Extracting topics in texts: Towards a fuzzy logic approach

Mohand Boughanem **Henri Prade** **Ourdia Bouidghaghen**
IRIT, CNRS - University of Toulouse
118 route de Narbonne, 31062 Toulouse Cedex 9, France
bougha@irit.fr prade@irit.fr bouidgha@irit.fr

Abstract

The paper presents a preliminary investigation of potential methods for extracting semantic views of text contents, which go beyond standard statistical indexation. The aim is to build kinds of fuzzily weighted structured images of semantic contents. A preliminary step consists in identifying the different types of relations (is-a, part-of, related-to, synonymy, domain, glossary relations) that exist between the words of a text, using some general ontology such as WordNet. Then taking advantage of these relations, different types of fuzzy clusters of words can be built. Moreover, apart from its frequency of occurrence, the importance of a word may be also evaluated through some estimate of its specificity. The size of the clusters, the frequency and the specificity of their words are indications that enable us to build a fuzzy set of sets of words that progressively "emerge" from a text, as being representative of its contents. The ideas advocated in the paper and their potential usefulness are illustrated on a running example. It is expected that obtaining a better representation of the semantic contents of texts may help to better retrieve the texts that are relevant with respect to a given query, and to give some indication of what the text is about to a potential reader.

Keywords: semantic contents, fuzzy relation, information retrieval.

1 Introduction

Texts are fuzzy in many respects, even when their authors have not intended to be intentionally vague and have rather tried to be accurate in their statements. This seems to be due to the very nature of natural languages, which echo the way humans perceive the world. Most of the fuzzy features of texts have been pointed out very early after the introduction of the notion of a fuzzy set [46]. Several have led to some noticeable developments.

The meaning of words is fuzzy as soon as the classes of objects, or elements to which they refer have no crisp boundaries, or the items they intend to denote remain ill identified. As everybody knows, fuzzy sets provide useful representations of categories associated with gradual properties such as 'large', 'tall', 'young' or 'cheap', possibly modulated by linguistic hedges [23]. Linguistic categories involving multiple features, usually denoted by substantives, are also often fuzzy, even it is more difficult to assess membership functions to them. For instance, the class of birds is fuzzy in the sense that some birds are more typical than others, since they possess more characteristic properties of a bird, or possess them to a greater extent. However, it is not easy to define a degree of "birdiness". The way categories or concepts are used and perceived by the mind is often pervaded with fuzziness [24]. The grammaticality of sentences [3], [14], the semiotics of words [42], of texts [43], involve fuzzy notions, as already pointed out about thirty years ago. See also Rieger [33]-[35] for the development of a fuzzy semantic view of texts based on the co-occurrence of words.

L. Magdalena, M. Ojeda-Aciego, J.L. Verdegay (eds): Proceedings of IPMU'08, pp. 1733–1740
Torremolinos (Málaga), June 22–27, 2008

The relevance of a text with respect to a group of keywords is also a fuzzy notion, as it is has been recognized early [11], [29], and then successfully developed, by also allowing for weighted queries [7]; see Kraft *et al.* [21] for an introductory survey.

Other uses of fuzzy logic have been more recently introduced in the processing of texts. Subasic and Huettner [40] look for subjective features of text content, such as emotions, in analyzing news reports and movie reviews, by performing a "fuzzy semantic typing" (they use a fuzzy thesaurus as a similarity relation on a set of affect categories, for expanding affect sets and obtaining a fuzzy affect representation). St-Jacques *et al.* [39] use fuzzy logic techniques for grouping similar features together and partitioning different ones in the lexical space of a dictionary.

The idea of categorical perception as a fuzzy clustering process, which underlies the previously cited work, will also pervade the approach proposed in the present paper, although in a different way. Existing approaches use statistical methods in order to characterize the content of a text [36]. However, it seems that a better view could be obtained by going beyond a simple counting, and taking into account clusters of words belonging to the same lexical field. Ontologies have become available resources for identifying relations between words in a text. Moreover, words in a cluster may be more or less specific, and thus contribute differently to the reader's perception of the topics of a text.

