**Л.С. Ломакина, А.С. Суркова**

# ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ АНАЛИЗА И МОДЕЛИРОВАНИЯ ТЕКСТОВЫХ ДАННЫХ

**Монография**

**Воронеж**
**Издательство «Научная книга»**
**2015**

**УДК 004.912**
**ББК 32.81**
**Л 74**

**Рецензенты:**

**Турлапов В.Е.** д-р. техн. наук (Нижегородский государственный университет им. Н.И. Лобачевского)
**Яхно В.Г.** д-р физ.-мат. наук (Институт прикладной физики РАН)

**Л 74** **Ломакина, Л.С.** Информационные технологии анализа и моделирования текстовых структур: Монография / Л.С. Ломакина, А.С. Суркова. – Воронеж: Издательство «Научная книга», 2015. – 208 с.

**ISBN 978-5-98222-863-5**

В книге излагаются актуальные проблемы построения моделей текстовых данных и создания на их основе систем обработки и анализа текстов с учетом структурных особенностей текстов. Основные задачи анализа текстовых данных – кластеризация, классификация, идентификация – рассматриваются с позиций принципов системного представления текстов, нечеткой логики и обучающихся систем. Большое внимание уделено описанию универсальных методов анализа текстов, которые могут применяться для решения различных прикладных задач. Приводятся примеры конкретных алгоритмов и результатов вычислительных экспериментов по построению систем тематической кластеризации и идентификации основных характеристик текста.

Книга рассчитана на научных и инженерно-технических работников в области прикладной информатики и автоматической обработки текста, а также полезна для студентов, магистрантов и аспирантов, специализирующихся в указанной области.

Рис. 47. Табл. 27. Библиогр.: 131 назв.

**УДК 004.912**
**ББК 32.81**
**Л 74**

**ISBN 978-5-98222-863-5** 

## СОДЕРЖАНИЕ

# Введение

Важность, значимость и необходимость анализа и обработки текстовых и других слабоструктурированных информационных данных постоянно возрастают. В связи с широким распространением систем электронного документооборота, социальных сетей, блогов, сетевых информационных порталов, персональных сайтов это становится особенно важным и как техническая задача, и как значимая часть взаимодействия людей в современном информационном мире.

Одной из основных форм представления информации является текстовая форма, наряду с графической, звуковой, а также видео информацией. Если первоначально первостепенными проблемами считались задачи, связанные с обеспечением сбора, хранения, поиска и предоставления данных, то в последнее время, при упрощении доступа к разнообразным коллекциям текстовых документов, появляются новые задачи анализа и обработки текстовых данных. К традиционным проблемам добавляются новые, связанные, например, с большими объемами текстовых данных в различных социальных сетях и других информационных, поисковых и аналитических приложениях Интернета.

Область, занимающаяся проблемами обработки все увеличивающегося объема текстовой информации, получила название *Text Mining*. На русский этот термин можно перевести как Интеллектуальный анализ текстов аналогично уже устоявшемуся понятию *Data Mining* – интеллектуальный анализ данных. Предметную область *Text Mining* как согласованную совокупность задач можно описать следующим образом. К числу традиционных проблем можно отнести задачи определения авторства, автоматического извлечения ключевых слов, аннотирования и реферирования, кластеризации и классификации по тематическим категориям и т.д. К недавно возникшим задачам относятся проблема анализа текстов в глобальной сети для обеспечения безопасности и выявления потенциально опасных или нежелательных сообщений, а также задачи, связанные с многоязыковыми текстами и проблемой «переводного» плагиата и заимствования.

При всем достаточно большом количестве книг и статей, посвященных задачам *Text Mining* необходимо отметить отсутствие литературы на русском языке, узкую направленность имеющихся материалов, при этом в большинстве случаев приходится каждый раз заново решать возникающие частные задачи.

Целью настоящей монографии является рассмотрение с единых позиций общих вопросов, связанных с подходами к моделированию и обработке текстовых данных при решении разнообразных прикладных задач.

Также важной особенностью книги является обсуждение возможностей и перспектив современных информационных технологий при анализе текстовых данных.

