

# A Text Mining Approach towards Knowledge Management Applications

Alexandre L. Gonçalves<sup>1</sup>, Fabiano Beppler<sup>1,2</sup>, Alessandro Bovo<sup>1,2</sup>, Vinícius Kern<sup>1,2</sup>, Roberto Pacheco<sup>1,2,3</sup>

<sup>1</sup>*Stela Institute, Florianópolis, SC, Brasil*

*{a.l.goncalves, fbeppler, alessandro, kern}@stela.org.br*

<sup>2</sup>*Knowledge Engineering and Management Pos-Graduation, Federal University of Santa Catarina, Florianópolis, SC, Brazil*

<sup>3</sup>*Department of Computing and Statistics, Federal University of Santa Catarina, Florianópolis, SC, Brazil*  
*{rpacheco}@inf.ufsc.br*

## Abstract

*The recognition of entities and their relationships in document collections is an important step towards the discovery of latent knowledge as well as to support knowledge management applications. The challenge lies on how to extract and correlate entities, aiming to answer key knowledge management questions, such as; who works with whom, on which projects, with which customers and on what research areas. The present work proposes a knowledge mining approach supported by information retrieval and text mining tasks in which its core is based on the correlation of textual elements through the LRD (Latent Relation Discovery) method. Our experiments show that LRD outperform better than other correlation methods. Also, we present an application in order to demonstrate the approach over knowledge management scenarios.*

## 1. Introduction

The knowledge has become an important strategic resource for organizations. The generation, codification, management and sharing of the organization knowledge is essential for the innovation process. To know who works with whom, on which projects, with which customers and on what research areas is an important step towards the understanding of intra or extra-organizational relationships.

In the last years the amount of documents has been increased considerable, as much in organizations as in the Web. We state that documents, instead of organizational databases, are the primary resource to reveal latent knowledge, once they keep registered relevant textual patterns (entities), such as, people, organizations and projects, and how such entities are related to each other.

In this work, we present an entity-based knowledge mining approach to support knowledge management tasks. Thus, through the combination of extraction and retrieval of information and text mining tasks, we intend to unveil different levels of connectivity among entities through the projection of collaborative networks or knowledge maps. Such networks are useful tools to provide insights, for instance, about people relationships, that can be spontaneous (i.e., they have common interests and act based on this) or inducted (i.e., they have worked or are working together in a couple of projects).

The rest of the paper is organized as follows. We present the work background in Section 2. Our text mining approach is presented in Section 3. Results are reported in Section 4. Section 5 presents a knowledge management application and finally, we conclude the paper and discuss future work in Section 6.

## 2. Background

Named Entity Recognition (NER) has been applied with success in the identification and classification of textual elements, such as, people, organization, places, monetary values and dates taken into account document collections [Grover et al. 2002], [Cunningham 2002], [Zhu et al. 2005b], [Brin, 1998], [Soderland, 1999], [Ciravegna, 2001]. As result of the process, for each document a set of entities is extracted. Thus, applying co-occurrence-based methods the connectivity among entities can be achieved in order to indicate insights toward knowledge discovery.

Co-occurrence methods are important, for instance, in the identification of collocations<sup>1</sup>

---

<sup>1</sup> Natural sequence of words, which possibly, identifies candidate concepts to be extracted from written information.

[Manning and Schütze 1999], information retrieval through vector expansion [Gonçalves et al. 2006] and also as the core for the current proposed. Such methods aim to correlate textual elements aiming to unearth latent relationships. In this context are  $t$  test, *chi-square* ( $\chi^2$ ), phi-squared ( $\phi^2$ ) [Conrad and Utt 1994], [Church and Gale 1991], *Z score* [Manning and Schütze 1999], [Vechtomova et al. 2003], Mutual Information (MI) [Church and Hanks 1990] or derivations of Mutual Information (VMI) [Vechtomova et al. 2003]. Also, more empirical methods have been applied such as CORDER [Zhu et al. 2005a] and *Latent Relation Discovery* (LRD) [Gonçalves et al. 2006].

First of all, the result of the process obtained through the correlation of entities provides direct relationships among entities. However, it is only useful for preliminary analysis on knowledge management applications. Additionally, indirect relationships can also be achieved through clustering algorithms. Such techniques have been used in a wide range of application domains, such as, information retrieval, data mining, machine learning and pattern recognition. They have as main target the grouping of similar objects in the same class [Hair et al. 1998], [Johnson and Wichern 1998], [Halkidi et al. 2001].

