of choice]. *Prikaspiyskiy zhurnal: upravlenie i vysokie tekhnologii* [Caspian Journal: Management and High Technologies], 2014, no. 1, pp. 23–33.

7. Kureychik V. V., Sorokoletov P. V., Shcheglov P. S. Analiz sovremennogo sostoyaniya avtomatizirovannykh sistem priobrete-niya i predstavleniya znaniy [Analysis of the current state of automated systems for the acquisition and representation of knowledge]. *Izvestiya Yuzhnogo federalnogo universiteta. Tekhnicheskie nauki* [Proceedings of the Southern Federal University. Technical Sciences], 2008, no. 9, pp. 120–125.

8. Nildz B., Anderson Kh., et al. *Printsipy bukhgalterskogo ucheta* [Accounting principles], 2nd ed. Moscow, Finansy and statitistika Publ., 1994. 496 p.

9. Shleer S., Mellor S. *Obektno-orientirovannyy analiz: modelirovanie mira v sostoyaniyakh* [Object-Oriented Analysis: Modeling the World in the states], Kiev, Dialektika Publ., 1993. 236 p.

10. Dieng R., Giboin A., Tourtier P., Corby O. Knowledge acquisition for explainable, multiexpert, knowledge-based design systems. *European Knowledge Acquisition Workshop*, 1992, pp. 298–317.

УДК 004.912:[615.06+001.821]

MINING DRUG-DRUG INTERACTIONS FROM TEXTS OF SCIENTIFIC ARTICLES

Статья поступила в редакцию 27.11. 2014, в окончательном варианте 02.03. 2015

Kamaev Valeriy A., D.Sc. (Engineering), Professor, Volgograd State Technical University, 28 Lenin Ave., Volgograd, 400005, Russian Federation, e-mail: cad@vstu.ru

Melnikov Mikhail P., post-graduate student, Volgograd State Technical University, 28 Lenin Ave., Volgograd, 400005, Russian Federation, e-mail: m.p.melnikov@gmail.com

Vorobkalov Pavel N., Ph.D. (Engineering), Associate Professor, Volgograd State Technical University, 28 Lenin Ave., Volgograd, 400005, Russian Federation, e-mail: pavor84@gmail.com

Detection of drug-drug interactions (DDI) can cause serious consequences during treatment. A quick search of such interactions can provide doctors with information which is essential for making right decisions. Detection of DDIs is a time-consuming task. Natural language processing for text mining of scientific articles can be used to do DDI information more accessible for doctors. Nowadays there are some databases containing large amount of biomedical articles. Therefore computational performance of classification method applied for identification restricts usage of such methods. The main purpose of the research is to find a method of fast retrieval of DDI information from biomedical texts. In this article, we investigate up-do-date research works in the area of natural language processing for detection of DDIs. Many of investigated methods require much time to perform on large text corpuses. For developing and testing of DDI extraction methods we've created a text corpus containing examples of articles with and without DDI information. We propose a fast text mining approach to DDI articles classification using term frequency–inverse document frequency (tf-idf) statistic. Tf–idf is a numerical statistic that is intended to reflect how important a word is to a document in a corpus. To implement and test the classification algorithm, we've developed the text classification system. As a result, our approach is able to achieve reasonably high F1 score value (measure of binary classification) in DDI articles classification while still keeping short run-time. After all, we consider how to improve the developed algorithm for increase its precision and recall. When these improvements will be made the software realization of the algorithm may be used by experts in DDI area to search new DDI evidences in scientific publications.

Keywords: information retrieval, drug-drug interaction, machine learning

ПОИСК ХАРАКТЕРИСТИК ВЗАИМОДЕЙСТВИЙ ЛЕКАРСТВЕННЫХ СРЕДСТВ В ТЕКСТАХ НАУЧНЫХ СТАТЕЙ

Камаев Валерий Анатольевич, доктор технических наук, профессор, Волгоградский государственный технический университет, 400005, Российская Федерация, г. Волгоград, пр. им. В.И. Ленина, 28, e-mail: cad@vstu.ru

Мельников Михаил Павлович, аспирант, Волгоградский государственный технический университет, 400005, Российская Федерация, г. Волгоград, пр. им. В.И. Ленина, 28, e-mail: m.p.melnikov@gmail.com

