Hierarchical co-occurence relations

T.D. Gedeon [1] and L.T. Kóczy [1,2]
E-mail: {tom | koczy}@cse.unsw.edu.au

[1] Department of Information Engineering
School of Computer Science & Engineering
The University of New South Wales
Sydney NSW 2052 AUSTRALIA

[2] Dept. of Telecommunication & Telematics
Technical University of Budapest
Budapest H-1521 HUNGARY

ABSTRACT

We introduce a method using fuzzy similarity (equivalence) and tolerance (compatibility) relations, that allows the "concentric" extension of searches based on the hierarchical co-occurrence of words and phrases. This is to solve the problem of automatic indexing and retrieval of documents where user queries may not include any words occuring in the documents that should be retrieved. Various methods will be proposed and illustrated, with the intention of real application in legal document collections.

KEYWORDS: Fuzzy relations, Information Retrieval, Search Extension.

1. Introduction

Almost every document has some hierarchical structure with respect to the importance of the words or concepts occurring in it. It can be assumed that every document has a *title* which most likely contains relevant information concerning the contents. Most documents also have *sub-titles*, and some have a collection of *keywords* at the beginning of the text (as above in this paper). A number of approaches useful for automatic indexing of the context can be found in [1] and [2]. For example, the frequency-keyword approach, where all of the informative words in the text are keywords. In this paper we will restrict the usage of the term 'keyword' to words occurring on higher logical hierarchical levels. Hence to avoid confusion we will use the expression frequency-word approach. Also, we will use the word 'document' to also mean 'document segment' or 'text unit'.

The basic idea of automatic indexing based on co-occurrence is that words or phrases occurring frequently together in the same document or paragraph are connected in their meaning in some way. Certainly this will not mean that such words are necessarily synonyms or have related meanings, as antonyms may occur together just as often as do synonyms, not considering more sophisticated semantic connections.

The simplest idea is to check words using the frequency-word approach, and instead of linking documents with words, establishing a matrix or co-occurrence graph indicating the mutual co-occurence of pairs of words and phrases. A finer model will be introduced where the degree of co-occurrence is described by a membership degree in the sense of fuzzy logic.

A more sophisticated approach is the hierarchical approach. In this, the presumed semantic structure of the documents is taken into consideration by assuming (reasonably) that the title is descriptive of the contents of the paper. Thus, the words occurring in the title will be very important for the whole of the contents of the document, except obviously unimportant words like articles, or connectives.

Similarly, the sub-title of each section, sub-section, etc. of the document is assumed to be descriptive of the contents of the relevant sub-unit. In this sense, there is a hierarchical semantic structure in the document that contains at least two levels (1: title and eventual keywords, 2: text), but possibly more than two (e.g. 1:

0-7803-4778-1/98 $10.00 © 1998 IEEE

title and keywords, 2: sub-titles, 3: texts) that can be represented by a tree graph as in Figure 1. In the case of sub-sub-titles, etc., the number of levels increases in a similar way.)

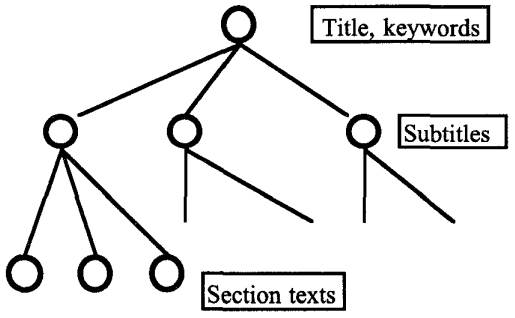

Figure 2. Document hierarchical structure.

If more than two (keyword and general word) levels are considered in the model, it will be necessary to introduce additional terminology: "keywords" for the words occurring in the title and the "Keywords" section (and maybe in the introductory and conclusion part of the whole document), "sub-keywords" for the terms occurring in the sub-titles (and corresponding introductions and conclusions), etc., and "words" for the lowest level comprising the contents of the whole document. Let us denote the set of keywords for a given collection of documents $D=\{D_1, D_2, ..., D_n\}$ by $K(D)$, and if there is a further hierarchy of the keyword levels, by $K_1(D)$, $K_2(D)$, etc., and the set of all significant words by W. Then it is advisable to define these sets so that $K_1(D) \subset K_2(D) \subset ... \subset K_m(D) \subset W$ where m denotes the number of hierarchical levels taken into consideration ($m \geq 1$).

