

**Information retrieval in institutional repositories using the summarization technique derived from the selection of Cassiopeia attributes****Recuperação de informação em repositórios institucionais utilizando a técnica de sumarização a partir da seleção de atributos do Cassiopeia**

DOI:10.34117/bjdv6n11-286

Recebimento dos originais:08/10/2020

Aceitação para publicação:15/11/2020

**Luanna Azevedo Cruz**

Mestra em Educação

Instituto Federal de Educação, Ciência e Tecnologia da Bahia – IFBA

Endereço:Estrada Vicinal para Tenda, s/n, Barro Vermelho, Seabra/BA, Brasil

E-mail:luannaacruz@gmail.com

**Marcus Vinicius Carvalho Guelpeli**

Doutor em Computação

Universidade Federal dos Vales do Jequitinhonha e Mucuri – UFVJM

Endereço:Rodovia BR 367, 5000, Diamantina/MG, Brasil

E-mail:marcus.guelpeli@ufvjm.edu.br

**ABSTRACT**

The large volume of available text documents arising from the increase in scientific output creates a need for researching and implementing methods that facilitate information search and retrieval in academic text bases, such as institutional repositories. This study's objective is thus to analyze whether the application of the summarization technique, based on the method of selecting attributes (words) of the Cassiopeia model (implemented in the PragmaSUM summarizer), in academic texts, is helpful for retrieving information by reducing information overload and improving the accuracy of user search results. The research was developed in steps: elaboration of the reference collection; implementation of a search engine; execution of standard information retrieval; evaluation of information retrieval using the precision metric; and data analysis from Friedman ANOVA and Kendall's Coefficient of Concordance statistical tests. Results revealed that summarization, mainly performed with high compression rates (80% and 90%), reduced information overload and increased the accuracy of the results presented to the user, allowing quality information retrieval in academic texts. Furthermore, it simplified the indexing process, attenuated high dimensionality and promoted faster information retrieval.

**Keywords:** Information retrieval, Institutional repository, Text Mining, Summarization, Cassiopeia model.

**RESUMO**

Com a quantidade elevada de documentos textuais disponíveis, a partir do aumento das produções científicas, surge a necessidade do estudo e implementação de métodos que facilitem a busca e

recuperação de informação em bases de textos acadêmicos, como os repositórios institucionais. O objetivo desta pesquisa é analisar se a aplicação da técnica de sumarização, a partir do método de seleção de atributos (palavras) do modelo Cassiopeia (implementado no sumarizador PragmaSUM), em textos acadêmicos, auxilia na recuperação de informação, diminuindo a sobrecarga de informação e melhorando a precisão dos resultados retornados ao usuário. A pesquisa foi desenvolvida em etapas: elaboração da coleção de referência; implementação de um buscador; execução da recuperação de informação; avaliação da recuperação de informação por intermédio da métrica *precision*; e análise dos dados a partir dos testes estatísticos ANOVA de Friedman e Coeficiente de Concordância de Kendall. Os resultados obtidos mostraram que a sumarização, efetuada principalmente com altas taxas de compressão (80% e 90%), diminuiu a sobrecarga de informação e aumentou a precisão dos resultados apresentados ao usuário, permitindo qualidade na recuperação de informação em textos acadêmicos. Além disso, simplificou o processo de indexação, atenuou a alta dimensionalidade e promoveu maior agilidade na recuperação de informação.

**Palavras-chave:** Recuperação de informação, Repositório institucional, Mineração de Textos, Sumarização, Modelo Cassiopeia.

## 1 INTRODUCTION

As technology advances and digital information and communication technology innovations (DICTI) develop and become more widespread, users have started to create digital documents composing a large repository of culture and knowledge. Thus, the continuous development of DICTI has been generating new possibilities and challenges in different fields of knowledge. Regarding the field of education in particular, following the increase in scientific output, universities and research institutions have been underscoring the structure of mechanisms equipped with technology, such as institutional repositories. The idea is to increase visibility, accessibility, preservation and dissemination of the research being developed. For the development and implementation of institutional repositories, software packages may provide tools that present features for organizing, storing and retrieving information<sup>1,2</sup>.

According to Leite et al.<sup>3</sup>, repositories offer a scientific information service in a digital environment, and present, as information base, complete and digital texts that are stored and presented for user access. Thus, texts such as articles, dissertations and theses are used as sources of information and can improve the teaching-learning process, while also encouraging innovation and contributing to universal access to knowledge. However, the number of documents available has been observed to be growing significantly. This expansion of information has generated problems such as information overload, lack of organization and unstructured data. Thus, dealing with high volumes of available textual information and locating them quickly and accurately started becoming a challenge for users<sup>4</sup>.

In this perspective, in order to facilitate information access, according to Baeza-Yates and Ribeiro-Neto<sup>5</sup>, information organization methods for subsequent search and retrieval are being used. This is the context in which Information Retrieval (IR) is found – a Computer Science field aiming to detect documents that meet the user's needs within a collection of documents (also called corpus), thereby enabling information access, retrieval and analysis.

Currently, users perform IR in a variety of contexts: (1) online, from user queries (search phrases) – systems such as Google ([www.google.com](http://www.google.com)) and Yahoo ([br.yahoo.com](http://br.yahoo.com)) provide search possibilities on billions of indexed documents, stored in millions of computers, and scroll through web page hyperlinks; (2) retrieval of personal information using search tools available on operating systems such as Windows 10 ([www.microsoft.com](http://www.microsoft.com)) search fields, or in email programs that provide search fields; and (3) in the context of corporate and institutional research, based on the retrieval of document archives stored in specific databases, such as institutional repositories, digital libraries and document collections. In these repositories, searching for the document by its subject is the main form of retrieval performed by users and, from the query entered, the IR system can perform full-text or metadata (are data about data, or information about information) searches. In repositories in general and documentation centers, the IR process can be developed because the whole universe is understood and accessible, as they follow the same standards<sup>2,6,7</sup>.

The processes of interaction between user and documents are performed by information retrieval systems (IRS); i.e., search tools. Thus, the main objective of an IRS is to retrieve documents that are suitable for the user. For this reason, the IRS must create a representation of the corpus texts and present them to the user in a way that permits a quick selection of the items that completely or partially satisfy the user's information needs, formalized by a query. Thus, the IRS must retrieve all the relevant documents and the least possible amount of irrelevant data. However, relevance is debatable due to its subjective nature, for a document may be considered relevant to a certain and irrelevant to another<sup>8,9</sup>.

