UNIVERSIDADE FEDERAL DOS VALES DO JEQUITINHONHA E MUCURI

Programa de Pós-graduação em Educação

Luanna Azevedo Cruz

Modelo para recuperação de informação em repositórios institucionais utilizando a técnica de sumarização a partir da seleção de atributos do Cassiopeia

Diamantina
2019

Luanna Azevedo Cruz

Modelo para recuperação de informação em repositórios institucionais utilizando a técnica de sumarização a partir da seleção de atributos do Cassiopeia

Dissertação apresentada ao Programa de Pós-Graduação em Educação da Universidade Federal dos Vales do Jequitinhonha e Mucuri, como requisito para obtenção do título de Mestre.

Orientador: Prof. Dr. Marcus Vinicius Carvalho Guelpeli

Diamantina
2019

Elaborado com os dados fornecidos pelo(a) autor(a).

C957m Cruz, Luanna Azevedo.
Modelo para recuperação de informação em repositórios institucionais utilizando a técnica de sumarização a partir da seleção de atributos do Cassiopeia / Luanna Azevedo Cruz, 2019.
91 p.: il.

Orientador: Marcus Vinicius Carvalho Guelpeli

Dissertação (Mestrado – Programa de Pós-Graduação em Educação) - Universidade Federal dos Vales do Jequitinhonha e Mucuri, Diamantina, 2019.

1. Recuperação de informação. 2. Repositório institucional. 3. Mineração de Textos. 4. Sumarização. 5. Modelo Cassiopeia. I. Guelpeli, Marcus Vinicius Carvalho. II. Título. III. Universidade Federal dos Vales do Jequitinhonha e Mucuri.

CDD 020

Ficha Catalográfica – Sistema de Bibliotecas/UFVJM
Bibliotecária: Jullyele Hubner Costa – CRB6/2972

LUANNA AZEVEDO CRUZ

Modelo para recuperação de informação em repositórios institucionais utilizando a técnica de sumarização a partir da seleção de atributos do Cassiopeia

Dissertação apresentada ao MESTRADO EM EDUCAÇÃO, nível de MESTRADO como parte dos requisitos para obtenção do título de MESTRA EM EDUCAÇÃO

Orientador (a): Prof. Dr. Marcus Vinicius Carvalho Guelpeli

Data da aprovação : 08/11/2019

Prof.Dr. MARCUS VINICIUS CARVALHO GUELPELI - UFVJM

Prof.Dr. ALEXANDRE RAMOS FONSECA - UFVJM

Prof.Drª MARIA LUCIA BENTO VILLELA - UFVJM

Prof.Dr. RENATO DOURADO MAIA - UNIMONTES

DIAMANTINA

MINISTÉRIODAEDUCAÇÃO

UNIVERSIDADEFEDERALDOSVALESDOJEQUITINHONHAEMUCURI

DIAMANTINA – MINASGERAIS

PRÓ-REITORIADEPESQUISAEPÓS-GRADUAÇÃO

ATESTADO DE DEFESA POR VIDEOCONFERÊNCIA

Atesto para os devidos fins que no dia 08 de novembro de 2019, às 08h, nas dependências da UFVJM – em (Diamantina), foi realizada a defesa de dissertação do(a) discente Luanna Azevedo Cruz com o trabalho intitulado ***"Modelo para recuperação de informação em repositórios institucionais utilizando a técnica de sumarização a partir da seleção de atributos do Cassiopeia"*** sob minha orientação no Programa de Pós-graduação em Educação (PPGed).

Na qualidade de presidente da banca, atesto queo(a) Prof. (a) Dr. (a) Renato Dourado Maia, docente da Universidade Estadual de Montes Claros - UNIMONTES, participou através de videoconferência.

Em virtude da participação remota do membro da banca acima indicado, eu, Prof. Dr. Marcus Vinicius Carvalho Guelpeli, enquanto servidor público, no gozo de fé pública, assino no lugar desse na Ata de Defesa e na Folha de Aprovação da referida defesa.

Por ser verdade, dou fé e assino o presente atestado.

Diamantina, 08 de novembro de 2019.

AGRADECIMENTOS

Ao meu orientador, professor Marcus, pela paciência, disposição e ensinamentos durante a realização do trabalho.

Aos professores Alexandre, Maria Lúcia e Renato, pelas contribuições e sugestões.

Aos meus pais, Walter e Vanda, por todo amor e apoio, e por serem minha referência de vida.

Às minhas irmãs Ludymilla, Vanessa e Amanda, pela amizade e estímulo em todos os momentos.

Ao meu noivo Maurício, maior incentivador desta jornada, pelo amor, força e dedicação.

Aos colegas do mestrado e do grupo de pesquisa MTPLNAM pelo suporte.

Aos meus amigos de Diamantina e Seabra, pelos momentos de distração, apoio e carinho. Em especial, Jesyka, Marco e Hércules, pela parceria durante a pesquisa.

À SRE Diamantina pelo incentivo.

RESUMO

Um grande repositório de cultura e conhecimento vem sendo formado a partir de documentos digitais criados por usuários das mais diversas áreas. No contexto educacional, textos acadêmicos compõem bases textuais, como os repositórios institucionais, que são fonte de informação e auxiliam no processo de ensino e aprendizagem. No entanto, assimilar e lidar com o elevado volume de informação disponível, localizando-as de forma rápida e precisa, passou a ser um desafio. Neste sentido, a área de Recuperação de Informação atua com o propósito de detectar, dentre uma coleção de documentos, os que satisfazem às necessidades do usuário. Porém, problemas como quantidade excessiva de documentos retornados e falta de relevância e precisão dos resultados apresentados dificultam o processo de recuperação de informação. Técnicas de Mineração de Textos podem auxiliar nesse processo, por meio da extração de dados, descoberta de padrões, associações e regras, realização de resumo, e análises em documentos de texto. Dessa forma, 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), num *corpus* de textos acadêmicos, auxilia na recuperação de informação, diminuindo a sobrecarga de informação e melhorando a precisão dos resultados retornados ao usuário. A seleção de atributos do modelo consiste em um método de redução da alta dimensionalidade e dados esparsos. A pesquisa foi desenvolvida em seis etapas que compreenderam as seguintes ações: levantamento bibliográfico; preparação do *corpus* e sumarização dos textos acadêmicos; implementação de um buscador; execução da recuperação de informação padrão e com a seleção de atributos do modelo Cassiopeia; avaliação da recuperação de informação por intermédio das métricas *precision, recall,* e *F-measure;* e, por fim, 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.

ABSTRACT

A great repository of culture and knowledge has been formed from digital documents created by users from various áreas. In the educational context, academic texts make up textual bases, such as institutional repositories, which are a source of information and assist in the teaching and learning process. However, assimilating and dealing with the high volume of information available, locating it quickly and accurately, has become a challenge. In this sense, the Information Retrieval area acts with the purpose of detecting, among a collection of documents, those that satisfy the user's needs. However, problems such as excessive amount of returned documents and lack of relevance and accuracy of the results presented make the information retrieval process difficult. Text Mining Techniques can assist in this process by extracting data, discovering patterns, associations and rules, summarizing, and analyzing text documents. Thus, the objective of this research is to analyze if the application of the summarization technique, based on the attributes selection method (words) of the Cassiopeia model (implemented in the PragmaSUM summarizer), in a corpus of academic texts, helps in the retrieval of information, reducing information overload and improving the accuracy of results returned to the user. Model attribute selection consists of a method of reducing hogh dimensionality and sparse data. The research was developed in six steps that comprised the following actions: bibliographic survey; corpus preparation and summarization of academic texts; implementation of a search engine; performing standard information retrieval and selecting attributes from the Cassiopeia model; assessment of information retrieval using precision, recall, and F-measure metrics; and, finally, data analysis from Friedman ANOVA and Kendall Agreement coefficient statistical tests. The results showed that summarization, mainly performed with high compression rates (80% and 90%), reduced the information overload and increased the accuracy of the results presented to the user, allowing quality information retrieval in academic texts. In addition, it simplified the indexing process, attenuated the high dimensionality and promoted faster information retrieval.

Keywords: Information Retrieval. Institutional repository. Text Mining. Summarization. Cassiopeia model.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CAPES – Coordenação de Aperfeiçoamento de Pessoal de Nível Superior

EA – Ensino aprendizagem

EE – Educação especial

EP – Educação permanente

EPE – Educação pré-escolar

FE – Filosofia da educação

HE – História da educação

PE – Psicologia educacional

POE – Política educacional

RI – Recuperação da Informação

SA – Sumarização Automática

SE – Sociologia da educação

SRI – Sistema de Recuperação de Informação

TE – Tecnologia educacional

SUMÁRIO

1 INTRODUÇÃO

Com os avanços tecnológicos e o uso massivo das novas tecnologias digitais de informação e comunicação (NTDICs) os usuários passaram a criar documentos digitais, compondo um grande repositório de cultura e conhecimento. Desse modo, o contínuo desenvolvimento das NTDICs tem criado possibilidades e desafios em diferentes campos do conhecimento. Na área educacional, em especial, com o crescimento das produções científicas, as universidades e instituições de pesquisa têm dado ênfase à estruturação de mecanismos dotados de tecnologia, como os repositórios institucionais, com o intuito de ampliar a visibilidade, acessibilidade, preservação e disseminação das pesquisas desenvolvidas. Para o desenvolvimento e implantação dos repositórios institucionais, pacotes de *software* podem fornecer ferramentas que apresentam funcionalidades de organização, armazenamento e recuperação da informação (MARCONDES *et. al,* 2005; MIRANDA e MOURA, 2017).

Segundo Leite *et al.* (2012), os repositórios oferecem serviço de informação científica, em ambiente digital, e apresentam, como base informacional, textos completos e digitais que são armazenados e disponibilizados para acesso dos usuários. Assim, textos como artigos, dissertações e teses, são utilizados como fonte de informação e podem contribuir com melhorias no processo de ensino aprendizagem e, ao mesmo tempo, permitir inovação e auxiliar na universalização do acesso ao conhecimento. No entanto, percebe-se que a quantidade de documentos disponibilizados vem crescendo de maneira significativa, e essa expansão informacional deu origem a problemas como sobrecarga, falta de organização e desestruturação das informações. Dessa forma, lidar com o elevado volume de informação textual disponível, localizando-as de forma rápida e precisa, passou a ser um desafio para os usuários (MARCONDES *et. al,* 2005; SAYÃO *et. al,* 2009; MIRANDA e MOURA, 2017).

Nesta perspectiva, com o objetivo de facilitar o acesso às informações, métodos de organização da informação, para posterior busca e recuperação, vêm sendo utilizados. É nesse âmbito que se insere a Recuperação de Informação (RI), área da Ciência da Computação que visa a detectar, dentre uma coleção de documentos, também chamada de *corpus,* os que satisfazem às necessidades do usuário, facilitando o acesso, recuperação e análise de informações (BAEZA-YATES e RIBEIRO-NETO, 2013).

Atualmente, os usuários efetuam a RI em diversos contextos: (1) na *web*, a partir de consultas efetuadas pelo usuário, sistemas como, por exemplo, *Google*[1] e *Yahoo*[2], fornecem possibilidades de pesquisas sobre bilhões de documentos indexados, armazenados em milhões de computadores, além de percorrer os *hiperlinks* das páginas *web*; (2) na recuperação de informações pessoais, com o uso de ferramentas de pesquisa disponíveis em sistemas operacionais como, por exemplo, os campos de pesquisa do *Windows* 10[3], ou em programas de *e-mail* que fornecem pesquisa nos campos de busca; e (3) no âmbito da pesquisa corporativa e institucional, a partir da recuperação em coleções de documentos armazenados em bases de dados especificas como, por exemplo, os repositórios institucionais, bibliotecas digitais e acervos documentais. Nesses repositórios, a busca de documentos por assunto é a principal forma de recuperação utilizada pelos usuários e, a partir da consulta inserida, o sistema de RI pode realizar buscas em textos completos ou em metadados[4]. O foco deste trabalho permeia essa última abordagem da RI, com efetivação de busca em textos completos e não estruturados. Nos repositórios em geral e centros de documentação, o processo de RI pode ser desenvolvido em virtude de todo o universo ser conhecido e estar acessível, seguindo os mesmos padrões (MANNING, RAGHAVAN e SCHÜTZE, 2009; SILVA, SANTOS E FERNEDA, 2013).

Os processos de interação entre o usuário e os documentos são efetuados pelos sistemas de recuperação de informação (SRIs), ou seja, ferramentas de busca (FERNEDA, 2012). Dessa maneira, o principal objetivo de um SRI é efetuar a recuperação de documentos que sejam úteis para o usuário, e para isso deve criar uma representação dos textos do *corpus* e apresentá-los ao usuário de maneira que seja possível uma rápida seleção dos itens que satisfazem total ou parcialmente à sua necessidade de informação, formalizada por meio de uma consulta. Assim, um SRI deve recuperar todos os documentos relevantes e a menor quantidade possível de irrelevantes. Contudo, a relevância é uma temática bastante debatida devido a seu caráter subjetivo, uma vez que um documento considerado relevante para determinado usuário pode não ser para outro (FERNEDA, 2003; ARANHA, 2007).

A qualidade dos resultados retornados ao usuário pode ser verificada por meio da avaliação da RI, que é o processo no qual se associa uma métrica quantitativa aos resultados produzidos por um sistema de RI, em resposta a um conjunto de consultas do usuário. Tal

[1] Disponível em: http://www.google.com
[2] Disponível em: http://br.yahoo.com
[3] Disponível em: http://www.microdoft.com/pt-br/software-download/windows10
[4] Metadados são dados sobre dados, ou informação sobre informação. Eles podem ser criados pelos autores do trabalho ou automaticamente pelos sistemas, e fornecem informações sobre o documento digital, como por exemplo, palavras-chave, título, autores e resumo (SAYÃO *et. al,* 2009).

avaliação é simples e possibilita a repetição do experimento a custos relativamente baixos, quando comparados à complexidade e dispêndio necessários para a execução de testes com usuários (MANNING, RAGHAVAN e SCHÜTZE, 2009; BAEZA-YATES e RIBEIRO-NETO, 2013).

A avaliação da qualidade é importante para que problemas relacionados à RI possam ser identificados e solucionados. De acordo com Baeza-Yates e Ribeiro-Neto (2013), muitas vezes os resultados apresentados ao usuário não são precisos e a quantidade de textos retornados é elevada, o que gera sobrecarga de informação e, consequentemente, dificulta a assimilação de informações. Tais aspectos são confirmados por Dias e Carvalho (2017), que destacam a sobrecarga como uma das principais preocupações na apresentação dos resultados obtidos por meio de SRIs. Grainger e Potter (2014) complementam ao afirmarem que quando os usuários realizam uma pesquisa por documentos, apenas 10% deles estão dispostos a ir além da primeira página de resultados e somente 1% chega a navegar até a terceira página. Outra importante questão reside na possibilidade dos documentos recuperados possuírem termos redundantes e desnecessários para a RI, já que, nos SRIs, os documentos longos têm maior chance de corresponder à consulta simplesmente pelo tamanho que possuem, o que não significa que sejam relevantes para a expressão de busca (SILVA, SANTOS e FERNEDA, 2013).

É nesse contexto que a Mineração de Textos (MT) opera, auxiliando nos processos de RI, extração de dados, resumos textuais, descoberta de padrões, associações e regras, e realização de análises em documentos de texto. Além disso, a adequada organização das coleções de textos agiliza os processos de busca e recuperação da informação. Assim, com a finalidade de extrair conhecimento e tratar informações textuais, técnicas de MT, como a sumarização, podem ser aplicadas nas coleções. A sumarização trata da redução do conteúdo textual, produzindo sumários sem que haja perda de informatividade e sentido (CARRILHO JÚNIOR, 2007; SILVA, 2009; REZENDE, MARCACINI e MOURA, 2011; GUELPELI, 2012).

Dessa forma, considerando a importância dos repositórios institucionais como fonte de informação e meio de divulgação do conhecimento produzido por uma comunidade científica, a presente pesquisa consistiu em analisar se a aplicação da técnica de sumarização, a partir do método de seleção de atributos (palavras) do modelo Cassiopeia, num *corpus* de textos acadêmicos, auxilia na recuperação de informação, diminuindo a sobrecarga de informação e melhorando a precisão dos resultados retornados ao usuário. A seleção de atributos proposta no modelo Cassiopeia trata-se de um novo método para definição do corte de Luhn (método de

redução da dimensionalidade e dados esparsos). Tal método é implementado no sumarizador utilizado nesta pesquisa, o PragmaSUM, com o propósito de produzir sumários com maior informatividade (ROCHA e GUELPELI, 2017).

1.1 Objetivos

Analisar se a aplicação da técnica de sumarização, a partir do método de seleção de atributos do modelo Cassiopeia (implementado no sumarizador PragmaSUM), num *corpus* de textos acadêmicos que compõem repositórios institucionais, auxilia na recuperação de informação, diminuindo a sobrecarga de informação e melhorando a precisão dos resultados retornados ao usuário.

Para alcançar o objetivo geral, os seguintes objetivos específicos deverão ser atingidos:

- Aplicar a técnica de sumarização, realizada pelo sumarizador PragmaSUM a partir da seleção de atributos do modelo Cassiopeia, no processo de recuperação de informação de textos acadêmicos que compõem repositórios institucionais;
- Implementar um sistema de recuperação de informação, ou seja, um buscador para RI textual;
- Realizar testes executando o processo de RI de duas maneiras: a primeira, seguindo o modelo padrão de RI; e a segunda, incluindo no processo a sumarização com o método de seleção de atributos do modelo Cassiopeia;
- Avaliar a recuperação de informação utilizando as métricas *precision, recall e F-measure*, em ambas execuções, e comparar os resultados dos dois modelos analisando os valores obtidos a partir das métricas;
- Verificar se a sumarização, com a seleção de atributos do modelo Cassiopeia, reduz a sobrecarga de informação e melhora a precisão dos resultados apresentados ao usuário.

1.2 Problema

O processo de recuperação de informação textual começa quando o usuário inicia uma pesquisa em um sistema de RI, buscando satisfazer suas necessidades de informação. Ao realizar uma consulta, em repositórios institucionais, o usuário espera que o sistema retorne os textos relevantes ao seu interesse. Porém, a quantidade excessiva de documentos retornados, o que causa sobrecarga de informação, e a falta de relevância e precisão dos resultados apresentados dificultam a assimilação e obtenção de informação útil.

1.3 Hipótese

A aplicação da técnica de sumarização, a partir do método de seleção de atributos do modelo Cassiopeia (implementado no sumarizador PragmaSUM), num *corpus* de textos acadêmicos disponibilizados em repositórios institucionais, auxilia na recuperação de informação, diminuindo a sobrecarga de informação e melhorando a precisão dos resultados retornados ao usuário.

1.4 Estrutura da Dissertação

Esta dissertação foi estruturada em cinco capítulos, a saber:

- O capítulo 2 – Fundamentação Teórica – apresenta o embasamento teórico que fundamenta a pesquisa. Serão mostrados assuntos relacionados à recuperação de informação, tais como sistemas, modelos e avaliação da RI, e a técnica de sumarização, destacando o modelo Cassiopeia e o sumarizador automático PragmaSUM.
- Os passos metodológicos serão apresentados no capítulo 3 – Metodologia – e compreendem a execução de seis etapas: levantamento bibliográfico referente aos temas envolvidos; preparação do *corpus* e sumarização, utilizando o sumarizador automático PragmaSUM, dos textos acadêmicos do domínio educacional; implementação de um buscador com funcionalidades de busca e indexação; execução da recuperação de informação padrão e utilizando o modelo Cassiopeia; avaliação da recuperação de informação por intermédio das

métricas *precision*, *recall e F-measure;* e, por fim, análise estatística dos resultados.

- Os Resultados e Discussões são elucidados no capítulo 4.
- Finalmente, no capítulo 5 – Considerações Finais – serão apresentadas as conclusões do trabalho, bem como limitações, contribuições e trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão apresentados os principais conceitos que fundamentam esta pesquisa. Inicialmente serão abordados assuntos relacionados à recuperação de informação, tais como sistemas, modelos e avaliação da RI, e em seguida, que envolvem Mineração de Textos: sumarização automática, seleção de atributos do modelo Cassiopeia, e o sumarizador PragmaSUM.

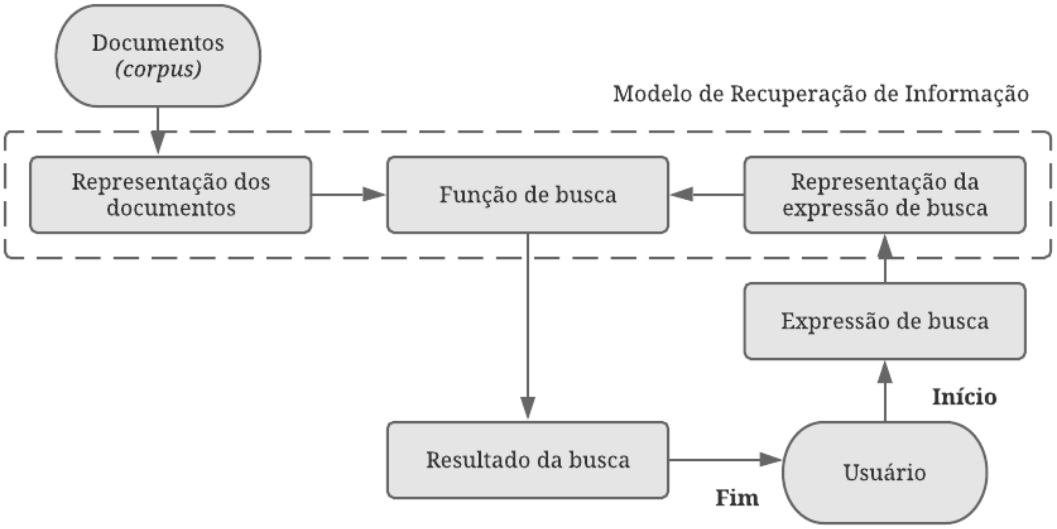
2.1 Recuperação de Informação – RI

A Recuperação de Informação é uma subárea da Ciência da Computação que trata da representação, armazenamento, organização e acesso à informação. O termo "*Information Retrievel*" (Recuperação de Informação) foi definido por Mooers (1951 *apud* Ferneda, 2003, p. 11), da seguinte forma: é o nome do processo no qual "um possível usuário de informação pode converter a sua necessidade de informação em uma lista real de citações de documentos armazenados que contenham informações úteis a ele". Ainda segundo o autor, a RI trata dos "aspectos intelectuais da descrição da informação e sua especificação para busca, e de qualquer sistema, técnicas ou máquinas que são empregadas para realizar esta operação".

Nesse contexto, a RI se concentra em facilitar a aquisição de itens de interesse do usuário, tais como documentos, páginas *web*, objetos multimídia, etc. Assim, segundo Manning, Raghavan e Schütze (2009), a RI busca encontrar materiais (geralmente documentos) de natureza não estruturada, como textos disponíveis em coleções, e que satisfaçam a necessidade de informação do usuário.

