# A PROPOSAL FOR A MATCHING FUNCTION FOR AN INFORMATION RETRIEVAL SYSTEM

**V. Pačovski[1], M. Kon-Popovska[2],**

[1]Department of Mathematics, Faculty of Civil Engineering, Rade Koncar 12

[2]Institute of Informatics, Faculty of Natural Sciences & Mathematics,

Sts. Cyril and Methodius University, Skopje, Macedonia

matematika@gf.ukim.edu.mk, margita@ii.edu.mk

**Abstract.** The paper considers the idea of proposing a matching function between document and the query, based, on a objective (statistical) parameters. The function also allows additional qualitative ranking of a subset of retrieved documents, which contain all the words of a query. The function is tested on a test-collection containing over 1000 rather short documents (common news) which is continually growing. Paper represents a part of own continuous research in Information retrieval and as a starting point, we presume previous results: the database of documents is a welldefined structure and computed quotients of relevance, as a kind of measurement for the strength of a document-word relation as well as word-subject relation are available. The further research, will discuss the possibilities of interaction with the user, from the point of implementing the weighted retrieval and enabling additional intervention upon the set of retrieved documents, using user's judgement (the relevance feedback) to improve the quality of retrieval.

**Key words:** relevancy, matching, retrieval, ranking, evaluation, matching function, information retrieval

## 1. Introduction

Paper represents a part of own continuous research in Information retrieval and as a starting point, we presume previous results: the database of documents is a well-defined structure (with removed stop-words) and computed quotients of relevance, as a kind of measurement for the strength of a document-word relation as well as word-subject relation, are available.

It has been tested with acceptable response time on a rather small collection of documents, and has yet to be tested on a large collection of documents. In this work, we presume previous result presented in [1]: the database of documents

is a SQL compatible, well-defined structure (with stop-words in Macedonian language removed and optimized) and computed quotients of relevance [2].

The system is continually being tested on a test-collection containing over 1700 rather short documents (common news), which is continually growing (up to 30 documents per day) and established (also continually growing) dictionary of over 18.000 words.

## 2. Relevancy

Relevancy is a concept most directly connected with the user's information need. It is the only factor that can represent the strength of the relationship between documents from a database and a user.

The most important factors that affect relevancy are the following. First of all, relevancy is a personal concept, which by definition means that it is **subjective**. Two users may, with the same request, have different opinions about the relevancy of the same retrieved document. Next, it is **dynamic**. In one retrieval session, a user can asses a document as a relevant one, and later, as an irrelevant, and vice versa. Besides, the documents a user has retrieved and read, can affect his/hers judgment, and in fact, affects his/hers way of thinking about the relevancy of the documents that are not yet retrieved or viewed. There is also a question of **complexity**. Namely, relevancy is determined not only from the subject, but also from the credibility (of the source or the author), specificity, accuracy, the date of the document, clarity, etc.

Finally, relevancy depends on **the document representation** (the document surrogate, or the internal document representation - IDR). In other words, the user judges the relevancy of a document by looking at a portion (or a fragment) of a document that can be accessed. Even the way in which that fragment is presented can affect the user's judgment about that document.

## 3. General considerations

When we search a collection of documents we actually try to retrieve as much as possible relevant, and as little as possible (or better yet, none) non-relevant documents. And, since there is no outside factor to do the job for us, we are forced to use our imperfect knowledge to "guess" if the document is relevant or not.

Without entering philosophical paradoxes on the topic of relevancy, we will simply presume that relevancy can be approximated through the summary data of the document and its relations with other documents. This is quite a reasonable assumption because its alternative is that the user reads the whole text and judges for him/her.

We have to consider three types of connections: word-document, word-subject and document-subject.

To explain the way of calculating the relevancy, i.e. the process of computing **the quotients of relevancy**, we will limit ourselves to one subject and a set of documents, which have some kind of connection with that subject. The last assumption does not mean that some of them are not relevant to other subjects.

For that purpose, let's see what is known. We know the number of documents that have some connection with the subject, then (within documents), the frequencies of words, which enables a calculation of global relevancies of words within the subject.

## 4. The information need of a user or matching

The user expresses his information need by entering a request or a questionnaire in a form of a sentence or a sequence of key words. However, after an IDR (internal document representation) of a request (a query or a questionnaire) has been created, there is still a problem of matching that IDR with the IDR of documents and a degree of similarity.

