А.В. Добров

ТЕХНОЛОГИИ ИНТЕЛЛЕКТУАЛЬНОГО ПОИСКА И СПОСОБЫ ОЦЕНКИ ИХ ЭФФЕКТИВНОСТИ

A.V. Dobrov

TECHNONOGIES OF INTELLECTUAL INFORMATION RETRIEVAL AND TECHNIQUES EVALUATING THEIR EFFECTIVENES

В данной статье осуществляется попытка изучить существующие теории и реализации интеллектуальных информационно-поисковых систем (ИИПС) и выделить набор факторов, определяющих их эффективность. Рассматриваются принятые на сегодняшний день методики оценки эффективности информационного поиска. На основании различных реализаций идеи ИИПС предлагается комплексный лингвистический подход к выявлению факторов, влияющих на эффективность решения этой задачи, и методика оценки релевантности в применении к ИИПС, основанная на концептуальном рейтинговании и авторубрикации.

In this article, an attempt is made to examine the existing theories and implementations of intellectual systems of information retrieval (ISIR), and to determine the factors that affect their effectiveness. The present conventional techniques of evaluation of the information retrieval effectiveness are discussed. A complex linguistic approach to determining the factors that affect the effectiveness of solutions of this problem, and a technique to evaluate the relevance, based on the conceptual rating and automatic classification, are proposed on the basis of the the different implementations of ISIR.

Ключевые слова: информационный поиск, интеллектуальный поиск, семантический поиск, релевантность, точность, полнота, эффективность

Keywords: information retrieval, intellectual search, semantic search, relevance,

precision, recall

Термин «информационный поиск» (ИП) (англ. Information retrieval) был введён К. Муром (Calvin Mooers) в 1948 году в его докторской диссертации. Согласно Е. Гарфилду, в «Zator Technical Bulettin», основанной К. Муром в 1947 году, первым опубликованным определением считается следующее: ИП — это «...поиск информации, местонахождение которой или само ее существование a priori неизвестно...»[1] («…finding information whose location or very existence is a priori unknown…» [Garfield 1997: 1]). В более поздних публикациях К. Мур его уточнял, устанавливая основные аспекты области ИП — «интеллектуальные аспекты описания информации и ее детализации для поиска, а также любые системы, методики или машины, применяемые для выполнения этой операции» («the intellectual aspects of the description of information and its specification for search, and also whatever systems, technique, or machines that are employed to carry out the operation» [Mooers 1951: 25]).

К области ИП были изначально отнесены именно интеллектуальные аспекты обработки информации. Тем не менее, первые информационно-поисковые системы (далее — ИПС), создавались для обеспечения доступа к научно-исследовательским документам, а не для их интеллектуальной обработки. Интеллектуальные аспекты области ИП оставались теоретическими, а конкретные ИПС стали обходиться без них.

С появлением сети Интернет ИПС получили широкое распространение, а с развитием ИПС эта сеть становилась все более доступной. На сегодняшний день ИПС — это основное средство доступа к информации во «всемирной паутине» и, в связи с ее существенной ролью в области СМИ, — одно из основных средств доступа к информации вообще. «Развитие Интернет начиналось как средство общения и удаленного доступа (электронная почта, telnet, FTP). Но постепенно эта сеть превратилась в средство массовой информации …». [Семенов 2009, 4.5.14

[1] Здесь и далее перевод цитат из англоязычных источников — наш (А. Д.)

