The Cassiopeia Model: A study with other algorithms for attribute selection in text clusterization

Marcus V. C. Guelpeli[1], Ana Cristina Bicharra Garcia[1], António Horta Branco[2]
[1]Departamento de Ciência da Computação
Universidade Federal Fluminense – UFF
Rio de Janeiro, Brasil
[2]Departamento de Informática
Faculdade de Ciências da Universidade de Lisboa-
FACUL,Lisboa, Portugal
{mguelpeli,bicharra}@ic.uff.br, Antonio.Branco@di.fc.ul.pt

***ABSTRACT**: This work proposes a study of algorithms used for attribute selection in text clusterization in the scientific literature and selection of attribute in the Cassiopeia model The most relevant contributions include the use of summarized texts as an entrance in pre-processing stage of clusterization, language independence with the use of stop words and the treatment of high dimensionality, a problem that is inherent to Text Mining. Hence, our intention is to achieve an improvement in the measurement of clusters as well as to solve the problem of high dimensionality.*

Keywords: Text Mining, Knowledge Discovery, Summarization, Clusterization., Attribute Selection and Agglomerative Hierarchical

Received: 11 March 2011, Revised 29 April 2011, Accepted 4 May 2011

1. Introduction

One of the greatest problems when it comes to accessing information is the precise identification of subjects included in a given textual document. This search is normally conducted manually. For human beings, this type of search is fairly straightforward. However, automated systems find this task extremely difficult and computationally costly.

For the automatic recovery of information to work the searches must be conducted so as to approximate natural language as much as possible. Human language that is less deterministic, more flexible and open-ended, offers the user the possibility of formulating complex issues with greater ease, thereby allowing them to locate the most relevant documents. However, language's semantic wealth imposes a fair share of limitations to automated search systems.

This field presents challenges in regards to the enormous amount of information available and there is a need for the development of new means of accessing and manipulation large quantities of textual information. A specific problem in the field is the surplus of information, which in turn is connected to the localization of relevant information, the identification and extraction of knowledge embedded in the important information that was found. After identifying the relevant information, it is clear that it was not found in isolation, but accompanied by a range of other information, or spread out in a number of documents, and, hence, one needs to analyze the content of these pieces of information and filter or extract the data that is truly important.

A field called Knowledge Discovery from Texts – or KDT [6], [22], [21] and [14], is concerned with the process of recovering,

filtering, manipulating and summarizing knowledge that has been extracted from large sources of textual information and then presenting this knowledge to the end user by using a series of resources, which generally differ from the original resources. By employing Text Mining (TM) techniques in the field of KDT, according to [9] we are able to transform large volumes of information – which tend to be unstructured – into useful knowledge that is many times innovative, even for companies that make use of the information. The use of TM allows us to extract knowledge from rough (unstructured) textual information, providing elements that support Knowledge Management, which refers to a method of reorganizing the way in which knowledge is created, used, shared, stored and evaluated. Text Mining in knowledge management takes place in the transformation of content from information repositories to knowledge that can be analyzed and shared by the organization [25]. Text Mining is a field within technological research whose purpose is the search for patterns, trends and regularities in texts written in natural language. It usually refers to the process of extracting interesting and non-trivial information from unstructured texts. In this way, it looks to transform implicit knowledge into explicit knowledge [5]. The TM process was inspired in the Data Mining process, which consists of the "non-trivial extraction of implicit information, previously unknown and potentially useful in data" [7]. It's an interdisciplinary field that encompasses Natural Language Processing, specifically Computational Linguistics, Machine Learning, Information Recovery, Data Mining, Statistics and Information Visualization. For [13], TM is the result of the symbiosis of all these fields. There are many aims when it comes to applying the process of TM: the creation of summaries, clusterization (of texts), language identification, extraction of terms, text categorization, management of electronic mail, management of documents and market research and investigation.

The focus of this work is to use text clusterization, which is a technique that is employed when one does not know the classes of elements in the available domain and, hence, the aim is to automatically divide elements into groups according to a given criterion of affinity or similarity. Clusterization aids in the process of knowledge discovery in texts, facilitating the identification of patterns in the classes [9].

The aim of this work is to compare the Cassiopeia model with other clusterization methods described in the literature, in which the attribute is identified in the pre-processing phase by word frequency and, according to [10], this is the most important phase of clusterization and the one that will determine its success and thereby affect knowledge discovery.

This work is organized as follows. In Section 2, the Cassiopeia model is described. In Section 3, the simulation methodology is explained. Section 4 shows the results obtained in the experiments and Section 5 presents the conclusion and future works.

2. The Cassiopeia Model

The Cassiopeia Model, as illustrated in Figure 1, is comprised of two processes: Summarization and Clusterization.

