A Tool to Analyse Spatial Distribution of Science Research Activities Based on Toponym Resolution in Text

Jianxia Ma[1,2]*, Hanqing Ma[2], Shaoxiong Liu[2], Yingguang Zhao[2], Na Li[2]
[1]Cold and Arid Regions Environmental and Engineering Research Institute, Chinese Academy of Sciences,
Donggang West Road 320, Lanzhou, Gansu, China,
East China Normal University, Shanghai, China
[2] Scientific Information Center for Resources and Environment, CAS,
8 Middle Tianshui Road, Lanzhou 730000, Gansu Province, China
*Corresponding author, e-mail:majx@lzb.ac.cn

***Abstract*—Spatial distribution of research activities in geo-sciences is related not only to the distribution of authors but also to the distribution of research area. In the present work we developed a tool to analyze spatial distribution of science research activities based on toponym resolution in text of scientific papers. The idea is to identify and annotate toponym referred in research papers automatically, then to geo-code them, at last to carry out an analysis of the distribution of research area and authors in related subjects in geo-sciences. We carried out some experiment with the tool. Firstly, we extracted geographical names from research articles in a pre-built Chinese documentary database on some scientific subjects with words segment software and gazetteer. Secondly, combining with Google map's geo-coding API to get coordinates of the places, we exported the place names and coordinates into arcGIS. Consequently, we tried to analyze the spatial distribution of authors and research area in ecological footprint in China. The results show that, based on toponym resolution from large-scale article collection, we can analyze hot areas and blank areas of scientific research on some subjects and the distribution of the researchers of the subjects. And we presented a redesign of the system framework. The improvement is focused on leveraging a geographical knowledge database and some rules for disambiguation, integrating Conditional Random Fields in toponym resolution in Chinese text.**

Keywords-toponym resolution; geo-parsing; geo-coding; information analysis; text-mining; digital gazetteer; Geographic Information System (GIS)

I. BACKGROUND

With the development of Internet, digitalized information is booming. It is reported that there is almost 70% information in the internet contains toponym [1], and half of the catalogue items in library of University of California includes one or more subject headings related to place names [2]. Linking the textual resources and map is one of the development trends of digital library.

Recently, many scholars and applications have begun to show results of scientific papers' citation analysis combined with GIS visually. All of their researches are based on addresses of authors given by the authors directly. There are few reports on the analysis of distribution of research area based on text-mining in research papers, especially written in Chinese. In earth science, resources and environment related fields, research is closely related with some location. It is inefficient to read the articles one by one while annotate the research area by hand to get the understanding of the distribution of research area. In doing so, it is not easy to grasp where the research blanks and hot spots are.

Then how to analyze geographical feature in magnanimous textual collections and mine the hidden knowledge efficiently? How to relate textual resources in digital library with Map and GIS? The key is the toponym resolution in the research articles. The process of toponym resolution includes two tasks, namely Geo-Parsing and Geo-Coding [3].

In the present work, a tool to analyse spatial distribution of research activities based on toponym resolution in Chinese text was designed. It uses words segment software and gazetteer. Then coordinates of the places are obtained by combining Google map's geo-coding API. The place names and coordinates of the places are subsequently exported into arcGIS. Consequently, we tried to analyze the distribution of authors and research area in ecological footprint in China.

In addition, an improvement of the framework of the tool is presented. The improvement focused on leveraging a geographical knowledge database and some rules for disambiguation, integrating CRF in toponym resolution in Chinese text to record annotations of geo-referenced information of all of the items in the database.

The rest of this paper is organized as follows. The second section gives a review of related studies in toponym identification especially in papers written in Chinese. The third section makes a description of the framework of the analysis tool including function, methods we used and an effect analysis. The fourth part gives an example how we use the tool to carry out spatial analysis in ecological footprint in research

978-1-61284-848-8/11/$26.00 ©2011 IEEE

papers in Chinese. In the fifth section, a summary and discussion of future work is given.

II. INTORDUCTION TO RELATED STUDY

The process of toponym resolution includes two important tasks, i.e. Geo-Parsing and Geo-coding.

A. Geo-parsing

Geo-parsing consists of detecting and extracting the geographic names referred in the unstructured text of an article or a Web page using Named Entity Recognition (NER) techniques.

