ФОРМИРОВАНИЕ ЗАПРОСОВ К ПОИСКОВОЙ МАШИНЕ ДЛЯ ИЗВЛЕЧЕНИЯ ЗНАНИЙ ИЗ ИНТЕРНЕТА

COMPOSITION OF QUERIES TO SEARCH ENGINE FOR KNOWLEDGE RETRIEVAL FROM INTERNET

А.Л. Воскресенский

Специальная (коррекционная) общеобразовательная школа-интернат № 101

для глухих и слабослышащих детей, Москва

avosj@yandex.ru

Г.К. Хахалин

khakhalin@got.mmtel.ru

Рассматриваются методы формулирования запросов при поиске новой информации в Интернете, например при создании новых обучающих курсов.

Представлены построенные на результатах реальных экспериментов математические модели зависимостей числа найденных документов и их релевантности от учета при поиске синтаксиса и морфологии слов запроса, а также специальных мер (базирующихся на контексте запроса и не требующих словаря синонимов) включения в запрос синонимичных значений отдельных составляющих запроса.

1. Введение

Вариативность содержания обучения, способствуя созданию оптимальных условий для учащихся с учетом их психофизического состояния и уровня умственного развития, накладывает дополнительную нагрузку для учителя. Пополнение школьных хранилищ данных релевантными документами из Интернета затруднено отсутствием у учителей навыков поиска: обычно вызывают затруднение подбор ключевых слов, формулировка запроса, а полученные в результате выполнения запроса списки ссылок часто приводят учителей в состояние шока. Просмотрев несколько верхних ссылок и убедившись, что полученные тексты не отвечают их представлениям, они уходят от компьютера со словами «сделаем это позже». В таких случаях им остается ждать создания Semantic Web [1] или, пока он не создан, найти такой способ задания запроса, который позволил бы для существующих поисковых машин достичь приемлемого результата.

Ниже описывается эксперимент по определению факторов, влияющих на эффективность поиска документов, содержащих новую информацию по определенной предметной области.

2. Проблемы поиска информации

Современные поисковые системы позволяют по по запросу в форме естественно-языкового выражения находить документы, не содержащие слов запроса. Такие системы называются системами «семантического» поиска. Они отличаются от формально-логических поисковых систем, но базовый поисковый механизм и для тех и для других систем фактически один и тот же — это поиск документов по формально-логическому выражению [2, 3]. Запрос в семантических системах подвергается некоторым преобразованиям. Эти преобразования суть «расширения» запроса. Один из типов расширений сводится к нечеткому поиску, когда запрос расширяется близкими по написанию словами. Второй тип относится к расширению запроса по тезаурусу [4, 5], что позволяет расширить запрос близкими по смыслу словами, используя разные типы смысловых связей. Третий тип — это выделение специальных конструкций имен, дат и пр. Используются и другие преобразования. Однако в конечном итоге расширенный запрос должен быть сведен к запросу на формальном языке поисковой системы, который вряд ли доступен в полном объеме для обычных пользователей.

Одной из проблем поиска является выбор ключевых слов для запроса. Мало того, что пользователь может выбрать для запроса не совсем точный набор ключевых слов, этот набор, вследствие присущей языку полисемии, может не совпадать с ключевыми словами, принятыми для выбранной тематики и предметной области, в результате чего необходимые документы не попадут в выборку. Использование расширенного поиска также не решает в полной мере данную проблему.

Очевидно, что при формировании запроса для получения выборки, содержащей максимум релевантных документов, желательно использовать средства синтаксического анализа с перефразировками. Но сейчас в полном объеме эти средства нет возможности реализовать, т.к. это требует полной перестройки всей структуры современных машин

поиска (поисковых механизмов, индексных массивов, преобразований запросов и т.д.). В отличие от расширения запроса в предлагаемом подходе применяется его «сужение» по определенной методике, позволяющей учитывать в некоторой степени синтаксис запроса. Похожий подход используется в системе ALEX [6] при построении терминологических словарей.

Другой проблемой поиска является ранжирование его результатов. Релевантные с точки зрения пользователя документы далеко не всегда находятся в начале выборки. Это объясняется тем, что в списке результатов поиска учитывается частота вхождений в документ слов запроса, тогда как для пользователя важна смысловая связь с интересующей его предметной областью. Но во многих документах ссылка на предметную область дается зачастую только один раз, как, например, в текстах докладов научных конференций, к которым предъявляются жесткие требования по объему.

Учитывая большие размеры списков ссылок в результатах выполнения запроса (они могут соответствовать миллионам документов) и реальные возможности человека, многие релевантные документы, находящиеся в конце списка, остаются невостребованными.

