Fuzzy Pseudo-thesaurus Based Clustering of a Folkloristic Corpus

S. Szaszkó*, L. T. Kóczy*,†, T. D. Gedeon‡
*Budapest University of Technology and Economic, Hungary
†Széchenyi István University, Győr, Hungary
‡The Australian national University, Canberra, Ausztrália
{szaszko,koczy}@tmit.bme.hu, tom.gedeon@anu.edu.au

Abstract— **Automatic thesaurus extraction is essential for modern information retrieval. We develop a method for fuzzy pseudo-thesaurus based on word pair co-occurrence in documents. In this study it is presented, that considering the Word Frequency Degree counted on the whole corpus makes the obtained pseudo-thesaurus usable. Such parameters were found with which most of the obtained pairs of words were validated to be related by human expert. Among the extracted pairs and groups of words the relationship is often looser than synonymy, but they identify the frequently repeated topics of the corpus. We suggest the use of groups of closely related words for the definition of different topics and based on this clustering of the documents were performed.** [1]

Keywords— **Fuzzy information retrieval, fuzzy thesaurus**

I. 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.

Natural languages use many similar terms for the same or similar concepts. In order to most of the documents belonging to the same topic, special dictionaries have to be set up. A thesaurus is a collection of terms (words) which describe the same concept. With the use of the thesaurus we can discover connections among documents, which do not necessarily contain the same words, or retrieve relevant documents which do not necessarily include any of the query words.

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 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.

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 their 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) explicitly 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.

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 [1][8][9]. This means that the occurrence frequencies of certain words in the title and sub-titles, the abstract and introduction or conclusion parts most of the 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 frequencies of occurrence into possibility measures (fuzzy membership degrees) will be presented.

There are several example for supporting IR by thesauri. [7] examine word cooccurring in a 40 word wide window. [5] and [6] construct thesauri by transposing the term-by-document marix. Applying fuzzy logic to automated information retrieval is not new. Some of the most important advances in this field are summarised in [4]. In several points of this paper, reference will be made to concepts introduced in this work.

II. KEYWORD OCCURRENCE FREQUENCIES AND POSSIBILITY DEGREES

If analyzing a collection of documents related to a certain topic (e.g. folkloristic beliefs) 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. In the context of folkloristic beliefs such a relative stop words are hard to identify. Because of this the set of relative stop words will become empty. These texts are usually rather short and they contain only relevant, often enigmatically short expressions – except the proper stop words.

[1] Partly supported by National Scientific Research Fund, OTKA T048832 and Széchenyi University main direction research direction grant 2005.

By the omission of stop words (and relative stop words, if relevant) the set of significant words is obtained which m be used for a further analysis. 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 Figure 1.

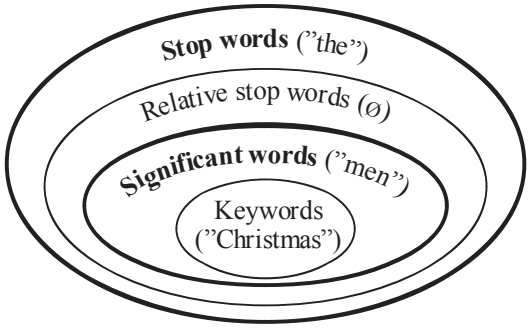

Fig. 1: Categories of words in documents

In the Fig. 1 the four categories of words can be seen: stop word like "the", relative stop words an empty set, and "men" as a general example for a significant word and finally, "Christmas" for a keyword, as beliefs connected to Christmas time form a large and important subgroup of our collection.

It is a crucially important issue how occurrence frequencies can be and are transformed into fuzzy membership degrees, which may be interpreted as important and relevant measures, satisfying the properties of possibilistic 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 absolute importance. 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, generally somewhere between 2 and 20 word counts. It does not matter much whether a word occurs in a document 10 or 12 times, it is likely that this document will be rather important from the point of view of 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 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 like beliefs and superstitions these numbers might vary. Especially it can never be expected that words occur more than a few (two, tree or four) times.

