Федеральное агентство по образованию

Государственное образовательное учреждение высшего профессионального образования

Вятский государственный гуманитарный университет

Реферат

Выделение ключевых слов в текстовых документах

Киров 2008 г.

# Введение

Для решения задач классификации текстов, документы в непосредственном виде не подходят для интерпретации классификатором. Поэтому необходимо применение процедуры индексации, которая переводит текст в удобное представление. В частности стандартные методы классификации, такие как нейронные сети и деревья решений, требуют представление документа в виде вектора из n взвешенных признаков. Построение такого вектора называется выделением признаков, а само представление векторной моделью документа.

Различия в подходах заключаются:

1) в понимании, что такое термин;

2) в способах определения веса термина.

Обычно признакам соответствуют все или некоторые отдельно взятые слова документа. В ряде экспериментов было обнаружено, что чуть даже более сложное представление менее эффективно. В частности, некоторые авторы пробовали использовать группы слов (стилистические, синтаксические) в качестве признаков. Д.Д. Льюис [5] довольно убедительно утверждает, что, скорее всего, причиной неутешительных результатов является то, что методы индексирования на основе фраз обладают худшими статистическими характеристиками по отношению к методам на основе одиночных слов, хотя их семантические качества гораздо выше.

Одним из самых распространенных методов перехода к математической модели документа, является «метод ключевых слов». Ключевое слово – слово в тексте, способное в совокупности с другими ключевыми словами представлять текст. Суть метода в следующем. Для каждого класса текстов создается список характерных для него слов, тогда каждый текст можно представить в виде вектора частот появлении в нём слов из данного списка [9]. Возникает проблема поиска и выделения из текста слов, которые будут для него ключевыми. Огромный объем информации, который подлежит обработке, делают особенно актуальной задачу автоматического выделения ключевых слов. Причем от чистоты этого выделения напрямую зависит точность классификации.

Целью работы является рассмотреть методы отделения ключевых слов в текстовых документах.

В работе, в основном, затрагиваются статистические методы, основанные на законах Ципфа, а также модель TF\*IDF.

В экспериментальном разделе описывается модуль, основанный на модели TF\*IDF, приведена оценка его возможностей для выделения ключевых слов из коллекции документов.

**1. Статистика в текстах: законы Ципфа**

Во всех текстовых документах, созданных человеком, можно выделить статистические закономерности. В любом языке есть слова, которые встречаются чаще, чем остальные, но не имеют значения. Есть слова, которые встречаются реже, но имеют намного большее смысловое значение.

В 1949 году Джордж Ципф (George Kingsley Zipf) гарвардский профессор-лингвист и филолог, работая над принципом наименьшего усилия, сформулировал несколько закономерностей. Данные законы получены не на основе математических выводов, а на основе анализа статистики частоты слов текстах на многих языках, то есть эмпирически.

В то время, когда Ципф сформулировал подмеченные им закономерности распределения частоты слов, законом они не считались – еще не было компьютеров и нельзя было провести точные расчеты, подтверждающие выявленные закономерности. В последующем были проведены многочисленные исследования, которые подтвердили и уточнили подмеченные закономерности. Также ведущую роль в обосновании законов сыграли работы Б. Мандельброта.

В частности Ципф положил, что слова с большим количеством букв встречаются в тексте реже коротких слов. Основываясь на этом постулате, Ципф вывел два универсальных закона.

## 1.1 Первый закон Ципфа («ранг – частота»)

Измерим количество вхождений каждого слова в текст и возьмем только одно значение из каждой группы, имеющей одинаковую частоту. Расположим частоты по мере их убывания и пронумеруем, порядковый номер частоты назовем рангом частоты (обозначим ранг слова ). Наиболее часто встречающиеся слова будут иметь ранг 1, следующие за ними – 2 и так далее.

Тогда очевидно, что вероятность встретить произвольное, заранее выбранное слово будет равна отношению количества вхождений этого слова к общему числу слов в тексте ( – количество вхождений слова, - количество слов в тексте).

 (1.1)

Ципф обнаружил следующую закономерность: произведение вероятности обнаружения слова в тексте на ранг частоты, есть число постоянное (С).

 (1.2)

Закон показывает, распространенность слова в тексте изменяется по гиперболе, в зависимости от количества вхождений. Например второе по используемости слово встречается примерно в два раза реже, чем первое, третье – в три раза реже, чем первое, и так далее.

Значение константы в разных языках различно, но внутри одной языковой группы остается примерно неизменной, какой бы текст мы ни взяли. Для русских текстов константа Ципфа приблизительно равна 0,08, для английских текстов 0,1. На рисунке 1.1. показано распределение частот слов в зависимости от ранга по материалам русской Википедии.

