JISTEM - Journal of Information Systems and Technology Management
Revista de Gestão da Tecnologia e Sistemas de Informação
Vol. 11, No. 2, May/Aug., 2014 pp. 415-436
ISSN online: 1807-1775
DOI: 10.4301/S1807-17752014000200011

AUTOMATED TEXT CLUSTERING OF NEWSPAPER AND SCIENTIFIC TEXTS IN BRAZILIAN PORTUGUESE: ANALYSIS AND COMPARISON OF METHODS

Alexandre Ribeiro Afonso
Cláudio Gottschalg Duque
University of Brasília (Universidade de Brasília–UnB), Brasília, DF, Brazil

ABSTRACT

This article reports the findings of an empirical study about Automated Text Clustering applied to scientific articles and newspaper texts in Brazilian Portuguese, the objective was to find the most effective computational method able to cluster the input of texts in their original groups. The study covered four experiments, each experiment had four procedures: 1. *Corpus Selections* (a set of texts is selected for clustering), 2. *Word Class Selections* (Nouns, Verbs and Adjectives are chosen from each text by using specific algorithms), 3. *Filtering Algorithms* (a set of terms is selected from the results of the preview stage, a semantic weight is also inserted for each term and an index is generated for each text), 4. *Clustering Algorithms* (the clustering algorithms Simple K-Means, sIB and EM are applied to the indexes). After those procedures, clustering correctness and clustering time statistical results were collected. The sIB clustering algorithm is the best choice for both scientific and newspaper corpus, under the condition that the sIB clustering algorithm asks for the number of clusters as input before running (for the newspaper corpus, 68.9% correctness in 1 minute and for the scientific corpus, 77.8% correctness in 1 minute). The EM clustering algorithm additionally guesses the number of clusters without user intervention, but its best case is less than 53% correctness. Considering the experiments carried out, the results of human text classification and automated clustering are distant; it was also observed that the clustering correctness results vary according to the number of input texts and their topics.

Keywords: Text Mining; Text Clustering; Natural Language Processing; Brazilian Portuguese; Effectiveness.

Manuscript first received/*Recebido em:* 01/01/2013 Manuscript accepted/*Aprovado em*: 20/05/2014

Address for correspondence / Endereço para correspondência

Alexandre Ribeiro Afonso is doctoral-degree student in the Information Science program (Faculdade de Ciência da Informação – FCI), University of Brasília (Universidade de Brasília–UnB), Brasília - DF, Brazil. E-mail: rafonso.alex@gmail.com

Cláudio Gottschalg Duque works as professor and researcher in the Information Science program (Faculdade de Ciência da Informação – FCI), University of Brasília (Universidade de Brasília–UnB), Brasília - DF, Brazil. Campus Universitário Darcy Ribeiro, Faculdade de Ciência da Informação,Edifício da Biblioteca Central, Entrada Leste, Brasília, DF- Brazil. CEP: 70.919-970. Phone Number: 55(61)3107-2632. E-mail: klaussherzog@gmail.com

1. INTRODUCTION

Automated text clustering systems have been developed and tested as an experimental and scientific activity. The purpose of text classification automation is to be as effective as humans when classifying texts in knowledge fields. A previous effective automatic clustering over a set of documents contributes to an effective automatic or manual information retrieval.

The main difference between clustering and categorization systems is that the clustering systems do not utilize any formal knowledge (like ontologies or thesauri) for training previously the system; instead, it works as an unsupervised learning system (Manning at al., 2008).

An automatic text clustering process could be divided in four main stages: *Corpus Selection, Word Class Selections, Filtering Algorithms* and *Clustering Algorithms*; during the experiments, we applied different procedures for each stage described to find the best combination of procedures which produced correct textual clustering by consuming less time, both for newspapers and scientific texts in Brazilian Portuguese.

Clustering Algorithms have been developed for general use, for all languages, but they have been tested mainly for English, and many studies about text clustering, using corpora in English as input, have been described over the last decade. However, different natural languages could produce different levels of correctness in clustering results, since each natural language has specific structures and properties (such as morphological and syntax peculiarities) with different levels of complexity in their use (number of repetitions of words in newspaper texts, number of synonyms, use of idiomatic expressions, and terminologies).

Some authors have described the impact of the linguistic characteristics over classification and information retrieval systems. For example, Rossel and Velupillai (2005) investigated the impact of using phrases in the vector space model for clustering documents in Swedish in different ways. Stefanowski and Weiss (2003) consider the problem of web search results clustering in the Polish language, supporting their analysis with results acquired from an experimental system named *Carrot*. Basic, Berecek and Cvitas (2005) argue that text processing algorithms and systems in English and other world languages are well developed, which is not the case with Croatian language, they affirm that the quality of input data strongly influences clustering and classification results.

Another fact *very important* to be noted is that the format (news or scientific) and the content (History, Geography, Pharmacy, etc.) of the corpus could produce different results for clustering experiments. This verification was performed during our second experiment.

All these cultural and linguistic particularities and the evident studies about the impact of the language over information retrieval systems make us to think about the scientific and newspaper communication in Brazil as having particular features. When analyzing the results of a clustering process, specifically for texts written in Brazilian

Portuguese, we can identify the best and not best procedures for each clustering stage, also observing their advantages and failures. Considering this observation, in future works, new procedures for text clustering can be proposed.

The goal of our study was to verify whether an automated clustering process could create the correct clusters for two text corpuses: a scientific corpus having five knowledge fields (Pharmacy, Physical Education, Linguistics, Geography, and History) and a newspaper corpus having five knowledge fields (Human Sciences, Biological Sciences, Social Sciences, Religion and Thought, Exact Sciences). Therefore, we had two corpuses already classified by humans and we wanted to measure the effectiveness of the clustering process (clustering correctness and clustering time) by using the human classification as reference for correctness. Then, using a statistical method, we searched for the combinations of clustering stages that produced the best clustering correctness values and the shortest clustering time values.

Through the following sections, we describe in detail each experiment stage (*corpus selections*, *word class selections*, *filtering algorithms*, and *clustering algorithms*), the procedures for each stage, the statistical methods adopted for clustering correctness and time measurement, statistical analysis, and results.

2. TEXT CLUSTERING IN BRAZILIAN PORTUGUESE: LITERATURE REVIEW

2.1 Text clustering approaches

In this article, we focus on a non-hierarchical text clustering method as opposed to hierarchical text clustering approach. Non-hierarchical text clustering is applied when the goal is to produce text clusters which do not fit in a specific knowledge hierarchy; it means that each text is only inside a specific cluster, a text would not be grouped in two or more clusters at the same time (Markov & Larose, 2007). When a hierarchy is necessary to organize the texts, the hierarchical approach is able to group, for example, two related clusters inside a major cluster such as taxonomy. Some scientific fields like Medicine are structured in low level sub-fields and a hierarchical approach could be well applied. But, when the user needs to group the documents in major scientific areas (for example, Geography, Linguistics, Pharmacy, etc.), a non-hierarchical clustering method would be better applied because the fields do not have a high level of related terminological characteristics, or the fields/areas are independent.