All methods and techniques discussed so far are the basis for the proposed approach on achieving knowledge management applications. Knowledge management is seemed as systematic and disciplined actions in which organization can take advantage to get some return [Davenport and Pruzak 1997]. According to Schreiber (2002), knowledge management is an important tool for the enhancement of the organizational knowledge infrastructure. In this scenario, the information technology has an important role in the process of transformation of the knowledge, from tacit to explicit [Marwick 2001]. Thus, we state making explicit entities and their relationships through information extraction and retrieval, and text mining techniques is an important step toward knowledge management applications, such as, communities of practice [Lesser and Storck 2001], [Wenger 1998], expertise location [Marwick 2001] and competency management [Dawson 1991], [Hafeez et al. 2002].

### 3. Proposed Approach

The proposed approach (Figure 1) is an extension of the traditional knowledge discovery from textual database model. Traditionally, textual elements are extracted and applied in the data mining phase aiming to reveal useful patterns [Mooney and Nahm 2005]. Our approach is concentrated as much in the extraction of textual elements (i.e., entities and concepts) as in the correlation of such elements. Thus the extraction and correlation of

textual elements are the basis for the data mining and information retrieval phases aiming to promote support to knowledge management applications.

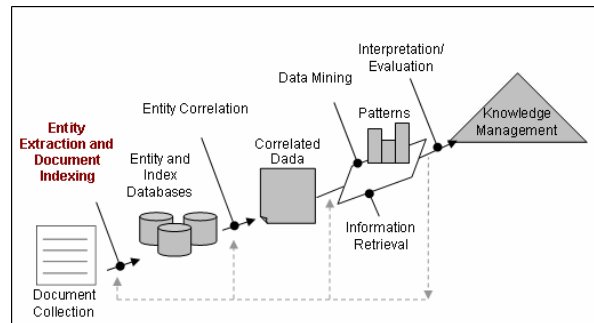


Figure 1: Text Mining approach toward Knowledge Management Applications

Next sections discuss the main phases of the approach, including the extraction and correlation of entities, the composition of the entity database and its use by information retrieval and clustering techniques to support knowledge management applications.

#### 3.1. Entity Extraction

The entity extraction phase called *Named Entity Recognition* (NER) aims to discover proper names, their variations and classes [Cunningham 2002], [Grover et al. 2002]. A named entity (NE) can be defined as a textual element which represents an object in a physical or abstract world. Formally, an entity can be defined as a vector  $E$  composed of description, class and additional information, i.e.,  $E = \{\text{description, class, <additional information>}\}$ . Additional information may indicate, for instance, the positions where such patterns occur in the document.

The NER process is mainly composed of two components, that is, lexical structures and patterns. The lexical structures are essential for the process and represent the knowledge base [Guthrie 1996]. Every class taken into account during the NER phase (e.g.: person, organization and project) is associated with a lexical table. Each lexical table stores a set of words that identify itself (e.g.: the person lexical table would have names such as “John” and “Smith”). Additionally, patterns are common in written language and they represent a sequence of words which can be classified (e.g.: an entity similar to “<content> Institute” is classified as “Organization” or “Institute”, or “<content>, Dr” is classified as a “Person/Doctor”). In this direction, the utilization of regular expressions is an important tool toward the identification of patterns which may be named/classified.

Table 1: List of intra-document weights for each relation among entities with seven entities and three documents

Document Id	Source Entity (SE)	Freq	Target Entity (TE)	Freq	Distance	Partial Relation Strength
1	E1	4	E2	2	2.0000	0.4387
1	E1	4	E3	3	2.0731	0.4938
1	E1	4	E4	1	2.3634	0.3094
1	E1	4	E7	1	2.8540	0.2562
1	E2	2	E3	3	2.3412	0.3123
1	E2	2	E4	1	2.6887	0.1632
1	E2	2	E7	1	3.0805	0.1424
1	E3	3	E4	1	2.2642	0.2584
1	E3	3	E7	1	2.5654	0.2280
1	E4	1	E7	1	3.8074	0.0768
2	E1	2	E3	2	2.0000	0.5850
2	E1	2	E4	1	2.1462	0.4088
2	E1	2	E5	2	2.7925	0.7771
2	E3	2	E4	1	2.6887	0.3263
2	E3	2	E5	2	2.2925	0.9465
2	E4	1	E5	2	3.3877	0.5542
3	E2	3	E6	2	2.0975	0.7827
3	E2	3	E7	2	2.7770	0.3511
3	E6	2	E7	2	3.3877	0.4270

### 3.2. Entity Correlation

Correlation methods have been widely applied to the identification of collocations. Additionally, methods may also be applied to any textual element to indicate proximity, or in our context, the relation strength. To the present work, textual elements are regarded as entities.

