УДК 004.021

DOI: 10.15827/0236-235X.117.100-105

Дата подачи статьи: 30.01.17

2017. Т. 30. № 1. С. 100–105

МЕТОДЫ АВТОМАТИЗИРОВАННОГО АНАЛИЗА КОРОТКИХ НЕСТРУКТУРИРОВАННЫХ ТЕКСТОВЫХ ДОКУМЕНТОВ

П.Ю. Козлов, аспирант, originaldod@gmail.com

*(Смоленский филиал Национального исследовательского университета МЭИ,
Энергетический проезд, 1, г. Смоленск, 214013, Россия)*

В работе рассматриваются задачи автоматизированного анализа текстовых документов в органах исполнительной и законодательной власти. Выделяется группа признаков для классификации текстовых документов, приводятся их типы, методы анализа и рубрицирования. Определяется перечень типов документов, которые необходимо классифицировать.

Для анализа коротких неструктурированных текстовых документов предлагается использовать метод классификации на основе весовых коэффициентов, экспертной информации, нечеткого логического вывода, для которого усовершенствована вероятностная математическая модель, разработан способ обучения и экспериментально подобрано соотношение весовых коэффициентов. Предварительно разработанный метод необходимо обучить.

На этапе обучения слова тезауруса для каждой предметной области разбиваются на три типа: уникальные, редкие и общие, и в зависимости от типа словам присваиваются весовые коэффициенты. Для поддержания актуальности весовых и частотных коэффициентов предлагается использовать динамическую кластеризацию. Разработанный метод позволяет анализировать описанные документы, а также учесть динамичность тезауруса рубрик.

Представлена схема работы системы автоматизированного анализа неструктурированных текстовых документов, написанных на естественном языке, различных типов: длинные, короткие, очень короткие. В зависимости от типа документа используется соответствующий метод анализа, который имеет наилучшие показатели точности и полноты при анализе текстовых документов данного типа. В качестве синтаксического анализатора используется парсер Malt-Parser, обученный на национальном наборе русского языка.

Результатом работы всей системы можно считать базу знаний, в которую попадают все извлеченные знания и их отношения. База знаний постоянно пополняется и используется работниками исполнительной и законодательной власти для обработки поступающих запросов.

Ключевые слова: автоматизированный анализ текстов, короткие неструктурированные тексты, динамичный тезаурус.

Органы исполнительной и законодательной власти становятся все более открытыми для населения, и это приводит к росту количества заявлений, жалоб и предложений. Значительная часть этих обращений поступают в электронном виде, что обуславливает необходимость их автоматизированной обработки. По каждому обращению необходимо выполнить следующие действия:

- принять обращение, определить суть проблемы и специалиста для решения данной проблемы;
- отправить обращение конкретному специалисту;
- принять некоторые меры по решению указанной проблемы;
- написать ответ с отчетом о проделанной работе.

Специфика автоматизированных систем обработки текстовых документов такого рода в нестационарности тезауруса ключевых слов, которые с выходом новых правовых документов кардинально изменяются, к тому же поступающие от населения обращения являются неструктурированными и короткими, что затрудняет статистический анализ.

Анализ литературных источников [1–5] показал, что текстовые документы можно классифицировать по четырем признакам: структурированность, объем, значимость ключевых слов и синтаксическая связность (рис. 1).

Поступающие в региональную систему обращения граждан обычно относятся к типу неструктурированных, коротких или очень коротких связных текстовых документов. Эти признаки выделены на рисунке 1.

Для классификации (кластеризации) текстовых документов разработано большое количество методов и алгоритмов.

Рассмотрим возможности и проблемы использования известных методов для анализа документов указанного типа.

Из представленных на рисунке 2 типов методов для анализа коротких неструктурированных документов при наличии динамического изменения характеристик рубрик в наибольшей степени подходят методы на основе нейросетей, вероятностные и генетические.

В то же время известные варианты данных методов имеют ограничения при автоматизированном анализе обращений граждан: нейросетевые методы достаточно сложны в обучении и связаны с построением большой трудномасштабируемой сети и сложным выбором метрики [6–8]; вероятностные классификаторы на коротких и неструктурированных текстовых документах на естественном языке не дают достаточную точность и эффективность [9–12]; генетические алгоритмы трудно обучаются и тяжело перестраиваются под динамические характеристики тезауруса [13].

