2009 IEEE INTERNATIONAL CONFERENCE ON NATURAL LANGUAGE PROCESSING AND KNOWLEDGE ENGINEERING (IEEE NLP-KE'09)

CASSIOPEIA: A Model Based on Summarization and Clusterization used for Knowledge Discovery in Textual Bases

Marcus V. C. Guelpeli [1] Ana Cristina Bicharra Garcia[1] António Horta Branco[2]

[1]Departamento de Ciência da Computação Universidade Federal Fluminense ,UFF,Brasil
[2]Departamento de Informática da Faculdade de Ciências da Universidade de Lisboa,UL, Portugal

This work proposes the creation of a model called Cassiopeia, whose aim is to allow for knowledge discovery in textual bases in distinct and/or antagonic domains by using a process of summarization and clusterization to obtain these pieces of knowledge. By proposing the Cassiopeia model, we hope to obtain a better cohesion of the clusters and to make feasible the issue of high dimensionality in knowledge discovery in textual bases.

Index Terms — **Text mining, Knowledge Discovery, Summarization and Clusterization.**

I. INTRODUCTION

The increase in textual information, which result to a great extent from the ease we now have in generating and storing information in electronic media, and the subsequent difficulty in recovering this information has resulted in the advent of what is called information overload [29].

The upsurge in the volume of non-structured information is typical of the current times because the internet, as a repository of information and a great generator of new knowledge, has caused daily increases in textual information in a largely disorganized fashion. This quasi anarchical structure brought with a major problem of (informational) organization, which came about as a result of human beings' difficulty in storing large amounts of information.

Nevertheless, for some time, people did not consider the organization of this textual information important. However, studies have shown quite the opposite, indicating that these non-structured pieces of information, if effectively organized, could be used intelligently in a number of fields, thereby allowing for a competitive advantage or used to support decision-making processes [3].

It was from this competitive and organization vision that the field of Text Mining (TM) emerged. Its main purpose is to extract patterns or to infer some type of knowledge from a set of texts. TM is the execution of various processes in various stages, in a sequential and interactive manner, which transforms or organizes a given number of documents into a systematic structure. These qualities allow it to be later used in an efficient and intelligent manner [30].

Text Mining is new, multidisciplinary field that includes knowledge from areas such as Computing, Statistics, Linguistics and Cognitive Science, inspired by Data Mining, which searches for emerging patterns in structured databases. TM aims to extract usefully knowledge from non-structured or semi-structured data.

There are three major fields within TM, according to [2], [15], [16] and [30]: Information Extraction from texts (IE), Information Recovery from texts (RI) and Knowledge Discovery from Texts (KDT).

In this work, the main focus is the field of Knowledge Discovery from texts, which, according to [4], [27], [22] and [12] refers to the process of recovering, filtering, manipulating and summarizing the knowledge extracted from large sources of textual information, and present it to the end user by making use of a variety of resources, which usually differ from the original ones.

This work proposes the creation of a model called Cassiopeia, which will be able to provide knowledge discovery in textual bases in distinct and/or antagonic domains. The Cassiopeia model uses the process of clusterization to obtain this knowledge, thereby providing better cohesion in its cluster. One of the biggest problems found in the field of TM is high dimensionality. In order to circumvent this, the Cassiopeia model uses the summarization process to reduce the size of the texts to be manipulated in the clusterization process.

This paper is organized as follows. In Section 2, we present the Cassiopeia model with its processes of Summarization and Clusterization. Section 3 presents the methodology used in the simulation. Section 4 discusses the results obtained in the experiments. Finally, section 5 presents the conclusions and suggests future works.

II. THE CASSIOPEIA MODEL

The Cassiopeia model, illustrated in Figure 1, is formed by two processes: Summarization and Clusterization.

