Asociación Argentina
de Mecánica Computacional

Mecánica Computacional Vol XXIX, págs. 9425-9438 (artículo completo)
Eduardo Dvorkin, Marcela Goldschmit, Mario Storti (Eds.)
Buenos Aires, Argentina, 15-18 Noviembre 2010

TEXT MINING APPLIED TO ONLINE NEWS

Mauricio Onoda[a], Valeria M. Bastos[a,b], Cristian K. Santos[a], Marcello P. A. Fonseca [a], Victor S. Bursztyn[a], Alexandre G. Evsukoff [a], Nelson F. F. Ebecken [a]

[a]Departamento de Engenharia Civil, COPPE/UFRJ - Centro de Tecnologia - Bloco B - Sala B 101 - Cidade Universitária, Rio de Janeiro - RJ – Brazil, http://www.ntt.eng.br

[b]Departamento de Tecnologia em Análise de Sistemas e de Ciência da Computação - Centro Universitário Estadual da Zona Oeste - UEZO, Rio de Janeiro - RJ – Brazil,

Keywords: Web mining, Business applications, Knowledge discovering

Abstract. This paper describes a work that includes the development and implementation of a practical and efficient methodology to construct a knowledge extraction environment that contemplates the search of information from Portuguese language Web sites. The application has much functionality in text mining, such as similarities and differences identification between pages and sites, content classification and document clustering, which can be applicable to competitive intelligence tools.
The application conception has as origin the exploration evaluation environment of literal informations that still come back toward the availability of a tool that deals with only part of problem. Thinking about the increasing availability of information in the Web, it was possible to elaborate a proposal of an environment that presents these solutions in an integrated form, supplying results analysis, according to the user indication.

1 INTRODUCTION

Nowadays, the Web offers information that is worth gold. It is a tool that is becoming increasingly popular for research and pursuit of knowledge. Furthermore, it allows small and medium businesses to use their content and strong applicant, in favor of opinion polls or even to know their customers and competitors.

Increasingly, major newspapers and magazines publish their news on the internet, serving a large audience mainly in developed countries. With this, companies have at their disposal and with a low-cost access to major media to disseminate information, enabling strategic actions are taken soon.

From this approach it is possible to monitor the dissemination of news in several sites associated with media journalism, through the research of one or more key terms, making a textual database can be exploited using text mining techniques. The results can be manipulated allowing the identification of various types of useful information, such as:

- Identify individuals and companies that are involved in news stories in the media;
- Identify the degree of exposure of any person or company in the news circulating on the Internet;
- List the most common terms used across different news;
- Analyze the content of news, by identifying groups of terms most frequently in the news;
- Monitor the image of a particular customer or product, identifying relationships with virtual communities or blogs;
- Manage image crisis in order to prevent or prepare it in an appropriate and effective for taking decisions or actions that minimize negative impacts on brands and businesses;
- Monitor storm of news groups, through the identification of relevant terms that appear in the news over time;
- Research new concepts review, behaviors or products.

This type of research can be conducted for any subject, person or company, whose relevance is to find out. The important thing is to have a significant number of texts associated with the chosen theme and, thereafter, apply some algorithms on the set of information, enabling monitoring of the dynamics of business, the research of new trends and discovering striking facts.

This paper describes the development and implementation of a methodology to build an environment of knowledge extraction, seeking information on news sites in Portuguese.

2 METHODOLOGY OF KNOWLEDGE EXTRACTION

The methodology applied in this work consists of several processing steps, where each of them draws a set of relevant information, not only for the process that will run in the next step, but also for the analyst who can evaluate each result.

For each method used, there is a specific treatment for the resultant information. But, in all cases, the semantic space construction is necessary, extracting relevant terms that are contributing during the results analysis phase.

The Web site pages are pre-processed, and existing terms in each page are stored in a database, with its stem. Some treatments, as frequencies' calculation are carried through during the site process reading. Selected terms are those on the tags that define the texts in the

pages. Although identified terms as stopword – frequent used terms, such as conjunctions, prepositions and articles, however without semantics relevance - are not considered in the next phase, all are stored in the database.

Linear (TF) frequency forms the space vector representation of each term and is used in the next process applied on the pages or sites contents (Liu, 2001). So, the analyst will be able to identify what is important among the terms used for comparative site, and in documents clustering process.

This methodology also uses the spectral detection of communities as a tool to discover potential relationships in textual documents. Based on Newman´s algorithm (Newman, 2004a), this technique can be applied in a case study represented by a documents´ collection (Santos, 2009a; Santos, 2009b).

2.1 Pre-processing

Data entry module executes the pages pre-processing of the selected site, and is used, basically, for reading the information contained in the site pages indicated by the user. This process is called web crawler (Chang, 2001), where user has the option to select a new URL or read a URL already visited. Obtained information is stored in the database, differentiated for the date and hour (timestamp) of reading the site.

