Identification of Descriptions of Scientific-Technical Effects in Patent Documents

Grigoriy Vereschak[1], Dmitriy Korobkin[1[0000-0002-4684-1011]], Sergey Fomenkov[1[0000-0001-9907-4488]], Sergey Kolesnikov[1[0000-0002-6910-7151]]

[1] Volgograd State Technical University, Lenin av. 28, Volgograd, Russia

dkorobkin80@mail.ru

Abstract. This paper describes a method of searching descriptions of scientific-technical (in particular, chemical) effects from US patent documents (USPTO). Chemical effects representations (chemical phenomena) according to the National Center for Biotechnology Information classification are used. The algorithms of primary processing of patent database, extracting key terms from the descriptions of chemical effects, extracting significant features from the text of patents, search the most relevant patents based on queries generated from descriptions of chemical effects were developed. Feature Extraction technology of Spark MLlib is used for the extraction of significant features from patents. Semantic text processing (NLP) is used to identify key features from descriptions of chemical effects and to compose a search query based on them. The search for the most relevant patents containing descriptions of chemical effects is performed base on the generated queries. The software is developed as an application for Linux-systems, its efficiency has been tested on a set of test tasks.

Keywords. Chemical phenomena, Natural Language Processing, Feature Extraction, MLlib, RDD

1 Introduction

The existing global patent database with more than 20 million can serve as a source of information [1,2] for the initial stages of designing new technical solutions [3]. One of the possible approaches to the generation of new systems are methods based on the use of scientific and technical effects [4] including chemical ones from patent documents.

A lot of creative solutions of technical problems are based on the use of chemical effects, which makes the task of automating the process of searching for descriptions of chemical effects in English-language (the largest part of the world patent database) patent documents. Such chemical effects can be biocatalysis, halogenation, etc., which are related to biochemical effects according to the classification of the National Center for Biotechnology Information [5].

The United States Patent and Trademark Office (USPTO) [6] provides free access to file storage patent, which contains zip archives with XML files (patent text) and

In: P. Sosnin, V. Maklaev, E. Sosnina (eds.): Proceedings of the IS-2019 Conference, Ulyanovsk, Russia, 24-27 September 2019, published at http://ceur-ws.org

related TIFF documents (images). In this paper, USPTO Bulk Download storage is used as a source of patent descriptions due to free access to information and convenience of XML format for parsing procedure.

It was decided to use the technology of distributed computing [7] because there is a need to process large amounts of information (hundreds of thousands and millions of patents).

It uses a distributed scalable file system (HDFS) [8], which designed to work with Big Data and provides high bandwidth access to data.

Used Spark MLlib [9] - a library of machine learning methods, supplied with the implementation of the algorithm required to solve the problem of extracting key features of the text. The RDD scheme [10] is involved as the main concept of Spark, which provides processing an arbitrary collection of objects as in a relational table. It can be distributed in memory, on disk or be completely virtual and it provides fast and scalable parallel data processing.

2 The developed methods

2.1 Algorithm of primary processing of patent database

The patent database downloaded from USPTO Bulk Downloads and containing full-text descriptions of patents with images is a tar archive file. There are several directories inside the main archive. Catalogs with patents are divided into three types: Design, Util, Plant. Inside each of these directories are patents archived into zip files. Inside each zip file, there is an XML file containing the text of the patent, TIFF images, and possibly chemical formula presentation files.

It is necessary to perform a recursive search of the necessary zip archives in all directories and subdirectories before parsing XML files to extract patent texts. If the archive name does not match the required template, which is set in the following format: US********-########, where US is the country, ******** - document number, ######## - date of publication, such archive is filtered. The processing of the primary patent database results in a directory containing the XML files of patent documents. The catalog containing the XML-files of patent documents is obtained as a result of the initial processing of the patent database.

