УДК 004.89 **DOI** 10.17150/2500-2759.2016.26(5).805-810

КЛАСТЕРИЗАЦИЯ ИНФОРМАЦИИ БАЗ ДАННЫХ ОБОРУДОВАНИЯ

А. В. Боровский[1], Е. Е. Раковская[1], А. Л. Бисикало[2]
[1] Байкальский государственный университет, г. Иркутск, Российская Федерация
[2] Иркутский государственный университет, г. Иркутск, Российская Федерация

Информация о статье

Дата поступления
17 мая 2016 г.

Дата принятия к печати
13 июня 2016 г.

Дата онлайн-размещения
31 октября 2016 г.

Ключевые слова

Text Mining; кластеризация; базы данных оборудования; датчики; частота слов; декомпозиция сингулярного значения

Аннотация

В статье описывается метод кластеризации информации баз данных оборудования на основе анализа естественно-языковых текстов. С помощью этого метода облегчается задача сортировки входных данных, а также поиска и устранения неверного соответствия данных. Анализ проводился с применением технологий Data Mining и Text Mining. По исходным данным, представляющим собой фрагменты описаний назначения приборов, строились таблицы частот слов-вхождений в документ с учетом их «важности» и семантики исходных текстов. Использование декомпозиции сингулярного значения позволило уменьшить размерность матрицы и выделить пространство наиболее изменчивых признаков. Полученное множество признаков применяется для нахождения различных группировок текстов — кластеризации, классификации и т. д. Также показывается возможность применения методов интеллектуального анализа текстов для кластеризации информации баз данных оборудования и приводятся первоначальные результаты проведенного исследования.

CLUSTERING OF INFORMATION FROM OF THE EQUIPMENT DATA BASES

Andrei V. Borovsky[1], Elena E. Rakovskaya[1], Artem L. Bisikalo[2]
[1] Baikal State University, Irkutsk, Russian Federation
[2] Irkutsk State University, Irkutsk, Russian Federation

Article info

Received
May 17, 2016

Accepted
June 13, 2016

Available online
October 31, 2016

Keywords

Text Mining; clusterization; equipment data base; detectors; word frequency; singular value decomposition

Abstract

This paper describes the method for clustering information of the equipment data bases using analysis of the natural language texts. This method helps sort input data and search and eliminate incorrect data consistency. The analysis has been carried on the basis of the Data Mining and Text Mining technologies. The input data, presented in the form of fragmented descriptions for the equipment use, was employed to build frequency tables for the words included in the document taking into consideration their 'importance' and semantics of the original texts. The singular value decomposition enables us to reduce the dimension of the matrix and distinguish the set of the most variable attributes which was used to find various text groups — clasterization, classification etc. We show that methods for the intellectual analysis of texts can be used for clasterization of information of the equipment data bases and present the initial results of the research.

Развитие науки и технологий на сегодняшний день обуславливает наличие огромного разнообразия измерительного оборудования, а также приборов и средств автоматизации, контроллеров и т. д., которые применяются в различных отраслях промышленности [1]. Большое количество производителей комплексов, приборов, множество модификаций имеющихся устройств, многообразие стандартов (государственные, отраслевые

и др.) делают сложной задачу выбора технического оснащения для функционирования автоматизированных систем.

Комплектация оборудования осложняется также тем, что информация о технических устройствах содержится в различных источниках: стандартах, справочниках, каталогах, Интернет-ресурсах и т. д. Она может быть частично структурирована, например, иметь вид реестров, таблиц, спецификаций, списков, а может быть представлена в виде текстовых документов с описанием характеристик оборудования.

При определении необходимого технического ресурса существует проблема группируемости информации об оборудовании в наборы, кластеры и другие классификации, что предполагает разработку критериев и методов оценивания степени близости (сходства) данных с учетом их разнородности — структурированные или неструктурированные данные, имеющие качественные или количественные признаки и т. д. [2–4].

Для проведения процедур обработки текстов данные определенным образом формализуются. В задачах кластеризации естественно-языковой информации классической формальной интерпретацией является векторная модель представления текстов (VSM — vector space model) [5–9]. Здесь документ показывается как вектор, а векторное пространство определяется словарем. В данной модели последовательность слов в тексте полностью игнорируется, а сам текст отображается как «мешок слов» или «мешок, полный слов». При помощи «мешка слов» можно определить, какие векторы документов схожи.