Целью данной работы является оценка факторов, влияющих на полноту и точность поиска. При этом, как и в работах, связанных с оценкой пользовательских характеристик программных продуктов (см., [7]), авторы используют методы планирования эксперимента для получения математических моделей влияния оцениваемых факторов.

3. Описание эксперимента

3.1. Постановка задачи

Если исходить из принципа композиции Фреге, что значение фразы является функцией значений ее частей и способа комбинирования этих частей, то значение слова (или группы слов) можно определить по формам и расположению окружающих слов, т.е. по контексту. Тогда, если в поисковом запросе опустить некоторое слово или группу слов (исключаемый элемент текста обозначим через X), то в результате выполнения запроса получим документы, в которых на месте пропущенного X будут стоять элементы текста, имеющие тот же самое (или близкое) значение, что и в X.

Другими словами, в результате поиска будут автоматически получены документы одной предметной области, содержащие элементы текста со значениями, которые близки к смыслу исходного документа (на основании которого составлялся запрос), но в какой-то степени и отличающиеся от него. Соответственно, эти документы могут содержать знания, отличающиеся от знаний, содержащихся в исходном документе, т.е. новые знания.

3.2. Планирование и выполнение эксперимента

Рассматривались две функции отклика: число найденных в результате запроса документов (Y_1) и число релевантных документов, содержащихся в первых 50-ти найденных документах (Y_2). Выбор второй функции определялся невозможностью для экспериментаторов реально оценить релевантность всех полученных документов. При этом предполагается, что поисковая машина переносит большинство релевантных документов в начальную часть списка ссылок.

Использовалась поисковая машина Яндекс. Выбор определялся, во-первых, тем, что «контекстные операторы» в формальном языке запросов реализуются только при координатном индексе (как у Яндекса или Рамблера), а во-вторых, формальный язык запросов Яндекса в большей степени обеспечивает корректность задания уровней исследуемых факторов, чем, например, язык запросов Рамблера.

Эксперимент проводился по исследованию влияния трех факторов:

A — наличие пропущенного фрагмента текста;

B — учет порядка слов в запросе;

C — учет морфологических форм слов.

Все три фактора — качественные и принимают значения «-» (фактор отсутствует) или «+» (фактор присутствует). Такому факторному эксперименту соответствует план 2^3 [8], включающий в себя восемь комбинаций вариантов запросов:

(1) — все три фактора минимальны; запрос без пропусков, связь слов по условию OR, слова представлены квазиосновами;

a — максимум фактора A, минимум других факторов; в запросе заданы квазиосновы слов контекста, соединенных условием OR (порядок слов не имеет значения), X исключен из запроса;

b — X участвует в запросе, слова в запросе соединены условием AND (учитывается порядок слов), слова представлены квазиосновами;

ab — X исключен, в остальном соответствует варианту b;

c — X присутствует, связь слов по условию OR, слова представлены словоформами;

ac — X исключен, связь слов по условию OR, слова представлены словоформами;

bc — X присутствует, связь слов по условию AND, слова представлены словоформами;

abc — X исключен, связь слов по условию AND, слова представлены словоформами.

Задание требуемых значений факторов обеспечивается языком запросов Яндекса [9]. Когда связь слов осуществляется по условию AND (т.е. учитывается порядок слов), верхний уровень фактора A задается оператором /(+2 +5), означающим, что слова запроса, между которыми стоит этот оператор, в результатах поиска могут разделять от одного до четырех других слов. Так формируется пустой слот, в который могут попасть слова и выражения, сино-

нимичные исключенному элементу X. Образно этот слот можно назвать «смысловой ловушкой».

Для оценки адекватности моделей каждый из опытов включал в себя две независимые реплики. Первая реплика была представлена запросом *«географические атласы стран Европы»* (X = *«атласы»*), вторая — *«поиск новых знаний в интернете»* (X = *«знаний»)*. Тематически эти две фразы не связаны друг с другом.

На рис. 1 и 2 представлены образцы результатов выполнения запросов (реплики 3-2 и 5-2, соответственно). Рандомизация результатов достигалась путем случайного порядка задания запросов (см. табл. 1). Результаты показаны в табл. 2.

