Разработка модели информационного портрета пользователя для персонифицированного поиска

Алексей Владимирович
Широков
avshirokov@yahoo.com

Аннотация

В статье рассматривается проблема улучшения поисковых и навигационных систем. Для ее решения предлагается использовать информационный портрет пользователя. Описывается метод построения и применения такого портрета, основанный на ключевых словах. Приводятся результаты экспериментов, показывающих его работоспособность. Делается вывод о том, что системы, использующие информационный портрет дают лучшие результаты, чем системы без персонификации.

1. Введение

В связи с увеличением объема данных, размещенных в Интернете и корпоративных базах данных, актуально стоит проблема повышения качества поисковых и навигационных систем. Сейчас поисковая система может судить об информационной потребности пользователя только лишь по ключевым словам, заданным в запросе. То есть решается задача нахождения значения функции

$$v = f(V, q),$$

где V — множество всех узлов гипертекста, v — множество узлов, отобранных по запросу, q — условия пользовательского запроса. При этом, система не располагает сведениями о контексте, в котором пользователь употребляет эти слова. Да и самому пользователю бывает трудно подобрать ключевые слова, которые точно описывают интересующую его тематику. Поисковая система предоставляет пользователю результаты, отсортированные по степени релевантности к запросу. При этом учитывается содержание документов или ссылочная структура между ними, но не учитывается круг интересов и индивидуальные особенности пользователя, ведущего поиск. Чтобы сделать поиск более адекватным, предлагается учитывать информационный портрет пользователя. То есть решать задачу вычисления функции

$$v = f(V, q, p),$$

где p — информационный портрет пользователя. Это позволит поисковой системе лучше представить контекст употребления ключевых слов и правильнее ранжировать результаты. В применении такого подхода, основными задачами являются:

1. Разработка методики создания информационного портрета пользователя.
2. Использование для первоначального поиска и последующей навигации информационного портрета пользователя.
3. Создание комплекса программ, эффективно использующего модель информационного портрета.

Исследование посвящено алгоритмам построения информационного портрета и его использованию для повышения эффективности поиска.

Информационным портретом пользователя (ИПП) будем называть набор параметров и их значений, описывающих сферу интересов пользователя, интересующие его области знаний.

Обычно, под ИПП понимается вектор, элементами которого являются понятия с указанием веса, характеризующего степень интересности понятия пользователю.

Такой портрет можно составлять методом предварительного анкетирования пользователя или учитывая активность пользователя при работе с информационными ресурсами. Первый способ считается слишком трудоемким, требующим с одной стороны составления тестовых наборов, охватывающих всю область знаний, а с другой стороны, больших усилий и временных затрат при прохождении пользователем этих тестов. Тем не менее, тесты можно использовать для первоначального приближенного формирования информационного портрета. Подход, учитывающий активность пользователя, лучше подходит для решения задачи формирования информационного портрета. Он не требует от пользователя большого количества дополнительных действий и потенциально обладает большей точностью.

Есть два направления: персональные агенты [2, 3, 4, 7], и персонификация поисковой системы [5].

В первом случае система строит портрет на основе просматриваемых пользователем документов из различных источников. Во втором, система отслеживает документы, отобранные пользователем в результатах, выданных по запросу конкретной поисковой системой.

В обоих случаях в основе могут лежать одни и те же алгоритмы построения ИПП. Отличие в том, что персональный агент строит векторы документов налету, а персонификатор поисковой системы может использовать векторы документов, определенные ранее при индексировании.

Обычно применяют способ построения ИПП, при котором система классифицирует выбранные

пользователем документы в соответствии с некоторой онтологической структурой [7]. Категории онтологии, в которые попало больше документов, составляют ИПП. Документ соотносится с категорией на основе скалярного произведения их n-мерных векторов. Реализации различаются способом построения n-мерного пространства и методом вычисления скалярного произведения. В качестве координат обычно используют ключевые слова и их веса.