В этой связи представляется целесообразным использование их модификаций.

Идея разработанного метода на основе весовых коэффициентов состоит в следующем:

- каждому слову, соответствующему предметной области, назначается вес;
- по умолчанию вес всех слов принимается равным;
- проводится обучение метода на некоторой первоначальной выборке документов, в ходе которого веса слов изменяются соответственно их значимости в контексте конкретной предметной области;
- проводится корректировка весовых коэффициентов экспертом и на начальном этапе, и в процессе работы системы, так как известна правильность результатов анализа.

Чтобы показать модифицированную статистическую часть метода, приведем пример стандартного вероятностного классификатора в виде формулы

$$J \leftarrow F(d_k) = \max_j \left(\sum_i \begin{cases} f_{i,j} \cdot |c_{ik}|, & \frac{|c_{ik}|}{|d_k|} \geq P_{i,j} \\ 0, & \frac{|c_{ik}|}{|d_k|} < P_{i,j} \end{cases} \right),$$

где $F(d_k)$ – максимальная сумма произведения частот употребления слов на количество их употребления в k -м документе, которая определяет предметную область J ; $f_{i,j}$ – частота употребления i -го слова в j -й предметной области; $|c_{ik}|$ – количество употреблений i -го слова в k -м текстовом документе; $|d_k|$ – количество слов в k -м документе; $P_{i,j}$ – порог употребления i -го слова в j -й предметной области.

Математическую модель разработанного метода автоматической классификации можно представить в виде формулы

$$J \leftarrow W(d_k) = \max_j \sum_{i=1}^{n_k} K_{i,k} \cdot w_{i,j} \cdot |c_{i,k}|,$$

где $W(d_k)$ – максимальная сумма произведений весовых коэффициентов на количество употреблений в k -м документе (весовой коэффициент, в свою очередь, зависит еще от синтаксического коэффициента важности и актуальности информации); $w_{i,j}$ – весовой коэффициент i -го слова для j -й предметной области; $|c_{i,k}|$ – количество употреблений i -го слова в k -м документе; $K_{i,k}$ – синтаксический коэффициент важности i -го слова в k -м документе (определяется синтаксической значимостью слова в предложении).

В документах встречаются слова (общие), которые употребляются почти во всех предметных областях, они не несут информацию о предметной области документа. Следовательно, их веса необходимо сделать намного меньше других. Слова, встречающиеся только в одной предметной области (уникальные), являются самыми значимыми, и их веса будут значительно больше других. Еще остаются редкие слова, которые не являются ни уникальными, ни общими. Они несут некоторую информацию о предметной области, поэтому им назначаются промежуточные значения весовых коэффициентов. Алгоритм обучения проводит анализ БД ключевых слов и разбивает их на три категории: уникальные, редкие и общие. Далее эксперт выбирает нужное соотношение и значение весовых коэффициентов заданных трех типов ключевых слов. Для разграничения редких и общих слов вводится порог встречаемости: если слово встретилось в документах только одной предметной области, то это уникальные слова, если не во всех и меньше порога, то редкие, все остальные слова являются общими.

В ходе экспериментов были получены оптимальные значения весовых коэффициентов и порога отбора общеупотребительных слов: вес уникальных слов = 50, редких = 10, общих = 1, а порог отбора общих слов составляет 80 %. При данных характеристиках алгоритма обучения метод показывает наилучшие результаты.

Рис. 2. Схема классификации методов анализа текстовых документов

Fig. 2. Classification scheme of text document analysis methods

Можно выделить следующие преимущества весового алгоритма перед частотным алгоритмом:

- отсутствие порога частоты употребления слов позволяет распознавать короткие предложения, содержащие большое количество сокращений, цифр и минимум одно ключевое слово;
- использование весов, а не частот употребления увеличивает возможности для обучения алгоритма;
- при экспертной настройке базы знаний слов предметных областей удобнее работать с весами, чем с частотами и порогами употребления;
- возможность использовать слова, не принадлежащие напрямую к предметной области, но приписанные к ней с маленькими весовыми коэффициентами.

