A Hybrid Method for Chinese Entity Relation Extraction

Hao Wang, Zhenyu Qi*, Hongwei Hao, and Bo Xu

Institute of Automation, Chinese Academy of Sciences, Beijing, China
{h.wang,zhenyu.qi,hongwei.hao,xubo}@ia.ac.cn

Abstract. Entity relation extraction is an important task for information extraction, which refers to extracting the relation between two entities from input text. Previous researches usually converted this problem to a sequence labeling problem and used statistical models such as conditional random field model to solve it. This kind of method needs a large, high-quality training dataset. So it has two main drawbacks: 1) for some target relations, it is not difficult to get training instances, but the quality is poor; 2) for some other relations, it is hardly to get enough training data automatically. In this paper, we propose a hybrid method to overcome the shortcomings. To solve the first drawback, we design an improved candidate sentences selecting method which can find out high-quality training instances, and then use them to train our extracting model. To solve the second drawback, we produce heuristic rules to extract entity relations. In the experiment, the candidate sentences selecting method improves the average F1 value by 78.53% and some detailed suggestions are given. And we submitted 364944 triples with the precision rate of 46.3% for the competition of Sougou Chinese entity relation extraction and rank the 4th place in the platform.

Keywords: Information Extraction, Entity Relation Extraction, Conditional Random Field Model, Knowledge Base.

1 Introduction

Entity relation extraction is one of the main tasks of information extraction. The input of this problem is multi-structured data, including structured data (infobox form), semi-structured data (tables and lists) and non-structured data (free text). And the output is a set of fact triples extracted from input data. For example, given the sentence "姚明出生于上海" (Yao Ming was born in Shanghai) as input, the relation extraction algorithm should extract "<姚明, 出生地, 上海>" (Yao Ming, birthplace, Shanghai) from it. These fact triples can be used to build a large, high-quality knowledge base, which can benefit to a wide range of NLP tasks, such as question answering, ontology learning and summarization.

Now massive Chinese information exists on the internet and the research of Chinese entity relation extraction will have important significance. But current research mainly focuses on the processing of English resource and the study conducted on

* Corresponding author.

C. Zong et al. (Eds.): NLPCC 2014, CCIS 496, pp. 357–367, 2014.

Chinese corpus is less. Compared to English language, Chinese language need word segmentation, and the proper nouns don't have the first letter capitalized, so the Chinese entity relation extraction is more difficult and more challenging.

In this paper, we propose a hybrid method for Chinese entity relation extraction. We adopt different methods to extract different frequency relation words. We first build a Chinese semantic knowledge base, using the corpus of Douban web pages, Baidu encyclopedia and Hudong encyclopedia. An improved selecting candidate sentences method trained by conditional random field model is used to extract high-frequency relation words of the knowledge base, and the method based on some simple rules and knowledge base is used to extract low-frequency relation words.

Specifically, our contributions are:

- We propose candidate sentences selecting method, which can reduce the mistakes introduced by automatic tagging training data and improve the extraction performance.
- It's hard to get enough training data for some rare relations. Here, we propose the method based on some simple rules and knowledge base to extract these low-frequency relation words.

The rest of the paper is organized as follows: Section 2 introduces related works of this paper. Section 3 introduces the construction of our Chinese semantic knowledge base, and the improved selecting candidate sentences method trained by conditional random field model for high-frequency relation words, and method based on knowledge base and simple rules for low-frequency relation words. Section 4 describes experimental results and detailed analysis of our methods. We conclude in Section 5.

2 Related Works

Since the late 1980 s, the MUC (Message Understanding Conference) [1] promoted the vigorous development of relation extraction, and made the information extraction to be an important branch in the field of natural language processing. In order to meet the increasing social demand, since 1999, the NIST (National Institute of Standards and Technology) organized ACE (Automatic Content Extraction) reviews [2], and automatically extracting entities, relations, and events in news corpus is the main content of this conference.