2.2 Non-hierarchical text clustering in Brazilian Portuguese

Our interest is to study the effect of the well-known text clustering technology (specifically, non-hierarchical text clustering methods) over a Brazilian digital repository where the texts are written in Brazilian Portuguese. Many possibilities for the four levels of a clustering experiment (*corpus selections*, *word class selections*, *filtering algorithms*, and *clustering algorithms*) were not tested yet for this language; so we execute new clustering experiments testing new corpora formats, filtering algorithms and using the well-known clustering algorithms.

The three articles described below (Maia & Souza, 2010), (DaSilva et. al, 2004), and (Seno & Nunes, 2008) are closely related to our work. They test and evaluate the effectiveness of traditional non-hierarchic text clustering algorithms by using scientific and newspaper corpora in Brazilian Portuguese.

The study reported by Maia and Souza (2010) is the most similar to our study. The study is about Automatic Text Categorization and Automatic Text Clustering in Brazilian Portuguese. The study regarding clustering methods aimed to compare the use of noun phrases and single terms as text representations for Simple K-Means clustering algorithm, trying to find the best linguistic representation. The researchers analyzed the effectiveness of the Simple K-Means clustering algorithm when clustering a corpus with 50 scientific texts about subareas of Information Science. The corpus was divided into 5 Information Science subareas (Historical and Epistemological Studies about Information; Knowledge Organization and Information Representation; Information Propagation, Mediation and Usability; Politics, Ethics and Information Economy; Management of Information Units) and the researchers aimed to evaluate the results from Simple K-Means when clustering texts indexed by noun phrases or terms. The manuscript also describes the same process for a newspaper corpus having 160 texts of 4 newspaper sections (informatics, tourism, world, vehicles). About the best clustering results, Maia and Souza (2010) describe 44% of clustering correctness using single terms from the scientific corpus, and 81% of clustering correctness using noun phrases from the newspaper corpus. They concluded that the use of noun phases is not better than the use of single terms as a representation for clustering tasks, since the correctness rate for noun phases and terms has an approximated percentage number, but the use of noun phases consumed a longer time. About the difference between text clustering and text categorization, considering the best results for correctness, they observed 8 percentage points difference between automatic text clustering and automatic text categorization for the scientific corpus, and 10 percentage points difference for the newspaper corpus, the text categorization process got the best values.

A related study for Brazilian Portuguese was performed by DaSilva et al. (2004). The researchers propose and evaluate the use of linguistic information in the preprocessing phase of text mining tasks (categorization and clustering). They present several experiments comparing their proposal for selection of terms based on linguistic knowledge with usual techniques applied in the field. The results show that *part of speech information* (in this paper, we use the term *word classes*) is useful for the preprocessing phase of text categorization and clustering, as an alternative for stop words and stemming.

We could also cite the study of Seno and Nunes (2008) as a related work. The paper presents some experiments on detecting and clustering similar sentences of texts in Brazilian Portuguese. They propose an evaluation framework based on an incremental and unsupervised clustering method which is combined with statistical similarity metrics to measure the semantic distance between sentences. Experiments show that this method is robust even to treat small data sets. It has achieved 86% and 93% of F-measure and Purity, respectively, and 0.037 of Entropy for the best case.

3. AUTOMATED TEXT CLUSTERING: DESCRIPTION OF THE EXPERIMENTS' STAGES

Our study was performed by analyzing the statistical results from four text clustering experiments. The four experiments had the same architecture; it means the four experiments are composed by the same sequential stages. In this section, we describe the possibilities for the four sequential stages: *Corpus Selections*, *Word Class Selections, Filtering Algorithms,* and *Clustering Algorithms*.

Corpus Selections: We chose two textual databases to extract the newspaper and the scientific corpus. The scientific corpus was taken from five scientific journals published by the digital library of Federal University of Goiás (UFG) in Brazil. We got the articles from five scientific fields (Pharmacy, Physical Education, Linguistics, Geography, and History). The text choices from the scientific digital library were a random process.

The newspaper corpus was taken from the "Lácio-Web Textual Database" described by Aluísio and Almeida (2006). It is produced by NILC (Núcleo Interinstitucional de Linguística Computacional), a Computational Linguistics research group in Brazil. The texts from Lácio-Web database are classified in five areas (Human Sciences, Biological Sciences, Social Sciences, Religion and Thought, Exact Sciences). Lácio-Web stores texts from newspapers that are currently active in Brazil. The text choices from the textual database were a random process.

The content of the newspaper texts was not modified after corpus selection, but a text partition was executed for the scientific corpus after corpus selection. Since the scientific texts are larger texts, we took from each scientific text: title, abstract, keywords, and the first page/column of the introduction; the first page was selected since it was observed that in scientific texts the introduction, generally, fills no more than one or two pages/columns (considering the scientific fields analyzed), this choice permits to keep a small set of terms but considering the content of the introduction. An initial conjecture was that the terms which identify the topic are more frequent in these first parts of the scientific text, this verification is part of the experiments done.

Although some texts from the Lácio-Web Database can be classified in more than one area, we did not permit text replicas in our corpus. When a text was found inside two or more clusters in our corpus, it was replaced by another one, but the text chosen could still be replicated by many clusters in Lácio-Web database. The original topics of the scientific articles are chosen according to the scope descriptions of the journals in their web pages (there is a section "Foco e Escopo" for each journal which describes the topics of interest). The scientific database is found at (http://www.revistas.ufg.br/) and the newspaper Lácio-Web database can be found at (http://www.nilc.icmc.usp.br/lacioweb/).

Both formats (newspaper and scientific texts) have public access and they are commonly found in libraries, and since our goal was to evaluate the current technology over digital libraries, we decided to test these specific text formats. We worked with at

most five knowledge areas for each format during the experiments; if we chose more than five knowledge areas it would create a very large study for a single research. Another reason for this number is that we divided the original set of texts in two corpuses (having three and five knowledge areas) to verify whether, when the number of knowledge areas and the number of texts increase, the clustering effectiveness decreases. The criterion for choosing the knowledge areas to construct the corpora is the approximation of the knowledge areas and their distances; it means we decided to choose two areas inserted in the same major area (e.g. History and Geography) and other two areas not directly related (e.g. Physical Education and Linguistics) and one more area was chosen randomly. The same criterion was set for the newspaper texts (Social Sciences and *Religion and Thought*) and (Human Sciences and Biological Sciences).

The two sets of data produced after this manual processes were the experiment corpora that follow the original human classifications from the two databases.

