ELSEVIER

Fuzzy Sets and Systems 126 (2002) 49–61

FUZZY sets and systems

www.elsevier.com/locate/fss

Fuzzy tolerance relations and relational maps applied to information retrieval ☆

László T. Kóczy[a, *], Tamás D. Gedeon[b, 1], Judit A. Kóczy[c, 2]

[a] *Department of Telecommunication and Telematics, Technical University of Budapest, Sztoczek u.2, Budapest H-1521, Hungary*

[b] *Department of Information Engineering, School of Computer Science and Engineering, University of New South Wales, Sydney 2052, Australia*

[c] *CONTROLLTraining Education Centre Ltd. Co., 23 Csalogány, Budapest H-1027, Hungary*

Received 25 March 1999; received in revised form 23 January 2001; accepted 2 February 2001

Abstract

One of the major problems in automatic indexing and retrieval of documents is that usually it cannot be guaranteed that the user queries include (all) of the actual words that occur in the documents that should be retrieved. Also it often happens that words with several meanings occur in a document, but in a rather different context from that expected by the querying person. In order to achieve better recall and higher precision, fuzzy tolerance and similarity relations have been 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 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. © 2002 Elsevier Science B.V. All rights reserved.

1. Introduction

An information retrieval system allows users to efficiently retrieve documents that are relevant to their current interests. The collection of documents from which the selected ones have to be retrieved might be extremely large and the use of terminology might be inconsistent. If the language of the documents is close to natural language (like in legal texts) this becomes especially obvious.

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

☆ This research was supported by the Australian Research Council.

* Corresponding author. Tel.: +36-1-463-4190; fax: +36-1-463-3107.

E-mail addresses: koczy@boss.ttt.bme.hu (L.T. Kóczy), tom@cse.unsw.edu.au (T.D. Gedeon), koczy@controll.hu (J.A. Kóczy).

[1] At time of this research Visiting Professor at the Department of Information Engineering, School of Computer Science and Engineering, University of New South Wales, Sydney, Australia.

[2] At time of this research Visiting Honorary Associate at the Department of Information Engineering, School of Computer Science and Engineering, University of New South Wales, Sydney, Australia.

PII: S0165-0114(01)00054-9

the topic searched for should be high, that is the set of relevant documents retrieved 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 which are not relevant for the given query, being equivalent with the expectation of obtaining an as small as possible retrieved document set (cf. [10]).

Automated keyword search is the most widespread approach to this problem; however, it is easy to recognise that documents not containing the actual keyword(s), but maybe its synonyms or some terms with a closely related but more specific meaning, might be similarly relevant for the search. If the keyword in the query is *Soft Computing* (*SC*), 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 be also retrieved.

On the other hand, if the query specifies the two keywords Fuzzy and Relation the humorous situation might occur that 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 previous studies 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 [5–7]. This means that the occurrence frequencies of certain words in the title, sub-titles, abstract or conclusion parts of documents might be characteristic for the occurrence frequencies of certain (other) words in the main body of the text. The frequency of word A in the title and word B in the text is called their hierarchical co-occurrence. 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 into possibility measures (fuzzy membership degrees) will be presented frequencies. (For a neural networks based estimation technique, see [1]. The content of this paper is mainly based on [8].)

Applying fuzzy logic to automated information retrieval is not new. Some of the most important advances in this field are summarised in [9]. In several points of this paper, reference will be made to concepts introduced in this work.

2. Keyword occurrence frequencies and possibility degrees

Both occurrence and co-occurrence of keywords can be expressed with the help of word counts in documents. If analysing a collection of documents related to a certain topic (e.g. legal documents) it will be found that some of the words occur quite frequently in all or most of them, thus these words are of no significance with regards to the contents of any particular document. The words which are common in any natural language document are called stop words, while those which might be significant in some context but have a role similar to that of the real stop words in a certain context will be called in this study *relative stop words*. As an example, let us consider the word "law" which would certainly occur rather frequently in any legal document, and would be not discriminative concerning the particular contents of such a document. 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 can be seen 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 (to be more exact, a word stem).

It is obvious that keywords searched in titles, abstracts, introductions, etc. will have a lower occurrence count than significant words in general (including keywords as special significant words) in the full text.

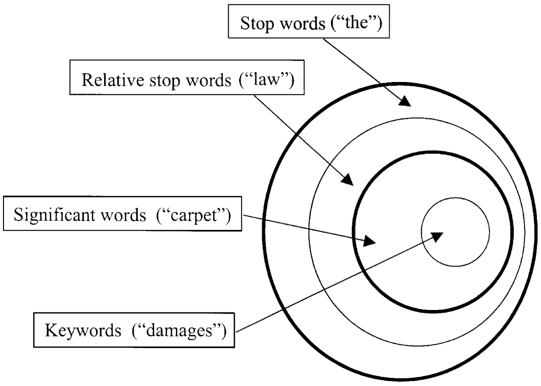


Fig. 1. Categories of words in documents.

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.

