

Поисковое взаимодействие: мифы, реальность и базовые принципы

© Николай Бузикашвили

Институт системного анализа РАН, Cognitive Technologies Ltd.

buzik@cs.isa.ru | cognitive.ru

Аннотация

В противоположность общепринятому мнению, данные о взаимодействии пользователя с онлайн-поисковыми системами свидетельствуют, что поисковое поведение пользователя одинаково в “традиционных” ИПС и в Веб. Пользователь, ищущий без посредника, реализует одни и те же тактики многошагового поиска, следующие из принципа минимума усилий. В условиях неопределенности поисковой среды, тактика минимальной модификации запроса приводит к результату за меньшее время. Поиск, выполняемый посредником, обычно следует другому принципу — гарантированного результата, и является менее эффективным. Помимо поведения пользователя, с тех же позиций, рассмотрено поведение поисковой среды, и даны оценки эффективности разных методов поиска.

1 Введение

Общим местом сегодня стал миф о примитивном пользователе Веб, в значительной мере сформированный работой [9] и анализом первых двух этапов проекта *Excite* [11,22,26]. Однако более широкая коллекция данных, в том числе использованных в [9], напротив, позволяет говорить об эффективном пользователе Веб, поведение которого не отличается от поискового поведения (ПП) пользователей других поисковых сред.

Согласно все более популярной точке зрения (напр., [4]), поведенческий принцип, реализуемый при поиске, — это принцип минимума усилий. Все наблюдения за поведением пользователя, *ищущего “для себя”*, подтверждают этот тезис.

В то же время, в классической схеме поиска с посредником, в роли которого традиционно выступает библиотечный работник, ПП посредника не подчиняется принципу минимума усилий. Когда неопределенность информационной потребности для посредника высока, он следует не этому прин-

ципу, а тому, что можно было бы назвать принципом гарантированного результата, состоящему в такой формулировке запросов, при которой результаты поиска содержали бы документы, удовлетворяющие как можно большему числу возможных интерпретаций информационной потребности патрона (заказчика). Однако классическая итеративная схема патрон-посредник, в которой посредник хорошо ориентируется в поисковой инструментари, но мало знаком с предметной областью, сегодня встречается гораздо реже неинтеративной схемы, в которой посредник имеет дело с понятной информационной потребностью и строит поиск в точности, как для себя.

ПП, следующее принципу минимума усилий, как правило, обеспечивает минимальный риск на каждом шаге и обычно минимизирует общее время достижения результата. Напротив, ПП, подчиняющееся принципу гарантированного результата, в среднем неоптимально и удлиняет поиск, когда истинная интерпретация информационной потребности является массовой, однако оно эффективно для редких интерпретаций.

ПП существует только во взаимодействии с конкретной поисковой средой. Интерфейсы/методы, используемые в поисковой машине (ПМ), мало пригодны для обработки многозначных запросов. Но нередко запрос может отвечать несхожим информационным потребностям. Эффективная схема поиска — это эффективная схема устранения разнообразия, прежде всего тематического. При вводе запроса разнообразие интерпретаций должно быть максимально сокращено. Но должно ли оно быть максимально представлено при отображении результатов, чтобы пользователь смог сократить его на следующем шаге поиска?

Коллизия неоднозначности интерпретаций запроса, с которой мы сталкиваемся на уровне взаимодействия пользователя и ПМ почти буквально воспроизводит коллизию неоднозначности интерпретаций информационной потребности из классической схемы поиска с посредником. И точно также, тактика ПМ, состоящая исключительно в покрытии возможно большего числа допустимых запросом интерпретаций, подобно тактике посредника, основанной на принципе гарантированного результата, в большинстве случаев приводит к удлинению сес-

сии, будучи, однако, эффективной в случае редких интерпретаций.

2 Поисковое поведение в разных интерактивных поисковых средах

Сегодня накоплен обширный эмпирический материал о поисковом поведении (ПП) пользователей интерактивных поисковых систем (ПС). Все данные распадаются на две несопоставимые по объему и полноте группы. Одни получены в результате внешнего наблюдения — протоколирования всех запросов к одной ПС. Другие — результат всестороннего наблюдения малой группы пользователей.

В первом случае отслеживается протокол (transaction log) поискового сервера. Наблюдение объективно, представительно, потенциально неограниченно, но поверхностно (так, неизвестно, удалось ли найти искомое). Во втором случае наблюдение производится со стороны пользователя, а наблюдаемые ведут самоотчеты и подвергаются анкетированию. Пользователь всесторонне вовлечен в исследование, и это искажает его поведение. Кроме того, и предмет поиска, и сами пользователи никак не представительны: группу из одного–трех десятков человек составляют либо коллеги и студенты авторов исследования, либо, как правило, студенты-медики, занятые подбором библиографии.

2.1 Миф о пользователе Веб

В [9] дано сравнение ПП в трех онлайн-средах: Веб, “традиционные” ПС (ТПС; заметим, в категорию ТПС попали как БД с исключительно булевыми запросами, так и ПС, поддерживающие еще и свободные запросы) и библиотечные каталоги (ОРАС — Online Public Access Catalogues). Эта работа вместе с наблюдениями пилотного [11], первого [22] и второго [26] этапов проекта *Excite* сформировали общепринятый сегодня миф об особенностях ПП в Веб. Клишируемый образ результатов проекта *Excite* состоит в том, что пользователь Веб ленив и нелюбопытен — делает короткие запросы (60% запросов состоит из 1-2 слов вместо “положенных в классических ИПС” 7–9), а не найдя подходящего результата на одном-двух начальных экранах выдачи, не желает углубляться в оставшиеся сотни. Он невежествен — редко и с ошибками применяет логические связки. Что неудивительно — ведь его никто не учит.

Таблица 1. Сравнение ПП в Веб, ТПС, ОРАС [9]

	Веб (1997/99)	ТПС (1993)	ОРАС (1993)
Число запросов в сессии	1–2	7–16	2–5
Число термов в запросе	2	6–9	1–2
Доля булевых запросов	8%	37%	1%

Подразумевается, что всякое различие — не в пользу пользователя Веб и, как следствие, “короткий запрос — плохой запрос”. Но, не зная о результатив-

ности поиска, невозможно понять, что скрывается за числом запросов или долей модификаций. Длинный запрос сам по себе не лучше короткого, а большее число модификаций едва ли не говорит об эффективности системы пользователь–ППМ.

