

JISTEM - Journal of Information Systems and Technology Management  
*Revista de Gestão da Tecnologia e Sistemas de Informação*  
Vol. 9, No. 2, May/Aug. 2012, pp.213-234  
ISSN online: 1807-1775  
DOI: 10.4301/S1807-17752012000200002

## INFORMATION RETRIEVAL SYSTEM USING MULTIWORDS EXPRESSIONS (MWE) AS DESCRIPTORS

**Edson Marchetti da Silva**

Federal University of Minas Gerais, MG, Brazil

**Renato Rocha Souza**

Getúlio Vargas Foundation - FGV, RJ, Brazil

---

### ABSTRACT

This paper aims to propose an alternative method for retrieving documents using Multiwords Expressions (MWE) extracted from a document base to be used as descriptors in search of an Information Retrieval System (IRS). In this sense, unlike methods that consider the text as a set of words, bag of words, we propose a method that takes into account the characteristics of the physical structure of the document in the extraction process of MWE. From this set of terms comparing pre-processed using an exhaustive algorithmic technique proposed by the authors with the results obtained for thirteen different measures of association statistics generated by the software Ngram Statistics Package (NSP). To perform this experiment was set up with a corpus of documents in digital format.

**Keywords:** Extraction of Expressions Multiwords, Measures of Association Statistics, Compared Search, Information Retrieval System, the Document Structure.

### 1. INTRODUCTION

Since the first computers appeared, one of their main purposes has been to collect, store and process large volumes of data to produce information. It is for computer systems to receive data, organize them and classify them, so they can be retrieved and presented to the user requesting to meet the demand for desired information. Since the 1960s some models have been proposed and implemented to manage the maintenance and retrieval of structured data. Among them we mention the Network Model, the Hierarchical Model and the Relational Model. All of them require that a structural scheme is designed to receive data by creating a strong bond between the semantic data and the exact location where it is stored, i.e., the metadata. In this type of solution to ensure that the extraction of information is deterministic, the data must

---

Manuscript first received/*Recebido em* 15/01/2012 Manuscript accepted/*Aprovado em:* 15/03/2012

Address for correspondence / *Endereço para correspondência*

*Edson Marchetti da Silva*, Universidade Federal de Minas Gerais, MG, Brasil, Msc em Administração pela Universidade FUMEC, Doutorando em Ciência da Informação pela UFMG – E-mail: [edson@div.cefetmg.br](mailto:edson@div.cefetmg.br)

*Renato Rocha Souza*, Fundação Getúlio Vargas, RJ, Brasil, Doutor em Ciência da Informação pela UFMG. E-mail: [rsouza.fgv@gmail.com](mailto:rsouza.fgv@gmail.com)

Published by/ *Publicado por:* TECSI FEA USP – 2012 All rights reserved.

necessarily be organized in a structured way and grouped according to their intrinsic characteristics and semantics. Therefore, these models are suitable only when dealing with data that can be organized this way, as is the case of information systems, which store their data supported by the technologies provided by Relational Database Management Systems and their extensions. However, most of the information generated by humans is not as structured as it is registered through language in written form. The big challenge, which still presents many open questions, is how to bring the computer close to the human form in dealing with information, that is, by the treatment of the natural language.

The quest to build a machine capable of communicating with humans in a natural way through the spoken or written language is something that Artificial Intelligence (AI) has been seeking for decades. AI is a research area, which according to Russell & Norvig (2004 p. 3-4) had its genesis with John McCarthy in 1956 and that historically has been working on two fronts: the first focused on systems that think and act as humans and the second focused on systems that think and act rationally. Research with the focus on the first approach, proved far more complex than it seemed. The second approach, which works with rationality, does what is right considering the data it has, it is far more successful, although limited to represent only some aspects of human nature.

According to Manning & Schütze (1999, p. 4-7) two schools of thought prevailed in language studies. The first, the empiricists, between 1920 and 1960, postulated that the experience is unique, or else at least the main form of construction of knowledge in the human mind. They believed that cognitive ability was in the brain and that no learning is possible from a *tabula rasa*, and that therefore, the brain had the ability to associate a priori pattern recognition and generalization, which combined with the rich human sensor capacity enabled language learning. The second, the rationalists, between the years 1960 and 1985 postulated that a significant part of the knowledge of the human mind is not derived from the senses, but previously established, presumably by genetic inheritance. This current of thought was based on the theory of innate faculty of language proposed by Noam Chomsky, which considers the initial structures of the brain as responsible for making every individual, from sensory perception, follow certain paths and ways to organize and generalize the information internally.

Currently, from the most diverse areas of knowledge, advances have been aimed at the ability of machines to represent and retrieve information. In this search, one of the main aspects is to develop the ability to interpret documents assigning semantic value to the written text. The area of Language Engineering and Natural Language Processing (NLP) is highlighted which through studies of morphology, syntax and semantic analysis, and statistical processing were designed to predict behavior of a textual content.

All these issues are still a useful field for the sciences. There is a ceaseless quest to articulate ways of representing knowledge in machinery to reduce the differences between computational and symbolic capacity of human thought. From what perspective should the issue address? This is a relevant and complex debate, waged by the most diverse areas, from human, social and exact sciences. The language is symbolic and a direct equivalence does not even exist between the signs in mind and the creation of a word that expresses its meaning in the different languages spoken around the world.

The language constructs semantic fields or areas of significance linguistically circumscribed. Vocabulary, grammar and syntax are geared to the organization of semantic fields. Thus the language constructs classification schemes to differentiate objects in gender or by number; ways to accomplish the stated reason for the opposition to be listed; modes indicate the degree of social intimacy, etc.. (Berger, 1985 p. 61)

The human mind is a particular view of an individual formed by social relations which constitute what we commonly call personality, which makes up its own set of beliefs and values. Added to this there are personal relationships, data and information kept in mind that form knowledge. Reflections for the production of human knowledge, or simply to produce answers to questions and needs: the mind does not process all the knowledge in the brain. The mind seeks to approach similar situations producing inferences, creating new relationships or seeking memories recorded in the memory. That is, a cut of one point in the context of the brain. Therefore, there is no guarantee of accuracy in the answers at any time. The computer works in a completely different context of the human brain. Therefore, current technology will never be able to simulate the human mind to its fullest. What can you get is an approximation of some human capabilities. According to Vygotsky's ideas, a clear understanding of the relationship between thought and language is necessary in order to understand how the intellectual development occurs.

The meaning of words is only a phenomenon of thought as it is embodied in the speech and is only a linguistic phenomenon connected with thought and enlightened by it. It is a phenomenon of verbal thought or speech signifier - a union of thought and language. (Vygotsky, p. 277-278).

We believe that by directing the efforts of science, in search of the semantic representation of knowledge for information retrieval, simulating the human mind is not the path that will give the best results. Therefore, these efforts result in the same "defects", or characteristics of the human form of processing information, uncertainty, etc. does not guarantee repeatability. So the best way to handle this problem is to reduce the language to the limitations of logic and thus guarantee the accuracy of what you want to express, rather than try to approximate the language of logic and enter the inaccuracy.

We propose as a common thread of this work the theoretical treatment of the text by reducing the content expressed in a natural language to a certain set of lexical compounds that have greater capacity to express the meaning of a textual content, Multi-Word Expressions (MWE), and use them as search descriptors in an Information Retrieval System (IRS).

## **Related Works**

Several studies aimed at identifying MWE were published, among them we highlight Dias Lopes and Guilloré (1999) aimed at the extraction of MWE independently of language, based solely on statistical methods; Silva Lopes (1999) that aims to extract *n*-grams from the analysis of a text in a local context called LocalMaxs; Portela, Mamede and Batista (2011) who take into account the morpho-syntactic text, and therefore require intensive use of computational resources, among others. We can also cite studies that apply the concept of MWE for automatic translation alignment through the use of lexical expressions to see how they would compare the same text in different languages which may provide clues relevant to the identification of these

expressions: Calzolari et al. (2002), Sag et al. (2002); Ramich (2009), Zhang, Yoshida, Tang and Ho (2009); Villavicencio, Ramisch, Machado, Caseli and Finatto (2010). Based on these studies the existence of a gap in relation to extraction of MWE is verified, which takes into account the intrinsic physical characteristics of the documents and which language is independent. It is from these ideas that we proposed to obtain MWE from a document base and use it to search keywords compared to the automated retrieval of similar documents.

