1999 IEEE International Fuzzy Systems Conference Proceedings
August 22-25, 1999, Seoul, Korea

The construction of fuzzy relational maps in information retrieval [1]

László T. Kóczy [2], Tamás D. Gedeon [3], Judit A. Kóczy [4]

[2] Dept. of Telecommunication and Telematics
Technical University of Budapest
Budapest H-1521 Hungary
(koczy@boss.ttt.bme.hu)

[3] School of Information Technology
Murdoch University
Perth 6150 Australia
(t.gedeon@murdoch.edu.au)

[4] CONTROLLTraining Education Centre Ltd. Co.
23 Csalogány, Budapest H-1027 Hungary
(koczy@controll.hu)

Abstract

One of the major problems in automatic indexing and retrieval of documents is that usually user queries will not include all the relevant words that occur in the documents which should be retrieved. Also it often happens that the same query word with different meanings occur in different contexts from that expected by the querying person. In order to achieve better recall and higher precision, fuzzy tolerance and similarity relations were introduced based on the counted or estimated values of (hierarchical) co-occurrence frequencies. This study addresses the problem of how these relations can be generated from the values of occurrence frequencies, especially as these are based on possibilistic rather than probabilistic measures, and also how the relations can be implemented by fuzzy relevance matrices.

1. Introduction

An information retrieval system allows users to efficiently retrieve documents that are relevant to their current interests. The collection of documents might be extremely large and the use of terminology might be inconsistent, especially if the language of the documents is close to natural language (like in legal texts).

There are two partially contradicting measures of the effectiveness of a high quality information retrieval system. On one hand it is expected that the recall of the topic searched for should be high, that is, the set of relevant retrieved documents be as large as possible.

On the other hand, it is also required that the precision be as high as possible, that is, no documents be retrieved at all which are not relevant for the given query, being equivalent with the expectation of obtaining an as small as possible retrieved document set (cf. [1]).

Automated keyword search is the most widespread approach, however documents not containing the actual keyword(s) will not be retrieved. If the keyword in the query is *Soft Computing*, documents on *Fuzzy Systems*, *Neural Networks* and similar topics will be unambiguously relevant, even if they do not mention the broader term SC a single time. Moreover, other parts of the same scientific community prefer to use the name *Computational Intelligence* with a rather similar meaning, so all documents related to the latter should also be retrieved. Conversely, if the query specifies the two keywords *Fuzzy* and *Relation*, a story about two young people that contains the sentence "By that time the *relation* between John and Mary became rather *fuzzy*." will be among the retrieved documents – clearly having nothing to do either with fuzziness in the sense of fuzzy logic, or with mathematical relations.

In some previous work we suggested the use of hierarchical co-occurrence frequencies as indicators of the importance of individual words and groups of words in the contents of given documents [2,3]. It is obvious that these frequencies are not probabilistic measures, as it is not the relative frequency of a certain word among all words of the document that directly measures its relevance, however these frequencies determine the possibility degrees of the documents in a somewhat indirect, certainly not linear and essentially non-additive way.

In the next section a method for transforming the counted or estimated [4] frequencies into possibility measures (fuzzy membership degrees) will be presented.

[1] Supported by the Australian Research Council.
[2] Visiting Professor, Dept. of Information Engineering UNSW.
[3] Paper submitted while at University of New South Wales (UNSW), and paper completed at Murdoch University.
[4] Visiting Honorary Associate, Dept. of Info. Eng. UNSW.

0-7803-5406-0/99/$10.00 ©1999 IEEE

2. Keyword occur. frequencies and possibility degrees

Both occurrence and co-occurrence of keywords can be expressed with the help of word counts in documents. For a collection of documents, the words which occur frequently in all or most of them are of no significance with regards to the contents of any particular document. Those common in any natural language document are called stop words, while those which might be significant in some contexts but not in the document collection being used will be called *relative stop words*. For example, the word "law" which could occur frequently in legal documents would be not discriminative concerning the particular contents. By the omission of stop words and relative stop words we obtain the set of significant words which might be used for a query. Some of these words might be more important than the rest and might be chosen as the set of keywords. In a hierarchical co-occurrence approach the titles and sub-titles, etc. might be checked only for keyword occurrences, while the rest of the documents for any significant word. An example for classifying words into these four categories is in Fig. 1.

In the figure the four categories of words can be seen: absolute and relative stop words (like "the" and "law" in this particular context, and "carpet" as a general example for a significant word and "damag(es)" for a keyword (stem).

