UDC 004.827

V.A. Krisilov, DEng, Professor,
E.A. Gorodnichaya,
Odessa National Polytechnic University

METHODS OF ASSESSING THE DATABASE QUERIES' RESULTS RELEVANCE

Introduction. Regularly observed is that the volumes of stored and processed data increase exponentially. This implies special requirements to the methods and tools of data searching and processing [1].

The relevance of query results represents one of indicators characterizing the quality of information retrieval. The notion of relevance does mean semantic matching between the search query and the result [2]. The relevance characterizes the extent to which the content found as a result of information retrieval, does satisfy respective information request. In various cases the relevance calculation approaches do differ [3…5]. Herein, we propose to consider the relevance as a quantitative measure of the search result compliance to the query. Low relevance of some query sample is a consequence of the uncertainty of request or the searched object's parameters values**.**

When searching objects, we do face two types of uncertainty causes: query uncertainty and object description uncertainty [6]. The query uncertainty may include semantic ambiguity of the text data and the object description uncertainty corresponds to the measurement uncertainty, text data uncertainty, characteristics processing error etc**.** One of the most common types of uncertainty is the uncertainty of objects' temporal characteristics description, e.g. dates of events, history exhibits dating, etc. Uncertainty of objects' temporal characteristics description does reveal in the cases where the events' time range is artificially expanded**.**

Analysis of recent research and publications. In [1] exposes the discussion on possibility of direct search using mobile phones to find some information on the Internet. The proposed search strategy allows to minimize the relevant documents' total volume and to rank the found documents aiming onto the system efficiency and accuracy improvement. The [2] examines the main factors influencing the relevance, closely considering one of the algorithms to determine the relevance of a document to the request formulated and the impact of search engines' own resources. The source [3] discusses the current methods of text fragments' relevance calculating on the basis of case models' analysis for the subsequent annotations construction in the form of extracts, i.e. annotations, consisting entirely of original text fragments sequence. Suggested is a new method of calculating the text fragments' relevance based on an assessment of the subjects' balance within the normalized subjects' space, obtained through non-negative matrices factorization, (used as the matrix decomposition in the latent semantic analysis model). The [4] is devoted to seeking an approach to finding solutions at knowledge bases using document metadata, when the document's relevance is estimated with a set of metrics that formalize these semantic networks' proximity. In [5] proposes a method for assessing the text response relevance in computer-based training systems. In [6] considered are the fuzzy database queries, query uncertainty and object description uncertainty**.**

The Aim of the Research consisted in developing a methodology to quantify the query results relevance. Proposed is to use fuzzy sets when describing objects and database queries to facilitate the relevance evaluation.

Main Body.

Describing temporal characteristics to evaluate the query relevance. So often only approximately known is when the searched event has occurred. The historical object's temporal characteristics cor-

DOI 10.15276/opu.1.45.2015.20

ISSN 2076-2429 (print)
ISSN 2223-3814 (online)

rect description essentially influences the historical events further representation. Both an unclear description of the temporal characteristics, and the use of different formats in the object description are hindering further analysis, search and evaluation of historical events' time period.

To describe the temporal characteristics various formats are used: an exact date / time, e.g., March 19, 1946; a time interval, e.g., 336…323 BC; various terms with different degrees of detail, e.g., the second half of 3rd century BC, the last third of 2nd century BC. Such temporal characteristics description makes difficult or ever impossible objects' searching and grouping by time characteristics.

To solve the problem, proposed is to describe the temporal characteristics of objects and queries in the form of fuzzy variables.

Here we admit (*PO*, *T*, *MTo*) set as the object's fuzzy variable, where *PO* — variable's name, *T* — universal set, *MTo* — fuzzy subset of *T* set. The query fuzzy variable correlates to (*PZ, T, MTz*) set, where *PZ* — variable's name, *T* — universal set, *MTz* — fuzzy subset of *T* set.

The fuzzy set of time characteristics *MT* is defined as a set of ordered pairs $MT=\{\mu_{MT}(t)/t\}$, where *MT* — fuzzy set time characteristics, $\mu_{MT}(t)$ — membership function, *t* — time response [7].

The characteristic membership function in most cases has a trapezoidal shape (Fig. 1). The smaller is values' difference between *a* and *b* temporal characteristics as well as *c* and *d*, the closer is the given fuzzy variable to the crisp one. If fuzzy variable becomes crisp one, the membership function takes a rectangular form, with $a=b$ and $c=d$. In most cases, the time characteristics getting a maximum fuzziness, the membership function takes a triangular shape, with $b=c$. I.e. comparing a triangular and a trapezoidal functions, provided they do cover the same time span, the triangular function has a larger uncertainty.

