ПОЛНОТЕКСТОВЫЙ ПОИСК В ЭЛЕКТРОННЫХ БИБЛИОТЕКАХ

А.А. Толстобров, В.Г. Хромых
Воронежский государственный университет, Воронеж
394000, г. Воронеж, Университетская площадь, 1
artol@main.vsu.ru

В статье рассмотрен подход к реализации полнотекстовых поисковых систем для библиотек Веб-страниц на основе принципов семантического индексирования. Предложен метод снятия противоречия между вычислительными требованиями и качеством поиска, основанный на "просеивании" результатов поиска согласованными методами.

FULL-TEXT SEARCHING IN ELECTRONIC LIBRARIES

A.A. Tolstobrov, V.G. Khromykh
Voronezh State University, Universitetskaya pl., 1, Voronezh, Russia, 394006
artol@main.vsu.ru

The article addresses an approach to full-text Web libraries search implementation. The proposed method of eliminating contradiction between computational complexity and search quality is based on "screening" of raw search results via the series of coordinated methods.

Введение

Действующие системы внеатрибутного поиска в полнотекстовых документальных базах данных основаны, в основном, на методах "булевского поиска", что не удовлетворяет возросшим требованиям к уменьшению трудоемкости нахождения данных. Выходящая за рамки традиционной теории баз данных, проблема доступа в документальных базах требует перехода от запросов с точными предикатами (парадигмы, поддержанной традиционными поисковыми системами), к концепции запросов по релевантности, выражающей некий "уровень соответствия" каждого из документов намерениям пользователя. Поскольку широкое определение релевантности тесно связано с семантикой текста, на полном уровне недоступной машинному анализу, существующие подходы к реализации массовых систем поиска основаны, как правило, на компромиссных методах расчета "функций ранжирования" простыми статистическими методами.

Структура Веб усугубляет данную задачу за счет увеличения количества доступных документов, отсутствием специальной подготовки поль-

зователей и сложившимся отношением к рассмотрению результатов поиска (характерные запросы, состоящие из 2-4 запросных слов; рассмотрение соответствия в первых 10-15 документах отчета и т.п.).

Принципиальные проблемы коммерческих "булевских" поисковых систем можно разделить на следующие категории.

- Несовершенство идентификации терминов из-за синонимии: запрос должен быть подмножеством (при дизъюнктивном поиске) документа, что ведет к уменьшению полноты результатов. Предполагается, что пользователь должен "угадать", какой из синонимов использован в документе, либо использовать конъюнктивный запрос, перечисляя вручную синонимы термина (автоматизация этого процесса путем использования словаря синонимов ведет к быстрому снижению точности [8]).
- К этой же категории можно отнести проблему грамматического словоизменения, особенно актуального для синтетических языков, таких, как русский, высокая флективность которого затрудняет использование хорошо изученных алгоритмов стемминга словоформ. Однако с разработкой полноценных морфоанализаторов данная проблема к настоящему моменту успешно решена.
- Несовершенство методов работы с омонимией вызывает трудности классификации документов, усугубляемые обычно малой длиной запроса и неспособностью системы автоматически выбрать "правильное" значение омонима ни в запросе, ни в индексе.
- С последним напрямую связана третья категория проблем, относящийся к методике построения индексов, где каждое слово в документе анализируется независимо от других, что ведет к потере значительной доли информации.

Системы латентного индексирования

Алгебраические модели поиска, основанные на векторном представлении текстов документов, и использующие в качестве меры близости векторные метрики, обладают двумя существенными недостатками. Во-первых, работа с матрицей такого объема крайне сложна из-за ее огромного объема. Во-вторых, такой подход плохо поддерживает синонимию — документы считаются семантически далекими друг от друга, если в них не имеется совпадающих слов.

