Пользовательская и системная сложность данных

Кураленок И.Е.

Санкт-Петербургский
Государственный Университет
ik@intellij.com

Уткин А.В.

Санкт-Петербургский
Государственный Университет
alexey.outkine@mail.ru

Аннотация

В предлагаемой работе проверяется гипотеза о переносимости численных результатов оценки систем текстового поиска на данные, отличающиеся от тестового корпуса. Проверив эту гипотезу, мы строим понятие сложности данных, позволяющее с большей точностью переносить результаты от одного тестового корпуса к другому. Для исследования этого вопроса мы использовали данные РОМИП разных лет, которые в наибольшей степени приближены к реальным условиям тестирования систем, что позволяет говорить об обоснованности выводов сделанных в работе.

1 Введение

В области текстового поиска долгое время единственным достоверным результатом оценки являлся ранжированный по той или иной метрике список систем участниц [6,7]. Численные результаты исследования эффективности считались бесполезными в условиях других данных (непереносимыми).

Вопрос о переносимости и повторяемости результатов является ключевым в любой оценке. Однако в информационном поиске, в связи с зависимостью результатов тестирования от субъективного мнения эксперта эти проблемы не достаточно исследованы. До 1998-го года единственным аргументом в пользу переносимости результатов был «большой» объем тестовых данных и «разумность» эксперта. В связи с этим в области появилась недоверие к результатам оценки выразившееся в крайне низком количестве реализаций на практике новых методов, хорошо показавших себя в лабораторных условиях [5].

На базе данных TREC (Text REtrieval Conference) за последние годы было сделано несколько фундаментальных открытий в области оценки [6,11,12], что значительно повысило ее авторитет. В частности, были показаны: стабильность основных метрик оценки к изменению состава группы экспертов, повторяемость результатов на базе различных запросов и т.п.. Но, несмотря на это, наиболее интересный с точки зрения пользователя и разработчика вопрос - «а будет ли это эффективно работать на моих данных?» - остается нерешенным по сей день. Исследования NIST, позволяют с уверенностью сказать лишь о том, что та или иная система поиска окажется «не хуже» другой на ваших данных. Переносимость же численных результатов, которые позволяют дать ответ на вопрос «а стоит ли тратить из-за этой разницы силы на реализацию чего-то нового», считается невозможной [2,8].

Кроме информации о потенциальной эффективности того или иного подхода, результаты оценки могут помочь и в улучшении качества поиска. Очевидный способ их применения – выбор наиболее подходящего для конкретного запроса метода поиска. Однако на пути такого применения стоит проблема значительного отличия эффективности системы на разных заданиях. Для того чтобы понять природу такого различия мы провели анализ поведения источника оценки – пользователя. Результатом нашего исследования стало введение понятия сложности данных, которое позволит предсказать скачки эффективности.

В представляемой работе были использованы данные Российского семинара по Оценке Методов Информационного Поиска (РОМИП [1]) разных лет, которые в наибольшей степени приближены к реальным условиям тестирования систем.

В первой части работы мы исследуем вопрос принципиальной переносимости численных результатов оценки, рассмотрев сходимость их усреднения при расширении тестовой коллекции. Во второй части мы исследовали соотношение объективных характеристик поведения пользователя и оценки систем в зависимости от поставленной задачи поиска. И, наконец, в заключении наметим некоторые выводы и возможные направления для дальнейшего развития этой темы.

2 Экспериментальные данные

Для исследования зависимости изменения результатов оценки при изменении тестовой коллекции мы использовали данные первого Российского семинара по Оценке Методов Информационного Поиска (РОМИП). Инициатива РОМИП [1] состоит в регулярном проведении семинаров, посвященных оценке качества русскоязычного текстового поиска. Методология

проведения РОМИП основывается на зарубежном опыте проведения подобных мероприятий – TREC, CLEF и т.п. и использует «лабораторную» парадигму [3, 10]. Такой выбор данных связан с рядом их достоинств:

- Доступность полных данных по оценке РОМИП, в частности детали работы экспертов, таких как время принятия каждого решения;
- В РОМИП принимают участие реальные поисковые системы, построенные на основе разных моделей поиска;
- запросы, использованные в тестировании, взяты из логов реальной поисковой системы.

Приближенность к реальности – наиболее важная составляющая в предлагаемом исследовании. Два последних из перечисленных факторов позволяют говорить о такой приближенности.