Естественно предположить, что задача формирования ИПП сводится к задаче автоматической категоризации текстов. Однако для обучения большинства систем категоризации требуется уже готовая структура категорий с достаточно большой обучающей выборкой по каждой категории. Такие структуры слишком общи и не всегда адекватны интересам пользователя. В связи с этим больший интерес представляют алгоритмы, которые эффективно решают свою задачу в условиях малых обучающих выборок, когда обучение начинается с одного-двух документов, и система классификации переобучается при добавлении нового документа.

Системы построения ИПП также различаются способом получения пользовательской оценки документов. Это может происходить в автоматическом или ручном режиме. Например, если пользователь просматривает документ дольше определенного времени, то считается, что пользователь оценил документ положительно [7]. В ручном режиме пользователь явно указывает, какие документы он считает соответствующими запросу [3]. В любом случае, система оценивания должна быть организована так, чтобы обеспечивать достаточную точность при небольших усилиях со стороны пользователя.

В исследовании ставится задача отработать взаимодействие пользователя и компонентов системы и проверить, можно ли достичь практически применимых результатов, используя один из подходов построения ИПП, основанный на ключевых словах.

2. Идея исследования

В качестве гипотезы примем утверждение, что применение даже простого алгоритма персонификации дает лучшие результаты по сравнению с поисковой системой без персонификации.

Для проверки этого утверждения, разрабатывается экспериментальный программный комплекс, реализующий метод построения ИПП, учитывающий пользовательские оценки.

В отличие от исследований, где для построения ИПП используются только положительные примеры без различий в величине оценки [3], предлагается учитывать как положительные, так и отрицательные примеры, оцененные по шкале с несколькими уровнями.

Создаваемая система персонификации не использует жестких заранее построенных онтологических структур [7], а строит гибкий по составу информационный портрет в процессе диалога с пользователем и только на основе предоставляемых им данных. Количество тематических категорий не ограничивается.

Сложность подобных исследований состоит в том, что нет готовых наборов данных для тестирования систем персонификации. Поэтому в предлагаемой работе одновременно с построением ИПП, сохраняются поисковые истории участников эксперимента. Часть такой истории используется для построения ИПП, а другая часть для оценки системы персонификации.

По собранным данным оценивается устойчивость формируемых портретов во времени и величина отличия друг от друга портретов разных пользователей. Эти характеристики покажут, способен ли предлагаемый алгоритм персонификации выявлять индивидуальные предпочтения пользователей.

Для сравнения систем работающих с учетом и без учета портретов пользователя, для интернет-страниц из поисковых историй вычисляются три оценки: по версии пользователя, по версии системы персонификации и по версии обычной поисковой системы. Все оценки нормируются в диапазоне от −1 до 1. Чем больше расхождение в оценках интернет-страниц участником эксперимента и оцениваемой системой, тем хуже работает система.

Кроме выводов относительно эффективности предложенного алгоритма персонификации, создаваемый программный комплекс рассматривается в качестве прототипа персонального поискового агента. С этой позиции методами интервьюирования исследуются удобство системы оценки интернет-страниц с точки зрения пользователя, и возможности практического применения ИПП.

Чтобы выявить все особенности работы системы персонификации и поискового поведения ее потенциальных пользователей, работоспособность системы проверяется в условиях обычной повседневной работы участников эксперимента с реальными интернет-страницами.

3. Описание методов, алгоритмов и экспериментов

3.1 Экспериментальный программный комплекс

Для проведения экспериментов был создан программный комплекс, состоящий из промежуточного прокси-сервера, базы данных для хранения журнала оценок, и набора программ-оценщиков, которые по различным методикам вычисляют оценки вновь поступающих документов.

Были разработаны три программы-оценщика для получения оценки документа: непосредственно

пользователем (userRater), по версии поисковой системы без персонификации (sengineRater), и с использованием ИПП (profileRater).

3.2 Получение пользовательской оценки

Участники эксперимента использовали обычные знакомые им интернет-браузеры. Браузеры настраивались на работу через специальный прокси-сервер. В начале сеанса работы, участники проходили идентификацию средствами прокси-сервера. Дальнейшая работа с Интернетом велась в обычном режиме. Технических ограничений на порядок и состав посещаемых страниц не налагалось.

