НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
июль–август 2018 Том 18 № 4 ISSN 2226-1494 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
July–August 2018 Vol. 18 No 4 ISSN 2226-1494 http://ntv.ifmo.ru/en

УДК 004.89

КОМПОЗИЦИЯ АЛГОРИТМОВ ТЕМАТИЧЕСКОЙ СЕГМЕНТАЦИИ ТЕКСТОВ КАК СРЕДСТВО ИНТЕЛЛЕКТУАЛИЗАЦИИ ПРОЕКТИРОВАНИЯ ТЕХНИЧЕСКИХ СИСТЕМ

Н.В. Добренко[a]

[a] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
Адрес для переписки: Graziokisa@yandex.ru

Информация о статье
Поступила в редакцию 18.04.18, принята к печати 18.05.18
doi: 10.17586/2226-1494-2018-18-4-690-694
Язык статьи – русский

Ссылка для цитирования: Добренко Н.В. Композиция алгоритмов тематической сегментации текстов как средство интеллектуализации проектирования технических систем // Научно-технический вестник информационных технологий, механики и оптики. 2018. Т. 18. № 4. С. 690–694. doi: 10.17586/2226-1494-2018-18-4-690-694

Аннотация
Рассматривается задача тематической сегментации протяженных текстов для поддержки работы проектировщика технических систем. На примере показано, что разные алгоритмы сегментации выделяют содержательно разные фрагменты текста, и композиция алгоритмов в классической форме, т.е. путем суммирования результатов с целью выделения одного наилучшего, представляется неправомерной. В то же время одновременная демонстрация нескольких вариантов тематической сегментации позволит читателю получить интегральное представление о структуре текста, облегчив тем самым выбор эффективной стратегии освоения текста. Описана построенная система визуализации тематической сегментации протяженных текстов, позволяющая пользователю выделять и анализировать не весь текст целиком, а только фрагменты, соответствующие его текущим информационным потребностям. Система позволяет одновременно просматривать результаты сегментации текста, выполняемые различными алгоритмами. Тем самым расширяются возможности пользователя по оперативному и эффективному анализу и освоению большого объема текстовой информации.

Ключевые слова
тематическая сегментация, композиция алгоритмов, система визуализации

Благодарности
Работа выполнена при поддержке НИР-ФУНД 617042 в Университете ИТМО.

ALGORITHM COMPOSITION OF TEXT THEMATIC SEGMENTATION AS INTELLECTUALIZATION INSTRUMENT FOR DESIGN OF TECHNICAL SYSTEMS

N.V. Dobrenko[a]

[a] ITMO University, Saint Petersburg, 197101, Russian Federation
Corresponding author: Graziokisa@yandex.ru

Article info
Received 18.04.18, accepted 18.05.18
doi: 10.17586/2226-1494-2018-18-4-690-694
Article in Russian

For citation: Dobrenko N.V. Algorithm composition of text thematic segmentation as intellectualization instrument for design of technical systems. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2018, vol. 18, no. 4, pp. 690–694 (in Russian). doi: 10.17586/2226-1494-2018-18-4-690-694

Abstract
The paper considers the problem of thematic segmentation of extended texts aimed at the support of technical systems designer operation. The example shows that different segmentation algorithms allocate meaningfully different text fragments, and the composition of algorithms in a classical form, that is, by summarizing the results in order to single out the best one, seems to be wrong. At the same time, the simultaneous demonstration of several versions of the thematic segmentation enables the reader to obtain an integral representation of the text structure, thereby facilitating the choice of an effective

strategy for mastering the text. The created system of thematic segmentation visualization of extended texts is described, providing the user to select and analyze not the whole text, but only fragments corresponding to his current information needs. The system gives the possibility to view simultaneously the results of text segmentation performed by various algorithms. Thus, the user's abilities for quick and efficient analysis and capturing of a large amount of textual information are enhanced.

Keywords
thematic segmentation, algorithm composition, visualization system

Acknowledgements
The work was supported by SRR-FUND 617042 in ITMO University.