The following must be considered here. 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 in a document very frequently are usually stop words (absolute or relative ones), and so they should be left out of consideration. For the remaining class of 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 highly likely that this document will be rather important for the querying person 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 mentioning (three or four or more) is an indicator that the word in question is an important word from the point of view of the document. With short documents these numbers might vary. It is quite different with keyword occurrences in titles or subtitles 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 fulfill the following conditions:

$$\sigma : fR^{+} \rightarrow [0,1],$$

$$\sigma(F_1) > \sigma(F_2) \Leftrightarrow F_1 > F_2,$$

$$\frac{\mathrm{d}^2(\sigma)}{\mathrm{d}F^2} \geqslant 0 \Leftrightarrow F \leqslant T_F$$

and

$$\frac{\mathrm{d}^2(\sigma)}{\mathrm{d}F^2} \leqslant 0 \Leftrightarrow F \geqslant T_F.$$

(Here T_F is a suitable threshold value.) In practice σ is not necessarily continuously differentiable, but its characteristics should be nevertheless "S-shaped". This function is rather similar to other frequency-membership mapping functions, e.g. to g in [9, p. 88]. In this paper some practically applicable concrete sigmoid functions will be tested and compared, defined by tabular values rather than transcendental expressions. However, in [9] basically grades of relevance are used instead of normalised fuzzy membership functions, which have the disadvantage of reflecting ordering, but no metric, i.e., no measurable degree of similarity and dissimilarity between two word/document pairs.

Although occurrence frequencies are integers, it is reasonable to introduce the sigmoid mapping over the whole positive half of the real lines, as in [7]. The importance degrees are introduced as convex combinations of occurrence counts (e.g. $F_{ij} = \lambda_1 T_{ij} + \lambda_2 C_{ij} + \lambda_3 L_{ij}$, where λ_i are real coefficients and T, C and L denote title-keyword, location-keyword and cue words related frequencies, respectively). These are the occurrence frequencies of the particular word in the title and subtitles; in the abstract, introduction and/or conclusion part of the document; and in the immediate neighborhood of such terms as "topic", "aim", "main point", etc.). So these fictitious ("equivalent") occurrence frequencies (which in reality are the weighted averages of various real frequencies) might assume any non-negative value. The typical characteristics of such a sigmoid function can be seen in Fig. 2.

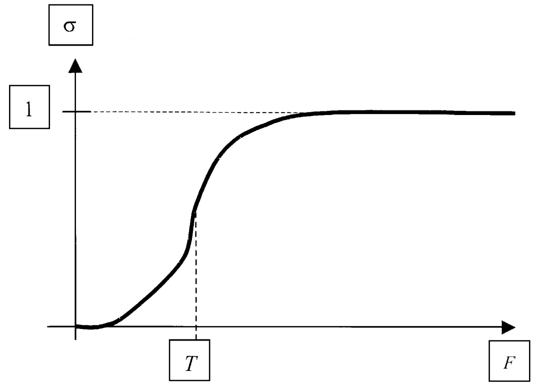


Fig. 2. Sigmoid function transforming occurrence frequencies into membership degrees.

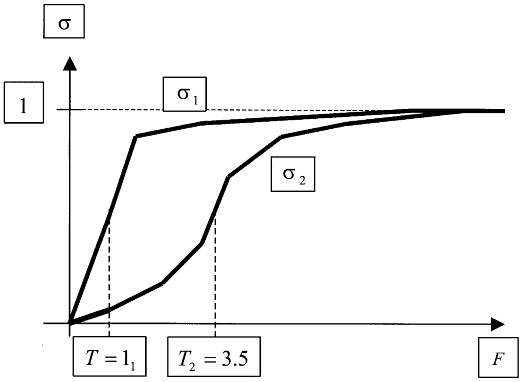


Fig. 3. Sigmoid curves for title/subtitle and text occurrence (σ_1 is steeper for titles, while the other is smoother as in the main body of the text higher occurrence is needed for rather high relevance).

More practical broken line functions with concrete values can be seen in Fig. 3. Here σ_1 is a mapping for title (subtitle) occurrences and σ_2 another one for text occurrences. The threshold values are obviously different. Depending on the length of the document, the number of levels of subtitles, etc., the characteristics of the sigmoid curve can change.

Membership degrees generated by the occurrence frequency transformation can be interpreted as possibility measures of a certain document being important for a querying person if the given word was included in the query keyword set. Although possibility has some similarities with probability, its axiomatic properties differ in an essential point: additivity does not hold [3]. It is easy to realise this when considering the sigmoids. Let us demonstrate this by the following table defining a sigma for integer values of F (this definition will be used throughout the paper in all of our examples):

3. An example of generating fuzzy document importance (relevance) 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)", using a traditional non-intelligent retrieval system. (The '*' indicates here that all appearances of the word stems were included.) As a result, 621 documents have been retrieved. In this paper we present a small representative example where the sizes of tables are small enough to be printed within the text. Because of this, only the last 20 documents will be considered: documents 602 to 621, denoted by $\{D_1,\ldots,D_{20}\}$. (This was a random selection from the retrieved set, the sequence corresponded only to the physical location of the documents in the original collection and had no connection with the contents.) 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.

