AUTOMATIC INDEXING OF TEXT DOCUMENTS: ESSENTIAL CRITERIA PROPOSAL

Graciane Silva Bruzinga Borges[1], & Gercina Ângela Borém de Oliveira Lima[2]
[1]Universidade Federal de Minas Gerais, Brasil.
[2]Universidade Federal de Minas Gerais, Brasil.

ABSTRACT
The volume of documents currently being published demands great effort in order to devise alternative indexing techniques for them; undertaking this process manually is a very slow process. In this research project, automatic indexing criteria that have been used most frequently since the 1950s have been identified. The work was divided into two main stages: (1) identification of the automatic indexing criteria found in the literature and, (2) a proposal for a set of ideal criteria for the automatic indexing process.

KEYWORDS: Automatic indexing; manual indexing; information representation; criteria for automatic indexing

1. Introduction

Indexing is the action of representing a document through an abbreviated description of its content with the goal of identifying its essence. This representation is realized through an analysis of the source text, which, necessarily, must be performed by specialists who pay close attention to methodologies and procedures. There are at least two ways of undertaking this process: manual indexing and automatic indexing. Investigations into traditional indexing identify alternatives that are aimed at improving the capacity of professionals to do the abstractions, and to ensure that the thematic representation is as close as possible to the author's original content. With regard to studies of automatic indexing, it is notable that its rise is due to the need to solve problems such as the slowness of manual indexing. It is for this reason that automatic indexing is seen as an alternative for speeding up this process, through the utilization of the resources offered by technology.

The limited time that an indexer has to perform a manual analysis of the subject can interfere with the quality of the indexing. It is for this reason that automatic indexing techniques are being proposed. Though some of them are not totally satisfactory, they may enhance the manual process by providing an initial extraction of terms, thus allowing the indexer to select those that are the most appropriate for representing the document. Another benefit that may result from this technique is a reduction in subjectivity, which is a characteristic inherent in the intellectual aspect of the task.

The literature in this area identifies other practical problems with manual indexing, such as: (1) different indexers attribute different terms to the same document; (2) the same indexer attributes different terms to the same document at different times; (3) the indexer's knowledge of the subject may affect the level of consistency achieved in the activity; (4) the ability of the indexer to remain current with the state-of-the-knowledge and (5) the ability to understand the language of the document being dealt with.

2. Manual indexing

The personal/intimate capacity to recognize what the document being analyzed deals with is a critical issue in the indexing process. For indexing purposes, the terms selected are the behavioral correlation between what the indexers think the document is about and "what the document is really about", as they would be the terms applied to the search for a specific document (Maron *apud* Guedes) [1]. According to Unisist [3], the process utilized to describe and identify a document according to its subject is called indexing. As it is an intellectual activity, it is natural that, in the day-to-day activities of indexers, there may be divergences between the terms attributed to a single document by professionals from different institutions and in different contexts. Therefore, according to Lancaster [4], a single publication may be represented by different sets of indexing terms, depending on the group of users it is aimed at and on the particular interests of this group.

According to Silva and Fujita (p. 136-137) [5], “the concept of indexing arose from the development of indices and is currently more closely linked to the concept of subject analysis”. With the need to recover information in an ever faster, more precise and more specialized manner, the practice of developing indices has come to focus more specifically on the content of each document. In this work, the manual indexing process was considered as having two main stages: subject analysis and the translation of this analysis, or that is to say, the translation of the document content into indexing terms.

2.1. Subject analysis

The aim of the subject analysis step is to determine what a document deals with, that is, its topic. To this end, reading and comprehension of the text are of paramount importance. However, the limited time available to the indexer and the ever growing amount of documents demanding treatment are worrisome factors that may compromise the quality of the activity. According to Lancaster (p. 20-21) [6], “the indexer is rarely given the luxury of being able to read a document from beginning to end”. In order to perform this step, one must consider the domain to which the document belongs, identifying the traits specific to the field of knowledge, be they cultural, terminological, historical or linguistic. To this end, the knowledge of the indexer about this domain is important to the quality of the analysis. This way, the activity can be performed in accordance with the context, since the document would not be considered an isolated entity, but rather a part of a whole (Hjorland) [7].

