УДК 004.031.4; 025.4.036

ПОКАЗАТЕЛИ ОЦЕНКИ КАЧЕСТВА ПЕРТИНЕНТНОСТИ РЕЗУЛЬТАТОВ АВТОМАТИЗИРОВАННОГО ПОИСКА В ИНФОРМАЦИОННЫХ СИСТЕМАХ

INDICATORS OF QUALITY ASSESSMENT RESULTS PERTINENCE AUTOMATED SEARCH IN INFORMATION SYSTEMS

С.Е. Савотченко
S.E. Savotchenko

Белгородский институт развития образования,
Россия, 308007, Белгород, ул. Студенческая, 14

Belgorod Institute of Education Development,
14 Studencheskaya St, Belgorod, 308007, Russia

e-mail: savotchenko@bsu.edu.ru

Аннотация. В данной статье проанализировано состояние проблемы повышения пертинентности информационного поиска в автоматизированных системах, в том числе и в глобальной сети. Для анализа качества автоматизированного информационного поиска предлагается использовать метод составления последовательности условно нормализованных запросов. Введены новые количественные показатели оценки качества пертинентности и релевантности выполнения запросов.

Resume. This article analyzes the problem of increasing pertinence of information search in the automated systems, including in global network. The method of preparation of the sequence of normalized conditional requests is proposed to use in analyzing the quality of the automated information retrieval. The new quantitative quality assessment indicators of pertinence and relevance of the query are determined.

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

Keywords: information search, pertinence, semantic links, paradigmatic relations, information retrieval systems.

В последнее время наметилась новая тенденция в развитии «интеллектуализации» поискового аппарата автоматизированных информационных системах (АИС). Она направлена на то, что бы улучшать не столько показатели релевантности, а, скорее, показатели пертинентности результатов поискового запроса пользователя. Согласно определениям ГОСТ 7.74-96: релевантность – соответствие полученной информации информационному запросу, пертинентность – соответствие полученной информации информационной потребности, то есть пертинентность определяет степень соответствия между ожиданиями пользователя и результатами поиска.

Пертинентность, в первую очередь, обусловлена субъективным мнением пользователя о том, в какой мере данный результат поиска (документ) удовлетворяет его информационную потребность, выраженную в вводимом им формализованном запросе с той или иной степенью полноты и точности [Пальчунов Д.Е., 2008]. Поскольку понятие релевантности является уже пертинентности, то выдаваемый АИС результат может оказаться релевантным данному запросу, но не удовлетворять информационную потребность пользователя, то есть не быть пертинентным. Основная причина этого заключается в многозначности естественного языка, поскольку пользователь на нем составляет свой, как правило, не учитывая возможности существования у одного понятия нескольких значений (или синонимов слов), хотя сам пользователь предполагает провести именно тематический поиск, то есть получить в результате выполнения своего запроса конкретные сведения в какой-либо области знания.

Технология тематического поиска использует иерархические классификационные системы, в которых вся область знаний разделяется на крупные предметные области (классы), которые, в свою очередь, подразделяются на более мелкие (подклассы), и т. д. Каждому классу и подклассу присваивается классификационный индекс. В результате формируется разветвленная упорядоченная структура, с помощью которой можно классифицировать все источники информации. В качестве примера такого вида классификационных систем можно указать международную универсальную

десятичную классификацию (УДК), международную десятичную классификацию М. Дьюи (ДКД), национальную библиотечно-библиографическую классификацию (ББК) и др.

Подобные виды классификации широко используются при автоматизированном поиске в электронных каталогах и других базах данных, организованных по иерархическому принципу, что позволяет намного повысить показатели качества информационного поиска. Однако, в базах данных с неопределенным количеством документов, к которым относятся информационно-поисковые системы (ИПС) Интернет, в силу ограниченности программно технических возможностей не применяется отбор документов по классификационному принципу, и проблема осуществления полноценного тематического поиска в них до сих пор не решена в полной мере.

Один из вариантов решения проблемы эффективности автоматизированного информационного поиска в последние годы заключается в использовании онтологий предметных областей [Gruber T. R., 1993], которые позволяют определить более четким образом смысловое содержание поискового запроса. Под онтологией понимается явная формальная спецификация концептуализации, разделяемая некоторым сообществом агентов [Плесневич Г.С., 2012], в широком смысле включающая словарь терминов соответствующей предметной области и связей между ними. Концептуализация означает наличие знаний о предметной области, т.е. фиксация понятий, которые классифицируют объекты этой предметной области и связи между ними. Агентами являются пользователи, автоматизированные системы.