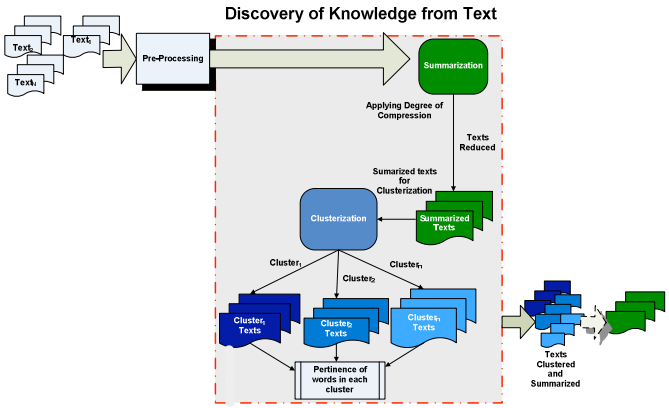

Fig.1. The Cassiopeia model.

The Cassiopeia model starts with text inputs for knowledge discovery. These texts go through the pre-processing phase, where they are prepared to be processed by a computer. This occurs by way of a technique called case folding, which puts all letters into small caps, as well as making other minor changes, such as removing all existing figures, tables and markings.

After undergoing this technique, the texts are in a format that is capable of being processed. In this pre-processing phase, one technique that is used is summarization. In the Cassiopeia model, the purpose of summarization is to decrease the number of words for the clusterization, which occurs in the processing phase. In this way, it tackles the issue of high dimensionality (a major issue in the field of TM), by not using the similarity matrix in the processing phase, as well as allowing stopwords to be maintained. These two factors – not using the similarity matrix and allowing stopwords – will be explained in greater detail in items A and B.

For summarization to occur in the Cassiopeia model, the degree of compression to be applied to each text is determined. This is a fixed number and is used by the Profile algorithm, defined in [5], which will also be covered in item A.

In the Cassiopeia model, the data analysis phase – in which the TM algorithms are applied for the use of techniques to generate knowledge based on information contained in a given text, as well as where knowledge extraction, which uses techniques to extract knowledge that is explicit in the text, occurs – does not take place only during processing. It also takes place in the summarization module, that is, in the pre-processing phase, as it used the Profile algorithm, which allows for this function.

After the texts are pre-processed, the processing phase is initiated, where the text is clustered, that is, placed in clusters based on similarity criteria, which will be explained in item B. Each of the clusters that are created possesses a word vector, known as cluster centroids, which contains highly relevant words that are pertinent to the clustered texts. By comparing the inputs of new texts, clusters, sub-clusters or even a fusion of these two may emerge [16].

Because of dimensionality, the word vectors adopt a similarity threshold, according to [29], which is an important point in the solution to the issue of high dimensionality. The reason for this threshold will be explained in greater detail in item B, but during the processing phase, it can undergo variations until it reaches a stabilizing value, i.e. the degree of pertinence of each word in each cluster, as illustrated in Figure 1.

These clusters are organized in a top-down hierarchical manner. Reclusterization occurs up until the moment in which the word vectors of each cluster become stable, i.e., until no further modification takes place.

After the end of the processing phase, the post-processing phase takes place, where each of clusters or sub clusters will contain, based on similarity, a set of summarized texts.

A. *Summarization in the Cassiopeia Model*

Text summarization is a process that aims to create a shorter version of an original text [18].

The need to simplify and summarize exists due to the increase in the volume of information available in the media, coupled with the shortage of time needed to read texts of various natures. As a consequence of this process, readers are unable to absorb all of the content matter of the original texts. Hence, the summary is a shorter version of the text, the purpose of which is to capture the author's main idea and convey it to the reader in a few short lines.

The summarization module of the Cassiopeia model uses as part of its composition the extraction and transposition of sentences, respecting their position in the original text (the superficial approach) and also adopts rules created by [9] to classify the original text (the deep approach), with a basis on the pragmatic profile[10].

The summarization module uses the texts that are classified by the Profile algorithm according to formality and temporality, based on the style rules established by [9].With this, the Profile algorithm determines the degree of automatic compression of the text that will be used to obtain the summary. Summarization using the Profile algorithm, according to [5] and [8], is formed by the phrases with the greatest frequency of words determined by the Profile algorithm. The Profile algorithm can be represented, very generally, as follows:

The Profile Algorithm:

1. Represents the text in vector form;
2. Calculates from the text: the number of sentences, the number of words, the average number of words per sentence and classify the sentences in groups, either LOW, NORMAL or HIGH.
3. Determines a heuristic based on temporality and Hovy formality to either eliminate or maintain sentences in the text that is to be summarized, based on each of the three groups, LOW, NORMAL or HIGH.