Word Class Selections: We first applied (for each text from both corpuses) a POS-Tagger algorithm (Part-of-Speech tagger); it inserts tags (word class tags like Noun, Verb, Adverb, Adjective, etc.) to every word inside a text. The POS-Tagger used is trained by the researchers of the NILC institute, as described by Aires (2000), the MXPOST Tagger application (Ratnaparkhi, 1996) was used to produce this Brazilian Portuguese Tagger. This POS-Tagger was chosen because the taggers for (Nouns, Adjectives and Verbs) are inserted, and the tagger does not differentiate the hierarchies of the nouns, it also was trained by using a Brazilian Portuguese corpus. Moreover, the tagger was developed and tested by natural language processing specialists from NILC.

After applying the POS-Tagger algorithm for each text of the corpus and selecting Nouns, Verbs and Adjectives, we applied a stemming algorithm for each tagged text. The stemming process was performed by using an application developed by NILC; it is described by Caldas (2001). The stemmer follows Porter's algorithm and works for Brazilian Portuguese by identifying the stem of words by incrementally removing their suffix/termination. After tagging and stemming, each text from the scientific and newspaper corpus is represented only by their stems and their tags (only Nouns, Verbs and Adjectives). This stemmer was chosen because it was developed specifically for Brazilian Portuguese, and it was developed by natural language processing specialists from NILC.

Filtering Algorithms: In this experiment stage, we applied some algorithms for creating a text index for each text. The index must contain the main stems from the previous w*ord class selections* stage and it has to represent the text by holding its semantic meaning. So, the index produced by this stage is a reduced set of stems also having a semantic weight for each stem. The semantic weight algorithm executes a mathematical function that returns the stem's semantic weight according to its semantic importance inside the text and the entire corpus. For the four experiments, we applied the mathematical function *IDF* (Inverse Document Frequency)-Transform (Markov & Larose, 2007):

IDF-T = fij*log (number of documents/number of documents with stem i) (I)

fij is the frequency of the stem *i* in document *j*. The *Weka* toolkit is a free software for data mining and text mining tasks, and we used *Weka* software to apply the *IDF-T*.

After inserting a semantic weight (*IDF-T*) for each stem of each text, we can apply one of three procedures for stem selections. The first procedure is to use the same set of stems returned by the *word class selections* stage without intervention. It means the index produced by the *filtering algorithms* stage is simply the set of stems selected during the *word class selections*: (nouns), (nouns, verbs), (nouns, adjectives), (verb, adjectives), (nouns, adjectives, verbs), or (nouns, adjectives, verbs - without tags) with their *IDF-T* semantic weights. The second filter procedure possible is to use an intelligent algorithm based on the *Genetic Search* Metaheuristic (Goldberg, 1989) combined with *Correlation-based Feature Subset Selection* (it is named cfsSubSetEval by *Weka*) for selecting the stems, both implemented by *Weka*. The third option is to filter the set of stems coming from the *word class selections* stage by using a frequency filter algorithm; this frequency filter algorithm selects the stems according to the stems' frequencies adopted by the researcher, for example if the researcher chose (frequency: 2) only the stems having at least two occurrences inside the corpus could be selected to the output index. These filtering procedures for selecting terms were chosen since they work over different paradigms (no specific filtering, an intelligent filtering, or a frequency filtering), the objective was to verify the impact of each one during the clustering experiments.

The indexes produced during this stage represent the input texts as a set of selected terms (text indexes) and they are clustered by the clustering algorithms during the next stage.

Clustering Algorithms: The clustering algorithms are executed in this final stage. Since the clustering algorithms do not receive any extra knowledge (like a hint for the topics or vocabulary) about the texts to be classified, it only works on the texts' content, and the clustering algorithms must preview the topic for each cluster, or even the number of clusters. A few clustering algorithms such as SKM (Simple K-Means) described by Manning and Schütze (1999), and sIB (Sequential Information Bottleneck) described by Slonin et al. (2002), are algorithms that ask for the number of clusters from the user. Other algorithms such as EM (Expectation Maximization) (Manning & Schütze, 1999), Evolutionary Searches (Jones at al., 1995), or Neural Networks Architectures (Kohonen, 1997), (Kohonen, 1998) can try to guess the number of clusters and the topic for each cluster, but guessing the number of clusters is a more difficult algorithmic task. Each clustering algorithm may find a different result, since they could apply different mathematical strategies and heuristics to compare the texts (represented by indexes) when clustering. Algorithms can use formal knowledge of the language (as grammar structures or textual features) when executing a clustering process, therefore natural languages having a high level of scientific description would get better results for text clustering.

The Data Mining toolkit *Weka* already has the three clustering algorithms we chose for the experiments (EM, SKM, and sIB), so we decided to use this tool again for the *clustering algorithms* stage of the experiments. Many options of running can be

chosen before starting the algorithm execution; we can choose an iteration number, the number of groups for clustering, and other restrictive alternatives for each algorithm. The three algorithms we tested make linear and excluding clusters, which means they are non-hierarchical.

After describing each stage, we can set a final architecture for the experiments, the figure below shows the final model adopted for the four experiments performed. The *evaluation methods* for measurement are described in the next section.