Then the parsing of the XML files begins to retrieve the description of the patent document. The XML document is scanned for the presence of <claim-text> tags, which contains patent claims. As a result of the preprocessing the patent database, key-value pairs are loaded into RDD, where the key is the number of the patent document and the value is the list with the description of the patent document.

Figure 1 shows the algorithm for preprocessing a patent database.

2.2 Algorithm for extracting key terms from the descriptions of chemical effects

Compiling a search query to a patent database is based on key terms that are extracted from descriptions of chemical effects according to the classification of the National Center for Biotechnology Information by means of semantic analysis [5]. Stanford 3

NLP software was used to extract key terms from a textual description of the effects [11].

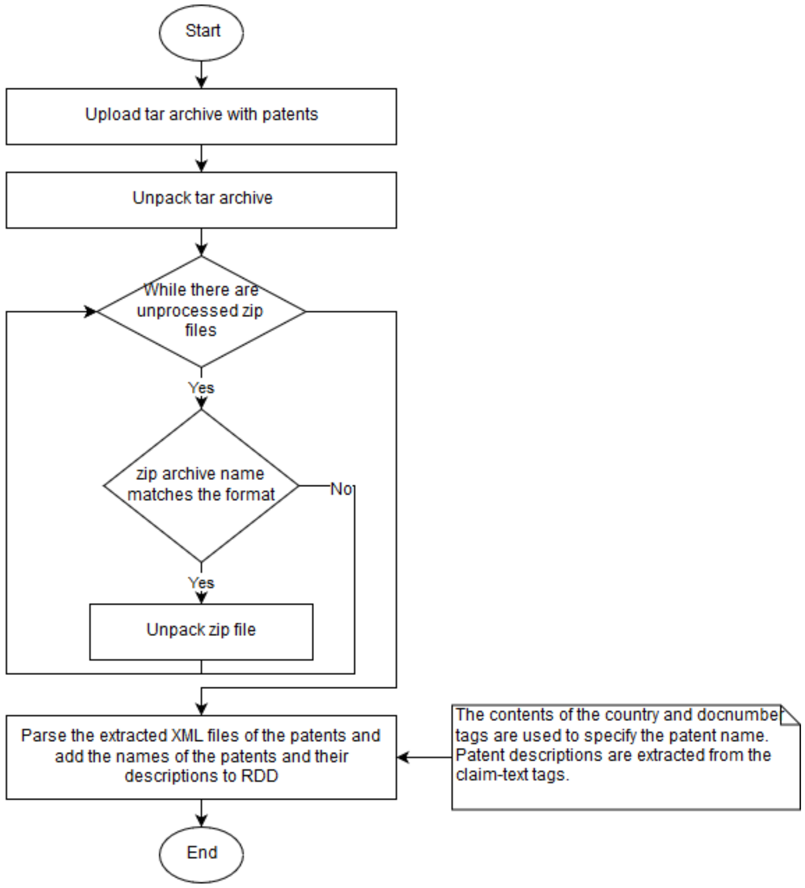
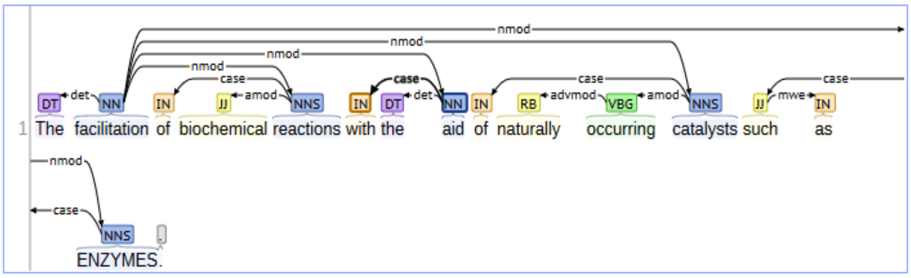

Start
Upload tar archive with patents
Unpack tar archive
While there are unprocessed zip files
Yes
zip archive name matches the format
No
Yes
Unpack zip file
Parse the extracted XML files of the patents and add the names of the patents and their descriptions to RDD
The contents of the country and docnumber tags are used to specify the patent name. Patent descriptions are extracted from the claim-text tags.
End

Fig. 1. Algorithm of primary processing of the patent database

It is necessary to identify only certain relationships in the sentence for the search for keywords: nmod, amod, advmod (nmod is a noun that plays the role of a supplement (nominal modifier); amod is an adjective or a verbal entity that acts as a definition; advmod is an adverb, plays the role circumstances).

