Автоматические операции с запросами к машинам поиска интернета на основе тезауруса: подходы и оценки[1]

Павел Браславский
Институт машиноведения УрО РАН
pb@imach.uran.ru

В работе рассматриваются предпосылки использования автоматических операций с запросами к машинам поиска (МП) интернета на основе тезауруса. В рамках предложенного подхода описаны операции перевода запроса, расширения запроса на основе шаблона, построения запроса на основе пути между двумя концепциями, а также ослабления запроса. Приведены примеры запросов, сформированных на основе тезауруса предметной области «Автоматический оптический контроль печатных плат», и соответствующие отклики МП Яндекс и Google. Обсуждаются результаты и даются рекомендации по практическому применению предложенных методов.

Введение

Запросы из ключевых слов – наиболее распространенный способ выражения информационной потребности пользователя при обращении к машинам поиска (МП) интернета. На форуме «Диалога», посвященном интернет-поиску, Илья Сегалович заметил, что запросы уже можно рассматривать как особую разновидность естественного языка.[2] Короткие запросы – характерная черта интернет-поиска. По сообщению Михаила Маслова, средняя длина запроса к МП Яндекс[3] за последнюю неделю февраля 2004 года составила 2,81 слова. Если сравнить этот показатель с данными 1997 и 1999 годов (1,2 и 2,7 слова соответственно)[4], то можно предположить, что рост средней длины запроса замедлился. Короткие запросы вкупе с большими объемами и разнообразием информации в Сети, а также нежелание большинства пользователей идти дальше первой-второй страницы выдачи (т.е. первых 10-20 ссылок) заставляют разработчиков МП уделять большое внимание механизмам ранжирования результатов. Одним из первых интернет-механизмов ранжирования был *DirectHit*, который отслеживал выбор пользователей и учитывал «мнение большинства» при формировании отклика на одинаковые запросы. В начале 2004 года был запущен поисковый сервис Eurekster[5], который ранжирует результаты МП AlltheWeb[6] в соответствии с

[1] Работа выполнена при поддержке РФФИ, грант № 03-07-90342
[2] http://www.dialog-21.ru/forum/actualtopics.aspx?bid=14
[3] http://www.yandex.ru
[4] http://www.yandex.ru/skazki/story_2.html
[5] www.eurekster.com
[6] www.alltheweb.com

предпочтениями участников определенного сообщества («микросоциальный» аналог *DirectHit*). Механизм *PageRank* и аналогичные подходы, основанные на анализе ссылочной структуры Веба, позволили существенно повысить качество откликов МП в ответ на короткие запросы. Еще один подход: МП Rambler[7] при ранжировании результатов учитывает популярность ресурсов в рейтинге Rambler Top100.

Однако, несмотря на совершенствование механизмов ранжирования, актуальной остается «словарная проблема» (*vocabulary problem*). Она состоит в том, что, с одной стороны, слова многозначны, а с другой – одни и те же концепции могут быть выражены различным образом. Поэтому успех интернет-поиска по-прежнему во многом зависит от удачно сформулированного запроса [6]. Вместе с тем, формулировка «хорошего запроса», соответствующего информационной потребности, часто становится непростой задачей для пользователя.

Еще один способ повышения качества (точности/полноты), а также удобства поиска – автоматические и полуавтоматические операции с запросами (модификация, расширение, изменение весов терминов). Наряду с методами, основанными на анализе коллекции документов или той ее части, которая выдается в ответ на первичный запрос (например, обратная связь по релевантности, *relevance feedback*), существуют методы на основе специальных словарей – тезаурусов. Тезаурусы могут быть построены автоматически на основе анализа совместной встречаемости слов, а также вручную. На начальном этапе развития информационного поиска тезаурусы служили для стандартизации словаря ИПС и экономии памяти. Впоследствии основной функцией тезаурусов стало повышение полноты поиска за счет объединения синонимичных и семантически близких терминов по OR. Методы расширения запросов с помощью тезаурусов широко обсуждались в литературе, сообщались противоречивые результаты. Однако построенные вручную тезаурусы обычно дают хорошие результаты в различных приложениях [7, 10, 15].[8]

Сегодня тезаурусы находят ограниченное применение в универсальных полнотекстовых МП интернета. Одна из причин – в том, что чрезвычайно трудно построить тезаурус, который соответствовал бы тематическому разнообразию информации, индексируемой универсальной МП. С другой стороны, полнота не является критическим параметром универсальных МП интернета. И, наконец, возможность использовать тезаурус при поиске в интернете вряд ли будет востребована большинством пользователей. Тем не менее, можно привести несколько примеров. Так, в конце 90-х годов прошлого века МП AltaVista[9] предоставляла сервис *AltaVista Refine*, который позволял устранять неоднозначность терминов запроса с помощью словаря совместной встречаемости слов [14]. В настоящее время Google[10] предлагает поиск по синонимам и грамматическим вариантам для ограниченного набора английских слов. Так, в ответ на запрос *'~cats'* будут найдены документы, содержащие слова *'cat'*, *'dogs'*, *'pets'* и *'kitten'*.

