Bare Demo of IEEEtran.cls for IEEE Computer Society Conferences

Changzhao Zhu
Shanghai JiaoTong University
Email: http://zczhhhh@sjtu.edu.cn

***Abstract*—For helping users get a better understand of a filed, I aimed at building a system which provides the academic filed profile. The system focuses on: 1) Extracting researcher profiles automatically from the Web; 2) Domain expert discovery and recommendation. 3)Domain core paper discovery and recommendation. So far, 2129 researcher profiles and 138 domain have been extracted and detected. Furthermore, I propose a model for commendation of core paper. In this paper, I describe the architecture and technology used in this system**

1. Introduction

Academic literature records the development and progress of science, recently academic information, including papers, authors and conference and the relationship between these entries, has played a more and more role in research and business, An effective understanding of the scientific field can not only effectively improve the quality and sharing of papers, but also effectively help researchers to conduct academic exchanges and shorten the industrialization cycle of scientific research results. Previous systems are aimed at papers recommendation and extracting users profile. However, these functions only provide minimal help for understanding the filed. A user who knows little a bout a new filed may want to learn about: 1) Who are the expert of this filed. 2)What are the core technology in the filed.3) How did this filed development. To address these problems, I extract scholars information from the Web, identify domain and discovery the expert by computing the probability distributions on different implicit topics with LDA model, recommend the core paper based on the reference network.

2. Related Work

2.1. Personal Profile Extraction

Several research efforts have been made for extracting person profiles. For example, Yu et al.[1] propose a two-stage extraction method for identifying personal information from resumes. The first stage segments a resume into different types of blocks and the second stage extracts the detailed information such as Address and Email from the identified blocks. A few efforts also have been placed on the extraction of contact information from emails or from the Web. For example, Kristjansson et al. [2] have developed an interactive information extraction system to assist the user to populate a contact database from emails. In comparison, profile extraction consists of contact information extraction as well as other different subtasks.

2.2. Topic Model

Considerable work has been conducted for investigating topic models or latent semantic structures for text mining. For example, Hofmann [3] proposes the probabilistic latent semantic indexing (pLSI) and applies it to information retrieval (IR). Blei et al. [4] introduce a three-level Bayesian network, called Latent Dirichlet Allocation (LDA). Some other work has been conducted for modeling both author interests and document contents together. For example, the Author model [5] is aimed at modeling the author interests with a one-toone correspondence between topics and authors. The Author-Topic model [6] [7] integrates the authorship into the topic model and can find a topic mixture over documents and authors.

3. Researcher's profile extracting

3.1. Data processing

The dataset is download from dblp(http://dblp.uni-trier.de/xml/i), DBLP is an integrated database system of computer English literature with author as the core of the research results in the computer field. The dataset includes papers, international journals and conferences. The original dataset has 14153306 authors, which is so large that beyond my capacity, finally, I choose the authors whos citation is over 300, and we get 2129authors and there work.

3.2. Process

Profile extraction is the process of extracting the value of property including position, email research interests in a person profile. Fortunately, The dblp dataset provide the research institution for every author in every article, so the process of extracting researchers profile is: 1) get a list of web pages by a search engine(the key words is the name and institution of a researcher. I use Google API) 2) Identify

the homepage/introducing page using a binary classifier. I use Support Vector Machines. The table shows the feature, whether the page contains the researchers work and so on. 3)I use use some special word such as Professor Ph.D Office Fax, email to extract positions, contact information of the researcher. 4) take the element in the page with ¡/image¿ token, and use face++ API to detect face. save the picture with face. If there is another same name with different institution. I will use face++ API to compare the two face picture to see if they are a same person.

In URL	Whether it contain edu Whether it contain a substring of name Whether it contain cs
In the page	Whether it contain the name Whether it contain the Institution Whether it contain the researchers work

4. Experts found

Experts found contain the domain detection, we may extract information about researchers interests in the researcher profile extracting step, but its not enough. Traditionally, information is usually represented based on the bag of words assumption. Jie Tang et.al[9] has proposed a unified topic model for simultaneously modeling the topical distribution of papers, authors, and conferences. My system is based on LDA modle, when modeling for each paper, according to the probability distribution of the current paper topic, to generate an implied topic, then according to the subject of the probability distribution of each entity, generate the paper associated with each word, the author, and entity of the meeting. Finally I get 138 domain tokens.

5. Core paper found

Previous paper recommendation systems only recommend those paper including key words with high citation. These paper are undoubtedly of high value, some papers are even the founder of a subfield. However, this is not a good solution to cross C domain problems. For example, if I want to learn about the recommendation filed with key word recommendation , it would be better if the system can return some core paper for data mining. The dblp dataset dont contain the reference relationship of articles, so I need to extract the information from the Web, due to the limit of resources and time. I extract reference relationships for 100000 articles in the dataset. The process is: 1) Find the paper with high citation in the domain.2) Search for it references and find the domains of these references.3) Query the filed dependency. 4) Recommend the reference which has high field dependency. The calculation of the dependency is:

$$d(f_1, f_2) = \frac{\sum N(if nodes N in f_1 or f_2)}{\sum e_N(if edge N has node in f_1, f_2)}$$

Acknowledgments

The authors would like to thank...

References

[1] K. Yu, G. Guan, and M. Zhou. Resume information extraction with cascaded hybrid model. In Proc. of ACL05, pages 499C506, 2005

[2] T. Kristjansson, A. Culotta, P. Viola, and A. McCallum. Interactive information extraction with constrained conditional random fields. In Proc. of AAAI04, 2004.

[3] K. Yu, G. Guan, and M. Zhou. Resume information extraction with cascaded hybrid model. In Proc. of ACL05, pages 499C506, 2005

[4] T. Hofmann. Probabilistic latent semantic indexing. In Proc.of SIGIR99, pages 50C57, 1999.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. Journal of Machine Learning Research, 3:993C1022, 2003.

[6] A. McCallum. Multi-label text classification with a mixture model trained by em. In Proc. of AAAI99 Workshop, 1999.

[7] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In Proc. of UAI04, 2004

[8] M. Steyvers, P. Smyth, and T. Griffiths. Probabilistic author-topic models for information discovery. In Proc. of SIGKDD04, 2004

[9] Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). ArnetMiner:extraction and mining of academic social networks. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp.990-998). DBLP.