Воробкалов Павел Николаевич, кандидат технических наук, доцент, Волгоградский государственный технический университет, 400005, Российская Федерация, г. Волгоград, пр. им. В.И. Ленина, 28, e-mail: pavor84@gmail.com

Взаимодействие лекарственных средств (ВЛС) может вызывать серьезные последствия во время лечения, при этом быстрый поиск информации о таких взаимодействиях может предоставить врачу необходимую для принятия решений информацию. Поиск информации об эффектах ВЛС является длительной задачей. Чтобы сделать такую информацию более доступной для врача, могут быть использованы методы машинного обучения в области обработки естественного языка. Современные библиографические базы данных содержат значительное количество научных статей в области медицины, а вычислительная сложность методов классификации, применяемых для определения статей нужной тематики, ограничивает использование таких методов. Главная цель данного исследования – поиск быстрого метода извлечения информации о ВЛС из текстов научных статей медицинской тематики. Были проанализированы результаты современных исследований в области применения методов обработки естественного языка для поиска ВЛС. При этом было выявлено, что многие из исследованных методов требуют значительных вычислительных затрат на больших объемах данных. Для разработки и тестирования эффективных методов поиска информации о ВЛС, был создан текстовый корпус, содержащий примеры статей – как содержащих так и не содержащих такую информацию. Был разработан быстрый метод автоматической классификации статей с использованием статистической меры «частота слова – обратная частота документа» (tf-idf). Эта мера используется для измерения степени важности слова для документа в корпусе текстов. Для тестирования предложенного алгоритма классификации была разработана специальная программная система. По результатам ее апробации на сформированной подборке текстов был сделан вывод о том, что предложенный метод позволяет достичь достаточно высоких значений F1 – меры измерения точности бинарной классификации, при этом метод не требует значительных вычислительных затрат. В результате проведенных исследований были намечены направления дальнейших улучшений алгоритма, которые могут повысить его точность. После практической реализации намеченных улучшений, модифицированное программное средство может быть использовано экспертами для поиска и описаний новых ВЛС.

Ключевые слова: информационные технологии, тексты на естественном языке, поиск информации, взаимодействие лекарственных средств, машинное обучение, автоматическая бинарная классификация, статистическая мера tf-idf, вычислительная эффективность.

General characteristics of the work problems. Detection of drug-drug interactions (DDIs) is an important practical challenge. One drug can increase, decrease, or change the therapeutic effect of another drug. Information about DDIs can help doctors to avoid potentially dangerous interactions of drugs. Other interactions can be used to improve the efficiency of treatment. A common source of DDIs information is commercial databases such as factsandcomparisons.com [3] and reference.medscape.com [7]. Being up-to-date is a crucial quality of such systems. New interactions between drugs are being detected continually, that's why information in drug monographs and Patient Information Leaflets can be out-dated. Scientific articles are a common source of new DDIs. Search of such interactions is a difficult task which requires qualified specialists in pharmacology or medicine. To find new articles which contain evidences of new DDIs demands looking

through thousands of scientific articles, filtering of the articles with DDI evidences, determining which drugs take part in the interaction and what is the type and significance of the described interaction. All these time-taking steps realized manually make almost impossible quick updating of DDI databases. Instant adding of new interactions and their correction remains an unresolved problem, that's why DDIs databases may lack the supporting scientific evidences and different databases can have unmatched DDI information [6]. Text mining of articles can solve this problem reducing time of detecting new articles, related to drug-drug interaction and thus giving experts an opportunity to focus more on information content. To use text mining algorithm we need authoritative and complete source of scientific articles and MEDLINE can be such source. MEDLINE (http://www.ncbi.nlm.nih.gov/pubmed) [12] is the biggest bibliographic database of life sciences and biomedical information. It provides over 13 millions article records with abstracts which are available for free for any users. The abstracts can be used as the main source of texts for DDIs text mining, but for such a large number of articles computational performance of classification method can be still a valuable factor.

In this article we briefly review the recent papers in the area of DDI automated search, which demonstrate different methods used for classification of articles. The considered algorithms demonstrate high precision and recall. Also we describe how we constructed corpus of articles which contain DDI evidences and randomly selected life science articles. Also there is a description of software architecture we developed for testing of developed algorithms. The main purpose of the research is to find an approach of fast retrieval of drug-drug interactions information from large databases of biomedical texts. For this task we consider usage of different methods and estimate their computational performance. After all we examine possible approaches to improve the developed algorithm, which can help increase its precision and recall. After these improvements the software realization of the algorithm may be used by experts in DDI area to search new DDI evidences in scientific articles.