В нашей работе мы рассматриваем автоматические операции с запросами к МП интернета на основе тезауруса в рамках подхода, который обладает следующими существенными особенностями:

---

[7] www.rambler.ru

[8] В работе [15] сообщается, что ощутимое повышение качества поиска по коллекции TREC достигалось при расширении *коротких* запросов. Короткие запросы были сформированы на основе краткого описания информационной потребности, их средняя длина составляла 11 (!) слов.

[9] www.altavista.com

[10] www.google.com

- тезаурус является компонентом автономной метапоисковой машины, т.е. не привязан к конкретной МП;
- тезаурус описывает терминологию узкой предметной области;
- основной элемент тезауруса – концепция (а не отдельный термин);
- концепции тезауруса связаны отношениями, семантика которых может быть различной (набор типов отношений не фиксируется).

Архитектура и функциональность метапоисковой машины описаны в [1, 7]. Структура и формат представления тезауруса описаны в [2]; актуальная версия формата находится по адресу http://imach.uran.ru/pb/thesaurus/thesaurus.xsd.

В этой работе мы рассматриваем четыре автоматических операции на основе тезауруса: перевод запроса, расширение запроса на основе шаблона, построение запроса на основе пути между двумя концепциями, а также ослабление запроса. Приводятся примеры запросов, сформированных на основе тезауруса предметной области «Автоматический оптический контроль печатных плат», а также соответствующие отклики МП Яндекс и Google. Делаются оценки эффективности и формулируются рекомендации по практическому использованию методов.

Пример тезауруса

Совместно с экспертом вручную был построен русско-английский тезаурус предметной области «Автоматический оптический контроль печатных плат». Как заметили авторы обстоятельного обзора 1996 года, «автоматический контроль печатных плат является наиболее зрелым промышленным приложением машинного зрения» [13]. Особенность данной предметной области – в том, что она находится на стыке двух дисциплин: технологии производства печатных плат и машинного зрения (computer vision), соответственно – термины «происходят» из разных терминосистем. Основным методом составления словаря была интроспекция, в качестве вспомогательного материала мы использовали обзорную статью [13], отечественные стандарты [3, 4], словари по машинному зрению и компьютерной графике [9, 12], словарь по технологии изготовления печатных плат [11].

На настоящий момент тезаурус содержит около 200 концепций, 800 английских и русских терминов и 700 однонаправленных связей между концепциями. В тезаурусе используются следующие типы связей: *род – вид*, *часть – целое*, *следует – предшествует*, *используется для – использует*, *носитель свойства – свойство* (асимметричные связи), а также *ассоциация*, *коррелят* (симметричные связи). «Концентраторами» (*hubs*) тезаурусной сети являются концепции *дефект* (26 смежных концепций), *печатная плата* (23 смежные концепции), *автоматический оптический контроль* (17 смежных концепций). На рисунке представлен фрагмент тезаурусной сети (чтобы не перегружать рисунок, для каждой пары отображена только одна связь). Работа над тезаурусом продолжается.

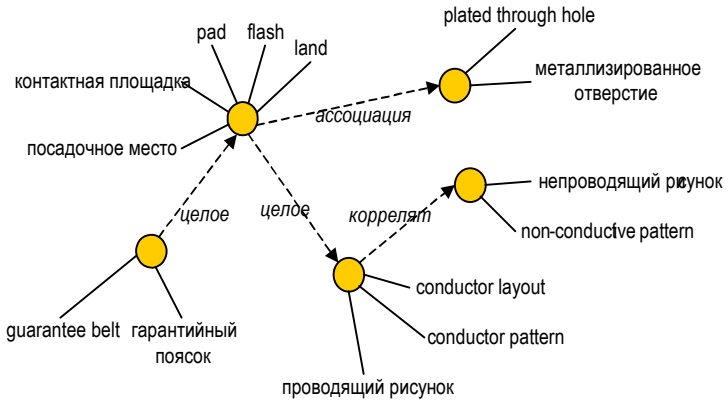

Рисунок. Фрагмент тезаурусной сети

Автоматические операции

Перевод и распределение разноязычных запросов между МП. Это наиболее простая и одновременно эффективная операция. Многие специальные термины представляют собой словосочетания, пословный перевод которых с помощью универсальных словарей дает плохие результаты. Так, если словари Lingvo[11] и Мультитран[12] «знают» перевод термина «печатная плата», то специфичные термины «золотая плата» или «пороговое разделение» отсутствуют в этих словарях. Безусловно, автоматической операции должна предшествовать ручная работа – занесение иноязычного эквивалента в тезаурус. В случае метапоисковой машины операция перевода может сопровождаться разделением разноязычных запросов между МП (например, русский запрос адресуется Яндексу, английский – Google).