Our proposal use the LRD algorithm [Gonçalves et al., 2006] as the method to establish relations among entities taking into account tree aspects: (a) **co-occurrence**: two entities co-occur if they appear in the same document; (b) **distance**: the distance of all co-occurrences intra documents is taken into account; and (c) **relation strength**: given an entity,  $E1$ , the relation strength between two entities  $E1$  and  $E2$  takes into account their co-occurrence, mean distance, and frequency in co-occurred documents as defined in Equation 1. The greater the mean distance is, the smaller the relation strength. Generally, the relation strength between  $E1$  and  $E2$  is asymmetric depending on whether  $E1$  or  $E2$  is the target.

$$R(E1, E2) = \hat{p}(E1, E2) \times \sum_i \left( \frac{f(\text{Freq}_i(E1)) \times f(\text{Freq}_i(E2))}{m_i(E1, E2)} \right), \quad (1)$$

where  $f(\text{Freq}_i(E1)) = \text{tfidf}_i(E1)$ ,  $f(\text{Freq}_i(E2)) = \text{tfidf}_i(E2)$ , and  $\text{Freq}_i(E1)$  and  $\text{Freq}_i(E2)$  are the numbers of occurrences of  $E1$  and  $E2$  in the  $i$ th document, respectively. The term frequency and inverted document frequency measure  $\text{tfidf}$  is defined as  $\text{tfidf}_i(j) = \text{tf}_i(j) * \log_2(N / df_j)$ , where  $\text{tf}_i(j) = f_i(j) / \max(f_j(k))$  is the frequency  $f_i(j)$  of entity  $j$  in the  $i$ th document normalized by the maximum frequency of any entity in the  $i$ th document,  $N$  is the number of documents in the corpus, and  $df_j$  is the number of documents that contain the entity  $j$ .

### 3.3. Composition of the Entity Database

As result of the entity extraction and correlation processes, a database with pairs of related entities is created. Each pair, represented by a source entity (SE) and a target entity (TE) is calculated based on LRD method. Given a pair  $\langle SE, TE \rangle$ , the relation strength is stored as shown in the Table 1. For this example, seven entities extracted from three documents were correlated.

The table is indexed and stored as an inverted index. So given a SE, it is possible to retrieve all related TEs sorted by their relation strength. For instance, the relation strength between the target entity (TE)  $E3$  and the source entity (SE)  $E1$  is computed using the relative frequency (number of documents in which the relation occur by the total number of documents) times the partial weights summation, i.e.,  $R(E1, E3) = 2/3 * (0.4938 + 0.5850) = 0.7192$ .

### 3.4. Information Retrieval and Pattern Generation

As mentioned, the entity database is indexed and through information retrieval techniques, answers to knowledge management questions, such as, which the related projects considering a particular SE, makes possible.

Such information is useful in order to produce an initial map about the document collection regarding entities and their relationships. However, it only enables the establishment of direct relationships, i.e., when entities co-occur in the same document. To overcome that, we apply an entity clustering phase in order to obtain more complex relationships. So far, we have applied a simplified but quite fast algorithm, defined as fast clustering to create these maps, called knowledge maps. Given a SE, the  $k$  most TEs are retrieved. It will compose all centroids (first level) of the map. For each centroid a new search is carried out up to a threshold (second level). By using Table 1, all entities from a particular cluster are connected, that is, intra-cluster correlation. The process is repeated among entities from different clusters in an inter-cluster correlation process. Before presentation phase, entities which occur in multiple clusters are merged. In this way, only one entity remains with multiple references for its clusters and other entities.

## 4. Evaluation

As the core of the work lies on the entity correlation, the present work proposes a model towards the evaluation of the relation strength among entities. Although, such task tends to be not easy, mainly due to the lack of standard data sets to tackle the correlation of textual element, we intend to compare the precision of LRD and other standard correlation methods.

