УДК 004.827

В.А. Крисилов, Е.А. Городничая

Одесский национальный политехнический университет, Одесса

ИНФОРМАЦИОННАЯ ТЕХНОЛОГИЯ ОЦЕНКИ И ПОВЫШЕНИЯ РЕЛЕВАНТНОСТИ РЕЗУЛЬТАТОВ ЗАПРОСОВ К БАЗАМ ДАННЫХ

Одним из наиболее часто встречающихся видов неопределенности является неопределенность временных характеристик описания объектов и формирования запросов к ним. Представленная информационная технология использует аппарат нечетких множеств для описания объектов и запросов к базам данных для облегчения поиска и группировки объектов по временным характеристикам. Также предложенная информационная технология оценки и улучшения релевантности результатов запросов позволяет количественно оценивает релевантность результатов запросов.

Ключевые слова: *нечеткие множества, релевантность, нечеткий запрос.*

Постановка проблемы

Объемы хранимой и обрабатываемой информации увеличивается экспоненциально. Это выдвигает особые требования к методам и средствам поиска и обработки информации [1, 2].

Одним из показателей, характеризующих качество поиска информации, является релевантность результатов запроса. Релевантность – семантическое соответствие поискового запроса и результата [3]. Релевантность характеризует степень соответствия содержания, найденного в результате информационного поиска, содержанию информационного запроса. В разных случаях релевантность вычисляется по-разному [4, 5, 6]. В данном случае, предлагается рассматривать релевантность, как количественную меру соответствия запроса и его результата. Невысокая релевантность выборки некоторого запроса является следствием неопределенности либо запроса, либо значений свойств объекта, по которому производится поиск.

При поиске объектов выделяют причины неопределенности двух видов: неопределенность запроса и неопределенность описания объекта [7]. К неопределенности запроса может относиться семантическая неопределенность текстовых данных, к неопределенности описания объекта – неточность измерений, неопределенность текстовых данных, погрешность обработки характеристик и др. Одним из наиболее часто встречающихся видов неопределенности является неопределенность описания временных характеристик объектов, например: сроки происхождения событий, датировки исторических экспонатов и т.д. Неопределенность описания временных характеристик объектов проявляется в случаях, когда для описываемых событий искусственно расширяется временной диапазон.

Анализ литературы. В [1] обсуждаются возможности применения прямого поиска при использовании мобильных телефонов для поиска информации в Интернете, предлагаемая стратегия поиска позволяет минимизировать выдачу релевантных документов и ранжировать ее для повышения эффективности и точности работы системы. В [2] рассматриваются различные средства и приемы поиска информации. В [3] рассматриваются основные факторы, влияющие на релевантность, один из алгоритмов определения релевантности документа запросу, влияние собственных ресурсов поисковых машин. В [4] рассматриваются актуальные методы вычисления релевантности фрагментов текста на основе анализа тематических моделей для последующего построения аннотаций в форме выдержек, т.е. аннотаций, полностью состоящих из последовательности фрагментов исходного текста, предложен новый метод вычисления релевантности фрагментов текста, основанный на оценке весов тематик в нормализованном пространстве тематик, получаемом с помощью факторизации неотрицательных матриц, которая используется в качестве матричного разложения в модели латентносемантического анализа. В [5] рассматривается подход к поиску решений в базах знаний с использованием метаданных документа, релевантность документа оценивается совокупностью метрик, формализующих близость указанных семантических сетей. В [6] предложен метод оценки степени релевантности текстового ответа в компьютерных обучающих системах. В [7] рассматриваются нечеткие запросы к базам данных, неопределенность запроса и неопределенность описания объекта. В [8] представлены основы нечеткой логики.

В данной работе предлагается использовать аппарат нечетких множеств для описания объектов и запросов к базам данных для улучшения релевантности.

Цель статьи – разработка информационной технологии для количественной оценки релевантности результатов запросов.

Описание временных характеристик

Часто лишь приблизительно известно, когда произошло интересуемое событие. От правильности

описания временной характеристики исторического объекта зависит дальнейшее представление исторических событий. Нечеткость описания временной характеристики, а также использование различных форматов при описании объекта затрудняет дальнейший анализ и поиск временного промежутка исторических событий.