State of the art. The proven importance of the DDI automated search explains why many scientists investigate this subject.

In [2] the authors construct a corpus from MEDLINE articles, using «Facts & Comparisons» Drug Interaction Facts database and their institution's care provider order entry system as a source of expert-reviewed drug interactions. They applied the LIBSVM (Library for Support Vector Machines) implementation of the SVM machine learning algorithm for classifying the articles. They show the advantage of using this method over using simple search queries in the MEDLINE database.

Another approach is described in [11]. The authors discover DDIs through «the integration of «biological domain knowledge» with «biological facts» from MEDLINE abstracts and «curated sources». They use «parse tree query language» requests for extracting information from natural language texts. After that they use AnsProlog programming language for reasoning. Using this approach over MEDLINE abstracts helped to find several potential DDIs which were not present in DrugBank database (www.drugbank.ca).

There are also researches that compare different methods of classification. In [4] authors use variable trigonometric threshold, support vector machine, logistic regression, naive Bayes and linear discriminant analysis (LDA) linear classifiers on articles containing DDIs. The learning and test sets are composed manually classifying articles finding publications with DDIs evidences. With 5 different feature transforms it's shown that all methods except LDA are effective and achieve high quality of classifying.

In 2011 in Huelva city (Spain) the conference «1st Challenge task on Drug-Drug Interaction Extraction» was held [8]. The corpus of the challenge consists of DDIs marked-up by a researcher with pharmaceutical background and annotated at the sentence level. The challenge organizers assert the advantages of kernel-based methods over classical machine learning classifiers.

Many of methods mentioned above are computationally intensive and require feasible time to perform on large text corpuses. As faster and requiring less computational resources we suggest

using of «term frequency–inverse document frequency» (tf-idf) statistic for classification. Because of its high performance this method was used in early search engines and showed its effectiveness on large document sets.

Corpus. We used «Facts & Comparisons» Drug Interaction Facts database (www.factsandcomparisons.com) for constructing a group of articles, containing DDIs. It's s a reputable and comprehensive source of drug information [2] containing over 1,800 detailed monographs. Each monograph has references to articles containing pharmacokinetics studies in which evidence for DDIs is reported. We randomly selected 186 drug-drug interactions from this database. Most of the references are listed by URL to article in MEDLINE. For each of the 186 DDIs, we included every reference to MEDLINE. So we identified 483 DDI articles and included their abstracts and bibliographic information in our corpus and labeled them as containing DDI information (DDI articles).

To construct group of articles without DDI information, we randomly selected 532 life science articles from MEDLINE database. It's known there is about 1 % prevalence of drug-drug interaction citations in MEDLINE's database [2]. Thus we consider that possible quantity of DDIs articles in this set as insignificant and ignored them. All these article's abstracts were added to the corpus as negative examples and labeled as not containing DDI information (not-DDI articles).

The whole corpus was randomly divided into three parts: training set (60 % articles); validation set (20 % articles) and test set (20 % articles). Characteristics of the corpus are shown in the table.

Table 1

Characteristics of annotations in the text corpus

Parameter	Value
Average number of words	186.59
Variance for number of words	8086.44
Standard deviation	89.92
Coefficient of variation	0.48

Classification. We extracted textual features from the abstract texts. To reduce influence of word-forms, every word was replaced with its stem using Stanford Core NLP library (nlp.stanford.edu) [10]. This approach doesn't take into account words polymorphism, but it allows quickly combine different word-forms into one stem. To ignore numbers we deleted all words, containing numerical symbols, removing the words, matching the following regular expression: «*[^0-9a-zA-Z]+*». Thus we presented every abstract as a word *vector* $d = \{w_0, w_1, \ldots, w_n\}$ where w_i is a word's stem.

Tf–idf is a numerical statistic that is intended to reflect how important a word is to a document in a corpus [9]. It helps to determine how unique the word is for the texts from the specified text class. Term frequency is calculated as follows:

$$tf(w, d) = n_i : N, \quad (1)$$

where n_i represents the total word w occurrences number in the document. N represents the total number of the words in the document.

