

# GRANULES OF WORDS FROM FUZZY RELATIONS AND SPECTRAL CLUSTERING

Patrícia F. Castro<sup>1</sup> and Geraldo B. Xexéo<sup>1,2</sup>

<sup>1</sup>*Departamento de Engenharia de Sistemas e Computação, COPPE/UF RJ, Rio de Janeiro, Brazil*

<sup>2</sup>*Departamento de Ciência da Computação, IM/UF RJ, Rio de Janeiro, Brazil*

**Keywords:** Granular Computing, Fuzzy Relation, Spectral Clustering.

**Abstract:** Conventional information retrieval systems have proven ineffective in dealing with information overload. One possible solution is to incorporate some features that allow users of these systems to custom handle this information. In order to enable systems of this kind, some of the characteristics of present-day systems should be reviewed. Among other features, all documents are described with the same level of detail. We believe that the redrafting of document models is the starting point for reform of these systems. The paradigm of granular computing has proven to be very suitable for the treatment of complex problems and can produce significant results in large-scale environments such as the Web. This paper explores the granulation process of words with a view to its application in the subsequent improvement in document representation. We use fuzzy relations and spectral clustering in this process and present some result.

## 1 INTRODUCTION

Possible solutions for the information overload problem involve processes such as information retrieval, filtering, and extraction, as well as classification, clustering, and summarizing of documents, with the aim of assisting people to locate, in a more efficient way, the documents that meet their information needs. These needs can be defined as discovering or deriving new information, finding patterns in such information or separating the information that is useful from that which is not.

(Yao, 2002) says that the incorporation of these features in current information retrieval systems give rise to the emergence of a new generation of such a system: the information retrieval support systems.

In order to enable systems of this kind, some of the characteristics of present-day systems should be reviewed and re-engineered. These systems use document representation schema that are very simple, as well as a retrieval method that is also quite simple. All the documents are described with the same level of detail. The representation and retrieval method are the same, regardless of the user's characteristics. The structure and semantics of information, as contained in the document and in the collection, are not taken into consideration.

A paradigm that arises from the treatment of

information, known as granular computing, has attracted the attention of many researchers. According to (Yao, 2007), granular computing gathers a set of theories, methodologies, techniques, and tools, that employ granules to solve complex problems. According to (Predycz, 2005), the granules permeate any human task. Humans are constantly abstracting and formulating concepts from these granules, processing these concepts and returning the results of such treatment. To give an example, we can make an analogy with the human capability of dealing with images. At no given moment do we consider the pixels individually. All the time we build groupings of these pixels using some semantics capable of conveying notions of texture, colour, etc. Similarly, when analyzing text, the words are not considered individually. Groupings of these words, representing some semantics, convey their contents.

Moreover, humans can perceive the real world through many levels of granularity (abstraction) and can easily alternate between these various levels. Consequently, people abstract and consider only that which serves a specific purpose and ignore that which is irrelevant (Yao and Zhong, 2002); (Hobbs, 1985); (Yao, 2007a). By being able to focus on different levels of granularity, different levels of knowledge can be obtained as well as a deeper

understanding of the structure that is inherent to each type of knowledge.

Granular reasoning is, therefore, essential for human intelligence and, according to (Zhong, 2008), it can have a significant impact on problem-solving methodologies, especially in large-scale environments such as the Web.

The granulation process is based on the decomposition of objects according to some kind of relationship whereby these objects stay together. The process is inherently fuzzy, vague and imprecise. This paper explores this process, through the use of fuzzy relations. Based on this kind of relationship we use a spectral clustering algorithm in the creation of the granules and present some results.

## 2 RELATED WORKS

Alternative techniques for creating document models have aroused the interest of many researchers. Some approaches are based on the same vector model (Liu, 1994) and others suggest alternative ways. (Doan et al., 2005) proposes the modeling of texts based on the theory of fuzzy sets. (Khalid, 2006) presents a new paradigm for mining documents that can exploit the semantic features of documents. (Ingersen et al., 2008) is based on the cognitive aspects of information retrieval in its proposal. (Fishbein, 2008), proposes a scheme based on Holographic Reduced Representations (HRR) to encode both the semantic structure and syntactic structure of documents. Finally, similar to the proposal presented in this paper, (Lin, 2007) employs granular computing concepts in the treatment of the problem. The granules are formed by sets of keywords with frequent co-occurrence. In this context, other techniques have been proposed. Latent Semantic Analysis (LSA) (Dumais, 1997) uses principal component analysis to find groups of words that co-occur. Topic models (Steyvers, 2007) is a statistical model for discovering abstract topics in document collections. Analysis of formal concepts (Ganter, 2005) also uses the evaluation of objects and their relationships in order to identify concepts or topics of interest. We present a new method for the analysis of co-occurrence of words. We also use an algorithm that proves very effective for capturing this type of relationship between words to form granules. We believe this is the greatest contribution of this work.