According to some studies, like Unisist (p. 83) [3] and Fujita (p. 64) [8], subject analysis is divided into three stages: (1) understanding the context of the document as a whole; (2) identifying the concepts that represent this content and (3) selecting concepts valid for recovery. According to Unisist [3], “in practice, these three stages overlap”. In the opinion of Fujita (p. 64) [8], this overlap occurs at the moment the document is read. The final stage ends when the articulation of the so-called *indexing phrase* is made, which is produced by the indexer in Natural Language – NL. Once all the intellectual processes of reading and understanding the text, and identifying and selecting concepts representative of the document in question are done, the indexer must state: *This document is about…* Once this statement has been made, the indexer can move on to the final step in the indexing process, the translation of the subject analysis into indexing terms.

2.2. Translating the subject analysis

Subject analysis translation aims at distilling the subject of the document into a set of indexing terms. This analysis will occur even in cases where no formal rules have been prescribed. Such rules can be stipulated to serve the interests of the institution or of the terminological control instrument. This control can be exercised through the use of a *controlled vocabulary*, and is often performed intuitively. Some of the main controlled vocabularies used within the scope of librarianship are: Taxonomy, Thesaurus, List of Subject Headers and Bibliographic Classifications. As this is an intellectual activity performed by an individual, even if this individual is specialized in the area, indexing is a complex activity in which one can encounter significant difficulties. Therefore, in the 1950s, studies on automatic indexing processes, which considered computational resources, were initiated with the goal of speeding up the subject analysis step of the indexing process.

3. Automatic indexing

Also known as *computer assisted indexing*, also as *semi-automatic indexing,* this type of indexing is considered an extraction model with statistical and probabilistic traits. Its origins coincide with the first attempts at joining the fields of information technology, statistics and documentation. To Moreiro González (p. 3 apud Bufrem) [9, 8],

> *[…] The essence of the process is the automatic identification of keywords within the text through the frequency with which they appear and its theoretical foundation lies in Zipf's law. New formulations of this law gave rise to other techniques for discriminating terms, which the author discusses, emphasizing statistical indexing of terms by frequency, known by the acronym IDF; Term Frequency Inverse Document Frequency (TFIDF); the N-grams method that modifies Zipf's law in order to enable the handling of composite words; and Stemmers, which use the frequency with which sequences of letters appear in the body of a text to extract word stems. Beyond these possibilities, the semantic relations between linguistic terms can be established by grouping and classification methods.*

According to Robredo [11], “the automatic indexing process is similar to the human reading-memorization process, insofar as its general principle is based on the comparison of each word in the text with a list of meaningless words”. This list must be established beforehand and the result of this comparison leads one to consider, through a process of elimination, if the remaining words in the text have meaning. The history of automatic indexing can be associated with the use of software for the generation of pre-coordinated indices (Robredo) [11]. In the opinion of Naves [12], examples of pre-coordinated languages include: “lists of subject headers (Library of Congress, Rovira, Wanda Ferraz), permuted indexes, chain indexes and bibliographic classifications (Dewey Decimal System – DDS, Universal Decimal System – UDS)”.

In the late 1950s, relatively simple methods for building indices from texts, especially from words found in the titles of documents, were developed. The Keyword in Context (KWIC) method was developed by H. P. Luhn in 1959 and constitutes a rotational index in which each keyword found in document titles becomes an entry in the index. The Keyword Out of Context (KWOC) method is similar to the KWIC method, but the keywords that are used as access points are repeated outside of their contexts, usually highlighted in the left margin of the page or used as subject headers. In addition to KWIC and KWOC, there are also Selective Listing in Combination – SLIC, created by J. R. Sharp in 1966, which organizes the sequence of terms from a document alphabetically and eliminates the redundant sequences; and, Preserved Context Indexing System – PRECIS, created by Dr. Derek Austin in 1968, which produces a printed index based on alphabetical order and in the systemic “alteration” of terms so that they occupy the entry position (Lancaster) [4]. Another relevant system developed was the Nested Phrase Indexing System – NEPHIS, created by T. C. Craven in 1977, which is an articulated subject index. In this model, the entry terms are reordered in such a manner that each one is linked to its original neighbour through a functional word or using special punctuation, thus maintaining a structure similar to that of a phrase, albeit, oftentimes a scrambled one.