The quality of the results shown to the user can be verified through an IR evaluation, which is the process in which a quantitative metric is associated with the results produced by an IR system, in response to a set of user queries. This way, the use of metrics and the comparison of results presented by the system, with results suggested by a manmade reference collection, are the most commonly applied evaluation procedures. The *precision* metric is one of the most widely used measurements and represents the fraction of retrieved documents that is relevant. Reference collections have the following

elements: documents; queries; and the human-made judgment of relevance, in which each document is classified as relevant or irrelevant in relation to a query<sup>5,6</sup>.

Quality assessment is important so that IR-related problems can be identified and possible solutions can be applied. According to Baeza-Yates and Ribeiro-Neto<sup>5</sup>, the results presented to the user are often inaccurate and the number of texts returned is high, which generates information overload and, consequently, thwarts efforts to assimilate information. These aspects are confirmed by Dias and Carvalho<sup>10</sup>, who highlight overload as one of the main concerns in presenting the results obtained through IR systems. Grainger and Potter<sup>11</sup> complement this by stating that when users perform a search for documents, only 10% of them are willing to go beyond the first page of results and only 1% even navigate to the third page. Another important issue is whether retrieved documents have redundant and unnecessary terms for RI. This is because, in IR systems, long documents are more likely to match the query simply due to their size, which does not mean that they are relevant to the search expression<sup>7,12</sup>.

It is in this context that Text Mining (TM) operates, as it assists IR processes, data extraction, text summaries, discovery of patterns, associations and rules, and text document analysis. Additionally, an adequate organization of text collections streamlines the processes of searching and retrieving information. Thus, in order to extract knowledge and treat textual information, TM techniques, such as summarization, can be applied to collections<sup>13</sup>.

Summarization involves reducing text content and producing abstracts without losing information and meaning; i.e., summaries are formed by words that significantly represent the text as a whole. The use of automatic summarizers is common, in which summarization algorithms can provide compression ratios in different percentages. This rate defines the percentage or size of the summary in relation to the original text. For example, a text summarized at a rate of 60% will yield a summary whose size is 40% of the original text. According to Guelpe<sup>14</sup>, automatic summarization reduces part of the stopwords and minor attributes in information repositories. This contributes to the decrease in high dimensionality and sparse data, which are common problem in IR caused by the use of a large number of words<sup>13,15</sup>. Considering this, Beyer et al.<sup>16</sup> state that low dimensionality is necessary to maintain the word's distinguishing ability and, consequently, allows the reduction of processing time. The most common technique for reducing high dimensionality and sparse data in the literature is the Luhn algorithm<sup>17</sup>. Luhn proposed defining an upper and lower threshold, so that words that fall outside the range are eliminated from the analysis. Therefore, the first cut aims to eliminate stopwords, and the second serves to decrease the number of very specific words found only once in documents<sup>8,18</sup>.

In this sense, the present research consisted in analyzing if the application of the summarization technique, based on the method of selecting attributes (words) of the Cassiopeia model in a corpus of academic texts, helps IR by reducing information overload and improving the accuracy of results shown to the user. The Cassiopeia model is a text group that proposes a medium cut in the frequency distribution of words. Thus, the selection of attributes of a text allows to reduce the dimensionality, maintaining the attributes that have greater capacity to represent the collection of documents<sup>14</sup>.

The selection of attributes proposed in the model is implemented in the summarizer used in this research, PragmaSUM, with the purpose of producing more informative summaries<sup>19</sup>. Thus, from predefined queries and from the batch-summarized academic text repository, the implemented search engine executed IR processes – searching, indexing, retrieving and ranking – the results showing documents according to the query. That having been done, the quality of the IR was assessed using the *precision* metric. The data obtained were statistically analyzed in order to verify the research proposal.

## **2 MATERIAL AND METHODS**

### **2.1 ELABORATION OF THE REFERENCE COLLECTION**

This phase refers to the elaboration of the IR test that involved preparation of the collections used in the research, definition of queries and relevance assessments.

According to Silva R. and Silva E.<sup>20</sup>, there is no consensus about the minimum size for the corpus to be defined as representative. Therefore, for this study, part of the corpus built by Aguiar et al.<sup>21</sup> were used, which is composed of 500 scientific articles, in Portuguese, distributed by ten knowledge areas of the educational domain. These areas belong to the larger field of Education and can be viewed in the table created by the government agency Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, Coordination for the Improvement of Higher Education Personnel): Special education; Permanent Education; Pre-school Education; Teaching and Learning; Philosophy of Education; History of Education; Education Policies; Educational Psychology; Sociology of Education and Educational Technology<sup>22</sup>.

Thus, each educational domain of the corpus includes 50 texts and five directories, organized by Aguiar et al.<sup>21</sup> as follows: Notes, includes files (.txt) with text statistics and external references; Original Articles, where original articles are stored in (.pdf) format; Texts, composed by the bodies of the texts in (.txt) format, i.e., the source text of the articles; Abstracts, contains the manual summaries of the articles in (.txt) format; and Keywords, contains the keywords defined by the authors of the articles, in (.txt) format.

The corpus size was reduced due to the relevance assessment of the documents. Thus, 300 articles were randomly selected, which composed a main collection of 300 academic texts to be used for the IR tests. From this central collection, the following collections have been defined: Collection of original texts, consisting of 300 original and complete articles, textual basis for “Standard IR Execution”, i.e., following the traditional IR model; Collection of source texts – textual basis for “IR execution with the Cassiopeia model” – made up of 300 source texts that underwent the capture and manipulation phase. This step involved removing bibliography, tables, footnotes, figures, page numbers and graphics, leaving only the textual body of the articles, that is, the source text in (.txt) format.

For searching the IR system interface, ten queries were defined from the domains to which the corpus documents belong. The search expressions selected were: “special education”, “permanent education”, “pre-school education”, “sociology of education”, “educational psychology”, “teaching and learning”, “philosophy of education”, “history of education”, “education policies” and “educational technology”.

Once that was completed, the relevance assessment of 300 texts began, by domain; which meant that each one of the 30 documents, belonging to a specific domain, was classified as relevant or irrelevant regarding to corresponding query. This assessment was conducted by text skimming the domain documents. Thus, for each one of the ten queries, fifteen relevant documents were selected. According to Lakatos and Marcone<sup>23</sup>, reading by skimming aims to capture the general inclination of the text by accounting for the title, subtitle and illustrations, if any; and by reading the abstracts and paragraphs while trying to find the methodology and essence of the work. As an example, the 30 documents of the Special Education domain were defined as relevant or irrelevant in relation to the query “special education”.