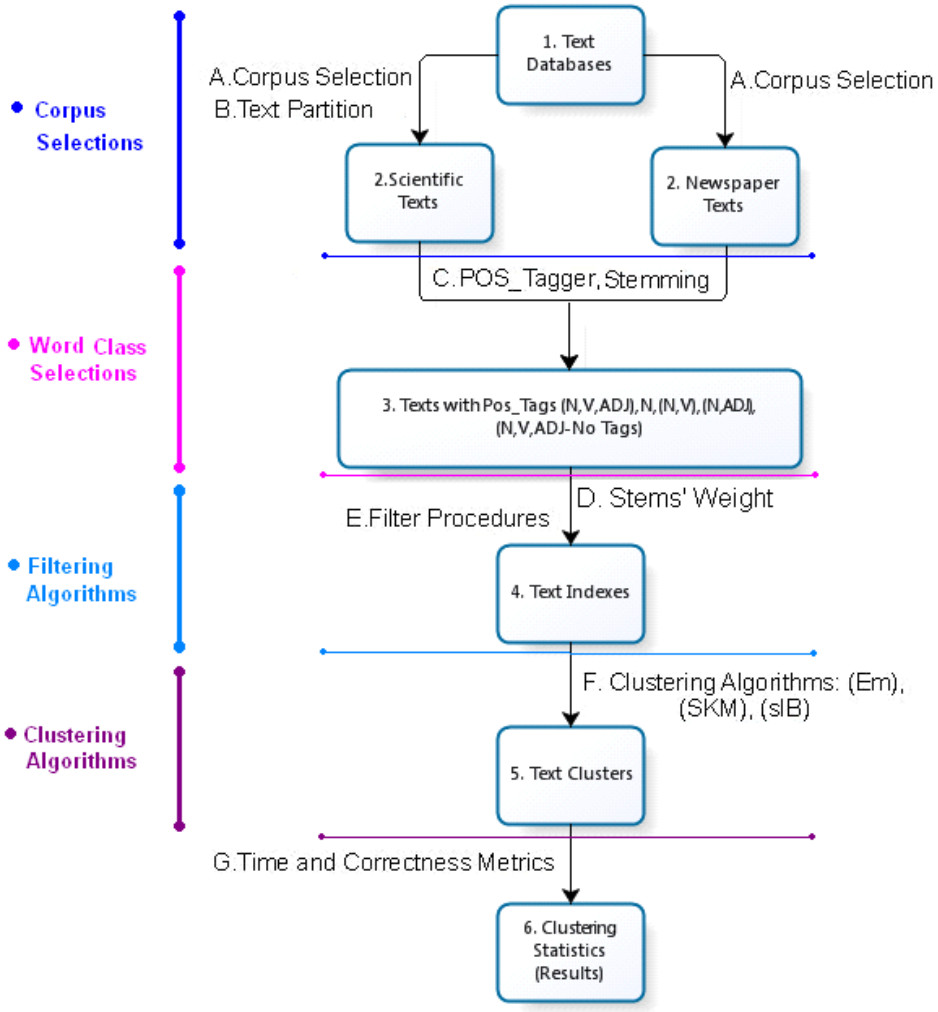

Figure 1: Experiment Stages and measurement

4. EVALUATION METHODS

4.1 Clustering correctness metrics

The usual metrics for clustering correctness are F-measure, Percentage Values of Correctness, Purity and Entropy. The usual metrics evaluate clustering algorithms' performance and the quality of the resulting clusters. F-Measure has been applied to measure the quality of the entire clustering process, while Purity and Entropy are used to measure the quality of the resulting clusters, and Percentage Values of Correctness is applied to measure the errors of the clustering results considering a human classification as reference, Percentage Values of Correctness returns the exact number of clustering errors from each cluster.

Song and Park (2006) only applied F-Measure, Seno and Nunes (2008) uses three metrics (F-measure, Purity and Entropy), and Maia and Souza (2010) uses Percentage Values of Correctness as metric, we notice the choice of each metric depends on the objectives of the experiment. Our decision was to use Percentage Values of Correctness as metric, since we wanted to compare the clustering results returned by the clustering algorithms and the previous human classification, the exact number of clustering mistakes from each experiment was necessary to choose the best clustering experiment. We also used an additional metric for measuring the number of deviated clusters, a Deviation Number (DN) which identifies the exact number of clusters created more than the expected number of clusters or less than the expected number of clusters. This metric value (DN) is calculated only for the EM algorithm since only this clustering algorithm guesses the number of clusters when executing a clustering process.

We realized that these two values (Percentage Values of Correctness and a Deviation Number - DN) are more informative than using a single statistical value about the distribution of the documents over the clusters created, as returned by (F-measure, Purity and Entropy metrics); the results of the two metrics described is more informative (for this specific set of experiments) then the usual averages since they return two different exact numbers and not a single average from a clustering result.

The result analysis is performed according to *Weka* outputs. For each clustering algorithm execution, *Weka* provides the output of *n* clusters (0 to *n*-1) and *m* texts classified (0 to *m*-1) for each cluster. We can see the clustering results through the result window as Cartesian Coordinates (x, y) and colors for each cluster. The next figure 2 below shows the result for the sIB algorithm execution for 3 clusters:

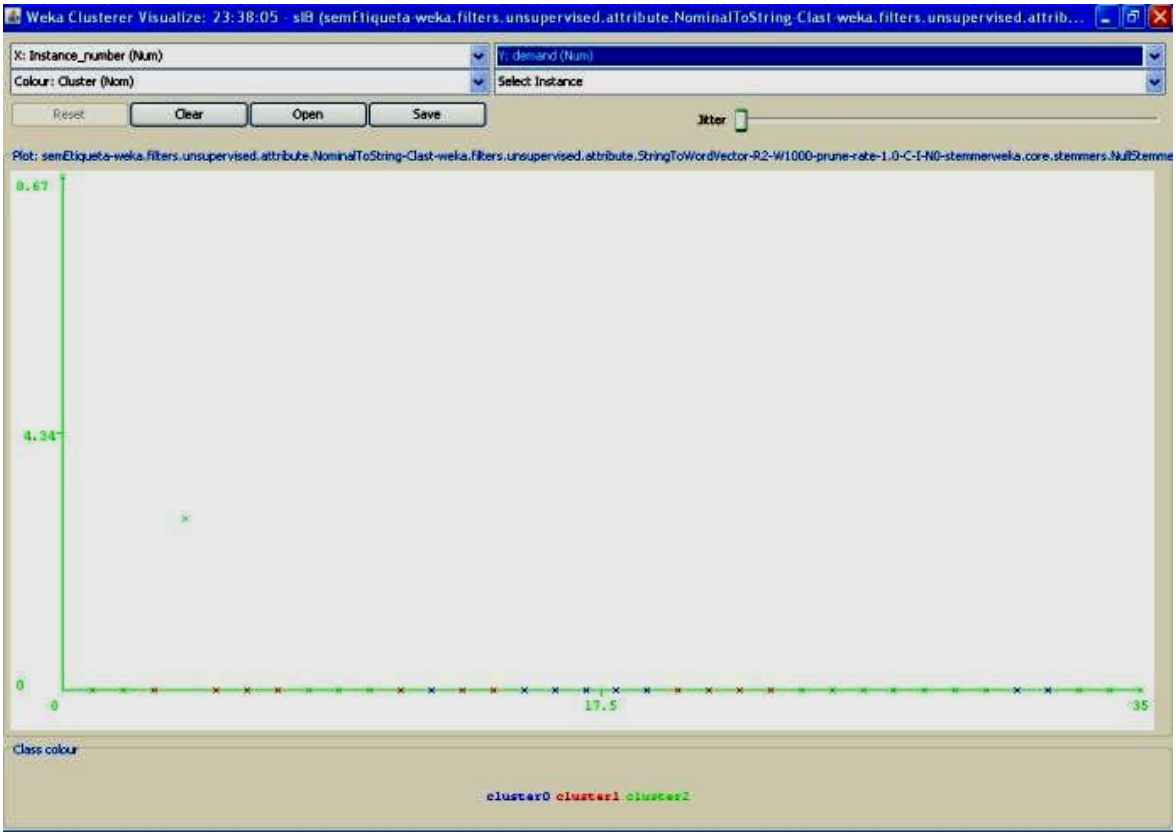

Figure 2: *Weka* image for sIB clustering algorithm output

The x-axis holds the number of the document; this sample experiment has 36 text documents. By our configuration, we know that the 0-11 numbers represent the documents about cluster 1 (for example, Pharmacy); numbers 12-23 represent cluster 2 (for example, Physical Education) and cluster 3 (for example, Linguistics) is represented by numbers 24-35.

The y-axis holds the corpus items (in this case, stems), and the symbol "*" indicates both a document (0-35) and a Boolean value (true or false) for a stem on the y-axis. If the document has the stem shown on the y-axis then the symbol "*" is aligned according to the y-axis. The color of the symbol "*" indicates the cluster where the document is found. Each cluster produced by the sIB, EM or SKM algorithm execution will get a specific color.

