The process of summarization in the pre-processing stage in order to improve measurement of texts when clustering

Marcus V. C. Guelpeli Ana Cristina B. 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 introduces the Cassiopeia model, which allows for knowledge discovery in textual bases, used for the purposes of text mining in distinct and/or antagonistic domains. 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. In the knowledge extraction, the texts are clustered and reclustered according to a similarity criterion. With the results obtained, the study hopes to show the impact of including summarization in the process of text clusterization. The experiments conducted in this study indicate that text clusterization using summaries is in fact much more effective than direct clusterization of texts in their entirety, as measured by internal and external measures traditionally employed in the field of text clusterization. Finally, the post-processing stage creates clusters of summarized texts with a high degree of informativity, a quality that is inherent to summarization. The clusters are highly esteemed with the indexed words. This fact is due to the process proposed by the Cassiopeia model, which allows for strong similarity among the clustered texts. In the future, this similarity will allow for the creation of categories based on the word indices of each cluster.**

Keywords- Knowledge Discovery; Text Mining; External and Internal Clustering Measures; Summarization; Similarity.

I. INTRODUCTION

State that, since information is generated and shared at heightened speed and amplitude, a new dynamic of reuse and production of new knowledge emerges. Therefore, processing this information becomes necessary since human ability to read and register is limited.

Large quantities of structured or non-structured information found mostly online have awakened interest in research for the development of increasingly sophisticated and high-speed tools for searching and recovering information. In order to extract knowledge to support decision-making, the field of text mining (TM) emerges using a competitive and organizational approach. The primary aim of text mining is extracting patterns or inferring some type of knowledge of a set of texts [5].

According to [8], the most comprehensive and non trivial process of identifying new, valid, potentially useful and understandable patterns in unstructured text is known as knowledge discovery in text -KDT, which is a main focus of this article. The field of KDT is quite broad and, according to [8], three main stages can be observed in this field: the pre-processing or data preparation stage, the data analysis or knowledge extraction stage, and the post-processing or discovery evaluation stage.

The model called Cassiopeia provides knowledge discovery in textual bases in antagonistic or different domains which are represented in this article by the following domains: journalistic and medical. In the knowledge extraction stage, the Cassiopeia model uses the text clustering technique in order to obtain knowledge as well as to provide better measurement among clusters at the final stage.

One of the biggest problems found in the TM field, consequently in KDT, is high dimensionality and sparse data. This is the reason why the Cassiopeia model includes summarization in the pre-processing stage with the function of reducing the size of texts that are going to be manipulated in the knowledge extraction stage. During the post-processing stage, the Cassiopeia model provides texts which are summarized and clustered with good measurement in its clustering, thus making it easier to assess knowledge discovery.

This article has been organized as follows. Section 2 presents the Cassiopeia model with the pre-processing stage using the Summarization process (for reducing dimensionality and sparse data) and the knowledge extraction stage where the Cassiopeia model uses the Clustering process. Section 3 presents the methodology. Section 4 discusses the results obtained in the experiments. Section 5 presents conclusions and suggestions for future research.

II. 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.

978-1-908320-00-1/11/$26.00 ©2011 IEEE

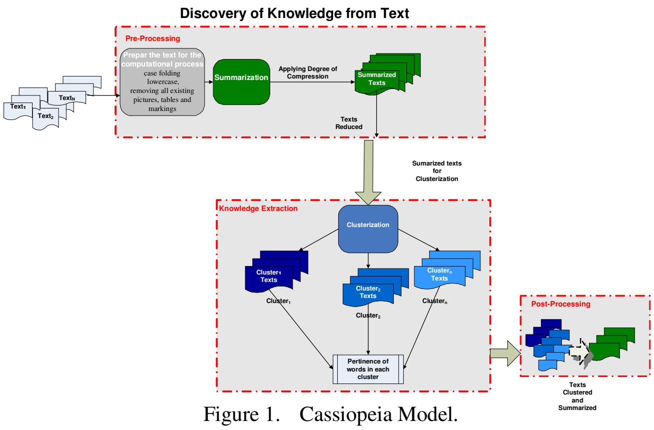

Discovery of Knowledge from Text
Pre-Processing
Text₁
Text₂
Text_N
Prepar the text for the computational process
case folding
lowercase,
removing all existing pictures, tables and markings
Summarization
Applying Degree of Compression
Summarized Texts
Texts Reduced
Sumarized texts for Clusterization
Knowledge Extraction
Clusterization
Cluster₁ Texts
Cluster₂ Texts
Clusterₙ Texts
Cluster₁
Cluster₂
Clusterₙ
Pertinence of words in each cluster
Post-Processing
Texts Clustered and Summarized

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. 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 IIA and IIB.