Прокси-сервер добавлял к каждому, проходящему через него html-документу, подвижную форму для оценки (рис. 1).

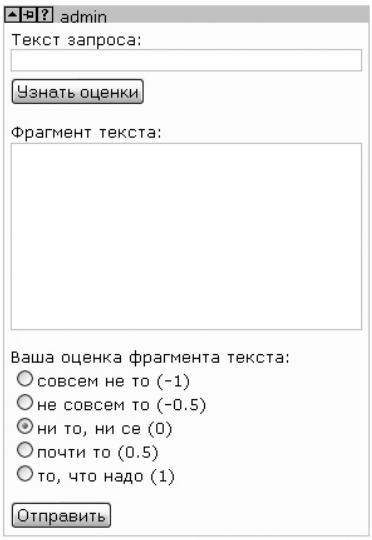

Рис. 1. Форма для оценки документа

Пользователь, имея в виду какой-либо запрос, с помощью поисковой системы или другими способами находил интернет-страницы (документы), просматривал их, выделял в них фрагменты текста, которые он хотел оценить, и выставлял оценки этим фрагментам. Оценка показывает насколько точно, по мнению пользователя, фрагмент соответствует запросу.

Оценки выставлялись по шкале:

- совсем не то (-1)
- не совсем то (-0,5)
- ни то, ни сё (0)
- почти то (0,5)
- то, что надо (1)

Данные формы отправлялись прокси-серверу, который заносил их в журнал оценок.

Таким образом, был накоплен журнал оценок, в котором хранятся:

1. Дата оценки.
2. Имя участника эксперимента.
3. Текст запроса.
4. Ссылка на документ (url).
5. Фрагмент текста.
6. Оценка соответствия фрагмента запросу.

Если в одном документе пользователь давал оценку нескольким фрагментам, то в журнал отдельными строками записывались все фрагменты и оценки.

Программа-оценщик userRater вычисляет общую оценку документа как среднеарифметическое значение оценок всех фрагментов этого документа.

Ручной метод сбора оценок был выбран потому, что он, хотя и требует от пользователя некоторых усилий, обладает множеством преимуществ. Позволяет отсекать несодержательную часть документов (рекламу, фреймы, элементы навигации). Дает возможность оценить разные фрагменты одного документа по-разному. Позволяет выразить содержание документа своими словами в поле «Фрагмент текста». Это полезно когда в документе нет подходящих ключевых слов, или документ насыщен графикой или флэш-элементами.

3.3 Сбор данных для исследования

В эксперименте участвовало 6 человек. Исходя из задач исследования, основной целью на этом этапе было не привлечение большого количества испытуемых, а получение достаточно длинных индивидуальных поисковых историй. Пользователи работали с системой как в свободном режиме, когда поиск велся по любой интересной пользователю теме, так и в режиме целевого задания, когда ведущий эксперимента задавал нескольким участникам один и тот же запрос, но описывал каждому из них различный тематический контекст.

Для исследования были отобраны истории по 14 запросам, с длинной более 20 оцененных фрагментов в каждой. Средняя длина истории составила 42,14 фрагмента.

3.4 Построение ИПП

Были разработаны два алгоритма построения портрета. В их основе лежит метод TF-IDF [6], модифицированный так, чтобы учитывать оценку пользователя и получать результат в условиях малых выборок и единственной категории.

Алгоритм Words

Все фрагменты, помещенные пользователем в журнал, объединяем в один супердокумент. С помощью парсера mystem.exe получаем из супердокумента все русские слова в нормальной форме. После удаления стоп-слов, подсчитываем веса слов в супердокументе по формуле:

$$w_i = tf_i \times r_i \times \log\left(\frac{1000000}{ipm_i}\right),$$

где tf_i — число вхождений слова в супердокумент, r_i — средняя пользовательская оценка слова в супердокументе, ipm_i — (instances per million), среднестатистическое для русских текстов число вхождений слова на миллион [1], $i = \overline{(1,n)}$, n — количество уникальных слов в супердокументе.

