Автоматизированное создание баз знаний для интеллектуальных систем с учетом лингвистической неопределенности

Е. Е. Котова[1], И. А. Писарев[2]
Санкт-Петербургский государственный электротехнический
университет «ЛЭТИ» им. В.И. Ульянова (Ленина)
[1]eekotova@gmail.com, [2] pisarevivan@yandex.ru

***Аннотация*. При проектировании онтологических баз знаний актуальными вопросами являются вопросы противоречия и/или избыточности, а также точности и интерпретируемости. Развитие, избыточность и насыщенность информационных ресурсов приводит к лингвистической неопределенности и затрудняет интерпретируемость и применение систем, основанных на знаниях, в человеко-машинных средах. Разные источники информации могут включать некоторые противоречия и/или избыточности. В работе представлена методология процесса проектирования тематических баз знаний. С целью уменьшения лингвистической неопределенности учитываются как экспертные знания, так и знания, извлеченные из информационных ресурсов. Веб-среда, реализующая методологию, направленна на потребность сотрудничества и поддержку обучения в проектировании баз знаний.**

Ключевые слова: базы знаний; онтологии; лингвистическая неопределенность; экспертные знания; информационные ресурсы

I. Введение

Базы знаний (KB) фиксируют факты о сущностях, их свойства, семантические отношения в форме триплетов «субъект-предикат-объект» (SPO). Доменно-ориентированные базы знаний, такие как DBpedia, Yago, Wikidata или Freebase [1–4], хранят миллиарды фактов, которые были частично автоматически извлечены из ресурсов/материалов/статей Википедии [5]. Развитие услуг, ориентированных на знания, инициирует задачи в контексте извлечения, управления и рассуждения с большими семантическими базами знаний.

Всемирная паутина является наиболее полным, но, как отмечают авторы [5] вероятно, самым сложным источником информации, к которой мы имеем доступ сегодня. Подавляющее большинство всей информация в общедоступной части сети Surface Web, состоит из неструктурированного текста или полу-структурированной информации в виде таблиц, списков, инфобоксов [5] и др.

Результатом запроса пользователя является набор ссылок на наиболее подходящие документы, выбранные поисковой системой из миллиардов доступных [6].

В структуре Википедии разработаны множество полезных функций для обнаружения и устранения неоднозначности именованных объектов [6], однако вместе с тем отмечается обилие неопределенности в интерпретируемости понятий: различные источники часто делают противоречивыми утверждения, в частности, специальные термины могут быть очень сложными для перевода [3].

Получение из множества документов целевой информации, точной, фактической – является для пользователя самостоятельной задачей, решение которой требует определенных базовых знаний.

Насыщенность информационных ресурсов в учебном процессе увеличивает когнитивную нагрузку в учебной деятельности студентов, что приводит к необходимости поиска способов ее снижения.

Появляются возможности для создания новых приложений [3]. Многие приложения в современных информационных технологиях используют онтологические базовые знания [2].

Представление знаний является достаточно развитой областью в направлении искусственного интеллекта и основывается на множестве моделей, начина от моделей концептуального представления знаний (фреймы, продукции, семантические сети) до языков семантического уровня представления знаний RDFS и OWL, составляющих основу описания онтологий [2].

Несмотря на подходы, основанные на автоматическом извлечении структур знаний из текстовых корпусов, использующие технологии разбора естественного языка, авторами отмечается, что наиболее успешные и широко используемые онтологии все еще созданы человеком [2].

В работе предлагается методология проектирования тематических баз знаний с применением методов Text Mining в автоматизированном режиме.

II. Средства анализа текстовой информации

Широкий спектр инструментов для решения задач лексического анализа текстовых форм в основном предназначен для анализа англоязычного контента.

Специальные инструменты, разработанные для задач обработки, интеллектуального анализа и исследования конкретных конструкций в тексте и дискурсе, используют разные подходы [7]. Краткий анализ примеров применения методов Text Mining и используемых инструментов представлен в табл. 1 (на основе публикаций [7–16]).