It is worth highlighting that this work allows for the maintenance of stopwords within the sentences, as opposed to other works in this field, which are forced to remove stopwords in order to decrease the processing volume.

Stopwords are a closed class of words that according to the literature [17], [25], [19], [14], [20] do not carry meaning. These classes are made up of articles, pronouns, interjections and prepositions. The maintenance of stopwords in the Cassiopeia model is important and decisive, as it leads to a non-dependency of language.

The strategy of maintaining the stopwords is only viable because the Cassiopeia model first conducts summarization, thereby considerably decreasing the processing of words in the text, then selects the characteristics of the words in the text, using relative frequency and, finally, employs a similarity threshold. The second and third processes are explained in detail in item B.

B. Clusterization in the Cassiopeia Model

Text clusterization is used when the elements of the available domain are unknown and, hence, the aim is to automatically divide the elements into clusters based on some criterion of affinity or similarity [16]. Clusterization aids in the process of knowledge discovery in texts, facilitating the identification of patterns within clusters [24].

For the Cassiopeia model, this type of discovery is extremely relevant, because it occurs in the identification of "interesting" clusters, which could potentially lead to the discovery of some useful piece knowledge.

The identification of clusters by of their characteristics, known as cluster analysis, is important for the Cassiopeia model, seeing as the texts are clustered by an evaluation of the similarities among them. This evaluation occurs in the three phases described below.

These three phases were proposed by [6] and [8], in the clusterization module of the Cassiopeia model, whose aim is to group together textual documents that have already been summarized.

First Phase - (Selection of characteristics): relative frequency is used to select the word characteristics in the text. Relative frequency determines the importance of a term according to the frequency in which it is found in the text. The more times a term shows up in the text, the more important it is for that text. It is calculated using equation (1) [26]. This formula normalizes the result of the absolute word frequency, making sure that small documents are not represented by small vectors and large documents by large vectors.

Because of normalization, all documents will be represented by vectors of the same size.

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

Where $F_{real}X$ equals the relative frequency of *X*, $F_{abs}X$ equals the absolute frequency of *X*, that is, the number of times that *X* shows up in the document, and *N* is the total number of words in the text.

Considered a vectorial-space, each word represents a dimension (hence, there are as many dimensions as there are unique words in the text). In this manner, this issues starts being dealt with during the summarization process, where the space of dimensionality is significantly reduced. Afterwards, it is once again dealt with in the first phase of this process, in which the word characteristic is selected in the text using relative frequency. This complements the problem of high dimensionality, where there is a threshold or similarity threshold [29], in which the words (characteristics) with importance (frequency) inferior to this value are simply ignored in the composition of the word vectors in the text.

Second phase - ***(Similarity Calculation):*** this phase identifies the similarity between the texts (characteristics selected in the first phase). For this phase, a measure of fuzzy similarity was used, set theoretic inclusion [1], in which the presences of words in both texts are compared. This fuzzy value represents the degree to which an element is included in the other text or the degree of similarity between them.

If the word appears in both texts, the value of one (1) is added to the counter; if it doesn't, zero (0) is added. In the end, the degree of similarity is a fuzzy value between 0 and 1, calculated by the average, that is, the total value of the counter (common words) divided by the total number of words in both texts (without counting repeated words).

The fact that a word is more important in one text or the other, as it may appear in different frequencies in each text, is not taken into consideration. This problem can be resolved, in part, by another function [23], which takes the average with fuzzy operators, which are similar to the above expect they use weights for the words. In this way, the fact the the words appear in the texts with difference levels of importance is taken into account. In this case, the weights of each word are based on the relative frequency. The similarity value is calculated by taking the average of the average weights of the words the texts have in common. That is, when a word appears in both documents, the average of their weights is summed, as opposed to using the value one (1). In the end, the average is calculated from the total number of words found in both documents.