Монография написана с позиций современных информационных технологий и содержит изложение различных аспектов методологии, технологии и реализации решения основных задач анализа и обработки текстовых данных. Таким образом, книга разделена на три части: методологические, теоретические и практические аспекты анализа и моделирования текстовых данных, в каждой части материал структурирован по главам.

Первая часть посвящена проблемам построения общей методологических аспектов анализа и моделирования текстовых данных. В главе рассмотрены особенностей основных задач – кластеризации, классификации и идентификации текстов. Также определены и описаны основные принципы, использование которых позволяет построить системы анализа и обработки текстовых данных. К таким принципам относятся принцип системного представления текстов, принцип нечеткой логики и принцип обучающихся систем. Рассмотрение отдельных задач с учетом основных принципов позволило представить единый методологический подход к рассматриваемым проблемам анализа и моделирования текстовых данных.

Вторая часть дает общее представление о возможностях решения основных задач анализа текстовых данных на основе сформулированных принципов. Рассмотрены основные подходы к системному представлению текстов: статистические, информационные, структурно-иерархические и другие и показаны возможности использования потокового представления текстов и использования алгоритмов сжатия при решении задач анализа текстов. Также во второй части изложены принципы и методы нечеткой логики и аспекты обучающихся систем, необходимые для корректного использования в задачах анализа и моделирования текстов.

В третьей части в разделе «Алгоритмическое и программное обеспечение анализа и обработки текстов» приведены конкретные алгоритмы анализа текстовых данных, основанные на системном представлении текстов. Рассмотрены базовые нечеткие методы кластеризации, такие как *fuzzy c-means* (*FCM*), *Kernel Fuzzy Clustering* и др. и предложенные модификации алгоритмов классификации и кластеризации на основе нечетких отношений. Также в данном разделе описаны алгоритмы решения основных задач анализа текстов на основе нейронных сетей, деревьев принятия решений и подхода Random forest.

Раздел «Практическая реализация алгоритмов решения основных задач» демонстрирует примеры практической реализации рассмотренных в предыдущем разделе алгоритмах, приведены результаты кластеризации,

классификации и идентификации текстовых данных. При анализе полученных результатов, сделаны выводы о перспективности и направлениях дальнейших исследований.

Следует отметить также необходимость использования английских терминов при обсуждении практически всех аспектов рассматриваемых проблем. Поскольку многие задачи и подходы к их решению обсуждаются и развиваются в нескольких направлениях и нескольких научных областях, то возникают терминологические противоречия. В данной сфере еще не завершено окончательное формирование терминологический и понятийной базы, что осложняется и различными трактовками определений при переводе на русский язык. В связи с этим для большинства вводимых понятий и терминов приводятся английский аналог.

# Библиографический список


1. Андреев А.М., Березкин Д.В., Морозов В.В., Симаков К.В. Метод кластеризации документов текстовых коллекций и синтеза аннотаций кластеров // Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008. с 220-229
2. Арапов М.В. Квантитативная лингвистика. – М.: Наука, 1988. – 184 с.
3. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Методы и модели анализа данных: OLAP и Data Mining. – СПб.: БХВ-Петербург, 2004. – 336 с.
4. Белозерова Н.Н. Можно ли поверить дискурс фракталом? //Язык и литература, 2002, №16. – Тюмень, изд-во ТюмГУ, 2002.
5. Белоногов Г.Г., Кузнецов Б.А. Языковые средства автоматизированных информационных систем. М., 1983. Мельников Г.П. Системный подход в лингвистике. // Системные исследования. Ежегодник. – М.: Наука, 1972.
6. Вапник В.Н., Червоненкис А.Я. Теория распознавания образов (статистические проблемы обучения). М., Наука. 1974. 416 с.
7. Верещагин Н.К., Успенский В.А., Шень А. Колмогоровская сложность и алгоритмическая случайность. – М.: МЦНМО, 2013.
8. Воронцов К.В. Машинное обучение (курс лекций). Режим доступа: http://www.machinelearning.ru/wiki/index.php?title=Mo
9. Гальперин И.Р. Текст как объект лингвистического исследования – М.: Наука, 1981. – 140 с.
10. Гладких А.В. Синтаксические структуры естественного языка в автоматизированных системах сообщений. – М.: Наука, 1985.
11. Городецкий Б.Ю. Компьютерная лингвистика: моделирование языкового общения //Новое в зарубежной лингвистике. Вып. 24. - М., 1989.
12. Дмитриев А.С. Хаос, фракталы и информация //Наука и жизнь, 2001, №5
13. Дюран Б., Одел П. Кластерный анализ. М., Статистика, 1977. – 128 с.
14. Егорушкин А. У каждого свой язык. //Компьютера – 2002, №21.
15. Еремеев А.П. Построение решающих функций на базе тернарной логики в системах принятия решений в условиях неопределенности // Изв. РАН. Теория и системы управления. 1997. N 5. - с. 138-143
16. Заде Л. Понятие лингвистической переменной и его применение к принятию приближенных решений. М.: Мир. – 1976. – 164 с.
17. Заде Л.А. Размытые множества и их применение в распознавании образов и кластер-анализе // Классификация и кластер. М.: Мир, 1980., с. 208-247
18. Колмогоров А.Н. Три подхода к определению понятия «Количество информации» // Новое в жизни, науке, технике. Сер. "Математика, кибернетика". 1991, №1, стр. 24-29
19. Костышин А.С. О применимости некоторых формальных методов для исследования полных строений текстов. //Материалы конференции КВАЛИСЕМ-2000 – Новосибирск, изд-во Новосибир. гос. пед. ун-та., 2000.