## 3 WORD GRANULATION

Granulation means forming aggregates of indiscernible objects. The indiscernibility between these objects can be treated by a similarity function. There are two terms used to denote the main types of similarity between words (Kozima, 1993) (Rapp, 2002): paradigmatic similarity and syntagmatic similarity. Despite this distinction, it is relatively rare in published works, and some studies (Kozima, 1993) refer to the first type as semantic similarity, and to the second (Rapp, 2002) as relatedness similarity. The focus of our work is essentially on the second kind. Both types are computed using different methods and are used in a wide variety of applications. Relatedness similarity is generally measured by employing some statistical or algebraic tool. In this paper we present a fuzzy approach for the analysis of this similarity.

## 4 FUZZY RELATION

In this section we present a brief review of the theory of fuzzy relations (Chakrabarti, 2003); (Haruechaiyasak, 2002); (Zadeh, 1993) used in the assessment of the similarity between words and the subsequent creation of granules.

**Definition 1.** A fuzzy relation between two finite sets  $X = \{x_1, \dots, x_u\}$  and  $Y = \{y_1, \dots, y_v\}$  is formally defined as a fuzzy binary relation  $f: X \times Y \rightarrow [0,1]$ , where  $u$  and  $v$  represent the number of elements in  $X$  and  $Y$ , respectively.

**Definition 2.** Given a set of index terms,  $T = \{t_1, \dots, t_i\}$  and a set of documents,  $D = \{d_1, \dots, d_j\}$ , each  $t_i$  is represented by a fuzzy set  $h(t_i)$  of documents;  $h(t_i) = \{F(t_i, d_j) \mid \forall d_j \in D\}$ , where  $F(t_i, d_j)$  is the membership degree of  $t_i$  in  $d_j$ .

**Definition 3.** The fuzzy relationship between words is based on the evaluation of co-occurrence of  $t_i$  and  $t_j$  in the set  $D$  and can be defined as follows:

$$RT(t_i, t_j) = \frac{\sum_k \min(F(t_i, d_k), F(t_j, d_k))}{\sum_k \max(F(t_i, d_k), F(t_j, d_k))} \quad (1)$$

A simplification of the fuzzy RT relation based on co-occurrence of words is given as follows:

$$r_{i,j} = \frac{n_{i,j}}{n_i + n_j - n_{i,j}} \quad (2)$$

where

- $r_{i,j}$  represents the fuzzy RT relation between

words  $i$  and  $j$

- $n_{i,j}$  is the number of documents containing both the  $i^{\text{th}}$  and  $j^{\text{th}}$  words
- $n_i$  is the number of documents containing the  $i^{\text{th}}$  word
- $n_j$  is the number of documents containing the  $j^{\text{th}}$  word

## 5 SPECTRAL CLUSTERING

The spectral clustering technique is characterized by exploring the similarity between all pairs of objects. This technique has proven to be much more effective than more traditional techniques such as the  $k$ -means method, for example, which considers only the similarity of the objects to the central elements of their groups (Ng, 2001).

Given  $n$  data points  $x_1 \dots x_n$ , the spectral clustering algorithm constructs a similarity matrix  $S \in \mathbb{R}^{n \times n}$ , where  $S_{i,j} \geq 0$  reflects the relationship between  $x_i$  and  $x_j$ . It then uses similarity information to group  $x_1 \dots x_n$  into  $k$  clusters. There are several variants of spectral clustering. Here we consider the commonly used normalized spectral clustering (von Luxburg, 2006).

## 6 EVALUATION

We have created two distinct corpora. The first corpus contains 200 articles on computational intelligence selected from Google Scholar. These articles are related to 10 distinct subjects: cognition, fuzzy systems, genetic algorithms, neural networks, data mining, knowledge management, machine learning, pattern recognition, optimization and logic. For the second corpus, 160 articles on text mining/information retrieval were selected from the same site. In this case, eight subjects were used: clustering, latent semantic analysis, information retrieval, ontology, semantics, fuzzy relations, concept extraction and topic models. Each corpus was subjected to a pre-processing step where stopwords and words not classified as nouns were removed through application of a tagger available in <http://dragon.ischool.drexel.edu/>. Next, we analyzed the fuzzy correlation between these words, by applying equation 2, presented in section 4. Words and their correlations were subjected to the spectral clustering algorithm with implementation available over

<http://www.mathworks.com/matlabcentral/fileexchange/26354-spectral-clustering-algorithms>.