Современные поисковые системы]

С развитием ИПС за ними стала закрепляться также и другая роль: современные ИПС предназначены не только для поиска, но и для хранения информации. На эту функцию ИПС указывают определения, приводимые в различных источниках, причем иногда — в первую очередь на нее. Например, В. П. Захаров определяет ИПС следующим образом: «Информационно-поисковая система (ИПС) — это упорядоченная совокупность документов (массивов документов) и информационных технологий, предназначенных для хранения и поиска информации — текстов (документов) или данных (фактов). Информационно-поисковыми системами являются любые определенным образом организованные хранилища информации. Причем информационно-поисковые системы могут быть и неавтоматизированными. Главное — это целевая функция: хранение и поиск информации». [Захаров 2005: 1]

На сегодняшний день объемы информации существенно возрасли, и особую актуальность приобретают лингвистические технологии, позволяющие находить результаты, релевантные запросам с точки зрения семантики. Качество ИПС все в большей степени зависит от того, насколько обоснованы с точки зрения содержания запросов выдаваемые результаты. Теория информационного поиска с самого начала была тесно связана с лингвистикой. Еще до появления диссертации К. Мура, В. Бушу казалось очевидным, что поиск информации должен быть организован на основе автоматического извлечения содержания текстов. С развитием ИПС разработчики стали склоняться к моделям, лишенным лингвистических компонентов. Основные известные современные ИПС не осуществляют синтаксический анализ и не выявляют содержание документа и запроса. Релевантность выдаваемых результатов вычисляется безотносительно к содержанию запроса, и предпочтение отдается наиболее популярным ресурсам (например, методика Google PageRank), или в первую очередь выдаются рекламные ссылки. В связи с этим растет количество проектов, направленных на разработку ИПС, которые учитывали бы языковое содержание документов и запросов (Powerset, Wolfram и т.д.). Такие ИПС называют «интеллектуальными» (далее —

ИИПС).

В данной статье осуществляется попытка изучить существующие реализации ИИПС и выделить факторы, влияющие на их эффективность.

Предыстория информационно-поисковых систем

В конце 1940-х годов американские ВС столкнулись с массивом немецких военных научно-исследовательских документов, который им не удавалось обработать. Эта проблема дала развитие идеям В. Буша, определившим многие направления области ИП, в частности — идее хранения и поиска информации по принципам человеческого интеллекта.

Статья Буша «Как мы можем думать» («As We May Think» [Bush 1945]) была опубликована в 1945 году в журнале «The Atlantic Monthly». Буш указал на то, что «лабиринт человеческого опыта усложняется чудовищными темпами, а средства, которыми мы пользуемся, чтобы пробраться через него к цели сиюминутной важности, — все те же, которыми пользовались в эпоху парусных судов» («The summation of human experience is being expanded at a prodigious rate, and the means we use for threading through the consequent maze to the momentarily important item is the same as was used in the days of square-rigged ships» [Bush 1945: 2]). Буш считал, что «если некий источник информации оказался ценным для сознания человека, то мы должны, насколько только можем, организовать его так, чтобы он отражал то, как это сознание работает»: «... if the data source was to be useful to the human mind we should have it represent how the mind works to the best of our abilities» [Wall 2009]. Буш призывал ученых создать единую базу знаний всего человечества, предназначенную для хранения и предоставления информации и построенную на принципе мыслительных связей. Эту систему он назвал «Memex», указывая на основной для этой системы новый вид индекса — индекс, организованный по принципам человеческой памяти (memory index).

Возникновение области информационного поиска

В 1950-ых годах в США возрастало беспокойство на фоне «гонки вооружений» с СССР. Оно дало рост множеству научных разработок и

мотивировало создание механизированных систем поиска литературных источников (А. Кент ([Kent 1962]) и др.) и изобретение индексов цитирования (Е. Гарфилд). В 1951 году в MIT[2] был проведен первый опыт компьютерного документального поиска («Probably the earliest experiment in computerized document retrieval was Bagley, in 1951, at the Massachusets Institute of Technology» [Doyle et al. 1975: 262]). В 1955 году А. Кент и его коллеги ввели и описали меры точности и полноты и предложили «систему оценки» ИПС, включающую в себя методы статистической выборки для оценки числа не найденных релевантных документов [Kent et al. 1955]. Эта система применяется до сих пор, но само по себе понятие «релевантность» неоднократно пересматривалось, и в данной статье будет осуществлена попытка модифицировать данную систему в применении к ИИПС.