According to Salton [13, 14] and Swanson [15], automatic indexing presents relative merits when compared with manual techniques. Researchers claimed it enabled automatic extraction of relevant keywords from texts, and that comparisons between such words and those designated by indexers found a 60 – 80% coincidence in the designated terms. From the 1970s onwards, research into the automatic indexing of textual documents was intensified. Two of the most important such experiments were based on the performance of the MEDLARS SRI, used in the National Library of Medicine, Washington D.C.; and, of the SMART experimental SRI, created by Gerard Salton while working at Cornell University (Salton) [14]. Literature points to some types of automatic indexing. *Automatic indexing through extraction* is one of them. In this process, words or expressions that appear in the text are extracted to represent its content as a whole. The principles employed attempt to emulate those that would be used by human indexers (Lancaster) [4].

In the 1950s, automatic indexing based on word occurrence frequency began with the work of Luhn, in 1957, and Baxendale, in 1958. Baxendale (*apud* Lancaster) [16] suggests that, rather than going through the process of analysing the whole text, only the “phrasal topic” and the “suggestive words” are analyzed. His studies demonstrated that only the first and last sentences needed to be processed since, in 85% of the cases, the first phrase corresponded to the phrasal topic, and in 7% of the cases, the last phrase corresponded to it. The phrasal topic is defined as the portion of the text that contains the most information about it.

Systems based on automatic extraction indexing essentially perform the following tasks: (1) counting words in a text; (2) comparing them to a list of prohibited words; (3) eliminating the meaningless words (articles, prepositions, conjunctions, etc.); and, (4) ordering the words according to their frequency. It has been noted that this type of indexing is limited in its ability to conduct the process in a consistent manner. A similar process, but one that includes a concern for the semantic aspects of the text, is *indexing through automatic designation*. This process frequently presents difficulties, since the representation of thematic content requires terminological control to develop a “profile” of the words or expressions that frequently occur in the documents for each designated term (O’connor *apud* Lancaster) [17].

Another type of automatic indexing mentioned in the literature is *automatic full text word identification*, which analyzes the whole document without taking textual semantics or word position within the sentence into consideration. There are also *automatic syntax indexing* that aims at analyzing the most relevant words in a sentence, as well as *automatic semantic indexing* which is based on the principle that the document already has formatting structures that indicate the semantics of the terms.

4. Automatic indexing analysis criteria used for treating textual documents

4.1. Methodology

The study method employed involves two steps: (1) *criteria identification*, which is subdivided into two stages: (a) defining the universe and the study sample, and (b) defining the empirical object through criteria systematization; and, (2) *Analysis of criteria combinations,* also subdivided in two stages: (a) selecting a second sample from the studied universe, and (b) interpreting the criteria.

4.2. Criteria identification – step 1

4.2.1 Defining the universe and selecting study sample no. 1

The study universe for this work comprises technical-scientific papers, dissertations, theses and books on automatic indexing that present the results of research in this field. The document should necessarily present the research methodology and state conclusive results about the relevance of the automatic indexing criteria used.

The study sample was composed of 103 domestic and international research papers published between the 1950s and 2008 on this subject. The bibliographic research was carried out according to the following strategy for document selection: (1) delimiting the main research goal and its purpose; (2) indicating the keywords used to delimit the subject in English and in Portuguese; (3) determining the types of documents that would be included in the sample; (4) selecting the sources of information for the bibliographical research; (5) delimiting the volume of the sample; (6) indicating the format of, and the support for, the selected documents; (7) determining the search strategy; (8) defining the portions of the documents to be considered for technical reading; and, (9) analyzing the selected sample. Once this sample was analyzed, it was possible to undertake the procedures described in the stage below.

4.2.2 Defining the empirical object and systematizing the criteria

The texts available in digital format were printed and the ones contained in periodicals related to this field were photocopied, thus enabling handling of the documents in the same way and facilitating access to them. Later, the documents were arranged in chronological order and read, starting with the most recent text.

Table 1: Observation guide no. 1

Aspect indicated in the table	**Composition data**
Criterion	Identifying each criterion in accordance with the terminology defined by the author(s):
Purpose	Identifying the main goal for using and/or developing the criterion.
Description	Characterizing the procedure for using the criterion.
Detailing/Examples	Specifying traits of the criterion and providing examples of their use.
Disadvantages	Identifying any disadvantage(s) observed in the use of the criterion as determined by the author(s).
Advantages	Identifying any advantage(s) observed in the use of the criterion as determined by the author(s).
Citations provided	Identifying the documents used directly in the development of the criteria system.

