The Cassiopeia Model:

Using summarization and clusterization for semantic knowledge management

Marcus V. C. Guelpeli Ana Cristina Bicharra Garcia
Departamento de Ciência da Computação
Universidade Federal Fluminense – UFF
Rio de Janeiro, Brasil
{mguelpeli,bicharra}@ic.uff.br

António Horta Branco
Departamento de Informática
Faculdade de Ciências da Universidade de Lisboa-FACUL,Lisboa, Portugal
Antonio.Branco@di.fc.ul.pt

***Abstract*— This work proposes a comparative study of algorithms used for attribute selection in text clusterization in the scientific literature with the Cassiopeia algorithm. The aim of the Cassiopeia model is to allow for knowledge Discovery in textual bases in distinct and/or antagonistic domains using both Summarization and Clusterizations as part of the process of obtaining this knowledge. Hence, our intention is to achieve an improvement in the measurement of clusters as well as to solve the problem of high dimensionality in the knowledge discovery of textual bases.**

Keywords- Text mining; Knowledge Discovery; Summarization and Clusterization.

I. 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 [7], [29], [26], [16] and [20] – 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 [20], 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 [31]. 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 [6] and [19]. 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" [8]. It's an interdisciplinary field that encompasses Natural Language Processing, specifically Computational Linguistics, Machine Learning, Information Recovery, Data Mining, Statistics and Information Visualization. For [15], 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 [11].

The aim of this work is to compare the Cassiopeia method 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 [12], 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.

II. The Cassiopeia Model

The Cassiopeia model illustrated in Figure 1 starts with text entry for knowledge discovery. These texts undergo the pre-processing phase, where they are prepared for the computational process, i.e., the case folding technique is employed, in which all letters are transformed into lower case, as well as other procedures, such as removing all existing pictures, tables and markings. The text thus presents a compatible format for being processed.

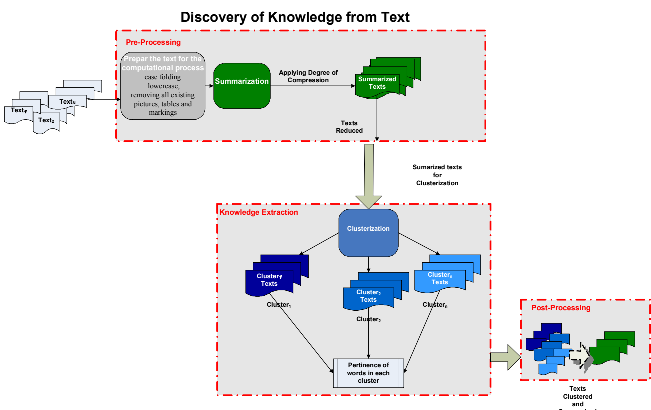

Figure 1. Cassiopeia Model.

Still in the pre-processing stage, the summarization process is employed with the aim of decreasing the number of words for the clustering process which occurs in the knowledge extraction stage. It thus renders high dimensionality and sparse data viable (a problem in the TM field) by not employing the similarity matrix shown in Figure 2 in the knowledge extraction stage. Moreover, it makes it possible to keep stopwords, thus providing language independence. Explanations regarding the lack of use of the similarity matrix and the keeping of stopwords are detailed in items II.A and II.B.

After the pre-processing phase is finalized, the Cassiopeia model begins the knowledge extraction stage, which employs the text clustering process by similarity. This process is detailed in item II.B.

The clusters then created possess a vector of words called cluster centroids, whose words are highly relevant for each cluster and where these are pertinent in relation to the clustered texts. After reclustering new texts, which occurs in the knowledge extraction stage, other clusters, subclusters or a fusion of these clusters may emerge [17]. According to [34], due to dimensionality the word vectors adopt a threshold that is also another important point for solving the problem of high dimensionality and sparse data in TM. However, due to reclustering, it can suffer alterations before arriving at its stabilization value, i.e., the degree of pertinence of each word in each cluster, as shown by Figure 1. The reason for this threshold is explained in item II.B.

These clusters are hierarchically classified top-down. Reclustering occurs until the moment the centroids of each cluster become stable, i.e., they do not undergo further alterations.

As soon as the knowledge extraction stage is over, the post-processing stage begins. In this stage, each one of the clusters or subclusters of texts in the Cassiopeia model contains, according to similarity, a set of summarized texts with a high degree of information content and with main ideas, a characteristic of summarized texts. These texts clustered by similarity are important for knowledge discovery in the Cassiopeia model.

A. *Summarization in the Cassiopeia Model employed in the pre-processing stage*

The pre-processing stage refers to cleaning the texts in order to facilitate the next stage.

In the Cassiopeia model, as shown by Figure 1, the texts are prepared to be processed, where case folding techniques are employed and the images, tables and markings are removed. This phase also sees the emergence of the first contribution of the Cassiopeia model, which is the addition of the summarization process.