The implementation requires information about the value of  $k$ . Initially, we adopted  $k$  values of 10 and 8 for the first and second corpus, respectively. The justification for this lies in the fact that we have chosen 10 subjects and, therefore, based on this choice, we can control the groups generated. In a second evaluation, we reduced these values by half: 5 and 4, in each corpus, respectively. The aim was to examine the clustering algorithm's ability to make generalizations of the words contained in their groups. The algorithm parameters were kept at their default values. Tables 1, 2, 3 and 4 present the most significant words found in each of the clusters generated in each of the scenarios described above.

Table 1: The 10 clusters/granules of corpus 1.

GRANULE	SUBJECT	KEYWORDS
1	machine learning	computer, aspect, behavior, intelligence, paradigm, years
2	----	exploration, benchmark, architecture, variation, characteristic, interaction, fact
3	neural network	extension, importance, neuron, goal, stability, property, choice
4	knowledge management	storage, knowledge, capability, management, path, business
5	Cognition	representation, theory, life, language, cognition
6	pattern recognition	conclusion, input, classification, region, element, application
7	genetic algorithm	population, fitness, optimum, member, algorithm, convergence, solution
8	----	importance, performance, definition, statistics, measurement
9	data mining	attention, data, concept, generalization, addition, relationship
10	----	extraction, example, relations, variable, analysis, satisfaction

Aiming to establish a comparison with a well-known approach, we submit on the same basis, an algorithm for latent semantic analysis (LSA). Tables 5 and 6 represent the degree of similarity between the granules generated with the technique proposed in this work and the concepts (granules) obtained with LSA. The contents of each cell in the table represent the percentage of similarity between the

granules and concepts. To facilitate the analysis, we highlighted the cells with the greatest similarity measures.

Table 2: The 5 clusters/granules of corpus 1.

GRANULE	SUBJECT	KEYWORDS
1	genetic algorithm / optimization	exploration, performance, fitness, operator, member, algorithm, convergence, solution, population, optimum, crossover
2	neural networks	extension, input, example, property, regression, analysis, neuron, procedure, realization, synthesis, vector, coefficient, manner, applicability
3	data mining / knowledge management	user, technique, topic, storage, knowledge, management, capability, information, methodology, data, business, database
4	cognition / logic	behavior, theory, life, paradigm, language, computer, principle, aspect, manipulation, intelligence
5	cognition	protocol, difference, relations, complexity, analysis, problem, role, system, cognition, method, application

Table 3: The 8 clusters/granules of corpus 2.

GRANULE	SUBJECT	KEYWORDS
01	semantic	evolution, entity, library, management, language, technology, ontology, domain, description, semantics
02	latent semantic analysis	subspace, combination, detection, decomposition, association, retrieval, matrix, effectiveness, vector, collection
03	clustering	example, prototype, constraint, tendency, algorithm, objective, possibility, principle, data, problem,
04	information retrieval	period, kind, property, relations, decomposition, retrieval, information, expansion, criterion, construction
05	concept extraction	extension, representation, evaluation, concept, strategy, selection, explanation, logic, interpretation, identification, text, baseline
06	ontology	mechanism, classifier, correlation, thesaurus, creation, ontology, context, integration, recognition, source, module.
07	fuzzy relations	membership, co-occurrence, set, binary
08	topic models	probability, language, processing, mixture, model, generator

Table 4: The 4 clusters/granules of corpus 2.

GRANULE	SUBJECT	KEYWORDS
01	semantic/ontology	development, evolution, entity, library, management, language, version, technology, ontology, methodology, domain, description, semantics, input, mechanism, classifier, correlation, thesaurus, creation, ontology, context, integration, identification, recognition, source, module.
02	latent semantic analysis/ concept extraction	item, user, basis, subspace, combination, detection, decomposition, association, retrieval, matrix, effectiveness, vector, collection, method, extension, representation, evaluation, concept, strategy, selection, explanation, addition, logic, interpretation, identification, text, baseline
03	clustering/ information retrieval	example, prototype, constraint, tendency, algorithm, objective, possibility, finding, principle, data, problem, difficulty, period, user, minimum, kind, property, relations, decomposition, retrieval, information, expansion, criterion, method, construction
04	topic models	probabilistic, language, processing, mixture, model, generative

## 7 RESULTS

Looking through Tables 1 and 3, the proposed technique combines words significant enough to present the topics in each corpus. In corpus 1, for computational intelligence, 7 topics are easily identified from the words associated with their clusters/granules. In corpus 2, on text mining / information retrieval, we achieved better results, because the eight subjects that make up the corpus are easily identified. The results presented in Tables 2 and 4 show that the technique performs well against the ability of granule generalization contained in the corpus.