The paper is organized as follows. Section 2 provides a review on the use of ontologies in concept-based information retrieval, especially for query expansion. Section 3 presents an approach for building a representation of the contents of a text at different levels of details, by progressively identifying clusters of appropriate words. It is illustrated and discussed in Section 4 on a detailed example. The concluding remarks outline the use of such a granulation procedure [47] in query evaluation, and the possible use of small worlds hierarchies [12]-[13] of words in place of ontologies.

2 Information retrieval and ontologies

Most of classical information retrieval (IR) models are based on “bag of words” approaches expressing the fact that both documents and queries are represented using basic weighted keywords. The performances of such models suffer from the so-called *keyword barrier* [26] due to term ambiguity and vocabulary mismatch. Indeed in such approaches relevant documents are not retrieved if they do not share terms with the query. The main limitation of this oversimplified view of document contents is that it does not take into account the topical content of documents [2]. Various approaches have been proposed to go beyond the simplistic bag of words approaches. They attempt to identify and extract word sense or concepts occurring in the documents. Two main approaches have been undertaken in IR: semantic indexing and concept-based indexing.

Semantic indexing is basically based on word sense disambiguation (*WSD*). It consists in associating the extracted words of document or query, to words of their own context [22], [37], [45]. Kovertz and Croft [22] have studied the relationships between sense mismatches amongst query terms and their occurrences in the collection. They conclude that co-location and co-occurrence between query terms provide some elements of disambiguation. Sanderson [37] evaluates the effect of term ambiguity by introducing ambiguous terms in collections of documents. He has showed that queries with few terms (one to two terms) are more affected by ambiguity than longer ones. Schütze and Pederson [38] have proposed to identify the context of every term of a given collection by clustering terms, based on the commonality of neighbor words. The idea is that words used in the same sense will share similar neighbors. So, by building a vector space representation based on this co-occurrence, it is possible to identify the different contexts of words. Voorhees [44] exploits *WordNet* (hyponymy links for nouns) for sense disambiguation. The proposed approach computes a *semantic distance* [31] between words to be compared in order to identify the right sense. A detailed state of art WSD can be found in [18].

Other semantic approaches attempt to extract term sense (single words or phrases) either by analyzing the syntax and semantics of the text [1], or by using pair terms distribution in the collection of documents, as it is done for instance in latent semantic indexing (LSI) [10].

LSI consists in representing the documents and the queries in a semantic concept space instead of word space. The concept space is automatically built by exploiting the associations among terms in a large collection of texts.

In concept-based IR, sets of words, names, noun phrases are mapped into the concepts they encode [15]. This text-concept mapping is driven by conceptual structures, which can be general or domain specific. They include dictionaries, thesauri and linguistic ontologies, and they can be either manually or automatically generated, or they may pre-exist [16]. WordNet and EuroWordNet are examples of (thesaurus-based) ontologies widely employed to improve IR systems. Automatic approaches attempt to generate these structures using either natural language processing or statistics and fuzzy reasoning [9], [16], [25].

A conceptual structure can be represented using distinct data structures: trees, semantic networks, conceptual graphs, etc. In [28] and [41] the use of conceptual graphs for representing documents and queries is discussed. The authors propose a method for measuring the similarity of phrases represented as conceptual graphs. Gonzalo *et al.* [15] propose an indexing method based on WordNet synsets. The vector space model is employed, using synsets as indexing space instead of word forms. In a similar spirit Richardson and Smeaton [32] have proposed to represent the documents and the queries on the basis of concepts names extracted from WordNet. The particularity of this approach compared to the previous ones concerns the way the query-document matching is carried out. Indeed they propose to measure the similarity between query and document by considering the semantic similarity between all pairs of concepts of the document and the query. These similarities are then summed up and normalized by the number of concepts in the document. Boughanem *et al.* [6] have proposed a more general conceptual IR model, which represents documents and queries as sub-trees of concepts (nodes) issued from an ontology. The document and query representations are not only set of concepts occurring in their contents but they are completed by intermediate concepts (nodes). Query evaluation is then based on the computation of a degree of inclusion of the query tree in the document one.