Evaluating the query and result relevance.

We shall distinguish key relevance types according to the type of object found upon request: the object is not fully consistent with the requirement subject; the object is fully compliant; the object partially corresponds to the query.

1. *The found object is completely inconsistent with the query* (Fig. 1). This occurs when the query does not result in finding any object which coincides with the request's at least one value, i.e. the functions of the object and the query does not intersect. In this case proposed is to calculate the degree of remoteness between the found object and the query:

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

where *DR* — degree of divergence between the found object and the request parameters;

i — coefficient indicating that the query temporal characteristics belong to the request's fuzzy variable;

j — coefficient indicating that the temporal characteristics belong to the object's fuzzy variable;

a_i, b_i, c_i, d_i — parameters of query fuzzy variable, satisfying the condition $a_i \le b_i \le c_i \le d_i$.

a_j, b_j, c_j, d_j — parameters of object fuzzy variable, satisfying the condition $a_j \le b_j \le c_j \le d_j$.

The greater is the divergence/remoteness between the found object and the query, the less such found object does match the respective query.

2. *The found object is completely consistent with the query*. This occurs when the query results in finding an object coinciding with all request's parameters i.e. the object is fully consistent with the query.

3. *The found object is partially consistent with the query:*

— The query fully absorbs the found object, i.e. the query resulted in finding an object that matches the request by all object parameters, but the request contains some parameters not represented with the found object. That can be due to the case when high uncertainty request formulated either the object has more precisely defined parameters than these requested.

— The found object does completely absorb the request, i.e. the query resulted in finding an object that matches the request by all parameters, but contains some parameters not represented at the request. This can occur when the object has a high uncertainty or the request has been more accurately formulated than the object's features.

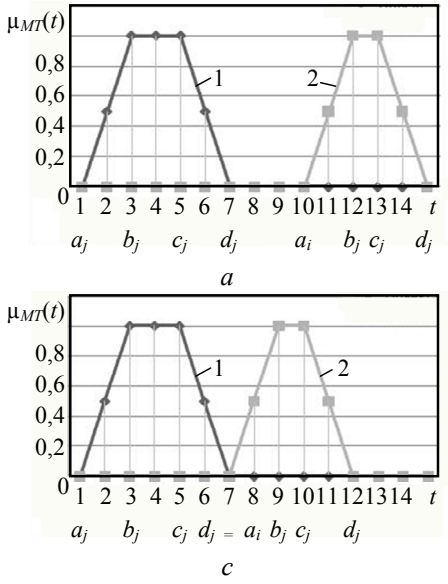

Fig. 1. Functions where the found object is completely inconsistent with the query: 1 — Object; 2 — Query; DR=5 (a); DR=3 (b); DR=2 (c);

— The found object does partially overlap the request, i.e. the query object found coincides with the query by several values of the request. In these cases when the found object corresponds to the requirement only partially, proposed is to calculate relevance as

$$P = \frac{S_I}{S_{NI}},$$

where P — relevance;

S_I — area of object-to-query intersection;

S_{NI} — area of non-coincidence region, located between the object and the query.

Now we proceed to series of transformations:

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

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

where S_{PO} — object area;

S_{PZ} — query area;

S — area of region covering both the object and the query;

a_k, b_k, c_k, d_k — parameters of intersection region satisfying the condition $a_k \le b_k \le c_k \le d_k$.

Therefore the query result relevance will be

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

The smaller is the relevance factor the lesser would be found object-to-query compliance index.

Fig. 2 shows the functions that completely covers the query object. In Fig. 2, *a* the area of query-to-object intersection is larger than in Fig. 2, *b*, as the request does completely cover the object, and the object area in Fig. 3, *a* is larger than the object area in Fig. 2, *b*. Additionally, the area where the object and the request do not intersect, in Fig. 2, *a* is smaller than in Fig. 2, *b*. Thus, the larger is the query-object intersection area and lesser the area in which the object and the request don't intersect the better relevance will be found.