Таблица I МЕТОДЫ И ИНСТРУМЕНТЫ TEXT MINING (ПРИМЕРЫ)

Название, разработчик	**Используемые языки**
LIWC [8, 9, 10] (https://liwc.wpengine.com)	Словари LIWS2015 на различных языках
WMatrix [11] (https://www.lancaster.ac.uk/scc/)	Wmatrix4 для текстов на английском языке
Coh-Metrix [12] (https://www.memphis.edu/iis/)	Английский язык
TAGME [13, 14] (https://tagme.d4science.org/tagme/)	Английский, немецкий, итальянский языки
Apache Stanbol [15, 16] (https://stanbol.apache.org/)	Английский, датский, голландский, немецкий, португальский, испанский, шведский языки
VOSviewer (www.vosviewer.com)	Создание терминов сетей совместного использования на основе текстовых данных на английском языке.
ATR4S [17] (http://www.ispras.ru)	Работа с русскоязычной тематикой. Набор инструментов для автоматического распознавания терминов.

Для автоматизированной обработки и анализа текстов на английском языке могут использоваться программные средства: OpenNLP (http://opennlp.apache.org/), NLTK (http://www.nltk.org/) [18], Apache UIMA (http://uima.apache.org/) [19, 20], FlexiTerm [21], TermSuite [22], JATE [23].

Из краткого обзора можно заключить, что средства для работы с текстами на русском языке представлены в меньшей степени. Отсутствуют рекомендации по применению конкретных методов Text Mining, а также нет исследований по их сравнению [17]. Это подтверждает актуальность постановки задачи разработки инструментария для автоматизированного создания онтологических баз знаний с применением методов Text Mining. Также следует заметить, что данные инструменты предоставляют возможности их использования для специалистов-профессионалов высокого уровня: экспертов, программистов, лингвистов, математиков, но меньше предназначены для обучения студентов вузов. Учебные программы подготовки специалистов технического профиля не включают изучение методов Text Mining. Потому необходима разработка инструмента, предоставляющего возможности работы с мультиязычными ресурсами, включая русскоязычный контент с целью обучения специалистов и проведения научных исследований.

III. МЕТОДОЛОГИЯ АВТОМАТИЗИРОВАННОЙ РАЗРАБОТКИ БАЗЫ ЗНАНИЙ В ВИДЕ СТРУКТУРЫ ОНТОЛОГИИ

Методология автоматизированного создание баз знаний основана на алгоритме сценария выполнения последовательных шагов анализа областей знаний.

Алгоритм сценария обработки документов для построения базы знаний понятийной структуры области знаний изображен на рис. 1.