In [8], a concept based approach, based on the use of a neural-net spreading-activation algorithm and heuristics to select concepts in a thesaurus, is proposed. Another approach to detect the topical structure of a set of documents is presented in [19]. It complements a work done in [17] and [5] on the use of hyponymy and meronymy relationships.

3 Identifying fuzzy clusters of words

In order to have a view of the contents of a text that is better than the one provided by pure tf·idf-like statistics [36], a natural idea, explored in the following, is to try also to take advantage of the use of an ontology such as WordNet [27], where different semantic relations between words are stored.

3.1 Relations between words

In WordNet, there are six main relations that may hold between a pair of words (or word expressions) *w* and w'. They are

- *w* S *w′*: *w* and *w′* are *synonyms*;
- *w* G *w′*: *w* is in the *glossary* definition of *w′*;
- *w* I *w′*: *w* specializes *w′* ("is-a" relation); then *w′* I^{-1} *w* reads *w′* generalizes *w*;
- *w* P *w′*: *w* is a *part of* w; conversely *w′* P^{-1} *w* reads *w′* is composed of *w*;
- *w* D *w′*: *w* and *w′* are in the same *domain*;
- *w* R w': *w* is *related to w′*.

Relations S, D, and R are symmetrical, while G, I and P are anti-symmetrical. Besides, S, I and D are transitive. Moreover, it is possible to define new relations from these relations by taking their unions: *w* $(R^i \cup R^j)$ *w′* would mean that *w* is in relation R^i or R^j with *w′*; one may also think of composing relations when they are not transitive: *w* $R^i o R^i$ *w′* iff ∃ w°, *w* R^i *w*° and *w*° R^i *w′*, but this would lead to relate words that are already semantically distant.

3.2 What we are looking for

Then, given any pair of words present in a text, belonging to the same considered category (e.g., the noun groups), one can find the relations that hold between them. In the following, for simplicity, we only deal with lemmatized noun groups, although the approach may be applied to other types of words such as verbs for instance. In case of polysemic words or expressions, the multiple senses present in the ontology will be kept distinct when using the relations.

Generally speaking, the most interesting words in a text (here "interesting" means liable to give information about the contents of the text) are those that are i) frequent, or ii) are in relation of some type with many words in the text, and iii) that are sufficiently specific. Thus, a word in a text will be associated with two evaluations:

- its number of occurrences in the text (maybe normalized), as usual; moreover, words in the same "synset" (these words are synonymous of each other in then sense of relation S) will be counted together;

- its specificity, estimated through its "depth" in WordNet in the conceptual tree induced by the "is-a" relation. Note that the measure of the specificity of a word is absolute, since it is estimated by its depth in the ontology. But, it remains somewhat close in spirit to the idea of idf, since a specific word is often less frequent than a more general one at least in a very broad (maybe virtual) corpus made of a large variety of texts. However, it may be qualitatively different of an idf measure on limited corpora of rather specialized texts.

The third criterion “being in relation with *many* words in the text” involves the idea of clusters of words, which is now discussed.

3.3 Defining clusters from relations

Clusters of words present in the text that are in relation X of some type can be identified. Formally speaking, a cluster, which is not a singleton, is such that there is a path between any pair of words in the cluster made of a sequence of X-related words of the cluster.

Associated with a cluster are its size and its global frequency in the text, computed as the cumulated frequency in the text of the words in the cluster. Inside a cluster, each word has a level of "centrality", computed as the number of words in the cluster with which it is in direct relation.

Note that for a transitive relation, such as the *is-a* relation, two related words present in the text may be chained in the ontology through intermediary words that are not present in the text. These intermediary words might be used to expand the text and/or the query in a retrieval process, as done in [4] where it leads to some improvements. However, it does not seem that it is necessary if one just want to reflect the contents of a text. However, we may somewhat enlarge the above idea of cluster by considering that in case of the *is-a* relation, two words present in the text may be still put in the same cluster if they have a "close" common ancestor (not present in the text) in the hierarchy of the ontology. However, isolated words that have no close common ancestors should remain isolated.

3.4 Progressive selection of significant words

At this step, we have identified clusters of words. Each cluster has a "weight", which is the cumulated frequency of its words. Each word in a cluster is itself associated with three pieces of information: its frequency, its specificity (its "depth"), and the number of words to which it is related in the cluster.

