К. Р. Пиотровская

ТЕКСТ-МАЙНИНГ: ПЕРСПЕКТИВЫ РАЗВИТИЯ

Обсуждается перспектива интеграции современных систем текст-майнинга открытого доступа и языка программирования R на примере пакета tm.

Ключевые слова: текст-майнинг, дата-майнинг, язык программирования R, кластеризация, категоризация, автоматическое реферирование.

X. Piotrowska

A Survey of Text mining

*We give a survey on text mining facilities in R and explain how typical application tasks can be carried out using our framework. We present the **tm** package, which provides a framework for text mining applications within R.*

Keywords: text-mining, R, count-based evaluation, text clustering, text classification, summarization.

В последнее десятилетие наблюдается неуклонный рост исследований в области лингвистического анализа и переработки текста, например в области стилеметрии [11; 21; 12; 2], где с помощью статистических методов исследуется авторский стиль в норме и патологии, или в поисковых системах при изучении рейтинга новостных документов, материалов блогов, для моделирования поведения пользователя [23; 1].

Популярность последнего возросла с обнаружением большого количества ценной информации, скрытой от непосредственного наблюдения в текстах, которые оказались недоступны в классических структурированных форматах данных, а успехи электронного документооборота способствовали появлению новых концепций для автоматической обработки текстов. Постепенно наиболее инновационные методы структурного и квантитативного лингвистического анализа текста пополняют инструментарий нового направления исследований — текст-майнинг (ТМ), который представляет широкую перспективу теоретических подходов и методов для обработки входной текстовой информации и является междисциплинарной областью научной деятельности на стыке дата-майнинга, автоматической переработки текста, описательной статистики и информатики.

Сегодня практически каждый статистический компьютерный продукт обладает чертами ТМ, а многие известные дата-майнинг-пакеты предлагают решения по задачам автоматической переработки текста. Основные возможности этих систем сводятся к следующим шести задачам [8]: предварительная обработка данных, ассоциативный анализ, классификация, кластеризация, категоризация, автоматическое реферирование. Важной для ТМ-систем является API (Application Programming Interfaces) — характеристика, которая показывает перспективы расширения возможностей системы с помощью плагинов.

Предварительная обработка данных призвана выполнить требования по качеству и структуре данных, предъявляемые к репозиториям знаний, и заключается в преобразовании неструктурированного пользовательского запроса или текста в структурированный формат с помощью процедур стемминга, удаления стоп-слов и приведения регистра. Затем преобразованный в структурированный формат текст сравнивается с текстами, поступающими из базы данных. Одним из ключевых форматов здесь является формат RDF (Resource

Description Framework) [17]. Решение задач стандартизации форматов документов и выполнение семантических операций над ними способствовали появлению нового направления, получившего название «Web-семантика» [5].

Ассоциативный, или латентно-семантический, анализ текста заключается в выделении ключевых слов и идентификации ассоциативных отношений между ключевыми словами и дескрипторами и базируется на инструментарии факторного анализа.

Кластеризация в ТМ рассматривается как процесс выделения компактных подгрупп объектов с близкими свойствами: система должна самостоятельно найти признаки и разделить объекты по подгруппам. Результатом кластеризации является таксономия, или визуальная карта, которая обеспечивает эффективный охват больших объемов данных. Результаты кластеризации применяются при реферировании больших массивов текстов, при определении взаимосвязанных групп документов, например, новостных документов, e-mail фильтрации и автоматической маркировке документов в бизнес-библиотеках, при упрощении процесса просмотра при поиске необходимой информации, при нахождении уникальных документов из коллекции и выявлении дубликатов или очень близких по содержанию документов [27].

Поскольку кластеризация разрешает определить группы объектов, то, как правило, она предшествует категоризации, при решении которой используются статистические корреляции для построения правил размещения документов в определенные категории, причем в связи с большим объемом количества объектов и их атрибутов требуются интеллектуальные механизмы оптимизации. Этот процесс применяется при решении таких задач, как, например, группировка документов в internet-сетях и на web-сайтах, размещение документов в определенные папки, сортировка сообщений электронной почты, избирательное распространение новостей подписчикам.