В 60-е годы началась работа Дж. Сэлтона (Gerard Salton). Под его началом была разработана ИПС «SMART» (Salton's Magic Automatic Retriever of Text — «волшебный автоматический текстовый поисковик Сэлтона»). В этой системе были введены многие лингвистические технологии. Сэлтон является автором книги «Теория индексации» («A Theory of Indexing» [Salton 1987]), на положениях которой до сих пор строится работа большинства ИПС. В [Сэлтон 1973] излагаются данные «Теории индексации», на основе которых «рассматривается структура информации и методы … построения словарей; … методы анализа содержания, основанные на использовании … словарей, а также на статистических свойствах текстов и методах синтаксического анализа; способы оценки, разработанные для проверки эффективности многих из предлагаемых методов» [Сэлтон 1973: 14, 15]. В этой работе демонстрируется ключевая роль проблематики лингвистических технологий: поиск релевантных запросу документов «затрудняется отсутствием решения многих важных теоретических вопросов: Что именно представляет смысл и основное содержание документа? Какие лингвистические приемы используются для передачи основного содержания? ... Как можно выделить единицы контекстуальных отношений, если они существуют? и т.д.» [там же: 13,14]. В

[2] Массачусетский Технологический Институт (Massachusetts Institute of Technology)

работе определяется степень взаимосвязи между тем, в какой мере применяется та или иная технология, и тем, как это влияет на эффективность поиска.

Поисковые машины новейшего периода и интеллектуальный поиск

В 1990 году появилась ИПС «Арчи» — «первая поисковая машина» («The first search engine» [Li 2002]), работавшая через FTP. Арчи называют «дедушкой всех поисковых машин» («the grandfather of all search engines» [Baker 2009]). Он впервые позволил пользователям находить файлы по названиям в сети Интернет. Он включал в себя «систему сбора информации» («data gatherer» [Li 2002]), «систему поиска по регулярным выражениям для обнаружения имен файлов» («regular expression matcher for retrieving file names» [там же]) и поисковые индексы, хранившие списки, полученные системой сбора информации. В качестве результата поиска выдается «список FTP-адресов файлов или каталогов, соответствующих критериям отбора, указывается размер файлов, дата последней модификации и имя каталога, где этот файл лежит» [Семенов 2009: 4.5.13]. Поисковый запрос может интерпретироваться как подстрока имени файла, как само имя, или как регулярное выражение для поиска его подстроки. Поиск осуществлется только по именам файлов, и критерии смыслового соответствия основываются на простейших строковых подстановках.

В 1994 году Д. Янг и Д. Фило создали Yahoo, а Lycos представил ИПС, предлагающую ссылки на темы, связанные с поисковым запросом — это была первая web-система с элементами интеллектуального поиска, разработанная доктором М. Молдином.

ИПС AltaVista начала работать в 1995 году. Эта ИПС впервые предложила расширенную систему поиска и принимала языковые запросы на «естественном языке» — например, могла обработать запрос «*Как пройти в библиотеку?*», вместо «*библиотека*». Синтаксический анализ она не осуществляла, но иногда эффективно имитировала.

В 1997 году Сергеем Брином и Л. Пейджем в Стэнфордском Университете была запущена ИПС Google. История Google подробно описана в [Вайз, Малсид

2007]. Из технологий интеллектуального поиска Google реализует только ограниченный поиск по синонимам на английском языке: «Если нужно найти не только сами слова из запроса, но и их синонимы, поставьте тильду ("~") непосредственно перед словом» [Справка Google 2009]. Существует множество курьезов, связанных с недостатком его лингвистической базы. Например, на запрос «*~president*» система выдает множество результатов с высвеченным словом «*bill*» (счет, чек) (возможно, имеется в виду Bill Gates), со словом «*bush*» (куст) или выражением «*George W. Bush*». С точки зрения лингвистики, эти результаты не синонимичны запросу и обусловлены смешением понятий синонимии и гипо-гиперонимии, накладывающегося на игнорирование синтаксической и семантической структуры документов и запросов.

