Yugoslav Journal of Operations Research
2 (1992), Number 1, 107-112

AN EVALUATION OF TEXTUAL DOCUMENTS INDEXING METHODS

Dragan MIHAJLOVIĆ, Danilo OBRADOVIĆ

Computer Control and Measurements Institute
Faculty of Technical Sciences, University of Novi Sad
Trg Dositeja Obradovića 6, 21000 Novi Sad, Yugoslavia

Abstract. The paper presents results of indexing methods evaluation of titles and key words of textual documents in Serbian language. The following indexing methods are evaluated: — automatic indexing with single words, — automatic indexing with compressed single words and — manual indexing with key words. Evaluation is performed on the basis of the computation of a linear correlation coefficient between the actual relevance and the formal computer relevance. The actual relevance of two documents is computed by finding the set of common words in both of them. The formal relevance is obtained by means of finding of intersection of the document searching characteristics.

Key words and phrases: indexing, manual indexing, automatic indexing, indexing evaluation

1. INTRODUCTION

The appearance of the computer supported information and documentation systems (INDOC-systems) lead to the textual documents automatic indexing algorithms. One group of algorithms is based on dictionaries or thesaurus, the other group uses statistical properties of the text and the third group follows morphological analysis of the text. In this research special attention is paid to the comparison of different methods of automatic document indexing as well as to the comparison of the automatic and manual document indexing.

The indexing evaluation methods can be divided into two basic groups [1].

a) The expert indexing evaluation — based on the comparison of the basic subject of indexing text of the document and its document searching characteristics (DSC).

b) The indexing evaluation — based on the searching results expressed through the criterion of completeness, exactness and the information searching particularity.

The first group is obvious and simple, but its application do not guarantee satisfying results.

The second group incorporates certain problems of separating the influence of indexing from other agents (e.g. exactness and completeness of the fund completion) influencing completeness, exactness and information searching particularity. All the difficulties associated with the INDOC-system evaluation are faced within this second group of methods.

The aim of the paper is to study the different indexing methods in order to find out their shortcomings and their advantages.

This research has been limited to comparison of the following indexing methods:

— Method A — automatic indexing with single words. Each word of the document title enters DSC except uninformative words (conjunctions, exclamations, pronouns, particles etc.);

— Method B — automatic indexing with compressed single words. The compressed forms of the document title words enter the DSC. Uninformative words do not enter the DSC. The compressed form of the word is obtained using the particular compression algorithm [5].

— Method C — man's manual indexing. DSC is a list of key words.

The indexing evaluation is done under the following circumstances:

1) Evaluation is to be performed only on the basis of the full documents texts comparison and their DSC. A document text and its DSC is in Serbian.

2) Human participation in the research procedure should be minimized.

3) Significant statistical results should be reached by analysing a sufficiently small set of documents.

4) The efficiency of different indexing methods is to be valued on the same set of documents.

5) As a basis of the comparative analysis, the $R_{\alpha\beta}$-criterion (coefficient of the linear correlation between real relevance α^* and formal-computer relevance β^*) is to be used.

2. THE EXPERIMENT DESCRIPTION

A sufficiently satisfactory description of the INDOC-system regarding the user is bidimensional function $\varphi(\alpha^*, \beta^*)$, known as the basic statistical INDOC-system model [2], [3], [4]. The bidimensional function $\varphi(\alpha^*, \beta^*)$ of the random variables α^* and β^* (α^* — real relevance, β^* — formal relevance) represents probability of the (α^*, β^*) pair appearance through random choice of documents.

As the basis of the indexing methods comparative analysis, the linear correlation coefficient, $R_{\alpha\beta}$, of the random variables α and β is used. In the study [2] it is demonstrated that under the same conditions, increase of linear correlation coefficient, $R_{\alpha\beta}$, of the random variables α and β, leads to the increase of system quality indicators (completeness, exactness, etc.).

In the studies [2] and [4] it is shown that, if one uses the "zero indexing algorithm" (automatic indexing with zero effectivity) which forms DSC by a random choice of the terms from the full document text, the correlation coefficient can be expressed as:

$$R_0 = \sqrt{l'/l} \tag{1}$$

where:

l — is the number of the full text terms,

l' — is the number of the DSC terms.

To form $\varphi(\alpha^*, \beta^*)$ the following method is used: a text of a document, from the explored set of documents, forms the information request, while the other texts of documents are checked on the real and formal relevance. This is repeated with all documents from the explored set as an information request. In such a way $C = n(n-1)/2$ the number of pairs of variables (α^*, β^*), are obtained, where n represents the number of the analysed set of documents.