Расширение запроса на основе шаблона. Операция состоит в применении шаблона к опорной концепции (опорную концепцию выбирает пользователь). Шаблон определяет используемые поля элемента *termEntry* (кроме основного поля *term*, могут быть использованы *acronym, variant* и *cognate,* подробнее см. [2]), типы связей вместе с соответствующим оператором (AND, OR или ANDNOT), глубину расширения и языковые опции. Термины опорной концепции дополняются терминами соседних концепций тезауруса. Обход в ширину производится по указанным типам связей на заданную глубину. Термины внутри концепций по умолчанию объединяются с помощью OR. Языковые опции определяют, будут ли запросы на разных языках разделяться. Объединение терминов разных концепций с помощью OR нацелено на повышение полноты поиска, в то время как AND и ANDNOT доставляют более строгие запросы. Предложенная нами структура тезауруса обладает большой гибкостью; предполагается, что разработчик настраивает схему тезауруса под предметную область и конкретные задачи. Шаблоны для расширения запросов должны соответствовать этой схеме (учитывать гранулярность описания концепций и набор используемых типов связей).

Путь между двумя заданными концепциями тезауруса. Запрос формируется из терминов концепций, составляющих кратчайший путь между концепциями, указанными пользователем (если такой путь найдется). В [7] описывается аналогичный метод расширения запросов (*correlated search*) и сообщается, что он доставляет хорошие результаты.

---

[11] Мы использовали ABBYY Lingvo 7.0.

[12] http://www.multitran.ru

Ослабление запроса. Как показали наши предварительные эксперименты, запросы, построенные с помощью автоматических операций на основе тезауруса, часто оказываются очень строгими. Поэтому мы предлагаем еще один тип операции – ослабление запроса. Наш подход аналогичен описанному в [10]. Запрос может быть ослаблен с помощью последовательного удаления кавычек (поиск по отдельных словам вместо поиска по фразе), добавления квазисинонимов терминов (однокоренных слов, вариантов, сокращений), замены операторов AND на OR.

Примеры запросов

В таблицах 1 и 2 приведены примеры запросов, полученных в результате применения описанных автоматических операций к концепциям тезауруса предметной области «Автоматический оптический контроль печатных плат», а также соответствующие отклики МП Яндекс и Google. Из-за ограниченности ресурсов мы не могли провести основательную количественную оценку качества поиска. Оценки точности приведены только для коротких списков откликов. В приведенных примерах мы используем неагрессивные методы расширения запросов (шаблоны глубины 1 – добавляются только термины соседних концепций – и расширение на основе пути длины 2). Это связано как с гранулярностью тезауруса (большинство концепций содержит больше двух терминов одного языка), так и с тем обстоятельством, что Google ограничивает длину запроса десятью словами (остальные слова отсекаются). Эксперименты проводились в конце марта 2004 года.

Таблица 1. Запросы к МП Яндекс

Запрос	Ссылки	Комментарий
Простые запросы		
"золотая плата"	5	Релевантных документов нет.
"пороговое разделение"|бинаризация	945	
"пороговое разделение"	11	9 релевантных документов (точность 82%).
бинаризация	927	
Расширение на основе шаблона		
"золотая плата" && ("сравнение с эталоном"|"сравнение с образцом") && (АОК|"автоматический оптический контроль")	0	
полутоновый && монохромный && halftone && "gray-level" && grayscale && monochrome	15	Один документ – строгое соответствие (словарь).
Путь между двумя концепциями		
("пороговое разделение"|бинаризация) && "обработка изображений" && ("автоматический оптический контроль" | "автоматический визуальный контроль")	0	
Ослабление запроса		
(+золотая +плата) && (АОК| (+автоматический +оптический +контроль)|(+автоматический	737	

+визуальный +контроль))		
((+пороговое +разделение)|бинаризация) && (+обработка +изображений) && (+автоматический +оптический +контроль)	23	3 документа – строгое соответствие, из них один релевантный документ (точность 4%).