Third Phase - (Agglomerative Hierarchical Method): the third phase used the agglomerative hierarchical method which defines the number of previous clusters by analyzing the constructed dendograms. Using the *Clicles* algorithm, according to [13], you can identify groups of texts by specifying some kind of relationship rule to create clusters based on the similarity analysis of the textual terms. In this way, according to [8], the Clicks algorithm is able to construct more cohesive clusters.

The use of the summarization module along with these three phases of clusterization allows the Cassiopeia model to not resort to the use of the similarity matrix (Table I), which is the crucial point of high dimensionality within the field of Text Mining, as the similarity matrix grows exponentially with the text base [28].

TABLE I.
SIMILARITY MATRIX.

	$Text_1$	$Text_2$	$Text_3$	$Text_N$
$Text_1$	1.0	0.3	0.2	0.7
$Text_2$	0.3	1.0	0.5	0.3
$Text_3$	0.7	0.5	1.0	0.3
$Text_n$	0.6	0.3	0.3	1.0

III. SIMULATION METHODOLOGY

For the experiments with the Cassiopeia model proposed in this work as well as the Eurekha Categorizer created by [29] and [30], the following Corpuses were used: TeMário [21] and UBM Notícias[1].

The TeMário corpus 2006 is an extended version of the 2004 TeMário Corpus produced by the Interinstituional Nucleus of Computational Linguistics at the Federal University of São Carlos. All texts originnate from the online edition of the daily newspaper Folha de São Paulo and they are distributed among 7 sections: Brazil, Daily Life, Finances, Special Topics, World, Opinion and Everything. The 2004 version of TeMario 2004 contained 100 texts and the current version, TeMario 2006, was extend by another 150 texts, totalling 250 texts in the Portuguese language.

The UBM – Notícias corpus is composed of journalistic texts produced between 2007 and 2008. The texts are news articles from the Barra Mansa University Center from the most diverse academic activities, with a grand total of 35 sections and 456 texts.

To evaluate the performance of the Cassiopeia model and of the Eurekha categorizer, four metrics were chosen. They are:

Recall(*R*): number of sentences in the automatic summary that are present in the reference summary / number of sentences in the automatic summary;

Precision(*P*): number of sentences in the automatic summary that are present in the reference summary / number of sentences in the reference summary;

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

Where n is the number of news articles in cluster P, *Sim* is the similarity ratio and each p is a member of cluster P.

$$\textit{Coupling (Cp):} \quad \frac{\sum_{i>j} Sim(C_i, C_j)}{n(n-1)/2} \tag{3}$$

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

As a comparison to the Cassiopeia model, we selected the Eurekha clusterizer created by [29], [30]. The reason for this choice is the fact that Eurekha uses the same graphotheoretical algorithm (*Clicles*) and works with fuzzy logic, albeit with a different fuzzy function and different approaches. Eurekha uses the fuzzy function to calculate similarity, which takes into account the differences and similarities between the documents. Cassiopeia, on the other hand, is based on the idea of text similarity in relation to the cluster. In the Cassiopeia model, the pre-processing phase uses summarization to reduce the number of words (as described in detail in item A), whereas Eurekha excludes stopwords. In the processing phase, Eurekha uses a similarity matrix that calculates the similarity between the texts. Cassiopeia, however, does not use a similarity matrix and, instead, uses word vectors, comparing more frequenly appearing words (described in item B).

In [7], qualitative analyses of the constructed clusters were made for the Cassiopeia model and for Eurekha. The following Corpuses were used Temário, Distribution 1.0 and Really Simple Syndication (RSS)[2] and Reuters-215782[3], obtaining other results and studies such:

[1]UBM Notícias are journalistic texts from the press relations organ of Barra Mansa University Center. Barra Mansa University is a higher education institution and the texts are produced by the Nucleus of Social Communication, under the responsibility of journalist Renata Nery. For more information, visit www.ubm.br.

[2] These files, which come from a variety of RSS channels from Terra Networks Brasil S/A, were collected on a daily basis during the period comprising February and March 2008.

[2]*Corpus* extracted from Terra Networks Brasil S/A.

[3] The complete collection has 1578 texts, although these files were not entirely available for use. Hence, we only used the 100 texts that were available online.

As variation analysis, to test differences between average samples and between linear combinations of the averages.