The main idea is the following. If a certain word or phrase frequently occurs together with another in documents, the two may have some connected meaning or significance. If a word or phrase $\{w_i\}$ occurs frequently in a document, and the keywords $\{W_j\}$ (in the title, etc.) are certain other words, the document content words would belong to the class of related concepts of the keywords. The more frequent the co-occurrence the more it is likely that any user querying for any $\{W_j\}$ will be interested in documents containing $\{w_i\}$ in the text - even if the queried word does not appear in the title of these latter documents, and maybe not even in the text. Noting our concept of "hierarchical co-occurrence," we observe that it is likely that $\{W_j\} \subset \{w_i\}$, however, even $\{W_j\} \cap \{w_i\} = \varnothing$ cannot be excluded!

As an example let us take somebody who is interested in articles on Soft Computing or Computational Intelligence. In many overview articles on these subjects, the term Fuzzy Logic will occur frequently. However, it is very likely that in an article on Fuzzy Logic none of the terms Soft Computing or Computational Intelligence will occur. In this case it is quite clear that the connection between SC and FL is hierarchical in the meaning, and the structure of many documents will follow this, as shown in Fig. 2.

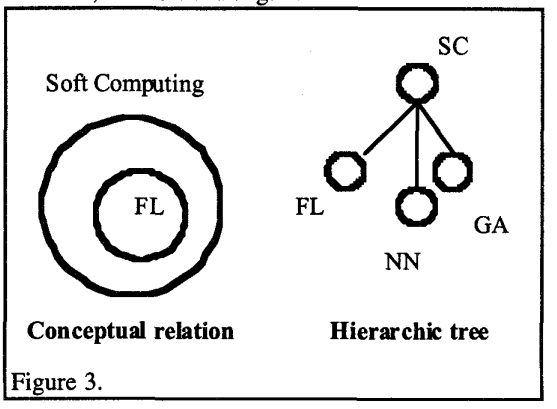

Figure 3.

The left side of the diagram expresses that Fuzzy Logic is a special branch of Soft Computing, and so, it is a subset of the topic marked by the keyword SC. The right hand side shows that articles on SC include those related to Fuzzy Logic, Neural Networks, Genetic Algorithms. In the next section we introduce a model that is suitable for finding documents not containing the words "Soft Computing" but dealing e.g. with Fuzzy Logic, by querying for "Soft Computing", and not asking for "Fuzzy Logic" at all.

2. Fuzzy relations

In this section we provide a short overview of fuzzy relations, particularly on a few important types of fuzzy and crisp relations that we will refer to in later sections. Further details on fuzzy relations can be found in [3].

A fuzzy set A is always defined in terms of a universe of discourse $X=\{x\}$ and a mapping μ_A from this set to the unit interval $[0,1]$: $\mu_A: X \to [0,1]$, where $\mu_A(x)$ is called the membership function of the fuzzy set A, and its

concrete values for any $x=x_0$ are the membership grades of x_0 in A. A fuzzy relation is a fuzzy set of the Cartesian product of two or more sets as the universe, so e.g. a binary fuzzy relation R is defined by the mapping $\mu_R: X \times Y \to [0,1]$ where $X=\{x\}$, $Y=\{y\}$ and consequently $X \times Y = \{(x,y)\}$. The special case when $Y=X$ is the binary relation over the Cartesian square of a given universe.

Binary fuzzy relations of $X \times X$ are categorized according to their properties in a manner similar to ordinary (crisp) relations. The crisp equivalence relations ($\equiv$) as defined fulfil three properties: reflexivity ($x \equiv x$ is always true), symmetry ($x \equiv y \Rightarrow y \equiv x$), and transitivity ($x \equiv y \wedge y \equiv z \Rightarrow x \equiv z$). The fuzzy analog of equivalence is called the *similarity relation* ($\cong$), and essentially the same three properties hold: reflexivity ($\mu_{\cong}(x,x)=1$), symmetry $\mu_{\cong}(x,y)=\mu_{\cong}(y,x)$), though transitivity has to be formulated in a somewhat different manner ($\mu_{\cong}(x,z) \geq \min\{\mu_{\cong}(x,y), \mu_{\cong}(y,z)\}$).

