Technology for Extracting Geographical Names from Text Documents Based on the PostgreSQL

© Oleg Zhizhimov
Institute of Computational Technologies of SB RAS,
Novosibirsk, Russia
zhizhim@mail.ru

Abstract. Extracting geographical names from arbitrary text documents is important in the tasks of processing large arrays of documents and linking their content to a specific geographic region. In the simplest form, the model for extracting geographical names from the text looks like a sequence of actions with the text, while at each stage its task is solved. Among these tasks, there are undoubtedly: text parsing, analyzing text elements, processing synonyms and abbreviations, bringing the text elements to normal form from possible word forms and grammar rules, comparing text elements with the elements of dictionaries of geographical names, adding special tags to the text for unambiguous identification geographical names. The proposed work describes a technology that implements the above tasks on the basis of a freely distributed PostgreSQL DBMS. In this case, the standard configuration is used, all the server part settings are performed within the framework of the documented procedures. GeoNames Gazetteer database, Open Street Map (OSM) databases, OKATO and КЛАДР classifications are used as an authoritative database of geographical names.

Keywords: geographical names, full-text search, model of extraction of names, text processing, PostgreSQL, geographical search.

1 Introduction

The purpose of this work is to create a model for extracting geographical names from arbitrary text with its indexing by geographic attributes, for example, by geographical coordinates, with the possibility of further organizing the geometric search.

It should be noted that the existing software systems for accessing textual information resources do not have the necessary functionality for storing and processing geographic data. The provision of their required functionality is complicated by the lack of uniform standards for the search and presentation of data related to the geographical aspect that would be associated with existing geographic information systems (GIS), that is, with systems for which the geographic aspect of information is the main [1]. Hence the relevance and prospects of creating a technology that provides processing of geographic information in "non-geographic" general information systems [2].

2 Model and Algorithms

If you very briefly describe the proposed model of fixing geographic content in a text data array for subsequent indexing, it will look like this.

- The first thing to do when processing an arbitrary text is to disclose all the abbreviations. In the text, the abbreviations for their unabridged values are replaced. This procedure is essential for further analysis, because in the texts, geographical names are usually accompanied by abbreviated notation of the type of geographic object. This requires not only a simple mechanical substitution of values in accordance with the reduction dictionary, but also an analysis of the accompanying content. In particular, the abbreviation «г.» can be perceived only as «год», but also as «город», depending on the surrounding words. The formalized rules, according to which the abbreviations are disclosed, form a special dictionary of abbreviations.
- The text obtained as a result of the above procedure is divided into separate words (tokenization) with the fixation of the sequence number of each word in the source text. It also removes the stop words defined in the special dictionary and brings the rest of the words to normal form in accordance with the morphological vocabulary, which can reduce many different linguistic forms of the word to one lexeme.
- The next desired, but not mandatory, step is the disclosure of the transfers. The fact is that in different texts there are often various enumerations of geographical names with a group indication of the type of object. For example, the text "... studies were conducted in the Novosibirsk, Kemerovo and Omsk regions" for unambiguous fixation of geographical objects requires its transformation to the form "... studies were conducted in the Novosibirsk region, the Kemerovo region and the Omsk region".
- After completing the above procedures, you can fix geographic objects - assign special labels to the appropriate word combinations or replace the corresponding combination of words with a special

---

Proceedings of the XX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2018), Moscow, Russia, October 9-12, 2018

label. The first option is necessary in case of further text indexing for both geometric and full-text search, and the second one is for indexing geographic objects only for geometric search. A special label can be a unique identifier of a geographic object in a database of geographical names. Formally, the whole procedure is reduced to the replacement of normalized lexemes by special labels with an object identifier or to labels with lexemes. The correspondence of lexemes and labels is contained in a special geographic dictionary.

- Finally, the last step is to solve the problem of polysemy of geographic names. For example, more than 40 geographical objects (based on [5]) can be placed in a well-defined form of the "Советский район". However, among all possible it is necessary to choose the one that best matches the surrounding context. There are several possible solutions to the conflict:
  - On the basis of hierarchical relationships, the decision to identify an object among the competing ones is taken on the basis of an analysis of the hierarchical links of the fully-identified objects adjacent to the text. Hierarchical relations (administrative subordination, geographic location, etc.) are generally present in geographic names databases. Moreover, object identifiers of some databases store this hierarchy in the value of the identification code, for example, the OKATO directory [3]. In particular, for the city of Karasuk, the OKATO code 50217501 contains information about the Karasuk district (OKATO 50217000) and the Novosibirsk region (OKATO 50000000).
  - On the basis of geometric parameters - the decision to identify an object among the competing ones is taken on the basis of minimizing the distance to the completely identified objects next to the text. The distance is calculated based on the coordinates of the objects present in the geographic names database. In this case, different versions of the decision criterion are possible.