Широкое применение онтологии нашли в АИС, относящихся к различным сферам деятельности, например, к программной инженерии, электронному обучению, бизнес-информатике и к др. Использование онтологий фактически реализует семантические связи между понятиями поискового запроса, может привести к повышению уровня пертинентности результатов автоматизированного информационного поиска. Поэтому активно проводятся исследования и появляются разработки методик повышения качества информационного поиска с помощью онтологии по таким направлениям как применение семантических технологий в сети Интернет и электронных библиотеках [Савотченко С.Е., Жуков П.С., 2013].

В электронных библиотеках, как в базах данных с определенным количеством документов, достигнуты определенные в данном направлении. В течение ряда лет ведутся исследования в области анализа семантики связей между данными, по которым осуществляется поиск, в результате которых появились комплекс онтологий SPAR, семантический раздел в модели научных данных CERIF. В данном направлении ведется активная работа по проекту SKOS (Simple Knowledge Organization System), в котором предлагается модель связывания научных данных, адаптированная для компьютерной обработки, включающая контролируемые структурные словари семантических значений для связывания научных данных [Хорошевский В.Ф., 2012].

В последние годы существенно повысилось практическое применение онтологий в ИПС сети Интернет, например, в Google для классификации веб-сайтов. Онтология товаров и услуг с их характеристиками разработана компанией Amazon. Другой пример – онтология UNSPSC (United Nations Standard Products and Services Code – система ООН стандартных кодов для товаров и услуг).

Как правило, онтологии состоят из описаний на формальных языках представления знаний, которые должны использовать формальные представления понятий. В качестве таких формальных языков в ИПС выступают информационно-поисковые языки (ИПЯ). Для отражения семантических связей между понятиями в ИПЯ используются парадигматические отношения, представляющие собой объективно существующие смысловые отношения между лексическими единицами (ЛЕ), которые устанавливаются и фиксируются в словаре, исходя из потребностей информационного поиска.

Одним из средств поиска с учетом парадигматических отношений является тезаурус, в котором термины иерархически связаны между собой парадигматическими отношениями типа синонимия, род-вид, целое-часть, ассоциация. В качестве терминов выступают понятия (слова или словосочетания), которые согласно современным представлениям являются наиболее информативными и наиболее устойчивыми смысловыми единицами. Смысл термина в тезаурусе передается в основном путем соотнесения его с другими терминами с помощью установления семантических отношений между ним и этими терминами [Загорулько Ю.А. и др., 2012].

В таком контексте становится важным анализ семантической структуры поискового запроса, которая представляет собой совокупность понятий, выявленных в предметной области знаний и связанных между собой парадигматическими отношениями. Сформированную структуру запроса, состоящую из текстовых форм наименований понятий, следует привести к формализованной форме ее представления, т.е. осуществить нормализацию слов и словосочетаний, которую возможно произвести согласно описанному в ГОСТ 7.25-2001 принципу.

Применение онтологий, тезаурусов в автоматизированном поиске может трактоваться как путь к развитию интеллектуальных поисковых систем, позволяющих производить поиск с более высокой пертинентностью. Поскольку различные ИПС сети Интернет имеют свои преимущества и недостатки, то в связи их развитием в данном направлении использования возникает необходимость количественного анализа качества автоматизированного информационного поиска, осуществляемого

по реализуемым в ИПС поисковым алгоритмам и методам, а также построения математических моделей для оценки его эффективности. В первую очередь для этого следует определить соответствующие показатели, характеризующие пертинентность результатов поиска с различных сторон.

Для количественного анализа способности ИПС проводить пертинентный поиск предлагается использовать методику, основанную на использовании семантических связей при организации поисковых запросов [Савотченко С.Е., Логинова А.Е., 2012.]. В ее основе лежит процедура составления последовательности условно нормализованных запросов Q_m, лексические единицы которой связаны парадигматическими отношениями и формируются согласно информационно-поисковому тезаурусу: Q_m = {д, с, вр, вц, нч, нв, а}, где [Савотченко С.Е., Проскурина Е.А., 2013.]:

i=0=(д) – заглавный дескриптор, называемый запросом базового уровня (ведущая ЛЕ),
i=1=(с) – ЛЕ, которая является синонимом к (д),
i=2=(вр) – ЛЕ, которая является вышестоящим родовым к (д),
i=3=(вц) – ЛЕ, которая является вышестоящим целым к (д),
i=4=(нч) – ЛЕ, которая является нижестоящим частичным к (д),
i=5=(нв) – ЛЕ, которая является нижестоящим видовым к (д),
i=6=(а) – ЛЕ, которая является ассоциацией к (д).