Gazetteers based extraction is one of the most popular techniques for Geo-Parsing. Clough [4] and Tobin [5] reported their geo-referencing systems which firstly used information extraction techniques to identify place names in textual documents and to resolve the place names against gazetteers. This approach is simple and allows efficient implementations, however, with a loss of precision in toponym extraction. And it's a tedious job to get a full covered gazetteer.

Natural language processing is generally based on statistical models. Various statistical models including Hidden Markov Models (HMMs) [6], Maximum Entropy Models (MEMs) [7], Maximum Entropy Markov Models (MEMMs) [8], Conditional Random Field (CRF) [9] were discussed in many documents for extraction of geographic names, and some have gotten great success in the CoNLL. However, these approaches require lots of training and are corpus dependent.

B. Geo-coding

Geo-coding is the key step to correlate textual information to maps. Gazetteer or the geographical knowledge base is the key component. Linda Hill [10] mentions three core elements in digital gazetteer, the placenames, type categories of places/features, and the footprint. A well-designed digital gazetteer can support geo-entity identification, toponym disambiguation and geo-coding. By now, the famous digital gazetteers includes ADL Gazetteer [11], Getty TGN [12]、GeoName [13]. And some digital map services, including Google Map [14], Microsoft Bing map [15], Yahoo PlaceFinder [16], Baidu Map [17] provide API for geo-coding.

C. Chinese Toponym Extraction

Unlike English, there is no blank to mark word boundaries in Chinese text. With the development of Chinese word segmentation in the past ten years, a lot of research has been carried out regarding Chinese topnonym extraction. The previous research focused on syntax rules and word segmentation[18,19]. Li reported max-margin Markov networks was used to identify unknown geographical names in Chinese text [20]. Yu proposed a cascaded hidden Markov model aimed to incorporate person name, location name, and organization name recognition into an integrated theoretical frame [21]. Maximum Entropy Model was used in some experiments [22]. Supporting Vector Machine was also used in geo-entity identification [23]. Recently, CRFs has been widely used in toponnym identification and part-of speech (POS) tagging with good performance [24, 25].

III. CURRENT FRAMEWORK OF THE ANALYSIS TOOL

We developed an analysis tool based on toponym resolution in text which supports the analysis of spatial distribution of research area. It can be used to analyse the distribution of the authors of the research fields too. The function and methods used in the tool are introduced as follows. The framework of the tool is given in Fig. 1.

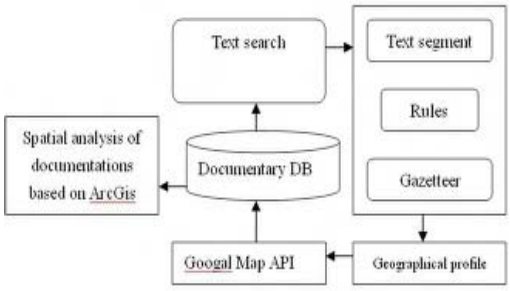

Figure 1. Framework of the analysis tool

A. Documentary Database Preparation

Firstly we built up document database with metadata on a special topic. In the database, we are interested in four fields, i.e. authors' affiliation and address, title and abstract. Authors' affiliation and address show the distribution of researchers, while toponyms in title and abstract are related to the research area. Then the tool read the items and input these four fields to the Geo-recognition module one by one.

B. Geo-recognition in Text

1) Geo-extraction from authors' affiliation and address fields.

The addresses of authors in the metadata were extracted with string match and some rules (for example, before some special character or punctuation). This procedure is simple, and the precision and recall is preferable.

2) Geo-recognition from unstructured text.

Normally, there are some places mentioned in title and abstract of Geo-related articles. There is no obvious punctuation marking the toponym in unstructured text.

An open-source Chinese segment package, IKAnalyzer was used at this stage. IKAnalyzer is combined with dictionary-based and grammar-based word segmentation. It provides a dictionary of 220,000 words, and supports extended dictionary with API as well. We extended the geographical name list with Sougou's dictionary [26] including 114395 Chinese Administrative districts names.

C. Geo-coding with Google Map

After extracting the place names from authors' affiliations, titles and abstracts, we sent the toponyms to Google Map's Geo-Code API [27], then got the coordinates of the toponyms. By now, the article has a geographical profile with the toponym and coordinates.

