Representing a Web Page as Sets of Named Entities of Multiple Types – A Model and Some Preliminary Applications

Nan Di, Conglei Yao, Mengcheng Duan
Dept of Computer Science and Technology
Peking University
Beijing, 100871, P.R. China
{dinan, ycl, dmc}@net.pku.edu.cn

Jonathan J. H. Zhu
Dept of Media and Communication
City University of Hong Kong
Kowloon, Hong Kong
j.zhu@cityu.edu.hk

Xiaoming Li
State Key Laboratory of Advanced Optical Communication Systems & Networks
Peking University
Beijing, 100871, P.R. China
lxm@pku.edu.cn

ABSTRACT
As opposed to representing a document as a "bag of words" in most information retrieval applications, we propose a model of representing a web page as sets of named entities of multiple types. Specifically, four types of named entities are extracted, namely person, geographic location, organization, and time. Moreover, the relations among these entities are also extracted, weighted, classified and marked by labels. On top of this model, some interesting applications are demonstrated. In particular, we introduce a notion of person-activity, which contains four different elements: person, location, time and activity. With this notion and based on a reasonably large set of web pages, we are able to show how one person's activities can be attributed by time and location, which gives a good idea of the mobility of the person under question.

Categories and Subject Descriptors
H.1.1 [**Models and Principles**]: Systems and Information Theory

General Terms
Algorithms, Measurement

Keywords
Web Content Mining, Web Page Model, Named Entity

1. INTRODUCTION
With the explosive growth of the Web, it has become increasingly necessary for users to utilize automated applications in finding the desired information from a large number of web pages. Web content mining can be broadly defined as the discovery and analysis of desirable information from the page content of the web. The prevailing representation of the web page content is the bag-of-words model, whereas we think that named entities in the content are more important and informative for in-depth mining. We propose a novel named-entity-based model for page content, which consists of multiple types of named entities along with relations between them. Specifically, we utilize four types of named entities, person, location, organization and time, with the relations between them to represent the web page. Furthermore, we weight, classify and label the relations. With this model, we can construct a series of interesting applications, such as a system tracking the entity's activities by time line and geographical space. In the following of this article, we will illustrate the details of this model, along with the related key techniques and two interesting applications.

WWW 2008, April 21--25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

2. AN OVERVIEW OF THE MODEL AND A PRELIMINARY APPLICATION

2.1 Named Entity Discovery
As for the complexity and diversity of the Web, traditional named entity (NE) discovery methods, such as rule-based method and model-based method, do not always work due to the lack of scalability. In our research, we combine these two types of methods to make the extraction more scalable. Taking person entity as an example, we combine a rule-based method, a model-based method, and a Bayesian method to make the NE extraction more scalable. For the rule-based method, we use three types of information as rules. For the model-based method, we utilize the HMM model. As for the Bayesian method, we assume the first name and the last name of a candidate person name are independent, and use naive Bayesian method to assess a confidence to the candidate person name. Our method not only fits on person entity but also other types of name entities. Readers can refer to [1] for more details.

2.2 Time Discovery
We consider two kinds of time that are important for a web page: one is the page-born time, which indicates when the page is available on the web; the other is page-content time, which indicates when the event described in this page happened. Although we can use the LMT (last modified time) to measure the former, we think it's not desirable for two reasons: first, a great number of pages don't have LMTs or their LMTs are randomly set; second, the time we are interested in is when the page can be accessed on the web rather than when it's born. Therefore, we choose to use the time when our crawler finds this page as the page-born time. Comparing to page-born time, page-content time is more valuable and more difficult to obtain. We try to uncover it from both the URL and the page text. A page's URL may contain date information. We choose a set of URLs manually from the pages crawled and learn the time patterns from these URLs. Then we can obtain the time information from URL by applying these patterns. We also observe that many news web pages contains the page-content times, which are often located just right after the title. Based on this observation, we divide one page into a sequence of text quarks by VIPS[2], and extract the candidate time from small quarks. We compare the time from both URL and text, and choose the earlier one as the page-content time.

2.3 Relation Discovery
After the extraction of the four types of entities, we utilize the unsupervised method to explore the relations among these entities. Given the redundancy of web pages, we employ entity cooccurence to locate the entity pairs which might own one relation. Previous relation

analysis researches[3] only use the context over the cooccurence as the representation of this pair. It works well in a small corpus, whereas in the case of web content mining its performance is not acceptable. Thus, we use search engine to extend the representation: we choose some key terms from the cooccurence context using term distance as the selection criterion, then construct a query using these terms and send it to the search engine. We use the retrieved pages' content as the representative text. To the relation classification, we use the label propagation algorithm[4] based on graph.