The mapping from occurrence frequencies or counts to possibilistic membership degrees is thus generally 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 conditions which are given in [3]. In practice σ is not necessarily continuously differentiable, but its characteristics should be nevertheless "S-shaped".

III. Fuzzy Pre-processing of a Folkloristic Corpus

Hungary has a very rich folkloristic tradition and especially in the last century a successful work has been done to research and preserve this heritage. For instance, nowadays many young people learn the traditional dances of villages, in many cities there are parties with folk music and dance. The 3rd CIOFF World Folklorida, the "Olympia of Folk Art" took place in Hungary very recently, an obvious sign of international appreciation of this work.

Quite a few Hungarian cultural anthropologists were collecting beliefs and superstitions mainly in the 20th century. There are about 27 000 documents on paper in the National Museum. Unfortunately the classical techniques of anthropology are not able to process this amount of data and usually studies analyze only 6 to 10 documents. Of the above collection, there exists a digitized database of 2704 Hungarian belief texts, suitable for computerized analysis, which has been processed by principal component analysis[10].

In order to distinguish the different dialects one word is spelled in different ways, sometimes even special characters are used to note the pronunciation down. The other problem is the use of old style language, a big part of the vocabulary of the corpus is not in use anymore. So first of all different appearances of the same words had to be collected into pre-process dictionary. In this way we also solved the problem of Hungarian language being an agglutinative language, it puts many different tags at the end of the words.

After all 1837 significant words remain in the pre-process dictionary. Special attention was paid for negation. Hungarian language puts *nem* (not) word before the verb for its negation, so if the software found the *nem* word in the text, it considered it together with the next word and it searched that whole string in the pre-process dictionary.

At this phase of the research we did not try to solve the problem of words with more than one meaning like *fog*, which can mean tooth and also to catch. The context has to be analyzed to select the actual meaning which is quite a complex task.

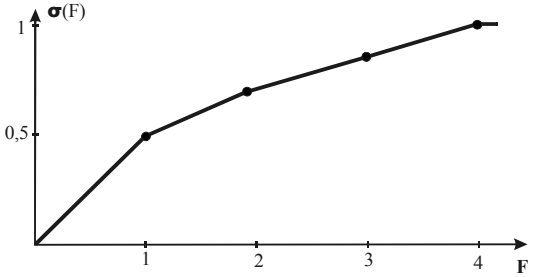

Fig. 2 Sigmoid curves used for the sort documents

The sigmoid function σ(F) used is shown in Figure 2. The documents are usually 2 to 5 lines longs, just very few of them exceed the size of half a page so even a single occurrence is quite significant. Thus in case of a single occurrence the membership degree is already 0.5. Less than 0.1% of the significant words appear more than 4 times, so in case a word occurs for times in a document the membership degree is 1.

By the end of the pre-processing matrix **M** was obtained whose size is 2704x1837, the words (W) and the documents (D) are listed at the edges. The fuzzy degrees stored in **M** show how much a given word is related to a document.

A. Word Frequency Degree

Let us define a degree in the following way:

$$I_w = \sum_{i=1}^{N} \sigma_{d_i W} \tag{1}$$

The Word Frequency Degree shows a given word how significant is in collection of the documents, in the whole corpus.

IV. Establishing Fuzzy Pseudo-thesauruses by Co-occurrence

In order to define synonymy take the set of (all) concepts [4]. The concepts may be abstract ones, which cannot be found in the real world. Now take the set of words. A word belongs to a concept with a fuzzy membership degree. One word may belong to several concepts and several words may belong to one concept.

Two words are synonyms if they belong to the very same concepts with the same degrees. The fuzzy thesaurus lists not just the synonyms, but also the degree of synonymy. For example if word A and word B belong to some concepts which are common, but they also belong to some different concept they are synonyms with a degree which is less than 1. If A and B are not related to any common concept they are synonyms with the degree of zero.