1) *Text Summarization Concept*

Summarization can be defined as reduced texts that transmit main and more relevant ideas of an original text, clearly and objectively, without losing information [4]. The need for simplification and summarization is justified by the increase in the volume of information available in means of communication (mainly the Internet) and by the lack of time to read texts of different kinds [9]. As a consequence of this process, readers are unable to absorb the whole content of the original texts. Therefore, an abstract is a summary that aims to capture the main idea the author intended to portray and transmit it in a few lines to the reader.

Automatic summarization (AS) used in the Cassiopeia Model is extractive following the empirical approach, also known as the superficial approach. This technique uses statistical or superficial methods that identify the most relevant segments of the source text, producing the summaries by means of juxtaposition of extracted sentences, without any modification in relation to the order of the original text.

Another important factor worth highlighting in this stage of the Cassiopeia model is the maintenance of stopwords, which is rendered viable with the addition of the summarization process. Stopwords are closed sets of words and they are formed by articles, pronouns, interjections and prepositions. The Cassiopeia model keeps the stopwords, unlike other research studies in this field which remove them in order to diminish the volume of words in the knowledge extraction stage. Keeping stopwords represents an important achievement for the Cassiopeia model as it becomes independent of language.

B. Text clustering technique employed in the knowledge extraction stage of the Cassiopeia Model

The knowledge extraction stage is basically the application of text mining algorithms. The algorithms and techniques employed in this stage can be divided into two categories, according to [17]: knowledge generation that uses techniques for generating knowledge from information contained in a certain text and knowledge extraction that uses techniques to extract knowledge that is explicit in the text. The Cassiopeia model employs the clustering technique to generate knowledge from textual documents.

According to [5], text clustering is an entirely automatic process that divides a collection into clusters of texts with similar content. The way in which the Cassiopeia model conducts clustering is described in the three phases below, as suggested by publications by [10] and [11].

a) First Phase - (Identification of Attributes)

The Cassiopeia model selects word characteristics in the text using relative frequency. It defines the importance of an expression according to how often the term is found in the text. The more a term appears in the text, the more important this term is for that text. Relative frequency is calculated by equation (1). This formula normalizes the result of absolute frequency of words, preventing small documents from being represented by small vectors and large documents from being represented by large vectors.

$$F_{rel} X = \frac{F_{abs} X}{N} \qquad (1)$$

With normalization, all documents are represented by vectors of the same size. Whereas $F_{real}X$ 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, which is the word, appears in the document and N is equal to the total number of words in the text.

Considered a spatial vector, each word represents a dimension (there are as many dimensions as there are different words in the text). This is a high dimensionality and sparse data problem that is common in TM and starts being treated by the Cassiopeia model during the pre-processing phase, where the summarization is conducted, thus causing significant reduction in the dimensionality space and sparse data.

b) Second phase - (Selection of Attributes)

The Cassiopeia model identifies similarity using a similarity measure. The Cassiopeia model uses [3] set theoretic inclusion, a simple inclusion measure that evaluates the presence of words in two compared texts. If the word appears in both texts, the value one (1) is added to the meter; if not, zero (0) is added. At the end, the degree of similarity is a value between 0 and 1, calculated by the average; i.e., the total amount of the meter (commons) divided by the total number of words in both texts (without counting repetitions). The fact that one word is more important in certain texts or how often it appears is not taken into consideration in this calculation. In the Cassiopeia model, this problem is solved with another function described by [23], which calculates the average, however using weights for each word. Therefore, it considers the fact that words appear with different importance in the texts. In this case, the weight of the words is based on relative frequency. The similarity value is calculated by the average between the average weights of common words. That is, when the word appears in both documents, the average weights are added up instead of adding the value one (1). At the end, the average of the total number of words in both documents is calculated.

In this phase, once again there is an attempt to minimize the problem caused by high dimensionality and sparse data. The Cassiopeia model uses a similarity threshold [34], where words (characteristics) whose importance (frequency) is lower than the similarity value are simply ignored in order to compose the vector of words in the text. The Cassiopeia model also defines a maximum number of 50 positions (truncation) for the vectors [34].

With similarity calculations, the definition of the similarity threshold, and vector truncation, the Cassiopeia model defines the selection of attributes.

In this phase of the clustering process there is a cohort that represents the average frequency of words obtained with similarity calculations. Then the organization of vectors proceeds in decreasing value, as represented by Figure 3. The Cassiopeia model employs a similarity threshold that undergoes a change in the clustering and a variation from 0.1 to 0.7 and a vector truncation with 50 positions and a vector truncation with 50 positions with 25 words to the left of the frequency average and the 25 words to the right.

