Экономия времени как мера качества поисковой системы

© И.Е. Куралёнок[1], М.А. Скачков[2], О.В. Басков[2]

[1]Яндекс, г. [2]Санкт-Петербург

[2]Санкт-Петербургский государственный университет

solar@yandex-team.ru, skvmichael@yandex.ru, ov.japh@gmail.com

Аннотация

Данная статья описывает подход к оценке качества поисковых систем, основанный на времени удовлетворения пользователями их информационной потребности. Статья включает описание математической модели, результаты её применения к экспериментальным данным, имитирующим логи поисковой системы, а также сравнение полученных результатов с оценками поисковых систем, основанными на других пользовательских метриках.

1 Введение

Существует масса способов оценки качества поисковых машин. Все способы можно условно разделить на две категории. К первой категории отнесём способы оценки качества экспертами, вручную оценивающими релевантность того или иного результата выдачи, ко второй – способы оценки при помощи автоматизированных метрик, построенных на пользовательских логах поисковой системы. На сегодняшний день логи – ключевой инструмент принятия решения в индустриальном поиске. В качестве примера автоматизированных пользовательских метрик можно привести следующие:

- доля кликов на первый результат выдачи поисковой машины по всем запросам,
- средняя позиция первого клика,
- средняя глубина просмотра выдачи поисковой системы,
- средняя позиция последнего клика.

Встречаются также гибридные подходы, в основе которых лежит предсказание метки релевантности того или иного документа по поведению пользователей на этом документе или в процессе поиска [4, 5, 7]. Такие подходы имеют свои недостатки, поскольку достаточное количество статистики таким образом можно собрать только для определённого класса наиболее частотных пар «запрос – документ».

Труды 12[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010, Казань, Россия, 2010

Традиционным подходом оценки качества поисковой машины является Кренфилдский подход (Cranfield), основанный на сравнении результатов поиска с эталонным результатом. Более полная классификация способов оценки качества поисковых машин, а также подробное описание Кренфилдского подхода приведены в [1].

В нашем понимании одна из главных целей поисковой системы – экономия времени пользователей при поиске информации в интернете. Мы предлагаем подход к оценке, основанный на времени удовлетворения пользователями их информационной потребности, и хотим его сравнить с Кренфилдским подходом, а также с другими подходами, базирующимися на автоматизированных пользовательских метриках. Данный подход не попадает под вышеприведенную классификацию. У поисковой системы есть много других пользовательских свойств кроме времени поиска, но в данном исследовании мы их рассматривать не будем.

Работа организована следующим образом. В части 2 приводится описание эксперимента, который был поставлен для моделирования поведения пользователей поисковой системы. В части 3 описывается математическая модель предлагаемого нами подхода к оценке. Результаты применения этой модели к данным, полученным из проведённого эксперимента, приводятся в части 4. В частях 5 и 6 полученные выводы о качестве поисковых систем сравниваются с Кренфилдской оценкой и оценками, основанными на других пользовательских метриках.

2 Исходные данные

Основная цель исследования – узнать, как влияет поисковая система на время удовлетворения пользователем своих информационных потребностей. Определение этой степени влияния является сложной задачей. При её решении возникает несколько требующих внимания моментов:

- как смоделировать пользователя поисковой системы, чтобы обеспечить повтор?
- как измерять степень влияния поисковой системы на время поиска?

Мы поставили эксперимент следующим образом. В качестве участников эксперимента были взяты реальные люди. Задачи составлялись на основе репрезентативной выборки сессий пользователей одной

Рис. 1. Максимальный просмотренный результат поиска

из коммерческих поисковых систем. Для каждой сессии была сформулирована информационная потребность, которую пытался удовлетворить пользователь в этой сессии. Список таких информационных потребностей и был взят в качестве заданий. Участники эксперимента в непринуждённой домашней обстановке не спеша выполняли задания, сформулированные в виде одного предложения, например, «выяснить, сколько стоит тур в Париж на 10 дней». Свободно распоряжаясь своим временем, участники могли выполнять задания не сразу и периодически отвлекаться, позже возвращаясь к нему. Находя ресурсы, которые, по их мнению, полностью или частично удовлетворяли поставленному заданию, участники отмечали их специальным маркером. Кроме того, все совершаемые участниками действия (например, движения мыши, клики, прокрутка страницы, формулировка запросов) записывались для дальнейшего анализа в файлы специального формата. Единственным ограничением, наложенным на участников, было то, что они были обязаны проводить поиск только в указанной поисковой системе.