Помимо лидирующих ИПС, существует множество менее крупных проектов, нацеленных на создание ИИПС. Наиболее примечательными из них представляются «Nigma», «Wolfram|Alpha», «Hakia» и «Powerset».

В 2005 году была запущена альфа-версия ИИПС Nigma. Утверждается, что NIGMA осуществляет авторубрикацию результатов поиска для обеспечения их фильтрования. В действительности в качестве рубрик выдаются подстроки, часто встречающиеся в документах совместно с запросом. Например, запрос «*музыка*» относится к рубрикам «*википедия*», «*без регистрации*», «*портал*», «*новинка*», «*скачать игры*», «*интернет магазины*», и т.д. Такие рубрики, как «*искусство*», в выдаче при этом не встречаются. Синтаксический анализ система не осуществляет и реализует самый обычный логический язык запросов.

«Wolfram|Alpha», запущенная в 2009 году, — это фактологическая ИПС. Она основывается «на собственной базе знаний, которая содержит данные о математике, физике, астрономии, …» [WolframAlpha 2009]. Wolfram преобразовывает единицы различных систем измерения. Расчет ответа на основании собственной базы привел к ряду курьезов: «на момент открытия, запрос president of russia 1999 выдавал имя Аслана Масхадова» [там же]. Декларируется, что Wolfram осуществляет естественноязыковой анализ запросов, однако это не так: система лишь выделяет в запросе известные ей подстроки.

Проект «Hakia» стартовал в 2004 году, и его целью было создать ИПС, способную конкурировать с Google благодаря семантическому поиску. До сих пор, однако, Hakia не включает в себя компонентов, осуществляющих семантический анализ. Вместо них задействуются технологии QDEX (query index) и SemanticRank. QDEX — это индекс поисковых запросов, привязанных к концептам онтологии (неясно, чем они отличаются от дескрипторов), SemanticRank — это алгоритм, устанавливающий связи между «концептами» статистическими методами [Hakia Takes On Google With Semantic Technologies 2007].

Проект «Powerset» стартовал в 2005 году. Его цель — создание «естественноязыковой поисковой машины, которая читает и понимает каждое предложение во Всемирной Паутине» («a natural language search engine that reads and understands every sentence on the Web» [Powerset (company) 2009]). В действительности индексируется только английская часть Википедии. Powerset заслуживает особого внимания, так как использует синтаксические парсеры (напр., XLE [Parsing Miss South Carolina's Statement 2007]) и учитывает поверхностные синтаксические структуры. Терминальные узлы синтаксических деревьев Powerset приводит к дескрипторам, позволяющим учитывать синонимию. Эффективность этого метода не всегда очевидна. На запрос «*free software*» («*свободное программное обеспечение*») в качестве результатов выдаются выражения «*independent programs*» («*независимые программы*»), «*license-free software*» («*программное обеспечение без лицензии*»). Поверхностный синтаксис не позволяет разграничивать лексические значения. В первом примере между значениями узлов синтаксического дерева не были установлены семантические отношения, в результате выбор значений оказался произвольным: у слов «free» и «independent» есть общее значение 'независимый (о человеке)', — однако и в запросе, и в найденном результате определяемые существительные («programs» и «software») не обозначают человека ни в одном из значений. Благодаря применению синтаксического анализа Powerset выдает пропорционально больше осмысленных результатов, чем его аналоги, однако поверхностные синтаксические связи не тождественны понятийным, и не могут гарантировать содержательную релевантность результатов поиска.

