УДК 81’322.2

Максименко О.И.
Московский государственный областной университет

ИНФОРМАЦИОННО-ПОИСКОВЫЕ СИСТЕМЫ: ОЦЕНКА МЕТОДОМ НЕЧЁТКОЙ ЛОГИКИ

Аннотация. Информационно-поисковые системы (ИПС), использующие поиск на естественном языке, требуют особых методов оценки. В статье приводится ряд способов оценки подобных систем, а также описывается структура ИПС КАСКАД и преимущества одного из возможных вариантов оценки эффективности ИПС методом нечёткой логики (на примере данной ИПС). Для количественной оценки качества поиска используются одновременно два частных критерия: критерий полноты и критерий точности поиска. Зависимость оценки качества ИПС от коэффициента полноты наглядно проиллюстрирована на рисунке. Проведённый эксперимент позволил сделать вывод о том, что оценка качества ИПС определяется не только техническими характеристиками самой ИПС, но и зависит от формулировки запроса.
Ключевые слова: информационно-поисковые системы, нечёткая логика, экспертные оценки.

O. Maksimenko
Moscow State Regional University

INFORMATION RETRIEVAL SYSTEMS: ESTIMATION BY FUZZY LOGIC

Abstract. Natural language-based information retrieval systems demand specific estimation methods. In the article a number of estimation methods of these systems are presented, the structure of information retrieval system KASKAD is described and advantages of one of the possible variants of estimating search systems efficiency using the fuzzy logic method are given. Two criteria are used for quantitative evaluation of search system quality – the search completeness and search precision. The dependence of search system quality evaluation on the criterion of completeness is illustrated on the graph. The experiment showed that the information retrieval system evaluation is determined not only by its technical characteristics but by wording of the request for information.
Key words: information retrieval systems fuzzy logic, expert estimation.

В последнее время существенно возросла скорость появления и накопления новых данных в самых разных областях человеческой деятельности. В связи с этим появилась необходимость создания информационных систем, которые не только содержали бы в себе средства поиска тех или иных фактических данных, но и помогали бы находить с их помощью нетривиальные способы решения разнообразных проблем.

Информационный поиск – это процесс выделения из общего информационного потока той информации, которая связана с указанной в инфор-

мационном запросе темой или содержит необходимые факты и сведения. Автоматизация процесса поиска осуществляется с помощью информационно-поисковых систем (ИПС).

Первые ИПС работали преимущественно с информацией фактического характера, например, характеристиками объектов и их связей. Затем появилась возможность обрабатывать текстовые документы на естественном языке (ЕЯ), изображения и другие виды и форматы представления данных. Несмотря на то, что принципы хранения данных в системах обработки фактической и документальной (текстовой) информации схожи, алгоритмы обработки в них заметно различаются. Поэтому в зависимости от характера информационных ресурсов, которыми оперируют ИПС, принято различать два крупных класса систем – документальные (ДИПС) и фактографические (ФИПС).

В автоматизированных научно-технических информационных системах наиболее широкое применение получили ДИПС, что отражает факт ведущего положения документа в процессе распространения информации. Описание содержания документа с помощью информационно-поискового языка (ИПЯ) представляет собой поисковый образ документа (ПОД), а описание содержания запроса – поисковый образ запроса (ПОЗ). Правила составления поисковых образов документов и запросов являются правилами перевода текстов с естественного языка (ЕЯ) на ИПЯ. При наличии массива документов и соответствующих им ПОД поиск отвечающих запросу документов сводится к сопоставлению поисковых образов документов и запроса. Для оценки степени их соответствия формулируется *критерий смыслового соответствия*, который представляет собой пару, состоящую из меры формальной релевантности (меры близости) поисковых образов документа и запроса и порогового значения меры близости, при превышении которого документ признаётся формально релевантным соответствующему запросу. Известны следующие меры формальной релевантности: скалярное произведение векторов запроса и документа, мера Танимото (нормированное скалярное произведение), мера косинуса угла между векторами запроса и документа и др. [1]. Однако на практике формальная релевантность вовсе не означает содержательного соответствия выданного документа запросу. Если ИПЯ неточно выражает смысл документов и запросов, то может оказаться, что близкие по смыслу документы и запросы будут обладать разными поисковыми образами, и, наоборот, у далёких по смыслу документов поисковые образы окажутся сходными. В этом случае не все документы, формально соответствующие запросу, будут на самом деле соответствовать ему по смыслу, образуя информационный шум. По тем же причинам может оказаться, что часть документов, релевантных запросу, не будет выдана; в этом случае говорят о потерях информации. Есть также понятие *пертинентность* – соответствие документа информационной потребности, что фактически определяется только пользователем, знающим, чего он хочет.