D. *Spatial Analysis of Research Activity Based on Toponym Resolution from Documents*

Once we searched and traversed the database and assigned geographical profile for each item, we can export the data into ArcGIS. Then we can carry out spatial distribution analysis of research activity on some subjects based on documentation. In this work, we carried out an analysis in the field of ecological footprint study.

Firstly we searched in CNKI and downloaded papers related to ecological footprint published from 2000-2010. There are 1858 items in the document collection. There are 1745 items have clear authors' affiliations and addresses. With the tool we developed, we identified 98% author's affiliation and address. But when it comes to geo-coding, 81.26% can be returned with coordinates with Google map's geo-code API. In combination with Google earth, we increased the rate of geo-coding to 94.96%.

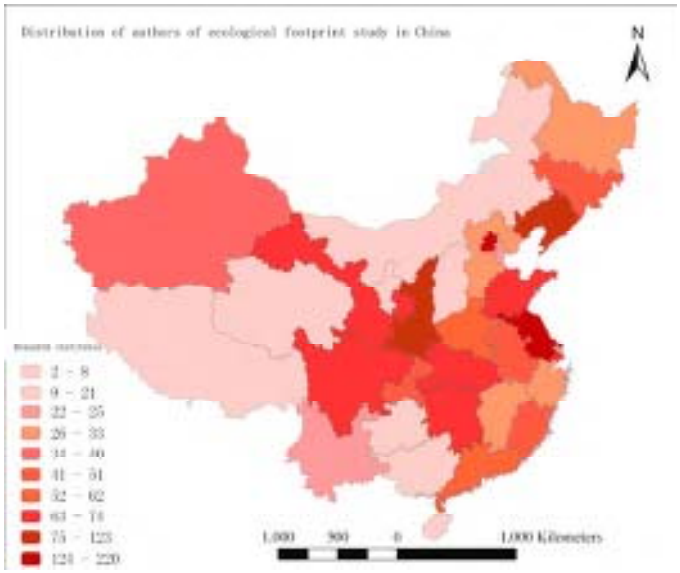

Figure 2. Distribution of authors of ecological footprint study in China

There are 1495 papers related to explicit research area. There are 1763 geographical names according to manual annotation in abstract. With the tool, the recall of the toponym is 87.18%, the precision is 88.16%, the F value is 87.67%. According to geo-coding, 84.72% addresses, i.e. 1148 toponyms can be geo-coded.

We analysed the result. Because our gazetteer covered almost all of the administrative district above Xiang，research area of ecological footprint study mainly related to district above "Xian"，the recall and precision is acceptable. On the other hand, there are also some study area in ecological footprint we cannot get coordinate from Google Map, for example "Northwest District", "Heihe River Basin", "urban cluster of the Pearl River Delt", "Buyunshan Natural Reserves" , and so on. The rate of geo-code of research area is lower than that of the author's address.

Based on the data, we made 2 maps with ArcGIS. Fig. 2 is about the distribution of authors who have carried out study in ecological footprint. Fig. 3 shows the distribution of research area of ecological footprint.

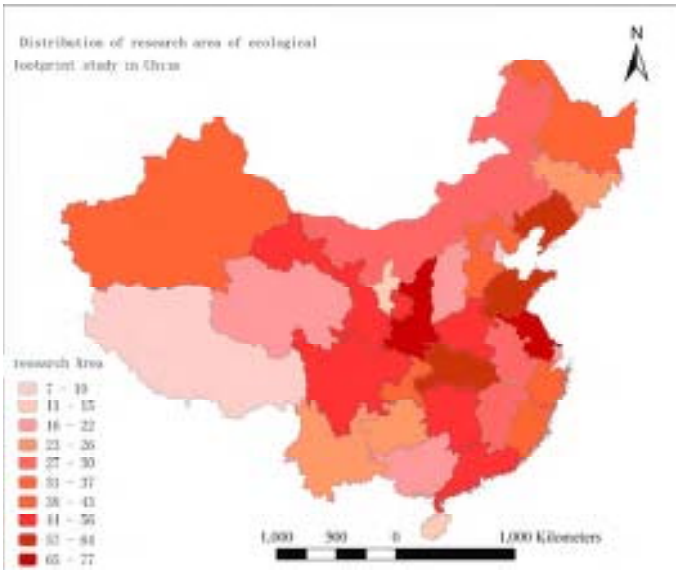

Figure 3. Distribution of research area of ecological footprint study in China

According to Fig. 2, the authors in ecological footprint study mainly come from Beijing, Jiangsu, Shanxi， and Liaoning.