If we want to obtain this thesaurus there are two main questions:

1st: What to use instead of the set of concepts?

2nd How to choose degree of the synonymy when it is not 0 or 1?

There are not many possible answers for the first question if we want an automatic method to generate the thesaurus. We substitute the set of documents for the set of concepts. Because this is not a very good substitution the thesaurus obtained is called pseudo-thesaurus.

There are plenty of possibilities to define the degree of synonymy between pairs of words. In the next two subsections we will see that the obvious choices do not give good result, so in 4.3 a special weight factor is introduced.

Establishing the fuzzy pseudo-thesaurus:

Step 1: *Co-occurrence degree calculation*

$$\mu'_{ij}(D) = \min(\sigma_{W_i D}, \sigma_{W_j D})$$

$$\mu_{ij} = \frac{1}{C}\frac{1}{s}\sum_{z=1}^{N} \mu'_{ij}(D_z) \tag{2}$$

where C is a constant which keeps μ_{ij} in the range of [0,1], C is independent while s a weight may be dependent from i and j. The first idea how to choose C can be N, the number of the documents, but in this way the values of μ_{ij} are very small. We get more reasonable values if

$$C = \max_{i,j}\left(\frac{1}{s}\sum_{z=1}^{N} \mu'_{ij}(D_z)\right). \tag{3}$$

Step 2: *Suitable α-cut*

If the number of the significant words is M, then the co-occurrence degrees (μ_{ij}) form a matrix size of MxM, let call it **W**. The words are listed on both sides from 1 to M. Since $\mu_{ij} = \mu_{ji}$ this matrix **W** can be represented by an undirected graph.

Choose a suitable α for which the *α-cut* leaves about 30 to 40 nodes in the graph. This is a representation of a pseudo-thesaurus.

Step 3: *Searching maximal cliques*

An edge means that the connected two nodes representing words are "synonyms" in this broader term ("related" in the meaning). If a set of nodes are fully connected, than they are called a clique, and they are supposed to be related to the same broad concept.

Step 4: *Fuzzy clique*

Many times among the found maximal cliques there are a few which have many common nodes. Since we chose α arbitrarily it is reasonable to check if these close cliques describe the same broad concept and they can be aggregated. We take the cliques which have just one different node and investigate these different nodes. If there is an edge between them on level α'=0.7α cut we aggregate the cliques.

A. Weight s=*1*

Here the measure of co-occurrence is simply proportional with the sum of the co-occurrence. Column I_w shows in Table 1 that just very frequent words remained in the α-cut. The most frequent words of the corpus (number 18, "go" and 26, "do") have the most edges.

TABLE 1
LIST OF WORDS IN CASE WEIGHT *S*=1

NR	HUN.	ENGLIH	I_w
1	ad	giv(e)	74.7
2	asszony	woman, wif(e)	92.1
3	este	evening	103.3
4	fent	Above	69.1
5	fog	tooth/catch	123.5
6	férj	husband	66.1
7	gyermek	Child	160.8
8	György	George	31.8
9	haza	Home	42.8
10	3	3	83.7
11	ház	House	173.6
12	karácsony	Christmas	90.4
13	kicsi	small, little	110.9
14	Legény	young man, fellow, lad	41.7
15	Luca	Lucia (St.)	52.8
16	Lány	girl, lass	142.4
17	Meghal	die	90.8
18	Megy	go, walk	218.4
19	Mise	(holy) mass	27.6
20	Mond	say	108.3
21	Nap	day/Sun	172.3

22	Ront	spoil, bewitch	48.7
23	Sok	much, many	86.0
24	Szent	Saint (St.)	37.8
25	Tehén	cow	97.1
26	Tej	milk	61.3
27	Tesz	Do, put	177.8
28	Tojik	lay eggs	32.3
29	Tyúk	hen	77.4
30	Víz	water	98.9
31	Éjfél	midnight	40.5
32	Éjjel	night	87.2
33	Év	year	94.7

