Automatic Geotagging of Russian Web Sites

Alexei Pyalling, Michael Maslov, Pavel Braslavski
Yandex
Vavilova 40
119991 Moscow, Russia
+7 (495) 974 35 55
{pyal, maslov, pb}@yandex-team.ru

ABSTRACT
The poster describes a fast, simple, yet accurate method to associate large amounts of web resources stored in a search engine database with geographic locations. The method uses location-by-IP data, domain names, and content-related features: ZIP and area codes. The novelty of the approach lies in building location-by-IP database by using continuous IP blocks method. Another contribution is domain name analysis. The method uses search engine infrastructure and makes it possible to effectively associate large amounts of search engine data with geography on a regular basis. Experiments ran on Yandex search engine index; evaluation has proved the efficacy of the approach.

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

General Terms
Algorithms, Design, Experimentation, Verification.

Keywords
Geotagging, Geographic Information Retrieval.

1. INTRODUCTION
Recently location aspects of web resources and their owners have become very important for many Internet surfers. This shift is marked by academic research in this area and emergence of local online search services.

Yandex search engine (www.yandex.ru) indexes resources in the post-soviet country domains and Russian-language resources elsewhere. At the time of writing, Yandex has indexed more than 600 million pages of more than 2.5 million web sites; about 95% of them belong to Russia. Although most Internet activities can be observed in large cities like Moscow and Saint-Petersburg, the Internet in Russia and other post-soviet countries develops mainly on account of the remote areas. This fact makes geographic information retrieval an important issue for Yandex search engine.

The issue is partly solved by Yandex manually-edited directory (http://yaca.yandex.ru). Presently, the directory contains around 87,000 entities with manually assigned geography; about 48,000 of them have Russian city attributions. The geography attribute combines different semantics of localities: 1) provider location (the physical location of the resource owner); 2) content location (the geographic location that the content of the web resource is about); 3) serving location (the area that the web resource reaches) [2]. Manually assigned values can be inherited by subdomains or individual pages of the site. Editors disallow value inheritance for specific domains (e.g. free hosting services or public domains). Approximately 140,000 sites additionally got Russian city attributes through such inheritance from the directory (*extended manual classification, EMC*). However, the current geotag coverage of Yandex database is insufficient, which drives us to investigate automatic methods for massive geography association with already indexed web resources. We use EMC as validation set for the methods presented below.

WWW 2006, May 23–26, 2006, Edinburgh, Scotland.
ACM 1-59593-323-9/06/0005.

The industry-derived issue defines a very pragmatic approach: the methods must be efficient and sound and must make extensive use of the already available data. The poster reports on progress in automatic geotagging of Russian sites on city level.

2. DATA AND METHODS
In the literature we find various methods that make use of location-by-IP data, domain names, as well as site content (location references like city names, telephone area codes and zip codes) for geotagging (related works are not cited due to space limitations). The main idea of our approach is to efficiently combine multiple sources of geographic information.

For city detection we developed two kinds of methods dealing with 1) site content and 2) site-level data (domain name and IP address). The methods were combined in a workflow shown in Figure 1. The number of classified sites, precision (P) and recall (R) values calculated by EMC are provided for each classification step. Dashed arrows reflect the fact that classification results are combined with input data for subsequent processing, so classification results are accumulated along the workflow.

1. Contend-based classifier (CBC). The method uses not original documents but their search index representations. While this does not allow us to precisely extract addresses from pages, it greatly increases algorithm efficiency. We compiled a list of six-digit ZIP codes for 12,000 locations in Russia [3] and a list of telephone area codes for 2,000 locations [1] along with location names. Two query templates were developed. The first is aimed at finding web pages with both ZIP code and respective location name. The second is focused on extracting pages with area code, location name and address elements like street or telephone number designators in close proximity. If several references are retrieved from a site, then it is required that the majority of them refer to the same location.

2. Domain label classifier (DLC). The method is based on domain label analysis. First, we assume that a domain label equal to a city name transliteration is a good indicator of site-city affiliation. Input data analysis allows us to sift out ‘good’ transliteration

variants: if the majority of known sites with a given domain label belong to the corresponding city, then we assume, that all sites with the domain label belong to the city (for instance Tver city sites: tver.eparhia.ru, tver.marketcenter.ru, www.tver.ru). Second, we look for city-specific domain labels, i.e. if the majority of known sites with the label belong to the same city, then the label is 'good'. Such labels are usually city nicknames or abbreviations (e.g. *nsk* – Novosibirsk, *dolgopa* – Dolgoprudny)

3. Domain name hierarchy classifier (DNHC). The idea is to find 'good' city domains whose subdomains are likely to belong to the same city, e.g. *spb.ru* and *omskcity.com* (Saint Petersburg and Omsk, respectively). Note that DNHC is used twice in the workflow (see Figure 1).