Приведенный анализ ИИПС показывает, что их эффективность не всегда высока. Для качественной оценки данного параметра представляется необходимым выявить факторы, влияющие на него. Оценка эффективности ИПС часто основывается на сравнении результатов поиска с *информационной потребностью* пользователя (данный параметр называют пертинентностью: «пертинентность (от англ. pertinent — подходящий, относящийся к делу) — соответствие содержания документа фактической информационной потребности» [Панков, Захаров 1996: 335]). Такая оценка субъективна: информационная потребность пользователя в полной мере известна только ему самому, и поисковый запрос может ей вовсе не соответствовать. Для объективной оценки эффективности ИПС принято опираться на меры соответствия выдаваемых результатов формулировке запроса (а не интересу пользователя) — меры точности и полноты, введенные А. Кентом. Наиболее известной формулой для оценки эффективности ИПС является формула Ван Рисбергена (также известная как F-мера или F_1-мера) [Информационный поиск 2009]:

$$F_1 = \frac{2 * P * R}{P + R}$$, *где* ***P*** *— мера точности, а* ***R*** *— мера полноты.*

Мера Рисбергена определяет общую эффективность как **гармоническое среднее** его точности и полноты. Оценка параметров эффективности ИИПС должна осуществляться на результатах рубрикации документов и поиска по ним. Совокупности результатов поиска и рубрикации объемны и не могут быть полностью проанализированы за одну серию испытаний. Чаще всего оценка производится выборочным методом, при оценке параметров эффективности для каждого документа с вероятностью 99% и погрешностью 5% оценка мер точности и полноты для всей ИИПС с вероятностью 95% окажется достаточной для выявления качественных изменений эффективности работы системы.

Так как данные поиска и рубрикации неоднородны, выборка должна быть бесповторной и репрезентативной: количество данных определенного типа или тематики в выборке должно относиться к ее объему так же, как и общее их количество в исходной совокупности к ее объему. Кроме того, должно выполняться

требование ***случайности*** выборки:

1. Выборка должна быть построена так, чтобы любой объект в пределах совокупности имел равные возможности быть отобранным;
2. Выборка должна быть сформирована так, чтобы любое сочетание из n объектов (где n – количество объектов, или случаев, в выборке) имело равные возможности быть отобранным для анализа.

Для выполнения требования случайности часто используют генератор случайных чисел: случайным образом выбирается номер объекта в исходной совокупности. При выполнении условия случайности выборки распределение выборочных данных стремится к нормальному. Минимальный объем выборки определяется по формуле

$n = \frac{t^2 * N * S^2}{\Delta_{\bar{x}}^2 * N + t^2 * S^2}$ [Васнев 2002], где n — объем выборки, t — коэффициент кратности средней ошибки, зависящий от доверительной вероятности в соответствии с t-критерием Стьюдента (для p=0,95 t=1,96), N — объем исходной совокупности, S — среднеквадратическое отклонение значения параметра от среднего в выборке (для величин, шкалируемых от 0 до 1, S≤0,5; поэтому 0,5 принимается как гарантирующее значение), $\Delta_{\bar{x}}$ — предельное значение ошибки (погрешность).

Для оценки параметров эффективности работы системы в конкретном случае минимальный объем выборки n составляет $\frac{1,7 * N}{0,0025 * N + 1,7}$ (1); тогда доверительная вероятность равна 99%, а погрешность — 5%. Выборочный метод при этом в принципе имеет смысл, если N≥27, так как для N=26 по указанной формуле n≥25,042, а так как n — натуральное число, n≥26; если же N<27, то n=N.

Для оценки параметров эффективности работы всей системы в целом достаточно вычислять минимальный объем выборки по формуле $\frac{0,96 * N}{0,0025 * N + 0,96}$ (2), в этом случае доверительная вероятность составляет 95%, а погрешность — 5%. Выборочный метод при этом имеет смысл, если N≥21, так как для N=20 по

указанной формуле n≥19,001, а так как n — натуральное число, n≥20, и для N<20, опять же, n=N.

Таким образом, для вычисления эффективности ИИПС достаточно провести серию из n испытаний, в рамках каждого из которых вычисляется формула Рисбергена. Сложность состоит в вычислении мер точности и полноты, не вполне поддающихся объективной оценке.