The percentage values of correctness are calculated by analyzing the results displayed by figure 2, it shows us a *Weka* experiment result using the sIB algorithm, for 3 clusters (Pharmacy, Physical Education, and Linguistics), with 12 documents queued for each area (0 to 35), as shown by the x-axis. Each color represents a cluster generated by the system (sIB asks the user for the number of clusters to be generated). Cluster 0 is blue, cluster 1 is red, and cluster 2 is green. We consider the last 12 elements as green, with 7 mistakes; the 12 elements in the middle are blue, having 2 mistakes; and the 12 elements at the beginning are red, with 6 mistakes. The error summation is 15, so the experiment correctness percentage is 58.3% of correct classifications. We apply this calculus to all the experiments performed.

For a clustering experiment using the EM algorithm we use two metrics: we check whether the number of clusters generated by the EM algorithm execution is correct or, if not correct, we calculate a deviation number (DN) given by the absolute value function (it always returns a number equal or greater than zero):

$$DN = (| \textit{Number of Expected Clusters} - \textit{Number of Output Clusters} |) \quad \text{(II)}$$

Then we calculate the percentage values of correctness as described for the sIB and SKM algorithms.

4.2 Clustering time metrics

The time unit chosen for clustering time measurement is minutes, since the executions are long, and they could take from seconds to hours to complete. Minutes are an intermediate metric, and for this reason, we decided to use this time unit. When the time taken by an execution is broken, like 2 minutes and 35 seconds, we chose to extend the number to the next unit of minutes, but if the past seconds are less than 30, we round down to the nearest minute unit. For example, 2 min 29 sec. becomes 2 minutes, but 2 min. 30 sec. becomes 3 min.

It is difficult to measure and compare exact time, since the computational conditions could vary even when the machine architecture, hardware configuration, operating system, and compiler are the same. For example, let's suppose we expect the time results for two machines having the same machine architecture and running the same clustering algorithm for the same experiment, but for any reason the number of processes running on machine one is greater than machine two. As a consequence the time results will have different values when clustering is finished, since machine one needs to control more processes than machine two. For this reason, the time values we got here are just an approximation under a few specific conditions.

For these experiments, we use a machine having a Pentium Dual Core T4500 Intel processor. It has 2 GB Main Memory; the Operating System is Microsoft Windows XP Professional 2002 SP 3; the *Weka* version is 3.6.3; and the Java Development Kit (JDK) installed is jdk1.6.0_04.

We took notes only for the time consumed by *Weka* to build the clusters. It means we just measured the time taken by the *clustering algorithms* stage: SKM, sIB, and EM clustering algorithms.

5. RESEARCH QUESTIONS

Question 1: Considering the four experiments performed, and their different configurations for the four stages, what is the experiment which produces best results (the one that produces the best clustering correctness value and consumes less time) having a **scientific corpus** as input?

Question 2: Considering the four experiments performed, and their different configurations for the four stages, what is the experiment which produces best results

(the one that produces the best clustering correctness value and consumes less time) having a **newspaper corpus** as input?

6. EXPERIMENTS' DESCRIPTION AND RESULTS

6.1 Experiment 1

6.1.1 Configurations of the experiment 1A (newspaper corpus)

Corpus Selection: Newspaper texts from Lácio-Web corpus, related to Biological Sciences, Exact Sciences, and Human Sciences. 15 texts from each area. Total: 45 texts.

Word Class Selections: We experimented different sets of word tags (4 options kept the word class tags and 1 option did not use the tags): (nouns), (nouns, verbs), (nouns, adjectives), (nouns, adjectives, verbs), (nouns, adjectives, verbs - without the tags).

Filtering Algorithms: 2 options: A *Weka* intelligent filter named *Genetic Search* with the method for evaluating stems named cfsSubSetEval. We kept the default values for (GeneticSearch/ cfsSubSetEval). Another try with no filter. The formula for Lexical Item Weights was only *IDF-T*.

Clustering Algorithms : EM, SKM and sIB algorithms.

Notice there are 30 executions for each possible combination above (5 Sets of Tags * 2 Filters Procedures * 3 Clustering Algorithms).

6.1.2 Configurations of the experiment 1B (scientific corpus)

Corpus Selection: Scientific Journals from UFG Digital Library from 3 different journals, about Pharmacy, Physical Education, and Linguistics. 12 papers from each journal. Total: 36 texts.

Word Class Selections: We tried many different sets of tags (4 options kept the word class tags and 1 option without the tags): (nouns), (nouns, verbs), (nouns, adjectives), (nouns, adjectives, verbs), (nouns, adjectives, verbs - without the tags).

Filtering Algorithms: 2 options: A *Weka* intelligent filter named *Genetic Search* with the method for evaluating stems named cfsSubSetEval. We kept the default values for (GeneticSearch/ cfsSubSetEval). Another try with no filter. The formula for Lexical Item Weights was only *IDF-T*.

Clustering Algorithms : EM, SKM and sIB algorithms.

Notice there are 30 executions for each possible combination above (5 Sets of Tags * 2 Filters Procedures * 3 Clustering Algorithms).

For both Experiments 1A and 1B, the result having the best value of correctness for each clustering algorithm is chosen as the best (only one combination for each algorithm is chosen).

6.2 Experiment 2

Experiment 2 aimed to verify the clustering correctness percentage when we inserted more input texts and clusters; a more diverse corpus was compiled. We verified how the percentage correctness error rate increases when adding more textual elements.

For this experiment, we only tested the combinations that reached the best values for Experiments 1A and 1B, since we wanted to verify whether the best values of clustering correctness would be kept from the experiments 1A and 1B, and how much additional time it would consume. It means we redid the experiments 1A and 1B, but having 5 knowledge areas and 5 clusters and choosing new documents for the five knowledge areas.

6.2.1 Configurations of the experiment 2A (newspaper corpus)

For Experiment 2A, which works with newspaper texts, we added two additional sets: Social Sciences, and *Religion and Thought* texts (15 additional texts for each area). The set of POS-Tags and filters were the same from the best values obtained during Experiment 1A. Therefore, the algorithms had to produce 5 clusters instead of 3 clusters, and the total number of texts was 75 texts for the newspaper corpus.

6.2.2 Configurations of the experiment 2B (scientific corpus)

For Experiment 2B, which works with scientific texts, we added History and Geography texts (12 additional texts for each area). The set of POS-Tag and filters were the same from the best values gotten during Experiment 1B. Therefore, the algorithms had to produce 5 clusters instead of 3 clusters, and the total number of texts was 60 texts for the scientific corpus.

The total number of executions for this experiment is 6: (1 filter * 3 algorithms * 1 tag set) = 3 executions for each corpus.

6.3 Experiment 3

This experiment verified the effect of taking the Nouns out of the text indexes. Our goal was to check the importance of the Nouns for text clustering. If the importance of the Nouns for clustering is high, their absence will produce poorer clusters. We again modified a few features from Experiment 1. For this experiment, two corpuses were used as before, a newspaper corpus, the same from Experiment 1A, and a scientific corpus, the same from Experiment 1B.