Векторная модель представляет документы матрицей слов и документов:

$$M = |F| \cdot |D|,$$

где F — это множество признаков документов, состоящее из лексем; D — множество документов.

При выборе признаков F необходимо учитывать неинформативность часто встречающихся слов, а также исключить редкие слова. Неинформативными могут быть также слишком короткие или длинные слова [6]. Количественные характеристики «неинформативности» слов вычисляются экспериментально с учетом специфики текста.

Для данных, имеющих нечисловую природу, существует проблема их сопоставления, сравнения. В целом ряде теоретических и практических исследований применяются коэффициенты подобия. Самыми распространенными из них являются:

1. Коэффициент, выражающий равноважность нулевых и единичных признаков

$$K = \frac{P}{z} \ (0 \le K \le 1),$$

где P — общее число совпадающих признаков; z — общее число признаков, по которым идет сравнение.

2. Коэффициент Рао

$$K = \frac{p(1,\ 1)}{z},$$

где $p(1,\ 1)$ — число совпадающих единичных признаков у обоих объектов.

3. Коэффициент Джекарда

$$K = \frac{p(1,\ 1)}{p(1,\ 1) + Q} \ (0 \le K \le 1).$$

где Q — общее число несовпадающих признаков.

4. Коэффициент Дейка (придает вдвое больший вес совпадающим признакам)

$$K = \frac{2p(1,\ 1)}{2p(1,\ 1) + Q} \ (0 \le K \le 1).$$

Различные коэффициенты подобия, рассчитанные для одних и тех же объектов, будут различны по величине. Выбор того или иного коэффициента определяется характером решаемой задачи и во многом является субъективным.

С учетом специфичности обработки естественно-языковых текстов применяются следующие метрики сходства:

1. Обратная (инверсная) частота документов (Inverse Document Frequency)

$$idf(t_i) = \frac{N}{df(t_j)}, \qquad (1)$$

где N — число всех документов; $df(t_j)$ — частота документов с термином t_j.

2. Частота термина t_j в документе d_i (Term Frequency)

$$tf_{d_i}(t_j). \qquad (2)$$

3. Комбинация из приведенных локальных (2) и глобальных (1) признаков *TF-IDF* (Term Frequency Inverse Document Frequency)

$$TF - IDF = tf_{d_i}(t_j) \log \frac{N}{df(t_j)}.$$

Для расчета частоты термина в документе и инверсной частоты документов применяются лексические группы, представляющие из себя устойчивые словосочетания, или, например, математические формулы, имена людей, аббревиатуры и т. д. [10; 11]. Для кластеризации оптимальный выбор коэффициентов подобия определяется полученными результатами, дающими наилучший набор документов с точки зрения семантики.

ISSN 2500-2759

В рамках концепции TF-IDF каждому признаку f_k в документе d_j присваивается вес, который вычисляется по формуле

$$w_{k,j} = \frac{(1+\log(N_{j,k}))\log\frac{|D|}{N_k}}{\sqrt{\sum_{s\neq k}(\log(N_{j,s})+1)^2}},$$

где $N_{j,k}$ — количество появлений признака f_k в документе d_j; $|D|$ — мощность множества D; N_k — количество появлений признака во всех документах.

Значимым критерием веса является частота появления признака в тексте с учетом информативности, т. е. чем чаще слово появляется в тексте, тем выше его вес.

Собственно кластеризация текстовой информации проводится после сингулярного разложения матрицы «слово-на-документ» и выделения базовых параметров, которые обладают наибольшей степенью изменчивости в выбранной совокупности слов и документов [12; 13]. Различные методы кластеризации базируются на определении мер расстояния в метрическом пространстве. Применение такого рода коэффициентов является наиболее естественным и легко интерпретируемым. Полученные кластеры выражаются в явном виде как множество точек в евклидовом пространстве со своими геометрическими характеристиками — длиной или протяженностью (например, расстояние точки кластера до центроида, расстояние между кластерами и т. д.) [14].