Так как вторая часть предлагаемого исследования посвящена анализу работы эксперта, рассмотрим процедуру оценки более подробно. В нашем исследовании использовались дорожки по поиску по Веб-коллекции (Web) и коллекции нормативных документов (Legal). Веб-коллекция содержала более 728 000 документов, общим объемом более 7 Гб. Коллекция нормативных документов содержала 61 000 документов, общим объемом 1,6 Гб. Для оценки поисковых систем использовалось следующее количество запросов: Web 2003 – 54, Web 2004 – 67, Web 2005 – 75, Legal 2004 – 91, Legal 2005 – 83. Оценка производилась по методу «общего котла» [2, 4].

Оценку каждой коллекции проводили от 5 до 13 асессоров. Документы, выданные поисковыми системами, объединялись в общий пул. Задача асессора состояла в том, чтобы оценить соответствие каждого документа из пула запросу. Перед началом работы каждый эксперт должен был ознакомиться с единой инструкцией по порядку проведения оценки. Каждый запрос оценивали не менее двух асессоров, независимо друг от друга [1].

Оценка проводилась с использованием специально разработанной для этого рабочей среды. Особенностью этой среды [1] было то, что сохранялась информация о каждом действии пользователя, включая информацию о моменте совершения этого действия. Это позволило нам определить время, потраченное асессором на принятие каждого решения.

Полученные результаты экспертной оценки использовались для построения стандартных метрик области [1,6] в частности:

Точность (Precision) – доля релевантных документов в ответе системы.

$$P = \frac{relevant}{total}$$

Точность на уровне α - P_α (Precision at cut-off level) – точность ответа, содержащего первые α документов.

R-Точность (R-Precision) – точность на уровне числа всех известных релевантных документов для этого запроса.

$$R_p = P_{total\_relevant}$$

Средняя точность (Average Precision) – Для каждого релевантного документа d вычисляется точность на уровне $level$ равному порядковому номеру документа в ответе. Эти значения усредняются по общему количеству релевантных документов для данного запроса.

$$averageP = \sum_{level:d_{level} \in relevant} P_{level}$$

11-точечный график по методике TREC (11-point TREC) - 11-точечный график полноты/точности отражает изменение точности в зависимости от требований к полноте и дает более полную информацию, чем единая метрика в виде одной цифры. По оси x на графике откладываются значения полноты, по оси y – значение точности при условии, что рассматривается начальный отрезок результатов запроса, на котором достигается заданный уровень полноты. Для запроса, для которого известно n релевантных документов, полнота может принимать дискретные значения 0, 1/n, 2/n, ... , 1. Для того, чтобы можно было получать единый график полноты/точности для множества запросов

- рассматриваются фиксированные значения полноты 0.0, 0.1, 0.2, ... , 1.0 (всего 11 значений);
- используется специальная процедура интерполяции точности для данных фиксированных значений полноты;
- для множества запросов производится усреднение точности для заданных уровней полноты.

Интерполированное значение точности равно максимальному значению точности при уровне полноты большем или равным заданному.

3 Исследование переносимости усредненных результатов оценки

В предлагаемом эксперименте мы исследуем стабильность усредненных результатов оценки к изменению набора запросов. Если численные результаты окажутся стабильными при увеличении множества запросов, можно говорить о том, что при достаточном наборе запросов численные результаты будут повторяться независимо от состава этого

набора, что и является главным признаком переносимости.

В качестве исследуемой характеристики был выбран график зависимости точности от полноты по версии TREC, как наиболее известная и полная характеристика работы поисковой системы на всех уровнях глубины ответа.

В работе [6] показывается, что важнейшую роль в стабильности оценки играет количество заданий, использованных для их вычисления. Поэтому мы исследовали зависимость изменения усредненных графиков от количества запросов в использованном для их вычисления тестовом наборе. Для получения достаточного количества наборов был использован следующий подход:

- были выбраны по 100 наборов запросов мощности от 5 до 45 с шагом 5;
- каждый набор случайно выбирался из множества подмножеств доступных запросов (54 запроса дорожки поиска РОМИП'2003) заданной мощности в условиях равномерного распределения вероятности выбора между всеми возможными вариантами.