Недостатки весового алгоритма по сравнению с частотным алгоритмом:

- возрастает сложность процесса обучения;
- отсутствие порога распознавания предложения увеличивает вероятность распознавания документа, который относится к предметной области, неизвестной нашей системе;
- появление уникального слова с большим весом в другой предметной области сильно увеличит шанс на ошибочное распознавание документа.

Проблема динамического тезауруса в разработанном методе решается путем анализа сдвига кластеров предметных областей и своевременного запуска переобучения и подстройки метода. Методы динамической классификации изложены в [14–17].

Нечеткий логический вывод рассмотрен в статьях [18] и [19].

На рисунке 3 представлена схема функционирования системы автоматической классификации текстовых документов при использовании разработанного метода, предполагающего организацию обратной связи с экспертами.

Текстовый документ поступает на вход системы и проходит следующие этапы анализа: регистрация в системе, где получает уникальный номер

и озаглавливается необходимыми тэгами для дальнейшей работы; сегментация на слова, предложения, абзацы; морфологический анализ, определяющий лексические характеристики слов и морфемы; классификация документа по трем характеристикам и определение метода его анализа или обозначение документа как непригодного и возврат его экспертам для анализа (если документ короткий или очень короткий и неструктурированный, то используется разработанный метод анализа, в про-

Рис. 4. Схема базы знаний

Fig. 4. A knowledge base diagram

тивном случае – вероятностный); уточнение предметной области; синтаксическая разметка документа (для разработанного метода) и проверка синтаксической связности; поиск знаний, определение важности документа и адресата, отправка документа ответственному лицу; сохранение полученных знаний в базу знаний для дальнейшего использования. База знаний реализуется в виде фрейм-овой реляционной БД, которая хранит извлеченные понятия и отношения между ними. На рисунке 4 представлена схема базы знаний.

Предметная область, к которой принадлежит извлеченный фрейм Frame, представлена в виде сущности Class. Relation описывает отношения между фреймами, а также задает тип отношений. Value предназначен для хранения характеристик фреймов и соотносится со слотами и их типами [20].

В качестве ПО для БД используется MS SQL Server 2008, функционала которого достаточно для реализации поставленных задач. Программирование остальных этапов анализа, за исключением MaltParser, осуществляется на Microsoft Visual C#.

Для морфологического анализа предварительно составляется морфологический словарь на Национальном корпусе русского языка.

Для поддержания актуальности тезауруса каждый раз происходит динамическое отслеживание рубрик и их уточнение.

Основываясь на недостатках разработанного и имеющихся методов, необходимо изменить весь процесс анализа текстовых документов, как предложено на рисунке 3. Весовой метод необходимо применять только при анализе коротких текстовых документов, следовательно, добавляется этап проверки длины текстового документа, который определяет метод анализа: один из известных методов или описанный выше.

Литература

1. Батура Т.В., Мурзин Ф.А., Проскуряков А.В. Программный комплекс для анализа данных из социальных сетей // Программные продукты и системы. 2015. № 4. С. 188–197.
2. Schutze H., Hull D.A., Pedersen J.O. A comparison of classifiers and document representations for the routing problem. Proc.

SIGIR-95, 18th ACM Inter. Conf., Seattle, USA, 1995, pp. 229–237.

3. Ng H.T., Goh W.B., Low K.L. Feature selection, perceptron learning, and a usability case study for text categorization. Proc. SIGIR-97, 20th ACM Intern. Conf. 1, Philadelphia, USA, 1997, pp. 67–73.

4. Dagan I., Karov Y., Roth D. Mistake-driven learning in text categorization. Proc. EMNLP-97, 2nd Conf., Providence, USA, 1997, pp. 55–63.

5. Joachims T. Text categorization with support vector machines: learning with many relevant features. Proc. ECML-98, 10th Europ. Conf., Chemnitz, Germany, 1998, pp. 137–142.

6. Lam S.L., Lee D.L. Feature reduction for neural network based text categorization. Proc. DASFAA-99, Taiwan, 1999, pp. 195–202.