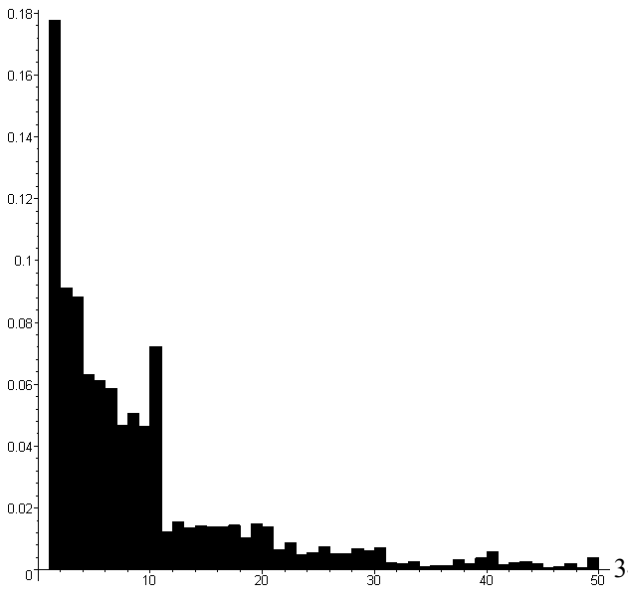
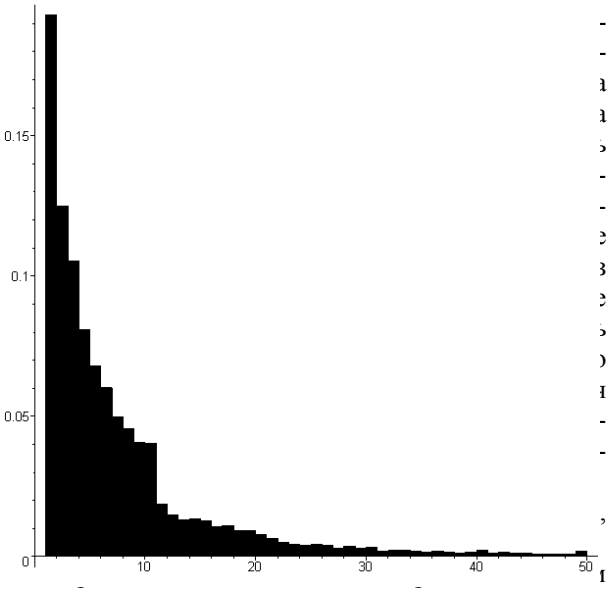


Рис. 2. Распределение кликов

мента получилось 4337 сессий, записанных в файлы по одной сессии на файл.

Полученные данные обрабатывались следую-



поисковой системы можно однозначно понять информационную потребность пользователя, что, вообще говоря, не всегда возможно;

- понятие об удовлетворении информационной потребности зависит от пользователя и может быть смоделировано выборкой участников, смещение которой не повлияет на результаты эксперимента.

Чтобы понять, похоже ли поведение участников эксперимента на поведение пользователей, были построены следующие графики:

- максимально просмотренный результат поиска (рис. 1);

- распределение кликов по результатам поиска (рис. 2);

- распределение запросов по времени (рис. 3).