Um processo de RI é a "especificação formal de três elementos: a representação dos documentos, a representação da necessidade de informação por meio de uma expressão de busca e como esses dois elementos serão comparados na função de busca", como pode ser verificado na Figura 1 (FERNEDA, 2012, p. 20).

Figura 1– Representação do Processo de Recuperação de Informação



Fonte: Ferneda, 2012.

Um documento é "todo artefato que representa ou expressa um objeto, uma ideia ou uma informação por meio de signos gráficos e icônicos (palavras, imagens, diagramas, mapas, figuras, símbolos), sonoros e visuais" (LE COADIC, 2004, p. 5). Os documentos podem ser representados por meio da extração e análise de seu conteúdo, permitindo que sejam obtidos conceitos que os identificam. Assim, o documento pode ser uma unidade de recuperação, ou seja, o item básico que pode ser recuperado como resultado de uma consulta (FERNEDA, 2012).

A expressão de busca é a maneira que o usuário, componente central do processo de RI, utiliza para comunicar sua necessidade de informação com o sistema e, normalmente, é formada por termos que representam tal necessidade. A dificuldade reside em evitar a recuperação de documentos irrelevantes para que o esforço em selecionar os itens úteis seja minimizado. Assim, na RI é necessário que, dado um documento, seja possível calcular o quão similar ele é em relação à consulta e aos demais documentos que compõem o *corpus* (CARRILHO JÚNIOR, 2007; FERNEDA, 2012).

A representação da expressão de busca deve ser semelhante à utilizada na representação dos documentos para que seja possível realizar uma comparação entre a consulta e os documentos que compõem o *corpus.* Tal comparação é feita pela função de busca, que recupera os itens que provavelmente oferecem a informação que o usuário pesquisou. Por fim, o resultado da busca consiste nesse conjunto de itens que é apresentado ao usuário (FERNEDA, 2012; BAEZA-YATES e RIBEIRO NETO, 2013).

Dessa forma, ao realizar uma consulta, em um mecanismo de busca, o usuário espera que o sistema retorne os textos relevantes ao seu interesse. No entanto, "nenhum sistema

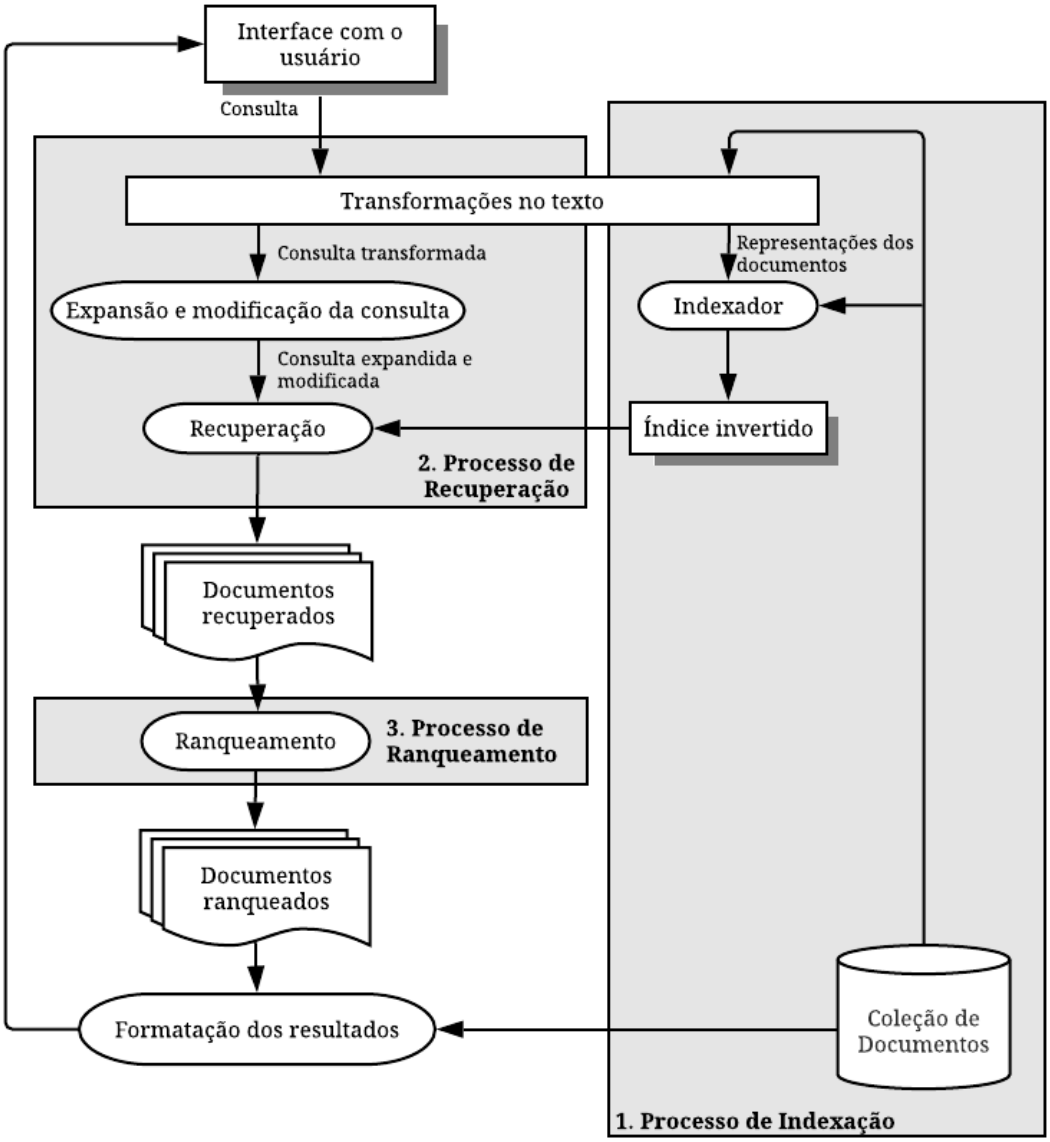
de RI pode oferecer respostas perfeitas a todos os usuários o tempo todo" (BAEZA-YATES e RIBEIRO-NETO, 2013, p. 4). Adicionalmente, alguns problemas dificultam o processo de recuperação de informação, tais como: quantidade excessiva de documentos retornados, o que causa sobrecarga de informação; falta de relevância e precisão dos resultados apresentados; e baixa tolerância dos usuários em verificar a lista de documentos recuperados, analisando somente os primeiros resultados mostrados (SPINK *et al*., 2001; MANNING, RAGHAVAN e SCHÜTZE, 2009).

2.2 Sistema de Recuperação de Informação – SRI

Segundo Ferneda (2012, p. 13), o objetivo principal de um SRI é "representar o conteúdo dos documentos do *corpus* e apresentá-los ao usuário de uma maneira que lhe permita uma rápida seleção dos itens que satisfazem total ou parcialmente a sua necessidade de informação". Assim sendo, o foco é recuperar todos os documentos relevantes para o usuário e o menor número possível de documentos irrelevantes, ou seja, ajudar os usuários a encontrar informações de seu interesse. Para isso, é necessário interpretar o conteúdo desses textos e ordená-los, considerando o quão relevante são para a consulta feita pelo usuário. A relevância, na RI, é uma temática bastante discutida devido ao seu caráter abstrato, que dificulta a criação de estruturas que garantam que os resultados de uma expressão de busca sejam relevantes para o usuário, e subjetivo, porque usuários diferentes podem ter visões distintas sobre o que é ou não relevante (SILVA, SANTOS e FERNEDA, 2013; BAEZA-YATES e RIBEIRO-NETO, 2013; BARTH, 2013).

Para realizar a recuperação de informação, um SRI funciona a partir da execução dos processos de indexação, recuperação e ranqueamento, respectivamente, como é exibido na Figura 2.

Figura 2 – Processos de indexação, recuperação e ranqueamento de documentos



Fonte: Baeza-Yates e Ribeiro-Neto, 2013, p. 9. Adaptado.

Para implantar um SRI, o primeiro passo consiste em obter o *corpus*, ou seja, a coleção de documentos selecionados e organizados seguindo critérios necessários à pesquisa (SILVA R. e SILVA E., 2013). Tal coleção pode ser particular ou coletada na *web* para posteriormente ser armazenada em um repositório.

Após a definição do *corpus,* o **processo de indexação** é o primeiro a ser executado pelo SRI. Os textos da coleção são transformados com a aplicação de operações textuais, como: remoção de palavras com pouca significação (*stopwords)*; normalização de termos reduzindo-os a seus radicais (*stemming*); e seleção de termos que serão utilizados como termos de indexação, para representação dos documentos. Após a preparação, os textos precisam ser indexados, ou seja, um índice do texto deve ser criado para que os processos de recuperação e

ranqueamento sejam executados com maior agilidade. O índice é um conjunto de termos que indicam onde a informação de interesse pode estar localizada e permite rapidez no acesso aos dados e processamento das consultas (FERNEDA, 2003).

Segundo Baeza-Yates e Ribeiro-Neto (2013), o índice invertido é a estrutura de índice mais utilizada na RI, pois possui todas as palavras diferentes do *corpus* e, para cada uma delas, a lista de documentos que as contêm. Aranha (2007, p. 57) define o índice invertido como:

> [...] uma estrutura de dados com um registro para cada palavra. Nesse registro, existe a informação sobre os documentos em que ela ocorre, o número de ocorrências e a posição em cada um deles. O índice invertido contém ainda algoritmos estatísticos e probabilísticos para computar rapidamente a relevância dos documentos. A estrutura de dados contendo o índice é armazenada no disco rígido, já que normalmente requer muito espaço. Ela ainda tem funcionalidades de atualização para adicionar um novo documento dentro da base.

Um exemplo simples de seguimento de arquivo de índice invertido pode ser visualizado na Tabela 1.

Tabela 1 – Exemplo de índice invertido

Palavra ou termo	**Ocorrências (identificador do documento)**
Computador	1, 5, 6, 8
Impressora	2, 3, 7
Mouse	3, 6, 7, 8
Monitor	1, 2, 5, 9, 12

Nesse caso, o *corpus* é formado pelas palavras computador, impressora, *mouse* e monitor. Para cada palavra são listados os documentos nos quais elas ocorrem. Assim, se o usuário buscar "impressora", os documentos '2', '3' e '7' são exibidos. Da mesma forma, a busca por "computador e *mouse*" ("computador AND *mouse"*) retorna o conjunto de documentos indexados por ambos os termos, resultando nos documentos de identificadores '6' e '8'.

Quando a indexação é finalizada, o **processo de recuperação** pode ser iniciado. Assim, o usuário, por meio da *interface,* realiza uma consulta que é analisada e transformada por operações, tais como: correção ortográfica; remoção de *stopwords*, para reduzir a estrutura de indexação; e *stemming*, para diminuir as variantes da mesma palavra raiz para um conceito comum. As *stopwords*, como artigos, preposições e conjunções, são inúteis para a recuperação,

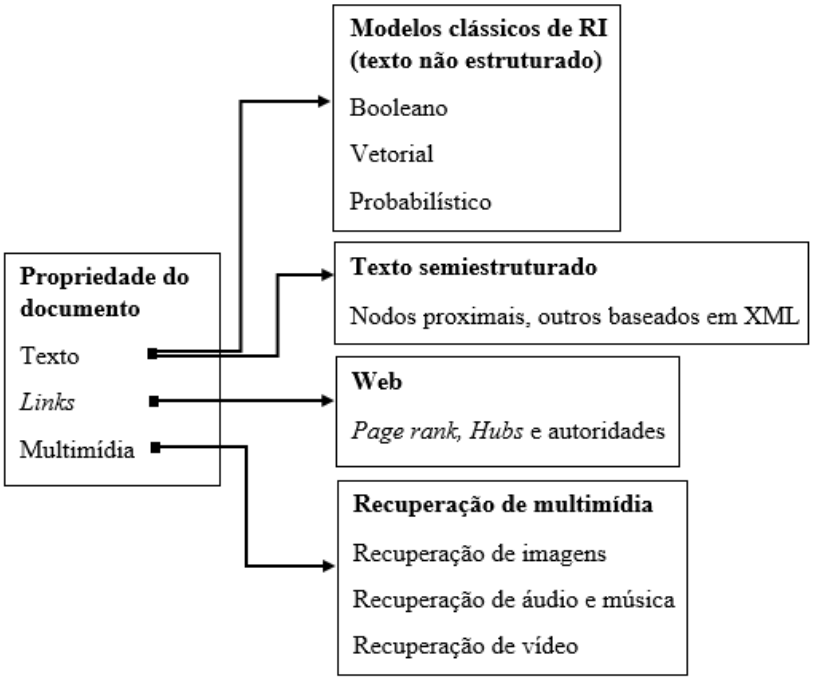
por serem muito frequentes entre os documentos de uma coleção (MANNING, RAGHAVAN e SCHÜTZE, 2009).

Após o processamento da consulta e verificação do índice, a coleção de documentos recuperados é formada e o **processo de ranqueamento** é iniciado. Os documentos são ranqueados (ordenados) e os que apresentam maior probabilidade de serem relevantes são identificados e retornados ao usuário (BAEZA-YATES e RIBEIRO-NETO, 2013).

2.3 Modelos de recuperação de informação

Segundo Baeza-Yates e Ribeiro-Neto (2013, p. 21) "a modelagem em RI é um processo complexo que tem como propósito gerar uma função de ranqueamento, ou seja, uma função que atribui escores a documentos em relação a uma consulta". Os modelos de RI são essencialmente baseados em texto, ou seja, para o ranqueamento, em relação à consulta do usuário, são utilizados os textos dos documentos. Porém, *links* e objetos multimídia podem ser empregados (MANNING, RAGHAVAN e SCHÜTZE, 2009). A Figura 3 ilustra uma taxonomia de modelos de RI.

Figura 3 – Taxonomia de modelos de RI



Fonte: Baeza-Yates e Ribeiro-Neto, 2013, p. 9. Adaptada.

Para documentos baseados em textos devem ser considerados os textos não estruturados e os semiestruturados. Os modelos clássicos de RI (booleano, vetorial e probabilístico) envolvem textos não estruturados, ou seja, são modelados como uma sequência de palavras. Já os modelos que lidam com texto semiestruturado (nodos proximais e outros em xml), apresentam itens estruturais do texto, como títulos e seções. Na *Web,* tais modelos não são suficientes devido à existência de *links* entre as páginas, o que exige modelos específicos como *Page rank*[5] e o *Hubs & Autoridades*[6]. Por fim, diferentes estratégias de recuperação são utilizadas para dados multimídia, como imagens, áudio e música, e vídeo (BAEZA-YATES e RIBEIRO-NETO, 2013; BARTH, 2013). O foco desta pesquisa é a recuperação de informação em documentos textuais, logo, os modelos clássicos serão apresentados a seguir.

2.3.1 Recuperação de informação clássica: modelos booleano, vetorial e probabilístico

Os modelos clássicos de RI, booleano, vetorial e probabilístico, consideram que cada texto é descrito por uma coleção de palavras-chave, denominadas termos de indexação. O modelo booleano é considerado o mais fraco, devido à inexistência de uma completa conformidade entre consulta e documentos. Com relação aos modelos vetorial e probabilístico, o primeiro mostra-se robusto e melhor em coleções genéricas, ao fornecer uma função de ranqueamento simples e eficaz (MANNING, RAGHAVAN e SCHÜTZE, 2009; BARTH, 2013; CARDOSO, 2004).

O modelo booleano, baseado na teoria de conjuntos e álgebra booleana, é simples e possui semântica precisa. Nele, os termos de indexação apresentam pesos binários, nos quais 1 (um) indica a presença e 0 (zero) a ausência do termo no documento. As buscas são formuladas por meio de expressão booleana convencional, na qual os termos são ligados com o uso de operadores lógicos (E, OU e NÃO). Apesar da simplicidade e formalismo do modelo, fatores como a não existência de ranqueamento, que pode levar à recuperação de muitos ou poucos documentos, e a necessidade de criação de consultas booleanas pelo usuário, são considerados desvantagens (BAEZA-YATES e RIBEIRO-NETO, 2013; FERNEDA, 2003).

No modelo vetorial, os termos de indexação e os termos da expressão de busca possuem pesos não binários, usados para definir o grau de similaridade entre a consulta do

[5] Ponderação baseada em *links* que simula um usuário navegando aleatoriamente na *web.*
[6] Autoridades são páginas que possuem muitos *links* apontando para elas, enquanto *hubs* são páginas que apresentam muitos *links* de saída (BAEZA-YATES E RIBEIRO-NETO, 2013).

usuário e os documentos da coleção. Trata-se de um modelo simples e rápido no qual os documentos recuperados são ordenados de acordo com o grau de similaridade, de forma decrescente. O modelo supõe a formação de um par entre documentos e termos da consulta, e efetua o ranqueamento dos documentos, o que propicia respostas mais precisas que o booleano. De acordo com Baeza-Yates e Ribeiro-Neto (2013), alguns aspectos fazem com que o modelo vetorial seja um método popular continuamente utilizado, tais como: melhoria da qualidade da recuperação devido à ponderação de termos; uso da estratégia do par termo-documento, que possibilita a recuperação de documentos próximos às condições da busca; e o cálculo do grau de similaridade (FERNEDA, 2003; BARTH, 2013).

No modelo probabilístico, uma consulta é um subconjunto de termos de indexação e o documento é representado por um vetor de pesos que definem a existência ou não de termos de indexação. O objetivo é tentar estimar a probabilidade de um documento ser relevante de acordo com uma consulta do usuário, considerando as informações disponíveis ao sistema. Conforme Batista Júnior (2006, p. 16), tal modelo "assume que os termos são independentes e cada documento é representado por um vetor binário em que cada elemento indica a ausência ou presença de um termo da coleção no documento", ou seja, os documentos são recuperados de acordo com a probabilidade de relevância. O ponto forte do modelo é o ranqueamento dos documentos de maneira decrescente, conforme a probabilidade de relevância. No entanto, o método não considera a frequência de ocorrência do termo de indexação e apresenta necessidade de estimar a separação dos documentos em relevantes e irrelevantes (COOPER, 1994; BARTH, 2013).

2.4 Avaliação da Recuperação de Informação

Avaliar a RI significa verificar a qualidade dos resultados retornados ao usuário, ou seja, o quão bem o sistema atende à sua necessidade de informação. Segundo Baeza-Yates e Ribeiro-Neto (2013, p.106), avaliação da recuperação é "um processo sistemático no qual se associa uma métrica quantitativa aos resultados produzidos por um sistema de RI em resposta a um conjunto de consultas de usuário".

Desse modo, o uso de métricas e a comparação dos resultados apresentados pelo sistema com os resultados sugeridos por uma coleção de referência, criada por humanos, são os procedimentos de avaliação mais comumente utilizados pela simplicidade e possibilidade de repetição do experimento a custos relativamente baixos, quando comparados à complexidade e dispêndio necessários para execução de testes com usuários. Os autores também afirmam que

essas coleções apresentam importantes vantagens, tais como: rapidez da avaliação; facilidade de reprodução dos experimentos; e, possibilidade de criação de diferentes coleções com focos específicos em determinadas necessidades de informação.

A organização das coleções de referência, ou seja, coleções de teste padronizadas, envolve os seguintes elementos: um conjunto de documentos; um conjunto de consultas (expressões de busca); e o julgamento de relevância, feito por humanos, no qual cada documento é classificado como relevante ou irrelevante em relação a uma consulta. Claramente, isso é viável para coleções pequenas (LUGO, 2004; MANNING, RAGHAVAN e SCHÜTZE, 2009).

As métricas quantitativas, *precision* (precisão) e *recall* (revocação), são as medidas mais utilizadas na avaliação da RI e são calculadas conforme Equações 1 e 2, respectivamente (MANNING, RAGHAVAN e SCHÜTZE, 2009).

Precision é a fração dos documentos recuperados que é relevante, isto é:

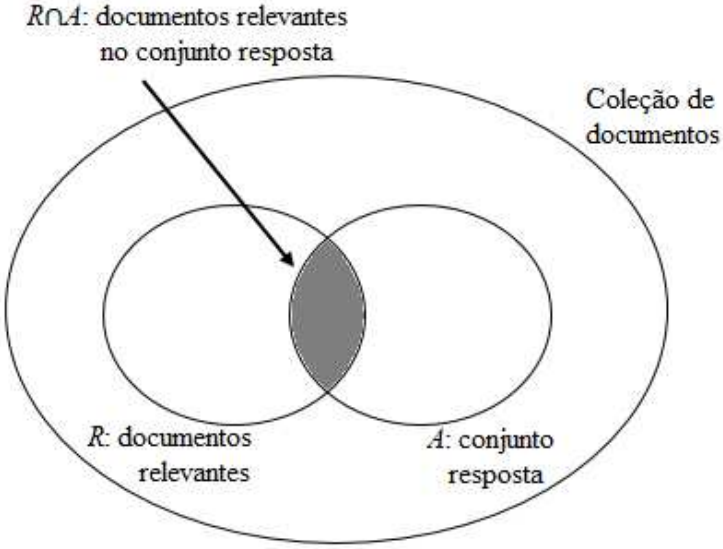
$$Precision = \frac{número\ de\ documentos\ relevantes\ recuperados}{número\ de\ documentos\ recuperados} \quad . \qquad (1)$$

Recall é a fração dos documentos relevantes que foi recuperada, ou seja:

$$Recall = \frac{número\ de\ documentos\ relevantes\ recuperados}{número\ total\ de\ documentos\ relevantes} \quad . \qquad (2)$$

Assim, conforme a Figura 4, uma requisição de informação (consulta) *I* possui seu conjunto *R* de documentos relevantes (coleção de referência). Quando tal consulta é processada, um conjunto de documentos *A* é recuperado (conjunto resposta).

Figura 4 – Precisão e revocação para uma requisição de informação I



Fonte: Baeza-Yates e Ribeiro-Neto, 2013, p. 111.

Segundo Manning, Raghavan e Schütze (2009), o uso de uma só medida que combina *precision* e *recall* pode ser pertinente. A *F-measure* (medida-F), média harmônica entre *precision* e *recall*, é uma dessas opções e é calculada da seguinte maneira (Equação 3):

$$F\text{-}measure = 2 * \frac{Precision * Recall}{Precision + Recall} \quad . \tag{3}$$

A *F-measure* apresenta valores no intervalo de [0,1]. Quanto mais próximos de 0 (zero) piores os resultados e quanto mais próximos de 1 (um) melhores (MANNING, RAGHAVAN E SCHÜTZE; 2009; BARTH, 2013).