Автоматическое реферирование заключается в составлении коротких изложений материалов и аннотаций и обычно опирается на два направления автоматического реферирования — квазиреферирование и краткое изложение семантики текста. Квазиреферирование основано на выявлении наиболее информативных фраз в документе и формировании из них рефератов с помощью статистических методов, базирующихся на оценке информативности различных элементов текста по частоте их появления в тексте, с помощью позиционного метода или маркеров содержательности. Альтернативный метод базируется на автоматизации синтаксического разбора предложений, построения семантической матрицы и выделения семантически связных групп предложений.

Статистический контекст ТМ в научных исследованиях, в образовании и в бизнес-решениях имеет широкий спектр применения и включает использование статистических методов для автоматизации юрисдикции [10], техники латентно-семантического анализа в биоинформатике [9], определения плагиата, кросс-языкового автоматизированного поиска [16] или построения адаптивных спам-фильтров.

В таблице ниже дается перечень наиболее часто используемых текст-майнинговых коммерческих систем и систем открытого доступа, а также реализованных в них основных операций [22].

К коммерческим системам относятся: текст-управляемая бизнес-аналитика **Clearforest**, софт для извлечения ключевых понятий и релевантных предложений **Copernic Summarizer**; средство для поиска документов, инструменты для текст-майнингового поиска и анализа **dtSearch**, **Insightful Infact**, **Inxight**; рабочее место для дата- и текст-майнинга SPSS **Clementine**; комплект для извлечения знаний в тексте SAS **Text Miner**; средство для

кластеризации и категоризации текстов **TEMIS**; сервис для статистического анализа текста **WordStat**.

Из данных, приведенных в таблице, видно, что большинство коммерческих реализаций не поддерживает расширений плагинами, что приводит к монолитной, жесткой структуре.

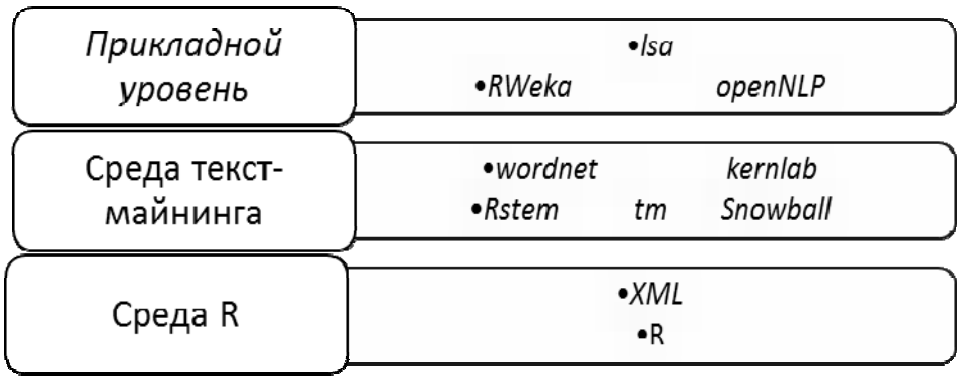
ТМ-системы и реализованные в них функции анализа текста

Программный продукт	Preprocess	Associate	Cluster	Summarize	Categorize	API
Коммерческие программы						
Clearforest	+	+	+	+		
Copernic **Summarizer dtSearch**	+			+		
Insightful **Infact**	+	+		+		
Inxight	+	+	+	+	+	+
SPSS **Clementine**	+	+	+	+	+	+
SAS **Text Miner TEMIS**	+	+	+	+	+	
WordStat	+	+	+	+	+	
	+	+	+	+	+	
Программы в свободном доступе						
GATE	+	+	+	+	+	+
RapidMiner	+	+	+	+	+	+
Weka/**KEA**	+	+	+	+	+	+
R/**tm**	+	+	+	+	+	+