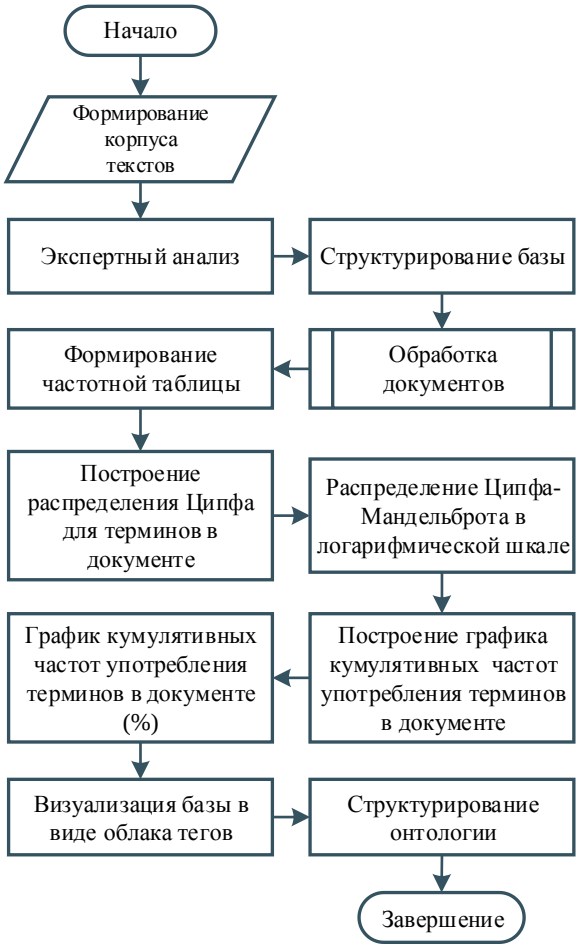

Рис. 1. Алгоритм сценария обработки документов

Контекст повышает структуру поиска [6], поэтому в первую очередь необходимо определить границы области знаний для формирования корпуса текстов. Источники информации важны как контекст информации и все первичные источники сохраняются в отдельном репозитории. Рекомендованные форматы текста doc, docx, pdf. Иные форматы преобразуются в заданные. Уточнение смысла запроса рассматривается как дополнительно преимущество пользователя [6] на этапе экспертного анализа. Предусмотрены два режима работы пользователя: рабочий и методический, которые реализованы в соответствующих интерфейсах. Рабочий интерфейс поддерживает все шаги по сценарию обработки информации. В методическом интерфейсе раскрываются и объясняются все алгоритмы и применяемые методы Text Mining. Пользователь может переключаться между рабочим и методическим интерфейсами и получать все необходимые знания. Например, поясняется, что такое «Латентно-семантический анализ» или «Облако тегов» с комментариями из источников информации.

Списки частот слов доступны в рабочем интерфейсе. Может быть представлен полный список частот, или сокращен в зависимости от поставленной задачи (например, удалением экспертом редко встречающихся слов, или с похожими свойствами). Цель состоит в том, чтобы уменьшить лингвистическую неопределенность, рассматривая термины с очень похожими определениями (свойствами) и удаляя те слова, которые не нужны для интерпретации и построения понятийной модели. Списки частот могут быть отсортированы по алфавиту или по частоте. Используются наиболее распространенные инструменты для предварительной обработки текстовых данных. Отмечается, что эти процедуры являются этапами предварительной обработки, но могут отличаться в разных языках [24].

Предложенный алгоритм отличается совместным применением правил морфологического анализа и анализа частот на основе базы данных общей лексики очень большого размера (Google NGram), что позволяет повысить точность при создании тематических словарей терминов.

IV. ПРИМЕРЫ УЧЕБНЫХ РАБОТ СТУДЕНТОВ ПОСТРОЕНИЯ ПОНЯТИЙНОЙ СТРУКТУРЫ ТЕМАТИЧЕСКОЙ ОБЛАСТИ ЗНАНИЙ

Примеры приведены для области знаний «Интеллектуальные методы анализа и обработки данных».

По сценарию необходимо подобрать корпус текстов (Source n) и провести его анализ. Обработка происходит в индивидуальном рабочем интерфейсе системы. Фрагменты частотных таблиц, полученных на примере трех источников, представлены в табл. 2–4.

Таблица II ФРАГМЕНТ ЧАСТОТНОЙ ТАБЛИЦЫ ИСТОЧНИКА 1

N	LogN	LogF	Термин	Частота (F)	Fc	Fcnorm
1	0	1,113943	данные	13	13	6,467662
2	0,30103	1	решение	10	23	11, 44279
3	0,477121	0,954243	объект	9	32	15,9204
4	0,60206	0,90309	метод	8	40	19, 9005
…	…	…	…	…	…	…

Таблица III ФРАГМЕНТ ЧАСТОТНОЙ ТАБЛИЦЫ ИСТОЧНИКА II.

N	LogN	LogF	Термин	Частота (F)	Fc	Fcnorm
1	0	1,518514	данные	33	33	13,63636
2	0,30103	1,230449	анализ	17	50	20,66116
3	0,477121	1,146128	метод	14	64	26,44628
4	0,60206	1,079181	классификация	12	76	31,40496
…	…	…	…	…	…	…

Таблица IV ФРАГМЕНТ ЧАСТОТНОЙ ТАБЛИЦЫ ИСТОЧНИКА III.

N	LogN	LogF	Термин	Частота (F)	Fc	Fcnorm
1	0	2,536558	данные	344	344	9,673791
2	0,30103	2,338456	метод	218	562	15, 80427
3	0,477121	2,178977	задача	151	713	20, 05062
4	0,60206	2,123852	анализ	133	846	23,79078
…	…	…	…	…	…	…

Полученные результаты показывают, что наиболее часто употребляемым термином в источниках, связанных с областью знаний «Интеллектуальные методы анализа и обработки данных», является «ДАННЫЕ». Также можно заметить, что в первой пятерке частоты употребления терминов во всех трех источниках являются термины «АНАЛИЗ, МЕТОД».