Document pre-processing includes the elimination of “stopwords” and the application of stemming algorithms. Nevertheless, the most important issue is the counting of frequencies that will actually be stored in the BoW.

The *BoW* is a table, of which the lines are related to the documents and the columns are related to the words (terms) that appear in the entire collection. The collection of document is thus represented by the set $D = \{D_i, i = 1...N\}$ and the set of all terms is denoted as $T = \{T_j, j = 1...M\}$.

Although it is generally more interesting to store the *BoW* using special data structures due to dimensions involved, mathematically it can be considered as a $N \times M$ sparse matrix of which an element x_{ij} represents an index that related the term $T_j \in T$ in the document $D_i \in D$. The vector representation $\mathbf{x}_i = (x_{i1}, \ldots, x_{iM})$ is usually adopted in information retrieval, such that classical results of linear algebra can be employed (Michael, 1999). This has been called the vector space representation of documents.

Terms are converted to its canonic form, e.g. verbs in imperative form: “estava”, “estou”, estive” to “estar”. After, terms are reduced to its stem, through the application of "stemming" algorithm adapted for the Portuguese language (Orengo, 2001), that performs significantly better than Portuguese Porter algorithm version (Porter, 1980).

According to (Santos, 2008), the most usual metric to compute indexes in the BoW is the TF-IDF (term frequency–inverse document frequency) index, which is based on information theory and defines the importance of a term in the document set.

The *TF-IDF* index is the product of two factors: the term frequency and the logarithm of the inverse document frequency. The frequency of a term T_j in the document D_i is the number of occurrences of T_j in D_i, divided by the total number of terms in the document:

$$TF(D_i, T_j) = \frac{n_{ij}}{\sum_{k=1}^{M} n_{ik}} \tag{1}$$

where n_{ij} is the number of occurrences of the term T_j in the document D_i.

The second factor is inverse document frequency of the term T_j, computed as:

$$IDF(T_j) = \log\left(\frac{N}{N_j}\right) = \log N - \log N_j \tag{2}$$

where N_j is the number of documents that contains the term T_j at least once (or other pre-defined value).

The *TF-IDF* index, which is actually stored at the *BoW* is computed as:

$$BoW(D_i, T_j) = x_{ij} = TF(D_i, T_j).IDF(T_j) \tag{3}$$

The *TF-IDF* index results in a weighted frequency such that a very frequent term that appears in almost every document will have a low *IDF* while a term that occurs in a few documents will have higher *IDF* value. The composed *TF-IDF* index may be interpreted as the importance of the term T_j to the document D_i.

When the documents in the document set are related to a great number of subjects, the *BoW* computed by the TF-IDF index will generally result in a sparse matrix since many terms do not appear in all documents in the collection.

2.2 Spectral Clustering of Document Networks

Many systems can be represented as networks. That is, a set of nodes joined in pairs by edges. The study of networked systems has experienced particular interest in the last decade (Barabási, 2003; Newman, 2003; Boccaletti et all, 2006). One issue that has received a considerable attention is the identification of the community structure in networks (Newman, 2006a; Newman, 2006b; Newman, 2004a; Newman, 2004b; Leicht, 2007; Karrer, 2007; Palla, 2005; Radicchi, 2004). In this work community structure is used to cluster documents represented as a document network, as described next.

Document networks

The document corpora can be regarded as a complex network, of which the nodes are related to the documents and the edges are weighted by the similarities among them.

The network is formally denoted as a undirected and weighted graph $G = (V, E)$, such that the affinity matrix is symmetric, real valued and its elements represent the similarity between two documents. When constructing similarity graphs the goal is to model the local neighborhood relationships between the data points, in this case, documents.

Several similarities metrics may be used to compute the affinity matrix allowing different results. In this works the Gaussian function is used as similarity metric, such that each element of affinity matrix $\mathbf{A}$ is defined as:

$$a_{ij} = \begin{cases} \exp\left(-\dfrac{\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2}{2\sigma^2}\right), & if \ \ i \neq j \\ 0 & otherwise \end{cases} \tag{4}$$

where σ is a dispersion parameters that controls the spread of the similarity function.

Most of the algorithms in community detection don't apply to weighted networks, such that two nodes are connected if the similarity of the corresponding data points is greater than epsilon (a threshold defined ad-hoc). This approach is often called the ε-neighborhood approach (Luxburg, 2007).

In the next section, algorithms to detect community structures in networks are presented for clustering the document networks.