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. According to [18], 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..

A. *Summarization in the Cassiopeia Model employed in the pre-processing 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 [5]. 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.

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 Cassiopeia model employs the clustering technique to generate knowledge from textual documents.

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 [5].

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 [11], 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 [18], 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.

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 in equation (2). 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, as shown in Figure 2.

According to [14], the Zipf curve shown in Figure 2 (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 [14] 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 threshold and can be appreciated in detail in researches by [14], [16] and [10].

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 defined by [18] as containing similarity values among all the elements of the referred dataset. According to [17], 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 2, 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.

The lack of similarity matrix 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(a) and phase 2(b). In order to keep this structure of vectors (centroids), the Cassiopeia model uses the Hierarchical Clustering method and Cliques algorithm, discussed in phase 3(c).

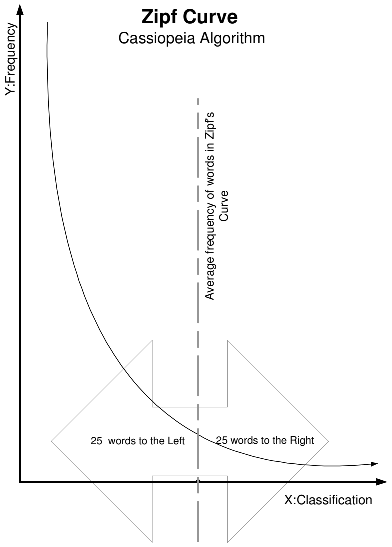


Figure 2. Selection of Attribute in the Cassiopeia model.

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

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

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

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 3, 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. Algorithm 1 describes the steps of the Clique Algorithm.

In this case, according to [5], the clusters tend to be more cohesive and better quality, once the elements are more similar or close.

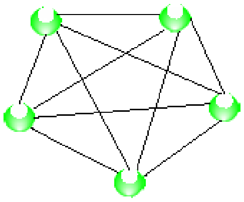


Figure 3. Graphic representation of the Clique Algorithm..

1.Cliques Algorithm:

1. Select next Object and add it to a new cluster;
2. Look for a similar object;
3. If this object is similar to all of the objects in the cluster, add it;
4. Stop criterion: while there is at least one object not allocated, come back to Step 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.

According to [18] and [10] 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[1,] which is guaranteed by summarization used in the pre-processing stage.

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 [16] 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.

In Portuguese, the texts are included among journalistic and medical domains, totalizing 200 original texts, or 100 texts per domain.

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,

[1] 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.

Hematology, Neurology, Oncology, Orthopedics and Pediatrics.

In the journalist domain, the corpus *TeMário* 2004 [13] 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, journalistic and medical domains, totalizing 200 original texts. 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). 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%.

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 [9] which selects, to compose the extract, the sentences that include the most commonly used words in the original text. The Gist_Average_Keyword by [12], 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 [12] 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 [7] 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.

C. Metrics

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 [15], 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 [15] 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

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

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.

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

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

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

2) Internal Metrics

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

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 $\mathrm{P_j}$ are members of cluster *P*.

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

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 $\mathrm{C_j}$ is the centroid of cluster P*i* and n_a is the number of clusters present in *P*.

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

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.