The difficulties in obtaining the expert evaluations of the real relevance (for $n = 100$, $C = 4950$), could be overcome by automatic determination of real relevance as the full documents intersection.

The above consideration will be explained on the following example.

Example: The Entering Documents Structure

Let T be defined as a title of document e.g.

Informacioni sistem gradova i opština
(Communities and Cities Information System),

let K be defined as key words of document e.g.

gradovi i opštine; informacioni sistem;
kancelarijsko poslovanje; simpozijum; 1984; Beograd.
(Cities and communities; information system;
office management; Symposium; 1984, Belgrade).

and let L be defined as

$$L = T \cup K.$$

For β^* (see [2] and [4]) determination of the suitable DSC intersections are used. For the document analyzed they are as follow:

— The title words set:

$$T = \{\text{informacioni, sistem, gradova, opština}\},$$

— The compressed title words set [5]:

$$S = \{\text{informkion, sist, grad, opštin}\},$$

— The key words set:

$$K = \{\text{gradovi i opštine, informacioni sistem, kancelarijsko poslovanje, simpozijum, Beograd}\}.$$

The numerical information as well as uninforming words (conjunctions, prepositions, particles etc.) are filtered out through the stop dictionary containing about 850 words, and they do not enter the document on the whole and the document's DSC.

In this experiment 73 documents were analysed.

During the given documents set handling the pair sets (α^*, β^*) are obtained for each pair L and T; L and S; L and K.

On the basis of [1] the following equations are obtained:

$$\varphi(\alpha^*) = \sum_{\beta^*=0}^{\beta_M} \varphi(\alpha^*, \beta^*), \qquad \varphi(\beta^*) = \sum_{\alpha^*=0}^{\alpha_M} \varphi(\alpha^*, \beta^*), \tag{2}$$

$$M[\alpha^*] = \sum_{\alpha^*=0}^{\alpha_M} \varphi(\alpha^*) \cdot \alpha^*, \qquad M[\beta^*] = \sum_{\beta^*=0}^{\beta_M} \varphi(\beta^*) \cdot \beta^*, \tag{3}$$

$$\sigma_\alpha^2 = \sum_{\alpha^*=0}^{\alpha_M} (\alpha^* - M[\alpha^*])^2 \varphi(\alpha^*), \qquad \sigma_\beta^2 = \sum_{\beta^*=0}^{\beta_M} (\beta^* - M[\beta^*])^2 \varphi(\beta^*), \tag{4}$$

$$M[\alpha^* \cdot \beta^*] = \sum_{\alpha^*=0}^{\alpha_M} \sum_{\beta^*=0}^{\beta_M} \varphi(\alpha^*, \beta^*) \alpha^* \cdot \beta^*, \tag{5}$$

the correlation coefficient is determined:

$$R_{\alpha\beta} = \frac{M[\alpha^* \cdot \beta^*] - M[\alpha^*] \cdot M[\beta^*]}{\sigma_\alpha \cdot \sigma_\beta}. \tag{6}$$

In the Eqs. (2)–(5) the maximal value of α^* and β^* are included (α_M and β_M).

The linear correlation coefficient value obtained from the Eq. (6) is compared with the least effective indexing coefficient obtained from Eq. (1), the average lengths for different DSC-s of the full text, that are a cardinality of the respective sets, are also determined:

$$|\overline{L}| = 11.98, \quad |\overline{T}| = 5.16, \quad |\overline{S}| = 4.73, \quad |\overline{K}| = 5.15. \tag{7}$$

3. ANALYSIS OF THE EXPERIMENTAL RESULTS

First of all, one has to justify whether the real relevance of the text as a document is obtained by using α^* the length of intersection $L_i \cap L_j$:

$$\alpha^* = |L_i \cap L_j|.$$

Form the set of 75 documents, the 200 pairs of documents is selected to be analyzed. The mutual relevance α' is expertly evaluated having values from 1 to 10. Such evaluations are also performed using Eqs. (2)–(6) with $\alpha^* = |L_i \cap L_j|$. The linear correlation coefficient between expertly determined real relevance and theoretically obtained real relevance is found to be:

$$R_{\alpha\alpha'} = 0.624. \tag{8}$$

Such a relatively large positive correlation of α and α' justifies the real relevance determination as the full text intersections.