Среди хорошо известных средств дата-майнинга, находящихся в открытом доступе и поддерживающих операции текст-майнинга, можно назвать такие комплексы, как классификатор **Weka**; **KEA** [26], позволяющий при хорошей API поддержке и широкой пользовательской базе выделять ключевые слова; **GATE** [7], строящий онтологии и алгоритмы машинного обучения; **RapidMiner** [19], система для извлечения знаний и дата-майнинга**.** Тем не менее многие существующие системы открытого доступа при извлечении тех или иных сведений из текста имеют тенденцию к узкоспециализированным решениям. Например, **Shogun** [24], средство для определения строковых ядер, или средство **Bow** [18], система статистического анализа, моделирования и автоматической переработки текста.

Одним из перспективных альтернативных направлений представляется организация гибких интегрированных сред с учетом возможностей ТМ-инфраструктур открытого доступа и статистического многообразия методов, реализованных под среду программирования R — программную среду с открытым исходным кодом, развиваемую в рамках проекта GNU, являющуюся свободным программным обеспечением и получившую широкую популярность среди специалистов, которые занимаются анализом и визуализацией данных. Язык R активно применяется ведущими зарубежными компаниями, такими как Google,

Pfizer, Merck, Bank of America и Shell, а также ведущими университетами мира. Эта программная среда насчитывает более 5000 пакетов расширений для самых разных областей знаний [3; 20]. Его поддержка включена в коммерческие (SPSS, Statistica, Oracle Data Mining) и открытые пакеты программного обеспечения (Gretl). В то же время в отечественной научно-практической литературе вопрос применения языка R для решения различных научных и образовательных задач и, в частности, задач в области прикладной лингвистики, остается нераскрытым. Одним из пакетов, расширяющих возможности среды R в области TM, является пакет **tm** [10]. Он позволяет исследователям применять многочисленные из накопленных методик к текстовым структурам данных. Многие передовые TM-методы, например определение строковых ядер или латентный семантический анализ, могут стать более доступными либо при интеграции с другими пакетами расширений R, таких как **kernlab** [15] или **lsa** [25], либо через интерфейсы естественноязыковой обработки открытого доступа. В частности, Weka позволяет подключить методы стемминга [13] и токенизации, в то время как OpenNLP [14] поможет с разбиением на лексемы и предложения и организует разметку по частям речи. Можно подключить эту функциональность в различных точках инфраструктуры **tm**, например, для предварительной обработки с помощью способов трансформационных методов, для генерации матрицы терминов документа и т. п. Эти возможности обеспечиваются модульной структурой описываемой среды (см. рисунок).



Интеграция R-пакетов и TM-среда

Использование пакета **tm** позволяет организовать более гибкую интеграцию статистических методов, реализованных под среду R и TM-инфраструктур открытого доступа.

СПИСОК ЛИТЕРАТУРЫ

1. *Большакова Е. И., Клышинский Э. С., Ландэ Д. В., Носков А. А., Пескова О. В., Ягунова Е. В.* Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: Учебное пособие. М.: МИЭМ, 2011. 272 с.

2. *Пашковский В. Э., Пиотровская В. Р., Пиотровский Р. Г.* Психиатрическая лингвистика. М.: URSS, 2013. 158 с.

3. A language and environment for statistical computing // R Development Core Team; R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/

4. *Adeva J., Calvo R.* Mining Text with Pimiento // IEEE Internet Computing, 2006. Vol. 10. No 4. C. 7–35.

5. *Berners-Lee T., Hendler J., Lassila O.* The Semantic Web // Scientific American. 2001. C. 34–43.

6. *Bierner G., Baldridge J., Morton T.* OpenNLP: A Collection of Natural Language Processing Tools. 2007. URL http://opennlp.sourceforge.net/

7. *Cunningham H., Maynard D., Bontcheva K., Tablan V.* GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications // Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics. Philadelphia, 2002.

