SUPORTE AO MAPEAMENTO SISTEMÁTICO: UM APOIO À PESQUISA BIBLIOGRÁFICA

Pedro Henrique Conilh de Beyssac Ramos

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Sistemas e Computação.

Orientador: Geraldo Bonorino Xexéo

Rio de Janeiro
Junho de 2016

SUPORTE AO MAPEAMENTO SISTEMÁTICO: UM APOIO À PESQUISA BIBLIOGRÁFICA

Pedro Henrique Conilh de Beyssac Ramos

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Examinada por:

Prof. Geraldo Bonorino Xexéo, D.Sc.

Prof. Jano Moreira de Souza, Ph.D.

Prof. Eduardo Soares Ogasawara, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
JUNHO DE 2016

Ramos, Pedro Henrique Conilh de Beyssac

Suporte ao mapeamento sistemático: um apoio à pesquisa bibliográfica/Pedro Henrique Conilh de Beyssac Ramos. – Rio de Janeiro: UFRJ/COPPE, 2016.

XVI, 131 p.: il.; 29, 7cm.

Orientador: Geraldo Bonorino Xexéo

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2016.

Referências Bibliográficas: p. 98 – 102.

1. Mapeamento Sistemático. 2. Revisão Sistemática. 3. Revisão Bibliográfica. I. Xexéo, Geraldo Bonorino. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

Gloria in excelsis Deo

Agradecimentos

Primeiramente, gostaria de afirmar que esse trecho de agradecimento é certamente a parte mais relevante desse documento. As palavras aqui descritas são destinadas às pessoas pelas quais tenho profunda gratidão. Em seguida, serão apresentados os capítulos de minha contribuição acadêmica visando: retribuir o feito pela sociedade que financiou meus estudos e honrar a instituição de ensino que me deu suporte para tal realização.

Que todo mérito por esse estudo seja primeiramente dado a Deus, pois sem sua força e auxílio nada teria sido realizado. Esse estudo só foi possível graças aos anjos que Ele colocou em minha vida para ajudar em minha caminhada. Por isso, venho agradecer a todos que fazem parte da minha vida e dizer que cada um deles também foi essencial pela finalização desse trabalho.

Nominalmente venho agradecer a cada pessoa que fez parte dessa trajetória: ao meu pai, José Airton, meu maior exemplo de caráter, persistência e de que o valor de um homem não se mede por diplomas, notas ou títulos. Ele foi um dos maiores responsáveis por prover o suporte necessário para que esse estudo fosse concluído; à minha namorada Paloma que nunca deixou de me apoiar, sendo fundamental para essa conclusão. Mesmo nos momentos mais difíceis esteve disposta a fazer sacrifícios; à Eunice e Rosângela; a todos meus familiares, em especial a: Dagot, Marie-Agnès, Marie-Cécile, Thierry, Mathieu, Kátia, Emmidy, Zyon, Alessandra e Renata; aos meus Amigos mais próximos que não desistiram de me apoiar e acreditar em mim: Alessandra Machado, André Ramos, Arthur Mello, César Barbosa, Daniel Nunes, Diego Souza, Erick Regis e Pâmela Cristine, Fábio Venâncio, Fernando Magalhães, Gabriel Bié, Gabriel Mannarino, Gustavo Daniel, Gustavo Lima, Heloíse e Maria Cecília, Juliano Rodrigues, Marcos Petrúngaro, Roberto Tadeu, Rodrigo Coelho, Vanius Farias, Victor Furtado, Wellington Mascena e Ygor Canalli; aos que não hesitaram em ajudar: Bárbara Pimenta, Danielle Caled, Fabrício Pereira, Fellipe Braida, Fernanda Ribeiro, Hugo Rebelo, Luís Felipe, Marcelo Arêas, Matheus Emerick, Raul Sena, Talita Ribeiro, Victor Vidigal e Vitor Silva; a todos os voluntários dos experimentos; aos companheiros de projeto COPPETEC; ao coorientador e amigo: Fellipe Duarte; ao orientador: Geraldo Xexéo; aos professores e membros da banca: Jano Moreira e Eduardo Ogasawara; ao professor: Geraldo Zimbrão;

ao CEFET/RJ/Maracanã, em especial aos professores: Eduardo Bezerra, Gustavo Guedes e Renato Mauro; a todos os queridos funcionários do PESC: Adilson, Ana Paula, Carol, Gutierrez, Itamar, Patrícia, Rosa e Solange; à Marinha do Brasil, em especial a: Albert Lucena, Anderson Vidipó, Bruno Torres, Délcio José, Daniel Marques, Leonardo Pires, Marcelo Castro, Raphaela Pedreira e Thiago Teixeira e a todos aqueles que porventura possa não ter citado aqui, mas que de alguma forma colaboraram durante minha caminhada, muito obrigado!

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

SUPORTE AO MAPEAMENTO SISTEMÁTICO: UM APOIO À PESQUISA BIBLIOGRÁFICA

Pedro Henrique Conilh de Beyssac Ramos

Junho/2016

Orientador: Geraldo Bonorino Xexéo

Programa: Engenharia de Sistemas e Computação

A atividade de revisão bibliográfica trata-se de uma tarefa de grande importância por gerar uma melhor compreensão do conhecimento envolvido em determinados assuntos. Não se trata de um processo realizado somente em pesquisas científicas, apesar de ser nessa área que se torna fundamental sua realização. Entretanto, a execução de revisões demanda tempo e esforço. Apesar da relevância envolvida e relativa dificuldade, a literatura existente demonstra que a ciência que visa apoiar esse processo ainda ainda caminha de forma primitiva. Basicamente, o suporte existente engloba ferramentas que atuam de forma passiva, apresentando as informações geridas, porém não alertando o pesquisador sobre aspectos importantes como, por exemplo, documentos não referenciados.

O presente estudo, destina-se a apresentar um método que tem como objetivo auxiliar a realização de um mapeamento sistemático ou de uma visualização de um domínio do conhecimento (KdViz - *knowledge domain visualization*) com fins de verificação da literatura já explorada, buscando apontar possíveis ausências de referências. Através da rede de citações existente entre documentos, utiliza-se a combinação do algoritmo *Hiperlink-Induced Topic Search*, popularmente conhecido como HITS, com o método de *Louvain* e de métodos para busca e recuperação da informação a fim de efetuar posterior heurística sugestiva.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

SUPPORT TO SYSTEMATIC MAPPING: AN AID TO BIBLIOGRAPHIC RESEARCH

Pedro Henrique Conilh de Beyssac Ramos

June/2016

Advisor: Geraldo Bonorino Xexéo

Department: Systems Engineering and Computer Science

The literature review is an activity of great importance for generating a better understanding of the knowledge involved in certain subjects. This is not a process only performed in scientific research, although it is fundamental in this area. However, the implementation of revisions takes time and effort. Despite the relative difficulty and relevance involved, the literature shows that the science that aims to support this process still walks primitively. Basically, the existing support includes tools that act passively, showing the information managed, but not alerting the researcher on important issues as, for example, documents not referenced .