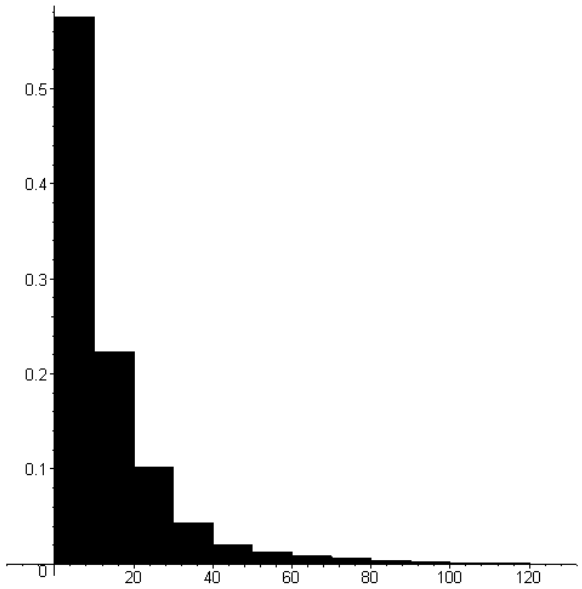


Рис. 3. Распределение всех запросов по времени (в абсолютных величинах)

Поскольку в выдаче поисковой системы Rambler на каждой странице присутствуют 15 результатов, а в поисковых системах Google и Yandex – по 10 результатов, данные статистики рассматривались только для заданий, выполненных в поисковых системах Google и Yandex.

Эти графики выглядят естественно и соответствуют ранее полученным результатам. Например, в исследовании [2], посвящённом вопросу о том, как люди просматривают страницу результатов выдачи поисковой системы, приводится распределение просмотров результатов, которое примерно соответствует полученному графику распределения кликов.

3 Модель

Чтобы измерить степень влияния поисковой системы на время выполнения задания, необходимо смоделировать процесс поиска. Это очень сложная задача. Упростим её, введя некоторые дополнительные предположения. Ясно, что время выполнения задания зависит от поисковой системы, пользователя, задания и от других аргументов, т. е. $t = f(s,u,c,...)$, где s – поисковая система, u – пользователь, c – задание.

В основе модели лежит предположение, что функцию времени можно разложить на независимые сомножители, т. к. нам кажется, что сложность задания и собственная скорость пользователя при поиске не зависит от поисковой системы:

$$t = f_1(s) f_2(u) f_3(c) f(...).$$

Функция $f(...)$ не зависит от пользователя, поисковой системы и задания и является свойством самого эксперимента. Поэтому будем полагать её константой. За неимением лучших предположений, функции $f_1(s), f_2(u), f_3(c)$ будем считать линейными, т. е. долю влияния поисковой системы, задания и пользователя на время поиска будем считать одинаковой.

Будем полагать, что время, которое пользователь тратит на нахождение информации по его заданию, прямо пропорционально сложности задания C и обратно пропорционально качеству поисковой системы S и коэффициенту C который характеризует скорость пользователя в поиске нужной информации с помощью поисковой системы (является свойством пользователя):

$$t \propto C/(SU). \quad (1)$$

Выразим отсюда коэффициент качества поисковой системы:

$$S \propto C/(tU).$$

По сделанным выше предположениям коэффициент пропорциональности является константой для любых значений S, C, U. Для упрощения задачи будем искать коэффициенты качества поисковых систем S не в чистом виде, а их соотношение для разных систем. Таким образом, коэффициент пропорциональности сократится. Поэтому для простоты вычислений будем считать его единицей изначально. Тогда формула (1) примет вид

$$t = C/(SU).$$

Для удобства возьмём натуральный логарифм

$$\ln t = \ln C - \ln S - \ln U.$$

Запишем это уравнение для каждой сессии, полагая, что их количество равно I,, количество заданий равно J, количество пользователей равно K, количество поисковых систем равно M:

$$\ln t_i = \ln C_j - \ln S_m - \ln U_k.$$

Имеем: $b_i = x_j - x_{j+k} - x_{j+k+m}$, $b = Ax$. Матрица A системы в каждой строке имеет 3 ненулевых значения: одну 1 и две -1. Очевидно, эта система имеет достаточно много строк и не имеет решения в чистом виде, поэтому будем искать ближайшее решение методом наименьших квадратов:

$$x^* = \arg\min(\|Ax - t\|).$$

Попарные соотношения экспонент от величин x^*_{j+k+m} дадут искомые результаты.