Можно рассчитывать на получение удовлетворительных практических результатов кластеризации с помощью мер расстояния только в тех случаях, когда объекты обнаруживают выраженную тенденцию к проявлению кластеризационных свойств.

Ранее перечисленные теоретические обоснования применены для кластеризации информации базы данных оборудования.

Рассмотрим набор имеющихся данных. Запись базы данных состоит из наименования оборудования, его характеристики, типа, вида, фирмы и т. д., т. е. оборудование имеет качественные и количественные признаки. Для решения конкретной задачи выбора оборудования, например, датчиков давления с учетом назначения (измерение избыточного давления, абсолютного давления, давления-разрежения, разности давлений и т. д.), применялись технологии Data Mining и Text Mining. Для проведения кластеризации использовалась выборка данных, включающая в себя наименование датчиков, их стандартное обозначение, назначение средства измерения (табл. 1).

Таблица 1

Данные для анализа (фрагмент)

Наименование	Обозначение	Назначение средств измерения
1	2	3
Датчики давления	PTE5000, P1E, P1A	Измерение и непрерывное преобразование избыточного давления жидкостей и газов в нормированный выходной сигнал постоянного тока или напряжения
Датчики давления	ДЛ 001	Измерение избыточного давления, формирование и передача параметров измеряемого давления в виде цифрового сигнала по интерфейсу RS-485
Датчики давления	ДПС 025	Измерение быстропеременных давлений с амплитудой от 0,12 до 506 Мпа при статическом давлении от 2 204 до 125 Мпа в жидких и газообразных средах
Датчики абсолютного давления	Bm 222M	Измерение абсолютного давления газообразных сред

Работа проводилась в три этапа.

1. Препроцессинг (загрузка данных, морфологический и лексический анализ текстовых данных). В анализе использовалась информация о назначении средств измерения (см. табл. 1, стлб. 3). Из текста удалялись незначащие с точки зрения семантики слова (предлоги, местоимения), отбрасывались окончания из имеющихся лексем, удалялись слова, которые часто или редко присутствовали в текстах («порог встречаемости» 95 % и выше и 5 % и ниже соответственно). Например, «датчики», «давления», «предназначены», «ДЛ 001», «ДПС 025», «Bm 222M». Проблема синонимии решалась определением слов-синонимов, например, «дифференциальное давление» и «разность давлений», а также учетом данных выражений как одного и того же элемента.

2. Индексирование, «оцифровка» текста. Согласно данным о частоте вхождений отдельных слов в текстовый документ (и/или обратной частоте встречаемости) была построена матрица вхождений слов, проведено ее сингулярное разложение (для уменьшения размерности) и определены значимые параметры (табл. 2).

ISSN 2500-2759

Таблица 2

Результаты сингулярного разложения матрицы вхождений слов («слово-на-документ»)

Номер документа	Параметр 1	Параметр 2	Параметр 3	Параметр 4
1	0,128 575	–0,022 851	0,206 190	0,050 006
2	0,122 473	–0,090 507	0,327 987	0,004 621
3	0,022 780	–0,016 572	0,013 855	–0,138 132
4	0,024 011	–0,015 705	–0,000 803	–0,045 551
5	0,098 613	–0,025 327	0,088 437	–0,052 134
6	0,108 098	–0,028 416	0,124 753	0,047 647
7	0,022 780	–0,016 572	0,013 855	–0,138 132
8	0,221 776	–0,148 382	–0,011 217	0,128 207
9	0,147 844	0,087 663	0,177 324	0,030 825
10	0,209 182	–0,107 492	–0,033 142	–0,498 496

3. Кластеризация. По имеющейся информации была проведена кластеризация с применением метода *K*-средних. В качестве исходных данных использовались параметры 1–4 (см. табл. 2). Метрики расстояния вычислялись как евклидово расстояние (табл. 3).