This study is intended to present a method that aims to support the systematic mapping process or knowledge domain visualization (KDViz) activity. The method helps to verificate the literature explored and seek to identify possible missing references. We use the combination of the algorithm: Hyperlink-Induced Topic Search, popularly known as HITS, Louvain method and information retrieval through the citation network of existing documents to make further suggestive heuristic.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Algoritmos

Capítulo 1

Introdução

1.1 Motivação

A criação de normas para inúmeras atividades realizadas no dia a dia, permitiu a sociedade se organizar e definir métodos e processos em diversas áreas do conhecimento. As normas criadas possibilitaram a refutação de resultados, assim como maior controle no andamento de tarefas. Em especial, isso permitiu o desenvolvimento do raciocínio crítico científico, ou metodologia científica (MEADOWS e DE LEMOS LEMOS, 1999).

A ciência caminha tendo seus avanços registrados em diferentes veículos, como revistas, jornais, congressos e artigos científicos. Porém, independente do meio na qual circulam tais informações, a construção do conhecimento deve ser fundamentada em um método científico. Esse envolve seguir a metodologia científica, a qual engloba, entre outros aspectos, a realização de uma pesquisa bibliográfica acerca do que está sendo pesquisado para se propor um possível avanço (MAIA, 2008).

Uma vez começado um trabalho de pesquisa, com o intuito de compreender o estado da arte da respectiva área de conhecimento de atuação, faz-se necessário realizar um estudo prévio dos trabalhos existentes disponíveis para a comunidade. Muitas vezes, a literatura servirá de argumento para justificar o estudo atual, seja para apresentar possíveis melhorias, provar falhas ou reforçar o já demonstrado previamente. Sem ela não é possível ter parâmetros que qualifiquem os resultados, o que dificultará a obtenção de conclusões acerca do estudo (CIENTÍFICO, 2004).

ALVES (2013) destaca a importância da revisão da literatura para a definição, compreensão e avaliação do problema a ser estudado:

> "A má qualidade da revisão de literatura compromete todo o estudo, uma vez que esta não se constitui em uma seção isolada mas, ao contrário, tem por objetivo iluminar o caminho a ser trilhado pelo pesquisador, desde a definição do problema até a interpretação dos resultados".

Um aspecto que pode resultar no comprometimento citado é a dificuldade em se obter o conjunto mais correto possível de bibliografias relacionadas a um determinado estudo. Isto se dá devido a fatores como :

i) utilização de termos diferentes para denominar um mesmo assunto;

ii) a geração acelerada de produção científica;

iii) ausência de um único concentrador de informações sobre novas criações;

iv) e o idioma no qual um certo documento foi escrito.

Contudo, por mais que se siga um método rigoroso de pesquisa bibliográfica, sempre haverá a chance de não se alcançar determinadas publicações acerca do tema pesquisado. O que pode acarretar o surgimento de viés nos resultados das pesquisas devido à falta de conhecimento necessário. Portanto, é de grande importância que os estudos envolvam além de um bom método para pesquisa bibliográfica, uma forma de verificação do resultado obtido, a fim de minimizar possíveis negligências de estudos correlatos.

Algumas tentativas foram feitas como a de FABBRI *et al.* (2013) que aborda o problema usando mineração de texto para, a partir de *strings* de buscas, descobrir novas referências. Um outro exemplo é a abordagem de CHEN *et al.* (2009) que auxilia na visualização do conhecimento existente (*Knowledge Domain Visualization* - KDViz (BÖRNER *et al.*, 2003)). Através das referências incluídas nos documentos, essa abordagem fornece suporte à visualização do conhecimento envolvido.

Em suma, até o presente momento, apesar dos estudos encontrados se proporem a fazer sugestão de referências utilizando parte dos conteúdos envolvidos, esses não utilizam a estrutura do grafo. Tampouco utilizam tal representação a fim de expandir o grafo inicial com intuito de tentar sugerir novas referências através da análise da estrutura de relacionamentos descoberta na expansão.

1.2 Problema

Este trabalho propõe um método para indicar referências relevantes que foram esquecidas, negligenciadas ou não foram encontradas pelo pesquisador ao longo de seu estudo. Logo, existem dois aspectos que devem ser considerados ao se abordar o presente problema: o **aspecto temporal** e o **alcance da descoberta**.

O aspecto temporal determina se, em um certo momento do tempo, a bibliografia proposta contempla toda a bibliografia relevante ao propósito da busca. Caso a bibliografia atenda ao proposito do estudo, considerando as bibliografias mais atuais sobre seu tema naquele momento, ela é dita como sendo uma "bibliografia completa". Logo a discussão deste aspecto deve responder a duas perguntas:

1) Atualmente a bibliografia proposta está completa?

Onde uma bibliografia é completa se durante a realização da pesquisa o autor considerou todo o universo relevante acerca do tema que está tratando. Isto é, a fim de que seu estudo tenha levado em consideração o estado da arte. O que é fundamental para evitar que resultados e conclusões já obtidas não sejam desconsiderados ou descobertos " novamente ".

2) Ao se passarem inúmeros anos, essa bibliografia ainda estará "completa"?

Um estudo tende a se tornar desatualizado, ao ser finalizado, por conta de novos trabalhos que agregam conhecimento que antes eram desconhecidos. Isto faz com que os estudos finalizados se tornem base para um novo que irá corroborar, refutar ou superar os seus resultados.

No caso de estudos que acabam se tornando conhecimento base para outros, significa dizer que seus conteúdos poderiam ser reescritos igualmente no presente e ainda assim estariam com suas palavras válidas, pois sua bibliografia ainda estaria completa. Porém, no caso dos estudos superados, esses se tornam incompletos, uma vez que novos estudos surjam acerca do que fora publicado, comprometendo reescritas exatas dos estudos sem considerar o novo conjunto bibliográfico.

O outro aspecto que deve ser considerado ao se abordar o tema central desse estudo, trata-se do alcance da descoberta. Esse aspecto representa as características de um conjunto bibliográfico utilizado em um trabalho. Para explicá-lo deve-se responder a duas perguntas:

3) Quão restrita a um tema deve ser a pesquisa?

Um estudo pode envolver diversas áreas do conhecimento. Algumas vezes pode até transitar por ciências diferentes, como é o caso da bioinformática, neurociência computacional, processamento de linguagem natural, entre outras. Ao escrever sobre temas como os citados, é comum que se fale sobre mais de um campo da ciência, por exemplo: ao se elaborar um estudo sobre bioinformática, dificilmente o autor deixará de abordar aspectos sobre genética, tampouco de computação. Afinal, trata-se de um uma área híbrida, que não está restrita a apenas um campo da ciência,que envolve biologia e computação. Nomearemos este aspecto de "largura" envolvida na busca. Logo é esperado que:

> Quanto maior for a largura de uma bibliografia, maior será a quantidade de áreas envolvidas.

4) Quão profundo deve-se pesquisar em um certo tema?

Há estudos que demandam de conhecimento superficial sobre determinados temas, enquanto outros, requerem um esforço maior a fim de se obter conhecimentos mais específicos de uma área. Esse esforço é consequência de ser necessário buscar, em alguns momentos, através de referências dos estudos iniciais, por outras referências que possam expor mais detalhadamente certos pontos iniciais.