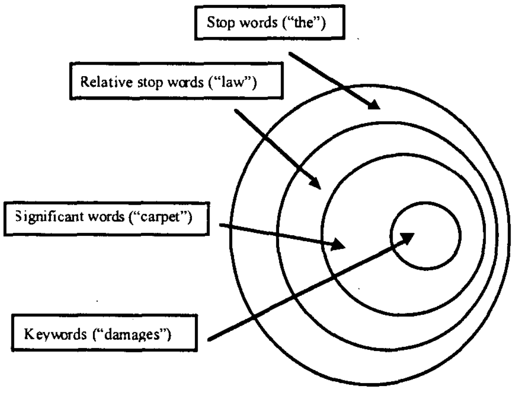

Figure 1. Categories of words in documents

Clearly, keywords in titles, abstracts, introductions, will have a lower occurrence count than significant words in general in the full text. It is a crucially important issue how occurrence frequencies can be transformed into fuzzy membership degrees, which are essentially (possibilistic) fuzzy importance or relevance measures.

Membership degrees or fuzzy measures range from 0 to 1, where 0 expresses the total lack of importance, and 1 stands for absolutely important. Words occurring very frequently are usually stop words, and so they should be left out of consideration. For the remaining significant words it is generally true that higher occurrence frequencies indicate higher importance degrees as well. Although the connection between occurrence frequency (word count) and importance degree is strictly monotonic, it is certainly not proportional. The critical domain is somewhere what can be defined as "a few occurrences", depending on the type and size of the document, somewhere between 2 and 20 word counts. It does not matter much whether a word occurs in a document 20 or 22 times, it is likely that this document will be rather important for the query in both cases. On the other hand, one or two occurrences of a word might be coincidental or might indicate that the subject is touched upon only very superficially, while repeated mention is an indicator that the word in question is an important word for the document. With short documents these numbers might vary. This is quite different with keyword occurrences in titles, ..., where even a single occurrence usually indicates high importance.

The mapping from occurrence frequencies or counts to possibilistic membership degrees is thus a sigmoid function, with its steep part around the "critical" area of occurrences – the concrete values depending on the expected lengths and types of documents, and the category of environment (title, text, etc.). These sigmoids $\sigma(F)$ have to fulfil the following conditions:

$$\sigma{:}\mathrm{R}^{+}\rightarrow[0,1], \quad \sigma\ F_1 > \sigma\ F_2 \Leftrightarrow F_1 > F_2$$

$$\frac{d^2\ \sigma}{dF^2} \geq 0 \Leftrightarrow F \leq T_F \quad \text{and} \quad \frac{d^2\ \sigma}{dF^2} \leq 0 \Leftrightarrow F \geq T_F \qquad (1)$$

In practice is not necessarily continuously differentiable, but its characteristics should be nevertheless "S-shaped". Although occurrence frequencies are integer numbers, it is reasonable to introduce the sigmoid mapping over the whole positive half of the real lines, as e.g. in [2] the importance degrees are introduced as convex combinations of occurrence counts (e.g. $F_{ij} = \lambda_1 T_{ij} + \lambda_2 C_{ij} + \lambda_3 L$, where $_i$ are real coefficients and *T*, *C*, *L* denote title-keyword, location-keyword and cue words related frequencies, resp.) and so these fictitious occurrence frequencies might assume any non-negative value. The typical characteristics of such a sigmoid function can be seen in Fig. 2.

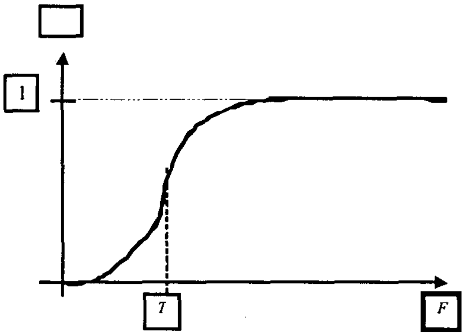

Figure 2. Sigmoid function transforming occurrence frequencies into membership degrees

More practical broken line functions with concrete values can be seen in Fig. 3. Here is a mapping for title (subtitle) occurrences and another one for text occurrences. The threshold values are obviously different.

Depending on the length of the document, the number of levels of subtitles, ..., the sigmoid curve can change.

