References

1. Ermakov A.E., Pleshko V.V. *Semanticheskaia interpretatciia v sistemakh kompiuternogo analiza teksta* [*The semantic interpretation in the computer analysis of the text*]. Informatcionnye tekhnologii [Information technology], 2009, no. 6, pp. 2-7.
2. Ermakov A.E. *Avtomaticheskoe izvlechenie faktov iz tekstov dose: opyt ustanovleniia anaforicheskikh sviazei* [*Automatic extraction of facts from text files: the experience of the establishment of anaphoric relations*]. Kompiuternaia lingvistika i intellektualnye tekhnologii : tr. mezhdunar. konf. «Dialog'2007» [Computational Linguistics and Intellectual Technologies: Third Intern. Conf. «Dialog'2007»], 2007, pp. 131-135.
3. Kiselev S.L., Ermakov A.E., Pleshko V.V. *Poisk faktov v tekste estestvennogo iazyka na os-nove setevykh opisanii* [*Search the facts in the text of natural language based on network descriptions*]. Kompiuternaia lingvistika i intellektualnye tekhnologii : tr. mezhdunar. konf. «Dialog'2004» [Computational Linguistics and Intellectual Technologies: Third Intern. Conf. "Dialog'2004."], 2004, pp. 72-75.
4. CoreNLP official site. Available at: http://nlp.stanford.edu /software/corenlp.shtml (accessed 2015).
5. Calais official site. Available at: http://www.opencalais.com/about (accessed 2015).
6. NetOwl Extractor official site. Available at: https://www.netowl.com /entity-extraction/ (accessed 2015).
7. Ontosminer official site. Available at: http://www.ontos.com /products/ontosminer/ (accessed 2015).
8. Link Parser official site. Available at: http://www.abisource.com /projects/link-grammar/ (accessed 2015).
9. Agichtein Eugene, Gravano Luis. *Snowball: extracting relations from large plain-text collections.* In Proceedings of the fifth ACM conference on Digital libraries, 2000, pp. 85–94.
10. Minard Anne-Lyse, Ligozat Anne-Laure, Grau Brigitte. *Multi-Class SVM for Relation Extraction from Clinical Reports.* Proceedings of Recent Advances in Natural Language Processing, 12-14 September 2011, pp. 604-609.
11. Dmitriev A.S., Zaboleeva-Zotova A.V., Orlova Y.A., Rozaliev V.L. *Automatic identification of time and space categories in the natural language text.* Applied Computing 2013: proceedings of the IADIS International Conference (Fort Worth, Texas, USA, October 23-25, 2013), IADIS (International Association for Development of the Information Society), 2013, pp. 187-190.
12. Gildea D., Daniel J. *Automatic Labeling of Semantic Roles.* Computational Linguistics, 2002, Vol. 28, No 3, pp. 245-288.
13. Lao Ni, Subramanya Amarnag, Pereira Fernando, Cohen William W. *Reading The Web with Learned Syntactic-Semantic Inference Rules.* Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 12-14 July 2012, pp. 1017–1026.
14. Stenchikova S., Dilek Hakkani-Tur, Gokhan Tur. *QASR: Spoken Question Answering Using Semantic Role Labeling.* State University of New York, 2004, No 3, pp. 11–17.
15. Wenlei M., Wesley W. *The phrase-based vector space model for automatic retrieval of free-text medical documents.* Data & Knowledge Engineering, 2007, pp. 76–92.

УДК 004.912

МЕТОДЫ АДАПТАЦИИ ТЕКСТОВОЙ ИНФОРМАЦИИ ДЛЯ ЛИЦ С ОГРАНИЧЕННЫМИ ВОЗМОЖНОСТЯМИ ПО ЗРЕНИЮ[1]

Статья поступила в редакцию 04.11.2015 г., в окончательном варианте 15.11.2015 г.

Орлова Юлия Александровна, кандидат технических наук, кандидат педагогических наук, доцент, Волгоградский государственный технический университет, 400005, Российская Федерация, г. Волгоград, пр. им. Ленина, 28, e-mail: yulia.orlova@gmail.com

[1] Исследование выполнено при финансовой поддержке РФФИ в рамках научных проектов № 13-07-00351, 14-07-97017, 15-07-07519, 15-07-05440.

Работа посвящена вопросу адаптации текстовой информации для восприятия лицами с ограниченными возможностями здоровья по зрению. Рассматривается извлечение ключевых сущностей из текстов тематически близких новостных статей и их визуализация. Для решения задачи агрегации новостей выбран алгоритм шинглов и проведена его модификация в части введения параллельного выполнения операций сравнения значений хеш-функций. Представлены несколько вариантов такой параллельной реализации: с использованием технологий CUDA, Open CL и Google App Engine. Оценены параметры реализации алгоритма (время работы, ускорение достигаемое по сравнению с последовательной обработкой), применительно к задаче анализа новостных текстов. Представлена программная реализация комплексного анализа новостного текста, основанная на комбинации смыслового анализа и последующего аннотирования текста материала с представлением ее в сжатом виде в формате Mind Map.