Проектировщик современных технических систем сталкивается с необходимостью оперативного освоения максимально широкого спектра научной и технической информации, релевантной конкретной задаче проектирования. Для интеллектуализации этого процесса может быть использована тематическая сегментация (ТС) [1], которая позволяет пользователю анализировать не все найденные документы, а только фрагменты, содержащие релевантную информацию.

Роль ТС при изучении текстового материала вытекает из теоретического анализа структуры и моделей понимания текста [2, 3]. Согласно [3], автор при написании текста формирует его структуру в виде иерархии макро- и микропропозиций, каждой из которых соответствует отдельный сегмент (топик) текста. Читатель пытается воспроизвести структуру топиков, которую заложил автор, и выделить ту комбинацию топиков, которая соотносится с его индивидуальными целями и ситуационными моделями. В этом случае у читателя формируется понимание (интерпретация) текста. Очевидно, что эффективная ТС может предложить читателю исходную структуру топиков в качестве опорной, чтобы он смог построить свою интерпретацию с минимальными затратами ресурсов – выбрать для чтения то, что нужно, или убрать то, что заведомо не нужно.

Объектом исследования в настоящей работе являются тексты, относящиеся к жанру научной прозы, такие как монографии, учебники, научные статьи, тематикой которых являются аспекты технологий различных предметных областей. Решение задачи ТС применительно к этим текстам имеет свою специфику – малый объем выборки, сравнительно большая протяженность, разнообразие языков оригинала, тематическое единство с плавными переходами от одного топика к другому, сложная структура топиков.

Для решения задачи ТС в литературе представлен широкий спектр подходов [4–6]. В наших предыдущих исследованиях [7–10] проводилась сравнительная оценка эффективности различных алгоритмов машинного обучения для ТС протяженных текстов. В качестве примера в табл. 1 приведены характеристики качества ТС текста [11] алгоритмами, показавшими наибольшую эффективность. Они представляют весь спектр подходов к ТС:

- анализ текста как линейной последовательности векторов слов для каждого абзаца (алгоритм TextTiling);
- анализ текста как линейной последовательности векторов слов для каждого абзаца, соответствующих топикам, выделенным в результате сингулярного разложения матрицы «слово–документ» (алгоритм LSA (Latent semantic analysis));
- выявление в тексте групп абзацев, соответствующих заданному числу тем, на основе порождающих моделей текста (алгоритм LDA (Latent Dirichlet Allocation)).

Алгоритм	Precision	Recall	F-мера
LSA	0,447	0,636	0,539
TextTiling	0,833	0,556	0,667
LDA	0,438	0,918	0,596

Таблица 1. Характеристики качества тематической сегментации текста

Характеристики текста и использованных алгоритмов представлены в табл. 2.

Размер текста	60282 печ. зн., 108 абзацев
Размер абзаца	13–764 слов
Язык текста	старофранцузский
Алгоритмы	LSA и LDA – реализованы с применением пакета genism языка Python, TextTiling – реализован самостоятельно; оптимальное число тем n=3 для алгоритма LDA определено в соответствии с методикой [4]

Таблица 2. Характеристики текста и использованных алгоритмов

Результаты работы отобранных алгоритмов визуально представлены на рис. 1 в сопоставлении с границами топиков, которые были исходно заданы автором, а также с экспертной разметкой.

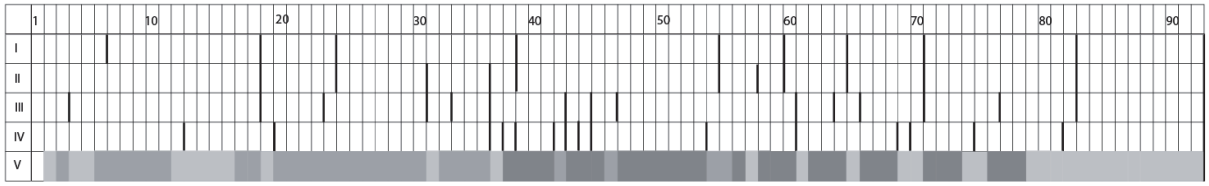

Рис. 1. Сопоставление вариантов сегментации текста: шапка – номера абзацев, строка I – авторская сегментация, строка II – экспертная сегментация, строка III – сегментация алгоритмом LSA, строка IV – сегментация алгоритмом TextTiling, строка V – сегментация алгоритмом LDA; для строк I–IV указаны границы топиков, для строки V оттенками серого показана принадлежность абзацев к одной из тем