Community structure in networks

A community structure in a network is defined as a group of vertices that have a high density of edges within them, with a lower density of edges between groups. Formally, for all nodes i in the community C the number of connections node belonging its own community k_i^{in} is larger than k_i^{out}, the number of connections it has to the rest of the network (Radicchi, 2004), such that:

$$k_i^{in} > k_i^{out}, \forall i \in C. \tag{6}$$

Further, they define a community in a weak sense, such that the sum of internal connections is larger than the sum of external ones:

$$\sum_{i \in C} k_i^{in} > \sum_{i \in C} k_i^{out} \tag{7}$$

(Newman, 2004a) defined a quantitative measure to evaluate an assignment of nodes into communities called modularity, that can be used to compare different assignments of nodes into communities quantitatively. The network modularity Q is defined as:

$$Q = \sum_i (e_{ii} - a_i^2) \tag{8}$$

where the index i runs over all communities. The fraction of all links connecting nodes in group i and j is denoted by e_{ij}. Hence, e_{ii} is the fraction of all links lying within group i. The fraction of all links connecting to nodes in group i is denoted by $a_i = \sum_j e_{ij}$. One can interpret a_i^2 as the expected fraction of internal links in group i, if the network was random and the nodes were distributed randomly into the different groups.

If the number of within-community edges is no better than random, then the value $Q = 0$. A value $Q = 1$, which is the maximum, indicate strong community structure. In practice however, values typically fall in the range from about 0.3 to 0.7 (Newman, 2004a).

The modularity matrix approach can be optimized by eigenvalues analysis as described next.

Spectral modularity optimization

The algorithm that has been proposed by Newman for community structure detection (Newman, 2004b) uses a new matrix $\mathbf{B}$ that represents a characteristic matrix for network in terms of its spectral properties, called modularity matrix. This approach is a reformulation of the modularity function Q, presented in (8) and that can be conveniently written in its generalized matrix form as:

$$Q = \frac{1}{4m} \mathbf{s}^{\mathrm{T}} \mathbf{B}^{(g)} \mathbf{s} \tag{9}$$

where $\mathbf{s}$ is the column vector whose elements are the $s_i \in \{-1,1\}$, that represents the elements belonging to group 1 if $s_i = 1$, and to group 2 if $s_i = -1$. The total number of edges in the networks m is defined as:

$$m = \frac{1}{2} \sum_i d_i \tag{10}$$

The network vertices degree d_{ij} is defined as (5), and the real symmetric generalized modularity matrix $\mathbf{B}^{(g)}$ has elements:

$$B_{ij}^{(g)} = B_{ij} - \delta_{ij} \sum_{k \in g} B_{ij} \tag{11}$$

where δ_{ij} is the Kronecker δ-symbol, $k \in g$ represents the k elements belonging to g group, and the traditional modularity matrix $\mathbf{B}$ has its elements defined as:

$$B_{ij} = A_{ij} - \frac{d_i d_j}{2m} \tag{12}$$

This formulation allows maximize the modularity by choosing an appropriate division of the network based on the signs of eigenvector elements related to largest (positive) eigenvalues of $\mathbf{B}^{(g)}$, computed by the eigendecomposition:

$$\mathbf{B}^{(g)} = \mathbf{V} \boldsymbol{\Lambda} \mathbf{V}^{\mathrm{T}} \tag{13}$$

where $\mathbf{V}$ is the orthogonal matrix of the eigenvectors and $\boldsymbol{\Lambda}$ is a real diagonal matrix of the eigenvalues. Not all eigenvalues must be computed such that efficient algorithms can be used.

The elements which sign are positives stay on a cluster and those with negative ones on the other, as seen above. The procedure follows subdividing the network repeatedly and computing the modularity function upon the partitions. If exists no division of the sub network that will increase the modularity of the network, then there is nothing to be gained by dividing the sub network and the procedure must be broken.

This happen when there is no positive eigenvalues to the matrix $\mathbf{B}^{(g)}$, providing the termination check of the subdivision process trough the leading eigenvalue, make the network indivisible.

This spectral approach is very interesting because there is no necessity to known in advance the number k of cluster, once it is determined by the procedure itself, without the necessity of the another algorithm like k-means for example, as happen with another spectral algorithms.

The Newman's algorithm is composed of these steps:

1. Compute the affinity matrix $\mathbf{A}$ like (4);
2. Compute the generalized modularity matrix $\mathbf{B}^{(g)}$ for all network elements like (11);
3. Make the eigendecomposition of the $\mathbf{B}^{(g)}$ and discover the leading eigenvalue;
4. Construct the first division of the network based on signs of the eigenvector related to the leading eigenvalue;
5. Compute modularity Q of the partitioning based on (9);
6. Verify if the leading eigenvalue is positive;
7. In the case of the leading eigenvalue to be positive then go to step 2 and continue the procedure;
8. In the case of the leading eigenvalue to be negative then the procedure must be left alone and to be finished;

3 DEVELOPMENT ENVIRONMENT

The development environment is a system composed of several text mining tools that can generate different types of results such as statistical analysis, content analysis, network analysis and perception analysis, and provide the documents and a data base of terms found , as shown in Figure 1.