Even though not all, but many of the edges give really meaningful pairs of words. The maximal cliques show some logic behind, but usually some very frequent words appears as an odd-one-out.(see Figure 3)

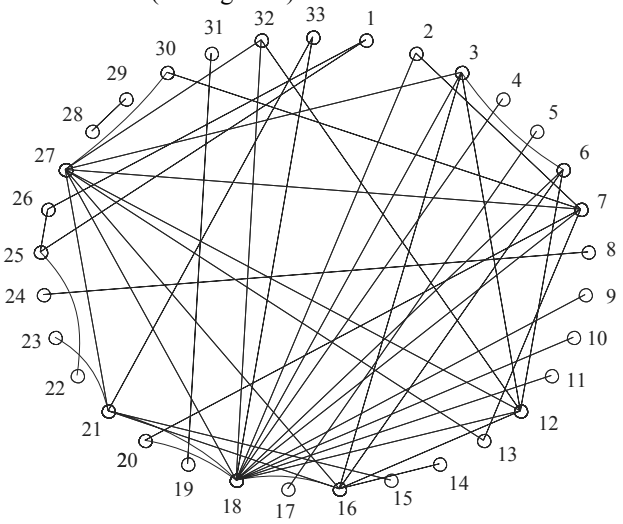

Fig. 3: Graph of a pseudo thesaurus, weight s=1

B. Weight s=$max(I_{WA}, I_{WB})$

To avoid the dominancy of frequent words lets divide the co-occurrence measure, μ_{ij} by word frequency degree (I_w), which is greater from word j and word j. In this case C=1 because I_w is never smaller than the sum of μ'_{ij}-s.

The highest not empty α-cut has 90 nodes. For all words in the α-cut I_w=0.5, which means that all appear just once in the hole corpus. The relations found here have no significance because of the low occurrence.

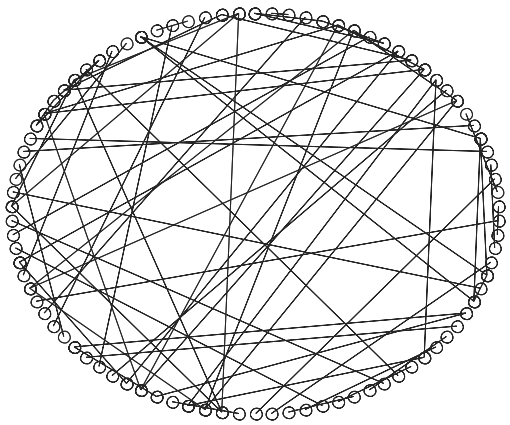

Fig. 4. Graph a of pseudo thesaurus, weight s= max(I_{WA} ,I_{WB})

C. Weight s=*1+ (max(IWA ,I_{WB}))/20*

By 4.1. and 4.2. we guess that weight s should be between 1 and max(I_{WA} ,I_{WB}). After several tests weight s=1+ (max(I_{WA} ,I_{WB}))/20 was proved to be the most efficient to identify concepts. As its can be seen in Table 2, these words represent a wider range of I_w values.

