Web Resource Geographic Location Classification and Detection

Chuang Wang[*2], Xing Xie[1], Lee Wang[3], Yansheng Lu[2], Wei-Ying Ma[1]

[1]Microsoft Research Asia, 5F, Sigma Center, No. 49, Zhichun Road, Beijing, 100080, P.R. China

[2] Department of Computer Science, Huazhong University of Sci. & Tech., Wuhan, 430074, P.R. China

[3] Microsoft Corporation, One Microsoft Way, Redmond, WA 98052, USA

{xingx,wyma}@microsoft.com {chwang, ysl}@mail.hust.edu.cn leew@microsoft.com

ABSTRACT

Rapid pervasion of the web into users' daily lives has put much importance on capturing location-specific information on the web, due to the fact that most human activities occur locally around where a user is located. This is especially true in the increasingly popular mobile and local search environments. Thus, how to correctly and effectively detect locations from web resources has become a key challenge to location-based web applications. In this paper, we first explicitly distinguish the locations of web resources into three types to cater to different application needs: 1) *provider location*; 2) *content location*; and 3) *serving location*. Then we describe a unified system that computes each of the three locations, employing a set of algorithms and different geographic sources.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *search process, retrieval models, information filtering*

General Terms: Algorithms, Experimentation, Performance

Keywords: Location-based web application, web location, geographic location, provider location, content location, serving location

1. INTRODUCTION

Intuitively, web resources (including web pages, sites, etc.) have geographic features [1,2,3,5]. For example, a web page with information about or within a special geographic scope, such as web resources listings on houses for sale in a given region, could be regarded as a local page with a certain location.

With rapid pervasion of the web into users' daily lives, in the increasingly popular mobile and local search environments, location-based web applications are emerging. The common principle of these applications is to detect the geographic attribute from web resources then match it with current user's location. User location can be acquired from his/her context environment or be provided explicitly. Therefore, how to correctly and effectively deduce the web location, taking full advantage of relevant geographic sources, is the key to the success of these location-based web applications.

Having investigated different needs on web locations from a number of location-based web applications, we conclude that there are at least three types of web locations that may co-exist in the same web resource. In this paper, we first explicitly distinguish the locations of web resources into three types, namely *provider location*, *content location* and *serving location*, to cater different application needs. We also introduce a unified system employing a set of algorithms to compute the three types of locations by extracting geographic information from the web resource content, mining hyperlink structures and user logs. Experimental results on large samples of web data show that our solution outperforms existing type-less approaches.

[*] This work was done when the first author worked as an intern at Microsoft Research Asia.

WWW 2005, May 10-14, 2005, Chiba, Japan.
ACM 1-59593-051-5/05/0005.

2. WEB LOCATION CLASSIFICATION

2.1 Location Type Definitions

A web resource may have the following three different types of locations:

- **Provider location:** *The physical location of the provider (organization, corporation or person) owning the web resource*. This location is crucial to web geographic information retrieval and navigation such as online map and Yellow Pages services.
- **Content location**: *The geographic location that the content of the web resource is about*. As a spatial attribute of the web content, this location can be used to classify and organize web resources to better satisfy user' information needs. One of its applications is location-based search.
- **Serving location:** *The geographic scope that the web resource reaches*. Knowing the serving location of a web resource can benefit many business applications such as local advertisements and e-commerce.

In real world, every entity, whether it provides "products" or "services" to its intending users, has its physical location and its affecting geographic scope, namely provider location and serving location, respectively. Furthermore, the two locations usually have different geographic scopes and scales. Generally, provider location is a point location where the provider locates or multiple point locations if the provider entity is geographically distributed. Serving location is often a shaped region or multiple regions where its "products" or "services" can reach. Note that our proposed provider location is also different from presented host location in [5] that is derived from the geographic sources for Hosts rather than for web content. Content location is the spatial feature of web content and is usually considered more meaningful at page level, even at block level, while both provider location and serving location are spatial attributes of the entity behind web resources and are often detected at the site level.

We will take MSN site as an example to further illustrate these three types of locations. On the site, our algorithms found that

provider location of the site is "Microsoft Corporation One Microsoft Way Redmond, Washington 98052, USA", namely the physical location of Microsoft Corporation. And the computed serving location is "Global" due to msn.com is a general web site with world-wide user reach. The content location of MSN's New York local page [6] is "New York, NY, USA".

2.2 Location Detection by Type

As introduced in [5], there exist various available geographic sources to derive web location, such as place name, postal code, phone number, hyperlink, and languages, etc. In our approach, different types of location are calculated according to different algorithms and geographic sources. Figure 1 shows the workflow of our novel location detection system by the three location types. We first utilize the extracted address segments to acquire the provider location. Then all extracted geographic references will be used to compute the content location. The serving location is estimated based on the content locations computed in the above step, plus the geographic information carried from inbound hyperlinks and/or from user logs.

