Automatic Geotagging of Russian Web Sites

Alexei Pyalling, Michael Maslov, Pavel Braslavski
Yandex

ABSTRACT

In this work a fast, simple, yet accurate method to associate large amounts of web resources stored in a search engine database with geographic locations is presented. 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

METHODS

Direct data extraction from the Web

Queries to search engine were generated to extract blocks of information looking like address blocks. **(Fig 1).** Query templates were optimized to extract maximum number of correctly determined sites. This allowed to **determine 30%** of sites present in Yandex database with the **accuracy ~90%**.

Propagation of the results

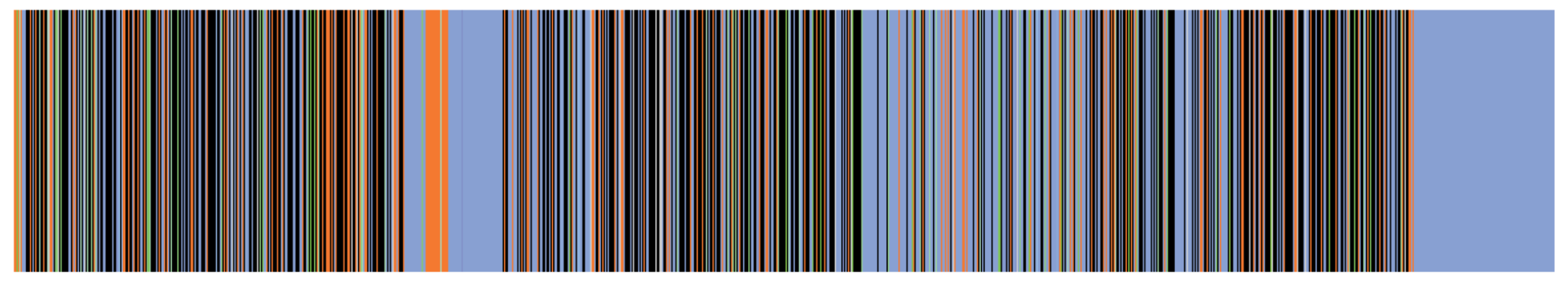
Three hypothesis were used to increase the number of determined sites

1. Geographically local resources forms **continuous blocks in IP space** (use the same local provider)**(Fig. 2)**.
2. Site domain **names often have a hierarchical structure**. Top level name determines the country (ru, ua, de), next level name determines the city (Omsk, Samara, Kiev)**(Fig 3)**
3. There is a number of **keywords** presence of which in the domain name can be **used as a site's city indicator** (direct transliteration of the city (Kiev), city nicknames (nnov - Nizhnii Novgorod, spb - Sankt Petersburg))**(Fig. 4)**

Fig. 1 A sample query

<table>
<tr><td>ZIP
142432 &&</td><td>City
Chernogolovka</td><td>Telephone code
8252 &&</td><td>City
Chernogolovka &&</td><td>Address hints
(street | e-mail)</td></tr>
</table>

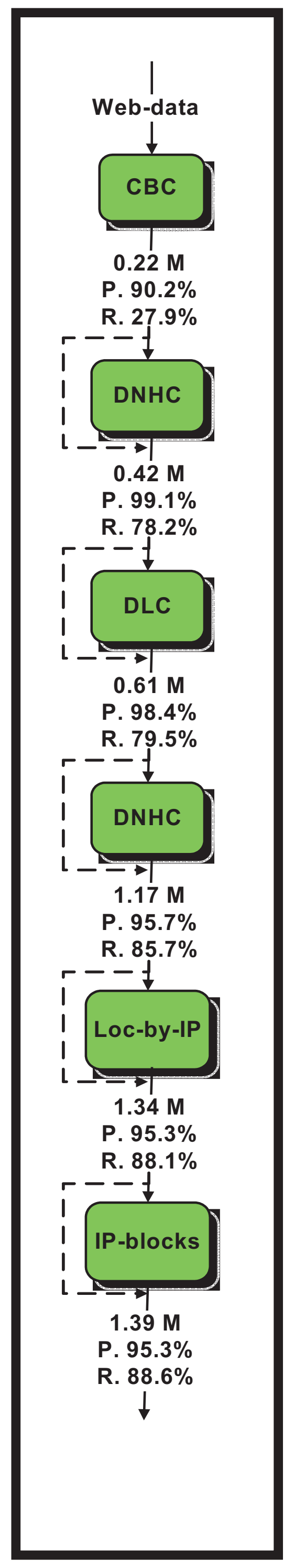
Fig. 2 Colored geotags for sorted by ip sites list