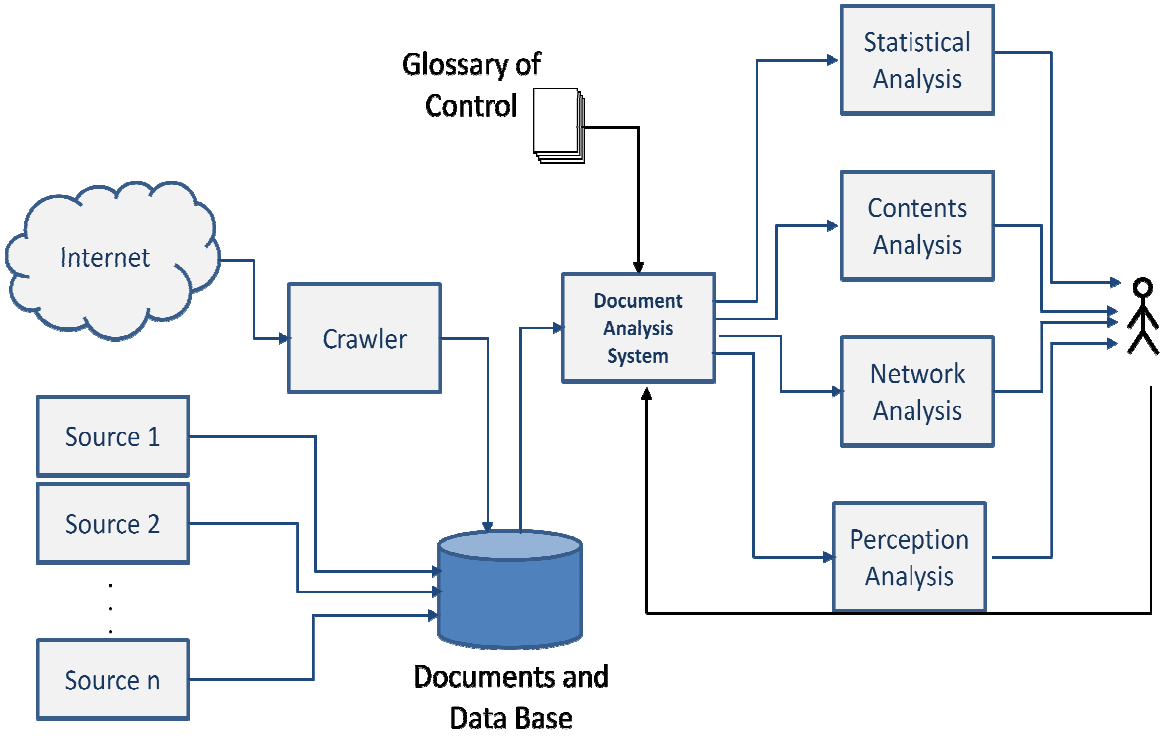

Figure 1: Document Analysis System

The main benefits of the system are:

- Monitoring of words associated with products, brands, people, etc. the media with an automatic evaluation of the perception of these words by the media;
- Monitoring of main issues under discussion in the media monitored containing selected words;
- Analysis of observed relationships among the most frequently words encountered;
- Documents automatic classification for analysis and decision support;
- Automatic clipping.

And the main results can be defined as:

- Statistical Analysis: classical analysis about the most frequent presenting distributions, historical, etc.
- Content Analysis: analysis of subjects stored in the given period of time; classifying new documents according to the content identified, identification and detection of new subjects.
- Network Analysis: assessment and discovery of relationships (links) between interesting entities (companies, people, etc..) Interest, identification of communities.
- Perception Analysis: evaluating potential (positive or negative) of a document (news article, etc..) to the business objectives, according to its content and trial experts previously stored in the system.

From the methodology identified in the previous section, the document analysis system is prepared to perform various activities on the text documents, as shown in the sequence of steps given below:

- Load the data, storing the information in a database;
- Preliminary cleaning of the texts, removing stopwords and correcting some terms with misspellings;
- Total count of documents and terms;
- Counting of documents and terms for a period, if necessary;
- Statistical analysis, generating a list of names, organizations and terms most frequently used;
- Content analysis, generating the array of terms and identifying one or more important terms for the analysis of co-occurrence and checking the level of detail (ε) necessary;
- Network analysis in terms and related documents, identifying potential communities by subject;
- Perception Analysis, by identifying documents that express potential (positive and negative).