Так как в журнале хранятся оценки не для

отдельных слов, а для целых фрагментов, при подсчетах r_i считается, что каждое слово фрагмента имеет такую же оценку, как и весь фрагмент в целом.

Таким образом, портрет представляет собой вектор *W*, координатами которого являются веса слов, вычисленные с учетом пользовательских оценок.

Алгоритм Querys

В супердокумент объединяем фрагменты, полученные для одного запроса. Далее, вычисляем веса слов в супердокументе так же, как в алгоритме Words.

Таким образом, портрет представляет собой набор векторов запросов, координатами которых являются веса слов, вычисленные с учетом пользовательских оценок. Такие векторы запросов удобно называть категориями ИПП. Текст запроса становится названием категории.

При добавлении каждого нового оцененного фрагмента, ИПП перестраивается.

Уже в начале эксперимента, после построения нескольких ИПП, стало ясно, что алгоритм Words дает хорошие результаты только для одного запроса. В случае, когда пользователь работает с несколькими запросами, наибольшие веса получают слова, встречающиеся сразу во всех запрашиваемых областях знаний, портрет теряет свою выразительность и требуется проводить деление портрета на категории. Поэтому, в качестве основного, далее использовался алгоритм Querys.

Применение ipm вместо классического IDF по числу документов, во-первых, позволяет, используя априорную характеристику редкости слова, эффективно вычислять веса слов уже для первого фрагмента, то есть на малых выборках, во-вторых, дает возможность строить портрет одной категории одного пользователя в условиях, когда нет других категорий или портретов пользователей для сравнения, и вычислить IDF проблематично.

3.5 Оценка документа системой персонификации

Сходство документа и портрета вычисляется следующим образом:

1. Получаем полный текст документа.
2. По алгоритму Words строится вектор документа W_{doc}. При этом, оценки r_i принимаются равными единице.
3. Из ИПП берется вектор очередной категории $W_{profile}$.
4. По словам, входящим одновременно и в W_{doc} и в $W_{profile}$, вычисляется оценка сходства этих векторов:

$$\mathrm{sim}(W_{profile}, W_{doc}) = \frac{W_{profile} \cdot W_{doc}}{|W_{profile}| \times |W_{doc}|} =$$

$$= \frac{\sum_i (w_i^{profile} \times w_i^{doc})}{\sqrt{\sum_i (w_i^{profile})^2} \times \sqrt{\sum_i (w_i^{doc})^2}},$$

где $i = \overline{(1, n)}$, n — количество совпадающих уникальных слов.

5. Переходим к пункту 3, пока не оценим сходство с каждой категорией.

В результате получаем список категорий ИПП с оценкой сходства документа с каждой из них. На практике, пользователя чаще интересует соответствие документа конкретному запросу. Поэтому для экономии ресурсов и для простоты восприятия оценок пользователем, в проведенных экспериментах, программе profileRater на вход подавался текст запроса, в соответствии с которым на шаге 3 выбирался $W_{profile}$, а шаг 5 опускался. Таким образом, оценка соответствия документа и портрета, по версии системы персонификации, вычислялась по формуле:

$$r_{profile} = sim(W_{profile}, W_{doc}).$$

Значения функции sim нормированы от -1 (документ не соответствует портрету) до 1 (документ соответствует портрету).

3.6 Оценка документа поисковой системой

Оценка документа на соответствие запросу по версии поисковой системы без персонификации вычислялась исходя из позиции документа в линейном списке всех документов, возвращенных по запросу.

$$r_{sengine} = 1 - \frac{d - 1}{dcount},$$

где d — номер позиции документа, *dcount* — общее количество документов в списке.

При этом:

- Если документ не был проиндексирован поисковой системой, то $r_{sengine} = 0$.
- Если документ был проиндексирован, но не выдан по запросу, $r_{sengine} = -1$.
- Для получения линейного списка документов, в параметрах поиска указывалось: *pag=u&rd=0*.
- Чтобы узнать, проиндексирован ли документ поисковой системой, в параметрах поиска указывалось: *ras=1&text=&site=<ссылка на документ>*.
- Запрашивались только первые 1000 позиций в результатах поиска.