## 2.2 IMPLEMENTATION OF SEARCH ENGINE

At this stage, the search engine was implemented to provide an IR system to perform IR in the collection of academic texts within the educational domain. The search engine is responsible for the processes of searching, indexing, retrieving and ranking the documents that will be shown to the user. Thus, in this research, we chose to use *Apache Solr* version 7.2.1. *Solr* is a full-text search server for textual database IR that uses both the boolean and vector space models<sup>24,25</sup>.

In order to improve the appearance and facilitate IR execution and assessment, an interface linked to *Solr* was developed with the following components: Text box for consultation; Search button, which initiates the query entered in the text box; Selection of core field of the textual collection – Core

Original for the “**Collection of original texts**” and standard IR execution, and *Core\_sum* for “**Collection of Source Texts**”; i.e., IR execution with the Cassiopeia model; Area for the results, in which the academic texts retrieved from the search are listed; *Precision* metric automatic calculation area, where for each query the metric values are presented, which are also recorded in log (.xls format), together with the queries made and the response time (in seconds); and Pagination, which is displayed if the number of items retrieved exceeds the result limit determined per page.

Thus, the technological solutions used to implement the proposal were the C# programming language and free software *Microsoft Visual Studio Community 2017* (visualstudio.microsoft.com), version 15.7.4.

### 2.3 EXECUTION OF INFORMATION RETRIEVAL – IR

In the “**IR Execution**” step the IR process occurred in two ways: “Standard IR Execution” and “IR Execution with the Cassiopeia Model”. The objective was to analyze the research proposal through comparison with traditional recovery processes. The main difference between the executions is related to the collections obtained in the “Elaboration of Reference Collection” phase, which should be chosen from the core selection field. The same search engine was used in both runs, so the search, indexing and ranking processes work the same way.

#### **Execution of Standard IR**

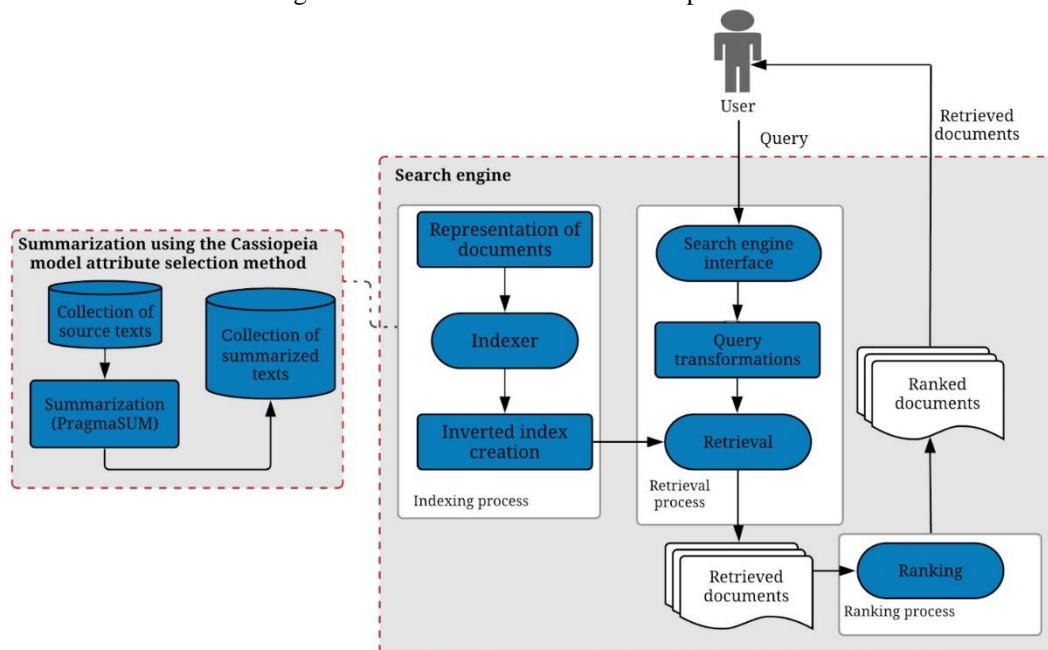
The execution of standard IR is performed by general IR systems, in which there are preparation methods, classifications or data mining of the original corpus texts. Therefore, such execution followed traditional IR procedures, according to Baeza-Yates and Ribeiro-Neto<sup>5</sup>, taking the following steps:

- (1) IR execution is initiated when a user enters the query terms into the search field of the search engine for documents that meet their information interests. The core, which is searched and in which documents are retrieved, must be selected. In this case, the “original core” is chosen. All ten queries were entered in the search engine’s interface to proceed with the tests
- (2) *Solr* transforms the query by removing stopwords and making spelling corrections, and checks the text collection for creating the representation of documents for the indexer.
- (3) The inverted index is created and, after the search processing, retrieval begins by returning the retrieved set of documents.
- (4) The academic texts returned are ranked in order of relevance, using the Boolean and vector space models proposed in *Solr*.
- (5) Finally, the academic texts are displayed to the user in the results area.

## IR Execution with the Cassiopeia Model

This execution, here called “IR with Cassiopeia model”, refers to the IR execution in which the corpus source texts went through the summarization process, using the Cassiopeia’s attribute selection method, as shown in Figure 1. The summarizer which implements this method is PragmaSUM<sup>19</sup>. Thus, the collection of summarized source texts is used by the IR search engine.

Figure 1. IR Execution with the Cassiopeia model



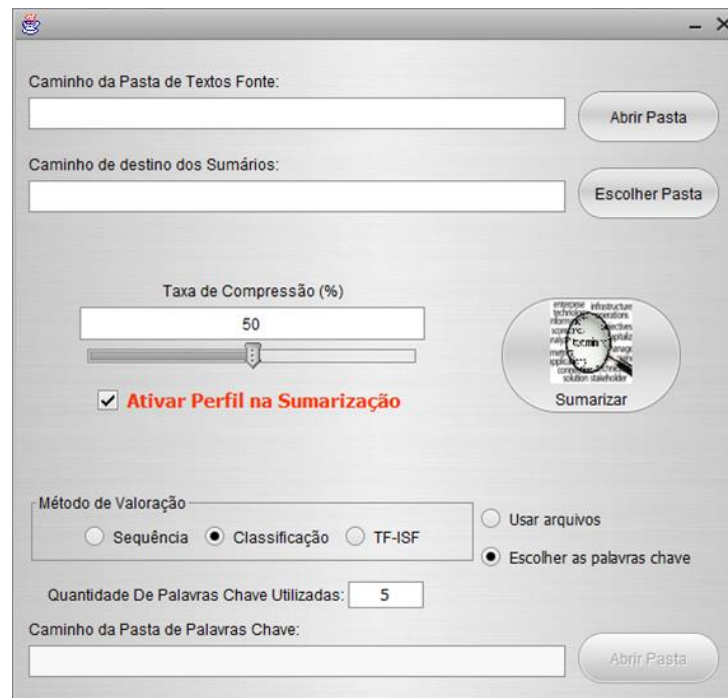
To implement the research proposal (Figure 1), the texts contained in the “Collection of source texts” were summarized by PragmaSUM, giving rise to a set of summarized texts. PragmaSUM summarization comprises preprocessing and processing.