Source: *produced by the authors.*

Next, an *observation guide,* which defined the guiding aspects for this specific activity, was created and used as a research tool (table 1). Observation guide No. 1 allowed the execution of two activities: (1) emphasizing, in the texts of sample No. 1, the aspects indicated in Table 1; (2) producing, for each criterion, a table to display the aspects indicated in Observation guide No. 1. Through this procedure, two results were achieved:

Result 1: a list of sixteen criteria identified from Study Sample No. 1, thus defining the *empirical object* of this research: (1) *Word phrase formation, (2) Goffman transition formula, (3) Absolute frequency of word occurrence in text, (4) Word co-occurrence relative frequency, (5) Word co-occurrence simple frequency, (6) Word occurrence relative frequency in text, (7) Word identification (comparison using dictionary), (8) Word stemming, (9) Stop-list/Stop-words, (10) Words highlighted in the text, (11) Numeric weight, (12) Term weighting, (13) Zipf's first law, (14) Zipf's second law or Zipf-Booth's law, (15) Phrasal topic, (16) Semantic vocabulary/concept headers vocabulary/thesaurus*
Result 2: systematizing the 16 criteria by filling out Observation guide No. 1 for each of them.

4.3. Analysis of Criteria Combinations – Step 2

4.3.1 Selecting study sample No. 2

The first stage of the second step was the selection of 12 (twelve) texts from study sample No. 1, thereby obtaining study sample No. 2. According to Marconi and Lakatos [18], Lakatos [19], and Mattar [20], sampling is the process through which a subset to be studied – a sample – is selected from among a whole group – a population – such that data about the group may be obtained by studying that subset. For a sample to be representative, each item from the population must have the same chances of being selected; that is, of being included in the sample. There are pre-defined sampling types, and, for this study, we chose to employ a *non-probabilistic sampling method* – subjective, with no statistical basis, defined by personal criteria derived from professional experience and knowledge of the sector being studied, usually corresponding to 10% to 15% of the target population (Marconi and Lakatos; Lakatos; Mattar) [18, 19, 20].

4.3.2 Criteria interpretation

In order to verify the practical use of the criteria identified in step 1, an analysis of the studies from study sample No. 2, using these criteria and conducted by other researchers in the field, was carried out. For comparative purposes, a summary of the studies, produced by tabulating the data obtained, is presented here. Data tabulation followed observation guide No. 2, detailed in Table 2.

Table 2: Observation guide No. 2

Aspect indicated in the table	**Composition data**
Research	Title of the study.
Goal	As stated by the author(s) of the study.
Researcher(s)	Name(s) of the author(s).
Timeframe	Timeframe, in years, during which the study was conducted.
Place	Country in which the study was conducted.
Criteria adopted	Criteria used in the research, using the same terminology and numbering indicated in 3.2.1.
Software used	Names of software used in the study.
Comparison with manual indexing	Whether or not the comparison was made: [yes] if it was and [no] if it was not.
Type of document	Nature of the document analyzed.
Field of focus	Field of knowledge the document focuses on.
Results	Whether the result was satisfactory or unsatisfactory, according to assessments by the researcher(s).
Numbering of the text in sample No. 1	Number corresponding to the text, indicated in sample No. 1

Source: *produced by the authors*

The present study did not aim at covering the subject exhaustively but, rather, only enough elements to support the choice of the best criteria to meet the goals of this work. Therefore, each of the twelve texts selected from sample No. 2 were systematized according to Observation guide No. 2. The selected text references are presented in Table 3. The number preceding each text refers to its sequential position in sample No. 1.