As shown in Fig. 3, the hot areas of ecological footprint study are in Shanxi and Jiangsu Province. There are little research have been carried out in Ningxia, Xizang, Tianjin, Hainan and Qinghai. According to Yang Kaizhong's research [29], Ningxia, Shanxi and Guizhou's ecological civilization level rank the last three. Fig. 3 shows in the future we should pay much more attention to these three provinces in the study of ecological footprint too.

IV. REDESIGN OF THE FRAMEWORK

When we tried to analyse another paper collection in Palynology, we found that the recall and precision were decreased. That is mainly because 1) There are many physical geographical entities in this field's papers. But the gazetteer we used mainly covered administrative district names, and some natural geographical entities are not included in the gazetteer. 2) The gazetteer is just a name list, there is no semantic relationship. So when we extract a place name "Ping'an Xiang", we can not decide which one it is. 3) In the stage of geo-coding, with Google map's API we can not get coordinates for place names lower than "Xiang" and few physical geographical entities can be geo-coded. At the same time, we found that some place names appeared in the full-text were marked with coordinates by the authors. So it's necessary to redesign the framework, to better the effect of identification of geo-entities which are not included in the gazetteer.

After some comparison, we chose CRF as the solution. In order to improve the effect of disambiguation, we improved the structure of gazetteer as a geographical knowledge base with semantic relationship to support toponym disambiguation, and

enriched the geographical knowledge base dynamically from the geo-entity resolution procedure from unstructured text.

The improved framework of the analysis tool is as Fig. 4.

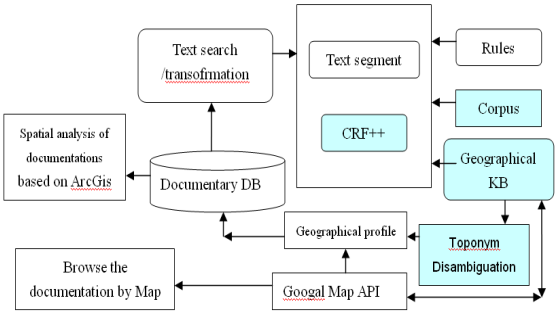

Figure 4. Improved framework of the analysis tool

A. *CRF++ Based Toponym Identification*

The redesigned framework includes CRF++ as the main module of geo-entity identification. In order to obtain better results, we made a corpus related to physical geography. In this corpus we annotate the toponym with "BIEO" system. "B" means Begin of a toponym, "I" means inter of a toponym, "E" means end of a toponym. O means other words.

Because CRF model highly depends on features, features selection is an important part of the model [28]. We selected 2 left and 2 right characters for feature extraction. And we considered the single character feature, single character's part-of-speech feature, the word as a sign of a geographical name, the word before a geographical name, the word next to a geographical name, the word in gazetteer. When we extracted the corpus, we can add the terms to the geographical knowledge base according to the probability as the toponym, the word before or next to the toponym. At the same time the corpus can be used to train the estimation parameters of values. The feature template using CRF is shown as below. Feature selection is an iteration process. The best one could be gained according to the balance of training time and experimental performance.

Character(n) (n=-2，-1，0，+1，+2)

POS(n)(n=-2,-1,0,+1,+2)

Location(n) (n=-2,-1,0,+1,+2)

LeftWord(n) (n=-2,-1,0,+1,+2)

RightWord(n) (n=-2,-1,0,+1,+2)

Dic(n) (n=-2,-1,0,+1,+2)

B. *Geographical Knowledge Base with Semantic Relationship Supporting Toponym Disambiguation*

During the toponym identification, the geographical knowledge base can be the proof of geo-entity match. Since one place has different names, the alternative names for one toponym in the geographical knowledge base will support the normalization of toponym. In the case of multi-place sharing the same names, the geographical knowledge base can support toponym disambiguation through the hierarchical relationship. For example, there are different places named "Ping'anXiang" in China. If "Gansu Province"is referred in the context of "Ping'an Xiang"， We can identify the toponym as the "Ping'an Xiang" in "Tianshui City, Gansu Province" according to the geographical knowledge base where there is an item shows the relationship "Gansu Province>Tianshui City>Zhangjiachuan Huizu Zizhixian>Pinganxiang". We can also make toponym disambiguation according to the geo-entity's feature type. In the step of toponym geo-coding, according to different scales and needs, some toponym can be mapped to the coordinate of the central point, while others can be mapped to polygon. Furthermore, as mentioned before, we should add the list for generic words marking toponym, words list before toponym, words list next to toponym. In order to support these function during the toponym identification and disambiguation, the schema of geographical knowledge base are designed as follows.