4 Результаты эксперимента

В ходе эксперимента описанная модель была применена к исходным данным следующим образом. В качестве меры для определения невязки была взята евклидова мера. В качестве минимизирующего алгоритма выбран метод сопряжённых градиентов из-за простоты вычисления градиента линейной функции. Полученные коэффициенты качества поисковых систем приведены в табл. 1.

Из-за предположений, допущенных в модели (мы опустили коэффициент пропорциональности), эти коэффициенты нельзя рассматривать как некоторую метрику и сравнивать их в чистом виде с по-

добными коэффициентами, полученными другими методами. Необходимо рассмотреть их соотношение. Для удобства выберем одну из поисковых систем, относительно которой будем нормировать полученные коэффициенты. Нами была выбрана поисковая система Rambler. Итоговые соотношения приведены в табл. 2.

Таблица 1. Значение коэффициентов качества поисковых систем

Поисковая система	$\ln S_k$	S_k
Google	-11.0840	$1.54*10^{-5}$
Yandex	-11.2117	$1.35*10^{-5}$
Rambler	-11.3528	$1.17*10^{-5}$

Таблица 2. Отношения коэффициентов качества поисковых систем

Google / Rambler	1.308
Yandex / Rambler	1.151
Rambler / Rambler	1.000

Из табл. 2 видно, что Google выступает лучше, чем Rambler, в смысле времени продолжительности сессии в 1.308 раза. Это означает, что в поисковой системе Google участники эксперимента в среднем быстрее решали поставленные задачи, чем в поисковой системе Rambler. Качество системы Yandex выше качества системы Rambler в 1.151 раза. Таким образом, на момент проведения эксперимента поисковую систему Google можно считать лучшей в смысле времени поиска поисковой системой из рассматриваемых, в то время как поисковая система Rambler выступает хуже всех рассматриваемых систем.

5 Связь с Кренфилдской оценкой

Поскольку участники эксперимента отмечали специальным маркером те ресурсы, которые, по их мнению, полностью или частично удовлетворяли поставленному заданию, можно говорить о Кренфилдских оценках. Сравним полученные результаты с оценками качества поисковых систем, основанными на одной из таких оценок – *mean reciprocal rank* (MRR). Напомним, что reciprocal rank (RR) определяется как величина, обратная позиции первого корректного документа (в нашем случае – первого документа, отмеченного как хороший), а MRR является средним значением RR по всем запросам. Таким образом, MRR является величиной, которой можно характеризовать поисковую систему. Значения MRR для рассматриваемых поисковых систем приведены в табл. 3.

Таблица 3. Значение mean reciprocal rank для поисковых систем

Поисковая система	MRR
Google	0.648
Yandex	0.514
Rambler	0.484

Поскольку наша модель не определяет коэффициенты качества поисковой системы напрямую, а лишь только их соотношения, вычислим отношения MRR для рассматриваемых систем, при этом нормирование проведём относительно поисковой системы Rambler.

Таблица 4. Отношения коэффициентов качества поисковых систем для модели на основе Кренфилдской статистики MRR

Google / Rambler	1.338
Yandex / Rambler	1.062
Rambler / Rambler	1.000

Эти результаты отличаются от результатов, полученных для модели на основе времени продолжительности сессии. При этом, если для поисковой системы Google отличия незначительны, то поисковую систему Yandex данная модель оценивает значительно ниже, чем модель на основе времени. Однако модель на основе MRR ранжирует поисковые системы так же, отдавая первое место поисковой системе Google, второе – Yandex и третье – Rambler.

6 Связь с другими пользовательскими метриками

Рассмотрим связь предложенной модели с моделями оценки качества поисковых систем на основе других метрик. Для исследования нами были выбраны 4 пользовательские метрики:

- средняя позиция первого клика;
- доля коротких сессий (короткой будем называть сессию, длившуюся менее 1 минуты);
- вероятность клика на первую ссылку в выдаче поисковой системы;
- средняя позиция последнего клика.

Значения этих метрик на данных нашего эксперимента приведены в табл. 5.

