УДК 004.051

С.Е. Савотченко, В.А. Стукалов

СЕМАНТИЧЕСКАЯ МЕРА РЕЗУЛЬТАТОВ ПОИСКОВЫХ ЗАПРОСОВ

Савотченко Сергей Евгеньевич, *доктор физико-математических наук, доцент, окончил физический факультет Харьковского государственного университета, в настоящее время профессор кафедры «Информационные технологии» Белгородского института развития образования. Имеет статьи в области математического моделирования, информационных технологий и автоматизированных информационных систем. [e-mail: savotchenko@hotbox.ru].*

Стукалов Вадим Андреевич, *аспирант кафедры «Библиотековедение, Библиографоведение и книговедение» Белгородского государственного института искусств и культуры. Имеет статьи в области автоматизированных библиотечных информационных систем. [e-mail: svadglo@gmail.com].*

Аннотация

В работе впервые введены понятия семантической меры и парциальной семантической меры. Получены результаты вычисления таких характеристик на примере десяти популярных поисковых систем. Предложенный подход к оценке качества информационно-поисковых систем позволил сформировать критерий проверки способности выдавать пертинентные документы.

Ключевые слова: дескриптор, информационный поиск, семантические связи, парадигматические отношения, информационно-поисковые системы.

Sergei Evgenevich Savotchenko, *Doctor of Physics and Mathematics, Associate Professor, graduated from the Physics Faculty of Kharkov State University, Professor at the Department of Computer Science at Belgorod Institute of Education Development; an author of articles in mathematical modeling, information technology and automated information systems. e-mail: savotchenko@hotbox.ru.*

Vadim Andreevich Stukalov, *Post-graduate Student at the Department of Library Science, Bibliography, and Bibliology at Belgorod State Institute of Arts and Culture; an author of articles in the field of integrated library systems. e-mail: svadglo@gmail.com.*

Abstract

The concepts of semantic measure and partial semantic measure are introduced in this paper for the first time. The results of such characteristics computing on the example of ten popular search systems are obtained. The proposed approach to evaluate the quality of information search systems has allowed to generate a criterion for testing the ability of pertinent documents issue.

Key words: descriptor, information search, semantic relations, paradigmatic relations, information retrieval systems.

Введение

К одной из основных задач информационного поиска относится отображение из имеющегося множества информации такого подмножества, которое наилучшим образом соответствовало информационной потребности пользователя [1, 2]. Возникает естественная проблема определения меры такого соответствия.

В последнее время наметилась новая тенденция в развитии поисковых систем. Она направлена на то, чтобы улучшать не столько показатели релевантности, а скорее пертинентности результатов поискового запроса пользователя. Согласно определениям ГОСТ 7.73-96, релевантность – соответствие полученной информации информационному запросу; пертинентность – соответствие полученной информации информационной потребности, то есть пертинентность определяет степень соответствия между ожиданиями пользователя и результатами поиска.

Обычно пользователь при работе с информационно-поисковыми системами (ИПС) сталкивается с оценкой полноты и точности результатов информационного поиска [3]. Одним из способов повышения точности и пертинентности информационного поиска является систематизация. Повышению полноты и точности поиска способствует технология построения запросов, основанная на соответствующей систематизации предметных областей [4]. Систематизировать понятия позволяет введение парадигматических отношений между лексическими единицами (ЛЕ) в системах [5].

В связи с такого рода развитием ИПС возникает необходимость количественного анализа качества информационного поиска, осуществляемого по реализуемым в ИПС

поисковым алгоритмам и методам, а также построения математических моделей для оценки эффективности информационного поиска. В первую очередь, для этого следует определить соответствующие показатели, характеризующие пертинентность поиска с различных сторон [6–8].

