УДК 004.827

ИНФОРМАЦИОННАЯ ТЕХНОЛОГИЯ ОЦЕНКИ И ПОВЫШЕНИЯ РЕЛЕВАНТНОСТИ РЕЗУЛЬТАТОВ ЗАПРОСОВ

Д. т. н. В. А. Крисилов, Е. А. Городничая

Одесский национальный политехнический университет
Украина, г. Одесса
katherine.gorodnichaya@ukr.net

Одним из наиболее часто встречающихся видов неопределенности является неопределенность временных характеристик описания объектов и формирования запросов к ним. Предложена информационная технология оценки и повышения релевантности результатов запросов с помощью их количественной оценки.

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

Объемы хранимой и обрабатываемой информации увеличивается экспоненциально. Это выдвигает основные требования к методам и средствам поиска и обработки информации.

Одним из показателей, характеризующих качество поиска информации, является релевантность результатов запроса – семантическое соответствие поискового запроса и результата [1]. Релевантность характеризует степень соответствия содержания, найденного в результате информационного поиска, содержанию информационного запроса. В разных случаях релевантность вычисляется по-разному. В данном случае предлагается рассматривать релевантность как количественную меру соответствия запроса и его результата. Невысокая релевантность некоторого запроса является следствием неопределенности либо запроса, либо значения свойств объекта, по которому производится поиск.

Специалисты выделяют различные причины неопределенности при поиске объектов: неопределенность запроса и неопределенность описания объекта, например семантическая неопределенность текстовых данных, неточность измерений, погрешность обработки количественных характеристик и некоторые другие.

Одним из наиболее часто встречающихся видов неопределенности является неопределенность временных характеристик описания объектов, например сроки выполнения договоров, сроки происхождения событий, датировки исторических экспонатов и т. д. Это проявляется в виде искусственного расширения временных диапазонов, описываемых событий.

В данной работе предлагается использовать аппарат нечетких множеств для описания объектов и запросов к базам данных для повышения релевантности. Целью данной работы является 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 до н. э. — 10 июня 323 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 этой 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.

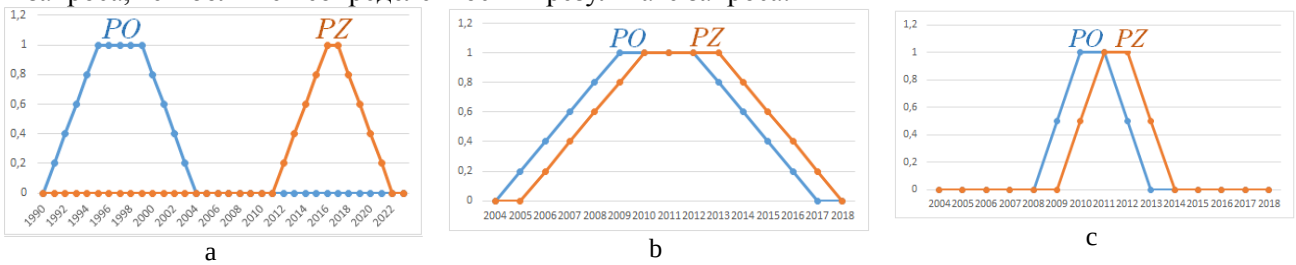
Нечеткое множество временных характеристик MT определяется как множество упорядоченных пар $MT=\{\mu_{MT}(t)/t\}$, где MT – нечеткое множество временных характеристик, $\mu_{MH}(t)$ – функция принадлежности, t – временная характеристика [2].

Релевантность результата запроса

В данном случае предлагается рассматривать релевантность как количественную меру соответствия нечеткой переменной запроса и нечеткой переменной объекта, и для его количественного измерения использовать расстояние Евклида [3]

$$e(PO,PZ)=\sqrt{\sum_{i=1}^{n}(\mu_{MTo}(t_i)-\mu_{MTz}(t_i))^2}. \quad (1)$$

Чем меньше расстояние между функциями объекта и запроса, тем лучше релевантность. Объект и запрос полностью совпадают в случае когда *e(PO,PZ)*=0. Чем меньше расстояние между PO и PZ, тем лучше релевантность результата запроса. Чем больше площадь пересечения функций объекта и запроса, тем больше неопределенность в результате запроса.



a b c

Рис. 1. Функции соответствия объекта запросу:
a – e(PO,PZ)=3,4; *b – e(PO,PZ)*=0,6; *c – e(PO,PZ)*=1

На рис. 1, a функции объекта (PO) и запроса (PZ) не пересекаются, т. е. найденный объект не соответствует запросу, в данном случае его репрезентативность равна 3,4. На рис. 1, b и рис. 1, c функции запроса и объекта пересекаются, т. е. запрос частично соответствует найденному объекту. На рис. 1, b репрезентативность лучше, чем на рис. 1, c, т. к. на рис. 1, b больший объем данных объекта соответствует данным запроса, а репрезентативность зависит от площади пересечения функций объекта и запроса, но при этом на рис. 1, b больше неопределенность запроса и объекта, по сравнению с рис. 1, c.

Представленная информационная технология количественно оценивает релевантность результатов запросов по расстоянию Евклида, а также использует аппарат нечетких множеств для описания объектов и запросов к базам данных для повышения релевантности.

ИСПОЛЬЗОВАННЫЕ ИСТОЧНИКИ

1. Капустин В. А. Основы поиска информации в Интернете. Методическое пособие.– СПб.: Институт «Открытое общество». Санкт-Петербургское отделение, 1998.– 13 с.
2. Макеева А. В. Основы нечеткой логики. Учебное пособие для вузов.– Н. Новгород: ВГПУ, 2009.– 59 с.
3. Рыжов А. П. Элементы нечетких множеств и ее приложений – Москва, 2003.– 81 с.

V. A. Krisilov, K. A. Gorodnichaya

Information technology for assessment and increase of the relevance of query results.

One of the most common types of uncertainty is the uncertainty of the temporal characteristics of describing objects and querying them. The authors propose an information technology for assessment and increase of the relevance of the query results by using quantitative assessment relevance of query results.

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