To better describe the experiments the work is structured into the following sections which are presented in the following contents: Section 3 - theoretical framework about MWE, Section 4 - methodology, Section 5 - Results and conclusions; Section 6 - Recommendations for future work.

## 2 THEORETICAL FRAMEWORK

As noted by Zhang et al. (2009), the ability to express the sense of a word depends on the other words that accompany it. When a word appears accompanied by a set of terms, the greater the chances of this set to have a significant meaning. This indicates that not only a word but also the contextual information is useful for processing information. It is from this simple and direct idea that research on MWE is motivated. Thus it is expected to capture relevant semantic concepts from the text expressed by MWE.

Although there are many papers on the subject, there is no formal definition of consensus in the literature on MWE. We can consider that MWE are formations composed of two or more adjacent words that occurring in a frequency above a threshold when combined it have a greater semantic expressiveness than when each of its terms are set separately. For Sag et al. (2002) MWE are "idiosyncratic interpretations that cross the boundaries (or spaces) between words" (p. 2). A further description found in the literature is shown below.

The term multiword expression has been used to describe a large number of different constructions, but closely related, such as support verbs (give a demonstration, give a lecture), nominal compounds (Military Police), institutionalized phrases (bread and butter) and many others. [...] IN encompasses a large number of buildings, such as fixed expressions, noun compounds and verb-particle constructions. (Villavicencio *et. al.*, 2010, p. 16.)

According to Ranchhod (2003, p. 2) the fixed expressions are linguistic objects that have differences in terminology and the absence of criteria for the analysis that led them to be regarded as exceptional linguistic objects can not be integrated into the grammar of languages. However, there has been a growing interest, especially in NLP, after all these fixed forms are so numerous in any type of text, therefore, they can not be ignored. Therefore, these characteristics make the relevant MWE treatment a lexical resource, the informational inputs of which are important for many applications related to the NLP, such as an automatic translation of text summarizing, etc. In this sense, Villavicencio *et. al.* (2010) point out that many studies have sought ways of automation in lexical acquisition. These studies seek to understand the formation of lexical resources, an area still in need of research.

To Sag (2002, p. 4) MWE can be classified into:

- Fixed Expressions are those that do not present morphosyntactic and crunches do not allow internal modifications. They challenge the conventions of grammar and compositional interpretation, as treating them as word for word would not have the representation of the phrase, which has its own meaning given by composition.
- Semi-Fixed Expressions are those that have restrictions on word order and composition, but admit any lexical variations in bending, the reflexive form and choice of determinants. This type of MWE is categorized into three subgroups: the non-decomposable idioms, the noun compounds, and proper names. The first category occurs when two or more words together form an expression that has a new meaning, different from that obtained by the words in isolation. Example "kick the bucket", which has the compound meaning the idea of "give up". In this case there is variability of the idiom. The second category: compounds nominal are similar to the non-decomposable expressions, being syntactically unchangeable and in most cases they can be inflected in number. The third category: proper names are syntactically highly idiosyncratic. Take for example the compound "Holy Spirit"; it may be related to the federal state of Brazil, and it can be a surname, etc.
- Expressions Syntactically Flexible are expressions that admit syntactic variations in the position of its components. The following types of variations are possible: verb-particle constructions, constructions consisting of a verb and one or more particles that are semantically idiosyncratic or compositional; decomposable idioms. The light-verb construction is a verb regarded as being semantically weak subject to a variability syntactic solution, including passivation. They are highly idiosyncratic, because there is a notorious difficulty in predicting which light-verb combines with which noun.
- Institutionalized Expressions are compositional expressions (collocation), which vary morphologically or syntactically and that typically have a high statistic occurrence.

According to Moon (1998 cited by Villavicencio et al.) MWE are lexical units formed by a broad continuum between the compositional groups and non-compositional or idiomatic. In this context it is understood by those compositional expressions from the characteristics of these components which determine characteristics of the whole. And non-compositional idioms whose meaning or set of words has nothing to do with the meaning of each part. Given these characteristics, in dealing with MWE as words separated by space, they will surely bring anomalies to the process of IR.

Among the different approaches that deal with NLP, they highlight those dealing with MWE and use the symbolic methods by Calzolari et al. (2002) and a statistical approach by Evert and Krenn (2005). Both seek to interpret the textual content written in a natural language, but follow different paths to get results and computational costs of different contents. Thus the advantages and disadvantages of each method depend on the context for which they are being used. The symbolic approach seeks to find the meaning of syntactic, morphological and pragmatic texts based on a controlled dictionary of words and a set of rules aimed at interpretation. In this case, processing is strongly dependent on the language and the domain of the corpus. While the statistical approach seeks to give treatment to the text by recognizing behavior patterns based on the frequency of co-occurrence of words. The MWE are a set of words that co-occur with a frequency above chance.



Calzolari et al. (2002, p. 1934) corroborate the classification presented by Sag (2002) and even include an "etc" at its end. That is, as the authors themselves define, MWE are used to describe different but related phenomena, which can be described as a sequence of words which act as a unit at any level of language analysis and which have some or all of the following behaviors: reduced syntactic and semantic transparency, reduction or absence of compositionality, more or less stable, capable of violation of any rule syntax; high degree of lexicalization (depending on pragmatic factors), high degree of conventionality. Also according to these authors, MWE are located at the interface between grammar and lexicon. They also have some of the causes of the difficulties encountered in theoretical and computational framework for the treatment of MWE, as the difficulty of establishing clear boundaries for the field of MWE, the lack of computational lexicons of reasonable size to assist in NLP, before the multilingual perspective, often can not find a direct lexical equivalence; generalization of lexical difficulty (general and terminology) to a specific context.

The work Calzolari et al. (2002) uses a focused approach in MWE that is productive on the one hand and, on the other shows that regularities that can be generalized to classes of words with similar properties. In particular they seek to find grammatical devices that allow the identification of new MWE motivated by the desire for recognition as possible in the automated acquisition of MWE. In this sense, the research of these authors studied in depth two types of MWE: support verbs and compound nouns (or nominal complex). For according to them these two types of MWE are at the center of the spectrum of compositional variation where the internal cohesion together with a high degree of variability in lexicalization and language-dependent variation can be observed.

The approach used by Evert and Krenn (2005) is based on the calculus of statistical measures of association of the words contained in the text. In empirical tests, these authors used a subset of eight million words extracted from a corpus consisting of a newspaper written in German. The proposed approach was divided into three steps. In the first extracts the tuples from the corpus source contain Lexical pronouns (P), nouns (N) and verbs (V). These data are grouped in pairs (N + P, V) and placed in a contingency table, represented by a three-dimensional structure, where each pair is disposed in a plane P + N V and the third axis is assigned to the frequency information represented by four cells. Thus a comparison is made between all pairs extracted from the lexical corpus with their sentences, accounting for each sentence, one of four possibilities: there are PN and V; there is PS, there is not V; there is not PS and there is V; there are not PS and V. That is, one unit is added whenever one of the possibilities occurs.

The second step the association measures are applied to the frequencies collected in the previous step. This process results in a list of pairs of MWE candidates with their association scores calculated and ordered from the most strongly associated to the less strongly associated. The "n" top candidates on the list are selected for use in the next step.

The third step is the evaluation of the list of MWE generated by a human expert. Thus, the approach proposed by these authors is characterized by an extraction of semi-automatic MWE. In order to minimize the intellectual work of an expert, these authors propose the use of a technique of extracting a random sample, representative of the corpus rather than the complete set of documents.

Research conducted by Villavicencio et al. (2010) seeks to extract the MWE combining two different approaches: the approach and the approach based on associative lexical alignment. At first, the association measures are applied to all bigrams and trigrams generated from the corpus and the result of these measures is used for evaluation. The second approach draws MWE in an automated way based on the alignments of lexical versions of the same content written in Portuguese and English. To combine the results obtained, the authors used two approaches to Bayesian networks.