4 CASE STUDY

This case study was conducted to evaluate and monitor online news involving Petrobras, the Brazilian oil industry leader and one of the largest integrated energy companies in the world. For this, we collected 8335 news on the internet between January and April/2010. The words used on search engines (crawler) were "oil" and "Petrobras".

The result was a database with 3,129,712 words and 8,018 terms. The total number of words refers to all words found in all documents. The total number of terms refers to the total number of distinct words found. The preprocessing step, performed by programs developed specifically for this purpose, consisted of the exclusion of stopwords and stemming. Like all the news was formal texts, there was no need to use spellcheck or a dictionary. After preprocessing, the number of terms dropped to under 2,406, or 30% of the total.

The study was divided into two parts: a statistical analysis and content analysis. Although the text mining technique described in this paper can be applied only on content analysis, statistical analysis is part of the proposed methodology, since it provides other information, also important in the knowledge extracting process.

As the initial step of statistical analysis, terms that are names of persons and names of organizations were identified in documents. From this, it was generated, respectively, the statistics of occurrence as shown in Figure 2 and Figure 3. In these figures there is the total number of documents containing each term and the total number of occurrence of each term. In the figures, there is a high occurrence of the name of president Luiz Inacio Lula da Silva and current presidential candidate, Dilma Roussef, and Hugo Chavez and Barack Obama. Interestingly, the president of Petrobras, José Sergio Gabrielli appears with less frequency until the US secretary of state Hillary Clinton.