From Experiment 1A and 1B, all the configurations described were kept, except the POS-Tags. Here we only used the tag set: (Verb, Adj); so we kept the configurations for *corpus selections*, *filter algorithms*, *clustering algorithms* but this time we tested only Verbs and Adjectives as tags for the stage *word class selections*.

The total number of executions for this experiment is 12 (2 filters * 3 algorithms * 1 tag set) = 6 executions for each corpus. The result having the best value of correctness for each clustering algorithm was chosen as the best.

6.4 Experiment 4

This experiment used another kind of filtering procedure for the *filter algorithms* stage, so we tried a filter based on stem frequencies. It means that instead of using the text indexes produced during Experiment 1, this time we selected the main stems from the texts to be clustered according to their frequency inside the corpus. The goal was to verify whether the filter's variation causes a notable effect over clustering performance (correctness and time).

We tried different frequencies 2, 3, 4, 5 for each term of the corpus for this new filter, since we were trying to find the best one. During these executions we only chose two sets of tags: (Nouns) and (Nouns, Verbs, Adjectives – without tags) for the *Word Class Selections* stage. These word classes were chosen since, initially, we had the hypothesis that (Nouns) were the best alternative to identify a topic from a text, and the set (Nouns, Verbs, Adjectives - without tags) has been applied to many works for text mining, indexing, and clustering. We performed these new tests for the newspaper corpus (Experiment 4A) and also for the scientific corpus (Experiment 4B).

When we specify we tried different frequencies 2, 3, 4, 5 for the experiment, it means that we only considered stems having frequency 2, 3, 4, or 5 inside the entire corpus for each execution test.

The number of execution tests for each corpus was ((4 possible frequencies for filtering * 3 clustering algorithms * 2 tag sets) = 24 possibilities), or 48 executions considering both newspaper and scientific corpus. The result having the best value of correctness for each clustering algorithm was chosen as the best.

7. EXPERIMENTAL ANALYSIS AND RESEARCH ANSWERS

The three pairs of charts below summarize the best results of clustering correctness and time achievement for experiments *(1-4)A* (newspaper corpus) and *(1-4)B* (scientific corpus), which we described in the last section. In the first graphic pair, we use a percentage notation "%" for clustering correctness, in the second pair a time notation in minutes. The last pair of charts shows the deviation numbers (DN) for the EM algorithm. The blue charts are about *Experiments A* (newspaper corpus) and red graphics about *Experiments B* (scientific corpus). Notice the results are always only for the last experiment stage (*Clustering Algorith*ms), so the results for time and correctness are always from the sIB, SKM and EM algorithms.

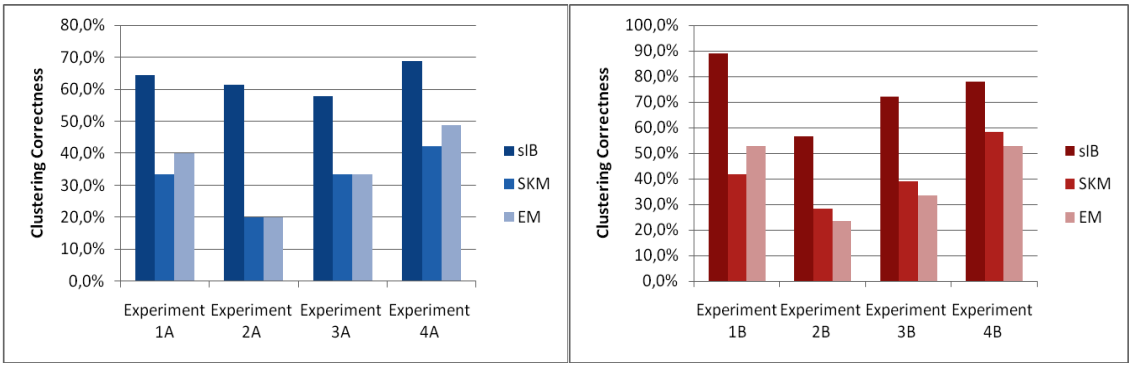

Figure 3: Two charts show the clustering correctness values gotten by each algorithm for each experiment (1-4) described. Percentage values of correctness on y-axis and the experiment name on x-axis.

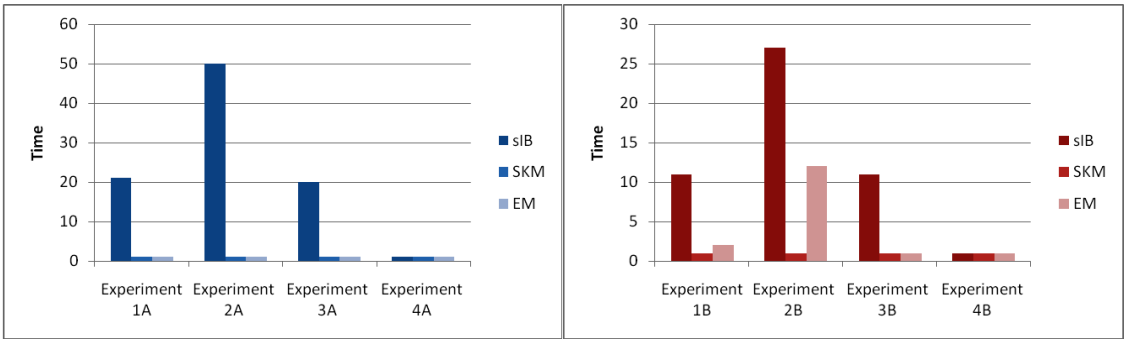

Figure 4: Two charts show the time consumed by each algorithm.

Time in minutes on y-axis and the experiment name on x-axis.

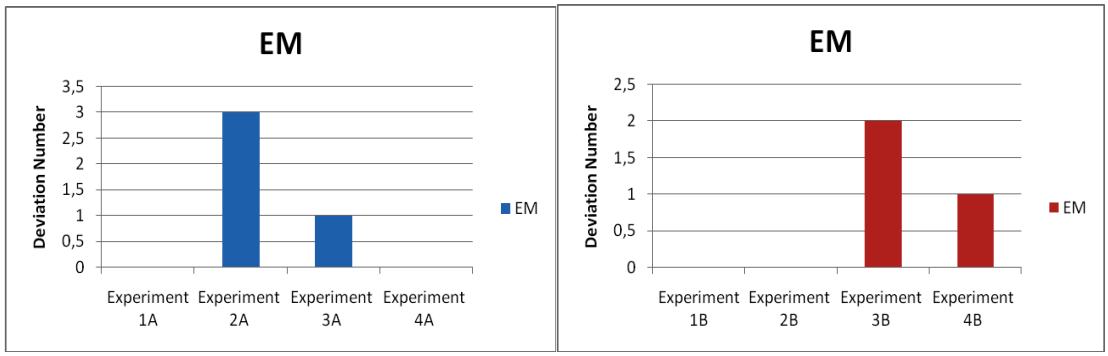

Figure 5: Two charts show the *Deviation Number* (DN) only for the EM algorithm.