Hypotheses test: used for the collected samples in the simulation of the Cassiopeia model and of Eurekha in the Temário corpus, Reuters corpus and RSS_Terra.

The recall and precision metrics were used.

The results presented in this work, however, are based on larger and more up-to-date journalistic corpuses. The metrics also differ, with Cohesion Equation 2 and Coupling Equation 3, as proposed by [11]. The study of the metrics (equations 2 and 4) is the main purpose of this study. According to [11] the best situation between these two metrics would be an increase in cohesion and a decrease in coupling. This could be understood here as the attainment of greater cohesion among texts in the clusters and a lower amount of coupling between the clusters.

Due to the use of these metrics (Equations 2 and 3), a new study was used involving averages, standard deviation, variance and the Pearson variation coefficient, (C_v). Since the Pearson variation coefficient is a relative measure of dispersion – as opposed to standard deviations, which is a measure of absolute dispersion – it is understood in descriptive statistics that standard deviation on its own has many limitations. Hence, this work considers dispersion or variability of date in relative terms to its average value, that is, using the Pearson variation coeffecient shown in Equation 4.

$$C_v = \frac{\sigma}{\overline{X}} * 100 \qquad (4)$$

Where C_v is the Pearson variation coefficient, σ is the standard deviation of the data in the series and $\overline{X}$ is the average of the data in the series. Multiplication turns the resulting value into a percentage *(%)*.

A relevant question for this work is: by using the Pearson variation coefficient, which of the two samples generated by Eurekha and by Cassiopeia has greater cohesion, that is, which is more homogenous? Since it is impossible to answer this question using standard deviation, as it is a measure of absolute dispersion, it is necessary to calculate the C_v of both series. The series that shows less variation – in other words, the one with the lowest the value of C_v – will be the one that contains the highest level of homogeneity between the texts in the cluster (a lower degree of dispersion means greater cohesion). For the coupling metric, on the other hand, the greater the C_v the better the result, indicating a lower level of homogeneity between the clusters (a greater degree of dispersion means less coupling).

IV. RESULTS OBTAINED IN THE EXPERIMENTS

In the first simulation the TeMário 2006 corpus was used. A number of 35 clusters were obtained with Eurekha and with the Cassiopeia model. This number of clusters was obtained automatically by Cassiopeia and with Eurekha, there was a need to adjust the value to get to this number of clusters. Table II illustrates the results showing the standard deviations, the averages, its variations and its Pearson variation coefficients.

TABLE II
RESULT OF THE PEARSON COEFFICIENT VARIATIONS OF EUREKHA AND CASSIOPEIA USING THE TEMÁRIO CORPUS

	Eurekha			Cassiopeia		
	Recall	Precision	Cohesion	Recall	Precision	Cohesion
Standard Deviation	3,60	13,65	0,04	6,17	26,89	0,01
Average	8,71	95,29	0,06	7,43	67,23	0,04
Variance	12,97	186,33	0,00	38,08	722,95	0,00
C_V	41%	14%	66%	83%	40%	36%

It is clear from the C_v that the sample from Eurekha obtained a lower variation percentage in the Recall and Precision metrics, which were better than in the Cassiopeia model. However, for this study, the main point of interest is the cohesion in the texts in the obtained clusters and the coupling between clusters.

In analysing the cohesion metric, Cassiopeia obtained superior results, showing a lower variation among its sample. In other words, the texts in the clusters from the Cassiopeia model are more cohesive than the clusters in Eurekha, whose sample showed greater dispersion in comparison to Cassiopeia. This dispersion can be visualized in Figure 2, which shows a greater homogeneity in the Cassiopeia sample. Hence, it is clear, statistically, that there is greater cohesion among texts from the clusters from Cassiopeia