The main evaluation approaches can be defined as: (a) **Quantitative** methods judge whether results achieved by the model, based on quantifiable parameters, are suitable. For example, a classic method for analyzing hierarchical agglomerative clustering is the cophenetic correlation coefficient [Sokal and Rohlf, 1962], [Halkidi et al., 2001]. The Square Error Criterion is commonly used to evaluate the efficiency of numerical data clustering [Duda and Hart, 1973]; (b) **Gold standard** approaches compare the learned model to an “ideal” model produced *a priori* by domain experts. These are typical in information retrieval, text categorization and information extraction, e.g., MUC [DARPA, 1995], TREC<sup>2</sup>, SMART<sup>3</sup> e Reuters<sup>3</sup>. Their primary disadvantage is that standard collections are expensive to produce. Moreover, they are intrinsically subjective since they are based on expert opinion,; and (c) **Task oriented** evaluations examine algorithms in the context of applications. They are concerned whether the learning algorithm has produced a model that properly works. Tonella et al. [2003] discuss some of the problems associated with such approach including its cost and the need for careful design in order to minimize subjectivity.

Due to the problem context the gold standard and task oriented approaches would be more suitable. However, in general it demands high costs regarding time and people to create datasets as well as it introduces a bias due to the subjectivity. Intending to overcome such limitations, we propose a valuations model discussed below.

#### 4.1 Results

In our experiment we have used the LRD method and compared it with other four statistics methods (MI, VMI, Phi-squared and Z score) in the relation strength establishment among entities.

The evaluation is based on a set of 2500 papers from the “Semantic Web” and “Ontology” areas. For each paper the NER process is applied and the result stored as a vector. Each element of the vector represents an entity composed of description, class (Person, Organization and Research Area) and its positions through the document. From the 2301 extracted entities, 970 are organizations, 914 are people and 417 are research areas. In order to avoid subjectivity, the relation between an SE and its TEs is firstly defined by calculating the joint frequency through a traditional search engine.

We state that the joint frequency extracted from documents in the Web which mention the relation  $R\langle E1, E2\rangle$  is an indicative of relatedness. So, given an entity  $E1$ , a search engine available on the Web was used to establish the joint frequency with its

related entities. The joint frequency is also an indicative of order or importance. As result, each SE will produce a list with its related TEs and used during the evaluation phase. To analyze the precision, *Spearman’s* correlation method was considered (Equation 2).

$$RA = 1 - \frac{6 \sum_i (R_{i,CM} - R_{i,SE})^2}{N^3 - N}, \quad (2)$$

where  $-1 \leq RA \leq 1$  (1 indicates perfect correlation),  $R_{i,CM}$  indicates the order of the entity for a specific correlation method,  $R_{i,SE}$  the order obtained through the search engine and  $N$  the number of entities used in the query.

For each entity, the top 10 related entities in all correlation methods LRD ( $M1$ ), *Phi-squared* ( $M2$ ), MI ( $M3$ ), VMI ( $M4$ ), and *Z score* ( $M5$ ) were used. Nevertheless, different entities can be selected by a particular method. Generally, it tends to produce lists TEs greater than 10. Also, the study is based on different window sizes, being 50, 100, 200 and no window. Window size is used to validate the relation between two entities, that is, if two entities which co-occur in the same document are out of some specified range, the relation is not valid.

Table 2 presents an example using the “Semantic Web” term (entity in the research area class) and window size of 50. In order to avoid excessive penalties for a particular method which selects TEs beyond the  $N$  threshold the ranking is normalized. If a specific TE is not selected by some method, the index representing the order is  $N+1$ . If the entity order, from  $M1$  to  $M5$ , is equal to  $N+1$  and the index generated by search engine is also lesser than  $N+1$ , or the entity position from  $M1$  to  $M5$  is different of  $N+1$ , the partial *Spearman’s* index  $(R_{i,CM} - R_{i,SE})^2$  is calculated for the  $i$ th pair, otherwise, the correlation value is not taken into account. Best results were achieved by LRD, *Phi-squared* and *Z score*, with 0.930, 0.601 and 0.372, respectively.

Table 3 presents the summarized Spearman’s correlation index (regarding all 2301 entities) for *organization*, *person* and *research area* classes as well as the average value. The LRD algorithm achieve the best results followed by the *Phi-squared* and *Z score* methods, while the MI and VMI methods present the worst results. Among the three classes, the *person* class has the worst performance. The MI and VMI problem is due to their deficiency to deal with low frequencies. In this case high values are generally attributed to relations with little importance.

