УДК 622:528.9:912:004
DOI: 10.25635/2313-1586.2020.01.113

Суханов Владимир Иванович
доктор технических наук, доцент, профессор
Центра ускоренного обучения,
Институт радиоэлектроники
и информационных технологий
Уральского федерального университета,
620002, Екатеринбург, ул. Мира, 19
e-mail: cyx-fat@mail.ru

Sukhanov Vladimir I.
Doctor of Technical Sciences,
Professor of Center for Accelerated Learning,
Institute of Radioelectronics
and Information Technologies,
Ural Federal University,
620002 Ekaterinburg, 19 Mira Str.
e-mail: cyx-fat@mail.ru

КОМПЬЮТЕРНЫЙ АНАЛИЗ СОДЕРЖАНИЯ ГОРНОЙ ЭНЦИКЛОПЕДИИ

Аннотация:

Рассмотрены задачи разработки отечественного свободного программного обеспечения горно-геологических информационных систем с открытым кодом для предприятий горной промышленности. Приведены основные сведения о программной реализации и результатах статистической обработки статей Горной энциклопедии для отбора базовых терминов при создании онтологий горного производства. Базовым программным обеспечением выбраны кросс-платформенная СУБД PostgreSQL, язык Python 3, пакеты pymorphy2, matplotlib. Инструментальные средства позволяют реализовать необходимые технологические операции чтения статей энциклопедии, выделение информации об их именах, ссылках на другие статьи, накопление статистики в базе данных, формирование графического представления результатов и другие операции.

Ключевые слова: онтологии, горное производство, геоинформатика, программирование, статистики, PyQT, PostgreSQL

COMPUTER CONTENT ANALYSIS OF THE MOUNTAIN ENCYCLOPEDIA

Abstract:

The paper considers development roblems of the domestic free software of mining and geological information systems with an open code for mining industry enterprises. The basic information about the software implementation and the statistical processing results of the Mining Encyclopedia articles to fulfill the selection of basic terms while the creation of ontologies of mining. The basic software were: cross-platform database PostgreSQL, Python 3, packages pymorphy2, matplotlib. Instrumental tools allow us to implement the necessary technological operations of reading articles of the encyclopedia, allow the allocation of information about their names, links to other articles, the accumulation of statistics in the database, the formation of a graphical representation of the results, and other operations

Keywords: ontology, mining, geoinformatics, programming, statistics, PyQT, PostgreSQL.

Введение

Переход к цифровой организации предметных знаний горной промышленности для полномасштабного внедрения методов интеллектуальных технологий требует изучения внутренней структуры знаний, их представления на машинно-читаемом языке. Значимым направлением представления знания являются онтологии [1] как множество понятий и утверждений об этих понятиях, на основе которых можно формализовать представление знаний конкретной предметной области независимо от использования этих знаний, описывающих классы, отношения, функции и индивиды. Успешное применение онтологий в горной промышленности иллюстрируется рядом работ, выполненных в основном зарубежными исследователями [2 – 4].

Создание частных онтологий – трудоемкий процесс, требующий участия носителей знаний о предметной области, в совершенстве знающих структуру и причинно-следственные связи между понятиями и отношениями между ними. Не исключая участия специалистов-экспертов, можно воспользоваться накопленными за десятилетия знаниями, аккумулированными в учебной и справочной литературе. На роль таких изданий подходят энциклопедические словари, в составлении которых принимают участие профессионалы – ученые и практики. Что касается горной промышленности в сети Интернет

в открытом доступе существуют ресурсы [5, 6], являющиеся электронными изданиями как официальных разделов Большой советской энциклопедии, так и отраслевых справочников.

Для автоматизированного построения онтологий на основе текстов статей горной энциклопедии нужно решить следующие задачи:

1) выбор архитектуры программно-аппаратного комплекса;

2) выбор базового программного обеспечения и средств разработки;

3) выделение полезной информации из содержимого электронного документа статьи энциклопедии;

4) навигация между статьями по ссылкам в документах;

5) морфологический, синтаксический и семантический анализ текстов статей для извлечения понятий и отношений между ними;

6) выбор диалоговых и программных интерфейсов взаимодействия с пользователями и смежными информационными системами;

7) программная реализация модулей и интерфейсов.

Программная реализация модулей и интерфейсов — сложный, затратный эволюционный процесс с привлечением большого числа специалистов: аналитиков, кодировщиков, тестировщиков, менеджеров и программистов.

Анализ средств разработки

Open Geospatial Consortium (OGC) – международная добровольная организация по разработке стандартов и рекомендаций в области геоинформационных сервисов [7], которой подготовлены разнообразные стандарты и рекомендации по представлению и манипулированию геопространственными данными, на основе которых разработано большое количество общесистемного и прикладного программного обеспечения с открытым кодом, благодаря которому разработка доверенных горно-геологических систем для открытой разработки месторождений посильна для небольших коллективов разработчиков.