Для описания временных характеристик используются различные форматы: указание точной даты/времени, например: 19 марта 1946 года; указание временного интервала, например: 336 г. до н. э. – 323 г. до н. э.; использование различных терминов с разной степенью подробности, например: вторая половина III ст. д.н.э, последняя треть II века д.н.э. Такое описание временных характеристик существенно затрудняет или делает невозможным поиск и группирование объектов по временным характеристикам. Для решения проблемы предлагается описывать временные характеристики объектов и запросов в виде нечетких переменных.

Под нечеткой переменной объекта будем понимать тройку (PO, T, MTo), где PO – название переменной, T – универсальное множество, MTo – нечеткое подмножество множества T. Под нечеткой переменной запроса будем понимать тройку (PZ, T, MTz), где PZ – название переменной, T – универсальное множество, MTz – нечеткое подмножество множества T.

Нечеткое множество временных характеристик MT определяется как множество упорядоченных пар MT={μMT(t)/t}, где MT – нечеткое множество временных характеристик, μMH(t) – функция принадлежности, t – временная характеристика. [8]

Характеристическая функция принадлежности в большинстве случаев имеет трапециевидную форму (рис. 1). Чем меньше разница между a и b, а также c и d, тем ближе нечеткая переменная к четкой. В случае, если нечеткая переменная становится четкой, то функция принадлежности принимает прямоугольный вид, при этом a=b и c=d. В большинстве случаев, если временная характеристика задана с максимальной нечеткостью, функция принадлежности принимает треугольную форму, при этом b=c. Т.е., при сравнении треугольной и трапециевидный функций, в случае если треугольная и трапециевидная функция покрывают одинаковый временной диапазон, треугольная функция будет иметь большую неопределенность.

Оценка релевантности запроса и результата

Выделим несколько видов соответствия найденного объекта запросу: найденный объект полностью не соответствует требованию, найденный объект полностью соответствует требованию, найденный объект соответствует требованию частично.

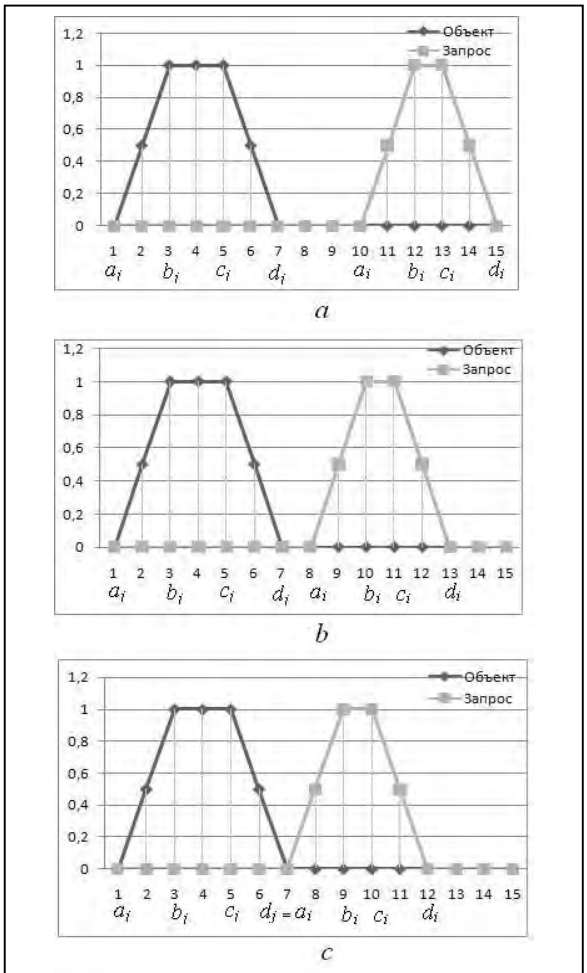


Рис. 1. Функции, в которых объект полностью не соответствует запросу: a – DR=5; b – DR=3; c – DR=2

1. Найденный объект полностью не соответствует требованию

Это происходит, когда в результате запроса не найдено ни одного объекта, который совпадал бы с запросом хотя бы по одному значению, т.е. функции объекта и запроса не пересекаются. В данном случае предлагается вычислять степень удаленности найденного объекта и запроса по формуле (1).

$$DR = \frac{(|\ b_i - c_j\ | + |\ a_i - d_j\ |)}{2}, \qquad (1)$$

где DR – степень несоответствия найденного объекта запросу; i – коэффициент, который указывает, что временные характеристики принадлежат нечеткой переменной запроса; j – коэффициент, который указывает, что временные характеристики принадлежат нечеткой переменной объекта; a_i, b_i, c_i, d_i – параметры нечеткой переменной запроса, для которых выполняется условие $a_i \le b_i \le c_i \le d_i$; a_j, b_j, c_j, d_j – параметры нечеткой переменной объекта, для которых выполняется условие $a_j \le b_j \le c_j \le d_j$.