## 1.2 Второй закон Ципфа («количество – частота»)

Первый закон не учитывает факт того, что разные слова могут входить в текст с одинаковой частотой. Ципф установил, что частота и количество слов, входящих в текст с этой частотой, также имеют зависимость. Если построить график, отложив по оси абсцисс частоту вхождения слова, а по оси ординат – количество слов в данной частоте, то получившаяся кривая будет сохранять свой вид для всех без исключения текстов.

Как и для первого закона, это утверждение верно в пределах одного языка. Однако и межъязыковые различия невелики. На каком бы языке текст ни был написан, вид кривой Ципфа останется неизменной. Может немного отличаться лишь коэффициент гиперболы.

Дж. Ципфом и другими исследователями было установлено, что гиперболическому распределению подчиняются не только все естественные языки мира (рис. 1.2.), но и другие явления социального и биологического характера: распределения ученых по числу опубликованных ими статей, городов США по численности населения, населения по размерам дохода в капиталистических странах, биологических родов по численности видов и другие.

Также важным является тот факт, что и документы внутри какой-либо отрасли знаний могут распределяться согласно этому закону.

Законы Ципфа позволяют находить ключевые слова.

Воспользуемся первым законом Ципфа и построим график зависимости ранга от частоты. Исследования показывают, что наиболее значимые для текста слова лежат в средней части графика (рис. 1.3.). Этот факт имеет простое обоснование. Слова, которые попадаются слишком часто, в основном оказываются предлогами, местоимениями. Редко встречающиеся слова тоже, в большинстве случаев, не имеют решающего смыслового значения.

От установки ширины зависит качество отделения значимых слов. Если большую ширину диапазона, то в ключевые слова будут попадать вспомогательные слова; если установить узкий диапазон – можно потерять смысловые термины. Поэтому в каждом отдельном случае необходимо использовать ряд эвристик, для определения ширины диапазона, а также методиками, уменьшающих влияние этой ширины.

Одним из способов, например, является предварительное исключение из исследуемого текста слов, которые изначально не могут являться значимыми и, поэтому, являющиеся «шумом». Такие слова называются нейтральными или стоповыми (стоп-словами).

Для русского текста стоповыми словами могли бы являться все предлоги, частицы, личные местоимения. Есть и другие способы повысить точность оценки значимости слов.

Некоторые слова могут встречаться почти во всех документах некоторой коллекции и, соответственно, оказывать малое влияние на принадлежность документа к той или иной категории, а значит не быть ключевыми для этого документа. Поэтому очевидно, что, рассматривая всю коллекцию документов, мы повысим информативность выделения ключевых слов.

# 2. Глобальная статистика, модель TF\*IDF

Выше отмечалось, к коллекции документов тоже применимы законы Ципфа. Для понижения значимости слов, которые встречаются почти во всех документах, вводят инверсную частоту термина IDF (inverse document frequency) – это логарифм отношения числа всех документов () к числу документов содержащих некоторое слово t (2.1.). Значение этого параметра тем меньше, чем чаще слово встречается в документах базы данных. Таким образом, для слов, которые встречаются в большом числе документов IDF будет близок к нулю (если слово встречается во всех документах IDF равен нулю), что помогает выделить важные слова.

 (2.1)

Параметр TF (term frequency) – это отношение числа раз, которое некоторое слово t встретилось в документе d, к длине документа (2.2.). Нормализация длиной документа нужна для того, чтобы уравнять в правах короткие и длинные (в которых абсолютная встречаемость слов может быть гораздо больше) документы.

 (2.2.)

Коэффициент TFIDF равен произведению TF и IDF. TF играет роль повышающего множителя, IDF – понижающего. Тогда весовыми параметрами векторной модели некоторого документа можно принять коэффициенты TF\*IDF входящих в него слов [10].

Для того чтобы веса находились в интервале (0, 1), а векторы документов имели равную длину, значения TF\*IDF обычно нормализуются по косинусу.

Отметим, что эта формула оценивает значимость термина только с точки зрения частоты вхождения в документ, тем самым не учитывая порядок следования терминов в документе и их синтаксическую роль; другими словами, семантика документа сводится к лексической семантике входящих в него терминов, а композиционная семантика не рассматривается.

Ключевыми в данном случае будут являться слова набравший наибольший вес. Слова с малым весом, вообще можно не учитывать при классификации.

Проиллюстрируем на простом примере.

Пусть коллекция состоит из 2 предложений.

1). Мама мыла мылом Машу.

2). Мама мыла, мыла раму.

3). В магазине купила мама мыло.