Ключевые слова: новостной текст, нечеткие дубликаты, шинглы, аннотирование, mind map, технологии распараллеливания вычислений, смысловой анализ, лица с ограниченными возможностями здоровья, CUDA, Open CL, Google App Engine

METHODS OF ADAPTATION OF TEXT INFORMATION FOR PERSONS WITH DISABILITIES OF PERCEPTION IMPAIRED

Orlova Yuliya A., Ph.D. (Engineering), Ph.D. (Pedagogics), Associate Professor, Volgograd State Technical University, 28 Lenin Ave., Volgograd, 400005, Russian Federation, e-mail: yulia.orlova@gmail.com

The work is devoted to the question of adaptation text information to persons with disabilities. Discusses the retrieval of the key entities from the texts thematically similar news articles and their visualization. To solve the problem of aggregating news selected algorithm shingles and its modification carried out regarding the introduction of parallel execution of operations comparison of values of hash functions. Presents several variants of this parallel implementation: using technologies like CUDA, Open CL and Google App Engine. The estimated parameters of the algorithm (time, acceleration achieved compared to sequential processing), in terms of the analysis of news texts. Presents software realization of complex analysis of news text, based on a combination of semantic analysis and subsequent annotation of text material with a view in its compressed form in the format of a Mind Map.

Keywords: news text, fuzzy duplicates, shingles, annotation, mind map, technologies of parallel computing, semantic analysis, persons with disabilities, CUDA, Open CL, Google App Engine

Введение. В последнее время в России все большее внимание уделяется лицам с ограниченными возможностями здоровья (ОВЗ). Создаются условия для их передвижения, обучения и коммуникации. Одним из видов таких ограничений являются нарушения зрения. С целью создания необходимых условий для их жизнедеятельности и обучения разрабатываются специальные меры: организационные, инженерно-технические, аппаратно-программные и пр. При этом возможны различные подходы: адаптация среды обучения и / или жизнедеятельности; разработка и использование технических средств, компенсирующих недостатки зрения; представление информации в виде, приемлемом для лиц с нарушениями зрения. Последнее направление разработано относительно слабо. Поэтому целью данной работы является рассмотрение моделей и методов адаптации текстовой информации для представления ее лицам с ОВЗ по зрению.

Общая характеристика проблематики работы. На основании федерального закона № 273-ФЗ «Об образовании в Российской Федерации» образовательные организации должны обеспечить возможность инклюзивного образования по адаптированным образователь-

ным программам для лиц с ОВЗ. Инклюзивное образование предполагает обеспечение равного доступа к образованию для всех обучающихся с учетом разнообразия особых образовательных потребностей и индивидуальных возможностей. Адаптированная образовательная программа – образовательная программа, адаптированная для обучения лиц с ОВЗ с учетом особенностей их психофизического развития, индивидуальных возможностей и, при необходимости, обеспечивающая коррекцию нарушений развития, социальную адаптацию указанных лиц. Организация и осуществление образовательной деятельности по образовательным программам высшего образования – программам бакалавриата, специалитета, магистратуры, подготовки научно-педагогических кадров в аспирантуре (адъюнктуре) подразумевают создание безбарьерной среды для лиц с ОВЗ: с нарушениями зрения, с нарушениями слуха, с ограничением двигательных функций.

Рассмотренные в данной работе модели и методы адаптации текстовой информации для лиц с ОВЗ по зрению могут применяться для автоматизации сбора и визуализации новостной информации с официальных сайтов образовательных организаций и образовательных web-ресурсов.

Новостные агрегаторы – это одни из самых востребованных ресурсов в Интернете [4]. Новости читают все, однако для людей с ОВЗ по зрению практически отсутствуют удобные средства и возможности работы с новостными ресурсами и лентами новостей (кроме масштабирования ими изображений на экране дисплея) [4]. Например, ленты новостей с образовательных сайтов становятся малодоступными для лиц с ОВЗ, т.к. разработчики дизайна сайтов заинтересованы в уменьшении размера шрифта с целью размещения большего объема отображаемой информации.

На решение этой проблемы направлены последние издаваемые законы и нормативные акты министерства образования РФ (в том числе закон № 273-ФЗ). При этом соответствующее решение для лиц с ОВЗ по зрению на большинстве Интернет-ресурсов пока либо отсутствует, либо тривиально ограничено увеличением шрифта основного текста.