7. Ruiz M., Srinivasan P. Hierarchical text categorization using neural networks. Information Retrieval, 2002, vol. 5, no. 1, pp. 87–118.

8. Yang Y., Liu X. A re-examination of text categorization methods. Proc. of SIGIR-99, 22nd ACM Inter. Conf., Berkeley, USA, 1999, pp. 42–49.

9. Козлов П.Ю. Сравнение частотного и весового алгоритмов автоматического анализа документов // Научное обозрение. 2015. № 14. С. 245–250.

10. Lewis D.D. Naive (Bayes) at forty: The independence assumption in information retrieval. Proc. ECML-98, 10th Europ. Conf., Chemnitz, Germany, 1998, p. 415.

11. Heckerman D. A tutorial on learning with bayesian networks. Learning in graphical models, 1999, pp. 301–354.

12. de Campos L.M., Romero A.E. Bayesian network models for hierarchical text classification from a thesaurus. Inter. Jour. of Approximate Reasoning, 2009, vol. 50, no. 7, pp. 932–944.

13. Wong M.L., Cheung K.S. Data mining using grammar based genetic programming and applications. Kluwer Acad. Publ., 2002, 228 p.

14. Гимаров В.А., Дли М.И., Круглов В.В. Задачи распознавания нестационарных образов // Изв. РАН. Теория и системы управления. 2004. № 3. С. 92–96.

15. Гимаров В.А., Дли М.И., Круглов В.В. Временная изменчивость образов // Вестн. МЭИ. 2003. № 2. С. 91–98.

16. Гимаров В.А., Дли М.И., Круглов В.В. Задачи динамической кластеризации // Системы управления и информационные технологии. 2005. Т. 18. № 1. С. 18–21.

17. Гимаров В.А., Дли М.И. Нейросетевой алгоритм классификации сложных объектов // Программные продукты и системы. 2004. № 4. С. 51–55.

18. Круглов В.В., Дли М.И., Голунов Р.Ю. Нечеткая логика и искусственные нейронные сети. М.: Наука. Физматлит, 2001. 224 с.

19. Круглов В.В., Дли М.И. Интеллектуальные информационные системы: компьютерная реализация логического вывода. М.: Физматлит, 2002. 256 с.

20. Хабаров С.П. Представление знаний с применением фреймов. URL: <http://www.habarov.spb.ru/bz/bz07.htm> (дата обращения: 23.01.2017).

AUTOMATED ANALYSIS METHOD OF SHORT UNSTRUCTURED TEXT DOCUMENTS

P.Yu. Kozlov¹, Postgraduate Student, originaldod@gmail.com

¹ Smolensk Branch of the Moscow Power Engineering Institute, Energetichesky proezd 1, Smolensk, 214013, Russian Federation

Abstract. The paper considers the problem of an automated analysis of text documents in the executive and legislative authorities. It provides a characteristics group in order to classify text documents, their types, methods of analysis and rubricating. There is a list of the types of documents that need to be classified.

To analyze short unstructured text documents the authors propose to use a classification method based on weighting factors, expert information, fuzzy inference with a developed probabilistic mathematical model, a way of learning and experimentally chosen ratio of weight coefficients. The pre-developed method should be trained.

During learning the thesaurus words for each domain are divided into three types: unique, rare and common. The words are allocated with weights depending on the type. In order to maintain the relevance of weight and frequency coefficients it is proposed to use dynamic clustering. The developed method allows analyzing the disclosed documents, as well as taking into account thesaurus heading agility.

The paper presents a scheme of automatic classification system for unstructured text documents written in natural language. There might be various types of text documents: long, short, very short. Depending on the document type the system uses a corresponding method of analysis, which has the best indicators of accuracy and completeness of such text document analysis. MaltParser is a parser which is used here and trained on a national set of the Russian language. The result of the whole system work is a knowledge base, which includes all extracted knowledge and attitudes. The knowledge base is constantly updated and used by employees of the executive and legislative authorities to handle incoming requests.

Keywords: analysis automated analysis of texts, short texts unstructured, dynamic thesaurus.