Compatibility relations are reflexive and symmetric, but not necessarily transitive, so they form a wider class than equivalence. The fuzzy analog is called the *tolerance relation* ($\approx$), and it has the first two properties as above: reflexivity ($\mu_{\approx}(x,x)=1$) and symmetry ($\mu_{\approx}(x,y)=\mu_{\approx}(y,x)$).

A convenient way to represent binary fuzzy relations of finite element universes is the use of matrices, where columns and rows correspond to the elements of the component universes X and Y and elements of the matrix are the membership degrees themselves, and can be visualised by bipartite graphs.

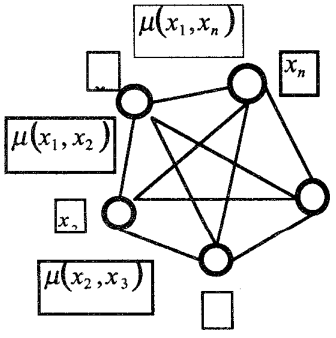

Figure 4.

Similarly, relations on $X \times X$ can be described with quadratic matrices, where for example, the similarity and tolerance relations have only 1s in the diagonals, and are symmetric. The graphical visualisation of such a matrix is shown in Fig. 3.

Selecting an arbitrary $\alpha \in [0,1]$ in such a fuzzy graph, the α-cut of the graph contains only those edges where the membership degree is at least α. If X_i is a node of the graph G representing a similarity relation, the set of all nodes $E(X_i)=\{X_j \in G \mid \mu(X_i,X_j) \geq \alpha\}$ represents the *equivalence* (similarity) *class* of X_i. From the properties of the similarity relation it is clear that $X_j X_k \in E(X_i) \Rightarrow \mu(X_j,X_k) \geq \alpha$ and also that $X_i \in E(X_i)$. Consequently, similarity relations generate α-partitions of the graph.

Tolerance relations behave differently, as tolerance is not transitive. While every node is necessarily an element of its own *tolerance cluster*: $X_i \in T(X_i)$, and other nodes are not necessarily connected by edges to each other with at least the same degree of membership as the defining node is to other nodes in the class. The α-cuts of tolerance classes of the nodes will usually not be complete graphs themselves. On the other hand, if the maximal sub-graph $C_\alpha(X_i)$ of $T(X_i)$ containing X_i itself is selected, where every node has at least α membership degree (α-clique), the set of maximal sub-graphs will form a cover of G, so that $\bigcup_i C_\alpha(X_i)=G$, usually $i \neq j \Rightarrow C_\alpha(X_i) \cap C_\alpha(X_j) \neq \varnothing$.

The graphs will thus not usually be empty, as some nodes of G belong to two or more *compatibility classes* simultaneously. An example is shown in Table 1 and Fig. 4.

	X_1	X_2	X_3	X_4	X_5	X_6
X_1	*1.0*	**0.7**	0.2	0.5	0.3	**0.8**
X_2	**0.7**	*1.0*	0.0	0.6	0.1	**0.9**
X_3	0.2	0.0	*1.0*	**0.7**	0.2	**0.7**
X_4	0.5	0.6	**0.7**	*1.0*	**0.8**	**0.8**
X_5	0.3	0.1	0.2	**0.8**	*1.0*	**0.9**
X_6	**0.8**	**0.9**	**0.7**	**0.8**	**0.9**	*1.0*

Table 1. Membership values $\mu(X_i,X_j)$

The relation represented by G is not a similarity relation as it is not transitive. Let us take for example

$\{X_3, X_4, X_5\}$. here $\mu(X_3, X_5) \geq \min\{\mu(X_3, X_4), \mu(X_4, X_5)\}$, that is, $\min\{\mu(X_3, X_4), \mu(X_4, X_5)\} = \min\{0.7, 0.8\}$ but $\mu(X_3, X_5) = 0.2$ which is less than 0.7, contradicting the transitive property of similiarity relations. On the other hand, all of the elements in the diagonal of the matrix are all 1-s, hence the relation is reflexive, and the matrix is symmetrical (the relation is symmetrical itself), consequently G represents a tolerance relation.