Units of deviation on y-axis and the experiment name on x-axis.

Question 1 - Notable patterns about the scientific corpus (experiments: 1B, 2B, 3B and 4B – red graphics)

Best stages' combination for the scientific corpus considering correctness and time: 77.8% correctness in 1 minute by the fourth experiment.

Configuration for the best result (considering the best rate of correctness and less time):

Corpus selections: A corpus having three not related topics, Pharmacy, Linguistics and Physical Education.

Word class selections: (Nouns + Verbs + Adjectives - No Tags).

Filtering algorithms: (Frequency Filter: Frequency 5).

Clustering algorithm: sIB.

Question 2 - Notable patterns about the newspaper corpus (experiments: 1A, 2A, 3A and 4A – blue chart)

Best stages' combination for the newspaper corpus considering correctness and time: 68.9% correctness in 1 minute by the fourth experiment.

Configuration for the best result (considering the best rate of correctness and less time):

Corpus selections: A corpus having three not related topics, Biological Sciences, Exact Sciences, and Human Sciences.

Word class selections: (Nouns)

Filtering algorithms: (Frequency Filter: Frequency 3)

Clustering algorithm: sIB

For the newspaper experiments, we must observe that a few texts selected for tests can be found inside two or more groups from the Lácio-Web Database. The corpus for Experiments 1, 3, and 4 has 31.1% of the texts previously classified inside more than one group. The corpus for Experiment 2 has 9.3% of the texts previously classified inside more than one group.

8. CONCLUSION

This study tried to determine certain performance patterns when executing automated text clustering. We verified whether the best results found for automated text clustering and human text classification are approximated, we chose a realistic, experimental, and statistical point of view for finding clustering correctness and time values. Many works on this topic can be found for the English language, but we could

not find many works for Brazilian Portuguese, so this was the motivation for our research.

The first problem about evaluation of text clusters produced by a clustering algorithm is the concept of correctness: what is correct clustering? Or, what is a good cluster? Is the best clustering process the one that got the best time, or the best grouping or both? We accepted the predefined classified databases used as having correct classification, and they were produced by humans, so this classification follows a specific criterion, but is this specific criterion the best one? What is the best choice to the user: a system able to find the correct number of clusters having a low rate of clustering correctness, or having a high rate of clustering correctness over additional clusters? We can see the problem of evaluating text clustering is something relative, and the concept of correctness and quality for clustering depends on one's expectations. During this study, we chose a specific criterion for evaluation as we described during the experiments' description.

It was noticed that there is a combinatorial property for this kind of investigation, and the combinatorial choices are broad, with many different trees of possible investigation for the experiment stages: *corpus selections*, *word class selections*, *filtering algorithms*, and *clustering algorithms*. So we decided to investigate a few trees and specific branches for these trees, using the current linguistic tools available for Brazilian Portuguese. We see the possible investigations are very broad, and simple details during the stages of the experiments can produce very different results. Although we did 126 clustering tests during the four experiments, we investigated only a small part of the possibilities, under some exact conditions. So many attempts can still be performed.

Considering the results, regularity can be observed between the newspaper and the scientific experiments: the sIB algorithm always got the best clustering correctness values for both corpuses. We can notice the rate of clustering correctness for newspaper experiments is equal to or lower than the values obtained for the scientific corpus for all tests, except for Experiments sIB-2A, EM-3A, and EM-4A. One possible reason for this best result is the scientific vocabulary, since it is formal and rich, and it uses specific language which gives better clues for clustering algorithms.

Although we have gotten a reasonable result for the tuple (scientific texts, frequency filter, Nouns + Verbs + Adjectives - without tags, sIB algorithm) as described, the sIB algorithm asks for the number of clusters to be created and, in most cases, the user does not know the nature of the texts and topics or even the number of clusters; moreover sIB algorithm got a reasonable result only when producing three clusters for the scientific corpus. Some algorithms (like the EM algorithm) try to find the number of cluster and their terminological features, but EM does not yet have a high level of correctness, and much work must be done in this way to improve the results.

The second experiment (which works with five clusters) got a very worse result. We can see from Experiments 1 and 2 (figure 3) that the differential values of clustering correctness between the columns (Experiment 1A and Experiment 2A), for each clustering algorithm, have lower values than the differential values between the

columns (Experiment 1B and Experiment 2B) for each clustering algorithm. Perhaps it happens because of the scientific corpus chosen, since we added History and Geography texts for the new Experiment, 2B. So, we suspect that these two scientific fields have a similar vocabulary, which would cause an algorithmic clustering confusion over these two clusters, or that the two fields do not have a very formal or technical vocabulary powerful enough for clear text discrimination in relation to the other fields. Therefore, this fact opens a new hypothesis for the higher percentage difference between the columns (Experiment 1B and Experiment 2B).

Something important to observe is the fact that texts have a multi-classification nature and according to Ranganathan (1967), a textual topic could be analyzed from many different viewpoints (classes and facets). Even though works and studies about automatic text classification and clustering have been using a deterministic classification (each document inside only one group), a more realistic classification should permit a document to be inserted into more than one group, at least for some knowledge fields from the newspaper corpus. We realized that scientific texts are better classified using a deterministic approach than newspaper texts, since most of time a formal vocabulary is present for each scientific field. This formal language could not always be identified for newspaper texts, so an experimental possibility is to redo the tests considering a multi-clustering possibility for each text, but we should also consider that a multi-clustering approach could generate difficult searching and maintenance for documents, at least for databases having many clusters and having the same document inside many clusters.

Without getting a semantic or pragmatic level of analysis it would be hard to produce a good and practical clustering near to human classification. Therefore, we believe that general linguistic tools are important for clustering success. A Wordnet database (Fellbaum, 1998) for Brazilian Portuguese and other general language descriptions would play a very important role in acceptable clustering, at least for informative newspaper texts.

The evolution of the algorithms, scientific natural language descriptions and Bibliometrics is a key for text clustering improvement.

REFERENCES

Aires, R. V. X. (2000). *Implementação, Adaptação, Combinação e Avaliação de Etiquetadores para o Português do Brasil. [Implementation, Adaptation, Combination and Evaluation of Brasilian Portuguese Taggers.]* Unpublished MsC Thesis. Universidade de São Paulo, São Paulo, Brazil.

Aluísio S. M., & Almeida G. M. B. (2006). O que é e como se constrói um corpus? Lições aprendidas na compilação de vários corpora para pesquisa lingüística. [What is a corpus and how to build a corpus? Lessons learned during the compilation of various corpora for linguistic research.] *Calidoscópio*, **4**(3), 155-177.

Basic, B. D., Berecek B., & Cvitas A. (2005). Mining textual data in Croatian, in *Proceedings of the 28th International Conference,* MIPRO 2005*,* Business Intelligence Systems. (pp. 61–66).