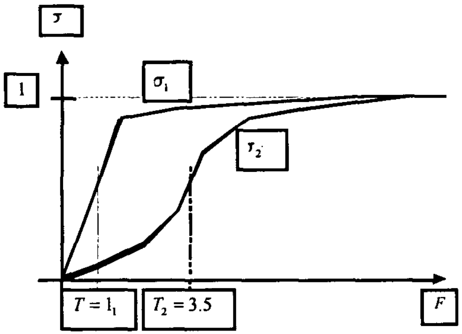

Figure 3. Sigmoid curves for title/subtitle and text occurrence

Membership degrees generated by the occurrence frequency transformation can be interpreted as possibility measures of a certain document. Although possibility has some similarities with probability, its axiomatic properties differ in an essential point: additivity does not hold. It is easy to realise this when considering the sigmoids.

Let us demonstrate this by the following table defining a certain sigma for integer values of F:

F	0	1	2	3	4	5	6	7	8	9
σ	0	0.1	0.2	0.4	0.7	0.9	0.95	0.98	0.99	1.00

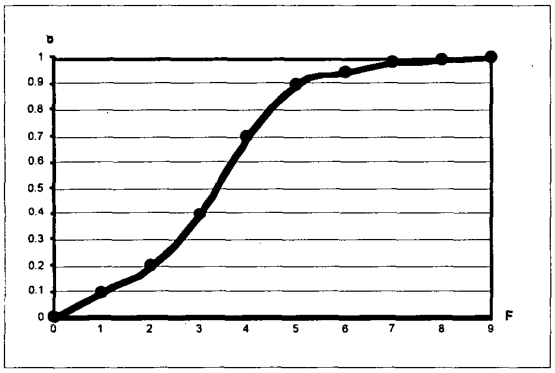

Figure 4. Sigmoid curve with typical occurrence frequencies

3. An example of generating fuzzy document importance degrees from occurrence counts

In the following a very simple example will be presented. We have done a simple query on the legal data base http://www.AustLII.edu.au with the following keyword combination: "(bond* or deposit*) not (no appearance)".As a result, 621 documents have been retrieved. Our example will deal with documents 602 to 621, denoted by $\{D_1,...,D_{20}\}$. We have data for further queries restricted to this collection of 621 documents regarding 100 (key)words. In the example 18 out of these 100 will be presented, according to Table 1:

W	word stem	W	word stem
1	agreement	10	material
2	bedroom	11	occasion
3	carpet	12	premis
4	compensation	13	reasonable
5	damag	14	replac
6	evidenc	15	set
7	follow	16	view
8	liability	17	landlord
9	loss	18	tenant

Table 1. Keyword stems used for the queries in the example

Occurrence frequencies of the above word stems in the collection of documents $\{D_1,...,D_{20}\}$ are shown in Table 2. Based on the occurrence frequency – importance degree transformation sigmoid defined in Fig. 4, the frequencies in Table 2 are transformed into possibilistic importance degrees shown in Table 3.

The 18 words have been selected more or less randomly. However, the last two words ("landlord" and "tenant") were intentionally chosen as they can be expected to appear with rather high counts, because of the type of legal documents that formed the original collection of 621 documents.

It is no surprise that these words show up in almost every document an occurrence count equal to or greater than 9, which was chosen in σ as the threshold value for importance possibility equal to 1. The importance degrees are less than 1 for V_{17} in $)_9$ and $_)$, and for V_{18} in $)_7, D_{14}$ and $Ɔ_{18}$, these degrees being 0.7 and 0.9, and 0.98, 0.95 and 0.99. Even these degrees are at least equal to 0.9, except $\sigma_{17,9} = \sigma(W_{17}, D_9) = 0.7$, in a document that contains anyway a rather low total frequency count of the words in question, compared to most others. Because of this, these two words have to be considered to be relative stop words, and in the further investigations they will be left out completely, as meaningless in this context.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	6	84	0	1	15	2	9	0	5	0	6	17	1	1	4	13	0
2	0	0	0	0	0	0	2	3	0	0	4	9	1	0	0	1	0
3	2	0	9	4	0	0	4	8	0	0	29	0	0	0	0	0	0
4	3	0	1	0	0	4	0	0	0	3	3	4	1	0	1	0	1

Table 2. Occurrence frequency counts of chosen words in the selected collection of documents of the example

W p	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	0.95	1.00	0.00	0.10	1.00	0.20	1.00	0.00	0.90	0.00	0.95	1.00	0.10	0.10	0.70	1.00	0.00
2	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.40	0.00	0.00	0.70	1.00	0.10	0.00	0.00	0.10	0.00
3	0.20	0.00	1.00	0.70	0.00	0.00	0.70	0.99	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.40	0.00	0.10	0.00	0.00	0.70	0.00	0.00	0.00	0.40	0.40	0.70	0.10	0.00	0.10	0.00	0.10