The statistical approach for the extraction of MWE through the co-occurrence of words in texts has been used in several recent works, among them: Pearce (2002); Kreen and Evert (2005); Pecina (2006); Ramisch (2009) and Villavicencio et al. (2010). These studies use various statistical techniques that seek to identify MWE as a set of adjacent words that co-occur with a frequency greater than expected in a random sequence of words in a corpus. Thus the associative approach is nothing more than the use of a set of association measures that aim to identify the candidate expressions for MWE. Among the techniques used include: coefficient of Pearson Chi Square; Dice coefficient; Pointwise Mutual Information – PMI; Poisson Stirling among others.

The lexical approach alignment checks if MWE found in a document written in certain a language also occurs in the corresponding version written in another language. In order to perform a review, the documents need to be aligned by matching the words expressed between the different versions in different languages. However, for the alignment to be possible, it is necessary that the documents are analyzed based on their morphology processed by a preprocessing tagging. Thus the parts of speech are used as additional information in the identification process of MWE. In the research carried out by Zhang et al. (2009) a method called Enhanced Mutual Information and Collocation Optimization (EMICO) is proposed to extract MWE focused on named entities. These compounds are characterized by being contiguous containing from two to six words describing more stable syntactic pattern concepts. These authors employ this technique in processing and text mining techniques, comparing it with traditional indexing vector space model speculating that the use of MWE for semantic interpretation of the text produces better results than the statistical and semantic models that deal with individual words.

In seeking to make sense of a text from their relevant parts, other strategies have been adopted. In this line the use of noun phrases stands out as search descriptors, addressed by the work of Kuramoto (1996) and Souza (2005) and researcher Maia (2010) who seeks to use the phrases to group documents. The method of identification of noun phrases uses an approach based on language, in the words of the text which are pre-labeled to identify them in grammatical classes as a basis for extracting phrases. However, the identification of phrases requires an in-depth analytical processing of sentences which demands a comprehensive rules-based computer processing which depends on the language. In the context of this research, which aims at seeking a test case of IR, through the use of parts of the text as semantically relevant keywords for the search process compared to a computational cost that makes possible the response time for text processing to online, we chose the use of MWE that are easier to obtain and language-independent. These aspects lead us to suppose that the proposed technique is more appropriate for the context to retrieve similar documents from a corpus of MWE extracted from a document used as a reference for the search. The goal is to get the semantic meaning of the document represented by the MWE and use them as descriptors of the search process.

### 3 METHODOLOGY

The purpose of this study is to test the automatic retrieval of documents from a corpus, from a document used as a reference, considering the intrinsic physical characteristics of textual content in order to compare the use of different techniques. In this sense, MWE will be extracted from a reference document for use with search keywords in a IR system. This methodology allows the user to search an alternative. In that, instead of informing keywords as part of the search, the user will be responsible for informing a document. In other words the search will be made from bigrams extracted from a document. This alternative strategy simplifies the user's work, which is known to use documents on the topic of interest to serve as the basis of the compared search in the recovery of similar documents.

Figure 1 shows a proposed software structure diagram which can be presented as a module of addition compared search, highlighted, which can be added to conventional systems of word search.

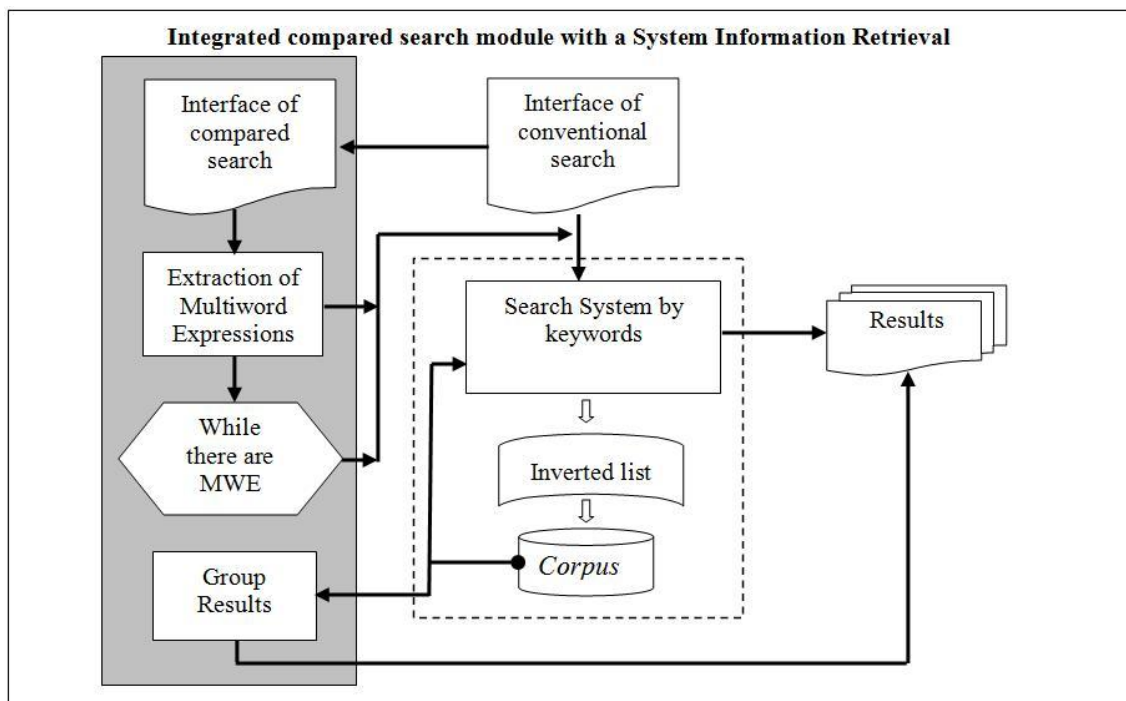


Figure 1 - Module of compared search integrated with a SRI.

Source: Prepared by the authors.

According to Sarmiento (2006), a text is not just a random jumble of words. The order of the words in the text is what makes sense. Therefore, the study of co-occurrence of words brings important information. This may indicate that the words are directly related by affinity or compositionality or indirectly by similarity. Therefore, the empirical base of linguistics is to find from the frequency of co-occurrences observing significant dependencies between terms. Evert (2005 cited by Sarmiento) points these four groups of measures:

- Tests of statistical significance;
- Coefficients of association;
- Based on concepts of information theory;
- Based on various heuristics.



To perform the experiment, two software components were implemented: One called Server and one called Client. The Server is responsible for indexing the corpus and providing a consultation service for information retrieval. The Client is responsible for receiving the reference document for MWE extraction search, sending a request to query and return the response with similar documents. This paper proposes a heuristic called Heudet to identify MWE. Then these tools are described in detail.

### 3.1 Converting a PDF document in the list of standardized terms

For a document to be processed by the proposed application, it is necessary to be in a text format, encoded in ASCII. That is, it is necessary to convert the binary format, typical of the software in which it was recorded, in a plaintext format. In this research all incoming documents are in PDF format, protected or not. To perform the conversion into text format, the TET PDF software was used. This software consists of a Dynamic Link Library (DLL) that was coupled in software components developed in C++ by the authors. The process of converting the PDF document page by page was performed in order to identify the header of the pages. To perform the segmentation of the PDF document pages, the Adolix software was used. All sub-steps of this process are executed by both software components drawn up for the experiment. Figure 2 shows an outline of the steps taken in this process.

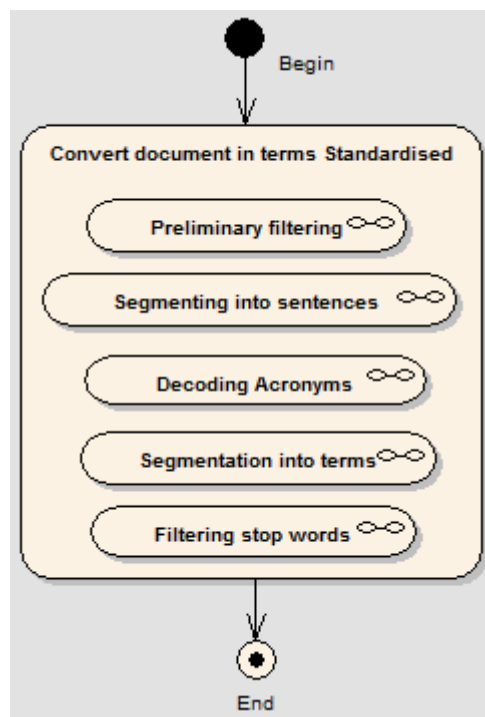


Figure 2 - Process of converting documents in standardized terms.  
Source: Prepared by the authors.