4. Location-by-IP (Loc-by-IP). We use an in-house database IPREG associating hosts' IP addresses with respective locations. IPREG had been compiled from Internet registry records for other purposes. IPREG is validated in the workflow, i.e. only 'good' IP address blocks of IPREG are kept.

5. IP blocks classifier (IP-blocks). City sites are often hosted by local providers who are not necessarily listed in IPREG or similar databases. Consequently, resources belonging to the same city often form continuous blocks in the IP address space. The method is based on determining such 'good' continuous IP blocks, i.e. where the majority of known sites in the block belong to the same city.

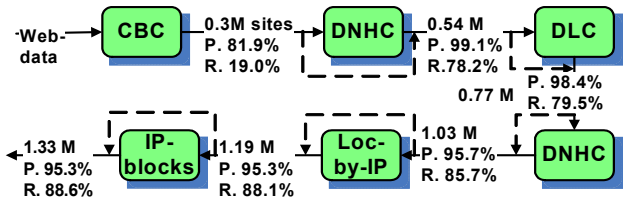

Figure 1. Classification workflow

As shown in Figure 1, the first DNHC deployment increases significantly both recall and precision according to EMC. Subsequent steps do not lead to a drastic increase in EMC-calculated quality (there is even a slight decline in precision), however, the number of classified sites grows substantially (due to less popular sites not presented in EMC).

As a result, using the workflow we could associate ~1.3 million Russian sites of ~2 million presented in Yandex database with respective Russian cities.

3. EVALUATION

Performance of the algorithm working with good, highly referenced sites can be taken from the comparison with EMC data. To test overall performance of the algorithm under stress conditions was generated a test set. We compiled a list of randomly chosen 1,200 web sites, not more than one per second level domain. All the sites in the list were tagged automatically with city labels or got 'region zero' tag (i.e. the city could not be resolved by the algorithm). The list was presented to Yandex directory editors for manual tagging under the customary policy. The data obtained from manual tagging allowed to divide the test set into three categories: 1) geographically local sites, 2) good, not 'garbage' sites (i.e. not doorways, not sites under construction, not obsolete or cyber squatter sites) , and 3) the full set of sites.

The results of evaluation of algorithm for all of these categories are summarized in Table 1. The first column corresponds to the subset of local sites (1). Zero attribute of automat for this set was treated as no assignment – loss in the recall of the algorithm, no loss in precision. In the second and third columns, the automatically assigned zero tag was interpreted as 'no geography' attribution. The case is somewhat questionable, since the classifier was not designed to differentiate between local, global and 'garbage' sites; 'region zero' rather means that the city detection technique applied was not successful. As a result the precision and the recall factor for this cases are practically the same.

Table 1. Evaluation Results

	Local sites	Local + non-local sites	Full sample (+ 'garbage')
# of sites	723	1048	1200
Precision	0.917	0.722	0.688
Recall	0. 751	0.696	0.667
F1	0.826	0.709	0.677

4. CONCLUSION AND FUTURE WORK

The poster describes a number of methods aimed at solving site geotagging task. The methods make use of various sources of information like IP-by-location database and domain name, as well as content-based information: direct search of ZIP and area codes on site pages. The methods use search engine infrastructure and make it possible to effectively associate large amounts of search engine data with geography on a regular basis.

A novel approach was developed for associating IP addresses with locations based on site content. The methodology yields better accuracy for geotagging purposes compared to traditional methods based on Internet registry records parsing. Another contribution is methods based on exhaustive analysis of site domain names.

The performed evaluation proved the suitability of the approach for industrial needs. However, the evaluation demonstrated that the main problem is to distinguish local sites from national or global sites. This issue will be addressed in the nearest future: a classifier to distinguish sites without geographic context will be developed.

5. ACKNOWLEDGMENTS

The authors would like to thank the editors of the Yandex directory who took the trouble to manually assess the results of the automatic classification.

6. REFERENCES

[1] Long Distance Codes, Rostelecom, http://www.rt.ru/tools/references/longdistance/

[2] Wang, C., Xie, X., Wang, L., Lu, Y., and Ma, W.Y. Web Resource Geographic Location Classification and Detection. In *WWW 2005*, May 10-14, 2005, Chiba, Japan, 1138-1139.

[3] ZIP Codes Reference of the Russian Postal Service, http://info.russianpost.ru/html/ops.html