Точность поиска — это нормированная мера, измеряющаяся вещественной величиной в диапазоне от 0 до 1, и определяющая «отношение количества релевантных выданных документов к общему числу документов в выдаче» [Панков, Захаров 2006: 336]. Точность авторубрикации — это нормированная мера, определяющая для одного текста отношение количества корректно привязанных к нему рубрик к общему количеству рубрик, объективно релевантных данному тексту, или «количество истинных положительных значений (…), деленное на общее количество элементов, отнесенных к положительному классу (…)» («the number of true positives (…) divided by the total number of elements labeled as belonging to the positive class (…)» [Precision and recall 2009]). Полнота поиска — это нормированная мера, определяющая «отношение количества выданных релевантных документов к общему числу релевантных документов в исходном информационном массиве» [Панков, Захаров 2006: 336]. Полнота рубрикации — это нормированная мера, определяющая для одного текста отношение количества корректно привязанных к нему рубрик к общему количеству рубрик, объективно релевантных данному тексту, или, в общем случае, «количетво истинных положительных значений, деленное на общее количество элементов положительного класса» («the number of true positives divided by the total number of elements that actually belong to the positive class» [Precision and recall 2009]). Точность и полнота могут быть измерены для реакции ИПС на один запрос или текст, причем количество выданных результатов должно быть большим нуля. Формулы для определения данных параметров таковы:

$Precision = \frac{|D_{rel} \cup D_{retr}|}{|D_{retr}|}$, $Recall = \frac{|D_{rel} \cup D_{retr}|}{|D_{rel}|}$, где Precision — точность, Recall — полнота, D_{rel} — множество записей, релевантных формулировке запроса, D_{retr} —

множество выданных записей.

Чем выше значение параметра точности поиска и рубрикации, тем меньше «информационного шума», то есть нерелевантных результатов. Чем выше значение параметра полноты поиска, тем меньше «недоступных данных», то есть релевантных данных, не предоставляемых пользователю. Точность и полноту можно оценить выборочным методом. Например, если количество результатов составляет 1000000, то, согласно (1), случайной выборки объемом 680 результатов достаточно для того, чтобы с вероятностью 99% значение параметра по выборке не отличалось от значения параметра по всем выдаваемым результатам более чем на 0.05.

Проблему при определении параметров точности и полноты составляет множество D_{rel}: предполагается, что во всей совокупности документов можно выделить «строго» релевантные и «строго» нерелевантные анализируемому запросу. Тем не менее, релевантность документа запросу — это мера, зависимая от множества параметров, каждый из которых может иметь разный вес. Для систематизации стратегий интеллектуального поиска и оценки параметра релевантности предлагается следующий набор признаков, определяющих критерии соответствия результата запросу:

- целостность: соответствие результата самому запросу или пересечение соответствий его частям (учитывается в большинстве ИПС)
- область поиска (для поиска по пересечению частей): в предложении / в абзаце / в документе (чаще всего не учитывается)
- форма-смысл (для поиска по запросу или его части): поиск по формальному соответствию или семантический поиск (не учитывается в большинстве ИПС)
- сохранение формы главного слова: буквальное или грамматическое соответствие (практически нигде не разграничивается)
- сохранение порядка слов (для поиска по формальному соответствию): порядок слов совпадает с запросом / сохраняет смысл порядка слов запроса / произволен при сохранении синтаксической структуры (практически нигде не

учитывается)

- сохранение концептуальной эквивалентности: концептуальная эквивалентность (синонимы) или концептуальное наследование (род-вид) (не учитывается ни в одной из известных ИПС)

Эта система признаков может определить распределение коэффициентов вычисления релевантности для каждой стратегии. В общем случае релевантность вычисляется как rel(x,q) = K * (1-Dc(x,q)), где x — результат, q — запрос, K — коэффициент релевантности стратегии (от 0 до 1), Dc — семантическое растояние (по декартовой мере) между концептуальным рейтингом запроса и концептуальным рейтингом результата. Под концептуальным рейтингом понимается множество пар «концепт — вес», полученное путем авторубрикации. Установить K для каждой стратегии можно только экспериментальным путем, но представляется очевидным, что K напрямую зависит от свойств стратегии:

