

# Research Statement

JIANG Jing

School of Information Systems, Singapore Management University

Tel: (65) 6828-0785; Email: jingjiang@smu.edu.sg

07 February 2012

## Background

Written languages still remain as arguably the most common means of communication for people. With the explosion of user-generated Web content in recent years, textual data is becoming even more abundant. Social media such as blogs, microblogs, forums, wikis and online social networks are adopted by a growing population, and one of the most important types of data from these social media is text. There is therefore an urgent need to study how to make sense out of such large volumes of diverse and noisy text, or more specifically, how to find, extract, understand and summarize the information contained in the text. For example, blog and microblog readers may want to quickly identify recent trendy topics that are attracting attention in the blogosphere. Given an unfamiliar trendy topic, a reader may want to read a brief summary of the topic and see its related topics such as related people and organizations. Companies often look for consumer feedback of their products from online reviews. Governments may also want to find out the general public's opinions towards their policies expressed in online social media.

All these information needs cannot be simply satisfied through traditional search, which addresses the problem of finding relevant items given a set of search terms. To handle more complex information needs such as the ones mentioned above, we need text mining, extraction and summarization techniques. My central concern as a researcher is to develop effective and robust techniques to achieve a deeper understanding of text and to apply these techniques to critical, real-world information management problems.

## Research Areas

### Adaptive relation extraction

Relation extraction is the task of finding the relations between persons, organizations and other entities from natural language text. It is an important task in text mining and has many applications in structured search, text summarization, question answering, etc. Previous work on relation extraction often relies on supervised learning methods and manually annotated training data, which cannot easily scale up when we deal with large and diversified real text collections such as those from the Web. A recent line of my research is to explore adaptive approaches to relation extraction. I have previously systematically studied the feature space for relation extraction [9]. I further

applied the adaptation techniques I proposed in [10] to the problem of weakly supervised relation extraction where only a handful of labeled examples are used for training [8]. I found that the automatic adaptation method coupled with some minimal human guidance could substantially improve relation extraction performance. More recently, together with my collaborators, I studied how to extract specific relation descriptors such as “CEO” and “director” under a general relation type such as “employment” [4]. This work allows us to create a relatively small training data set to handle a large, diverse set of possibly unseen relation descriptors.

### Unsupervised information extraction

Information extraction aims to convert unstructured free text into structured information. While traditionally information extraction is done in a supervised manner, ultimately, we would like to perform unsupervised information extraction, i.e. information extraction without human annotations for training.

One example is how to define a semi-structured template for entity summaries. Imagine a Wikipedia user who wants to add a new article about a newly-founded start-up company. If there is a template that lists the major facts such as founder and headquarters one should include in a company summary, it becomes much easier to edit such an article. In [6], my student and I proposed an extension to the widely-used Latent Dirichlet Allocation (LDA) model for constructing semi-structured entity summary templates. Our model does not require any human supervision. It exploits the latent structures inherently contained in existing Wikipedia articles. With this model, we are able to construct summary templates that are organized into aspects and contain sentence patterns with empty slots to be filled in.

In [7], my student and I developed a model that could separate words that describe aspects and words that bear opinions. This model is very useful for automatic product review summarization, especially to discover more informative descriptions. For example, after discovering the “ambience” aspect in restaurant reviews, this model can further discover descriptions such as “cozy” and “romantic” for this aspect as opposed to simply “good” or “bad.”

My current focus is to develop more effective and robust models for unsupervised information extraction template induction. A preliminary work on template slot discovery has shown some promising early results [3].

### Text mining on social media

With the growth of the social Web, many textual information management problems have also moved to the social media. Through working with my students and collaborators, recently I started exploring interesting text mining problems on social media, in particular, microblogging sites such as Twitter.

Twitter is currently the most popular microblogging site, where people broadcast short messages on personal experiences, breaking news, opinions, etc.

In [5] we studied the problem of identifying influential users on Twitter. We adopted a PageRank-like algorithm and found that our extended algorithm could outperform Twitter's current method and the standard PageRank method. In [1] we used topic modeling to conduct a novel systematic comparison of the topic distributions and other characteristics of Twitter and New York Times. The differences we discovered suggest that Twitter search should have a different focus than traditional Web search, e.g. Twitter could be a good information source for topics such as celebrities. Since searching Twitter is an under-explored territory, our work sheds light on future research in this area. In [2], we studied how to identify key phrases that indicate trendy topics from Twitter. While key phrase extraction has been well studied for traditional text collections such as news articles, the special properties of Twitter pose many new challenges.

Currently, there are a few directions my students and I are focusing on. One direction is summarizing online opinions towards socio-political issues, which has not been well studied compared with product review mining and summarization. We are also studying how to extend topic modeling methods to handle large amounts of text streams such as Twitter feeds, and in particular, how to detect bursty topics from them.

### **Selected Publications and Outputs**

1. Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan and Xiaoming Li, "Comparing Twitter and Traditional Media using Topic Models." In Proceedings of the 33rd European Conference on Information Retrieval, pages 338-349, 2011.
2. Xin Zhao, Jing Jiang, Jing He, Yang Song, Palakorn Achanauarp, Ee-Peng Lim and Xiaoming Li, "Topical Keyphrase Extraction from Twitter." In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 379-388, 2011.
3. Cane Wing-ki Leung, Jing Jiang, Kian Ming A. Chai, Hai Leong Chieu and Loo-Nin Teow, "Unsupervised Information Extraction with Distributional Prior Knowledge." In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 814-824, 2011.
4. Yaliang Li, Jing Jiang, Hai Leong Chieu and Kian Ming A. Chai, "Extracting Relation Descriptors with Conditional Random Fields." In Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 392-400, 2011.
5. Jianshu Weng, Ee-Peng Lim, Jing Jiang and Qi He, "TwitterRank: Finding Topic-Sensitive Influential Twitterers." In Proceedings of the Third ACM International Conference on Web Search and Data Mining, pages 261-270, 2010.

6. Peng Li, Jing Jiang and Yinglin Wang, "Generating Templates of Entity Summaries with an Entity-Aspect Model and Pattern Mining." In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 640-649, 2010.
7. Xin Zhao, Jing Jiang, Hongfei Yan and Xiaoming Li, "Jointly Modeling Aspects and Opinions with a MaxEnt-LDA Hybrid." In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 56-65, 2010.
8. Jing Jiang, "Multi-Task Transfer Learning for Weakly-Supervised Relation Extraction." In Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, pages 1012-1020, 2009.
9. Jing Jiang and ChengXiang Zhai, "A Systematic Exploration of the Feature Space for Relation Extraction." In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, pages 113-120, 2007.
10. Jing Jiang and ChengXiang Zhai, "A Two-Stage Approach to Domain Adaptation for Statistical Classifiers." In Proceedings of the 16th ACM Conference on Information and Knowledge Management, pages 401-410, 2007.