Let us choose $\alpha = 0.7$ and take the α-cut of G. The edges are indicated in Table 1 by bold numbers and italics (for the diagonal). In Fig. 4 all edges above the boundary are indicated with their respective degrees of membership, while the remaining edges are shown without degrees.

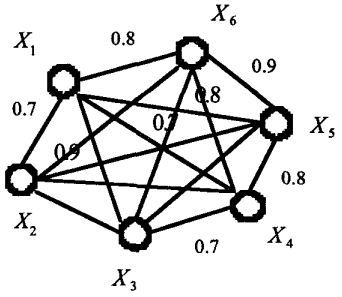

Figure 5.

Let us now construct the compatibility classes of the relation. (Note that searching for compatibility classes is an NP-complete task that needs a very long time for larger graphs [4]. There exist some faster approximate algorithms, however here we just assume that compatibility classes have been found already. This can be done since establishing the compatibility classes has

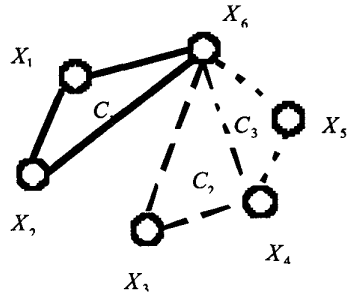

Figure 6. Set of compatibility classes

to be done once only, before the information retrieval service is started, in order to have a "logical map" of the knowledge in the data base in question.) The maximal compatibility classes in G_α ($\alpha = 0.7$) are $C_\alpha = \{C_1 = \{X_1, X_2, X_6\}, C_1 = \{X_3, X_4, X_6\}, C_1 = \{X_4, X_5, X_6\}\}$.

3. Fuzzy relations by co-occurence and importance measures

In this section we introduce a way of establishing complex relations based on the absolute and relative simple and weighted word counts in documents, and parts of documents.

The basic hypothesis is that the frequency of occurrence of significant words in a certain document is connected with the importance of that word in the document. Another additional assumption will be that pairs of words occurring frequently in the same document or the same part of a document are connected in their meaning (they might be synonymous, antonymous, or otherwise related).

In [1], [2] attempts have been made to find ways to index documents automatically, based on word frequencies is the main. In [5] the concept of *fuzzy importance degree* (also referred to as "measure") was introduced. If the [0,1]-normalized frequency of word w_i in the title/keyword section of document D_j is denoted by T_{ij} (keyword frequency, or title-keyword frequency), the normalized frequency of the same in the introduction/conclusion parts of the document is L_{ij} (location-keyword frequency), and the frequency in connection with cue words is C_{ij}, finally, if these three factors are weighted by $\lambda_1, \lambda_2, \lambda_3$ (where $\lambda_1 + \lambda_2 + \lambda_3 = 1$), the normalized fuzzy importance degree is calculated by the convex combination of the three frequencies: $F_{ij} = \lambda_1 T_{ij} + \lambda_2 C_{ij} + \lambda_3 L_{ij} = 1$.

Obviously, F_{ij} is a fuzzy membership degree that expresses the connection of w_i and D_j ($\mu(w, D_j)$). If the hierarchical structure of the document is taken into consideration as illustrated in Fig. 1, fuzzy importance degrees of level one (being F_{ij} itself), and levels two and so on, can be introduced (e.g. ${}^k F_{ij}^2 = \lambda_1 {}^k T_{ij}^2 + \lambda_2 {}^k C_{ij}^2 + \lambda_3 {}^k L_{ij}^2$, where the right superscripts indicate that level 2 titles such as sub-titles, and level 2 introductions and conclusions, and cue words located in some significant parts of the sub-sections were calculated; and the left superscripts refers to the index of the sub-document, i.e. meaning "part k" in this case.)

Another way of expressing the importance of a word in the document is just calculating its normalized frequency in the whole text ($K_{ij}=v(w_i,D_j)$) which we will call the fuzzy occurrence degree. As a matter of course, the frequency within any sub-section, sub-sub-section, etc. can be calculated, and so the frequencies ${}^kK_{ij}^2$, etc. can be determined. From now on it will be assumed that both fuzzy importance degrees: the normalized keyword frequencies F_{ij}; and the normalized word frequencies of (overall) occurrence K_{ij} obtained by automatic analysis of the relevant document and its sub-sections are available.

