edges or outgoing edges:

$$ConnectedTriple_{in} = \sum_{\forall c \in V \text{ with } e_{ci}, e_{cj} \in E, e_{ij} \notin E} w_{ci} + w_{cj} \tag{4}$$

$$ConnectedTriple_{out} = \sum_{\forall c \in V \text{ with } e_{ic}, e_{jc} \in E, e_{ij} \notin E} w_{ic} + w_{jc} \tag{5}$$

4 Experimental result

4.1 LDA training data

The LDA based topic model data training is a list of documents, each document is a list of paper titles in a conference that extract from DBLP. In this problem, we use 10.245 conferences as the training data. The estimation of number of hidden topic k is very important. If k is too large each topic feature will be narrow, if k is too small each topic feature will be wide. We experiment the parameters k with 50 topics, 100 topics and 150 topics. The testing is performed on the data set includes 724 links that have been labeled by manual. In our experiment, we randomly divided test data into three sets to determine whether the parameters are consistent with local data. We set the threshold θ for the measure, if the similarity score is greater than the threshold we will conclude that the two entities will be linked. We need to inspect different thresholds to find the optimal threshold.

Fig. 3 show the accuracy obtained with different value of number of hidden topic and threshold. Note that accuracy reported is percentage of both true positive and negative. Base on the result we chose number of hidden topic k=100 topics and threshold θ=0.7. Fig.4 shows that the threshold is consistent with local data.