Один из методов, позволяющих преодолеть эти недостатки, является метод LSI (Latent Semantics Indexing; название дано потому, что метод ставит задачей выявление латентных, скрытых факторов, присутствующих в корпусе текстов). Согласно этому методу, пространство термов декомпозицией сингулярным разложением (отбрасываются наименее значимые сингулярные значения) приводится к пространству ортогональных факторов (некоррелируемых «индексных термов»). Поскольку результирующие

факторы играют роль «приведенных» термов, декомпозиция «сближает» документы из одинаковых предметных областей. Данный метод позволяет эффективно решить проблемы синонимии, и в значительной мере - омонимии (полисемии) термов; значительным достоинством метода является то, что искажения, возникающие в результате несовершенности анализа омонимии, при увеличении длины запроса не накапливаются, а взаимно корректируются. После ранжирования результатов на первых местах в массиве результатов оказываются документы, наиболее релевантные запросу, причем предсказуемость релевантности оказывается значительно выше, чем получаемая традиционными методами.

Неполнота имеющихся теоретических обоснований эффективности метода [7] и попытки применения строгого статистического подхода привели к созданию модификации LSI, известной как PLSI (Probabilystic LSI) [5].

Рассмотрим гипотезы "документ d из множества D наилучшим образом соответствует фактору z" и "слово w из множества W наилучшим образом соответствует фактору z". Вероятность того, что первая гипотеза верна, $P(d_j \mid z_i)$; для второй гипотезы — $P(w_j \mid z_i)$. Совместная вероятность осуществления гипотез $P(d,w) = P(d)P(w \mid d)$; $P(w \mid d) = \sum_{z \in Z} P(w \mid z)P(z \mid d)$. В рамках допущений PLCI (в частности, отсутствие корреляции между в d и w), $P(d,w) = \sum_{z \in Z} P(z)P(d \mid z)P(w \mid z)$. Вероятности гипотез определяются максимизацией функции правдоподобия $L = \sum_{d \in D} \sum_{w \in W} n(d,w) \log P(d,w)$ (слово w встречается n раз в документе d).

Принципиальное отличие метода от LSI заключается в выборе целевой функции для определения оптимальной декомпозиции. Помимо строго вероятностного определения результирующих факторов, метод дает возможность строгого выбора размерности пространства латентных факторов.

Проблемы реализации систем на базе PLCI

Несмотря на то, что данный метод показывает ожидаемые результаты [5] и перспективен с точки зрения использования в системах поиска, следующие проблемы препятствуют его массовой реализации.

- Вычислительная сложность [8], быстро растущая с увеличением корпуса документов.
- Повышенная восприимчивость метода к "Документам-ловушкам" — документам компилятивного и рекламного характера, чей текст специальным образом составлен так, чтобы соответствовать популярным запросам [6].

Предложенное решение

Предлагаемый подход устраняет вышеупомянутые противоречия по трем направлениям.

1. Уменьшение корпуса документов, подлежащих ранжированию методами PLCI, путем предварительного отсечения документов, для которых вероятность соответствия данному запросу мала.
2. Уменьшение вычислительной сложности PLCI путем использования более простого метода оценивания-максимизации L (со сходимостью $P(d|z)$ и $P(w|z)$ после нескольких десятков итераций) [7].
3. Использование анализа структуры графа Веб для выявления связей между документами.

Один из простых методов [1] определяет обобщенный индекс цитирования как

$$PR(A) = (1-d)/N + d\sum_{i=1}^{n}\frac{PR(T_i)}{C(T_i)}$$

где $PR(A)$ — ранг документа A, $PR(T_i)$ — ранг документа T_i, ссылающегося на A, $C(T_i)$ — количество исходящих ссылок с документа T_i, d — множитель (0...1).

Зависимость ранга страницы от характера ее связей описывается как "чем больше исходящих ссылок имеется на странице T, тем меньше выигрывает страница A от такой ссылки". В модели "среднего пользователя", который начинает со случайной страницы T и следует по случайной ссылке по ней с вероятностью d. PR страницы — это вероятность попадания на нее пользователя; сумма PR по всей базе равна единице.