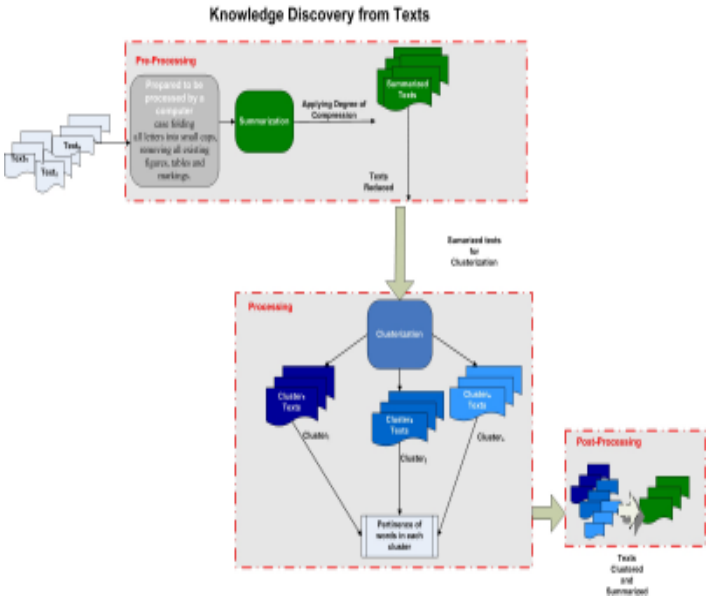

Figure 1. The Cassiopeia Model

The Cassiopeia model starts with text entries for knowledge discovery. These texts go through the pre-processing phase during which they are prepared for the computational process. In this stage, the case folding technique is employed, placing all letters in small caps, as well as other details such as removing all existing figures, tables and markers.

According to [8] and [9] this transforms the texts into a format that is compatible and ready to be processed. The pre-processing stage also includes summarization, the purpose of which, in the Cassiopeia model, is to decrease the number of words for clusterization, which occurs in the processing stage. In this way, the issue of high dimensionality (a problem in the field of TM) is addressed, avoiding the use of the similarity matrix in the processing phase, as well as allowing for the permanence of stopwords, which address the issue of language independence. Not using the similarity matrix and the permanence of stopwords are issues that will be explained in items 2.1 and 2.2.

After concluding the pre-processing phase, the processing stage begins, where the texts are clustered – that is, organized into clusters based on a similarity criterion – as described in detail in item 2.2.

The clusters that are created have a vector of words called cluster centroid which contains words with a high level of relevance to each cluster, that is, words which are pertinent in relation to the clustered texts. With the reclusterization of new texts, which takes place during processing, clusters, sub-clusters or even a fusion of clusters may appear [15]. The word vectors, due to the issue of dimensionality, adopt, according to [26], a similarity threshold, which is yet another important point involved in solving the problem of high dimensionality in TM. The reason for this threshold will be explained in item 2.2., but in case of reclusterization, it can suffer variations until it reaches its stabilization value, that is, the degree of pertinence of each word in each cluster, as shown in Figure 1.

The clusters are organized hierarchically (top-down). Reclusterization occurs up until the moment in which the centroids in each cluster reach stability, that is, when they no longer go through alterations. After the processing phase, it is time to start the post-processing stage, where each one of the text clusters or sub-clusters will contain by similarity a set of summarized texts with a high level of informativity and with the main ideas outlined, which is typical of summarized texts.

2.1 Summarization during pre-processing to decrease dimensionality and improve the measurement of text clusterizations. Summaries are reduced texts that convey the most important and most relevant ideas of the original text in a way that is clear and straightforward without loss of informativity [3].

This need to simplify and summarize occurs due to the increase in the volume of information available in the media and the short amount of time available for reading a wide range of texts [8] and [9]. As a consequence of this process, there is an inability on the part of readers to absorb the content matter of original texts. Hence, the summary is a shortened version of the text whose aim is to grasp the author's main point and convey it in a few sentences to the reader.

The Automatic Summarization (SA) used in the Cassiopeia Model is extractive and follows the empirical approach defined by [18], also known as the superficial approach. This technique uses statistical or superficial methods that identify the most relevant follow-ups of the source-text, producing extracts by juxtaposing the extracted sentences, without any modification in terms of the order of the original text. For this simulation, we selected both professional summarizers and those from the literature. The details regarding these summarizers, as well as their algorithms, will be presented in item 3.

The most relevant point of this work and the main focus of the study is the use of summarization as an integral part of the process of text clusterization, since, as well as the decreasing volume of processing, it confers a significant gains in the mensuration of text clusterizations. This is what this study hopes to show and the results can be observed in item 4.

2.2 Clusterization in the Cassiopeia Model

According to [4], text clusterization is a totally automatic process that separates a collection into groups of texts with similar content. The identification of clusters by way of their characteristics, which is known as cluster analysis, is important in the Cassiopeia model because the texts and clustered by evaluating the similarity between them. This evaluation is described in the three phases below.

The three phases of the Cassiopeia model were proposed by [8], the purpose of which is to cluster documents that have been previously summarized.

First Phase - (Attribute Identification): the characteristics of the words in the text are selected using relative frequency. This defines the importance of a term according to the frequency in which the term is found in the text. The more a term appears, the more important it is found to be in any given text. The relative frequency is calculated using equation (1). This formula normalizes the result of the absolute word frequency, avoiding situations in which shorter documents are represented by small vectors and large documents by large vectors. After normalization, all documents are represented with vectors of the same size

$$F_{real}X = \frac{F_{abs}X}{N} \quad (1)$$

Where $F_{real}X$ is equal to the relative frequency of X, $F_{abs}X$ is equal to the absolute frequency of X, i.e., the number of times X appears in the document, and N is equal to the total number of words in the text.

Each word is considered a vectorial space and, as such, represents one dimension (there are as many dimensions in a text as there are words). In this way, the first step towards handling this problem takes place in summarization, where the space of dimensionality is significantly reduced. It is later dealt with a second time in the first phase of this process, where the characteristic of the words are selected by using relative frequency.

Second Phase - (Similarity Calculation): in this phase, the model identifies the similarity between texts (as per the characteristics selected in the first phase). In order to do this, a measure of fuzzy similarity was used – a measure of set theoretic inclusion [2], which evaluates the presence of words in the two texts being compared. this fuzzy value represents the degree to which an element is included in another, i.e., the degree of equality between them. If the word appears in both texts, a value of one (1) is added to the counter; if it doesn't, then zero (0) is added.