Чем больше степень удаленности найденного объекта и запроса, тем больше найденный объект не соответствует запросу.

2. Найденный объект полностью соответствует требованию

Это происходит, когда в результате запроса найден объект, который совпадает с запросом по всем значениям, т.е. объект полностью соответствует запросу.

3. Найденный объект соответствует требованию частично

Найденный объект соответствует требованию частично:

3.1. Запрос полностью поглощает найденный объект, т.е. в результате запроса найден объект, который совпадает с запросом по всем значениям объекта, но при этом в запросе есть значения, которых нет в найденном объекте. Это может происходить в случае, когда запрос обладает высокой неопределенностью либо объект задан более точно, по сравнению с запросом.

3.2. Найденный объект полностью поглощает запрос, т.е. в результате запроса найден объект, который совпадает с запросом по всем значениям запроса, но при этом имеет значения, которых нет в запросе. Это может происходить в случае, когда объект обладает высокой неопределенностью либо запрос задан более точно, по сравнению с объектом.

3.3. Найденный объект частично перекрывает запрос т.е. в результате запроса найден объект, который совпадает с запросом по некоторым значениям запроса. В случаях, когда найденный объект соответствует требованию частично, предлагается вычислять релевантность по формуле (2).

$$P = S_I / S_{NI}, \qquad (2)$$

S_{NI} – площадь области, в которой объект и запрос не пересекаются, и которая находится между объектом и запросом.

Проведем ряд преобразований:

$$P = \frac{S_I}{S - S_I} = \frac{S_I}{S_{PO} + S_{PZ} - 2S_I} =$$

$$= \frac{\frac{d_k - a_k + c_k - b_k}{2}}{\frac{d_j - a_j + c_j - b_j}{2} + \frac{d_i - a_i + c_i - b_i}{2} - 2\left(\frac{d_k - a_k + c_k - b_k}{2}\right)} =$$

$$= \frac{d_k - a_k + c_k - b_k}{d_j - a_j + c_j - b_j + d_i - a_i + c_i - b_i - 2d_k + 2a_k - 2c_k + 2b_k}, \qquad (3)$$

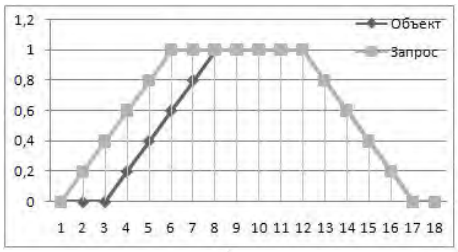
где S_{PO} – площадь объекта; S_{PZ} – площадь запроса; S – площадь области, которая покрывает объект и запрос; a_k, b_k, c_k, d_k – параметры пересекающейся области, для которых выполняется условие $a_k \le b_k \le c_k \le d_k$..

Таким образом, релевантность результата запроса будем рассчитывать следующим образом:

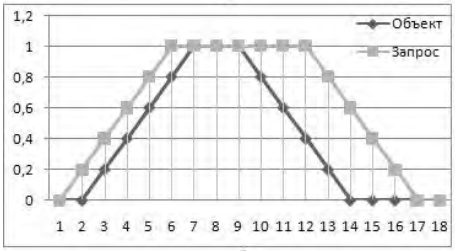
$$P = \frac{d_k - a_k + c_k - b_k}{d_j - a_j + c_j - b_j + d_i - a_i + c_i - b_i - 2d_k + 2a_k - 2c_k + 2b_k}. \qquad (4)$$

Чем меньше релевантность, тем меньше найденный объект соответствует запросу.