Table 3. Possibilistic importance degrees of chosen words in the selected collection of documents of the example

Having established the fuzzy importance degrees of each of the 20 documents for the 16 meaningful words in question, a few examples for simple queries will be shown. For illustrating the use of fuzzy importance degrees a few "concentrically" widening ad hoc categories of retrieved documents will be defined: *Very Important Documents* ($\sigma = 1$), *Rather Important Documents* ($1 > \sigma \geq 0.9$), *Reasonably Important Documents* ($0.9 > \sigma \geq 0.7$), *Somewhat Important Documents* ($0.7 > \sigma \geq 0.4$) and *Tangentially Important Documents* ($0.4 > \sigma > 0$).

Query 1. "*damag*" W_5

Very Important Documents: D_7

Rather Important Documents: D_{18}

Tangentially Important Documents*:* $D_6, D_8, D_{11}, D_{17}, D_{20}$

Query 2. "*occasion*" W_{11}

Tangentially Important Documents*:*

$D_2, D_5, D_7, D_9, D_{15}, D_{18}, D_{19}, D_{20}$

Comparing these two queries, an important difference can be noted: While for "damag" a document was found that had a very high occurrence count (and another one had a rather high occurrence), for the other word, "occasion" not a single document could be found where the possibility of importance reached 0.5. Even though the number of documents is large where the queried word occurs at all, none of them seems to have real relevance for the word. It is reasonable to introduce the notion of *maximum degree of importance* of a whole collection of documents, which is defined as the t-conorm of membership degrees σ_{ij} for word W_i for all *j*:

$$\omega_i(D) = \omega(W_i, D) = \bigcup_{j=1}^{d} \sigma_{ij} \text{ where } D = \{D_1, ..., D_d\} \cdot \qquad (2)$$

The most often used t-conorms are the max and the algebraic conorm, the latter can be given in closed form by using De Morgan's Law (see [5]).

An advantage of the latter is that it takes into consideration all documents in the collection, if however the number of documents with positive degree is large, becomes rather close to 1, even if the individual degrees are small (see [6]). Because of this, can be considered to be a relative measure of maximum importance, by which various collections of documents can be compared with each other, from the point of view of a given query word. Below, the max type overall degree of importance will be given for the above two query words: $\omega_5^M = 1$ and $\omega_{11}^M = 0.2$.

Another similar measure is the *average frequency of occurrence*, which can be defined as

$$\alpha_i(D) = \frac{|\chi(W_i)|}{d}, \qquad (3)$$

where χ denotes the indicator function of occurrence/no occurrence, and its cardinality is the number of places where it assumes 1. The average occurrence frequencies for the two query words are $\alpha_5 = 0.35$ and $\alpha_{11} = 0.4$.

In the following we discuss the problem of a simple joint query.

Query 3. "*bedroom*" W_2

Very Important Documents: D_{12}

Reasonably Important Documents: D_{11}

Somewhat Important Documents: D_8

Tangentially Important Documents: $D_7, D_{13}, D_{16}, D_{19}$

Query 4. "*carpet*" W_3

Very Important Documents: D_3, D_{11}, D_{20}

Rather Important Documents: D_8

Reasonably Important Documents: D_4, D_7

Tangentially Important Documents: $D_7, D_8, D_{13}, D_{16}, D_{19}$

If the two words are queried jointly, in the sense that the occurrence counts of both words are added, the frequencies shown in the upper half of Table 4 will be obtained. The lower half contains the importance degrees, which are in some of the documents obviously different from the sum of the two importance degrees: for the 7th document we have 0.95 rather than 0.9, for the 8th document we have 1 instead of 0.95+0.4, which would anyway be >1, and in the 11th document, the importance degree 0.7 is completely absorbed by the other as this latter is 1.