“Традиционные ПС, ОРАС и ПМ Веб различаются по используемым в них интерсхемам поиска и базам проиндексированных документов. Следует ли из этих различий различие в поведении пользователей? Ответ на этот вопрос дает сравнение характеристик поискового поведения во всех трех типах систем.” [9]. И хотя, в силу непредставительности данных о ПП в ТПС, ответ оказался ненадежным, не менее важно, что причины этих не слишком достоверных различий ПП остались так и не выяснены. Именно, по-прежнему неизвестно, какие факторы или их сочетания и в какой степени определяют ПП. И если в вопросе были названы *факторы поисковой среды*, то удовлетворившее всех и абсолютно никак не обоснованное объяснение гипотетических различий сегодня сводится к единственному и не входящему в эту группу фактору — неким особенностям самих пользователей: “пользователи Веб — это совершенно особый вид пользователей” [1]. Портреты пользователей ТПС и Веб скорее напоминают карикатуру. Так, портрет пользователя ТПС — взрослый, образованный, прошедший специальную поисковую подготовку. Портрет пользователя Веб — лицо неопределенного социального и образовательного статуса, в том числе ученики начальной школы, подготовки не прошедшие и ищущие игрушки и порносайты. При этом подразумевается, что первый портрет — портрет более умного и более эффективного пользователя.

2.2 Характеристики поискового поведения

Для описания ПП используются две группы характеристик: *количественные* и *качественные*. К первой группе относятся такие характеристики как длина сессии, длина запроса, доля булевых запросов и т.д., а также временные характеристики (время, затраченное на формулировку запроса, на просмотр результата и т.п.).

Вторая группа включает набор тактик модификации запроса [2,3,6], описывающих последовательный поиск не как последовательность *запросов*, а как последовательность *модификации* запроса, заданного на предыдущем шаге поиска. Помимо того, что это язык описания многошагового поиска, он также протоязык для формальных моделей, позволяющий обсуждать ПП в терминах формализма конечных состояний, сложности запроса и сложности его изменения.

В [2,3] перечислены более 20 тактик, распадающиеся на тактики задания начального запроса и тактики модификации очередного запроса. Тактики исходного запроса: мягкая — запрос из одного термина, жесткая — запрос из нескольких термов. Тактики модификации: 1) пересечение — запрос пополняется оператором *И* и новым термом; 2) варьирование — замена одного или более термов запроса, общее

число термов остается неизменным; 3) сокращение — из запроса выбрасывается один или более термов; 4) введение синонимов — в запрос вводятся синонимы.

Понятно, именно качественные характеристики и определяют тип поискового поведения. основной признак *поведения* — его *тактика*. Если у поведения в двух средах при совпадении всех численные признаков будут различны тактики, поведение различно. Напротив, различие в числовых показателях при совпадении тактик никак не позволяет говорить о различии поведения.

2.3 Наблюдения

Обратимся к данным о более сопоставимых по интерфейсу (и, согласно табл. 1, более разнящихся по ПП) средах — ТПС и Веб. Каталоги ОПАС отличаются от них принципиальным образом¹.

2.3.1 Анализ протоколов запросов в Веб и ТПС

Проект *Excite* [9,10,11,22,23,26] — не самый масштабный (в [21] изучено 153 миллионов запросов к ПМ *Alta Vista* в августе–сентябре 1998 г.), но наиболее продуманный и, что важно, лонгитюдный.

Проект выполнен в 1997–2001 гг. Он состоял в анализе протокола запросов, пришедших в некоторые промежутки времени на сервер *Excite*. Протоколирование производилось четырежды — в 1997 (“0-й”/пилотный и первый замеры), в 1999 и 2001 гг.

Исследование CSTR. В [12] проанализированы протоколы запросов к коллекции технических отчетов по вычислительным наукам *CSTR (Computer Science Technical Reports)*, входящей в состав Электронной Библиотеки Новой Зеландии. База *CSTR* полнотекстовая, содержит 50 тыс. документов, никак не рубрицированных и не аннотированных, т.е., за исключением специализации и отчасти объема, аналогична базе поисковой машины Веб.

Пользователи *CSTR* — специалисты по CS, т.е. именно те, кого принято считать наиболее грамотными “поисковиками”. Тем самым, согласно мифу о пользователе Веб, следовало бы ожидать, что ПП пользователей *CSTR* должно радикально отличаться от ПП неразвитых пользователей Веб.

Использовались два интерфейса — булевых и свободных запросов. Сначала (46 недель) интерфейсом по умолчанию был булев (25 тыс. запросов), затем — свободный (15 недель, 8 тыс. запросов). В обоих случаях 66% запросов было сделано в интерфейсе, заданном по умолчанию. В обоих случаях доля булевых запросов составила порядка 20%.²

¹ Правда, проведенное в 2001 г. исследование 127 студентов не обнаружило никаких различий в их поведении в Веб и в ОПАС при поиске по таким широким темам как образование и развлечения [15]. С учетом большего разнообразия спецификаций, предоставляемых ОПАС, и того, что сравнивались не только тактики, но и в числовые характеристики, это неожиданный результат.

² 68% запросов в булевом интерфейсе, принятом по умолчанию, и 44% запросов в булевом интерфейсе, выбранном

Как видно из Табл. 2, за исключением доли булевых запросов, характеристики ПП наиболее продвинутых пользователей ТПС *идентичны* характеристикам ПП разношерстных пользователей Веб.