3.7 Стабильность ИПП

Одной из задач исследования было определить, насколько быстро можно построить стабильный ИПП в контексте одного запроса. Для этого проводился анализ изменения ИПП. Сравнивались векторы категории ИПП до, и после добавления очередного, оцененного пользователем, фрагмента. Векторы сравнивались следующим образом.

1. По алгоритму Words строился вектор $W_{profile1}$ категории до добавления фрагмента и вектор $W_{profile2}$ категории после добавления фрагмента.
2. Оценка сходства этих векторов вычислялась по формуле:

$$\mathrm{sim}\left(W_{profile1}, W_{profile2}\right) = \frac{W_{profile1} \cdot W_{profile2}}{\left|W_{profile1}\right| \times \left|W_{profile2}\right|}$$

,

При этом если слово входило в вектор $W_{profile1}$ и не входило в $W_{profile2}$, то вес слова в векторе $W_{profile2}$ принимался равным нулю.

На графиках (рис. 2 – рис. 4) представлена оценка стабильности категории ИПП в зависимости от размера добавляемого фрагмента. По правой оси y измеряется сходство векторов (sim), по левой — количество слов в добавляемом фрагменте (words_added).

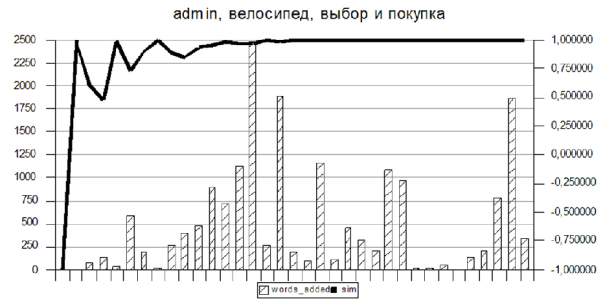

Рис. 2.

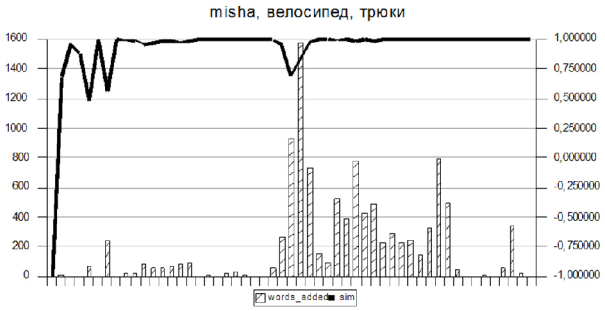

Рис. 3.

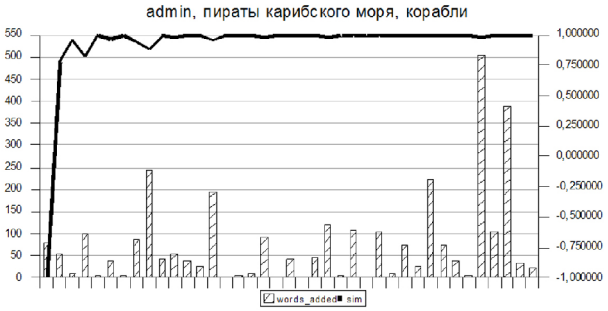

Рис. 4.

3.8 Зависимость ИПП от контекста

Чтобы узнать, отражает ли ИПП, построенный по предложенному алгоритму, индивидуальный контекст пользователя, сравнивались категории ИПП, полученные в режиме целевого задания, то есть, составленные по одному и тому же запросу разными пользователями.

Векторы сравнивались так же, как и при анализе стабильности ИПП.

Отметим, что в этом эксперименте участники пользовались одной и той же поисковой системой, поэтому ими были оценены практически одни и те же документы, что минимизирует влияние исходного оцениваемого материала на различие портретов.