Washington University developed TextRunner System [3, 4], the representative of the free text entity relation extraction, and then they released the WOE System [5], which is using Wikipedia for open information extraction. The basic idea of these two systems is first to identify the entity, and then regards the verb between the two entities as relationship. It would have a lot of dislocation, some illogical extraction result, as well as the discontinuous extraction. Continuous analysis and improvement of their previous work, then this group published the second generation open information extraction systems, for example, REVERB [6], R2A2 [7], and OLLIE [8]. Compared to the first generation, the effect of these systems had obvious promotion. The basic idea is first to identify the relationship, and then to identify entities, and specific versions is the result of improvements on the details.

The Intel China research center developed a Chinese named entity extraction system [9], and the relation between these entities can also be extracted. This system obtains rules by memory-based learning algorithm, and then these rules are used to extract named entity and the relationship between them. There have been many previous works of extracting Chinese entity relation by training conditional random field model, and some work use the online encyclopedia corpus [10, 11].

3 Chinese Entity Relation Extraction

Using the structured data part of Baidu encyclopedia and Hudong encyclopedia, we can build a Chinese semantic knowledge base, and this part will be described in detail in section 3.1. Different methods of extracting entity relationship should adapt to the frequency of relation words in our knowledge base.

The improved candidate sentences selecting method trained by conditional random field model, introduced in section 3.2, can be used to extract high-frequency relations. For some low-frequency relations, the method based on simple rules and knowledge base can be used, and we will discuss this case specific to our method for the competition of Sougou Chinese entity relation extraction, and this part will be described in detail in section 3.3. If a relationship is low-frequency in our knowledge base, it is hard to get enough training corpus for the improved conditional random field model, so we need specific treatment for different cases and make full use of the domain knowledge.

3.1 Chinese Semantic Knowledge Base Construction

We gathered the Baidu encyclopedia web data before July 2012 and Hudong encyclopedia web data before October 2012, which are composed of entities, corresponding infobox knowledge and unstructured content. We extract the infobox knowledge from these corpus and represent them in triples format <arg1, rel, arg2> , arg1 and agr2 are represent of entities and rel is represent of relation, like<中国, 首都, 北京>, and then store these triples in our knowledge base.

中文名	卧虎藏龙	主　演	周润发，杨紫琼，章子怡，张震
外文名	Crouching Tiger, Hidden Dragon	片　长	120 min
制片地区	中国，美国	上映时间	法国：2000年5月16日
导　演	李安	分　级	USA:PG-13
编　剧	王度庐，王蕙玲	对白语言	汉语普通话
类　型	爱情，动作，冒险，剧情	色　彩	彩色
		奖　项	奥斯卡最佳外语奖

Fig. 1. An Infobox from Baidu Encyclopedia

Fig 1 shows the infobox knowledge of Baidu encyclopedia, and we can easily extract many triples from the XML files. And the extraction of structured data in Hudong encyclopedia is similar to this.

When we determine to extract a relationship, we should first traverse our knowledge base to get the frequency of this relation word. If the frequency number is greater than 500, the corresponding relation is high-frequency; otherwise we regard it as low-frequency relation. The following two sections will introduce different methods to extract these two kinds of relation.

3.2 Candidate Sentences Selecting Method

When the relation word is high-frequency in our knowledge base, we will traverse the knowledge base to get the corresponding <arg1, rel, arg2> triples. And these triples are used to locate candidate sentences in wiki-page of Baidu encyclopedia and Hudong encyclopedia corpus to extract this relation. The sentences chose as candidate sentences contain information for the relation word. For example, a sentence is "李安导演的《卧虎藏龙》诠释了中国武侠的魅力", and the relation word we concerned is "导演", and then this sentence will be selected as candidate sentence.

Here, four different approaches are proposed to estimate whether a sentence is a candidate sentence for a triple of the target relation word.

We explore two methods to score a sentence:

$$\text{score} = bArg\ 1 * (b\ \text{Re}\ l + 1) * bArg\ 2 \tag{1}$$

$$\text{score} = (bArg\ 1 + 1) * (b\ \text{Re}\ l + 2) * bArg\ 2 \tag{2}$$

We define three variables: bArg1, bRel, bAgr2. For a triple <arg1, rel, arg2> and a sentence, if the arg1 or its alias name appears in the sentence, bArg1 = 1, otherwise bArg1 = 0; if the rel or its alias name appears in the sentence, bRel = 1, otherwise bRel = 0; if the arg2 or its alias name appears in the sentence, bArg2 = 1, otherwise bArg2 = 0.