Table 3: Study sample no. 2

TEXT NO. 1: [1958] BAXENDALE, P. B. MACHINE-MADE INDEX FOR TECHNICAL LITERATURE: AN EXPERIMENT. IBM JOURNAL OF RESEARCH AND DEVELOPMENT, N. 2, P. 354-361, 1958. **TEXT NO. 2: 1959] MARON, M. E.; KUHNS, J. L.; RAY, L. C. PROBABILISTIC INDEXING: A STATISTICAL APPROACH TO THE LIBRARY PROBLEM. IN: NATIONAL MEETING OF THE ASSOCIATION FOR COMPUTING MACHINERY, 14., ACM, 1959, CAMBRIDGE, MASSACHUSETTS. PROCEEDINGS... NEW YORK, NY: ACM, 1959. P.1-2.**
TEXT No. 3: [1960] SWANSON, Don R. Searching natural language text by computer. **Science**, v. 132, n. 3434, p. 1099-1104, 1960. **TEXT No. 6:**[1969] EDMUNDSON, H. P. New methods in automatic extracting. **J. ACM**, v. 16, n. 2, p. 264-285, Apr. 1969.
TEXT No. 7:[1970] SALTON, Gerard. Automatic text analysis. **Science**, v. 168, n. 3929, p. 335-343, 17 Apr. 1970. **TEXT No. 9:** [1973] SALTON, Gerard. Recent studies in automatic text analysis and document retrieval. **Journal of the Association for Computing Machinery**, v. 20, n. 2, p. 258-278, April 1973.
TEXT No. 16:[1982] ROBREDO, Jaime. Otimização dos processos de indexação dos documentos e de recuperação da informação mediante o uso de instrumentos de controle terminológico.**Ci. Inf.**, Brasília, v. 11, n. 1, 1982. p. 3-18**.** **TEXT No. 24:**[1989] SALTON, Gerard; SMITH, Maria. On the application of syntactic methodologies in automatic text analysis. In: BELKIN, N. J.; RIJSBERGEN, C.,J. Van (Eds.). ANNUAL INTERNATIONAL ACMSIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 12., 1989, Cambridge, MA. **Proceedings...** New York, NY, v. 23, n. SI, Jun. 25-28, 1989. p. 137-150.
TEXT No. 54: [1998] MOENS, Marie-Francine; DUMORTIER, Jos. Automatic abstracting of magazine articles: the creation of 'highlight' abstracts. In: ANNUAL ACM CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 21., ACM SIGIR, 1998, Melbourne, Australia. **Proceedings...** New York, NY: ACM, 1998. p. 359-360. **TEXT No. 55:**[1998] ROBREDO, Jaime; CUNHA, Murilo Bastos da. Aplicação de técnicas infométricas para identificar a abrangência do léxico básico que caracteriza os processos de indexação e recuperação da informação. **Ci. Inf.**, Brasília, v. 27, n. 1, p. 11-27, jan./abr. 1998.
TEXT No. 80:[2004] HONORATO, Daniel F. et al. Utilização da indexação automática para auxílio à construção de uma base de dados para a extração de conhecimento aplicada à doenças pépticas.In: I WORKSHOP DE COMPUTAÇÃO, 1., 2004, Palhoça. **Anais...** Palhoça: WORKCOMP-SUL, 2004. p. 1-9. **TEXT No. 101:**[2007] OLIVEIRA, Elias et al. Um modelo algébrico para representação, indexação e classificação automática de documentos digitais. **Rev. Bras. Biblio. Doc.**, Nova Série, São Paulo, v. 3, n. 1, p. 73-98, jan./jun. 2007.

Source: produced by the author

Table 4 was created to present the comparison between the criteria used and the texts of Study No. 2. The data used to construct this table came from the tables built for the twelve texts selected from sample No. 2. Those data are presented in this fashion in order to provide visualization of the quantity of texts that used each criterion identified.

Table 4: Usage of index criteria in each text of Study No. 2 sample

	Res. Proj. 1	Res. Proj. 2	Res. Proj. 3	Res. Proj. 4	Res. Proj. 5	Res. Proj. 6	Res. Proj. 7	Res. Proj. 8	Res. Proj. 9	Res. Proj. 10	Res. Proj. 11	Res. Proj. 12	Projects utilise criteria	%	Name of Criterion
	1950s		1960s		1970s		1980s		1990s		2000s				
Criterion 1	X		X	X	X	X		X	X		X		8	67%	Word phrase formation
Criterion 2							X						1	8%	Goffman Transition Formula
Criterion 3				X	X			X	X			X	5	42%	Absolute frequency of word occurrence in the text
Criterion 4		X			X					X			3	25%	Relative frequency of word co-occurrence
Criterion 5					X					X			2	17%	Simple frequency of term co-occurrence
Criterion 6		X					X						2	17%	Relative frequency of word occurrence in the text
Criterion 7			X	X	X	X		X	X	X	X	X	9	75%	Word identification (Comparison using a dictionary)
Criterion 8					X	X		X	X		X		5	42%	Word stemming
Criterion 9						X				X	X	X	4	33%	Stop-list / stop-words
Criterion 10				X									1	8%	Words highlighted in the text
Criterion 11		X	X	X		X						X	5	42%	Numerical weight
Criterion 12				X	X			X	X	X			5	42%	Term weighting
Criterion 13							X	X					2	17%	Zipf's first law
Criterion 14							X	X					2	17%	Zipf's second law or Zipf-Booth's law
Criterion 15	X												1	8%	Phrasal topic
Criterion 16			X		X	X	X						4	33%	Semantic vocabulary / conceptual headings vocabulary / Thesaurus