In the end, the degree of similarity is a fuzzy value between 0 and 1, calculated by the average, that is, the total value of the (common) counters divided by the total number of words in both texts (disregarding repetitions). The fact that a word is more important in one text or another, appearing with different frequencies in each text, is not taken into account. This problem can be resolved, in part, by employing another function [17], that takes the average using fuzzy operators, which are reminiscent of the original, but which place different weights on the words.

Thus, the fact that words appear with different degrees of importance in the texts is taken into account. In this case, the weights of each word are based on the relative frequency. The similarity value is calculated by the median of the average weights of the common words. In other words, when a word appears in both documents, the average of their weights is summed together instead of adding the value of one (1). In the end, the average is calculated using the total words in both documents. After the similarity calculation has been conducted, the Cassiopeia algorithm is used to organize the vectors in a decreasing way, as shown below in Figure 2.

1. Establish the average frequency of the words in the document based on the Zipf Curve.

$$\int(k;s;N) = \sum_{n\text{-}1}^{N} \frac{\frac{1}{n^s}}{k^s} \quad (2)$$

Where: *N* is the number of elements; *k* stands for classification; *s* is the value of the exponent that characterizes the distribution

2. Choose the 25 words to the left of the average and the 25 words to the right of it.

Third Phase - (Agglomerative Hierarchical Method): This last phase employs the Agglomerative Hierarchical method that, by analyzing constructed dendograms, defines the previous number of clusters. The Clicks algorithm is used to identify the text clusters by setting some sort of relationship rule that will creat clusters based on the similarity analysis of the terms in the text. In this way, according to [9], the Clicks algorithm is able to construct more cohesive clusters. The employment of the summarization module and then these three phases of the clustering module ensures that the Cassiopeia module is able to go without the use of the similarity matrix (which is a critical point of the high dimensionality within the field of TM, since the similarity matrix grows exponentially in relation to its text base[24].

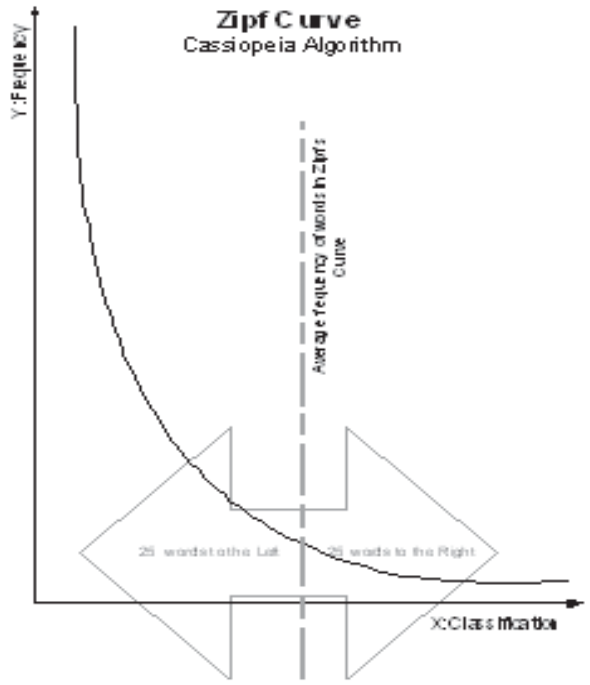

Figure 2. Selection of Attribute in the Cassiopeia model

3. Simulation Methodology

3.1 Corpus

This experiment included as a corpus original texts, meaning texts that had not been summarized, in both Portuguese and English. In Portuguese, there are texts from the journalistic, legal and medical domains, totaling 300 original texts, 100 for each domain. In the legal domain, the articles were extracted from the website (www.direitonet.com.br). The texts were extracted from nine categories, as classified by the website itself: Environmental, Civil, Constitutional, Consumer, Family, Criminal, Welfare, Procedural and Labor. Texts belonging to the medical domain were obtained from the database Scientific Electronic Library Online - SciELO Brasil, on the website (www.scielo.br) and they were extracted from ten categories classified by the website: Cardiology, Dermatology, Epidemiology, Geriatrics, Gynecology, Hematology, Neurology, Oncology, Orthopedics and Pediatrics. Texts from both the legal and the medical domain are scientific and were extracted from specialized databases. For the journalism domain, we used the TeMário 2004 corpus [19], which has texts originating from the online newspaper Folha de São Paulo and spanning the five sections of the paper: Special, International, World, Opinion and Politics.

For texts in the English language, there was also a variation in the domains, but in this case, we selected 200 original texts from the journalistic and medical domains. We did not locate any legal texts that complied with our established criteria which was that they be free of charge. The journalism texts were extracted from the news agency Reuters (www.reuters.com) for the period ranging from April 27 to April 30 2010 and they fall into ten different categories: Economy, Entertainment, G-20, Green Business, Health, Housing Market, Politics, Science, Sports and Technology. The medical texts used were taken from the Scientific Electronic Library Online – SciELO, from their webiste (www.scielo.org). These texts were published between April 9 and 17 2001 and fit into ten categories, as classified on the website: Cardiology, Dermatology, Epidemiology, Geriatrics, Gynecology, Hematology, Neurology, Oncology, Orthopedics and Pediatrics Cardiologics.