Thus, given a set of texts in .txt format, in the preprocessing phase PragmaSUM clears the texts in order to reduce the amount of words, preparing them for computational processing. In order to obtain information by allowing qualitative and quantitative gain in processing, PragmaSUM uses the technique of reducing high dimensionality and sparse data proposed in the Cassiopeia model, i.e., the attribute selection method. This selection is made as follows: the summarizer averages the frequency of the words contained in the source text and selects the word that has the closest frequency to the average; the 24 words above and 25 below the average are chosen from a cut with the 50 words that will be used in the valuation of the text items. Only the words contained at the peak of the relevant words will be used<sup>14,19</sup>.



The processing step involves the execution of the summary, which in PragmaSUM can occur individually or in batches. In the first, only one source text is summarized at a time, with the choice of using or not words to personify the summary, according to user preferences. The second consists of a batch summarization (Figure 2), in which large amounts of text can be summarized at one go, as described below<sup>19</sup>.

Figure 2. PragmaSUM: batch summarization



Source: Rocha & Guelpele, 2017.

The source-text and destination folders are selected in the fields “Path of Source Text Folder” and “Path of Summary Destination Folder”, respectively. The compression ratio is selected, from 0 to 100%. The option “Activate Profile in Summary” is checked if the personification of the summary is of interest to the user. With profile activation, the fields “Valuation Method”, “Number of Key Words Used” and “Path of Keyword Folder” are displayed. “Valuation Method”, which permits one of the following methods: Sequence – words are chosen in the order in which they appear in the source text; Classification – words are ordered according to their frequency in the source text (more frequent words have higher punctuation); TF-ISF – use of the Term Frequency-Inverse Sentence Frequency (TF-ISF) algorithm to classify sentences (sentences) of the text. According to Rocha and Guelpele<sup>19</sup>, the TF-ISF value determines that each sentence has an associated score, given by the value of all its words. Thus, it is the criterion for selecting the sentences that will be part of the summary. “Number of Key Words

Used”, which defines the number of key words used in the summarization. “Path of Keyword Folder”, where the folder that contains the keyword files are selected.

Thus 16 batch summarizations were concluded with the 300 source texts from the “Collection of source texts”, combining compression ratios (50%, 70%, 80% and 90%) and number of keywords (three, four, five or none). The sentence valuation method used was the TF-ISF algorithm, as it presents the best results for the summaries performed in the same corpus<sup>19</sup>. Therefore, each of the 16 summarizations gave rise to 16 collections of summarized texts.

The IR execution with the Cassiopeia model followed the standard IR steps; however, “Core\_sum” was selected in the interface core field. Therefore, for IR performed with the Cassiopeia model, these 16 summarized collections were used; i.e., ten queries were inserted in the IRS and executed in each collection.

#### 2.4 EVALUATION OF INFORMATION RETRIEVAL – IR

The standard IR and Cassiopeia model evaluations were performed by calculating the *precision* metric for each of the ten queries inserted in the IR system, in each collection (one original and 16 summarized). *Precision* values range from 0 (zero) to 1 (one): the closer to 1, the better; the closer to 0, the worse.

#### 2.5 STATISTICAL ANALYSIS OF THE DATA

The definition of the appropriate statistical test was based on the diagram proposed by Callegari and Jacques<sup>26</sup>. Considering that the data obtained are independent, ordinal samples and present abnormal distribution, it was possible to identify that the nonparametric statistical tests, such as Friedman’s ANOVA and Kendall’s coefficient of concordance are the most appropriate, since they permit analyzing whether the samples of an experiment present significant differences in their distribution<sup>14</sup>. These tests were performed using software called BioStat ([www.mamiraua.org.br](http://www.mamiraua.org.br)).

The ANOVA statistical test by Friedman compares and sorts the results of three or more related samples and calculates the average order for each. According to Campos<sup>27</sup>, the test does not directly use numerical data, but rather the positions occupied by them, after sorting the data by ascending value. Kendall’s coefficient of concordance test aims to normalize Friedman’s ANOVA and, according to Viali<sup>28</sup>, may be useful in trial-related reliability studies. The test verifies the degree of association between variables and generates an assessment of whether or not they agree with the ranks of the experiments. The closer to zero, the lower the agreement; the closer to one, the higher the agreement<sup>14,26</sup>.

Thus, the data obtained from the execution and IR evaluation, which is the *precision* metric values, were tabulated and statistically analyzed. Statistical analysis was performed between: (1) collections, by compression ratio applied in summarization (degrees of freedom = 4); and, (2) all 17 collections (degrees of freedom = 16). First, Friedman's ANOVA statistical test, with a significance level of 0,05 ( $\alpha = 0,05$ ), was applied to check if there is 95% certainty that there is a significant difference between the experiments (collections). Once this was done, the Kendall's tau coefficient (W) was verified: from the rank of the experiments, the coefficient of concordance (W) is calculated to evaluate the agreement between the produced ranks.