Далее для выбранных наборов строились графики TREC. Затем по наборам с равным количеством запросов строились усредненные графики и считались среднеквадратические отклонения от этого графика во всех 11 точках. Для всех этих точек вычислялось среднее значение с учетом отношения отклонения к абсолютной величине на усредненном графике. Полученные значения усредненного отклонения составили искомый график зависимости отклонения от мощности набора запросов.

Однако убывания абсолютных величин этого графика оказывается недостаточно для однозначного ответа на вопрос о сходимости рассмотренных характеристик. Этот факт связан с тем, что для исследования разброса результатов были использованы случайные наборы запросов из одного и того же ограниченного множества, что предполагает наличие пересечения в наборах запросов. Таким образом, при увеличении количества запросов так же увеличивается и ожидание пересечения этих выборок, что приводит к методическому уменьшению разброса конечных характеристик. В связи с этим, для исследования вопроса сходимости графиков необходимо так же построить и график теоретического уменьшения величины разброса характеристик.

С целью построения этого графика была рассмотрена следующая модель:

Пусть есть набор значений $\{x_i\}_{i=1}^{n}$, *из этих значений произвольным образом выбираются* k, *причем все выборки равновероятны. Необходимо найти среднеквадратическое отклонение случайной величины* $\xi = \frac{1}{k}\sum_{i=1}^{k} x^i$.

Для решения поставленной задачи найдем математическое ожидание ξ:

$$\mu(\xi) = \mu\left(\frac{1}{k}\sum_{i=1}^{k} x^i\right) = \frac{1}{k}\sum_{i=1}^{k}\mu(x^i) = \frac{1}{n}\sum_{i=1}^{n} x^i \qquad (1)$$

и дисперсию этой величины:

$$\begin{aligned} &D(\xi) = \mu(\xi - \mu(\xi))^2 = \\ &\mu(\xi^2 - 2\xi\mu(\xi) + \mu(\xi)^2) \\ &= \mu(\xi^2) - \mu^2(\xi) \\ &= \mu\left(\frac{1}{k^2}\left(\sum_{i=0}^{k} x^{i^2} + \sum_{i=0}^{k}\sum_{j=0, j\neq i}^{k} x^{i^2}x^{j^2}\right)\right) - \mu^2(\xi) \\ &= \frac{1}{k^2}\frac{C_{n-1}^{k-1}}{C_n^k}\sum_{i=1}^{n} x_i^2 + \frac{1}{k^2}\frac{C_{n-2}^{k-2}}{C_n^k}\sum_{i=1}^{n} x_i \sum_{i=1}^{n}\left(\frac{1}{n}\sum_{j=1}^{n} x_j - x_i\right) - \mu^2(\xi) \\ &= \left(1 - \frac{k-1}{n-1}\right)\frac{1}{k}\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) + \frac{k-1}{k}\frac{n}{n-1}\mu^2(\xi) - \mu^2(\xi) \end{aligned} \qquad (2)$$

можно показать, что:

$$\left(1 - \frac{k-1}{n-1}\right)\frac{1}{k} \equiv \left(1 - \frac{n}{n-1}\frac{k-1}{k}\right) \qquad (3)$$

таким образом:

$$\begin{aligned} D(\xi) &= \left(1 - \frac{n}{n-1}\frac{k-1}{k}\right)\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \mu^2(\xi)\right) \\ &= \left(1 - \frac{n}{n-1}\frac{k-1}{k}\right)A \end{aligned} \qquad (4)$$

Полученное выражение является зависимостью теоретического значения дисперсии характеристики от количества использованных для ее вычисления запросов. Для преобразования полученной дисперсии в среднеквадратическое отклонение достаточно извлечения корня. Если все выборки равновероятны, то зависимость среднего отклонения любой их серии принадлежит семейству (4). В нашем случае мы хотим показать, что наблюдаемое среднее отклонение стремится к нулю быстрее теоретического. Для этого достаточно показать, что экспериментальная зависимость скользит по семейству (4) с уменьшением коэффициента A, который однозначно определяется каждой точкой экспериментального графика. График зависимости коэффициента семейства от количества заданий можно видеть на рис. 1.