2.5 *Apache Solr*

O *Apache Solr*[7] é um servidor de pesquisa para recuperação de informação em bases de dados textuais que apresenta as seguintes características (GRAINGER e POTTER, 2014; CORREIA, 2016; SOLR, 2018):

a) Escalável e tolerante a falhas: distribui o trabalho de indexação e processamento de consulta em múltiplos servidores;
b) Código aberto, fácil instalação e configuração;
c) Otimização da pesquisa: rápido para executar consultas complexas;

[7] *Apache Solr*. Ferramenta de busca e indexação. Disponível em: http://lucene.apache.org/solr.

d) Projetado para lidar com grandes volumes de documentos;
e) Centrado em texto: otimizado para pesquisar textos em linguagem natural, como *e-mails*, páginas *web*, documentos em formato .pdf, etc.;
f) Classificação dos resultados por relevância: os documentos são retornados em ordem de relevância, de acordo com a consulta do usuário;
g) Opção de busca facetada, agrupamento dinâmico e pesquisa geoespacial;
h) Fácil integração com banco de dados.

Além disso, o *Solr*: possui *Application Programming Interfaces* (APIs) que facilitam o uso de diferentes linguagens de programação; permite configuração externa, proporcionando adaptação a aplicativos sem necessidade de codificação; e apresenta vasta arquitetura de *plug-ins* em casos de personalização avançada (CORREIA, 2016; SOLR, 2018).

Basicamente, o *Solr* usa o *Apache Lucene* para fornecer estruturas de dados, indexar documentos e realizar pesquisas. A unidade básica de informação do *Solr* é o documento, um conjunto de dados que descreve algo. Esses documentos são compostos por campos, que são informações mais específicas e podem conter diferentes tipos de dados (SOLR, 2018).

Segundo Grainger e Potter (2014), o *Solr* funciona da seguinte maneira: executa análise textual dos documentos e consultas para identificar palavras textualmente semelhantes; entende e combina sinônimos; remove *stopwords*; e pontua cada resultado com base em quão bem ele corresponde à consulta realizada pelo usuário. Para indexação, a plataforma cria o índice invertido, conforme descrito na seção 2.2.

O ranqueamento é feito por meio do cálculo da pontuação de relevância para cada documento e, em seguida, os resultados da pesquisa são classificados de acordo com a pontuação, da maior para a menor. Tal pontuação é baseada na verificação de similaridade, com uso dos modelos booleano e vetorial. Primeiramente é utilizado o modelo booleano, para filtrar todos os documentos que não correspondem à consulta feita pelo usuário. Em seguida, é feita aplicação do modelo vetorial para pontuar e desenhar a consulta como um vetor, criando também um vetor adicional para cada documento. Basicamente, quanto mais próximos os dois vetores, maior a similaridade entre eles (GRAINGER e POTTER, 2014).

2.6 Mineração de Textos – MT

Segundo Wives (2004) e Ferneda (2003), a descoberta e análise de grupos textuais são processos importantes para estruturação, organização e recuperação de informações. Isso porque os indivíduos coletam e armazenam uma elevada quantidade de dados textuais, que precisam ser estudados, conhecidos e organizados de forma a fornecer informações que lhes deem conhecimento para a execução de uma tarefa. Rezende, Marcacini e Moura (2011) complementam ao afirmarem que o processo de busca e recuperação da informação pode ser agilizado quando as coleções textuais são organizadas de maneira inteligente. É neste contexto que a Mineração de Textos (MT) está inserida, auxiliando nos processos de RI, extração de dados, resumo de documentos, descoberta de padrões, associações e regras, e realização de análises em documentos de texto.

Conforme Carrilho Jr. (2007, p. 12), o principal objetivo de se minerar textos é "descobrir conhecimento novo e inovador a partir de massas de texto livre, isto é, na forma natural que conhecemos e lidamos diariamente, agregando valor comercial a empresas e organizações". Dessa forma, uma grande quantidade de textos não estruturados pode ser transformada em informação útil, a partir da aplicação de técnicas de MT como sumarização, clusterização e classificação, dentre outras (ARANHA E PASSOS, 2006). Nesta pesquisa será detalhada a técnica de sumarização, realizada pelo sumarizador PragmaSUM a partir da seleção de atributos do modelo Cassiopeia.

2.6.1 Sumarização automática – SA

Conforme mencionado, técnicas de MT, como a sumarização, podem auxiliar nos processos de RI. Segundo Batista Júnior (2006), se um sumário (resumo) é composto por um conjunto de palavras significativas, então pode ser uma boa fonte para a seleção de termos no processo de indexação. A sumarização trata da redução do conteúdo textual, produzindo sumários sem que haja perda de informatividade e sentido. Os elementos chave do texto são mantidos com o objetivo de adquirir ganhos em desempenho com relação à busca por informação útil. Assim, a sumarização pode ser expressa como a "tarefa de identificar o que é relevante no texto e, então, traçar o novo enredo, a partir do conteúdo disponível, preservando sua ideia central, sem transgredir o significado original pretendido" (RINO e PARDO, 2003, p.2; CARRILHO JÚNIOR, 2007).

É comum a utilização de sumarizadores automáticos para realização da sumarização. Eles são sistemas computacionais que têm como foco criar uma representação reduzida dos itens importantes de um texto, ou seja, sumários, para utilização por usuários humanos. Nesses sumarizadores, os algoritmos de sumarização podem fornecer taxa de compressão em diferentes percentuais. Essa taxa define a porcentagem ou tamanho do sumário em relação ao texto original. Por exemplo, um texto sumarizado a uma taxa de 60% dará origem a um sumário com tamanho referente a 40% do texto original (RINO e PARDO, 2003).

Segundo Guelpeli (2012), a SA reduz parte das *stopwords* e atributos pouco significativos existentes nos repositórios de informação, contribuindo com a diminuição da alta dimensionalidade e dados esparsos. De acordo com Wives (2001), a dimensionalidade é um conceito importante na RI, visto que cada documento, pertencente a uma coleção, possui palavras que o representam. Assim, "uma pequena coleção de textos pode facilmente conter milhares de termos, muitos deles redundantes e desnecessários, que tornam lento o processo de extração de conhecimento e prejudicam a qualidade dos resultados" (REZENDE, MARCACINI E MOURA, 2011, p. 9).

A coleção de documentos textuais, no espaço amostral[8], é representada por uma matriz documento-termo ($C = d \times t$), exibida na Figura 5. Cada linha retrata um documento (d), e cada coluna um termo (t) da coleção. A matriz contém os graus de similaridade entre todos os elementos de um conjunto de dados (MANNING *et al.,* 2009).

Figura 5 – Matriz de documento-termo

	t_1	t_2	⋯	t_M
d_1	a_{11}	a_{12}	⋯	a_{1M}
d_2	a_{21}	a_{22}	⋯	a_{2M}
⋮	⋮	⋮	⋱	⋮
d_N	a_{N1}	a_{N2}	⋯	a_{NM}

Fonte: Manning *et al.,* 2009 *apud* Guelpeli, 2012.

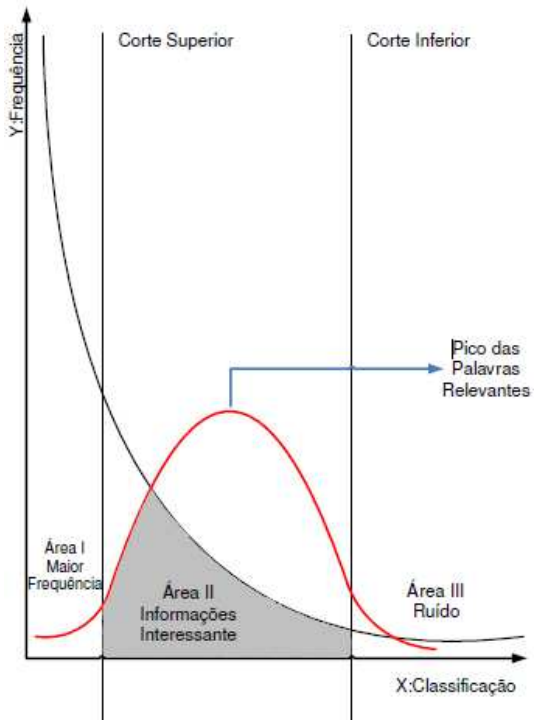
[8] Dado um experimento E, o espaço amostral é o conjunto de todos os resultados possíveis do experimento.

Desse modo:

> "[...] d_i corresponde ao i-*ésimo* documento, t_j representa o j-*ésimo* termo e a_{ij} é um valor que relaciona o i-*ésimo* documento com o j-*ésimo* termo, e pode ser calculado usando um determinado termo que está presente ou não, em um dado documento, ou mesmo um valor que indica a importância ou distribuição do termo ao longo da coleção de documentos" (Guelpeli, 2012, p. 37).

Segundo Beyer *et al.* (1999), a baixa dimensionalidade é necessária para manutenção da capacidade de distinção da palavra e, consequentemente, permite a redução do tempo de processamento. A técnica de redução da alta dimensionalidade e dados esparsos mais usual na literatura é o corte de Luhn (LUHN, 1958), exibido na Figura 6, que se baseia na Lei de Zipf[9] (ARANHA, 2007; GUELPELI, 2012).

Figura 6 – Curva de Zipf com os Cortes de Luhn



Fonte: Guelpeli, 2012.

Conforme pode ver verificado na Figura 6, Luhn propôs a definição de um limite superior e um limite inferior de corte, de forma que as palavras que estiverem fora do intervalo são eliminadas da análise. Assim, o primeiro corte objetiva eliminar *stopwords*, e o segundo

[9] A frequência de ocorrência de alguns eventos está relacionada à função de ordenação. Para o uso textual, ao somar a frequência de palavras e ordenar de forma decrescente, surge a Curva de Zipf. Nas ordenadas *Y* tem-se o valor dessa frequência, e nas abscissas *X,* o valor da posição de ordenação relativa da palavra, no qual a palavra que tem a maior frequência aparece na primeira posição (ROCHA, 2017, p. 30).

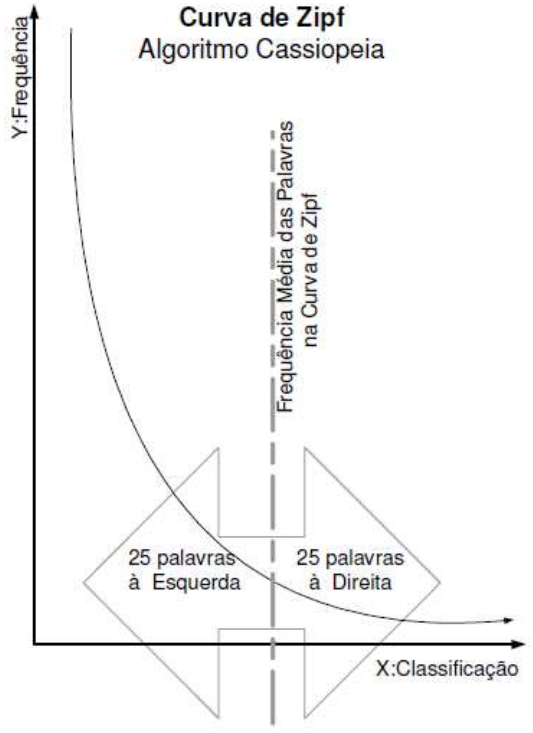
serve para diminuir o número de palavras muito específicas, encontradas apenas uma única vez nos documentos. Com isso tem-se três áreas: na área I, encontram-se as informações triviais ou básicas, com maior frequência; na área II, as informações interessantes; e, na área III, os ruídos (informações pouco importantes) (ARANHA, 2007; GUELPELI, 2012).

2.6.2 Seleção de atributos no modelo Cassiopeia

O modelo Cassiopeia é um agrupador de texto que apresenta um novo método para definição do corte de Luhn, ou seja, para seleção de atributos de um texto. Tal seleção permite reduzir a dimensionalidade, mantendo os atributos que possuem maior capacidade de representar a coleção de documentos (NOGUEIRA, 2009; GUELPELI, 2012).

Conforme Guelpeli (2012), o novo método de corte de Luhn propõe um corte médio na distribuição da frequência das palavras, como pode ser visualizado na Figura 7. Para identificação dos atributos, o modelo verifica as características das palavras no documento, utilizando a frequência relativa. Assim, a importância de uma palavra é definida conforme a frequência em que é encontrada no documento.

Figura 7 – Seleção dos atributos no modelo Cassiopeia



Fonte: Guelpeli, 2012.

A frequência relativa atribui pesos para as palavras, que são utilizados para calcular a média sobre o total de palavras do documento. A seleção dos atributos envolve o uso de um tamanho máximo de 50 posições para os vetores de palavras, realizando um corte que representa a frequência média das palavras obtidas com os cálculos e, em seguida, realiza a organização dos vetores de palavras. As 50 palavras do vetor, ordenadas em ordem decrescente, são divididas pelo modelo Cassiopeia, mantendo 25 posições do vetor à direita e 25 à esquerda da frequência média (GUELPELI, 2012). Dessa forma, o novo método para definição do corte de Luhn, é detalhado:

a) Calcular a frequência relativa: quantas vezes cada palavra aparece no documento, dividido pelo número total de palavras do documento;
b) Ordenar as palavras em ordem decrescente de frequência (da maior para a menor);
c) Achar a frequência relativa média das palavras, somando as frequências relativas e dividindo pelo número total de palavras do documento;
d) Encontrar a primeira palavra cuja frequência é mais próxima à média;
e) Marcar esta palavra e escolher, incluindo-a, mais as 24 anteriores (esquerda);
f) Marcar esta palavra e escolher as 25 posteriores (direita);
g) Montar o vetor em ordem decrescente com as 50 palavras escolhidas.

Desse modo, são eliminadas *stopwords*, palavras que tem maior frequência no texto e as que não tem tanta relevância. Mantem-se, então, somente as palavras existentes no pico de palavras relevantes.

2.6.3 PragmaSUM

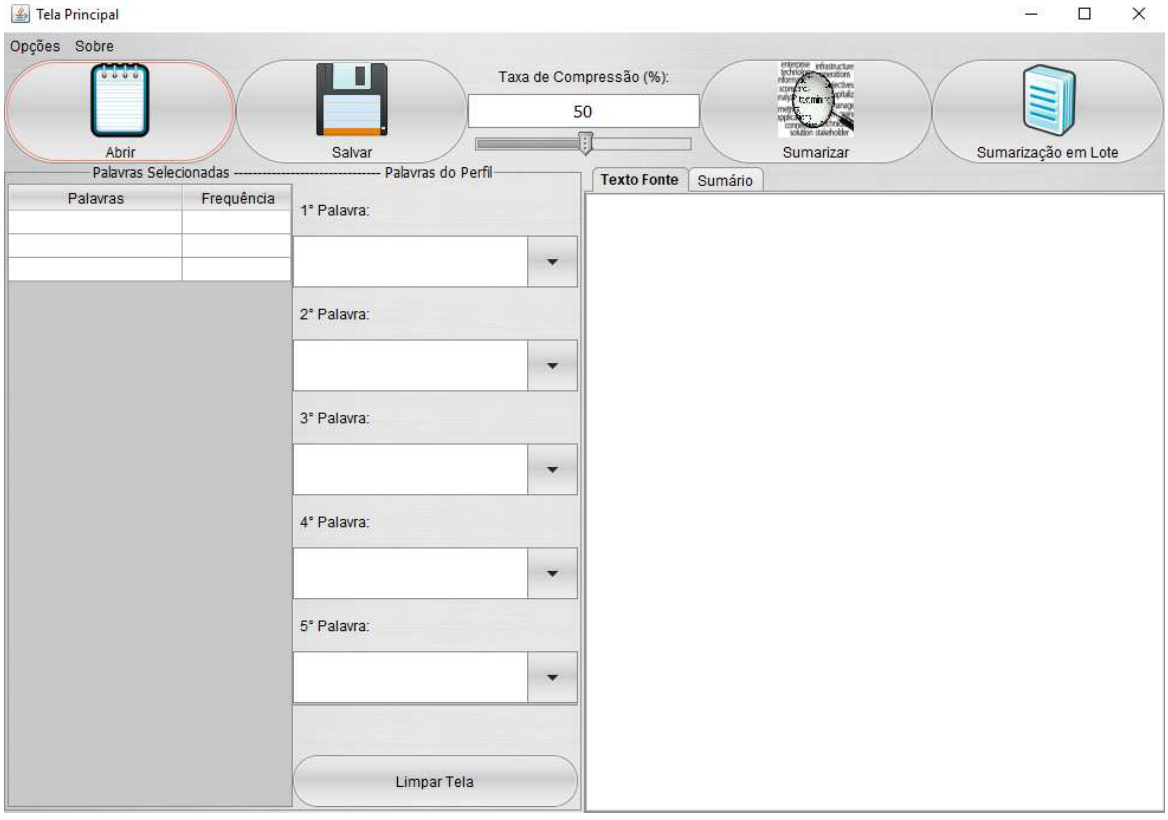
O PragmaSUM é um sumarizador automático de textos, desenvolvido por Rocha (2017), independente de idioma e domínio. Isso significa que não depende de intervenção humana e pode ser avaliado fora do domínio específico.

A sumarização no PragmaSUM compreende pré-processamento e processamento. Assim, dado um conjunto de textos no formato .txt, na fase de pré-processamento o PragmaSUM executa a limpeza dos textos, com o objetivo de reduzir a quantidade de palavras, preparando-os para o processamento computacional. Para obter informatividade, permitindo ganho qualitativo e quantitativo no processamento, o PragmaSUM utiliza a técnica de redução da alta dimensionalidade e dados esparsos (novo método para definição do corte de Luhn) proposta no modelo Cassiopeia, ou seja, o método de seleção de atributos (palavras) do modelo

(Figura 7). Para isso, o sumarizador calcula a média da frequência das palavras contidas no texto fonte e seleciona a palavra que possui a frequência mais próxima da média; as 24 palavras acima e as 25 abaixo da média são escolhidas, a partir de um corte com as 50 palavras que serão usadas na valoração dos itens do texto. Somente as palavras contidas no pico das palavras relevantes serão utilizadas (ROCHA e GUELPELI, 2017).

A etapa de processamento envolve a execução da sumarização que pode ocorrer individualmente ou em lote, conforme ilustra a Figura 8. Na primeira, apenas um texto fonte é sumarizado por vez, com a opção do uso ou não de palavras para personificação do sumário, seguindo as preferências do usuário. A segunda consiste na sumarização em lote, na qual grandes quantidades de textos podem ser sumarizadas de uma só vez (ROCHA e GUELPELI, 2017).

Figura 8 – Tela principal do PragmaSUM



Fonte: Rocha e Guelpeli, 2017.

Na sumarização individual o usuário abre o texto-fonte que deseja sumarizar, seleciona a taxa de compressão e as palavras do perfil[10], que podem ser inseridas de acordo com a escolha do usuário. Após acionar o botão "Sumarizar" o sumário é exibido na tela.

[10] São palavras que podem ser inseridas pelo usuário para valoração das frases.

A sumarização em lote (Figura 9) ocorre conforme descrito a seguir:

Figura 9 – PragmaSUM: sumarização em lote

Fonte: Rocha e Guelpeli, 2017.

a) As pastas de textos-fonte e de destino dos sumários são selecionadas nos campos "Caminho da Pasta de Textos Fonte" e "Caminho de destino dos Sumários", respectivamente;
b) A taxa de compressão é escolhida, de 0 a 100%;
c) A opção "Ativar Perfil na Sumarização" é marcada, caso a personificação da sumarização seja de interesse do usuário. Com a ativação do perfil, os seguintes campos são exibidos:

    - "Método de Valoração", que permite optar por um dos métodos: (01) Sequência – as palavras são escolhidas na ordem em que estão no texto fonte; (02) Classificação – as palavras são ordenadas de acordo com sua frequência no texto fonte (palavras mais frequentes apresentam maior pontuação); (03) TF-ISF –uso do algoritmo *Term Frequency* -

Inverse Sentence Frequency (TF-ISF) para classificação das sentenças (frases) do texto. Segundo Rocha (2017), o valor do TF-ISF determina que cada sentença tem uma pontuação associada, dada pelo valor de todas as suas palavras. Assim, ele é o critério para selecionar as frases que farão parte do sumário;

- "Quantidade de Palavras Chaves Utilizadas", no qual se define a quantidade de palavras chave a serem utilizadas na sumarização;
- "Caminho da Pasta de palavras Chave", no qual se seleciona a pasta que contém os arquivos de palavras-chave.

2.7 Trabalhos correlatos

Para maior conhecimento acerca do tema da pesquisa e análise dos estudos já realizados, alguns trabalhos relacionados à área de RI e sumarização foram verificados. No primeiro deles, Araújo Júnior e Tarapanoff (2006) trataram da precisão no processo de busca e recuperação da informação com o uso da mineração de textos. Para isso, foi feita a comparação entre indexação manual e uma ferramenta de mineração de textos, por meio da análise do índice de precisão de resposta no processo de busca e recuperação da informação. Os autores verificaram que não há ganho significativo na precisão ao se aplicar a ferramenta de mineração de textos em relação à indexação manual.

No mesmo ano, Batista Júnior (2006) investigou a aplicação de técnicas de sumarização automática na recuperação de informação. O objetivo foi verificar a contribuição dos extratos gerados para as etapas de indexação e realimentação de pseudo-relevantes da RI. Os resultados mostraram que os extratos gerados não foram úteis para a indexação.

Em 2009, Ferneda (2009) apresentou uma forma de aplicação dos algoritmos genéticos em sistemas recuperação de informação. As possíveis representações de um mesmo documento são consideradas um tipo de "código genético" desse documento e as buscas realizadas pelos usuários são consideradas o "meio ambiente" no qual os documentos estão inseridos. De acordo com o autor, os experimentos geraram resultados promissores na aplicação de algoritmos genéticos na recuperação de informação na *Web*, apresentando uma possibilidade para futuras implementações de sistemas com características evolutivas.

Botelho (2011) realizou um estudo de caso para avaliar a recuperação de informações em sistemas *online.* Para isso, considerou as expectativas dos usuários, a confiabilidade e atualidade do sistema; as características da informação, e os fatores

quantitativos e qualitativos da demanda. A pesquisa indicou que a forma básica das informações é flexível, de forma que para cada um dos arquivos existentes é possível a indexação por mais de um dicionário.

Em sua tese, Guelpeli (2012) contribuiu com a área de recuperação de informação ao propor o uso da técnica de agrupamento de documentos como apoio à busca e recuperação de textos em grandes bases textuais. A hipótese explorada consistiu na ideia de que seria possível melhorar o desempenho dos agrupadores de documentos, por meio da inclusão de sumarização, na fase de pré-processamento, e do uso do processo de agrupamento hierárquico, na fase de processamento. Para avaliar tal hipótese, o modelo Cassiopeia foi desenvolvido, integrando uma proposta poliestruturada na pesquisa.

Brito (2015) apresentou uma proposta de modelo de recuperação da informação baseado em conteúdo de arquivos de legendas de filmes e séries, utilizando *Apache Lucene* para recuperação da informação e a ferramenta OGMA (nome dado em homenagem ao Deus Celta Ogma), para análise de textos. Com o trabalho foi possível propor um modelo para pesquisa utilizando palavras-chaves, a classificação de filmes e séries por gênero e a descoberta por títulos parecidos.