Вид словаря тогда будет следующим:

|  |  |  |  |
| --- | --- | --- | --- |
| Слово | Всего | Встретилось в предложениях | IDF |
| Мама | 3 | 3 | 0 |
| мыть | 3 | 2 | 0,18 |
| мыло | 2 | 2 | 0,47 |
| Маша | 1 | 1 | 0,47 |
| рама | 1 | 1 | 0,47 |
| магазин | 1 | 1 | 0,47 |
| купить | 1 | 1 | 0,47 |

Вид векторов

|  |  |  |
| --- | --- | --- |
| 1 | 2 | 3 |
| Слово | TF | TF\*IDF | Слово | TF | TF\*IDF | Слово | TF | TF\*IDF |
| Маша | 0,25 | 0,12 | рама | 0,25 | 0,11 | магазин | 0,25 | 0,12 |
| мыло | 0,25 | 0,12 | мыть | 0,5 | 0,09 | купить | 0,25 | 0,12 |
| мыть | 0,25 | 0,05 | мама | 0,25 | 0 | мыло | 0,25 | 0,12 |
| мама | 0,25 | 0 |  |  |  | мама | 0,25 | 0 |

Влияние TF видно во втором векторе. Так как слово «мыть» встречается 2 раза, он выше, чем у остальных слов. Однако из-за того, что это слово встречается и в других документах, у него ниже параметр IDF, поэтому его общий вес в векторе будет ниже, чем у слова «рама». Так влияет параметр IDF.

Слово «мама» же вообще можно не учитывать в векторном представлении. Так как оно встречается во всех предложениях коллекции, его значение TF\*IDF всегда будет равно нулю.

Заметим, что все слова примера мы приводим к нормальной форме (лематизируем). Существуют противоречивые мнения относительно полезности данного шага в текстовой категоризации. Некоторые исследования (Baker, McCallum) отмечают снижение эффективности при использовании морфологической обработки, хотя в основном многие прибегают к ней, поскольку это способствует значительному сокращению размерности пространства.

Еще одним способом к сокращению словаря является возможный учет синонимии, так что слова – синонимы, объявляются одним термином словаря

Конечно, при данном подходе есть вероятность попадания в ключевые слова случайных специальных терминов, редких слов и имен собственных и другого «шума». Поэтому необходимо в предобработку текста включать алгоритмы повышающее качество отбора. Эвристики такого отбора чаще зависят от конкретно взятого случая.

Модель TF\*IDF является, пожалуй, наиболее популярной. Однако используются и другие индексирующие функции, включая вероятные способы индексирования [3] и методики индексирования структурированных документов [4]. Иные функции индексации могут потребоваться в тех случаях, когда изначально обучающее множество не дано и документную частоту не удаётся посчитать. В этих случая TF\*IDF меняют на эмпирические функции [2].

# 3. Экспериментальная оценка статистического анализа текста по модели TF\*IDF

Для оценки выделения ключевых слов с помощью модели TF\*IDF был разработан модуль, реализующий данный алгоритм. Целью эксперимента является оценка алгоритма.

В качестве входных примеров было использованы две коллекции документов. Коллекция COMPUTER включает в себя 450 статей по общекомпьютерной тематике (материал из электронной версии журнала «Компьютера»), коллекция ANIMAL включает 190 статей о животных (материал из Википедии).

|  |  |  |
| --- | --- | --- |
| Название | Количество документов | Суммарный объем |
| COMPUTER | 450 | 12,6 Мб |
| ANIMAL | 190 | 4,1 Мб |

Для каждого документа строилась векторная модель, в качестве ключевых брались 20 слов, набравших наибольший вес.

По каждому документу из коллекции проводилась экспертная оценка от 0 до 10 баллов (0 – ни одно из слов не может являться ключевым, 10 – все слова ключевые для данного документа). Данные по каждой коллекции усреднялись.

Эксперимент проходил в две стадии.

В первой стадии, для каждого документа коллекции была произведена следующая предварительная обработка.

* лематизация – приведение слова к нормальной форме (проводилась с помощью парсера mystem от компании Yandex);
* удаление стоповых слов (союзы, предлоги, некоторые наречия, одиночные буквы и цифры).

В ходе эксперимента были получены следующие результаты.

|  |
| --- |
| COMPUTER |
| Балл | Количество оценок | Количество, в процентах |
| 0–2 | 0 | 0 |
| 3–5 | 77 | 17,11 |
| 6–8 | 324 | 72,00 |
| 9–10 | 49 | 10,89 |
| Средняя оценка | 6,73 |

|  |
| --- |
| ANIMAL |
| Балл | Количество оценок | Количество, в процентах |
| 0–2 | 0 | 0 |
| 3–5 | 6 | 3,16 |
| 6–8 | 132 | 69,47 |
| 9–10 | 52 | 27,37 |
| Средняя оценка | 7,87 |

На второй стадии, помимо предобработки, проводившейся на первой стадии, были произведены дополнительные меры. Список стоповых слов был расширен некоторыми словами, не несущими смысловой нагрузки (например: глаголы быть, мочь), не входивших в первоначальный список.