Result 3: the most widely used criteria in the automatic indexing process.
With this information, one can identify those criteria which are the most used and which are, consequently, the most frequently combined with others. Since most studies point to satisfactory results, the factor most relevant for reaching a conclusion was the number of times the criteria appeared in relation to the total number of texts analyzed. The results obtained from the analyses conducted throughout this work are presented below. These results make it possible to propose a set of criteria that may be considered ideal for the automatic indexing process and that, according to the goals of the study, is intended to solve the problem originally proposed.

4.4. Discussion and presentation of results

Based on the analysis of data in table 4, it is possible to evaluate some aspects that are relevant to the use of the automatic indexing criteria, chosen from the literature based on sample No. 1.
Of the total of sixteen criteria selected, 50% presented a usage rate higher than 30% in the total number of studies analyzed, that is, to twelve studies. These criteria are presented in table 5.

Table 5: List of the criteria most extensively used in the studies indicated in sample No. 1

Criterion number	No. of studies that used the criterion	Percentage	Name of the criterion
Criterion 7	9	75%	Word identification (Comparison using a dictionary)
Criterion 1	8	67%	Word phrase formation
Criterion 12	5	42%	Term weighting
Criterion 11	5	42%	Numeric weight
Criterion 8	5	42%	Word stemming
Criterion 3	5	42%	Absolute frequency of word occurrence in the text
Criterion 16	4	33%	Semantic vocabulary/concept headers vocabulary/ thesaurus
Criterion 9	4	33%	Stop-list / stop-words

Source: *produced by the authors.*

Criterion No. 3, *absolute frequency of word occurrence in the text*, is believed to be relevant to the analysis of text documents. This criterion was used in five of the twelve studies analyzed, or 42%. Despite this criterion usually being perceived as limited, since it considers only the number of times each word occurs in the analyzed text, it shows a considerable usage rate in 5 of the 6 decades analyzed. The *absolute frequency of word occurrence in the text* is directly correlated to three other criteria:

- Relative frequency of word co-occurrence, which had a 25% usage rate;
- Simple frequency of word co-occurrence, which had a 17% usage rate;
- Relative frequency of word occurrence in the text, which had a 17% usage rate.

Indeed, relative frequency of occurrence and relative and simple frequencies of co-occurrence are more robust criteria than is simple frequency of occurrence, because they consider not only the number of times each word appears in the text, but also its occurrence in the database as a whole and the relationships between the words that compose the document. Therefore, measurement of the absolute frequency of occurrence of a word in a text came to be used in conjunction with other criteria that considered linguistic aspects of the text, such as: criterion No. 7, *word identification (comparison using a dictionary),* which presented a 75% usage rate; and criterion No. 16, *semantic vocabulary / concept headers vocabulary / thesaurus*, with only a 33% usage rate.

One might assume that the use of *absolute frequency of word occurrence in the text,* coupled with other criteria that take semantic aspects into consideration, may minimize the use of other, purely statistical, criteria.

When looking at criterion No. 16, *semantic vocabulary / concept headers vocabulary / thesaurus,* one realizes that, although this criterion is among the most extensively used, its application is still in its early stages given its huge potential for treating the semantic aspects of the text. Contrary to expectations, criterion No. 9, *stop-list / stop-words,* presented a usage rate of only 33% in the analyzed sample. A high usage rate was expected for this criterion, as well as for criterion No. 16, since it was among the first to be developed in the field. However, it is possible that the authors of the texts analyzed may have omitted its use exactly because its relevance is already well-known among researchers in the field.