На рис. 2 изображены графики закона Ципфа-Мандельброта для трех источников. На основе графиков, можно сделать вывод, что третий источник обладает самым большим частотным значением, что говорит о его полноте раскрытия темы. На основе трех графиков распределений Ципфа-Мандельброта по линиям трендов и подобию углов наклона можно сделать вывод о наличии некоторых общих статистических свойств.

В итоге работы над корпусом текстов, представляющими область «Интеллектуальные методы анализа и обработки данных» студентами была составлена учебная онтология в редакторе онтологий Protégé (https://protege.stanford.edu/), изображенная на рис. 3.

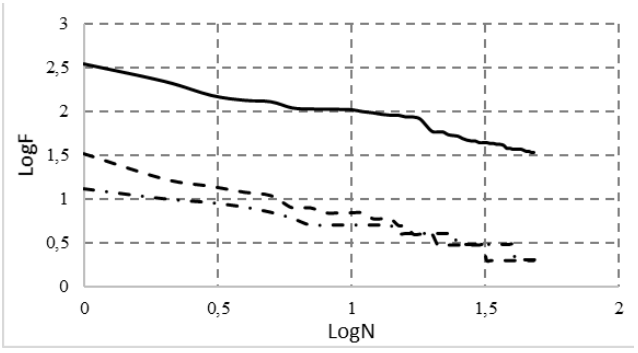

Рис. 2. Графики Ципфа-Мандельброта в логарифмическом масштабе

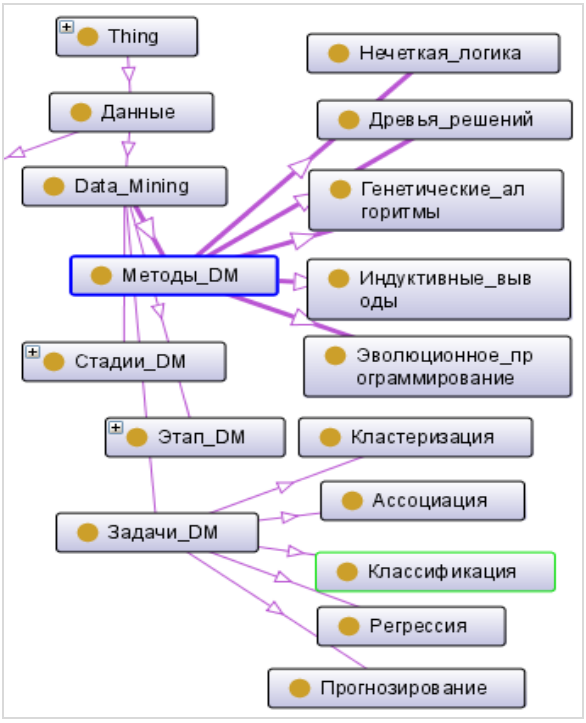

Рис. 3. Пример онтологии

V. Заключение

Особенностью подхода является реализация сценариев автоматизированного построения тематических баз знаний методами Text Mining с интеллектуальной поддержкой экспертного анализа понятийной структуры с целью уменьшения лингвистической неопределенности [25, 26].

Последовательность действий включает формирование корпуса текстов, преобразование форматов документов, формирование списка словоформ, частотного словаря, добавление определений в словарь терминов, построение распределений Ципфа и Ципфа-Мандельброта, построение облаков тэгов, добавление семантических отношений и редактирование онтологии. Особенностью работы является реализация сетевой системы автоматической обработки и анализа документов на русском и английском языках ОнтоМАСТЕР-Сценарий в соответствии со стандартами Semantic Web. В настоящее время система находится на стадии бета-тестирования и используется в экспериментальных целях в учебном процессе.

Выражение признательности

Авторы выражают благодарность к.т.н., доценту Писареву А.С. за помощь в реализации программ.

Список литературы

[1] Auer S. et al. Dbpedia: A nucleus for a web of open data //The semantic web. – Springer, Berlin, Heidelberg, 2007. – Pp. 722-735.

[2] Suchanek F. M., Kasneci G., Weikum G. Yago: a core of semantic knowledge //Proceedings of the 16th international conference on World Wide Web. – ACM, 2007. – Pp. 697-706.