The Stanford NLP parser accepts a natural language sentence as input and returns the semantic relationships found between words in the sentence (Figure 2). Each such relationship consists of the index of the main word (token), the index of the dependent token, and the type of relationship.

The token which is a dependent word in the found relationship is considered a key term because only the semantic relationships given above are considered. The name of the processed chemical effect and the key terms found are added to the chemical effects database. The algorithm for extracting keywords from descriptions of chemical effects is shown in Figure 3.


1 The facilitation of biochemical reactions with the aid of naturally occurring catalysts such as
ENZYMES.
DT NN IN JJ NNS IN DT NN IN RB VBG NNS JJ IN NNS
det
nmod
case
amod
advmod
mwe

Chemical effect: «Biocatalysis».
Chemical effect description: «The facilitation of biochemical reactions with the aid of naturally occurring catalysts such as ENZYMES».
Key Terms: «facilitation», «biochemical reactions», «naturally occurring catalysts», «enzymes».

Fig. 2. The semantic relationships between words in the sentence

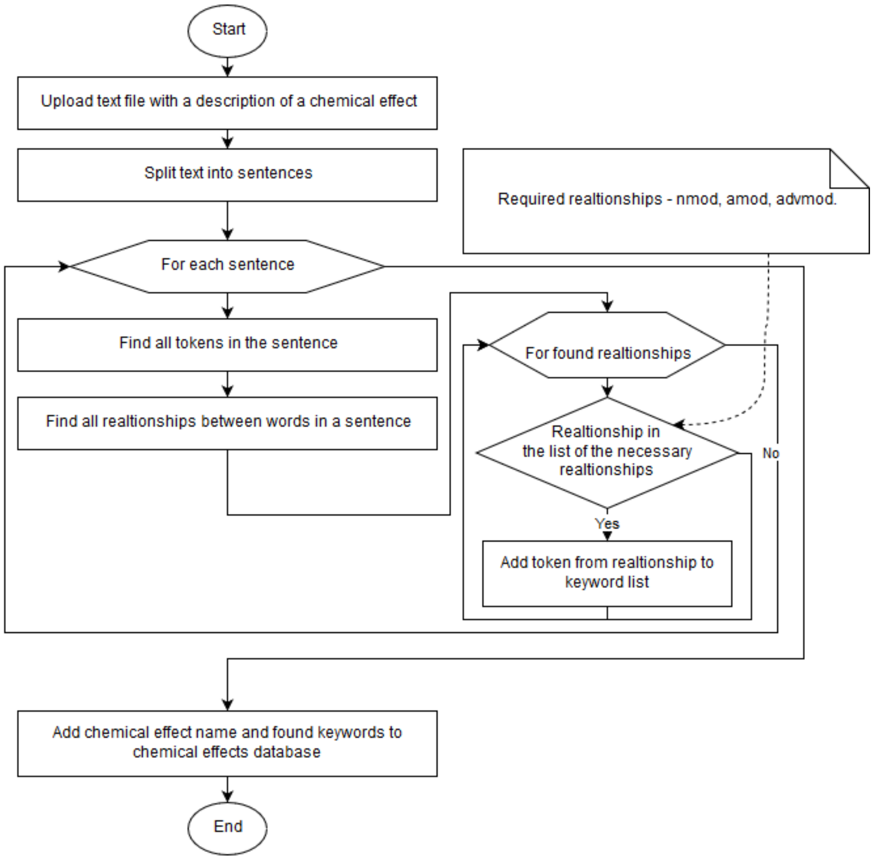

Start
Upload text file with a description of a chemical effect
Split text into sentences
Required realtionships - nmod, amod, advmod.
For each sentence
Find all tokens in the sentence
Find all realtionships between words in a sentence
For found realtionships
Realtionship in the list of the necessary realtionships
No
Yes
Add token from realtionship to keyword list
Add chemical effect name and found keywords to chemical effects database
End

Fig. 3. Algorithm for extracting key terms from descriptions of chemical effects.

2.3 Algorithm for extracting significant features from the text of patents