1 Определения показателей

Для сравнительного анализа механизмов информационно-поисковых языков (ИПЯ) в различных ИПС целесообразно использовать количественные показатели, характеризующие результаты выполнения запросов, отражающих основные смысловые связи, такие, как: отношения иерархии – вышестоящее родовое, вышестоящее целое, нижестоящее видовое, нижестоящее часть; отношения тождества – учет синонимов; отношения ассоциации [6]. В качестве запросов тогда предлагается составить специальным образом последовательность лексических единиц, все члены которой будут связаны четкими парадигматическими отношениями: $Q_m(i)$, где $i=0=(\text{д})$, $i=1=(\text{с})$, $i=2=(\text{вр})$, $i=3=(\text{вц})$, $i=4=(\text{нч})$, $i=5=(\text{нв})$, $i=6=(\text{а})$; (д) – заглавный дескриптор, называемый запросом базового уровня; (с) – ЛЕ, которая является синонимом к (д); (вр) – ЛЕ, которая является вышестоящим родовым к (д); (вц) – ЛЕ, которая является вышестоящим целым к (д); (нч) – ЛЕ, которая является нижестоящим частичным к (д); (нв) – ЛЕ, которая является нижестоящим видовым к (д); (а) – ЛЕ, которая является ассоциацией к (д).

Нас интересует количество релевантных документов $A_i\ (Q_m,\ S_l)$, выдаваемых на i-ю ЛЕ последовательности запросов Q_m в ИПС S_l [7]. Характеристики семантических связей в ИПС представляют собой показатели, определяемые выражениями [8]:

$$J_{ij}\left(Q_m, S_l\right) = A_i\left(Q_m, S_l\right) / A_j\left(Q_m, S_l\right). \quad (1)$$

Введем определение семантической меры результатов поисковых запросов – взвешенное среднее гармоническое значение показателей полноты семантических связей (1):

$$F_s\left(Q_m, S_l\right) = 1 \Big/ \sum_k \nu_k / J_k\left(Q_m, S_l\right), \quad (2)$$

где ν_k – веса, такие, что $\sum_{k=1}^{n} \nu_k = 1$, n – количество усредняемых показателей, суммирование производится по выделенной группе пар чисел, например, $k = \{1\ 0, 2\ 0, 3\ 0, 4\ 0, 5\ 0, 6\ 0, 2\ 3, 4\ 5, 1\ 6\}$, тогда $n = 9$. Мера называется равновесной, если все веса одинаковы: $\nu_k = 1/n$.

Средние показатели:

1) средние значения по запросам:

$$\overline{F}_s\left(S_l\right) = \sum_m F_s\left(Q_m, S_l\right);$$

2) средние значения по ИПС:

$$\overline{F}_s\left(Q_m\right) = \sum_l F_s\left(Q_m, S_l\right);$$

3) общая средняя:

$$\overline{F}_s = \sum_m \overline{F}_s\left(Q_m\right) = \sum_l \overline{F}_s\left(S_l\right) = \sum_m \sum_l F_s\left(Q_m, S_l\right).$$

Для того чтобы выявить, учитывается ли в ИПС пертинентность, а не только релевантность выдаваемых документов, целесообразно использовать специальный вариант меры (2), определяемой выражением:

$$F_{16}\left(Q_m, S_l\right) = \frac{1}{\dfrac{\nu}{J_{10}\left(Q_m, S_l\right)} + \dfrac{1-\nu}{J_{60}\left(Q_m, S_l\right)}}, \quad (3)$$

где вес $\nu \in [0;1]$. Данная характеристика представляет собой аналог хорошо известной F-меры (меры Ван Ризбергена) [1, 2], но предназначена для характеристики соотношения между параллелями парадигматических связей в результатах поисковых запросов, а не баланса между точностью и полнотой.

Мера (3) характеризует способность ИПС выдавать документы по синонимам и ассоциациям к запрошенному дескриптору. При этом интересна зависимость данной меры от веса ν, поскольку при $\nu = 0$: $F_{16}(Q_m, S_l) = J_{60}(Q_m, S_l)$, а при $\nu = 1$: $F_{16}(Q_m, S_l) = J_{10}(Q_m, S_l)$. При промежуточных значениях веса ν мера $F_{16}(Q_m, S_l)$ характеризует распределение предпочтений ИПС выдачи между синонимами и ассоциациями к запрошенному дескриптору. В идеальном случае, когда показатели долей документов с синонимами и с ассоциациями одинаковы, мера $F_{16}(Q_m, S_l)$ будет постоянной. На практике допустимо малое отклонение от такого постоянного значения. В этом случае можно считать, что рассматриваемые ИПС способны выдавать пертинентные документы.

