Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК_____

Азарко Владислав Вячеславович

Алгоритмы и методы обработки больших объёмов текстовой информации

АВТОРЕФЕРАТ
на соискание степени магистра технических наук
по специальности 1-40 80 02 «Системный анализ, управление и обработка информации»

_____

Научный руководитель
Гуринович Алевтина Борисовна

кандидат физико-математических наук

_____

Минск 2019

ВВЕДЕНИЕ

В современном мире текст является наиболее распространенным средством обмена информацией. Примерами текстовой информации являются электронные письма, большие текстовые документы, HTML-файлы, комментарии и сообщения в социальных сетях, комментарии в интернет магазинах и многое другое. На основании приведённых примеров можно сделать вывод, что текстовые данные присутствуют во всех сферах виртуальной деятельности человека, поэтому их анализ может быть полезен для бизнеса, государственных и некоммерческих организаций.

Анализ текста определяется как процесс обнаружения скрытых зависимостей и знаний в неструктурированных текстовых документах. Анализ текста также известен как процесс поиска знаний в текстовом интеллектуальном анализе данных. Проблемами анализа текстовой информации занимаются такие науки как компьютерная лингвистика и обработка естественного языка.

Компьютерная лингвистика (также: математическая или вычислительная лингвистика, англ. computational linguistics) — научное направление в области математического и компьютерного моделирования интеллектуальных процессов у человека и животных при создании систем искусственного интеллекта, которое ставит своей целью использование математических моделей для описания естественных языков.

Обработка естественного языка (Natural Language Processing, NLP) — общее направление искусственного интеллекта и математической лингвистики. Оно изучает проблемы компьютерного анализа и синтеза естественных языков. Применительно к искусственному интеллекту анализ означает понимание языка, а синтез — генерацию грамотного текста. Решение этих проблем будет означать создание более удобной формы взаимодействия компьютера и человека.

Как можно заметить из приведённых выше определений, обе науки используют для решения задач анализа текстовой информации математические методы. Математические методы, применяемые в других областях искусственного интеллекта, находят своё применение и в анализе текстов. Целью своей работы я ставлю рассмотрение основных задач анализа текстовой информации и сравнение математических моделей и методов их решения.

Магистерская диссертация выполнена самостоятельно, проверена в системе «Антиплагиат». Процент оригинальности соответствует норме, установленной кафедрой. Цитирования обозначены ссылками на публикации, указанные в «Списке использованных источников».

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность данной работы заключается в необходимости изучения и выбора наилучших методов интеллектуальной обработки больших объёмов текстовой информации, что позволит более эффективно анализировать текстовые данные в таких сферах как:

- оптимизация поисковых запросов;
- управление текстовыми базами знаний;
- кибербезопасность;
- управление рисками;
- анализ мнений.

Для успешного решения задач в перечисленных выше областях необходимы быстрые и качественные алгоритмы интеллектуальной обработки текста, что повышает актуальность исследований в этой области.

Целью магистерской диссертации является сделать обзор наиболее часто применяемых методов интеллектуальной обработки текстовой информации, описать качественные характеристики работы этих методов, а также провести их сравнение.

В ходе работы над данной магистерской диссертацией были выполнены следующие задачи:

- показать сферу применения интеллектуальной обработки больших объёмов текстовой информации;
- обозначить актуальность исследований в области обработки естественного языка и обработки текстов;
- описать основные задачи обработки текстовой информации;
- описать алгоритмы интеллектуального анализа текстов;
- изучить результаты научных исследований в этой области;
- описать критерии сравнения алгоритмов;
- провести сравнение алгоритмов классификации текстовых документов.

Результаты данной магистерской диссертации были опубликованы в сборнике 55-ой научной конференции аспирантов, магистрантов и студентов БГУИР.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** рассмотрено современное состояние проблемы обработки больших объёмов текстовой информации и интеллектуального анализ текста в целом, определены основные направления исследований, а также дается обоснование актуальности темы диссертационной работы. Сформулированы ее цель и задачи, даны сведения об объекте исследования и обоснован его выбор, представлены положения, выносимые на защиту, а также, структура и объем диссертации.