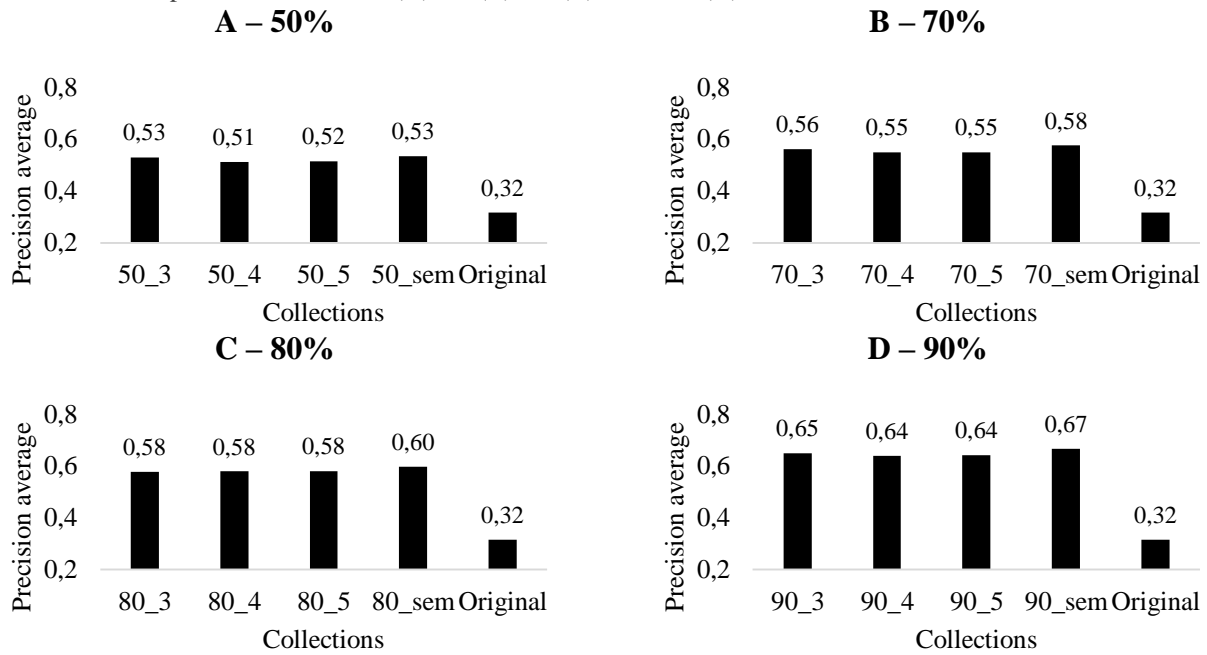
### **3 RESULTS AND DISCUSSION**

The results are displayed from the comparison between collections, by compression rate, followed by the comparison between all 17 collections. Due to the large volume of data generated, all achieved results are available at <https://bit.ly/2Tj2oEk>.

#### **3.1 STANDARD IR AND IR WITH THE CASSIOPEIA MODEL: COMPARISON BETWEEN COLLECTIONS, BY COMPRESSION RATIO**

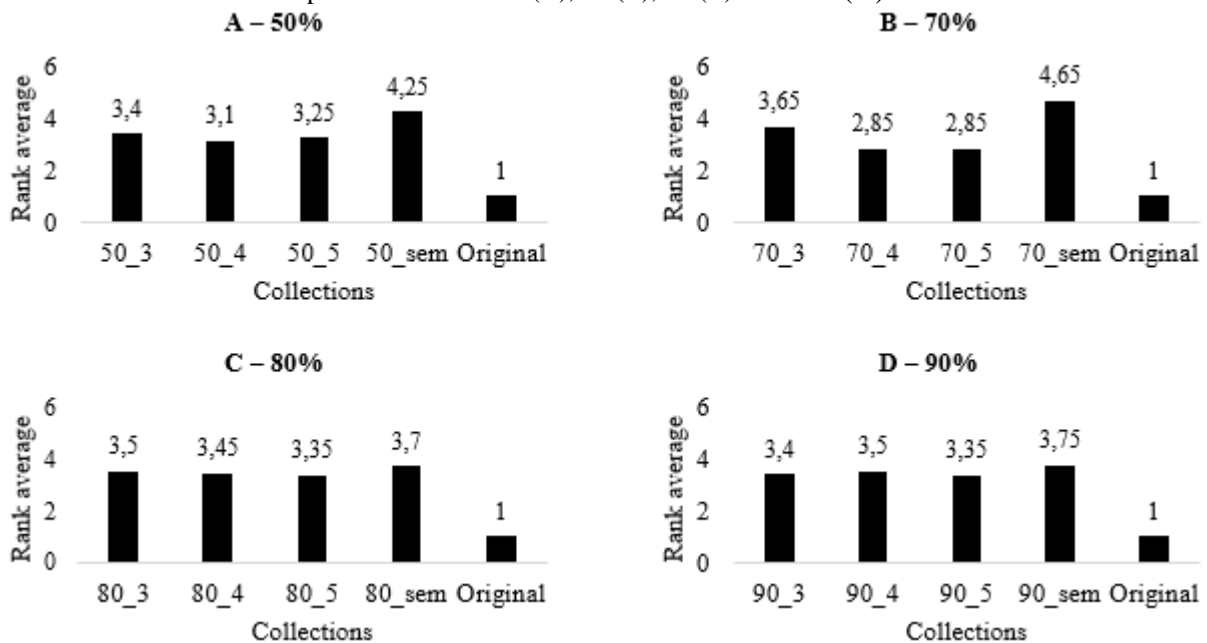
For comparing between "Original" and summarized collections (by compression ratio – 50, 70, 80 and 90%), Figure 3 shows the average *precision* values (obtained from ten queries performed in the collection). Kendall's coefficient of concordance (W) revealed that the collection samples display concordance levels at 69%, 85%, 59% and 58%, respectively.

Figure 3 – Average *Precision* values (by compression ratio) for all queries conducted in the “Original” and summarized collections with compression ratios of 50 (A), 70 (B), 80 (C) and 90% (D)



Collections “50\_sem”, “70\_sem”, “80\_sem” and “90\_sem”, when compared to the “Original” collection and collections summarized with the same compression ratio produced the numerical value of the highest *precision* value. This can be confirmed in Figure 4, which illustrates the average precision ranks.

Figure 4 – Average *precision* ranks (by compression ratio) for all queries conducted in the “Original” and summarized collections with compression ratios of 50 (A), 70 (B), 80 (C) and 90% (D)



Considering all the ratios, non-keyword summarized collections displayed higher numerical average rank value. According to Rocha and Guelpeli<sup>19</sup>, the summarizer presents good results with the use of keywords, so that the summaries are composed of important sentences. However, regarding the use of three, four, five, or zero keywords in the summarization, IR results with the Cassiopeia model revealed that the amount of keywords used in summarization did not influence IR. This can be concluded by observing that, throughout all ratios, the largest numerical difference within ranks occurred between the summary collections with no keywords and the “Original” collection (Figure 4). Keywords from academic texts are used for text representation and as search engine indexers; however, often authors do not make the best choices of these words. This issue is believed to have influenced the higher performance of summary collections without the use of keywords, since collections summaries were created using these keywords.

Table 1 shows the significance analysis (Friedman’s ANOVA) between the ranks of the “Original” collection and summarized collections with compression ratios of 50, 70, 80 and 90%. When analyzing the summarized collections with the same compression ratio, it was found that there is no statistical difference between them (Significance = ns).