<sup>2</sup> <http://trec.nist.gov/>

<sup>3</sup> [http://www.dcs.gla.ac.uk/idom/ir\\_resources/test\\_collections/](http://www.dcs.gla.ac.uk/idom/ir_resources/test_collections/)

Table 2: Establishment of the Spearman’s correlation for the “Semantic Web” entity and its most related pair taken into account 5 methods, where SE=search engine, order=order established via SE, from  $M1$  to  $M5$  the order established by correlation methods (LRD, Phi-squared, MI, VMI and Z score) and from  $R1$  to  $R5$  the Spearman’s partial value

$$(R_{i,CM} - R_{i,SE})^2 \text{ for each correlation methods.}$$

Related entities	SE	Order	M1	M2	M3	M4	M5	R1	R2	R3	R4	R5
Xml	3.240.000	1	6	11	11	11	11	25	100	100	100	100
Rdf	3.140.000	2	2	11	11	11	11	0	81	81	81	81
ontology	1.530.000	3	1	11	11	11	11	4	64	64	64	64
networks	1.470.000	4	10	3	11	11	11	36	1	49	49	49
web services	1.460.000	5	4	11	11	11	11	1	36	36	36	36
Owl	732.000	6	8	11	11	11	11	4	25	25	25	25
knowledge management	713.000	7	11	6	11	11	6	16	1	16	16	1
agent	623.000	8	5	11	11	11	11	9	9	9	9	9
interoperability	547.000	9	7	1	11	11	11	4	64	4	4	4
information systems	474.000	10	11	8	11	11	11	1	4	1	1	1
environments	471.000	11	11	4	2	4	11	*	49	81	49	*
reasoning	439.000	12	9	11	11	11	11	9	*	*	*	*
patterns	401.000	13	11	9	3	10	1	*	16	100	9	144
Daml	302.000	14	3	11	11	11	11	121	*	*	*	*
User interface	262.000	15	11	11	11	6	2	*	*	*	81	169
simulation	253.000	16	11	10	5	11	8	*	36	121	*	64
knowledge representation	237.000	17	11	7	11	11	10	*	100	*	*	49
hypertext	231.000	18	11	11	10	5	3	*	*	64	169	225
intelligent systems	164.000	19	11	11	8	7	11	*	*	121	144	*
Trees	157.000	20	11	11	11	9	7	*	*	*	121	169
electronic commerce	140.000	21	11	2	4	11	9	*	361	289	*	144
hypermedia	123.000	22	11	11	11	8	11	*	*	*	196	*
problem solving	118.000	23	11	11	6	11	5	*	*	289	*	324
relational database	83.500	24	11	5	1	2	4	*	361	529	484	400
database management	75.600	25	11	11	11	3	11	*	*	*	484	*
ontology engineering	68.000	26	11	11	9	1	11	*	*	289	625	*
system architecture	44.300	27	11	11	7	11	11	*	*	400	*	*
<b>Spearman</b>								230	1308	2668	2747	2058
								0,930	0,601	0,186	0,161	0,372

Table 3: Spearman values between -1 and 1 for the organization, person and research area classes as well as the average value for different window configurations

Spearman [-1,1]		Organization	Person	Research Area	Average
No window	LRD	0.4231	0.4496	0.4236	0.3979
	Phi-squared	0.1487	-0.0291	0.0834	0.1306
	MI	0.0797	-0.1525	-0.0071	0.0515
	VMI	0.0797	-0.1525	-0.0071	0.0515
	Z Score	0.1487	-0.0291	0.0834	0.1306
Window (50)	LRD	0.3286	0.2572	0.2866	0.2739
	Phi-squared	0.0598	-0.0562	-0.0048	-0.0180
	MI	0.0157	-0.1715	-0.0700	-0.0542
	VMI	0.0768	-0.1367	-0.0142	0.0173
	Z Score	0.1126	-0.0764	0.0198	0.0231
Window (100)	LRD	0.3423	0.2788	0.3242	0.3515
	Phi-squared	0.1073	-0.0519	0.0488	0.0910
	MI	0.0624	-0.1551	-0.0196	0.0339
	VMI	0.0990	-0.1344	0.0151	0.0808
	Z Score	0.1161	-0.0884	0.0345	0.0759
Window (200)	LRD	0.3847	0.3452	0.3759	0.3980
	Phi-squared	0.1380	-0.0231	0.0923	0.1620
	MI	0.0803	-0.1298	0.0111	0.0827
	VMI	0.0824	-0.1317	0.0276	0.1320
	Z Score	0.1075	-0.0844	0.0549	0.1417

## 5. Knowledge Management Scenarios

Tools for analysis of entities and their relationships represent important resource towards knowledge management applications, such as, communities of practice, social networks, expertise location and competency management.