Поэтому в данной работе для расширения возможностей коммуникации с окружающим миром для лиц с ОВЗ по зрению разработаны методы отображения текстовой информации в адаптированной форме – аннотирование и представление в виде графической записи MindMap [9, 14]. Основной идеей работы является объединение различных источников новостей за счет алгоритма поиска нечетких дубликатов [5], далее аннотирования текстов и отображения их визуального представления [1, 15]. Альтернативой для лиц с ОВЗ по зрению может быть устное воспроизведение новостной информации (или аннотаций к ней) на основе звукосинтезирующих программ. Однако такое решение имеет ряд недостатков, главные из которых это необходимость установки сторонних программ или постоянного доступа в Интернет; достаточно низкая скорость звукового воспроизведения текстов и, соответственно, восприятия информации.

Выделение тематически близких текстов. Существует несколько основных методов установления тематической близости документов [2, 6]. В рамках данной работы были рассмотрены TF, TF-IDF, TF-RIDF, Long Sent, Heavy Sent, Shingles, Lex Rand [6]. В результате экспериментального сравнения качества работы каждого из методов было принято решение использовать алгоритм шинглов. Этот алгоритм был предложен в 1997 году Бродером [13]. Он основан на представлении документа в виде последовательностей фиксированной длины N, состоящих из соседних слов. При этом на последовательности могут накладываться ограничения – например, слова должны находиться в одном предложении. Такие последовательности в одних источниках называют «шинглами», в других «N-граммами» [3]. Два документа считаются похожими, если множества их N-грамм существенно пересекаются. Аналогично можно оценить похожесть двух предложений, или же предложения и текста.

Д. Фетерли была предложена модификация алгоритма шинглов в которой документ представлялся 84 шинглами [3, 6]. Выбор из всего множества шинглов происходит по следующей схеме: для всех шинглов документа рассчитывается значение 84 хеш функций. Для каждой хеш функции выбирается шингл с максимальным значением хеш функции. Затем эти 84 шингла разбиваются на 6 групп по 14 шинглов. Такие группы называются «супершинглами». Далее документ представляется всевозможными попарными сочетаниями из 6 супершинглов, которые называются «мегашинглами». Число таких мегашинглов равно 15 (число сочетаний из 6 по 2). Два документа сходны по содержанию, если у них совпадает хотя бы один мегашингл.

В качестве важного направления использования данного алгоритма рассматривалось выявление заимствований в текстах (включая научные статьи, диссертации, дипломные работы и пр.) [5]. Однако выяснилось, что описанный алгоритм является «неустойчивым» при модификации текста в случае заимствований. Также алгоритм не обнаруживает заимствования в случае малого совпадения документов. Описанный алгоритм применим для отбора документов при поиске, если критерием отбора документов является обнаружение масштабных заимствований. Для поиска конкретных заимствованных фрагментов алгоритм не применим.

Для повышения производительности и скорости работы при сохрании высокой точности результатов в данной работе используется модификация алгоритма шинглов, предложенного в [6]. Модифицированный алгоритм, так же как исходный алгоритм, состоит из пяти этапов с введением распараллеливания на пятом шаге.

1. Канонизация текста. Она приводит оригинальный текст к единой форме без предлогов, союзов, знаков препинания, HTML тегов. Из текста удаляются прилагательные, а существительные приводятся к именительному падежу, единственному числу [3].

2. Разбиение на шинглы, т.е. выделенные из анализируемого текста последовательности из 10 слов, идущих друг за другом.

Sh1 = word0 word1 word2 … word8 word9

Sh2 = word1 word2 word3 … word9 word10

Sh3 = word2 word3 word4 … word10 word11 и т.д.

Таким образом, для текста получается набор шинглов, равный количеству слов минус длина шингла плюс один. Действия по каждому из первых двух пунктов алгоритма выполняются для каждого из сравниваемых текстов.

3. Вычисление контрольных сумм хеш-функциями. Сравниваемые тексты представляются в виде набора шинглов и контрольных сумм, рассчитанных через 84 уникальные между собой хеш-функции [9, 11]. Для каждого шингла рассчитывается 84 значения контрольной суммы через разные функции (SHA1, MD2, MD4, MD5, CRC32 и т.д.).

4. Выборка сигнатуры сравниваемых текстов. В каждом наборе вычисленных контрольных сумм (фактически это столбы таблицы, где строки это шинглы, а столбцы - 84 значения хеш-функции) выбирается минимальное значение и включается в сигнатуру, характеризующую весь текст.

5. Сравнение и определение результата. Сравниваются значения каждой хеш-функции, входящей в сигнатуру каждого из исследуемых текстов. Количество операций сравнения двух текстов при таком подходе сокращается до 84-х. Однако, проблема алгоритма заключается в количестве таких сравнений. Увеличение количества текстов для сравнения характеризуется экспоненциальным ростом количества операций, что критически отражается на вычислительной эффективности. Данный этап улучшен введением распараллеливания.

Параллельная реализация алгоритма шинглов с использованием технологий CUDA и Open CL. В алгоритме шинглов основную вычислительную нагрузку несет вычисление кэша. В результате экспериментальной оценки работы алгоритма с помощью четырех

разных реализаций, для каждой из них была получена сводная таблица с оценкой времени вычислений (в секундах). Число итераций можно рассматривать как число анализируемых новостных текстов.