If the importance degree of each significant word in each document in a full or sample collection is available, the *fuzzy co-occurrence degrees* can be calculated. By co-occurrence the similarity or logical equivalence of the importance degrees or (normalized) relative frequencies will be understood. Fuzzy logical equivalence can be defined in various ways (all of these being extensions of the Boolean logical equivalence operation $A \equiv B \triangleq (A \wedge B) \vee (\neg A \wedge \neg B)$). Here, two straightforward definitions of fuzzy equivalence will be used. The first is based on the Zadeh-style fuzzy operators $\mu_{\neg A}(x)=1-\mu_A(x)$, $\mu_{A\wedge B}(x)=\min\{\mu_A(x),\mu_B(x)\}$ and $\mu_{A\vee B}(x)=\max\{\mu_A(x),\mu_B(x)\}$ where $\bar{\neg}$, $\bar{\wedge}$ and $\bar{\vee}$ are for fuzzy negation, conjunction and disjunction respectively.

The fuzzy equivalence has the form:
$\mu_{A\sim B}(x)=\max\{\min\{\mu_A(x),\mu_B(x)\},\min\{1-\mu_A(x),1-\mu_B(x)\}\}$

The second definition for fuzzy equivalence is based on the algebraic fuzzy operations: $\mu_{A\wedge B}(x)=\mu_A(x)\mu_B(x)$, and $\mu_{A\vee B}(x)=\mu_A(x)+\mu_B(x)-\mu_A(x)\mu_B(x)$ (negation being identical with the above).

(Henceforth, complicated notation will be simplified such that the fuzzy logical operation will not be differentiated by the wave above the operator, as it is usually clear from the context if it is a fuzzy operation, further, membership functions will be usually denoted just by the symbol of the fuzzy set or statement, so e.g. the algebraic fuzzy disjunction being written simply as $A \vee B=A+B-AB$. For more details on fuzzy operators and operations see [3].)

When introducing *hierarchical co-occurrence* the following is meant: first the hierarchical structure and the document indexing/analysis structure in that particular model are determined. (Determine the number of levels in the document. Determine the weights λ_i. For each hierarchical level and within it, for each section, sub-section, etc. determine the text unit in question, and if necessary, its special location parts, like the introduction, etc.) Then for each text unit determine the *fuzzy importance degree* and the *fuzzy occurrence degree* as well. The *fuzzy equivalence* of these two degrees will result in the *hierarchical fuzzy co-occurrence degree* of the given document, section, etc. Its formal definition is as follows:

$$H_{i_1 i_2 j} \triangleq F_{i_1 j} \equiv K_{i_2 j} \quad \text{for the main text,}$$

and $${}^kH^l_{i_1 i_2 j} \triangleq {}^kF^l_{i_1 i_2 j} \equiv {}^kK^l_{i_1 i_2 j}$$

for sub-section number k in level l, all for keyword W_{i_1} and word w_{i_2} in document D_j.

If a sample collection of documents is fixed for training the information retrieval system, the average degrees of hierarchical fuzzy co-occurrence can be calculated by

$$H_{ij} \triangleq \frac{\sum_{k=1}^{n} H_{ijk}}{n}$$

where n is the number of documents in the sample collection, and i stands now for the subscript of the keyword, j for that of the general text word in question. The average non-hierarchical co-occurrence degree can be calculated similarly.

4. Hierarchical co-occurence queries

By using the fuzzy importance and co-occurrence degrees, and the fuzzy relation classes discussed in the previous sections, we establish a complex hierarchical relational map of a sample document collection. After deciding the levels and weighting factors and doing the keyword and word counts in the collection, these frequencies are normalized for the unit interval [0,1]. Then the normalized indices can be interpreted as fuzzy membership degrees and used directly in the formulae given earlier. As a result, the following relations and corresponding graphs will be established:

- Keyword co-occurences relation/graph G_w (established from the normalized co-occurrences N_{ij}^w);
- Word co-occ. G_w (from normalized co-occ. N_{ij});

- Fuzzy importance degree (keyword-document occ.) G_{WD} (from the fuzzy importance degrees F_{ij});
- Word-document occ. G_{wD} (from normalized occ. K_{ij});
- Hierarchical co-occ. G_{Ww} (from hier. Co-occ. H_{ij}); and
- Further hierarchical co-occ. relations – multilevel models.