Porém, a mesma análise pode ser feita para cada vez que uma nova referência for encontrada. Caracterizando-se assim por uma busca realizada através de conceitos envolvidos em uma mesma área. A característica que representa quão detalhado pode ser um estudo em uma área específica, dá-se o nome de "profundidade".

É interessante destacar que os aspectos (largura e profundidade) estão envolvidos de forma proporcionalmente inversa no resultado geral de uma busca, conforme apresentado na figura 1.1. Por exemplo: se uma busca retorna 10 documentos específicos de uma determinada área, obtemos um resultado geral bem profundo. Porém se a cada 2 documentos existir 1 área distinta representada nesse resultado, então o resultado geral ficou menos profundo e com uma largura maior.

Além disso, uma vez que ao se incrementar a busca em um aspecto eleva-se a quantidade de documentos, consequentemente isso torna mais difícil de se incrementar o outro aspecto simultaneamente. Então, para o resultado geral, pode-se combinar os dois aspectos de três formas: largura elevada e baixa profundidade, pouca largura com alta profundidade e aspectos equilibrados.

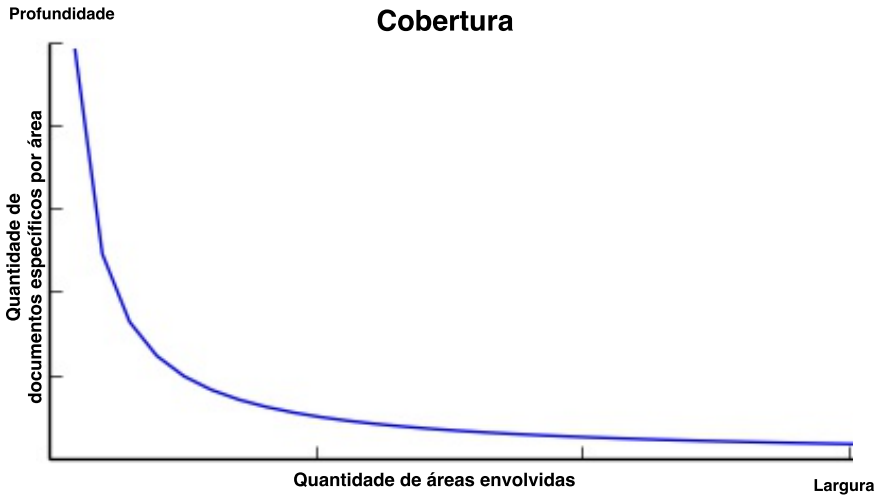

Figura 1.1: Relação lagura × profundidade

A figura 1.2 ilustra um exemplo de estado, em um instante de tempo fixo, em que a bibliografia de determinado estudo se encontra. A partir da figura pode-se pensar em uma modelagem do problema como uma floresta de grafos direcionados onde:

- os **vértices** representam os documentos envolvidos, como por exemplo: artigos, publicações de patentes, livros, revistas ou qualquer documento referenciado por documentos envolvidos na pesquisa analisada que estejam acessíveis.
- as **arestas** representam as referências dos documentos(vértice de destino) que foram citados no vértice de origem.

Observa-se ainda que a bibliografia gerada pelo estudo, no momento em que a figura 1.2 foi gerada, está incompleta. Isto é, existiam documentos relevantes para o estudo (em vermelho) que não foram contemplados ou encontrados pelo pesquisador. Para o propósito do presente trabalho chamaremos o conjunto de documentos que não foram contemplados, em vermelho na figura, de **Vértices não explorados** enquanto o conjunto de documentos que foram contemplados, em azul, serão chamados de **Vértices conhecidos**. Logo, um problema interessante a ser estudado é:

> Identificar **Vértices não explorados** da bibliografia de uma pesquisa a partir de conjunto inicial de **Vértices conhecidos** da pesquisa.

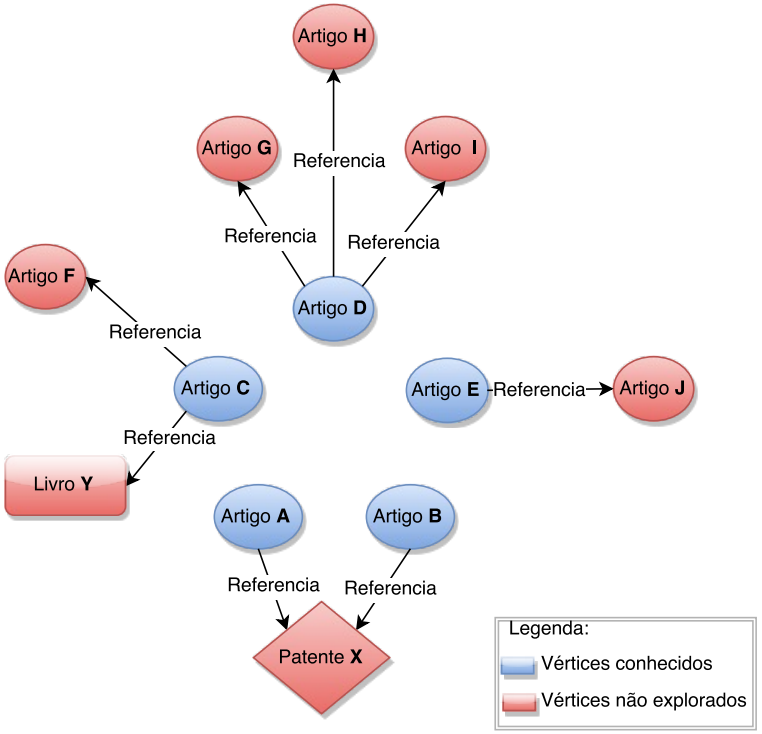

Figura 1.2: Estado da bibliografia de um estudo em um instante t

1.3 Proposta

O presente trabalho explora estratégias, do aspecto **alcance de descoberta**, de largura e profundidade combinadas com o aspecto **temporal** para propor um

método que identifique **Vértices não explorados** a partir de conjunto inicial de **Vértices conhecidos** da bibliografia de uma pesquisa. Para tanto, as duas primeiras perguntas da seção 1.2 serão utilizadas como ponto de avaliação do método proposto[1]. Enquanto as duas perguntas subjacentes serão utilizadas para se discutir, e combinar, estratégias de largura e profundidade que atendam ao propósito em que o método será utilizado que chamaremos de **propósito da pesquisa**.

Portanto a hipótese do presente trabalho é:

> A partir de um **propósito de pesquisa** definido e de um conjunto inicial de **Vértices conhecidos** o método proposto identifica **Vértices não explorados** da bibliografia de uma pesquisa.

Dito isto, o método proposto busca alcançar o objetivo mencionado a partir de três pontos chaves a serem considerados:

1. A relevância de cada referência envolvida: utilizando o algoritmo HITS (*Hiperlink-Induced Topic Search*) (KLEINBERG, 2000), também conhecido como *Hubs and Authorities*.

2. Agrupamento de referências por áreas semelhantes: utilização do método de *Louvain* (BLONDEL *et al.*, 2008) direcionado a encontrar comunidades que servirão de critérios de agrupamento.