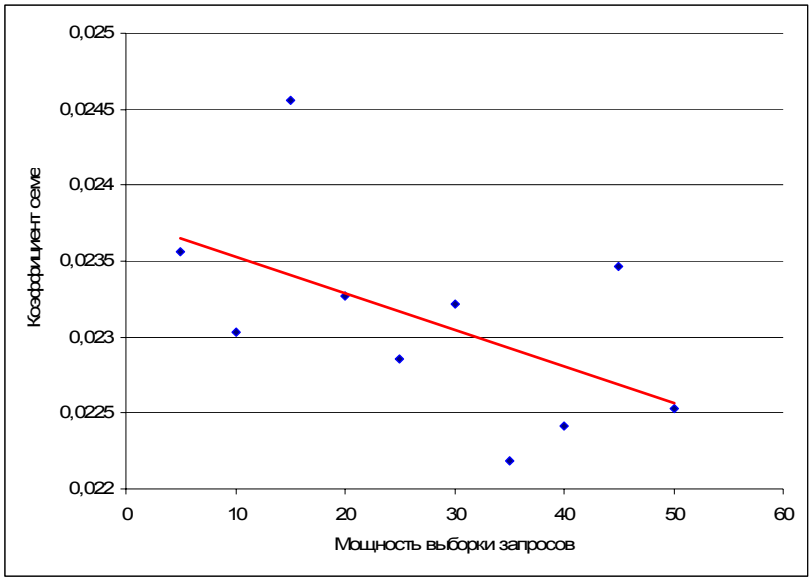

Рис. 1 Зависимость коэффициента семейства (4) от мощности набора запросов

Для большей наглядности результатов на графике приведена прямая наименьшей суммы квадратов отклонения, из которой явно следует, что отклонение наблюдаемых значений сходится быстрее теоретического, что позволяет говорить об общей сходимости усредненных результатов оценки при расширении множества запросов. Что в свою очередь опровергает тезис о невозможности переноса численных характеристик оценки на другие данные.

4 Оценка сложности данных

На данный момент в области информационного поиска не существует четкого понимания того, что такое сложность данных, как можно оценить эту характеристику и как можно ее применять при решении задач информационного поиска.

Однако известно несколько фактов, которые можно связать со сложностью данных. Известно, что на определенные запросы поисковая система отвечает лучше, чем на другие. Иными словами, некоторые запросы являются менее, а другие - более сложными для поисковой системы. Благодаря экспериментам, проводимым в рамках таких инициатив, как TREC, CLEF, NTCIR и РОМИП [1,3,7] были отработаны методики оценки эффективности работы поисковых систем. Таким образом, мы можем оценить системную сложность данных (сложность данных для поисковых систем), используя общеизвестные метрики эффективности.

С другой стороны, известно, что в зависимости от сложности данных пользователю требуется разное время, чтобы понять, соответствует ли найденный документ запросу или нет. Это позволяет нам оценить пользовательскую сложность данных.

Таким образом, мы имеем два взгляда на такую характеристику, как сложность данных. Первый взгляд – с точки зрения системы (системная сложность данных), второй – с точки зрения пользователя (пользовательская сложность данных). В рамках нашего исследования мы пытались выявить связь между системной и пользовательской сложностью данных. Наличие такой связи позволит говорить о единой сложности данных и применять эту сложность для получения наиболее вероятных результатов эффективности той или иной методики на конкретном запросе.

В нашем исследовании, для оценки системной сложности данных мы использовали метрики приведенные в п.2. В качестве оценки пользовательской сложности данных было принято среднее время, которое понадобилось асессору для принятия решения о соответствии документа запросу. Эта величина требовала поправок, сглаживающих субъективность асессоров [3]. Во-первых, было установлено, что оценку «не релевантен» асессор ставит в среднем в два раза быстрее, чем оценку «релевантен». Во-вторых – асессоры имеют разную скорость работы.

5 Отношение между пользовательской и системной сложностью данных

Для исследования связи между пользовательской и системной сложностью данных мы вычислили коэффициент корреляции между значениями системных метрик (точности, точности на уровне 5, точности на уровне 10, r-точности, средней точности) и значениями метрики пользовательской сложности (средневзвешенное время принятия решения асессором).