Таблица 2. Данные протоколов *Excite* [11,22,23] и *CSTR* (ЭБ Новой Зеландии) [12]

	<i>Excite</i>				<i>CSTR</i>
	1997 этап 0	1997 этап 1	1999 этап 2	2001 этап 3	1996/7
Тематика	Любая				CS
Число запросов (тыс.)	51	1000	1000	1000	32
Число сессий (тыс.)	18	211	326	262	16
Ср. число запросов за сессию	2,8	2,5	1,9	2,3	2,5
Доля модифицированных запросов	22%	52,0%	39,6%	44,6%	40%
Среднее число термов запроса	2,3	2,4	2,4	2,6	2,4
Доля булевых запросов	9%	5%	5%	10%	20%
Ср. число пролистных страниц результата	2,3	1,7	1,5	1,7	1–2

2.3.1 Полные наблюдения в ТПС

Продолжительность сессии. Согласно [9], сессия в среднем включает 7 запросов. Однако по [24], она составляет 3, по [19] — 2–3 запроса.

Число термов в запросе С одной стороны, принято считать, что число термов в “классическом” запросе велико: 7,9 у пользователя-новичка и 14,4 у опытного [5], либо, в случае знакомой темы, 8,7 у пользователя-новичка, 7,2 у опытного, а в случае незнакомой — 9,7 у новичка 9,7 и 9,0 у опытного [7]. Эти данные получены на небольших группах пользователей библиографических систем. Заметим, помимо огромного разброса, данные [5] и [7] дают противоположную зависимость длин запросов от квалификации пользователя (что, при видимой противоречивости, скорее отражает немонотонность изменения этой характеристики с ростом квалификации пользователя).

Однако с другой стороны, согласно [19], запрос включает 1–3 термина. По [24], запросы без синонимов содержат 3 термина, запросы с синонимами — 5,5 термов.

Один и тот же студент, ища в Веб ишемическую болезнь сердца, делает запрос из 3 термов (а ища порносайт — из одного). И при поиске ишемической болезни сердца в обширной базе он не столкнется с необходимостью удлинить запрос вдвое, включая в него синонимы. Необходимость расширить запрос синонимами возникает либо при стремлении к полноте (редкий случай, характерный для специальных задач поиска, например, для сбора

пользователем, не содержали логических связей. В результате, 79% всех запросов при умалчиваемом булевом и 82% при умалчиваемом свободном были небулевы.

библиографии), либо при недостаточном числе релевантных документов, что, в свою очередь, может оказаться заметным только при массовом несоответствии запросов и базы (как если бы пользователями ИНИОН поменялись местами с пользователями расположенной напротив Центральной Медицинской Библиотеки), либо малом объеме базы (что сегодня неактуально).

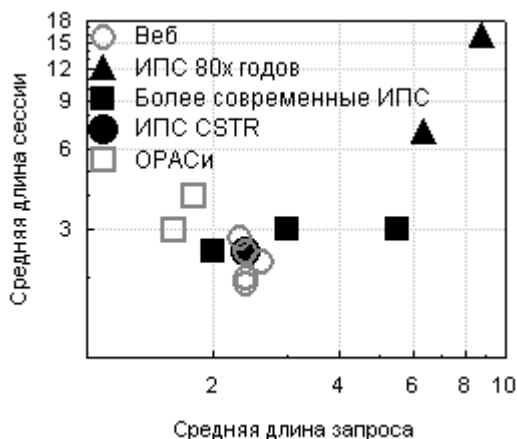
Доля булевых запросов. Приведенные в [9] (табл.1) 37% булевых запросов в ТПС против 5-10% в ПМ Веб впечатляют, пока неизвестно, что это данные единственного исследования [20], в котором оставшиеся 63% запросов состояли из одного термина. Т.е., с одной стороны, небулево задание запросов вообще было невозможно. С другой стороны, доля запросов из одного термина оказалась в два с лишним раза выше, чем в Веб (!), что, по логике мифа, едва ли свидетельствует о продвинутой пользовательской традиции ПС.

2.4 Демифологизация пользователя Веб

Пользователей этих ПС никто не учит [13], запрос в них содержит всего лишь 1-2 термина [9,25], длина сессии 2-3 запроса [19], булевых запросов мало, а тактики просты [25]. Именно так звучит миф о ПП в Веб. Однако приведенные здесь утверждения относятся к поведению вовсе не в Веб, а в ОРАС и ТПС.

Как было показано выше, более полные и более свежие исследования ПП пользователей ТПС не обнаруживают никаких заметных различий в числовых характеристиках с ПП пользователей Веб.

Рисунок 1. Числовые характеристики ПП в Веб, ИПС и ОРАС



Выводы о специфичности ПП в Веб основаны на сравнении ПП в ТПС и ПП в Веб, прежде всего на первых двух этапах проекта *Excite*. Поистине гипнотическое действие оказали результаты второго этапа. Под их впечатлением авторы проекта усмотрели “драматическую тенденцию”, свидетельствующую об “еще большем упрощении ПП — запросы и сессии укорачиваются, а модификации запроса все более сводятся либо к добавлению в предшествующей

запросу нового термина, либо к отбрасыванию термина из предыдущего запроса” [26].

Сведения о ПП в ТПС, на которых зиждется миф о пользователе Веб [9], — результат трех полных наблюдений малых групп. Большинство позиций, по которым сравнивалось поведение в Веб и ТПС, основаны на данных лишь одного из трех наблюдений (в остальных соответствующие данные отсутствуют). Но как мы видели, другие исследования говорят о гораздо меньшем различии количественных признаков ПП в ТПС и в Веб.

Статистика модификации запросов в проекте *Excite* [22]: 32% модификаций состояло в замене одного или более терминов, 42% — в пополнении запроса дополнительными терминами и 26% — в удалении из запроса одного или более терминов.

В [24] описано поведение студентов в новой для них ТПС *Lisa*. Если в первой сессии наблюдалось разнообразие тактик, то в последней оно свелось к заданию одного термина в качестве исходного запроса (в первой сессии не менее часто начальный запрос не менее часто содержал несколько терминов), а из тактик модификации запроса остались *И-* и *ИЛИ-*расширение запроса одним термином, замена или отбрасывание одного термина, т.е. *те же тактики*, что у “плохого” пользователя Веб [26].