В целях количественного описания уровня качества поиска ИПС исторически первыми были предложены следующие частные критерии оценки

системы: полнота – *r* (от *recall ratio*) и точность – *p* (от *precision ratio*).

Эффективность поиска информации в ИПС во многом зависит от качества запроса пользователя. В отличие от среды баз данных, в ИПС нет чёткого представления пользовательских запросов. Существует несколько способов повышения эффективности поиска путём модификации представления запроса.

Рассмотрим вариант оценки эффективности поиска ИПС на примере автоматизированной многоязычной базы данных (АМБД) КАСКАД, в разработке которой мы принимали непосредственное участие. Данная АМБД предназначена для обеспечения зарубежных пользователей информацией об отечественных программных продуктах. Информационное и лингвистическое обеспечение АМБД формирует базу данных (далее – БД) по программным средствам (ПС) на русском языке, осуществляет поиск данных по запросу на естественном языке (ЕЯ) (русском или английском) и производит перевод русскоязычной БД на английский язык.

Описания ПС выполняются в соответствии с анкетой, которая содержит сведения о наименовании ПС, его назначении, аннотацию, а также данные об организации-разработчике, годе окончания разработки, адресные данные, координаты для связи и пр. Основу АМБД составляет русский вариант БД по отечественным ПС, предназначенным для зарубежных пользователей. Средства доступа к ней обеспечивают возможность поиска по отдельным параметрам описания ПС с помощью подсказок и по запросам, сформулированным на ЕЯ (русском или английском). Между русскоязычной и англоязычной БД сохраняется полное взаимнооднозначное соответствие – все файлы имеют одинаковую структуру, состав, наименование, но хранятся в разных каталогах.

Использование СМП для получения БД на иностранном языке или автоматических словарей даёт возможность быстро осуществлять перевод данных на иностранные языки (что становится экономически выгодным для БД, насчитывающих сотни и тысячи документографических единиц), а также позволяет использовать единообразную научно-техническую терминологию, что, в свою очередь, существенно улучшает поисковые возможности системы. При установлении взаимнооднозначного соответствия между терминами русского и английского языков появляется возможность производить поиск в базе и выдавать его результаты на любом языке. Язык интерфейса и выходных документов определяется пользователем. Поскольку в системе используются автоматические словари, и структурой АМБД предусмотрена синонимичность файлов на различных языках, можно производить поиск по запросу на ЕЯ только в русскоязычной базе. При этом запрос может быть сформулирован на английском языке. Затем с помощью автоматического словаря он переводится на русский язык. После этого запрос обрабатывается по русскоязычной базе, а результаты поиска (по номерам найденных документов) выводятся на языке исходного запроса.

Лингвистическое обеспечение АМБД базируется на автоматическом выделении основ ключевых слов (дескрипторов) в текстовых полях (пол-

ное наименование и аннотация ПС). Для формирования словаря основ ключевых слов использовался словарь окончаний, разработанный при создании СМП АСПЕРА, и составленные предварительно вручную словари стоп-слов и стоп-основ, которые включают предлоги, союзы, наречия, некоторые глаголы и неинформативные существительные.

Для предоставления большей свободы при формулировке запроса частотный словарь пополняется словами-синонимами, которым ставится в соответствие одинаковая частота. Запрос на русском языке преобразуется с помощью частотного словаря в последовательность дескрипторов, после чего проверяется вхождение дескрипторов запроса в документы. Границы значений частот для формирования критерия выдачи определяются для высокочастотных, среднечастотных и низкочастотных основ.

При наличии частотных словарей основ ключевых слов критерием выдачи для выбора релевантных документов служит:

– вхождение слов запроса в документ;

– совпадение частот слов запроса в документе;

– попадание частот слов, входящих в запрос и в документ, в указанную категорию частоты (высокая, средняя, низкая).

Русско-английский словарь состоит из трёх частей:

– словаря основ русских слов;

– словаря английских эквивалентов слов;

– словаря словосочетаний.

Словарь русских основ состоит из записей, упорядоченных по алфавиту. Каждая запись представляет собой собственно основу слова, грамматическую характеристику (до семи наборов грамматических признаков без разделителей), относительный адрес перевода в словаре эквивалентов. В качестве основы слова может быть представлена квазиоснова или словоформа, если длина основы меньше 4-х символов.