IP block construction method for sorted by IP sites list

Fig. 4 Using nicknames for site resolution (all sites are resolved as Chernogolovka's sites)

ftp.chg.ru	Sites used for the rule construction	chg-neformat.nm.ru	Correctly tagged sites
sgi.chg.ru		www.chg-info.com	
doctor.chg.ru		www.chg.net.ru	
50-let.chg.ru		rock-chg.chat.ru	
albom.chg.ru		chtc.chg.su	
portal.chg.ru		dj.chg.su	
www.chg.ru		fvol.chg.su	
garik.chg.ru		forum.chg.su	
netserv3.chg.ru		dingo-chg.narod.ru	
baby.chg.ru		skoch-chg.narod.ru	
chg.europortal.ru	Poorly tagged sites	chg-bus.narod.ru	
chg.maxtrader.ru		chg.fastbb.ru	
chg.fatal.ru		apache.chg.ru	

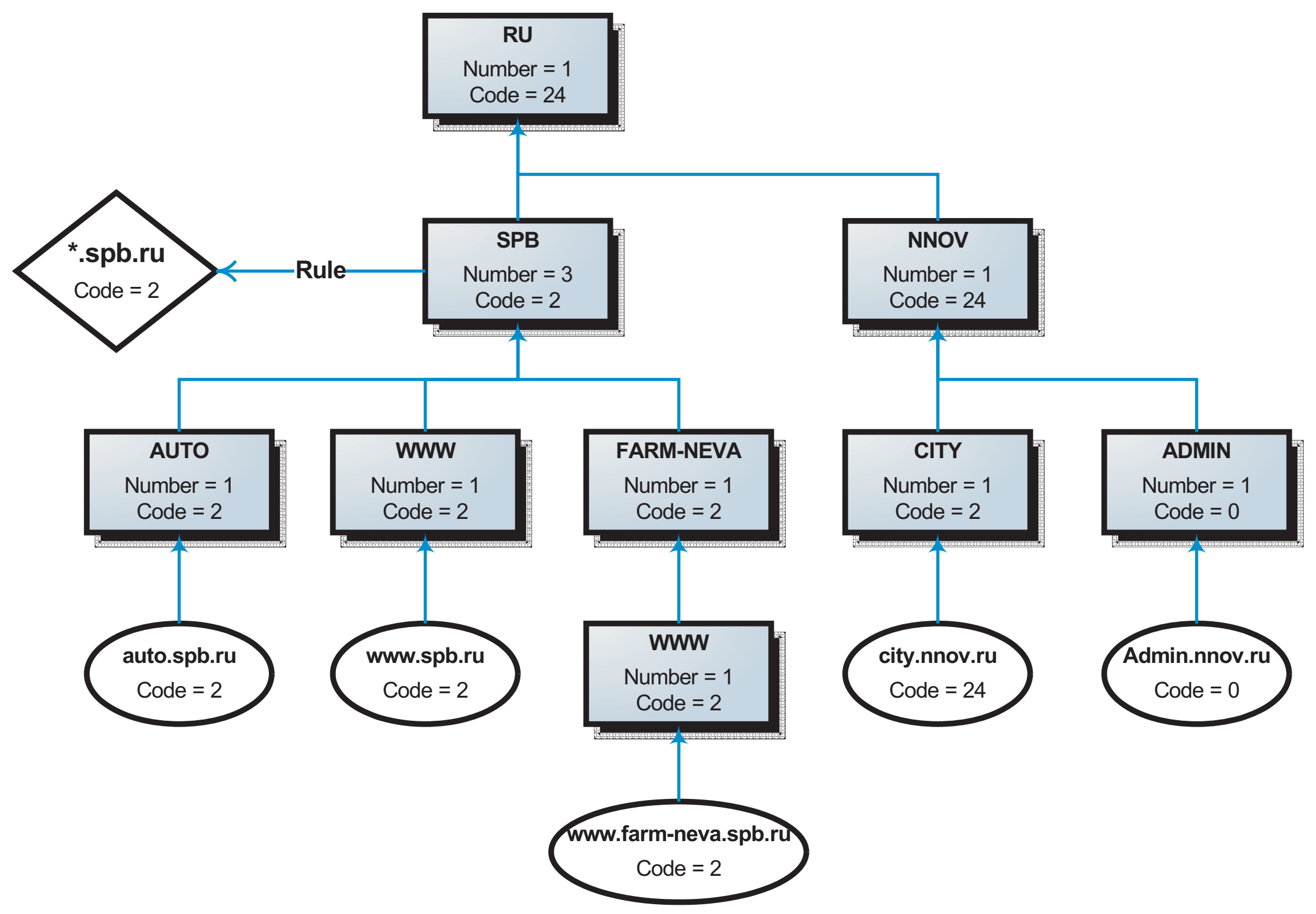


ALGORITHM DESCRIPTION

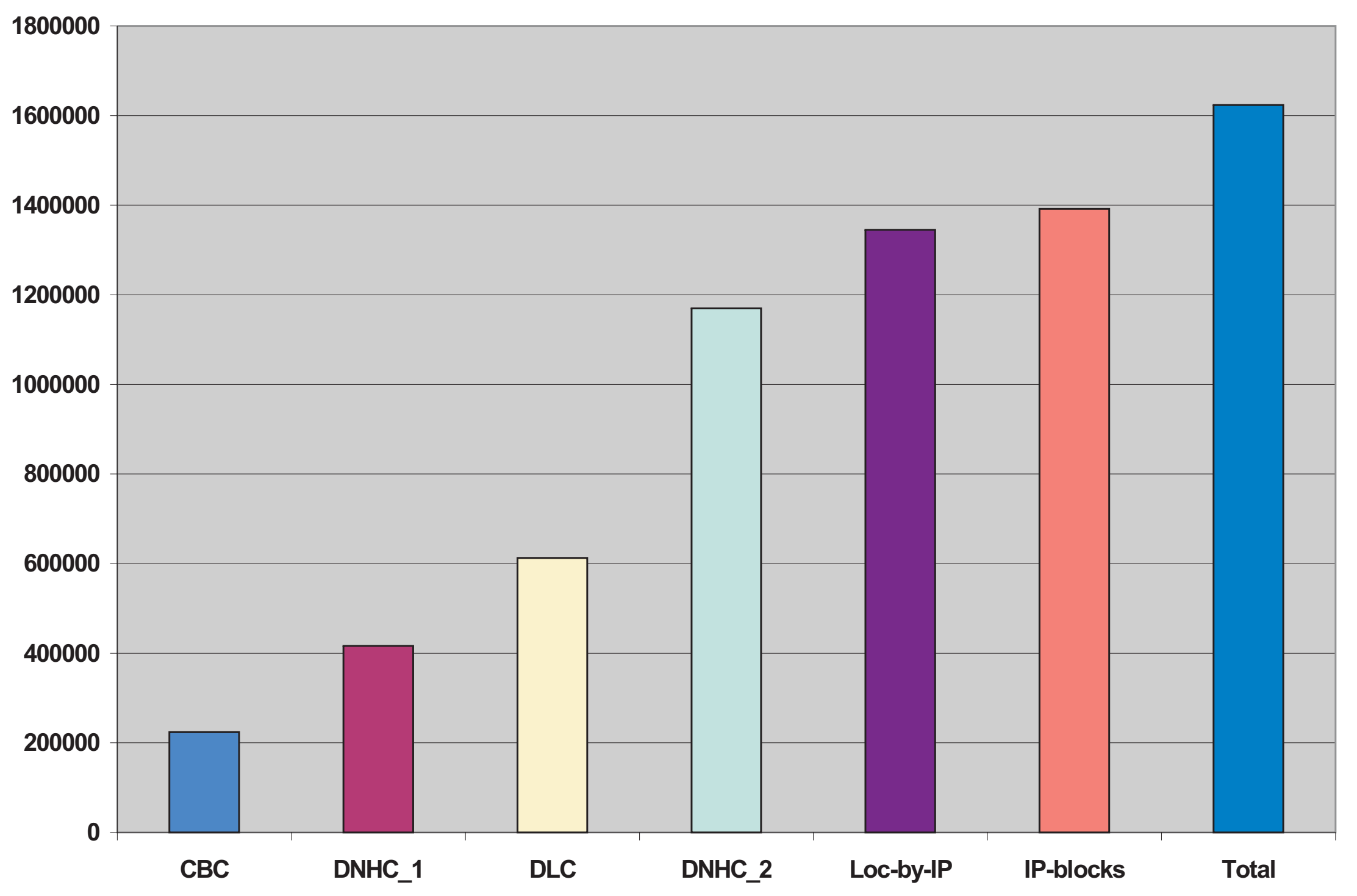
In the literature we find various methods that make use of location-by-IP data, domain names, as well as site content (location references like city names, telephone area codes and zip codes) for geotagging (related works are not cited due to space limitations). The main idea of our approach is to efficiently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