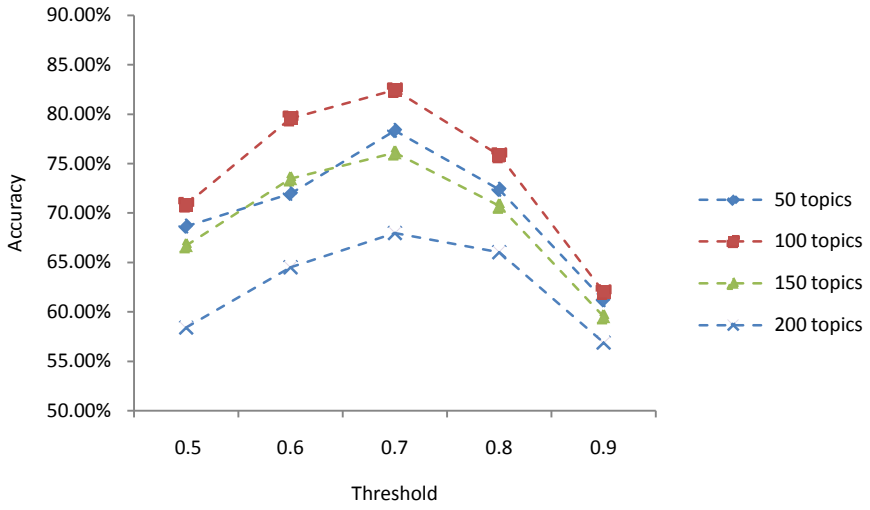

Fig. 3. LDA-based similarity accuracy

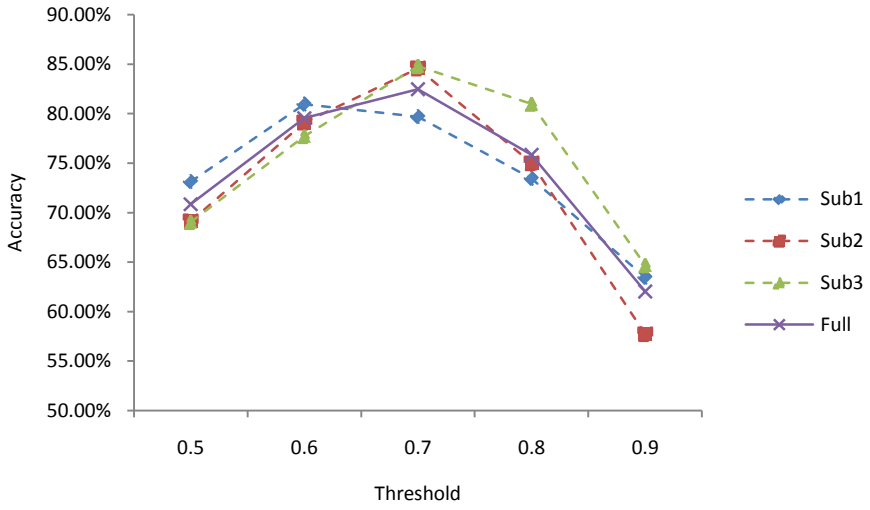

Fig. 4. LDA-based similarity accuracy in sub datasets

Fig. 5 compares the accuracies of LDA topic model and pLSA topic model also TFIDF weight and shows that LDA gave a better result than the others.

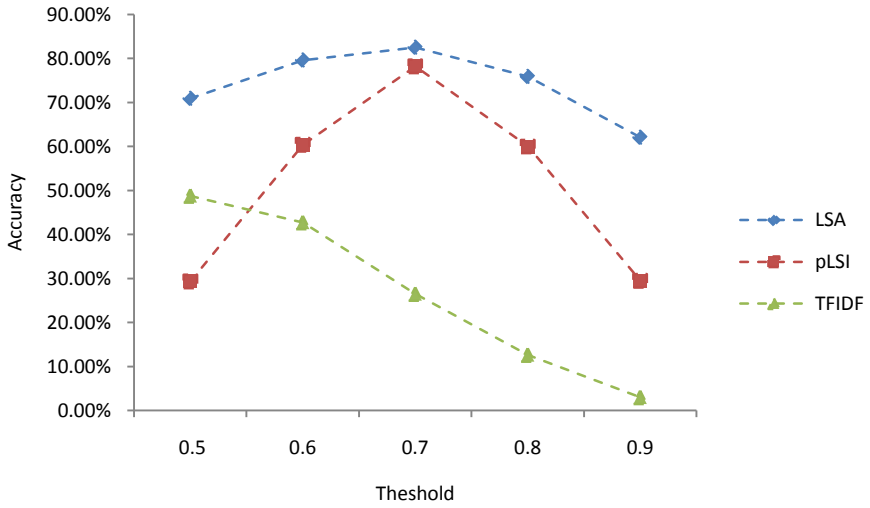

Fig. 5. Accuracies of LDA and pLSA, TFIDF

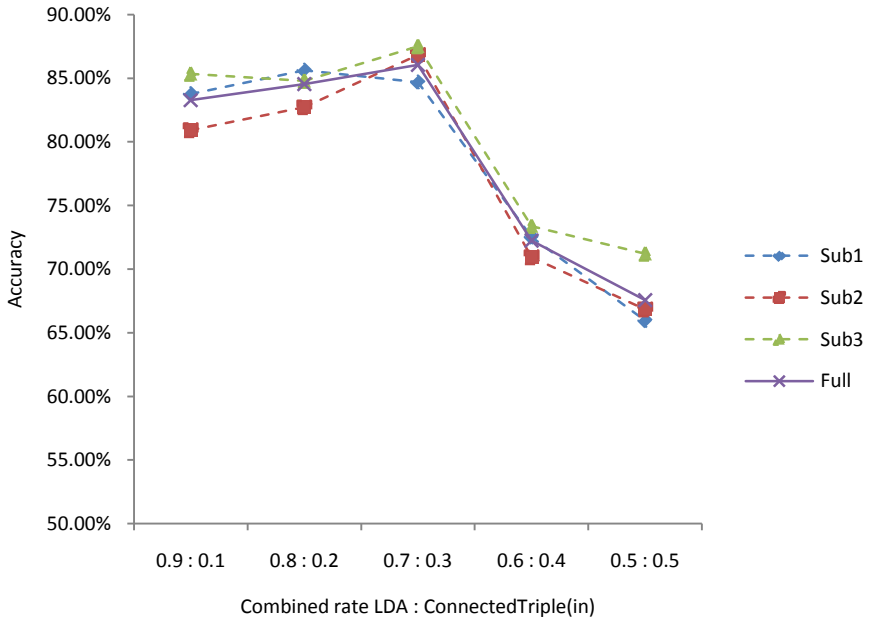

Fig. 6. Combining LDA topic similarity and ConnectedTriple similarity

Fig. 6 show the accuracy obtained with different combined rate between LDA topic similarity and ConnectedTriple similarity in different sub datasets. We find the optimal rate is 0.7:0.3. Base on these results we chose the threshold θ=0.7 and the rate

of combining two similarity measures is 0.7:0.3. Fig. 7 shows that the combining of LDA base similarity and ConnectTriple similarity gave a better result.