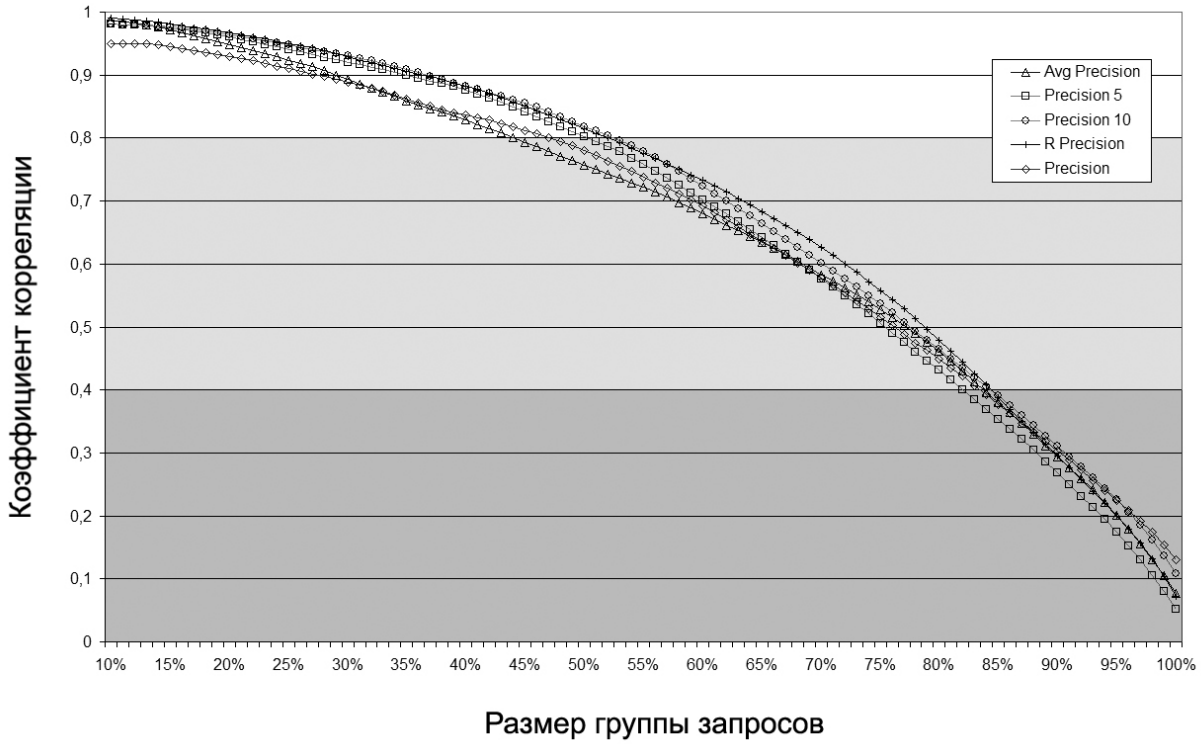

Рис.2 Коэффициент корреляции системной и пользовательской сложности данных, вычисленный по объединению дорожек, по различным системным метрикам.