Таблица 1

Сводная таблица времени расчета разных подходов

Итерации	50	100	150	200	250	300	350	400	450	500
Последовательный код	0,39	1,14	2,23	3,69	5,54	7,74	10,29	13,21	16,50	20,15
OpenMP	0,34	0,71	1,44	2,21	3,32	5,47	7,69	9,86	11,93	14,26
CUDA	0,13	0,23	0,51	0,73	1,10	1,42	1,89	2,21	2,82	3,42
OpenCL	0,11	0,17	0,27	0,34	0,52	0,62	0,87	1,04	1,35	1,69

Полученные результаты для наглядности представлены в графической форме на рисунке 1.

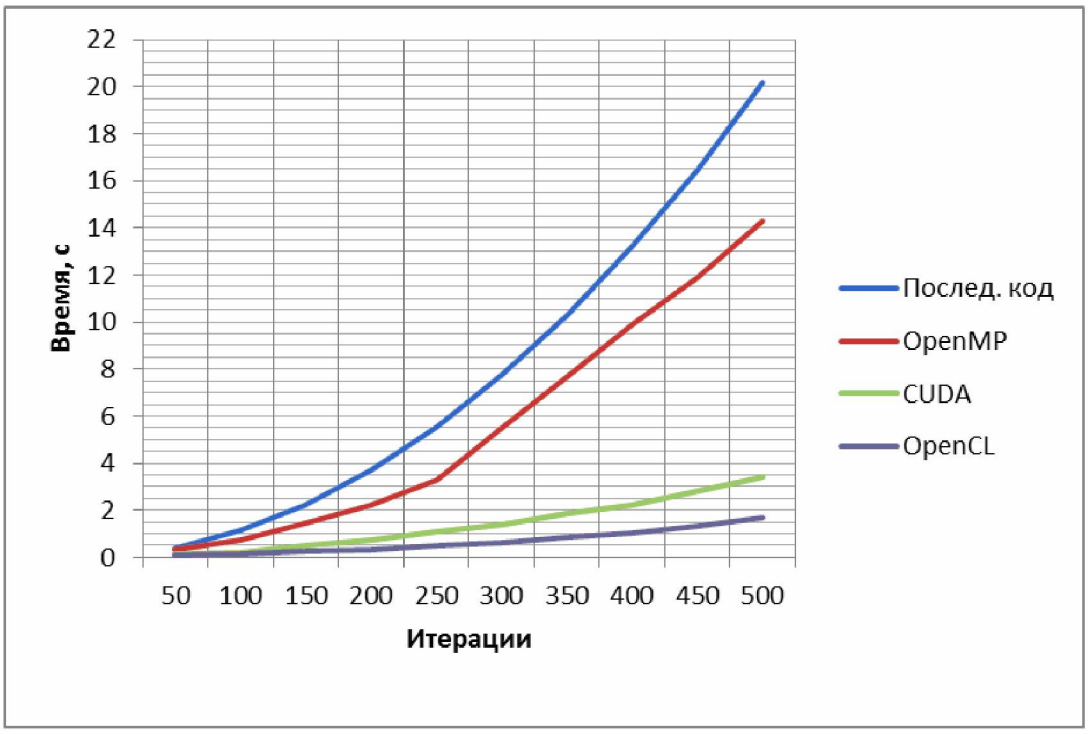

Рис. 1. Зависимость времени обработки от числа текстов

Тестирование проводилось на устройстве с CPU «Intel core i5 3.0 GHz», видеокартой (Cuda) NVIDIA GeForce GTX 650Ti 2GB и OpenCL AMD Radeon 7870 2GB.

По сравнению с последовательным алгоритмом для параллельной реализации алгоритма шинглов с использованием технологии Cuda среднее ускорение составило 5.10; а с применением технологии OpenCL –10.12. При этом параллельно выполнялся только подсчет и сравнение хешей, а нормализация текста выполнялась последовательно. OpenCL превосходит CUDA в 1,93 раза, что в данном случае объясняется тем, что тестируемая с OpenCL видеокарта незначительно превосходит своего конкурента по вычислительной мощности.

Параллельная реализация алгоритма шинглов с использованием технологии Google App Engine. Google App Engine – сервис хостинга сайтов и web-приложений на серверах Google с бесплатным именем <имя_сайта>.appspot.com, либо с собственным именем, задействованным с помощью служб Google. Приложения, разворачиваемые на базе App

Engine, должны быть написаны на Python, Java, PHP. Для проведения тестирования алгоритм был переписан на языке Python.

Тестирование работы программы проводилось на 30 текстах размером от 46 до 50 Кб. В тестовой выборке содержалось по 10 текстов на 3 новостные темы. Сначала запускалась программа при помощи Google App Engine, затем локально (вычислительный узел – одно ядро Core i5 3.3 GHz). Время разделения текстов по 3-м тематикам приведено в таблице 2.