Inverse document frequency is calculated as follows:

$$idf(w, D) = |D| : (d_i \supset w), \quad (2)$$

where $|D|$ represents the total number of documents among the entire training documents corpus. $|(d_i \supset w)|$ represents the total number of documents containing the word w.

Tf-idf statistic is calculated as follows:

$$tfidf(w, d, D) = tf(w, d) \times idf(w, D). \quad (3)$$

For classification we used terms confidence and support [13]. Confidence is similar with term frequency: if term frequency determines how important the word w is for document, confi-

dence determines importance of a document for a class of texts. The value of confidence is calculated as follows:

$$conf(w, c_m) = N(w, c_m) : N(w, all), \quad (4)$$

where $N(w, c_m)$ represents the total number of documents containing word w_j among the DDI articles category. $N(w, all)$ represents the total number of documents containing feature word w_j among the entire training documents corpus.

Support is a frequency of documents containing the word w. Thus it's a document frequency. The value of support is calculated as follows:

$$\sup(w, D) = idf(w, D)^{-1} = N(w, all) : |D|. \quad (5)$$

We modified the classification method «one-word location» based on presence or absence of «feature» word in the text [11]. In this case, feature word is determined as a word with confidence and support values bigger than some threshold values. Thus the classification works as follows:

$$class = DDI : conf(w, c_m) \geq threshold \wedge \sup(w) \geq threshold, \quad (6)$$

$$class \neq DDI : conf(w, c_m) < threshold \vee \sup(w) < threshold. \quad (7)$$

Because of low results of such method for our corpus we used the following modification of algorithm. Previously we calculate the values of confidence and support for every word in the corpus. Then we count the number of feature words in every abstract for current values of confidence and support thresholds. If the number of characteristic words in the abstracts is bigger than some particular value M, the article is classified into DDI articles as follows:

$$class = DDI : |w : conf(w, c_m) \geq Tc \wedge \sup(w) \geq Ts| \geq M, \quad (8)$$

$$, class \neq DDI : |w : conf(w, c_m) \geq Tc \wedge \sup(w) \geq Ts| < M, \quad (9)$$

where T_c is confidence threshold and T_s is support threshold.

As can be seen we have three parameters which affect the result - whether abstract is classified as DDI or not. We choose *F1* score value as classification quality characteristic. In that way we have three parameters which can be changed for optimization of *F1* score. The F1 score can be interpreted as a weighted average of the precision and recall, where an *F1* score reaches its best value at «1» and worst score at «0». Because of low computational complexity of classification with beforehand calculated confidence and support values, a method of grid search optimization is suitable. The method's idea is in the following: we loop all possible values of optimization parameters with particular steps. It guarantees finding of global maxima in given borders. Confidence and support thresholds were looped with 0.01 steps in borders [0, 1], when M parameter was looped with step 1 in borders [1, 10]. The choice of upper border for M parameter is explained by the length of abstracts and frequency of characteristic words in the texts. We present a part of looped values of thresholds and M in the table 2.

Table 2

Optimization parameters and classification quality measures

Optimization parameters			**Classification quality measures**				
T_c	T_s	M	*Type I error*	*Precision*	*Type II error*	*Recall*	*F1 score*
0,95	0,03	2	0	1	0,72	0,28	0,44
0,85	0,03	2	0,14	0,86	0,55	0,55	0,67
0,75	0,03	2	0,31	0,69	0,40	0,60	0,64
0,86	0,02	2	0,11	0,89	0,43	0,57	0,69

The optimal values for learning set are the following: confidence threshold = 0.86, support threshold = 0.02, M = 2. With these values we achieved *F1* score value 0.69 (precision = 0.89, re-

call = 0.57). Applying these values to test set we get *F1* score value 0.68 (precision = 0.80, recall = 0.60). It indicated that learning model is neither under, nor overfitted (table 3).