3.2 Summarizers

The criteria used to select the summarization algorithms in these experiments were defined as those that could define compression percentages per word; in order for this to occur, they should be able to perform compressions of 50%, 70%, 80 and 90%.

For the summarization process in Portuguese, three summarizers described in the literature were used: Supor [16] which

selects sentences that include the most frequent words from the source-text to compose the extract; Gist_Average_Keyword [20], in which the punctuation of the sentences can occur by one of two simples statistical methods, the keywords method or the average keywords method; and Gist_Intrasentença [20], which is applied to all sentences in by excluding the stopwords.

For the summarization process in the English texts, three summarizers were use, two of which are professional and one from the literature that can be found on the internet. Copernic and IntellexerSummarizerPro are professional summarizers whose algorithms are considered black-box. SewSum [12], the summarizer from the literature, uses a language-specific lexicon to map the inflected forms of words in the content to their respective roots.

3.3 Metrics

The process of clustering by similarity is, by definition, a non-supervised processed and, as such, there are no predetermined classes or examples that can indicate the characteristics of the data set.

According to [11], the evaluation of clusterization can be distributed in three major categories: External or Supervised Metrics, Internal or Non-Supervised Metrics and Relative Metrics, which will not feature in this work.

For the supervised or external metrics, the results of clusterization are evaluated using a structure of predetermined classes that reflect the opinion of a human specialist. For this type of metric, measures such as Precision, Recall and F-Measure are used [23].

In the case of non-supervised or internal metrics, to evaluate the results one uses only information contained in the generated groups; in other words, no external information is used. The most common measures used to achieve this, according to [23] and [1], are cohesion, coupling and silhouette coefficient, a harmonic measure of these two metrics.

With the purpose of validating the results in this experiment, both external and internal measures will be used. The measures were defined as follows:

3.3.1 *External Metrics*

$$Recall(R): \frac{tlcd}{tgcd} * 100 \quad (3)$$

Where *tlcd* is the local total of the dominant category of *cluster i* and *tgcd* is the global total of the dominant category of *cluster i* in the process.

$$Precision(P): \frac{tlcd}{te} * 100 \quad (4)$$

Where *tlcd* is the local total of the dominant category of *cluster i* and *te* is the total number of elements in *cluster i*.

$$F\text{-}Measure(P): 2 * \frac{Precision * Recall}{Precision + Recall} \quad (5)$$

3.3.2 Internal Metrics:

$$Cohesion(C): \frac{\sum_{i>j} sim(P_i, P_j)}{n(n\text{-}1)/2} \quad (6)$$

Where *n* is the number of texts in the cluster *P*, Sim is the similarity calculation and each P_i is a member of the cluster *P.*

$$Coupling\ (A): \frac{\sum_{i>j} sim(C_i, C_j)}{n_a(n_a\text{-}1)/2} \quad (7)$$

Where *C* is the centroid of a given cluster present in *P*, *Sim* is the similarity calculation and n_a is the number of clusters present in *P*.

$$Silhouette\ Coefficient\ (S): \frac{b(i) - a(i)}{max(a(i), b(i))} \quad (8)$$

Where *a(i)* is the average distance between the *i-eth* element of the group and the other elements belonging to the same group. *B(i)* the lowest possible distance between the *i-eth* element of the group and any other group that does not contain the element. The Silhuette Coefficient of a group is the arithmatic average of the coefficients calculated for each element belonging the group, which is shown in

$$\overline{s} = \frac{1}{N}\sum_{i=1}^{N} S \qquad (9)$$

The value of S is located in range between 0 and 1.

3.3.3 Attribute Identification Method

This is where the characteristics of the words in the text are selected using methods that have received the most attention in works related to the field of non-supervised attribute selection in textual documents. They are described below.

3.3.3.1 Ranking by Term Frequency- (RTF)

Ranking by frequency uses the concept of TF as scoring measure for a given attribute, giving more value to the term that appears most frequently throughout the entire collection. This count is usually normalized to avoid a bias towards longer documents so as to place a measure of importance of *i* whiten the given document *d j*. Hence, one gets the term frequency, defined as follows:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \qquad (10)$$

$N_{i,j}$ is the number of occurences of the term that is under consideration (*t i*) in the document *d j*, and the denominator is the sum of the number of occurences of all the terms in document *d j* in other words, the size of the document | *d j* |.

3.3.3.2 Ranking by Document Frequency- (RDF)

This method calculates the number of documents in which the terms appear. It takes into account the fact that the terms that appear in few documents are not relevant to the collection and therefore can be ignored. Formally, this can be obtained as follows:

$$w_{t,d} = (1+ \log tf_{t,d}).\log_{10}(\frac{N}{df_t}) \qquad (11)$$

- *dft* is inverse measure of the informativity of t.
- $dft \leq N$.
- *idf* (inverse document frequency) of *t*.
- We use log (*N/dft*) instead of *N/dft* to soften the effect of the *idf*.

3.3.3.3 Inverse Document Frequency- (TFIDF)

$$idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|} \qquad (12)$$

- |D|: represents the total number of documents in the corpus or collection;
- $|\{d : t_i \in d\}|$: Number of documents in which the term t appears i which is $n_{i,j} \neq 0$ If the term is not in the corpus, this will lead to a division by zero. Hence, the common usage is $1+|\{d : t_i \in d\}|$.

To measure the importance of a term *i* in a document *j*, the following calculation is used: $tf\text{-}idf_{i,j}$ where $tf\text{-}idf_{i,j} = tf_{i,j} * idf_i$.[19].