According to [27], the Zipf[36] curve shown in Figure 3 (an adaptation for the Cassiopeia model) has three distinct areas. The area defined as I is where trivial or basic information is found, with greater frequency; area II is where interesting information is found; and area III is where noises are found. The stopwords are found in area I. It is common for clustering jobs to achieve a first cohort, based on the Zipf curve denominated by [27] as area I, to remove the stopwords, which are the most commonly found words. This occurs still in the pre-processing stage. Then there are many techniques for creating the second cohort. The variation of these cohorts is known as Lunh[18] threshold and can be appreciated in detail in researches by [27],[28], [29] and [22].

The first cohort causes the algorithms to become language-dependent, since it needs a list of these stopwords for each language. This first cohort is necessary because clusterers work with a similarity matrix (shown in Figure 2), defined by [34] as containing similarity values among all the elements of the referred dataset. According to [33], not conducting this cohort (removal of stopwords) would cause a high dimensionality and sparse data problem, which grows exponentially with its text base and generates a crucial problem in the TM field.

The selection of attributes in the Cassiopeia model shown in Figure 3, formalized in equation 2 and added by summarization at the pre-processing phase guarantees the elimination of stopwords, thus rendering language independence. In addition to language independence, another important factor in the Cassiopeia model is that it does not use a similarity matrix, and this assures a possible solution for the problem of high dimensionality and sparse data in TM.

Figure 2. Similarity Matrix obtained through Eurekha[1] [34] using corpus TeMário [25].

The lack of similarity matrix shown in Figure 2 has been replaced, in the Cassiopeia model, by the use of vectors in each cluster called centroids, thus avoiding the calculation for distance, which is common for clusters that use the similarity matrix. The Cassiopeia model calculates similarity using the procedures cited in item II.B. of this article during phase 1 and phase 2. In order to keep this structure of vectors (centroids), the Cassiopeia model uses the Hierarchical Clustering method and Cliques algorithm, discussed in phase 3.

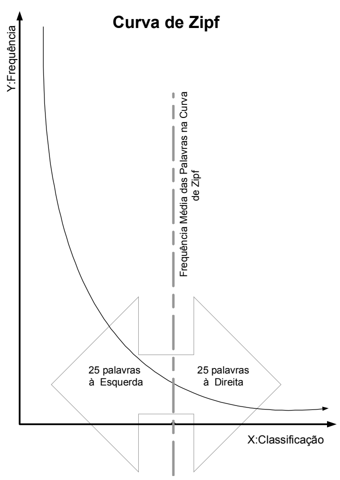

Figure 3. Selection of Attribute in the Cassiopeia model.

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

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

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

[1] The *Eurekha* analyzes the content of texts and identifies those that contain the same subject. These documents with similar content are assigned to a single cluster. At the end of the review process, the software offers the user the different clusters found and their respective documents.

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

c) Third Phase

The Cassiopeia model uses the Hierarchical Clustering method that, by analyzing dendograms built, defines the previous number of clusters. With the Cliques algorithm, which belongs to the a graph-theoretic class, whose graph formed is illustrated in Figure 4, the elements are only added to a cluster if its degree of similarity is greater than the threshold defined for all elements present in the clusters and not only in relation to the central element.

In this case, according to [11], the clusters tend to be more cohesive and better quality, once the elements are more similar or close. Algorithm 1 describes the steps of the Cliques Algorithm.

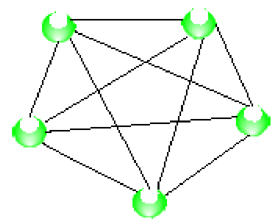

Figure 4. Graphic representation of the Clique Algorithm.

1 – Clique Algorithm:

1. Select next Object and add it to a new cluster;
2. Look for a similar object;
3. Se este objeto for similar a todos os *objetos do cluster, adicioná-lo;*
4. *Enquanto houver objetos, voltar ao passo 2;*
5. *Return to step 1.*

C. Cassiopeia model and post-processing phase

In the post-processing stage, the Cassiopeia model ends with summarized texts clustered by similarity. The Cassiopeia model employs descriptive analysis throughout its knowledge extraction stage, which, according to [22] produces new knowledge based on textual data obtained from standards that can be interpreted by humans. [22] states that these descriptive activities use unsupervised algorithms, which extract data patterns from unlabeled data. [22] concludes that the main task of MT is to obtain association rules, clustering, and summarization.

According to [17] and [34], in this phase it is possible to obtain the assessment of knowledge discovery, analyzing the resulting clusters with the texts contained in each cluster. According to [2], a problem generated by high dimensionality in TM is understanding the extracted knowledge.

The Cassiopeia model allows for an easier way to obtain knowledge compared to other text clustering because its texts are summarized, i.e., they have a much smaller number of sentences and these have a much greater degree of information[2,] which is guaranteed by summarization used in the pre-processing stage.