Также была отделена некоторая часть слов, согласно законам Ципфа. Для каждого документа был построен вектор статистики входящих в него слов, и убирались слова с низкой оценкой. Параметры сокращения выбирались эмпирически и составили примерно 5%. Слова с высокой оценкой не убирались, так как слова не несущие смысловой нагрузки, но часто встречающиеся в документе, в большинстве своем отделились на этапе удаления стоповых слов.

На втором этапе получены следующие результаты.

|  |
| --- |
| COMPUTER |
| Балл | Количество оценок | Количество, в процентах |
| 0–2 | 0 | 0 |
| 3–5 | 64 | 14,22 |
| 6–8 | 338 | 75,11 |
| 9–10 | 48 | 10,67 |
| Средняя оценка | 6,8 |

|  |
| --- |
| ANIMAL |
| Балл | Количество оценок | Количество, в процентах |
| 0–2 | 0 | 0 |
| 3–5 | 0 | 0 |
| 6–8 | 135 | 71,05 |
| 9–10 | 55 | 28,95 |
| Средняя оценка | 8,07 |

Сводные диаграммы по обоим этапам (рисунки 3.1, 3.2).

Полученные результаты показывают, что метод вполне справляется с отделением ключевых слов. Однако часто высокие позиции занимают слова, не являющиеся основными для документа.

Предобработка документов, снижение уровня «шума» в документе привела к повышению качества выделения ключевых слов.

# Заключение

Законы Ципфа описывают любой текст на основе частотного анализа вхождения слов в текст. Однако этого явно недостаточно для оценки документа в коллекции. Модель TF\*IDF позволяет перейти к математической, векторной модели текста, выделить список ключевых слов.

В ходе проведенного эксперимента показана возможность применения модели на реальных примерах. Найдены некоторые эвристические приемы, позволяющие улучшить выделение ключевых слов: расширение списка стоповых слов, статистическое отделение незначащих слов.

Из преимуществ метода следует отметить высокую производительность, гибкость к данным.

Однако у этого метода есть существенный недостаток: при построении вектора не учитывается порядок слов, контекст, то есть важная семантическая составляющая текста.

Из возможных перспективных улучшений метода отметим:

* автоматизация выбора эвристик для расширения стоп-листа;
* автоматизация выбора параметра при отбрасывании не несущих смысловой нагрузки слов по законам Ципфа;
* при построении учитывать расположение слов в документе;
* объединение, разбиение текста для возможно более качественного построения векторного представления.

# Библиографический список

1. Apte, C., Damerau, F.J., Weiss, S.M., Automated learning of decision rules for text categorization. ACM Transactions on Information Systems 12, 3, 233–251., 1994
2. Dagan, I., Karov, Y., Roth, D., Mistake-driven learning in text categorization. In Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (Providence, US, 1997), pp. 55–63., 1997
3. Fuhr, N., Govert, N., Lalmas, M., and Sebastiani, F., Categorisation tool: Final prototype. Deliverable 4.3, Project LE4–8303 «EUROSEARCH», Commission of the European Communities, 1998
4. Larkey, L.S., Croft, W.B., Combining classifiers in text categorization. In Proceedings of SIGIR‑96, 19th ACM International Conference on Research and Developmentin Information Retrieval (Zurich, CH, 1996), pp. 289–297., 1996
5. Lewis, D.D., An evaluation of phrasal and clustered representations on a text categorization task. In Proceedings of SIGIR‑92, 15th ACM International Conference on Researchand Development in Information Retrieval (Kobenhavn, DK, 1992), pp. 37–50., 1992
6. Salton, G. and McGill, M.J. Introduction to modern information retrieval. McGraw-Hill, 1983.
7. T. Joachims A probabilistic analysis of the rocchio algorithm with TFIDF for text categorization In Proc. of the ICML'97, 143–151, 1997.
8. Андреев А.М. Березкин Д.В. Сюзев В.В., Шабанов В.И. Модели и методы автоматической классификации текстовых документов // Вестн. МГТУ. Сер. Приборостроение. М.: Изд-во МГТУ. – 2003. – №3.
9. Андреев А.М., Березкин Д.В., Морозов В.В., Симаков К.В. Автоматическая классификация текстовых документов с использованием нейросетевых алгоритмов и семантического анализа НПЦ «ИНТЕЛЛЕКТ ПЛЮС»