Таблица 2

Время расчета и ускорении, достигаемое за счет использования Google App Engine

Итерации	5	10	15	20	25	30
Время с App Engine	9,558	18,454	27,792	36,450	45,416	54,436
Время локально	10,212	20,414	30,530	40,518	50,888	63,166
Ускорение	1,068	1,106	1,099	1,112	1,120	1,160

Таким образом, за счет введения распараллеливания процессов обработки можно существенно увеличить скорость объединения новостных текстов в тематические кластеры [8, 14]. Google App Engine позволяет достичь ускорения порядка 1,2 раза по сравнению с локальным запуском приложения с последовательным выполнением процессов. При этом необходимо учесть, что для выполнения операций с Google App Engine выделялось число вычислительных узлов не более четырех. При выделении большего количества вычислительных мощностей для работы с текстами больших размеров можно получить значительное ускорение по сравнению с использованием локальной машины.

Далее для каждого такого кластера осуществляется выделение ключевых фраз и слов, которые используются для построения визуального представления новости.

Извлечение из текста ключевых фраз и представление их в виде интеллект-карты новости (mind map). В основе построения новостных текстов заложен принцип «перевернутой пирамиды», который требует размещения основной информации в самом начале материала и последующее ее раскрытие далее по тексту в деталях. Заголовок новости отражает ее тему и содержит не более 10 слов. Основные факты отражены в 1–2 абзацах текста (лид), 3-ий и последующие абзацы раскрывают детали происходящего (новостной информации). Таким образом, для содержимого новости справедлива формула: (Who?+What?+Where?+Why?+When?+ How?) – закон «пять W и одно H» [2, 8].

Оптимальным в силу особенностей построения и удобства визуального представления новости является использование графового метода [7, 10, 12, 15]. В комбинации с графовым методом автором статьи разработан также собственный алгоритм подсчета веса ключевых слов, который более подробно описан в [8] (рис. 2).

Алгоритм заключается в поиске ключевых предложений (для установления повышающего коэффициента для сущностей из этого предложения) и далее нахождении в цикле ключевых слов из всех выделенных сущностей. Кандидатами в ключевые слова являются уникальные слова (персоны, организации, места и прочие). Также в кандидаты добавляются слова, которые не удалось определить при помощи морфологического словаря, и слова-сущности в именительном падеже. Пороговое значение относительной частоты для отнесения сущности к ключевым словам экспериментально установлено равным 0,2 × количество сущностей [2, 8, 14].

На рисунке 3 представлен скриншот главного окна разработанной программы, осуществляющей комплексный анализ новостного текста и его визуализацию. Для выбранной с новостного сайта статьи на основе алгоритма определения ключевых сущностей и предложений строится ее аннотация.