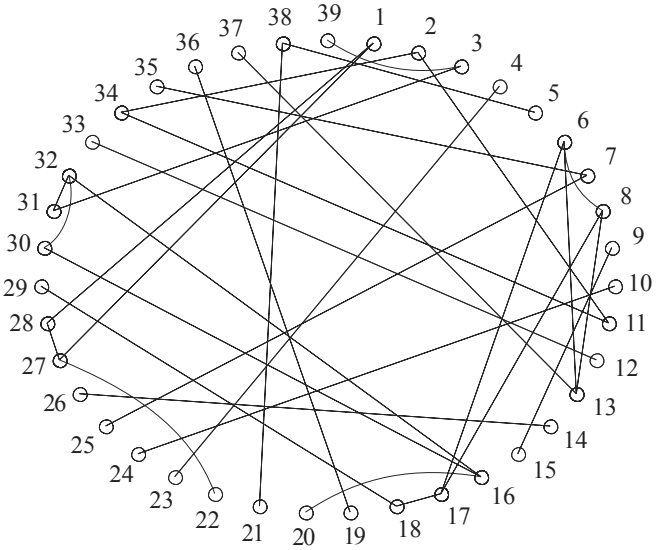

Fig. 5. Graph of a pseudo thesaurus, weight s=1+ (max(I_{WA} ,I_{WB}))/20

TABLE 2
LIST OF WORDS IN CASE WEIGHT S=1+ (MAX(I_{WA} ,I_{WB}))/20

NR	HUN.	ENGLISH	I_w
1	Ad	giv(e)	74.7
2	Bal	left	28.4
3	Csibe	chicken	35.2
4	Csörög	clatter	5.5
5	Cédula	peace of paper	9.5
6	Este	evening	103.3
7	Fecske	swallow	16.8
8	Férj	husband	66.1
9	gyermek	child	160.8
10	György	George	31.8
11	jobb	right	30.9
12	jön	come	91.1
13	Karácsony	Christmas	90.4
14	kenyér	bread	51.4
15	kicsi	small	110.9
16	Luca	Lucia (St.)	52.8
17	lány	girl, lass	142.4
18	megy	go, walk	218.4
19	mise	(holy) mass	27.6
20	nap	day, sun	172.3
21	Név	name	43.3
22	Ront	spoil, bewitch	48.7
23	szarka	magpie	6.7
24	szent	saint, St.	37.8
25	szeplő	freckle	10.3
26	Süt	bake	33.6
27	tehén	cow	97.1
28	Tej	milk	61.3
29	Tesz	do, put	177.8
30	Tojik	lay eggs	32.3
31	tojás	Egg	43.8
32	Tyúk	Hen	77.4
33	vendég	Guest	29.8

34	viszket	itch	25.5
35	Vér	blood	17.6
36	Éjfél	midnight	40.5
37	Éjjel	night	87.2
38	Ír	write	20.7
39	Ültet	plant, seat	21.3

The graph in Figure 5 is not dominated by any node, very frequent words like go, do, day (number 18, 29, 20) have not more than two edges. We managed to avoid that some very frequent words are connected to most of the other nodes like in Figure 3. It is logical that words with high are related to more concepts (this is why they are so frequent), but some of them should not be in most of the maximal cliques.

TABLE 3
MAXIMAL CLIQUES

giv(e) (1)	cow (27)	milk (28)
left (2)	right (11)	itch (34)
evening (6)	*husband (8)*	*Christmas (13)*
evening (6)	*husband (8)*	*girl, lass (17)*
Lucia (St.) (16)	lay eggs (30)	hen (32)

chicken (3)	egg (31)
chicken (3)	plant, seat (39)
clatter (4)	magpie (23)
peace of paper (5)	write (38)
swallow (7)	freckle (25)
swallow (7)	blood (35)
Child (9)	small (15)
George (10)	Saint, St. (24)
come (12)	guest (33)
charismas (13)	night (37)
bread (14)	bake (26)
Lucia (St.) (16)	day, sun (20)
girl, lass (17)	go, walk (18)
go, walk (18)	do, put (29)
(holy) mass (19)	mindnight (36)
name (21)	write (38)
Spoil, bewitch(22)	cow (27)
egg (31)	hen (32)

Table 3 lists the maximal cliques of the graph of Figure 5, these sets of word describe broad concepts which are important in this corpus. The 3rd an 4th line of the Table 3 differ just in the last word. Take a lover α-cut of Matrix **W**. We can find that already at α'=0.8α there is an edge between Christmas(13) and girl(17), so the two maximal cliques can be aggregated, they form the fuzzy maximal clique: evening(6), husband(8),Christmas(13) and girl(17). It is easy to imagine beliefs which are about how a girl can find a husband at Christmas Eve.