Table 3

Top 15 feature words

Word	Confidence	Support
curve	1,000	0,057
withdrawal	0,950	0,021
volunteer	0,951	0,063
metabolite	0,880	0,026
half	0,882	0,053
elimination	0,941	0,035
ritonavir	1,000	0,021
contraceptive	0,961	0,027
cmax	1,000	0,021
adverse	0,909	0,023
dose	0,883	0,160
randomize	0,896	0,049
coadministration	1,000	0,030
absorption	0,919	0,038
twice	0,905	0,022
life	0,900	0,051

The list of feature words for these thresholds values includes the following: withdrawal, volunteer, elimination, adverse, dose, randomize, pharmacodynamic, pharmacokinetic, mg, and other words. These words are obviously connected to pharmacology and DDI subject.

To check if the method's accuracy depends on the subject of the articles from the negative group at the next we added pharmacology articles to the random MEDLINE articles in the test set. We manually selected 62 abstracts from «British Journal of Pharmacology» not related to DDI. In this case *F1* score value still is 0,69 (precision = 0,81, recall = 0,60).

It is worth noting that the proposed approach is quite universal and is weakly dependent on the subject area, for example it doesn't take into account such information as drug groups or some semantic connections between drugs.

Also we can calculate computational complexity of this algorithm. For that we should consider complexity of different steps of learning and classification stages.

The algorithm for calculating tf-idf values is shown at the figure 1.

Because the average time complexity of adding new values and search in hash table is *O(1)* [1], the complexity of the whole algorithm is O(n), where n is a number of stems in the learning set. After that the values of confidence and support are calculated for each stem. There are 7942 unique stems in our training set. Adding new abstracts into training set can add new words into this dictionary, but this amount can't be considerable, because the vocabulary of scientific articles isn't infinite and is kept within reasonable limits. The numbers of articles containing DDI information should be saved for each stem. It simplifies further extensions of the learning set.