Occurrence frequencies of the above word stems in the collection of documents $\{D_1,\ldots,D_{20}\}$ 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 with 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 W_{17} in D_9 and D_{14}, and for W_{18}

Table 1
Keyword stems used for the queries in the example

W	Word stem
1	Agreement
2	Bedroom
3	Carpet
4	Compensation
5	Damag
6	Evidenc
7	Follow
8	Liability
9	Loss
10	Material
11	Occasion
12	Premis
13	Reasonable
14	Replac
15	Set
16	View
17	Landlord
18	Tenant

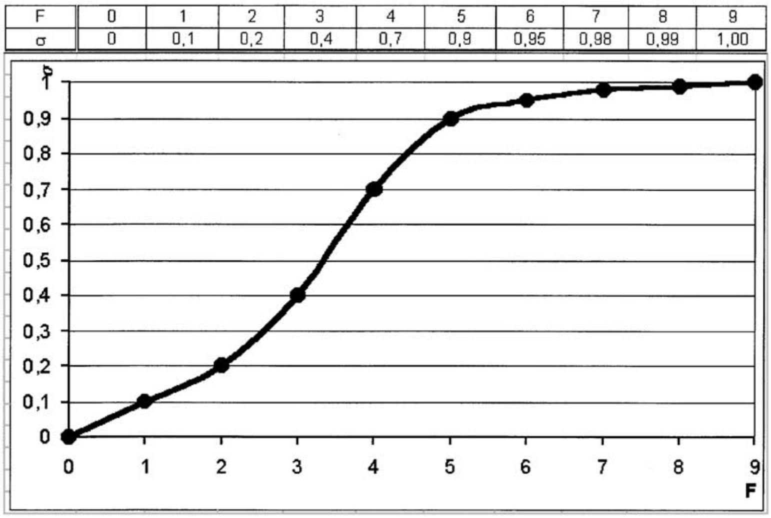
in D_7, D_{14} and D_{18}, these degrees being 0.7 and 0.9, and 0.98, 0.95 and 0.99. Even these degrees are at least equal to 0.9, except $\sigma_{17,9} = \sigma(W_{17}, D_9) = 0.7$, in a document that contains anyway a rather low total frequency count of the words in question, compared to most others. 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.

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* ($1 > \sigma \geqslant 0.9$), *Reasonably Important Documents* ($0.9 > \sigma \geqslant 0.7$), *Somewhat Important Documents* ($0.7 > \sigma \geqslant 0.4$) and *Tangentially Important Documents* ($0.4 > \sigma > 0$). As a matter of course, the threshold values can be adapted to any concrete application.

Query 1. "*damag*" W_5
Very Important Documents: D_7
Rather Important Documents: D_{18}
Reasonably Important Documents: $\emptyset$
Somewhat Important Documents: $\emptyset$
Tangentially Important Documents:
$D_6, D_8, D_{11}, D_{17}, D_{20}$.

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

$W \backslash D$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	6	84	0	1	15	2	9	0	5	0	6	17	1	1	4	13	0	7	9	0
2	0	0	0	0	0	0	2	3	0	0	4	9	1	0	0	1	0	0	2	0
3	2	0	9	4	0	0	4	8	0	0	29	0	0	0	0	0	0	0	0	51
4	3	0	1	0	0	4	0	0	0	3	3	4	1	0	1	0	1	0	6	0
5	0	0	0	0	0	1	21	2	0	0	2	0	0	0	0	0	1	6	0	1
6	12	23	9	1	5	0	2	7	2	6	7	5	1	1	5	1	12	6	7	19
7	1	7	0	0	1	1	2	3	3	1	4	1	1	0	1	4	1	0	0	0
8	0	5	0	1	4	0	0	0	0	0	0	1	0	0	0	0	0	0	1	2
9	0	1	2	0	0	1	1	0	0	0	0	1	0	0	1	0	0	0	0	0
10	0	1	0	0	0	0	0	0	0	0	0	0	1	0	2	0	0	0	0	0
11	0	1	0	0	1	0	1	0	2	0	0	0	0	0	1	0	0	1	1	1
12	15	31	4	4	14	2	13	9	4	5	13	28	5	2	1	5	2	19	40	1
13	1	17	2	0	0	0	0	0	0	2	1	4	0	0	0	0	0	0	9	0
14	2	2	4	0	0	0	1	2	4	0	0	0	2	0	0	0	0	0	0	9
15	1	12	1	0	0	2	4	0	0	0	1	1	0	0	0	0	7	4	1	0
16	5	3	1	0	0	0	0	0	0	0	3	0	0	0	1	0	2	2	3	0
17	22	40	28	12	19	9	14	10	4	29	17	38	13	5	32	16	23	14	54	32
18	44	42	18	16	26	11	7	16	12	32	38	54	27	6	25	12	21	8	42	43



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

Fig. 4. Example for sigmoid curve with typical occurrence frequencies.