Two examples:

"Karácsony estélyén doboskát sütnek s a leány az elsôvel kiszalad és amely legénnyel találkozik legelôször az lesz a férje."

Christmas Eve "doboska" cakes are prepared and the
daughter of the house runs out with the first piece and the young man whom she meets first will become her husband.

"Karácsony estélyén a lánynak egy öl fát kell felvenni és ha a számuk páros, akkor férjhez meg, ha páratlan, akkor nem megy."

Christmas Eve the girls should pick up a bunch of wood, if the number of pieces is even she will be married, if it is odd, she will not be married (namely, next year).

V. CONCEPT BASED CLUSTERING

If we check the pairs of words in Fig. 5. or in the pseudo-thesaurus of Table 3 it can be seen that the relationships rarely indicate real synonymy, most of the cases paired words cannot be interchanged. The point of departure in establishing the thesaurus was to find the words belonging to the same concept. In the sets of words found the words are linked in meaning as it was shown in the end of chapter IV.

The sets of words of the pseudo-thesaurus describe concepts in a broader understanding. These sets can be understood as definition of various topics. After learning the main topics of the corpus we can search for the documents belonging to given topics.

This last step is actually a task of categorization because the categories are given by topics. But these topics are established by an automatic method unlike the case of classical categorization, where the categories are defined by human experts, so the whole process can rather be considered a kind of clustering since the documents are grouped without any a prior information. The number of clusters can be controlled by choosing the level of the α-cut during establishment of the thesaurus.

A. Clustering the Folkloristic Corpus

In chapter IV we used very high values of α in order to gain small pseudo-thesauri which can be analyzed and represented in this study. For a potential clustering bigger maximal cliques are needed so we used lower α-cut. After establishing the thesaurus with the weight parameter of chapter IV.*C* eventually 327 maximal cliques were found. Out of these 3 cliques contain 6 words, 7 cliques contain 5 words, 47 cliques contain 4 words, 91 cliques contain 3 words and 179 cliques contain 2 words.

Let us define the similarity measure between a cluster and a document:

$$S_{C,D} = \prod_{i \in A_c} \sigma_{Wi,D} \tag{4}$$

where A_c is the set of words belonging to clique C.

The similarity of all the documents to a particular cluster can be calculated by multiplying the correspondent columns of matrix **M**:

$$\overline{S}_C = \prod_{i \in A_c} \overline{M}_i \qquad \overline{\overline{M}} = \left[\overline{M}_1, \overline{M}_2, ... \overline{M}_i ... \overline{M}_N\right] \tag{5}$$

Similarity S_c indicates how much a document covers the topic of a cluster. By the above definition S=0 if not all the words of the maximal clique appear in the document. It is logical that the topic of the document can be somehow close to the topic of the cluster even if a word of the clique does not appear. Let us say that a word which is not contained by a document still characterizes for the document with very low possibilistic measure (σ). This is somehow true as documents

usually do not contain all the words which are characteristic for their topic.

Zero elements of matrix M are changed for this low value for the calculation of the similarity measure. In case of our corpus the lowest σ is 0.5 (in case of single occurrence). Let us choose one fifth value, 0.1 in case no occurrence.

By calculating (5) for all the maximal cliques matrix **S** is obtained. On the edges of **S** cliques and documents are listed. The final clusters can be obtained by a suitable α-cut of **S**.

Choosing α:

Les us take a cluster defined by two words. Certanly a document containing just a single piece of the defining words one time should not be in the cluster. For this α=0.5*0.1=0.05 and cut is strict.

Fig. 6 shows present the member of clusters with various cardinalities. Only four clusters remain empty, mainly those which were defined by 5-6 words. It is quite encouraging that the documents are distributed among categories quite equally. The average number of the document per cluster is 17 and only 39 topics have got more documents than 34.