ISSN 2076-2429 (print)
ISSN 2223-3814 (online)

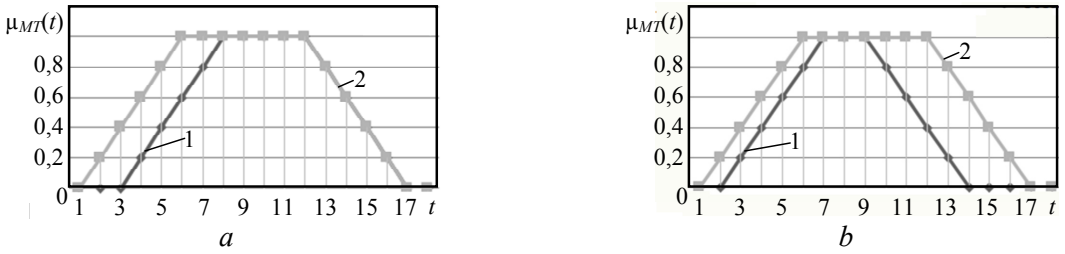

Fig. 2. Functions where the found object is completely consistent with the query: 1 — Object; 2 — Query;

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

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

The Fig. 3 shows the functions, where the object does completely cover the query. In Fig. 3, *a* the area of query-to-object intersection is larger than in Fig. 3, *b*. Additionally, the area where the object and the request do not intersect, in Fig. 3, *a* is smaller than in Fig. 3, *b*. The area on which the object and the request do not intersect, in Fig. 3, *b* is greater than in Fig. 2, *b*, thus the query relevance shown in Fig. 3, *b* is worse than in Fig. 2, *b*.

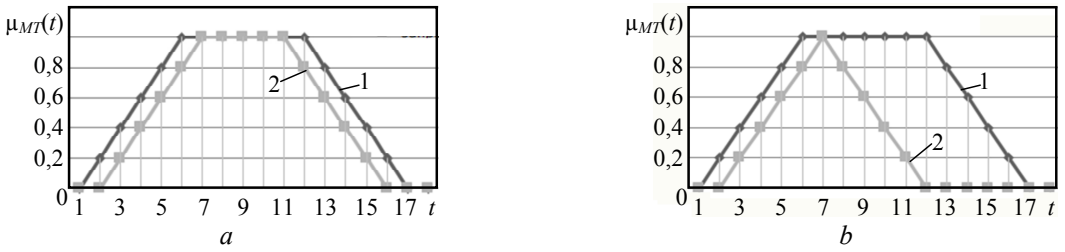

Fig. 3. Functions where the found object is completely consistent with the query: 1 — Object; 2 — Query;

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

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

Fig. 4 shows the functions, at which the object is partially covering the query. The best relevance of the examples presented, is attribute to the query represented in Fig. 4, *e* as it has the largest area of object-to-request intersection as well as the smallest area in which the object and the request do not intersect. The worst relevance case in a query displayed in Fig. 4, *d*, as it has the smallest area of the object-to-request intersection as well as the biggest area in which the object and the request do not intersect. At Fig. 4, *a* and Fig. 4, *b* the intersection areas are the same, but in Fig. 4, *a* the relevance is better, since the area in which the object and the request do not intersect, in Fig. 4, *a* is much less than in Fig. 4, *b*.

Results. In this paper some particular cases of the correspondence between object and query are considered. For objects not fully complying with the required parameters, it is proposed to calculate the degree of the found object's and request's remoteness by the formula (1). The results confirm that the larger is the distance between the found object and the request, the greater is the degree of the found object's non-matching to the request. For objects partially compliant to the request, it is proposed to calculate the relevance using formula (2). The research evidenced that the less relevant query object is, the lesser such found object corresponds to the request. The effected study includes a search by the archaeological museum's exhibits that relate to the ancient department (Ancient Greece). Upon request, it was necessary to find artifacts dated of the 3rd century BC. As a result the whole found sam-

pling of thirteen objects included two objects, fully complying with the request: Terracotta 'Tanagra' figure of a woman wearing a sunhat (3rd century BC) and Red-figure Pelike. Attica (330-320 BC), and two objects that partially match the request: Aphrodite. Terracotta (4th – 3rd century BC) and Vessel in the form of a horse's head (3rd – 2nd century BC). Untrained users who conducted an automated search of objects spent about 3 minutes on familiarization with the search principle, filling the query data and search properly.