На рис. 2 представлены функции, в которых запрос полностью покрывает объект. На рис. 2, a площадь пересечения объекта и запроса больше, чем на рис. 2, b, т.к. запрос полностью покрывает объект, а площадь объекта на рис. 3, a больше площади объекта на рис. 2, b. Кроме того, площадь, где объект и запрос не пересекаются, на рис. 2, a меньше, чем на рис. 2, b. Таким образом, чем больше площадь пересечения объекта и запроса и меньше площадь, на которой объект и запрос не пересекаются, тем лучше релевантность.



a



b

Рис. 2. Функции, в которых запрос полностью покрывает объект:

$$a - P = \frac{17 - 3 + 12 - 8}{17 - 3 + 12 - 8 + 17 - 1 + 12 - 6 - 32 + 6 - 24 + 16} = \frac{18}{4} = 4,5;$$

$$b - P = \frac{14 - 2 + 9 - 7}{14 - 2 + 9 - 7 + 17 - 1 + 12 - 6 - 28 + 4 - 18 + 14} = \frac{14}{8} = 1,75;$$

На рис. 3 представлены функции, в которых объект полностью покрывает запрос. На рис. 3, a площадь пересечения объекта и запроса больше, чем на рис. 3, b. Кроме того, площадь, где объект и запрос не пересекаются, на рис 3, a меньше, чем на рис. 3, b. При этом площадь, на которой объект и запрос не пересекаются, на рис. 3, b больше, чем на рис. 2, b, поэтому релевантность запроса, показанная на рис. 3, b, хуже, чем на рис. 2, b.

На рис. 4 представлены функции, в которых объект частично покрывает запрос. Самой лучшей релевантностью, из представленных примеров, обладает запрос, который отображен на рис. 4, e, т.к. у него самая большая площадь пересечения объекта и запроса, а также самая маленькая площадь, на которой объект и запрос не пересекаются. Самой худшей релевантностью обладает запрос, который отображен на рис. 4, d, так как у него самая маленькая площадь пересечения объекта и запроса, а также самая большая площадь, на которой объект и запрос не пересекаются. На рис. 4, a,

b площади пересечения объекта и запроса одинаковые, но на рис. 4, а релевантность лучше, т.к. площадь, на которой объект и запрос не пересекаются, на рис. 4, а гораздо меньше, чем на рис. 4, b.

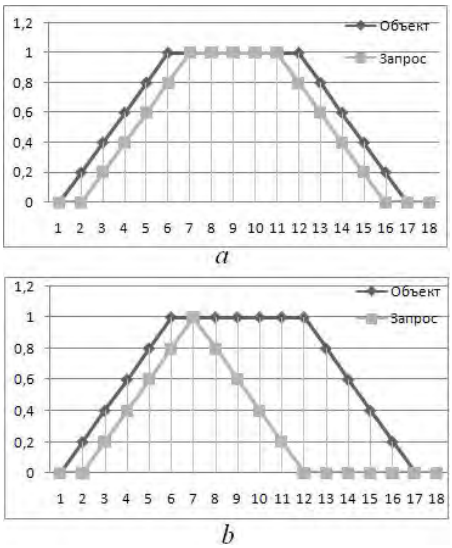


Рис. 3. Функции, в которых объект полностью покрывает запрос:

$$a - P = \frac{16-2+11-7}{17-1+12-6+16-2+11-7-32+4-22+14} = \frac{18}{4} = 4,5;$$

$$b - P = \frac{12-2+7-7}{17-1+12-6+12-2+7-7-27+4-14+14} = \frac{10}{12} = 0,83;$$

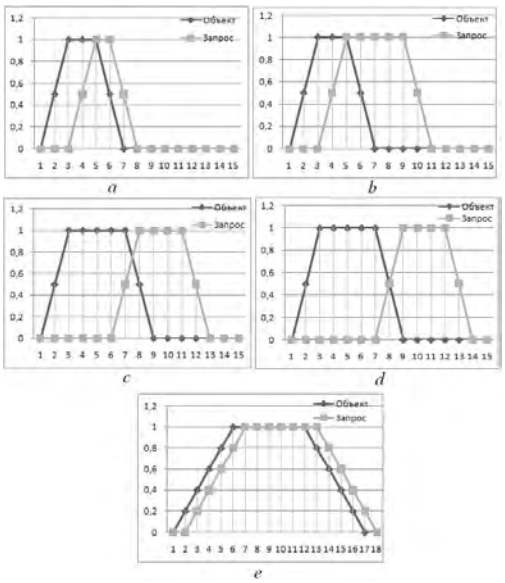


Рис. 4. Функции, в которых объект частично покрывает запрос: a – P=0,67; b – P=0,33; c – P=0,19; d – P=0,1; e – P=5

Выводы

В данной работе были рассмотрены частные случаи соответствия объекта запросу. Для объектов, которые полностью не соответствуют предъявленным требованиям, предлагается вычислять степень удаленности найденного объекта и запроса по формуле (1). Результаты исследований подтверждают, что чем больше степень удаленности найденного объекта и запроса, тем больше найденный объект не соответствует запросу. Для объектов, которые соответствуют требованию частично, предлагается вычислять релевантность по формуле (4). В ходе исследования соответствия запроса объекту было выяснено, что чем меньше релевантность, тем меньше найденный объект соответствует запросу.