В **первой главе** рассматривается проблема интеллектуального анализа текста. Рассматриваются основные направления развития в области анализа текста, сфера применения и основные задачи интеллектуального анализа текстовой информации. Также рассматриваются основные этапы процесса анализа текстовых документов и методы их математического описания.

Во **второй главе** производится рассмотрение задачи классификации документов как основной задачи интеллектуального анализа текстов. В этой главе даётся постановка задачи и описание основных методов её решения. В главе приводится описание критериев сравнения методов решения задачи классификации, а также обзор исследований в этой области.

В **третьей главе** рассматривается одно из практических применений задачи классификации текстовых документов, а именно – задача анализа тональности текста. В главе даётся постановка задачи и описывается её решение двумя наиболее эффективными способами: методом опорных векторов и свёрточными нейронными сетями. В главе подробно описываются этапы решения задачи и описывается эксперимент по сравнению методов её решения. На основании результатов эксперимента даётся сравнительная характеристика применения метода опорных векторов и свёрточных нейронных сетей для решения задачи анализа тональности текстовых документов.

В **приложениях** приведены листинги программного кода алгоритмов решения задачи оценки тональности текстовых документов с использованием метода опорных векторов и свёрточных нейронных сетей.

ЗАКЛЮЧЕНИЕ

Результатом данной магистерской диссертации стало сравнение методов интеллектуального анализа больших объёмов текстовой информации.

В ходе работы над данной магистерской диссертацией были выполнены следующие задачи:

– описана сфера применения интеллектуальной обработки больших объёмов текстовой информации;

– обозначена актуальность исследований в области обработки естественного языка и обработки текстов;

– описаны основные задачи обработки текстовой информации;

– описаны алгоритмы интеллектуального анализа текстов;

– изучены результаты научных исследований в этой области;

– описаны критерии сравнения алгоритмов;

– проведено сравнение алгоритмов классификации текстовых документов.

Результаты данного исследования могут быть применены при разработке автоматизированных систем обработки текстовой информации, создании экспертных и рекомендательных систем в различных отраслях экономики.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

[1] Ju R. An Efficient Method for Document Categorization Based on Word2vec and Latent Semantic Analysis. 2015 IEEE Int. Conf. on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing. Liverpool, UK, 2015, pp. 2276–2283.

[2]. Sentiment analysis algorithms and applications: A survey/ Medhat W., Hassan A., Korashy H - Ain Shams Engineering Journ. 2014

[3] Polyakov I.V. A Text classification problem and features set/ I.V. Polyakov, T.V. Sokolova, A.A. Chepovsky - Vestn. NGU 2015, vol. 13, iss. 2, pp. 55–63.

[4] Tarasov D.S. Deep Recurrent Neural Networks for Multiple Language Aspect-Based Sentiment Analysis. Computational Linguistics and Intellectual Technologies: Proc. of Annual Int./ D.S. Tarasov - Conf. “Dialogue-2015”. Moscow, Russia, 2015, vol. 2, iss. 14 (21), pp. 65–74.

[5] Moraes R. Document-level sentiment classification: An empirical comparison between SVM and ANN./ R. Moraes, J.F. Valiati, W.P. Gavião Neto - Expert Systems with Applications. 2013, no. 40, pp. 621–633.

[6] Ghiassi M. Automated text classification using a dynamic artificial neural network model / Ghiassi M., Olschimke M., Moon B - Expert Systems with Applications. 2012, no. 39, pp. 10967–10976.

[7] Zhang X. Character-level Convolutional Networks for Text Classification/ Zhang X., Zhao J., LeCun Y - Proc. of the Neural Information Processing Systems Conf. (NIPS 2015). Montreal, Canada, 2015

[8] Yang Y. An evaluation of statistical approaches to text categorization. Information Retrieval Jour. 1999, vol. 1, iss. 1, pp. 69–90.