IV. EXPERIMENTAL RESULTS

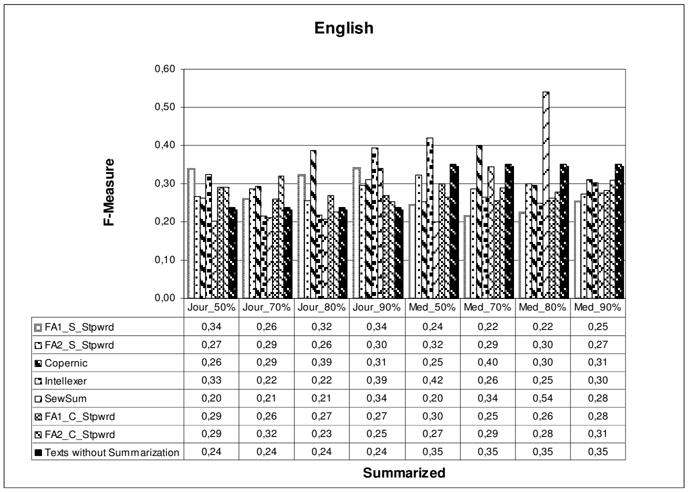
Due to the large quantity of metrics used to measure clusters of the Cassiopeia model and, consequently, the results, graphs of the harmonic standard, in the external metric *F-Measure* and in the internal measure *Silhouette Coefficient*. Although they are not included in this study, it is worth highlighting that results were produced for Portuguese and English in all domains: journalistic and medical, for different levels of compression with 50%, 70%, 80% and 90% and with all measures described in item C.

Figures 4 and 5 show results of text clustering obtained by the Cassiopeia model, using F-Measure and Silhouette Coefficient, in English regarding journalistic domains and medical using all compressions (50%, 70%, 80% and 90%) and the seven summarization algorithms[2] in addition to the texts without summarization[3].

Regarding the journalistic domain shown in Figure 4, the summarization algorithms mostly allow the Cassiopeia model to generate text clusters with F-Measure values, greater than the text clusters without summarization.

There are a few exceptions: for a compression of 50% the exception was SewSum; for 70% compression it was SewSum and Intellexer, for 80% of compression it was SewSum, Intellexer and FA2_com_Stopword, and for 90%, all the algorithms, summarization allowed the Cassiopeia model to obtain clusters with F-Measure clusters larger than clusters with unsummarized texts.

In the medical domain, in compressions of 50%, 70% and 80%, only one summarization algorithm in each compression allowed the Cassiopeia model to generate clusters of F-Measure value greater than texts without summarization. With 90% none of the summarization algorithms made it possible to generate clusters that had F-Measure values greater than the unsummarized texts.



	Jour_50%	Jour_70%	Jour_80%	Jour_90%	Med_50%	Med_70%	Med_80%	Med_90%
▩ FA1_S_Stpwrd	0,34	0,26	0,32	0,34	0,24	0,22	0,22	0,25
□ FA2_S_Stpwrd	0,27	0,29	0,26	0,30	0,32	0,29	0,30	0,27
◩ Copernic	0,26	0,29	0,39	0,31	0,25	0,40	0,30	0,31
□ Intellexer	0,33	0,22	0,22	0,39	0,42	0,26	0,25	0,30
□ SewSum	0,20	0,21	0,21	0,34	0,20	0,34	0,54	0,28
⊠ FA1_C_Stpwrd	0,29	0,26	0,27	0,27	0,30	0,25	0,26	0,28
⧅ FA2_C_Stpwrd	0,29	0,32	0,23	0,25	0,27	0,29	0,28	0,31
■ Texts without Summarization	0,24	0,24	0,24	0,24	0,35	0,35	0,35	0,35

Figure 4. Measures obtained in clusters generated by the Cassiopeia model in 100 interaction steps. The results are the accumulated average of the F-Measure, in English, within the journalistic and medical domain. Measures obtained in clusters generated by the Cassiopeia model in 100 interaction steps. The results are the accumulated average of the F-Measure, in English, within the journalistic and medical domain.

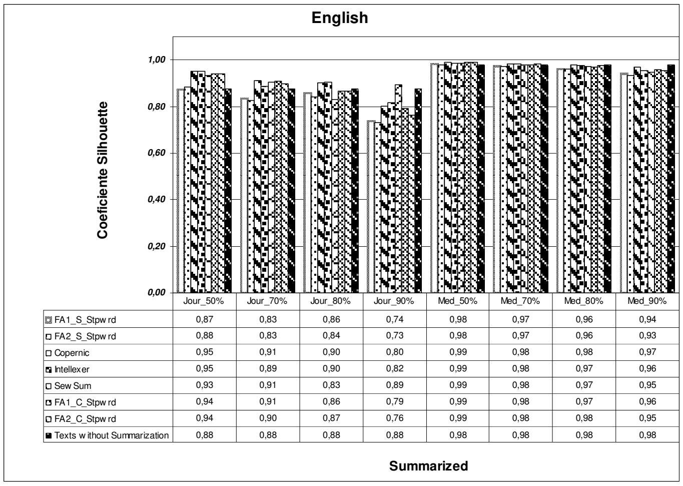
For the journalistic domain, shown in Figure 5, with summarization algorithms of 50% and 70% compression, the text clusters produced in the Cassiopeia model improved with Silhouette Coefficient values greater than unsummarized text clusters.