Поскольку мера $F_{16}(Q_m, S_l)$ содержит только часть слагаемых меры $F_s(Q_m, S_l)$, то ее можно назвать парциальной или частичной.

При малых значениях веса, близких к нулю, то есть когда $\nu \to 0$ $(\nu << 1)$, зависимость меры $F_{16}(Q_m, S_l)$ от него упрощается и становится линейной:

$$F_{16}\left(Q_m, S_l\right) = J_{60}\left(Q_m, S_l\right) \cdot \left\{1 + \nu \cdot \left(1 - \frac{J_{60}\left(Q_m, S_l\right)}{J_{10}\left(Q_m, S_l\right)}\right)\right\}.$$

При значениях веса, близких к единице, то есть когда $\nu \to 1$, величина меры $F_{16}(Q_m, S_l)$, также линейной, зависит от веса:

$$F_{16}\left(Q_m, S_l\right) = J_{10}\left(Q_m, S_l\right) \cdot \left\{1 + (1-\nu) \cdot \left(1 - \frac{J_{10}\left(Q_m, S_l\right)}{J_{60}\left(Q_m, S_l\right)}\right)\right\}.$$

2 Методика проведения исследования

Для проведения исследований была выделена следующая группа показателей $\{J_{10}, J_{20}, J_{30}, J_{40}, J_{50}, J_{60}, J_{23}, J_{45}\}$. Использованные последовательности запросов Q_m, члены которой составлены на основе информационно-поискового тезауруса (ГОСТ 7.25-2001), приведены в таблице 1.

Таблица 1

Последовательности запросов

Вид	Q_1	Q_2	Q_3	Q_4	Q_5
д	музей	линейная алгебра	языкознание	библиотека	обучение
с	галерея	алгебра Банаха	лингвистика	книгохранилище	воспитание
вр	учреждение культуры	математическая наука	гуманитарные науки	учреждение культуры	педагогический процесс
вц	музейное дело	высшая алгебра	филология	центральная библиотечная система	образование
нч	экспонат	линейное уравнение	семантика	школьная библиотека	заочное обучение
нв	музей-заповедник	матричная алгебра	психолингвистика	книжный фонд	лекционное занятие
а	искусство	определитель	алфавит	библиотекарь	ученик

Для проведения исследований были выбраны десять наиболее популярных русскоязычных ИПС: S_1={nigma.ru}, S_2={qip.ru}, S_3={mail.ru}, S_4={bing.com}, S_5={ngs.ru}, S_6={yandex.ru}, S_7={google.ru}, S_8={rambler.ru}, S_9={aport.ru}, S_{10}={ru.yahoo.com}. Даты ввода запросов: 15.03.12–27.03.12.

Методика проведения исследований следующая. В строке поиска ИПС S_1 вводится первая ЛЕ последовательности Q_1 (вид отношения – (д) из таблицы 1). Количество выданных по этому запросу документов есть величина $A_1(Q_{1(д)}, S_1)$. Затем в этой же ИПС вводится второй член последовательности Q_1 (вид отношения – (с) из таблицы 1). Количество выданных по этому запросу документов есть величина $A_2(Q_{1(с)}, S_1)$. И так далее для всех членов последовательностей всех запросов по таблице 1 во всех указанных ИПС, в результате чего получается необходимый набор величин $A_i(Q_m, S_l)$. Затем с помощью этих величин вычисляются показатели семантических связей (1) и далее значения меры (2) и их средние.

3 Результаты

Для вычисления семантических мер десяти Интернет-ИПС по n=8 показателям взяты все веса v_k = 0,125, и формула (2) принимает вид:

$$F_s(Q_m,S_l)=\frac{8}{\sum_{i=1}^{6}\frac{1}{J_{i0}(Q_m,S_l)}+\frac{1}{J_{23}(Q_m,S_l)}+\frac{1}{J_{45}(Q_m,S_l)}}.$$

Результаты проведенных наблюдений представлены на рисунках 1 и 2. В рассматриваемом случае общая средняя $\overline{F}_s$=0,182087.