The algorithm for fixing geographic objects in arbitrary text is shown in Figure 1.

3 Reference books and dictionaries

The listed information resources contain the source data on the basis of which the own database of geographical objects described below is formed.

- OKATO - All-Russian classifier of objects of administrative-territorial division [3].
- KLADR - address classifier of the Russian Federation [4].
- GeoNames is a database containing over 10 million geographical names and information about more than 7.5 million of their unique characteristics [5]. Among the characteristics: the names of places in various languages, latitude, longitude, altitude above sea level. All of these characteristics are categorized, so that each characteristic of a geographic feature belongs to one of nine classes. And each of these categories, in turn, is divided into subcategories, the total number of which is more than 600. In addition to names in different languages are stored the geographical coordinates, height above sea level, population, administrative subdivision and postal codes. Unfortunately, the database contains duplicates, errors in names and other inaccuracies.
- The OSM (Open Street Map) [6] database is an open database of geographic features that includes their geometric and geographic characteristics.
- Getty the geographic names thesaurus (TGN) [7] - contains geographic names with point coordinates, including retrospective ones. The lack of Russian names are given in transcription.
- State catalog of geographical names ROSREESTR [8] - contains a complete register of official geographical names by region with point coordinates.

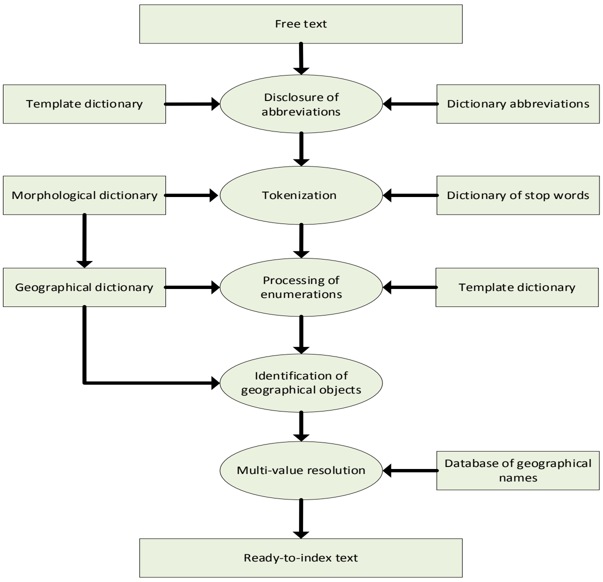
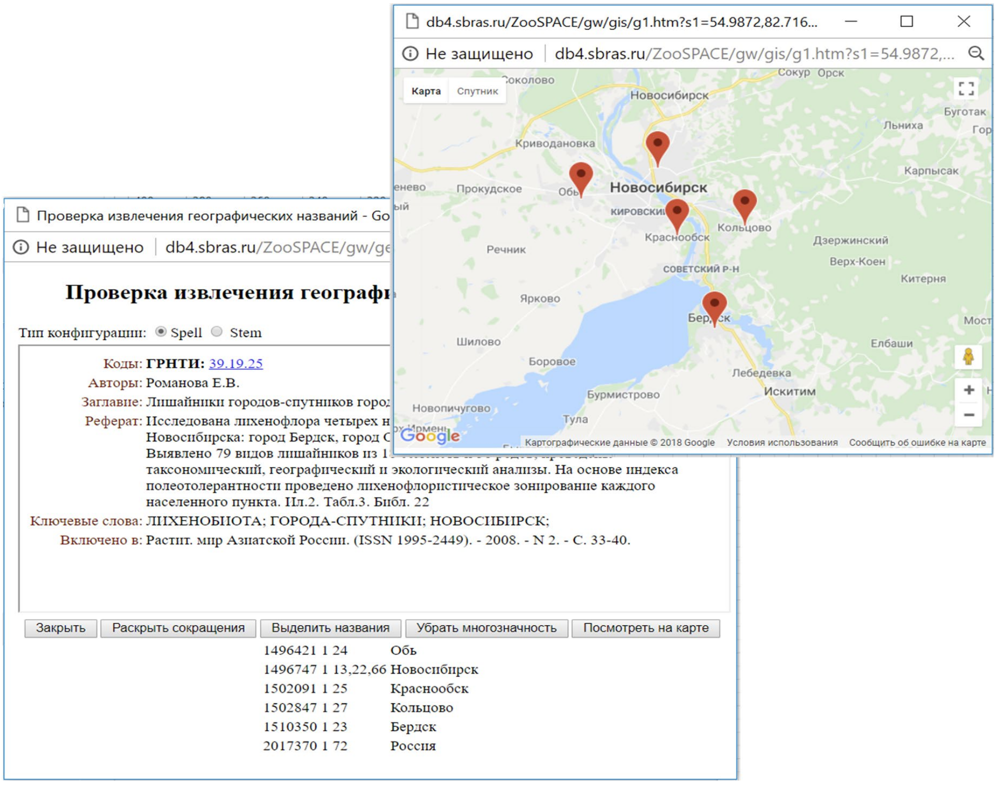