Рекомендация [8] является частью материалов по разработке семантического Веб, описывающего язык представления онтологий. Язык веб-онтологий OWL призван обеспечить представление, которое может быть использовано для описания классов и отношений между ними, присущих веб-документам и приложениям. Этот документ демонстрирует, как использовать язык OWL, чтобы

– формализовать область определения классов и свойства этих классов;

– определить индивиды и назначить их свойства;

– уточнить классы и индивиды в соответствии с формальной семантикой OWL.

Наиболее сложной частью проекта являются инструменты анализа текстов русского языка. В открытом доступе есть описания проблем и реализаций парсеров русского языка [9, 10, 11] с открытым кодом. Предлагаемая в [10] технология основана на собственных словарях русского языка, поставляемых вместе с пакетом программ. Технология [11] использует словари, построенные на основе открытого корпуса русского языка OpenCorpora [12], что делает ее более предпочтительной.

Для реализации этих технологий предлагается использовать следующее программное обеспечение: для работы с базами данных – СУБД PostgreSQL, для администрирования БД и просмотра данных – инструмент PgAdmin, для программирования модулей проекта целесообразно использовать язык программирования Python 3 и пакеты *psycopg2, pymorphy2, matplotlib* [13 - 15].

Структура словарной статьи

Горная энциклопедия является открытым ресурсом, размещенным в сети Интернет по адресу http://www.mining-enc.ru/. Словарные статьи представлены страницами, связанными между собой ссылками, которые используются читателем при просмотре вложенных разделов гипертекста. Текст словарной статьи энциклопедии содержит большое количество тегов языка HTML для размещения различного рода вспомогательной информации (структуры документа, заголовков, рекламы, меню, реквизитов издания и др.) и содержательной информации, размещенной в тегах — структурных разделах документа, например

* КАРЬЕР … добыче полезных ископаемых … в земной коре, … *

В приведенном фрагменте тег *<img>* выполняет вставку в документ изображения, задаваемого атрибутом *src*. Тег *<a>* создает ссылку со страницы документа на другую статью энциклопедии. Поскольку атрибут *title* не является обязательным, при автоматическом анализе таких фрагментов возникают трудности в именовании статей-ссылок. В этом случае роль имени статьи может выполнять текст ссылки, в нашем примере — «*земной коре*».

Основные решения анализатора

Для количественного анализа связности статей энциклопедии достаточно в базе данных определить две таблицы: статей и ссылок. Таблица статей имеет следующие столбцы:

- ключ id — для внешних ссылок;
- название — атрибут *title*;
- адрес — атрибут *href*.

Таблица ссылок имеет столбцы:

- ключ главной статьи — *idfrom;*
- ключ ссылки на статью — idto.

Этой информации вполне достаточно для проведения статистического анализа связности статей энциклопедии.

Чтение статьи энциклопедии выполняется фрагментом

```
try:
   with urllib.request.urlopen(href) as response:
      html = response.read()
      mainRead(self, html.decode("utf-8"))  # Обработка через
        память ПК
except Exception:
   print('Нет текста статьи '+ href)
```

Последующий анализ текста статьи с выделением нужной информации выполняется модулем на языке Python. Разбор текста статьи выполняется стандартным «парсером» *HTMLParser* библиотеки языка Python. Парсер отыскивает фрагменты текста статьи, содержащиеся внутри тега , находит и анализирует ссылки в тегах <a> и формирует из них список ссылок. Основную работу выполняет следующий метод:

```
def handle_data(self, data):
    global lstRefs
    if self.need:
```

```
if (self.tag == 'a'):
    attr = dict(self.attrs)
    refTitle = data
    refHref = ''
    if 'href'in attr.keys():
        refHref = attr['href']
    if 'title'in attr.keys():
        refTitle = attr['title'].replace(' ','_')
    if refTitle:
        lstRefs.append((refTitle, refHref, data))
    self.tag = ''
```

При определении имени статьи предпочтение отдается содержимому атрибута *title* при его наличии. В противном случае используется атрибут *data* — текст ссылки в статье. Полученный список *lstRefs* используется для добавления записей в таблицы базы данных статей и ссылок в модуле анализа статьи.

Результаты обработки энциклопедии

Общее число статей — 4265. Общее число ссылок — 83423. Гистограммы распределения числа ссылок приведены ниже (рис. 1).