References

1. Batura T.V., Murzin F.A., Proskuryakov A.V. A software package for a social network data analysis. *Programmnye produkty i sistemy* [Software & Systems]. 2015, no. 4, pp. 188–197 (in Russ.).
2. Schutze H., Hull D.A., Pedersen J.O. A comparison of classifiers and document representations for the routing problem. *Proc. 18th ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR-95)*. Seattle, 1995, pp. 229–237.
3. Ng H.T., Goh W.B., Low K.L. Feature selection, perceptron learning, and a usability case study for text categorization. *Proc. 20th ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR-97)*. Philadelphia, 1997, pp. 67–73.
4. Dagan L., Karov Y., Roth D. Mistake-driven learning in text categorization. *Proc. 2nd Conf. on Empirical Methods in Natural Language Processing (EMNLP-97)*. Providence, RI, 1997, pp. 55–63.
5. Joachims T. Text categorization with support vector machines: learning with many relevant features. *Proc. 10th European Conf. on Machine Learning (ECML-98)*. Chemnitz, Germany, 1998, pp. 137–142.
6. Lam S.L., Lee D.L. Feature reduction for neural network based text categorization. *Proc. DASFAA-99*. Taiwan, 1999, pp. 195–202.
7. Ruiz M., Srinivasan P. Hierarchical Text Categorization Using Neural Networks. *Information Retrieval*. 2002. vol. 5, no. 1, pp. 87–118.
8. Yang Y., Liu X. A re-examination of text categorization methods. *Proc. 22nd ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR-99)*. Berkeley, CA, 1999, pp. 42–49.
9. Kozlov P.Yu. Comparing frequency and weight algorithms for automatic document analysis. *Nauchnoe obozrenie* [Scientific Review]. 2015, no. 14, pp. 245–250 (in Russ.).
10. Lewis D.D. Naive (Bayes) at forty: The independence assumption in information retrieval. *Proc. 10th European Conf. on Machine Learning (ECML-98)*. Chemnitz, Germany, 1998, p. 415.
11. Heckerman D. A Tutorial on Learning With Bayesian Networks. *Learning in graphical models*. 1999, pp. 301–354.
12. de Campos L.M., Romero A.E. Bayesian network models for hierarchical text classification from a thesaurus. *Int. Jour. of Approximate Reasoning*. 2009, vol. 50, no. 7, pp. 932–944.
13. Wong M.L., Cheung K.S. *Data Mining Using Grammar Based Genetic Programming and Applications*. Kluwer Academic Publ., 2002, 228 p.
14. Gimarov V.A., Dli M.I., Kruglov V.V. Unstable image recognition problems. *Izvestiya Ross. akademii nauk. Teoriya i sistemy upravleniya* [Jour. of Computer and Systems Sciences International]. 2004, no. 3, pp. 92–96 (in Russ.).
15. Gimarov V.A., Dli M.I., Kruglov V.V. Temporary image variation. *Vestnik MEI* [MPEI Vestnik]. 2003, no. 2, pp. 91–98 (in Russ.).
16. Gimarov V.A., Dli M.I., Kruglov V.V. Dinamic clustering problems. *Sistemy upravleniya i informatsionnye tekhnologii* [Control Systems and IT]. 2005, vol. 18, no. 1, pp. 18–21 (in Russ.).
17. Gimarov V.A., Dli M.I. A neural network algorithm for classification of complex text objects. *Programmnye produkty i sistemy* [Software & Systems]. 2004, no. 4, pp. 5–56 (in Russ.).
18. Kruglov V.V., Dli M.I., Golunov R.Yu. *Nechetkaya logika i iskusstvennyye neyronnyye seti* [Fuzzy logic and artificial networks]. Moscow, Nauka, Fizmatlit Publ., 2001.
19. Kruglov V.V., Dli M.I. *Intellektualnye informatsionnye sistemy: kompyuternaya realizatsiya logicheskogo vyvoda* [Intelligent Systems: Computer Implementation of a Logical Inference]. Moscow, Fizmatlit Publ., 2002.
20. Khabarov S.P. *Predstavlenie znaniy s primeneniem freymov* [Knowledge Representation Using Frames]. Available at: <http://www.habarov.spb.ru/bz/bz07.htm> (accessed January 29, 2017).