Еще в 1991 г. Уайлдемут [25], наблюдая ПП студентов-медиков в ТПС, показала, что *наиболее распространенное ПП состоит в использовании простейших тактик. Точно та же картина наблюдается во всех без исключения исследованиях ПП, как в Веб, так и в ТПС.*

3. Факторы, определяющие различие ПП

Сегодня любой пользователь специализированной или локальной ПС заведомо является и пользователем Веб. Доля среди пользователей ТПС тех, кто сначала познакомился с Веб, возрастает. Как мы видели, пользователи ТПС ведут себя так же как пользователи Веб. И можно полагать, что пользователи ТПС демонстрируют одно и то же поведение в ТПС и Веб, причем их ПП в Веб не отличается от ПП других пользователей Веб.

Хотя пользователь Веб и оправдан, вопрос о важности факторов, определивших различие количественных характеристик ПП, остался открытым. Попытаемся внести ясность. И начнем с факторов, *внешних* по отношению к поисковой среде:

- *пользователь*: конкретный, специализированный, усредненный. Здесь “специализированный” означает некоторую, например, профессиональную общность;
- *тема/предметная область поиска*: конкретная тема, специализированная, усредненная. Здесь “специализированная” означает некоторую предметную область (медицина, computer science, отдых) и “усредненная” — всю совокупность тем.

Чтобы, сравнивая поведение в разных поисковых средах, избежать дополнительных допущений, хо-

телось бы располагать данными, соответствующими комбинациям одних и тех же значений названных факторов, то есть того как пользователи одного типа ищут объекты одного и того же типа. Если ПП студентов-медиков, ищущих некую медицинскую тему в ТПС, отличается от поведения студентов-медиков, ищущих ту же тему в Веб, то действительно, поведение (студентов-медиков) в ТПС отличается от поведения в Веб. Однако если ПП студентов-медиков в ТПС отличается от ПП домохозяйки в Веб, это не дает основания говорить о различии ПП в ТПС и в Веб.

Однако комбинации внешних по отношению к поисковой среде факторов, для которых выполнялись наблюдения в Веб, *не пересекаются* с комбинациями, относящимися к исследованиям ПП в ТПС (Табл. 3). Отсутствуют как одинаковые, так и сходные комбинации. Единственным частичным исключением служит исследование ПП в *CSTR*, где специализированные пользователи искали специализированные темы и делали это точно так же, как усредненные пользователи Веб ищут усредненные темы.

Таблица 3. Комбинации внешних к поисковой среде факторов, представленные в исследованиях ПП в Веб и ТПС*

Пользователь	Тема	Конкретная	Специализированная	Усредненная
Конкретный		●	●	* ([1])
Специализированный		●	● ([12])	
Усредненный				*

* Обозначения: ● — исследования ПП в ТПС, * — исследования ПП в Веб. Типичные комбинации факторов приведены без указания ссылок.

Традиционно ([17,18,9,1]) в качестве факторов, определяющих ПП, называются факторы поисковой среды:

- *время реакции* системы;
- *интерфейс/методы поиска*;
- *объем базы*. Сюда же включим *число/долю* релевантных документов в базе.

Какие же факторы и в какой степени важны? (1) Время реакции было существенно в ТПС 60х-70х годов прошлого века. (2) Различия в булевом и свободном интерфейсе, как вытекает из данных исследования *CSTR*, не имеют значения. С одной стороны, это достаточно неожиданный результат. С другой стороны, он скорее означает не то, что интерфейсные различия несущественны вообще, а то, что различия между двумя этими интерфейсами невелики. Оба они являются оконными интерфейсами недиалогового задания запроса. Запрос уточняется только по результатам очередного шага поиска. Если бы в одной из поисковых сред запрос диалогово уточнялся без обращения к поиску, вероятно, различия в характеристиках ПП были бы заметнее. (3) Число релевантных документов является

пороговым фактором — оно не обязано быть большим, но должно быть *достаточным*. И только когда это число недостаточно, пользователи вынуждены расширять запрос синонимами. Но такая ситуация значима только для старых, недостаточно больших баз.

Таким образом, учитывая, что экстравагантные интерфейсы (методы поддержки формулировки запроса) не имеют массового применения, факторами поисковой среды, влияющими на числовые характеристики ПП, сегодня могут быть только размер базы и методы поиска и представления его результатов.

4. Ленивый пользователь и ответственный посредник

Согласно все более популярной точке зрения, поведенческий принцип, реализуемый пользователем при поиске, — это *принцип минимума усилий* (напр., [4]). Действительно, все наблюдения за поведением пользователя, *ищущего “для себя”*, подтверждают этот тезис. В этом случае ПП, следующее принципу минимума усилий (ПМУ), как правило, обеспечивает минимальный риск на каждом шаге и минимизирует общее время достижения результата.

В то же время, в схеме поиска с посредником, роль которого традиционно отводится работнику библиотеки, ПП посредника не подчиняется ПМУ. Посредник не очень хорошо понимает высказанную ему информационную потребность, но сознает, что она может иметь множество интерпретаций, о которых он не знает, но ни одну из которых он не хотел бы упустить в запросе.

Библиотечный работник — это “классический” посредник и подавляющее большинство работ по поиску с посредником посвящены описанию взаимодействия патрона-читателя и посредника-библиотекаря. Однако сегодня гораздо более массовой схемой поиска с посредником является другая, в которой патрон-начальник поручает посреднику-подчиненному найти нечто. Если в классической схеме взаимодействие патрона и посредника итеративно (в процессе выполняемого посредником поиска патрон может уточнять для посредника свою информационную потребность), то в неклассической — однократно (посредник получает задание, которое уже не уточняется). Другое отличие неклассической схемы — то, что посредник обычно ориентируется в теме не хуже патрона и, как результат, ПП такого посредника отличается от ПП классического, а именно следует, как и ПП в схеме без посредника, принципу минимума усилий.

Чем выше уверенность в понимании того, что нужно патрону, тем меньше отличий ПП посредника от поиска “для себя”. Чем ниже эта уверенность, тем более посредник стремится *минимизировать риск ошибочной интерпретации* изложенной ему информационной потребности. Исполнитель исходит из того, что изложенное допускает заведомо

более одной интерпретации и стремится сформулировать запрос так, чтобы тот покрывал возможно больше интерпретаций. При этом исполнитель исходит из того, что ему известны заведомо не все возможные интерпретации. В результате, поиск ориентирован не на точность, а на полноту. Исполнитель следует *принципу гарантированного результата*.