Таблица 3

Кластеризация текстов по назначению средств измерения

Номер документа	Номер кластера	Расстояние до центроида
1	2	0,136 164
2	2	0,157 787
3	4	0,154 317
4	4	0,101 715
5	4	0,128 476
6	4	0,210 561
7	4	0,154 317
8	1	0,323 563
9	2	0,206 398
10	3	0,239 210
11	3	0,239 210
12	4	0,127 917
13	4	0,119 651
14	2	0,130 624
15	4	0,225 074
16	2	0,135 535
17	4	0,132 897
18	1	0,343 508
19	4	0,194 602
20	4	0,172 081
21	1	0,711 361
22	1	0,781 289
23	2	0,327 861
24	1	0,336 144
25	3	0,260 736
26	4	0,154 317
27	1	0,362 601
28	1	0,362 601
29	2	0,251 835
30	3	0,183 747
31	3	0,116 732
32	4	0,166 318

Таким образом, были получены 4 кластера, характеризующих семантическое сходство специфических текстов — коротких описаний назначения датчиков давления (рис.).

Кластер 1 — датчики давления, применяемые для измерения избыточного, абсолютного давления, а также разности давлений; кластер 2 — датчики давления, применяемые для измерения избыточного давления; кластер 3 — датчики, применяемые для измерения и преобразования измеряемого параметра; кластер 4 — датчики для измерения давления газообразных и/или жидких сред.

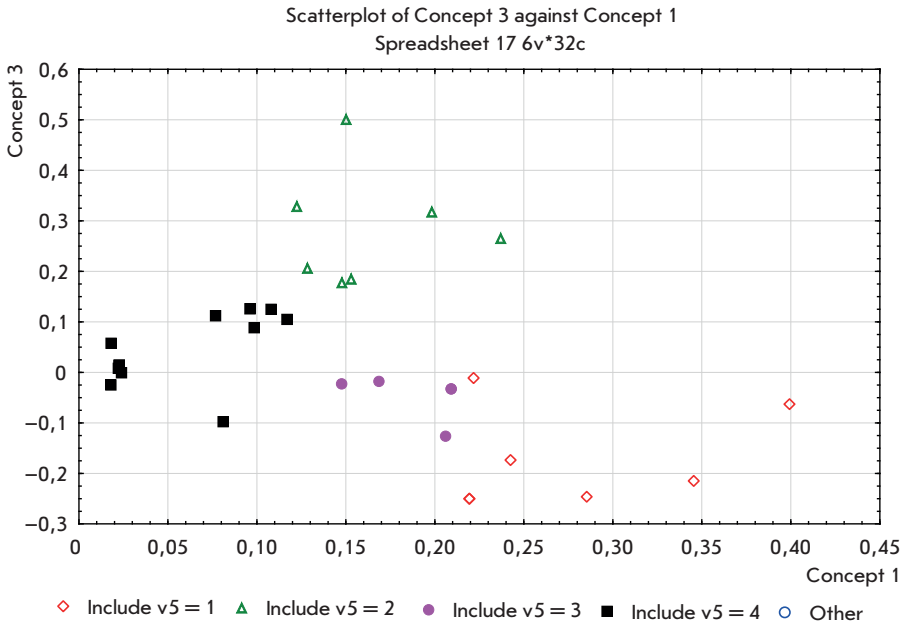
Полученные кластеры достаточно локализованы, множества точек образуют группировки, но имеется некоторая пространственная «близость» кластеров (см. рис.). Невозможно понять, к какому набору данных относятся, например, пограничные точки кластеров 1 и 3. Это объясняется следующими причинами:

1. Исходные данные для кластеризации не имеют явно выраженных пространственных компактных группировок, отличающихся друг от друга, что затрудняет определение количества кластеров, а также само разбиение на группы.

2. Описания средств измерения в определенном смысле являются однотипными. Текстовые блоки частично структурированы, имеют похожую синтаксическую схему. Почти всегда слова, из которых состоят документы, встречаются в тексте один раз, а количество слов-терминов ограничено.

Особенностью проведенного анализа является то, что полученные группировки характеризуют различные стороны применения датчиков. Например, кластер 2 включает датчики для измерения избыточного давления, а кластер 4 — датчики для измерения давления жидких или газообразных сред.