Устанавливая динамический порог попадания в результаты поиска в соответствии с распределением рангов страниц, удается удалить большинство "документов-ловушек" [6]. Возрастание точности результатов запроса происходит с увеличением объема корпуса документов.

Практическая реализация

В основе подхода, предложенного для практической реализации системы, лежит идея последовательного многоуровневого просеивания и классификации массива документов по согласованным алгоритмам.

1. На первом уровне (уровне "отсева" информации) производится изъятие "заведомо нерелевантных" документов из массива результатов поиска; для согласования с последующими этапами вводится функция оценки, в которой большей стоимостью обладает ложное отсечение "нужного" документа, чем ложный пропуск ненужного. Данный метод основан на высокопроизводительных технологиях булевского отбора, реализованных с

использованием морфологической нормализации словоформ и полисемантического тезауруса.

2. На втором уровне (уровень "сборки мусора") на основании сравнения ОИЦ с динамически вычисляемым пороговым значением производится предварительная обработка и учет "документов-ловушек".

3. На третьем уровне производится ранжирование документов в соответствии с их релевантностью запросу на основании критериев близости, определяемых согласно модели PLSI.

Оценка качества поиска

Принятый в литературе метод оценки качества поиска по кривой "полнота-точность" [8] использовать для обработки экспериментальных результатов не удалось в связи с характером использованной коллекции русскоязычных документов [6], не обеспечивающей сравнимого числа релевантных документов по различным запросам. В связи с этим использован подход, предложенный в [11]. Среди первых M документов, вошедших в результаты запроса q из множества запросов Q, вычисляем $r(q)$ — общее число документов в коллекции, релевантных q. Интегральная оценка записывается в виде

$$Qual = |Q|^{-1} \sum_{i \in R_q} i^{-1} \sum_{q \in Q} \left(1 + r(q)\right)^{-1} .$$

Данная оценка учитывает число релевантных документов в корпусе наряду с рангом каждого документа из набора релевантных. Экспериментальные результаты сравнения системы с описанной в [6], на коллекции в 850 000 документов, представлены ниже.

Общее число запросов	41
Уменьшение *quality*	6
Увеличение *quality*	35
Среднее качество поиска по методу [6]	0.110
Среднее качество поиска по описанному методу	0.131
Среднее время поиска по методу [6], сек	20,2
Среднее время поиска по методу [6], сек	21,6

Из приведенных результатов, в частности, следует, что замена метода "тонкого" ранжирования LSI на PLCI в реализации поиска дает некоторое улучшение качества поиска без существенных потерь общей вычислительной производительности.

Литература

[1] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, 1998.

[2] Nando de Freitas, Kobus Barnard. Bayesian Latent Semantic Analysis. Technical report, 1999.

[3] Кураленок И.Е., Некрестьянов И.С. Оценка систем текстового поиска. Труды третьей всероссийской научной конференции «Электронные библиотеки», октябрь 2001

[4] Thomas Hofmann. Probabilistic latent semantic indexing. Proc. of the SIGIR'99, 1999.

[5] Hofmann, T. “Probabilistic Latent Semantic Analysis”, Proc. Uncertainty in AI, UAI 1999, Stockholm.

[6] А.А. Толстобров, В.Г Хромых. Многоуровневые системы поиска информации. Труды Всероссийской научно-методической конференции Телематика-2002, июнь 2002.

[7] Hofmann, T. “Probabilistic Latent Semantic Indexing”, Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR’99).

[8] Michael W. Berry, Susan T. Dumais, Todd A. Letsche “Computational Methods for Intelligent Information Access”

[9] Christos H. Papadimitriou, Prabhakar Raghavan, Hisao Tamaki, S. Vempala Latent Semantic Indexing: A Probabilistic Analysis. ACM Sigmod Conference Proceedings, 1998.