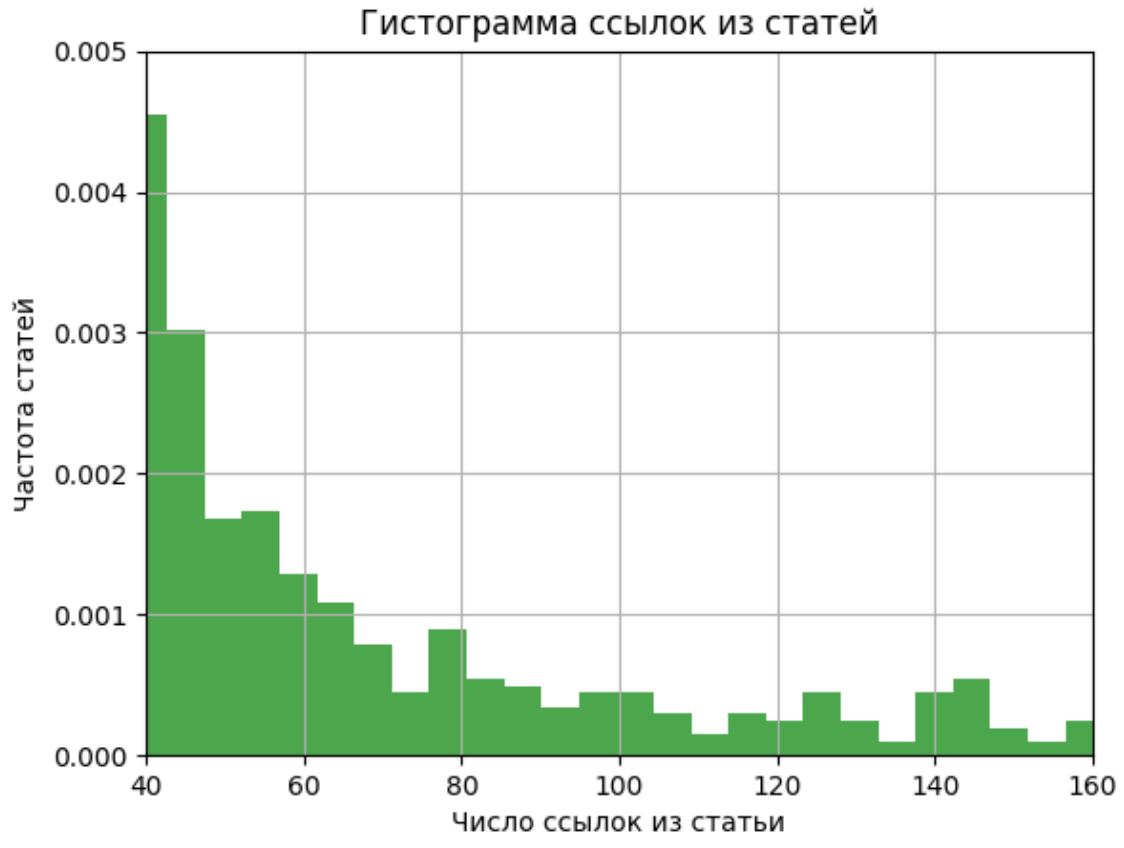

Рис. 1 – Гистограммы распределения числа ссылок

Для иллюстрации статистических параметров текстов статей ниже приведены фрагменты данных о числе ссылок из статей на другие статьи «Горной энциклопедии»: <Имя статьи> – <Количество ссылок на другие статьи>:

Индия — 332	Испании — 234	драгоценным камням — 211
Мадагаскар — 201	Европа — 195	ФРГ — 186
Полезные_ископаемые — 182	Куба — 179	Узбекская ССР — 179
Иран — 177	Филиппины — 174	Рудные_месторождения – 171
. . . . .		
Шахтная_горная_технология — 139	Мировой_океан — 138	Туркменскую ССР — 138
Индонезия -137	Канада — 135	Бразилия — 131
Земля — 131		
. . . . .		
горное дело — 94	палеозойские — 94	Рассеянных_элементов_руды – 93
Минералогия — 90	угли ископаемые — 89	Технологическая_схема_горнодобывающего_предприятия — 88
. . . . .		
Щебень — 17	Якутуголь — 17	альмандин — 17
анкерами — 17	буровыми скважинами — 17	Брикетирования – 17
. . . . .		
хризотилового — 1	цинка — 1	шашек-детонаторов — 1
. . . . .		
Геохронология — 0	Известковый_туф — 0	Кирка — 0
Линза — 0	Масштаб — 0	Минеральные ресурсы — 0

Иллюстрация числа ссылок на имеющиеся в энциклопедии статьи приведена в выборке ниже: <Имя статьи> – <Количество ссылок на эту статью>:

