Qualitative and Quantitative Methods in Libraries (QQML) 2:477 –484, 2014

Technical relevance of keyword searches in full text databases

Erzsébet Tóth[1], Béla Lóránt Kovács[2]

[1]Affilation: University of Debrecen, Faculty of Informatics
[2]Affilation: University of Debrecen, Faculty of Informatics

Abstract: In our presentation we show a method, which makes possible to measure precision of keyword searches executed in full text databases. This method analyses how much information on average is expressed by the context of keywords in connection with a specific keyword in the database. Since the average information content depends on other elements of the database, thus we can consider this method objective. Using this method we can create user types, which categorize people who carry out various searches. We place those individuals in the first category who search for novelties, so they want to find texts with high average information-content. Those persons belong to the second category who search widespread relationships of meaning, as they wish to obtain texts with a low average information-content. In our test we determine the technical relevance of search results, but we take into consideration the user needs through the created user types. Previously we suppose that the average information-content of a textual document reflects its precision. In our analysis we examine this hypothesis in more details. Another interesting question, which emerges during our analysis is how we can use notions of technical relevance and technical precision at keyword searches in a full text database. Finding appropriate answer to this question makes possible that objective and really mathematical methods would appear in the relevance measurement of keyword searches in order to check rather subjective methods.

Keywords: technical relevance, average information content**,** full text databases, measurement, keyword searches

1. Definitions

Precision and recall are two fundamental measures of the effectiveness of information retrieval systems. Precision is the percentage of relevant retrieved documents out of the total number of documents retrieved by the system on a query. The relevance of a document is judged by the user formulating the query,

Received: 24.4.2014 / Accepted: 12.5.2014 ISSN 2241-1925

and is a subjective measure (Salton, 1989). In other words, precision is the measure of the system's capability to retrieve relevant documents and at the same time to withhold irrelevant ones. Theoretically recall is the percentage of relevant retrieved documents out of all relevant documents (including relevant retrieved and relevant not retrieved documents), but in practice it means the total number of the actually accessible relevant documents (Ungváry, 2001 pp. 196-197). Here we note that precision and recall are in an inverse relationship with each other, therefore the ideal state can not be reached at all. Thus we can say that more complete a search is, more imprecise it is. If we increase recall, the precision will reduce and vice versa: the increase of precision will result in the decrease of the recall (Horváth and Sütheő, 2003 p. 180). We determined to use the terminology of technical relevance to the more widely used term relevance in order to avoid the complicated issues of defining relevance (see for example Saracevic, 1998 and Mizzaro, 1998).

Various views of relevance have been developed in the field of information science. Among these views we call your attention to the system's view of relevance. According to Saracevic (1975, p. 327) it is 'a result of the thinking that relevance is mostly affected by the internal aspects and manipulations of the system'. Corresponding to this we need to make a distinction between system-based relevance and user-based relevance. The former means that the system determines whether a search result is relevant or not (e.g. when we search by Boolean operators the technical relevance of the retrieved items is 100% in each case, because of Boolean operator matching used by the system). The latter refers to that a user makes a judgment about the relevance of the same item (Horváth and Sütheő, 2003 pp. 156-157).

In our paper we use a system's point of view, and introduce an objective measure, which searches for the presence of the query terms in the document. First we present a technical definition of relevance: 'a document is defined as technically relevant if it fulfills all the conditions posed by the query' (Bar-Ilan, 2000 p. 441). So it means that 'all search terms and phrases that are supposed to appear in the document do appear, and all terms and phrases that are supposed to be missing from the document – terms preceded by a minus sign or a NOT operator – do not appear in the document' (Bar-Ilan, 2002 p. 310). Technical relevance is considered to be an objective measure. We can easily and quickly calculate technical relevance by a computer program instead of using human relevance judgment. This relevance judgment is rather subjective, because it depends on the subject's expertise who checks the content of the document. 'Technical precision is defined as the percentage of technically relevant retrieved documents out of the total number of retrieved documents' (Bar-Ilan, 2000 p. 441). The notion of technical relevance has been first discussed in (Bar-Ilan, 1999).