20. Кофман А. Введение в теорию нечетких множеств. – М., Радо и связь, 1982. 432 с.
21. Ландэ Д.В., Снарский А.А., Безсуднов И.В. Интернетика. Навигация в сложных сетях: модели и алгоритмы. – М.: Либроком, 2009. – 264 с.
22. Леоненков А.В. Нечеткое моделирование в среде Matlab и fuzzyTECH. – С.Пб.: BHV-Санкт-Петербург, 2003. – 736 с.
23. Леонтьева Н.Н. Динамика единиц в семантических структурах. //Труды Международного семинара Диалог-2002 по компьютерной лингвистике и ее приложениям. Том 1. Теоретические проблемы. – М., 2002.
24. Лингвистический энциклопедический словарь. – М.: 1990.
25. Ломакин Д.В., Панкратова А.З., Суркова А.С. Золотая пропорция как инвариант структуры текста. // Журнал «Вестник Нижегородского университета им. Н.И. Лобачевского», 2011, №4, с. 196-199.
26. Ломакина Л.С., Мордвинов А.В., Суркова А.С. Построение и исследование модели текста для его классификации по предметным категориям. // Системы управления и информационные технологии, 2011, №1(43), с. 16-20.
27. Ломакина Л.С., Родионов В.Б., Суркова А.С. Иерархическая кластеризация текстовых документов. // Системы управления и информационные технологии, 2012, № 2(48), с. 39-44.
28. Ломакина Л.С., Суркова А.С. Автоматизированные информационно-поисковые системы. Задачи. Принципы. Методология: учеб. пособие / Л.С. Ломакина, А.С. Суркова; Нижегор. гос. техн. ун-т им. Р.Е. Алексеева. – Н. Новгород, 2011. – 109 с.
29. Ломакина Л.С., Суркова А.С., Буденков С.С. Кластеризация текстовых данных на основе нечеткой логики // Системы управления и информационные технологии, №1(55), 2014. – С. 73-77.
30. Мартыненко Г.Я. Основы стилеметрии. – Л.: Изд-во ЛГУ, 1988. – 176 с.
31. Мельничук А. С. Понятие системы и структуры языка. // Вопросы языкознания, 1970, №1. С. 27
32. Мельчук И.А. Русский язык в модели «Смысл ⇔ Текст». Москва-Вена, 1995.
33. Минский М. Фреймы для представления знаний. - М.: Энергия. 1979.
34. Москальская О.И. Грамматика текста. М.: Наука, 1981.
35. Москальчук Г.Г. Структура текста как синергетический процесс. М.: УРСС, 2003. Москальчук Г.Г. Структурная организация и самоорганизация текста. – Барнаул, 1998.
36. Мурзин Л.Н., Штерн А.С. Текст и его восприятие – Свердловск, 1991.
37. Нариньяни А.С. Автоматическое понимание текста - новая перспектива // Труды международного семинара Диалог-97 по компьютерной лингвистике и ее приложениям. - Москва, 1997, с. 203-208.
38. Негуляев В.А. Исследование коммуникативных микроструктур патентного текста и их роли для автоматической обработки информации. //Вычислительная лингвистика. М.: Наука, 1976.
39. Нечеткие множества в моделях управления и искусственного интеллекта / Под ред. Д.А.Поспелова. – М., Наука, 1986. 312 с.