Результаты сравнения похожести категорий ИПП для разных пользователей представлены в таблицах 1 и 2.

Таблица 1. Категория: пираты карибского моря

Участник	misha	oleg	admin	
Контекст	фильм	история	корабли	
	1	–0,312187	–0,219952	misha
	–0,312187	1	0,364948	oleg
	–0,219952	0,364948	1	admin

Таблица 2. Категория: велосипед

Участник	misha	oleg	admin	
Контекст	трюки	история	выбор и покупка	
	1	0,241430	0,262715	misha
	0,241430	1	0,149957	oleg
	0,262715	0,149957	1	admin

Видно, что в первом случае, участники oleg и admin были единодушны в своем негативном отношении к интернет-страницам о фильме, и система персонификации смогла выявить, что их понимание запроса отличается от контекста участника misha, который напротив, высоко оценил документы с информацией о фильме.

Во втором случае, система персонификации сформировала три портрета, которые скорее близки, чем различны, однако и тут можно говорить о том, что построенные системой портреты участников admin и misha оказались более похожи, чем портреты участников oleg и admin.

В обеих, представленных таблицах, результаты измерений совпадают с ожидаемым взаимным расположением портретов, исходя из контекстов, заданных при постановке эксперимента.

3.9 Влияние ИПП на результаты поиска

По запросным историям вычислялись различия в оценках каждой интернет-страницы: участником (R_user); поисковой системой без персонификации (R_sengine); системой персонификации, использующей ИПП, сформированный только по предыдущим оцененным фрагментам (R_profile); системой персонификации с использованием

Таблица 3. Оценки интернет-страниц

num	url	R_user	R_sengine	R_profile	R_full_profile
2	http://mykeira.com/films.shtml	-1,0000	-1,0000	-0,7677	-0,5316
3	http://pirates.film.ru/synopsis.html	-1,0000	-1,0000	-0,9271	-0,3581
4	http://piratesofthecaribbean.movieblog.ru/	-1,0000	0,9940	-0,7592	-0,6814
5	http://www.raskraska.com/raskraski/53/	-0,5000	0,0000	-0,8637	-0,4446
6	http://ru.akella.com/Game.aspx?id=137	0,0000	-1,0000	-0,5270	-0,6007
7	http://www.ezoloto.ru/index.php?level=2&	-0,5000	0,0000	-0,8298	-0,1178
8	http://www.lenta.ru/news/2007/03/18/pira	-1,0000	0,9830	-0,8720	-0,5454
9	http://games.1c.ru/pirates_of_the_caribbe	-0,5000	-1,0000	-0,5857	-0,1951
10	http://oper.ru/news/read.php?t=10516019	-1,0000	-1,0000	-1,0000	-1,0000
11	http://piratesofthecaribbean.movieblog.ru/	0,0000	-1,0000	-0,8516	-0,5146
12	http://www.mir66.ru/competitions/view_cc	-1,0000	0,9250	-0,6687	-0,2622
13	http://filmz.ru/film/2682.htm	-1,0000	0,7870	-0,5675	-0,2158
14	http://pirates.my1.ru/	1,0000	0,9460	-0,4134	0,0360
15	http://pirates.my1.ru/publ/3-1-0-20	1,0000	-1,0000	-0,2989	0,1872
16	http://pirates.my1.ru/publ/3-1-0-19	1,0000	-1,0000	-0,3318	0,0574
17	http://pirates.my1.ru/publ/4-1-0-18	1,0000	-1,0000	-0,3473	0,4054
18	http://pirates.my1.ru/publ/4-1-0-16	1,0000	-1,0000	-0,2541	0,1777
19	http://pirates.my1.ru/publ/4-1-0-15	1,0000	-1,0000	-0,1820	0,1166
20	http://pirates.my1.ru/publ/4-1-0-14	1,0000	-1,0000	-0,2404	0,1452
21	http://pirates.my1.ru/publ/4-1-0-10	1,0000	-1,0000	0,1138	0,5927
22	http://pirates.my1.ru/publ/4-1-0-9	1,0000	-1,0000	0,1683	0,6715
23	http://pirates.my1.ru/publ/4-1-0-8	0,5000	-1,0000	0,2851	0,3347
24	http://pirates.my1.ru/publ/2-1-0-3	1,0000	-1,0000	0,3683	0,4207
25	http://pirates.my1.ru/publ/2-1-0-4	1,0000	-1,0000	0,6536	0,5145
26	http://pirates.my1.ru/publ/1-1-0-1	1,0000	-1,0000	0,2957	0,1227
27	http://www.jamaicatravel.ru/	-0,5000	0,9160	0,3274	-0,0056