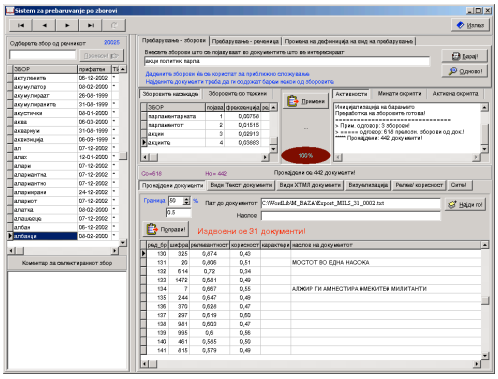

Figure 1: A list of retrieved documents

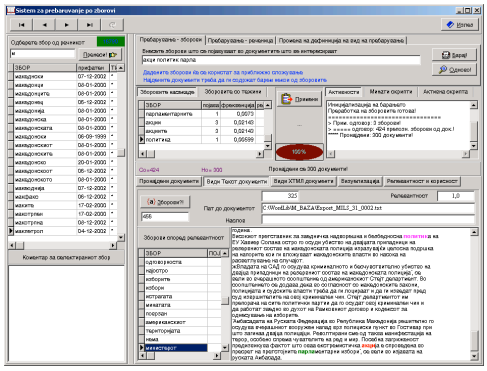

Figure 2: A view on a retrieved document

Standard approach in the majority of existing systems is a direct comparison with the database of IDR of documents. One of the problems with it is the enormous number of comparisons. What makes it even more difficult is the constant increase of the number of documents, which keep coming into the database. As a result, it can be expected that, for a typical request entered in different time intervals, the system will take more and more time for reply.

What seems to be a logical solution is to find some kind of an indirect (hidden) connection, which, at least at the beginning, will enable reduction of the search space. Such a connection is proposed here.

Namely, instead of direct comparison of IDR of the request and documents, while calculating the quotient of relevancy, the following is proposed. Considering that we have the words of a request and we have previously obtained the relevancies of words to the documents and to the topics, it sounds reasonable to use the connection word-topic for retrieving appropriate, potentially relevant documents.

The reasoning is as follows: The user, at least at the beginning, has very vague information need, so the shortest way to give him/her good information is to discover which the most relevant topics for his request are.

Based on his request, the system calculates the quotients of relevancy of those words, taken as a group, related to the topics, and the topics with highest relevancy are suggested to the user as a list. The user, as a feedback, will then give the system a chance to reduce the search space of IDR of documents. In that sub-space, if needed, this procedure can be repeated to further narrow the search space to subtopics.

### 4.1 Absolute quotient of relevancy, or a word-subject relation

One of the assumptions we base our considerations on is that one word has an exact quantity of information for every topic. For some, that amount of information is minimal and can be neglected, but for others it can be very big, meaning that the word contains maximum information. The ideal case is when a word contains all the information about the topic, but as far as we know, within the limits of existing natural languages, that is not possible.

The amount of information for a subject (topic) contained in one word can be expressed in various metrics. Further in the text we will use probability metrics, i.e. the amount of information will be represented through values from the interval [0,1]. Those values will describe the word-subject relation and we will name them **absolute quotients of relevancy**.

With their help, we are calculating the relevancy of a document to a certain subject.

### 4.2 Theoretical considerations

Let **K** be a topic (subject). The collection of documents connected with that topic is $\mathbf{K_d}$, and $\mathbf{D_i}$ is the **i**-th document, and $\mathbf{A_j}$ is **j**-th word in that topic.

We see the subject (topic) as a disjunctive union of documents (which are viewed as sets of words, where each word can belong to more than one document).

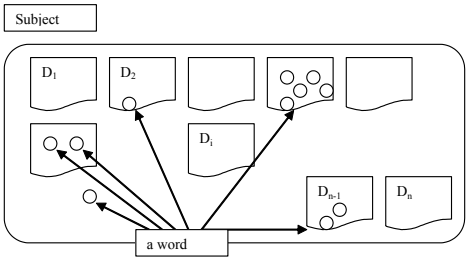


Figure 3: Visual representation of the idea for calculation of the relevancy

The probability of a document (as a probability of an independent event) within that topic can be calculated by the following formula:

$$\mathbf{p(D_i)} = \frac{|\mathbf{D_i}|}{\sum_{k=1}^{n} |\mathbf{D_k}|} \qquad (1)$$

where $|...|$ is a cardinality of a document, the number of occurrences of all words in the document. The sum of all these probabilities is 1.