The last four criteria for which high usage rates were verified might also present a relationship among themselves. Criterion No. 1, word phrase formation with a 67% usage rate and criterion No. 8, word stemming with 42%, are criteria linked to the structure of the formation of a word. The first one verifies the relationships found among words positioned close together, forming sentences that are rich in content representative of the text. The second one considers the stem of each word either to eliminate or to take into consideration a group of words containing the indicated stem. This verification is performed based on a previously defined list of word stems that must be discarded and/or taken into account after checked by the software. Even today, these two criteria are considered extremely relevant to the analysis of text documents since verification of grammatical structure is the basis for performing semantic analyses that will be needed at a later time.

Finally, the last two criteria, numeric weight and term weighting, which coincidentally present a 42% usage rate, can be associated. Both represent aspects of the attribution of the degree of importance to a given word in the text. The first criterion is based on the determination of special values for groups of words already defined as relevant to that specific field of work. The second criterion is aimed at defining the portions of the text that may contain words that are potentially representative of the document, such as the title, the abstract and the conclusion. Currently, it is believed that these two criteria are relevant to the analysis of text documents since they predict a reduction of text analysis as a whole to the analysis of specific portions of the text and to the consideration of words with a high degree of relevance to the subject being addressed. The other 50% of the criteria that present a usage rate below 30%, in relation to the total number of studies analyzed, are presented in table 6.

Table 6: List of least used criteria in the studies in sample No. 1

Criterion number	**No. of studies that used the criterion**	**Percentage**	**Name of criterion**
Criterion 15	1	8%	Phrasal topic
Criterion 10	1	8%	Words highlighted in the text
Criterion 2	1	8%	Goffman transition formula
Criterion 14	2	17%	Zipf's second law or Zipf-Booth's law
Criterion 13	2	17%	Zipf's first law
Criterion 6	2	17%	Relative frequency of word occurrence in the text
Criterion 5	2	17%	Simple frequency of word co-occurrence
Criterion 4	3	25%	Relative frequency of word co-occurrence

Source: *produced by the authors.*

Following analysis of table 4, the comments below can be made. Three of the criteria presented, criterion No. 2, Goffman transition formula with only an 8% usage rate, and criteria 13 and 14, Zipf's first and second laws or Zipf-Booth's law, respectively, both with 17%, can be related to each other by the fact that they are both based on statistical analysis of the text. It can therefore be seen that criteria 13 and 14 are no longer necessary, since, as indicated previously, the combination of a frequency analysis criterion with other criteria for linguistic treatment characteristics can replace the excessive use of other statistical criteria.

Another poorly represented criterion in sample No. 2 was criterion No. 10, words highlighted in the text, with a usage rate of only 8%. Although this consideration for *parser* analysis may present some satisfactory results, it is not sufficiently consistent to be recommended in the final results of this study. The last criterion that we analyzed was No. 15, phrasal topic with an 8% usage rate, meaning it was used in only one of the twelve studies in the sample. This criterion deserves a lot of attention as it is one of the precursors of the field.

Finally, from the meticulous analysis performed throughout the study on these criteria, we propose a set of nine criteria to be understood as ideal for the development of indexing software to treatment text documents. We believe these nine criteria can provide extraction of meaningful terms from indexed documents, yielding results similar to those obtained manually:

- Word phrase formation
- Absolute frequency of word occurrence in the text
- Word identification (comparison using a dictionary)
- Word stemming
- Stop-list / stop-words
- Numeric weight
- Term weighting
- Semantic vocabulary / conceptual headers vocabulary / thesaurus

5. Final considerations

Indexing is the process that makes the connection between the documents that are available in the system and the documents that may be retrieved by users, according to their needs. This activity has increased markedly, since the publication of text documents has undergone considerable growth. Today, there is a constantly expanding production and search for knowledge, which creates a situation in which it is necessary to organize data in a systemic fashion in order to make it available to the user in an appropriate manner. Deficiencies and difficulties have been found in the manual indexing process, which reaffirmed the need for studies that seek alternatives to this process.

The importance of taking semantic aspects of the text into consideration, so that the indexing is conducted in a more contextualized and consistent manner, has also been noted. It is believed that the use of controlled vocabularies, as a taxonomy built into the software developed for the automatic indexing process, can contribute to and strengthen this process when associated with the semantic aspects. However, this was not the focus of this work. Therefore, the semantic aspect of the automatic indexing process makes it possible for relevant future studies to be conducted in this field.