Таблица 4. Расхождения оценок

Участник	Запрос	Контекст	R_sengine	R_profile	R_full_profile
oleg	пираты карибского моря	история	0,64	0,36	0,29
misha	пираты карибского моря	фильм	0,37	0,26	0,22
admin	пираты карибского моря	корабли	0,67	0,17	0,18
kate	проектирование музейных экспозиций		0,37	0,25	0,22
admin	самолеты первой мировой войны	история, конструкция	0,7	0,29	0,47
admin	велосипед	выбор и покупка	0,8	0,19	0,2
misha	велосипед	трюки	0,82	0,25	0,33
oleg	велосипед	история	0,77	0,38	0,3
oleg	интерфейс		0,59	0,4	0,35
kate	современная музейная экспозиция		0,34	0,24	0,24
oleg	сделать клетку для кроликов		0,58	0,28	0,25
misha	австрийская кухня рецепты		0,59	0,21	0,21
misha	мастер-класс гитара		0,5	0,18	0,17

полного окончательного варианта ИПП (R_full_profile).

Пример фрагмента запросной истории участника oleg по запросу «пираты карибского моря» с оценками представлен в таблице 3.

Величина расхождений в оценках между участником и различными системами вычислялась по формуле:

$$diff\left(R_{user}, R_{sengine}\right) = \frac{\sum_i \sqrt{\left(r_i^{user} - r_i^{sengine}\right)^2}}{2 \times n},$$

где r_i^{user} — оценка i-той интернет-страницы пользователем, $r_i^{sengine}$ — оценка i-той интернет-страницы поисковой системой без персонификации, $i = \overline{(1, n)}$, n — длина истории.

Значения функции diff нормированы от 0 (оценки совпадают) до 1 (оценки противоположны). Величины расхождений оценок между участником и системами представлены в таблице 4.

Таблица 4 показывает, что система с персонификацией для всех запросов показывает меньшее расхождение с оценками пользователей.

4. Выводы и обсуждения результатов

Результаты исследования подтверждают гипотезу о том, что применение системы, учитывающей ИПП, позволяет получать лучшие с точки зрения пользователя результаты, чем использование поисковых систем без персонификации.

Такой эффект достигается за счет обучения системы персонификации. Выделенный фрагмент таблицы 3 показывает, как после получения нескольких оцененных пользователем фрагментов, система приспосабливается к контексту пользователя и выдает более адекватные оценки, в то время как оценки интернет-страниц системой без персонификации естественно остаются неизменными.

Из графиков (рис. 2 – рис. 4) видно, что после первоначальных значительных колебаний, портрет становится достаточно стабильным. Стабилизация обычно происходит на 10–11 фрагменте. В дальнейшем, только относительно большие фрагменты могут существенно повлиять на состав

слов и соотношения их весов в ИПП.

Часто пользователи, задавая один и тот же запрос, могут подразумевать разные смысловые контексты. Таблицы 1 и 2 показывают, что предложенной системе удается выявлять эти индивидуальные различия. Значит, существует принципиальная возможность улучшения поисковых и навигационных систем с помощью применения ИПП.