В информационных файлах БД каждая запись (документ) представляет собой последовательность полей, каждое из которых является символьным текстом или кодом, значение которого приведено в соответствующем словаре. Идентификатором документа является первый код – регистрационный номер документа.

Как уже говорилось, поиск по запросам на ЕЯ производится только в русскоязычной базе. На первом этапе запрос преобразуется в последовательность дескрипторов (по частотному словарю основ ключевых слов). Если ни одного слова запроса в словаре не найдено, пользователю предлагается повторить формулировку запроса, воспользовавшись в качестве подсказки словарём дескрипторов. Запрос на английском языке предварительно переводится (пословно) на русский язык. Далее работа ведётся аналогично, но в качестве подсказки приводится английский словарь дескрипторов. Результаты поиска эшелонируются в соответствии с критерием выдачи (вхождение всех дескрипторов запроса в документ; получение пересечения номеров документов по всем дескрипторам; вхождение дескрипторов с высокой или низкой частотой и т. п.) [2].

Для количественной оценки качества поиска АИПС необходимо ис-

пользовать одновременно два частных критерия: критерий полноты и критерий точности поиска. Другим вариантом является использование интегральных критериев, которые представляют собой числовые оценки качества работы ИПС. При использовании частных критериев могут возникать проблемы, когда, например, у одной выдачи критерий полноты больше, чем критерий точности, а у другой, наоборот, критерий точности больше, чем критерий полноты. В этом случае совместное использование критериев приводит к противоречивой ситуации, когда невозможно отдать предпочтение одной из выдач. Решить указанную проблему позволяет нечёткая экспертная система, входными переменными которой являются критерии полноты и точности поиска, а выходной – уровень качества поиска тестируемой ИПС. В зависимости от предпочтений эксперта, составлявшего базу правил логического вывода системы, всегда будет сделан однозначный вывод о качестве конкретного поиска.

Как упоминалось выше, оценка качества ИПС КАСКАД проводилась вариантом нечёткой экспертной системы. Рассматриваемый вариант системы использует одновременно два критерия оценки качества: критерий полноты поиска и критерий точности поиска. Блок-схема экспертной системы показана на рисунке 1. Входные переменные описываются лингвистическими переменными (ЛП) «Коэффициент полноты» и «Коэффициент точности», выходная переменная описывается ЛП «Качество ИПС». Базовые терм-множества входных ЛП одинаковы и состоят из трёх нечётких переменных: для ЛП «Коэффициент полноты» – это $T_R = \{A_1 =$ небольшой, $A_2 =$ средний, $A_3 =$ большой$\}$, для ЛП «Коэффициент точности» – $T_P = \{B_1 =$ *н*ебольшой, $B_2 =$ средний, $B_3 =$ большой$\}$. Нечёткие переменные A_i определены на отрезке $X = (0, 1)$. Областью определения нечётких переменных B_i является отрезок $Y = (0, 1)$.

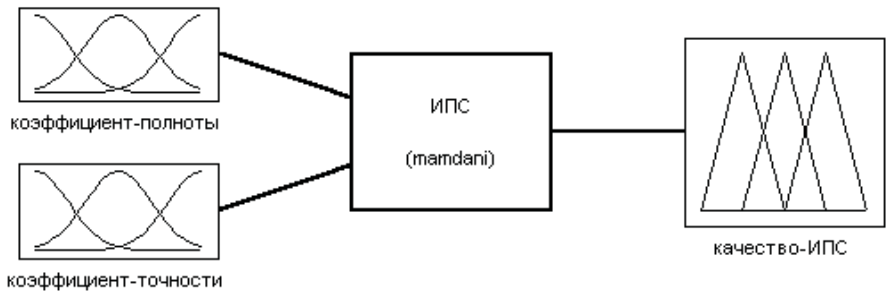

Рис. 1. Блок-схема нечёткой экспертной системы оценки качества ИПС.

Для получения оценки качества поиска ИПС были использованы два известных алгоритма нечёткого вывода: *Mamdani* и *Larsen*. В базе правил вывода входные переменные (предпосылки) объединяются с помощью связки «И», которая реализуется с помощью операции МИНИМУМ. Минимальное значение функций принадлежности нечётких подмножеств входных переменных, определяющих одно и то же нечёткое подмножество выходной

переменной, используется для его модификации при выполнении операции логического вывода (импликации). Композиция нечётких подмножеств заключений правил осуществлялась с помощью операции МАКСИМУМ. Приведение к чёткости выполнялось центроидным методом. Заметим, что база правил вывода реализует симметричный алгоритм, который даёт одинаковую оценку качества при значениях критериев полноты и точности равных, например, $R = a$ и $P = b$, и при значениях $R = b$ и $P = a$ (рис. 2).