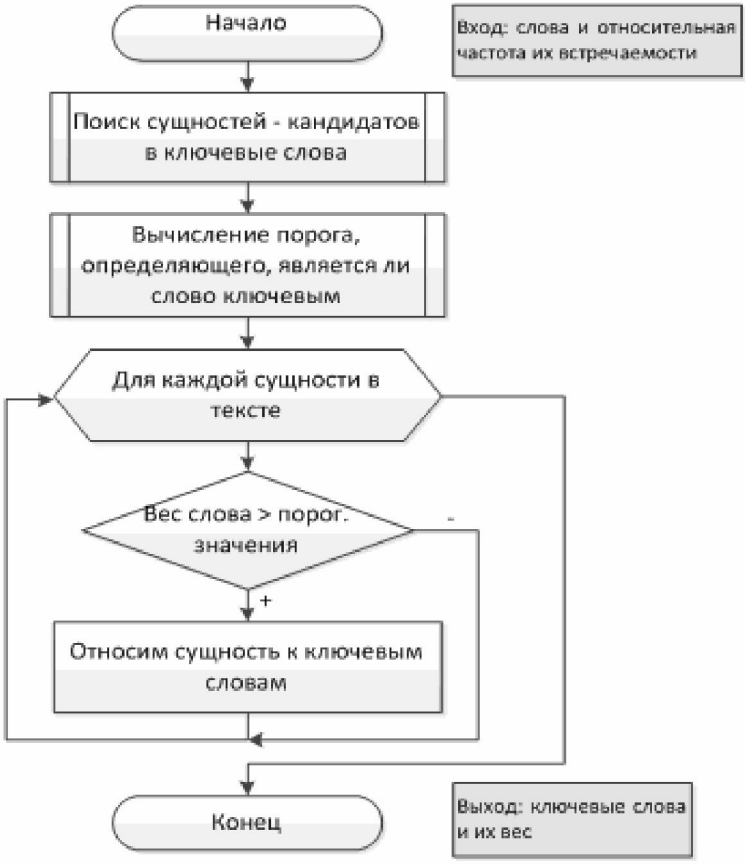

Рис. 2. Алгоритм поиска ключевых слов

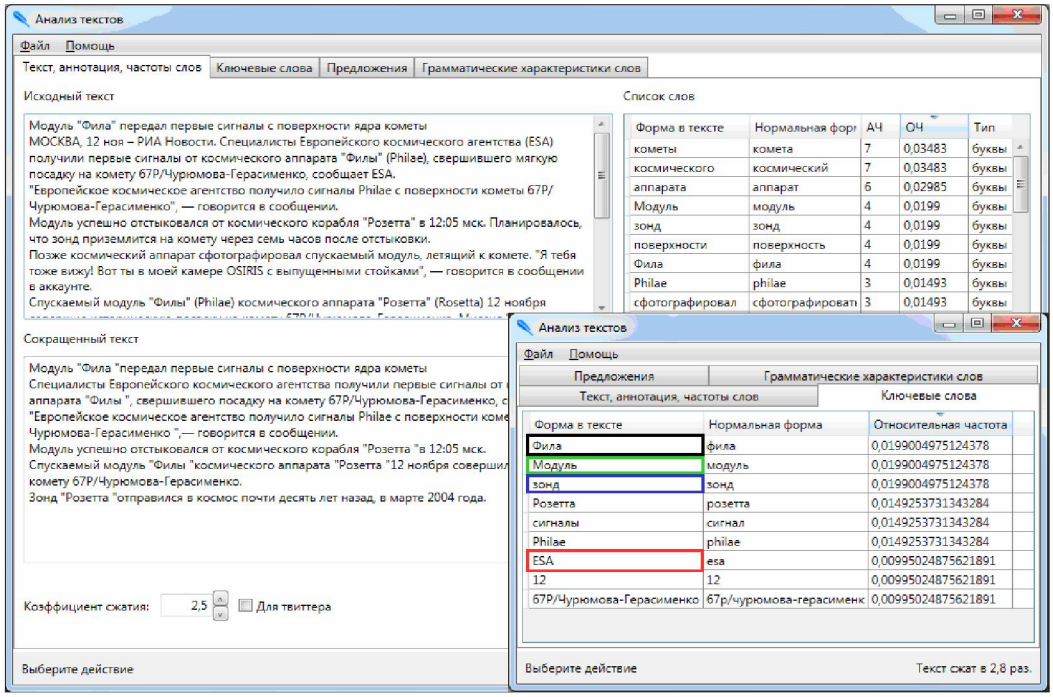

Рис. 3. Результаты построения аннотации текста

Вычисляются ключевые сущности (иначе – тематические узлы новости). Они выделены цветом на рисунке 4. Далее строится иерархия ключевых слов, и на основе этой иерархии отображается mind map исходной статьи [10, 17]:

Рис. 4. Представление содержания текста в виде mind map

Заключение. Исходя из результатов сравнения скорости разделения текстов по тематикам, можно сделать вывод, что в зависимости от вида решаемой задачи и используемых аппаратных технологий, можно применять различные параллельные реализации алгоритма шинглов. При этом на наших тестовых выборках технология OpenCL значительно превосходила другие варианты.

Использование автоматизированной методики извлечения ключевых слов позволило повысить качество обработки новостных Интернет-статей. Отметим, что для параметра времени при автоматизированной обработке учитывается не только непосредственно время анализа системой, но и время, необходимое для окончательной корректировки текстов. Качество результата же оценивалось по таким критериям: сохранение ключевых фактов; связность ключевых сущностей; сохранение синтаксической структуры текста после удаления незначащих частей. Каждый из названных критериев оценивался экспертами по шкале от 0 до 10 баллов, затем для оценки качества (адекватности) извлеченных ключевых сущностей аннотации находилось среднее арифметическое для перечисленных трех показателей по каждому тексту. В итоге по сравнению с ручным способом, время определения ключевых сущностей уменьшилось как минимум в 2 раза, а качество обработки новостей осталось на том же уровне, как и при анализе текста человеком.

Данная работа может быть использована для адаптации текстовой информации новостного потока образовательной организации для людей с ограниченными возможностями здоровья по зрению.

Список литературы

1. Автоматизация составления портретных изображений по естественно-языковому описанию / Ю.А. Орлова, А.В. Долбин, Е.В. Кипаева, В.Л. Розалиев // Известия ВолгГТУ. Сер. Актуальные проблемы управления, вычислительной техники и информатики в технических системах. - Волгоград, 2015. - № 2 (157). - С. 71-76.

2. Автоматизированный подход к определению авторства текста / А.В. Муха, В.Л. Розалиев, Ю.А. Орлова, А.В. Заболеева-Зотова // Известия ВолгГТУ. Серия "Актуальные проблемы управления, вычислительной техники и информатики в технических системах". Вып. 17 : межвуз. сб. науч. тр. / ВолгГТУ. - Волгоград, 2013. - № 14 (117). - С. 51-54.