Тем не менее, у предложенного метода построения ИПП, есть недостатки, которые происходят от использования в качестве основы ключевых слов. Требуется предоставить довольно много оцененных фрагментов для получения адекватных прогнозных оценок новых документов. В выделенном фрагменте таблицы 3 показано, что первоначально, исходя из предыдущего опыта, система негативно оценила документ, который понравился пользователю. Только после предоставления одиннадцати положительных примеров система дала оценку, близкую к оценке участника эксперимента. Это происходило из-за того, что интернет-страницы, даже имея общую тематику, могли существенно отличаться составом ключевых слов. В таких условиях простой метод ключевых слов не применим. Требуется учитывать априорные знания о семантическом поле каждого слова, например, используя синонимы, триггерные пары [3] или более совершенные методы.

Из-за излишней стабильности ИПП, новым понятиям сложно сразу получить высокие веса. Они начинают оказывать существенное влияние, только после того, как встретятся большое число раз, сравнимое со статистикой по уже имеющимся в портрете словам. Следует модифицировать алгоритм для обеспечения нужного баланса между стабильностью и гибкостью портрета. Например, введя функцию забывания.

ИПП можно передавать поисковой системе для уточнения контекста поиска. Пробные эксперименты, в ходе которых с помощью расширенного языка запросов первоначальный пользовательский запрос дополнялся несколькими наиболее значимыми словами из портрета, показали, что таким способом персональный поисковый агент может предоставлять пользователю улучшенные результаты поиска, используя уже существующие поисковые системы. Автоматическое изменение строки запроса является перспективным направлением для дальнейших исследований.

ИПП можно применять и для опережающей оценки интернет-документов, которые может посмотреть пользователь.

В качестве прототипа системы был опробован вариант, когда пользователь мог узнать оценки документов, выбрав команду «Получить оценки» на форме для оценки (рис. 1).

Система составляла список ссылок, расположенных на текущей странице и отправляла его для оценки на прокси-сервер. Там программа-оценщик profileRater вычисляла, насколько интернет-документ соответствует ИПП. Оценки документов, представленных на текущей странице ссылками, показывались во всплывающем окне, при наведении курсора на ссылку. Это позволяло пользователю принять решение о том, стоит ли переходить к данному документу или можно сэкономить время и посмотреть только документы с высокими оценками.

В ходе опроса, участники эксперимента высказали мнение, что система выставления оценок была интуитивно понятна и удобна в работе.

Таким образом, исследование показало, что, хотя вопрос об оптимальном методе построения ИПП остается открытым и требует дальнейшего изучения, применяя ИПП, можно существенно улучшить поисковые и навигационные системы.

5. Литература

[1] Шаров С. А. Частотный словарь. Вторая версия частотного списка // РосНИИ ИИ. [Электрон. ресурс]. – 2001. – Режим доступа: http://www.artint.ru/projects/frqlist.asp

[2] Chen C. PVA: A Self-Adaptive Personal View Agent / C. Chen, M. Chen, Y. Sun // Journal of Intelligent Information Systems. – 2002. – P. 173-194.

[3] Chen L. Web Mate: A Personal Agent for Browsing and Searching / L. Chen, K. Sycara // In proceedings of the 2nd International Conference on Autonomous Agents and MultiAgent Systems, AGENTS '98, ACM. – 1998. – P. 132-139.

[4] Marais J. Supporting cooperative and personal surfing with a desktop assistant / J. Marais, K. Bharat // UIST'97, ACM. – 1997.

[5] Qiu F. Automatic identification of user interest for personalized search / F. Qiu, J. Cho // WWW'06, ACM Press. – 2006.

[6] Salton G. Introduction to Modern Information Retrieval / G. Salton, M. J. McGill // McGraw-Hill, New York. – 1983.

[7] Trajkova J. Improving Ontology-Based User Profiles / J. Trajkova,// M.S. Thesis EECS, University of Kansas. – 2003.

Model of user informational profile for personalized search

Alex Shirokov

The article touches upon a problem of improving Internet search engines and navigators. Author makes a suggestion that using of user informational profile is a feasible solution. He presents a method of assembling and using such profile, based on a modification of TF-IDF algorithm. Experimental results show that the method works rather well. As a conclusion, author demonstrates that system with personalization achieves much better results than unpersonalized one.