The great advantage of the use of technical relevance is that it can be evaluated for very large sets of full text documents, and can be easily checked by applying a simple pattern-matching algorithm. Its drawback is that it does not estimate the importance or the authority of the document. However, human judgment using a no binary relevance scale can assess the importance of a document (Bar-Ilan, 2002 p. 310). Another shortcoming of technical relevance is that it cannot make a difference between documents providing extensive and useful information on the search topic, and between documents in which the search topic mentioned only superficially, e.g. in a footnote of the full text article. Computing technical relevance provides a quick and easy method to differentiate between the documents ’about the search topic’ and textual sources that clearly do not correspond to the query, but we must also be aware of its limitations (Bar-Ilan, 2004, p. 208). In the following section we will formally define new measures based on the notion of technical relevance.

In our paper we analyze full text databases and we also consider Benczúr's previous results to be significant in this field (1988). As preliminaries we mention that he applied Kolmogorov's algorithmic approach to define the measure of information stored in database systems. Kolmogorov calls this measure information quantity. Benczúr defines the information quantity of relation *r* to be the algorithmic complexity of *x(r)*. Here *x(r)* indicates the canonical form of relations (Benczúr, 1988 p. 6). He refers to the basic definitions and theorem of the algorithmic complexity measure introduced by Kolmogorov (Kolmogorov, 1965). According to Benczúr the simplicity of a relation *r* can be measured by a ratio, which must reflect some constraints. A relation *r* satisfies some constraints that we regard its functional dependency. He used this observation to define the simplicity of a constraint or dependency (Benczúr, 1988 pp. 7-8).

2. Information needs

In the literature there is a cognitive approach to the interpretation of the information need. Mackay (1960) and Taylor (1968) say that the reason for understanding an information need comes into being when one becomes aware of a mental state of current incompleteness. Wersig (1971) describes this state of incompleteness as a problematic situation. In this sense the user finds himself in a real life situation in which he has recognized his own inadequacy. When the user understands his information need, soon he will be motivated to obtain the information necessary to solve the problems or uncertainty driven by the situation. Thus the formation of our information need depends on the real life situation we get into. In addition to this, we mention that Belkin, Oddy and Brooks named information need differently anomaly. They applied the concept of the ‘anomalous state of knowledge’, which is known as the ASK hypothesis: ‘…an information need arises from a recognised anomaly in the user's state of knowledge concerning some topic or situation…’ (1982 p. 62).

Consequently we can create two user types for individuals with various information needs in our analysis. First category consists of users who search for novelties, so they want to find textual documents with high average information content on a certain topic. For example these persons can be research physicians who search for articles by keywords. Those individuals belong to the second category that search widespread relationships of meaning, as they wish to receive texts with low average information content. For example we can think of physicians who execute searches by using simple keywords. These created user types reflect for us the information needs that users have during online search.

3. Formulas

After that, the question is how to measure the information-content of the document in a keyword search. If we know the information-content a document carries in connection with a keyword, we can compare that with the requirements put up for it. During the search, we will arrange the documents according to their information value. This value is called informativity (Kovács and Takács, 2013) in the literature. It is used to show how many bits of information a word in the document carries in connection with the keyword in the given field of science. It tells us how usual or unusual environment a certain document provides in connection with the document. Therefore, informativity is much more telling than the complete information-content of the document, because that depends on the number of the documents' words, as well. In order to obtain informativity, first we have to clarify how to calculate the information value of the documents. If we search documents based on keywords, then a document's information value is summed by the information value of all the other words contained by the document. The information value (I) of a word (x) can be calculated with the help of the following formula (Wiener, 1948):

$$I(x) = -\log_2 p(x). \tag*{1.}$$

The p-value is the probability of a word's occurrence in a document where the searched keyword (y) appears (Rényi, 1989). We can calculate this according to the following formula: the number of the occurrences of word x in a document where word y occurs is divided by the number of all words in a document where word y occurs.

