УДК 025.4.03

Анализ вопросов автоматизации поиска информации

Н.Ю. Рязанова [1]

[1] МГТУ им. Н.Э. Баумана, Москва, 105005, Россия

Рассмотрены вопросы полнотекстового поиска и анализа текстовой информации для построения современных поисковых систем. Проанализированы методы оценки релевантности найденной информации поисковому запросу на естественном языке.

E-mail: ryaz_nu@mail.ru

Ключевые слова: *поисковая система, полнотекстовый поиск, релевантность запросу.*

Поиск информации начинается с формулирования запроса, отражающего цель поиска. Как известно, в ответ на запрос любая интернет-поисковая система выводит список ссылок с краткими выдержками из найденных документов. Очевидно, что качество поисковой системы с точки зрения пользователя зависит от того, насколько точно полученные выдержки отражают содержание запроса. Оценка степени соответствия полученных документов запросу или содержательной релевантности (адекватности) является, по своей сути, субъективной оценкой. Для автоматизации процесса поиска информации вводится понятие формальной релевантности. Процесс информационного поиска может быть представлен последовательностью шагов, приведенной на рис. 1.

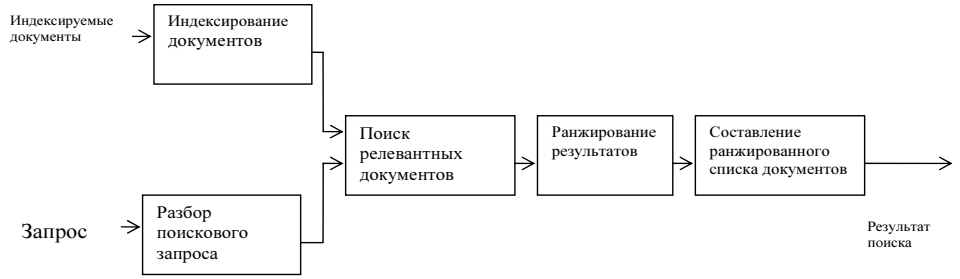

Рис. 1

К переменным параметрам относятся тематика поиска и запрос.

Формализация ранжирования. Вопросы полнотекстового поиска естественным образом связаны с анализом текста. Эмпирические законы, которые отражают характерные особенности любых текстов,

созданных человеком, были сформулированы в 1949 г. лингвистом и филологом Дж.К. Зипфом (George Kingsley Zipf) в результате статистического анализа текстов. Законы Зипфа коротко можно сформулировать следующим образом:

– в каждом языке есть слова, которые встречаются чаще, чем остальные, но не имеют значения;

– есть слова, которые встречаются реже, но имеют намного большее смысловое значение.

Первый закон Зипфа. Очевидно, что слова входят в анализируемые тексты разное число раз. Эта величина называется частотой вхождения. Если сгруппировать слова по частоте вхождения, то получается подмножество слов, встречающихся в тексте примерно одинаковое число раз. Слова, которые встречаются в тексте максимальное число раз, составляют первое подмножество, и далее по убыванию частоты вхождения. Порядковый номер подмножества называется рангом частоты. Вероятность встретить в тексте заданное слово определяется как отношение частоты вхождения к числу слов в тексте:

$$\textit{Вероятность} = \textit{Частота вхождения слова} / \textit{Число слов}$$

Дж.К. Зипфом была обнаружена интересная закономерность: произведение вероятности обнаружения слова в тексте на ранг частоты (*Вероятность* × *Ранг частоты*) является величиной, близкой к постоянной. Следовательно, в соответствии с первым законом Зипфа, если самое часто встречающееся слово находят в тексте, например, 100 раз, то второе по частоте появления в тексте слово будет встречаться приблизительно 50 раз.

Второй закон Зипфа. Частота и количество слов, входящих в текст с этой частотой, связаны между собой. Если построить график, отложив по оси X частоту вхождения слова, а по оси Y — количество слов, встречающихся с данной частотой, то полученная кривая будет сохранять свои параметры для всех без исключения созданных человеком текстов на одном языке.

Более того, законы Зипфа имеют универсальный характер, т. е. справедливы для всех естественных языков (рис. 2) [1]. На каком бы языке текст ни был написан, форма кривой Зипфа останется неизменной.

Алгоритм ранжирования TF-IDF. На законах Зипфа базируется алгоритм ранжирования, который получил название Term Frequency — Inverse Document Frequency (TF-IDF). В алгоритме оценивается частота вхождения слова (TF): как отношение числа вхождений некоторого слова к общему количеству слов документа. Таким образом, оценивается важность слова в пределах отдельного документа:

$$\mathrm{TF} = \frac{n_i}{\sum_k n_k},$$

где n_i — число вхождений слова в документ; $\sum_k n_k$ — общее число слов в документе.