The idea is to provide an image of the contents of a text under the form of the set of the fuzzy subsets of significant words in each cluster. Note that each fuzzy subset is not necessarily normalized. Indeed, the core of this structure is made of representatives of the clusters having one of the highest cumulative frequencies. The representatives that are chosen in each cluster are the words that are the ones that are the most central, or the most specific, provided that they are sufficiently frequent. Clearly, the exact balance of these three criteria should be a matter of experiments. Such a selection procedure may be iterated on the remaining words, adding progressively new layers of subsets of less and less significant words in the clusters that are progressively selected. At the beginning, only the most significant words in the “heaviest” clusters are selected, then more clusters are considered (leading to non normalized fuzzy subsets), and more words in each cluster.

4 Illustrative example

The following text is an excerpt from a web page http://www.iopcfund.org/erika.htm. It is used for illustrating the ideas presented in Section 3.

"*The Erika broke in two off the coast of Brittany, France, whilst carrying approximately 30 000 tonnes of heavy fuel oil. Some 19 800 tonnes were spilled. The sunken bow section contained 6 400 tonnes of cargo and the stern a further 4 700 tonnes. Compensation is available to any individual, business, private organisation or public body who has suffered pollution damage as a result of the Erika incident. Compensation is payable under the 1992 Civil Liability and Fund Conventions as*

enacted into French law. The total claims arising out of this incident by far exceeded the amount of compensation available, some €185 million or £125 million. In order to enable the 1992 Fund to make substantial payments to claimants, the French Government and the French oil company Total SA undertook to pursue their claims only if and to the extent that all other claimants were compensated in full, the claim by Total SA to rank after the Government's claim .Legal actions have been taken against the shipowner, his insurer and the 1992 Fund by 796 claimants. Out-of-court settlements have been reached with some 440 of these claimants. Actions by some 261 claimants (including 142 salt producers) are pending. The total amount claimed in the pending actions, excluding the claims from the French State and Total SA, is €59 million (£40 million)…"

This text contains 40 nouns (38 correspond to an entry in WordNet). Their frequency in the text and their WordNet depth are given below.

N	Word	Frequency in the document	Depth in WordNet IS-A hierarchy
1	Coast	1	5
2	Brittany	1	7
3	France	1	9
4	Tonne	4	8
5	Fuel oil	1	8
6	Bow	1	8
7	Section	1	4
8	Cargo	1	7
9	Stern	1	8
10	Compensation	3	12
11	Individual	1	5
12	Business	1	7
13	Organisation	1	5
14	Body	1	6
15	Pollution	1	7
16	Damage	1	10
17	Result	1	4
18	Incident	2	6
19	Liability	1	7
20	Fund	3	10
21	Convention	1	10
22	Law	1	5
23	Claim	5	8
24	Amount	2	6
25	Payment	1	10
26	Claimant	5	7
27	Government	2	7
28	Oil company	1	8
29	Extent	1	6
30	Million	4	8
31	Legal action	1	9
32	Shipowner	1	7
33	Insurer	1	9
34	Out-of-court	1	8
35	Action	2	9
36	Salt	1	7
37	Producer	1	8
38	State	1	8
39	Erika	2	no entry
40	Total SA	2	no entry