With the help of the value above mentioned we can calculate a document's informativity. The value of informativity (Y) can be calculated in the following way by modifying Kovács and Takács's formula (2013):

$$Y(x_1) = -\sum_{i=1}^{n} \frac{1}{n-1} \log_2 p(x_i) \tag*{2.}$$

In the above formula n indicates the number of words in a document, therefore this formula helps to tell how many bits of information a word in the document carries on average in connection with the keyword. Naturally the keyword can occur many times in the document, thus its further occurrences also have information value. Nevertheless, the first occurrence of the keyword does not carry information as it is seen in the formula. After that, our task is to arrange the documents according to what informativity they have.

The calculation of the document's informativity (Y) keyword by keyword (y) can be done with the following method:

$$\begin{array}{lll} y_1 & \rightarrow & Y_1^1 \ldots Y_m^1 \\ \ldots & & \\ y_k & \rightarrow & Y_1^k \ldots Y_m^k \end{array}$$

Figure 1.

On the graph above k is the number of keywords, while m is the number of documents. According to this graph, we assign all the relevant documents from 1 to m (according to the Y value attached to them) to every keyword. Following this, we will be able to arrange the documents according to their informativity.

Keyword by keyword, we arrange the informativity (Y) of certain documents in the following way:

$$\begin{array}{lll} y_1 & \rightarrow & Y_{i1}^1 \ldots Y_{im}^1 \\ \ldots & & \\ y_k & \rightarrow & Y_{i1}^k \ldots Y_{im}^k \end{array}$$

Figure 2.

On the graph above, we can see that at each keyword, we arranged the documents into an increasing order according to how much informativity they have in connection with the keyword. From now on, with this knowledge we will be able to search only those documents whose informativity approaches the lower or the upper value. In order to do this we can introduce a limit below or above which we can get in the case of the searched documents.

FIGURE We choose a limit for each keyword, on the graph below the sign for value Y will be index s:

$$\begin{array}{lll} y_1 & \rightarrow & Y_{i1}^1 \ldots Y_{is}^1 \ldots Y_{im}^1 \\ \ldots & & \\ y_k & \rightarrow & Y_{i1}^k \ldots Y_{is}^k \ldots Y_{im}^k \end{array}$$

Figure 3.

On the graph we arranged the informativity of each document for each keyword ($y_1 ... y_k$). Within these values we can set a limit which can narrow the matches.

For example, the Y_{is}-value at a keyword can be a lower limit which arranges the five documents carrying the biggest informativity into one set. If somebody searches among the documents of a certain field of science, for those five documents which provide the most unusual environment for the searched keyword, he will be able to find the searched document this way.

Therefore, value Y_{is} is a limit which sets a distinction among the technically relevant documents. This distinction shows whether a document's words carry more, or less average information than a given value in connection with the keyword. Thus, value Y_{is} makes the search more precise. Those documents which can be grouped by this value into the set of searched documents are considered technically precise. Consequently, we have set a distinction between technical relevance and technical precision. Technically precise documents are a subset of technically relevant documents. We consider all documents technically relevant, which contain the searched keyword. Within this set we consider all documents technically precise which have an informativity higher or lower than Y_{is} – naturally depending on whether we search for a document of higher or lower value than this.

If somebody searches for documents within a given field of science, which use the searched keyword in a generally accepted environment, he will search for documents of low informativity. Thus, the words of these documents will carry low average information-content in connection with the given keyword. This is why value Y_{is} is going to function here as an upper limit. If somebody searches for documents which use the searched keyword in an unusual - or new - environment within the given field of science, then he will search for documents of high informativity. Thus, the words of these documents will carry high average information-content in connection with the given keyword. This is why value Y_{is} is going to function here as a lower limit. The users with two different information requests, described at the beginning of this paper, are able to use the notion of technical precision based on the above mentioned. If he only searches for the five most unusual documents in connection with a keyword, it is enough for him to set a Y_{is} value which only provides access for the set of these.