Em 2017, Monteiro *et al.*, (2017) estudaram sobre os sistemas de recuperação de informação e o conceito de relevância nos mecanismos de busca, com enfoque em semântica e significação. No trabalho, verificou-se que a relevância, no contexto dos mecanismos de busca, engloba, especialmente na relação sistema-usuário, a coleção de entidades descritas, a personalização e a contextualização.

Silva (2018) implementou um sistema de recuperação de informação para uso corporativo, desenvolveu um *framework* de RI com arquitetura para expansão de novos métodos, e analisou as técnicas de uso de tesauro e a teoria de Semântica Distribucional para construir uma Análise de Contexto Local (ACL). Segundo o autor, os resultados dos experimentos validaram a abordagem e mostram-se com uma performance competitiva e qualificada para as soluções geradas.

Bandim e Correa (2019) realizaram um experimento que consistiu na aplicação do processo de indexação automática em um *corpus* formado por 60 artigos científicos, e posterior avaliação da qualidade na indexação, verificando os índices de consistência, precisão, revocação e medida-F. No processo foram utilizados o Tesauro Brasileiro em Ciência da Informação e o *software* SISA (Sistema de Indexação Semiautomático). Os resultados mostraram a influência do Tesauro nos resultados de indexação, ainda que as relações de termo geral pouco contribuíram para a qualidade na indexação automática.

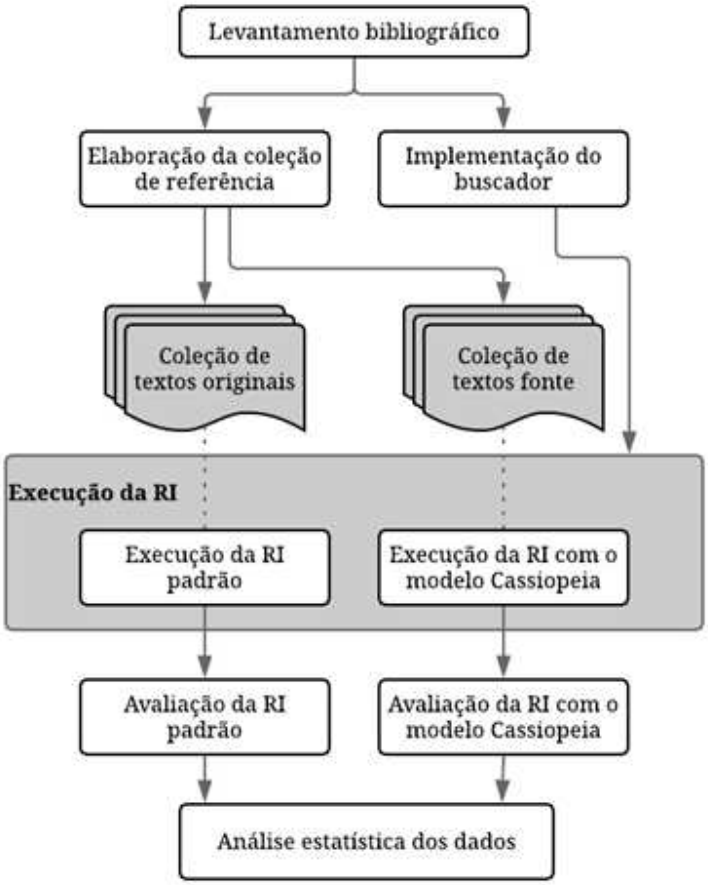
A partir dos trabalhos correlatos supracitados e da revisão de literatura realizada, verificou-se que esta pesquisa apresenta características que fornecem contribuições para a área de RI. Isso pode ser aferido considerando que a proposta consistiu em aplicar a técnica de sumarização, a partir da seleção de atributos do modelo Cassiopeia, em textos acadêmicos disponibilizados em repositórios institucionais, como contribuição para solução dos problemas relacionados à recuperação de informação. O trabalho ainda possui a especificidade de oferecer duas ferramentas para fins de estudo e pesquisa na área de RI: um *corpus*, em português, com julgamentos de relevância; e um buscador para testes e avaliação da RI.

3 METODOLOGIA

Este capítulo descreve a metodologia utilizada para realização do estudo. A pesquisa pretende gerar conhecimentos para aplicação prática, dirigidos à solução de problemas específicos, logo, é classificada, quanto à natureza, como aplicada. Quanto aos objetivos e procedimentos, trata-se de uma pesquisa exploratória e bibliográfica. Exploratória por proporcionar maior familiaridade com o problema, com vistas a torná-lo mais explícito; e bibliográfica, uma vez que implicou no estudo teórico de artigos, livros, teses e dissertações, para enriquecer os conhecimentos acerca do tema. A abordagem para tratamento da coleta de dados é quantitativa, visto que as métricas utilizadas para avaliar a recuperação de informação geram dados numéricos que foram tabulados e analisados (GIL, 2002).

Para melhor compreensão, a Figura 10 exibe as etapas metodológicas executadas e que serão detalhadas a seguir.

Figura 10 – Estrutura metodológica



3.1 Levantamento Bibliográfico

A primeira etapa, “Levantamento Bibliográfico”, consistiu na catalogação e leitura de artigos, dissertações, teses e livros, para melhor compreensão acerca dos temas envolvidos na pesquisa.

3.2 Elaboração da coleção de referência

A fase de “Elaboração da coleção de referência”, ou seja, elaboração do teste para RI, envolveu as seguintes atividades: preparação da coleção de documentos, determinação das coleções utilizadas na pesquisa, e definição das consultas e realização dos julgamentos de relevância.

3.2.1 Preparação da coleção de documentos

Segundo Carrilho Júnior (2007), em ambientes de pesquisa e desenvolvimento, é comum a utilização de *corpus* previamente coletados e preparados. Neste trabalho, foi utilizada parte do *corpus* construído por Aguiar, Rocha e Guelpeli (2017), que é formado por 500 artigos científicos, em português, separados em dez áreas de conhecimento do domínio Educacional. Tais áreas pertencem à grande área de Educação e podem ser visualizadas na tabela da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES[11]): Educação Especial; Educação Permanente; Educação Pré-escolar; Ensino-aprendizagem; Filosofia da Educação; História da Educação; Política Educacional; Psicologia Educacional; Sociologia da Educação e Tecnologia Educacional.

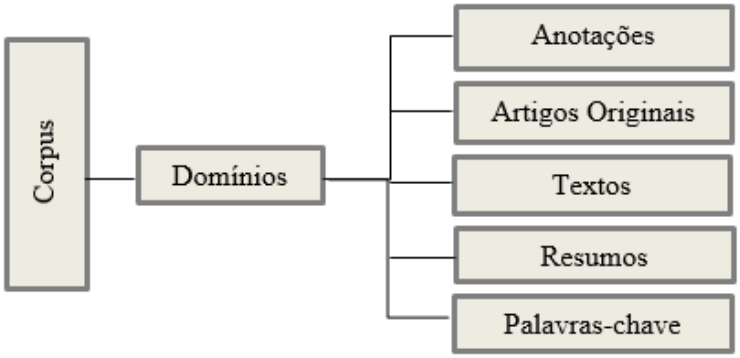
Cada domínio educacional do *corpus* possui 50 textos e cinco diretórios (Figura 11), organizados por Aguiar, Rocha e Guelpeli (2017) da seguinte maneira:

a) Anotações – possui arquivos (.txt) com as estatísticas e referências externas do texto;
b) Artigos Originais – onde estão armazenados os artigos originais em formato (.pdf);
c) Textos – composto pelos corpos dos textos em formato (.txt), ou seja, o texto fonte dos artigos;

[11] Disponível em: http://www.capes.gov.br/avaliacao/instrumentos-de-apoio/tabela-de-areas-do-conhecimento-avaliacao.

d) Resumos – contém os resumos manuais dos artigos, em formato (.txt);
e) Palavras-chave – possui as palavras-chave definidas pelos autores dos artigos, em formato (.txt).

Figura 11 – Diagrama simplificado do corpus Educacional



Fonte: Rocha, 2017. Adaptado.

Para facilitar a identificação, os arquivos existentes nos diretórios "Textos", "Resumos" e "Palavras-chave" foram renomeados utilizando a ferramenta *Advanced Renamer*[12], da seguinte forma:

a) "Textos": iniciais do domínio, seguido da palavra artigo e do número correspondente à ordem na pasta;
b) "Resumos": iniciais do domínio, seguido da palavra "resumo" e do número correspondente à ordem na pasta;
c) "Palavras-chave": iniciais do domínio, seguido da palavra "chave" e do número correspondente à ordem na pasta.

Por exemplo, se o artigo 01 do domínio "Educação Especial" possui a chave 01 e o resumo 01, então os arquivos receberam os seguintes nomes: EEArtigo01, EEchave01 e EEresumo01, respectivamente.

[12] *Advanced Rename*, utilitário de renomeação de arquivos em lote. Disponível em: https://www.advancedrenamer.com/

3.2.2 Coleções utilizadas na pesquisa

Segundo Silva R. e Silva E. (2013), não há consenso sobre o tamanho mínimo para que um *corpus* seja definido como representativo. Logo, para esta pesquisa foi criado um *corpus* menor, a partir do *corpus* Educacional desenvolvido por Aguiar, Rocha e Guelpeli (2017). A redução do tamanho do *corpus* foi realizada devido à necessidade de efetuar o julgamento de relevância dos documentos. Desse modo, foram escolhidos, aleatoriamente, 30 artigos de cada um dos dez domínios educacionais, formando uma coleção principal composta por 300 textos acadêmicos a serem utilizados para os testes de RI. A partir da coleção central, as seguintes coleções foram definidas:

a) Coleção de textos originais – base textual para "Execução da RI Padrão", ou seja, seguindo o modelo tradicional de RI. Composta por 300 artigos originais e completos, contidos no diretório "Artigos Originais";
b) Coleção de textos fonte – base textual para "Execução da RI com o modelo Cassiopeia", formada por 300 textos fonte que passaram pela fase de captura e manipulação. Tal etapa envolveu a retirada de bibliografia, tabelas, notas de rodapé, figuras, números de páginas e gráficos, restando somente o corpo textual dos artigos, ou seja, o texto fonte em formato (.txt). Nessa coleção foram aplicadas as sumarizações, efetuadas com diferentes taxas de compressão e quantidade de palavras-chave, com uso do sumarizador PragmaSUM, que gerou 16 diferentes representações para a coleção, como será detalhado na seção 3.4.2.1.

Neste trabalho, as estatísticas do *corpus* de textos fonte (coleção de textos fonte), exibidas na Tabela 2, foram coletadas com o uso da ferramenta *Finecount 3.0 Free*[13]. Cada um dos dez domínios educacionais (EE – Educação Especial, EP – Educação Permanente, EPE – Educação Pré-escolar, EA - Ensino-aprendizagem, FE – Filosofia da Educação, HE – História da Educação, POE – Política Educacional, PE – Psicologia Educacional, SE – Sociologia da Educação e TE – Tecnologia Educacional) apresenta 30 textos fonte.

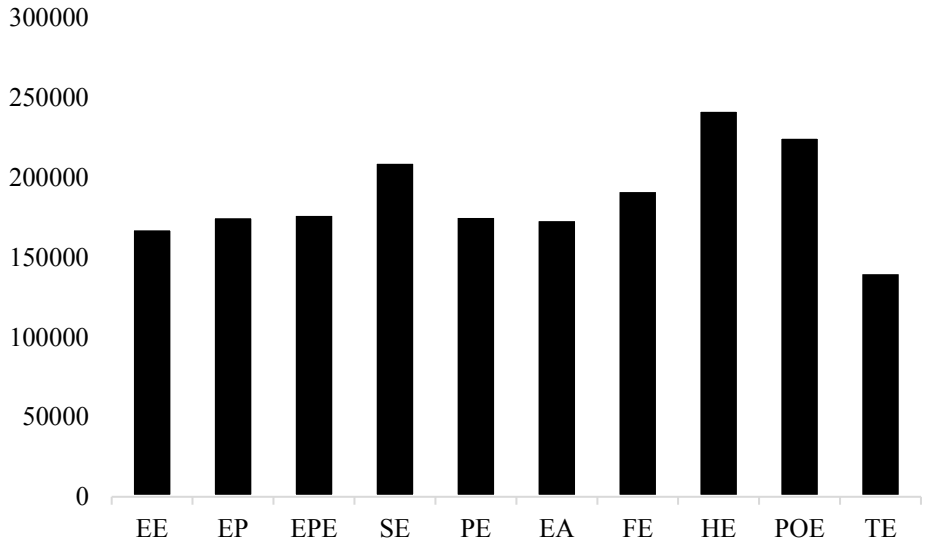
[13] *Software Finecount 3.0 Free*. Disponível em: http://www.finecount.eu/download/

Tabela 2 – Estatísticas da coleção de textos fonte do *corpus*

Arquivos	Caracteres	Caracteres e espaços	Palavras	Palavras e numerais	Média de palavras por texto
EE	829.492	985.243	164.318	168.103	5.603,44
EP	859.916	1.021.986	173.138	175.695	5.771,27
EPE	894.128	1.067.817	174.110	177.137	5.803,67
SE	1.034.710	1.228.778	206.725	209.788	6.890,84
PE	879.568	1.039.608	172.508	175.853	5.750,27
EA	856.996	1.017.458	171.005	173.890	5.700,17
FE	914.476	1.091.432	189.968	192.081	6.332,27
HE	1.162.045	1.381.725	236.895	242.328	7.896,50
POE	1.094.113	1.297.605	221.154	225.460	7.371,80
TE	689.062	818.126	137.496	140.660	4.583,20
Total	9.214.506	10.949.778	1.847.317	1.880.995	6.157,72
Desvio padrão	138.835,13	165000,37	29.423,58	29.998,30	971,83
Média geral	921.450,6	1094977,8	184.731,7	188.099,5	6.170,34

Considerando que, segundo Aluísio e Almeida (2006), o balanceamento de um *corpus* é importante para validar e proporcionar confiabilidade, a Figura 12 exibe a porcentagem de palavras e numerais, existentes em cada um dos domínios, em relação ao total de palavras e numerais do *corpus* (1.880.995). É possível perceber que os domínios possuem pequena variação na quantidade de palavras e numerais: História da Educação, Política Educacional e Sociologia da Educação apresentam a maior quantidade, enquanto Tecnologia Educacional a menor.

Figura 12 – Quantidade de palavras e numerais do *corpus*



3.2.3 Definição das consultas e julgamentos de relevância

Para realização das buscas na *interface* do SRI, foram definidas dez consultas a partir dos domínios aos quais pertencem os documentos do *corpus.* As expressões de busca escolhidas foram: “educação especial”, “educação permanente”, “educação pré-escolar”, “sociologia da educação”, “psicologia educacional”, “ensino aprendizagem”, “filosofia da educação”, “história da educação”, “política educacional” e “tecnologia educacional”.

Feito isso, iniciou-se o julgamento de relevância dos 300 textos, por domínio, ou seja: cada um dos 30 documentos, pertencentes a um domínio específico, foram classificados como relevantes ou irrelevantes em relação à consulta correspondente. Tal julgamento foi realizado a partir da leitura, por *skimming*[14], dos documentos do domínio. Assim, para cada uma das dez consultas foram selecionados quinze documentos relevantes.

Como exemplo, os 30 documentos do domínio Educação Especial foram definidos como relevantes ou irrelevantes em relação à consulta “educação especial”, como ilustra a Tabela 3.

Tabela 3 – Exemplo de julgamento de relevância dos documentos em relação à consulta

Consulta/Domínio	Documento	Julgamento de relevância (0 ou 1)
“educação especial”	EEArtigo01	0
“educação especial”	EEArtigo02	1
...	...	...
“educação especial”	EEArtigo30	1

Dessa forma, os documentos do domínio Educação Especial (consulta “educação especial”) foram classificados em relevantes (1) e irrelevantes (0). No exemplo, os artigos “EEArtigo02” e “EEArtigo30” são relevantes para a consulta “educação especial”.

3.3 Implementação do buscador

Nesta etapa, foi feita a implementação do buscador, com o objetivo de fornecer um SRI para efetuar a recuperação de informação na coleção de textos acadêmicos do domínio educacional. O buscador é responsável pelos processos de busca, indexação, recuperação e ranqueamento dos documentos que serão retornados ao usuário.

[14] Segundo Marcone e Lakatos (2003), a leitura por *skimming* tem como objetivo captar a tendência geral do texto, valendo-se dos títulos, subtítulos, ilustrações, se houver; e da leitura de resumos e parágrafos, tentando encontrar a metodologia e a essência do trabalho.

Assim, optou-se pela utilização do *Apache Solr* versão 7.2.1. O *Solr* pode ser instalado em qualquer sistema operacional em que o *Java Runtime Environment* (JRE) versão 1.8 ou superior esteja disponível (SOLR, 2018). Segundo Correia (2016), o *Apache Solr* é amplamente utilizado devido à extensa e atualizada documentação disponível. A quantidade de recursos avançados também é um fator relevante que fez com que empresas como a NASA, *Apple*, *Netflix* e *Disney* passassem a aplicar o *Solr* como ferramenta de busca (BUCHLER, 2018).

3.3.1 Instalação e Configuração do Apache Solr

A instalação do *Solr* requer que o pacote baixado no site *Apache Lucene* seja extraído em um diretório. A iniciação da ferramenta é feita por meio do comando "*solr start*" executado no *prompt (*cmd*)* do *Windows*. Isso iniciará o *Solr*, em segundo plano, na porta 8983. O console de administração pode ser acessado localmente no endereço http://localhost:8983/solr.

Os principais arquivos de configuração do *Solr* são (GRAINGER e POTTER, 2014; CORREIA, 2016):

a) solr.xml – define propriedades relacionadas à administração, *log*, fragmentação, dentre outras;
b) solrconfig.xml – apresenta as principais configurações do *Solr*. Nele os parâmetros ligados à pesquisa foram definidos, como por exemplo, em quais campos serão feitas as consultas e qual o tamanho do texto de resultados que será apresentado na tela;
c) schema.xml – define a estrutura dos dados que serão indexados e armazenados, bem como índice, campos e tipos de campo. O arquivo foi configurado para: transformar letras maiúsculas em minúsculas; dividir o texto em *tokens*, ou seja, em palavras, tratando os espaços em branco; executar *stemming;* e remover *stopwords* e itens duplicados. O parâmetro referente a sinônimos dos termos foi definido somente para a pesquisa e não para a indexação, para não haver sobrecarga com palavras sinônimas;
d) data-config.xml – responsável pelas *queries.*

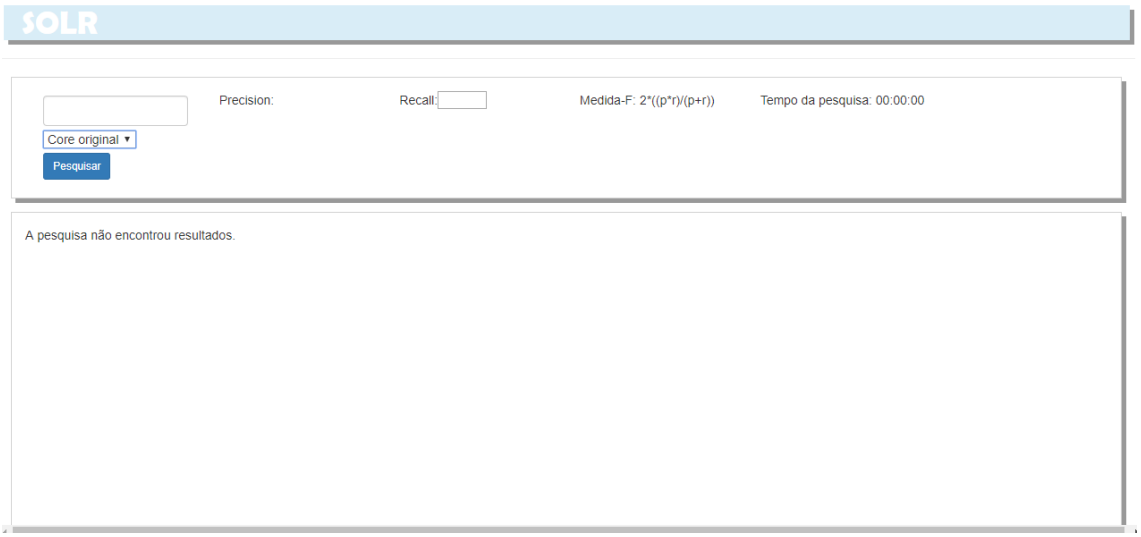
O último passo de configuração foi a criação de núcleos ou *cores*, para indexação e pesquisa. Cada *core* armazena uma coleção específica com os documentos a serem pesquisados e recuperados pelo buscador. Foram criados dois *cores*: "*Core* original", para a "Coleção de

textos originais", ou seja, de textos completos, e execução da RI padrão; e "*Core*_sum", para as variações da "Coleção de textos fonte" (execução da RI com o modelo Cassiopeia).

Para facilitar a visualização, execução e avaliação da RI, uma *interface* (Figura 13) vinculada ao *Solr* foi desenvolvida com os seguintes componentes:

a) Caixa de texto: campo de busca para realização da consulta;
b) Botão de pesquisa: executa a pesquisa inserida na caixa de texto;
c) Campo de seleção do núcleo ou *core*: no qual a coleção textual desejada deve ser selecionada;
d) Área de resultados: onde são listados os textos acadêmicos recuperados a partir da busca;
e) Área de cálculo automático das métricas *precision, recall e F-measure*: para cada consulta são apresentados os valores das métricas. Cada pesquisa realizada também gera um *log* em formato (.xls) para registro dos dados das simulações efetuadas, tais como: consultas, tempo de resposta (em segundos) e valor das métricas;
f) Paginação: é exibida caso a quantidade de itens recuperados exceda o limite de resultados determinados por página.

Figura 13 – Tela principal do buscador *Solr*



Assim, as soluções tecnológicas selecionadas para implementar a proposta foram: a linguagem de programação C# e o *software* livre *Microsoft Visual Studio Community* 2017[15],

[15] Visual Studio. Ferramenta de desenvolvimento. Disponível em: https://visualstudio.microsoft.com/pt-br/.

versão 15.7.4, que é um *Integrated Development Environment* (IDE) repleto de recursos gratuitos para estudantes e desenvolvedores individuais.

O SRI foi instalado em um computador com a seguinte configuração: processador *Intel*(R) *Core*(TM) i7-7500U CPU @2.70GHz, 2.90 GHz; 8 Gb de memória RAM; e sistema operacional de 64 bits, *Windows 10 Home Single Language*. Logo, as execuções e avaliações da RI, efetuadas no SRI e que geraram o registro dos tempos de resposta (em segundos) e dos valores das métricas, foram realizadas no mesmo equipamento e em condições básicas de controle, tais como: uso do mesmo navegador (*Google Chrome*[16]) e verificação de não ocorrência de processos desnecessários ao funcionamento básico do computador.