The Feature Extraction technology (technology of extracting significant features) from MLlib is the basis of the algorithm, namely the TF-IDF algorithm. It is necessary to prepare the input data before you begin to identify significant features. The input data for this algorithm are the patent descriptions stored in RDD. Each record in RDD is a key-value pair, where a key is a patent number, value is a list containing patent descriptions. First of all, you need to combine descriptions into a single text and convert RDD to another data type - Data Frame.

TF - IDF (word frequency - inverse frequency) is a feature vectorization technique widely used in text mining that reflects the importance of a term for a document in a document corpus. The TF-IDF algorithm [9] involves splitting the source text into tokens. For each patent description punctuation is removed, the case of words is aligned. The output is a table containing the patent number and a list of words included in its description after applying the tokenization procedure. It is necessary to remove stop words (prepositions, conjunctions, particles, pronouns, introductory words) to reduce the list of tokens and improve speed and efficiency. Stop words have the lowest IDF values.

The CountVectorizer provided by MLlib is used to find TF values instead of HashingTF. HashingTF works twice as fast but does not allow you to access the words in the list of tokens by index. The MLlib tools then calculate the IDF and TF-IDF measures. The output is a vector representation of tokens and measures of their significance. Figure 4 shows an algorithm for extracting significant features from the text of patents.

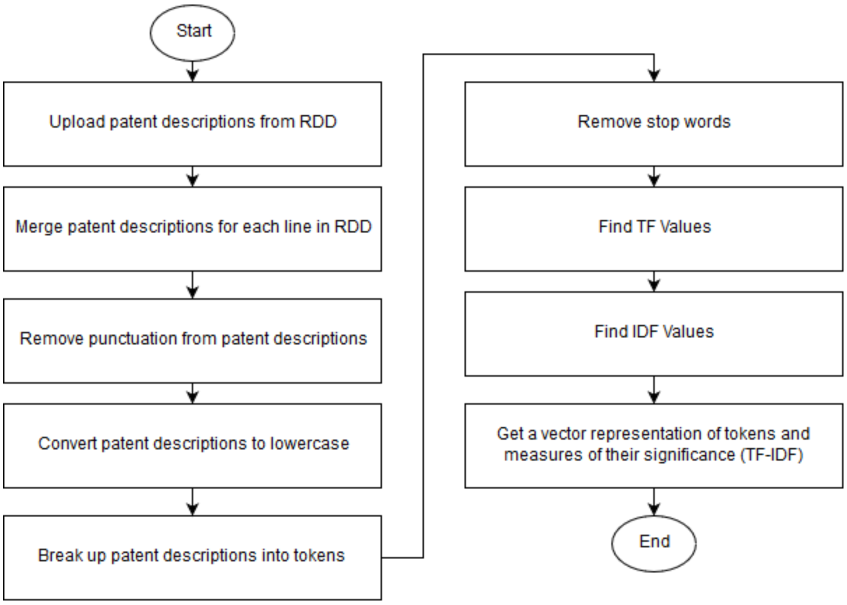

Start
Upload patent descriptions from RDD
Merge patent descriptions for each line in RDD
Remove punctuation from patent descriptions
Convert patent descriptions to lowercase
Break up patent descriptions into tokens
Remove stop words
Find TF Values
Find IDF Values
Get a vector representation of tokens and measures of their significance (TF-IDF)
End