With respect to corpus 1 which was tested, we give special emphasis to the grouping of words that describe the topics of genetic algorithms/optimization and data mining/ knowledge management. Such topics are strongly related. The proposed technique shows consistency since it captures these relationships by grouping the words contained in their respective documents. LSA identified 13 clusters of words for corpus 1 text and 10 clusters for corpus 2.

Table 5: Equivalence between granules and concepts for corpus 1.

		LSA												
		1	2	3	4	5	6	7	8	9	10	11	12	13
G R A N U L E	1	0.39	0.53	0.50	0.49	0.45	0.42	0.92	0.42	0.42	0.45	0.51	0.32	0.56
	2	0.96	0.41	0.55	0.43	0.45	0.60	0.42	0.56	0.38	0.44	0.32	0.45	0.34
	3	0.46	0.67	0.43	0.40	0.45	0.56	0.50	0.89	0.42	0.34	0.23	0.56	0.42
	4	0.58	0.67	0.76	0.40	0.40	0.54	0.45	0.78	0.23	0.34	0.56	0.56	0.92
	5	0.34	0.45	0.76	0.23	0.40	0.54	0.45	0.78	0.95	0.34	0.56	0.56	0.23
	6	0.39	0.34	0.50	0.87	0.45	0.42	0.67	0.42	0.42	0.67	0.51	0.68	0.45
	7	0.46	0.24	0.43	0.46	0.45	0.56	0.78	0.45	0.42	0.34	0.25	0.56	0.42
	8	0.78	0.85	0.36	0.32	0.45	0.60	0.42	0.15	0.38	0.44	0.47	0.45	0.39
	9	0.39	0.34	0.36	0.87	0.45	0.42	0.67	0.47	0.68	0.67	0.51	0.68	0.76
	10	0.45	0.48	0.50	0.35	0.47	0.42	0.67	0.42	0.65	0.67	0.90	0.68	0.45

Table 6: Equivalence between granules and concepts for corpus 2.

		LSA									
		1	2	3	4	5	6	7	8	9	10
G R A N U L E	1	0.39	0.53	0.96	0.49	0.32	0.42	0.51	0.42	0.42	0.45
	2	0.43	0.41	0.32	0.43	0.45	0.60	0.42	0.96	0.38	0.44
	3	0.46	0.67	0.23	0.88	0.68	0.56	0.50	0.47	0.42	0.34
	4	0.58	0.67	0.56	0.40	0.56	0.57	0.45	0.78	0.23	0.87
	5	0.34	0.45	0.56	0.23	0.56	0.54	0.45	0.78	0.95	0.34
	6	0.39	0.34	0.51	0.33	0.68	0.82	0.67	0.42	0.42	0.67
	7	0.46	0.24	0.25	0.46	0.56	0.56	0.89	0.45	0.42	0.34
	8	0.78	0.92	0.47	0.32	0.45	0.60	0.42	0.15	0.38	0.44

Despite the greater number of groups, we can see that in all groups of words created with the technique presented in this work, both corpus are defined by an LSA equivalence. Thus, we understand that the techniques are equivalent in terms of effectiveness. Although not measured in terms of processing time for each technique, we observed that the technique proposed here performs better than LSA.

## 8 CONCLUSIONS

The paper explored the granulation process based on fuzzy relations of co-occurrence and spectral clustering. The methodology was presented and some preliminary results were shown. These results demonstrate the real applicability of the proposal. Our next step will be to explore the ability of this technique in the generalization and specialization of granules. This feature will allow the construction of building ontologies with these granules. We also intend to study a way to allow

overlap between the granules produced. The clustering algorithm used does not allow this overlap and we understand that this feature will produce granules much more significant than those produced with the current method. We believe the introduction of such features will enable the representation of documents whose handling is closer to the human way of dealing with granules, as described in the introduction.

## ACKNOWLEDGEMENTS

The authors would like to thank the financial support of CNPq, CAPES, FAPERJ and Fundação Coppetec.