Таблица 2. Запросы к МП Google

Запрос	Ссылки	Комментарий
Простые запросы		
“golden PCB”	94	17 релевантных документов (точность 18%), остальные посвящены материнским платам с золотым цветом диэлектрического материала.
binarization OR thresholding	~81600	
binarization	~8200	
thresholding	~113000	
Расширение на основе шаблона		
"golden PCB" ("reference comparison" OR "template matching") "automatic optical inspection"	1	Документ релевантный.
halftone "gray-level" grayscale monochrome	208	2 первые позиции – словари.
Путь между двумя концепциями		
"automatic optical inspection" "image processing" (binarization OR thresholding)	15	4 релевантных научные статьи (точность 27%).
Ослабление запроса		
(AOI OR "automatic optical inspection") "image processing" (binarization OR thresholding)	127	Падение точности за счет аббревиатуры *AOI = area of interest*, используемой в обработке изображений.
("golden PCB" OR "golden board") ("reference comparison" OR " template matching") AOI	23	3 – научные статьи, 20 – технические описания с пяти сайтов (точность – 100%).
golden AOI (PCB OR board) (reference comparison) OR (template matching)	37	13 релевантных документов (точность 35%).

Заключение

Эксперименты с автоматически модифицированными запросами демонстрируют значительные вариации качества поиска. Несмотря на присутствие хороших результатов, предложенный подход не обладает устойчивостью по отношению к различным исходным концепциям, методам модификации, МП, языку. Этот результат согласуется с данными исследования [5]. Как показывают эксперименты, эффективность поиска можно повысить с помощью дополнительного анализа откликов (формата и размера файла, URL). Эти возможности мы планируем реализовать в разрабатываемой метапоисковой машине.

При этом не вызывает сомнений, что автоматические методы существенно облегчают формирование сложных специализированных запросов, повышают удобство поиска. Эти обстоятельства приводят нас к предложению рассматривать автоматически сформированные выражения скорее как подсказку пользователю, чем как готовые запросы. Таким образом, автоматические операции с запросами могут поддерживать интерактивность поискового процесса и способствовать обучению пользователей навыкам интернет-поиска.
Ручные процедуры создания тезауруса могут быть узким местом предлагаемого подхода. Поэтому в дальнейшем мы планируем изучить возможность полуавтоматического построения тезауруса предлагаемого вида. Проведенные эксперименты говорят также о полезности введения дополнительного поля в элемент *termEntry* для устранения многозначности термина (с помощью ANDNOT). Значение этого поля может быть получено с помощью статистических методов.

Литература

1. Альшанский Г.А., Браславский П.И., Титов П.В. Формирование информационных запросов к машинам поиска интернета на основе тезауруса: семантико-ориентированный подход // Труды VIII Международной конференции по электронным публикациям "EL-Pub2003". 8 – 10 октября 2003 года, Новосибирск, Академгородок, http://www.ict.nsc.ru/ws/elpub2003/5964/
2. Браславский П.И. Тезаурус для расширения запросов к машинам поиска Интернета: структура и функции // Компьютерная лингвистика и интеллектуальные технологии. Тр. Междунар. конференции Диалог’2003 (Протвино, 11-16 июня 2003 г.). М.: Наука, 2003. С. 95-100.
3. Платы печатные. Основные параметры конструкции: ГОСТ 23751-86. М.: Изд-во стандартов, 1986.
4. Платы печатные. Термины и определения: ГОСТ 20406-75. М.: Изд-во стандартов, 1975.
5. Alemayehu N. Analysis of Performance Variation Using Query Expansion // Journal of the American Society for Information Science and Technology, 2003. 54(5). P. 379–391.
6. Aula A. Query Formulation in Web Information Search. In Isaías, P. & Karmakar, N. (Eds.) Proceedings of IADIS International Conference WWW/Internet 2003. 2003. Vol. I. P. 403-410.
7. Bodner R., Song F. Knowledge-based approaches to query expansion in information retrieval. In McCalla, G. (Ed.), Advances in Artificial Intelligence. New York: Springer, 1996. P. 146-158.
8. Braslavski P., Alshansli G., Shishkin A. ProThes: Thesaurus-based Meta-Search Engine for a Specific Application Domain. In Proc. of the 13th World Wide Web Conference, 2004, May 18-22, New York, NY, USA. In press.
9. Edinburgh Online Graphics Dictionary, http://homepages.inf.ed.ac.uk/rbf/grdict/grdict.htm
10. Gauch S., Smith J.B. An Expert System for Automatic Query Reformulation. In Journal of the American Society of Information Science. 1993. 44 (3). P. 124-136.
11. Glossary of Printed Circuit Board Terms, http://www.pwtpcbs.com/glossary
12. Haralick R. M., Shapiro L. G. Glossary of Computer Vision Terms // Pattern Recognition. 1991. Vol. 24. No. 1. P. 69-93.
13. Moganti M., Ercal F., Dagli C. Automatic PCB Inspection Algorithms: a Survey // Computer Vision and Image Understanding. 1996. 2 (63).
14. Schwarz C. Web Search Engines // Journal of the American Society for Information Science. 1998. 49 (11). P. 973–982.
15. Voorhees E. M. Query Expansion Using Lexical-Semantic Relations. In Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, 1994. P. 61-69.