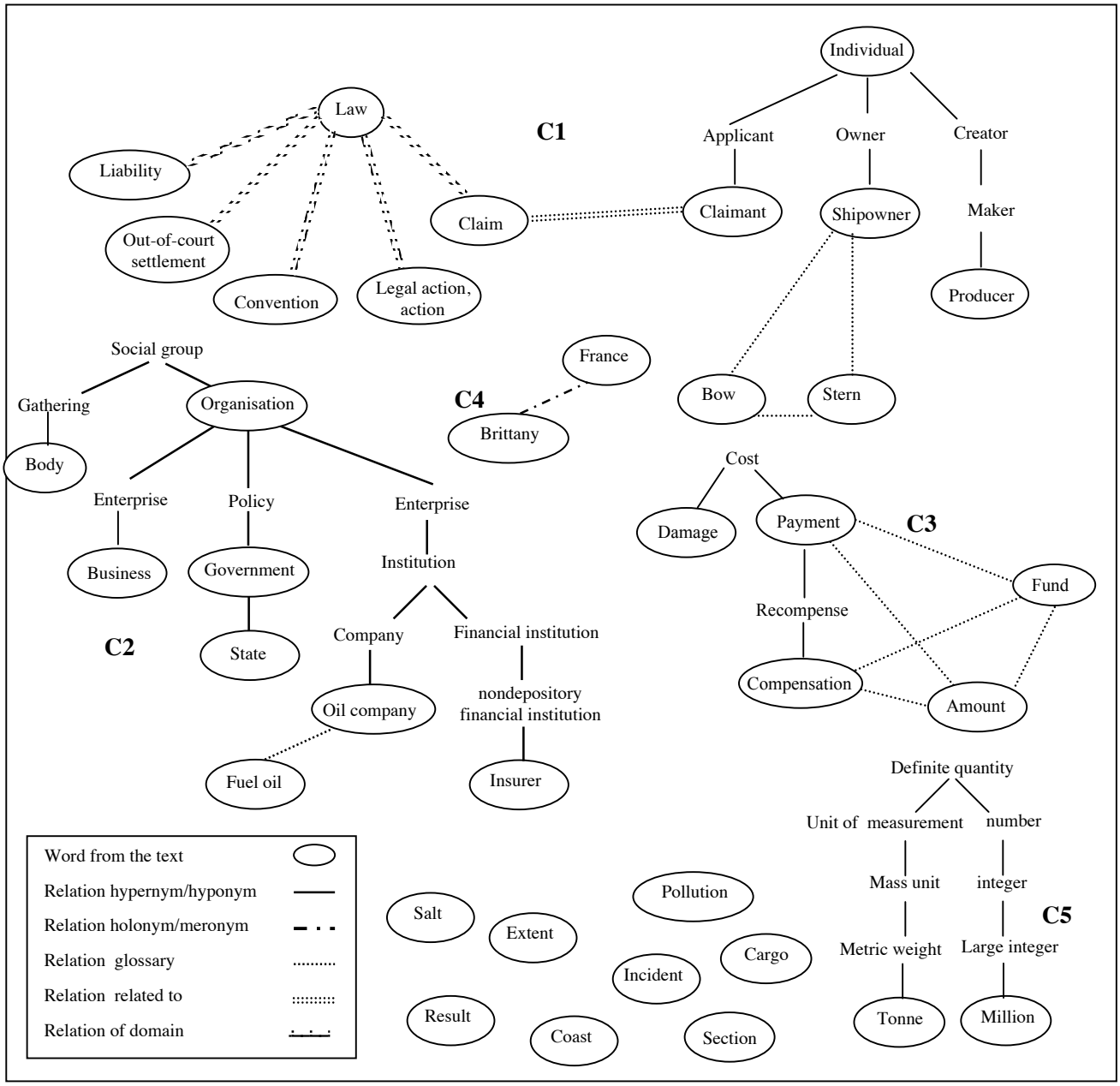
Table 1: Word frequencies in the text and their depth in the IS-A hierarchy of WordNet.

Note that for polysemic words (e.g. '*claim'* has 6 senses), we use a disambiguation method [4] in order to select the appropriate sense of the word in the text. The relations G, I, P, D, or R that hold between nouns of the text, give birth to the following clusters illustrated in Figure 1.

As seen in Figure 1, one can distinguish five clusters that are not singletons. Roughly speaking there are three big clusters C1, C2 and C3, and two small clusters C4 and C5. Note that some clusters (cluster C5 in Figure 1) gather words that have a "close" common ancestor in the ontology, although they are not related through a path only made of words present in the text.

We apply the procedure outlined in Section 3. At each iteration of the procedure, the clusters having the highest cumulative frequencies are selected. For each cluster selected as shown below, words that are the most central or the most specific providing that they are frequent enough are selected as representatives of the cluster. Table 2 gives an example of this procedure on the first cluster C1. The frequency (F), centrality (C) and specificity (S) values for the words in C1 are roughly estimated on a three level scale ('large', 'medium' and 'small').