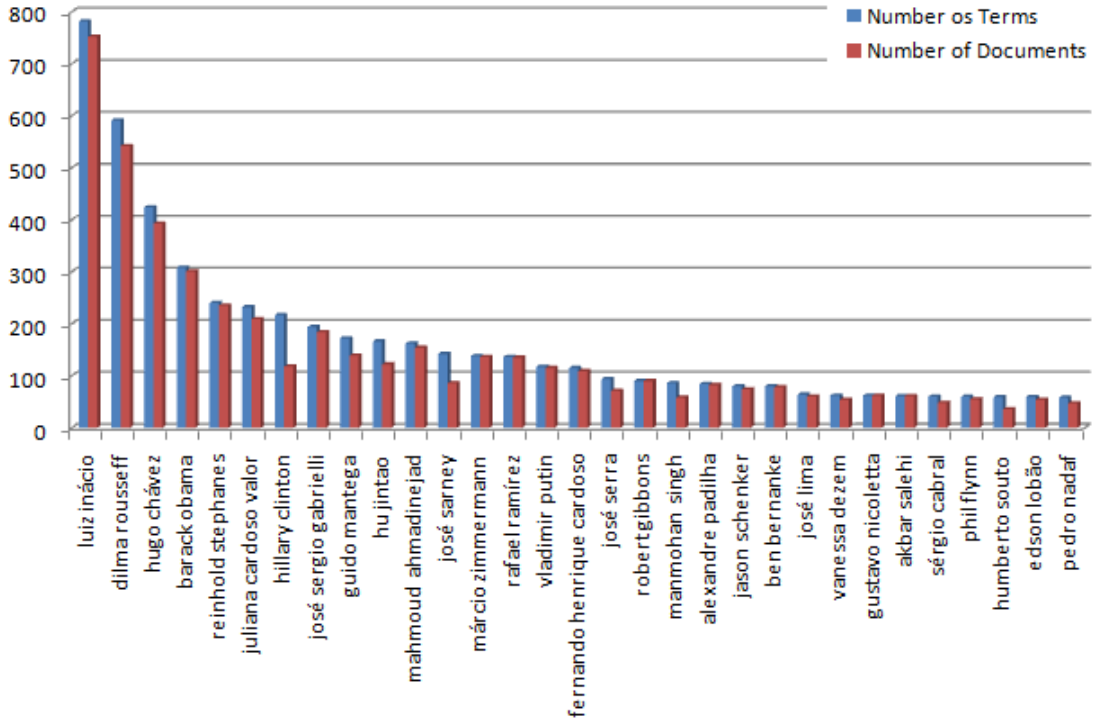

Figure 2: Statistics of personalities

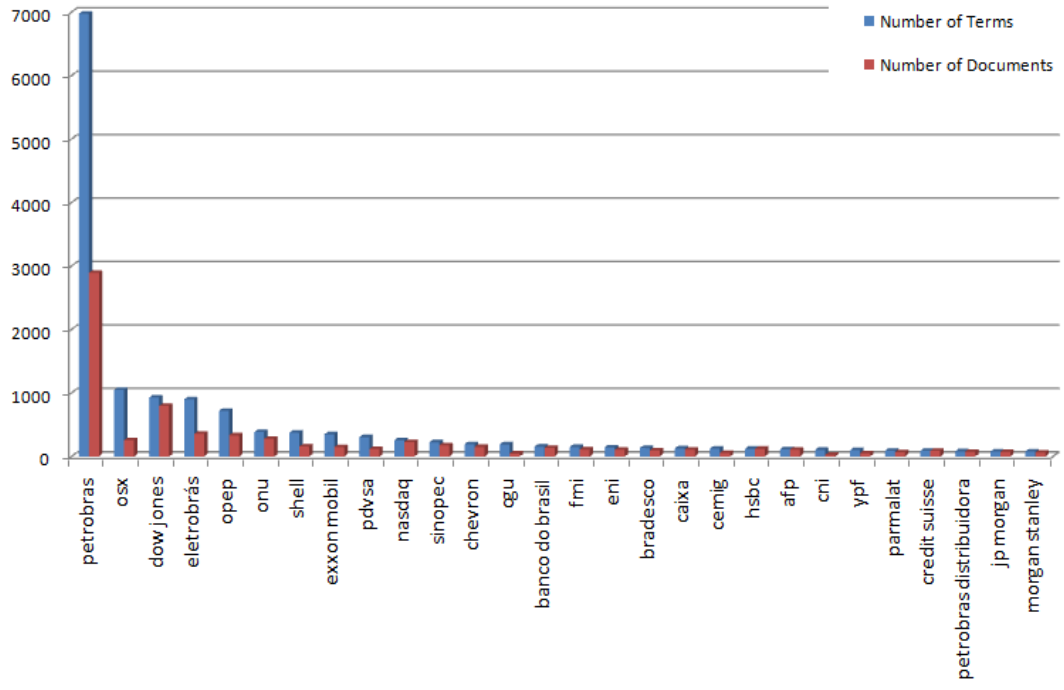

Figure 3: Statistics of organizations

In Figure 3, there is a large disparity between the "Petrobras" with other organizations. In second place with more occurrences appears OSX, but in a reduced number of documents (about 200).

From the occurrences it was possible to analyze the degree of exposure, which is the ratio between the total number of occurrences of a term and the number of documents in which the

Completing the phase of statistical analysis, the graph of co-occurrence of terms (Figure 6) was generated. In this graph, nodes correspond to the most relevant terms and size of each node is proportional to the number of documents in which the term appears. The thickness of the connections between two nodes is proportional to the number of documents that the two words occur together. The graph has been filtered to display only the main connections to avoid a figure very polluted. In this figure, the term "Petrobras" is strongly linked to the term "energy".

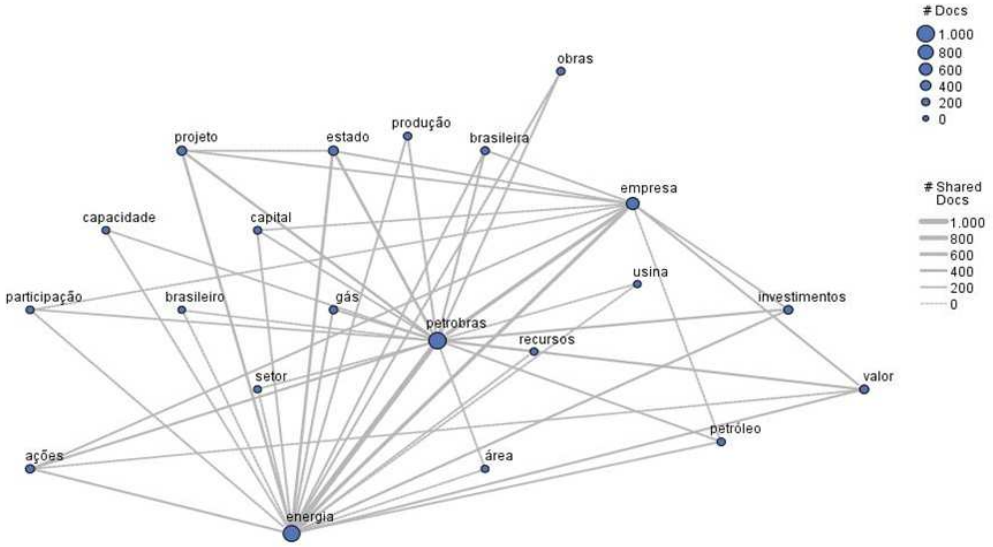

Figure 6: Co-occurrence of terms

The content analysis, the second part of this case study, included the evaluation of the main issues that were addressed in the media during the period analyzed. The proposed methodology allows you to perform a segmentation of the complete collection of documents in more or less related groups that address similar issues. The algorithm used can be adjusted to a more detailed or more general clustering, according to the needs of the analysis. Furthermore, it is possible to partition some of the groups obtained for a more detailed assessment of any particular group. The result of content analysis is presented in a network, where each node represents a document that is positioned in the figure due to its similarity with other documents.