4. Results Obtained in the Experiments

Due to the very large volume of experiment results obtained through measures proposed in item 3, harmonic measures will be shown in external measures (the *F-Measure*) and in internal measures (the *Coeficiente Silhouette*). It is worth noting that, although they were not shown in this article, results were generated for all measures, in Portuguese and English, for all domains – journalistic, legal, and medical – and for different compression levels of 50%, 70%, 80% and 90%.

As in all corpus simulations in Portuguese and English, three summarizers were used (Gist_Keyword, Gist_Intra and Supor) for the Portuguese language (Copernic, Intellexer and SweSum), for the English language, and four compressions (50%, 70%, 80% and 90%). There are thus comparison possibilities in twelve possible results for each language and in each domain.

Figure 3 shows text clustering results in the corpus in Portuguese, using the Cassiopeia model (with and without stopwords) and the literature methods RDF, RTF and TFIDF. Summarizers were used in Portuguese: Gist_Keyword, Gist_Intra and Supor, with 50%, 70%, 80% and 90% compression, in the journalistic, legal and medical domains. Results are composed of the average sum of the averages obtained from each summarizer, using the harmonic F-Measure throughout the 100 interactions.

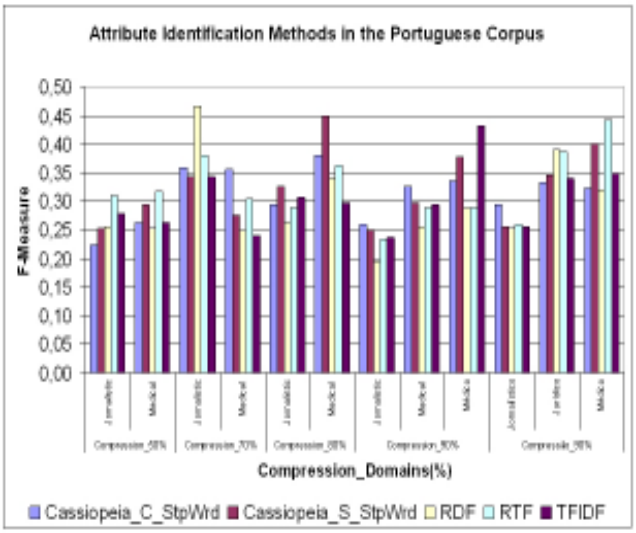

Figure 3. Text clustering values in the Corpus in Portuguese with internal measure using the harmonic FMeasure

In Figure 4, it is possible to observe improved performance of the Cassiopeia model (with stopwords) in text clusters compared with the Cassiopeia model (without stopwords) and literature methods. With 50% of all samples, the Cassiopeia model (with stopwords) reached this percentage for different summarizers and compression levels, using the FMeasure.

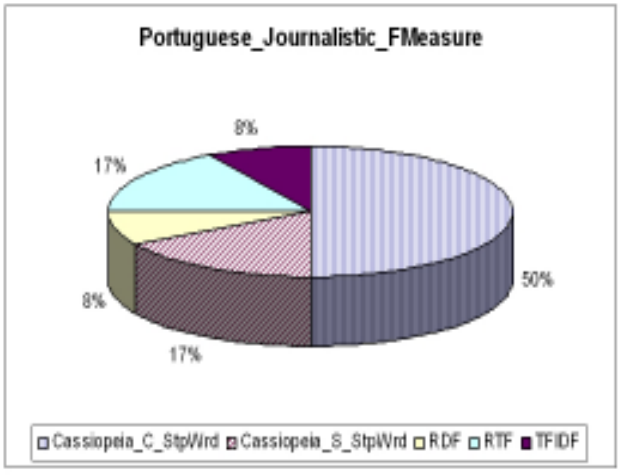

Figure 4. Percentages obtained by the Cassiopeia model and the RDF, RTF and TFIDF models in text clusters in the Corpus in Portuguese in the journalistic domain with the harmonic FMeasure

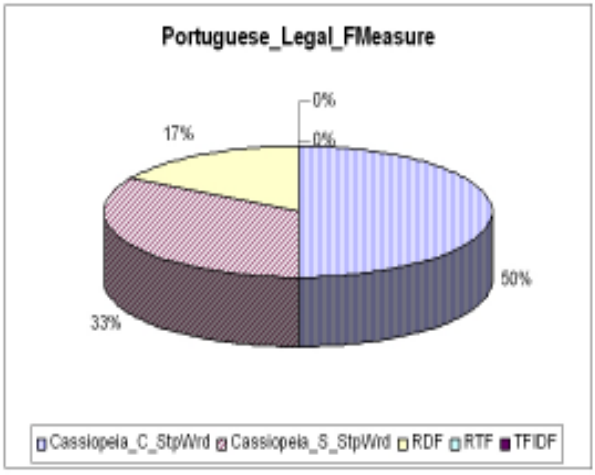

Figure 5. Percentages obtained by the Cassiopeia model and the RDF, RTF and TFIDF models in text clusters in the Corpus in Portuguese in the legal domain with the harmonic FMeasure

In Figure 5, it is possible to observe improved performance of the Cassiopeia model (with stopwords) in text clusters compared

with the Cassiopeia model (without stopwords) and literature methods. With 50% of all samples, the Cassiopeia model (with stopwords) reached this percentage for different summarizers and compression levels, using the FMeasure.

In Figure 6, it is possible to observe improved performance of the RTF method in text clusters compared with the Cassiopeia model (with and without stopwords) and literature methods. With 33% of all samples, RTF method reached this percentage for different summarizers and compression levels, using the FMeasure.