The exceptions are: for a 50% compression, function FA1_sem_Stopword and for a 70% compression, functions FA1 and FA2_sem_Stopword. With 80%, only two summarization algorithms improved the clusters generated by the Cassiopeia model; thus the Silhouette Coefficient values increased. With 90%, only one increased.

In the medical domain, by increasing compression of summarization algorithms, clusters were obtained with decreased Silhouette Coefficient values. With 50% compression, only one summarization algorithm displayed clusters with Silhouette Coefficient smaller than unsummarized text clusters. With 70%, there are two; with 80%, there are five; and with 90% there are all.

[2] Summarization algorithms were explained in item 3.2. They are used during pre-processing in the Cassiopeia model in order to summarize texts. Compression of 50%, 70%, 80% and 90% were used. The results of these summarizations are long texts which are alternately clustered by the Cassiopeia model. These clusters are evaluated by external and internal measures.

[3] Unsummarized texts do not suffer any compression. They are clustered by the Cassiopeia model and produce text clusters that are evaluated by external and internal measures. They are comparative parameters with summarized texts.

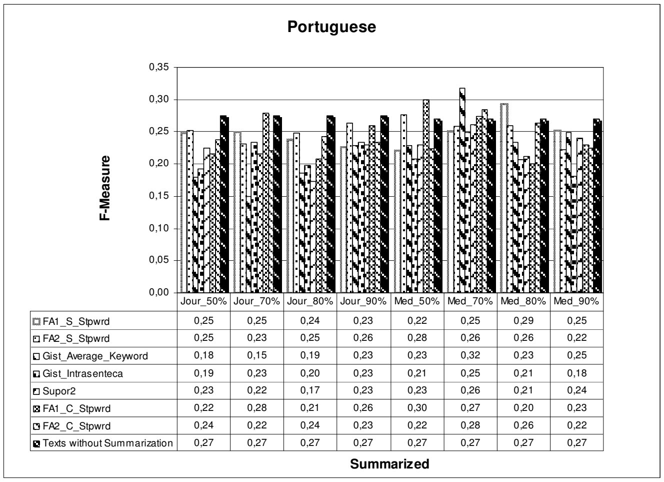


	Jour_50%	Jour_70%	Jour_80%	Jour_90%	Med_50%	Med_70%	Med_80%	Med_90%
▨ FA1_S_Stpw rd	0,87	0,83	0,86	0,74	0,98	0,97	0,96	0,94
□ FA2_S_Stpw rd	0,88	0,83	0,84	0,73	0,98	0,97	0,96	0,93
□ Copernic	0,95	0,91	0,90	0,80	0,99	0,98	0,98	0,97
▨ Intellexer	0,95	0,89	0,90	0,82	0,99	0,98	0,97	0,96
□ Sew Sum	0,93	0,91	0,83	0,89	0,99	0,98	0,97	0,95
□ FA1_C_Stpw rd	0,94	0,91	0,86	0,79	0,99	0,98	0,97	0,96
□ FA2_C_Stpw rd	0,94	0,90	0,87	0,76	0,99	0,98	0,98	0,95
■ Texts w ithout Summarization	0,88	0,88	0,88	0,88	0,98	0,98	0,98	0,98

Figure 5. Measures obtained in clusters generated by the Cassiopeia model in 100 interaction steps. The results are the accumulated average of the Silhouette Coefficient, in English, within the journalistic and medical domain. Results are shown for each compression percentage of 50%, 70%, 80% and 90% using the seven summarization algorithms and the whole unsummarized text.

Figures 6 and 7 displays the results of text clusters obtained by the Cassiopeia model, using F-Measure and Silhouette Coefficient, in Portuguese, in the journalistic and medical domains. Using all compressions (50%, 70%, 80% and 90%), seven summarization algorithms and the unsummarized text.

Figure 6 shows that summarization algorithms, for the most part, are not able to increase the F-Measure value in clusters obtained by the Cassiopeia model compared to clusters obtained with unsummarized texts. The only exception was function FA2_com_Stopword with a compression of 70%, as the value of clusters in the FMeasure increased.