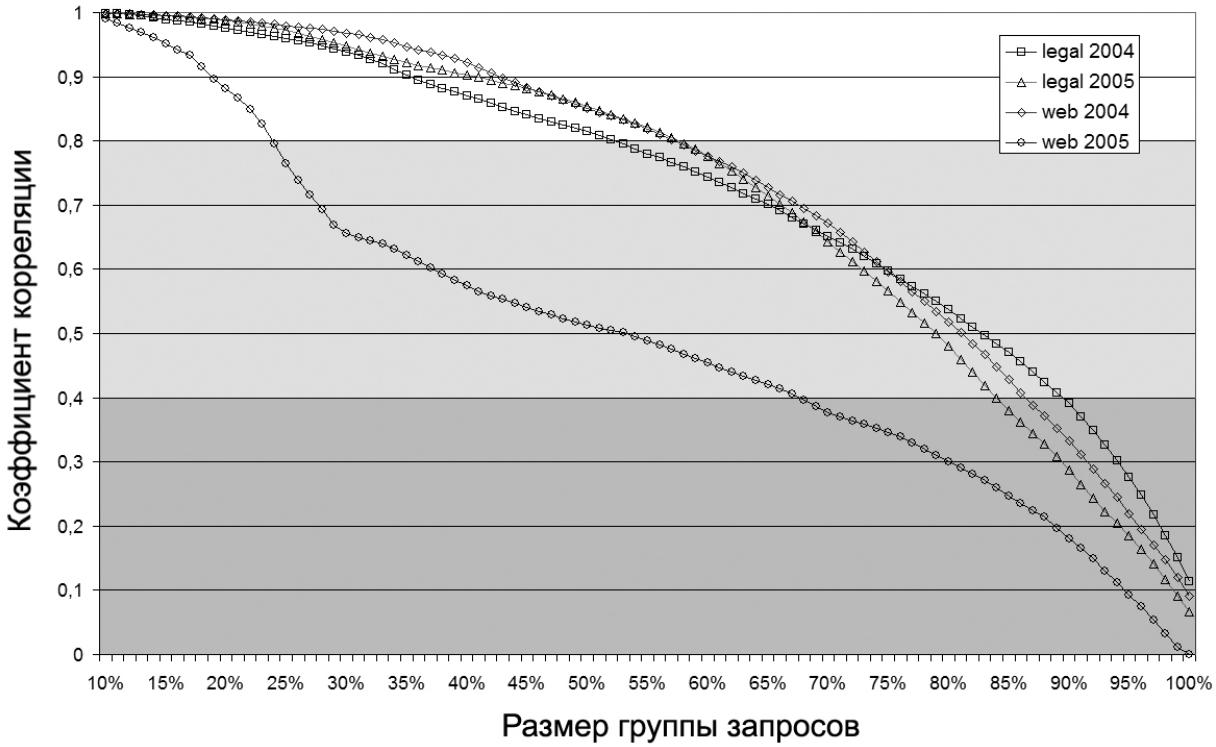

Рис.3 Коэффициент корреляции системной сложности данных (метрика - средняя точность) и пользовательской сложности данных, вычисленный по отдельным дорожкам.

Коэффициенты корреляции вычислялись по группам запросов разного размера. На каждом шаге находилась оптимальная группа запросов заданного размера, дающая максимальный коэффициент корреляции.

На рис. 2 изображена зависимость коэффициентов корреляции от размера группы запросов (выраженного в процентах от общего количества запросов в дорожке). Коэффициент корреляции вычислялся по различным системным

метрикам по объединению дорожек Legal 2004, Legal 2005, Web 2004, Web 2005, имеющему в сумме 316 запросов.

Было установлено, что приблизительно для 80-85% запросов наблюдается корреляция системной и пользовательской сложности данных (коэффициент корреляции > 0,4), а для 45-55% запросов корреляция является сильной (коэффициент корреляции > 0,8). Как видно из графика, данная закономерность имеет место для всех системных метрик.

Из рис.3 видно, что закономерность сохраняется и для каждой коллекции в отдельности, несмотря на то, что коллекции нормативных документов, в отличие от веб-коллекций, принадлежат к узкой тематике. Также интересен тот факт, что размер группы запросов, дающей определенный уровень корреляции в объединении дорожек приблизительно равен сумме размеров групп с таким уровнем корреляции в каждой дорожке по отдельности. Это говорит о масштабируемости выявленной закономерности.

6 Анализ результатов

Интересен вопрос, почему коэффициент корреляции убывает с увеличением размера группы запросов. Из графиков видно, что наиболее сильно коэффициент корреляции убывает при увеличении размера группы запросов с 80% до 100% от запросов дорожки. Иными словами 12%-20% запросов показывают обратную связь пользовательской и системной сложности данных. У этого может быть несколько причин. Во-первых, существуют запросы, под которые «заточены» некоторые поисковые системы – например, запросы определенного типа или относящиеся к определенной тематике. При этом сложность данных для пользователя на таких запросах остается высокой. Во-вторых, сами пользователи могут быстрее принимать решение о релевантности документа, например, если они лучше знакомы с определенной тематикой.

Из рис.3 видно, что на дорожке Web 2005 связь между пользовательской и системной сложностью данных выражена хуже, чем на других дорожках. Это может быть вызвано тем, что дорожка Web 2005 имела очень большое пересечение запросов с дорожками РОМИП Web 2003 и Web 2004. И в качестве асессоров привлекались одни и те же люди. Таким образом, асессоры быстрее принимали решения, что и повлияло на корреляцию. Это подтверждается сравнением среднего времени принятия решения асессором при оценке дорожки Web 2004 и Web 2005 – оценки 2005 году ставились в среднем на 30% быстрее.

7 Заключение

В представленной работе мы затронули две проблемы оценки систем текстового поиска: вопрос о переносимости усредненных результатов на отличные от оригинальных данные, и проблема значительного отличия эффективности системы на разных заданиях.