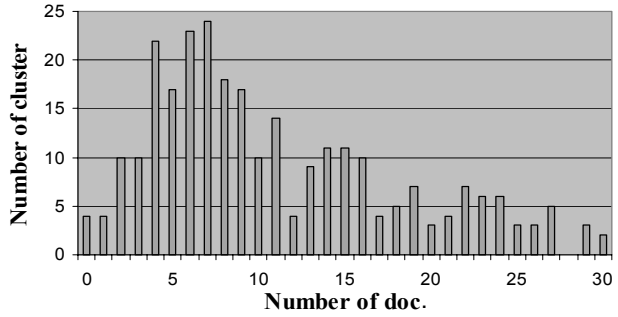

Fig. 6. Result of clustering

Just 59 documents are assigned to more than 10 topics, most of them belong to one or two groups. A weakness of the result is that more than 37 % of the documents are not assigned to topics. This might be because of the great number of the very sort documents. All those covered more than 30 topics are much longer than the average length of documents.

Example:

Take those documents containing the words of the following maximal clique***: piece of paper, write, name and dumpling***. All of the documents thus found describe beliefs in the following topic: If you put piece of papers into raw dumplings with names of different people you can find out with whom you are going to be married. The suitable day is usually Christmas' Eve and New Year's Eve, but the day of St. Lucia and Andrew are also mentioned. The number of the dumplings varies from 3 to 30, but always the one which comes onto top of the water is suppose to contain the significant name.

CONCLUSIONS AND FURTHER STUDY

Based on some further studies on textual document retrieval a method of automatic thesaurus generation has been proposed and an appropriate weight factor was introduced. The method was applied on a collection of Hungarian folkloristic beliefs. This way a pseudo-thesaurus was generated. The sets of words thus found in the thesaurus were meaningful for the contents of the beliefs analyzed, and helped to understand the course topics of beliefs.

New clustering method was introduced based on the topics identified by the pseudo-thesaurus. The documents inside different groups are really linked in content.

Methods should be tested with different corpuses in order to identify and resolve side effect, like there are too many unclustered documents, the importance of the small set of words is high.

ACKNOWLEDGEMENT

We express our thanks to S. Darányi and F. Kiss for providing access to the folkloristic corpus and the pre-process dictionary used in for this study.

REFERENCES

[1] 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.

[2] G. Klir, T. Folger, "Fuzzy Sets" Uncertainty and Information, Prentice-Hall, Englewood Cliofs, NJ, 1988

[3] L T. Kóczy, T. D. Gedeon and J. A. Kóczy, "Fuzzy tolerance relations and relational maps applied to information retrieval" *Fuzzy Sets and Systems 126* (2002) 49–61

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

[5] Y. Qiu, H. Frei, "Comcept Based Query Expansion" SIGIR conference, 1993

[6] G Salton, C. Buckley, "Improving Retrieval performance by Relevance Feedback" Jurnal of A. S. for Information Science, 1990, 288-297

[7] H. Schütze, J. O. Pedersen, "A Cooccurrence-based Thesaurus and Two Application to Information Retrieval" Information Retrieval and Management, 1997, Vol. 33, 307-318

[8] S. Szaszkó, L. T. Kóczy "Identifying Concept in Folcloristic Corpus by Fuzzy Pseudo-thesaurus" *Eurofuse* 2004, Warsaw, 522-532

[9] S. Szaszkó, L. T. Kóczy "What Lectures Note About, Identifying Concepts by Fuzzy Pseudo-thesaurus" *EESTEC-IEEE Conference*, 2004, Italy

[10] Voigt, V. - Preminger, M. * Ládi, L. * Darányi, S. "Automated motif identification in folklore text corpora. Folklore." Electronic Journal of Folklore Vol. 12., Tartu, 1999 126-141 pp. Also available at: http://haldjas.folklore.ee/folklore/vol12/motif.htm