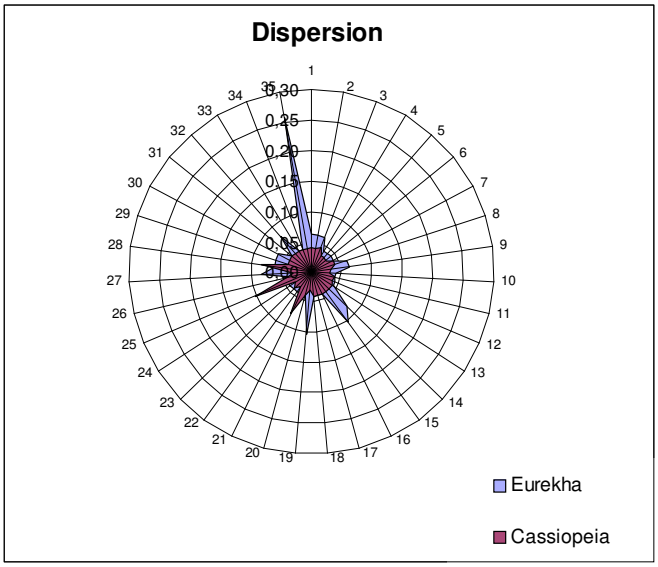

Fig. 2. Illustration of the sample dispersion of the Cassiopeia model and of Eurekha using the TeMário corpus.

Table III illustrates the result of the coupling metric, where the Cassiopeia model obtained a more significant Cv value than Eurekha. However, as explained in section 3, with regards to the coupling metric, this is not desirable. Hence, in regards to the coupling metric in this sample, Cassiopeia achieved inferior results as compared to Eurekha.

TABLE III
RESULT OF THE COUPLING OF THE CLUSTERS FROM EUREKHA AND CASSIOPEIA USING THE TEMÁRIO CORPUS

Eurekha		Cassiopeia	
Average	0,055565361	Average	0,039338114
Standard Deviation	0,038854203	Standard Deviation	0,014849045
Centroids	0,938367347	Centroids	0,342040817
Coupling	0,052131519	Coupling	0,019002268
C_V	70%	C_V	38%

The second simulation was conducting using the UBM Notícias corpus was used. A total of 53 clusters were obtained by Eurekha and the Cassiopeia model. Once again, it was necessary to adjust values in Eurekha to get to this number of clusters. Table IV shows the results, indicating the standard deviations, averages, variances and Pearson variation coefficients.

TABLE IV
RESULT OF THE PEARSON VARIATION COEFFICIENT OF EUREKHA AND CASSIOPEIA USING THE UBM NOTICIAS CORPUS

	Eurekha			Cassiopeia		
	Recall	Precision	Cohesion	Recall	Precision	Cohesion
Standard Deviation	11,18	25,63	0,01	5,97	26,82	0,00
Average	14,81	78,45	0,04	8,77	66,00	0,03
Variance	125,00	656,98	0,00	35,60	719,23	0,00
C_V	75%	33%	33%	68%	41%	16%

As shown in the table, the sample from the Cassiopeia model with the Pearson variation coefficient obtained a lower variation percentage in the Recall, Precision and Cohesion, which is better than the results from Eurekha. Nevertheless, the emphasis and central aims of this work is cohesion and coupling.

In the cohesion metric the Cassiopeia model once again shows more cohesion than results obtained with Eurekha. Figure 3 shows that Cassiopeia obtained lower degrees of dispersion in its sample, thereby indicating the the clustered texts are more homogenous, i.e., more cohesive.

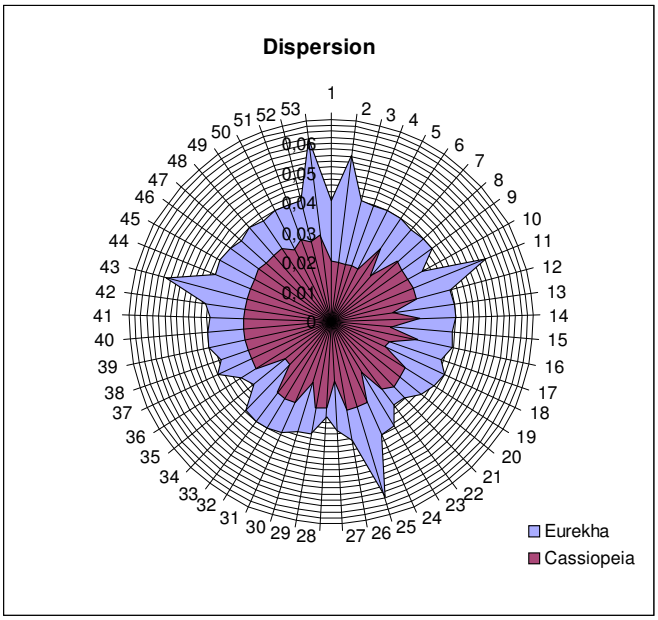

Fig. 3. Illustration of the dispersion of the sample of the Cassiopeia model and of Eurekha using the UBM Notícia corpus