Взаимодействуя с ПМ, исполнитель, ищущий для себя, имеет дело с сочетанием двух видов неопределенности: распределения искомым документов и собственно поведения ПМ (интерпретацией ею запроса). Исполнитель-посредник также сталкивается с неопределенностью поисковой среды, с той только разницей, что поведение ПМ для него более предсказуемо, чем для обычного пользователя, а распределение документов, представляющих мало-знакомую посреднику область (посредник — специалист в поиске “вообще”, но не специалист в предметной области), напротив, менее предсказуемо. Но помимо неопределенности среды, посредник имеет дело еще и с неопределенностью поисковой потребности пользователя, и эта неопределенность является для него основной.

4.1 Поиск без посредника и принцип минимума усилий

Задача поиска для человека относится к числу вспомогательных действий, она не самоценна. Так, если математик может потратить годы на отыскание более красивого доказательства теоремы, то едва ли кто-то согласится потерять и минуту на конструирование “красивого” запроса. Человек, *ищущий для удовлетворения собственной поисковой потребности*, следует принципу минимума усилий. Этот принцип применим на всех уровнях ПП — от выбора источника (беседа с коллегой, либо просмотр литературы, либо поиск в базе) до конкретных действий по модификации запроса в последовательном интерактивном поиске. В то же время, ПМУ — скорее концептуальный принцип (и именно в таком качестве, как принцип, способный объяснить все, он и был предложен Ципфом [27]). Всякий раз он нуждается в конкретизации того, что же понимать под “усилием”, т.е. в не всегда возможном переходе к измеримой величине. Простейшим, с технической точки зрения, но не всегда адекватным коррелятом усилия служит затраченное время. При мгновенной реакции ПМ все время в сессии уходит на формулировку запросов и просмотр результатов.

ПМУ реализуется в ленивых тактиках — пользователь “говорит” в запросе меньше, чем мог бы сказать, если бы затратил дополнительное время на формулировку запроса. Это существенно *интерактивный* стиль — очередной запрос (в том числе, первый) не рассматривается пользователем как последний. Процедура поиска для пользователя априори многошаговая.

На первый взгляд, ленивая тактика является тактикой *минимального риска*. Это не совсем так и здесь возможны две противоположные ситуации.

Первая, когда ленивая тактика является разумно осторожной. Именно, если уверенность в том, что более сложная модификация запроса не ухудшит результат, превышает некоторый порог, пользователь прибегнет к этой сложной модификации. Если же уверенность недостаточно высока, будет использована более простая модификация, для которой либо уверенность выше порога, либо модификация предельно проста (включение одного термина). Пусть результат оказался ухудшен. Тогда в случае простейшей модификации пользователь на следующем шаге должен вместо ухудшившей модификации произвести новую, а в случае сложной модификации ему неизвестно, какая именно из ее компонент вызвала ухудшение и поэтому ни одна из них не может быть исключена из дальнейшего использования. Поэтому вместо введенной модификации можно использовать любую ее часть, а тем самым шаг оказался бесполезным. Чтобы сократить бесполезные шаги, пользователь вносит сложные модификации только когда достаточно уверен в их пользе. Это тактика минимального риска.

Другой случай. Положим, что пользователь нашел в PDF-документе ссылку на заведомо нужную статью. Тогда тактика минимального риска состоит в том, чтобы руками набрать авторов и название. При безошибочном наборе спецификаций и при условии, что эти спецификации согласуются со спецификациями документа в базе, процедура поиска потребует одного шага. Ленивая же тактика состоит в том, чтобы набрать только часть спецификации, почти достаточную, на взгляд пользователя, для того, чтобы статья попала в начало списка. Понятно, такая тактика не есть тактика минимального риска и в среднем требует больше одного шага.

Однако последняя ситуация нетипична. Кроме того, учитывая, что в базе нужный документ может иметь несколько иные спецификации, и что при наборе пользователь может допускать ошибки, полное специфицирование может оказаться не минимизирующим риск: если документ, точно совпадающий с заданной в запросе спецификацией, не найден, это не означает, что искомого документа нет в базе. И тогда требуется дополнительный шаг поиска.

Таким образом, на практике, поведение, подчиняющееся ПМУ, почти всегда является и поведением минимального риска.

4.2 Поисковое поведение посредника и принцип гарантированного результата

ПП посредника, выполняющего поиск для “патрона” отличается от поиска “для себя”. Именно, посредник не следует ПМУ. Его поведение отвечает “принципу максимальной полноты запроса”, или *принципу гарантированного результата* (ПГР), понимаемому в том смысле, посредник старается не упустить в запросе ни одной интерпретации полученного задания, при этом не только не будучи уверен, какая из интерпретаций важна для патрона, но даже не зная, какие интерпретации возможны. В

таким случае мы имеем дело с вариантом принципа минимума риска.

Патрон и посредник имеют разные модели мира, и предметная область, интересующая патрона обычно мало знакома посреднику. Соответственно посредник не может локализовать в своей модели мира информационную потребность патрона. Как результат, наблюдения за ПП посредника свидетельствуют, что он делает более широкие запросы в более общих терминах, чем нужно патрону [14]. Тактика максимальной полноты запроса является следствием не некоей “злокозненной” методики поиска, исповедуемой посредниками (в [14] приводится пример того, что вместо заказанного ему типа лодок посредник указал в запросе лодку вообще), а неспособности посредника достаточно четко соотнести потребность патрона с тем знанием о мире, которым располагает посредник.

Посредник говорит в своем запросе все, что он знает. Это стиль *неинтерактивен* и соответствует ситуации, когда очередной запрос (в том числе, начальный) рассматривается как последний. Посредник формулирует запросы так, чтобы гарантировать попадание в [обозреваемую часть] результатов поиска по возможности всех интерпретаций воспринятой им информационной потребности патрона во всех возможных контекстах. Эта тактика ориентирована на полноту охвата интерпретаций, а не точность поиска. Однако чем точнее соотносит посредник информационную потребность пользователя, тем дальше его манера поиска от тактики максимальной полноты и ближе к ленивой тактике.