Table 1. Significance analysis (Friedman’s ANOVA –  $\alpha = 0,05$ ), by compression ratio, for the “Original” collection and summarized collections ranks, with rates of 50, 70, 80 and 90%

Compression ratio	Comparisons	Difference	Significance (p)
50%	r1 (50_3) e 2 (50_4)	3	ns
	r1 (50_3) e 3 (50_5)	1,5	ns
	r1 (50_3) e 4 (50_sem)	8,5	ns
	r1 (50_3) e 5 (Original)	24	< 0,05
	r2 (50_4) e 3 (50_5)	1,5	ns
	r2 (50_4) e 4 (50_sem)	11,5	ns
	r2 (50_4) e 5 (Original)	21	< 0,05
	r3 (50_5) e 4 (50_sem)	10	ns
	r3 (50_5) e 5 (Original)	22,5	< 0,05
	r4 (50_sem) e 5 (Original)	32,5	< 0,05
70%	r1 (70_3) e 2 (70_4)	8	ns
	r1 (70_3) e 3 (70_5)	8	ns
	r1 (70_3) e 4 (70_sem)	10	ns
	r1 (70_3) e 5 (Original)	26,5	< 0,05
	r2 (70_4) e 3 (70_5)	0	ns
	r2 (70_4) e 4 (70_sem)	18	ns
	r2 (70_4) e 5 (Original)	18,5	ns
	r3 (70_5) e 4 (70_sem)	18	ns

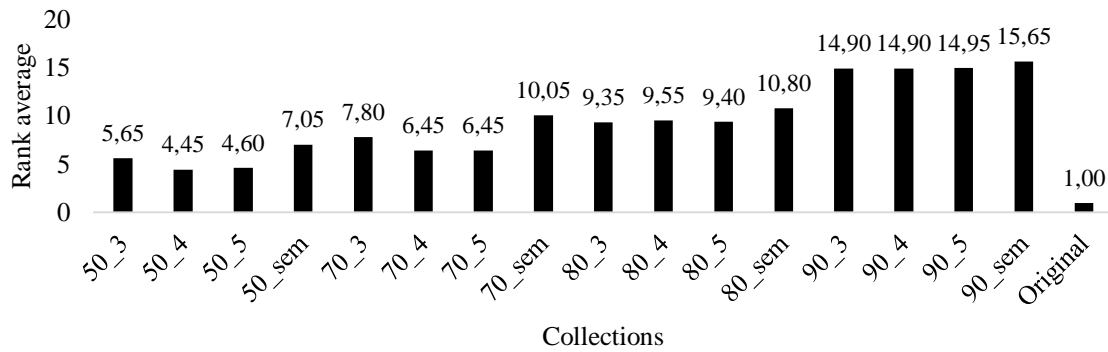
	r3 (70_5) e 5 (Original)	18,5	ns
	r4 (70_sem) e 5 (Original)	36,5	< 0,05
80%	r1 (80_3) e 2 (80_4)	0,5	ns
	r1 (80_3) e 3 (80_5)	1,5	ns
	r1 (80_3) e 4 (80_sem)	2	ns
	r1 (80_3) e 5 (Original)	25	< 0,05
	r2 (80_4) e 3 (80_5)	1	ns
	r2 (80_4) e 4 (80_sem)	2,5	ns
	r2 (80_4) e 5 (Original)	24,5	< 0,05
	r3 (80_5) e 4 (80_sem)	3,5	ns
	r3 (80_5) e 5 (Original)	23,5	< 0,05
	r4 (80_sem) e 5 (Original)	27	< 0,05
90%	r1 (90_3) e 2 (90_4)	1	ns
	r1 (90_3) e 3 (90_5)	0,5	ns
	r1 (90_3) e 4 (90_sem)	3,5	ns
	r1 (90_3) e 5 (Original)	24	< 0,05
	r2 (90_4) e 3 (90_5)	1,5	ns
	r2 (90_4) e 4 (90_sem)	2,5	ns
	r2 (90_4) e 5 (Original)	25	< 0,05
	r3 (90_5) e 4 (90_sem)	4	ns
	r3 (90_5) e 5 (Original)	23,5	< 0,05
	r4 (90_sem) e 5 (Original)	27,5	< 0,05

The comparison between standard IR (“Original” collection) and the Cassiopeia model IR (summarized collections) revealed that, by means of significance analysis, only the collections “70\_4” and “70\_5” are not statistically different from the “Original” (Significance = ns). Therefore, within the 19 summarized collections, 14 are superior to the “Original” (Significance < 0,05, highlighted in red); i.e., there is a significant difference in the precision of the results returned to the user (Table 1). Hence, for these IR collections obtained through the Cassiopeia model, the user’s effort in analyzing the retrieved documents will be smaller, given that fewer irrelevant items were retrieved.

### 3.2 STANDARD IR AND IR WITH THE CASSIOPEIA MODEL: COMPARISON BETWEEN THE 17 COLLECTIONS

The comparison among all collections (one “Original” and 16 summarized collections – 16 degrees of freedom), Figure 5 shows the average ranks (Friedman’s ANOVA) of *precision* (obtained through ten queries conducted in the collections) for all 17 collections. Kendall’s coefficient of concordance (W) revealed a concordance rate of 73% among the collections.

Figure 5 – Average of precision ranks for all queries conducted in the collection



As compression ratio increased, the average numerical values of the *precision* ranks also increased for summarized collections. The “Original” collection revealed the smallest numerical average within ranks.

The statistical significance of comparing the ranks of all collections (Table 2) demonstrated that summarized collections with a compression ratio of 90% (r13 to r16 – collections “90\_3”, “90\_4”, “90\_5” and “90\_sem”) were superior to the summarized collections with 50% and to two with com 70% (“70\_4” and “70\_5”) (row “Summarized Collections”). However, there was no statistical difference between summarized collections displaying 80% and 90% compression. Additionally, all summarized collections with rates of 80 and 90% and the collection “70\_sem” were superior to the “Original” collection (row “Original and Summarized Collections”).