## REFERENCES

- Chakrabarti, S. 2003. *Mining the Web: Discovering Knowledge from Hypertext Data*. s.l.: Morgan Kaufmann, 2003.
- Doan, S, Ha, S and Horiguchi, S. 2005. A Fuzzy-Based Approach for text Representation in Text Categorization. *14th IEEE International Conference on Fuzzy Systems*. 05 2005, pp. 1008-1013. ISBN: 0-7803-9159-4.
- Dumais, S and Landauer, T. 1997. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review*. 1997. Vols. 104, No.2, pp. 211-240.
- Fishbein, J. 2008. Integrating Structure and Meaning Using Holographic Reduced Representation to Improve Automatic Text Classification. *Master Thesis*. University of Waterloo : s.n., 2008.
- Ganter, B, Stumme, G and Wille, R. 2005. Formal Concept Analysis: Foundations and Applications. *Lecture Notes in Artificial Intelligence*. s.l.: Springer-Verlag, 2005. ISBN 3-540-27891-5.
- Haruechaiyasak, C, Shyu, M and Chen, ML. 2002. Web Classification Based on Fuzzy Association. *Proceedings of the 25th Annual International Computer Software and Applications Conference (COMPSAC'02)*. 2002.
- Hobbs, J. 1985. Granularity. *Proceeding of the Ninth International Joint Conference on Artificial Intelligence*. 1985, pp. 423-435.
- Ingersen, P, Skov, B and Larsen, B. 2008. Inter and Intra-document Context Applied in Polyrepresentation for Best Match IR. *Information Processing and Management: an International Journal*. 2008, Vol. 44, pp. 1673-1683.
- Khalled, S. 2006. A Semantic Graph Model for Text Representation and Matching in Document Mining. *PhD Thesis*. University of Waterloo. Canadá : s.n., 2006.
- Kozima, T. 1993. Similarity Between Words Computed by Spreading Activation on an English Dictionary. *Proceedings of the 6th Conference of the European Chapter of the ACL*. 1993, pp. 232-239.
- Lin, T. 2007. Granular Computing and Modeling the Human Thoughts in Web Documents. *Lecture Notes in Artificial Intelligence. Proceedings of the 12th International Fuzzy Systems Association World Congress on Foundations of Fuzzy Logic and Soft Computing*. Cancun, Mexico : s.n., 2007. pp. 263-270. ISBN: 978-3-540-72917-4.
- Liu, G. 1994. The Semantic Vector Space Model (SVSM): A Text Representation and Searching Technique System Sciences. *Information Systems: Collaboration Technology Organizational Systems and Technology. Proceedings of the Twenty-Seventh Hawaii International Conference*. 1994, Vol. IV, pp. 928-937.
- Ng, A and Jordan, M. 2001. On Spectral Clustering: Analysis and an Algorithm. *Advances in Neural Information Processing Systems*. 2001, Vol. 14.
- Predycz, W. 2005. *Knowledge-Based Clustering. From Data to Information Granules*. Hoboken, New Jersey : John Wiley & Sons, 2005.
- Rapp, R. 2002. The Computation of Word Associations: Comparing Syntagmatic and Paradigmatic Approaches. *Proceedings of COLING-02*. 2002.
- Steyvers, M and Griffiths, T. 2007. Probabilistic Topic Models. [book auth.] T Landauer, et al. *Latent Semantic Analysis: A Road to Meaning*. s.l.: Laurence Erlbaum, 2007.
- von Luxburg, U. 2006. *A tutorial on Spectral Clustering*. Technical Report 149 : Max Planck Institute for Biological Cybernetics, 2006.
- Yao, Y. 2007a. A Ten-year Review of Granular Computing. *Proceedings of IEEE International Conference on Granular Computing*. 2007a, pp. 734-739.
- Yao, Y and Zhong, Y. 2002. Granular Computing using Information Tables. [book auth.] T Y Lin, Y Yao and L A Zadeh. *Data Mining, Rough Sets and Granular Computing*. Heidelberg : Physica, 2002, pp. 102-124.
- Yao, Y. 2002. Information Retrieval Support Systems. *The 2002 IEEE World Congress on Computational Intelligence, Honolulu, Hawaii, USA*. 2002, pp. 773-778.
- . 2007. The Art of Granular Computing. *LNAI 4585*. s.l.: Springer, 2007, pp. 101-102.
- Zadeh, L. 1993. Fuzzy Sets. *Readings in Fuzzy Sets for Intelligent Systems*. 1993.
- Zhong, N, et al. 2008. Towards Granular Reasoning on the Web. *Proceedings of the 2008 Workshop on New Forms of Reasoning for Semantic Web: Scalable, Tolerant and Dynamic (NEFORD 2008), the 3rd Asian Semantic Web Conference (ASWC2008)*. 2008.