40. Нечеткие множества и теория возможностей. Последние достижения / Под ред. Р.Р.Ягера. – М.: Радио и связь, 1986.– 408 с.
41. Орлов Ю.К. Обобщенный закон Ципфа-Мандельброта и частотные структуры информационных единиц различных уровней. //Вычислительная лингвистика. М.: Наука, 1976.
42. Поликарпов А.А. Циклические процессы в становлении лексической системы языка: моделирование и эксперимент – М., 2001.
43. Пономаренко И.Н. Фрактал в структуре художественного текста //Русский язык: исторические судьбы и современность. II Международный конгресс русистов-исследователей. – М., 2004.
44. Прангишвили И.В. Системный подход и общесистемные закономерности. – М.: СИНТЕГ, 2000, 528с.
45. Прикладная статистика: Классификации и снижение размерности: Справ. изд. / Под ред. С. А. Айвазяна. - М.: Финансы и статистика, 1989. - 607 с.
46. Романов А.С. Методика идентификации автора текста на основе аппарата опорных векторов // Доклады ТУСУРа. – 2009. – № 1 (19), ч. 2. – С. 36–42
47. Руспини Э.Г. Последние достижения в нечетком кластер-анализе // Нечеткие множества и теория возможностей: Последние достижения / Под ред. Р.Р. Ягера; - М: Радио и связь, 1986.— с. 114-132.
48. Рутковская Д., Пилиньский М., Рутковский Л. Нейронные сети, генетические алгоритмы и нечеткие системы. - М.: Горячая линия - Телеком, 2006. - 452 с.
49. Рыжов А.П. Элементы теории нечетких множеств и измерения нечеткости. М., Диалог-МГУ, 1998.
50. Сайт, посвященный литературоведческой атрибуции: http://corneille-moliere.com/.
51. Сафронова Ю.Б. Некоторые системно-количественные характеристики лексико-семантических парадигм разных видов. //Уч. зап. ТГУ. Вып. 745. 1986, с.129-138.
52. Скороходько Э.Ф. Семантические сети и автоматическая обработка текста. Киев, Наукова думка, 1983. – 218 с.
53. Сметанин Ю. Г., Ульянов М. В. Мера символьного разнообразия: подход комбинаторики слов к определению обобщенных характеристик временных рядов // Бизнес-информатика №3 (29) 2014. с. 40–48.
54. Сметанин Ю.Г., Ульянов М.В. Мера символьного разнообразия – характеристика временных рядов. // Materials of the III International Scientific Conference «Information-Management Systems and Technologies». Odessa. 2014, 19-21.
55. Соколов О.М., Пионтковская Т.А. Система лингвистических конструктов, формирующих фреймы на глагольно-предикативной основе. //Системные исследования. Ежегодник. - М.: Наука, 1991. с.124-141.
56. Солганик Г.Я. Стилистика текста. – М., 1997.
57. Солнцев В.М. Язык как системно-структурное образование. М.: Наука, 1977.