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

$W\backslash D$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
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	0.98	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	0.00	0.20	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	0.00	0.00	1.00
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	0.00	0.95	0.00
5	0.00	0.00	0.00	0.00	0.00	0.10	1.00	0.20	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.10	0.95	0.00	0.10
6	1.00	1.00	1.00	0.10	0.90	0.00	0.20	0.98	0.20	0.95	0.98	0.90	0.10	0.10	0.90	0.10	1.00	0.95	0.98	1.00
7	0.10	0.98	0.00	0.00	0.10	0.10	0.20	0.40	0.40	0.10	0.70	0.10	0.10	0.00	0.10	0.70	0.10	0.00	0.00	0.00
8	0.00	0.90	0.00	0.10	0.70	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.20
9	0.00	0.10	0.20	0.00	0.00	0.10	0.10	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00
10	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.20	0.00	0.00	0.00	0.00	0.00
11	0.00	0.10	0.00	0.00	0.10	0.00	0.10	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.10	0.10	0.10
12	1.00	1.00	0.70	0.70	1.00	0.20	1.00	1.00	0.70	0.90	1.00	1.00	0.90	0.20	0.10	0.90	0.20	1.00	1.00	0.10
13	0.10	1.00	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.10	0.70	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
14	0.20	0.20	0.70	0.00	0.00	0.00	0.10	0.20	0.70	0.00	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.00	1.00
15	0.10	1.00	0.10	0.00	0.00	0.20	0.70	0.00	0.00	0.00	0.10	0.10	0.00	0.00	0.00	0.00	0.98	0.70	0.10	0.00
16	0.90	0.40	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.40	0.00	0.00	0.00	0.10	0.00	0.20	0.20	0.40	0.00
17	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.70	1.00	1.00	1.00	1.00	0.90	1.00	1.00	1.00	1.00	1.00	1.00
18	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00	1.00	1.00	1.00	0.95	1.00	1.00	1.00	0.99	1.00	1.00

Query 2. "*occasion*" W_{11}
VeryImportantDocuments: $\emptyset$
RatherImportantDocuments: $\emptyset$
ReasonablyImportantDocuments: $\emptyset$
SomewhatImportantDocuments: $\emptyset$
TangentiallyImportantDocuments:
$D_2, D_5, D_7, D_9, D_{15}, D_{18}, D_{19}, D_{20}$.

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 where the queried word occurs at all is large, none of them seems to have real relevance to this 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

$$\omega_i(D) = \omega(W_i, D) = \bigcup_{j=1}^{d} \sigma_{ij}$$

where $D = \{D_1, \ldots, D_d\}$.

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 [6])

$$\omega_i^{\mathrm{M}}(D) = \max_{j=1}^{d}\{\sigma_{ij}\}$$

$$\omega_i^{\mathrm{A}}(D) = 1 - \prod_{j=1}^{d}(1 - \sigma_{ij}).$$

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, ω becomes rather close to 1, even if the individual degrees are small (see [4]). Because of this, ω 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:

$$\omega_5^{\mathrm{M}} = 1 \quad \text{and} \quad \omega_{11}^{\mathrm{M}} = 0.2.$$

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

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

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

$$\alpha_5 = 0.35 \quad \text{and} \quad \alpha_{11} = 0.4.$$

In the following we discuss the problem of a simple joint query. As an example, let us choose two words, W_2 and W_3 ("bedroom" and "carpet"). First, the single queries are presented.

Query 3. "*bedroom*" W_2
VeryImportantDocuments: D_{12}
RatherImportantDocuments: $\emptyset$
ReasonablyImportantDocuments: D_{11}
SomewhatImportantDocuments: D_8
TangentiallyImportantDocuments:
$D_7, D_{13}, D_{16}, D_{19}$.

Query 4. "*carpet*" W_3
VeryImportantDocuments: D_3, D_{11}, D_{20}
RatherImportantDocuments: D_8
ReasonablyImportantDocuments: D_4, D_7
SomewhatImportantDocuments: $\emptyset$
TangentiallyImportantDocuments:
$D_7, D_8, D_{13}, D_{16}, D_{19}$.

There are various logical ways to perform a joint query. If two words, in this example, W_2 and W_3, have similar relevance for the query (they might be, e.g., synonyms or complementary terms), every document containing *any of the terms* is relevant. In such a case the two words are queried jointly, in the sense that the occurrence counts of both words are added. In the above example, the frequencies shown in the upper half of Table 4a 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. In this approach a bounded sum operator was applied. There are many alternatives, such as, e.g., the max operator expresses another form of "OR-ness".