Результаты проведенных экспериментов позволяют сделать вывод, что численные результаты могут быть перенесены на другие данные. Так же полученные результаты заставляют обратить более пристальное внимание на численные характеристики оценки и выдвинуть дополнительные требования к вводимым метрикам.

Нам представляется, что переносимость характеристик оценки зависит не только, и не столько от тестируемой системы, сколько от соотношения характеристик данных, на которых производился контрольный запуск и тех, на которых мы хотим оценить производительность. Таким образом, для вычисления точных характеристик работы системы возможно достаточно знаний свойств данных, на которых производится запуск и знание полной информации о тестировании в рамках каких либо других данных.

В качестве упомянутых характеристик данных может выступать введенные нами понятия сложности данных. В рамках работы мы ввели понятия пользовательской и системной сложности данных, основанных на наиболее распространенных метриках. В результате экспериментов проведенных на данных РОМИП разных лет мы отметили неизменную корреляцию пользовательской и системной сложности на подавляющем большинстве запросов. Более того, было отмечено, что процентное соотношение «хороших» и «плохих» запросов сохраняется, как в рамках одной дорожки в разные годы, так и при различных формулировках исходной задачи (в случае разных дорожек). Последнее наблюдение и высокое качество экспериментальных данных позволяет говорить об универсальности выявленной закономерности и возможности введения общего понятии сложности данных для поиска.

Важность выявленной закономерности подчеркивается тем, что нам удалось статистически связать два абсолютно разных понятия: пользовательскую сложность, основанную на способности человека к восприятию текстовых данных и их анализу, и системную сложность, базирующуюся на оценке качества автоматического анализа тех же текстов с точки зрения асессора. Обе эти характеристики по-разному связаны с пользователем (в нашем случае, исполняющим роль оценщика), однако предположить столь сильную связность этих характеристик было сложно.

Несмотря на длинную историю развития, область оценки систем текстового поиска остается одной из наименее исследованных областей прикладной математики. Представленная работа предлагает новый взгляд на область применения и цели оценки. Полученные результаты говорят о перспективах предложенного подхода.

Литература

[1] И. Кураленок, И. Некрестьянов, Е. Павлова. «РОМИП 2003: Опыт организации.» Труды РОМИП-2003, под ред. Некрестьянова И., СПб, Россия, октябрь 2003 http://romip.narod.ru

[2] Кураленок И. Некрестьянов И. «Оценка систем текстового поиска». Программирование, Москва, Россия, июнь 2002

[3] И. Некрестьянов, М. Некрестьянова, А. Нозик. «Анализ “лабораторной” парадигмы оценки систем поиска.» http://company.yandex.ru

[4] И. Некрестьянов, М. Некрестьянова, А. Нозик. «К вопросу об эффективности метода “общего котла”». RCDL 2005.

[5] И.В. Сегалович «Как работают поисковые системы» Мир Internet. 2002, #10 http://www.dialog-21.ru/direction_fulltext.asp?dir_id=15539

[6] C. Buckley, E. M. Voorhees. «Evaluating evaluation measure stability.» In Proc. Of the SIGIR’2000 p. 33-40, 2002.

[7] Harman D. «What we have learned, and not learned, from trec». In Proc. of the BCS IRSG'2000, pages 2-20, 2000.

[8] C. Van Rijsbergen. «Foundation of evaluation». Journal of Documentation, 4(30):pages 365-373, 1974.

[9] K. Sparc Jones, editor. «Infromation Retrieval Experiment». Butterworth, London, 1981.

[10] E. M. Voorhees. «The philosophy of Information.» Revised Papers from the Second Workshop Language Evaluation Forum on Evaluation of Cross-formation Retrieval Systems, pp. 355 – 370, 2001.

[11] E. Voorhees. «Variations in relevance judgments and the measuremennt of retrieval effectiveness.» In Proc. of the SIGIR'98, pages 315-323, August 1998.

[12] J. Zobel. «How reliable are large-scale information retrieval experiments? » In Proc. of the SIGIR'98, pages 308-315, August 1998.

System and User Data Complexity

I. E. Kuralenok, A. V. Utkin

The aim of this work is to examine complexity of data as the data characteristic which could be used in solving information retrieval problems. In this research, we investigate the problem of data complexity evaluation.