3. Expansão dos **Vértices conhecidos** iniciais: A partir das informações presentes nos vértice combinadas com consultas em sistemas de busca que indexam tanto os **vértices conhecidos** quanto os **vértices não explorados**.

É interessante destacar que o HITS gera pontuações que possibilitam quantificar a relevância relativa a cada vértice de acordo com a estrutura do grafo formado. O algoritmo permite estabelecer dois tipos de pontuação por vértice: medida de autoridade e medida de concentração. Sendo a primeira medida a ser utilizada pelo presente algoritmo devido à sua característica de representar vértices importantes conforme demonstrado pelo próprio autor do algoritmo HITS em um estudo aplicado ao cenário de páginas na *internet* (KLEINBERG, 1999).

Além disso, a utilização do método de *Louvain* possibilita encontrar agrupamentos de acordo com a densidade de arestas entre os vértices. O que, em outras palavras, permite agrupar de acordo com o grau de relacionamento entre os vértices envolvidos, resultando em uma provável separação por assunto.

Por fim, a Expansão dos **Vértices conhecidos** permite descobrir, a partir das informações dos vértices conhecidos, novos vértices que fazem parte de uma rede implícita de autoridades, em certos assuntos, mesmo que essa informação não esteja presente na bibliografia inicial.

[1] e serão quantificadas e avaliadas nos experimentos

1.4 Contribuições

As contribuições destes trabalhos podem ser resumidas aos itens abaixo:

- Motivar discussões e comparações para promover maior evolução do tema envolvido, que segundo a literatura existente, encontra-se em um estado de imaturidade (MARSHALL e BRERETON, 2013b).
- Criação de um método computacional que auxilie em pesquisas bibliográficas e seja capaz de prover um auxílio para descobrir referências negligenciadas em estudos em andamento e concluídos.
- Aplicação do método proposto à construção de uma ferramenta que permita a pesquisadores utilizá-la a fim de auxiliar em suas buscas e validações de seus materiais bibliográficos.

1.5 Organização

Este estudo está organizado em 6 capítulos. O capítulo 1 propôs uma discussão acerca do tema com fins motivacionais e informativos. O capítulo 2, aborda conceitos básicos para o entendimento dos demais tópicos, utilizando uma metalinguagem, busca explicar o estado da arte em ferramentas e técnicas de apoio ao processo de revisão da literatura. Além disso, disserta sobre como esse tipo de tópico pode ser escrito, ter sua produção facilitada e até mesmo verificada. Concluindo, será explicado o conceito existente utilizado para a tentativa de melhoria do processo abordado. Em seguida, há o capítulo 4 no qual será explicada a arquitetura utilizada pelo trabalho. O capítulo 5 irá apresentar os resultados obtidos para posterior disclusão no capítulo 6 onde será dissertado sobre as conclusões. O capítulo 6.5 destina-se as ideias que foram deixadas para possíveis implementações futuras. Por fim, são mostradas as referências, sementes de toda disclusão envolvida nesse estudo.

Figura 1.3: Mapa mental da organização envolvida

Capítulo 2

Revisão da Literatura

2.1 Revisão da literatura, tipos e conceitos

Os conceitos de revisão da literatura e a pesquisa bibliográfica apresentam uma proximidade que pode levar a uma confusão que, consequentemente, criará uma dificuldade de se diferenciar os dois. Apesar de sua proximidade, são termos que representam processos distintos e requerem níveis de conhecimento bem diferentes.

Uma pesquisa bibliográfica tem como objetivo obter informações sobre bibliografias existentes, atuando como um processo de construção para auto conhecimento e dispensando conhecimento prévio sobre a área pesquisada (CALDAS, 1986). Por outro lado, realizar uma revisão da literatura existente exige um aporte prévio de conhecimento. A revisão da literatura é um procedimento no qual se adiciona de fato conhecimento à área pesquisada, através da dissertação de conceitos aprendidos, agregando conclusões, interpretações e questionamentos(CENDÓN *et al.*, 2000). Portanto a revisão da literatura engloba a pesquisa bibliográfica.

A revisão de literatura vai além de ser um título de capítulo em trabalhos finais, dissertações de mestrado ou teses de doutorado. Trata-se de um processo que pode ser realizado de diversas formas. Porém, segundo BOTELHO *et al.* (2011), pode-se dividí-lo em duas principais categorias: as revisões narrativas e as revisões sistemáticas.

As revisões narrativas, como explicado por CORDEIRO *et al.* (2007), são revisões que não estão sujeitas a critérios rigorosos de busca, tampouco seguem um método de busca que vise responder a uma questão específica. Nesse tipo de revisão, o critério de escolha dos artigos é determinado apenas pelo próprio autor, que devido à falta de regras para o processo pode acabar por comprometer a imparcialidade do estudo através de conceitos próprios. Portanto, não se trata de uma metodologia muito adequada à realidade científica, uma vez que para poder validar um estudo a comunidade científica precisa poder reproduzir o trabalho realizado, o que também

inclui sua revisão da literatura.

A partir da necessidade de se realizar um processo que pudesse ser reproduzido a fim de ser validado, surge a revisão sistemática (KITCHENHAM *et al.*, 2009). Essa categoria visa responder a certas questões relacionadas ao tema central de pesquisa com intuito de realizar um levantamento dos trabalhos mais relevantes produzidos na área. Através de uma metodologia rigorosa, o método provê uma forma confiável e reprodutível dos resultados obtidos a partir de sua aplicação para realização de uma revisão da literatura(FABBRI *et al.*, 2013).

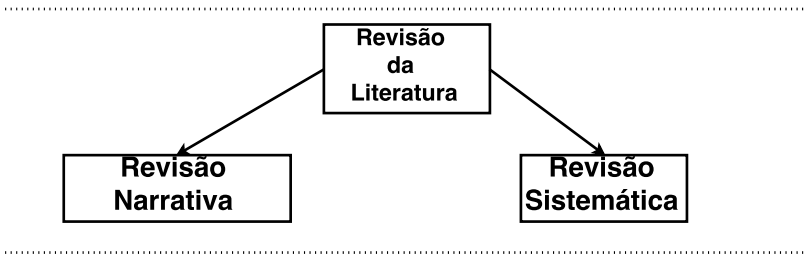

Figura 2.1: Estrutura das categorias segundo BOTELHO *et al.* (2011)

2.2 Revisões sistemáticas

A revisão da literatura de forma sistemática pode ser dividida entre quatro principais métodos(BOTELHO *et al.*, 2011) :

i) Revisão Sistemática

ii) Meta-Análise

iii) Revisão Qualitativa

iv) Revisão Integrativa

A revisão sistemática, do inglês: *Systematic Literature Review* (SLR), é o metodo mais simples entre os quatro métodos citados, e KITCHENHAM (2004) define as etapas, a serem seguidas rigorosamente, para realizar corretamente a pesquisa bibliográfica. Tal método busca, através da pesquisa bibliográfica sistemática, cobrir de forma profunda os conceitos envolvidos no tema foco. Logo, a revisão sistemática trata de uma busca para estabelecer o estado das evidências encontradas acerca do assunto em foco(PETERSEN *et al.*, 2008).