3. Алгоритм шинглов для веб-документов, поиск нечетких дубликатов текстов, сравнение текстов на похожесть [Электронный ресурс]. – Режим доступа: http://www.codeisart.ru/part-1-shingles-algorithm-for-web-documents/

4. Васьковский, Е.Ю. Системный анализ вопросов, связанных с востребованностью информации на web-сайтах / Е.Ю. Васьковский, Ю.М. Брумштейн // Прикаспийский журнал: управление и высокие технологии. – 2015. – № 1. – С. 59–74.

5. Заболеева-Зотова, А.В. Автоматизация семантического анализа текста технического задания: монография / А.В. Заболеева-Зотова, Ю.А. Орлова. – Волгоград: ИУНЛ, 2010. – 155 с

6. Зеленков, Ю.Г. Сравнительный анализ методов определения нечетких дубликатов для Web-документов / Ю.Г. Зеленков, И.В. Сегалович // Труды IX Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2007. – 9 с.

7. Ландэ, Д.В. Использование графов горизонтальной видимости для выявления слов, определяющих информационную структуру текста / Д. В. Ландэ, А. А. Снарский, Е. В. Ягунова // Труды XV Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2013, Ярославль, 14-17 октября 2013 г. – 7 с.

8. Солошенко, А.Н. Автоматизация реферирования новостных Интернет-текстов / А.Н. Солошенко, Ю.А. Орлова, В.Л. Розалиев // Известия ВолгГТУ. Серия "Актуальные проблемы управления, вычислительной техники и информатики в технических системах". Вып. 18 : межвуз. сб. науч. тр. / ВолгГТУ. - Волгоград, 2013. - № 22 (125). - С. 81-86.

9. Смирнова, М.О. Демонстрация процесса поиска информации с использованием хеширования/ Смирнова М.О. // Прикаспийский журнал: управление и высокие технологии. - 2011. - № 3. - С. 25-30.

10. Усталов, Д.А. Извлечение терминов из русскоязычных текстов при помощи графовых моделей // Теория графов и приложения = Graphs theory and applications : материалы конференции. – 2012. – С. 62–69

11. Automatic identification of time and space categories in the natural language text / А.С. Дмитриев, А.В. Заболеева-Зотова, Ю.А. Орлова, В.Л. Розалиев // Applied Computing 2013 : proceedings of the IADIS International Conference (Fort Worth, Texas, USA, October 23-25, 2013) / IADIS (International Association for Development of the Information Society), UNT (University of North Texas). – [Fort Worth (Texas, USA)], 2013. – P. 187-190.

12. Anisimov, A.V. A method for the computation of the semantic similarity and relatedness between natural language words / A.V. Anisimov, O.O. Marchenko, V.K. Kysenko // Cybernetics and Systems Analysis, July 2011. – Volume 47, Issue 4. - Pp. 515-522

13. Broder, A. On the resemblance and containment of documents. Compression and Complexity of Sequences / A. Broder // SEQUENCES'97, IEEE Computer Society, 1998, Pp. 21-29

14. Establishing Semantic Similarity of the Cluster Documents and Extracting Key Entities in the Problem of the Semantic Analysis of News Texts / А.Н. Солошенко, Ю.А. Орлова, В.Л. Розалиев, А.В. Заболеева-Зотова // Modern Applied Science. - 2015. - Vol. 9, No. 5. - С. 246-268.

15. Furu Wei. A document-sensitive graph model for multi-document summarization / Furu Wei, Wenjie Li, Qin Lu, Yanxiang He // Knowledge and Information Systems, February 2010. Volume 22, Issue 2. – Pp. 245-259

16. Processing of Spatial and Temporal Information in the Text / А.С. Дмитриев, А.В. Заболеева-Зотова, Ю.А. Орлова, В.Л. Розалиев // World Applied Sciences Journal (WASJ). - 2013. - Vol. 24, Spec. Issue 24 : Information Technologies in Modern Industry, Education & Society. - С. 133-137.

17. Sheng-Tun Li. Constructing tree-based knowledge structures from text corpus / Sheng-Tun Li, Fu-Ching Tsai // Applied Intelligence, August 2010. – Volume 33, Issue 1. – Pp. 67-78

References

1. Automation of constructing iconic imagery in natural language description / Y.A. Orlova, A.V. Dolbin, E.V. Kipaeva, V.L. Rozaliev // Izvestia VolgSTU. Ser. Actual problems of control, computer science and Informatics in technical systems. - Volgograd, 2015. - № 2 (157). - C. 71-76.

2. An automated approach for determining authorship of text / V.A. Mucha, V.L. Rozaliev, Y.A. Orlova, A.V. Zaboleeva-Zotova // Proceedings of VSTU. Series "Actual problems of control, computer science and Informatics in technical systems". Vol. 17 : Intercollege. SB. nauch. Tr. / VolgSTU. - Volgograd, 2013. - № 14 (117). - pp. 51-54.