Результаты табл. 1 показывают, что все отобранные алгоритмы демонстрируют достаточно близкие значения F-меры. Однако анализ рис. 1 показывает, что каждый алгоритм принципиально выполняет ТС по-разному, выделяя те или иные характерные особенности структурной организации текста. Например, для алгоритма TextTiling характерны ошибки сдвига границ, а для алгоритма LDA – мелкие «вкрапления» в текущий сегмент из других сегментов. В то же время алгоритм LDA, несмотря на отдельные вкрапления, демонстрирует реально существующий в тексте переход между тремя основными темами. Алгоритмы LSA и LDA выделяют зоны текста (абзацы 42–46), для которых характерна «списочная» структура (параллельный тип связности).

Таким образом, разные алгоритмы выделяют содержательно разные фрагменты текста, и композиция алгоритмов в классической форме, т.е. путем суммирования результатов с целью выделения одного наилучшего, представляется неправомерной. В то же время одновременная демонстрация нескольких вариантов ТС позволит читателю получить интегральное представление о структуре текста, облегчив тем самым выбор эффективной стратегии освоения текста. Это особенно важно для текстов на чужом для автора языке, где неудачное выделение границ фрагмента, подлежащего переводу, приводит к большим затратам временных и интеллектуальных ресурсов.

Таким образом, возникает задача визуализации результатов ТС анализируемого протяженного текста. Анализ литературы (см. [12, 13], а также обзор[1]) показывает, что средства визуализации в настоящее время находят все более широкое применение для интеллектуальной поддержки различных технологических процессов. В рамках настоящей работы разработана система ТС текстов, в которую входит утилита для визуализации результатов ТС. Архитектура системы в нотации UML в виде диаграммы развертывания представлена на рис. 2.