Segundo BOTELHO *et al.* (2011) e CORDEIRO *et al.* (2007) a meta-análise é um método que possui embasamento matemático e estatístico para realização do processo de revisão. Esse método também realiza resumos e agrupamentos de forma a se calcular a participação dos conceitos envolvidos na pesquisa. Nas palavras de

GLASS (1976) a meta-analise é "a análise estatística de uma coleção de resultados de estudos individuais, com o objetivo de integrar os resultados".

A revisão qualitativa segundo BOTELHO *et al.* (2011) :

> "sintetiza **exclusivamente** os estudos primários qualitativos, podendo diferir em abordagens e níveis de interpretação."

Onde entende-se o estudo primário como: "um estudo empírico que investiga uma questão específica de pesquisa" (KEELE, 2007). Estudo esse no qual o autor foi pioneiro no trabalho com os dados envolvidos.

O método de revisão integrativa agrupa estudos empíricos e teóricos(BOTELHO *et al.*, 2011) de forma mais abrangente acerca de determinados temas. Não se restringido a estudos primários. Esse tipo de revisão permite ao pesquisador realizar levantamento sobre o que já foi feito dissertando sobre os resultados de outros estudos, que podem até mesmo ser revisões, agregando suas conclusões e ideias.

A figura 2.2 organiza a *big picture* do métodos de revisão da literatura:

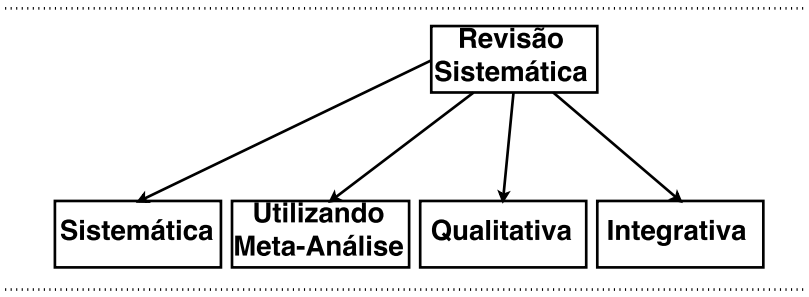

Figura 2.2: Estrutura dos métodos existentes segundo BOTELHO *et al.* (2011)

Todos os métodos citados agregam pontos interessantes à revisão de literatura existente. Porém o presente estudo irá focar no método sistemático devido ao seu maior rigor metodológico. Todo esse rigor científico e maior compatibilidade com a ciência, pode ser percebido através do estudo de KEELE (2007).

Como já explicado na seção 2.1, para se escrever uma revisão da literatura, seja ela utilizando qualquer um dos métodos citados, é necessário que antes já tenha havido um aprendizado sobre a área tema. Para isso, faz-se necessária a realização de uma pesquisa bibliográfica. No entanto, a pesquisa bibliográfica, ao estar pautada em auxiliar na revisão da literatura, acaba por ter características do método adotado. No caso da sistemática, a revisão acaba por focar no estado das evidências (documentos) existentes, como citado por PETERSEN *et al.* (2008). E, portanto, a revisão sistemática apresenta um comportamento guloso na direção de um assunto para busca de novos vértices (documentos) para a bibliografia o que favorece o aspecto de profundidade discutido na seção 1.2.

Em contrapartida, há uma forma de busca chamada mapeamento sistemático section 2.4, também conhecida por estudo de escopo (KEELE, 2007). Trata-se de uma forma de estudo que visa buscar de forma ampla (analogamente, como conhecidamente em ciência da computação, ao que uma busca em largura executa) o conhecimento existente ao redor de um tema sem se aprofundar nos detalhes de cada trabalho encontrado (PETERSEN *et al.*, 2008).

Portanto, percebe-se que tanto a pesquisa bibliográfica quanto o mapeamento sistemático possuem caráter exploratório com objetivo de observar os estudos já realizados em determinadas áreas. De forma geral, os dois métodos de busca diferenciam-se pelo tipo de objetivo a alcançar, tipo de busca realizada e o processo utilizado (PETERSEN *et al.*, 2008). Ainda segundo PETERSEN *et al.* (2008), há a recomendação da utilização de um método sendo complementar ao outro, vide figura 2.3. Fica claro que primeiramente um mapeamento sistemático pode ser usado como ferramenta para se ter uma visão mais ampla do que se busca e em seguida pode-se realizar a pesquisa bibliográfica que trará um caráter mais específico ao resultados encontrados.

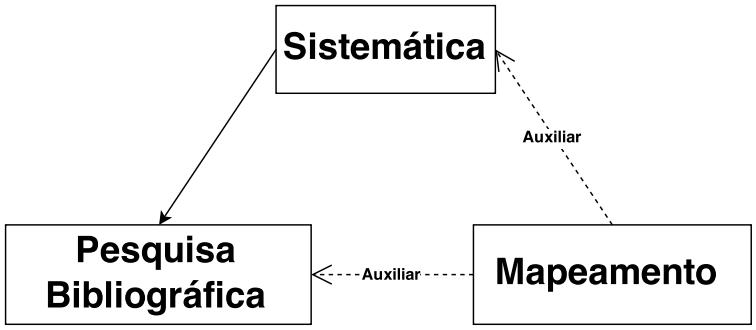

Figura 2.3: Sub-método obrigatório e opcional

2.3 Pesquisa bibliográfica

A pesquisa bibliográfica, como citada anteriormente em 2.1, trata-se de uma das etapas para realização da revisão da literatura. Antes de dissertar acerca de um assunto, é necessário ter conhecimento sobre os trabalhos existentes na área, termos e conceitos periféricos.

O tipo de pesquisa a ser executada pode variar de níveis mais informais, como o caso de trabalhos colegiais iniciais, nos quais muitas vezes os alunos ainda não dispõem de maturidade suficiente para entender uma metodologia a ser rigidamente seguida, a níveis mais rígidos, como trabalhos científicos nos quais se necessita seguir um método que dê respaldo as afirmações feitas. E uma dessas formas é de forma sistemática, ou seja, seguindo passos preestabelecidos para se alcançar um objetivo.

Passos esses, planejamento e condução da revisão, descritos por KITCHENHAM (2004) em seu processo de revisão sistemática da literatura:

- Planejamento

  Desenvolvimento do protocolo de revisão

- Condução da revisão

  Identificação dos estudos

  Seleção dos estudos

  Estudo de Avaliação da Qualidade

  Extração dos Dados

  Síntese dos Dados

- Publicação dos Resultados (Etapa de Revisão da Literatura)

Como pode-se extrair de KITCHENHAM (2004) e observar através da figura 2.4: através das fases descritas acima, o pesquisador irá primeiramente executar o planejamento de sua pesquisa, isto é, definir que tipo de perguntas pretende responder. Por Exemplo: quais algoritmos já foram utilizados para resolução de um certo problema? Ou então: que ferramentas existem para lidar com uma questão existente?

