Рычагов С.А. Использование латентно-семантического анализа для автоматической классификации текстов // Международный журнал информационных технологий и энергоэффективности. – 2017. – Т.2 №2 с. 28-33

Международный журнал информационных технологий и энергоэффективности

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/

УДК: 004.423.4 (043.3)

ИСПОЛЬЗОВАНИЕ ЛАТЕНТНО-СЕМАНТИЧЕСКОГО АНАЛИЗА ДЛЯ АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ ТЕКСТОВ

Рычагов С.А.

Филиал федерального государственного бюджетного образовательного учреждения высшего образования «Национальный исследовательский университет «МЭИ» в г. Смоленске (214013, г. Смоленск, пр-д Энергетический, 1); e-mail: rychagov_s@rambler.ru

Статья посвящена изложению подхода к использованию латентно-семантического анализа для автоматической классификации текстов. В ходе исследования сделано заключение о том, что применение латентно-семантического анализа снижает нагрузку по сравнению со стандартными векторными методами, а также эффективность метода повышается, если предварительно обрабатывать входные данные для фильтрации шума.

Ключевые слова: латентно-семантический анализ, классификация.

USAGE OF A LATENT-SEMANTIC ANALYSIS FOR AUTOMATIC CLASSIFICATION OF TEXTS

Rychagov S.A.

Branch of the federal state budget educational institution Higher Education "National Research University" MPEI" In Smolensk city (214013, Smolensk, Pr-d Energetichesky, 1); e-mail: rychagov_s@rambler.ru

The article is devoted to an approach to the use of latent-semantic analysis for automatic classification of texts. In the course of the study, it was concluded that the use of latent-semantic analysis reduces the load compared with standard vector methods, and the effectiveness of the method is enhanced if the input data for noise filtering is pre-processed.

Keywords: latent-semantic analysis, classification.

В настоящее время актуальной проблемой является задача автоматической классификации текстов. Эта задача получила большое распространение в связи с увеличением числа документов, хранящихся в электронном виде, и необходимости их упорядочить. В качестве примера областей, где эта задача приобретает особую актуальность, можно рассмотреть автоматизированную оценку свободных развернутых

ответов, классификацию текстов с целью упорядочения данных в пределах научной области, поиск похожих по смыслу текстов в поисковых системах.

В настоящее время существует много различных методов классификации текста. Большая их часть использует вероятностный подход, нейронные сети или деревья решений. Требованиями к методу классификации в виду больших объемов обрабатываемых данных являются эффективность и масштабируемость, а также способность обхода зашумленности данных излишней бесполезной информацией. [1]

Метод латентно-семантического анализа является одним из самых эффективных и перспективных методов классификации. Основная идея метода заключается в следующем: если в исходном вероятностном пространстве, которое состоит из векторов слов (вектор – предложение, абзац, документ и т.п.), между двумя любыми словами из двух различных векторов может не наблюдаться никакой зависимости, то после некоторого алгебраического преобразования данного векторного пространства эта зависимость может появиться, причем величина этой зависимости будет определять силу ассоциативно-семантической связи между этими двумя словами. [2, 3]

Исходной информацией для метода является матрица, в строках которой содержатся термы, а в столбцах – тексты, документы. Эта матрица описывает данные, используемые для обучения системы. Ее элементы содержат веса, учитывающие частоту, с которой используется каждый из термов в каждом тексте.

В стандартном алгоритме не предусматривается предварительная обработка исходных данных, которая, тем не менее, может сильно уменьшить размерность матрицы и повысить эффективность работы метода.

Предварительную обработку можно осуществить с помощью выполнения следующих действий:

- удалить строки, которые соответствуют стоп-словам (стоп-слова (иначе называемые шумовыми) – это слова, знаки, символы, которые самостоятельно не несут никакой смысловой нагрузки и просто игнорируются поисковыми системами при осуществлении ранжирования или индексации сайтов);
- удалить строки, которые содержат слова, встречающиеся только один раз в тексте выборки;
- привести слова к исходной форме;
- в некоторых тематиках резонно удалять имена собственные и числовую информацию, если они в рамках данной тематики не несут смысловой нагрузки.

Наиболее распространенный вариант ЛСА основывается на использовании разложения матрицы весов по сингулярным значениям (Singular Value Decomposition). [4] С помощью него любую матрицу можно разложить на множество ортогональных матриц, линейная комбинация которых является достаточно точным приближением к исходной матрице.

Согласно теореме о сингулярном разложении в самом простом случае матрица может быть разложена на произведение трех матриц:

$$A = USV^T , \quad (1)$$

где A – исходная матрица;

U и VT – ортогональные матрицы;

S – диагональная матрица, значения на диагонали которой называются сингулярными коэффициентами матрицы A. [5]

Основная идея заключается в том, что если матрица A является терм-документной, то матрица A*, содержащая только k первых линейно-независимых компонент, отражает основную структуру различных зависимостей, присутствующих в исходной матрице. Структура зависимостей определяется весовыми функциями термов.