Following each of these substeps is presented in details.

#### 3.1.1 Preliminary filtering of the contents of the documents

After transforming the document text into PDF, preliminary filtering was performed in order to remove parts of the contents considered as noise. The adopted

heuristic evaluates the content that occurs repeatedly in all the pages from the top of each page. This extract, called header, is filtered and therefore eliminated in the converted text. Another filtering process that is performed at this stage the removal of the references, to include terms such as: name of authors and works, which often lie outside the central theme of the document.

### 3.1.2 Segmenting the string into sentences

After the elimination of the parts considered as noise, the goal in this substep is to perform parsing, ie, to process the string extracted from the document and separate the block into sentences and terms.

The accuracy of this process is of fundamental importance, so that an error does not propagate at this point in further processing. As described by Mikheev (2002, p. 290). Processing division of a text into sentences is a simple task in most cases. All it takes is to consider how separators characters: full stop, exclamation mark, question mark. However, there are some exceptions, for example, when the endpoint is used between numbers, abbreviations, or even when in both things at once. Therefore, some care should be taken, since an error in the separation of the sentences may generate failure to identify MWE. Take the example where an error in the process of separation of sentences leads to misidentification of MWE. When considering the sentences below as a single sentence, the term "information science" could be interpreted as MWE. While, in fact, there is no such a semantic meaning in the text, for the words and information sciences are not connected semantically; when considering the structure of the text, the fact that the terms are placed in separate sentences is take into account: On the internet we can find a lot of information. Science works to improve quality of life.

To handle these exceptions, we use a strategy similar to that adopted by Mikheev (2002) that considers the local context of the document and applies a small set of rules for making the disambiguation. However, in the context of this work these rules could be relaxed without affecting the final result. The process of separating the text into sentences and words to create the vocabulary words is known as tokenization. Manning, Raghavan & Schütze (2009, p. 22-26) define tokenization as the task of receiving as input a given sequence of characters in a document and split it into parts called tokens, while discarding those characters that indicate the points of separation.

After the text is broken up into sentences, they must be broken into words in order to become or not a term in the vocabulary. The characters usually used to indicate the separation of the words are the comma, the hyphen and blank space. But they can not be considered as separators on an unrestricted basis. For example, the comma can be used to separate whole numbers from decimals in the European model of numerical representation, or the thousands in the Saxon model; the hyphen may be used to divide syllables of a word at the end of a line, or compounds that can be found in different spellings, in the case of the blank space, the problem occurs when it is used to separate the names, in which case the terms should not be separated because they made a semantic sense.

To mitigate these problems we used some strategies described below. In the case of the comma it is discarded, so the numerical representations are expressed only by numbers without separators. In the case of the hyphen in the Portuguese language such as: "infraestrutura<sup>1</sup>", "infra-estrutura<sup>2</sup>" or "infra estrutura<sup>3</sup>"; by making a Google search

<sup>1</sup> Under the new Portuguese orthographic agreement in effect as of 2009.

for three terms two different results are found. When searching for “infra-estrutura” or “infra estrutura”, approximately 3,960,000, links were found while “infraestrutura” found approximately 3,420,000 responses. Therefore this is still an open question. In this study we will ignore the hyphen, thus, words spelled with a hyphen will be treated as a single word, and syllable breaks of the dash, when removed, will regroup the word. In the case of the blank space, the problem is found in the contents with proper nouns, such as “New York”, because the semantic meaning in this case must be made by the two words together, not as two separate entries in the vocabulary. In this work, this problem becomes irrelevant, because if these words are relevant in the context of the document, they will become a bigram and will be found only if the sequence in the document collection. Therefore, during the process of converting a text, a treatment was carried out byte by byte characters where the following tasks were performed: (1) To identify and convert all accented characters, which are represented by multibyte characters, the usual Portuguese, transforming them into non-accented characters while preserving the original spelling of the text of uppercase and lowercase letters, (2) to remove the hyphens, (3) to remove the dot (.) that is used to abbreviate words, (4) to remove the periods (.) and commas (,) used as separators of numbers, (5) to remove expressions such as “[...]” “(...)”; (6) Delete all ASCII bytes whose value is less than 1 or greater than 126. All these steps were performed in order to minimize the error parser separation of sentences.

Thus, the rules adopted to consider the existence of a delimiter sentence were: (1) If you find any of the following characters: question mark, exclamation point, (2) If after the (.) period there is a line breaking character, a new paragraph, end of a text or a capital letter. All characters used as separators are eliminated from sentences.

### *3.1.3 Decoding Acronyms*

A very common practice of writing, especially in science, is the use of abbreviations. Typically, the first appearance terms are shown in full with the letters that make up the acronym in each term presented in uppercase followed by the acronym itself with capital letters separated or not by a period between brackets. From this premise, in this sub step the goal is to identify acronyms in order to build a table of acronyms used in each document, and add to the part in full text whenever when the acronym occurs. This strategy is important to be adopted, since the content expressed in the text only as an acronym would not be interpreted as MWE. While in fact this kind of content is usually high in semantic content to express the meaning of the document, and when it is placed in full, depending on its frequency of occurrence, it makes this set of terms become MWE.

### *3.1.4 Segmentation sentences into terms*

In this sub step, the goal is to separate the sentences into terms in order to create the vocabulary of terms. Tokens, ie, the pieces that were targeted, normally go through a standardization process before they become a term of the vocabulary. Normalization aims to reduce the number of dictionary entries. In this sense all words are transformed into lowercase.

---

<sup>2</sup> Spelled before the agreement.

<sup>3</sup> Spelled incorrectly, but that could be found.

### 3.1.5 Secondary Filtering – Stop Words

In this substep, after breaking the documents into a word sequence, a new filter is executed. The goal is to remove the vocabulary words that appear very frequently in all documents and, therefore, have little power of discrimination. Manning, Raghavan & Schütze (2009, p. 27). defined stop words as common words that seem to have little value to select the corresponding documents. These words usually belong to the class of articles, prepositions and some conjunctions. These authors explain that a strategy that can be used to determine the list of stop words is to count the number of times each term appears in the collection of documents, and to verify, often manually, which the semantic relevance of the term is in relation to field of documents being indexed. Those considered relevant are included in the list of stop words. For the purpose of this paper the use of the list of stop words contributes positively. For example content, "information science", treated without the filter would be a stop word trigram, after filtering it, it would be transformed into a bigram. After removal of the stop word, the size of each returned term is verified after performing the break of the sentence into words, and those with only one character are the discards.

## 3.2 Server features

This software component aims to index the corpus and provide a document recovery service accessed via the network through a number of IP and communication port. This component performs the following steps:

1. to convert the document into standard terms (described in Section 4.1);
2. to indexing terms;
3. to provide a recovery service of documents by searching for keywords.

Steps 2 and 3 will be detailed below.

### 3.2.1 Index terms

The purpose of this step is to build an inverted list of standardized terms pointing to the documents in which they are referenced. Additionally, we use the technique described by positional index of Manning, Raghavan & Schütze (2009, p. 41-43). This technique consists in adding to the inverted structure list the position or positions controlled from a numerical sequence containing the position where the term has been found in the document. That is, how much of the sentence and how many words within the sentence. This allows to perform searches where it is desired to find an expression containing consecutive terms of a single sentence, as it is necessary for identifying MWE. It should be noted that in the search time, it is necessary to perform the search, separately, each of the terms of expression, and from the result returned for each one of them it is possible to verify if they are consecutive. Figure 3 shows a sketch of the data structure used by this technique. Where:  $\{t_1, t_2, t_3, \dots, t_n\}$  represent the vocabulary terms;  $\{d_1, d_2, d_3, \dots, d_n\}$  represent the documents;  $\{p_1, p_2, p_3, \dots, p_n\}$  represent the position of the sentence and word within the sentence in which a particular term was found in a document, and,  $\{r_1, r_2, r_3, \dots, r_n\}$  represent a reference to where the document is stored.

