Автоматический метод оценки тематической содержательности документов*

© Владыкин А.

Санкт-Петербургский государственный университет информационных технологий, механики и оптики
vladykin@gmail.com

Аннотация

Для эффективной работы с постоянно растущим объемом данных необходимы методы и средства поиска, фильтрации, структуризации и категоризации информации. В настоящей работе описывается метод автоматической оценки тематической содержательности документа, который можно использовать для повышения релевантности результатов поиска за счет их классификации и ранжирования по предлагаемым критериям.

1. Введение

Мы являемся свидетелями бурного роста объема цифровой информации. Нарастающими темпами увеличивается Интернет. В значительной степени его рост обусловлен увеличением числа пользователей и реализацией концепции Web 2.0 с ее идеей о генерируемом пользователями контенте.

Для эффективной работы с этим колоссальным объемом данных необходимы методы и средства поиска, фильтрации, структуризации и категоризации информации. Классическая задача информационного поиска — поиск документов, удовлетворяющих запросу, в рамках некоторой

* Автор выражает благодарность научному руководителю Некрестьянову И.С. за ценные замечания по данной работе.

коллекции документов — в настоящее время решается весьма успешно. Однако количество найденных по запросу документов зачастую слишком велико, чтобы пользователь мог просмотреть каждый из них и найти то, что ему действительно подходит.

Указанная проблема привела к разработке различных методов ранжирования и классификации/кластеризации результатов поиска. В настоящей работе описывается метод автоматической оценки тематической содержательности документа, который можно использовать для повышения релевантности результатов поиска за счет их классификации и ранжирования по предлагаемым критериям.

2. Представление результатов поиска

2.1 Ранжирование

Ранжирование — упорядочение документов по релевантности. Ранжирование результатов направлено на то, чтобы лучшие из найденных документов оказались в начале списка результатов, и пользователю не пришлось просматривать весь список.

В настоящее время наиболее популярными являются методы ранжирования, основанные на ссылочной структуре Сети. Приведем следующие примеры: HITS [4], PageRank [5], тематический индекс цитирования, используемый Яндексом [1].

Эти методы объединены следующей идеей, пришедшей из научного мира и используемой для определения значимости трудов какого-либо ученого. «Авторитетность» (а через нее и релевантность) веб-страницы определяется количеством ссылок на нее с других страниц. Однако для действительно точного определения значимости важно не только количество ссылок на них, но и качество этих ссылок.

PageRank работает на основе только гиперссылок, рекуррентно вычисляя вероятность попадания пользователя на каждую страницу в соответствии с моделью случайного блуждания. HITS и тематический индекс цитирования отличаются от PageRank учетом тематики ссылающихся страниц: ссылка между страницами одной тематики более весома, чем ссылка между страницами разной направленности. Отличие HITS от тематического индекса цитирования заключается в разделении страниц на «концентраторы» (hubs) и «авторитеты» (authorities). Для авторитетов характерно большое число входящих ссылок, для концентраторов — большое число исходящих.

2.2 Кластеризация

Информационная потребность пользователя не всегда может быть точно выражена словами. Кроме того, большинство запросов поисковых систем состоит из одного-трех слов, что порождает неоднозначность и приводит к появлению среди результатов поиска документов на совершенно разные темы.

Кластеризация результатов поиска направлена на то, чтобы разделить множество документов на независимые подмножества и наглядно представить его структуру в виде иерархической структуры [7].

2.3 Классификация

Подобно кластеризации, задачей классификации является представление множества документов в виде иерархической структуры категорий. Отличием является то, что категории известны и заранее зафиксированы.

Классификации результатов поиска обычно проводятся по одному из двух признаков: тема документа либо жанр документа.

Тема документа — это предмет обсуждения (например, автомобили) [2].

Жанр характеризует особенности изложения темы документа. Примеры: много ссылок (коллекция ссылок), технический текст (научная статья), картинки почти без текста (реклама), короткий ответ на конкретный вопрос (форум технической поддержки) и т. д. [3]

К жанровым также относят классификации степени объективности изложения и позитивности/негативности отношения автора.

3. Оценка тематической содержательности

3.1 Постановка задачи

Опыт поиска информации в различных источниках показывает, что документы, найденные по запросу, обычно делятся на три категории. Условно назовем эти категории «статья», «обзор» и «шум».

Статья — содержательный текст, соответствующий запросу; для него характерно подробное изложение темы, наличие определений, описание свойств и т. п. Пример (для запроса «алгоритм»): словарная или энциклопедическая статья об алгоритмах, соответствующая глава из книги по дискретной математике.

Обзор — текст по теме, но с малой подробностью изложения; характеризуется упоминательно-перечислительным использованием слов из поискового запроса. Пример (для того же запроса «алгоритм»): оглавле-

ние книги или план учебного курса по алгоритмам, список статей по computer science, список вопросов к экзамену по программированию.

Шум — нерелевантные документы, в которых слова из поискового запроса встречаются случайно и не по теме. Документы данной категории обычно характеризуются низкой частотой употребления искомых слов, оказываются в хвосте списка найденных документов и поэтому «всплывают» только тогда, когда релевантных документов мало.