We define the event $\mathbf{A_j}$ : "**j**-th word has occurred". That occurrence depends on partial occurrences of that word within a document $D_j$

$$\mathbf{p(A_j / D_i)} = \frac{|\mathbf{A_{ji}}|}{|\mathbf{D_i}|} \qquad (2)$$

where $|A_{ji}|$ is a number of occurrences of the word $A_j$ in i-th document. Now, the total probability of the event $A_j$ (according to the formula for total probability) is:

$$\mathbf{p(A_j)} = \sum_{i=1}^{n} \mathbf{p(D_i)} \cdot \mathbf{p(A_j / D_i)} \qquad (3)$$

Thus obtained values are taken as a starting point for calculating the relevancy. Namely, between these probabilities we seek the highest:

$$\mathbf{p_0} = \max_{j=1,m} \mathbf{p(A_j)} \qquad (4)$$

Quotients of relevancy for all words are obtained as a division between their probabilities and that highest probability.

$$k_j = \frac{p(A_j)}{p_0} \qquad (5)$$

Here, we can see that at least one word will have a quotient of relevancy equal to 1. Namely, that was the word with the highest probability. Thus obtained quotients will be named **quotients of absolute relevancy**.

In this context it is interesting to see what the meaning is of the probability $p(D_i / A_j)$ that can be calculated by a Beyes formula. This probability signifies the probability that the word, which has appeared came from i-th document. We consider that probability as a **reverse probability**, meaning how strong the connection of that word with certain documents is. It can also be used as a kind of weight (for weighted retrieval see it discussed forward) and also enables ranking of documents.

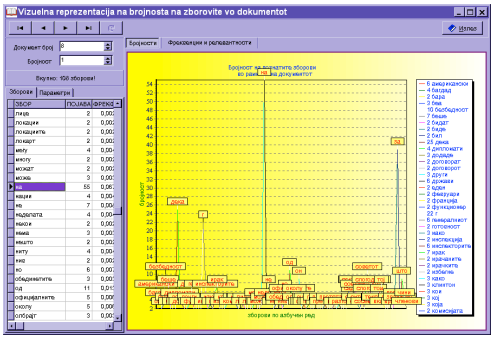


Figure 4: Visual representation of the occurrence of words within a document

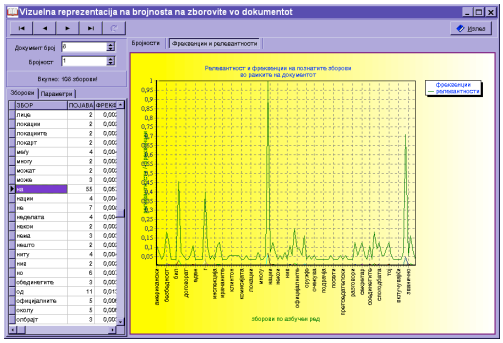


Figure 5: Graphic presentation of relevancy of words within a document

### 4.3 Tables of absolute and reverse relevancy

The reason of existence of these tables of relevancy is based on a consideration that the relevancy of a word for a document from which it has been extracted and for a given subject, has components which are absolute, meaning that they are independent from the user's judgment, and depend only upon the subject, i.e. the treatment of the system.

So, in the table of absolute relevancy we have a triplet:

word / subject / quotient of relevancy

where the quotient of relevancy is the real number calculated above.

In this way we obtain a rather efficient presentation of strength of the relation word-subject**.**

Likewise, in the table of reverse relevancy, we have a quadruplet:

word / document / quotient of relevancy / subject

that enables a kind of ranking of documents according to their importance for certain word, after the subject has been defined.

### 4.4 Matching function

As a starting point, we presume previous results: the database of documents is a well-defined structure (with removed stop-words) and computed quotients of relevance, as a kind of measurement for the strength of a document-word relation are available.

Let **D** be a collection of documents, $\mathbf{D_i}$ – one of them, and **r** a request i.e. a set of **n** words. The document $\mathbf{D_i}$ contains some of them, we presume **m,** and the **k**-th has $\mathbf{n_k}$ number of occurrences within the document, and $\mathbf{w_k}$ – weight within the document in a sense mentioned previously (see formula 5). Then, the relevancy $\mathbf{r_i}$ of the $\mathbf{D_i}$, to the request **r**, as a probability measure of similarity (meaning, a number between 0 and 1), will be computed according to the following assumptions:

- all words are equally important, (meaning, that they participate at most with a quotient **1/n** in matching),
- their occurrences can and should be a factor which will decide, but will not prevail;

So, we have to include the number of occurrences, but at most within limits **1/n**. One of the methods, suggested here, is a kind of normalization, by including additional quotient which is computed by division of the number of occurrences of a certain word with the maximal number of occurrences of all the words contained in that document i.e. $\frac{\mathbf{n_k}}{\max\limits_{1\le k\le m}(\mathbf{n_k})}$. So, the **1/n** part will be given to only one of the words (the one which has the greatest number of occurrences).