[9] A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 22(10):1345–1359, 2010

[10] Multi-Domain Sentiment Dataset [Электронный ресурс] – Режим доступа: http://www.cs.jhu.edu/~mdredze/datasets/sentiment/

[11] Айвазян С.А. Прикладная статистика: Классификация и снижение размерности / С.А. Айвазян, В.М. Бухштабер, И.С. Юнюков - М.: Финансы и статистика, 1989.

[12] Knowledge Discovery Through Data Mining: What Is Knowledge Discovery? - Tandem Computers Inc., 1996.

[13] Кречетов Н. Продукты для интеллектуального анализа данных. / Н. Кречетов - Рынок программных средств, N14-15_97, с.32-39.

[14] General Systems Theory - The Skeleton of Science / Boulding K.E. - Management Science, 2, 1956.

[15] Гик Дж., Прикладная общая теория систем/ Дж. Гик - М.: Мир, 1981.

[16] Киселев М. Средства добычи знаний в бизнесе и финансах / М. Киселев, Е. Соломатин - Открытые системы, № 4, 1997, с.41-44.

[17] Дюк В.А. Обработка данных на ПК в примерах/ В.А. Дюк - СПб: Питер, 1997.

[18] Нейский И.М. Методика адаптивной кластеризации фактографических данных на базе Fuzzy C-means и MST / И. М. Нейский

[19] Корунова, Н. В. Кластеризация документов проектного репозитария на основе нейронной сети кохонена / Н. В. Корунова [Электронный ресурс]. - Режим доступа: http://nsmv2008.ulstu.ru/docs/klasterzacij_dokumentov.pdf

[20] Елизаров, С. И. Разработка и Исследование методов и алгоритмов кластеризации для систем анализа данных / С. И. Елизаров [Электронный ресурс]. - Режим доступа: http://www.eltech.ru/education/aspir/SIElizarov.doc

[21] Кластерный анализ [Электронный ресурс]. - Режим доступа: http://ru.wikipedia.org/wiki/Кластерный_анализ

[22] Yoshua Bengio, Learning Deep Architectures for AI, 2009

[23] Deep Learning in Neural Networks: An Overview - Jurgen Schmidhuber, 2014.

[24] Ian Goodfellow, Yoshua Bengio and Aaron Courville, Deep Learning, MIT Press, 2016.

[25] Ю. В. Рубцова. Построение корпуса текстов для настройки тонового классификатора / Рубцова Ю.В. - Программные продукты и системы, 2015, №1(109), —С.72-78.

[26] С. Короткий. "Нейронные сети: Основные положения / Короткий С. - СПб, 2002. 357 с.

[27] Фролов А.А. Информационные характеристики нейронных сетей/ А.А. Фролов, И.П. Муравьев - М.: Наука, 2005, 160 с.

[28] Минский М., Пейперт С. Перцептроны/ Минский М. Мир, 2001. 234 с.

[29] Хайкин С. Нейронные сети: полный курс / С. Хайкин - издательский дом «Вильямс», 2006. – 1104с.

[30] Дипломные проекты (работы). Общие требования. СТП 01-2013. – Минск, БГУИР. – 2013. – 174 с.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТЫ

Азарко, В. В. Применение Microsoft Orleans в разработке автоматизированных систем обработки информации / В. В. Азарко, Е. С. Высоцкий // Информационные технологии и управление: материалы 54-й научной конференции аспирантов, магистрантов и студентов, Минск, 23 – 27 апреля 2018 г. – Минск: БГУИР, 2018. – С. 42.

Азарко, В. В. Сравнение методов классификации текстов / В. В. Азарко // Информационные технологии и управление: материалы 55-й научной конференции аспирантов, магистрантов и студентов, Минск, 22 – 26 апреля 2019 г. – Минск: БГУИР, 2019. – С. 54.