Таблица 4. ПП исполнителя-патрона (“поиск для себя”) и исполнителя-посредника (“поиск под чужую информационную потребность”)

Исполнитель	Способен определить пертинентность	Цель исполнителя	Принцип поиска на каждом шаге
сам патрон	да	точность итогового результата	ПМУ
Посредник	частично	полнота запроса	ПГР

4.3 Диполь “ленивый пользователь — ответственный посредник”?

Если ПП в поиске без посредника и ПП “классического” посредника-библиотекаря выглядят как своего рода поведенческий диполь, на одном конце которого реализуется ПМУ, а на другом — ПГР, то массовые на сегодня формы ПП “неклассического” посредника в эту схему не укладываются. Нетрудно заметить, что разница между ПП “для себя” и ПП посредника вытекает не из роли последнего, а из его способности локализовать информационную потребность в своей модели мира. При поиске для себя, даже если предметная область плохо знакома исполнителю, его потребность согласуется с этим уровнем знания и локализована в его модели мира.

Зависимость формулируемого исполнителем запроса от способности локализовать потребность немонотонна³. Как видно из Табл.5, когда информационная потребность абсолютно не ясна (не локализована), исполнитель в точности воспроизведет сказанное патроном. Когда потребность ясна приблизительно, исполнитель сформулирует запрос шире высказанной потребности. Наконец, когда потребность предельно (но, возможно, ошибочно) ясна исполнителю, запрос будет отражать именно эту интерпретацию, т.е. будет уже буквально высказанной потребности.

Таблица 5. ПП посредника в зависимости от способности локализовать внешнюю для него информационную потребность

Способность локализовать потребность	Формулируемый посредником запрос	Принцип поиска	Характерная ситуация и ее типичность
отсутствует	точно воспроизводит высказанную потребность	минимума риска [минимума усилий]	специальная потребность (уникально)
частичная	запрос шире потребности	гарантированного результата	читатель-библиотекарь (немассово)
полная	запрос уже буквально высказанной потребности	минимума усилий	начальник-подчиненный (массово)

6. Оптимальные стратегии и два типа поиска

Опыт подсказывает, что, сделав что-то небрежно, нам приходится переделывать сделанное заново, и что, торопливость на каждом шаге, может только удлинить процесс в целом. Так ли это применительно к поисковому поведению?

Рассмотрим следующую модель последовательного поиска, формулируемую в терминах, близких к языку модификаций запросов Бейтс-Фидель [2,3,6]. Пусть пользователь может назвать любую комбинацию любых слов. В ответ он получает список как-то соотносящихся с этой комбинацией документов. Некоторые комбинации являются “магическими”, т.е. такими, что обозреваемая часть списка содержит достаточное число (например, 1) искомым документов. Цель поиска состоит в том, чтобы *отгадать одну из магических комбинаций*. Пользователь способен сравнить результаты двух последовательных шагов поиска (т.е. оценить, привела ли модификация запроса на очередном шаге к ухудшению

³ Проявление аналогичной немонотонности мы встречали в п. 2.3.1: согласно [5], запросы опытного пользователя длиннее, чем у новичка, а согласно [7], дело обстоит наоборот. Видимо, “опытный пользователь” из [5] оказался бы в [7] где-то между новичком и [еще более] опытным пользователем, поскольку, хотя и в силу противоположных причин, у обоих запросы короче.

ила улучшению результата). Пользователь имеет право на очередном шаге поиска добавлять, удалять или заменять любое число слов из текущей комбинации. Такая постановка позволяет рассмотреть следующие критерии *последовательного* поиска.

Критерий минимального числа добавленных слов (включая слова исходного запроса). Во-первых, стратегия, состоящая в минимальной модификации (добавлении, удалении или замене в точности одного слова) является оптимальной, какой бы ни была неопределенность среды. Помимо того, в специальном случае отсутствия неопределенности (пользователю заранее известна одна из магических комбинаций, как например, в случае с упомянутой выше ссылкой в PDF-файле) оптимальны стратегии, состоящие во вводе любого числа слов на каждом шаге отгадывания (в частности, всех слов магической комбинации на первом же шаге).

Критерий минимального числа шагов. Чем выше неопределенность, тем меньше слов запроса может быть модифицировано на каждом шаге. При высокой неопределенности стратегия минимальной модификации является оптимальной.

Критерий минимального суммарного времени. Простейшей мерой усилия является затраченное время. Если положить время реакции системы и время реакции пользователя (время на просмотр результатов и модификацию запроса) нулевыми, имеем критерий минимума шагов. Более реалистично допущение, что время реакции пользователя зависит от сходства результатов, полученных на последовательных шагах. В этом случае предпочтительность простых модификаций повышается в сравнении с критерием минимума шагов.

Итак, подчиняющееся ПМУ ленивое поведение, которое состоит в минимальной модификации запросов, субоптимально в условиях неопределенности среды. Правила модификации запросов, в отличие от числа шагов и средней длины запроса, не зависят от цели поиска (точность или полнота).

В отличие от поведения, подчиняющегося ПМУ, поведение, подчиняющееся ПГР (т.е. ПП классического посредника) в среднем требует больше термов, больше шагов и больше суммарного времени.

7. Многозначность запроса

Рассматривая схему поиска с посредником, мы обсудили неопределенность интерпретаций поисковой потребности патрона, т.е. неопределенность того “запроса”, который он адресуется посреднику. Аналогичная коллизия множественности допустимых интерпретаций возникает при обработке запроса поисковой средой.

7.1 Основные типы поиска и интерфейсы

Второй участник поискового взаимодействия (ПС) представим в виде интерфейса и используемых поисковых схем, включающих способ отображения результатов (линейный список, дерево). Кратко оха-

рактируем те схемы поиска, которые важны с точки зрения взаимодействия.

Поиск по релевантности. Основной для всех ПМ. В качестве меры релевантности используются оценки контрастности (TF*IDF-оценки) термов запроса. Имеет множество модификаций [16].