Fig. 4. Algorithm for extracting significant features from the text of patents

2.4 Algorithm for search the most relevant patents based on queries generated from descriptions of chemical effects

The search for relevant patents from the patent database loaded into RDD is performed based on a search query consisting of key terms extracted from the texts of chemical effects. Filtering is used to distinguish relevant patents from the General patent database, where the main condition of the filter is the presence of key terms of the desired chemical effect in the key features of the patent under consideration. For ranking the list of found patents a measure of relevance is introduced, which is defined as the ratio of the number of matched key terms to the total number of key terms from the query.

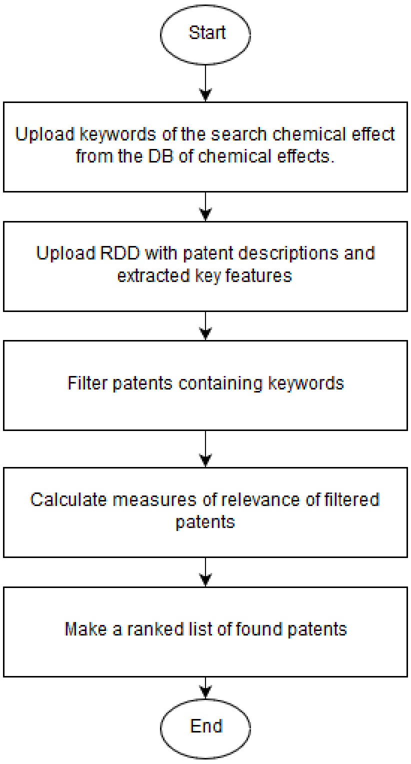
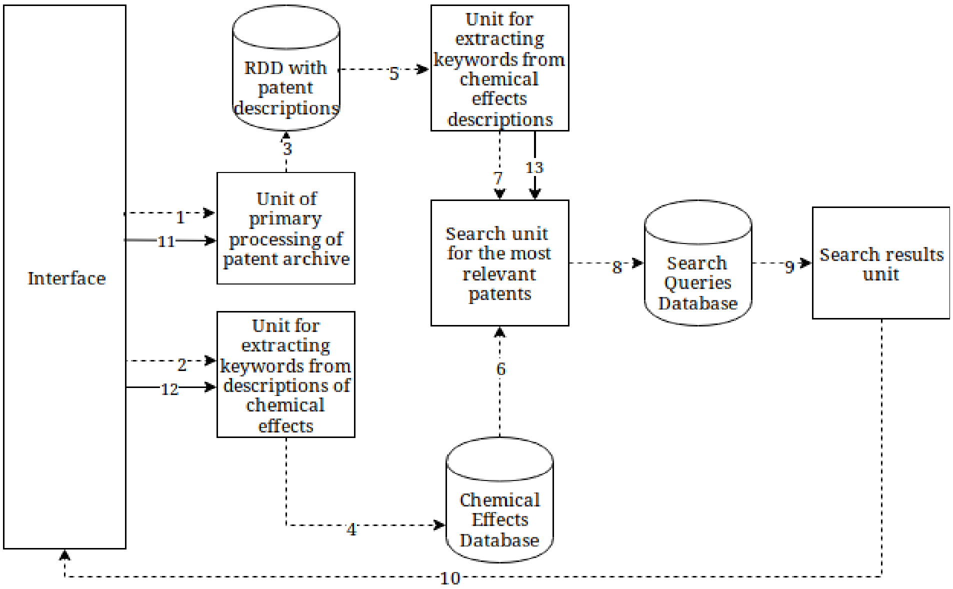

Start
Upload keywords of the search chemical effect from the DB of chemical effects.
Upload RDD with patent descriptions and extracted key features
Filter patents containing keywords
Calculate measures of relevance of filtered patents
Make a ranked list of found patents
End

Fig. 5. The search algorithm for relevant patents based on queries generated from descriptions of chemical effects

3 Developed software

The architecture of the search module for descriptions of chemical effects in USPTO patent documents is shown in Figure 6.