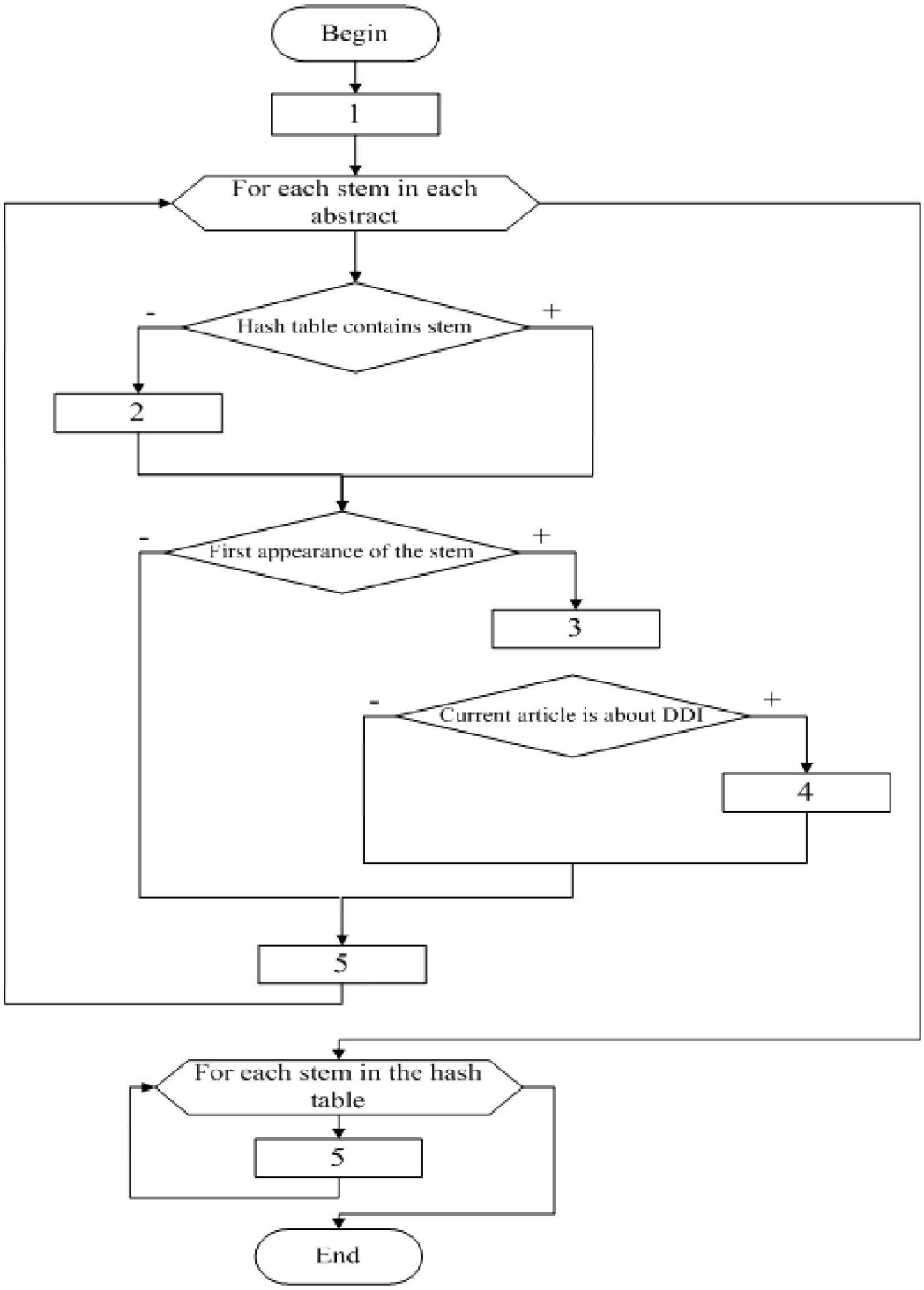

Fig.1. Tf-idf calculation algorithm, where (1) – create an empty hash table where the key is the word stem and the value is structure containing: number of DDI abstracts, containing the stem; total number of documents, containing the stem. (2) – add stem to the hash table. (3) – increment the total number of documents. (4) – increment number of DDI abstracts. (5) - calculate values of confidence and support

The next step is the learning. After the values of tf-idf have been calculated the optimal values of thresholds should be found. The time for learning depends on the constant number of iterations we chose before start of learning and the size of the learning set. Thus the complexity of the algorithm we used for learning is $O(n)$, where n is the number of stems in the learning set.

The process of classifying of one abstract doesn't require feasible time for calculation. We should count how many feature words contains the abstract to determine whether the abstract describes DDI interaction or not. Using hash table of feature words this algorithm's complexity is $O(m)$, where m is the number of the words in the abstract which is classified.

The next operation which computational complexity should be considered is the extension of the learning set. If we add new abstracts into the learning set values of confidence and support should be recalculated. The complexity of this operation is $O(kp)$ where k is the number of stems in the new abstracts and p is the number of the words in the dictionary.

As we can see complexity of all steps of the algorithm is low, the algorithm isn't computationally intensive and doesn't require feasible time to perform even on large text corpuses. For example the learning on the test set lasts about 4 minutes (Core i5 1.4GHz processor, one thread computation), classification of an articles abstract is instant (about 2 milliseconds).

To implement and test the classification algorithm we developed the text classification system. This system should meet the following requirements.

- It should be simple to use; it should use text user interface; the user types which action the system should perform, specifies which file contains the input data and where the system should place the output data.
- It should be able to work with different input data.
- The system should be easily scalable – it should allow adding new algorithms and modifying already implemented algorithms without changing core and other mechanisms of the system.

To meet these requirements we developed the following system architecture (fig. 2). Each function of the system corresponds to one command typed in the console.

As programming language for this system we have chosen Java. First of all this decision has been made because Stanford CoreNLP library is implemented using this programming language. The other reasons were its full cross-platform development support and its popularity among researchers in NLP area.