58. Суркова А.С. Идентификация текстов на основе информационных портретов // Вестник Нижегородского университета им. Н.И. Лобачевского, 2014, № 3 (1), с. 145–149
59. Суркова А.С., Родионов В.Б. Алгоритм разбиения неструктурированного множества текстовых объектов // Научно-технический вестник Поволжья.– Казань, 2013г. - №5. – с .298-300
60. Турыгина Л.А. Моделирование языковых структур средствами вычислительной техники – М., 1988
61. Уфимцев Р. Тексты как фракталы. Режим доступа: http://www.cognitivist.ru/er/kernel/prologi_16_texts_as_fractals.xml.
62. Хмелев Д.В. Классификация и разметка текстов с использованием методов сжатия данных. Краткое введение: http://compression.graphicon.ru/download/articles/classif/intro.html.
63. Хьетсо Г. и др. Кто написал "Тихий Дон"? (Проблема авторства «Тихого Дона») – М., 1989
64. Цыпкин Я.З. Основы теории обучающихся систем, М., Наука, 1970, 252 с.
65. Шевелёв О.Г. Методы автоматической классификации текстов на естественном языке: Учебное пособие. – Томск: ТМЛ-Пресс, 2007. 144 с.
66. Шеннон К. Э. Математическая теория связи // Работы по теории информации и кибернетике/ Пер. С. Карпова. – М.: ИИЛ, 1963. – 830 с.
67. Шрейдер Ю.А., Шаров А.А. Системы и модели – М.: Радио и связь, 1982. – 152 с.
68. Штовба С.Д. Введение в теорию нечетких множеств и нечеткую логику. http://matlab.exponenta.ru/fuzzylogic/book1/index.php.
69. Шульгин В.И. Основы теории цифровой связи. Харьков, ХАИ, 2008, 184 с.
70. Advances in Fuzzy Clustering and its Applications. Editor(s): J. Valente de Oliveira, W. Pedrycz. John Wiley & Sons, Ltd. 2007. 434 p.
71. Alrabaee S., Saleem N., Preda S., Wang L., Debbabi M. OBA2: An Onion approach to Binary code Authorship Attribution // Digital Investigation, 2014, №11, P. S94–S103
72. Bennett C.H., Gacs P., Li M., Vitanyi P.M.B., Zurek W. Information Distance // IEEE Transactions on Information Theory, 44:4(1998), pp. 1407-1423
73. Berry M. W., Kogan J. Text Mining. Applications and Theory. – Wiley, 2010. 207 p.
74. Bezdek J.C. [et al.]. Fuzzy models and algorithms for pattern recognition and image processing. Springer Science + Business Media, Inc. 2005. 776 p.
75. Bloom C. New techniques in context modeling and arithmetic encoding // IEEE DCC, Los Alamitos CA, March 1996, March, p. 426.
76. Bolshakov I.A., Gelbukh A. Computational linguistics: models, resources, applications, Mexico, 2004, 186 pp.
77. Bolshoy A., Volkovich Z., Kirzhner V., Barzily Z. Genome Clustering: From Linguistic Models to Classification of Genetic Texts. Springer, 2010. 206 p.
78. Brocardo M.L., Traore I., Saad S., Woungang I. Authorship Verification for Short Messages using Stylometry // Proceedings of the IEEE International Conference on Computer, Information and Telecommunication Systems (CITS 2013), 2013, pp. 1-6.

79. Cavnar W.B. N-Gram-Based Text Filtering For TREC-2 // Proceedings of the Second Text Retrieval Conference (TREC-2). – NIST, Gaithersburg, Maryland. – 1993, pp. 171-180
80. Cavnar W.B., Trenkle J.M. N-Gram-Based Text Categorization // Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval. – Las Vegas. – 1994. P. 161–175
81. Cavnar W.B., Vayda A.J. N-gram-based matching for multi-field database access in postal applications // Proceedings of the 1993 Symposium On Document Analysis and Information Retrieval. University of Neveda. – Las Vegas. – 1993.
82. Chen Hs., Chau M. Web Mining: Machine learning for Web Applications // Annual Review of Information Science and Technology, 2004, vol. № 38, pp.289-329
83. Cilibrasi R., Vitanyi P.M.B. Clustering by compression // IEEE Trans. Inf. Theory, 2005. Vol. 51, no. 4, 1523–1545.
84. Clarke B., Fokoue E., Zhang H.H. Principles and Theory for Data Mining and Machine Learning. Springer Science, LLC, 2009, 781 p.
85. Cooley R., Mobasher B., Srivastava J. Web mining: information and pattern discovery on the World Wide Web // Proceedings of the 9th ZEEE International Conference on Tools with Artificial Intelligence, 1997, 558-567.
86. D'Urso P. Fuzzy Clustering of Fuzzy Data // Advances in Fuzzy Clustering and its Applications (eds. J. V. de Oliveira, W. Pedrycz). 2007. - pp. 155-192
87. Doyle J., Keselj V. Automatic Categorization of Author Gender via N-Gram Analysis // Proceedings of The 6th Symposium on Natural Language Processing, SNLP'2005. http://web.cs.dal.ca/~vlado/papers/SNLP05J.pdf.
88. Dua S., Du X. Data Mining and Machine Learning in Cybersecurity. New York, 2011. 224 p.
89. Etzioni O. The world-wide web: Quagmire or gold mine? // Communications of the ACM, 1996, № 39(11), p.65-68
90. Feldman R., Sanger J. The text mining handbook. Advanced Approaches in Analyzing Unstructured Data. – Cambridge University Press. 2007. 410 p.
91. Goldberg D.E. Genetic algorithms in search, optimization, and machine learning. Reading, MA: Addison-Wesley. 1989
92. Granovetter M. The Strength of Weak Ties // American Journal of Sociology, 1973, Vol. 78, No. 6., pp 1360—1380
93. Gries S.Th., Newman J., Shaoul C. N-grams and the clustering of registers. //Empirical Language Research Journal, 2011, №5.1.
94. Hathaway, R.J. and Bezdek, J.C. Switching regression models and fuzzy clustering // IEEE Transactions on Pattern Analysis and Machine Intelligence. 1993, 1 (3): pp.195–204.
95. Hoppner F., Klawonn F., Kruse R., Runkler Th. Fuzzy Cluster Analysis: Methods for Classification, Data Analysis, and Image Recognition. New York, John Wiley & Sons, 1999.
96. Juola P. Authorship Attribution // Foundations and Trends in Information Retrieval. Vol. 1, No. 3 (2006) 233–334