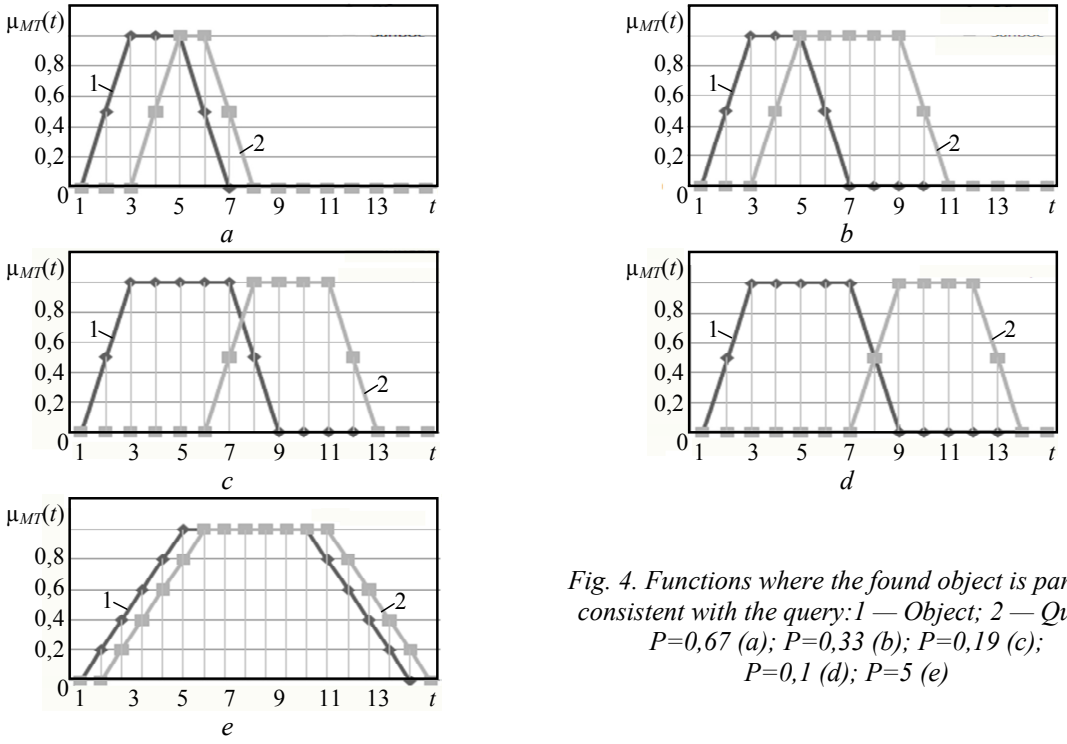

Fig. 4. Functions where the found object is partially consistent with the query:1 — Object; 2 — Query; P=0,67 (a); P=0,33 (b); P=0,19 (c); P=0,1 (d); P=5 (e)

Conclusions. In this paper described is the methodology of quantifying the query results relevance. The suggested information technology uses fuzzy sets to describe objects and databases query to facilitate searching and objects grouping by temporal characteristics, as well as the evaluation of query results relevance. This methodology includes a description of the three types of found object's compliance to the query: the found object is fully inconsistent with the request, the found object is fully compliant, the found object does partially correspond to the request. The presented method allows quantitative evaluation of the queries results' quality; for objects that are fully inconsistent with the required specification, calculated is the degree of remoteness between the found object and the request; for objects partially matching the request, calculated is the relevance index.

Література

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. — 2007. — № 8. — С. 23 — 26.
2. Людкевич, С.А. Основные факторы, влияющие на релевантность [Электронный ресурс] / С.А. Людкевич, Е.С. Есипов. — 2003. — Режим доступа: http://www.promo-techart.ru/analysis/relevants.htm (Дата обращения: 15.09.2014).
3. Машечкин, И.В. Методы вычисления релевантности фрагментов текста на основе тематических моделей в задаче автоматического аннотирования / И.В. Машечкин, М.И. Петровский, Д.В. Царёв // Вычислительные методы и программирование: новые вычислительные технологии. — 2013. — Т. 14, № 1. — С. 91 — 102.

ISSN 2076-2429 (print)
ISSN 2223-3814 (online)

4. Карпенко, А.П. Многокритериальная оценка релевантности документов корпоративной онтологической базы знаний на основе их ролевой кластеризации [Электронный ресурс] / А.П. Карпенко, В.А. Трудоношин // Наука и образование. — 2013. — № 11. — С. 311 — 328. — Режим доступа: http://dx.doi.org/10.7463/1113.0637857 (Дата обращения: 15.09.2014).
5. Badorina, L.N. Method of the relevance degree estimation of the text answer in computer training systems / L.N. Badorina // Вісник НАУ. — 2007. — Т. 31, № 1. — С. 70 — 72.
6. Коновалов, Д.П. К вопросу нечётких запросов к реляционным базам данных / Д.П. Коновалов // Перспективы развития информационных технологий. — 2010. — № 2. — С. 87 — 92.
7. Time series analysis, modeling and applications: A computational intelligence perspective / ed. by W. Pedrycz and S.-M. Chen. — Heidelberg: Springer, 2013. — 404 p.