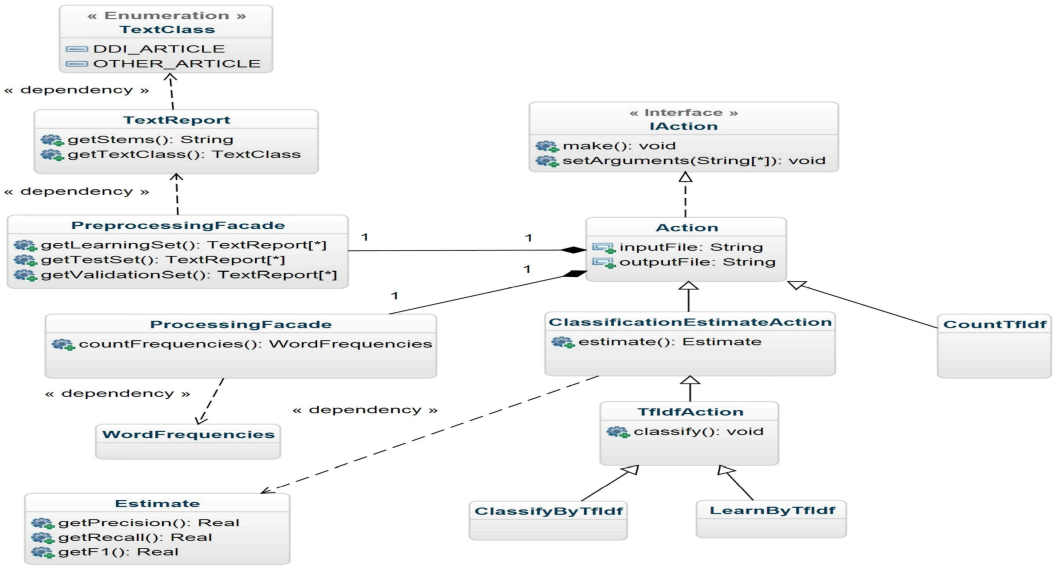

Fig.2. Classification system class diagram

Further works. The techniques which can be used for improvement of accuracy can be split into three groups: replacing classification method, changing features and feature values, manipulating with learning sets.

There are several different classification methods that can be used instead of modified «one-word location» method we are using now.

- Logistic regression.
- Linear discriminant analysis.
- Support vector machines (SVM).
- Binomial Naive Bayes.

Still because we target large article corpuses computational performance of these methods should also be considered.

Improvement possibilities that require changing features and feature values are as follows.

- Filtering of drugs names. The list of feature words includes such words as *phenytoin* and *aspirin*, which are named entities and they obviously can't improve the classification accuracy.
- After deleting or replacing named entities, using the dictionary with drug names, this number can be considered as an additional parameter.
- Using bigrams and monograms instead of just unigrams. Using bigram textual features together with unigrams has shown its effectiveness [8]. This improvement can significantly improve classification accuracy.
- Include Medical Subject Heading (MeSH) terms. Every article at MEDLINE has a list of Medical Subject Heading words. These words can be used as textual features too.
- Convert strings with numbers into '#' [4]. This action can help to increase classification's accuracy in case of using bigrams, because in this case, for example, such bigram as «# mg.» can become a feature word.
- Delete short textual features (those with a length of less than 2 characters) [4]. It can exclude prepositions and conjunctions from the text features.
- Delete infrequent features (which occurred in less than 2 documents) [4]. It can exclude named entities and other rare words which can't improve classification accuracy.

Improvement techniques that simply change learning set are as follows.

- Increasing number of DDI articles in learning set. The size of the learning set can influence the method's accuracy [5].
- Including DDI articles from other sources to learning set. To diversify the learning set there may be included articles from the sources besides «Facts & Comparisons» Drug Interaction Facts database (factsandcomparisons.com).

Conclusions. The suggested method of DDI articles classification demonstrates its stability under various conditions, but its accuracy should be significantly improved. If the value of precision is high enough, the value of recall is low. In this case «low» and «high» estimates mean algorithm's applicability to practical tasks. 19 % of incorrectly classified DDI articles can be filtered by human editor but 40 % of articles with DDIs lost by classification algorithm are still too much.

Acknowledgement.

Funding: This work was supported by Ministry of Education and Science of Russia as part of the basic part Project 2586, task №2014/16.

Список литературы

1. Cormen Thomas H. Introduction to Algorithms / Thomas H.Cormen, Charles E. Leiserson, Ronald L. Rivest, Clifford Stein. – 3rd ed. – Massachusetts Institute of Technology, 2009. – P. 253–280.

2. Duda S. Extracting drug–drug interaction articles from MEDLINE to improve the content of drug databases / S. Duda, C. Aliferis, R. Miller, et al. // AMIA Annual Symposium Proceedings Archive. – 2005. – P. 216–220.

3. Facts & Comparisons. – Available at: http://www.factsandcomparisons.com/facts-comparisons-online/.

4. Kolchinsky A. Evaluation of Linear Classifiers on Articles Containing Pharmacokinetic Evidence of Drug-Drug Interactions / A. Kolchinsky, A. Lourenko, L. Li, et al. // Pacific Symposium on Biocomputing. – 2013. – No. 18. – P. 409–420.