Such applications have in common the use of entity relations to express different objectives. While communities of practice are broader, that is, all entity classes can be projected together, social networks have its focus on person class. Communities of practice are therefore a more general class of application, being easily configured to achieve social networks.

According to Alani et al. [2003], communities of practice are established by groups whose members are interested in a particular job, procedure, or work domain. Lesser and Stoch [2001] define communities as groups engaged in sharing and learning, based on common interests. On the other hand, organizations are more and more aware about the knowledge and skills of collaborators as their most valuable resource. Know who knows what is now a critical activity. Equally important is to know who knows whom, both inside and outside organizations. Thus, such networking identification may be useful as much in the optimization of

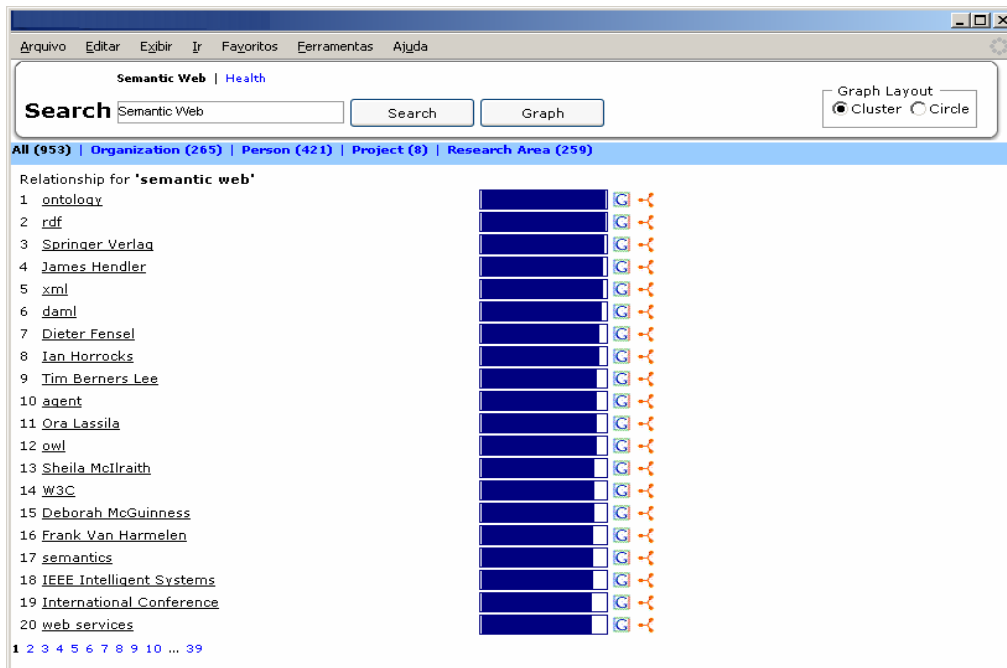


Figure 2: Entities and their relationships regarding all classes

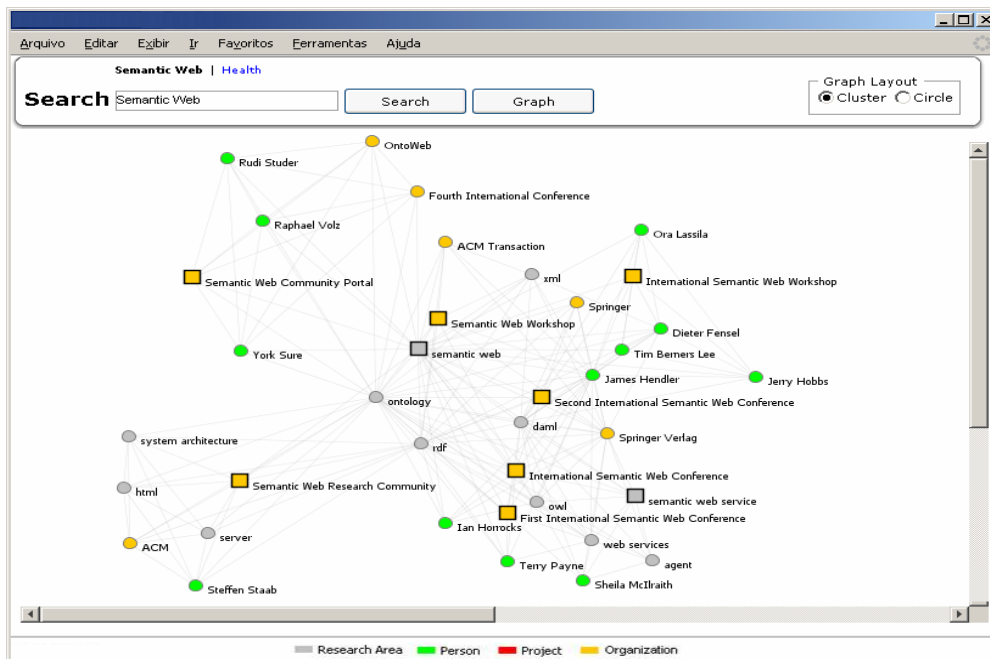


Figure 3: Clusters of entities and their relationships

project resources as in the creation of new business opportunities.