Ведем следующие обозначения: $A_i^r(Q_m, S_l)$ – количество релевантных документов, выданных на i-ую ЛЕ последовательности запросов Q_m в ИПС S_l, $A_i^p(Q_m, S_l)$ – количество пертинентных документов, выданных на i-ую ЛЕ последовательности запросов Q_m в ИПС S_l.

Используя данные величины, определим характеристики семантических связей в ИПС:

1) доля релевантных документов по запросу i среди всех найденных по запросу j:

$$J_{ij}^r(Q_m, S_l) = \frac{A_i^r(Q_m, S_l)}{A_j(Q_m, S_l)}, \quad (1)$$

2) доля пертинентных документов по запросу i среди всех найденных по запросу j:

$$J_{ij}^p(Q_m, S_l) = \frac{A_i^p(Q_m, S_l)}{A_j(Q_m, S_l)}, \quad (2)$$

3) доля релевантных документов по запросу i среди найденных релевантных по запросу j:

$$J_{ij}^{rr}(Q_m, S_l) = \frac{A_i^r(Q_m, S_l)}{A_j^r(Q_m, S_l)}, \quad (3)$$

4) доля пертинентных документов по запросу i среди найденных пертинентных по запросу j:

$$J_{ij}^{pp}(Q_m, S_l) = \frac{A_i^p(Q_m, S_l)}{A_j^p(Q_m, S_l)}. \quad (4)$$

Наибольший интерес с точки зрения оценки качества ИПС представляют собой показатели, у которых второй индекс j=0=(д), то есть группа $J_{i0}^k(Q_m, S_l)$, k={r, p, rr, pp}. В этом случае такая группа оценивает полноту отражения пертинентных и релевантных соответственно документов в определенных видах семантических связей по отношению к заглавному дескриптору. Если для ИПС S_l выполняются соотношения $J_{i0}^p(Q_m, S_l) > J_{i0}^r(Q_m, S_l)$, то можно считать, что данная ИПС способна выдавать в большей мере пертинентные документы на запрос пользователя, чем релевантные.

С точки зрения пользователя, которого в большей степени интересуют пертинентные документы, оставшаяся часть непертинентных документов, найденных ИПС S_l, представляют собой информационный шум. Количественную характеристику пертинентного информационного шума можно определить выражением:

$$B_i^p(Q_m, S_l) = \frac{\left|A_0(Q_m, S_l) - A_i^p(Q_m, S_l)\right|}{A_0(Q_m, S_l)} = \left|1 - J_{i0}^p(Q_m, S_l)\right|. \quad (5)$$

Аналогично можно ввести количественную характеристику релевантного информационного шума, то есть доля нерелевантных документов, найденных ИПС S_l:

$$B_i^r(Q_m, S_l) = \frac{\left|A_0(Q_m, S_l) - A_i^r(Q_m, S_l)\right|}{A_0(Q_m, S_l)} = \left|1 - J_{i0}^r(Q_m, S_l)\right|. \quad (6)$$

Следует отметить, что на практике определение количества релевантных документов $A_i^r(Q_m, S_l)$ может быть проведено достаточно просто средствами автоматизированного подсчета доли интересующих ЛЕ в документах, выдаваемых в результатах поиска. В то время как автоматизированное определение количества пертинентных документов $A_i^p(Q_m, S_l)$ может вызывать определенные затруднения и решение данной проблемы требует проведения отдельных

исследований. Простейший способ определения данной величины может заключаться в использовании экспертных мнений, однако он не является автоматизированном в такой степени, как способ определения количества релевантных документов.

Совокупность определенных выше характеристик может служить для оценки качества АИПС на предмет реализации «интеллектуального поиска», то есть возможностей системы выдавать в большей степени пертинентные документы. Расчеты таких показателей могут быть включены в методики проведения независимых экспертиз качества различных ИПС [Савотченко С.Е., Проскурина Е.А., 2015], что существенно повысит их результативность.

Список литературы
References

Gruber T. R. 1993. A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition Journal, 5: 199-220.

Загорулько Ю.А. Боровикова О.И., Загорулько Г.Б. 2012. Построение многоязычных тезаурусов средствами семантической технологии. В кн.: Открытые семантические технологии проектирования интеллектуальных систем: материалы II Междунар. научн.-техн. конф. Минск, БГУИР: 181-188.