Bezerra, G. B., Barra, T. V., Ferreira, H. M., & Von Zuben, F. J. (2006). A hierarchical immune-inspired approach for text clustering. *Selected papers based on the presentations at the 2006 conference on Information Processing and Management of Uncertainty,* IMPU, Paris, France. (pp. 131-142).

Biderman, M. T. C. (2001). O Português Brasileiro e o Português Europeu : Identidade e contrastes. [The Brazilian Portuguese and European Portuguese: Identity and contrasts.] *Revue belge de philologie et d'histoire,* **79** (3), 963-975.

Caldas Junior, J., Imamura, C.Y.M., & Rezende, S.O. (2001). Avaliação de um Algoritmo de Stemming para a Língua Portuguesa. [Evaluation of a Stemming Algorithm for Portuguese Language.] *The Proceedings of the 2nd Congress of Logic Applied to Technology,* São Paulo, Brazil. (pp. 267-274). São Paulo, Brazil: SENAC/Plêiade.

Camargo, Y. B. L. (2007). *Abordagem lingüística na classificação de textos em português. [A linguistic approach to the classification of texts in Portuguese.]* Unpublished MsC Thesis. COPPE - Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil.

DaSilva, C. F., Vieira, R., Osório, F. S., & Quaresma, P. (2004). Mining Linguistically Interpreted Texts. In *Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora*.

Fellbaum, C. (1998). *WordNet. An electronic lexical database.* Cambridge, MA: MIT Press.

Furlanetto, M. M. (2008). Neological Formations in Brazilian Portuguese: a Discursive View. *Fórum Lingüístico,* **5** (2), 1-22, Florianópolis, Brazil.

Goldberg D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley.

Hall M. A. (1998). *Correlation-based Feature Subset Selection for Machine Learning.* Hamilton, New Zealand. (University of Waikato Ph.D. dissertation).

IBGE (2000). *Brasil: 500 anos de povoamento [Brazil: 500 years of settlement]*. Rio de janeiro: IBGE.

Jones, G., Robertson, A. M., Santimetvirul, C., & Willett, P. (1995). Non-hierarchic document clustering using a genetic algorithm. *Information Research*, **1**(1), Retrieved 15 April, 2012 from http://InformationR.net/ir/1-1/paper1.html.

Hammouda, K. M., & Kamel. M. S. (2004). Efficient phrase-based document indexing for web document clustering. *IEEE Transactions on Knowledge and Data Engineering*, **16**(10), 1279–1296.

Kohonen, T. (1997). *Self-Organizing Maps.* 2nd ed., Berlin: Springer-Verlag.

Kohonen, T. (1998). Self-Organization of Very Large Document Collections: State of the Art. *Proceedings of ICANN98, the 8th International Conference on Artificial Neural Networks*, Skövde, Sweden, 2-4 September, 1998, **8**(1), 65-74: Springer.

Maia, L. C., & Souza, R. R. (2010). Uso de sintagmas nominais na classificação automática de documentos eletrônicos. [The use of noun phrases in automatic classification of electronic documents.] *Perspect. Ciênc. Inf.*, **15**(1), 154-172.

Manning, C. D, Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo, Delhi: Cambridge University Press.

Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: The MIT Press.

Markov, Z., & Larose D. T. (2007). *Data Mining the Web: Uncovering Patterns in Web Content, Structure and Usage*. Hoboken, New Jersey: John Wiley ; Sons, Inc.

Palmeira, E., & Freitas, F. (2007). Ontologias detalhadas e classificação de texto: uma união promissora. [Detailed ontologies and text classification: a promising union.] *ENIA 2007: VI Encontro Nacional de Inteligência Artificial*. Rio de Janeiro, Brazil, July, 03-06, 2007. Rio de Janeiro: Instituto Militar de Engenharia.

Ranganathan, S. R. (1967). *Prolegomena to Library Classification*. London: Asia Publishing House.

Ratnaparkhi, A. (1996). A Maximum Entropy Model for Part-of-Speech Tagging. *Proceedings of the First Empirical Methods in NLP Conference*. University of Pennsylvania, May 17-18, 1996. (pp. 133-142).

Reis, João José (2000). Presença Negra: conflitos e encontros. In Brasil: 500 anos de povoamento [Black Presence: conflicts and encounters]. Rio de Janeiro: IBGE, 2000. pp: 91.

Rossel, M., & Velupillai, S. (2005). The Impact of Phrases in Document Clustering for Swedish. *Proceedings of the 15th NODALIDA conference*, NoDaLiDa 2005, Joensuu, Finland. (pp.173-179).

Seno, E. R. M., & Nunes, M. D. V. (2008). Some Experiments on Clustering Similar Sentences of Texts in Portuguese. In Teixeira, A., StrubeDeLima, V. L., CaldasDeOliveira, L., Quaresma, P. (Eds.), *Lecture Notes in Artificial Intelligence, Vol. 5190. 8th International Conference on Computational Processing of the Portuguese Language*, PROPOR 2008, Aveiro, Portugal, September 08-10, 2008. (pp. 133-142). Berlin, Germany: Springer-Verlag.

Silva. A. S. (2006). Sociolinguística cognitiva e o estudo da convergência/divergência entre o Português Europeu e o Português Brasileiro. [Cognitive Sociolinguistics and the study of convergence / divergence between European Portuguese and Brazilian Portuguese.] *Veredas :Revista de Estudos Lingüísticos*, **10** (2006): Universidade Federal de Juiz de Fora.

Slonin, N., Friedman N., & Tishby, N. (2002). Unsupervised document classification using sequential information maximization. *Proceedings of the 25th International ACM*

SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland, August 11-15, 2002. (pp. 129-136). New York: ACM Press.

Song W., & Park S. C. (2006). Genetic Algorithm-based Text Clustering Technique. In Licheng Jiao, Lipo Wang, Xinbo Gao, Jing Liu, Feng Wu (Eds.), *Lecture Notes in Computer Science, Vol. 4221. Advances in Natural Computation, Second International Conference*, ICNC 2006, Xi'an, China, September 24-28, 2006. (pp. 779-782). Berlin: Springer-Verlag.

Stefanowski, J., & Weiss, D. (2003). Web search results clustering in Polish: experimental evaluation of Carrot. Advances in Soft Computing, Intelligent Information Processing and Web Mining, *Proceedings of the International IIS: IIPWM´03 Conference*, Zakopane, Poland, vol. 579 (14). (pp. 209-22).

Viera, A.F.G., & Virgil, J. (2007). Uma revisão dos algoritmos de radicalização em língua portuguesa. [A review of stemming algorithms for Portuguese Language.] *Information Research*, **12**(3), paper 315. Retrieved 15 April, 2012 from http://InformationR.net/ir/12-3/paper315.html.

Witten I. H., & Frank E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques, 2º Ed.* Amsterdam, Boston, Heidelberg, London, New York, Oxford, Paris, San Diego, San Francisco, Singapure, Sydney, Tokyo: Elsevier.