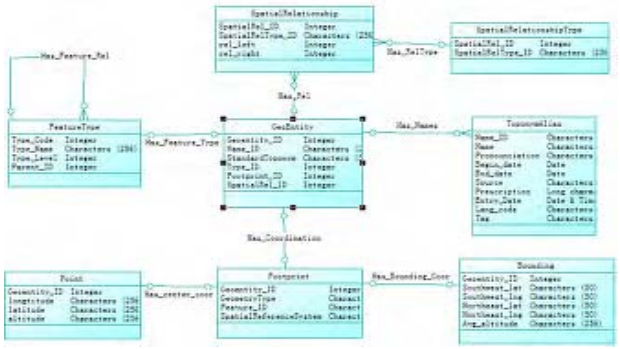

Figure 5. Schema of geological knowledge base

In the schema, table Geo-entity has a field Name_id connecting with table ToponymAlias to cope with cases like one place with different names. The field StandardToponym is the standard name of the geo-entity. Table Footprint means the coverage of the geo-entity, according to different scale we can choose point or bounding box. Table FeatureType means the feature type of the geo-entity. Here we divided the types of toponym into natural geographical entity and human geographical entity. The administrative district is down to the 5th level, i.e. "Xiang"，while the type of natural geographical entity is based on "Chinese Geography Thesaurus addendum 4".The spatial relationship between geo-entity is designed by table SpatialRelationshipType and SpatialRelationship. The relationship includes Adjoin-to, is-part-of, has-part-of, overlay, and so on. We reused some toponym resources for example "Dictionary of toponomys People's Republic of China", Sougou word list, "Chinese Geography Thesaurus addendum 4" and the open API of Google Map.

V. CONCLUSION

A tool to analyze spatial distribution of science research activities based on toponym resolution in text of scientific papers was developed. With this tool, we carried out an experiment on spatial distribution analysis of science research activities in ecological footprint. Firstly we extracted geographical names from research articles in a pre-built Chinese documentary database on ecological footprint with words segment software and gazetteer.

Secondly, combining with Google map's geo-coding API to get coordinates of the places, we exported the place names and coordinates into arcGIS. Consequently, we tried to analyze the spatial distribution of authors and research area in ecological footprint in China. The experiment shows that it is possible to analyze distribution of research activities based on automatic identification and annotation of the geo-entity in large-scale textual collections. The tool is useful for the science decision maker to allocate research resources.

Because the gazetteer we used in the tool doesn't cover some of physical geographical names, we redesign the framework leveraging a geographical knowledge database and some rules for disambiguation, integrating Conditional Random Fields in toponym resolution in Chinese text. On the spatial analysis based on toponym resolution in unstructured text, further research and experiment is needed and actually is on-going. We need much more corpus to be trained, need to adjust the feature template to get better efficiency. We also need to take into consideration of other valuable heuristics to improve the toponym resolution. A systematic evaluation of the method we have taken should be carried out as well.

REFERENCES

[1] "MetaCarta Corporate brochure," http://www.metacarta.com/docs/corporate_brochure_06_05, accessed at November 12, 2009.

[2] V. Petras, "Statistical Analysis of Geographic and Language Clues in the MARC Record," http://metadata.sims.berkeley.edu/papers/Marcplaces.pdf, accessed at January 16, 2010.

[3] J. L. Leidner, "Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names," Ph. D. Thesis, University of Edinburgh, Edinburgh, 2007.

[4] P. Clough. "Extracting metadata for spatially-aware information retrieval on the Internet," Proc. ACM Workshop on Geographic Information Retrieval (GIR), pp. 25-30, 2005.

[5] R. Tobin, C. Grover, K. Byrne, J. Reid, and J. Walsh, "Evaluation of Georeferencing," Proc. the 6th Workshop on Geographic Information Retrieval, 2010, doi: 10.1145/1722080.1722089.