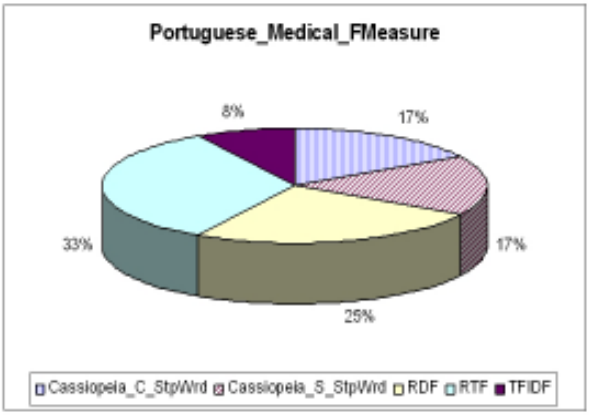

Figure 6. Percentages obtained by the Cassiopeia model and the RDF, RTF and TFIDF models in text clusters in the Corpus in Portuguese in the medical domain with the harmonic FMeasure

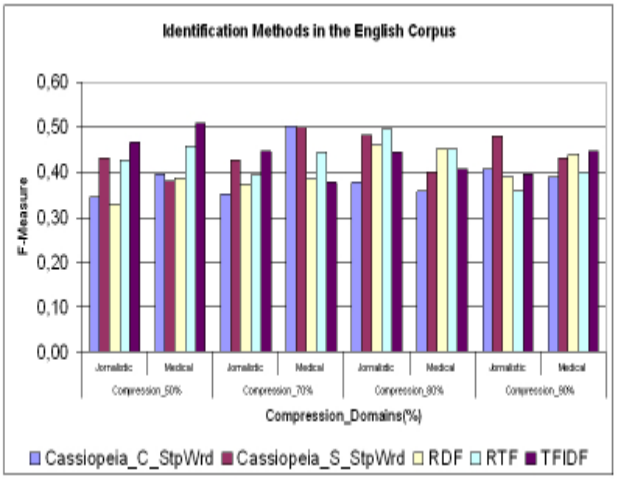

Figure 7. Text clustering values in the Corpus in English with internal measure using the harmonic FMeasure

Figure 7 shows text clustering results in the corpus in English, using the Cassiopeia model (with and without stopwords) and the literature methods RDF, RTF and TFIDF. Summarizers were used in English: Copernic, Intellexer and SweSum, with 50%, 70%, 80% and 90% compression, in the journalistic and medical domains. Results are composed of the average sum of the averages obtained from each summarizer, using the harmonic FMeasure throughout the 100 interactions.

In Figure 8, it is possible to observe improved performance of the TFIDF method in text clusters compared with the Cassiopeia model (with and without stopwords) and literature methods. With 42% of all samples, TFIDF method reached this percentage for different summarizers and compression levels, using the FMeasure.

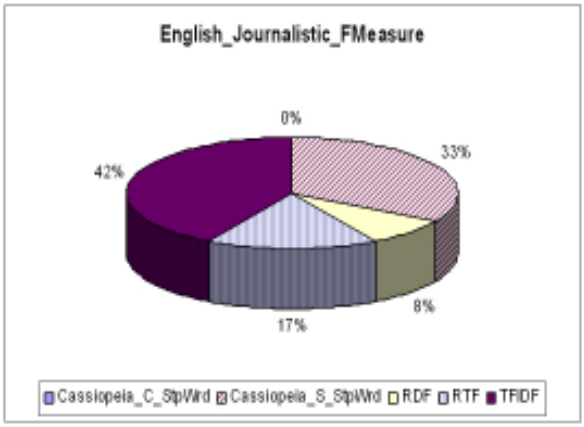

Figure 8. Percentages obtained by the Cassiopeia model and the RDF, RTF and TFIDF models in text clusters in the Corpus in English in the journalistic domain with the harmonic FMeasure

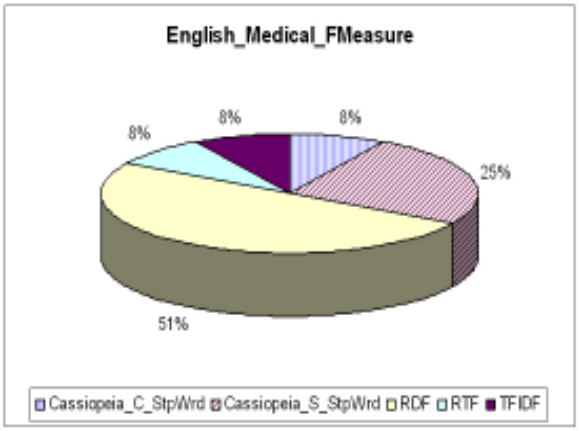

Figure 9. Percentages obtained by the Cassiopeia model and the RDF, RTF and TFIDF models in text clusters in the Corpus in English in the medical domain with the harmonic FMeasure

In Figure 9, it is possible to observe improved performance of the RDF method in text clusters compared with the Cassiopeia model (with and without stopwords) and literature methods. With 51% of all samples, RDF method reached this percentage for different

summarizers and compression levels, using the FMeasure Figures 10 and 11 show the text corpus cluster in Portuguese and English, respectively, using the Cassiopeia model (with and without stopwords) and literature methods RDF, RTF and TFIDF using summarizers Gist_Keyword, Gist_Intra and Supor in Portuguese and summarizers Copernic, Intellexer and SweSum in English, with 50%, 70%, 80% and 90% compression, in journalistic, legal and medical domains in Portuguese, and journalistic and medical domains in English. Results are the averages of the average sums obtained in each summarizer, using the harmonic Silhouette Coefficient method throughout the 100 interactions.