References

1. Lukina, A.G. (2007). Requirements to systems for searching information in the internet with the use of a mobile phone as a final device. *Nauchno-Technicheskaya Informatsiya: Seriya 1*, 8, 23-26.
2. Lyudkevich, S. and Esipov, E. (2003, November). The main factors that determine relevance. *PromoTechart*. Retrieved from http://www.promo-techart.ru/analysis/relevants.htm
3. Mashechkin, I.V., Petrovskiy, M.I. and Tsarev, D.V. (2013). Methods of text fragment relevance estimation based on the topic model analysis in the text summarization problem. *Numerical Methods and Programming*, 14(1), 91-102.
4. Karpenko A.P. and Trusonoshin V.A. (2013). Multi-criteria estimation of the relevancy of documents in the enterprise ontological knowledge base using thematic clusterization. *Science and Education*, 11. DOI: 10.7463/1113.0637857
5. Badorina, L.N. (2007). Method of the relevance degree estimation of the text answer in computer training systems. *Proceedings of the National Aviation University*, 31(1), 70-72.
6. Konovalov, D.P. (2010). On the question of fuzzy queries to relational databases. *Perspektivy Razvitija Informacionnyh Tehnologij*, 2, 87-92.
7. Pedrycz, W. and Chen, S.-M. (Eds.). (2013). *Time Series Analysis, Modeling and Applications: A Computational Intelligence Perspective*. Heidelberg: Springer.

АНОТАЦІЯ / АННОТАЦИЯ / ABSTRACT

В.А. Крісілов, К.О. Городнича. **Методика оцінки релевантності результатів запитів до баз даних**. Збільшення обсягів збереженої і оброблюваної інформації висуває особливі вимоги до методів і засобів пошуку та обробки інформації. Одним з показників, що характеризують якість пошуку інформації, є релевантність результатів запиту. Метою є розробка методики для кількісної оцінки релевантності результатів запитів. Оцінка релевантності результатів запитів важлива для коректного пошуку інформації. Представлена методика виділяє і оцінює результат запиту для трьох видів відповідності знайденого об'єкта запиту. Для об'єктів, які повністю не відповідають висунутим вимогам, пропонується обчислювати ступінь віддаленості знайденого об'єкта і запиту. Представлена методика використовує апарат нечітких множин для описання об'єктів і запитів до баз даних з метою полегшення пошуку і групування об'єктів за часовими характеристиками, а також дозволяє кількісно оцінювати релевантність результатів запитів для коректного пошуку інформації.

Ключові слова: нечіткі множини, релевантність, нечіткий запит.

В.А. Крисилов, Е.А. Городничая. **Методика оценки релевантности результатов запросов к базам данных.** Увеличение объемов хранимой и обрабатываемой информации выдвигает особые требования к методам и средствам поиска и обработки информации. Одним из показателей, характеризующих качество поиска информации, является релевантность результатов запроса. Целью является разработка методики для количественной оценки релевантности результатов запросов. Оценка релевантности результатов запросов важна для корректного поиска информации. Представленная методика выделяет и оценивает результат запроса для трех видов соответствия найденного объекта запросу. Для объектов, полностью не соответствующим предъявленным требованиям, предлагается вычислять степень удаленности найденного объекта и запроса. Представленная методика использует аппарат нечетких множеств для описания объектов и запросов к базам данных с целью облегчения поиска и группировки объектов по временным характеристикам, а также позволяет количественно оценивает релевантность результатов запросов для корректного поиска информации.

Ключевые слова: нечеткие множества, релевантность, нечеткий запрос.

V.A. Krisilov, E.A. Gorodnichaya. **Methods of assessing the database queries' results relevance.** Increase in the volume of stored and processed information imposes special requirements to methods and tools for information search and processing. One of the indicators characterizing the quality of information retrieval is the query results relevance. This article

purpose is to develop a methodology to quantify the relevance of query results. The query results' relevance evaluation is important for the correct information retrieval. The presented method identifies and evaluates the query results for the three types of found object compliance to the request. For objects that do not completely correspond to the required specification, it is proposed to calculate the diversity factor between the found object and the query, as in some cases it is impossible to find an object that would at least partially satisfy the requirements. The presented method uses fuzzy sets to describe objects and queries to databases in order to facilitate objects' searching and grouping by temporal characteristics, as well as allows to evaluate quantitatively the query results' relevance for the correct information retrieval assessing.

Keywords: fuzzy sets, relevance, fuzzy query.

Received October 15, 2014