Поиск по новизне. В построении ранжировки учитываются два фактора: релевантность документа запросу и “новизна” документа относительно уже включенных в список. Результирующее упорядочение определяется весами релевантности и новизны. Но 1) эта схема не гарантирует, что документ с высокой новизной, но представляющий незначительную группу, попадет в обозримую часть списка; 2) восприятие линейного списка, построенного по свертке критериев релевантности и новизны, затруднительно из-за психологически ощутимой немонотонности.

Расширение запроса. Здесь исходный запрос дополняется какими-то дополнительными терминами. Простейший пример — автоматическое пополнение синонимами. Поскольку синонимия зависит от предметной области, возможны забавные результаты.

Обратная связь. В как-то построенной ранжировке документов пользователь помечает один или несколько документов, релевантных его потребности. ПМ должна определить, что отличает эти документы и построить по этим признакам новую ранжировку. Поскольку ПМ может усмотреть вовсе не те отличия, которые подразумевает пользователь, переупорядочивание может ухудшить ранжировку.

Поиск по категориям. Обычно состоит в продвижении пользователя по тематическому дереву. Использует предварительно заданный рубрикатор и требует рубрицирования всех индексируемых документов.

7.2 Пример: запрос “звёзды”

Как реагируют реальные ПМ на запрос “звезды”? Это слово имеет много контекстов. Естественно было бы ожидать, что по столь многозначному запросу ПМ предложит не только список, упорядоченный по TF*IDF-релевантности, но и какую-то категоризацию. Однако надежду найти смыслы, отличные от персоналий, названий и (иногда) собственно звезд оставляет только *Yahoo* (табл. 6).

Таблица 6. Первая страница результатов поиска “звезд”

ПМ	Первый экран результатов поиска	Наличие категоризации
<i>Yahoo</i> (“stars”)	персоналии, название клуба, астрономия	Есть, достаточно логичная и удобная
<i>Google</i> (“stars”)	персоналии, издания, астрономия (1)	Вырожденная
<i>Google</i> (“звезды”)	персоналии, названия	Сильно вырожденная
<i>Excite</i> (“stars”)	Предложения купить название звезды, персоналии, название клуба	Есть, большая, но непредставительная и невнятная
<i>AltaVista</i> (“stars”)	Примерно как в <i>Excite</i>	Нет

7.3 Многозначность запроса

Во всех схемах поиска, кроме упорядочения по релевантности, отражен факт многозначности запроса — один и тот же запрос представляет совершенно разные информационные потребности пользователей. Но пользователь нуждается в удовлетворении своей конкретной потребности, а вовсе не некоего усреднения. Ранжирование по релевантности и по свертке релевантность–новизна удовлетворяют усредненную потребность. В свою очередь, механизмы обратной связи/расширения запроса действенны только если нужная пользователю информация попадет в обозримую часть списка, что происходит далеко не всегда. Только поиск по категориям *потенциально* разрешает коллизию многозначности. Однако у поиска по категориям есть тот же недостаток, что у ПП классического посредника, — утяжеление процесса поиска. Как следствие, он может играть только вспомогательную роль, обеспечивая возможность найти документы, удовлетворяющие редкой интерпретации запроса.

8. Что ищет пользователь

Потребность пользователя состоит в том, чтобы найти объект, обладающий некоторыми свойствами (О-свойствами), во всех источниках, обладающих еще какими-то свойствами (И-свойствами). *И-свойство* — характеристика источника как целого. *О-свойство* — нечто, упоминаемое внутри источника и не являющееся И-свойством. Примеры И-свойств: тематическая, временная, географическая, принадлежность документа, авторство и т.д. Хотя И-свойства могут напрямую указываться внутри источника в таких “тэгах” как аннотация, ключевые слова, дата и т.п., тем не менее, это характеристики источника в целом. И-свойства часто *вообще не упоминаются* в источнике (слово “комбинаторика” почти не встречается в текстах по комбинаторике). По частоте, с которой в источнике присутствует название И-свойства, нельзя судить о степени, в которой источник этим свойством обладает.

Поэтому поиск по О-свойствам и поиск по И-свойствам не могут быть реализованы одинаково. TF*IDF-оценки хороши для поиска по О-свойствам, но неприменимы к И-свойствам. Однако в языке запросов современных ПМ нет способа разделить (различить) два типа свойств и их приходится задавать в одном выражении. Но, в силу указанной особенности И-свойств, TF*IDF-методы дадут для них очень плохой результат. Поиск источника с заданным И-свойством должен быть основан на том, отнесен ли источник к “классу”, совпадающему или в некотором смысле близкому к этому свойству.

Совершенно нормальны вырожденные ситуации, когда один из элементов запрашиваемой пары <О-свойства><И-свойства> пуст. Так, документальный поиск — это поиск исключительно по И-свойствам, а фактографический — по О-свойствам.

Источник может характеризоваться сразу несколькими свойствами, причем не обязательно, от-

носящимися к разным шкалам. Так учебник — составной источник, в котором отдельные главы обладают разными свойствами из одной тематической шкалы. А междисциплинарная статья вся целиком обладает несколькими свойствами.

Пользователь обращается к ПМ либо (1) с целью найти нечто конкретное (объект и/или источник, чаще объект), либо (2) с целью ориентировки (в И-и/или О-свойствах, чаще И-). Если результат запроса неудовлетворителен, поиск либо завершается, либо продолжается с уточненным запросом. Если удовлетворителен, поиск либо завершается, либо смещается. Последовательный процесс может включать несколько смен ориентировочного поиска на конкретный и наоборот, а стадия ориентировки не обязательно предшествует конкретному поиску.

С описанных позиций, ближе всего к идеалу поисковой среды ОПАС, поскольку предлагают явные поля спецификации И-свойств (тематика, авторство, название). Единственный недостаток в том, что свободное задание в этих полях ограничено.