W D	1	2	3	4	5	6
2&3	2	0	9	4	0	0

W p	1	2	3	4	5	6
2&3	0.2	0	1	0.7	0	0

Table 4. Added occurrence counts and importance degrees of the words "bedroom" and "carpet"

The retrieved documents are summarised:

Query 5. *"carpet"* OR *"bedroom"* $W_2 \cup W_3$

Very Important Documents: $D_3, D_8, D_{11}, D_{12}, D_{20}$
Rather Important Documents: D_7
Reasonably Important Documents: D_4
Tangentially Important Documents: $D_1, D_{13}, D_{16}, D_{19}$

4. Establishing co-occurrencemaps and fuzzy tolerance relations

Let us address now the problem of fuzzy co-occurrence graphs mapping the mutual relations of keywords into a set of fuzzy degrees. In [2] equivalence of two fuzzy sets is defined, which is usually expressed by the max-min and algebraic norms. Here the fuzzy degrees are represented by the occurrence degrees σ_{ij} . For each pair of words, a series of co-occurrence degrees can be calculated: one for each document in the collection. The *average co-occurrence* will be calculated by applying the arithmetic means aggregation operation for each pair.

Table 5 summarises all co-occurrence degrees in the previous example, using the above max-based definition of fuzzyequivalence.

W W	1	2	3	4	5	6
1	0.95	0.54	0.36	0.51	0.54	0.57
2	0.54	0.94	0.71	0.79	0.77	0.44
3	0.36	0.71	0.96	0.58	0.71	0.48
4	0.51	0.79	0.58	0.89	0.69	0.44

Table 5. Degrees of co-occurrence based on fuzzy equivalence

There are several facts that can be immediately noticed when looking at the table. It is interesting that self-equivalence is not 1, which can be explained by the axiomatic properties of fuzzy operations (cf. [6]). However, for practical purposes, reflexivity will be assumed in the establishing of fuzzy relational maps. Another fact is the symmetry of the table, which results from the symmetric property of the relation described.

In the following, some of the seemingly stronger connections will be pointed out. If self-equivalences are left out of consideration, for the remaining values, the 0.9-cut of the relation contains the following pairs:

$$R_{0.9} = \{\{W_9, W_{10}\}, \{W_9, W_{11}\}, \{W_{10}, W_{11}\}\}$$

All other words appear as isolated points in the relation graph. It is interesting that these three pairs identify a single 0.9-clique of the three words "loss", "material" and "occasion". If we consult Table 2, however, it turns out that all these three words have rather few occurrences. The maximum importance degrees are $\omega_9 = \omega_{10} = \omega_{11} = 0.2$, in all three cases, and the average occurrence frequencies are $\alpha_9 = 0.3, \alpha_{10} = 0.15$ and $\alpha_{11} = 0.4$.

Let us go down with the importance level now to 0.8. There are many resulting pairs, shown graphically below.

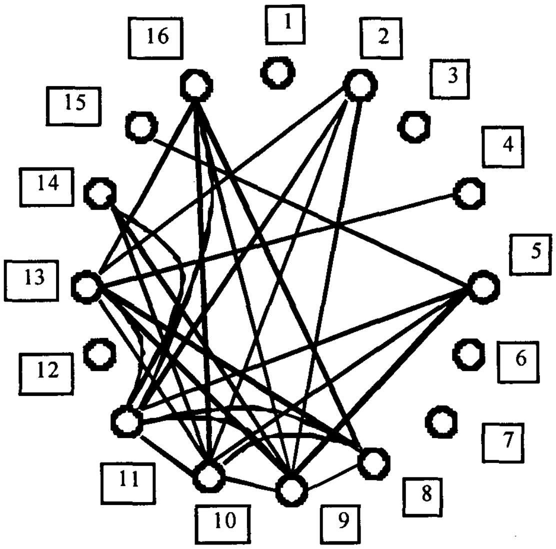

Figure 5. 0.8-cut of the tolerance relation in the example

The maximal tolerance classes found are (by indicating only the indices):
{2,9,10,11,13} = {bedroom, loss, material, occasion, reasonable}; {4,13} = {compensation, reasonable}; {5,9,10,11} = {damag, loss, material, occasion}; {5,15} = {damag, set}; {8,9,10,11,13,16} = {liability, loss, material, occasion, reasonable, view}; {9,10,11,14} = {loss, material, occasion, replac}

Meanwhile, the only compatibility class at the 0.9 possibility level was: { $\{W_9, W_{10}, W_{11}$ = {loss, material, occasion}.

It would be too far fetched to take any conclusion from these classes regarding the meaning or context of these word groups, as the sample of documents used in the example is too small. Let us accept these results anyway for the sake of the demonstration.

It is necessary to see however that in some of the above cases similarity follows from the fact that the words in question occur with low counts, and many overlapping 0 counts increase the degree of equivalence. Because of this, in the next we will modify the graph by multiplying every degree by the average occurrence counts of the two words in question. These frequencies are summarised in Table 6.