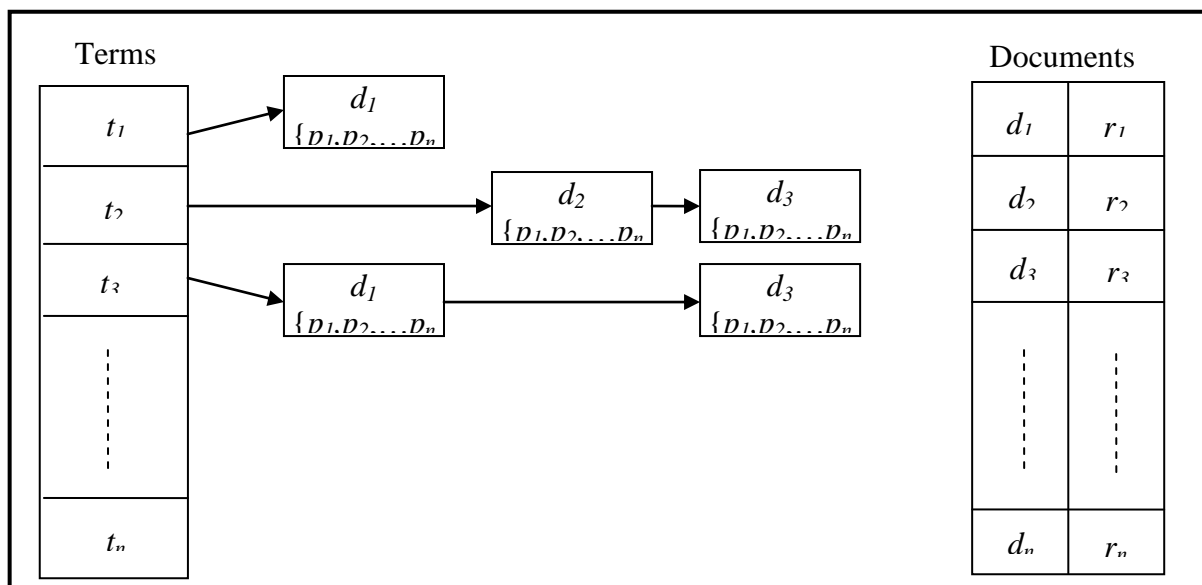


Figure 3 - Outline of the data structure used in the inverted list with a positioned index.  
Source: Prepared by the authors.

### 3.2.2 Provide consultation service

After the entire *corpus* has been processed and the documents are indexed in the volatile memory of the computer, the function of this stage is to provide a consultation service through a communication protocol between the two software components, the Server and Client. The communication protocol consists of sending the client a list of all bigrams extracted from the reference document search and return the response by the server with a reference link to similar documents found in the *corpus*. For each bigram the search for each of its separate terms will be processed. The results will be analyzed by checking the terms of each bigram found in the same sentence of the same document and adjacent. In this case the coefficient will be computed as relevant, otherwise the item response is discarded so that the next item can be analyzed.

### 3.3 Client features

This software component aims to consult the *corpus* from a document (PDF), which related documents exist. That is, a search process which will be extracted compared in all MWE found in the base document expressed using bigrams that will be sent to the service provided by the Server. Requisitions with the descriptors are sent by Client via a communication protocol TCP / IP network established via the Server. In the same way the answers are returned to the Client. This component performs the following actions:

1. to receive the document used as a reference search;
2. to extract MWE from the documents and generate a list of bigrams;
3. to send the request to the Server with the list of bigrams;
4. to return the search result.

#### 3.3.1 Receiving the document used in reference search



At this stage the goal is to develop a web application that serves as end-user interface compared to the process of document search. To develop the interface the PHP language was used and the software component created is called "Search". This interface is in charge of receiving the document and uploading it in order to make the call request to the Client through the document as a parameter processing. The PDF document will be converted into standardized terms (as described in Section 4.1). Figure 4 shows a sketch of the interface screen.

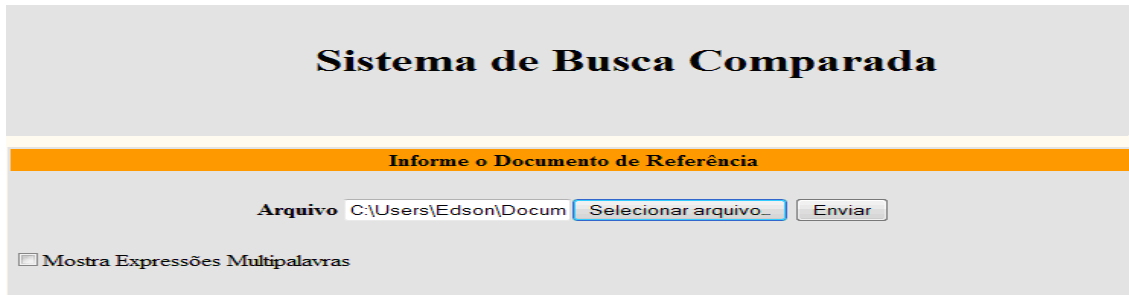


Figure 4 - Display the prototype, which reports the documents used in compared search. Source: Prepared by the authors.

### 3.3.2 Extract the bigrams

At this stage, the standard terms extracted from documents used as references for the search are identified by number and position of the sentence in the sentence in order to organize them into a data structure in memory that allows the extraction of MWE. The proposed structure is shown in Figure 5.

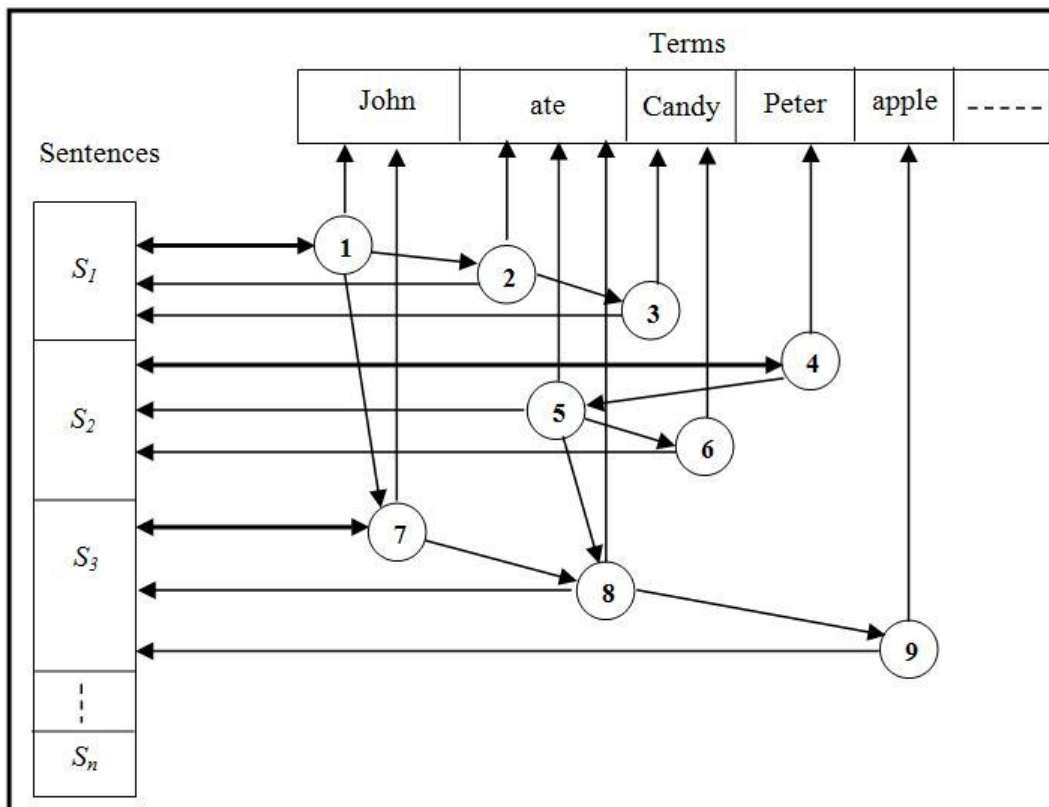


Figure 5 - Representation of the data structure created to extract MWE. Source: Prepared by the authors.