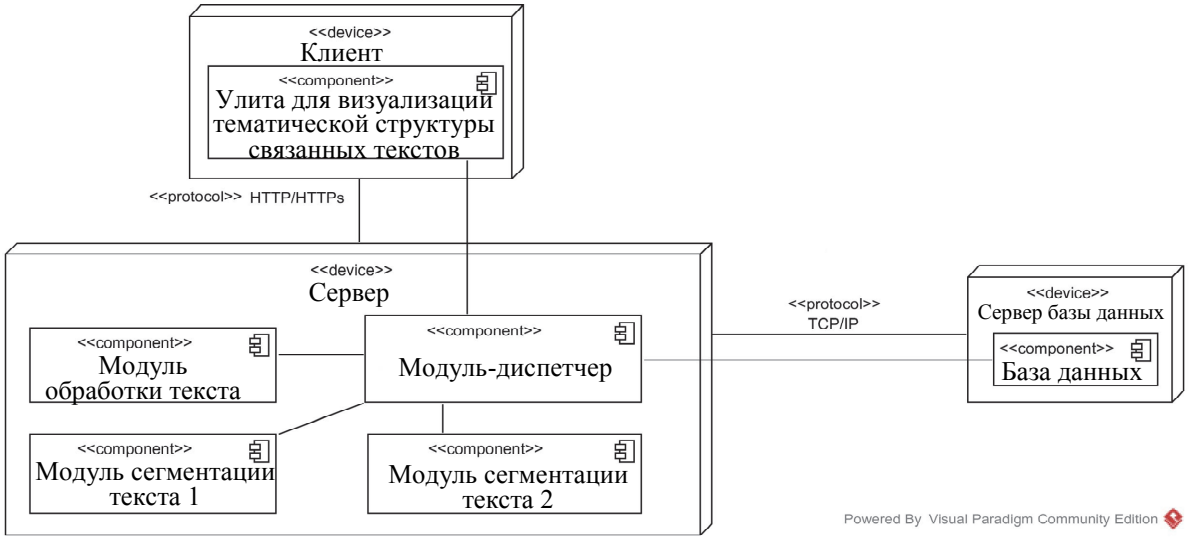

Рис. 2. Система тематической сегментации текстов с визуализацией результатов

Система построена по трехзвенной архитектуре с разнесением слоя сервера приложений на два узла. Клиент построен в виде Web-приложения. Web-клиент выполняет стандартную задачу интерфейсного ввода–вывода. На клиенте установлена утилита для визуализации тематической структуры связных текстов. В слой сервера приложений входят основной сервер и сервер базы данных.

[1] http://textvis.lnu.se

Архитектура сервиса является модульной, что позволяет добавлять новые алгоритмы машинного обучения. В состав сервера входят модуль-диспетчер, модуль предобработки текста, и также расширяемый набор модулей сегментации текста.

Модуль-диспетчер реализует функционал веб-сервера в аспекте обработки запросов на анализ текста и хранение результатов работы модулей сегментации текстов с целью предоставления их пользователям. В модуле предобработки предусмотрены различные варианты предобработки текста, в том числе лемматизация, удаление стоп-слов, отбор существительных, объединение коротких абзацев. Хранение анализируемых текстов и результатов их обработки происходит в базе данных. В модуле сегментации текста на основе алгоритма LDA автоматически выполняется формирование топа ключевых слов, который предъявляется пользователю в окне с ключевыми словами, что облегчает пользователю отбор интересующих его фрагментов текста. При сегментации текста другими алгоритмами для выделения ключевых слов может быть предусмотрен отдельный модуль (на рисунке не показан).

На рис. 3 представлен интерфейс визуализации тематической структуры связных текстов. Спецификой утилиты является работа с текстом не как с последовательностью ACII-символов, а как с последовательностью букв как графических символов, что позволяет избежать преобразования исходного текста в формат txt. Каждый абзац представляется как прямоугольный объект шириной 500 px и высотой, соответствующей длине абзаца как текстовой строки. Текущее положение в тексте задается пользователем путем клика по тексту и рассчитывается с точностью до абзаца. Для начальной отрисовки границ абзацев текста на линейке используется расчет пропорций всех абзацев, и происходит отрисовка границ на линейке статичной высоты. При нажатии пользователем на участок текста данные о позиции курсора сохраняются в глобальное хранилище. Все линейки, количество которых не ограничено, реализованы с помощью технологии реактивного программирования. Каждый компонент отрисовки линейки подключен к глобальному хранилищу, и при изменении в нем значения позиции курсора компонент перерисовывает курсор. Таким образом, абзац текста представляется как прямоугольный объект, имеющий некоторую высоту на экране, и справа к нему привязаны результаты сегментации, отображаемые на линейках. Такая реализация модуля визуализации позволяет связать позицию курсора с позицией абзаца независимо от прокрутки текста, т.е. курсор сдвигается одновременно с абзацем при прокрутке текста.

Таким образом, на стороне клиента (на экране визуализатора) появляется возможность демонстрировать одновременно несколько результатов сегментации (в виде линеек) для одного текста.