97. Kimbrell R.E. Searching for Text? Send and N-gram! // Byte, May 1998. – 1998. P. 297–312.
98. Klein D., Manning Ch.D. A generative constituent-context model for improved grammar induction. // Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002, pp. 128-135.
99. Kontostathis A., Edwards L., Leatherman A. Text mining and cybercrime // Text Mining. Applications and Theory. Ed. by Berry M. W., Kogan J. – Wiley, 2010. P.149-164.
100. Kosala R., Blockeel H. Web Mining Research: A Survey. ACM SIGKDD, 2000. Vol. 2, Issue 1, pp.1-15
101. Krsul I, Spafford EH. Authorship analysis: identifying the author of a program// Proc. 8th National Information Systems Security Conference, 1995, P. 514–524.
102. Lance G.N., Williams W.T. A general theory of classificatory sorting strategies. // Computer J., 1967, V.9, №4, P. 373-380.
103. Langdon G.G. and Rissanen J.J. 1981. Compression of black-white images with arithmetic coding. IEEE Trans.Commun.COM-29, 6(Jun.),858-867
104. Li M., Chen X., Li X., Ma B., Vitanyi P.M.B. The similarity metric // IEEE Trans. Inform. Th., 50:12(2004), 3250- 3264
105. Lomakina L.S., Rodionov V.B., Surkova A.S. Hierarchical Clustering of Text Documents // Automation and Remote Control, 2014, Vol. 72, No. 9, pp. 345-351.
106. Manning Ch.D., Raghavan P., Schutze H. Introduction to Information Retrieval. Cambridge University Press. 2008. 504 c.
107. Manning Ch.D., Schutze H. Foundations of statistical natural language processing. MIT Press., Cambridge. 1999. 680 c.
108. Miyamoto S., Ichihashi H., Honda K. Algorithms for Fuzzy Clustering. Methods in c-Means Clustering with Applications. Springer-Verlag Berlin Heidelberg. 2008. 247 p.
109. O'Connor B., Bamman D., Smith N.A. Computational Text Analysis for Social Science: Model Assumptions and Complexity // Second Workshop on Computational Social Science and the Wisdom of Crowds (NIPS 2011). https://people.cs.umass.edu/~wallach/workshops/nips2011css/papers/OConnor.pdf
110. Parpinelli, R.S., Lopes, H.S. and Freitas, A.A. An Ant Colony Algorithm for Classification Rule Discovery // Data Mining: a Heuristic Approach, Idea Group, 2002. pp. 191-208
111. Rendon E., Abundez I., Arizmendi A., Quiroz E. Internal versus External cluster validation indexes // International journal of computers and communications. 2011, Issue 1, Vol. 5. pp.27-34
112. Rijsbergen C. J. Information retrieval. 1979, 153 p.
113. Rissanen J.J., Langdon G.G. Universal modeling and coding // IEEE Trans. Inf. Theory IT-27. 1981. №1, pp. 12-23.
114. Rosenblum N., Zhu X., Miller B.P. Who wrote this code? Identifying the authors of program binaries // Proceedings of the 16th European conference on Research in computer security, 2011. Режим доступа: http://pages.cs.wisc.edu/~jerryzhu/pub/Rosenblum11Authorship.pdf