We choose the sentence of the highest score and the score must be greater than 0. Obviously, the sentence gaining by the first scoring method must have the arg1 and the arg2, and the sentence obtaining by the second method only must have the arg2.

Most of the time, there are many sentences with same highest score. Here, we propose two methods to get the final candidate sentences from these highest score sentences:

(a) Selecting the highest score sentence first appeared in an article;
(b) Selecting all highest score sentences.

After combination, we have four methods to obtain the candidate sentences, and then these sentences are used as training data.

Then we want to extract triples from the wiki-page content and the corresponding entity should not be in our semantic knowledge base. One simple idea is segmenting the article into sentences, and then extracts entity relation triples from all these sentences. Of course, there will be too many sentences need to extract, and we can reduce the candidate sentences.

Here, we introduce another method of getting less candidate sentences for extracting. Firstly, we will do word segmentation and part-of-speech tagging for the candidate sentences which are chose as training data, and then choose the nouns and verbs. Secondly, we do word frequency statistic for the selected word, and choose the top-n highest frequency words as key words. At last, these key words are used to determine the candidate sentences for extracting, and these candidate sentences are used as testing data.

For a triple <arg1, rel, arg2>, the rel is the relation word we concerned, and the arg1 is the entity of the wiki-page, and then we only need to get the arg2. So the Chinese entity relation extraction is converted to annotation problem. We train conditional random field model to label the arg2 in the testing data, and finally convert the annotation results to entity relation triples.

3.3 The Heuristic Rules Based Method

If a relation word is low-frequency in our knowledge base, we can't automatically get enough training data for statistical models. For this reason, we propose the heuristic rules based entity relation extraction algorithm, as shown in the following.

Algorithm1: The Heuristic Rules based Entity Relation Extraction Algorithm

Input: The target relations, some entities, corresponding categories and unstructured content
Output: Entity relation triples <arg1,rel,arg2>, arg1 and arg2 are entities and rel is the relation

1 Begin
2 Confirm the template <class1, rel, class2> of a target relation; here class1 and class2 are the categories of unknown entities. For example, a given relation "director", we can confirm class1 is movie or teleplay and class2 is people.
3 Produce an entity library, which contain entities and corresponding categories.
4 Get the keywords of target relation by using domain knowledge.
5 Select candidate sentences, which should contain keywords, and entities of class1 and class2 in our entity library.
6 Generate some simple rules to extract the entity relation.
7 End

Here we briefly explain the steps in this algorithm, combined with the competition of Sougou Chinese entity relation extraction. This competition involves 5 categories and a total of 17 relationships for extracting. And the movie category is similar with the teleplay category, and then we only need to discuss 4 categories and 12 relations. Some relation words are high-frequency and some are low-frequency, as shown in table 1. We adopt the method introduced in Section 3.2 to extract the high-frequency relation, and our heuristic rules based method is used to extract low-frequency relations.

Table 1. Categories and relation words of Sougou entity relation extraction competition

Frequency	Relation Words	Category
High-frequency	导演(director),演员(actor),编剧(writer),	Movie/ Teleplay
	作者(writer)	Book
Low-frequency	演唱者(singer), 作词(writer), 作曲(composer)	Song
	父母(parent), 兄弟姐妹(brother or sister), 夫妻(husband or wife)	People
	原著(original book), 原创音乐(Music soundtrack)	Movie/ Teleplay

The relation words which we concerned in people category are : “父母”, “兄弟姐妹” and “夫妻”. So we can confirm that class1 and class2 are both people category. We build an entity library containing all given people entities. The keywords for the relation word “父母” are : “父”, “爹”, “爸”, “子”, “女”, “母” and “妈”. The keywords for the relation word “夫妻” are : “结婚” , “完婚” , “闪婚” , “老公” , “老婆” , “妻子” , “丈夫” , “妻” , “夫” , “娶” and “嫁” . The keywords for the relation word “兄弟姐妹” are : “兄” , “哥” , “弟” , “姐” , “妹” and “姊”. Then we should select the sentences containing at least two people entities, and the sentences should have the keywords for different relation extraction.