The module is developed in Python 3.5 with used APIs: Stanford NLP, Apache Hadoop, Apache Spark. Used libraries: pyqt, psycopg2, numpy, lxml. Database created using PostgreSQL.

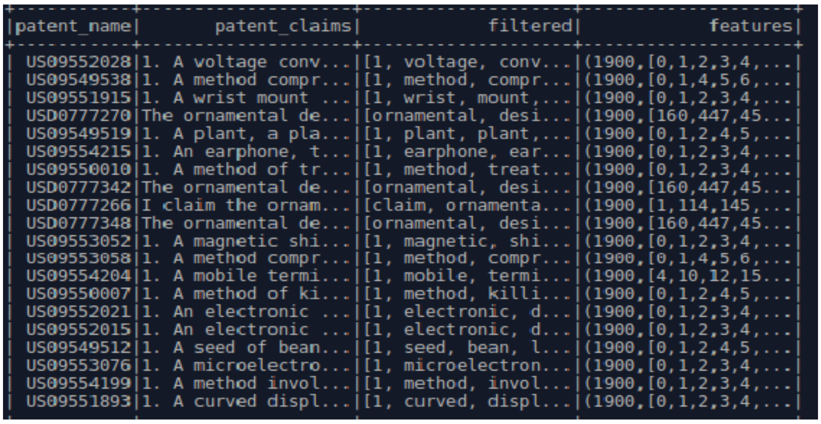

Interface
RDD with patent descriptions
Unit for extracting keywords from chemical effects descriptions
Unit of primary processing of patent archive
Unit for extracting keywords from descriptions of chemical effects
Search unit for the most relevant patents
Search Queries Database
Search results unit
Chemical Effects Database
1
2
3
4
5
6
7
8
9
10
11
12
13

1. Patent tar archive
2. File with the description of the chemical effect
3. Patent name and its description
4. Chemical effect name, its description, and keywords
5. Patent name and its description
6. Effect name and its keywords
7. Patent name, description, word weights
8. Effect name, patent name, patent description, relevance
9. List of patents
10. The text of the patent, its relevance
11. Loading patent archive
12. Loading chemical effect
13. Loading retrieved significant features of patents

Fig. 6. The architecture of the module

Figure 7 shows a use case diagram of the developed module in UML view.

The functions of the software module were tested. For example, RDD containing the names and descriptions of patent documents was used as input for the function of identifying sets of key features in patent documents. DataFrame table containing sets of key features with corresponding weight values was obtained as a result. You can see the records of the received table in the terminal (Figure 7).