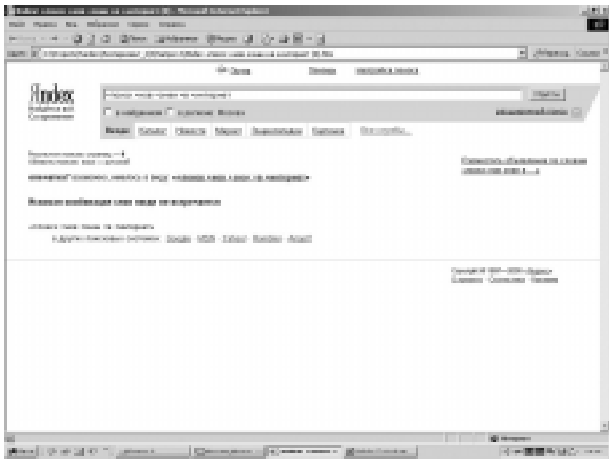

Рис. 1. Результаты выполнения запроса (реплика 3-2, минимальное значение функции отклика)

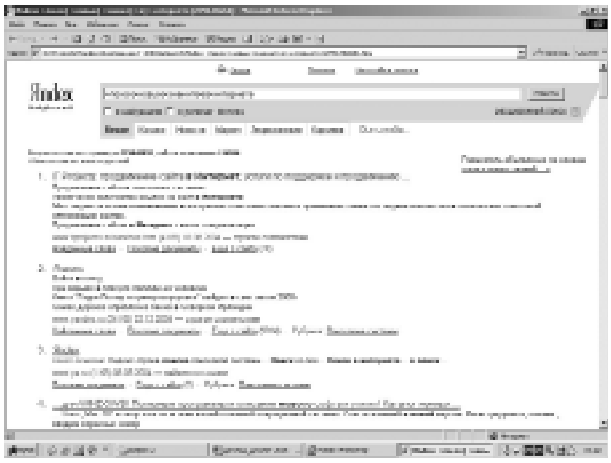

Рис. 2. Результаты выполнения запроса (реплика 5-2, максимальное значение функции отклика)

Номер п/п	Номер реплики опыта	Текст запроса
1	7-1	+географические +атласы +стран +Европы
2	6-2	+поиск|+новых|+в|+интернете
3	4-1	+географич /(+2 +5)стран +Европ
4	2-1	+географич|+стран|+Европ
5	3-2	+поиск +нов +знан +в +интернет
6	3-1	+географич +атлас +стран +Европ
7	1-1	+ географич|+атлас|+стран|+Европ
8	5-1	+географические|+атласы|+стран|+Европы
9	2-2	+поиск|+нов|+в|+интернет
10	4-2	+поиск +нов /(+2 +5)в +интернет
11	5-2	+поиск|+новых|+знаний|+в|+интернете
12	8-2	+поиск +новых +в +интернете
13	8-1	+географические +стран +Европы
14	7-2	+поиск +новых +знаний +в +интернете
15	6-1	+географические|+стран|+Европы
16	1-2	+поиск|+нов|+знан|+в|+интернет

Табл. 1. Порядок задания запросов к поисковой машине

Как видно из табл. 2, результаты весьма изменчивы (особенно Y_1), наблюдается расслоение результатов между репликами. Причиной этого может быть влияние неучтенных факторов, например, числа слов запроса. Для элиминирования этого расслоения была проведена нормализация функции отклика по каждой реплике:

$$Y_{ji} = (Y_{ji} - Y_{imin})/(Y_{imax} - Y_{imin}),$$

где j — номер опыта; i — номер реплики; Yjimin — минимальное значение функции отклика для i-ой реплики; Y_{imax} — максимальное значение функции отклика для i-ой реплики. Нормализованные значения функций отклика значительно более униформны, особенно для Y_1. По результатам эксперимента построены математические модели:

$$Y_{1t}=0{,}433-0{,}129A-0{,}865B-0{,}120C+0{,}129AB-$$
$$-0{,}126AC+0{,}120BC+0{,}126ABC$$

$$Y_{2t}=0{,}517+0{,}157A-0{,}039B-0{,}0004C+0{,}207AB-$$
$$-0{,}264AC+0{,}122BC-0{,}246ABC$$

Опыт	Исключен элемент X ?	Учет порядка слов	Слово--формы?	Реплика	Y_1	Y_1	Y_2	Y_2
(1)	-	-	-	1	48607206	0,975772	23	0,437500
				2	279573078	0,999740	15	0,789474
a	+	-	-	1	47980297	0,963187	21	0,375000
				2	279567901	0,999722	15	0,789474
b	-	+	-	1	18	0,000000	9	0,000000
				2	0	0,000000	0	0,000000
ab	+	+	-	1	1098	0,000022	33	0,750000
				2	1188	0,000004	19	1,000000
c	-	-	+	1	49814093	1,000000	13	0,125000
				2	279645655	1,000000	17	0,894737
ac	+	-	+	1	48880091	0,981250	12	0,093750
				2	14676	0,000052	15	0,789474
bc	-	+	+	1	147	0,000003	41	1,000000
				2	17	0,000000	5	0,263158
abc	+	+	+	1	17907	0,000359	30	0,656250
				2	254	0,000001	6	0,315789