When extracting the relation word “父母” , if the keywords are: “父” , “爹” , “爸” , “母” or “妈”, the entity before these keywords is the arg1 in triple <arg1, rel, arg2>, and the entity after these keywords is the arg2. But for the keywords of “子” and “女”, the entity before these keywords is the arg2, and the entity after these keywords is the arg1. When extracting the relation word of “兄弟姐妹” and “夫妻”, we don’t need to consider the order. We can get entity triples <arg1, rel, arg2> by using these simple rules.

We adopt the heuristic rules based entity relation extraction algorithm for the rest low-frequency relation words, and the detailed steps is similar to method described above.

4 Experimental Results and Analysis

We adopt the improved candidate sentences selecting method trained by conditional random field model to extract the high-frequency relation; section 4.1 will introduce the comparison and analysis of our experiment results. And section 4.2 will introduce our results for the competition of Sougou Chinese entity relation extraction.

4.1 The Comparison of Various Methods of Select Candidate Sentences

We select 5 categories and 2 relation words of every category, listed in Table 2.

Table 2. Five categories and two relation words of every category for extraction

Category	Relation Words
Geography	area, district
Movie	writer, director
Education	starting time, category
Book	publishing company, publication time
People	height, weight

In preparing for training data step, there are two methods to score a sentence based on the given triple <arg1, rel, arg2>:

(a) Score method 1: $score\ 1 = bArg1 * (b\,Rel + 1) * bArg2$

(b) Score method 2: $score\ 2 = (bArg1 + 1) * (b\,Rel + 2) * bArg2$

- If arg1 appears in this sentence, then bArg1 = 1, otherwise bArg1 = 0.
- If arg2 appears in this sentence, then bArg2 = 1, otherwise bArg2 = 0.
- If rel appears in this sentence, then bRel = 1, otherwise bRel = 0.

In preparing for training data step, two methods to get the final candidate sentences from these highest score sentences:

(a) Selecting the highest score sentence first appeared in an article;
(b) Selecting all highest score sentences.

In preparing the data for extracting step, two methods to extract triples from the wiki-page content (testing data):

(a) Choosing all the sentences in the wiki-page content;
(b) Selecting some sentences from the wiki-page content based on keyword matching.

Annotation:

- Label 1: score a sentence by score method 1, and then selecting all highest score sentences.
- Label 2: score a sentence by score method 1, and then selecting the highest score sentence first appeared in an article.
- Label 3: score a sentence by score method 2, and then selecting all highest score sentences.
- Label 4: score a sentence by score method 2, and then selecting the highest score sentence first appeared in an article.

(1) Choosing all the sentences in the wiki-page content, part of the experiment results:

Table 3. The extraction results of choosing all the sentences in the wiki-page content

Relation Words	Precision	Recall	F1	Relation Words	Precision	Recall	F1
Geo_area1	0.1570	0.1371	0.1463	Movie_Director1	0.1443	0.0833	0.1057
Geo_area2	0.1600	0.1421	0.1505	Movie_Director2	0.1633	0.0952	0.1203
Geo_area3	0.1486	0.1320	0.1398	Movie_Director3	0.1782	0.1071	0.1338
Geo_area4	0.1534	0.1371	0.1448	Movie_Director4	0.1458	0.0833	0.1061
Geo_district1	0.3118	0.2944	0.3209	EDU_Start_time1	0.3097	0.2909	0.3000
Geo_district2	0.3059	0.2640	0.2834	EDU_Start_time2	0.3117	0.2909	0.3009
Geo_district3	0.3333	0.3147	0.3238	EDU_Start_time3	0.3397	0.3212	0.3302
Ge_district4	0.3086	0.2741	0.2903	EDU_Start_time4	0.3333	0.3152	0.3240

(2) Selecting some sentences from the wiki-page content based on keyword matching, part of the experiment results:

Table 4. The extraction results of selecting some sentences from the wiki-page content based on keyword matching