1. релевантность соответствия целому запросу существенно выше, чем релевантность пересечения соответствий его частям
2. при поиске по пересечению частей релевантность тем выше, чем меньше частей, и чем более узкой является область поиска
3. релевантность соответствия по форме выше релевантности соответствия по смыслу и ниже релевантности буквального соответствия
4. релевантность соответствия с сохранением порядка слов выше, чем релевантность соответствия с сохранением смысла порядка слов, которая выше, чем релевантность соответствия без его сохранения
5. релевантность соответствия по концептуальной эквивалентности выше, чем по концептуальному наследованию

На основании предложенной методики оценки релевантности предлагается пересмотреть методики оценки точности и полноты. Так как значение релевантности является шкалируемым, выделять среди результатов множество релевантных бессмысленно: в некоторой степени релевантным окажется каждый результат.

Определим *выпадение* как среднее отклонение релевантности результатов,

декларируемой ИПС, от объективной релевантности выданных результатов запросу. Тогда точность определяется как разность единицы и выпадения:

$$Precision_{rel}(q) = 1 - \frac{\sum_{i=0}^{N_{RETR}} \left| rel(res_i, q) - rel_{decl}(res_i, q) \right|}{N_{RETR}}$$

В базовых формулах точности и полноты значится «количество релевантных выданных результатов» — мощность пересечения множества релевантных результатов с множеством найденных. Предлагается рассматривать данное значение как частный случай, при котором все декларируемые системой релевантности равны 1, и все объективные релевантности в исходном массиве равны либо 1, либо 0. Для общего случая предлагается оценивать более сложную величину — сумму релевантностей результатов. Тогда полноту можно переопределить как отношение суммы релевантностей результатов к сумме релевантностей всех имеющихся данных:

$$Recall_{rel}(q) = \frac{\sum_{i=0}^{N_{RETR}} rel(res_i, q)}{\sum_{i=0}^{N} rel(src_i, q)}$$

Можно обнаружить, что прежние формулы дают те же результаты, если rel принимает в качестве значений 0 либо 1. Обе формулы зависят от функции rel, определяющей декартово расстояние между концептуальными рейтингами запроса и результата, полученными путем авторубрикации, поэтому точность оценки эффективности ИИПС зависит от эффективности авторубрикации. Цепь вычислений замыкается на оценке эффективности авторубрикации одного концепта: она равна 1, если результат совпадает со входным концептом, и 0 во всех остальных случаях.

Таким образом, обзор истории развития идеи интеллектуального поиска показал низкий уровень эфективности современных ИИПС. Выявлена недосаточность принятых на сегодняшний день методик оценки эффективности ИП для оценки ИИПС. На основании различных реализаций идеи ИИПС предлагается комплексный лингвистический подход к выявлению факторов, влияющих на

эффективность решения этой задачи, и методика оценки релевантности в применении к ИИПС, основанная на концептуальном рейтинговании и авторубрикации. На основании предложенных методик предлагается уточненная методика оценки эффективности работы ИИПС.