8. *Davi A., Haughton D., Nasr N, Shah G., Skaletsky M., Spack R.* Review of Two Text-Mining Packages: SAS TextMining and WordStat. // The American Statistician. 2005. Vol. 59. No 1. C. 89–103.

9. *Dong Qw., Wang Xl., Lin L. Surayati Is., Razib M., Shahreen K.* Pairwise Protein Substring Alignment with Latent Semantic Analysis and Support Vector Machines to Detect Remote Protein Homology // Ubiquitous Computing and Multimedia Applications Communications in Computer and Information Science. 2011. Vol. 151. C. 526–546.

10. *Feinerer I., Wild F.* Automated Coding of Qualitative Interviews with Latent Semantic Analysis // Proceedings of the 6th International Conference on Information Systems Technology and its Applications. May 23–25, 2007. Kharkiv, Ukraine. H. Mayr, D Karagiannis (eds.). Lecture Notes in Informatics. Vol. 107. Bonn, 2007. C. 66–77.

11. *Gir´on J., Ginebra J., Riba A.* Bayesian Analysis of a Multinomial Sequence and Homogeneity of Literary Style // The American Statistician. 2005. Vol. 59. Issue 1. C. 19–30.

12. *Holmes D., Kardos J.* Who was the Author? An Introduction to Stylometry // Chance. 2003.Vol. 16. No. 2. C. 5–8.

13. *Hornik K.* Snowball: Snowball Stemmers. R package version 0.5, 2013.

14. *Hornik K., Zeileis A., Hothorn T., Buchta C.* RWeka: An R Interface to Weka. R package version. Version 3.7. 2011 http://CRAN.R-project.org/package=RWeka

15. Karatzoglou A., Feinerer I. Text Clustering with String Kernels in R. // Advances in Data Analysis (Proceedings of the 30th Annual Conference of the Gesellschaft fur Klassifikation e.V., Freie Universitat Berlin, March 8–10, 2006), R Decker, H. J. Lenz (eds.), Studies in Classification, Data Analysis, and Knowledge Organization. Springer-Verlag, 2007. C. 91–98.

16. *Li Y., Shawe-Taylor J.* Using KCCA for Japanese-English Cross-Language Information Retrieval and Classification // Journal of Intelligent Information Systems. 2006. Vol. 27 No. 2. C. 117–133.

17. *Manola F., Patrick J. Hayes, Peter F.* Patel-Schneider. RDF 1.1 Semantics.W3C Recommendation, 25 February 2014. URL: http://www.w3.org/TR/2014/REC-rdf11-mt-20140225.

18. McCallum 1996 McCallum A., Riedel S. Robust biomedical event extraction with dual decomposition and minimal domain adaptation //Proceedings of the BioNLP Shared Task 2011 Workshop, C. 46-50.

19 *Mierswa I., Wurst M., Klinkenberg R., Scholz M., Euler T.* YALE: Rapid Prototyping for Complex Data Mining Tasks // KDD '06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2006. C. 935–940.

20. *Mueller J.P.* ttda: Tools for Textual Data Analysis. R package version 0.1.1, 2006. URL http://wwwpeople.unil.ch/jean-pierre.mueller/

21. *Nilo J., Binongo G.* Who Wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution. Chance, Vol. 16, Issue 2, 2003. C. 9–17.

22. *Piatetsky-Shapiro G.* Data mining and knowledge discovery 1996 to 2005: overcoming the hype and moving from "university" to "business" and "analytics"// Data Mining and Knowledge Discovery. 2007. Vol. 15. No. 1. C. 99–105.

23. *Radlinski F., Joachims T.* Active Exploration for Learning Rankings from Click- through Data // Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. NY, 2007. C. 570–579.

24. *Sonnenburg S., Raetsch G, Schaefer C., Schoelkopf B.* Large Scale Multiple Kernel Learning // Journal of Machine Learning Research. 2006. Vol. 7. C. 1531–1565.