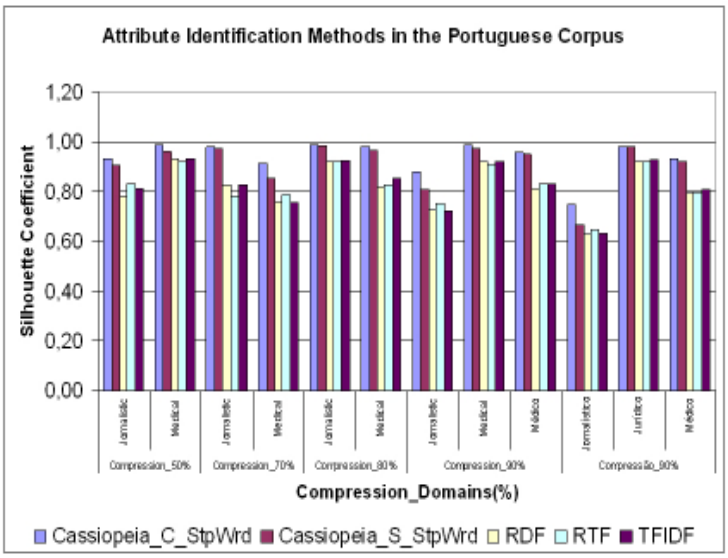

Figure 10. Text clustering values in the Corpus in Portuguese with external measure using the harmonic Silhouette Coefficient

In Figures 10 and 11 reveals a prevalence of 100% in all the samples of the best text clustering of the Cassiopeia model (with stopwords) in all languages, in each domain and in all compressions, compared with the Cassiopeia model (without stopwords) and literature methods RDF, RTF and TFIDF

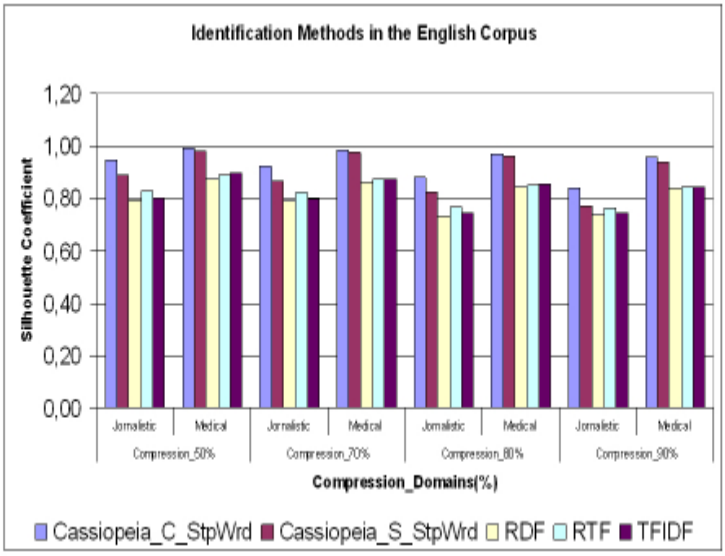

Figure 11. Text clustering values in the Corpus in English with external measure using the harmonic Silhouette Coefficient

5. Conclusion

When evaluating the results of external measures, we can observe, in Figures 3 and 7, the use of the F-Measure, which is a

harmonic measure of Recall and Precision. Good performance is observed in Portuguese language, as shown in Figure 3 and highlighted in Figures 4 and 5, where the Cassiopeia model obtained 50% of the sample with the best text clustering result. In Figure 6, equilibrium of the Cassiopeia model is observed, since, when added, the Cassiopeia model with stopwords and without stopwords obtained 34% of the whole sample but, in nominal values, the Cassiopeia model reached only 17%. In case of Figure 7, in the English language, we can observe that the Cassiopeia model did not obtain the best result. However, in Figures 8 and 9, we can observe that, in nominal values, the Cassiopeia model is just below the best value. In Figure 8, the sum of the Cassiopeia models reach 50%, while the best cluster value of TFIDF method is 42%.

With the results of internal measures, Figures 10 and 11 show the use of the Silhouette Coefficient measure, which is a harmonic measure of Cohesion and Linkage. In this case, we can see absolute predominance of the Cassiopeia method with stopwords as revealing the largest text clustering value among all samples, both for Figure 10 and Figure 11. This result was so expressive that the second best value in all the sample was the Cassiopeia model without stopwords.

Because Cassiopeia is an unsupervised model, these results were very significant since there was absolute prevalence in the internal measure that, as explained in item 3.3, uses only information contained in clusters generated to conduct result assessment. In other words, they do not use external information.

5.1.1 Future Works

A future possibility, or proposal, for the Cassiopeia model would be the inclusion of an autonomous learning module. We believe the inclusion of such a module would lead to even more significant results for the cohesion and coupling metrics.

Another factor that deserves future attention is the issue of post-processing in the Cassiopeia model. As the coupling indexes are highly estimated and the indexed words have a strong correlation with the texts in that cluster, it would be interesting to employ a technique to transform these words into categories and thereby further improve knowledge discovery in texts.

Reference

[1] Aranganayagil, S., Thangavel, K. (2007). Clustering Categorial Data Using Silhouette Coefficient as a Relocating Measure. *In*: International Conference on Computational Intelligence and Multimedia Applications, ICCIMA, Sivakasi, Índia, *In*: Proceedings. Los Alamitos: IEEE, p.13-17.