The retrieved documents are summarised in the following:

Query 5. "carpet" OR "bedroom" ($W_2 \cup W_3$)
VeryImportantDocuments: $D_3, D_8, D_{11}, D_{12}, D_{20}$
RatherImportantDocuments: D_7
ReasonablyImportantDocuments: D_4
SomewhatImportantDocuments: $\emptyset$
TangentiallyImportantDocuments:
$D_1, D_{13}, D_{16}, D_{19}$.

This query answers the question "Which documents are relevant for the terms carpet or bedroom?"

Table 4a
Added occurrence counts and importance degrees of the query “bedroom or carpet”

$W\backslash D$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2V3	2	0	9	4	0	0	6	11	0	0	33	9	1	0	0	1	0	0	2	51
2V3	0.2	0	1	0.7	0	0	1	1	0	0	1	1	0.1	0	0	0.1	0	0	0.2	1

Table 4b
Minimal occurrence counts and joint importance degrees of the query “bedroom and carpet”

$W\backslash D$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2&3	0	0	0	0	0	0	2	3	0	0	4	0	0	0	0	0	0	0	0	0
2&3	0	0	0	0	0	0	0.2	0.8	0	0	0.2	0	0	0	0	0	0	0	0	0

Obviously, if the joint query is understood in a conjunctive sense, the occurrence counts and relevance degrees must be combined in a different way, e.g., by taking the min conjunction of the two relevance degrees. In this case, documents containing *both terms* are sought. For the resulting values see Table 4b. (In this case, only three documents have been retrieved that contain both query words, and only one of them has a high degree of importance, 0.8).

4. Establishing co-occurrence maps 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 [7], the equivalence of two fuzzy sets is defined by $A\equiv B \mathrel{\hat{=}} (A\wedge B)\vee(\neg A\wedge\neg B)$, which is usually expressed by the max–min or algebraic norms as

$$\mu^{Z}_{A\tilde{\equiv}B}(x)=\max\{\min\{\mu_A(x),\mu_B(x)\},\\ \min\{1-\mu_A(x),1-\mu_B(x)\}\},$$

or

$$\mu^{A}_{A\tilde{\equiv}B}(x)=\mu_A(x)\mu_B(x)+[1-\mu_A(x)][1-\mu_B(x)]\\ -\mu_A(x)\mu_B(x)[1-\mu_A(x)][1-\mu_B(x)],$$

respectively. The two expressions are not equivalent, the max–min one (based on Zadeh’s original definitions) is easier to calculate; the algebraic one is, however, more sensitive. Because of computational reasons, the first one will be used in all examples throughout this paper (the superscript being omitted). 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:

$$\mu_{ij}=\mu_{W_i\equiv W_j}=\frac{1}{d}\sum_{k=1}^{d}\mu_{W_i\equiv W_j}(D_k).$$

Crisp equivalence relations necessarily satisfy three properties:

- Reflexivity ($a\equiv a$)
- Symmetry ($a\equiv b\Rightarrow b\equiv a$)
- Transitivity ($a\equiv b\wedge b\equiv c\Rightarrow a\equiv c$).

Fuzzy similarity/tolerance relations, especially those based on real data might only approximately satisfy these properties. Table 5 summarises all co-occurrence degrees in the previous example (the two contextual stop words having been omitted already), using the above max–min-based definition of fuzzy

Table 5
Degrees of co-occurrence based on fuzzy equivalence

$W\backslash W$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0.95	0.54	0.36	0.51	0.54	0.57	0.58	0.53	0.46	0.46	0.48	0.69	0.57	0.40	0.53	0.55
2	0.54	0.94	0.71	0.79	0.77	0.44	0.76	0.78	0.85	0.85	0.83	0.40	0.80	0.72	0.69	0.77
3	0.36	0.71	0.96	0.58	0.71	0.48	0.64	0.64	0.71	0.70	0.69	0.41	0.59	0.76	0.60	0.65
4	0.51	0.79	0.58	0.89	0.69	0.44	0.65	0.71	0.79	0.79	0.77	0.38	0.81	0.66	0.65	0.78
5	0.54	0.77	0.71	0.69	0.96	0.37	0.71	0.77	0.85	0.85	0.85	0.39	0.71	0.73	0.82	0.77
6	0.57	0.44	0.48	0.44	0.37	0.94	0.40	0.42	0.34	0.34	0.35	0.61	0.49	0.42	0.45	0.46
7	0.58	0.76	0.64	0.65	0.71	0.40	0.88	0.79	0.79	0.79	0.78	0.43	0.74	0.70	0.72	0.74
8	0.53	0.78	0.64	0.71	0.77	0.42	0.79	0.96	0.87	0.88	0.87	0.36	0.84	0.77	0.79	0.80
9	0.46	0.85	0.71	0.79	0.85	0.34	0.79	0.87	0.97	0.96	0.94	0.29	0.83	0.82	0.79	0.85
10	0.46	0.85	0.70	0.79	0.85	0.34	0.79	0.88	0.96	0.98	0.95	0.27	0.83	0.83	0.79	0.86
11	0.48	0.83	0.69	0.77	0.85	0.35	0.78	0.87	0.94	0.95	0.96	0.30	0.81	0.82	0.78	0.84
12	0.69	0.40	0.41	0.38	0.39	0.61	0.43	0.36	0.29	0.27	0.30	0.90	0.42	0.31	0.38	0.39
13	0.57	0.80	0.59	0.81	0.71	0.49	0.74	0.84	0.83	0.83	0.81	0.42	0.96	0.71	0.76	0.80
14	0.40	0.72	0.76	0.66	0.73	0.42	0.70	0.77	0.82	0.83	0.82	0.31	0.71	0.93	0.67	0.73
15	0.53	0.69	0.60	0.65	0.82	0.45	0.72	0.79	0.79	0.79	0.78	0.38	0.76	0.67	0.93	0.75
16	0.55	0.77	0.65	0.78	0.77	0.46	0.74	0.80	0.85	0.86	0.84	0.39	0.80	0.73	0.75	0.91