Summing, we get the formula:

$$\mathbf{r_i} = \sum_{k=1}^{m} \frac{\mathbf{n_k}}{\max\limits_{1\le k\le m}(\mathbf{n_k})} \cdot \frac{\mathbf{1}}{\mathbf{n}} \qquad (6)$$

It should be noted that the relevance of 1, in this case, will be given to the document that contains all of the words, and all those words have equal number of occurrences. Having in mind that it is ideal, which means too strict, we give the relevance of 1 by definition, to those documents that contain all the

words from a request, without considering their number of occurrences, and use the formula for the rest.

If we decide to use weighted retrieval, by assigning weights to the terms of the request (for example, the computed coefficients of relevancy $\mathbf{w_k}$, mentioned previously), then, they will be included instead of the quotient **1/n.** So, having in mind the formula (6), if we replace **1/n** with $\frac{\mathbf{w_k}}{\sum_{j=1}^{n} \mathbf{w_j}}$, we get the formula

$$\mathbf{r_i} = \frac{\mathbf{1}}{\sum_{j=1}^{n} \mathbf{w_j}} \cdot \sum_{k=1}^{m} \frac{\mathbf{n_k}}{\mathbf{\max_{1 \le k \le m}(n_k)}} \cdot \mathbf{w_k} \qquad (7)$$

Quotients $\frac{\mathbf{n_k}}{\mathbf{\max_{1 \le k \le m}(n_k)}}$ i.e. $\frac{\mathbf{w_k}}{\sum_{j=1}^{n} \mathbf{w_j}}$ can be used for secondary ranking of retrieved documents by certain words, according to their number of occurrences.

The further on-going research will attempt to improve this matching function by using in some way the previously computed quotients of relevance. Also, we will discuss the interaction with the user, from the point of implementing the weighted retrieval and enabling additional intervention upon the set of retrieved documents, using user's judgment (the relevance feedback) to improve the quality of retrieval.

## 5. Reevaluation or re-establishing the balance

Having in mind that an IR system has an uninterrupted input of information (new documents are entered, the dictionary grows, the stop-word list may change etc), our approach requires **periodical** reevaluation of relevancies.

But, here the word **periodical** does not necessary mean time period, because some subjects might not have a new document for some time. That is why it is safer (more recommendable) to accept that the reevaluation will be performed after N new documents have been entered in the system.

The number N can be a constant different for every subject. So, subjects with bigger input of documents will have higher constants, and subjects with smaller input will have smaller ones.

Reevaluation, i.e. re-computation of relevance quotients is performed in time intervals when the number of users in the system is expected to be minimal.

## 6. Conclusion

Having in mind that an IR system, by definition, is a unity of three elements: data structures, relevancy and a matching function, this kind of approach gives an adequate presentation of connections word-document-subject which enables a follow-up and comparison with user's judgment. How much the judgment of the users will correspond to supposed theoretically calculated quotients of relevancy remains to be seen experimentally. However, such experiments are in progress.

## 7. References

1. Pachovski, V. (2000), *Organizacija na podatoci i relevantnost vo sistemi za pribiranje informacii*, [Data Organization and Relevancy in Information Retrieval Systems], masters thesis, Institute for Informatics, Faculty of Natural Sciences, "St. Cyril and Methodius" University, Skopje
2. Pachovski, V. (2002) A proposition for a method of calculating relevancy in an Information Retrieval System, *Proceedings of VI Balkan Conference on Operational Research, A Challenge for Scientific and Business Collaboration,* Thessaloniki, 22-25 May (to appear)
3. Baeza-Yates, R, Ribeiro-Neto, B, (1999), *Modern information retrieval,* ACM press, New York, Addison-Wesley
4. Korfage, R.R. (1997), *Information Storage and Retrieval*, Wiley Computer Publishing, NY, USA
5. Van Rijsbergen, C.S. (1979), *Information Retrieval*, Butterworths, London
6. Carpineto, C, (2000) Order-theoretical ranking, JASIS, Vol. 51(7),587-602
7. Mizzaro, S. (1997) Relevance: The Whole History, *JASIS, Vol. 48(9),810–832*