The content analysis conducted in this study resulted in nine groups as shown in Figure 7, generated with the help of program Cytoscape. Each group is identified with a color and the most important terms of each group appear in the tag cloud where the font size is proportional to the importance of the word in the group.

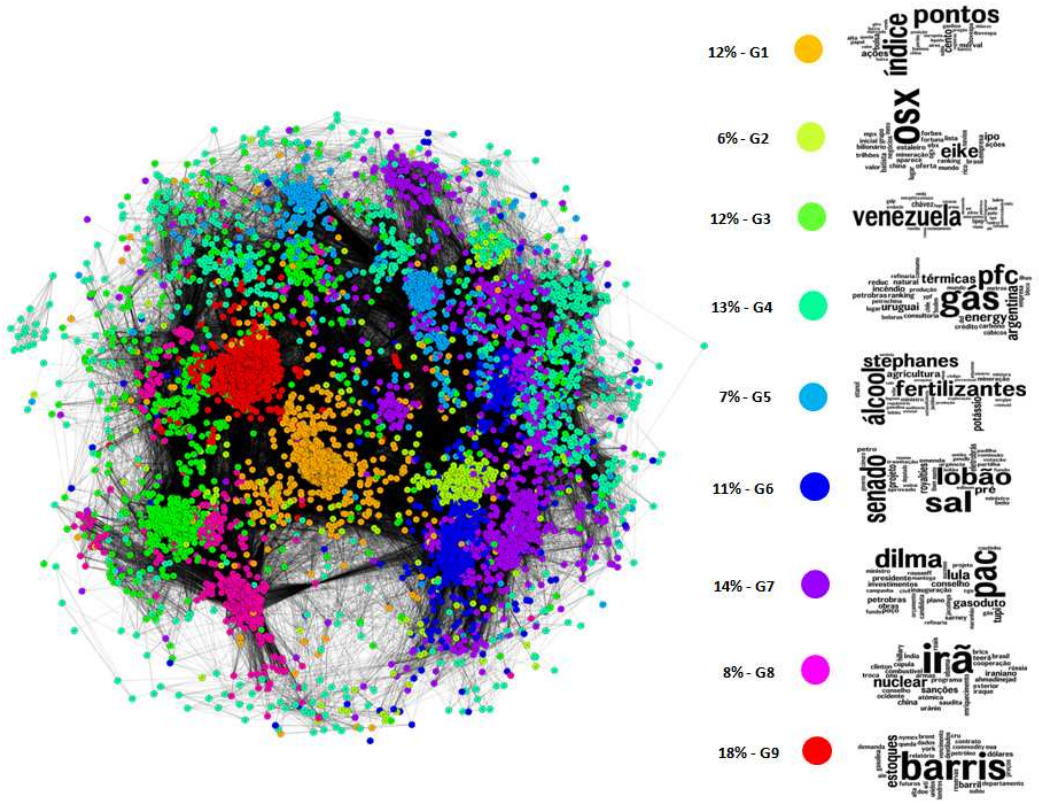

Figure 7: Content Analysis

Group analysis found explains a lot of facts found in the statistics. For example, the group G2 which deals mainly with the launch of the company's shares OSX, which occurred in March/2010. The OSX is a yard to be installed in Santa Catarina by the EBX group of businessman Eike Batista. The G8 group deals primarily with the nuclear deal negotiated with Iran over April/2010 by president Lula. This group may explain the large exposure of Hillary Clinton in this collection.

Besides the tag clouds, the clustering automatically choose the new that best represents each of the generated groups. Figure 8 shows a zoomed tag cloud of group 2 (G2) and their representative new.

A aposta do bilionário Eike Batista no negócio petrolífero do Brasil sofreu um revés, com a sua construtora de navios OSX reduzindo sua oferta inicial de ações (IPO, na sigla em inglês) após investidores mostrarem receio em pagar um preço considerado alto por uma empresa iniciante e ainda sem receita. A OSX, que tem 750 milhões de reais de dívida e que atualmente só tem um navio, reduziu a estimativa do valor total de seu IPO para até 3,31 bilhões de reais. No prospecto da oferta, a OSX disse que planeja usar 89 por cento dos recursos captados para construir plataformas e embarcações para a exploração e produção de petróleo. A maior demanda deve vir da OGX, petrolífera também controlada por Eike. A OSX procura se beneficiar de um esperado aumento na demanda por plataformas de petróleo e outros itens da indústria por parte não somente da OGX, mas também da Petrobras, em meio aos planos de exploração dos novos recursos petrolíferos em águas ultraprofundas do Brasil.