[6] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," Proc. IEEE, 77(2): pp. 257-286, 1989.

[7] D. Freitag, A. McCallum, and F. Pereira, " Maximum entropy Markov models for information extraction and segmentation," Proc. The Seventeenth International Conf. on Machine Learning. pp.591-598, 2000.

[8] A. McCallum and W. Li, " Early results for named entity recognition with conditional random fields feature induction and web-enhanced lexicons," Proc. the 7th Conference on Natural Language Learning, pp.188-191, 2003

[9] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields:Probabilistic models for segmenting and labeling sequence data," Proc. the 18th International Conf. on Machine Learning, pp. 282–289, 2001.

[10] L. L. Hill, "Core Elements of Digital Gazetteers: Placenames,Categories, and Footprints," Proc. the 4th European Conference on Research and Advanced Technology for Digital Libraries, pp. 280–290, 2000.

[11] "Alexandria Digital Library Project, Gazetteer Development," http://www.alexandria.ucsb.edu/gazetteer, accesed at November 2, 2010.

[12] The Getty Research Institute,"Getty Thesaurus of Geographic Names® Online," http://www.getty.edu/research/tools/vocabularies/tgn/index.html,bbb accesed at November 10, 2010

[13] GeoName Team,"GeoNames," http://www.geonames.org/, accesed at November 10, 2010.

[14] Google,"Google Maps," http://maps.google.com/, accesed at November 9, 2010.

[15] Microsoft, "BING MAPS," http://www.microsoft.com/maps/, accessed at November 9, 2010.

[16] Yahoo! Inc., "Yahoo!PlaceFinderGuide,"http://developer.yahoo.com/geo/placefinder/guide/, accesed at March 9, 2010.

[17] Baidu Map, "Baidu Map API,"http://openapi.baidu.com/map/index.html, accessed at March 9, 2010.

[18] X. Zhang, G. Lv, Z. Xie, and Y. Sun, "Extraction and Visualization of Geographical Names in Text," Proc. 24th of ICC, http://icaci.org/documents/ICC_proceedings/ICC2009/html/nonref/12_8.pdf, 2009, accesed at December 19, 2010 .

[19] X. Xiang, X. Shi, and H. Zeng, "Chinese named entity recognition system using statistics-based and rules-based method," Computer Applications, vol. 25(10), 2005, pp. 2404-2406.

[20] L. Li, Z. Ding, and D. Huang, "Recognizing location names from Chinese texts based on max-margin markov network," Proc. International Conference on Natural Language Processing and Knowledge Engineering, pp. 1-7, 2008.

[21] H. Yu, H. Zhang, Q. Liu, X. Lv, and S. Shi, "Chinese named entity identification using cascaded hidden Markov model," Journal on Communications, vol. 27 (2), 2006, pp. 88-95.

[22] J. Qian, J. Zhang, and T. Zhang, "Research on Chinese Person Name and Location Name Recognition Based on Maximum Entropy Model," Journal of Chinese Computer Systems, vol. 27(9) , 2006, pp. 1761-1765.

[23] L. Li, D. Huang, and C. Chen, "Identification of location names from Chinese texts based on support vector machine," Journal of Dalian University of Technology, vol. 47(3), 2007, pp. 433-438.

[24] X. Tang, X. Chen, and X. Zhang, "Research on Toponym Resolution in Chinese Text," Geomatics and Information Science of Wuhan University, vol. 2010(8), 2010, pp. 930-935.

[25] P. Lu, Y. Yang, Y. Gao, and H. Ren, "Hierarchical conditional random fields (HCRF) for Chinese named entity tagging," Proc. the 3rd International Conference on Natural Computation, pp.24-28, 2007.

[26] Sogou.com,"Sogou Dictionary," http://pinyin.sogou.com/dict/list.php?c=361&page=2, accessed at March 9, 2011.

[27] Google, "Google Geocoding API," http://maps.google.com/maps/api/geocode/, accessed at March 9, 2011.

[28] L. Ma, " A Study on Chinese Location Names Recognition Based on Conditional Random Fields," M. S. thesis, Dalian Univ. of Technology, Dalian, China, 2010.

[29] K. Yang, "Rank of Ecological Civilization of Provinces in China," http://news.sina.com.cn/c/sd/2011-03-31/115222215340.shtm, accessed at March 19, 2011.