Zagorul'ko Yu.A., Borovikova O.I., Zagorul'ko G.B. 2012. Postroenie mnogoyazychnykh tezaurusov sredstvami semanticheskoy tekhnologii. Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh sistem: materialy II Mezhdunar. nauchn.-tekhn. konf. [The construction of multilingual thesauri by means of semantic technologies. Proc.: Open semantic technologies of intelligent systems: Materials of the II International Scientific and Technological Conf.]. Minsk, BGUIR: 181-188. (in Russian).

Пальчунов Д.Е. 2008. Решение задачи поиска информации на основе онтологий. Бизнес-информатика. 1(3): С. 3-13.

Pal'chunov D.E. 2008. Reshenie zadachi poiska informatsii na osnove ontologii. Biznes-informatika. [The solution of information retrieval problem based on ontologies. Business Informatics]. 1(3): 3-13. (in Russian).

Плесневич Г.С. 2012. Формальные онтологии. В кн.: Открытые семантические технологии проектирования интеллектуальных систем: материалы II Междунар. научн.-техн. конф. Минск, БГУИР: 163-168.

Plesnevich G.S. 2012. Formal'nye ontologii. Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh sistem: materialy II Mezhdunar. nauchn.-tekhn. konf. [Formal ontology. Proc.: Open semantic technologies of intelligent systems: Materials of the II International Scientific and Technological Conf.]. Minsk, BGUIR: 163-168. (in Russian).

Савотченко С.Е., Логинова А.Е. 2012. Математический метод сравнительного анализа семантических особенностей информационно-поисковых систем. Теория и практика общественного развития6: 101-104.

Savotchenko S.E., Loginova A.E. 2012. Matematicheskiy metod sravnitel'nogo analiza semanticheskikh osobennostey informatsionno-poiskovykh sistem. Teoriya i praktika obshchestvennogo razvitiya. . [The mathematical method of comparative analysis of the semantic features of information retrieval systems. Theory and practice of social development]. 6: 101-104. (in Russian).

Савотченко С.Е., Жуков П.С. 2013. Моделирование информационного поиска в базе данных с учетом семантических связей. Автоматизация процессов управления. 2(32): 17-22.

Savotchenko S.E., Zhukov P.S. 2013. Modelirovanie informatsionnogo poiska v baze dannykh s uchetom semanticheskikh svyazey. Avtomatizatsiya protsessov upravleniya. [Modeling of information retrieval in a database based on semantic relations. Automation of Control Processes]. 2(32): 17-22. (in Russian).

Савотченко С.Е., Проскурина Е.А. 2013. Показатели семантических связей информационно-поисковых систем. Научные ведомости «БелГУ». Сер. История. Политология. Информатика. Вып. 25/1. 1(144): 145-151.

Savotchenko S.E., Proskurina E.A. 2013. Pokazateli semanticheskikh svyazey informatsionno-poiskovykh sistem. Nauchnye vedomosti «BelGU». Ser. Istoriya. Politologiya. Informatika. [Indicators of semantic relationships of information retrieval systems Belgorod State University Scientific bulletin. Ser. History. Political science. Computer science]. Vyp. 25/1. 1(144): 145-151. (in Russian).

Савотченко С.Е., Проскурина Е.А. 2015. Математические методы исследования семантических особенностей подсистемы поиска в автоматизированных информационных системах. Вестник Сибирского института бизнеса и информационных технологий. 1(13): 69-76.

Savotchenko S.E., Proskurina E.A. 2015. Matematicheskie metody issledovaniya semanticheskikh osobennostey podsistemy poiska v avtomatizirovannykh informatsionnykh sistemakh. Vestnik Sibirskogo instituta biznesa i informatsionnykh tekhnologiy. [Mathematical methods of research of semantic search subsystem features in automated information systems. Bulletin of the Siberian Institute of Business and Information Technology]. 1(13): 69-76. (in Russian).

Хорошевский В.Ф. 2012. Семантические технологии: ожидания и тренды. В кн.: Открытые семантические технологии проектирования интеллектуальных систем: материалы II Междунар. научн.-техн. конф. Минск, БГУИР: 143-158.

Khoroshevskiy V.F. 2012. Semanticheskie tekhnologii: ozhidaniya i trendy. Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh sistem: materialy II Mezhdunar. nauchn.-tekhn. konf. [Semantic Technologies: Expectations and Trends Proc.: Open semantic technologies of intelligent systems: Materials of the II International Scientific and Technological Conf.]. Minsk, BGUIR: 143-158. (in Russian).