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.

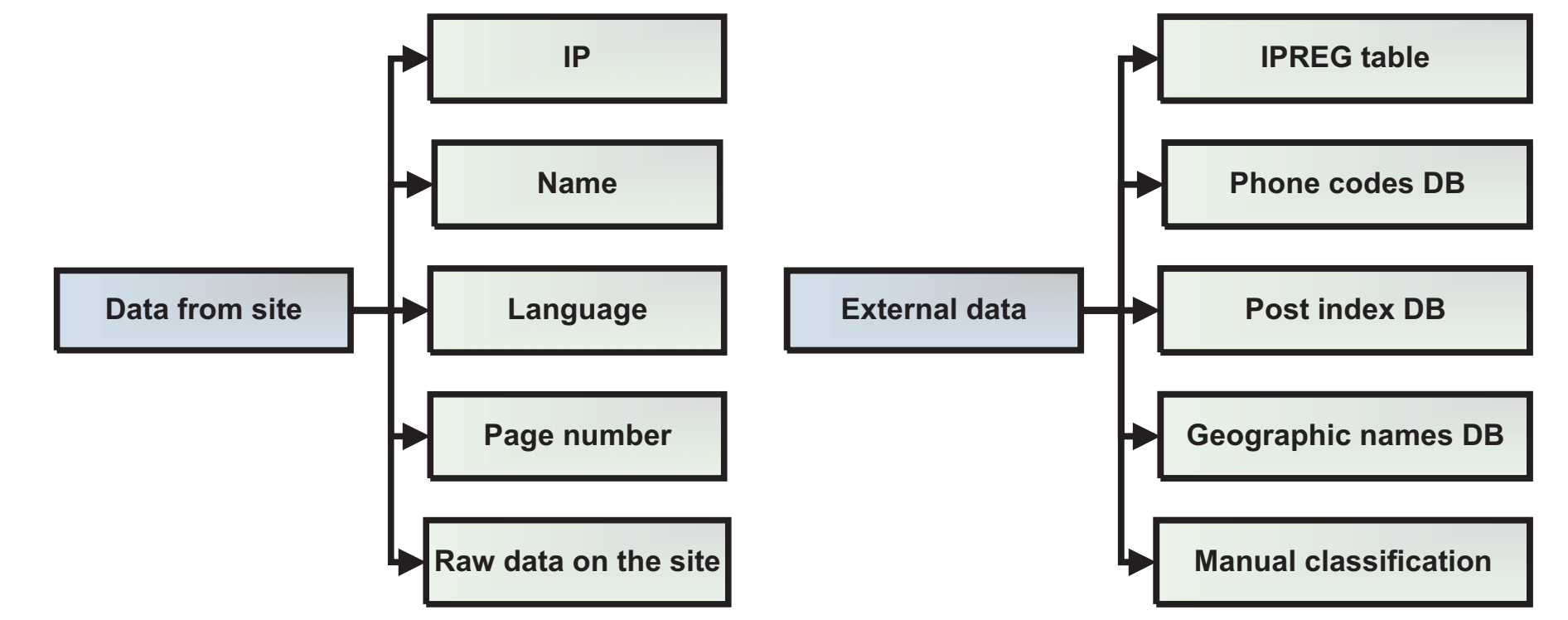
Fig. 3 Name ierarchy rules construction