3. Algoritm shinglov dlja veb-dokumentov, poisk nechetkih dub-likatov tekstov, sravnenie tekstov na pohozhest' (Shingles algorithm for web document retrieval, near-duplicate text finding, comparing the similarity of texts) Available at: http://www.codeisart.ru/part-1-shingles-algorithm-for-web-documents

4. Vaskovsky, E. Y. Systematic analysis of the issues related to the relevance of information on the web sites / E. Vaskovsky Y., Y. M. Bromstein // Caspian journal: management and high technologies. – 2015. – No. 1. – S. 59-74.

5. Zaboleeva-Zotova A.V., Orlova Ju.A. Avtomatizacija semanticheskogo analiza teksta tehnicheskogo zadanija: monografija (Automation of technical specifications text semantic analysis: a monograph). Volgograd, 2010, 155 p.

6. Zelenkov Ju.G., Segalovich I.V. Comparative analysis of Web-documents near-duplicate detection methods [Sravnitel'nyj analiz metodov opredelenija nechetkih dublikatov dlja Web-dokumentov].Trudy IX Vserossijskoj nauchnoj konferencii «Jelektronnye biblioteki: perspektivnye metody i tehnologii, jelektronnye kollekcii» RCDL'2013 (Proceedings of the IX Russian Scientific Conference RCDL'2007). 2007, 9 p.

7. Landje D.V., Snarskij A.A., Jagunova E.V.Using horizontal visibility graphs to identify words defining information structure of the text [Ispol'zovanie grafov gorizontal'noj vidimosti dlja vyjavlenija slov, opredeljajushhih informacionnuju strukturu teksta]. Trudy XV Vserossijskoj nauchnoj konferencii «Jelektronnye biblioteki: perspektivnye metody i tehnologii, jelektronnye kollekcii» RCDL'2013 (Proceedings of the XV Russian Scientific Conference RCDL'2013). Yaroslavl, 2013, 7 p.

8. Soloshenko, A. N. Automate the summarization of online news texts / A. N. Soloshenko, Y.A. Orlova, V.L. Rozaliev // From-vestia of VSTU. Series "Actual problems of management, computing engineering and Informatics in technical systems". Vol. 18 : Intercollege. SB. nauch. Tr. / Volgstu. - Volgograd, 2013. - № 22 (125). - Pp. 81-86.

9. Smirnova M.O. Demonstration of information retrieval using hashing. Prikaspiiyskiiy zhurnal: upravlenie i visokie tekhnologii. 2011. № 3. pp. 25-30.

10. Ustalov D.A. Term extraction from Russian texts using the graph models [Izvlechenie terminov iz russkojazychnyh tekstov pri pomoshhi grafovyh modelej]. Teorija grafov i prilozhenija : materialy konferencii (Graphs theory and applications : conference materials). 2012, pp. 62-69.

11. Automatic determination of time and place categories in NAT-Ural language text / A. S. Dmitriev, A. V. Zaboleeva-Zotova, Y.A. Orlova, V. L. Rozaliev // applied computing 2013 : proceedings of the IADIS international conference (Fort worth, Texas, USA, 23-25 October 2013) / IADIS (international Association for development of information society-tai), UNT (University of North Texas). – [Fort worth (Texas, USA)], 2013. – P. 187-190.

12. Anisimov, A.V., Marchenko, O.O., & Kysenko, V.K. (2011). A method for the computation of the semantic similarity and relatedness between natural language words. Cybernetics and Systems Analysis, 47 (4), 515-522.

13. Broder, A. On the resemblance and containment of documents. Compression and Complexity of Sequences / A. Broder // SEQUENCES'97, IEEE Computer Society, 1998, Pp. 21-29

14. Establishing Semantic Similarity of the Cluster Documents and Extracting Key Entities in the Problem of the Semantic Analysis of News Texts / A. N. Soloshenko, Y.A. Orlova, V.L. Rozaliev, A.V. Zaboleeva-Zotova // Modern Applied Science. - 2015. - Vol. 9, No. 5. - C. 246-268.

15. Furu Wei, Wenjie Li, Qin Lu, & Yanxiang He (2010). A document-sensitive graph model for multi-document summarization. Knowledge and Information Systems, 22 (2), 245-259.

16. Processing spatial-temporal information in the text / A.S. Dmitriev, A.V. Zaboleeva-Zotova, Y.A. Orlova, V.L. Rozaliev // world applied Sciences journal (WASJ). - 2013. - Vol. 24. spec. Question 24 : information technologies in modern industry, education and society. - S. 133-137

17. Sheng-Tun Li, & Fu-Ching Tsai (2010). Constructing tree-based knowledge structures from text corpus. Applied Intelligence, 33 (1), 67-78.