Figure 6. The algorithm for fixing geographic objects in arbitrary text

4 The prototype of the stand

For working out of technology of extraction of geographical names from texts, carrying out testing of algorithms and collecting information on errors the program stand in which the algorithms described above are realized was created.

As a system basis for the implementation of algorithms was chosen on the basis of DBMS


db4.sbras.ru/ZooSPACE/gw/gis/g1.htm?s1=54.9872,82.716...

Не защищено | db4.sbras.ru/ZooSPACE/gw/gis/g1.htm?s1=54.9872,...

Проверка извлечения географических названий - Go

Не защищено | db4.sbras.ru/ZooSPACE/gw/ge

Проверка извлечения географи

Тип конфигурации: ● Spell ○ Stem

Коды: **ГРНТИ:** 39.19.25
Авторы: Романова Е.В.
Заглавие: Лишайники городов-спутников город
Реферат: Исследована лихенофлора четырех н
Новосибирска: город Бердск, город С
Выявлено 79 видов лишайников из 1
таксономический, географический и экологический анализы. На основе индекса полеотолерантности проведено лихенофлористическое зонирование каждого населенного пункта. Ил.2. Табл.3. Библ. 22
Ключевые слова: ЛИХЕНОБИОТА; ГОРОДА-СПУТНИКИ; НОВОСИБИРСК;
Включено в: Растит. мир Азиатской России. (ISSN 1995-2449). - 2008. - N 2. - С. 33-40.

```
1496421 1 24        Обь
1496747 1 13,22,66 Новосибирск
1502091 1 25        Краснообск
1502847 1 27        Кольцово
1510350 1 23        Бердск
2017370 1 72        Россия
```

Figure 7. The application interfaces for testing the algorithms

PostgreSQL, which implements a full cycle of processing of text information with the ability to expand the basic functionality both through additional dictionaries and configurations, and writing additional modules in different programming languages [9].

The created prototype of the stand includes:

- A set of web server applications (PHP scripts) that run on the WEB server side. These applications communicate with the PostgreSQL database server and client applications. A separate server application is also a module for ZooSPACE [10] that allows you to analyze text data extracted from various bibliographic databases.
- A set of web client applications (Java scripts) that run on the WEB client side. These applications implement graphical user interface (GUI) functions to control the operation of the stand and to visualize the geographic features found on maps.

To ensure the operation of the stand created

1. **Dictionaries:**

- Dictionary abbreviations with templates based on regular expressions-using this dictionary reveals abbreviations in the input text (step 1).
- The stop word dictionary of the Russian language (russian.stop). This dictionary is included in PostgreSQL delivery and has not been changed in our configuration (step 2).
- Morphological dictionary of Russian language (ispell) with addition of geographical names and spelling rules for these names (ru_geo1.dict). A fragment of the file ru_geo1.dict:

```
. . .
абажур/K
. . .
Кольцово/M
Мошковский/A
Новосибирск-Южный/AEZ
. . .
```

- The geographical dictionary to replace the token for the combination of “label+token”. This dictionary (geor1.ths) corresponds to the thesaurus template (in terms of PostgreSQL thesaurus is a dictionary of substitutions: the left part of the symbol ":" is replaced by the right part, the presence of the symbol "*" in the first position of the right part prescribes not to control the right part of the morphological dictionary) and consists of:

```
. . .
Бердск: */gn/1510350
город Бердск: */gn/1510350
Советский район: */gn/490026, /gn/1491227
. . .
```