9. Поисковое взаимодействие и принцип максимума разнообразия

Поиск есть устранение разнообразия документов, удовлетворяющих (с точки зрения ПМ) сформулированной потребности. Отсюда 1) сама форма задания запроса с разделением полей И- и О-свойств должна сокращать разнообразие интерпретаций, 2) результаты поиска должны быть представлены пользователю так, чтобы ему было максимально просто сократить “остаточное разнообразие” множества документов, удовлетворяющих запросу. В то же время, в силу буквального воспроизведения коллизии множественности интерпретаций и (почти) полного совпадения принципов максимума разнообразия и гарантированного результата, реализации поиска в форме поиска по категориям должна отводиться вспомогательная роль. Это средство эффективно только для редких категорий.

10. Благодарности

Автор признателен В. Арлазарову за плодотворное обсуждение и частичное финансирование и А. Спинк за предоставленные материалы по проекту *Excite*.

Литература

- [1] Aula A. Query formulation in Web information search. *Proceedings of Int. Conf. on WWW/Internet*, pages 403-410, 2003.
- [2] Bates M. Information Search Tactics. *JASIS*, 30(4), pages 205-214, 1979.
- [3] Bates M. Where should the person stop and information interface start? *IPM*, 26(5), pages 575-591, 1990.
- [4] Bates M. Task Force Recommendation 2.3. Research and design review: Improving User Access

- to Library Catalog and Portal Information. Final Report. 2003.
- [5] Fenichel C. Online searching: Measures that discriminate among users with different types of experience. *JASIS*, 32, pages 23-32, 1981.
- [6] Fidel R. Moves in Online Searching. *Online Review*, 9(1), pages 61-74, 1985.
- [7] Hsieh-Yee I. Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers. *JASIST*, 44(3), pages 161-174, 1993.
- [8] Holscher C., Strube G. Web search behavior of internet experts and newbies. *Proc. Of 9th WWW conference*, pages 337-346. 2000.
- [9] Jansen B., Pooch U. A review of Web Searching Studies and a Framework for Future Research, *JASIST*, 52(3), pages 235-246, 2001.
- [10] Jansen B., Spink A. The Excite Research Project: A Study of Searching by Web Users. *ASIS Bulletin*, 27(1), pages 15-17, 2001.
- [11] Jansen B., Spink A., Saracevic T. Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web. *IPM*, 36(2), pages 207-227, 2000.
- [12] Jones S., Cunningham S., McNab R., Boddie S. A transaction log analysis of a digital library. *Int. J. on Digital Libraries*, 3(2), pages 152-169, 2000.
- [13] Meadow C., Marchionini G., Cherry J. Speculations on the Measurement and Use of User Characteristics in Information Retrieval Experimentation. *Canadian J. of Information and Library Science*, 19(4), pages 5-7. 1994.
- [14] Nordlie, R. Unmediated and Mediated Information Searching in the Public Library. ASIS 1996 Annual Conf. Proc., 1996, <http://www.asis.org/annual-96/ElectronicProceedings>
- [15] Niu X., Bizants A. The effects of search environment and task realism on search behavior. *Human Factors and Ergonomics Society 45th Annual Meeting*. 2001.
- [16] Ruthven I., Lalmas M., van Rijsbergen K. Combining and selecting characteristics of information use. *JASIST*, 53(5), pages 378-396, 2002.
- [17] Saracevic T. Modeling interaction in information retrieval: a review and proposal. *Proceedings of ASIS*, 33, pages 3-9, 1996.
- [18] Saracevic T., Kantor P., Chamis A., Trivison D. A study of information seeking and retrieving. I. Background and methodology. *JASIS*, 39(3), pages 161-176, 1988.
- [19] Seiden P., Szymborski K., Norelli B. Undergraduate Students in the Digital Library: Information Seeking Behavior in an Heterogenous Environment. *ACRL National Conference*, 1997. <http://www.ala.org/acrl/paperhtml/c26.html>
- [20] Siegfried S., Bates M., Wilde D. A profile of end-user searching by humanities scholars. *JASIS*, 44(5), pages 273-291, 1993.
- [21] Silverstein C., Henzinger M., Marais M., Moricz M. Analysis of a very large Web search engine query log. *ACM SIGIR Forum*, 33(3), 1999.
- [22] Spink A., Wolfram D., Jansen B., Saracevic T. Searching the Web: The Public and Their Queries. *JASIST*, 52(3), pages 226-234, 2001.
- [23] Spink A., Jansen B., Wolfram D., Saracevic T. From E-Sex to E-Commerce: Web Search Changes. *Computer*, March 2002, pages 107-109, 2002.
- [24] Vakkari P. eCognition and changes of search terms and tactics during task performance. *Proceedings of RIAO 2000*, pages 894-907, 2000.
- [25] Wildemuth B., Jacob E., Fullington A., de Blicke R., Friedman C. F A detailed analysis of end-user search behaviors. *Proceedings of 54 ASIS Annual Meeting*, pages 302-312, 1991.
- [26] Wolfram D., Spink A., Jansen B., Saracevic T. Vox Populi: The Public Searching of the Web. *JASIST*, 52(12), pages 1073-1074, 2001.
- [27] Zipf, G. Human behavior and the principle of least effort. Cambridge, MA: Addison-Wesley. 1949. (Сокращения: *JASIS[T]* — *Journal of the American Society for Information Science [and Technology]*, *JD* — *Journal of Documentation*, *IPM* — *Information processing and Management*)

IR Interaction: Myths, Reality and Basic Principles

Nikolai Buzikashvili

Counter to commonly accepted views observations of interactive seeking behavior in different search environments show that a user behaves precisely in the same manner in traditional IR systems and in the Web. Unmediated users resort to the same tactics of successive search based on the principle of least effort (PLE). On the contrary, an intermediary follows the principle of guaranteed result (PGR) rather than PLE. A decision-theoretic model of successive search is considered; optimal strategies of seeking are formulated for different criteria, and 'lazy user' seeking behavior based on PLE and seeking behavior based on PGR are compared with the optimal strategies. It is shown that PLE-based tactics usually yields suboptimal strategy under uncertainty of search environment. On the contrary, the PGR-based seeking is not effective on average.

Also IR system behavior is considered, in particular, from the point of view of diversity of queries interpretations. Technical solutions to reduce this diversity are proposed. At the same time, it is shown that IR based on maximum diversity/coverage principle is similar to PGR-based seeking and gives worse results except for uncommon interpretations.