Table V illustrates the result of the coupling metric, where Eerekha obtained a more significant C_v value than Cassiopeia. However, as explained in section 3, for the coupling metric this is not good. As such, in the coupling metric in this sample, Cassiopeia obtained better results than Eurekha.

TABLE V
RESULT OF THE COUPLING OF CLUSTERS OF EUREKHA AND CASSIOPEIA USING THE UBM NOTÍCIAS CORPUS

Eurekha		Cassiopeia	
Average	0,005934621	Average	0,00422515
Standard Deviation	0,038918725	Standard Deviation	0,02497673
Centroids	1,880816331	Centroids	1,11959183
Coupling	0,016793003	Coupling	0,00999636
C_V	15%	C_V	17%

V. CONCLUSION

The results from the simulation regarding the cohesion and coupling metrics are very significant, when one considers that the main purpose of this work is to show improved cohesion in texts distributed in clusters and the level of coupling within clusters. According to [11], the most desirable outcome in clusterization would be a high cohesion rate and a low coupling rate.

Finally, it must be noted that the recall and precision metrics are perhaps not the most appropriate for measuring the efficiency of a clusterizer, as the Cassiopeia model and Eurekha would need to be used as search engines for this, where the word would be used to formulate the search and it would be used to recover texts with greater similarity in relation to the formulated search. Only in this case would

recall and precision be more coherent.

C. *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.

The issue of the corpus is another detail worth looking at in future tests, in which there are other simulation possibilities: usage in other languages, with high and low correlations, and with a greater number of texts. The only problem is related to the need of these corpuses to be previously classified by specialists in order to facilitate comparisons and thereby validate the cohesion and coupling metrics discussed in this work.

REFERENCES

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

[2] Dagan, I., Feldman, R, e Hirsh, H, "Keyword-based browsing and analysis of large document sets". In Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval (SDAIR), Las Vegas, NV,1996.

[3] Ebecken, N. F. F. e Lopes, M C S. Costa, M. C. A,"Mineração de Textos." In:. Sistemas Inteligentes: Fundamentos e Aplicações. Manole: São Paulo, 2003.

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

[5] Guelpeli, M.V.C.; Garcia A.C.B, Automatic Summarizer Based on Pragmatic Profiles", International Conference WWW/Internet 2007- IADIS- Volume II pág. 149-153- ISBN: 978-972-8924-44-7 - mês de Outubro de 2007- Vila Real-Portugal .

[6] Guelpeli, M.V.C.; Garcia A.C.B, "An Analysis of Constructed Categories for Textual Classification Using Fuzzy Similarity and Agglomerative Hierarchical Methods", proceedings of The 2007 IEEE International Conference on SIGNAL-IMAGE TECHNOLOGY and INTERNET- BASED SYSTEMS. Shanghai, China 16 a 19 de December, 2007.

[7] Guelpeli, M.V.C.; Garcia A.C.B. Bernadini, F. C,"An Analysis of Constructed Categories for Textual Classification using Fuzzy Similarity and Agglomerative Hierarchical Methods." an Proceding. An Analysis of Constructed Categories for Textual Classification using Fuzzy Similarity and Agglomerative Hierarchical Methods, LNCS- Advanced Information and Knowledge Processing. Berlim: Springer Verlag, 2008

[8] 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, 2008, Vila Velha. Todas as Palavras da Sentença como Métrica para um Sumarizador Automático. Vila Velha : WebMedia, 2008. p. 287-291.

[9] Hovy, E.; Lin, C,"Automated Text Summarization in SUMMARIST." In: I. Mani and M. Maybury (1997.) Intelligent Scalable Text Summarization ACL 1997 Workshop, pp. 39-46. Madrid, Spain.