2. **Configuration FPS** (in terms of PostgreSQL) that defines a list of dictionaries and the order of processing of the text (rugeo1):

```
CREATE TEXT SEARCH DICTIONARY
rugeo_ispell (TEMPLATE = ispell,
 dictfile = 'ru_geo1', afffile =
'ru', stopwords = 'russian');
```

```
CREATE TEXT SEARCH DICTIONARY
tz_geo_1 (TEMPLATE = thesaurus,
 dictfile = 'geor1', dictionary =
'rugeo_ispell');

CREATE TEXT SEARCH CONFIGURATION
rugeo1 (PARSER = "default");

ALTER TEXT SEARCH CONFIGURATION
rugeo1 ADD MAPPING FOR hword WITH
tz_geo_1, rugeo_ispell, russian_stem;

ALTER TEXT SEARCH CONFIGURATION
rugeo1 ADD MAPPING FOR hword_part
WITH tz_geo_1, rugeo_ispell,
russian_stem;
```

The work of the algorithm for fixing geographical names can be illustrated by the example of processing a fragment of the text "В окрестностях города Новосибирска находятся: город Бердск, город Обь, поселок Краснообск и Наукоград Кольцово". As a result of query execution

```
SELECT plainto_tsquery('rugeo1', 'В
окрестностях города Новосибирска
находятся: город Бердск, город Обь,
поселок Краснообск и Наукоград
Кольцово');
```

get a response - marked-up text

```
'окрестность /gn/1496747 город новосибирск
находиться   /gn/1510350   город   бердск
/gn/1496421 город обь /gn/1502091 поселок
краснообск /gn/1502847 наукоград кольцово'
```

Other request

```
SELECT to_tsvector('rugeo1','В
окрестностях города Новосибирска
находятся: город Бердск, город Обь,
поселок Краснообск и Наукоград
Кольцово');
```

returns a list of tokens indicating their position in the text

```
'/gn/1496421':10 '/gn/1496747':3
'/gn/1502091':13 '/gn/1502847':17
'/gn/1510350':7 'бердск':9 'город':4,8,11
 'кольцово':19 'краснообск':15
'наукоград':18 'находиться':6
'новосибирск':5 'обь':12 'окрестность':2
'поселок':14
```

5 Conclusion

As a result of the work performed, a stand prototype was created for testing models and algorithms for extracting geographical names from unstructured text to build indexes for both text and geometric searches. Preliminary testing showed that the proposed technology provides a high degree of reliability of the results, provided that all directories contain information about the identified geographical features. The effectiveness of the technology depends on the completeness of the reference books.

Currently, the created directories contain information on geographical objects of the Novosibirsk region. In the future, it is planned to expand the range of supported regions.

Work is performed within the Integration Project of SB RAS (AAAA-A18-118022190008-8), Project for basic scientific research (AAAA-A17-117120670141-7) and RFBR project № 18-07-01457-a.

References

[1] Zhizhimov O.L., Mazov N.A. Problems of geographical reference of digital objects in digital libraries. Proc. XII All-Russian Sci. Conf. «Electronic libraries: Perspective Methods and Technologies, Electronic collections» (RCDL'2010). Kasan, p. 207–214. (2010).

[2] Barakhnin V.B., Zhizhimov O.L., Kupershtokh A.A., Skachkov D.M., Fedotov A.M. The Algoritm of Exstracting Place Names Representing Content from Text Documents. Vestnik NSU. Ser.: The Information technology, Vol.10, Iss.1, p.109-120. (2012).

[3] All-Russian classifier of administrative-territorial division objects (OKATO), http://protect.gost.ru/document.aspx?control=20&id=134377.

[4] Classifier of addresses of the Russian Federation (CLADR), http://kladr-rf.ru.

[5] The GeoNames geographical database. - http://www.geonames.org/.

[6] Open Street Map, http://wiki.openstreetmap.org.

[7] Getty Thesaurus of Geographic Names (TGN), - http://www.getty.edu/research/tools/vocabularies/tgn/index.html.

[8] State catalogue of geographical names, Rosreestr. - https://rosreestr.ru/site/activity/geodeziya-i-kartografiya/naimenovaniya-geograficheskikh-obektov/gosudarstvennyy-katalog-geograficheskikh-nazvaniy/.

[9] Bartunov J., Sigaev F. Introduction to full-text search in PostgreSQL, - http://citforum.ru/database/postgres/fts/bib.shtml.

[10] Zhizhimov, O.L., Fedotov, A.M., Shokhin, Y.I. The ZooSPACE platform- access organization to various distributed resources. Digital libraries: The Russian scien-tic e-magazine. - Vol.17. – Iss. 2. - ISSN 1562-5419 (2014).