[2] Cross, V. (1994). Fuzzy information retrieval, *Journal of Intelligent InformationSystems*, Boston, 3 (1) 29-56.

[3] Delgado, C. H., Vianna, C. E., Guelpeli, M. V. C. (2010). Comparando sumários de referência humano com extratos ideais no processo de avaliação de sumários extrativos. *In*: IADIS Ibero-Americana WWW/Internet 2010, Algarve, Portugal. p. 293-303.

[4] Fan, W., Wallace L. , Rich S., Zhang Z. (2006). Tapping the power of text mining, Communications of the ACM, V. 49.

[5] Fayyad, U., e Uthurusamy, R. (1999). Data mining and knowledge discovery in databases: Introduction to the special issue. Communications of the ACM, 39(11), November.Usama Fayyad. 1997. Editorial. Data Mining and Knowledge Discovery.

[6] Feldman, R. , Hirsh, H. (1997). Exploiting background information in knowledge discovery from text, *Journal of Intelligent Information Systems*, 9 (1) Julho/Agosto de.

[7] Frawley, W.J., Piatestsky, S.G., Matheus, C. (1992). Knowledge discovery in data bases: an overview, *AI Magazine*. Fall, p. 57-70.

[8] Guelpeli, M. V. C., Branco H. A., Garcia, A. C. B., (2009). CASSIOPEIA:A Model Based on Summarization and Clusterization used for Knowledge Discovery in Textual Bases. *In:* Proceedings of the IEEE NLP-Ke'2009 - IEEE International Conference on Natural Language Processing and Knowledge Engineering, Dalian.

[9] Guelpeli, M. V. C., Bernardini, F. C., Garcia, A. C. B. (2010). An Analysis of Constructed Categories for Textual Classification using Fuzzy Similarity and Agglomerative Hierarchical Methods, Emergent Web Intelligence: Advanced Semantic Technologies Series: Advanced Information and Knowledge Processing Edition., 2010, XVI, 544 p. 178 illus., Hardcover.

[10] Guyon, S., Gunn, M., Nikravesh, Zadeh, L. A. (2006). Editors, Feature Extraction: Foundations and Applications, Studies in Fuzziness and Soft Computing ; 207, p.137–165. Springer.

[11] Halkidi, M., Batistakis, Y., Varzirgiannis, M. (2001). On clustering validation techniques, *Journal of Intelligent Information Systems*, 17 (2-3) 107, 145.

[12] Hassel, M. (2007). Resource Lean and Portable Automatic Text Summarization, PhD-Thesis, School of Computer Science and Communication, KTH.

[13] Hearst, M.A. (1998). Automated discovery of wordnet relations. *In*: Christiane Fellbaum, editor, WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA.

[14] Keogh, E., Kasetty, S. (2002). On the need for time series data mining benchmarks: a survey and empirical demonstration, In: ACM SIGKDD, Edmonton, Canada, p.102-111.
[15] Loh, S. (2001). Abordagem Baseada em Conceitos para Descoberta de Conhecimento em Textos Porto Alegre: UFRGS.
[16] Módolo, M. (2003). SuPor: um Ambiente para a Exploração de Métodos Extrativos para a Sumarização Automática de Textos em Português, Dissertação de Mestrado. Departamento de Computação, UFSCar. São Carlos - SP.
[17] Oliveira, H. M. (1996). Seleção de entes complexos usando lógica difusa, Dissertação (Mestrado em Ciência da Computação – Instituto de Informática, PUC-RS, Porto Alegre.
[18] Pardo, T.A.S., Rino, L.H.M., and NunesS, M.G.V. (2003). GistSumm: A Summarization Tool Based on a New Extractive Metho. *In*: The Proceedings of the 6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken. Faro, Portugal.
[19] Pardo, T.A.S., Rino, L.H.M. (2004). TeMário: Um Corpus para Sumarização Automática de Textos, Relatórios Técnicos (NILC-TR-03-09). NILC – ICMC – USP. São Carlos, Brasil.
[20] Pardo, T.A.S. (2006). Estrutura textual e multiplicidade de tópicos na sumarização automática: o caso do sistema GistSumm Série de Relatórios do NILC. NILC-TR-10-06.
[21] Pottenger, W. M., Yang, T. (2001). Detecting emerging concepts in textual data mining, *In*: Michael Berry (ed.), Computational Information Retrieval, SIAM, Philadelphia, August.
[22] Tan, A. H. (1999). Text mining: the state of the art and the challenges. *In*: WORKSHOP ON KNOWLEDGE DISCOVERY FROM ADVANCED DATABASES, Proceedings. Heidelberg, p.65-70. (Lecture Notes in Computer Science, 1574.
[23] Tan, P. N., Steinbach, M., Kumar, V. (2006). Introduction to Data Mining. Addison-Wesley. 23.
[24] Vianna, D. S.(2004). Heurísticas híbridas para o problema da logenia, Tese de doutorado, Pontifícia Universidade Católica - PUC, Rio de Janeiro, Brasil.
[25] Velickov, S. (2004). TextMiner theoretical background. Acess *In*: <http://www.delft-cluster.nl/textminer/theory/>. 15 Mar 2011.
[26] Wives, L.K. (2002). Tecnologias de Descoberta de Conhecimento em Textos Aplicadas à Inteligência Competitiva. Exame de Qualificação- EQ-069 PPGC-UFRGS.