[10] Hovy, E.; C.Y. Lin; L. Zhou, "A BE-based multi-document summarizer with sentence compression." Proceedings of Multilingual Summarization Evaluation (ACL 2005), Ann Arbor, MI.

[11] Kramer, S.; Kaindll, H. "Coupling and cohesionmetrics for knowledge-based systems using frames and rules." ACM trans. Softw. Eng Methodol, New York, NY, USA, v13 n.3, p332-358, 2004.

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

[13] Kowalski, G, ".Information Retrieval Systems: Theory and Implementation" Boston: Kluwer Academic, 1997. 282 p.

[14] Larocca, J. N.1, Santos,A. D. S, Kaestner ,C. A.A. e Freitas A. A, "Generating Text Summaries through the Relative Importance of Topics." Lecture Notes in Computer Science Springer Berlin / Heidelberg Volume 1952/2000, ISSN0302-9743 (Print) 1611-3349 (Online) pp 300, 2000, Brazil.

[15] Loh S, "Descoberta de Conhecimento em Textos." Universidade Federal do Rio Grande do Sul-Instituto de Informática-Curso de Pós-graduação em Ciência da Computação. Exame de Qualificação- EQ-29 CPGCC-UFRGS, 1999.

[16] Loh, S, "Abordagem Baseada em Conceitos para Descoberta de Conhecimento em Textos." Porto Alegre: UFRGS, 2001. Requisito Parcial ao Grau de Doutor em Ciência da Computação, Instituto de Informática, Universidade Federal do Rio Grande do Sul, 2001.

[17] Marcu, D, "From Discourse Structures to Text Summaries." In I. Mani and M. Maybury (eds.), Proc. of the Intelligent Scalable Text Summarization Workshop, pp. 82-88. ACL/EACL'97 Joint Conference. Madrid, Spain.

[18] Mani, I.; Maybury, M.T. (1999). Advances in automatic text summarization. MIT Press, Cambridge, MA.

[19] Mittal,V. O.,Kantrowitz,M.,Goldstein,J.,Carbonell,J. G.(1999) Selecting Text Spans For Document Summaries: Heuristics And Metrics ,In Aaai/Iaai (1999), Pp. 467-473.

[20] Pardo, T.A.S.; Rino, L.H.M.; Martins A, "Coleção TeMário e a Avaliação de Sumarização Automática." Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional, NILC - ICMC-USP, Janeiro de 2006.

[21] Pardo, T.A.S.; Maziero, E. G., Nunes, M.G.V., Uzeda, V. R, " TeMário 2006: Estendendo o Córpus TeMário" Relatório Técnico número NILC-TR-07-06, Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional, NILC - ICMC-USP, Outubro de 2007.Disponível em http://www.nilc.icmc.usp.br/nilc/tools/corpora.htm acessado em 20 de Março de 2009.

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

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

[24] Rizzi, C.; Wives, L. K.; Engel, P. M.; Oliveira, J. P. M, " Fazendo Uso da Categorização de Textos em Atividades Empresariais". In: International Symposium on Knowledge Management/Document Management (ISKM/DM 2000), III, Nov, 2000.

[25] Sparck J. K, "Automatic Summarizing: factors and directions." In I. Mani and M. Maybury (eds.), Advances in automatic text summarization, The MIT Press, pp. 1-12.

[26] Salton, G,"Introduction to Modern Information Retrieval." New York: McGraw-Hill, 1983.

[27] Tan, A. H, "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).

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

[29] Wives, L.K, "Um estudo sobre Agrupamento de Documentos Textuais em Processamento de Informações não Estruturadas Usando Técnicas de "Clustering"."Porto Alegre: UFRGS, 1999. Dissertação (Mestrado em Ciência da Computação), Instituto de Informática, Universidade Federal do Rio Grande do Sul, 1999.

[30] Wives, L.K,"Tecnologias de Descoberta de Conhecimento em Textos Aplicadas à Inteligência Competitiva" Universidade Federal do Rio Grande do Sul-Instituto de Informática-Curso de Pós-graduação em Ciência da Computação.Exame de Qualificação- EQ-069 PPGC-UFRGS, 2002.