Парциальная семантическая мера $F_{16}(Q_m, S_l)$ была рассчитана по данным исследования десяти Интернет-ИПС. Графики ее зависимости от веса v представлены на рисунке 3.

Проведенные исследования показали, что, как видно на рисунке 3, для группы выделенных ранее шести ИПС $\{S_1, S_2, S_5, S_6, S_8, S_9\}$ зависимости $F_{16}(Q_m, S_l)$ от веса близки, что свидетельствует об идентичности их поисковых механизмов.

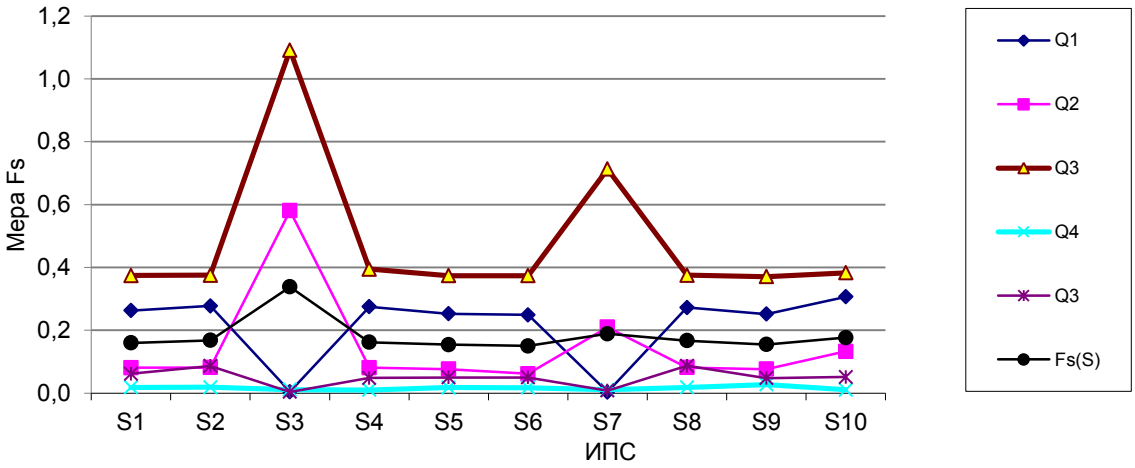

Рис. 1. Результаты вычисления меры (2) и средних значений по запросам для различных ИПС

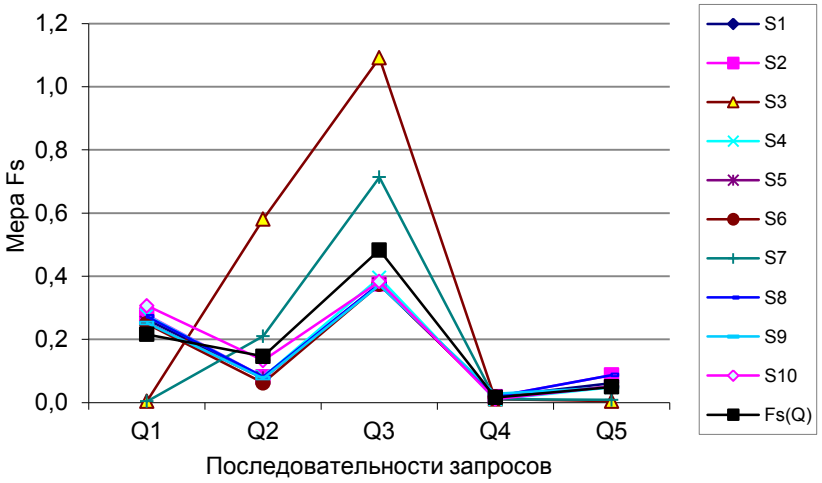

Рис. 2. Результаты вычисления меры (2) и средних значений по ИПС для различных запросов