Texts clustered during post-processing in the Cassiopeia model allow document recovery and the analysis of recovered documents may lead to obtaining similar documents, all this in summarized form and displaying high degree of information, thus providing knowledge discovery.

The Cassiopeia model still does not, but may subsequently facilitate, the identification of categories by means of the characteristics contained in each word vector. In academic literature, this technique is called cluster analysis, as defined by [27]. As this technique has still not been incorporated by the Cassiopeia model, it will be commented in Section 5A.

III. Methodology

This paper's contribution lies in the fact that the Cassiopeia model includes the text summarization in the pre-processing stage in order to improve clustering measurement and it additionally provides a new variation of Luhn [18] cohort thresholds. Therefore, measurement improvement occurs in the external and internal metrics in clusters of the Cassiopeia model. In order to verify this contribution, the following simulation methodology was created: the corpus used is described in item III.A of this article. The choice of summarizers is detailed in item III.B. In order to verify the contribution of this research, external and internal metrics were chosen that are commonly used in measuring the clustering process and commented in detail in item III.C.

Compression percentages were defined to be used in summarized texts. Each text was summarized by summarization algorithms of 50%, 70%, 80% and 90%.

The original texts, i.e., with summarization, were submitted to the knowledge extraction process of the Cassiopeia model. After each one of the texts obtained with the summarization algorithms with their respective compression percentages of 50%, 70%, 80% and 90%, they were also alternately submitted to the knowledge extraction process of the Cassiopeia model. Each one of these processes generated in the Cassiopeia model were submitted to a repetition of 100 steps to generate, at the end, a median average for each set of external or internal metrics obtained in clusterings. The results of the set of standards of all metrics (external and internal) were individually compared with texts lacking summarization and summarized texts with its compression percentages and summarization algorithms. These results are shown and analyzed in Section 4 of this paper.

A. Corpus

For this experiment, original texts and summarized texts extracted from their original versions were used as corpus, in Portuguese and English.

[2] The degree of information of a text is measured according to the world knowledge of the person it is destined for. In other words, a text possesses high degree of information when a more broad understanding of this text depends on the reader's cultural repertoire.

In Portuguese, the texts are included among journalistic, legal and medical domains, totalizing 300 original texts, or 100 texts per domain.

For the legal domain, the chosen texts were scientific articles taken from a legal website (www.direitonet.com.br) between Feb-09-2010 to Feb-14-2010 and are separated by the following fields: Environmental, Civil, Constitutional, Consumer, Family, Criminal, Social Security, Procedural and Employment Law.

In the medical domain, the texts are also composed of scientific articles taken from a scientific website (www.scielo.br) between Feb-09-2010 and Feb-14-2010 and are separated by the following fields: Cardiology, Dermatology, Epidemiology, Geriatrics, Gynecology, Hematology, Neurology, Oncology, Orthopedics and Pediatrics.

In the journalist domain, the corpus *TeMário* 2004 [25] was used with texts extracted from the online newspaper *Folha de São Paulo* and are distributed throughout five sections: Special, International, World, Opinion and Politics.

As for texts in the English language, there was also domain variation, but only for the journalistic and medical domains, totalizing 200 original texts. In the legal texts with the established criteria of this work, they were not found in a database with the same characteristics (mainly gratuity).

The journalistic texts were taken from *Reuters* news agency (www.reuters.com) between Apr-27-2010 to Apr-30-2010 and are separated by the following fields: Economy, Entertainment, G-20, Green Business, Health, Housing Market, Politics, Science, Sports and Technology.

The medical texts included scientific articles from a scientific website (www.scielo.org) between Apr-09-2010 and Apr-17-2010 and are separated by the following fields: Cardiology, Dermatology, Epidemiology, Geriatrics, Genecology, Hematology, Neurology, Oncology, Orthopedics and Pediatrics.

It is worth highlighting that these texts have already been classified, according to each field, into the textual databases where they were found. This classification is important because it serves as a reference for the external measure of the research using classification by specialists.

In order to complete this methodology there was language variation (English and Portuguese), then domain variation (journalistic and medical for both languages and legal just for Portuguese). Text summarizers were chosen according to the specifications detailed in item III.B, and there were seven for each language (note that random functions were executed three times on the same text and in each language). Finally, each summarizer used for each domain with 100 texts the compressions of 50%, 70%, 80% and 90%. For this experiment, a total of 30,000 summarized texts were used, 18,000 of which were in Portuguese and 12,000 of which were in English.

B. Summarizers

Professional and literature summarizers were chosen for the simulation. As criteria for choosing summarization algorithms of these experiments, we picked those which had the possibility of defining percentages of compression per word. Thus it was possible to have 50%, 70%, 80% and 90% of the original text.