Uma vez definido o objetivo da pesquisa, será conduzida a revisão bibliográfica propriamente dita. A revisão buscará executar em sequência: a identificação dos estudos existentes através da estratégia de busca a ser definida (escolha das *strings* de busca utilizando operadores booleanos. Por exemplo: revisão sistemática *AND* mapeamento sistemático), seleção dos estudos que o pesquisador julgar necessária uma análise mais profunda de acordo com sua relevância, avaliar a qualidade buscando através de parâmetros estabelecidos minimizar o viés existente na busca, realizar a extração dos dados definindo que tipo de dados irão servir de entrada para a posterior síntese de informações (Campos como data da publicação, título, autores ou qualquer metadado envolvido) e para finalizar a etapa de condução da revisão, haverá a síntese dos dados obtidos, essa síntese trata-se do processo de transformar dados em informação através do resumo das ideias encontradas. Nesse ponto, encerra-se o processo de pesquisa bibliográfica seguindo os conceitos descritos por CALDAS (1986)

Como última etapa há a condução da escrita em meios de comunicação como conferências, artigos periódicos, teses ou outros possíveis meios de divulgação científicos. Aqui vale ressaltar que essa etapa pode ser considerada parte da a revisão

da literatura por permitir que o autor da pesquisa agregue valores ao que foi encontrado através de todo entendimento adquirido (CENDÓN *et al.*, 2000).

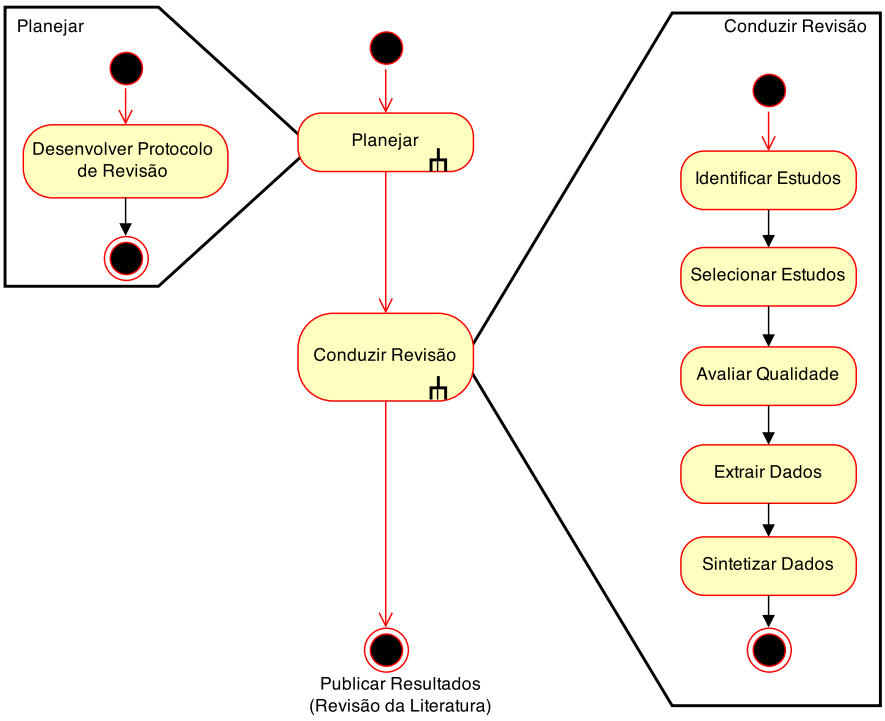

Figura 2.4: Etapas de pesquisa bibliográfica

Vale ressaltar que o presente estudo não busca estudar a fundo as etapas que constituem o processo sistemático, fica a cargo do leitor buscar entendê-las melhor no estudo de KITCHENHAM (2004). De forma resumida, entende-se que o processo descrito irá selecionar diversos documentos relacionados à área de estudo sendo pesquisada, os quais sofrerão uma triagem. Os documentos, que de fato forem julgados relevantes para o presente estudo, serão selecionados para uma leitura integral minuciosa.

O ato de realizar uma pesquisa bibliográfica, entre outros objetivos, destina-se a possibilitar o aprendizado acerca de conceitos básicos do que está sendo pesquisado. Isso proporciona ao pesquisador, foco no conteúdo e reflexões que virão pela frente. Porém, além disso, trata-se de um processo muito importante porque através dele também são encontrados os trabalhos correlatos.

Os trabalhos correlatos são aqueles que o autor pesquisou ou deveria ter pesquisado para saber o que havia de mais atual relacionado ao objetivo de sua pesquisa. Os melhores resultados encontrados são conhecidos popularmente por representarem o estado da arte.

Uma pesquisa que não utiliza o conhecimento do que já existe para se embasar, pode estar fadada ao fracasso (PETTICREW e ROBERTS, 2008). Isso se justifica pelo fato do autor pode estar por recriar algo ou produzir resultados já superados. Portanto, a descrição desses trabalhos acaba sendo parte essencial da justificativa de um trabalho em questão (WEBSTER e WATSON, 2002). O seguinte trecho de Richard Hamming ilustra bem essa questão:

"Perhaps the central problem we face in all of computer science is how we are to get to the situation where we build on top of the work of others rather than redoing so much of it in a trivially different way. Science is supposed to be cumulative, not almost endless duplication of the same kind of things".

Richard Hamming 1968 Turing Award Lecture

Uma maneira tradicional de realizar essa revisão, é a pesquisa pura através de livros e consultas diretas a documentos físicos em bibliotecas sem utilizar meios eletrônicos. Porém, o surgimento de ferramentas de busca textual para acadêmicos, como *Google Scholar*, *Citeseer*, *Ieee Xplore*, fez com que esse processo se tornasse um pouco menos árduo. Isso ocorre porque, através da rede mundial de computadores, um pesquisador é capaz de selecionar diversos textos, possivelmente relacionados ao seu tema, podendo decidir se a leitura é válida sem precisar se deslocar entre lugares físicos distintos.

Contudo, ainda assim as ferramentas citadas não eliminam o último passo do processo, ler o texto superficialmente para saber se a leitura integral é algo a se investir tempo ou se deve ser descartada. Esse processo de mineração pode acabar por despender tempo que será gasto sem gerar conteúdo ou conhecimento direto para o trabalho em questão. Trata-se de um processo cansativo e sujeito a falhas.

Todavia, com a tecnologia, não surgiram somente os buscadores textuais, mas também ferramentas que visam auxiliar e dar suporte a pesquisadores para área de pesquisa bibliográfica através do gerenciamento de referências e do auxilio para revisão sistemática da literatura. A organização de referências ao se realizar uma pesquisa bibliográfica é fundamental para não se perder a forma de citar a origem de onde um conceito foi retirado ou até mesmo o próprio documento. Portanto, ferramentas da seção 2.6 voltadas para esse propósito combinam armazenamento de referências já utilizadas, permitindo que o usuário possa fazer um *link* com os arquivos referentes. Com isso a forma de citar se torna padronizada e o acesso aos documentos originais de onde foram extraídas as citações é facilitado.