equivalence. (The use of the algebraic operation might lead to very small absolute membership values, inconvenient to handle.)

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. [4]). 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. It must be remarked that this table can be interpreted as the starting table for constructing a fuzzy thesaurus in the sense of [9]. However, an essential difference is that fuzzy similarity relations based on co-occurrence frequencies are not generated by experts' opinions, but by a fully automatic probability/possibility transformation. Further, it does not reflect actual similarity in the meaning, but simply relatedness in the given context of the chosen document collection, in our example the legal document set specified earlier. It might easily happen that antonyms have a very strong connection, or that nouns going usually with certain verbs in this type of legal text will be tightly related. We shall see some examples in the next sections.

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

$$\alpha_9 = 0.3, \quad \alpha_{10} = 0.15 \text{ and } \alpha_{11} = 0.4.$$

Let us go down with the importance level now to 0.8. Resulting pairs are

$$\begin{aligned} R_{0.8} = \{&\{W_2, W_9\}, \{W_2, W_{10}\}, \{W_2, W_{11}\}, \{W_2, W_{13}\},\\ &\{W_4, W_{13}\}, \{W_5, W_9\}, \{W_5, W_{10}\}, \{W_5, W_{11}\},\\ &\{W_5, W_{15}\}, \{W_8, W_9\}, \{W_8, W_{10}\}, \{W_8, W_{11}\},\\ &\{W_8, W_{13}\}, \{W_8, W_{16}\}, \{W_9, W_{10}\}, \{W_9, W_{11}\},\\ &\{W_9, W_{13}\}, \{W_9, W_{14}\}, \{W_9, W_{16}\},\end{aligned}$$

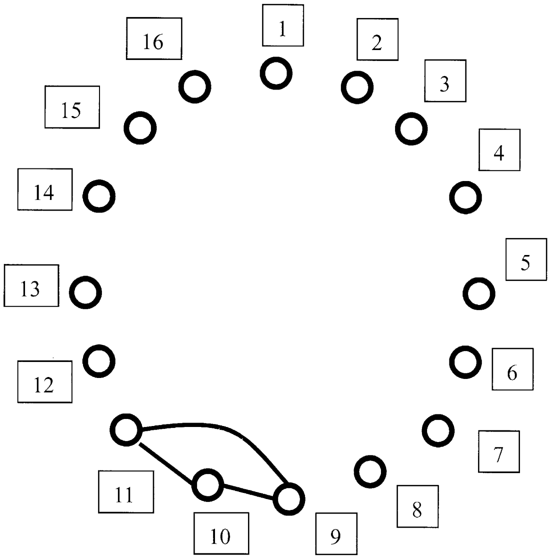


Fig. 5. 0.9-cut of the tolerance relation in the example.

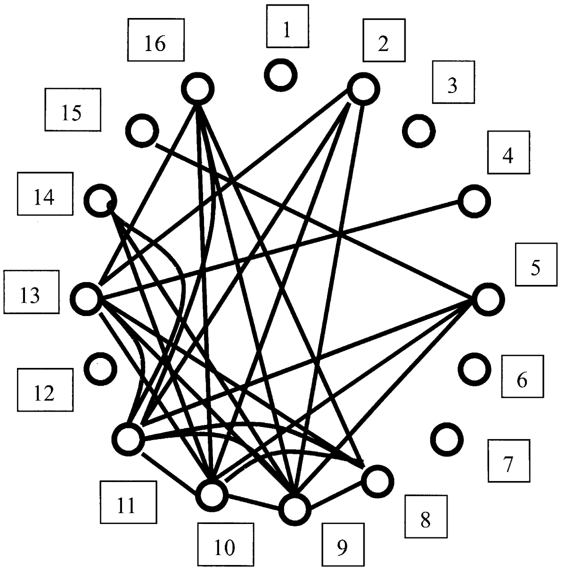


Fig. 6. 0.8-cut of the tolerance relation in the example.

$$\{W_{10},W_{11}\},\{W_{10},W_{13}\},\{W_{10},W_{14}\},$$
$$\{W_{10},W_{16}\},\{W_{11},W_{13}\},\{W_{11},W_{14}\},$$
$$\{W_{11},W_{16}\},\{W_{13},W_{16}\}\}.$$

To provide a better overview, the two cuts of the relation will be presented by graphs (see Figs. 5 and 6).