СССР — 1120	полезных ископаемых — 978	Минерал — 954
Горные_породы — 838	Руда — 759	Вода — 744
Полезные_ископаемые — 743	Плотность — 701	нефти — 686
шахтами — 554	США — 521	Пласт — 518
природному газу — 502	Добыча_полезных_ископаемых — 433	угля — 433
Карьер — 417	Земная_кора — 390	горных выработок — 381
. . . . .		
Бурильные_трубы — 25	Выклинивание — 25	Гвинея — 25
Думпкар — 25	Долерит — 25	Карпаты – 25
Коренная порода — 25	Концентрационный_стол — 25	Киноварь — 25
Нефтегазоносная_область — 25	Очистная_выемка — 25	Сейсмичность — 25
. . . . .		
турнодозера — 1	тектониты — 1	торфяниках — 1
торфяных болот — 1	термометр глубинный —	уваровит — 1
эксудации — 1	элеватором — 1	Экстрагирования – 1
Горная экология — 0	Зеркало скольжения — 0	Карьерная гидрогеология — 0

Заключение

Статья описывает основные средства и результаты компьютерного анализа содержимого «Горной энциклопедии», расположенной в сети Интернет. Основные выводы о проделанном исследовании следующие:

1. Горная энциклопедия является открытым ресурсом с большим объемом данных, поддающимся компьютерному анализу в силу хорошей структурированности текстов статей, подготовленных большим коллективом авторов в течение многих десятков лет.

2. Структура статей позволяет извлечь информацию о семантических связях описываемых в ней понятий и формализовать их реляционными таблицами простой структуры.

3. Отсутствие атрибута *title* в тегах ссылок на статьи в оригинальных текстах энциклопедии приводит к синтаксическому рассогласованию имен статей в базе данных при формировании ссылок с сохранением семантики. В приведенных примерах такие имена начинаются со строчной буквы.

4. Сформированные базы данных являются хорошей основой для автоматического формирования онтологий горных работ.

Литература

1. Онтологии и тезаурусы: модели, инструменты, приложения / Б.В. Добров, В.В. Иванов, Н.В. Лукашевич, В.Д. Соловьев [Электронный ресурс] - Режим доступа: http://www.intuit.ru/studies/courses/1078/270/info 12.05.2017

2. J. Du, R. He, V. Sugumaran. Clustering and ontology-based information integration framework for surface subsidence risk mitigation in underground tunnels. Cluster Comput (2016) Springer Science+Business Media New York 2016 [Электронный ресурс] - Режим доступа: http://crossmark.crossref.org/dialog/?doi=10.1007/s10586-016-0631-4&domain=pdf

3. Hu Z., Yao S., Liu Y. (2013) Research on Shared Ontology Model for Coal Mine Emergency Case. In: Lu W., Cai G., Liu W., Xing W. (eds) Proceedings of the 2012 International Conference on Information Technology and Software Engineering. Lecture Notes in Electrical Engineering. – Vol. 212. – P. 959 - 968. Springer, Berlin, Heidelberg

4. Bocharov V., Pivovarova L., Rubashkin V., Chuprin B. (2010) Ontological Parsing of Encyclopedia Information. In: Gelbukh A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2010. Lecture Notes in Computer Science. - Vol. 6008. Springer, Berlin, Heidelberg

5. Горная энциклопедия [Электронный ресурс] - Режим доступа: http://www.mining-enc.ru/. 17.06.2018

6. Горная энциклопедия [Электронный ресурс] - Режим доступа: http://basemine.ru/04/gornaya-enciklopediya/. 17.06.2018

7. Стандарты OGC [Электронный ресурс] - Режим доступа: http://gis-lab.info/wiki/Стандарты_OGC. 12.05.2018

8. OWL Web Ontology Language Guide [Электронный ресурс] - Режим доступа: https://www.w3.org/TR/owl-guide/ 17.06.2018

9. Парсим русский язык [Электронный ресурс] - Режим доступа: https://habr.com/post/148124/ 17.06.2018

10. Парсер - морфологический и синтаксический анализатор русскоязычных текстов [Электронный ресурс] - Режим доступа: http://www.solarix.ru/parser.shtml 17.06.2018

11. Морфологический анализатор pymorphy2. [Электронный ресурс] - Режим доступа: http://pymorphy2.readthedocs.io/en/latest/ 17.06.2018

12. Открытый корпус русского языка. [Электронный ресурс] - Режим доступа: http://opencorpora.org/ 17.06.2018

13. Psycopg [Электронный ресурс] - Режим доступа: http://initd.org/psycopg/. 20.09.2013

14. Морфологический анализатор pymorphy2 [Электронный ресурс] - Режим доступа: https://pymorphy2.readthedocs.io/en/latest/ 20.09.2017

15. Matplotlib: Научная графика в Python [Электронный ресурс] - Режим доступа: https://pythonworld.ru/novosti-mira-python/scientific-graphics-in-python.html. 20.09.2017