Search by keyword and hierarchical co-occurence:

Determine the set of words that match the keyword. All documents that match any of the matching words will be retrieved: $\Delta = G_{wD\tau_2}\left(G_{Ww\tau_1}(W_i)\right)$. The values of τ_1 and τ_2 could be different thresholds determining the level of matching. The method is illustrated in Fig. 6.

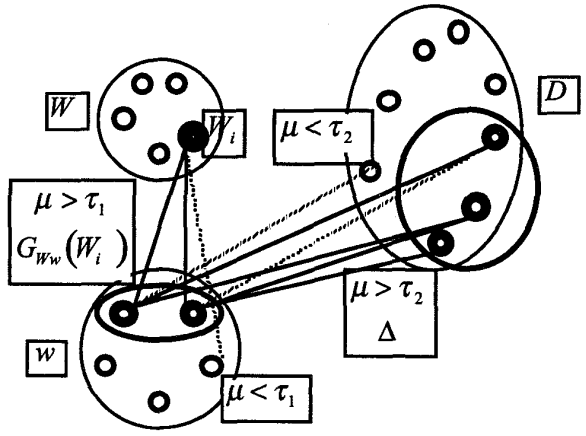

Figure 7.

In the figure the query keyword is indicated by a dark node. All matching words/documents (thick line nodes in *w/D*) are connected to it by solid lines, while a few words/documents having lesser membership value than the thresholds in the relation are shown by dashed connections. This latter in *D* is not considered to be matching and is left out of the class of retrieved documents Δ (the thick line nodes in *D*).

Search by keyword compatibility/equivalence relations and hierarchical co-occurrence

Determine the compatibility class in *W* and all matching words in *w*. All documents matching the image of the compatibility class of the original keyword will be retrieved: $\Delta = G_{wD\tau_2}\left(G_{Ww\tau_1}\left(C_{W_\sigma}(W_i)\right)\right)$.

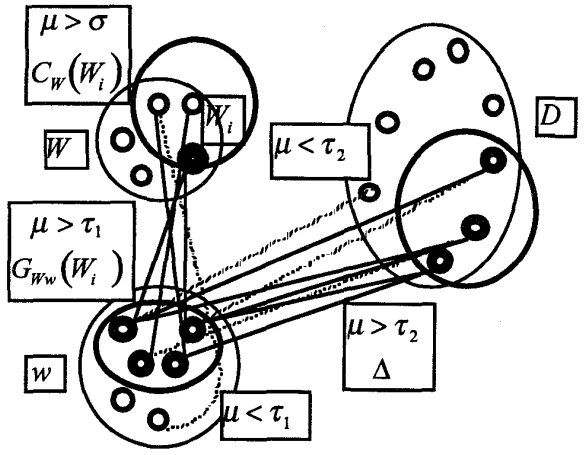

Figure 8.

The diagram notation is the same as before. Clearly a larger number of documents can be retrieved, which will increase the *recall* and reduce the *precision*. The setting of the respective τ values will be significant in determining the appropriate tradeoff.

References

[1] Gedeon, TD and Ngu, AHH: Index Generation is better than Extraction," *Proceedings International Conference on Non-Linear Theory*, pp. 771-774, Hawaii, 1993.

[2] Bustos, RA and Gedeon, TD: Learning synonyms and related concepts in document collections, in: Alspector, J, Goodman, R and Brown, TX (eds.): *Applications of Neural Networks to Telecommunications 2*, Lawrence Erlbaum, 1995, pp. 202-209.

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

[4] Garey, MR and Johnson, DS: *Computers and Intractability*. A Guide to the Theory of NP-Completeness, WH Freeman Co., San Francisco, 1979.

[5] Gedeon, TD, Singh, S, Kóczy, LT and Bustos, RA: Fuzzy relevance values for information retrieval and hypertext link generation, *Proceedings of EUFIT '96*, Aachen, 1996, pp. 826-830.