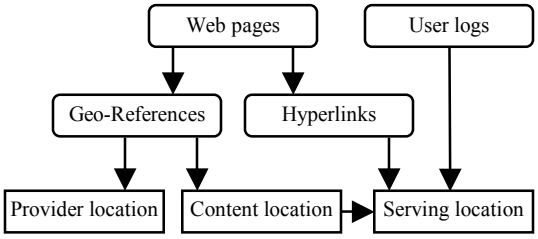

Figure 1. The flowchart of location detection system by location type.

Computing Provider Location Challenges of calculating the provider location of a web resource include: 1) accurately recognizing and extracting address strings from the web content, and then 2) correctly estimating whether extracted address strings are the provider location. Some simple rule such as address string templates, initializations, abbreviations and so on can facilitate this reorganization process. After an address string is recognized and extracted, we employ Support Vector Machine [4] to learn whether the string is a provider location in the binary classification setting. The following features were found be useful to our determination process: URL, title, anchor text, page content, referred frequency, hierarchical level, and even the spatial position of extracted address strings on the page.

Computing Content and Serving Location Since it is common that multiple geographic references may coexist in a given web page, the key problem is to estimate the representative location. Here, we use the similar algorithm in [1,3] to calculate content location. The basic idea is to utilize various extracted geographic references to deduce the location focus. The algorithm of calculating serving location is similar to that of content location, the difference lies in the input data and iterative procedure of the algorithm of calculating serving location. First, serving location is initialized to obtained content location. Then, based on the serving location value of the previous iteration and the serving locations of other sites that have inbound hyperlinks to the site of interest, serving location can be further refined using the above content location algorithm.

3. EXPERIMENTAL EVALUATIONS

Three frequent geographic references within the scope of USA, i.e. place name, postal code and telephone number, are exploited in our experiments. The benchmark data set is a collection of real web resources of major USA governmental sites whose top domains are .gov. These data were used in TREC2003 and has a wide geographic range covering all geographic levels within USA. Finally, three common measures: Precision, Recall, and Micro-F1 are reported.

Table 1. Summary of experimental results.

Location Type		Precision	Recall	Micro-F1
Provider Location	SVM-Linear	0.87	0.89	0.88
	SVM-Polynomial	0.93	0.88	0.90
	SVM-Sigmoid	0.92	0.90	0.91
	SVM-Gaussian	0.96	0.92	0.94
Content Location		0.95	0.80	0.87
Serving Location		0.93	0.91	0.92

Table 1 is the summary of our experimental results. For provider location, SVMs with several kernels, including linear, polynomial, sigmoid, and Gaussian, are tried in our experiments. The best performance with 0.94 Micro-F1 is achieved when using Gaussian kernel. In addition, the algorithm of content location and serving location also achieved encouraging performance with Micro-F1 0.87 and 0.92, respectively. Having analyzed the results, we find that the main contribution to the quality of our algorithm is that we have distinguished the locations of web resources into provider location, content location and serving location rather than mixing them into one location, and each location is computed by only considering its relevant geographic sources and intrinsic characteristics.

4. CONCLUSIONS

In this paper, we first defined the following web location types: *provider location*, *content location*, and *serving location*, and described their unique characteristics by examples. Different business applications need different types of web locations. Multiple locations often co-exist in the same web resource. Ignoring location types or choosing a wrong type to use could result in poor detection accuracy in web applications. Then, we proposed a unified system that computes all types of locations employing a set of effective location detection algorithms and only uses relevant data sources for each location type to achieve high accuracy and fast speed.

5. REFERENCES

[1] Amitay, E., Har'EI, N., Sivan, R., and Soffer, A. Web-a-where: geotagging web content. 27th Annual International ACM SIGIR Conference, Sheffield, UK, Jul. 2004.

[2] Buyukkokten, O., Cho, J., Garcia-Molina, H., Gravano, L., and Shivakumar, N. Exploiting geographic location information of web pages. ACM SIGMOD Workshop on the Web and Databases 1999, Philadelphia, USA, Jun. 1999.

[3] Ding, J., Gravano, L., and Shivakumar N. Computing geographic scopes of web resource. 26th International Conference on VLDB, Cairo, Egypt, Sep. 2000.

[4] Hearst, M.A. Trends and controversies: support vector machines. IEEE Intelligent Systems, 13(4), Jul. 1998, 18-28.

[5] McCurley, K. S. Geographic mapping and navigation of the web. 10th WWW Conference, Hong Kong, May 2001.

[6] MSN New York local page. http://local.msn.com/NewYork/