Не следует отождествлять «статьи» с «авторитетными источниками», на определение которых нацелены методы ссылочного ранжирования. Авторитетный источник может не содержать развернутого текста о себе точно так же, как содержательная «статья» может являться не авторитетным источником, а, скажем, рефератом в бесплатной онлайн-базе рефератов.

В следующей таблице приведены конкретные примеры статей и обзоров, взятые из английской Википедии. Все перечисленные документы найдены по запросу «algorithm». Для краткости вместо полных URL документов приведены только заголовки. В колонке «частота» указано количество вхождений ключевого слова «algorithm» в текст документа.

Документ	**Частота**	**Категория**
Algorithm	211	статья
Euclidean algorithm	64	статья
Extended Euclidean algorithm	43	статья
Genetic algorithm	89	статья
Parallel algorithm	25	статья
Big O notation	24	обзор
List of algorithms	208	обзор
List of important publications in computer science	73	обзор
List of terms relating to algorithms and data structures	49	обзор
Timeline of algorithms	51	обзор

В большинстве случаев интерес представляют содержательные документы, т.е. «статьи». Как видно из таблицы, различить статьи и обзоры

тривиальным подсчетом количества вхождений поискового запроса в текст документа не удается. Таким образом, необходима разработка более тонких критериев для определения содержательности документа.

3.2 Выбор признаков

Интуитивные соображения и анализ примеров подсказывают несколько характерных отличий «статей» от «обзоров».

1. «Обзоры» характеризуются б*о*льшим тематическим охватом, чем «статьи». Это означает, что используется разнообразная лексика, широкий диапазон релевантных понятий.
2. «Статьи» обладают б*о*льшей подробностью изложения, чем «обзоры». Это значит, что каждое использованное понятие обычно повторяется по несколько раз.
3. Текст «статьи» более равномерно «покрыт» релевантными терминами, чем текст «обзора».
4. Для «статей» более характерно включение релевантных понятий в развернутые предложения; для «обзоров» — в короткие предложения без глаголов.
5. Релевантные понятия в «статьях» чаще являются главными членами предложения; в «обзорах» — второстепенными.

Надежно отличить «статью» от «обзора», используя только один признак, не получится, однако в комбинации эти признаки могут дать приемлемую точность.

3.3 Реализация и результаты тестирования

На данный момент нами реализованы инструменты для вычисления количественных эквивалентов первых трех характеристик, перечисленных в п. 3.2. Составлен тестовый корпус документов, выбранных из английской Википедии по запросу «algorithm», и проведены оценки эффективности автоматической классификации.

Корпус был вручную размечен на «статьи» и «обзоры», снабжен автоматически вычисленными характеристиками 1-3 и использован для обучения и тестирования классификатора на основе метода опорных векторов. Обучение классификатора выполнялось алгоритмом C4.5 и методом опорных векторов в среде RapidMiner (http://rapid-i.com).

Результаты тестирования эффективности классификатора приведены в таблице. По столбцам — истинная категория документа, по строкам — предсказание автоматического классификатора.

	статья	**обзор**	точность
статья	29	12	70.73%
обзор	9	35	79.55%
полнота	76.32%	74.47%	

Средняя точность и полнота классификатора составляют 75%.

Далее планируется исследование поведения классификатора при учете характеристик 4-5 и проверка переносимости классификатора на другие предметные области.

4. Заключение

В данной работе показана необходимость разработки методов оценки тематической содержательности документов для увеличения релевантности выдачи поисковых систем. Предложена реализация автоматического классификатора документов по содержательности на «статьи» и «обзоры» и показана его работоспособность.

Особенностью классификатора является функционирование на основе только текста документа, без привлечения дополнительной информации наподобие сети гиперссылок. Это позволяет применять классификатор не только в Интернете, но и в других репозиториях документов, где гиперссылки отсутствуют.

Литература

[1] Индекс цитирования. http://help.yandex.ru/catalogue/?id=873431

[2] T. Bayer, U. Kressel, H. Mogg-Schneider. Categorizing paper documents. A generic system for domain and language independent text categorization. In *Computer Vision and Image Understanding*, 1998.

[3] A. Finn, N. Kushmerick, B. Smyth. Genre Classification and Domain Transfer for Information Filtering. In *Proceedings of ECIR-02, 24th European Colloquium on Information Retrieval Research*, 2002.

[4] J. M. Kleinberg, Authoritative sources in a hyperlinked environment. In *SODA '98: Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms*, pages 668-677, 1998.

[5] L. Page, S. Brin, R. Motwani, T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. *Stanford Digital Library Technologies Project*, 1998.

[6] D. Zhang, Y. Dong. Semantic, Hierarchical, Online Clustering of Web Search Results. In *Advanced Web Technologies and Applications*, 2004.

Autiomated Evaluation of Document Informativeness

Alexey Vladykin

Effective dealing with constantly increasing volumes of digital information is impossible without tools for information retrieval, filtering, structuring and categorization. This work describes a new method for automated evaluation of document informativeness, which may be used to increase relevance of retrieved documents by means of ranking and classification of search results according to proposed criteria.