Similarly to communities of practice, expertise location and competency management, have an important role inside organizations, mainly aiming the development of core competencies. According to Dawson [1991], core competencies are the combination of learnt skills or under development. It can foster or help on the establishment of business strategies [Hafeez et al., 2002]. So, methods used in the discovery and evaluation of core competencies have shown to be useful in a

couple of activities toward the management of the organizational intellectual capital.

Aiming to explore and provide ways to understand entities and their relationships we have developed some tools. The Figure 2 shows a tool in which by informing a SE, the most related TEs are retrieved taken into account some classes (e.g.: *organization*, *person*, *project* and *research area*) their relation strength. Also, the tool makes possible the analysis of the most relevant relations for each class, and promotes an easy way to inspect competencies or even to retrieve experts regarding a particular subject (*research area*).

Despite its utility such presentation can be enhanced by showing the same information graphically. We have applied a clustering algorithm to unveil latent relations and to group close entities. In order to facilitate the visualization, different classes are identified by different colors and shapes (squares represent main entities, the cluster centroids, whilst circles represent related entities).

Thus, through these representation ways (textual or graph) it is intended to provide insights to help on managing the organizational knowledge.

## 6. Conclusion and Future Work

We have shown a text mining approach based on co-occurrence in order to establish the relation strength among textual elements (entities, concepts). As result of the process databases representing documents and entities are created and indexed. Thus, through clustering and information retrieval tasks, we intend to unveil latent knowledge and support the decision making process on knowledge management area. Also, we have compared the *Spearman's* correlation between a search engine and other five methods. LRD has shown the best results taken into account different entity classes and window sizes.

Our future work is three-fold. First, we are working on refining proposed method aiming to improve metrics used to establish the relation strength among entities as well as to improve the clustering method. Second, benchmarks with other architectures, similar we have presented here, are intended. Finally, entities and their relations constitute primary resource for network analysis and ontologies. In this sense, improvements on knowledge management applications as those shown here are on the way.

## References

Alani, H., Dasmahapatra, S., O'hara, K. and Shadbolt, N. (2003) "Identifying communities of practice through ontology network analysis", *IEEE Intelligent Systems*, v. 18, n. 2, p. 18-25.

Bontcheva, K., Maynard, D., Tablan, V. and Cunningham, H. (2003) "GATE: A Unicode-based Infrastructure Supporting Multilingual Information Extraction", In *Workshop on Information Extraction for Slavonic and other Central and Eastern European Languages*. Held in conjunction with the 4<sup>th</sup> International Conference "Recent Advances in Natural Language Processing" (RANLP'2003), Bulgaria.

Brin, S. (1998) "Extracting Patterns and Relations from the World Wide Web", In *Proceedings of WebDB* (1998), pages 172-183.

Church, K. and Gale, W. (1991) "Concordances for parallel text", In *Proceedings of the Seventh Annual Conference of the University of Waterloo Centre for the New OED and Text Research*, pages 40-62.

Church, K. and Hanks, P. (1990) "Word association norms, mutual information, and lexicography", *Computational Linguistics*, v. 16, n. 1, p. 22-29.

Ciravegna, F. (2001) "Adaptive Information Extraction from Text by Rule Induction and Generalisation", In *Proceedings of IJCAI* (2001).

Conrad, J. G. and Utt, M. H. (1994) "A System for Discovering Relationships by Feature Extraction from Text Databases", *SIGIR*, p. 260-270.

Cunningham, H. Gate (2001) "A General Architecture for Text Engineering", *Computers and the Humanities*, v. 36, n. 2, p. 223-254.

DARPA (Defense Advanced Research Projects Agency) (1995), In *Proceedings of the Sixth Message Understanding Conference*, Morgan Kaufmann.

Davenport, T. H. and Prusak, L. (1997) "Information ecology: Mastering the information and knowledge environment", Oxford University Press.