O método de revisão sistemática acaba sendo facilitado pela existência de organização de referências acessadas durante pesquisas. Porém isso não é o suficiente, pois como já apresentado em seu método de pesquisa bibliográfica (seção 2.3), há muitos passos a serem seguidos e etapas rígidas a serem cumpridas. Devido a isso,

há a necessidade da existência do campo de pesquisa de ferramentas (seção 2.6) também voltadas para esse propósito a fim de permitir que o pesquisador realize um estudo mais coeso e ao máximo livre de viés.

Entretanto, ainda havendo as duas formas de auxílio citadas, persiste a questão do caráter não amplo da pesquisa bibliográfica na revisão sistemática. Devido a isso, uma pesquisa pode-se tornar custosa em termos de tempo ou até limitada a não visualizar possíveis relações com outros temas. Como mencionado na seção de revisões sistemáticas (seção 2.2), o mapeamento sistemático pode auxiliar nessa questão.

> *"Mapping Studies may be requested by an external body before they commission a systematic review to allow more cost effective targeting of their resources. They are also useful to PhD students who are required to prepare an overview of the topic area in which they will be working".* (KEELE, 2007)

2.4 Mapeamento sistemático

O Mapeamento Sistemático, ou estudo de escopo, trata-se de um método que visa identificar os estudos existentes acerca de um tema (KEELE, 2007). Segundo PETERSEN *et al.* (2008), "o propósito principal do mapeamento sistemático é prover uma visão geral da área de pesquisa, identificar a quantidade, tipo de pesquisa existente e resultados envolvidos". Uma de suas características é prover um estudo mais abrangente que o estudo de pesquisa bibliográfica. Sua ideia é analisar o universo que circunda as relações existentes entre as publicações.

Porém, não está claro, ainda, como o mapeamento sistemático deve ser realizado. Como citado por PETERSEN *et al.* (2008), o exemplo mais claro fora encontrado no estudo de BAILEY *et al.* (2007). Devido a isso, esse estudo irá considerar, assim como em DURELLI *et al.* (2010), as etapas definidas por PETERSEN *et al.* (2008) como as etapas a serem seguidas para se realizar um mapeamento sistemático (do inglês: *Systematic Mapping* (SM)):

- Definição do objetivo da busca
- Condução da busca
- Exibição dos documentos
- Classificação usando palavras chaves
- Extração dos dados e processo de mapeamento

Assim como em 2.3, o primeiro passo a ser definido é que tipo de questão deseja-se responder. Ou em outras palavras, qual escopo deseja-se cobrir. Por exemplo: que artigos falam acerca de mapeamento sistemático? Quais artigos falam sobre revisão sistemática? Quais autores trabalham com mapeamento sistemático? Quais Ferramentas existentes para área de revisão da literatura? Diversas são as possíveis perguntas que podem ser feitas pelo pesquisador nessa etapa.

Após definir o objetivo do estudo, a condução da busca se dará através da escolha de *strings* de busca. *Strings* essas formadas como em 2.3 no passo "Condução da revisão", por exemplo: (*Systematic mapping* and *Software engineering*) OR (*SM*). Essa passo poderá ser realizado de forma manual ou através de bases de dados com mecanismos de busca automatizados.

Em seguida, os documentos encontrados são mostrados para passarem por critérios de inclusão e exclusão definidos pelo autor da busca. Os critérios podem ser os mais diversos. Como por exemplo para exclusão: somente manter os artigos publicados entre os anos de 2010 e 2014. Ou, mostrar somente os artigos submetidos a conferências internacionais. Em teoria, o autor define o processo, porém não deixa explícito restrições a utilização de mecanismos automáticos para auxiliar nessa etapa.

A etapa de classificação usando palavras chaves, visa reduzir o tempo gasto para agrupar os documentos encontrados através de grupos. A ideia por trás dessa etapa é seguir dois passos descritos pelo autor PETERSEN *et al.* (2008) : "Primeiro, o autor da busca irá ler os resumos e irá procurar por palavras chaves e conceitos que reflitam a contribuição do documento. Enquanto estiver realizando esse passo, o autor da busca também irá identificar o contexto de pesquisa dos documentos. Quando isso estiver pronto, o conjunto de palavras chaves de diferentes documentos serão combinados para formar um conjunto de alto nível de conhecimento da natureza e contribuição dos estudos envolvidos".

Por fim, a extração de dados irá permitir o pesquisador inferir conhecimento acerca de todo dado coletado na mineração de informação envolvida. O objetivo nessa etapa é analisar as informações através de mapeamentos que facilitem a percepção das mesmas. Métodos estatísticos e visuais são bem-vindos nessa etapa.

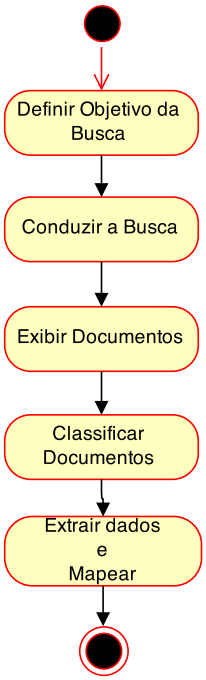

Figura 2.5: Método sistemático de mapear estudos

De forma geral percebe-se que o mapeamento sistemático se assemelha à pesquisa bibliográfica quando realizada de forma sistemática apresentada em 2.3. Porém, como citado por FABBRI *et al.* (2013), uma de suas principais diferenças é que no mapeamento sistemático os documentos não são lidos por inteiro, mas sim apenas seus resumos.

A diferença citada permite que o pesquisador otimize seu tempo e se foque mais na extensão dos estudos existentes, uma vez que passa a ser possível inspecionar uma maior quantidade de estudos. Nesse método o objetivo a ser alcançado é diferente do objetivo a ser atingido em uma pesquisa bibliográfica sistemática, portanto, isso se reflete nas etapas realizadas e profundidade dos dados pesquisados.

2.5 Visualização de domínio do conhecimento

O estudo da visualização de domínio do conhecimento, do inglês *Knowledge Domain Visualization* - KDViz, engloba múltiplos tipos de análises e representações de conhecimento, conforme abordado por: BÖRNER *et al.* (2003). Dentre os tipos de analises realizadas por essa área estão: a cientométrica, a bibliográfica e análises de citações, que podem ser melhor entendidas por: DOS SANTOS e KOBASHI (2009). Além dos tipos de análises mencionados, essa área está fortemente focada em prover formas de representação visual das informações analisadas.

O objetivo de estudos dessa área se caracteriza por ter um escopo mais amplo que o analisado pelo mapeamento sistemático. Seu objetivo é fazer uma análise das publicações envolvidas de forma tanto qualitativa quanto quantitativa. Seu foco é entender melhor áreas analisadas, observar tendências e compreender a dinâmica científica associada a um contexto.