3.4 Execução da Recuperação de Informação – RI

Na etapa de "Execução da RI" o processo de recuperação de informação ocorreu de duas formas, aqui denominadas "Execução da RI padrão" e "Execução da RI com o modelo Cassiopeia". O objetivo foi analisar a proposta da pesquisa por meio de comparação com os processos de RI tradicional.

A principal diferença entre as execuções está relacionada às coleções obtidas na fase de "Elaboração da coleção de referência" (seção 3.2.2 – Coleções utilizadas na pesquisa), que devem ser selecionadas no campo de seleção do núcleo ou *core*. O mesmo buscador foi utilizado nas duas execuções, logo, os processos de busca, indexação e ranqueamento funcionam da mesma forma. É importante destacar que os testes foram executados, de maneira controlada, pelos pesquisadores, sem participação de usuários.

3.4.1 Execução da RI Padrão

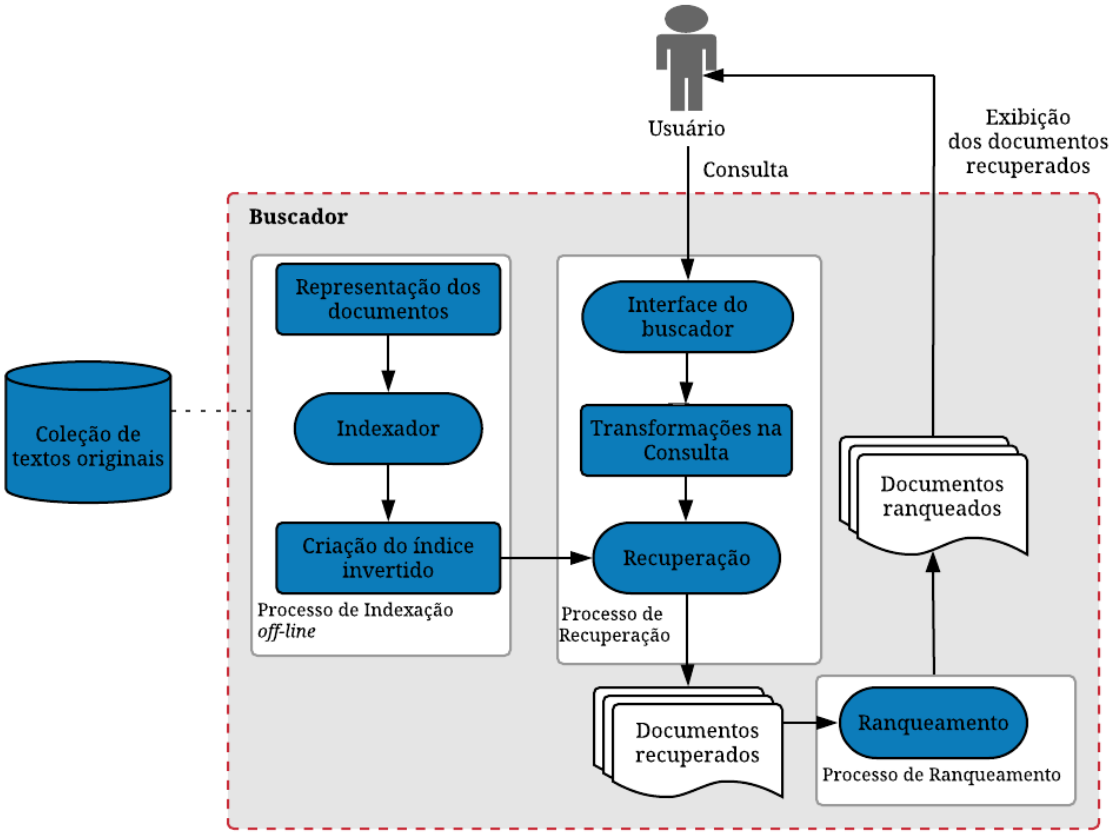
A "Execução da RI Padrão" refere-se à execução da recuperação de informação realizada pelos SRIs em geral, na qual não há aplicação de métodos de preparação, classificação ou mineração nos textos originais do *corpus*. Logo, tal execução seguiu os procedimentos tradicionais de RI, conforme exibido na Figura 14.

Para verificar a não existência de eventos aleatórios, primeiramente foram feitas 50 execuções da RI Padrão no "*Core* Original", ou seja, as dez consultas foram efetuadas 50 vezes cada, compondo um total de 500 execuções. O buscador apresentou o mesmo comportamento

[16] Disponível em: http://www.google.com/int/pt-BR/chrome

em cada uma delas, retornando os mesmos documentos, e manteve os valores de *precision*, *recall* e *F-measure,* o que prova a não existência de eventos aleatórios.

Figura 14 – Execução da RI Padrão



Fonte: Baeza-Yates e Ribeiro-Neto (2013). Adaptado.

A RI padrão é realizada conforme os passos apresentados a seguir:

a) A execução da RI é iniciada quando o usuário insere os termos da consulta no campo de pesquisa do buscador, para procurar documentos que atendam à sua necessidade de informação. O *core*, no qual serão pesquisados e recuperados os documentos, deve ser selecionado. Nesse caso, o "*Core* original" é escolhido. Todas as dez consultas foram realizadas na *interface* do buscador para proceder aos testes;

b) O *Solr* transforma a consulta removendo *stopwords* e efetuando correções ortográficas, e verifica a coleção de textos para criação da representação dos documentos para o indexador;

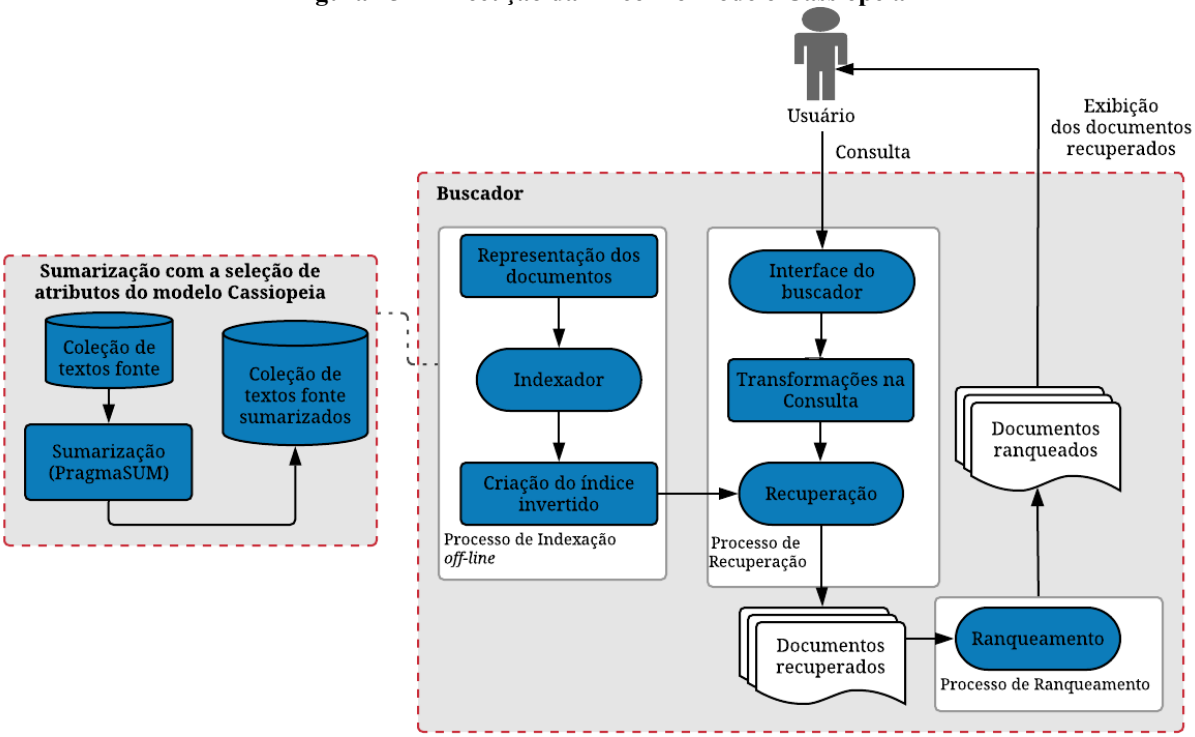
c) O índice invertido é criado e, após o processamento da busca, a recuperação é iniciada, retornando o conjunto de documentos recuperados;

d) Os textos acadêmicos retornados serão ranqueados por ordem de relevância, utilizando os modelos booleano e vetorial, propostos no *Solr*;
e) Por fim, os textos acadêmicos serão exibidos ao usuário, na área de resultados.

3.4.2 Execução da RI com o modelo Cassiopeia

Esta execução, aqui denominada "RI com o modelo Cassiopeia", refere-se à execução da RI na qual os textos fonte do *corpus* passaram pelo processo de sumarização, utilizando o método de seleção de atributos do modelo Cassiopeia, conforme Figura 15. O sumarizador que implementa tal método é o PragmaSUM. Assim, a coleção de textos fonte sumarizados é utilizada pelo buscador para a recuperação de informação.

Figura 15 – Execução da RI com o modelo Cassiopeia



A sumarização, utilizando o método de seleção de atributos do modelo Cassiopeia, foi adotada na pesquisa devido às possíveis contribuições para a área de RI. Conforme destacado por Guelpeli (2012), o método propõe uma forma de reduzir a elevada quantidade de atributos existentes em textos não estruturados (contidos em repositórios de informação), atenua a alta dimensionalidade e dá origem a sumários com boa informatividade.

3.4.2.1 Modelo Cassiopeia: sumarização com o PragmaSUM

Para implementação da proposta da pesquisa, os textos contidos na "Coleção de textos fonte" foram sumarizados pelo PragmaSUM, dando origem a um conjunto de textos sumarizados. A escolha do PragmaSUM se deve ao fato dele utilizar o método de seleção de atributos do modelo, ou seja, o algoritmo Cassiopeia, e ser o melhor sumarizador para o modelo, conforme testes realizados por Rocha (2017).

Foram executadas 16 sumarizações em lote, combinando taxas de compressão (50%, 70%, 80% e 90%) e quantidade de palavras-chave (três, quatro, cinco ou nenhuma), com os 300 textos fonte da "Coleção de textos fonte", o que deu origem a 16 coleções de textos sumarizados. Para cada sumarização os seguintes passos foram realizados:

a) Os 300 textos fonte foram inseridos no campo "Caminho da Pasta de Textos Fonte";
b) A pasta destino dos sumários foi definida no campo "Caminho de destino dos sumários";
c) A taxa de compressão foi escolhida (50%, 70%, 80% ou 90%);
d) A opção "Ativar Perfil na Sumarização" foi marcada, para que os campos "Método de Valoração", "Quantidade de Palavras Chaves Utilizadas" e "Caminho da Pasta de palavras Chave" fossem preenchidos:

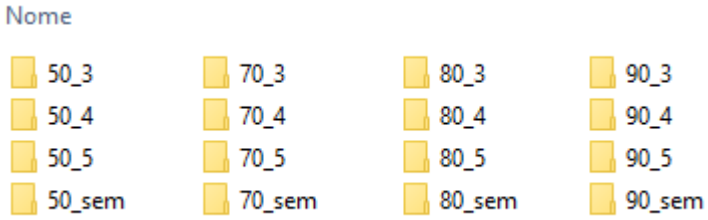
    - "Método de Valoração" escolhido: TF-ISF, por apresentar os melhores resultados para as sumarizações executadas por Rocha (2017), no mesmo *corpus*;

    - "Quantidade de Palavras Chaves Utilizadas": três, quatro, cinco e nenhuma, respectivamente;

    - "Caminho da Pasta de palavras Chave": uso dos 300 arquivos de palavras-chave, cada um correspondente a um texto, definidas pelos autores dos artigos, existentes nos diretórios "Palavras-chave", de cada um dos domínios.

Feito isso, bastou clicar no botão "Sumarizar" para que os sumários gerados fossem armazenados na pasta escolhida. Cada uma das 16 sumarizações em lote gerou um diretório com 300 arquivos de sumários, de todos os domínios, totalizando 16 pastas, como ilustra a Figura 16.

Figura 16 – Pastas com textos sumarizados



Cada pasta é uma coleção de textos acadêmicos sumarizados com uso de determinada taxa de compressão e quantidade de palavras-chave. Por exemplo, na pasta "50_3" encontram-se os textos sumarizados com taxa de compressão de 50% e uso de três palavras-chave. É nesses diretórios, e seus respectivos arquivos, que serão pesquisados e recuperados os documentos, de acordo com a consulta realizada no buscador.

3.4.2.2 Passos para execução da RI com o modelo Cassiopeia

O teste para identificar a ocorrência de evento aleatório também foi realizado na "RI com o modelo Cassiopeia". Foram feitas 50 execuções no "*Core*_sum", em cada uma das 16 coleções, ou seja, as dez consultas foram efetuadas 50 vezes, em cada uma das pastas, compondo um total de 8.000 execuções. O buscador apresentou o mesmo comportamento em cada uma delas, retornando os mesmos documentos, e manteve os valores de *precision*, *recall* e F-*measure,* o que prova a não existência de eventos aleatórios.

A execução da RI com o modelo Cassiopeia seguiu os passos da RI padrão (seção 3.4.1). Contudo, foram utilizadas as 16 coleções sumarizadas, ou seja, as dez consultas foram inseridas no SRI e feitas em cada uma das coleções. Para cada coleção foi necessário mudar a pasta no diretório do buscador (por exemplo, quando as dez buscas foram realizadas na coleção 50_3, essa pasta foi inserida no diretório). Feito isso, bastou selecionar o "*Core*_sum" no campo de seleção do núcleo ou *core* da *interface.*

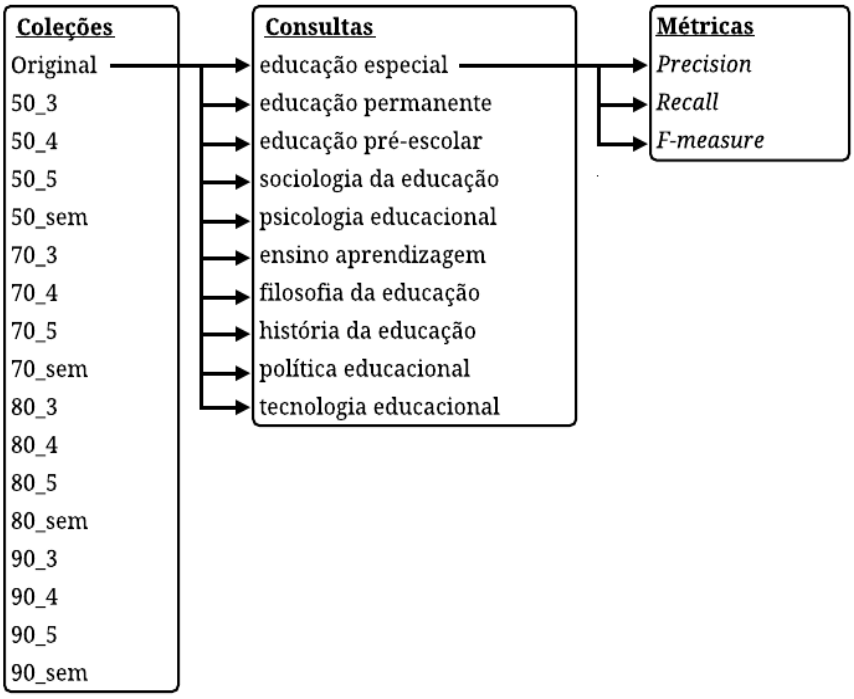
3.5 Avaliação da Recuperação de Informação – RI

As avaliações da RI padrão e com o modelo Cassiopeia foram realizadas com o uso das métricas *precision, recall e F-measure.* Segundo Baeza-Yates e Ribeiro-Neto (2013) e Manning, Raghavan e Schütze (2009), o processo de associar uma métrica numérica aos

resultados da consulta é amplamente utilizado pela simplicidade e possibilidade de repetição dos experimentos a custos baixos.

Os valores das métricas foram obtidos para cada expressão de busca inserida no SRI (“educação especial”, “educação permanente”, “educação pré-escolar”, “sociologia da educação”, “psicologia educacional”, “ensino aprendizagem”, “filosofia da educação”, “história da educação”, “política educacional”, “tecnologia educacional”), em cada coleção (uma original e 16 sumarizadas), conforme mostrado na Figura 17.

Figura 17 – Cálculo das métricas por coleção e consulta



Dessa forma, utilizando como exemplo a coleção “Original”, as dez consultas foram feitas na coleção de textos originais (300 artigos). Para cada expressão de busca, como “educação especial”, as métricas *precision, recall* e *F-measure* foram calculadas. O SRI exibe os valores das métricas na tela de resultados da busca e os registra em uma planilha de *log*.

3.6 Análise estatística dos dados

A análise estatística dos resultados, obtidos a partir de uma pesquisa, é uma importante ferramenta para validação de dados. Assim, testes estatísticos podem ser aplicados para comparar condições experimentais e proporcionar respaldo científico à pesquisa

(NORMANDO *et al.,* 2010). Segundo Callegari e Jacques (2007), os testes estatísticos são classificados em paramétricos e não paramétricos. Quando os dados apresentam distribuição conhecida os testes paramétricos devem ser utilizados. Nos testes não paramétricos não existe a necessidade de que a distribuição da variável na população seja conhecida.

Nesta pesquisa, a definição do teste estatístico adequado baseou-se no diagrama proposto por Callegari e Jacques (2007) (Figura 24, Anexo A). Considerando que os dados obtidos são amostras independentes, ordinais e apresentam distribuição anormal, foi possível identificar que os testes estatísticos não paramétricos, ANOVA de Friedman e coeficiente de concordância de Kendall, são os mais apropriados, uma vez que permitem analisar se as amostras de um experimento apresentam diferença significativa em sua distribuição (GUELPELI, 2012). Tais testes foram realizados com a utilização do *software* BioStat[17].

O ANOVA de Friedman compara e ordena os resultados de três ou mais amostras relacionadas e calcula a média das ordens para cada uma. Segundo Campos (2019, p. 58), o teste não "utiliza os dados numéricos diretamente, mas sim os postos ocupados por eles, após a ordenação por valores ascendentes desses dados. A ordenação numérica é feita separadamente em cada uma das amostras, e não em conjunto". Já o teste de coeficiente de concordância de Kendall tem o propósito de normalizar o ANOVA de Friedman e, de acordo com Viali (2008), pode ser útil em estudos de fidedignidade relativos a julgamentos. O teste verifica o grau de associação entre as variáveis e gera uma avaliação de concordância ou não com os *ranks* dos experimentos. Quanto mais próximo de zero, menor é a concordância, e quanto mais próximo de um, maior (CALLEGARI e JACQUES, 2007; GUELPELI, 2012).

Dessa forma, os dados obtidos a partir das execuções e avaliação das RIs, ou seja, os valores das métricas *precision, recall* e *F-measure*, foram tabulados e analisados estatisticamente. A análise estatística foi realizada entre: (1) as dez consultas efetuadas no SRI (graus de liberdade = 9); (2) as coleções "Original" e sumarizadas, por taxa de compressão aplicada (graus de liberdade = 4); e (3) todas as 17 coleções (graus de liberdade = 16). Primeiramente, foi feita a aplicação do teste ANOVA de Friedman com um nível de significância de 0,05 ($\alpha = 0,05$), para verificar se há 95% de probabilidade de que existe uma diferença significativa entre os experimentos (consultas ou coleções). Portanto, o objetivo é identificar se a precisão dos resultados retornados é igual para todas as consultas (análise por consulta) ou coleções (análise por coleção). O teste realiza as seguintes ações: produz um *rank* (r) entre os experimentos, ou seja, compara e ordena a métrica atribuindo valores do maior para

[17]Disponível em: http://www.mimaraua.org.br/downloads/programas

a menor; calcula a soma e média dos *ranks*, e a média, mediana e desvio padrão da métrica; efetua a análise de significância comparando os *ranks* dos experimentos (dois a dois); calcula a diferença entre eles; e, gera o valor-p (probabilidade condicional relacionada à significância de um resultado). Assim, se o valor-p $\leq \alpha$ as diferenças são estatisticamente significativas ($\leq$0,05), se valor-p $> \alpha$ as diferenças não são estatisticamente significativas (ns). Após a finalização do ANOVA de Friedman, foi realizada a verificação do coeficiente de concordância de Kendall: a partir do *rank* dos experimentos, o coeficiente de concordância (W) é calculado para avaliar a concordância entre os *ranks* produzidos.

4 RESULTADOS E DISCUSSÃO

Os resultados foram obtidos a partir da avaliação da RI, comparando a RI padrão, realizada na coleção de artigos originais e representada pela coluna “Original”, e a RI com o modelo Cassiopeia, executada em 16 coleções de artigos sumarizados. Devido à elevada quantidade de dados gerados (Apêndice A), e ao foco da pesquisa ser verificar a precisão dos resultados retornados, serão apresentados somente os resultados do *precision,* que variam de 0 (zero) a 1 (um): quanto mais próximos de 1, melhores; quanto mais próximos de 0, piores. Os dados serão exibidos de acordo com as comparações realizadas entre: as consultas efetuadas no SRI; as coleções, por taxa de compressão; e, todas as 17 coleções.