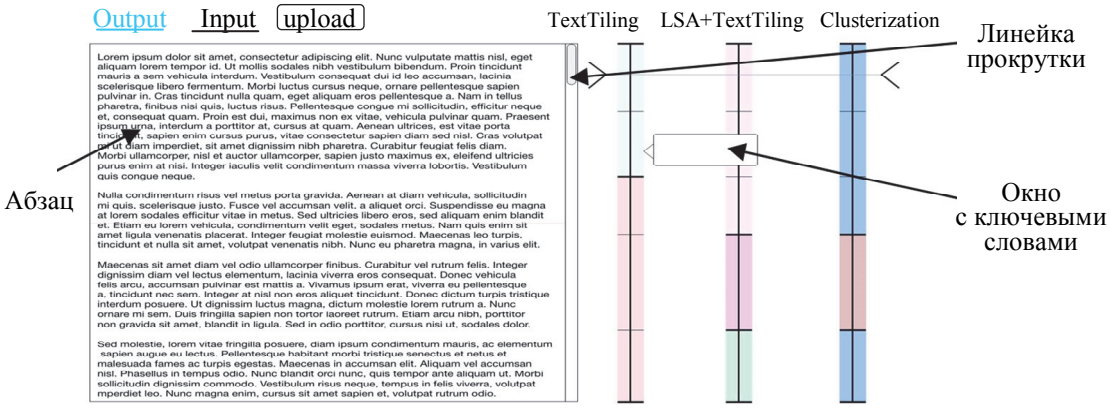

Рис. 3. Интерфейс поддержки визуальной композиции

Работа системы организована по следующему сценарию. Пользователь загружает текст, подлежащий сегментации, в систему, и выбирает желаемые алгоритмы сегментации. Система выполняет сегментацию и сохраняет ее результаты в базе данных. В текстовом окне интерфейса пользователю демонстрируется выбранный текст, который можно двигать с помощью линейки прокрутки. При клике на выбранный абзац активизируются линейки сегментации, соответствующие выбранным алгоритмам, и на них отображаются границы топиков, в которые входит выбранный абзац. Это позволяет пользователю более точно отбирать фрагменты текста, подлежащие изучению. Двигая текст посредством линейки прокрутки, пользователь может расширить зону анализа вплоть до границ текста. Результаты анализа сохраняются в базе данных и могут быть вызваны повторно.

Построенная система визуализации ТС протяженных текстов в явной форме демонстрирует семантическую структуру текста, что позволяет пользователю выделять и анализировать не весь текст целиком, а только фрагменты, соответствующие его текущим информационным потребностям. Тем самым расширяются возможности пользователя по оперативному и эффективному анализу и освоению большого объема текстовой информации.

Литература

1. Jurafsky D., Martin J.H. Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. Pearson Prentice Hall, 2009. 988 p.
2. Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. СПб.: Питер, 2000. 384 с.
3. ван Дейк Т.А., Кинч В. Статегии понимания связного текста. М., 1988.
4. Vorontsov K.V., Potapenko A.A. Additive regularization of topic models // Machine Learning. 2014. V. 101. N 1-3. P. 303–323. doi: 10.1007/s10994-014-5476-6
5. Boyd-Graber J., Chang J., Gerrish S., Wang C., Blei D. Reading tea leaves: how humans interpret topic models // Proc. 23rd Annual Conference on Neural Information Processing Systems (NIPS). Vancouver, Canada, 2009. P. 288–296.
6. Liu L., Tang L., Dong W., Yao S., Zhou W. An overview of topic modeling and its current applications in bioinformatics // SpringerPlus. 2016. V. 5. P. 1608. doi: 10.1186/s40064-016-3252-8
7. Боярский К.К., Гусарова Н.Ф., Добренко Н.В., Каневский Е.А., Авдеева Н.А. Исследование специфики применения алгоритмов тематической сегментации для научных текстов // Аналитика и управление данными в областях с интенсивным использованием данных. 2015. С. 181–189.
8. Бурая К.И., Грозин В.А., Гусарова Н.Ф., Добренко Н.В. Методы машинного обучения для выделения профессионально значимой информации из веб-форумов // Дистанционное и виртуальное обучение. 2015. № 12(102). С. 46–63.
9. Бурая К.И., Виноградов П.Д., Грозин В.А., Гусарова Н.Ф., Добренко Н.В., Трофимов В.А. Автоматическая суммаризация веб-форумов как источников профессионально значимой информации // Научно-технический вестник информационных технологий, механики и оптики. 2016. Т. 16. № 3(103). С. 482–496. doi: 10.17586/2226-1494-2016-16-3-482-496
10. Grozin V.A., Dobrenko N.V., Gusarova N.F., Ning T. The application of machine learning methods for analysis of text forums for creating learning objects. // Proc. Int. Conf. on Computational Linguistics and Intellectual Technologies. Moscow, 2015. V. 1. N 14. P. 202–213.
11. Ромме М. L' Art de la Marine, ou Principes et Préceptes Generaux de l'Art de Construire, d'Armer , de Manœuvrer et de Conduire des Vasseaux. La Rochelle, 1787. Chapitre VII.
12. Айсина Р.М. Обзор средств визуализации тематических моделей коллекций текстовых документов // Машинное обучение и анализ данных. 2015. Т. 1. № 11. С. 1584–1618.
13. Янина А.О., Воронцов К.В. Мультимодальные тематические модели для разведочного поиска в коллективном блоге // Интеллектуализация обработки информации. Тезисы докладов 11-й Международной конференции. Москва, 2016. С. 186–187.