To understand this structure, we will consider the document, the string is composed of  $S = \{ S_1, S_2, S_3 \}$  sentences shown below:

$S_1 \rightarrow$  John ate a candy.

$S_2 \rightarrow$  Pedro ate a candy.

$S_3 \rightarrow$  John ate an apple.

By considering that the reference document contains only the  $S_1$ ,  $S_2$  and  $S_3$  sentences, after performing the segmentation of text into sentences and words the result is a set of standardized terms  $V = \{ T_1, T_2, T_3, T_4, T_5 \}$ , as shown in table 1.

Table 1 – Standardized terms

Identification	Terms
$T_1$	John
$T_2$	ate
$T_3$	candy
$T_4$	Peter
$T_5$	Apple

Source: Prepared by the authors.

And finally, we consider that set of nodes  $N = \{ 1, 2, 3, 4, 5, 6, 7, 8, 9 \}$  representing each of the nine words of the text being arranged within the proposed structure.

Because the processing is performed in the sequence in which the sentences are read. By reading the  $S_1$  sentence, the terms  $T_1$ ,  $T_2$  e  $T_3$  are processed referenced by three nodes, 1, 2 e 3 respectively. By reading the  $S_2$  sentence, the terms  $T_4$ ,  $T_2$  e  $T_3$  are processed referenced by nodes 4, 5, 6 and so on.

After all sentences processed, the proposed structure allows us to identify what the existing phrases are in the string and the which sentences in which a term occurs.

To extract MWE, the algorithm goes through the sentences checking each word and which its adjacent words are: then there is the frequency of repetition at which adjacent terms occur. MWE with a frequency (Fr), number of repetitions, from a given parametrized value are considered as relevant. In this experiment we used three as the value of this parameter. The following is a pseudo-code with the steps of this process.

```

while (there are sentences) do
  term = nextTerm(Sentence)
  while (there are Adjacent) do
    Adjacent = findAdjacent()
    if unprocessed(Term)
      totAdjacent = countAdj(Term)
      if totAdjacent >= Nr
        Insert(Term, Adjacent)
      endif
    endif
  endwhile
endwhile

```

One point that should be highlighted is that although this processing is used to extract only bigrams, it does not mean that expressions with  $n$ -grams are not considered. In practice, the process can handle any number of consecutive terms that have a frequency equal to or above the observed quantity defined in the parameter. This can be done for any set of  $n$ -grams converted into pairs of bigrams. In the following example the Trigrams "University of Model Example" is transformed into two bigrams: "university model" and "model example". The term "of" is dropped in order to function as a stop word.

### 3.3.3 Send the request to the server with the list of bigrams

At this stage, the list of bigrams that expresses the semantic meaning of the document, in which one seeks to find similar documents, will be submitted through a request to the Server via a network communications protocol.

The list contains the bigrams and after the receipt by the Server, it will be split into pairs of terms. Peer-to-peer, each term is searched in the *corpus* to produce a list with the answers and documents in which these terms were found. The responses are identified by document number, sentence number and position of the word in the sentence. Thus, each MWE will be validated according to the responses received. The documents whose answers to the terms of MWE are not adjacent are discarded. The remaining responses are modulated according to the frequency of occurrence and the structural coefficient ( $Sc$ ) parameterized according to the shape of the spelling of the word in the text. The answers will eventually be sorted by relevance and presented only those corresponding to a percentage, defined by parameter, among the best responses. In other words, a cutoff point will be used where only those documents with better results than the percentage reported are to be presented as a response.

This processing can be better understood by observing the algorithm shown below, considering:

- $C$  *corpus* containing the documents.
- $B$  is the set of bigrams extracted from the reference document of the search.
- $Sc_a$  e  $Sc_b$  is structural coefficient of the term "a" and the term "b", respectively.
- $B = \{(t_{1a}, t_{1b}), (t_{2a}, t_{2b}), \dots, (t_{na}, t_{nb})\}$  – Bigrams formed by  $n$  pairs of terms.
- $R_a = \{(d_{1a}, s_{1a}, p_{1a}), \dots, (d_{na}, s_{na}, p_{na})\}$  – Answers search conducted of the  $i$ -ith term  $t_{1a}$  in the collection of documents  $C$ . Returns containing the triple where the terms were found:  $d$  = documents,  $s$  = sentence,  $p$  = position.
- $R_b = \{(d_{1b}, s_{1b}, p_{1b}), \dots, (d_{nb}, s_{nb}, p_{nb})\}$  – Same as before, only referring to the term "b" of the bigram.

The triple each of the terms "a" and "b" of bigram are compared to verify if they are adjacent.

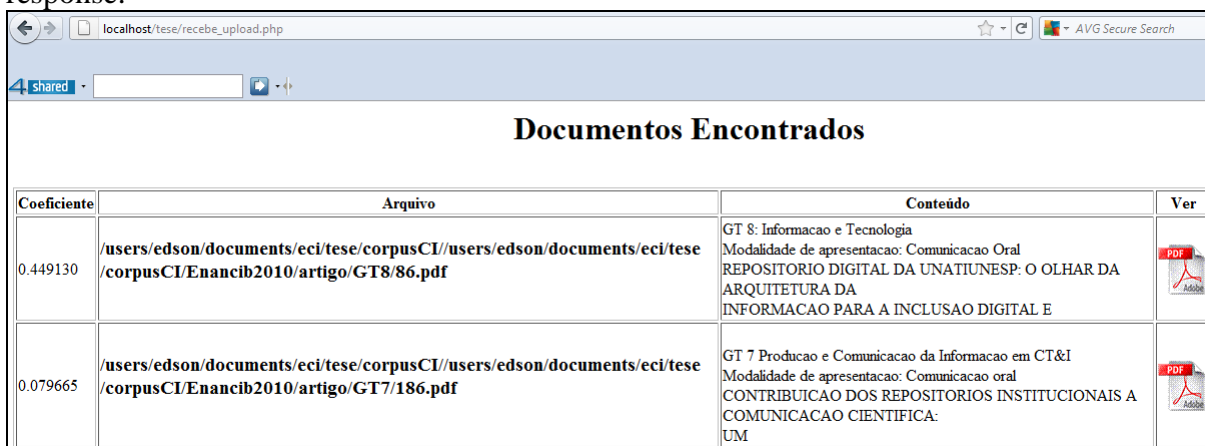
```

1 for x from 1 to n do
2    $R_a = \text{search}(t_{xa}, C)$ 
3    $R_b = \text{search}(t_{xb}, C)$ 
4   do
5     if ( $d_{xa} < d_{xb}$ ) then
6       next  $d_{xa}$ 
7     else
8       if ( $d_{xa} > d_{xb}$ ) then
9         next  $d_{xb}$ 
10      endif
11     endif
12   until  $d_{xa} = d_{xb}$ 
13   do
14     if ( $s_{xa} < s_{xb}$ ) then
15       next  $s_{xa}$ 
16     else
17       if ( $s_{xa} > s_{xb}$ ) then
18         next  $s_{xb}$ 
19       endif
20     endif
21   until  $s_{xa} = s_{xb}$ 
22   if ( $p_{xa}$  adjacent  $p_{xb}$ ) then
23      $\text{weightDoc}[I] = \text{weightDoc}[I] + t_{xa} * Sc_a + t_{xb} * Sc_b$ 
24   endif
25 endfor
26 sort(weightDoc)
27 showRelevant(documents)

```

### 3.3.4 Display the search result

At this stage the client will receive the Server response containing a reference search to access all documents that were considered similar. A page with these responses in order of relevance will be displayed allowing the user to query view the full document from a click on its reference. Figure 6 shows an outline of the screen response.