Табл. 2. Результаты выполнения эксперимента

Проверка коэффициентов моделей на значимость показала, что в первой модели значим только фактор B ($F = 7{,}87 > F_{a=0,05;1;8} = 5{,}32$), во второй модели все факторы по критерию Фишера не значимы. Это может быть объяснено тем, что распределение ссылок на релевантные страницы в результатах поиска не подчиняется нормальному распределению. В [9] рекомендуется для описания научной деятельности и ее результатов использовать распределение Ципфа. Во второй модели считаем незначимыми факторы B и C, т.к. их коэффициенты существенно меньше коэффициентов фактора A и взаимодействий между факторами. Исключив незначимые факторы, получим:

$$Y_{1t}=0{,}433-0{,}865B$$

$$Y_{2t}=0{,}517+0{,}157A+0{,}207AB-0{,}264AC+0{,}122BC-0{,}246ABC$$

3.3. Обсуждение результатов эксперимента

- Учет синтаксиса (порядка слов) ведет к уменьшению общего числа найденных страниц. Поскольку это не уменьшает числа релевантных ссылок на первых страницах результата поиска, можно сделать вывод, что учет синтаксиса снижает шум поиска.
- Создание «смысловых ловушек» в запросе увеличивает число релевантных результатов поиска. Морфология и синтаксис напрямую не влияют на релевантность, но при взаимодействии со «смысловыми ловушками» учет синтаксиса приводит также к увеличению релевантности, но требование включения в результат тех же словоформ, как и в запросе, приводит к ее уменьшению.
- Высокие значения нулевого коэффициента указывают на наличие неучтенных факторов или взаимодействий факторов. Необходимо построение модели, точнее описывающей исследуемые зависимости.

4. Выводы

Можно считать, что описываемый пробный эксперимент оказался удачным. Он показал, что:

1) на документах Интернета можно проводить эксперименты, результаты которых повторяемы и поддаются статистической оценке, при этом общие закономерности определяются, в основном, грамматическими особенностями языка;
2) поиск новых текстов, включающих неизвестные пользователю выражения, возможен, при этом использование контекста позволяет на стадии поиска обойтись без использования словаря синонимов;
3) при разработке методики построения «смысловых ловушек» можно получать новые знания из Интернета и для этого достаточно использовать способности обычного пользователя, владеющего общими навыками манипулирования естественно-языковыми запросами (выделение квазиоснов, элемента X и т.п.);
4) разработка методики требует дополнительных экспериментов с более представительным реестром запросов и с другими поисковиками.

Список литературы:

1) Тарасов В.А. SEMSEARCH: Интерактивная поддержка поиска в интернете средствами машинного обучения. // Труды Международной конференции КИИ'2002. М.: Физматлит, 2002. С. 756-760.

2) Харин Н.П. Некоторые особенности семантического поиска текстовой информации // Новости Искусственного Интеллекта. № 2, 2002. С. 22-25.
3) Ермаков А.Е. Полнотекстовый поиск: проблемы и их решение. // Мир ПК, № 5, 2001.
4) Лукашевич Н.В., Добров Б.В. Двуязычный информационный поиск на основе автоматического концептуального индексирования. // Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции Диалог'2003. М.: Наука, 2003. С. 425-432.
5) Браславский П.И. Автоматические операции с запросами к машинам поиска интернета на основе тезауруса: подходы и оценки. // Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции Диалог'2004. М.: Наука, 2004. С. 79-84.
6) Жигалов В.А., Жигалов Д.В., Жуков А.А., Кононенко И.С., Соколова Е.Г., Толдова С.Ю. Система ALEX как средство для многоцелевой автоматизированной обработки текстов. // Компьютерная лингвистика и интеллектуальные технологии: Труды Международного семинара Диалог'2002. Т.2: Прикладные проблемы. М.: Наука, 2002. С. 192-208.
7) Arapov, M., Voskresensky, A., Semenova, V. How to compare and evaluate OCR systems? (Our approach). // Proceedings of ELSNET GO EAST and IMACS Workshop on Integration of Language and Speech. 1995, Moscow, Russia, pp. 5 – 10.
8) Монтгомери Д.К. Планирование эксперимента и анализ данных. // Л.: Судостроение, 1980.
9) Детальное описание языка запросов. — http://www.yandex.ru/ya_detail.html
10) Хайтун С.Д. Наукометрия. Состояние и перспективы. // М.: Наука, 1983.