4.1 RI padrão e RI com o modelo Cassiopeia: comparação entre as consultas

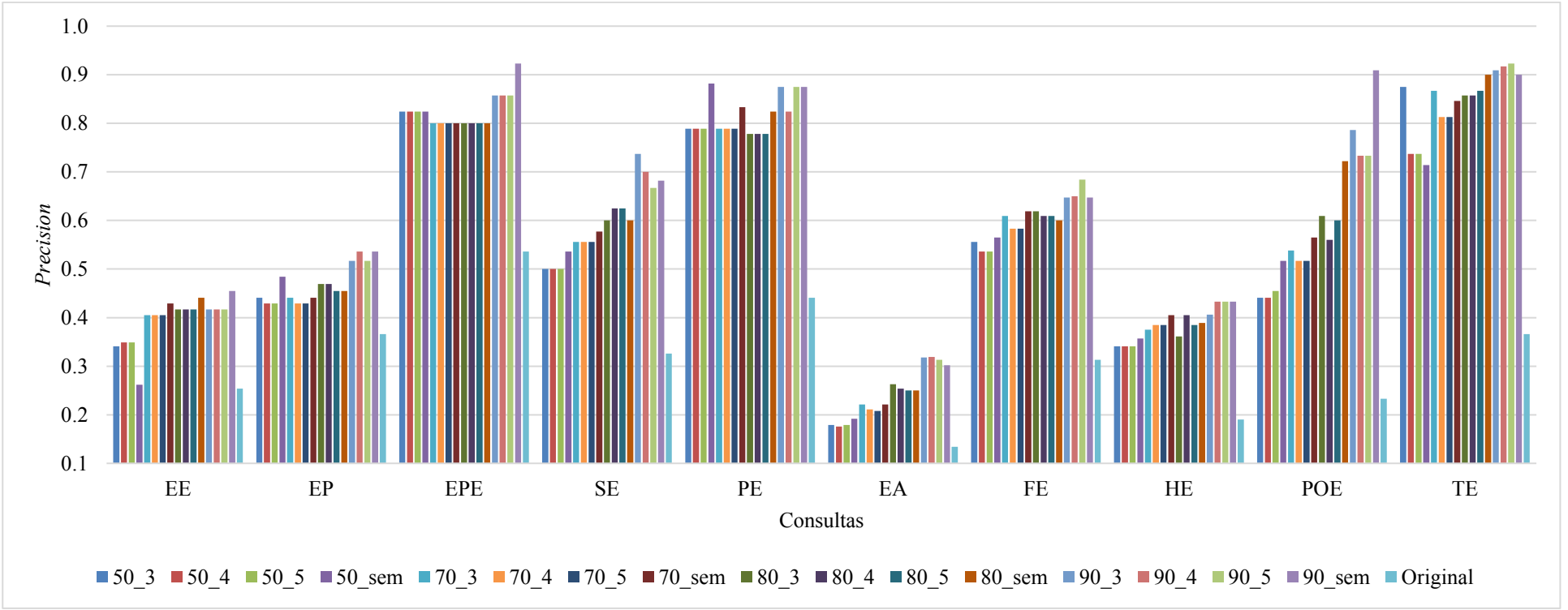
Para comparação entre as dez consultas realizadas nas coleções, a Tabela 4 exibe os valores (p) e *ranks* (r1 a r10 – ANOVA de Friedman) do *precision*, as médias de “p” e “r” e o coeficiente de concordância de Kendall (W).A Figura 18 ilustra os valores do *precision* para as consultas realizadas em todas as coleções.

A partir da análise dos dados por consulta, o coeficiente de concordância de Kendall (W) apontou que há 93% de certeza de que existe diferença significativa entre os experimentos (consultas). A expressão de busca “tecnologia educacional” apresentou os valores numéricos do *precision* e dos *ranks* (r1 a r10) mais elevados em 12 das 16 coleções sumarizadas, seguida de “educação pré-escolar” e “psicologia educacional” (Tabela 4). Isso significa que houve maior similaridade entre os termos destas consultas e os documentos que os contém, o que permitiu ao buscador retornar uma quantidade mais elevada de documentos relevantes, dentre os recuperados. A expressão “ensino aprendizagem” gerou as piores médias numéricas do *precision* e dos *ranks* para todas as coleções.

Tabela 4 – Valores (p) e *ranks* (r1 a r10 - ANOVA de Friedman) do *precision*, para as consultas efetuadas nas coleções

Consulta	**Educação especial (EE)**		**Educação permanente (EP)**		**Educação pré-escolar (EPE)**		**Sociologia da educação (SE)**		**Psicologia educacional (PE)**		**Ensino aprendizagem (EA)**		**Filosofia da educação (FE)**		**História da educação (HE)**		**Política educacional (POE)**		**Tecnologia educacional (TE)**	
Coleção	p	r1	p	r2	p	r3	p	r4	p	r5	p	r6	p	r7	p	r8	p	r9	p	r10
50_3	0,34	2,5	0,44	4,5	0,82	9	0,50	6	0,79	8	0,18	1	0,56	7	0,34	2,5	0,44	4,5	0,88	10
50_4	0,35	3	0,43	4	0,82	10	0,50	6	0,79	9	0,18	1	0,54	7	0,34	2	0,44	5	0,74	8
50_5	0,35	3	0,43	4	0,82	10	0,50	6	0,79	9	0,18	1	0,54	7	0,34	2	0,46	5	0,74	8
50_sem	0,26	2	0,48	4	0,82	9	0,54	6	0,88	10	0,19	1	0,57	7	0,36	3	0,52	5	0,71	8
70_3	0,41	3	0,44	4	0,80	9	0,56	6	0,79	8	0,22	1	0,61	7	0,38	2	0,54	5	0,87	10
70_4	0,41	3	0,43	4	0,80	9	0,56	6	0,79	8	0,21	1	0,58	7	0,39	2	0,52	5	0,81	10
70_5	0,41	3	0,43	4	0,80	9	0,56	6	0,79	8	0,21	1	0,58	7	0,39	2	0,52	5	0,81	10
70_sem	0,43	3	0,44	4	0,80	8	0,58	6	0,83	9	0,22	1	0,62	7	0,41	2	0,57	5	0,85	10
80_3	0,42	3	0,47	4	0,80	9	0,60	5	0,78	8	0,26	1	0,62	7	0,36	2	0,61	6	0,86	10
80_4	0,42	3	0,47	4	0,80	9	0,63	7	0,78	8	0,25	1	0,61	6	0,41	2	0,56	5	0,86	10
80_5	0,42	3	0,46	4	0,80	9	0,63	7	0,78	8	0,25	1	0,61	6	0,39	2	0,60	5	0,87	10
80_sem	0,44	3	0,46	4	0,80	8	0,60	5,5	0,82	9	0,25	1	0,60	5,5	0,39	2	0,72	7	0,90	10
90_3	0,42	3	0,52	4	0,86	8	0,74	6	0,88	9	0,32	1	0,65	5	0,41	2	0,79	7	0,91	10
90_4	0,42	2	0,54	4	0,86	9	0,70	6	0,82	8	0,32	1	0,65	5	0,43	3	0,73	7	0,92	10
90_5	0,42	2	0,52	4	0,86	8	0,67	5	0,88	9	0,31	1	0,68	6	0,43	3	0,73	7	0,92	10
90_sem	0,46	3	0,54	4	0,92	10	0,68	6	0,88	7	0,30	1	0,65	5	0,43	2	0,91	9	0,90	8
Original	0,25	4	0,37	7,5	0,54	10	0,33	6	0,44	9	0,13	1	0,31	5	0,19	2	0,23	3	0,37	7,5
Médias	0,39	2,85	0,46	4,24	0,81	9	0,58	5,97	0,79	8,47	0,23	1	0,59	6,26	0,38	2,21	0,58	5,61	0,82	9,38
Kendall - W	0,93																			

Figura 18 – Valores do *precision* para as consultas efetuadas nas coleções



4.2 RI padrão e RI com o modelo Cassiopeia: comparação entre as coleções, por taxa de compressão

Para verificação entre as coleções "Original" e sumarizadas (por taxa de compressão – 50, 70, 80 e 90%), a Tabela 5 mostra os valores (p) e *ranks* (r1 a r5 – ANOVA de Friedman) do *precision* (obtidos a partir das consultas realizadas nas coleções), as médias de "p" e "r" e o coeficiente de concordância de Kendall (W).

Tabela 5 – Valores (p) e *ranks* (r1 a r5) do *precision* para a coleção "Original" e as coleções sumarizadas por taxa de compressão (50%, 70%, 80%, 90%)

Coleção	**50_3**		**50_4**		**50_5**		**50_sem**		**Original**	
Consulta	p	r1	p	r2	p	r3	p	r4	p	r5
Educação especial (EE)	0,34	3	0,35	4,5	0,35	4,5	0,26	2	0,25	1
Educação permanente (EP)	0,44	4	0,43	2,5	0,43	2,5	0,48	5	0,37	1
Educação pré-escolar (EPE)	0,82	3,5	0,82	3,5	0,82	3,5	0,82	3,5	0,54	1
Sociologia da educação (SE)	0,50	3	0,50	3	0,50	3	0,54	5	0,33	1
Psicologia educacional (PE)	0,79	3	0,79	3	0,79	3	0,88	5	0,44	1
Ensino aprendizagem (EA)	0,18	3	0,18	3	0,18	3	0,19	5	0,13	1
Filosofia da educação (FE)	0,56	4	0,54	2,5	0,54	2,5	0,57	5	0,31	1
História da educação (HE)	0,34	3	0,34	3	0,34	3	0,36	5	0,19	1
Política educacional (POE)	0,44	2,5	0,44	2,5	0,46	4	0,52	5	0,23	1
Tecnologia educacional (TE)	0,88	5	0,74	3,5	0,74	3,5	0,71	2	0,37	1
Médias	0,53	3,4	0,51	3,1	0,52	3,25	0,53	4,25	0,32	1
Kendall - W	0,69									
Coleção	**70_3**		**70_4**		**70_5**		**70_sem**		**Original**	
Consulta	p	r1	p	r2	p	r3	p	r4	p	r5
Educação especial (EE)	0,41	3	0,41	3	0,41	3	0,43	5	0,25	1
Educação permanente (EP)	0,44	4,5	0,43	2,5	0,43	2,5	0,44	4,5	0,37	1
Educação pré-escolar (EPE)	0,80	3,5	0,80	3,5	0,80	3,5	0,80	3,5	0,54	1
Sociologia da educação (SE)	0,56	3	0,56	3	0,56	3	0,58	5	0,33	1
Psicologia educacional (PE)	0,79	3	0,79	3	0,79	3	0,83	5	0,44	1
Ensino aprendizagem (EA)	0,22	4,5	0,21	2,5	0,21	2,5	0,22	4,5	0,13	1
Filosofia da educação (FE)	0,61	4	0,58	2,5	0,58	2,5	0,62	5	0,31	1
História da educação (HE)	0,38	2	0,39	3,5	0,39	3,5	0,41	5	0,19	1
Política educacional (POE)	0,54	4	0,52	2,5	0,52	2,5	0,57	5	0,23	1
Tecnologia educacional (TE)	0,87	5	0,81	2,5	0,81	2,5	0,85	4	0,37	1
Médias	0,56	3,65	0,55	2,85	0,55	3	0,58	4,65	0,32	1
Kendall - W	0,85									

(continuação)

Coleção	**80_3**		**80_4**		**80_5**		**80_sem**		**Original**	
Consulta	p	r1	p	r2	p	r3	p	r4	p	r5
Educação especial (EE)	0,42	3	0,42	3	0,42	3	0,44	5	0,25	1
Educação permanente (EP)	0,47	4,5	0,47	4,5	0,46	2,5	0,46	2,5	0,37	1
Educação pré-escolar (EPE)	0,80	3,5	0,80	3,5	0,80	3,5	0,80	3,5	0,54	1
Sociologia da educação (SE)	0,60	2,5	0,63	4,5	0,63	4,5	0,60	2,5	0,33	1
Psicologia educacional (PE)	0,78	3	0,78	3	0,78	3	0,82	5	0,44	1
Ensino aprendizagem (EA)	0,26	5	0,25	3	0,25	3	0,25	3	0,13	1
Filosofia da educação (FE)	0,62	5	0,61	3,5	0,61	3,5	0,60	2	0,31	1
História da educação (HE)	0,36	2	0,41	5	0,39	3,5	0,39	3,5	0,19	1
Política educacional (POE)	0,61	4	0,56	2	0,60	3	0,72	5	0,23	1
Tecnologia educacional (TE)	0,86	2,5	0,86	2,5	0,87	4	0,90	5	0,37	1
Médias	0,58	3,5	0,58	3,45	0,58	3	0,60	3,7	0,32	1
Kendall - W	0,59									
Coleção	**90_3**		**90_4**		**90_5**		**90_sem**		**Original**	
Consulta	p	r1	p	r2	p	r3	p	r4	p	r5
Educação especial (EE)	0,42	3	0,42	3	0,42	3	0,46	5	0,25	1
Educação permanente (EP)	0,52	2,5	0,54	4,5	0,52	2,5	0,54	4,5	0,37	1
Educação pré-escolar (EPE)	0,86	3	0,86	3	0,86	3	0,92	5	0,54	1
Sociologia da educação (SE)	0,74	5	0,70	4	0,67	2	0,68	3	0,33	1
Psicologia educacional (PE)	0,88	4	0,82	2	0,88	4	0,88	4	0,44	1
Ensino aprendizagem (EA)	0,32	4,5	0,32	4,5	0,31	3	0,30	2	0,13	1
Filosofia da educação (FE)	0,65	3	0,65	3	0,68	5	0,65	3	0,31	1
História da educação (HE)	0,41	2	0,43	4	0,43	4	0,43	4	0,19	1
Política educacional (POE)	0,79	4	0,73	2,5	0,73	2,5	0,91	5	0,23	1
Tecnologia educacional (TE)	0,91	3	0,92	4,5	0,92	4,5	0,90	2	0,37	1
Médias	0,65	3,4	0,64	3,5	0,64	3,35	0,67	3,75	0,32	1
Kendall - W	0,58									

A Figura 19 apresenta as médias dos valores do *precision* para todas as consultas realizadas nas coleções “Original” e nas coleções sumarizadas (por taxa de compressão). A Figura 20 ilustra as médias dos *ranks* (ANOVA de Friedman) do *precision.*

Figura 19 – Médias dos valores do *precision* (por taxa de compressão) para todas as consultas realizadas na coleção "Original" e nas coleções sumarizadas com taxas de compressão de 50% (A), 70% (B), 80% (C) e 90% (D)

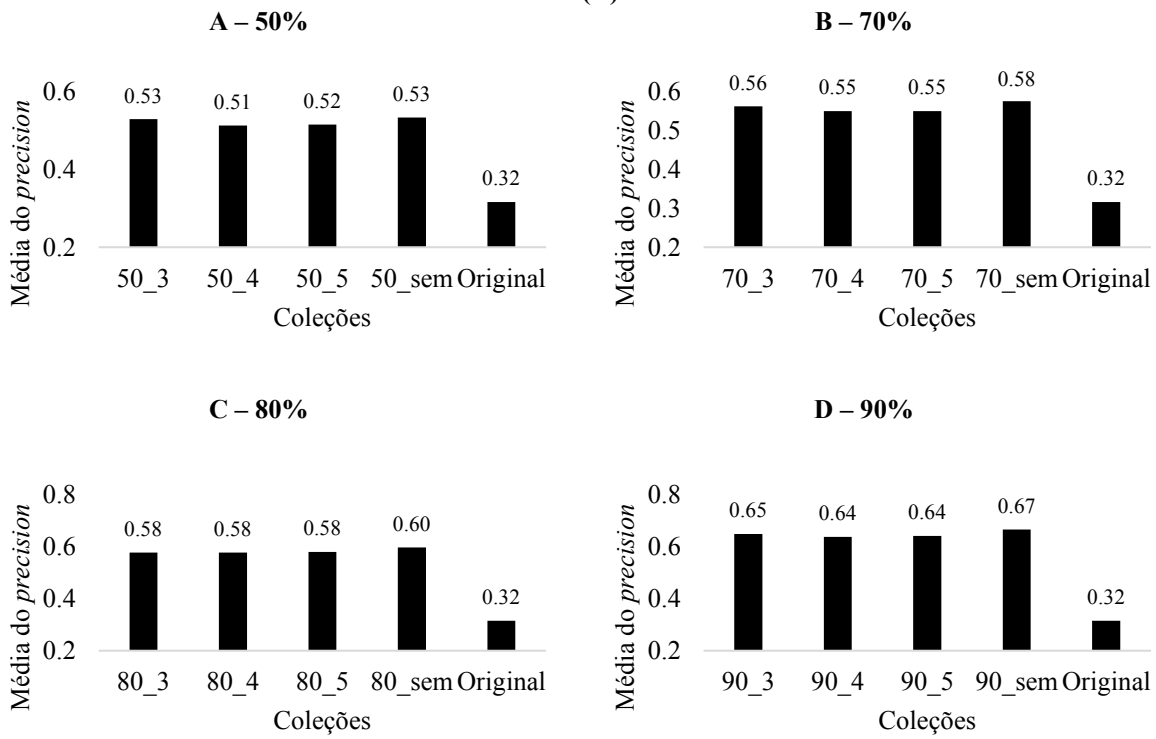
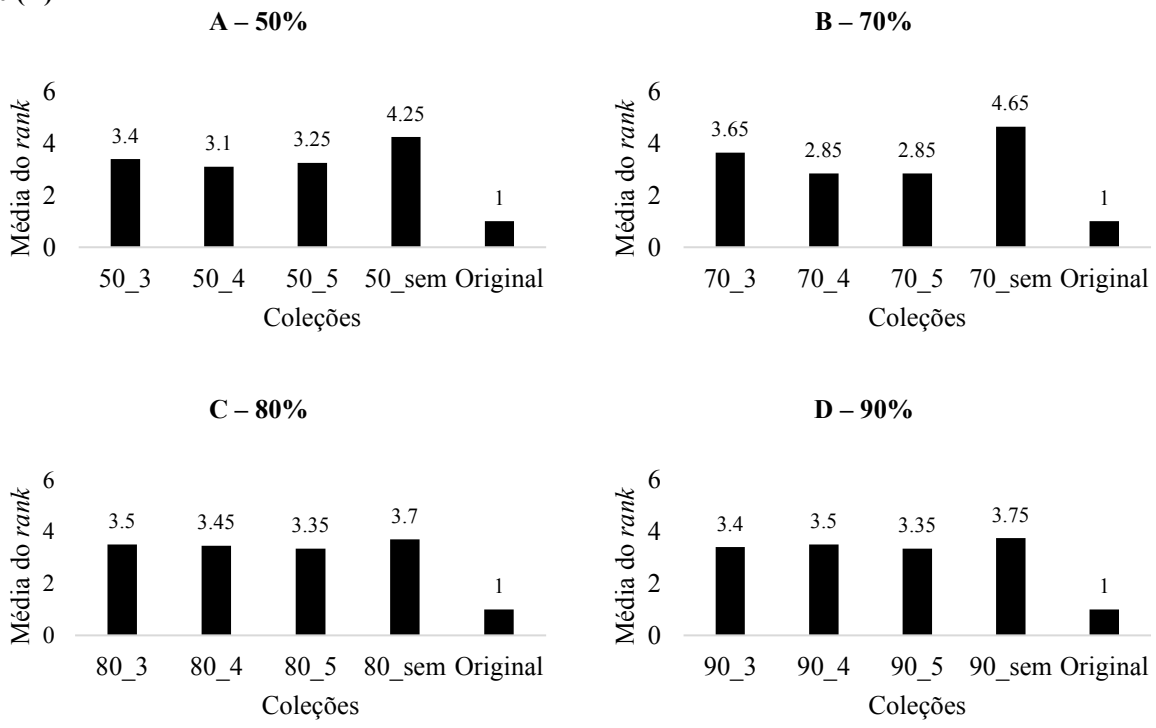


Figura 20 – Médias dos *ranks* do *precision* (por taxa de compressão) para todas as consultas realizadas na coleção "Original" e nas coleções sumarizadas com taxas de compressão de 50% (A), 70% (B), 80% (C) e 90% (D)



Desse modo, o coeficiente de concordância de Kendall apontou que para as coleções "Original" e as sumarizadas a taxa de compressão de 50%, as amostras das coleções têm um grau de concordância de 69% (Tabela 5). A coleção "50_sem", quando comparada com a coleção "Original" e as demais coleções sumarizadas com taxa de 50%, gerou os valores numéricos do *precision* mais elevados para sete das consultas ("educação permanente", "sociologia da educação", "psicologia educacional", "ensino aprendizagem", "filosofia da educação", "história da educação" e "política educacional) (Tabela 5) e apresentou maior média numérica dos *ranks* (Figura 20).

Nas coleções "Original" e sumarizadas com taxa de compressão de 70%, as amostras deram origem a um grau de concordância de Kendall de 85%. A coleção "70_sem", quando comparada com a coleção "Original" e as demais coleções sumarizadas com taxa de 70%, se destacou com os maiores valores numéricos do *precision* em seis consultas: "educação especial", "sociologia da educação", "psicologia educacional", "filosofia da educação", "história da educação" e "política educacional" (Tabela 5). Esta coleção também possui maior média numérica do *precision* e do *rank* (Figuras 19 e 20).

As amostras das coleções "Original" e sumarizadas com taxa de 80% apresentaram grau de concordância de Kendall de 59% (Tabela 5). A coleção "80_sem", quando comparada com a coleção "Original" e as demais coleções sumarizadas com taxa de 80%, possui maior valor numérico do *precision* em quatro consultas ("educação especial", "psicologia educacional", "política educacional" e "tecnologia educacional"), e apresentou maior média numérica do *precision* e dos *ranks* (Figuras 19 e 20).

Para as coleções "Original" e sumarizadas com taxa de 90% de compressão, o coeficiente de concordância apontou que as coleções têm um grau de concordância de Kendall de 58% (Tabela 5). A coleção "90_sem", quando comparada com a coleção "Original" e as demais coleções sumarizadas com taxa de 90%, possui maior valor numérico do *precision* em quatro consultas ("educação especial", "educação permanente", "educação pré-escolar" e "política educacional"). Esta coleção também possui maior média numérica do *precision* e do *rank* (Figuras 19 e 20).

Para todas as taxas, as coleções sumarizadas sem o uso de palavras-chave exibiram maior valor médio numérico do *precision* e do *rank* ("50_sem", "70_sem", "80_sem" e "90_sem") quando comparadas com a coleção "Original" e as demais coleções sumarizadas com mesma taxa (Figuras 19 e 20). Com relação ao uso de três, quatro, cinco, ou nenhuma palavra-chave na sumarização, conforme Rocha e Guelpeli (2017), o sumarizador apresenta bons resultados com o uso de palavras-chave, de forma que os sumários são constituídos por

sentenças importantes. No entanto, os resultados da RI com o modelo Cassiopeia revelaram que a quantidade de palavras-chave utilizadas na sumarização não influenciou na RI. Isso pode ser observado a partir da identificação de que em todas as taxas, a maior diferença numérica entre os *ranks* ocorreu entre as coleções sumarizadas sem uso de palavras-chave e a "Original" (Figura 20). As palavras-chave de textos acadêmicos são utilizadas para representação do texto e como indexadores em máquinas de busca; contudo, muitas vezes os autores não fazem as melhores escolhas dessas palavras. Acredita-se que essa questão influenciou no desempenho mais elevado das coleções sumarizadas sem uso das palavras-chave, visto que os sumários das coleções foram criados com o uso dessas palavras.

A Tabela 6 exibe a análise de significância (ANOVA de Friedman), entre os *ranks* da coleção "Original" e coleções sumarizadas com taxas de compressão de 50, 70, 80 e 90%. Analisando as coleções sumarizadas, verificou-se que não há diferença estatística entre as coleções sumarizadas com mesma taxa de compressão. Na comparação entre a RI padrão (coleção "Original") e a RI com o modelo Cassiopeia, considerando os resultados das coleções sumarizadas, por taxa de compressão, foi possível verificar, a partir da análise de significância, que apenas as coleções "70_4" e "70_5" não são estatisticamente diferentes da "Original". Assim, dentre as 16 coleções sumarizadas, 14 possuem significância menor que 0,05, ou seja, há diferença significativa na precisão dos resultados retornados ao usuário, entre estas coleções sumarizadas e a "Original" (Tabela 6). Logo, para essas coleções da RI com o modelo Cassiopeia, o esforço do usuário em analisar os documentos retornados foi menor, visto que uma quantidade maior de itens relevantes foi recuperada.