Relation Words	Precision	Recall	F1	Relation Words	Precision	Recall	F1
Geo_area1	0.3119	0.1726	0.2222	Movie_Director1	0.4833	0.1726	0.2544
Geo_area2	0.3736	0.1726	0.2361	Movie_Director2	0.5439	0.1848	0.2756
Geo_area3	0.2661	0.1675	0.2056	Movie_Director3	0.4110	0.1786	0.2490
Geo_area4	0.2623	0.1624	0.2006	Movie_Director4	0.5469	0.2083	0.3017
Geo_district1	0.4294	0.3706	0.3978	EDU_Start_time1	0.6993	0.6061	0.6494
Geo_district2	0.4000	0.2538	0.3106	EDU_Start_time2	0.7252	0.5758	0.6419
Geo_district3	0.4535	0.3959	0.4228	EDU_Start_time3	0.7329	0.6485	0.6881
Ge_district4	0.4031	0.2640	0.3190	EDU_Start_time4	0.7211	0.6424	0.6795

Table 3 and Table 4 show the effect of different options to the extraction results, and after comparison and analysis we can get the follow conclusion:

1, the real results should be better than shown above, because we use the number of web pages instead of the number of triples that need be extracted, and in fact some web pages don't have triples, so the number of triples that need be extracted is less than the number of web pages, and the actual recall rate will be higher.

2, the extraction results of different relation words vary a lot, some results are very good, but some are not. We can easy find this

3, in the annotation, four different labels represent four different methods to get candidate sentences to train the conditional random field model. After comparing the results, we can find that the method of label 2 can get the highest precision and the

method of label 3 can get the highest recall, and it is hard to conclude which method can get the highest F1 value. The method of label 2 can get accurate and related training data, so this method can achieve the highest precision. The method of label 3 can get abundant training data to achieve the highest recall.

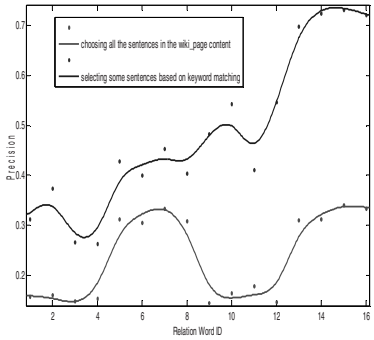

Fig. 2. The precision of different candidate sentences selecting methods

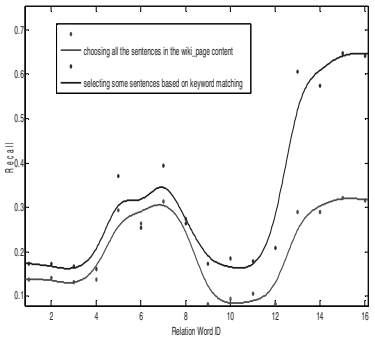

Fig. 3. The recall of different candidate sentences selecting methods

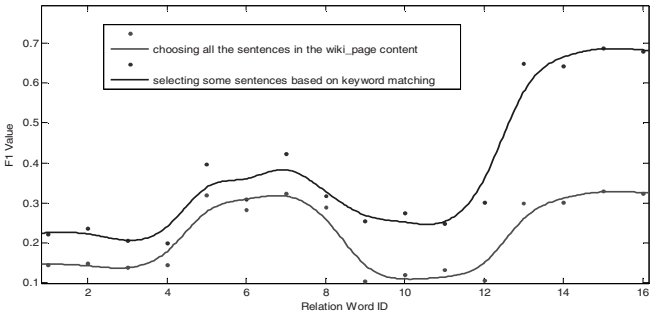

Fig. 4. The F1 Value of different candidate sentences selecting methods

When we apply the method of selecting some sentences based on keywords matching, the precision and recall have improved significantly (see Fig 2 and Fig 3). And the candidate sentences selecting method improve the average F1 value by 78.53%, and from Fig 4 we can find the F1 value increased obviously. So selecting candidate sentences is necessary, and our method has a good effect.

4.2 The Competition of Sougou Web-Based Entity Relation Extraction

The Sougou Company has their knowledge base of entity relation triples, and there is a test platform based on these triples for competitors to verify their results. But the results of test platform are not the final results, and the final results need sampling and

artificial validation to determine, and the organizing committee will give the final results later.