2.4 Entity-based Page Model

We then introduce the *entity-based page model*. This model represents each page using the persons, locations, organizations, relations and time which are extracted from the page: $\boldsymbol{P}=\{(p_{1,...,}p_i), (l_{1,...,}l_j), (o_{1,...,}o_k), (r_{1,...,}r_u), T\}$, where p_i is a person entity, l_j is a location entity, o_k is a organization entity, r_u is a relation pair and T is the page time. By combining all the page models, we can generate a *global entity model* of a set of web pages. It contains entity list from the page set, the entity relations based on cooccurence, the entity relations based on links between pages where the entities are found. This *global entity model* is denoted by $\boldsymbol{G}=\{EntityList, CoReList, LinkReList\}$, where *EntityList* is the entity list, *CoReList* is the cooccurence relation list, and *LinkReList* is the link relation list. Further analysis on this global model can answer some interesting questions such as: How many web pages are there that describe events happened at location *L*? What is the activity track of person *P* in the past year? Former researches can only get some specious answers from the bag-of-words model, whereas our new entity-based model pays more attention to the entities and relations, this effort can help us to dig more knowledge from the pages' content.

Figure 1. page-location mapping.

2.5 A Proof of Concept Application

We select the top 100 Chinese news sites in http://www.alexa.com as the seed sites and crawl each site down to four levels depth. This crawling process continues for 30 days, from October 1, 2007. We obtain 700K pages each day, and a total of 20M pages are crawled. After the crawling, we extract the persons, locations and organizations from them. The page-born time of each page is set as the crawled time, whereas the page-content time is extracted as described previously. To extract the entity relations among entities, we choose the top 1000 persons, from the totally 885108 persons, sorted by document frequency as the candidate entities. And 72562 relation pairs are found among these persons. The relation's weight is calculated and the relation pairs are classified into four categories: politics, culture, sports and entertainment. Based on the former process, our system provides two basic services now: page-location mapping and person friend searching. In page-location mapping we use Google Map API to show the number of pages that contains a particular location. In person friend searching, user can explore the relation set extracted from the top 1000 persons. The weight and type of each relation pair are available. We also show the change of weight during a period of time which usually implies the occurrence of some events at the inflexion.

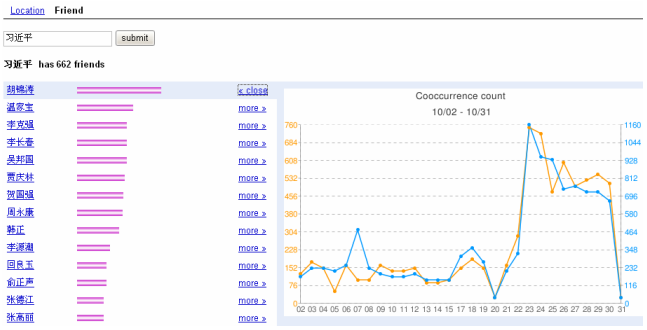

Figure 2. person friend search.

3. CONCLUSION AND FUTURE WORKS

In this paper, we briefly describe a new page model for web content mining. Instead of using all the terms in the text, our model pays more attention to various kinds of named entities in a page, such as person, location, organization and time. We develop a series of entity extraction methods to obtain these entities and their attributes. We explore the relations among these entities, calculate the weight, classify and label these relation pairs. By applying the methods to a set of pages crawled over 30 days, we implement a demonstration system which is able to show page-location mapping and be used for person friend search.

For future works, we have two research interests. One is to explore the evolution of entities and relations over time. A sudden change in the count of the entity appearance or in the relation weight may imply some important events on this entity or entity pair. Another area is to focus on person activity tracking which combines person, time, location and action together. And some advance issue such as conflicting relations, malicious information should also be considered.

Acknowledgements. This work is supported by NSFC (60773162) and 863 project 2006AA01Z196, HKSAR CERG (CityU 1456/06H) and City University of Hong Kong SRG (7001882).

4. REFERENCES

[1] Conglei Yao, Nan Di. Technique Report: Mining the whole set of person names from the Chinese Web. http://net.pku.edu.cn/~ycl/wdtr.pdf.

[2] Yu, S., Cai, D., Wen, J.-R. and Ma, W.-Y., Improving Pseudo-Relevance Feedback in Web Information retrieval Using Web Page Segmentation, In Proceedings of WWW' 03, pages 11-18.

[3] Takaaki Hasegawa, Satoshi Sekine, and Ralph Grishman. Discovering relations among named entities from large corpora. In Proceedings of ACL' 04, pages 415-422.

[4] Jinxiu Chen, Donghong Ji, Chew L. Tan, and Zhengyu Niu. Relation extraction using label propagation based semi-supervised learning. In Proceedings of ACL' 06, pages 129–136.