Coeficiente	Arquivo	Conteúdo	Ver
0.449130	/users/edson/documents/eci/tese/corpusCI/users/edson/documents/eci/tese/corpusCI/Enancib2010/artigo/GT8/86.pdf	GT 8: Informacao e Tecnologia Modalidade de apresentacao: Comunicacao Oral REPOSITORIO DIGITAL DA UNATIUNESP: O OLHAR DA ARQUITETURA DA INFORMACAO PARA A INCLUSAO DIGITAL E	
0.079665	/users/edson/documents/eci/tese/corpusCI/users/edson/documents/eci/tese/corpusCI/Enancib2010/artigo/GT7/186.pdf	GT 7 Producao e Comunicacao da Informacao em CT&I Modalidade de apresentacao: Comunicacao oral CONTRIBUICAO DOS REPOSITORIOS INSTITUCIONAIS A COMUNICACAO CIENTIFICA: UM	

Figure 6-screen response with the documents found.

Source: prepared by the authors

As can be seen, the screen displays four columns in the interface response: similarity coefficient, the link with the physical address of the document in the *corpus*, the first two hundred characters of text after being partially converted and filtered, and an icon to access the document complete.

### 3.4 Evaluation of bigrams extracted

To perform the empirical tests in a *corpus* composed of full articles was evaluated, published in major scientific meeting of the area of Information Science (ENANCIB) of 2010. All documents were obtained in Portable Document Format (PDF) and stored in a computerized system of files organized in folders and subfolders in a hierarchical way by the Working Groups (WG). The total corpus of 193 articles was typically containing between 20 to 25 pages, totaling 687,490 normalized terms, 7970 was different.

Figure 7 shows the frequency distribution of co-occurrence of bigrams found in the *corpus*.

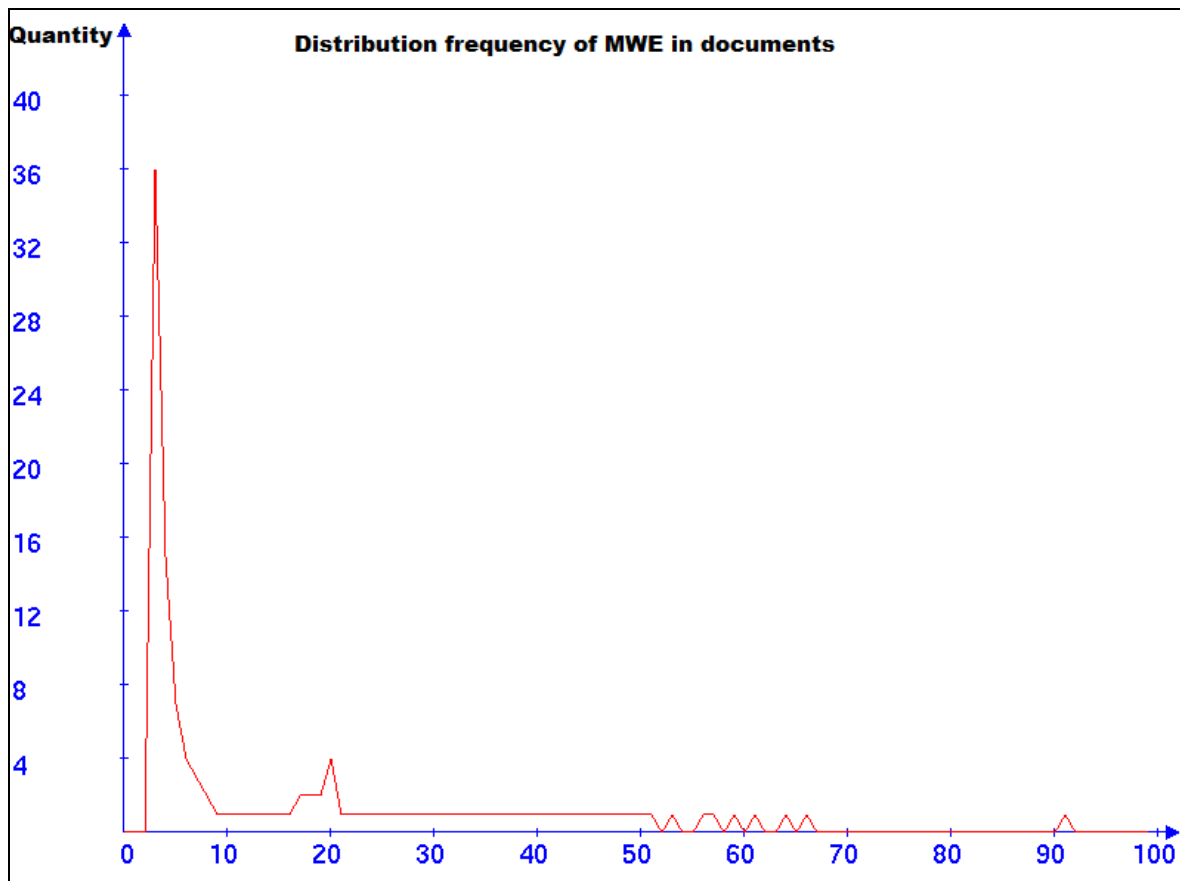


Figure 7 - Representation of the data structure created to extract MWE.

Source: Prepared by the authors.

After the standard corpus is indexed into memory, and it is necessary to compare the result of MS Heudet extracted by the technique proposed by the authors with the software NSP, extracted through thirteen different statistical techniques shown in Table 2. For each document of the corpus, fourteen files were generated with MWE, one for



each technique. The data files have been loaded into the MySQL database in order to facilitate the comparison of the MWE extracted by different techniques.

Table 2 – List of statistical association measures implemented by NSP.

Number	Measure Statistical Association (NSP)
01	Log-likelihood Ratio
02	Pointwise Mutual Information
03	Mutual Information
04	Poisson Stirling Measure
05	Left Fisher
06	Right Fisher
07	Fisher Twotailed Test
08	Phi Coeficcient
09	Tscore
10	Person's Chi Square Test
11	Coeficiente Dice
12	Jaccard Coeficcient
13	Odds Ratio

Source: Prepared by the authors.

#### 4 RESULTS AND CONCLUSIONS

The result of the comparison processing of the techniques is shown in Table 3. It is observed that all the techniques of statistical measures of association returned a similar number of MWE, an average of 15,063, although the relevance coefficient obtained by the techniques are different. While Heudet drew 14,755, and 14,343 of that total are common to the combined results of the thirteen measures taken by the NSP. Column (A) shows the total values of MWE extracted by fourteen different techniques. Column (B) shows the related quantities extracted from MWE, having one or more of their terms in accordance with one character, an average of 155 cases. These cases were discarded for technical Heudet. Therefore, these values are subtracted to NSP and values shown in column (C). Column (D) shows the amounts of the common MWE found when comparing Heudet with each of the techniques of the NSP. Finally in column (E), the net difference of the extracted MWE is presented by comparing Heudet with each of the techniques of the NSP. An average of 565 cases, and these are related to two situations: The first 223 cases corresponding to the difference between the average net values extracted by the NSP (14,908) and the amount extracted by the Heudet technique (14, 755), involving a gain in the identification of MWE compared to other NSP techniques, and the second, 343 cases drawn, mainly corresponding to MWE, bordering points of adjacent sentences. These cases were discarded by the Heudet technique because, in order for them to be considered as MWE, it is necessary that the bigrams are in the same textual element, the same sentence. This strategy is not adopted by other statistical techniques to consider the text as a bag of words.

Table 3 – Results of extraction of MWE.

Technical	(A) Quantity of MWE extracted	(B) Noise	(C) A – B	(D) Commum with heudet	(E) C – D
<b>Odds</b>	15054	155	14899	14324	575
<b>X2</b>	15055	155	14900	14329	571
<b>Os</b>	15062	154	14908	14344	564
<b>Jaccard</b>	15063	155	14908	14338	570
<b>Ll</b>	15063	154	14909	14345	564
<b>Tscore</b>	15063	153	14910	14345	565
<b>Phi</b>	15064	155	14909	14338	571
<b>Dice</b>	15064	155	14909	14339	570
<b>Twotailed</b>	15065	158	14907	14364	543
<b>Lfisher</b>	15065	158	14907	14365	542
<b>Pmi</b>	15067	159	14908	14333	575
<b>Tmi</b>	15068	154	14914	14349	565
<b>Rfisher</b>	15068	154	14914	14351	563
<b>Average</b>	15063	155	14908	14343	565
<b>Heudet</b>	14755	-	-	-	-

Source: Prepared by the authors.