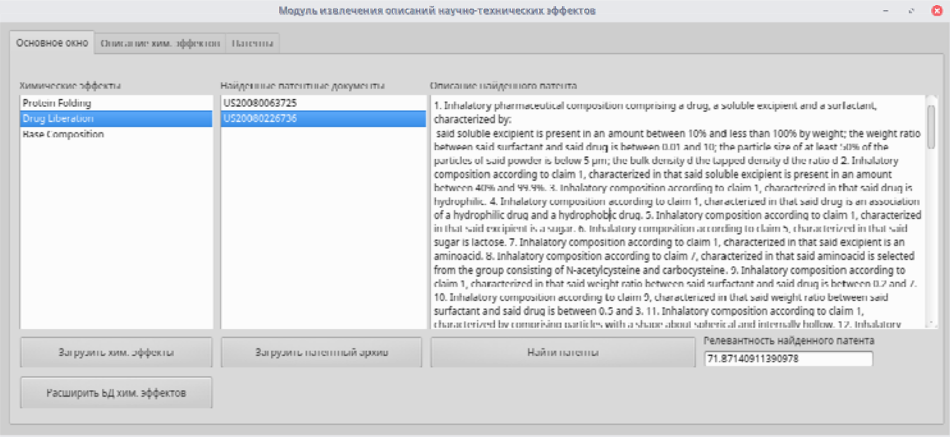


```
+------------+--------------------+--------------------+--------------------+
|patent_name|       patent_claims|            filtered|            features|
+------------+--------------------+--------------------+--------------------+
| US09552028|1. A voltage conv...|[1, voltage, conv...|(1900,[0,1,2,3,4,...|
| US09549538|1. A method compr...|[1, method, compr...|(1900,[0,1,4,5,6,...|
| US09551915|1. A wrist mount ...|[1, wrist, mount,...|(1900,[0,1,2,3,4,...|
| USD0777270|The ornamental de...|[ornamental, desi...|(1900,[160,447,45...|
| US09549519|1. A plant, a pla...|[1, plant, plant,...|(1900,[0,1,2,4,5,...|
| US09554215|1. An earphone, t...|[1, earphone, ear...|(1900,[0,1,2,3,4,...|
| US09550010|1. A method of tr...|[1, method, treat...|(1900,[0,1,2,3,4,...|
| USD0777342|The ornamental de...|[ornamental, desi...|(1900,[160,447,45...|
| USD0777266|I claim the ornam...|[claim, ornamenta...|(1900,[1,114,145,...|
| USD0777348|The ornamental de...|[ornamental, desi...|(1900,[160,447,45...|
| US09553052|1. A magnetic shi...|[1, magnetic, shi...|(1900,[0,1,2,3,4,...|
| US09553058|1. A method compr...|[1, method, compr...|(1900,[0,1,4,5,6,...|
| US09554204|1. A mobile termi...|[1, mobile, termi...|(1900,[4,10,12,15...|
| US09550007|1. A method of ki...|[1, method, killi...|(1900,[0,1,2,4,5,...|
| US09552021|1. An electronic ...|[1, electronic, d...|(1900,[0,1,2,3,4,...|
| US09552015|1. An electronic ...|[1, electronic, d...|(1900,[0,1,2,3,4,...|
| US09549512|1. A seed of bean...|[1, seed, bean, l...|(1900,[0,1,2,4,5,...|
| US09553076|1. A microelectro...|[1, microelectron...|(1900,[0,1,2,3,4,...|
| US09554199|1. A method invol...|[1, method, invol...|(1900,[0,1,2,3,4,...|
| US09551893|1. A curved displ...|[1, curved, displ...|(1900,[0,1,2,3,4,...|
```

patent_name – patent name;
patent_claims – patent description;
filtered – filtered words from the descriptions of the patents;
features – sparse-vectors containing a measure of the significance of filtered words.

Fig. 7. The storage structure of the extracted significant features

You can see a ranked list of patents that are relevant to the search query in Figure 8 as a result.



Модуль извлечения описаний научно-технических эффектов
Основное окно
Описание хим. эффектов
Патенты
Загрузить хим. эффекты
Загрузить патентный архив
Найти патенты
Расширить БД хим. эффектов