Литература

1. Вайз Д. А., Малсид М. Google. Прорыв в духе времени — М. : Эксмо, 2007.
2. Васнев С. А. 11.2. Виды выборки, способы отбора и ошибки выборочного наблюдения // Статистика [2002-2002]. Дата обновления: 16.09.2002. URL: http://www.hi-edu.ru/e-books/xbook096/01/index.html?part-011.htm (дата обращения: 29.07.2009).
3. Захаров В.П. Информационно-поисковые системы: Учебно-методическое пособие. — СПб., 2005.
4. Интернет // Википедия. [2005—2009]. Дата обновления: 06.10.2009. URL: http://ru.wikipedia.org/?oldid=18956654 (дата обращения: 06.10.2009).
5. Информационный поиск // Википедия. [2006—2009]. Дата обновления: 06.07.2009. URL: http://ru.wikipedia.org/?oldid=16894594 (дата обращения: 29.07.2009).
6. Панков И.П., Захаров В.П. Информационно-поисковые системы // Прикладное языкознание: Учебник. — Спб, 2006.
7. Семенов Ю.А. 4.5.13 Система поиска файлов Archie // Telecommunication technologies - телекоммуникационные технологии. [1997—2008]. Дата обновления: 19.03.2008. URL: http://book.itep.ru/4/45/arch4513.htm (дата обращения: 10.10.2009).
8. Семенов Ю.А. 4.5.14 Современные поисковые системы // Telecommunication technologies - телекоммуникационные технологии. [1997—2008]. Дата обновления: 20.08.2009. URL: http://book.itep.ru/4/45/retr4514.htm (дата обращения: 10.10.2009).
9. Справка Google: Расширенный поиск // Справочный центр Google. [2009-2009] Дата обновления: 4.01.2009. URL: http://www.google.ru/intl/ru/help/refinesearch.html (дата обращения: 16.10.2009)
10. Сэлтон Г. Автоматическая обработка, хранение и поиск информации. – М., 1973
11. Baker Loren. Timeline of Search Engine History // Search Engine Journal. [2003—2009]. Дата обновления: 15.09.2009. URL: http://www.searchenginejournal.com/search-engine-history/13152/ (дата обращения: 10.10.2009).
12. Bush Vannevar. As We May Think // The Atlantic Monthly, 1945-07 – 1945
13. Doyle Lauren B., Becker Joseph. Information Retrieval and Processing. – Melville, 1975.
14. Garfield, E. A Tribute To Calvin N. Mooers, A Pioneer Of Information Retrieval // The Scientist,

Vol 11, Issue 6, p. 9. – March 17, 1997

15. Hakia Takes On Google With Semantic Technologies // ReadWriteWeb. [2007 — 2007]. Дата обновления: 16.10.2009. URL: http://www.readwriteweb.com/archives/hakia_takes_on_google_semantic_search.php (дата обращения: 19.10.2009)
16. Information Retrieval // Wikipedia, The Free Encyclopedia. [2002—2009]. Дата обновления: 10.10.2009. URL: http://en.wikipedia.org/w/index.php?title=Information_retrieval&oldid=319063010 (дата обращения: 13.10.2009).
17. Kent Allen, Berry Madeline M., Luehrs Jr. Fred U., Perry J. W. Machine literature searching VIII. Operational criteria for designing information retrieval systems // American Documentation, Vol 6, Issue 2, pp 93-101. – April 1955
18. Kent Allen. Textbook on Mechanized Information Retrieval. – Interscience, New York, 1962
19. Li, Wei. The First Search Engine, Archie // Learning technologies timeline. [2002-2009]. Дата обновления: 21.09.2002. URL: http://www.isrl.illinois.edu/~chip/projects/timeline/1990archie.htm (дата обращения: 10.10.2009).
20. Mooers, C. N. Zatocoding applied to mechanical organization of knowledge // American Documentation, Vol. 2, Issue 1, pp 20-32. – 1951
21. Mooers, C. N. Mooers' Law; or why some retrieval systems are used and others are not. // Zator Technical Bulletin, 136. – Cambridge, MA: Zator Company, 1959
22. Salton G. A Theory of Indexing // CBMS-NSF Regional Conference Series in Applied Mathematics – Society for Industrial Mathematics, January 1, 1987
23. Wall Aaron. History of Search Engines: From 1945 to Google 2007 // http://www.searchenginehistory.com – 2009
24. WolframAlpha // Википедия. [2009—2009]. Дата обновления: 08.09.2009. URL: http://ru.wikipedia.org/?oldid=18312838 (дата обращения: 19.10.2009).