REFERENCES

[1] Maron, M. E. On Indexing, retrieval and the meaning of about. Journal of the American Society for Information Science, n. 28, n. 1, p. 38-43, 1977 apud Guedes, Vânia L. S. Estudo de um critério para indexação automática derivativa de textos científicos e tecnológicos. Ci. Inf., Brasília, v. 23, n. 3, p. 318-326, set./dez. 1994.

[2] Guedes, Vânia L. S. Estudo de um critério para indexação automática derivativa de textos científicos e tecnológicos. Ci. Inf., Brasília, v. 23, n. 3, p. 318-326, set./dez. 1994.

[3] Unisist. Princípios de indexação. Tradução de Maria Cristina M. F. Pinto. Rev. Esc. Biblio., Belo Horizonte, v. 1, n. 10, p. 83-94, mar. 1981. Título original: Indexing principles.

[4] Lancaster, F. W. Indexação e resumos: teoria e prática. Brasília: Briquet de Lemos, 2004. 452p.

[5] Silva, Maria R; Fujita, Mariângela S. L. A prática da indexação: análise da evolução de tendências teóricas e metodológicas. Transinformação, Campinas, v. 16, n. 2, p. 133-161, maio/ago. 2004.

[6] Lancaster, F. W. Indexação e resumos: teoria e prática. Brasília: Briquet de Lemos, 1993. 347p.

[7] Hjorland, Birger. The concept of 'subject' in Information Science. Journal of Documentation, v. 48, n. 2, p. 172-200, Jun. 1992.

[8] Fujita, Mariângela S. L. A identificação de conceitos no processo de análise de assunto para indexação. Rev. Dig. Biblio. Ci. Inf., Campinas, v. 1, n. 1, p. 60-90, jul./dez. 2003.

[9] Moreiro Gonzáles, José Antonio. El contenido de los documentos textuales: su análisis y representación mediante el lenguaje natural. Gijón: Ediciones Trea, 2004 apud Bufrem, Leilah S. A relação inescusável entre lingüística e documentação. Enc. Bibli: R. Eletr. Biblio. Ci. Inf., Florianópolis, n. 19, p. 83-94, 1º sem. 2005.

[10] Bufrem, Leilah S. A relação inescusável entre lingüística e documentação. Enc. Bibli: R. Eletr. Biblio. Ci. Inf., Florianópolis, n. 19, p. 83-94, 1º sem. 2005.

[11] Robredo, Jaime. A indexação automática de textos: o presente já entrou no futuro. In: Machado, U. D. (Org.). Estudos Avançados em Ciência da Informação, Brasília, DF: Associação dos Bibliotecários do Distrito Federal, 1982. v. 1, p. 235-274.

[12] Naves, Madalena M. L. Curso de indexação: princípios e técnicas de indexação, com vistas à recuperação da informação. Belo Horizonte: UFMG, Biblioteca Universitária, 2004. Material didático. 23p.

[13] Salton, Gerard. Automatic text analysis. Science, v. 168, n. 3929, p. 335-343, 17 Apr. 1970.

[14] Salton, Gerard. Recent studies in automatic text analysis and document retrieval. Journal of the Association for Computing Machinery, v. 20, n. 2, p. 258-27, Apr. 1973.

[15] Swanson, D. R. Searching natural language text by computer. Science, v. 132, n. 3434, p. 1099-1104, 21 Oct. 1960.

[16] Baxendale, P. B. Machine-made index for technical literature: an experiment. IBM Journal of Research and Development, n. 2, p. 354-361, 1958 apud Lancaster, F. W. Indexação e resumos: teoria e prática. Brasília: Briquet de Lemos, 2004. 452p.

[17] O'connor, J. Automatic subject recognition in scientific papers: an empirical study. Journal of the Association for Computing Machinery, n. 12, p. 490-515, 1965 apud Lancaster, F. W. Indexação e resumos: teoria e prática. Brasília: Briquet de Lemos, 2004. 452p.

[18] Marconi, M. D. A.; Lakatos, E. M. Técnicas de pesquisa: planejamento e execução de pesquisas, amostragens e técnicas de pesquisas, elaboração, análise e interpretação de dados. 3.ed. São Paulo: Atlas, 1996.

[19] Lakatos, Eva Maria. Fundamentos de Metodologia Científica. 3. ed. rev. e aum. São Paulo: Atlas, 1991.

[20] Mattar, F. N. Pesquisa de Marketing. São Paulo: Altas, 1996.