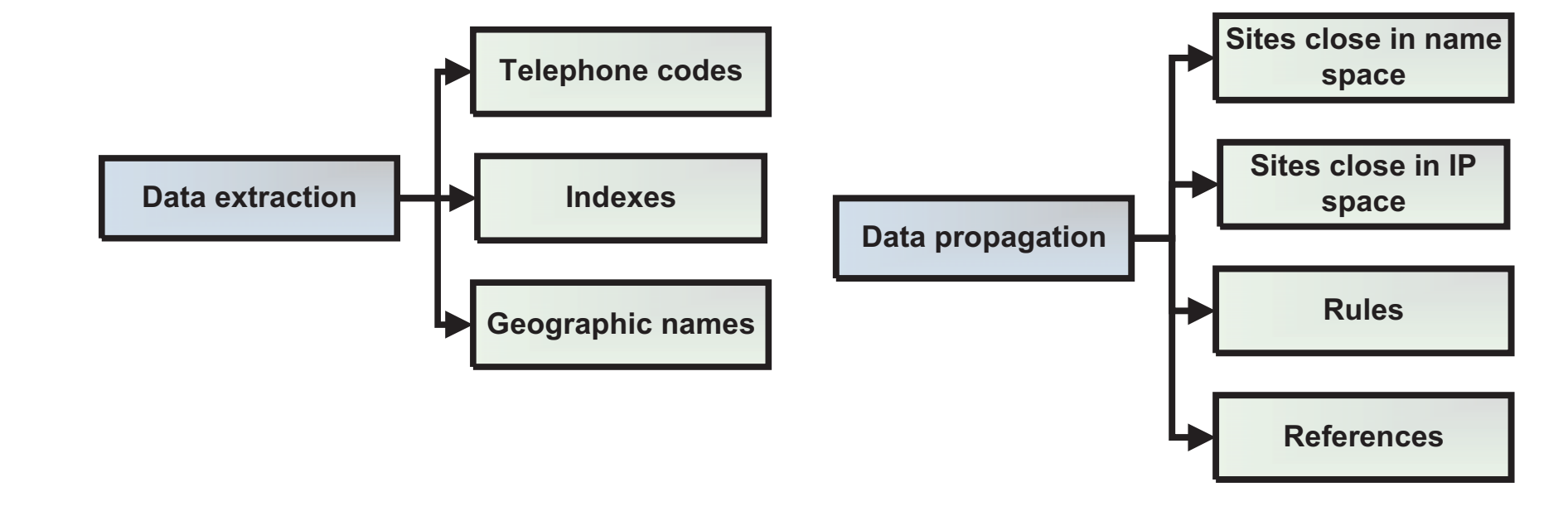
Overall performance of the algorithm
Total number of determined sites on different