Tabela 6 – Análise de significância (ANOVA de Friedman – α = 0,05), por taxa de compressão, entre os *ranks* da coleção "Original" e coleções sumarizadas com taxas de 50%, 70%, 80% e 90%

Taxa de compressão 50%			**Taxa de compressão 70%**		
Comparações	**Diferença**	**Significância (p)**	**Comparações**	**Diferença**	**Significância (p)**
r1 (50_3) e 2 (50_4)	3	ns	r1 (70_3) e 2 (70_4)	8	ns
r1 (50_3) e 3 (50_5)	1,5	ns	r1 (70_3) e 3 (70_5)	8	ns
r1 (50_3) e 4 (50_sem)	8,5	ns	r1 (70_3) e 4 (70_sem)	10	ns
r1 (50_3) e 5 (Original)	24	< 0,05	r1 (70_3) e 5 (Original)	26,5	< 0,05
r2 (50_4) e 3 (50_5)	1,5	ns	r2 (70_4) e 3 (70_5)	0	ns
r2 (50_4) e 4 (50_sem)	11,5	ns	r2 (70_4) e 4 (70_sem)	18	ns
r2 (50_4) e 5 (Original)	21	< 0,05	r2 (70_4) e 5 (Original)	18,5	ns
r3 (50_5) e 4 (50_sem)	10	ns	r3 (70_5) e 4 (70_sem)	18	ns
r3 (50_5) e 5 (Original)	22,5	< 0,05	r3 (70_5) e 5 (Original)	18,5	ns
r4 (50_sem) e 5 (Original)	32,5	< 0,05	r4 (70_sem) e 5 (Original)	36,5	< 0,05
Taxa de compressão 80%			**Taxa de compressão 90%**		
Comparações	**Diferença**	**Significância (p)**	**Comparações**	**Diferença**	**Significância (p)**
r1 (80_3) e 2 (80_4)	0,5	ns	r1 (90_3) e 2 (90_4)	1	ns
r1 (80_3) e 3 (80_5)	1,5	ns	r1 (90_3) e 3 (90_5)	0,5	ns
r1 (80_3) e 4 (80_sem)	2	ns	r1 (90_3) e 4 (90_sem)	3,5	ns
r1 (80_3) e 5 (Original)	25	< 0,05	r1 (90_3) e 5 (Original)	24	< 0,05
r2 (80_4) e 3 (80_5)	1	ns	r2 (90_4) e 3 (90_5)	1,5	ns
r2 (80_4) e 4 (80_sem)	2,5	ns	r2 (90_4) e 4 (90_sem)	2,5	ns
r2 (80_4) e 5 (Original)	24,5	< 0,05	r2 (90_4) e 5 (Original)	25	< 0,05
r3 (80_5) e 4 (80_sem)	3,5	ns	r3 (90_5) e 4 (90_sem)	4	ns
r3 (80_5) e 5 (Original)	23,5	< 0,05	r3 (90_5) e 5 (Original)	23,5	< 0,05
r4 (80_sem) e 5 (Original)	27	< 0,05	r4 (90_sem) e 5 (Original)	27,5	< 0,05

4.3 RI padrão e RI com o modelo Cassiopeia: comparação entre todas as 17 coleções

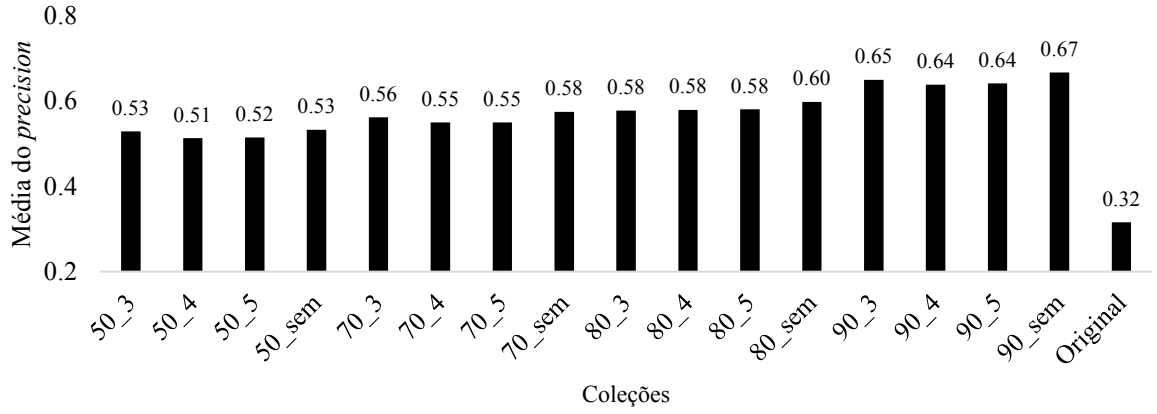
Para comparação entre todas as coleções (uma "Original" e 16 sumarizadas – 16 graus de liberdade), a Tabela 7 exibe os *ranks* (r1 a r17 – ANOVA de Friedman) do *precision* (obtidos a partir das consultas realizadas nas coleções), as médias e somas de "r", e o coeficiente de concordância de Kendall (W).

Tabela 7 – *Rank* (r1 a r17 – ANOVA de Friedman) do *precision* para todas as coleções

Coleção / **Consulta**	**50_3 r1**	**50_4 r2**	**50_5 r3**	**50_sem r4**	**70_3 r5**	**70_4 r6**	**70_5 r7**	**70_sem r8**	**80_3 r9**	**80_4 r10**	**80_5 r11**	**80_sem r12**	**90_3 r13**	**90_4 r14**	**90_5 r15**	**90_sem r16**	**Original r17**
Educação especial (EE)	3,00	4,50	4,50	2,00	7,00	7,00	7,00	15,00	11,50	11,50	11,50	16,00	11,50	11,50	11,50	17,00	1,00
Educação permanente (EP)	7,00	3,50	3,50	13,00	7,00	3,50	3,50	7,00	11,50	11,50	9,50	9,50	14,50	16,50	14,50	16,50	1,00
Educação pré-escolar (EPE)	11,50	11,50	11,50	11,50	5,50	5,50	5,50	5,50	5,50	5,50	5,50	5,50	15,00	15,00	15,00	17,00	1,00
Sociologia da educação (SE)	3,00	3,00	3,00	5,00	7,00	7,00	7,00	9,00	10,50	12,50	12,50	10,50	17,00	16,00	14,00	15,00	1,00
Psicologia educacional (PE)	7,50	7,50	7,50	15,50	7,50	7,50	7,50	13,00	3,00	3,00	3,00	11,50	15,50	11,50	15,50	15,50	1,00
Ensino aprendizagem (EA)	3,00	3,00	3,00	5,00	8,50	6,50	6,50	8,50	13,00	11,00	11,00	11,00	16,50	16,50	15,00	14,00	1,00
Filosofia da educação (FE)	4,00	2,50	2,50	5,00	10,00	6,50	6,50	12,50	12,50	10,00	10,00	8,00	15,00	15,00	17,00	15,00	1,00
História da educação (HE)	3,00	3,00	3,00	5,50	7,00	9,50	9,50	13,00	5,50	13,00	9,50	9,50	13,00	16,00	16,00	16,00	1,00
Política educacional (POE)	2,50	2,50	4,00	6,00	8,00	6,00	6,00	10,00	12,00	9,00	11,00	13,00	16,00	14,50	14,50	17,00	1,00
Tecnologia educacional (TE)	12,00	3,50	3,50	2,00	10,50	5,50	5,50	7,00	8,50	8,50	10,50	13,50	15,00	16,50	16,50	13,50	1,00
Média *Rank*	5,65	4,45	4,60	7,05	7,80	6,45	6,45	10,05	9,35	9,55	9,40	10,80	14,90	14,90	14,95	15,65	1,00
Soma *Rank*	56,50	44,50	46,00	70,50	78,00	64,50	64,50	100,50	93,50	95,50	94,00	108,00	149,00	149,00	149,50	156,50	10,00
Kendall - W	0,73																

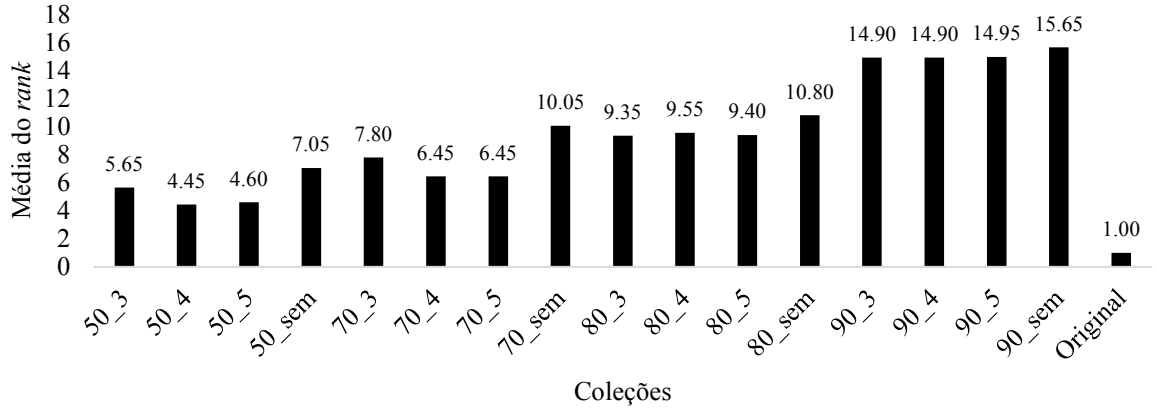
A Figura 21 mostra os valores médios do *precision* para todas as coleções.

Figura 21 – Médias dos valores do *precision* para todas as coleções



A Figura 22 ilustra as médias dos *ranks* (ANOVA de Friedman) para todas as 17 coleções.

Figura 22 – Médias dos *ranks* do *precision* para todas coleções



Ao efetuar a comparação entre as 17 coleções, o coeficiente de Kendall revelou uma concordância de 73% (Tabela 7). À medida em que a taxa de compressão aumentou, os valores numéricos médios do *precision* e dos *ranks* também aumentaram para as coleções sumarizadas (Figuras 21 e 22). A análise de significância, comparando os *ranks* entre todas as coleções (Tabela 8), demonstrou que as coleções que possuem taxa de compressão de 90% (r13 a r16 – coleções "90_3", "90_4", "90_5" e "90_sem") foram superiores às coleções sumarizadas com 50% e a duas com 70% ("70_4" e "70_5"). Contudo, não houve diferença estatística entre as coleções sumarizadas com 80% e 90% de compressão.

Tabela 8 – Análise de significância (ANOVA de Friedman – α = 0,05) entre os *ranks* das coleções

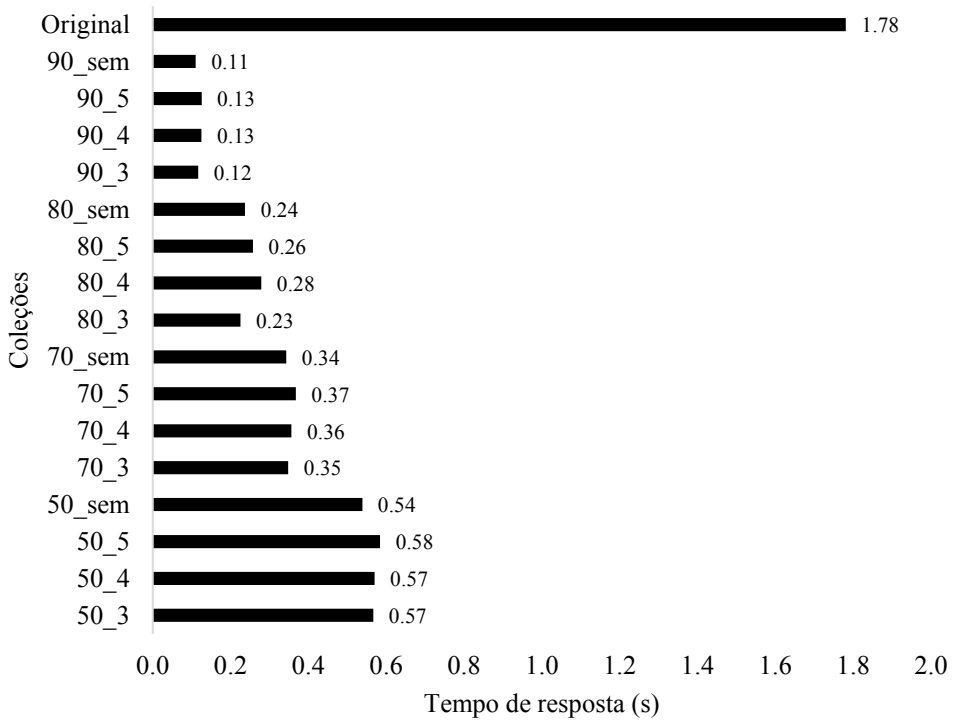
Comparações	Diferença	(p)	Comparações	Diferença	(p)	Comparações	Diferença	(p)
r1 e 2 =	12	ns	r4 e 5 =	7,5	ns	r8 e 9 =	7	ns
r1 e 3 =	10,5	ns	r4 e 6 =	6	ns	r8 e 10 =	5	ns
r1 e 4 =	14	ns	r4 e 7 =	6	ns	r8 e 11 =	6,5	ns
r1 e 5 =	21,5	ns	r4 e 8 =	30	ns	r8 e 12 =	7,5	ns
r1 e 6 =	8	ns	r4 e 9 =	23	ns	r8 e 13 =	48,5	ns
r1 e 7 =	8	ns	r4 e 10 =	25	ns	r8 e 14 =	48,5	ns
r1 e 8 =	44	ns	r4 e 11 =	23,5	ns	r8 e 15 =	49	ns
r1 e 9 =	37	ns	r4 e 12 =	37,5	ns	r8 e 16 =	56	ns
r1 e 10 =	39	ns	r4 e 13 =	78,5	< 0,05	r8 e 17 =	90,5	< 0,05
r1 e 11 =	37,5	ns	r4 e 14 =	78,5	< 0,05	r9 e 10 =	2	ns
r1 e 12 =	51,5	ns	r4 e 15 =	79	< 0,05	r9 e 11 =	0,5	ns
r1 e 13 =	92,5	< 0,05	r4 e 16 =	86	< 0,05	r9 e 12 =	14,5	ns
r1 e 14 =	92,5	< 0,05	r4 e 17 =	60,5	ns	r9 e 13 =	55,5	ns
r1 e 15 =	93	< 0,05	r5 e 6 =	13,5	ns	r9 e 14 =	55,5	ns
r1 e 16 =	100	< 0,05	r5 e 7 =	13,5	ns	r9 e 15 =	56	ns
r1 e 17 =	46,5	ns	r5 e 8 =	22,5	ns	r9 e 16 =	63	ns
r2 e 3 =	1,5	ns	r5 e 9 =	15,5	ns	r9 e 17 =	83,5	< 0,05
r2 e 4 =	26	ns	r5 e 10 =	17,5	ns	r10 e 11 =	1,5	ns
r2 e 5 =	33,5	ns	r5 e 11 =	16	ns	r10 e 12 =	12,5	ns
r2 e 6 =	20	ns	r5 e 12 =	30	ns	r10 e 13 =	53,5	ns
r2 e 7 =	20	ns	r5 e 13 =	71	ns	r10 e 14 =	53,5	ns
r2 e 8 =	56	ns	r5 e 14 =	71	ns	r10 e 15 =	54	ns
r2 e 9 =	49	ns	r5 e 15 =	71,5	ns	r10 e 16 =	61	ns
r2 e 10 =	51	ns	r5 e 16 =	78,5	< 0,05	r10 e 17 =	85,5	< 0,05
r2 e 11 =	49,5	ns	r5 e 17 =	68	ns	r11 e 12 =	14	ns
r2 e 12 =	63,5	ns	r6 e 7 =	0	ns	r11 e 13 =	55	ns
r2 e 13 =	104,5	< 0,05	r6 e 8 =	36	ns	r11 e 14 =	55	ns
r2 e 14 =	104,5	< 0,05	r6 e 9 =	29	ns	r11 e 15 =	55,5	ns
r2 e 15 =	105	< 0,05	r6 e 10 =	31	ns	r11 e 16 =	62,5	ns
r2 e 16 =	112	< 0,05	r6 e 11 =	29,5	ns	r11 e 17 =	84	< 0,05
r2 e 17 =	34,5	ns	r6 e 12 =	43,5	ns	r12 e 13 =	41	ns
r3 e 4 =	24,5	ns	r6 e 13 =	84,5	< 0,05	r12 e 14 =	41	ns
r3 e 5 =	32	ns	r6 e 14 =	84,5	< 0,05	r12 e 15 =	41,5	ns
r3 e 6 =	18,5	ns	r6 e 15 =	85	< 0,05	r12 e 16 =	48,5	ns
r3 e 7 =	18,5	ns	r6 e 16 =	92	< 0,05	r12 e 17 =	98	< 0,05
r3 e 8 =	54,5	ns	r6 e 17 =	54,5	ns	r13 e 14 =	0	ns
r3 e 9 =	47,5	ns	r7 e 8 =	36	ns	r13 e 15 =	0,5	ns
r3 e 10 =	49,5	ns	r7 e 9 =	29	ns	r13 e 16 =	7,5	ns
r3 e 11 =	48	ns	r7 e 10 =	31	ns	r13 e 17 =	139	< 0,05
r3 e 12 =	62	ns	r7 e 11 =	29,5	ns	r14 e 15 =	0,5	ns
r3 e 13 =	103	< 0,05	r7 e 12 =	43,5	ns	r14 e 16 =	7,5	ns
r3 e 14 =	103	< 0,05	r7 e 13 =	84,5	< 0,05	r14 e 17 =	139	< 0,05
r3 e 15 =	103,5	< 0,05	r7 e 14 =	84,5	< 0,05	r15 e 16 =	7	ns
r3 e 16 =	110,5	< 0,05	r7 e 15 =	85	< 0,05	r15 e 17 =	139,5	< 0,05
r3 e 17 =	36	ns	r7 e 16 =	92	< 0,05	r16 e 17 =	146,5	< 0,05
			r7 e 17 =	54,5	ns			

Adicionalmente, identificou-se que todas as coleções sumarizadas com taxas de 80 e 90%, e a coleção “70_sem” são estatisticamente diferentes da “Original”. Dessa maneira, notou-se que com elevadas taxas de compressão, que dão origem a textos com conteúdo

reduzido, uma maior quantidade de documentos relevantes foi retornada para o usuário, ou seja, houve maior precisão e menor sobrecarga de informação, auxiliando, assim, na recuperação de informação em textos acadêmicos. Isso é perceptível visto que, segundo Baeza-Yates e Ribeiro-Neto (2013), documentos longos têm maior chance de corresponder à consulta simplesmente pelo tamanho que possuem, o que não significa que sejam relevantes para a expressão de busca.

Para verificar a rapidez com que os documentos foram recuperados, o intervalo de tempo (s) entre o recebimento da consulta do usuário e a apresentação da resposta, ou seja, tempo de resposta, foi calculado automaticamente pelo SRI (para cada uma das consultas efetuadas em cada uma coleções da RI com o modelo Cassiopeia e RI padrão) e registrado em uma planilha de *log*, juntamente com os resultados das métricas. Assim, as médias do tempo de resposta das coleções sumarizadas e da coleção "Original" são exibidas na Figura 23.

Figura 23 – Média do tempo de resposta (s): comparação entre as coleções sumarizadas (RI com o modelo Cassiopeia) e a coleção original (RI padrão)



Dentre os conjuntos, nos quais a RI com o modelo Cassiopeia foi realizada, as coleções sumarizadas com taxa de compressão de 90%, que apresentaram maior precisão dos resultados retornados ao usuário, também mostraram o menor tempo de resposta. Ao analisar a agilidade da RI, foi possível perceber que à medida em que a taxa de compressão aumentou, o tempo de resposta diminuiu. A coleção "Original" (RI padrão), que gerou menor precisão dos resultados para o usuário, exibiu o maior tempo de resposta. Isso foi factível porque um SRI

efetua a representação do documento verificando o texto completo, ou seja, todas as palavras são utilizadas como termos para o índice, e devido à sumarização dos textos, houve redução da quantidade de termos para o índice e o índice invertido, o que simplificou o processo de indexação, permitindo maior agilidade no acesso aos documentos e processamento da consulta.

A partir disso, foi possível constatar que os resultados obtidos estão relacionados a algumas características das técnicas e ferramentas utilizadas na pesquisa. Primeiramente, de acordo com Rocha e Guelpeli (2017), o sumarizador PragmaSUM possui bom desempenho com altas taxas de compressão. Tal desempenho é atribuído ao uso conjunto: do método de seleção de atributos do modelo Cassiopeia, que gera sumários com boa informatividade; e do algoritmo TF-ISF que seleciona as frases que formam os sumários, verificando a importância de uma palavra em uma frase, o que possibilita a ponderação pela frequência dos termos, melhorando os resultados. Segundo Wives (2004) e Nogueira (2009), a seleção de atributos tem um fator decisivo para a boa qualidade e melhor desempenho da RI, visto que palavras pouco frequentes são muito discriminantes, ocupam espaço desnecessário no índice e não recuperam muitos documentos. Outra questão é que o fato dos textos fonte do *corpus* principal possuírem pequena variação de tamanho, o que vai ao encontro com o princípio de normalização do tamanho dos documentos, é importante fator para o ranqueamento na RI. Adicionalmente, essa qualidade também foi proporcionada pelo SRI *Apache Solr,* a partir de três principais vantagens do modelo vetorial: esquema de ponderação de termos, estratégia do par termo-documento que aproxima o documento às condições da consulta e ordenação dos documentos de acordo com o grau de similaridade em relação à consulta.

4.4 Hipótese

A metodologia de teste de hipótese, aqui adotada, considerou as amostras obtidas a partir das execuções e avaliação das RIs, que geraram os valores das métricas *precision, recall* e *F-measure*. A hipótese nula (H_0) consiste na afirmação de que a aplicação da técnica de sumarização, a partir da seleção de atributos do modelo Cassiopeia, num *corpus* de textos acadêmicos, não auxilia na recuperação de informação, logo, não diminui a sobrecarga de informação e não melhora a precisão dos resultados retornados ao usuário.

Formalmente, a hipótese nula pode ser representada por meio da Equação 4:

$$H_0: k_{\text{RI padrão}} = k_{\text{RI com o modelo Cassiopeia}}. \quad (4)$$

No qual:

H_0 = hipótese nula;

$k_{\text{RI com o modelo Cassiopeia}}$ = distribuição das *k* amostras para RI nos textos sumarizados com a seleção de atributos do modelo Cassiopeia;

$k_{\text{RI padrão}}$ = distribuição das *k* amostras para RI nos textos completos.