References

1. Jurafsky D., Martin J.H. *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. Pearson Prentice Hall, 2009, 988 p.
2. Gavrilova T.A., Khoroshevskii V.F. *Knowledge Base of Intelligent Systems*. St. Petersburg, Piter Publ., 2000, 384 p. (in Russian)
3. Van Dijk T.A., Kintsch W. *Strategies of Discourse Comprehension*. NY, Academic Press, 1983, 423 p.
4. Vorontsov K.V., Potapenko A.A. Additive regularization of topic models. *Machine Learning*, 2014, vol. 101, no. 1-3, pp. 303–323. doi: 10.1007/s10994-014-5476-6
5. Boyd-Graber J., Chang J., Gerrish S., Wang C., Blei D. Reading tea leaves: how humans interpret topic models. *Proc. 23rd Annual Conference on Neural Information Processing Systems, NIPS*. Vancouver, Canada, 2009, pp. 288–296.
6. Liu L., Tang L., Dong W., Yao S., Zhou W. An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, 2016, vol. 5, pp. 1608. doi: 10.1186/s40064-016-3252-8
7. Boyarskii K.K., Gusarova N.F., Dobrenko N.V., Kanevskii E.A., Avdeeva N.A. Specifics of applying topic segmentation algorithms to scientific texts. *Analitika i Upravlenie Dannymi v Oblastyakh s Intensivnym Ispol'zovaniem Dannykh*, 2015, pp. 181–189. (in Russian)
8. Buraya K.I., Grozin V.A., Gusarova N.F., Dobrenko N.V. Machine learning methods for extracting of professionally significant information from web forums. *Distantsionnoe i Virtual'noe Obrazovanie*, 2015, no. 12, pp. 46–63. (in Russian)
9. Buraya K.I., Vinogradov P.D., Grozin V.A., Gusarova N.F., Dobrenko N.V., Trofimov V.A. Automatic summarization of web forums as sources of professionally significant information. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2016, vol. 16, no. 3, pp. 482–496. (in Russian) doi: 10.17586/2226-1494-2016-16-3-482-496
10. Grozin V.A., Dobrenko N.V., Gusarova N.F., Ning T. The application of machine learning methods for analysis of text forums for creating learning objects. *Proc. Int. Conf. on Computational Linguistics and Intellectual Technologies*. Moscow, 2015, vol. 1, no. 14, pp. 202–213.
11. Ромме М. L' *Art de la Marine, ou Principes et Préceptes Generaux de l'Art de Construire, d'Armer, de Manœuvrer et de Conduire des Vasseaux*. La Rochelle, 1787. Chapitre VII.
12. Aysina R.M. Survey of visualization tools for topic models of text corpora. *Machine Learning and Data Analysis*, 2015, vol. 1, no. 11, pp. 1584–1618. (in Russian)
13. Ianina A.O., Vorontsov K.V. Multimodal topic modeling for exploratory search in collective blog. *Proc. 11th Int. Conf. on Intelligent Data Processing: Theory and Applications*. Moscow, 2016, pp. 186–187. (in Russian)

Авторы

Добренко Наталья Викторовна – ассистент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, Scopus ID: 56499375200, ORCID ID: 0000-0001-6206-8033, Graziokisa@yandex.ru

Authors

Natalia V. Dobrenko – assistant, ITMO University, Saint Petersburg, 197101, Russian Federation, Scopus ID: 56499375200, ORCID ID: 0000-0001-6206-8033, Graziokisa@yandex.ru