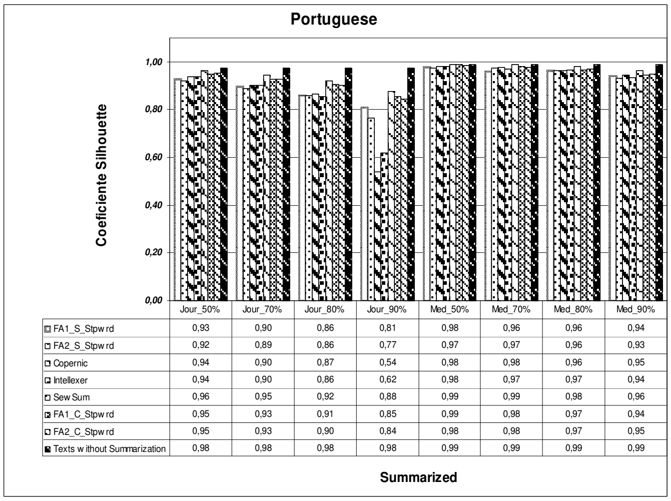


	Jour_50%	Jour_70%	Jour_80%	Jour_90%	Med_50%	Med_70%	Med_80%	Med_90%
▨ FA1_S_Stpwrd	0,25	0,25	0,24	0,23	0,22	0,25	0,29	0,25
□ FA2_S_Stpwrd	0,25	0,23	0,25	0,26	0,28	0,26	0,26	0,22
□ Gist_Average_Keyword	0,18	0,15	0,19	0,23	0,23	0,32	0,23	0,25
□ Gist_Intrasenteca	0,19	0,23	0,20	0,23	0,21	0,25	0,21	0,18
□ Supor2	0,23	0,22	0,17	0,23	0,23	0,26	0,21	0,24
▨ FA1_C_Stpwrd	0,22	0,28	0,21	0,26	0,30	0,27	0,20	0,23
□ FA2_C_Stpwrd	0,24	0,22	0,24	0,23	0,22	0,28	0,26	0,22
■ Texts without Summarization	0,27	0,27	0,27	0,27	0,27	0,27	0,27	0,27

Figure 6. Measures obtained in Clusters generated by the Cassiopeia model in 100 interaction steps. The results are the accumulated average of the F-Measure, in Portuguese, within the journalistic and medical domain. Results are shown for each compression percentage of 50%, 70%, 80% and 90% using the seven summarization algorithms and the whole unsummarized text.

In the medical domain, with 50% compression, two algorithms improved cluster values generated by the Cassiopeia model in the FMeasure in relation to the unsummarized text clusters. With 70% compression, only one; with 80% and 90%, none of the summarization algorithms managed to increase the F-Measure values of clusters obtained by the Cassiopeia model, if compared clusters generated with unsummarized texts.

For the journalistic domain shown in Figure 7, it was possible to observe that the summarization algorithms did not manage to increase the value of clusters generated by the Cassiopeia model measured by Silhouette Coefficient in relation to the unsummarized text.



	Jour_50%	Jour_70%	Jour_80%	Jour_90%	Med_50%	Med_70%	Med_80%	Med_90%
▨ FA1_S_Stpw rd	0,93	0,90	0,86	0,81	0,98	0,96	0,96	0,94
□ FA2_S_Stpw rd	0,92	0,89	0,86	0,77	0,97	0,97	0,96	0,93
□ Copernic	0,94	0,90	0,87	0,54	0,98	0,98	0,96	0,95
□ Intellexer	0,94	0,90	0,86	0,62	0,98	0,97	0,97	0,94
□ Sew Sum	0,96	0,95	0,92	0,88	0,99	0,99	0,98	0,96
▨ FA1_C_Stpw rd	0,95	0,93	0,91	0,85	0,99	0,98	0,97	0,94
□ FA2_C_Stpw rd	0,95	0,93	0,90	0,84	0,98	0,98	0,97	0,95
■ Texts w ithout Summarization	0,98	0,98	0,98	0,98	0,99	0,99	0,99	0,99

Figure 7. Measures obtained in clusters generated by the Cassiopeia model in 100 interaction steps. The results are the accumulated average of the Silhouette Coefficient, in Portuguese, within the journalistic and medical domains. Results are shown for each compression percentage of 50%, 70%, 80% and 90% using the seven summarization algorithms and the whole unsummarized text.