Аналогично случаю Кренфилдской оценки вычислим соотношения между полученными метриками, обращая внимание на зависимость качества поисковой системы от значения метрики (в случае метрик (2) и (3) зависимость прямая, в случае метрик (1) и (4) зависимость обратная). Результаты вычисления приведены в табл. 6.

Отсюда видно, что наиболее близкий к нашей модели оценки качества поисковой системы результат даёт метрика (1) – средняя позиция первого клика.

Таблица 5. Значения пользовательских метрик

		Google	**Yandex**	**Rambler**
1	**Средняя позиция первого клика**	1.838	2.165	2.434
2	**Доля коротких сессий (менее 1 минуты)**	0.181	0.168	0.154
3	**Вероятность клика на первую позицию**	0.559	0.435	0.422

Таблица 6. Соотношения качества поисковых систем для пользовательских метрик

	Средняя позиция первого клика	Доля коротких сессий	Вероятность клика на первую позицию	Средняя позиция последнего клика
Google / Rambler	1.324	1.175	1.325	1.327
Yandex / Rambler	1.124	1.091	1.031	1.217
Rambler / Rambler	1.000	1.000	1.000	1.000

7 Заключение и будущая работа

Мы ввели новую метрику качества поисковых систем, основанную на влиянии поисковой системы на одно из важнейших пользовательских свойств – время поиска необходимой информации. Мы исследовали вопрос связи полученного способа оценки поисковой системы с моделью, основанной на одной из Кренфилдских метрик – позиция первого просмотренного документа, отмеченного как удачный. При этом были получены результаты, дающие то же самое ранжирование поисковых систем по качеству, но отличающиеся по соотношению коэффициентов качества. Также были рассмотрены другие пользовательские метрики, из которых наиболее близкой к нашей модели является метрика, определяемая как средняя позиция первого клика на странице выдачи результатов поисковой системы.

Мы считаем недостаточным сравнение предложенной модели с Кренфилдскими метриками, в то время как данные метрики являются широко используемыми. В будущем мы планируем связать нашу модель с другими оценками, такими, как MAP [8, 9], nDCG [3], ERR [6].

Литература

[1] Кураленок И.Е., Некрестьянов И.С. Оценка систем текстового поиска// Программирование. – 2002. – Т. 28, №4. – С. 226-242.

[2] Google's Golden Triangle. Eye Tracking Study, 2005. – http://www.enquiroresearch.com/images/eyetracking2-sample.pdf.

[3] Yilmaz E., Kanoulas E., Aslam J.A. A simple and efficient sampling method for estimating AP and NDCG// SIGIR '08: Proc. of the 31st annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval.

[4] Cen R., Liu Y., Zhang M., Zhou Bo, Ru L., Ma S. Exploring relevance for clicks// CIKM '09: Proc. of the 18th ACM Conf. on Information and Knowledge Management.

[5] Dupret G., Liao C. A model to estimate intrinsic document relevance from the clickthrough logs of a web search engine// WSDM '10: Proc. of the third ACM Int. Conf. on Web Search and Data Mining.

[6] Chapelle O., Metlzer D., Zhang Ya, Grinspan P. Expected reciprocal rank for graded relevance// CIKM '09: Proc. of the 18th ACM Conf. on Information and Knowledge Management.

[7] Joachims T., Granka L.A., Pan B., Hembrooke H., Gay G. Accurately interpreting clickthrough data as implicit feedback// SIGIR '05. – P. 154-161.

[8] Buckley C., Voorhees E.M. Retrieval system evaluation// In E.M. Voorhees and D.K. Harman, editors, TREC: experiment and evaluation in information retrieval. – MITT Press, 2005.

[9] Turpin A., Scholer F. User performance versus precision measures for simple web search tasks// Proc. of the ACM SIGIR Int. Conf. on Research and Development in Information Retrieval, Seattle, WA, 2006. – P. 11–18.

Time saving as a quality measure of retrieval system

I.E. Kuralenok, M.A. Skachkov, O.V. Baskov

This paper describes an approach to assessing the quality of search engines based on time to satisfy users' information needs. The paper includes a description of the mathematical model, the results of its application to the experimental data simulating search engine logs and a comparison of the results with estimates of search engines based on other custom metrics.