Представленная информационная технология использует аппарат нечетких множеств для описания объектов и запросов к базам данных для облегчения поиска и группировки объектов по временным характеристикам, а также позволяет количественно оценивает релевантность результатов запросов.

Список литературы

1. Лукина А.Г. Требования к системам поиска информации в интернете при использовании мобильного телефона в качестве оконечного устройства / А.Г. Лукина // Научно-техническая информация. Серия 1: Организация и методика информационной работы – №8. – М.: Всероссийский институт научной и технической информации, 2007. – С. 23-26.

2. Капустин В.А. Основы поиска информации в Интернете. Методическое пособие / В.А. Капустин. – СПб.: Институт «Открытое общество». Санкт-Петербургское отделение, 1998. – 13 с.

3. Людкевич С.А. Основные факторы, влияющие на релевантность. [Электронный ресурс] / С.А. Людкевич, Е.С. Есипов. – «Промо Текарт», 2003. – Режим доступа: http://www.promo-techart.ru/analysis/relevants.htm.

4. Машечкин И.В. Методы вычисления релевантности фрагментов текста на основе тематических моделей в задаче автоматического аннотирования / И.В. Машечкин, М.И. Петровский, Д.В. Царев // Вычислительные методы и программирование. – Т.14. –М.: Вычислительные методы и программирование, 2013. – С. 91-102.

5. Карпенко А.П. Многокритериальная оценка релевантности документов корпоративной онтологической базы знаний на основе их ролевой кластеризации / А.П. Карпенко, В.А. Трудоношин // Наука и образование. МГТУ им. Н.Э. Баумана. Электрон. Журн. № 11. – М.: МГТУ им. Н.Э. Баумана, 2013 – С. 311–328. – [Электронный ресурс]. – Режим доступа: http://technomag.bmstu.ru/doc/637857.html.

6. Бадерина Л.Н. Метод оценки степени релевантности текстового ответа в компьютерных обучающих системах/ Л.Н. Бадёрина // Вісник національного авіаційного університету. – Т.1. – К.: Национальный авиационный университет, 2007. – С. 70-72.

7. Коновалов Д.П. К вопросу нечётких запросов к реляционным базам данных / Д.П. Коновалов // Перспективы развития информационных технологий №2. – Новосибирск: общество с ограниченной ответственностью "Центр развития научного сотрудничества", 2010 – С. 87-92.

8. Макеева А.В. Основы нечеткой логики. Учебное пособие для вузов / А.В. Макеева. – Н.Новгород: ВГПУ, 2009. – 59 с.

Поступила в редколлегию 10.06.2014

Рецензент: канд. техн. наук, доцент Р.О. Шапорин, Одесский национальный политехнический университет, Одесса.

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ОЦІНКИ ТА ПІДВИЩЕННЯ РЕЛЕВАНТНОСТІ РЕЗУЛЬТАТІВ ЗАПИТІВ ДО БАЗ ДАНИХ

В.А. Крісілов, К.О. Городнича

Одним з видів невизначеності, що найбільш часто зустрічаються, є невизначеність часових характеристик опису об'єктів і формування запитів до них. Представлена інформаційна технологія використовує апарат нечітких множин для опису об'єктів і запитів до баз даних для полегшення пошуку і угруповання об'єктів за часовими характеристиками. Також запропонована інформаційна технологія оцінки та покращення релевантності результатів запитів дозволяє кількісно оцінює релевантність результатів запитів.

Ключові слова: *нечіткі множини, релевантність, нечіткий запит.*

INFORMATION TECHNOLOGY ASSESSMENT AND IMPROVEMENT RELEVANT RESULTS OF DATABASE QUERIES

V.A. Krisilov, E.A. Gorodnichaya

One of the most common types of uncertainty is the uncertainty of the temporal characteristics of describing objects and querying them. The information technology represented uses fuzzy sets to describe objects and queries to databases in order to facilitate searching and grouping objects by time characteristics. The proposed information technology for assessment and improvement of the relevance of query results allows to quantitatively evaluates the relevance of query results.

Keywords: *fuzzy sets, relevance, fuzzy query.*