5. Kurczab R. The influence of negative training set size on machine learning-based virtual screening / R. Kurczab, Smusz S. Ling, A. Bojarski // Journal of Cheminformatics. – 2014. – Vol. 6.

6. Luis Tari Discovering drug-drug interactions: a text-mining and reasoning approach based on properties of drug metabolism / Luis Tari, Saadat Anwar, Shanshan Liang, James Cai, Chitta Baral. – Oxford University Press, 2010. – Vol. 26. – P. 547–553.

7. Medscape from WebMD. Drug interaction checker // Drugs.com. – Available at: http://reference.medscape.com/ drug-interactionchecker.

8. Proceedings of the 1st Challenge task on Drug-Drug Interaction Extraction / I. Segura-Bedmar, P. Martinez, D. Sanchez-Cisneros (ed.) // CEUR Work Proceedings. – Available at: http://ceur-ws.org/Vol-761/.

9. Rajaraman A. Data Mining / A. Rajaraman, J. D. Ullman // Mining of Massive Datasets. – 2011. – P. 1–17.

10. Stanford CoreNLP // The Stanford Natural Language Processing Group. – Available at: http://nlp.stanford.edu/software/corenlp.shtml.

11. Tari L. Discovering drug–drug interactions: a text-mining and reasoning approach based on properties of drug metabolism / L. Tari, S. Anwar, S. Liang, et al. // Bioinformatics. – 2010. – No. 26 (18). – P. 1547–1553.

12. U.S. National Library of Medicine // MEDLINE. – Available at: http://www.ncbi.nlm.nih.gov/pubmed.

13. Yun-tao Zhang An improved TF-IDF approach for text classification / Yun-tao Zhang, Ling Gong, Yong-cheng Wang // Journal of Zhejiang University Science. – Springer, 2005.

References

1. Cormen Thomas H., Leiserson Charles E., Rivest Ronald L., Stein Clifford *Introduction to Algorithms*, 3rd ed. Massachusetts Institute of Technology, 2009, pp. 253–280.

2. Duda S., Aliferis C., Miller R., et al. Extracting drug–drug interaction articles from MEDLINE to improve the content of drug databases. *AMIA Annual Symposium Proceedings Archive*, 2005, pp. 216–220.

3. *Facts & Comparisons*. Available at: http://www.factsandcomparisons.com/facts-comparisons-online/.

4. Kolchinsky A., Lourenko A., Li L., et al. Evaluation of Linear Classifiers on Articles Containing Pharmacokinetic Evidence of Drug-Drug Interactions. *Pacific Symposium on Biocomputing*, 2013, no. 18, pp. 409–420.

5. Kurczab R., Ling Smusz S., Bojarski A. The influence of negative training set size on machine learning-based virtual screening. *Journal of Cheminformatics*, 2014, vol. 6.

6. Luis Tari, Anwar Saadat, Liang Shanshan, Cai James, Baral Chitta *Discovering drug-drug interactions: a text-mining and reasoning approach based on properties of drug metabolism*, Oxford University Press, 2010, vol. 26, pp. 547–553.

7. Medscape from WebMD. Drug interaction checker. *Drugs.com*. Available at: http://reference.medscape.com/ drug-interactionchecker.

8. Segura-Bedmar I., Martinez P., Sanchez-Cisneros D. (ed.) Proceedings of the 1st Challenge task on Drug-Drug Interaction Extraction. *CEUR Work Proceedings*. Available at: http://ceur-ws.org/Vol-761/.

9. Rajaraman A., Ullman J. D. Data Mining. *Mining of Massive Datasets*, 2011, pp. 1–17.

10. Stanford CoreNLP. *The Stanford Natural Language Processing Group*. Available at: http://nlp.stanford.edu/software/corenlp.shtml.

11. Tari L., Anwar S., Liang S., et al. Discovering drug–drug interactions: a text-mining and reasoning approach based on properties of drug metabolism. *Bioinformatics*, 2010, no. 26 (18), pp. 1547–1553.

12. U.S. National Library of Medicine. *MEDLINE*. Available at: http://www.ncbi.nlm.nih.gov/pubmed.

13.Yun-tao Zhang, Ling Gong, Yong-cheng Wang An improved TF-IDF approach for text classification. *Journal of Zhejiang University Science*, Springer, 2005.