The only compatibility class at this possibility level is:

$$\{W_9,W_{10},W_{11}\} = \{\text{loss, material, occasion}\}.$$

The maximal tolerance classes found are (by indicating only the indices):

$\{2,9,10,11,13\} =$
$\{\text{bedroom, loss, material, occasion, reasonable}\}$,
$\{4,13\} = \{\text{compensation, reasonable}\}$,
$\{5,9,10,11\} = \{\text{damag, loss, material, occasion}\}$,
$\{5,15\} = \{\text{damag, set}\}$,
$\{8,9,10,11,13,16\} =$
$\{\text{liability, loss, material, occasion, reasonable, view}\}$,
$\{9,10,11,14\} = \{\text{loss, material, occasion, replac}\}$.

(A maximal tolerance or compatibility class is a maximal clique of a graph representing a—non-transitive—tolerance or compatibility relation.)

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, and let us investigate some new queries based on the here-established tolerance classes.

Query 6. "*loss*", "*material*"

The smallest tolerance class containing these two words is $\{W_9,W_{10},W_{11}\} = \{\text{loss, material, occasion}\}$ with 0.9 importance degree, and the output is

Very Important Documents : $\emptyset$
Rather Important Documents : $\emptyset$
Reasonably Important Documents : D_{15}
Somewhat Important Documents : D_2
Tangentially Important Documents :
$D_3, D_5, D_6, D_7, D_9, D_{12}, D_{13}, D_{18}, D_{19}, D_{20}$.

Note that documents no. 5, 9, 18, 19 and 20 are included in the "tangentially important" set, although none of the queried words occurs in them—this is where the use of the tolerance class brings in some "unexpected" suggestions for matches.

Table 6
Average occurrence counts of the words in the example

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0.75	0.35	0.35	0.5	0.35	0.95	0.7	0.3	0.3	0.15	0.4	1	0.35	0.4	0.5	0.4

Table 7
Equivalence degrees modified with average occurrence counts

$W\backslash W$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0.534	0.142	0.095	0.191	0.142	0.406	0.305	0.119	0.104	0.052	0.144	0.518	0.15	0.12	0.199	0.165
2	0.142	0.115	0.087	0.138	0.094	0.146	0.186	0.082	0.089	0.045	0.116	0.14	0.098	0.101	0.121	0.108
3	0.095	0.087	0.118	0.102	0.087	0.16	0.157	0.067	0.075	0.037	0.097	0.144	0.072	0.106	0.105	0.091
4	0.191	0.138	0.102	0.223	0.121	0.209	0.228	0.107	0.119	0.059	0.154	0.19	0.142	0.132	0.163	0.156
5	0.142	0.094	0.087	0.121	0.118	0.123	0.174	0.081	0.089	0.045	0.119	0.137	0.087	0.102	0.144	0.108
6	0.406	0.146	0.16	0.209	0.123	0.848	0.266	0.12	0.097	0.048	0.133	0.58	0.163	0.16	0.214	0.175
7	0.305	0.186	0.157	0.228	0.174	0.266	0.431	0.166	0.166	0.083	0.218	0.301	0.181	0.196	0.252	0.207
8	0.119	0.082	0.067	0.107	0.081	0.12	0.166	0.086	0.078	0.04	0.104	0.108	0.088	0.092	0.119	0.096
9	0.104	0.089	0.075	0.119	0.089	0.097	0.166	0.078	0.087	0.043	0.113	0.087	0.087	0.098	0.119	0.102
10	0.052	0.045	0.037	0.059	0.045	0.048	0.083	0.04	0.043	0.022	0.057	0.041	0.044	0.05	0.059	0.052
11	0.144	0.116	0.097	0.154	0.119	0.133	0.218	0.104	0.113	0.057	0.154	0.12	0.113	0.131	0.156	0.134
12	0.518	0.14	0.144	0.19	0.137	0.58	0.301	0.108	0.087	0.041	0.12	0.9	0.147	0.124	0.19	0.156
13	0.15	0.098	0.072	0.142	0.087	0.163	0.181	0.088	0.087	0.044	0.113	0.147	0.118	0.099	0.133	0.112
14	0.12	0.101	0.106	0.132	0.102	0.16	0.196	0.092	0.098	0.05	0.131	0.124	0.099	0.149	0.134	0.117
15	0.199	0.121	0.105	0.163	0.144	0.214	0.252	0.119	0.119	0.059	0.156	0.19	0.133	0.134	0.233	0.15
16	0.165	0.108	0.091	0.156	0.108	0.175	0.207	0.096	0.102	0.052	0.134	0.156	0.112	0.117	0.15	0.146

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 calculated importance degree by the average occurrence counts of the two words in question. This will take care of the normalisation problem, as well. These frequencies are summarised in Table 6.