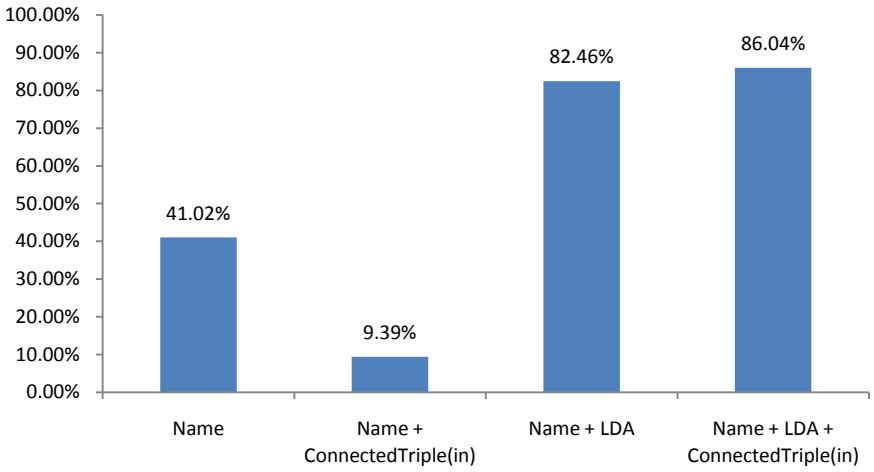

Fig. 7. Accuracy of combining similarity measures

5 Conclusions

We presented an approach to solve the name ambiguous problem in KAKEN and DBLP linking. Our solution is combining LDA based topic model and co-author network approach to improve accuracy of the links. We compared the LDA-based topic model with other models including pLSA and TFIDF weight. The paper has contributed a small part in solving the bibliographic data linking and applying for a specific problem. The results increase the accuracy from 41.02% to 86.04%.

References

1. Bizer, C., Heath, T., Idehen, K., Berners-Lee, T.: Linked data on the web. In Proceeding of the 17th international conference on World Wide Web WWW 08 (2008)
2. Bizer, C., Heath, T., Ayers, D., Raimond, Y.: Interlinking Open Data on the Web. In Demonstrations Track, 4th European Semantic Web Conference, Innsbruck, Austria (2007)
3. Hassanzaded, O., Consens, M.: Linked movie data base. In Proceedings of the WWW 09 Workshop on Linked Data on the Web, Madrid, Spain (2009)
4. Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Silk - a link discovery framework for the web of data. In Proceedings of WWW 09 Workshop on Linked Data on the Web, Madrid, Spain (2009)
5. Le, N. T., Ichise, R., Le, H. B.: Detecting Hidden Relations in Geographic Data. In Proceedings of the 4th International Conference on Advances in Semantic Processing, Florence, Italy (2010)

6. Biryukov, M.: Co-Author Network Analysis in DBLP: Classifying Personal Names. In 2nd International Conference on Modeling, Computation and Optimization in Information Systems and Management Sciences. Metz, France (2008)
7. Rosen-Zvi, M., Griffit, T., Steyvers, M., Smyth P.: The Author-Topic Model for Authors and Documents. In 20th Conference on Uncertainty in Artificial Intelligence, Banff, Canada, (2004)
8. Reuther, P., Walter, B., Ley, M., Weber, A., Klink, S.: Managing the Quality of Person Names in DBLP. In European Conference on Digital Libraries, pp. 508–511 (2006)
9. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet Allocation. In Journal of Machine Learning Research (JMLR) 3:993-1022 (2003)