For the analysed indexing methods the following values of $R_{\alpha\beta}$ are obtained:

$$R^A_{\alpha\beta} = 0.497966, \quad R^B_{\alpha\beta} = 0.689761, \quad R^C_{\alpha\beta} = 0.549889. \tag{9}$$

According to Eqs. (1) and (7) the least effective indexing coefficient is obtained as:

$$\begin{gathered} R^{A_0}_{\alpha\beta} = \sqrt{|\overline{T}|/|\overline{L}|} = 0.65628, \qquad R^{B_0}_{\alpha\beta} = \sqrt{|\overline{S}|/|\overline{L}|} = 0.62834 \\ R^{C_0}_{\alpha\beta} = \sqrt{|\overline{K}|/|\overline{L}|} = 0.65565. \end{gathered} \tag{10}$$

As a final DSC evaluation it is necessary to confirm that the value $R_{\alpha\beta}$ obtained from Eqs. (1) is within the confidence interval. It is known that the ρ_{xy}, the correlation coefficient of any x and y variables, with the 0.97725 probability is within the confidence interval:

$$r_{xy} - 2(1 - r^2_{xy})/\sqrt{N} < \rho_{xy} < r_{xy} + 2(1 - r^2_{xy})/\sqrt{N} \tag{11}$$

where r_{xy} is computed values of correlation coefficient, N is the number of x and y variable pairs.

In this case $N = C = 2628$ and the corresponding confidence intervals I are:

$$\begin{gathered} I(R^A_{\alpha\beta}) \in [0.468627. 0.522305] \qquad I(R^B_{\alpha\beta}) \in [0.669309, 0.710213] \\ I(R^C_{\alpha\beta}) \in [0.522673, 0.577105]. \end{gathered} \tag{12}$$

The following conclusion may be drawn:

1. If the Eq. (11) gives $\rho_{xy} \in r_{xy} \pm (0.02$ to $0.04)$ then it is sufficient to process 50 to 100 documents to obtain satisfactory precision.

2. If the confidence intervals of $R_{\alpha\beta}$, for all of methods analyzed, as defined in Eqs (12), do not coincide, there exists a statistically significant difference between the correlation coefficient obtained using three different indexing methods.

3. Although the indexing methods may be ordered as B, C, A according to descending values, of $R_{\alpha\beta}$, as it is shown in Eq. (9), this is not sufficient for the final DSC evaluation. due to the fact that those DSC-s are obtained from the different text lengths as shown in Eq. (7). In this case the correlation coefficients are to be considered for the worst corresponding cases [1], as it is given in Eq. (10). Namely,

it is proposed that the influence of different text lengths on $R_{\alpha\beta}$ values can be eliminated by forming the following W ratios:

$$W^A = R^A_{\alpha\beta}/R^{A_0}_{\alpha\beta} = 0.7588, \qquad W^A = R^A_{\alpha\beta}/R^{A_0}_{\alpha\beta} = 1.5915, \qquad W^A = R^A_{\alpha\beta}/R^{A_0}_{\alpha\beta} = 0.8387. \tag{13}$$

After reduction to the same conditions, from these results (see Eq. (13)) it may be concluded that the indexing by means of key words method (method B) is the best one, then the title compressed words method (method C) follows and at the end title single word indexing method (method A) is the least satisfactory one.

CONCLUSION

The experimental results obtained are used to establish a quality of different indexing methods in processing titles and key words of documents in Serbian language. The method of manual selection of key words is ranked the best followed by two automatic selection od DSC methods. In the case considered the greater precision and completeness in information retrieval is to be expected by using the manual indexing. That does not mean that automatic indexing methods with single and compressed single words are to be rejected. On the contrary, on account of number of advantages of the automatic indexing methods (e.g. reduced human participation in indexing, indexing methods (e.g. reduced human participation in indexing, indexing unification, faster indexing) it is justified to use automatic indexing methods also. The best approach would be to combine the manual and automatic indexing methods. Namely, the standard practice is that the authors are supplying key words besides the title of their text. These is then used as an input to the automatic indexing method based on the compressed single words. Thus, the efforts spent on the morphological analyses of information retrieval requests are significantly reduced.

REFERENCES

[1] И. И. Попов, А. Н. Павлов, *Экспериментальная оценка качества индексирования*. НТИ, сер. 2, 9 (1983), 13–19.

[2] И. И. Попов, *Некоторые модели оценки и оптимизации информационных систем: математический апарат моделирования*. НТИ, сер. 2, 3 (1981), 10–16.

[3] Р. А. Габриелян, А. Н. Павлов, И. И. Попов, Л. Е. Сарыханян, *Исследование алгоритмов индексирования*. НТИ, сер. 2, 9 (1982), 7–9.

[4] И. И. Попов, *Некоторые модели оценки и оптимизации информационных систем: оценка качества лингвистического обеспечения*. НТИ, сер. 2, 6 (1981), 7–14.

[5] D. Mihajlović, D. Obradović, *Jedan algoritam sažimanja srpskohrvatskih reči*. Informatica 3 (1982), 45–47.