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 ith and the jth word were applied. The resulting values will be considerably smaller as shown in Table 7.

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. Fig. 7 depicts the 0.3-cut of the new, weighted relation thus obtained.

If this relation is compared with the unweighted one, the astonishing fact will be noticed that the

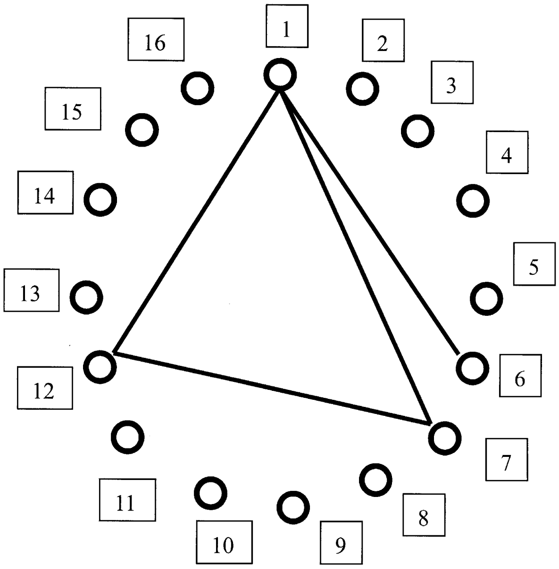


Fig. 7. 0.3-cut of the fuzzy tolerance relation obtained by weighting co-occurrence possibilities with the average occurrence counts.

graph of Fig. 7 is close to the logical complement of the one in Fig. 6. In an interesting way 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. 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, this is based on negative evidence, i.e., on the *lack of both words in most of the documents* and 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.

A series of case studies, involving various weightings and sigmoid function shapes, using the mentioned legal data base, and involving some legal expert opinion concerning the tolerance classes established can be found in [2].

5. Conclusions and further study

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, so that these degrees reflect real fuzzy membership or possibility. This is by the way connected to an interesting and unresolved problem, namely, how to transform probabilistic frequencies into possibilistic relevance degrees. It has been also shown 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. This way we presented a fully automatic method to establish a certain type of fuzzy thesaurus, one with predetermined mathematical properties. 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.

We found it 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, involving also checks with experts' assessments (cf. [2]). Further testing these graphs should be done on independent collections, and by further involvement of (e.g. legal) experts assessing the subjective degree of matching between the queried words or phrases and the retrieved documents, and also the connection of contents of the documents identified as belonging to the same relevance class. There are also some theoretical consequences of the comparison of fuzzy tolerance classes thus obtained and contents connections established by experts' assessments: the notion of fuzzy tolerance class might be generalized towards more flexible categories, fitting the structure of fuzzy tolerance maps obtained from certain document collections.

In the next step hierarchical co-occurrence relations must be established, based on the ideas in [5–7] 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. This research will be continued in the future. Another important point is that besides the examples in the projects, the same methods could be tested with standard test data available in the literature. This way experimental evaluation and comparison with other approaches will be possible.

References

[1] P. Baranyi, T.D. Gedeon, L.T. Kóczy, Improved fuzzy and neural network algorithms for frequency prediction in document filtering, TR97-02, Department of Information Engineering, School of Computer Science and Engineering, University of New South Wales, Sydney, 1997; J. Adv. Comput. Intell. (1998) 88–95.

[2] K. Chakrabarty, L.T. Kóczy, T.D. Gedeon, Information retrieval in legal documents by fuzzy relational charts, J. Amer. Soc. Info. Sci., submitted.

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

[4] L.T. Kóczy, Interactive σ-algebras and fuzzy objects of type N, J. Cybernet. 8 (1978) 273–290.

[5] L.T. Kóczy, T.D. Gedeon, Information retrieval by fuzzy relations and hierarchical co-occurrence, Part I, TR97-01, Department of Information Engineering, School of Computer Science and Engineering, University of New South Wales, Sydney, 1997.

[6] L.T. Kóczy, T.D. Gedeon, Information retrieval by fuzzy relations and hierarchical co-occurrence, Part II, TR97-03, Department of Information Engineering, School of Computer Science and Engineering, University of New South Wales, Sydney, 1997.

[7] K. Chakrabarty, L.T. Kóczy, T.D. Gedeon, Analysis of fuzzy relational charts in information retrieval, IETR99-01, School of Computer Science and Engineering, University of New South Wales, Sydney, 1999.

[8] L.T. Kóczy, T.D. Gedeon, J.A. Kóczy, The construction of fuzzy relational maps in information retrieval, IETR98-01, Information Engineering, School of Computer Science and Engineering, University of New South Wales, Sydney, 1998.

[9] S. Miyamoto, Fuzzy Sets in Information Retrieval and Cluster Analysis, Kluwer, Dordrecht, 1990, 259p.

[10] P. Wallis, J.A. Thom, Relevance judgments for assessing recall, Inform. Process. Manage. 32 (1996) 273–286.