Table 5. Some example of Sougou Web-based Entity Relation Extraction Competition

Category	Relation	Sentence	Triples
人物	父母	冉甲男与父亲冉平一起担任编剧的电影《画皮2》备受期待。	〈冉甲男，父母，冉平〉
	夫妻	林妲怡与蒋家第四代蒋友柏结婚，婚后息影。	〈林妲怡，夫妻，蒋友柏〉
	兄弟姐妹	曾维信的奶奶胡菊花，是胡耀邦的亲姐姐。	〈胡菊花，兄弟姐妹，胡耀邦〉
书籍	作者	《沙床》当代高校生活的青春忏情录作者：葛红兵。	〈沙床，作者，葛红兵〉
歌曲	作词	《幻想爱》是陈伟作词作曲，张韶涵演唱的一首歌曲。	〈幻想爱，作词，陈伟〉
	作曲		〈幻想爱，作曲，陈伟〉
	演唱者		〈幻想爱，演唱者，张韶涵〉
电影/电视剧	导演	李安导演的《卧虎藏龙》诠释了中国武侠的魅力。	〈卧虎藏龙，导演，李安〉
	编剧	电影海上烟云由柯枫自编自导。	〈海上烟云，编剧，柯枫〉
	原著	根据琼瑶原著《含羞草》改编的台湾电视连续剧《含羞草》。	〈含羞草，原著，含羞草〉
	演员	电视剧《龙堂》由著名演员张丰毅、陈小春主演。	〈龙堂，演员，张丰毅〉〈龙堂，演员，陈小春〉
	原声音乐	电影《大兵金宝历险记》主题曲是刘佳演唱的美丽国。	〈大兵金宝历险记，原声音乐，美丽国〉

Table 5 shows some extraction results for the competition. This competition involves 5 categories and a total of 17 relationships for extracting. And the movie category is similar with the teleplay category, so we combine these two categories. Then we only need to discuss 4 categories and 12 relations.

Finally we submitted a total of 364944 triples. The precision is 49.09% and we rank the fourth place.

5 Conclusion and Future Work

In this paper, we first build a knowledge base using the collected corpus, and then adopt different methods to get the <arg1, rel, arg2> triples for different frequency relation words of our knowledge base. In the experiment, a detailed comparison and analysis on some options of selecting candidate sentences are introduced, and then we participate in the competition of Sougou Chinese entity relation extraction and rank the fourth place in the test platform.

However, in our experiment, only lexical features are used for triples extraction, and the parser-based features are ignored. An interesting direction is how to combine the parser-based features into the previous work, and in turn improving the performance of our extraction work.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (NSFC) Grants No.61203281 and No.61303172.

References

1. Message Understanding Conference, http://en.wikipedia.org/wiki/Message_Understanding_Conference
2. Automatic Content Extraction, http://www.itl.nist.gov/iad/mig/tests/ace/
3. Etzioni, O., Banko, M., Soderland, S., et al.: Open information extraction from the web. Communications of the ACM 51(12), 68–74 (2008)
4. Yates, A., Cafarella, M., Banko, M., et al.: TextRunner: open information extraction on the web. In: Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, pp. 25–26. Association for Computational Linguistics (2007)
5. Wu, F., Weld, D.S.: Open information extraction using Wikipedia. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 118–127. Association for Computational Linguistics (2010)
6. Fader, A., Soderland, S., Etzioni, O.: Identifying relations for open information extraction. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1535–1545. Association for Computational Linguistics (2011)
7. Etzioni, O., Fader, A., Christensen, J.: Open information extraction: The second generation. In: Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, pp. 3–10. AAAI Press (2011)
8. Schmitz, M., Bart, R., Soderland, S., et al.: Open language learning for information extraction. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 523–534. Association for Computational Linguistics (2012)
9. Zhang, Y., Zhou, J.: A trainable method for extracting Chinese entity names and their relations. In: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, vol. 12 (2000)
10. Zeng, Y., Wang, D., Zhang, T., Wang, H., Hao, H.: Linking Entities in Short Texts Based on a Chinese Semantic Knowledge Base. In: Zhou, G., Li, J., Zhao, D., Feng, Y. (eds.) NLPCC 2013. CCIS, vol. 400, pp. 266–276. Springer, Heidelberg (2013)
11. Chen, Y., Chen, L., Xu, K.: Learning Chinese Entity Attributes from Online Encyclopedia. In: Wang, H., et al. (eds.) APWeb Workshops 2012. LNCS, vol. 7234, pp. 179–186. Springer, Heidelberg (2012)