115. Salton G. Automatic text processing. Addison-Wesley Publishing Company. 1989. 530 p.
116. Salton G., Wong A., Yang C.S. A vector space model for automatic indexing // Communications of The ACM - CACM, 1975, vol. 18, no. 11, pp. 613-620.
117. Sanderson, C., Guenter, S. Short text authorship attribution via sequence kernels, Markov chains and author unmasking: An investigation // Proceedings of the International Conference on Empirical Methods in Natural Language Engineering, 2006, pp. 482–491
118. Sato-Ilic M., Jain L.C. Innovations in Fuzzy Clustering. Theory and Applications. Springer, 2006. 152 p.
119. Schwartz R., Tsur O., Rappoport A., Koppel M. Authorship Attribution of Micro-Messages // Conference on Empirical Methods in Natural Language Processing (EMNLP), 2013, P. 1880-1891
120. Sebastiani F. Machine Learning in Automated Text Categorization // ACM Computing Serveys, 2002. Vol. 34. – №1. P. 1–47
121. Shannon, C.E.: A mathematical theory of communication. Bell System Technical J. 27, 379–423 (1948), русский перевод: Шеннон К. Э. Математическая теория связи // Работы по теории информации и кибернетике / Пер. С. Карпова. — М.: ИИЛ, 1963. — 830 с.
122. Stamatatos. E. A Survey of Modern Authorship Attribution Methods // Journal of the American Society for Information Science and Technology, 2009. № 60(3). P.538–556.
123. Stanko S., Lu D., Hsu I. Whose Book is it Anyway? Using Machine Learning to Identify the Author of Unknown Texts // Machine Learning Final Projects, 2013. http://cs229.stanford.edu/proj2013/StankoLuHsu-AuthorIdentification.pdf
124. Surkova A.S., Domnin A.A., Bulatov I.V., Tsarev A.A. Neural networks and decision trees algorithms – the base of automated text classification and clustering // American Journal of Control Systems and Information Technology. 2013, №2. p. 33-35.
125. Vapnik V.N. Statistical Learning Theory. John Wiley & Sons, Ltd. 1998. 736 p.
126. Vinh N.X., Epps J., Bailey J. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance // The Journal of Machine Learning Research. 2010, Vol. 11, pp. 2837-2854
127. Xie X L, Beni G. A validity validity measure for fuzzy clustering // IEEE Trans. PAMI, 1991, vol. 13(8), pp. 841-847.
128. Yang Y. An evaluation of statistical approaches to text categorization. // Journal of Information Retrieval, 1999, №1. P. 67-88.
129. Zadeh L.A. Similarity relations and fuzzy orderings // Information Sciences, 1971, vol. 3, no. 2, pp. 177-200.
130. Ziv J., Lempel A. A Universal Algorithm for Sequential Data Compression // IEEE Transactions on Information Theory, 1977, 23(3), pp. 337–343
131. Ziv J., Lempel A. Compression of Individual Sequences via Variable-Rate Coding // IEEE Transactions on Information Theory. 1978, 24 (5), 530–536.

Научное издание

**Ломакина** Любовь Сергеевна
**Суркова** Анна Сергеевна

**ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ АНАЛИЗА И МОДЕЛИРОВАНИЯ ТЕКСТОВЫХ ДАННЫХ**

Монография

Издание публикуется в авторской редакции

Дизайн обложки С.А.Кравец

Подписано в печать 30.12.2014. Формат 60x84 1/16.
Усл. печ.л. 13,2. Заказ 000. Тираж 500 экз.

ООО Издательство «Научная книга»
394077, Россия, г.Воронеж, ул. 60-й Армии, 25-120
http://www.sbook.ru/

Отпечатано с готового оригинал-макета
в ООО «Цифровая полиграфия»
394036, г. Воронеж, ул. Ф. Энгельса, 52.
Тел.: (473)261-03-61