25. *Wild Fr., Valentine Ch., Scott P.* Shifting interests: changes in the lexical semantics of ED-MEDIA // Journal of e-Learning 2010. Vol. 9. No 4. C. 549–562.

26. *Witten I., Frank E.* Data Mining: Practical Machine Learning Tools and Techniques. San Francisco: Morgan Kaufmann, 2011. 565 c.

27. *Zhao Y.* R and Data Mining: Examples and Case Studies. Academic Press, Elsevier, 2012. 256 c.

REFERENCES

1. *Bol'shakova E. I., Klyshinskij E. S., Landje D. V., Noskov A. A., Peskova O. V., Jagunova E. V.* Avtomaticheskaja obrabotka tekstov na estestvennom jazyke i komp'juternaja lingvistika: Uchebnoe posobie. M.: MIEM, 2011. 272 s.

2. *Pashkovskij V. Je., Piotrovskaja V. R., Piotrovskij R. G.* Psihiatricheskaja lingvistika. M.: URSS, 2013. 158 s.

3. A language and environment for statistical computing // R Development Core Team; R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/

4. *Adeva J., Calvo R.* Mining Text with Pimiento // IEEE Internet Computing. 2006. Vol. 10. No 4. C. 27–35.

5. *Berners-Lee T., Hendler J., Lassila O.* The Semantic Web // Scientific American, 2001. C. 34–43.

6. *Bierner G, Baldridge J., Morton T.* OpenNLP: A Collection of Natural Language Processing Tools. 2007 URL .http:// opennlp.sourceforge.net/

7. *Cunningham H., Maynard D., Bontcheva K., Tablan V.* GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications //Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics., Philadelphia, 2002.

8. *Davi A., Haughton D., Nasr N, Shah G., Skaletsky M., Spack R.* Review of Two Text-Mining Packages: SAS TextMining and WordStat // The American Statistician. 2005. Vol. 59. No 1. C. 89–103.

9. *Dong Qw., Wang Xl., Lin L. Surayati Is., Razib M., Shahreen K.* Pairwise Protein Substring Alignment with Latent Semantic Analysis and Support Vector Machines to Detect Remote Protein Homology // Ubiquitous Computing and Multimedia ApplicationsCommunications in Computer and Information Science. 2011. Vol. 151. C. 526–546.

10. *Feinerer I., Wild F.* Automated Coding of Qualitative Interviews with Latent Semantic Analysis // Proceedings of the 6th International Conference on Information Systems Technology and its Applications, May 23–25, 2007, Kharkiv, Ukraine, H. Mayr, D. Karagiannis (eds.). Lecture Notes in Informatics. Vol. 107. Bonn, 2007. C. 66–77.

11. *Gir´on J., Ginebra J., Riba A.* Bayesian Analysis of a Multinomial Sequence and Homogeneity of Literary Style // The American Statistician. 2005. Vol. 59. Issue 1. C. 19–30.

12. *Holmes D., Kardos J.* Who was the Author? An Introduction to Stylometry // Chance. 2003. Vol. 16. No. 2. C. 5–8.

13. *Hornik K.* Snowball: Snowball Stemmers. R package version 0.5, 2013.

14. *Hornik K., Zeileis A., Hothorn T., Buchta C.* RWeka: An R Interface to Weka. R package version. Version 3.7. 2011 http://CRAN.R-project.org/package=RWeka

15. *Karatzoglou A., Feinerer I.* Text Clustering with String Kernels in R. // Advances in Data Analysis (Proceedings of the 30th Annual Conference of the Gesellschaft fur Klassifikation e.V., Freie Universitat Berlin, March 8–10, 2006), R. Decker, H. J. Lenz (eds.), Studies in Classification, Data Analysis, and Knowledge Organization. Springer-Verlag, 2007. C. 91–98.