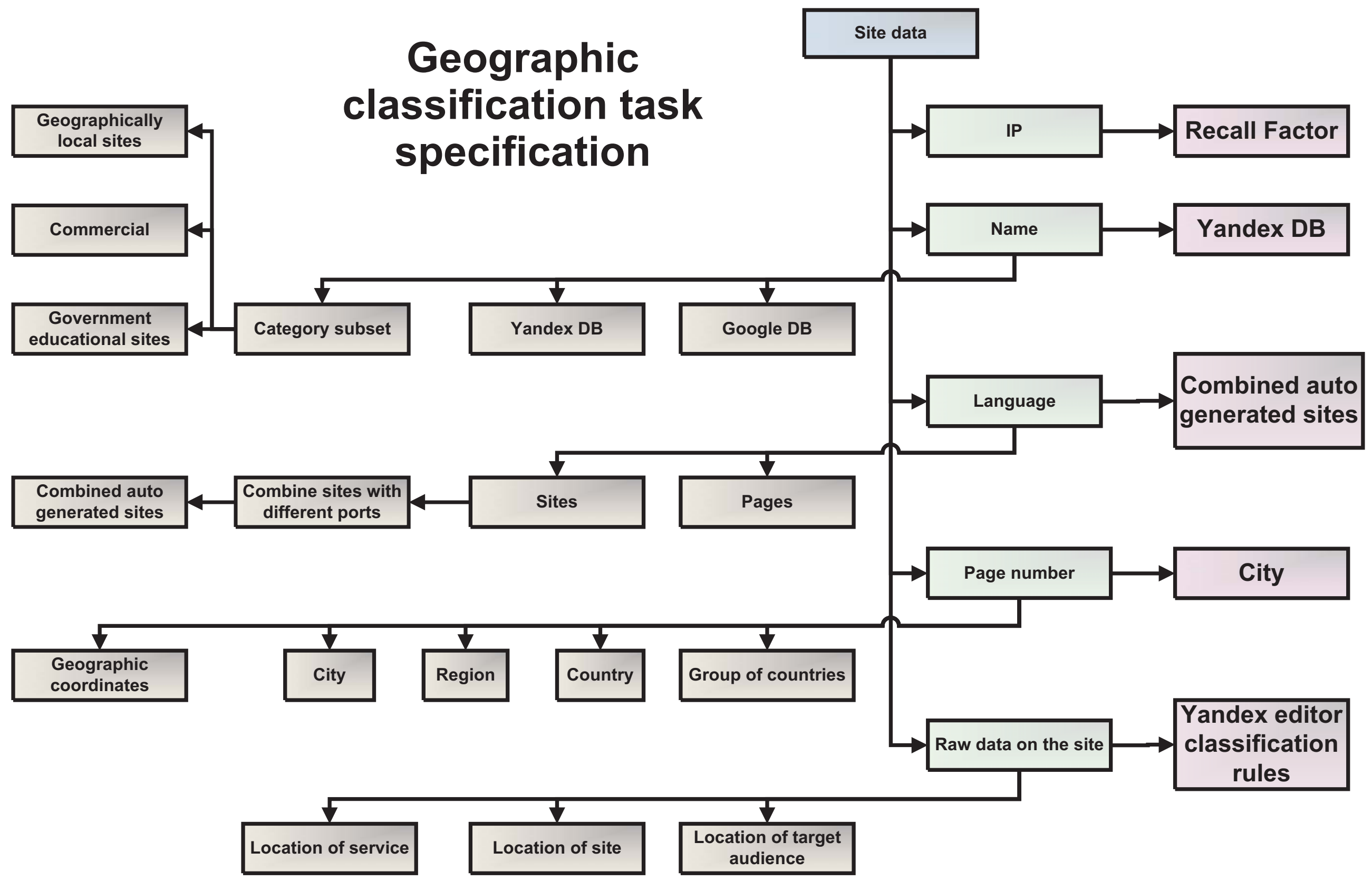
Data used for classification



Classification methods



Geographic classification task specification



RESULTS

In the current work a fast and accurate algorithm was constructed, capable to make regular updates of geographical information of search engine database sites.