Figure 8: Details of group 2 (G2)

The system found nine major groups of documents in the period, but this number is not definitive and depends on the degree of refinement of the analysis. The segmentation algorithm used can be adjusted to a greater or lesser level of granularity to make it more appropriate to the aims of the analysis.

5 CONCLUSIONS

This paper presented a methodology for analysis of news from Portuguese language Web sites. The differential of this study was considering a set of documents as a complex network, where nodes represent documents and edges are the weights according to the similarities among them. Thus, the cluster analysis of documents was processed as a detection of community structures in complex networks. The great advantage of this approach was the robustness to the high dimensionality of feature space and to the inherent data sparsity resulting from text representation in the vector space model. Furthermore, as a byproduct of the method itself, defines automatically the number of groups, which is a recurring problem in cluster analysis.

As an application, the approach of this work allows obtaining quantitative impact of the brand in electronic media as well as the most important relationships and may reveal a situation that deserves more attention. The methodology presented can be applied in a system of monitoring and support of strategic decisions.

Dynamic analysis of the groups is the main improvement of this methodology in the monitoring process.

REFERENCES

Barabási A.-L., Linked : how everything is connected to everything else and what it means for business, science, and everyday life, Plume, 2003.

Berry, M., Survey of Text Mining: Clustering, Classification, and Retrieval, Springer, 2003.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D.U., Complex networks: Structure and dynamics, *Physics Reports*, pp. 175-308, 2006.

Chang, J., Healey, M. J., McHugh, J. A. M., Wang, J. T. L., Mining the World Wide Web An Information Search Approach. *Kluwer Academic Publishers*, 2001.

Leicht, E.A., and Newman, M.E.J., Community structure in directed networks, *arXiv:0709.4500v1* 2007.

Karrer, B., Levina, E., and Newman, M.E.J., Robustness of community structure in networks, *arXiv:0709.2108v1*, 2007.

Liu, B., Ma, Y., Yu, P. S., 2001. Discovering Unexpected Information from Your Competitors' Web Sites. In *Proceedings of The Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD-2001), pp 144-153, 2001.

Luxburg, U. von, A tutorial on spectral clustering, *Statistics and Computing* 17(4), arXiv:0711.0189v1, pp. 395-416, 2007.

Michael, W. B., Zlatko, D. and Elizabeth, R. J., Matrices, Vector Spaces and Information Retrieval. *SIAM Review 1999*, pp.335-362, 1999.

Newman, M.E.J., The Structure and Function of Complex Networks, *SIAM Review*, pp. 167-256, 2003.

Newman, M.E.J., and Girvan, M., Finding and evaluating community structure in networks, *Phys. Rev.* E 69, 026113, 2004a.

Newman, M.E.J., Fast algorithm for detecting community structure in networks, *Phys. Rev.* E 69, 066133, 2004b.

Newman, M.E.J., Modularity e community structure in networks. *PNAS* 103(23):8577-8582, 2006a.

Newman, M.E.J., Finding community structure in networks using the eigenvectors of matrices, doi: 10.1103/*PhysRev*E.74.036104, 2006b.

Orengo, V. M., Huyck, C., 2001. A Stemming Algorithm for the Portuguese Language. In *Proceedings of Eighth International Symposium on String Processing and Information Retrieval* (SPIRE 2001), pp 186 – 193, 2001.

Palla, G., Derenyi, I., Farkas, I., and Vicsek, T., Uncovering the overlapping community structure of complex networks in nature and society, *Nature* 435, pp. 814-818, 2005.

Porter, M.F., An Algorithm for Suffix Stripping. In *Program***,** vol.14, n. 3, pp. 130-137, 1980.

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., and Parisi, D., Defining and identifying communities in networks, *Proc. Natl. Acad. Sci. USA* 101, 2658-2663, arXiv:cond-mat/0309488v2, 2004.

Santos, C. K. dos; Evsukoff, A. G.; Lima, B. S. L. P. Cluster analysis in document networks. *WIT Transactions on Information and Communication Technologies (Online)*, v. 40, pp. 95-104, 2008.

Santos, C. K. dos; Evsukoff, A. G.; Lima, B. S. L. P. Spectral clustering and community detection in document networks. *WIT Transactions on Information and Communication Technologies (Online)*, v. 42, p. 41-50, 2009a.

Santos, C. K., Evsukoff, A. G., Lima, B. L. S. P., Ebecken, N. F. F., Identificação de Relações Potenciais em Redes Sociais através da Detecção Espectral da Estrutura de Comunidades. *In 30º CILAMCE Congresso Íbero-Latino-Americano de Métodos Computacionais em Engenharia*, Armação de Búzios, Rio de Janeiro, 2009b.