Как правило, выбор k зависит от поставленной задачи. При выборе большого значения метод может потерять вычислительную мощность, а при выборе слишком маленького появляется риск не учитывать разницу между похожими термами. При выборе значения k автоматически можно, например, установить пороговое значение сингулярных коэффициентов и отбрасывать все строки и столбцы, соответствующие сингулярным коэффициентам, не превышающим данного порогового значения. [6, 7]

Схожесть между любой комбинацией термов и/или документов чаще всего вычисляют с помощью скалярного произведения их векторов. На практике хороших результатов можно добиться, используя коэффициент корреляции Пирсона. [8, 9]

Метод латентно-семантического анализа производит отображение документов и отдельных слов в семантическое пространство. В нем в свою очередь проводят последующие сравнения. Используются следующие допущения:

- документ является набором слов. В каком порядке они расположены – не важно, имеет значение только то, сколько раз слово встретилось в тексте;
- семантическое значение определяется набором слов, которые, как правило, идут вместе;
- необходимым допущением является единственность значения для каждого из слов.

Порядок использования метода латентно-семантического анализа рассмотрим на примере.

Пусть после предварительной обработки исходных данных имеем следующий набор документов (индексируемые слова подчеркнуты):

iPhone – телефон, обладающий высокой стоимостью;
российский депутат пользуется iPhone;
дороги в России в плохом состоянии;
повышается стоимость строительства дорог;
вдоль дорог проходит ремонт телефонных линий;
в Российском театре откладываются спектакли из-за ремонта;
стоимость продуктов в России сегодня не упала;
стоимость древесины в России завтра изменится;
в сервисных центрах осуществляют ремонт телефонов.

Сначала составляется частотная матрица индексируемых слов (таблица 1). В ней строки соответствуют индексированным словам, а столбцы — документам. В каждой ячейке матрицы указано сколько раз слово встречается в соответствующем документе.

Таблица 1 – Частотная матрица индексируемых слов

	T1	T2	T3	T4	T5	T6	T7	T8	T9
iPhone	1	1	0	0	0	0	0	0	0
Телефон	1	0	0	0	1	0	0	0	1
Стоимость	1	0	0	10	0	0	1	1	0
Дорог	0	0	1	1	1	0	0	0	0
Росси	0	1	1	0	0	1	1	1	0
Ремонт	0	0	0	0	1	1	0	0	1

Далее проводится сингулярное разложение полученной матрицы.

Singular values:

2.8556 2.1637 1.7592 1.5675 1.1772 0.4756

Matrix U:

0.2344 0.0628 -0.3400 -0.4644 0.6618 -0.4146
0.3310 -0.5841 -0.3099 -0.2210 0.0221 0.6355
0.5085 0.2822 -0.5887 0.2259 -0.4746 -0.1975
0.3404 -0.2270 0.1300 0.7712 0.4648 -0.0704
0.6033 0.4703 0.5263 -0.2172 0.0460 0.2975
0.3115 -0.5501 0.3846 -0.2064 -0.3438 -0.5401

Matrix V:

0.3761 -0.1105 -0.7040 -0.2931 0.1777 0.0491 -0.2990 -0.2990 -0.2280
0.2934 0.2464 0.1059 -0.4348 0.6013 -0.2463 0.2990 0.2990 0.2280
0.3305 0.1124 0.3731 0.3534 0.4339 0.4774 -0.3029 -0.3029 0.1196
0.2973 0.0255 -0.2607 0.6361 -0.0083 -0.5634 -0.0039 -0.0039 0.3476
0.3442 -0.6291 0.1164 0.2193 0.1215 0.0525 0.3068 0.3068 -0.4673
0.3204 -0.0369 0.5178 -0.2702 -0.2529 -0.5100 -0.2990 -0.2990 -0.2280
0.3893 0.3478 -0.0354 0.0056 -0.3640 0.2102 0.6514 -0.3486 -0.0598
0.3893 0.3478 -0.0354 0.0056 -0.3640 0.2102 -0.3486 0.6514 -0.0598
0.2250 -0.5242 0.0425 -0.2726 -0.2733 0.2006 -0.0078 -0.0078 0.6953

Сингулярное разложение выделяет ключевые составляющие матрицы, тем самым позволяя игнорировать шумы. Столбцы и строки соответствующие меньшим сингулярным значениям дают наименьший вклад в матричное произведение. Например, можно отбросить последние столбцы матрицы U и последние строки матрицы Vt, оставив только первые 2. Важно, что при этом гарантируется оптимальность полученного произведения.