4. Conclusions

In our paper, we aimed to answer the question of how the notions of technical relevance and technical precision can be used at keyword searches in a full-text database. To be able to do this, first, after Bar-Ilan, we defined these two notions which have been used as synonyms. Following this, we have created the types of users with two different sorts of information requests, who approached the texts in the databases with different needs. After this, we turned to the direction of measuring information in the case of information request. We investigated how to measure the document's information-content at a keyword search. We came to the conclusion that a document in connection with a keyword can be of interest when it comes to measuring its average information-content. This is the document's informativity, which shows how many bits of information one word in a document carries on average, in connection with the keyword. We can arrange the documents based on this, which enables us to give them to users with different information requests. By doing so, we have distinguished the notions of technical relevance and technical precision. The method described above, however, can be used not only for searches with keywords. With different formulas it can work for subject-heading searches as well. With minor modifications the method can be used at keyword searches, which do not look for full texts, but for smaller text-fragments in the database. This is why the notions of technical relevance and technical precision can serve as the theoretical basis for numerous further technological developments.

The publication was supported by the TÁMOP-4.2.2.C-11/1/KONV-2012-0001 project. The project has been supported by the European Union, co-financed by the European Social Fund.

References

Bar-Ilan, J. (2004). Search engine ability to cope with the changing web. In: *Web dynamics – adopting to change in content, size* / Mark Levene, Alexandra Poulovassilis. Berlin; Heidelberg : Springer 195-215.

Bar-Ilan, J. (2002). Methods for measuring search engine performance over time, *Journal of the American Society for Information Science and Technology*, Vol. 53. No. 4. 308-319.

Bar-Ilan, J. (2000). Evaluating the stability of the search tools Hotbot and Snap : a case study, *Online Information Review* Vol. 24. No. 6. 439-449.

Bar-Ilan, J. (1999). Search engine results over time - a case study on search engine stability, *Cybermetrics*, Vol. 2/3. Retrieved March 30, 2013 from:

http://www.cybermetrics.info/articles/v2i1p1.pdf

Belkin, N. J. – Oddy, R. – Brooks, H. (1982). ASK for information retrieval: part I., *Journal of Documentation*, Vol. 38. 61-71.

Benczúr, A. (1988). Information measurement in relational databases, *Lecture notes in computer science*, Vol. 305, 1-9.

Horváth, T. – Sütheő, P. (2003). *A tartalmi feltárás. In.: Könyvtárosok kézikönyve. 2. köt. Feltárás és visszakeresés*. Szerk. Horváth T. – Papp I. Budapest: Osiris, 2003.

Rényi, A. (1989). Valószínűségszámítás. 5. kiad. Budapest: Tankvk.

Kolmogorov, A.N. (1965). Three approaches to the definition of the concept of ’’information quantity’’ (in Russian), *Prolemü peredacsi informacii*, Moszkva Tom I., Vüp. 1., 3-11.

Kovács, BL – Takács, M (2013). New search method in digital library image collections. *Journal of Librarianship and Information Science*. (accepted)

Mackay, D. M. (1960). What makes a question? The listener, Vol. 62 (May 5), 789-790.

Mizzaro, S. (1998). How many relevances in information retrieval? Interacting with Computers, 10, 1998. 305-322. Retrieved March 10, 2013 from:

http://www.dimi.uniud.it/mizzaro/research/papers/IwC.pdf

Salton, G. (1989). *Automatic Text Processing*, Addison-Wesley, Reading, MA.

Saracevic, T. (1998). Relevance reconsidered. Information science: Integration in perspectives. In: *Proceedings of the Second Conference on Conceptions of Library and Information Science* (CoLIS 2), Copenhagen, Denmark 201-218.

Saracevic, T. (1975). Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, November-December, 1975, 321-343.

Taylor, R.S. (1968). Question negotiation and information seeking in libraries, *College and research libraries*, Vol. 29. 178-194.

Ungváry, R. (2001). Az információkeresés értékelése. In.: *Osztályozás és információkeresés: kommentált szöveggyűjtemény. 2. köt. Az információkeresés és elmélete*. Szerk. Ungváry R., Orbán É. Budapest: OSZK, 2001. Retrieved March 12, 2013 from: http://mek.oszk.hu/01600/01683/pdf/01683-2.pdf

Wersig, G. (1971). *Information-Kommunikation-Dokumentation: ein Beitrag zur Orientierung der Informations-Dokumentationswissenschaften*. München-Pullach: Verlag Dokumentation Saur KG

Wiener, N. (1948). *Cybernetics: or Control and Communication in the Animal and the Machine*, The MIT Press, Cambridge