The new representatives of clusters introduced at each iteration are denoted {C1: {w_1, …, w_i}, …, Ck:{w_1,…, w_j}}. At the first iteration, the stratum of subsets of words that first "emerge" from the text are: {C1: {claim, claimant}}. At the second iteration, we add: {C1:{legal action}, C3: {compensation, fund}}. At the third one, we add: {C1: {convention, law}, C2: {damage, amount}, C3: {government}}. At the fourth one we add: {C1: {stern, bow, shipowner}, C2: {oil company, state}}. The remaining words would appear in lower strata. These results are offering a rather good image of the contents of the text and provide some improvements with respect to the standard method that only considers words frequency.



"Figure 1 : Word clusters obtained by considering different relations between words."

It is worth noting that these results still be improved by taking into account the following remarks:

- one may not only consider nouns, but also verbs (for instance, it is clear that the verbal forms 'broke', 'spilled', 'sunken' in the first three sentences have quite specific senses that are of interest for the understanding).
- the selection of words that are representatives of a cluster is qualitative. Formal expressions (or even fuzzy rules) and experiments would be needed to express the exact balance of the three criteria (frequency, centrality and specificity) in the word selection process.
- in the procedure, only words present in the text can be selected. This sounds reasonable since one wants to represent the contents of the text. However, one may wonder if other words cannot do the job as well or even better. This might be in particular the case for "weak" clusters, when words are only linked through a "close" ancestor not present in the text.
- in case of clusters that are too large one may think of finding smaller clusters by forgetting some type of relations between words (e.g. the glossary one).

	Large			Medium			Small		
word	F	C	S	F	C	S	F	C	S
Law		x					x		x
Liability						x	x	x	
Out-of-court settlement						x	x	x	

Convention			x				x	x	
Legal action, action				x		x		x	
Claim	x				x	x			
Individual					x		x		x
Claimant	x				x	x			
Shipowner					x	x	x		
Producer						x	x	x	
Bow					x	x	x		
Stern					x	x	x		

Table 2: Fuzzy estimations of representative words in cluster C1.

5 Concluding Remarks

We have suggested new directions for extracting the significant words from a text. The idea of building 'concept clusters' has been already proposed in [20] as a new indexing method, without cluster 'summarization' and labeling procedure as above. Clearly, this work remains preliminary in several respects. The different options should be validated and made more precise through experiments. We have illustrated the procedure on a rather short text, whose length is the one of a paragraph. The procedure should not be applied in the same way to texts of any length. There are several reasons for that, first we have to be respectful of the local architecture of the text, and second to avoid the creation of clusters that would be too large and would include weakly related sub-clusters. However, there may exist pieces of information that may be significant at the scale of a full text (even if it is a book!), such as person names for instance.

What might be the exact benefit for query evaluation of a better image of the contents of texts, and how to use it? These are also open questions for further research. We may think of giving priority to those words that are in some sense in the core of the clusters having a high weight, while other words even if they are specific would be just optional, a kind of distinction that is reminiscent of the one proposed a long time ago in [30] between keywords that are necessarily appropriate and keywords that are only possible.

Another line of research would be to consider the use of different ontologies, or even of small worlds hierarchical structures [12], [13] built, e.g. from glossary relations between words.