Разложение такого вида называется двумерным сингулярным разложением:

Singular values:

2.8556 2.1637

Matrix U:

0.2344 0.0628
0.3310 -0.5841
0.5085 0.2822
0.3404 -0.2270
0.6033 0.4703
0.3115 -0.5501

Matrix V:

0.3761 -0.1105 -0.7040 -0.2931 0.1777 0.0491 -0.2990 -0.2990 -0.2280
0.2934 0.2464 0.1059 -0.4348 0.6013 -0.2463 0.2990 0.2990 0.2280

Согласно полученному разложению, отметим на графике точки, соответствующие отдельным текстам и словам (рисунок 1).

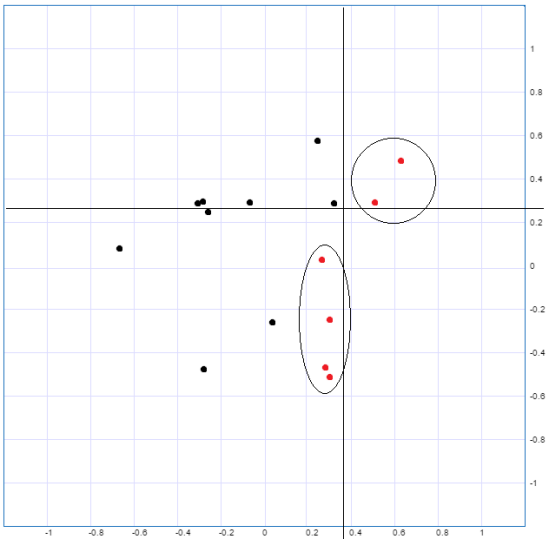

Рисунок 1 – Графическое представление разложения

Из рисунка 1 видно, что тексты образуют две группы, первая из которых связана с термами «Россия, стоимость», вторая – с остальными четырьмя. На практике, при наличии огромного количества исходных данных, значительно увеличится количество групп, пространство будет являться многомерным, однако сама суть метода останется без изменений.

Таким образом, использование латентно-семантического анализа для автоматической классификации текстов целесообразно, поскольку его применение значительно снижает нагрузку по сравнению со стандартными векторными методами. Эффективность метода повышается при использовании предварительной обработки входных данных для фильтрации шумовой информации.

Список литературы

1. F. Sebastiani Machine Learning in Automated Text Categorization ACM Computing Surveys (CSUR) 34 (1), 1-47
2. Некрестьянов И.С. Тематико-ориентированные методы информационного поиска / Диссертация на соискание степени к. ф-м.н. СПбГУ, 2000.
3. Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman (1990). «Indexing by Latent Semantic Analysis» (PDF). Journal of the American Society for Information Science 41 (6): 391–407.
4. Thomas Landauer, Peter W. Foltz, Darrell Laham Introduction to Latent Semantic Analysis Discourse Processes 25: 259–284.
5. В. В. Стрижов. «Информационное моделирование». Конспект лекций.
6. Dumais, S. T., Furnas, G. W., Landauer, T. K. and Deerwester, S. (1988). Using latent semantic analysis to improve information retrieval. In Proceedings of CHI'88: Conference on Human Factors in Computing, New York: ACM, 281-285
7. Голуб Дж., Ван Лоун Ч. Матричные вычисления. М.: «Мир», 1999.
8. Гмурман В. Е. Теория вероятностей и математическая статистика: Учебное пособие для вузов. — 10-е издание, стереотипное. — Москва: Высшая школа, 2004. — 479 с
9. Общая теория статистики: Учебник / Под ред. Р. А. Шмойловой. — 3-е издание, переработанное. — Москва: Финансы и Статистика, 2002. — 560 с.

References

1. F. Sebastiani Machine Learning in Automated Text Categorization ACM Computing Surveys (CSUR) 34 (1), 1-47
2. Nekrestyanov IS Subject-oriented methods of information retrieval / Thesis for a degree of Ph.D. SPbSU, 2000.
3. Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman (1990). "Indexing by Latent Semantic Analysis" (PDF). Journal of the American Society for Information Science 41 (6): 391-407.
4. Thomas Landauer, Peter W. Foltz, Darrell Laham Introduction to Latent Semantic Analysis Discourse Processes 25: 259-284.
5. V. V. Strizhov. "Information modeling". Lecture notes.
6. Dumais, S. T., Furnas, G. W., Landauer, T. K. and Deerwester, S. (1988). Using latent semantic analysis to improve information retrieval. In Proceedings of CHI'88: Conference on Human Factors in Computing, New York: ACM, 281-285
7. Golub J., Van Lown C. Matrix calculations. M .: Mir, 1999.
8. Gmurman VE Theory of Probability and Mathematical Statistics: A Textbook for Higher Education. - 10th edition, stereotyped. - Moscow: Higher School, 2004. - 479 s
9. General Theory of Statistics: Textbook / Ed. R. A. Shmoilova. - 3rd edition, revised. - Moscow: Finance and Statistics, 2002. - 560 pp.