Table 2. Significance analysis (Friedman’s ANOVA –  $\alpha = 0,05$ ) among collection ranks

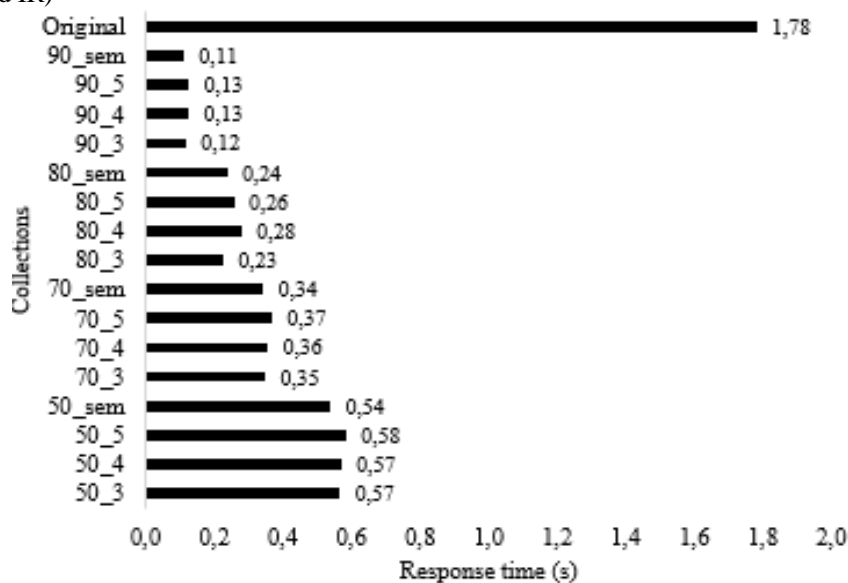
Summarized archives			Summarized archives and “Original”		
Comparisons	Difference	(p)	Comparisons	Difference	(p)
r1 (50_3) and r13 (90_3)	92,5	< 0,05	r8 (70_sem) and Original	90,5	< 0,05
r1 (50_3) and r14 (90_4)	92,5	< 0,05	r9 (80_3) and Original	83,5	< 0,05
r1 (50_3) and r15 (90_5)	93	< 0,05	r10 (80_4) and Original	85,5	< 0,05
r1 (50_3) and r16 (90_sem)	100	< 0,05	r11 (80_5) and Original	84	< 0,05
r2 (50_4) and r13 (90_3)	104,5	< 0,05	r12 (80_sem) and Original	98	< 0,05
r2 (50_4) and r14 (90_4)	104,5	< 0,05	r13 (90_3) and Original	139	< 0,05
r2 (50_4) and r15 (90_5)	105	< 0,05	r14 (90_4) and Original	139	< 0,05
r2 (50_4) and r16 (90_sem)	112	< 0,05	r15 (90_5) and Original	139,5	< 0,05
r3 (50_5) and r13 (90_3)	103	< 0,05	r16 (90_sem) and Original	146,5	< 0,05
r3 (50_5) and r14 (90_4)	103	< 0,05			
r3 (50_5) and r15 (90_5)	103,5	< 0,05			
r3 (50_5) and r16 (90_sem)	110,5	< 0,05			
r4 (50_sem) and r13 (90_3)	78,5	< 0,05			

r4 (50_sem) and r14 (90_4)	78,5	< 0,05
r4 (50_sem) and r15 (90_5)	79	< 0,05
r4 (50_sem) and r16 (90_sem)	86	< 0,05
r6 (70_4) and r13 (90_3)	84,5	< 0,05
r6 (70_4) and r14 (90_4)	84,5	< 0,05
r6 (70_4) and r15 (90_5)	85	< 0,05
r6 (70_4) and r16 (90_sem)	92	< 0,05
r7 (70_5) and r13 (90_3)	84,5	< 0,05
r7 (70_5) and r14 (90_4)	84,5	< 0,05
r7 (70_5) and r15 (90_5)	85	< 0,05
r7 (70_5) and r16 (90_sem)	92	< 0,05

Thus, it was noted that a larger amount of relevant documents was returned to the user when applying high compression rates, which produce texts with reduced content; in other words: there was greater accuracy and less information overload, thus assisting IR of academic texts. This is noticeable since, according to Baeza-Yates and Ribeiro-Neto<sup>5</sup>, long documents are more likely to match the query simply due to their size, but this does not mean that they are relevant to the search expression.

To check how quickly documents were retrieved, the time interval between receiving the user’s query and submitting the response – i.e. response time – was automatically calculated by the IR system (for each of the queries made in each of IR collections with the Cassiopeia model and standard IR) and recorded in a log worksheet, along with the metrics results. Thus, the response time averages of the summary collections and the “Original” collection are shown in Figure 6.

Figure 6 - Average Response Time: Comparison between Summarized Collections (IR with Cassiopeia Model) and Original Collection (Standard IR)





Among the sets in which IR with the Cassiopeia model was performed, the collections summarized with a compression ratio of 90%, which presented the highest accuracy of results returned to the user, also displayed the shortest response time. An analysis of the agility of IR observed that while the compression ratio increased, the response time decreased. The “Original” collection (standard IR), which generated the least accurate results for the user, displayed the longest response time (Figure 6). This was feasible because an IR system represents the document by checking the full text, i.e. all the words are used as terms for the index. Moreover, due to the text summarizations, there was a reduction in the number of terms for the index and inverted index, which simplified the indexing process, allowing faster access to documents and query processing.

From this point, it was possible to verify that the obtained results are related to some characteristics of the techniques and tools used in the research. First, according to Rocha and Guelpeli<sup>19</sup>, the PragmaSUM summarizer displays effective performance with high compression rates, attributed to the joint use of: the attribute selection method from the Cassiopeia model, which generates summaries with considerable information; and the TF-ISF algorithm that selects the sentences which form the summaries, verifying the importance of a word in a sentence, which weighs the term frequency, thereby improving results. According to Wives<sup>29</sup> and Nogueira<sup>18</sup>, attribute selection has a decisive factor for the good quality of the best IR performance and infrequent words are very discriminating, but take up unnecessary space in the index and do not retrieve many documents. Another issue is the fact that the source texts of the main corpus display little variation in size, which is in line with the principle of document size normalization, an important factor for ranking. Size is significant in IR because it improves the quality of document retrieval. Additionally, it is believed that this quality was also provided by IR system *Apache Solr*, by means of three main advantages of the vector model: term weighting scheme, term-document pair strategy that brings the document closer to the query conditions and document ordering according to the degree of similarity in relation to the query.

#### 4 CONCLUSION

In the comparison between the standard IR and the IR with the Cassiopeia model, by compression ratio, among the 16 summarized collections, 14 were statistically superior to the “Original” collection. However, there is no statistical difference between collections summarized with the same compression ratio.

In the analysis performed on the set of 17 collections, the study found that all summarized collections with rates of 80 and 90%, as well as the collection "70\_sem" were superior to the "Original" collection. Furthermore, as compression ratio increased, the average *precision* and *rank* values also

increased for summarized collections. Collections with a 90% compression ratio (“90\_3”, “90\_4”, “90\_5” and “90\_sem”) were higher than 50% compression summarized collections and two summarized collections with 70% (“70\_4” and “70\_5”). However, there was no statistical difference between the collections summarized with 80% and 90% compression.

By checking the use of keywords in the summarization, the study noticed that, for all compression ratios, collections summarized without the use of keywords displayed a higher *precision* rank average. Therefore, the use of three, four, five, or zero keywords did not influence IR.