That is, 14,343 corresponding to 96.21% of the extracted MWE are identical regardless of the technique you used. 223, corresponding to 1.5%, are different MWE, exclusive of the Heudet technique, which can be regarded as an accurate gain. The 342 remaining 2.29% of the corresponding MWE extracted by different NSP can be considered as noise that shows inaccuracy. The processing time for the extraction of the entire *corpus*, through the Heudet technique, consumed 197 seconds running on a UCP core™ 2 Duo T6400 2.0 Ghz notebook. Therefore, we conclude that the deterministic technique used for this specific purpose has advantages in terms of accuracy, simplicity and performance. Figure 8 shows an outline of the result.

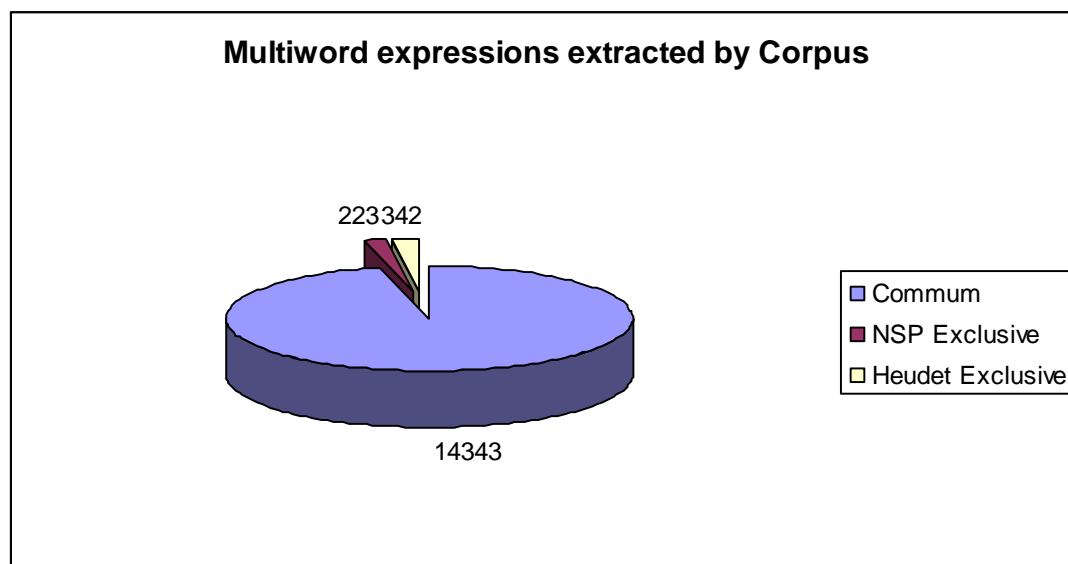


Figure 8 - Comparison of MWE obtained by various techniques.

Source: Prepared by the authors.

## 5 RECOMMENDATIONS FOR FUTURE WORK

The result obtained by Heudet technique is promising, because it showed better responses than those obtained for thirteen different statistical techniques exclusively. This can be explained by checking that statistical techniques do not consider the physical structure of the document during the extraction process of MWE. For them the text is a sequence of words in sentences that do not exist. The results could be further enhanced with the creation of new heuristics that aim to identify the inherent characteristics of the physical structure of the document that may assist in identifying MWE, further improving the results.

## REFERENCES

- BERGER, Peter L., LUCKMANN, Thomas. *A construção da realidade*. Petrópolis: Floriano de Souza Fernandes, 1985.
- CALZOLARI, Nicoletta FILLMORE, Charles J.; GRISHMAN, Ralph, IDE. Nancy; LENCI, Alessandro ; MACLEOD, Catherine ; ZAMPOLLI, Antonio 2002. Towards best practice for multiword expressions in computational lexicons. Em Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002), pp. 1934–1940, Las Palmas, Canary Islands.
- DIAS, Gaël ; LOPES, José Gabriel Pereira ; GUILLORÉ, Sylvie. Mutual expectation: a measure for multiword lexical unit extraction. In Proceedings of Vextal, 1999.
- EVERT, Stefan ; KREEN, Brigitte. 2005. Using small random samples for the manual evaluation of statistical association measures. *Computer Speech and Language*, 19(4):450–466.
- KURAMOTO, Hélio. Uma abordagem alternativa para o tratamento e a recuperação da informação textual: os sintagmas nominais. *Ciência da Informação*, Brasília v. 25, n. 2, mai/ago, p. 182-196, 1996.
- MAIA, Luiz Cláudio ; SOUZA, Renato Rocha. Uso de sintagmas nominais na classificação automática de documentos eletrônicos. *Perspectivas em Ciência Informação*, Belo Horizonte, v. 15, p. 154-172 , 2010.
- MANNING, Christopher D.; RAGHAVAN, Prabhakar ; SCHÜTZE, Hinrich. *An introduction to information retrieval*. Ed. Cambridge online, 2009.
- MIKHEEV, Andrei. Periods, capitalized words, etc. *Computacional Linguistics*, 28(3), 289-318, 2002.
- PEARCE, Darren. A comparative evaluation of collocation extraction techniques. Em of the Third (LREC 2002), Las Palmas, Canary Islands, Spain, May, 2002.
- PECINA, Pavel ; SCHLESINGER, Pavel. Combining Association Measures for Collocation Extraction. In *ACL'06*, page 652, 2006.
- PORTELA, Ricardo Jorge Rosa ; MAMEDE Nuno ; BATISTA, Jorge. Multiword Identificação. In *Terceiro Simpósio de Informática Portugal* pp. 110-199, 2011.

RANCHOLD, Elisabete M. O lugar das expressões ‘fixas’ na gramática do Português. in Castro, I. and I. Duarte (eds.), *Razão e Emoção*, vol. II, Lisbon: INCM, pp. 239-254, 2003.

RAMISCH, Carlos. Multiword terminology extraction for domain specific documents. Dissertação – Mathématiques Appliquées, École Nationale Supérieure d’Informatiques, Grenoble, 2009.

RUSSELL, Stuart J; NORVIG, Peter. *Inteligência Artificial*. Rio de Janeiro: Campus, 2004. 1021p.

SAG, Ivan A. ; BALDWIN, Timothy ; BOND, Francis ; COPESTAKE, Ann ; FLICKINGER, Dan. Multiword expression: a pain in the neck for nlp. Em *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing CICLing-2002*, volume 2276 of (Lecture Notes in Computer Science), pp. 1–15, London, UK. Springer-Verlag.

SARMENTO, Luís. Simpósio Doutoral Linguatca 2006. Disponível em: <http://www.linguatca.pt/documentos/SimposioDoutoral2005.html>: out. 2011

SILVA, Ferreira J. LOPES Pereira G. A local maxima method and fair dispersion normalization for extracting multi-word units from corpora. (1999). *Sixth meeting on Mathematics of Language*, pp. 369-381.

SOUZA, Renato Rocha. Sistemas de recuperação de informações e mecanismos de busca na web: panorama atual e tendências. *Perspectivas em Ciência Informação*, Belo Horizonte, v. 11, n. 2, Aug. 2006.

VIGOTSKY, Lev Semenovich. *Pensamento e Linguagem*. Edição eletrônica: Ed Ridendo Castigat, 1987.

VILLAVICENCIO, Aline ; RAMISCH, Carlos; MACHADO, André; CASELI, Helena de Medeiros; FINATTO, Maria José. Identificação de expressões multipalavra em domínios específicos. *Linguamática*, v. 2, n. 1, p. 15-33, abril, 2010.

ZHANG, Wen; YOSHIDA, Taketoshi; TANG, Xijin; HO, Tu-baq. Improving effectiveness of mutual information for substantial multiword expression extraction. *Expert Systems with Applications*, Elsevier, v. 36, 2009.