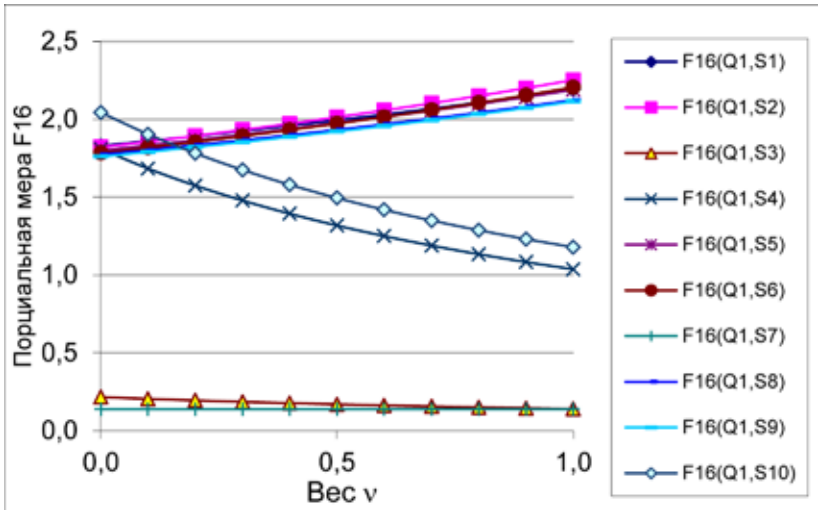

Рис. 3. Результаты вычисления меры (3)

Проведенные вычисления показывают, что парциальная мера практически не зависит от веса для ИПС $S_3 = \{\text{mail.ru}\}$ и $S_7 = \{\text{google.ru}\}$.

Заключение

Основные выводы по работе:

1) От формулировки последовательности запросов значения семантической меры результатов поисковых запросов слабо зависят.

2) Зависимость средней по запросам семантической меры результатов поисковых для большинства ИПС, за исключением S_3 и S_7, примерно одинакова и группируется около 0,16.

3) Можно выделить экспериментальное критическое значение меры $F_{sc} = 0{,}17$, для которого выполняется неравенство $\overline{F}_s(S_l) < F_{sc}$ для большинства ИПС, за исключением S_3, S_7 и S_{10}.

4) Для характеристики способности ИПС выдавать пертинентные документы (по синонимам и ассоциациям к запрошенному дескриптору) предлагается использовать парциальную семантическую меру результатов поисковых, как функцию веса. Критерий ее практического применения можно сформулировать в виде: если величина такой парциальной меры практически не зависит от веса и мало отклоняется от среднего значения, то можно считать, что исследуемая ИПС способна выдавать пертинентные документы по запросу.

СПИСОК ЛИТЕРАТУРЫ

1. Маннинг К., Рагхван П., ШтюцеХ. Введение в информационный поиск : пер. с англ. – М. : ИД Вильямс, 2011. – 528 с.

2. Теория ИПС. [Электронный ресурс]. – URL: http://www.likt590.ru/project/poisk/to.htm.

3. Официальные метрики РОМИП'2010 [Электронный ресурс]. –URL: http://romip.ru/romip2010/20_appendix_a_metrics.pdf.

4. Стулов А. Особенности построения информационных хранилищ // Открытые системы [Электронный ресурс]. – 2003. – № 4. – URL: http://www.osp.ru/os/2003/04/182942. – Загл. с экрана.

5. Савотченко С.Е., Жуков П.С. Моделирование информационного поиска в базе данных с учетом семантических связей // Автоматизация процессов управления. – 2013. – № 2(32). – С. 17–22.

6. Савотченко С.Е., Логинова А.Е. Математический метод сравнительного анализа семантических особенностей информационно-поисковых систем // Теория и практика общественного развития. – 2012. – № 6. – С. 101–104.

7. Савотченко С.Е. Проскурина Е.А. Корреляционный и дисперсионный анализ лингвистических особенностей поиска в интернете // Среднее профессиональное образование. – 2012. – № 12. – С. 38–40.

8. Савотченко С.Е. Проскурина Е.А. Показатели семантических связей информационно-поисковых систем // Научные ведомости «БелГУ». Сер. История. Политология. Информатика. – 2013. – Вып. 25/1, № 1(144). – С. 145–151.