Dawson, K. (1991) "Core competency management in R&D organizations", In *Technology Management: The New International Language*, Dunder Kocaoglu and Kiyoshi Niwa (eds.), New York, Institute of Electrical and Electronics Engineers, p. 145-148.

Duda, R. and Hart, P. (1973) "Pattern classification and scene analysis", Wiley, New York.

Gonçalves, A. L., Zhu, J., Song, D., Uren, V. and Pacheco, R. (2006) "LRD: Latent Relation Discovery for Vector Space Expansion and Information Retrieval", In *Proceedings of the 7<sup>th</sup> International Conference on Web-Age Information Management (WAIM 2006)*, J.X. Yu, M. Kitsuregawa, and H.V. Leong (Eds.): Lecture Notes in Computer Science (LNCS), Hong Kong, China, p. 122-133.

Grover, C., Gearailt, D. N., Karkaletsis, V., Farmakiotou, D., Paziienza, M. T. and Vindigni, M. (2002) "Multilingual XML-Based Named Entity Recognition for E-Retail Domains", In *Proceedings of the 3<sup>rd</sup> International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, pages 1060-1067.

Guthrie, L., Pustejowsky, J., Wilks, Y. and Slator, B. M. (1996) "The Role of Lexicons in Natural Language Processing", *Communications of the ACM*, v. 39, n. 1, p. 63-72.

Hafeez, K., Zhang, Y. and Malak, N. (2002) "Identifying core competence", *IEEE Potentials*, v. 49, n. 1, p. 2-8.

Hair Jr., J. F., Anderson, R. E., Tatham, R. L. and Black, W. C. (1998) "Multivariate data analysis". Prentice-Hall, Upper Saddle River, 5. ed., New Jersey.

Halkidi, M., Batistakis, Y. and Vazirgiannis, M. (2001) "On clustering validation techniques",

Journal of Intelligent Information Systems, v. 17, n. 2-3, p. 107-145.

Johnson, R. A. and Wichern, D. W. (1998) Applied multivariate statistical analysis, New Jersey: Prentice-Hall, 4<sup>th</sup> edition.

Lesser, E. L. and Storck, J. (2001) "Communities of practice and organizational performance", IBM Systems Journal, v. 40, n. 4, p. 831-841.

Manning, C. D. and Schütze, H. (1999), Foundations of statistical natural language processing, The MIT Press, Cambridge, Massachusetts.

Marwick, A.D. (2001) "Knowledge management technology". IBM Systems Journal, v. 40, n. 4, p. 814-830.

Mooney, R. J. and Nahm, Un Y. (2005) "Text Mining with Information Extraction". In: Proceedings of the 4<sup>th</sup> International MIDP colloquium, September 2003, Bloemfontein, South Africa, Daelemans, W., du Plessis, T., Snyman, C. and Teck, L. (Eds.), Van Schaik Pub., South Africa, p. 141-160, 2005.

Schreiber, G., Akkermans, H., Anjewierden, A., Hoog, R. de, Shadbolt, N., Velde, W. V. de and Wielinga, B. (2002), Knowledge engineering and management: The CommomKADS Methodology, The MIT Press, 3<sup>rd</sup> edition.

Soderland, S. (1999) "Learning Information Extraction Rules for Semi-Structured and Free Text. Machine Learning", n. 34, v. 1, p. 233-272.

Sokal, R. R. and Rohlf, F. J. (1962) "The Comparison of Dendrograms by Objective Methods", TAXON, v. 11, p. 33-40.

Tonella, P., Ricca, F., Pianta, E., Girardi, C., Di Lucca, G., Fasolino, A. R. and Tramontana, P. (2003) "Evaluation Methods for Web Application Clustering", In *Proceedings of the 5<sup>th</sup> International Workshop on Web Site Evolution*.

Vechtomova, O., Robertson, S. and Jones, S. (2003) "Query expansion with long-span collocates", Information Retrieval, v. 6, n. 2, p. 251-273.

Wenger E. (1998), Communities of practice, learning meaning and identity, Cambridge University Press, Cambridge, MA.

Zhu, J., Gonçalves, A., Uren, V., Motta, E. and Pacheco, R. (2005a) "Mining Web Data for Competency Management". In *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence (WI 2005)*, France.

Zhu, J., Uren, V. and Motta, E. (2005b) "ESpotter: Adaptive Named Entity Recognition for Web Browsing", In: *Proceedings of the 3<sup>rd</sup> Conference on Professional Knowledge Management (WM2005)*, Kaiserslautern, Germany.