Se a hipótese nula for considerada falsa, outra afirmativa deve ser verdadeira. Esta pesquisa propõe a hipótese alternativa H_1, representada na Equação 5: a aplicação da técnica de sumarização, a partir da seleção de atributos do modelo Cassiopeia, num *corpus* de textos acadêmicos, auxilia na recuperação de informação, diminui a sobrecarga de informação e melhora a precisão dos resultados retornados ao usuário.

$$H_1 = k_{\text{RI com o modelo Cassiopeia}} > k_{\text{RI padrão}}. \quad (5)$$

Portanto, a partir da análise dos resultados gerados, é possível concluir que a hipótese H_0 foi rejeitada e a hipótese H_1 aceita.

5 CONCLUSÃO

Com a quantidade elevada de documentos textuais disponíveis, particularmente na área educacional, a partir do aumento das produções científicas nas instituições, 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. Nessa conjectura, esta pesquisa centrou-se na utilização da mineração de textos como forma de contribuição na solução dos problemas relacionados à RI, tais como sobrecarga de informação e falta de precisão e relevância dos resultados apresentados ao usuário.

Dessa maneira, o principal objetivo deste trabalho foi analisar se a aplicação da técnica de sumarização, a partir do método de seleção de atributos do modelo Cassiopeia (implementado pelo sumarizador PragmaSUM), num *corpus* de textos acadêmicos que compõem repositórios institucionais, auxilia na recuperação de informação, diminuindo a sobrecarga de informação e melhorando a precisão dos resultados retornados ao usuário. A partir das avaliações das RIs, por meio das métricas *precision, recall e F-measure*, e análise estatística dos dados obtidos, constatou-se que a hipótese da pesquisa foi confirmada e colaborou, significativamente, com a recuperação de informação.

Na comparação entre a RI padrão e a RI com o modelo Cassiopeia, por taxa de compressão, dentre as 16 coleções sumarizadas, 14 foram estatisticamente superiores à "Original". Contudo, não há diferença estatística entre as coleções sumarizadas com mesma taxa de compressão.

Na análise realizada entre as 17 coleções, em conjunto, identificou-se que todas as coleções sumarizadas com taxas de 80 e 90%, e a coleção "70_sem" foram superiores à "Original". Além disso, à medida em que a taxa de compressão aumentou, os valores médios do *precision* e do *rank* também aumentaram para as coleções sumarizadas. Ademais, as coleções que possuem taxa de compressão de 90% ("90_3", "90_4", "90_5" e "90_sem") foram superiores às coleções sumarizadas com 50% de compressão e a duas coleções sumarizadas com 70% ("70_4" e "70_5"). Contudo, não houve diferença estatística entre as coleções sumarizadas com 80% e 90% de compressão.

Ao verificar o uso de palavras-chave na sumarização, foi possível perceber que para todas as taxas de compressão, as coleções sumarizadas sem o uso de palavras-chave exibiram maior valor médio do *rank* do *precision.* Portanto, o uso de três, quatro, cinco, ou nenhuma palavra-chave na sumarização não influenciou na recuperação de informação.

Nos casos em que a RI com o modelo Cassiopeia foi superior à RI padrão, o esforço do usuário em analisar os documentos retornados será menor, visto que uma quantidade maior de itens relevantes foi recuperada. Em relação à rapidez com que a RI foi executada, para as coleções sumarizadas, à medida em que a taxa de compressão aumentou, o tempo de resposta diminuiu, permitindo maior agilidade no acesso aos documentos e processamento da consulta. A coleção "Original" (RI padrão) apresentou o maior tempo de resposta, dentre todas as coleções.

Dessa maneira, a RI com o modelo Cassiopeia, especialmente com elevadas taxas de compressão que dão origem a textos com conteúdo reduzido, quando comparada à RI padrão, realizada nos textos acadêmicos completos, aumentou a precisão dos resultados retornados ao usuário, ou seja, recuperou uma quantidade mais elevada de documentos relevantes e menor de irrelevantes. Assim, reduziu a sobrecarga de informação; assegurou maior agilidade na RI, a partir da diminuição do tempo de resposta; simplificou o processo de indexação, devido à redução da quantidade de termos dos índices, mostrando-se pertinente, principalmente, para grandes coleções, caso em que o custo computacional é elevado; e, atenuou a alta dimensionalidade, por meio da diminuição do número de palavras irrelevantes, no momento do tratamento dos dados textuais.

Essas questões revelam a importância do conjunto de métodos, técnicas e ferramentas metodológicas, aqui empregadas, que auxiliaram na recuperação informacional de textos acadêmicos existentes em repositórios institucionais. O método de seleção de atributos do modelo Cassiopeia, em conjunto com o algoritmo TF-ISF, gerou sumários mais informativos, com as palavras mais importantes dos textos, o que viabilizou ganhos em desempenho com relação à busca por informação útil. Os textos fonte do *corpus* principal possuem pequena variação de tamanho, formalizando o princípio de normalização do tamanho dos documentos. Adicionalmente, o SRI (buscador *Solr*) possui caráter científico e didático, útil para estudos na área de RI, e utiliza o melhor modelo de RI para textos não estruturados: o vetorial, que pela simplicidade e agilidade permitiu melhoria na qualidade da recuperação de documentos próximos às condições da busca. Por último, a aplicação da sumarização em grandes coleções pode viabilizar a redução do *overhead* (processamento em excesso) de entrada/saída e atrasos de comunicação.

Com base nas conclusões, nas seções seguintes serão apresentadas contribuições, limitações e possíveis trabalhos futuros relacionados a esta pesquisa.

5.1 Contribuições

Diante dos resultados obtidos, a partir da utilização do modelo Cassiopeia na recuperação de informação, foi possível contribuir com:

- Um modelo para recuperação de informação que possibilita menor sobrecarga de informação e melhoria na precisão dos textos acadêmicos, existentes em repositórios institucionais, retornados ao usuário;
- A redução da quantidade de termos nos índices, com uso da sumarização a partir do método de seleção de atributos do modelo Cassiopeia;
- Simplificação do processo de indexação, com o uso dos resumos resultantes da sumarização;
- Um sistema para recuperação de informação em textos acadêmicos que compõem repositórios institucionais;
- Maior rapidez com que a recuperação da informação é executada;
- Um *corpus* do domínio educacional que apresenta julgamentos de relevância para testes de RI.

5.2 Limitações

A não utilização de avaliação baseada em usuários e de medidas orientadas a eles, como taxa de novidade e esforço da revocação, é uma das limitações desta pesquisa, visto que poderiam ser utilizadas como forma complementar na avaliação da recuperação informacional. Outro fator refere-se à utilização de um *corpus* relativamente pequeno, com dez expressões de busca. Tal questão é justificada pela inexistência de coleções de RI, em português, que apresentem os julgamentos de relevância para as consultas em SRIs.

5.3 Trabalhos futuros

Os resultados deste trabalho sugerem alguns caminhos de trabalho futuro, como: uso de diferentes sumarizadores automáticos, que não utilizam o método de seleção de atributos do Cassiopeia, no processo de criação de sumários; efetuar a recuperação de informação em coleções de grande porte; comparar a RI utilizando diferentes sistemas de RI; executar os testes por meio de um SRI desenvolvido com aprendizagem de máquina, para análise do conteúdo

textual em coleções maiores; e, por último, introduzir a sumarização automática nos sistemas de RI de forma que tanto o texto quanto o índice possam ser reduzidos.

REFERÊNCIAS

AGUIAR, L. H. G. DE; ROCHA, V. J. C.; GUELPELI, M. V. C. **Uma coleção de artigos científicos de Português compondo um *Corpus* no domínio educacional.** PLURAIS Revista Multidisciplinar, v. 2, n. 1, p. 60–74, 2017.

ALUÍSIO, S. M.; ALMEIDA, G. M. D. B. **O que é e como se constrói um corpus? Lições aprendidas na compilação de vários corpora para pesquisa linguística.** Calidoscópio, v. 4, n. 3, p. 155–177, 2006.

ARANHA, C. N. **Uma abordagem de pré-processamento automático para mineração de textos em português: sob o enfoque da inteligência computacional**. Tese (Doutorado) – Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2007.

ARANHA, C. N.; PASSOS, E. **A Tecnologia de Mineração de Textos**. RESI-Revista Eletrônica de Sistemas de Informação, 2006.

ARAÚJO JÚNIOR, R. H.; TARAPANOFF, K. **Precisão no processo de busca e recuperação da informação: uso da mineração de textos.** *Ci. Inf.* [*online*]. 2006.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Recuperação de Informação**: Conceitos e Tecnologia das Máquinas de Busca. 2. ed. Porto Alegre: Bookman, 2013. 590 p.

BANDIM, M.A.S.; CORREA, R.F. **Indexação automática por atribuição de artigos científicos em português da área de Ciência da Informação**. *Transinformação*, v.31, 2019

BARTH, F. J. **Uma introdução ao tema Recuperação de Informações Textuais**. Revista de Informática Teórica e Aplicada – RITA. Universidade Federal do Rio Grande do Sul (UFRGS), 2013.

BATISTA JÚNIOR, W. S. **Recuperação de informação com auxílio de extratos automáticos**. Dissertação de mestrado – Universidade Federal de São Carlos (UFSCar), 2006.

BEYER K.; GODSTEIN J.; RAMAKRISHNAN R.; SHAFT U. **When is "Nearest Neighbor" Meaningful?** In: Beeri C, Buneman P, editors. International Conference on Database Theory (ICDT), Jerusalem, Israel: Springer Verlag; p. 217-235, 1999.

BOTELHO, T. M. G. **Avaliação da recuperação de informações em sistemas "Online"; Considerações metodológicas**. SERPRO. 2011. Disponível em: <http://www.brapci.inf.br/index.php/article/download/17363>. Acesso em: 06 fev. 2018.

BRITO, A. G. **Proposta de modelo de recuperação da informação baseado em conteúdo de arquivos de legendas de filmes e séries**. Dissertação de mestrado. Universidade Fumec. Belo Horizonte, MG/Brasil, 2015.

BUCHLER, T. **Construa Aplicações utilizando o poderoso Apache Solr como *engine* de buscas**. MundoJ. Disponível em: <http://www.univale.com.br/unisite/mundo-j/artigos/43engine.pdf >. Acesso em: maio. 2018.

CALLEGARI-JACQUES, S. M. **Bioestatística: Princípios e Aplicações.** Porto Alegre: Artmed, 2007.

CAMPOS, G. M. **Estatística Prática para Docentes e Pós-graduandos**. Disponível em: <https://edisciplinas.usp.br/pluginfile.php/3223131/mod_folder/content/0/Estat>. Acesso em: 10 ago. 2019.

CARDOSO, O. N. P. **Recuperação de Informação**. INFOCOMP, [S.l.], v. 2, n. 1, p. 33-38, 2004. Disponível em: <http://infocomp.dcc.ufla.br/>. Acesso em: 07 fev. 2018.

CARRILHO JÚNIOR, J. R. **Desenvolvimento de uma metodologia para mineração de textos**. Dissertação de mestrado – Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), Rio de Janeiro, 2007.

COOPER, W. S. **The formalism of probability theory in ir: a foundation or an ncumbranc**e? In SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, pages 242–247, New York, NY, USA, 1994.

CORREIA, J. S. B. L. **Indexação de Documentos Clínicos**. Dissertação de mestrado. Universidade do Porto, Portugal, 85 p. 2016.

DIAS, M. P.; CARVALHO, J. O. F. **A Visualização da Informação e a sua contribuição para a Ciência da Informação**. DataGramaZero, Rio de Janeiro, v. 8, n.5, out. 2007. Disponível em: <http://www.dgz.org.br/out07/Art_02.htm>. Acesso em: jun/2019.

FERNEDA, E. **Recuperação de Informação: Análise sobre a contribuição da Ciência da Computação para a Ciência da Informação**. Tese de doutorado - Universidade de São Paulo – USP, São Paulo, Brasil, 2003.

FERNEDA, E. **Aplicando Algoritmos Genéticos na Recuperação de Informação.** DataGramaZero - Revista de Ciência da Informação - v.10 n.1, 2009.

FERNEDA, E. **Introdução aos Modelos Computacionais de Recuperação de Informação.** Rio de Janeiro: Ciência Moderna, 2012.

GIL, A. C. **Como elaborar projetos de pesquisa**. 4. Ed. – São Paulo, Atlas, 176 p. 2002.

GRAINGER, T.; POTTER, T. ***Solr in action***. Shelter Island: Manning, 666 p. 2014.

GUELPELI, M.V.C.; **Cassiopeia: Um modelo de agrupamento de textos baseado em sumarização**. Tese de doutorado – Universidade Federal Fluminense. Programa de Pós-graduação em Computação, Niterói, BR – RJ, Brasil, 2012.

HOWLAND, P. e PARK, H. ***Cluster*-Preserving Dimension Reduction Methods for Document Classification.** Book survey of text mining: clustering, classification, and retrieval Second. Editors BERRY, M. E CASTELLANO, M. Edition, Springer, Part I Clustering, pp 3- 24, 2007.

LAKATOS, E. M., MARCONI, M. A. **Fundamentos de metodologia científica**. 5. Ed.

São Paulo: Atlas. 2003.

LE COADIC, Y. F. A **Ciência da Informação**. 2.ed. Brasília: Briquet de Lemos, 2004.

LEITE, F., AMARO, B., BATISTA, T., & COSTA, M. **Boas práticas para a construção de repositórios institucionais da produção científica**. Brasília: Ibict. 2012.

LUGO, G. A. G. **Um Modelo de Sistemas Multiagentes para Partilha de Conhecimento utilizando Redes Sociais Comunitárias**. PhD thesis, Escola Politécnica da Universidade de São Paulo, abril 2004.

LUHN, H. P. **The automatic creation of literature abstracts.** IBM Journal of Research and Development, 2, pp. 159-165,1958.

MARCONDES, C. H. *et. al*. **Bibliotecas digitais**: Saberes e Práticas. Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT). EDUFBA. Salvador, BA, 2005.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. ***An introduction to information retrieval***. Draft: Cambridge University Press, 581 p. 2009.

MIRANDA, I. A. A.; MOURA, M. A. **Acesso aberto e gestão colaborativa de repositórios institucionais**: a experiência da UFMG. BiblioCanto, v. 3, n. 2, p. 37 – 50, 2017.

MOOERS, C. **Zatoeodmg applied to mechanical organization of knowledge**. American Documentation, v. 2, p. 20–32, 1951.

MONTEIRO, S. D.; FERNANDES, R. P. M.; DECARLI, G. C.; TREVISAN, G. L. **Sistemas de recuperação da informação e o conceito de relevância nos mecanismos de busca: semântica e significação.** Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, v. 22, n.50, p. 161-175, 2017.

NOGUEIRA, B. M. **Seleção não-supervisionada de atributos para Mineração de Textos**. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, São Paulo, Brasil, 2009.

NORMANDO, D. TJDERHANE, L. QUINTÃO, C. C. A. **A escolha do teste estatístico – um tutorial em forma de apresentação em PowerPoint**. Dental Press J. Orthod. V. 15, nº1, p. 101-106, 2010.

RINO, L.H.M.; PARDO, T.A.S. **A Sumarização Automática de Textos: Principais Características e Metodologias.** Anais do XXIII Congresso da Sociedade Brasileira de Computação, Vol. VIII: III Jornada de Minicursos de Inteligência Artificial (III MCIA), pp. 203-245. Campinas-SP, 2003.

ROCHA, V. J. C**. Pragmasum: novos métodos na utilização de palavras-chave na sumarização automática.** Dissertação de mestrado**.** Universidade Federal dos Vales do Jequitinhonha e Mucuri - UFVJM, Diamantina, MG, Brasil, 2017.

ROCHA, V. J. C.; GUELPELI, M. V. C. **Pragmasum: Automatic Text Summarizer Based On User Pr Profile.** International Journal of Current Research, v. 9, n. 7, p. 53935–53942, 2017.

REZENDE, S. O.; MARCACINI, R. M.; MOURA, M. F. **O uso da Mineração de Textos para Extração e Organização Não Supervisionada de Conhecimento.** Revista de Sistemas de Informação da FSMA n. 7 (2011) p. 7-21, 2011.

SAYÃO, L. F. *et. al*. **Implantação e gestão de repositórios institucionais: políticas, memória, livre acesso e preservação**. EDUFBA. UFBA. Salvador, 2009.

SILVA, R. D. L.; SILVA, E. M. **Mas o que é mesmo *Corpus*? – Alguns Apontamentos sobre a Construção de Corpo de Pesquisa nos Estudos em Administração**. XXXVII Encontro da ANPAD. Rio de Janeiro. RJ, 2013.

SILVA, L. O. **BOOKISH: uma ferramenta para contextualização de documentos utilizando mineração de textos e expansão de consulta.** Dissertação de Mestrado. Universidade Federal de Goiás – UFG, Goiânia, 2009.

SILVA, R. E. da; SANTOS, P. L. V. A. da C.; FERNEDA, E**. Modelos de recuperação de informação e web semântica: a questão da relevância**. Informação & Informação, v. 18, n. 3, p. 27 – 44, set./dez. 2013.

SILVA, F. T. da. **Luppar: um sistema de recuperação de informação para coleções fechadas de documentos**. Dissertação de Mestrado. Universidade Estadual do Ceará, Centro de Ciências e Tecnologia, Mestrado Acadêmico em Ciência da Computação, Fortaleza, 2018.

SOLR. **Apache Solr 7.2.1 Documentation**. Disponível em: <https://lucene.apache.org/solr/7_2_1/>. Acesso em: jul. 2018.

SPINK, A. *et al.* **Searching the web: the public and their queries**. Journal of the American Society for Information Science, v. 53, n. 2, p.226-234, 2001.

VIALI, L**. Testes de hipóteses não paramétricos**. Apostila. Instituto de Matemática, Departamento de Estatística – Universidade Federal do Rio Grande do Sul – UFRGS, Porto Alegre, 2008.

WIVES, L. K. **Técnicas de descoberta de conhecimento em textos aplicadas à inteligência competitiva**. PhD thesis, Instituto de Informática, UFRGS, Porto Alegre, 2001.

WIVES, L. K. **Utilizando conceitos como descritores de textos para o processo de identificação de conglomerados (clustering) de documentos** – Tese (doutorado) – Universidade Federal do Rio Grande do Sul – UFRGS. Porto Alegre, RS, Brasil, 136 p. 2004.

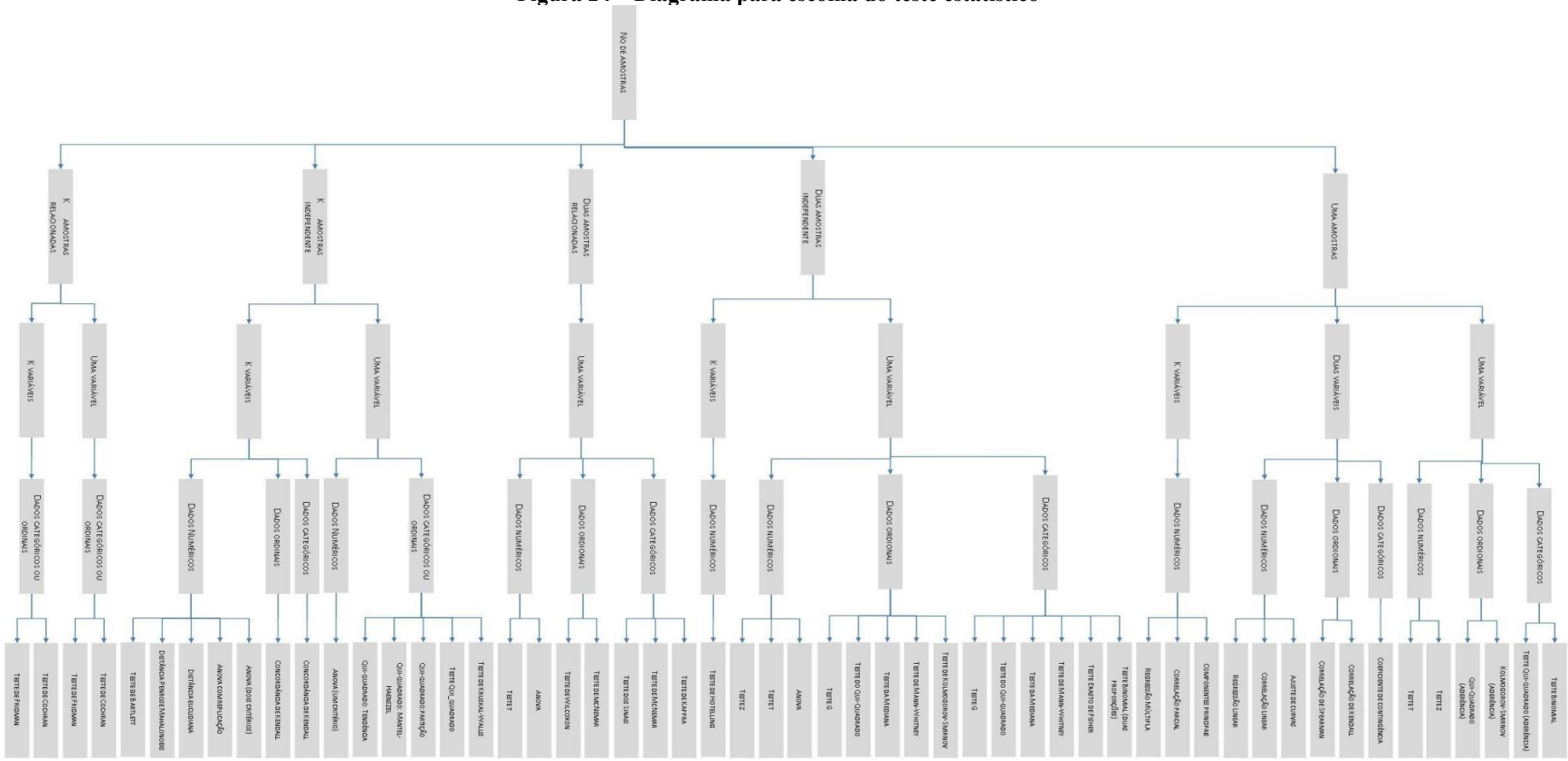
APÊNDICE A – RESULTADO DE TODAS AS MÉTRICAS

O Apêndice A apresenta os resultados das métricas *precision*, *recall* e *F-Measure,* gerados a partir da avaliação das RIs, e dos testes estatísticos realizados. Devido à elevada quantidade de informação gerada (tabelas e gráficos), o Apêndice pode ser acessado no *link* abaixo.

Link: https://drive.google.com/drive/folders/1yNukwcdY7TXBy-e1PNVP2x9Z75LJwF-g?usp=sharing

ANEXO A – DIAGRAMA PARA DEFINIÇÃO DO TESTE ESTATÍSTICO

Figura 24 – Diagrama para escolha do teste estatístico



Fonte: Callegari-Jacques, 2007.