[3] Vrandečić D., Krötzsch M. Wikidata: a free collaborative knowledge base. – 2014. https://ai.google/research/pubs/pub42240

[4] Bollacker K. et al. Freebase: a collaboratively created graph database for structuring human knowledge //Proceedings of the 2008 ACM SIGMOD international conference on Management of data. – AcM, 2008. – Pp. 1247-1250.

[5] International Conference on Current Trends in Theory and Practice of Informatics SOFSEM 2019: SOFSEM 2019: Theory and Practice of Computer Science. Pp. 50-53. https://link.springer.com/chapter/10.1007%2F978-3-030-10801-4_5

[6] Bunescu R., Paşca M. Using encyclopedic knowledge for named entity disambiguation //11th conference of the European Chapter of the Association for Computational Linguistics. – 2006. Pp. 9-16.

[7] Slater S. et al. Tools for educational data mining: A review //Journal of Educational and Behavioral Statistics. 2017. V. 42. No. 1. Pp. 85-106.

[8] Tausczik Y. R., Pennebaker J. W. The psychological meaning of words: LIWC and computerized text analysis methods //Journal of language and social psychology. 2010. V. 29. No. 1. Pp. 24-54.

[9] Pennebaker J. W., Francis M. E., Booth R. J. Linguistic inquiry and word count: LIWC 2001 //Mahway: Lawrence Erlbaum Associates. 2001. V. 71. No. 2001. P. 2001.

[10] Pennebaker J. W. et al. The development and psychometric properties of LIWC2015. 2015. 26 p.

[11] Wmatrix corpus analysis and comparison tool. http://ucrel.lancs.ac.uk/wmatrix/

[12] Graesser A. C. et al. Coh-Metrix: Analysis of text on cohesion and language //Behavior research methods, instruments, & computers. 2004. V. 36. No. 2. Pp. 193-202.

[13] Paolo Ferragina. http://pages.di.unipi.it/ferragina/

[14] Cornolti M., Ferragina P., Ciaramita M. A framework for benchmarking entity-annotation systems //Proceedings of the 22nd international conference on World Wide Web. ACM. 2013. Pp. 249-260.

[15] Apache Stanbol Components. (http://stanbol.apache.org/docs/trunk/components/)

[16] University profile maps. (http://www.vosviewer.com/university-profile-maps)

[17] Astrakhantsev N. ATR4S: toolkit with state-of-the-art automatic terms recognition methods in Scala //Language Resources and Evaluation. 2018. V. 52. No. 3. Pp. 853-872.

[18] Natural Language Toolkit. (http://www.nltk.org/)

[19] Cunningham H., Tablan V., Roberts A., Bontcheva K. Getting more out of biomedical documents with GATE's full lifecycle open source text analytics //PLoS computational biology. 2013. V. 9. No. 2. Pp.1- 14.

[20] Ferrucci D., Lally A., Gruhl D., Epstein E., Schor M., Murdock J. W., Welty C. Towards an interoperability standard for text and multi-modal analytics //IBM Res. Rep. 2006.

[21] Spasić I., Greenwood M., Preece A., Francis N., Elwyn G. FlexiTerm: a flexible term recognition method //Journal of biomedical semantics. 2013. V. 4. No. 1. Pp. 27.

[22] Cram D., Daille B. Terminology extraction with term variant detection //Proceedings of ACL-2016 System Demonstrations. 2016. Pp. 13-18.

[23] Zhang Z., Gao J., Ciravegna F. JATE 2.0: Java Automatic Term Extraction with Apache Solr //LREC. – 2016.

[24] Lucas C. et al. Computer-assisted text analysis for comparative politics //Political Analysis. 2015. V. 23. No. 2. Pp. 254-277.

[25] Kotova E. E., Pisarev I. A. Active internet technologies implementation in the students learning process support system //2016 XIX IEEE International Conference on Soft Computing and Measurements (SCM). IEEE, 2016. Pp. 302-304.

[26] Pisarev Kotova E. E., Pisarev A. S., Stash N. V. Structure of knowledge base of methods for processing hydroacoustic signals //2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus). IEEE, 2018. Pp. 1132-1135.

[27] Котова Е.Е., Писарев А.С., Писарев И.А. Программный комплекс разработки сценариев анализа данных ОнтоМАСТЕР-Сценарий. Свидетельство о государственной регистрации программы для ЭВМ № 2019613556 от 19 марта 2019 г.