In cases where IR with the Cassiopeia model was higher than the standard IR, the user’s effort to analyze the retrieved documents will be lower, given that a larger amount of relevant items has been retrieved. Regarding the speed with which IR was performed for summarized collections, response time decreased as the compression ratio increased. This allowed for faster access to documents and query processing. The “Original” collection (standard IR) presented the longest response time out of all collections.

Thus, IR with the Cassiopeia model, especially with high compression ratios that give rise to texts with reduced content, when compared to standard IR performed in full academic texts, the accuracy of the results increased, i.e., a larger number of relevant documents and fewer irrelevant ones were retrieved. Hence information overload decreased and greater agility in IR was ensured due to the reduction in response time; the indexing process became simpler due to the reduction in the number of index terms, proving to be relevant for large collections especially, in which case the computational cost is high; and high dimensionality was mitigated by decreasing the number of irrelevant words when managing textual data.

**REFERENCES**

1. Marcondes CH, Kuramoto H, Toutain LB, et al. *Bibliotecas digitais: Saberes e Práticas*. Salvador: EDUFBA, 2005, p. 345.
2. Sayão LF, Toutain LB, Rosa FG, et al. *Implantação e gestão de repositórios institucionais: políticas, memória, livre acesso e preservação*. Salvador: EDUFBA, 2009, p. 365.
3. Leite F, Amaro B, Batista T, et al. *Boas práticas para a construção de repositórios institucionais da produção científica*. Brasília: Ibict; 2012, p. 34.
4. Miranda IAA and Moura MA. Acesso aberto e gestão colaborativa de repositórios institucionais: a experiência da UFMG. *BiblioCanto* 2017; 37–50.
5. Baeza-Yates R and Ribeiro-Neto B. *Recuperação de Informação: Conceitos e Tecnologia das Máquinas de Busca*. 2th ed. Porto Alegre: Bookman, 2013, p. 612.
6. Manning CD, Raghavan P and Schütze, H. *An introduction to information retrieval*. Draft: Cambridge University Press, 2009, p. 581.
7. Silva RE, Santos PLVA and Ferneda E. Modelos de recuperação de informação e web semântica: a questão da relevância. *Informação & Informação* 2013; 27 – 44.
8. Aranha CN. *Uma abordagem de pré-processamento automático para mineração de textos em português: sob o enfoque da inteligência computacional*. Thesis. Pontifícia Universidade Católica do Rio de Janeiro, BRA, 2007.
9. Ferneda E. *Introdução aos Modelos Computacionais de Recuperação de Informação*. Rio de Janeiro (RJ, Brasil): Ciência Moderna, 2012, p. 155.
10. Dias MP and Carvalho JOF. Visualização da Informação e a sua contribuição para a Ciência da Informação. *DataGramZero*, <http://hdl.handle.net/20.500.11959/brapci/6137> (2007, accessed 08 May 2019).
11. Grainger T and Potter T. *Solr in action*. Shelter Island: Manning, 2014, p. 666.
12. Barth FJ. Uma introdução ao tema Recuperação de Informações Textuais. *Revista de Informática Teórica e Aplicada – RITA* 2013; 247 – 272.
13. Rezende SO, Marcacini RM and Moura MF. O uso da Mineração de Textos para Extração e Organização Não Supervisionada de Conhecimento. *Revista de Sistemas de Informação* 2011; 7–21.
14. Guelpeli MVC. *Cassiopeia: Um modelo de agrupamento de textos baseado em sumarização*. Thesis. Universidade Federal Fluminense, BRA, 2012.
15. Rino LHM and Pardo TAS. A Sumarização Automática de Textos: Principais Características e Metodologias. *Anais do XXIII Congresso da Sociedade Brasileira de Computação* 2003; 203–245.

16. Beyer K, Godstein J, Ramakrishnan R, et al. When is "Nearest Neighbor" Meaningful? In: Beeri C, Buneman P, editors. *International Conference on Database Theory (ICDT)*, Jerusalém, Israel: Springer Verlag, 1999, pp. 217–235.
17. Luhn HP. The automatic creation of literature abstracts. *IBM Journal of Research and Development* 1958; 159–165.
18. Nogueira BM. *Avaliação de métodos não-supervisionados de seleção de atributos para Mineração de Textos*. Dissertation. Universidade Federal de São Paulo, BRA, 2009.
19. Rocha VJC and Guelpe MVC. Pragmasum: Automatic Text Summarizer Based On User Profile. *International Journal of Current Research* 2017; 53935–53942.
20. Silva RDL and Silva EM. Mas o que é mesmo Corpus? – Alguns Apontamentos sobre a Construção de Corpo de Pesquisa nos Estudos em Administração. In: *XXXVII ANPAD Meeting*, Rio de Janeiro, Brasil, 07–11 september 2013, pp. 1–10. Brasil: ANPAD.
21. Aguiar LHG, Rocha VJC and Guelpe MVC. Uma coleção de artigos científicos de Português compondo um Corpus no domínio educacional. *PLURALS Interdisciplinar*, 2017; 60 –74.
22. Capes. Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). Brasília (DF, Brasil): Ministério da Educação, <http://www.capes.gov.br> (accessed 08 May 2019).
23. Lakatos EM and Marcone MA. *Fundamentos de metodologia científica*. 5. Ed. São Paulo: Atlas, 2003, p. 311.
24. Correia JSBL. *Indexação de Documentos Clínicos*. Dissertation. Universidade do Porto, PRT, 2016.
25. Solr. Apache Solr 7.2.1 Documentation. The Apache Software Foundation, [https://lucene.apache.org/solr/7\\_2\\_1](https://lucene.apache.org/solr/7_2_1) (2018, accessed 08 July 2019).
26. Callegari-jacques SM. *Bioestatística: Princípios e Aplicações*. 1th ed. Porto Alegre: Artmed, 2007, p. 264.
27. Campos GM. *Estatística Prática para Docentes e Pós-graduandos*. Ribeirão Preto: Universidade Federal de São Paulo, [http://www.forp.usp.br/restauradora/gmc/gmc\\_livro](http://www.forp.usp.br/restauradora/gmc/gmc_livro) (2000, accessed 15 August 2019).
28. Viali L. *Testes de hipóteses não paramétricos*. Apostila. Instituto de Matemática, Departamento de Estatística, Porto Alegre, BRA, 2008.
29. Wives LK. *Utilizando conceitos como descritores de textos para o processo de identificação de conglomerados (clustering) de documentos*. Thesis. Universidade Federal do Rio Grande do Sul, BRA, 2004.