16. *Li Y., Shawe-Taylor J.* Using KCCA for Japanese-English Cross-Language Information Retrieval and Classification // Journal of Intelligent Information Systems. 2006. Vol. 27. No. 2. C. 117–133.

17. *Manola F., Patrick J. Hayes, Peter F.* Patel-Schneider. RDF 1.1 Semantics.W3C Recommendation, 25 February 2014. URL: http://www.w3.org/TR/2014/REC-rdf11-mt-20140225.

18. *McCallum A., Riedel S.* Robust biomedical event extraction with dual decomposition and minimal domain adaptation // Proceedings of the BioNLP Shared Task 2011 Workshop, C. 46–50.

19. *Mierswa I., Wurst M., Klinkenberg R., Scholz M., Euler T.* YALE: Rapid Prototyping for Complex Data Mining Tasks // KDD ’06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2006. C. 935–940.

20. *Mueller J. P.* Ttda: Tools for Textual Data Analysis. R package version 0.1.1, 2006. URL http://wwwpeople.unil.ch/jean-pierre.mueller/

21. *Nilo J., Binongo G.* Who Wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution. Chance. Vol. 16. Issue 2. 2003. C. 9–17.

22. *Piatetsky-Shapiro G.* Data mining and knowledge discovery 1996 to 2005: overcoming the hype and moving from “university” to “business” and “analytics” // Data Mining and Knowledge Discovery. Vol. 15. No. 1. 2007. C. 99–105.

23. *Radlinski F., Joachims T.* Active Exploration for Learning Rankings from Click- through Data // Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. NY, 2007. C. 570–579.

24. *Sonnenburg S., Raetsch G., Schaefer C., Schoelkopf B.* Large Scale Multiple Kernel Learning // Journal of Machine Learning Research. 2006. Vol. 7. C. 1531–1565.

25. *Wild Fr., Valentine Ch., Scott P.* Shifting interests: changes in the lexical semantics of ED-MEDIA // Journal of e-Learning. 2010. Vol. 9. No 4. C. 549–562

26. *Witten I., Frank E.* Data Mining: Practical Machine Learning Tools and Techniques. San Francisco: Morgan Kaufmann, 2011. 565 c.

27. *Zhao Y.* R and Data Mining: Examples and Case Studies. Academic Press, Elsevier, 2012. 256 c.

УДК 004 + 004.946

А. В. Флегонтов

HI-TECH: ДИНАМИКА ВЗАИМОДЕЙСТВИЙ НАУКИ И ОБРАЗОВАНИЯ

Рассматривается динамика взаимодействия науки и образования в контексте развития высоких технологий. Отмечается базовая роль информационных технологий для развития высоких технологий. Рассматриваются тенденции и вектор развития сферы информационных технологий и ее роль в процессе трансформации науки и образования. Анализируются основные приоритетные направления исследований и разработок в области информационных технологий.

Ключевые слова: высокие технологии, информационные технологии, наука, образование.

УДК004.942 + 004946

A. Flegontov

Hi-Tech: Dynamics of Interactions of Science and Education

The dynamics of interaction of science and education in the context of the development of high technologies is discussed. The basic role of information technologies for the development of hi-tech is emphasized. The trends and development of information technology and its role in the transformation process of science and education are pointed out. The main priority directions of research and developments in the field of information technologies are described.

Keywords: hi-tech, information technology, science, education.

На современном этапе развития общества происходит значительная трансформация как науки, так и самого общества. При этом отмечено, что социальная динамика науки коррелируется с переходом от классической к неклассической и постнеклассической научным картинам мира с возрастанием роли неклассической и с появлением постнеклассической методологии. Это обстоятельство дает возможность рассмотрения и науки и общества как сложных самоорганизующихся систем.

В современных философских и науковедческих исследованиях фиксируется, что на протяжении XX в. менялись не только основания науки [3; 6], но и организация науки. Констатируется формирование качественно новой стадии развития науки и техники, а также их