For the summarization process in Portuguese, three summarizers were used, as found in literature:

The Supor by [20] which selects, to compose the extract, the sentences that include the most commonly used words in the original text. The Gist_Average_Keyword by [24], where the sentence score may be calculated by one of two simple statistical methods: the keyword method or the average keyword method. Gist_Intrasentence also by [24] conducts the exclusion of stopwords in all sentences. For the summarization process in English, three summarizers were used, one professional and another literature, which is available in the web: Copernic and Intellexer Summarizer Pro are professional summarizers and their algorithms are considered black boxes. SewSum by [14] is a literature summarizer. For each language, SewSum uses a lexicon for mapping flexed word forms, from the content to it respective root.

Four additional functions were also developed: FA1_S_Stopwords, FA2_S_Stopwords, FA1_C_Stopwords e FA2_C_Stopwords, all randomly choosing sentences from the original text, based on words. This occurs for texts in both English and Portuguese. The functions FA1_S_Stopwords and FA2_S_Stopwords remove the stopwords from the texts before the choice for reducing the number of words before the randomization process. As for FA1_C_Stpowords and FA2_C_Stpowords, these do not remove the stopwords. These functions adopt the two-method variation, due to the chosen percentage (%) and compression. Functions FA2 finish summarization when they achieve the compression percentage, regardless of where they are in the sentence. As for the FA1 functions, these are kept until the end of the sentence, thus not respecting the established compression percentage. Because they are random functions, each one received the summarization process three times in each text, in order to obtain a mean average. This number of repetitions was chosen based on observations of corpus tests, and no significant increase was verified in the average that could justify a larger number of summarizations using any one of the random functions.

C. Metrics

The process of clustering by similarity is by definition unsupervised. This way, there are no predefined classes and examples that indicate the characteristics of the data set.

According to [13] the evaluation of clusters may be distributed into three broad metric categories: External or Supervised Metrics; Internal or Unsupervised Metrics; and Relative Metrics, which are not used in this study.

For external or supervised metrics, the clustering results are assessed by a structure of predefined classes that reflects the opinion of a human specialist. For this kind of metric, according to [30], the following measures are used: *Precision*, *Recall*, and as a harmonic measure of these two metrics, *F-Measure*.

For internal or unsupervised metrics, the only information used are contained in the clusters generated to conduct the evaluation of results, i.e., external information is not used. The most commonly used standards for this purpose, according to [20] and [1], are Cohesion, Coupling, and as a harmonic measure of these two metrics, *Silhouette Coefficient.* With the aim of validating the results, this experiment used external and internal metrics and the following were defined.

1) External Metrics

Recall(R)
$$\frac{tlcd_i}{tg\,\mathrm{cd}_i} * 100 \quad (3)$$

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

Precision (P)
$$\frac{tlcd_i}{te_i} * 100 \quad (4)$$

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

F-Measure (F)
$$2 * \frac{Precision(P) * \mathrm{Re}\,call(R)}{Precision(P) + \mathrm{Recall(R)}} \quad (5)$$

2) Internal Metrics

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

Where *Sim* (P_i,P_j) calculates the similarity between texts *i* and *j* belonging to cluster *P, n* is the number of texts in cluster *P*, and P_i and P_j are members of cluster *P*.

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

Where *C* is the centroid of a certain cluster present in *P*, *Sim* (C_i, C_j) calculates the similarity of text *i* belonging to cluster *P* and text *j* does not belong to *P*, C_i centroid of cluster *P* and C_j is the centroid of cluster P*i* 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-th* element of the cluster and the other elements of the same cluster. Value *b(i)* is the minimum distance between the *i-th* element of the cluster and any other cluster that does not contain the element and *max* is the greatest distance between a(i) and b(i). The *Silhouette Coefficient* of a cluster is the mean average of the coefficients calculated for each element belonging to the cluster

Shown in *Equation 9* $\overline{s} = \frac{1}{N}\sum_{i=1}^{N} S$. Value S ranges from 0 to 1.

D. 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.

a) 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 [35]. This count is usually normalized to avoid a bias towards longer documents so as to place a measure of importance of *i t* withen the given document *d j*. Hence, one gets the term frequency, defined as follows:

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

$n_{i,j}$ 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* |.

b) 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:

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

Where *dft* is the document frequency of *t*: the number of documents that contain *t*.

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

c) 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].

IV. RESULTS OBTAINED IN THE EXPERIMENTS

Due to the large volume of data obtained by the results of the experiments, we have chosen to present only a few significant graphs focusing on the harmonic measures in the external metric, which is F-Measure, as well as in the internal metric, which is the Silhouette Coefficient. The results that are presented in figures 3, 4, 5, and 6 are the averages of the sums of the measures obtained from each summarizer in each of the simulated methods. Although they are not included in this work, it's worth highlighting that results were generated for all the metrics explained in item C, for both the Portuguese and English-language texts, in all the domains (journalistic, legal and medical), and for all different compression levels (50%, 70%, 80% and 90%). However, due to size constraints, it was impossible to include all of these results.