In the medical domain, two summarization algorithms with 50% compression improved the clusters generated by the Cassiopeia model measured by the Silhouette Coefficient, compared with the unsummarized text. With 70% compression, only one; with 80% and 90% compression, none of the summarization algorithms were able to increase the F-Measure values of clusters obtained with the Cassiopeia model.

V. Conclusion

By evaluating the results of external measures, where it is possible to observe in Figures 4 and 6 the use of the F-Measure, which is a harmonic measure of Recall and Precision and comparing clusters generated by the Cassiopeia model, using unsummarized texts, it is possible to conclude that, for the process of text summarization with 50% and 70% compression, the results were meaningful for measuring clusters. In Figure 4, above all, it is possible to note a slight decrease from the moment of compression increase, which seems to be consistent as text compression increases, there begins to be a loss of information, but this was not observed in Figure 6.

By evaluating the results of internal measures, it is possible to observe in Figures 5 and 7 the use of the Silhouette Coefficient measure, which is a harmonic measure of Cohesion and Coupling.

Comparing the clusters generated by the Cassiopeia model with the unsummarized texts, one may conclude that for the summarization process with 50% and 70% compression, the results of the clusters were also significant.

Figures 5 and 7 have similar behavior; i.e., one notes that with increase in text compression, there is a slight loss in the degree of information, which is reflected in clusters produced by the Cassiopeia model.

In a more general analysis, the results are coherent, since the domain, where rare or neologic words are more common in the Cassiopeia model, has an excellent performance, in the case of the medical domain. This observation does not apply to the lexicon-poor domain, where common words with high frequency, when summarized, lose information degree with increase in compression, which is the case of the journalistic domain. One can also highlight that the best results occur in English, due characteristics of the English language and/or good quality of English summarizers.

In final analysis, another extremely relevant factor is the agreement between external and internal measures. These results can be verified between the F-Measure and Silhouette Coefficient measures and can be observed comparing Figure 4 with Figure 6, Figure 5 and Figure 7. This is extremely positive for the Cassiopeia model, since both human assessment (external measure), and unsupervised assessment (internal measures) have similar results. This assures a good evaluation of the Cassiopeia model.

In conclusion, it is possible to ascertain, judging by the results obtained from the Cassiopeia model, that the contribution of including text summarization in the pre-processing stage and a new variation of the Luhn algorithm thresholds in the clustering process improves external and internal measures in its clusters. This contribution is observed in the interval from 50% to 70% (shown in this study). One can furthermore state that the results are sensitive to some variables such as: compression, domain, language, and summarization algorithms that directly influence the performance of these metrics in the Cassiopeia model.

A. *Future Research*

The selection of attributes is another way of testing the Cassiopeia model. These experiments have already been conducted and can be presented but were not considered due to the fact that this study focuses on contribution, regarding the impact of summarization, preceding the clustering process and its influences in measuring clusters This simulation possibility refers to the use of other forms of attribute selection that are used by other clustering mechanisms, such as: Ranking by Term Frequency (RTF), Ranking by Document Frequency (RDF) and Term Frequency, Inverse Document Frequency (TFIDF), which can be compared to the attribute selection method adopted by the Cassiopeia model.

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] Guelpeli, M. V. C. ; BernardinI, F. C. ; Garcia, A. C. B. “An Analysis of Constructed Categories for Textual Classification using Fuzzy Similarity and Agglomerative Hierarchical Methods”. Emergent Web Intelligence: Advanced Semantic Technologies .1st Edition., 2010, XVI, 544 p. 178 illus., Hardcover ISBN: 978-1-84996-076-2, 2010.

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

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

[8] Karanikas H. and. Theodoulious, B "Knowledge Discovery in Text and Text Mining Software," UMIST Department of Computation January 2002.

[9] 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.

[10] 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.

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

[12] 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

[13] Pardo, T.A.S. E 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.

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

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

[16] Ventura, J.M.J. “Extração de Unigramas Relevantes”. Dissertação apresentada na Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa para obtenção do grau de Mestre em Engenharia Informática. Lisboa ,Portugal, 2008.

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

[18] Wives, L.K.”Utilizando conceitos como descritores de textos para o processo de identificação de conglomerados clustering de documentos” – Tese (doutorado) – Universidade Federal do Rio Grande do Sul.Programa de Pós-graduação em Computação, Porto Alegre, Brasil, 2004.