В ИПС КАСКАД находилось на момент исследования 620 документов. Состав базы известен автору, поэтому тестирование проводилось на контрольном наборе документов. Был сформулирован первоначальный запрос на ЕЯ на выдачу документов следующего содержания: «Выдать информацию о разработчиках лингвистического обеспечения СМП». В

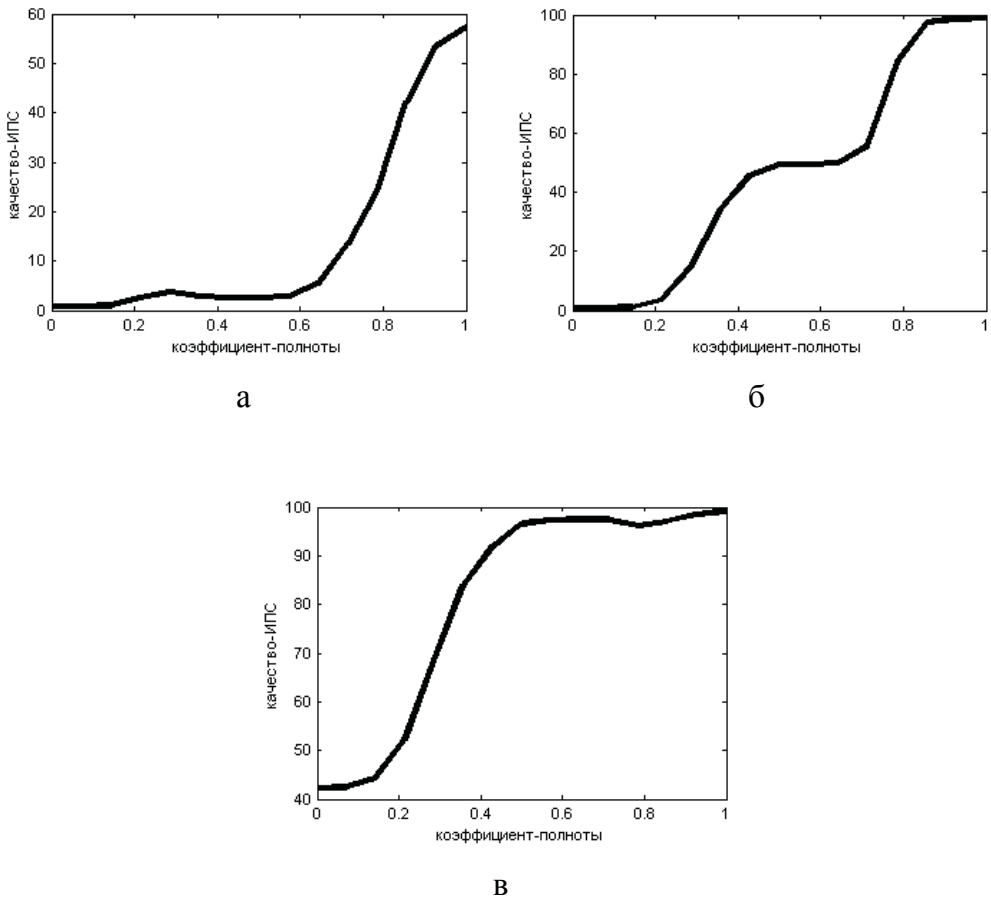

Рис. 2. Зависимость оценки качества ИПС от коэффициента полноты:
а) коэффициент точности «небольшой» (P = 0,2);
б) коэффициент точности «средний» (P = 0,5);
в) коэффициент точности «большой» (P = 0,9).