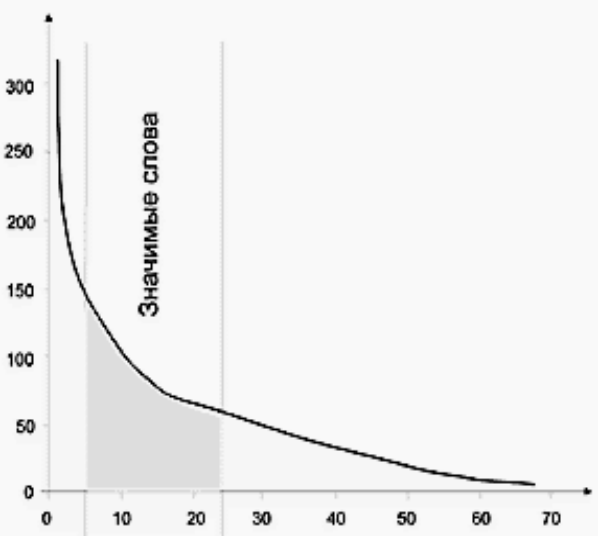

Рис. 2

Учет так называемой обратной частоты документа (*англ*. IDF), с которой слово встречается в документах коллекции, уменьшает вес широкоупотребительных слов:

$$\mathrm{IDF} = \log \frac{|D|}{|(d_i \supset t_i)|},$$

где $|D|$ — число документов в корпусе; $|(d_i \supset t_i)|$ — число документов, в которых встречается t_i (когда $n_i \neq 0$). Выбор основания логарифма в формуле не имеет значения, поскольку изменение основания приводит к изменению веса каждого слова на постоянный множитель, что не влияет на соотношение весов.

Таким образом, мера TF-IDF является произведением двух сомножителей: TF и IDF. Большой вес в соответствии с алгоритмом TF-IDF получают слова с высокой частотой использования в пределах конкретного документа и с низкой частотой употреблений в других документах [2].

В трудах Третьего российского семинара по оценке методов информационного поиска (РОМИП) отмечено, что алгоритм TF-IDF показал лучший результат по качеству поиска, которое оценивается по двум параметрам — полноте поиска и точности поиска. Полнота определяется как отношение числа выбранных в результате поиска

документов к общему числу документов, соответствующих запросу. Точность оценивается как отношение числа выбранных для показа документов, не соответствующих запросу (информационный шум), к общему числу показанных документов. Очевидно, что данные характеристики зависят друг от друга: увеличение точности приводит к уменьшению полноты и наоборот.

Анализ процесса сбора документов. Для практического осуществления поиска необходимо сформировать область поиска, которую создают в виде хранилища данных о документах. Для сокращения времени поиска выполняют предварительную индексацию информации в хранилище. С этой целью оформляют специальные индексные структуры и формализуют процесс отбора информации, ценной для поиска.

Анализ процесса сбора информации позволяет выделить три основные задачи. Первая заключается в выделении малоинформативных частей документов, таких как элементы навигации на странице, рекламные блоки и т. п. Вторая задача состоит в исключении из поискового индекса повторяющейся информации, а именно полностью или частично повторяющейся информации. Для решения обеих задач может быть выбран общий подход, так как в этих задачах необходимо определить повторяющиеся фрагменты текстов. Различаются они только анализируемой областью. Для первой задачи область определяется множеством страниц конкретного сайта, выбранного для анализа. Для второй задачи анализ выполняется для всего индекса системы, сформированного заранее. Третьей задачей является выявление тематики индексируемого документа. Система индексации выполняет такой анализ на основе предварительно созданной базы знаний. Для решения этой задачи в систему включается специальный модуль — классификатор.

Для поиска повторяющихся фрагментов или, другими словами, проверки текстов на уникальность используется метод шинглов (от англ. *shingles* — чешуя).

В алгоритме шинглов тексты разбивают на последовательности слов заданной длины (обычно от 5 до 10). Выделенные последовательности накладываются одна на другую со смещением в одно слово — «внахлест» — и сравниваются на предмет совпадения путем расчета контрольной суммы. Контрольных сумм будет столько, сколько в тексте слов, за вычетом длины шингла. Из всего множества контрольных сумм выбираются, например, только те, которые делятся на 25. Очевидно, что повтор даже одной последовательности слов является признаком дублирования. Число совпадений определяет схожесть двух текстов. Метод шинглов является достаточно надежным для поиска почти-дубликатов, он используется для выявления плагиата. Дубликаты удаляются из множества отобранных для представления документов.

СПИСОК ЛИТЕРАТУРЫ

1. Попов А.И. Поиск в Интернете — внутри и снаружи. URL: http://www.shipbottle.ru/projects/txt/internet_2_1998/index.shtml
2. Федоровский А.Н., Костин М.Ю. Mail.ru на РОМИП-2005 // Труды третьего российского семинара по оценке методов информационного поиска / под ред. И.С. Некрестьянова. СПб.: НИИхимии СПбГУ, 2005. С. 106–124.

Статья поступила в редакцию 25.10.2012