References

1. T. Adi, O. K. Ewell, P. Adi, High selectivity and accuracy with READWARE's automated system of knowledge organization, *8th Text Retr. Conf. TREC-8* Gaithersburg NIST onli. publ. 1999 ht tp://trec.nist.gov/pubs/trec8/t8_proceedings.html
2. L. Azzopardi, M. L. Girolami, C. J. van Rijsbergen. Topic-based language models for adhoc information retrieval. *Proc. Inter. Joint Conf. Neural Netw/.*, Budapest, 2004 3281- 3286
3. T. T. Ballmer, Fuzzy punctuation or the continuum of grammaticality. Memo ERL-M590, Univ. of California, Berkeley, 1976.
4. M. Baziz, M. Boughanem, G. Pasi, H. Prade, A fuzzy logic approach to information retrieval using an ontology-based representation of documents. In: *Fuzzy Logic and the Semantic Web*. (E. Sanchez, Ed.), Elsevier, 363-377, 2006.
5. M. Berland, E. Charniak. Finding parts in very large corpora. Proc. 27th Annual Meeting Assoc. for Comput. Linguistics, Univ. Maryland, 1999. http://www.aclweb.org/anthology-new/P/P99/
6. M. Boughanem, G. Pasi, H. Prade, A fuzzy set approach to concept-based information retrieval. *Proc. 10th Inter. Conf. on Information Proces. and Mgmt. of Uncertainty in Knowl.-Based Syst.* (IPMU'04), Perugia, July 4-9, 2004, 1775-1782.
7. D. A. Buell, D. H. Kraft, A model for a weighted retrieval system. *J. of American Society for Information Science*, 32, 211-216, 1981.
8. H. Chen, K. J. Lynch, K. Basu, and D. T. Ng, Generating, integrating, and activating thesauri for concept-based document retrieval. IEEE Expert, Special Series on Artif. Intellig. in Text-based Inform. Systems, 8(2), 25-34, April 1993.
9. P. Cimiano, *Ontology Learning and Population from text Algorithms, Evaluation and Applications*, Springer Verlag, 2006.
10. S. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, R. A. Harshman, Indexing by latent semantic analysis, J. of the Society for Information Science, 41(6), 391-407, 1990.
11. B. Demant, Fuzzy-Retrieval-Strukturen, *Angew. Inf., Appl. Inf.*, 13, 500-502, 1971.
12. B. Gaume, Mapping the forms of meaning in small worlds, *Inter. J. of Intellig. Syst.*, to appear
13. B. Gaume, K. Duvignau, O. Gasquet and M.-D. Gineste Forms of meaning, meaning of forms, *JETAI,* 14 (1), 61-74, 2002.
14. Y. Gentilhomme, Les ensembles flous en linguistique. *Cahiers de Linguistique Théorique et Appliquée* (Bucharest) 5, 47-63, 1968.
15. J. Gonzalo, F. Verdejo, I. Chugur, J. Cigarrán, Indexing with WordNet synsets can improve text retrieval, Proc. COLING/ACL '98 Workshop on Usage of WordNet for Natural Language Processing, 1998. http://vldbarc.org/dblp/db/journals/corr/corr9808.html#cmp-lg-9808002