1	2	3	4	5	6	7	8
0.75	0.35	0.35	0.5	0.35	0.95	0.7	0.3

Table 6. Average occurrence counts of the words in the example

In the next we apply these values as multiplicative factors on the original fuzzy equivalence degrees. The resulting values will be "weighted equivalences", where in the case of a pair$\{W_i, W_j\}$, the average occurrence counts of both the *i*th and the *j*th word were applied. The resulting values will be considerably smaller.

In this new table there are no large values, indicating that the small amount of random words and the small sample of documents was not really suitable to find out about semantic and contextual connections. When going down with the importance value, the 0.3-cut of this new relation results into the following tolerance groupings: {1,6} and {1,7,12}, that is {agreement, evidenc}and {agreement, follow, premis}.

There is only one larger clique of words for this low degree of importance in this case. Larger sets of words and larger document collections will expectably result in more enlightening word groups.

If this relation is compared with the unweighted one, the astonishing fact will be noticed that the graph of one is close to the logical complement of the other, namely the isolated points there (W_1, W_6, W_7 and W_{12}) are the ones, which are involved here in the highest possibility tolerance classes. The explanation can be found in the occurrence frequencies summarised in Table 2. These four words, but especially "agreement", "evidenc" and "premis" have high occurrence counts (see e.g. documents 1 and 2), and these induce many possibilities close to 1 in Table 3. While rows where the occurrence counts in many columns are zero, automatically generate high fuzzy equivalence values according to the formula at the beginning of this section ($\max\{\min\{0,0\},\min\{1-0,1-0\}\}=1$), and so, suggest some contextual connection (however based on negative evidence, i.e., on the *lack of both words in most of the documents*), rows with necessarily more random higher positive values in them produce only lower possibilistic tolerance connections among them. When the average occurrence weight comes into the formula, the rather meaningless equivalence of rare words will automatically loose weight and real equivalences emerge. It is one of the tasks of further research to find out, what should be the optimal weighting factor that does not hide the original connections based on absolute occurrence counts, but does not let rare words come too much into focus just because of their numerous occurrences.

5. Conclusions

In this study the simplest elements of fuzzy tolerance relation based intelligent queries were presented and illustrated. It has been shown that it is possible to transform occurrence frequency counts into possibilistic fuzzy importance degrees by using sigmoid type transformation functions, and that by using fuzzy logical equivalence functions, it is possible to determine fuzzy degrees expressing the possibility of two or more words occurring together in documents. Fuzzy relational maps express the connections among words and consequently help to find documents with hidden relations to the query. The average occurrence count was also introduced as a modifying factor that helps to exclude the assumption of semantic connection based overwhelmingly on negative evidence (the joint lack of occurrence in most documents). Some examples have been presented.

It will be necessary to extend investigations with larger sets of words (possibly with obvious connections among some of them), and larger document collections for generating the relational map. Testing these graphs should be done on independent collections, and by the involvement of experts assessing the subjective degree of matching between the queried words or phrases and the retrieved documents.

In the next step hierarchical co-occurrencerelations must be established, based on the ideas in [2,3] and following the practical approaches in this study. However, in that case the set of keywords and general important words must be necessarily even larger. A major problem is the computational complexity aspect of finding all compatibility (tolerance) classes in relational graphs of large size, which problem must be also addressed in future work.

References

[1] P. Wallis and J. A. Thom: Relevance judgments for assessing recall, *Information Processing and Management 32* (1996), pp. 273-286.

[2] L. T. Kóczy and T. D. Gedeon: Information retrieval by fuzzy relations and hierarchical co-occurrence, Part I, TR97-01, Dept. of Info. Eng., School of Comp. Sci. & Eng., UNSW, 1997, 18p.

[3] L. T. Kóczy and T. D. Gedeon: Information retrieval by fuzzy relations and hierarchical co-occurrence, Part II, TR97-03, Dept. of Info. Eng., School of Comp. Sci. & Eng., UNSW, 1997, 8p.

[4] P. Baranyi, T. D. Gedeon and L. T. Kóczy: Improved fuzzy and neural network algorithms for frequency prediction in document filtering, TR97-02, Dept. of Info. Eng., School of Comp. Sci. & Eng., UNSW, 1997, 21p.

[5] G. Klir and T. Folger: Fuzzy Sets, Uncertainty and Information, Prentice Hall, Englewood Cliffs, NJ, 1988.

[6] L. T. Kóczy: Interactive σ-algebras and fuzzy objects of type N, *J. of Cybernetics 8* (1978), pp. 273-290.