ISSN 2500-2759



Кластеризация текстов в соответствии с назначением средств измерения

В целом, результаты исследования показывают, что метод кластеризации может быть применен для автоматизации сортировки информации, поступающей в базу данных оборудования. Однако в случае неярко выраженных кластеров потребуется его сочетание с альтернативными методами, например, методом ключевых слов.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Ясенев В. Н. Информационные системы и технологии в экономике: учеб. пособие / В. Н. Ясенев. — М. : Юнити, 2008. — 560 с.

2. Прикладная статистика. Классификация и снижение размерности / С. А. Айвазян, В. М. Бухштабер, И. С. Енюков [и др.]. — М. : Финансы и статистика, 1989. — 606 с.

3. Мандель И. Д. Кластерный анализ / И. Д. Мандель. — М. : Финансы и статистика, 1988. — 176 с.

4. Классификация и кластер / под ред. Дж. Вэн Райзина. — М. : Мир, 1980. — 390 с.

5. Сэлтон Г. Автоматическая обработка, хранение и поиск информации / Г. Сэлтон. — М. : Сов. радио, 1973. — 560 с.

6. Барсегян А. А. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP / А. А. Барсегян, М. С. Куприянов, В. В. Степаненко. — СПб. : БХВ-Петербург, 2007. — 384 с.

7. Маннинг К. Д. Введение в информационный поиск / К. Д. Маннинг, П. Рагхаван, Х. Шютце. — М. ; СПб. ; Киев : Вильямс, 2011. — 520 с.

8. Salton G. Term-weighting approaches in automatic text retrieval / G. Salton, C. Buckley // Information Processing & Management. — 1988. — № 5 (24). — P. 513–523.

9. Larson R. R. Classification clustering, probabilistic information retrieval, and online catalog / R. R. Larson // Library Quarterly. — 1991. — Vol. 61 (2). — P. 133–173.

10. Леонтьева Н. Н. Автоматическое понимание текстов. Системы, модели, ресурсы : учеб. пособие / Н. Н. Леонтьева. — М. : Academa, 2006. — 202 с.

11. Катасёв А. С. Модели и методы формирования нечетких правил в интеллектуальных системах диагностики сложных объектов : дис. ... д-ра техн. наук : 05.13.18 / А. С. Катасёв. — Казань, 2014. — 257 с.

12. Боровиков В. П. STATISTICA. Искусство анализа данных на компьютере / В. П. Боровиков. — СПб. : Питер, 2003. — 686 с.

13. Халафян А. А. STATISTICA 6. Статистический анализ данных : учеб. пособие / А. А. Халафян. — М. : Бином-Пресс, 2007. — 512 с.

14. Паклин Н. Б. Бизнес-аналитика: от данных к знаниям : учеб. пособие / Н. Б. Паклин, В. И. Орешков. — СПб. : Питер, 2013. — 704 с.

REFERENCES

1. Yasenev V. N. *Informatsionnye sistemy i tekhnologii v ekonomike* [Information systems and technologies in economics]. Moscow, Yuniti Publ., 2008. 560 p.

2. Aivazyan S. A., Bukhshtaber V. M., Enyukov I. S. et al. *Prikladnaya statistika. Klassifikatsiya i snizhenie razmernosti* [Applied statistics. Classification and dimension reduction]. Moscow, Finansy i statistika Publ., 1989. 606 p.

3. Mandel' I. D. *Klasternyi analiz* [Cluster analysis]. Moscow, Finansy i statistika Publ., 1988. 176 p.

ISSN 2500-2759

4. Van Ryzin J. (ed.). *Classification and Clustering.* New York, Academic Press, 1977. (Russ. ed.: Van Ryzin J. (ed.). *Klassifikatsiya i klaster*. Moscow, Mir Publ., 1980. 390 p.).

5. Salton G. *Automatic Information Organization and Retrieval.* New York, McGraw-Hill, 1968. (Russ. ed.: Salton G. *Avtomaticheskaya obrabotka, khranenie i poisk informatsii*. Moscow, Sovetskoe Radio Publ., 1973. 560 p.).

6. Barsegyan A. A., Kupriyanov M. S., Stepanenko V.V. *Tekhnologii analiza dannykh: Data Mining, Visual Mining, Text Mining, OLAP* [Data analysis technologies: Data Mining, Visual Mining, Text Mining, OLAP]. Saint Petersburg, БХВ-Petersburg Publ., 2007. 384 p.