Figure 3 illustrates the external metric and the harmonic measure F-Measure, in which the Cassiopeia method displayed superior results in 7 out of 12 possible results considering domains and compressions, which corresponds to 58.33% of the entire sample.

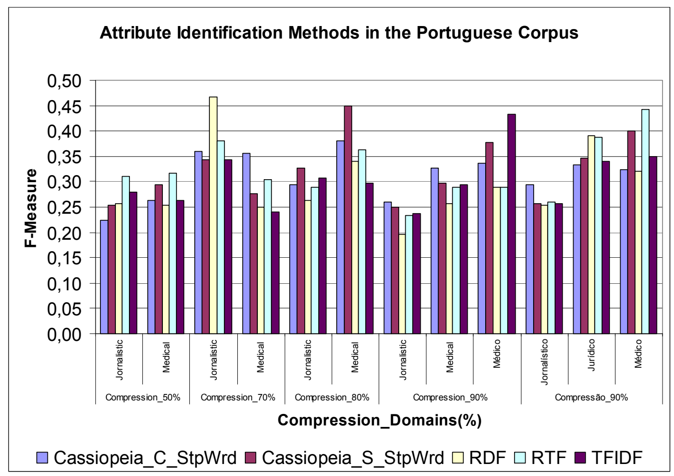

Figure 3. Values for the Cassiopeia Method both with and without Stopwords and of other methods in the literature, RDF, RTF and TFIDF, using the Gist_Keyword, Gist_Intra and Supor summarizers, at 50%,70%,80% and 90% compression, in the jornalistic, legal and medical categories in the Portuguese language. The results show the averages of the sums of the averages obtained by each summarizer in each of the simulated methods of the harmonic measure F-Measure taken throughout the 100 interactions.

In Figure 3, where the external metric and the harmonic measure F-Measure was used, the Cassiopeia method displayed superior results in 2 out of 8 possible results considering domains and compressions, which corresponds to 25% of the entire sample.

Figure 4 illustrates the external metric and the harmonic measure Silhouette Coefficient, in which the Cassiopeia method displayed superior results in all of the possible results considering domains and compressions, which corresponds to 100% of the sample. It's important to note that, in contrast to the results from Figure 3, in which the best results alternated between the Cassiopeia method with and without Stopwords, in

Figure 4, the superior results were obtained exclusively with the Cassiopeia method that uses stopwords.

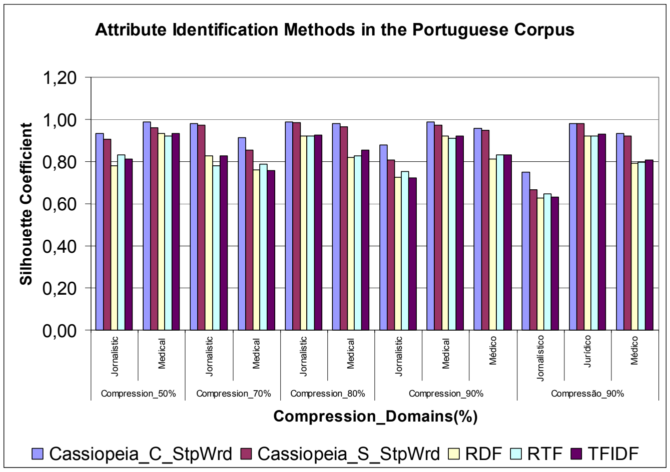

Figure 4. Values for the Cassiopeia Method both with and without Stopwords and of other methods in the literature, RDF, RTF and TFIDF, using the Gist_Keyword, Gist_Intra and Supor summarizers, at 50%,70%,80% and 90% compression, in the jornalistic, legal and medical categories in the Portuguese language. The results show the averages of the sums of the averages obtained by each summarizer in each of the simulated methods of the harmonic measure Silhouette Coefficient taken throughout the 100 interactions.

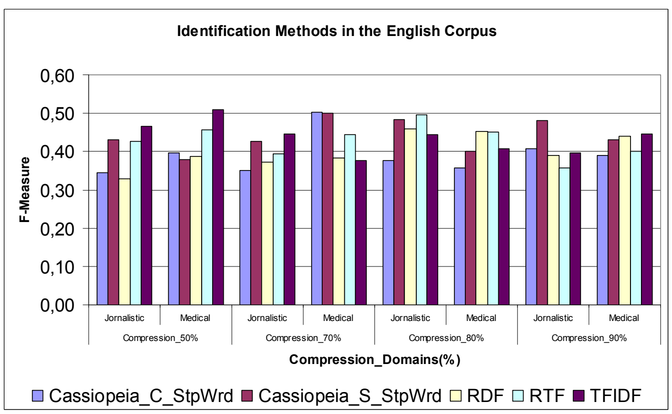

Figure 5. Values for the Cassiopeia Method both with and without Stopwords and of other methods in the literature, RDF, RTF and TFIDF, using the Copernic, Intellexer and SweSum summarizers, at 50%,70%,80% and 90% compression, in the jornalistic and medical categories in the English language. The results show the averages of the sums of the averages obtained by each summarizer in each of the simulated methods of the harmonic measure F-Measure taken throughout the 100 interactions.