Химические эффекты

Protein Folding
Drug Liberation
Base Composition

Найденные патентные документы

US20080063725
US20080226736

Описание найденного патента

1. Inhalatory pharmaceutical composition comprising a drug, a soluble excipient and a surfactant, characterized by:
said soluble excipient is present in an amount between 10% and less than 100% by weight; the weight ratio between said surfactant and said drug is between 0.01 and 10; the particle size of at least 50% of the particles of said powder is below 5 μm; the bulk density d the tapped density d the ratio d 2. Inhalatory composition according to claim 1, characterized in that said soluble excipient is present in an amount between 40% and 99.9%. 3. Inhalatory composition according to claim 1, characterized in that said drug is hydrophilic. 4. Inhalatory composition according to claim 1, characterized in that said drug is an association of a hydrophilic drug and a hydrophobic drug. 5. Inhalatory composition according to claim 1, characterized in that said excipient is a sugar. 6. Inhalatory composition according to claim 5, characterized in that said sugar is lactose. 7. Inhalatory composition according to claim 1, characterized in that said excipient is an aminoacid. 8. Inhalatory composition according to claim 7, characterized in that said aminoacid is selected from the group consisting of N-acetylcysteine and carbocysteine. 9. Inhalatory composition according to claim 1, characterized in that said weight ratio between said surfactant and said drug is between 0.2 and 7. 10. Inhalatory composition according to claim 9, characterized in that said weight ratio between said surfactant and said drug is between 0.5 and 3. 11. Inhalatory composition according to claim 1, characterized by comprising particles with a shape about spherical and internally hollow. 12. Inhalatory

Релевантность найденного патента
71.87140911390978

Fig. 8. Ranked list of found patents

4 Conclusion

The elements of the patent description are extracted from USPTO patents. As descriptions of scientific and technical effects are used representations of chemical effects (chemical phenomenon) according to the classification of the National Center for Biotechnology Information. Feature Extraction technology of Spark MLlib is used for the extraction of significant features from patents. Semantic text processing (NLP) is used to identify key features from descriptions of chemical effects and to compose a search query based on them. The search for the most relevant patents containing descriptions of chemical effects is performed on the basis of the generated queries. The software is developed as an application for Linux-systems, its efficiency has been tested and integrated with an automated information system of support of database of physical effects [12].

Acknowledgment

The reported study was funded by RFBR (research project 18-07-01086), RFBR and Administration of the Volgograd region (projects 19-47-340007, 19-41-340016).

References

1. Korobkin, D., Fomenkov, S., Kolesnikov, S., Kamaev, V., 2014. Synthesis of the physical principle of operation of engineering systems in the software environment CPN TOOLS. Research Journal of Applied Sciences. 2014. T. 9. № 11. pp. 749-752
2. Davydov, D., Fomenkov, S., 2002: The automated design of linear structures of the physical principles of action of technical systems. Mechanician (2), 33–35
3. Kravets, A.G., Korobkin, D.M., Dykov, M.A., 2015. E-Patent examiner: Two-steps approach for patents prior-art retrieval. In Proceedings of IISA 2015 - 6th International Conference on Information, Intelligence, Systems and Applications 6. DOI: 10.1109/IISA.2015.7388074
4. Korobkin, D., Fomenkov, S., Kolesnikov, S., 2013. Semantic network of Physical Effects descriptions in Natural Language context. In Proceedings of the IADIS International Conference WWW/Internet 2013, ICWI 2013, pp. 342-346
5. Chemical Phenomena. Available at: https://www.ncbi.nlm.nih.gov/mesh/68055598
6. Bulk Data Storage System. Available at: https://bulkdata.uspto.gov
7. Taylor, R. An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3040523
8. HDFS. Available at: https://ru.bmstu.wiki/HDFS_(Hadoop_Distributed_Filesystem)
9. Karau, H., Konwinski, A., Wendell, P., Zaharia, M. Learning Spark: Lightning-Fast Big Data Analysis. ISBN 10: 1449358624. ISBN 13: 9781449358624
10. MLlib is Apache Spark's scalable machine learning library. Available at: http://spark.apache.org/mllib
11. Stanford CoreNLP – Natural language software. Available at: https://stanfordnlp.github.io/CoreNLP
12. Fomenkov, S., Korobkin, D., Kolesnikov, S., Dvoryankin, A., Kamaev V., 2014. Procedure of integration of the systems of representation and application of the structured physical knowledge. Research Journal of Applied Sciences. 2014. V. 9. No 10. pp. 700703