7. Manning C. D., Raghavan P., Schutze H. *Introduction to Information Retrieval*. Cambridge University Press, 2010. 521p. (Russ. ed.: Manning C. D., Ragkhavan P., Shutze H. *Vvedenie v informatsionnyi poisk*. Moscow, Saint Petersburg, Kiev, Vil'yams Publ., 2011. 520 p.).

8. Salton G., Buckley C. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 1988, no. 5 (24), pp. 513–523.

9. Larson R. R. Classification clustering, probabilistic information retrieval, and online catalog. *Library Quarterly*, 1991, vol. 61 (2), pp. 133–173.

10. Leont'eva N. N. *Avtomaticheskoe ponimanie tekstov. Sistemy, modeli, resursy* [Automated understanding of texts. Systems, models, resources]. Moscow, Academy Publ., 2006. 202 p.

11. Katasev A. S. *Modeli i metody formirovaniya nechetkikh pravil v intellektual'nykh sistemakh diagnostiki slozhnykh ob"ektov. Dokt. Diss.* [Models and methods for making fuzzy rules in intelligent systems for diagnosing complex objects. Doct. Diss.]. Kazan, 2014. 257 p.

12. Borovikov V. P. *STATISTICA. Iskusstvo analiza dannykh na komp'yutere* [STATISTICA. The art of the computer aided data analysis]. Saint Petersburg, Piter Publ., 2003. 686 p.

13. Khalafyan A. A. *STATISTICA 6. Statisticheskii analiz dannykh* [STATISTICA 6. Statistical data analysis]. Moscow, Binom-Press, 2007. 512 p.

14. Paklin N. B., Oreshkov V. I. *Biznes-analitika: ot dannykh k znaniyam* [Business analytics: from the data to knowledge]. Saint Petersburg, Piter Publ., 2013. 704 p.

Информация об авторах

Боровский Андрей Викторович — доктор физико-математических наук, профессор, кафедра информатики и кибернетики, Байкальский государственный университет, 664003, г. Иркутск, ул. Ленина, 11, e-mail: andrei-borovskii@mail.ru.

Раковская Елена Евгеньевна — аспирант, кафедра информатики и кибернетики, Байкальский государственный университет, 664003, г. Иркутск, ул. Ленина, 11, e-mail: rakovskaya19@mail.ru.

Бисикало Артем Леонидович — кандидат химических наук, доцент, кафедра аналитической химии, Иркутский государственный университет, 664003, г. Иркутск, ул. К. Маркса, 1, e-mail: bisikalo.a@yandex.ru.

Библиографическое описание статьи

Боровский А. В. Кластеризация информации баз данных оборудования / А. В. Боровский, Е. Е. Раковская, А. Л. Бисикало // Известия Байкальского государственного университета. — 2016. — Т. 26, № 5. — С. 805–810. — DOI: 10.17150/2500-2759.2016.26(5).805-810.

Authors

Andrei V. Borovsky — Doctor habil. (Physical and Mathematical Sciences), Professor, Baikal State University, 11 Lenin St., 664003, Irkutsk, Russian Federation, e-mail: andrei-borovskii@mail.ru.

Elena E. Rakovskaya — PhD Student, Department of Computer Science and Cybernetics, Baikal State University, 11 Lenin St., 664003, Irkutsk, Russian Federation, e-mail: rakovskaya19@mail.ru.

Artem L. Bisikalo — PhD in Chemistry, Associate Professor, Department of Analytical Chemistry, Irkutsk State University, 1 Karl Marx St., 664003, Irkutsk, Russian Federation, e-mail: bisikalo.a@yandex.ru.

Reference to article

Borovsky A. V., Rakovskaya E. E., Bisikalo A. L. Clustering of information from of the equipment data bases. *Izvestiya Baykal'skogo gosudarstvennogo universiteta = Bulletin of Baikal State University*, 2016, vol. 26, no 5, pp. 805–810. DOI: 10.17150/2500-2759.2016.26(5).805-810. (In Russian).

ISSN 2500-2759