In Figure 6, which shows the external metric and the harmonic measure Silhouette Coefficient, the Cassiopeia method displayed superior results in all of the possible results considering domains and compressions, which corresponds to 100% of the sample. Once again, it is worth noting that, in contrast to the results from Figure 5, in which the best results alternated between the Cassiopeia method with and without Stopwords, in Figure 6, the superior results were obtained exclusively with the Cassiopeia method that uses stopwords.

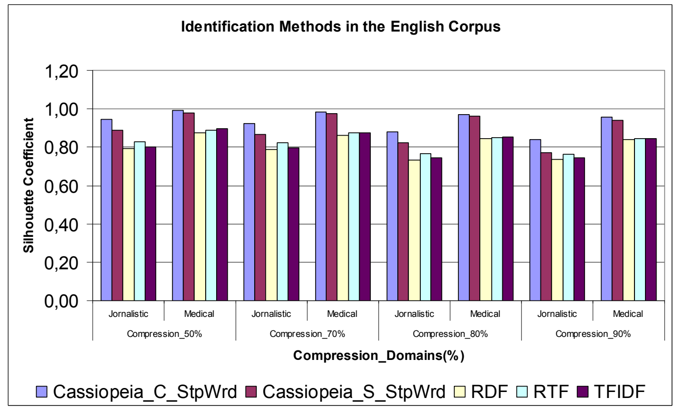

Figure 6. Values for the Cassiopeia Method both with and without Stopwords and of other methods in the literature, RDF, RTF and TFIDF, using the Copernic, Intellexer and SweSum summarizers, at 50%,70%,80% and 90% compression, in the jornalistic and medical categories in the English language. The results show the averages of the sums of the averages obtained by each summarizer in each of the simulated methods of the harmonic measure Silhouette Coefficient taken throughout the 100 interactions.

V. Conclusion

In order to evaluate the results of the Cassiopeia model using external metrics, we can observe figures 3 and 5 that show the use of the F-Measure, a harmonic measure of both Recall and Precision. Figure 3 shows a very good performance in the Portuguese language, where the best results came from the Cassiopeia model in more than 50% of the sample, thereby showing good regularity in the entire sample in comparison with other methods. In Figure 5, the results were fair, since the Cassiopeia method is on the same level as the methods in the literature.

The results of the internal metrics, observed in Figures 4 and 6, show the use of the Silhouette Coefficient, a harmonic measure of both Cohesion and Coupling. In this case, there was an absolute predominance of the Cassiopeia method as providing the best values among all samples. This was true for both Figure 4 and Figure 6. The best method was the Cassiopeia method with stopwords, which obtained the highest values in the entire sample, once again, in both figures 4 and 6. This result was so significant that the value that ranked second in the entire samples – in both figures 4 and 6 – was the Cassiopeia method without stopwords. Seeing as Cassiopeia is a non-supervised method, these results were very significant as there was an absolute prevalence in the internal metric, which, as we saw in item C, is the metric that uses only information contained in the generated groups in order to conduct an evaluation of the results, thereby eschewing the need for external information.

A. 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.

REFERENCES

[1] Aranganayagil, S. and Thangavel, K. "Clustering Categorial Data UsingSilhouette Coefficient as a Relocating Measure". In International conference oncomputational Intelligence and multimedia Applications, ICCIMA, Sivakasi, Índia. Proceedings. Los Alamitos: IEEE, p.13-17, 2007.

[2] Carvalho, V. O."Generalização no processo de mineração de regras de associação". Tese de Doutorado, Instituto de Ciência Matemáticas e de Computação. USP, São Paulo, Brasil, 2007.

[3] Cross, V. "Fuzzy information retrieval". Journal of Intelligent Information Systems, Boston, v.3, n.1, p.29-56, 1994.

[4] Delgado, C. H. and Vianna, C. E. and Guelpeli, M. V. C. "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, 2010.

[5] Fan, W. and Wallace L. and Rich S., and Zhang Z."Tapping the power of text mining" Communications of the ACM, vol. 49. 2006.

[6] 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.

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

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

[9] Guelpeli, M. V. C. ; BernardinI, F. C. ; Garcia, A. C. B. "Todas as Palavras da Sentença como Métrica para um Sumarizador Automático". In: Tecnologia da Informação e da Linguagem Humana-TIL, WebMedia, 2008. p. 287-291,Vila Velha, 2008.