16. N. Guarino, C. Masolo, G. Vetere, OntoSeek: Content-based access to the web. *IEEE Intelligent Systems*, 70-80, 1999.
17. M. Hirst, Automatic acquisition of hyponyms from large text corpora. *Proc. of COLLING-92*, Nantes, France, 539-545, 1992.
18. N. Ide, J. Véronis, Word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1), 1-40. 1998.
19. M. Y. Kan, J. L. Klavans, K. R. McKeown. Synthesizing composite topic structure trees for multiple domain specific documents, Tech. Report CUCS-003-01, Columbia Univ., 2001.
20. B.-Y. Kang, D.-W. Kim, S.-J. Lee, Exploiting concept clusters for content-based information retrieval, *Inform. Sciences*, 170, 443–462, 2005.
21. D. H. Kraft, G. Bordogna, G. Pasi, Fuzzy set techniques in information retrieval, In *Fuzzy Sets in Approximate Reasoning and Information Systems,* (J. C. Bezdek, D. Dubois, H. Prade, eds.), Kluwer, 469-510, 1999.
22. R. Krovetz, B. Croft, Lexical ambiguity and information retrieval, *ACM Trans. on Information Systems*, 10(2), 115-141, 1992.
23. G. Lakoff, Hedges: A study in meaning criteria and the logic of fuzzy concepts. *J. of Philosophical Logic*, 2, 458-508, 1973.
24. G. Lakoff, *Women, Fire and Dangerous Things. What Categories Reveal about the Mind*. The University of Chicago Press, 1987.
25. S. Loh, L. K. Wives, J. P. M. de Oliveira, Concept-based knowledge discovery in texts extracted from the Web, ACM SIGKDD Explorations, 2(1), 29-39, June 2000.
26. M. Mauldin, J. Carbonell, R. Thomason. Beyond the keyword barrier: Knowledge-based information retrieval. *Information Services and Use* 7(4-5), 103-117. 1987.
27. G. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. Miller, Introduction to WordNet: An on-line lexical database. *J. of Lexicography*, 3**,** 235–244, 1990.
28. M. Montes-y-Gómez, A. López-López, A. Gelbukh. Information retrieval with conceptual graph matching. Proc. DEXA'00, Greenwich, 2000. Springer, LNCS 1873, 312–321, 2000.
29. C. V. Negoita, On the notion of relevance in information retrieval. *Kybern.* 2, 161-165, 1973.
30. H. Prade, C. Testemale, Application of possibility and necessity measures to documentary information retrieval, In: *Uncertainty in Knowledge-based Systems* (B. Bouchon, R. R. Yager, eds.), Springer Verlag, LNCS 286, 265-274, 1987. Reprinted in: *Fuzzy Measure Theory* (Z. Y. Wang, G. J. Klir, eds.), 311-319, 1992.
31. P. Resnik, Using information content to evaluate semantic similarity in a taxonomy, *Proc. of the 14th Inter. Joint Conf. on Artificial Intelligence* (IJCAI'95), Montréal, 448-453.
32. R. Richardson, A. Smeaton, J. Murphy. Using WordNet as a knowledge base for measuring semantic similarity between words. Proc. of AICS Conf., Trinity College, Dublin, 1994. http://citeseer.ist.psu.edu/187048.html.
33. B. B. Rieger, Feasible fuzzy semantics, *Proc. 7th Inter. Conf. on Computa. Linguis.* (COLING 78), Bergen, (Heggstad, K., ed.), 41-43, 1978.
34. B. B. Rieger, The baseline understanding model. A fuzzy word meaning analysis and representation system for machine comprehension of natural language, *Preprints 6th Europ. Conf. on Artif. Intellig.* (ECAI/84), Amsterdam, (T. O'Shea, ed.), 748-749, 1984.
35. B. B. Rieger, Computing granular word meanings. A fuzzy linguistic approach to computational semiotics, In *Computing with Words* (P. P. Wang, ed.), John Wiley & Sons, New York, 147-208, 2001.
36. G. Salton, M. J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, 1987.
37. M. Sanderson. Word sense disambiguation and information retrieval, *Proc. of ACM SIGIR'94 Conf.*, 17, 142-151, 1994.
38. H. Schütze, J. O. Pederson, Information retrieval based on word senses, *Proc. 4th Annual Symp. on Document Analysis and Information Retrieval*, Las Vegas, 161 – 175, 1995.
39. C. St-Jacques, C. Barrière, H. Prade. Fuzzy logic tools for lexical acquisition. *Proc. 10th Inter. Conf. on Information Processing and Management of Uncertainty in Knowledge-Based Systems* (IPMU'04), Perugia, Italy, July 4-9, 2004, 2045-2052.
40. P. Subasic, A. Huettner, Calculus of fuzzy semantic typing for qualitative analysis of text, *ACM KDD 2000*, *Workshop on Text Mining*, Boston, 2000.
41. R. Thomopoulos, P. Buch, O. Haemmerlé. Representation of weakly structured imprecise data for fuzzy querying. *Fuzzy Sets and Systems*, 140, 111-128, 2003.
42. L. M. Vaina, Semiotics of with, *Versus*, 17, (Bompiani, Milano), 96-112, 1978.
43. L. M. Vaina, Fuzzy sets in the semiotic of text, *Semiotica*, 31, 261-272, 1980.
44. E. M. Voorhes, Using WordNet to disambiguate word senses for text retrieval, Proc. 16th Annual Inter. ACM SIGIR Conf. on Res. and Develop. in Inform. Retrieval, Pittsburgh, 171-180, 1993.
45. D. Yarowsky, Unsupervised word sense disambiguation rivaling supervised methods. *Proc. 33rd Annual Meeting Assoc. for Computat. Linguistics*, Cambridge, Ma, 189-196, 1995.
46. L. A. Zadeh, Quantitative fuzzy semantics, Information Sciences, 3, 159-176, 1971.
47. L. A. Zadeh, Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. *Fuzzy Sets and Systems*, 90, 111-127, 1997.