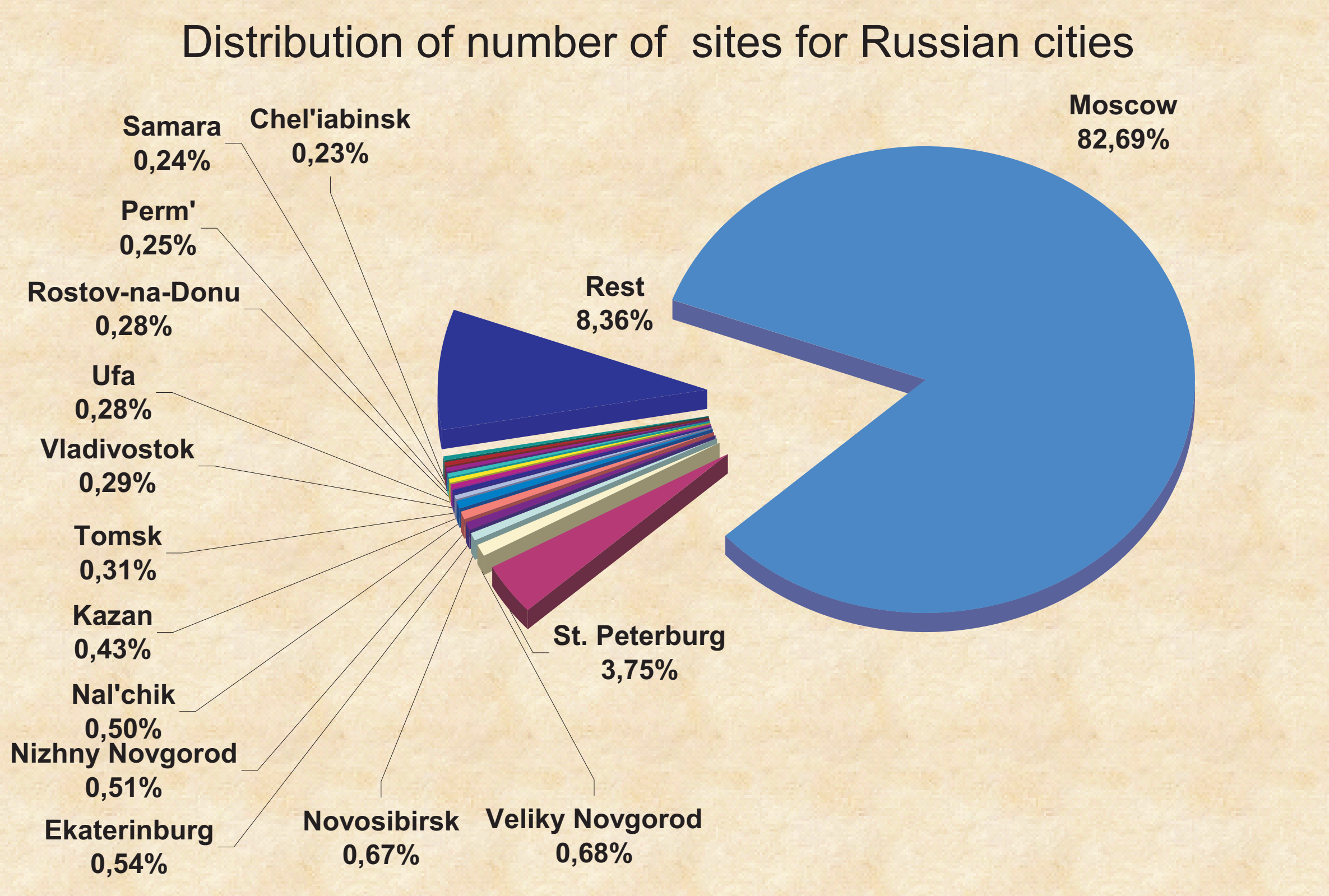
The novelty of the approach is in the usage of information about the distance between sites in the IP and name spaces, for geotag propagation. This allowed to triple the number of the resolved sites, without any loss in accuracy.

It was shown, that the task of ZIP - telephone codes extraction from the sites, does not necessarily need the extensive analyses of the page information. It can be done using standard search engine queries. This allowed to associate 30% of Yandex DB sites with the accuracy 90%.

The quality of geotagging for high referenced sites from Yandex DB can be estimated as recall factor ~ 90%, precision ~95%.

To test overall performance of the algorithm from Yandex database were randomly chosen 1200 sites. They were classified by the algorithm. The result was compared with classification of the same set done by the editors.

	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



1 623 619	Total
1 392 091	Determined
1 151 082	Moscow
52 223	Peterburg
9 421	Veliky Novgorod
9 260	Novosibirsk
7 564	Ekaterinburg
7 149	Nizhny Novgorod
7 024	Nal'chik
5 961	Kazan
4 289	Tomsk
3 848	Ufa
3 846	Rosotv-na-Donu
3 494	Perm'
4 052	Vladivostok
3 274	Samara
3 250	Chel'iabinsk
2 680	Irkutsk
2 632	Voronezh
2 420	Habarovsk
2 011	Yaroslavl'

BIBLIOGRAPHY

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