ответ на запрос было выдано 80 документов. При анализе запроса ИПС КАСКАД отбрасывает стоп-слова, т. е. неинформативные слова запроса; в нашем случае это «выдать информацию о…». Остальные слова запроса объединяются с помощью операции логического умножения (логическое И), происходит поиск в базе всех документов, в которые входят все слова запроса. Поскольку предлагаемая экспертная система оценки качества работает с коэффициентами полноты и точности, которые основаны на понятии релевантности, необходимо определить эти коэффициенты. Для этого эксперт проанализировал все выданные статьи на предмет их релевантности запросу. Оказалось, что было выдано 54 релевантных документа с точки зрения эксперта, таким образом, значение коэффициента точности составило P = 0,675. Всего в базе содержится 110 документов, посвящённых системам машинного перевода, следовательно, значение коэффициента полноты оказалось равным R = 0,49. Оценка качества ИПС КАСКАД составила Q = 54%.

Наличие информационного шума в выдаче, т. е. нерелевантных запросу статей объясняется тем, что одно лишь упоминание слов *разработчики*, *лингвистическое*, *обеспечение*, *СМП* в статье ещё не гарантирует её релевантности запросу. Повысить коэффициент точности, т. е. уменьшить количество выданных нерелевантных документов, можно путём применения специальной технологии формирования запроса к поисковой системе. Путём анализа нерелевантных документов в выдаче были определены ключевые слова, которые затем были включены в запрос в режиме отрицания (логическое НЕ). Такими словами оказались *цена*, *системные*, *требования*. В результате удалось добиться повышения коэффициента точности до значения P = 0,76 (общее число документов в выдаче составило 71 при 54 релевантных), коэффициент полноты не изменился. Оценка качества ИПС при этом возросла до Q = 69%. Теоретически методика отладки запроса позволяет достичь значений коэффициента точности, близких к 1.

Увеличить значение коэффициента полноты можно прежде всего путём добавления в запрос синонимов ключевых слов в режиме логического сложения (логическое ИЛИ). В нашем случае осуществлялось добавление следующих синонимов аббревиатуры *СМП*: *системы машинного перевода*, *программы-переводчики*, *системы МП*. Добавление словосочетания *системы машинного перевода* не изменило ситуации, так как в ИПС КАСКАД имеется небольшой словарь синонимов, и соответствующее словосочетание с самого начала принималось системой во внимание как синоним аббревиатуры *СМП*, имеющейся в запросе. Добавление синонимичного термина *программы-переводчики* улучшило ситуацию. В результате в ответ на запрос было выдано ещё 14 релевантных документов, причём количество нерелевантных документов в выдаче не изменилось. Отношение количества релевантных документов к общему числу документов в выдаче составило 68 к 85, т. е. коэффициент точности стал равен P = 0,8. Существенным образом удалось повысить значение коэффициента полноты, он стал равен R = 0,62. Оценка качества ИПС составила Q = 82%. Добавление синони-

ма *системы МП* вызвало следующий эффект: поскольку ИПС КАСКАД не имеет словаря словосочетаний, соответствующая фраза была интерпретирована как два слова, таким образом, поисковая система наряду с релевантными документами, в которые входят слова *система* и *МП*, выдала ряд нерелевантных документов, в которые входит слово *система* и которые не имеют отношения к системам машинного перевода. Следует заметить, что вновь выданных релевантных документов оказалось существенно больше, чем нерелевантных (27 и 6 соответственно). Таким образом, всего было выдано 118 документов, релевантных среди них стало 95. Коэффициенты полноты и точности оказались равными $R = 0{,}86$ и $P = 0{,}81$ соответственно. При этом качество системы было оценено как «хорошее», количественное значение составило $Q = 92\%$. Добиться дальнейшего повышения коэффициента полноты оказалось сложно, так как попытки его повышения приводили к уменьшению коэффициента точности (к выдаче большого количества нерелевантных документов), поэтому решено было остановиться на достигнутых значениях коэффициентов полноты и точности. Эксперимент, проведённый этим способом, был осуществлён и с другими вариантами запросов к ИПС (всего около 150), что дало в целом аналогичный результат.

Таким образом, можно сделать вывод о том, что оценка качества ИПС определяется не только техническими характеристиками самой ИПС (наличием словаря синонимов, словаря словосочетаний, мощностью базы данных), но, поскольку она зависит от формулировки запроса, оценка может быть улучшена путём модификации запроса. Проведённый анализ показал, что предлагаемая экспертная система может быть использована для формальной оценки качества поисковой системы.

ЛИТЕРАТУРА:

1. Корнеев В.В. и др. Базы данных. Интеллектуальная обработка информации / Корнеев В.В., Гареев А.Ф., Васютин С.В., Райх В.В. М.: Нолидж, 2000. 352 с.
2. Максименко О.И. Формализованная лингвистика. М.: Изд-во МГОУ, 2013. 189 с.