[10] 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, September 24-27, 2009.

[11] Guelpeli, M. V. C. and Bernardini, F. C. and 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 ISBN: 978-1-84996-076-2.

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

[13] Halkidi, M. Batistakis, Y and M. Varzirgiannis. "On clustering validation techniques". Journal of Intelligent Information Systems, 17(2-3):107{145, 2001.

[14] Hassel, M. "Resource Lean and Portable Automatic Text Summarization" PhD-Thesis, School of Computer Science and Communication, KTH, ISBN-978-917178-704-0.,2007

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

[16] Keogh, E.; Kasetty, S.(2002). "On the need for time series data mining benchmarks: a survey and empirical demonstration", In: ACM SIGKDD, Edmonton, Canada, 2002, p.102-111.

[17] Loh, S. "Abordagem Baseada em Conceitos para Descoberta de Conhecimento em Textos". Porto Alegre: UFRGS, 2001. qualificação doutorado, Instituto de Informática, Universidade Federal do Rio Grande do Sul, 2001.

[18] Luhn, H. P. "The automatic creation of literature abstracts.".IBM Journal of Research and Development, 2, pp. 159-165,1958.

[19] Makrehchi, M." Taxonomy-based Document ClusteringJournal of Digital Information Management",Volume 9,Issue 2, April, 2011,Pages 79-86.

[20] Módolo M. "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, 2003.

[21] Hourdakis, N.,Argyriou, M., Euripides G.M., Petrakis, E. E. M."Hierarchical Clustering in Medical Document Collections: the BIC-Means Method" Journal of Digital Information Management,Volume 8,Issue 2, April, 2010, Pages71-77

[22] Nogueira, B. M. and Rezende, S. O. "Avaliação de métodos não-supervisionados de seleção de atributos para Mineração de Textos" In: VII Concurso de Teses e Dissertações em Inteligência Artificial (CTDIA 2010) - São Bernardo do Campo, SP. v. 1. p. 1-10, 2010.

[23] Oliveira, H. M."Seleção de entes complexos usando lógica difusa".. Dissertação (Mestrado em Ciência da Computação) – Instituto de Informática, PUC-RS, Porto Alegre, 1996.

[24] Pardo, T.A.S. and Rino, L.H.M. and NunesS, M.G.V. "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.,2003.

[25] Pardo, T.A.S. and Rino, L.H.M. "TeMário: Um Corpus para Sumarização Automática de Textos" Relatórios Técnicos (NILC-TR-03-09). NILC – ICMC – USP. São Carlos, Brasil, 2004.

[26] Pottenger, W. M.; Yang, T.(2001). « Detecting emerging concepts in textual data mining", in:Michael Berry (ed.), Computational Information Retrieval, SIAM, Philadelphia, August 2001.

[27] Quoniam, L. "Bibliométrie sur des référence bibliographiques:methodologie" In: Desvals H.; Dou, H. (Org.). La ueille techndogique. pd : huiob 2M - 262., 1992.

[28] Quoniam, L. "Los x- en bibliomé Frie et dossier & travaux. Marseille".Université de Droit d'konomie et des Sciences d'Aix-Marseille 111, 1996.

[29] Tan, A. H. (1999). "Text mining: the state of the art and the challenges." In: WORKSHOP ON KNOWLEDGE DISCOVERY FROM ADVANCED DATABASES, 1999. Proceedings... Heidelberg, 1999. p.65-70. (Lecture Notes in Computer Science, 1574).

[30] Tan, P. N. and Steinbach, M. and Kumar, V. "Introduction to Data Mining". Addison-Wesley, 2006.

[31] Velickov, S.(2004). "TextMiner theoretical background."Acess in: <http://www.delft-cluster.nl/textminer/theory/>. 15 Mar 2011.

[32] Ventura, J.M.J. "Extração de Unigramas Relevantes". Dissertação Mestrado em Engenharia Informática. Lisboa ,Portugal, 2008.

[33] Vianna, D. S. "Heurísticas híbridas para o problema da logenia" Tese de doutorado, Pontifícia Universidade Católica - PUC, Rio de Janeiro, Brasil. 2004

[34] Wives, L.K."Utilizando conceitos como descritores de textos para o processo de identificação de conglomerados clustering de documentos" – Tese (doutorado), Porto Alegre, Brasil, 2004.

[35] Won, B.O. , Lee, I. "Framework For Mixed Entity Resolving System Using Unsupervised Clustering" Journal of Digital Information Management,Volume 8,Issue 6, December, 2010, Pages 362-368.

[36] Zipf, G. K." Human Behavior a and the Principle of Least Effort". Addison-Wesley, 1949.