Assim como para realizar o mapeamento sistemático é necessário seguir determinados passos previamente descritos, para realizar uma visualização do domínio do conhecimento também foram estabelecidos determinados passos. Esses são demonstrados pelo estudo BÖRNER *et al.* (2003) conforme exibidos pela figura 2.6 e podem ser entendidos como sendo realizados em cada passo as seguintes ações:

1) extrair dados. Para isso o pesquisador precisará ter acesso a uma ou mais bases de dados, assim como definir como realizará suas buscas (por exemplo: através de citações, termos ou referências);

2) definir qual será o foco da análise, como por exemplo, os autores envolvidos, os documentos, os meios onde foram publicados os documentos envolvidos ou termos utilizados;

3) definir que tipo de medidas serão utilizadas. São exemplos disso: co-citação, ano de publicação dos documentos e citação dos autores;

4) definir como será realizado o cálculo da similaridade entre as medidas previamente descritas. Para esse fim, pode-se utilizar diversas técnicas de mineração de dados como: *singular value decomposition* (SVD), *latentic semantic analisys* (LSA) e *vector space models*;

5) definir como os dados deverão ser processados para a apresentação. Isso quer dizer que se definirá se será realizada uma redução de dimensionalidade, separação dos dados em *clusters* ou uma análise escalar;

6) definir qual será a forma visual a representar as informações finais.

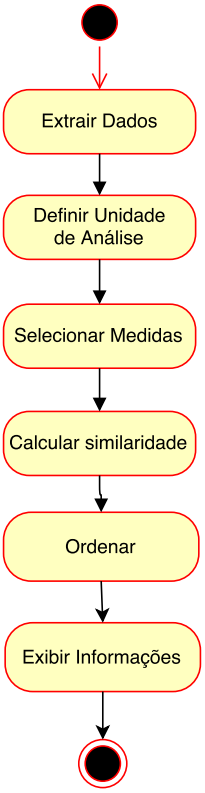

Figura 2.6: Passos para obtenção de visualização de um domínio do conhecimento

Porém esse tipo de estudo também pode ser aplicado com fins de mapeamento bibliográfico para auxiliar diretamente em pesquisas. Isso é possível devido à sua característica de combinar a representação visual e análise dos dados envolvidos.

A visualização de domínio do conhecimento se torna similar ao mapeamento sistemático quando voltada para fins de verificação da bibliografia já explorada. Essas duas formas de estudo diferem-se no nível de investigação sobre um assunto a ser realizado e o rigor metodológico utilizado.

Portanto para o caráter da presente dissertação, ambas as formas de estudos, da presente seção e da seção 2.4 sobre mapeamento sistemático, podem ser direcionadas a auxiliar na realização de pesquisas bibliográficas. O que torna o presente estudo relacionado a esses conceitos no âmbito de representação do conhecimento e auxílio para os tipos de análises citados.

2.6 Trabalhos correlatos

Uma vez explicado os conceitos e a teoria por de trás dos temas que circundam o objetivo desse trabalho, nessa seção serão abordados os trabalhos ferramentais ligados a esses assuntos. Serão abordadas suas vantagens e desvantagens a fim de contextualizar a proposta desse estudo.

Em diversos campos do conhecimento, faz-se notória a colaboração da tecnologia de informação e computação através de ferramentas de apoio, e com área de escrita científica não é diferente. Atualmente existem aplicações computacionais que visam ajudar o escritor na gerência para referência correta do que foi citado em seus textos (gerenciadores de referências), assim como na visualização da rede de conhecimento que envolve um artigo sendo escrito (SM, conforme descrito na seção 2.4) e também no auxilio direto à pesquisa bibliográfica seguindo um método sistemático, conforme descrito na seção 2.3.

Das três áreas de aplicações envolvidas, o gerenciamento de referências não possui uma seção a parte nesse estudo por se tratar de algo basicamente técnico. Sua ideia é fornecer suporte através de técnicas de armazenamento de dados e integração com editores de texto, navegadores e em alguns casos até redes sociais.

Visando suprir tais necessidades, existem ferramentas como Zotero (VANHECKE, 2008) e Bibsonomy (JÄSCHKE *et al.*, 2007), entre outras muitas citadas em *Defrosting the Digital Library: Bibliographic Tools for the Next Generation web* (HULL *et al.*, 2008). Essas ferramentas, apesar de simples em relação a parte teórica/científica, ecoam de forma grandiosa no meio acadêmico, uma vez que se tornam essenciais para o gerenciamento de referências adequado à escrita de documentos com uma lista média e longa de referências a ser mantida.

Diversos são os recursos oferecidos por essas ferramentas: criação de base de dados de referências, facilitando assim a escrita de referências em citações e ideias de outros, importação e exportação de dados, sincronia de dados em diferentes plataformas, compartilhamento de dados, indexação dos documentos referenciados, entre outras capacidades técnicas que facilitam o trabalho de um pesquisador e permitem que ele consiga focar esforço no que é realmente produtivo.

Porém, os gerenciadores são ferramentas que não permitem visualização do tipo de relação existente entre as referências utilizadas, tampouco estão relacionadas ao processo de revisão sistemática, servindo apenas de auxílio. São de certa forma estáticas no sentido de não realizar mapeamento dos dados através de imagens que permitam o pesquisador ter uma visão macro do mundo que está pesquisando.

Esse fato citado abre margem para a importância das ferramentas de mapeamento sistemático. Sua teoria é utilizada por ferramentas que podem utilizar uma ou mais bases de dados para traçar ligações implícitas entre diversos tipos de publicações científicas a fim de fornecer ao usuário uma visão macro do universo já publicado acerca de seu tema buscado. As possíveis ligações existentes entre as pesquisas formam um mapa do conhecimento envolvido.

Ferramentas voltadas para esse propósito se propõem a exibir ligações entre publicações através de suas citações, referências ou quaisquer metadados constituintes do documento indexado em sua base de dados, dando suporte a processos de revisão

sistemática da literatura ou a pesquisas bibliográficas.

Nesse nível de conhecimento envolvido, há ferramentas como *CiteSpace* (CHEN *et al.*, 2009), StArt (*State of the Art thought Systematic Review*) (FABBRI *et al.*, 2013), SLuRp (*Systematic Literature unified Review program*) (BOWES *et al.*, 2012) e PaperCube (BERGSTRÖM e ATKINSON, 2009) que utilizam de métodos computacionais a fim de exibir para o usuário informações que possibilitem interpretar a ciência por volta do que está sendo pesquisado através de variados mecanismos de exibição de dados acerca de cada área envolvida na busca.

Ainda poderia ser citada a ferramenta SLR *tool*, porém a mesma só funciona em um determinado sistema operacional e apenas em inglês e espanhol, como citado por FABBRI *et al.* (2013).

Junto à ferramenta StArt citada, advém o auxílio ao processo de SLR e SM. Sendo essa ferramenta uma das mais completas existentes até o presente momento para fins de SLR como mostrado em MARSHALL e BRERETON (2013a) e FABBRI *et al.* (2013).

A StArt visa auxiliar tanto no processo de mapeamento do conhecimento existente, quanto auxiliar na parte de seleção dos artigos relevantes ao usuário. Através de pontuações acerca da relevância de cada documento apontado na ferramenta, o usuário possui uma visão mais organizada e facilitada para classificar quais documentos são de fato relevantes e devem ser lidos integralmente e quais devem ser descartados.

A SLuRp, trata-se de uma ferramenta similar a StArt, que possui mecanismos, segundo BOWES *et al.* (2012), de exibição de síntese de dados variados, desde gráficos diversos a exibição tabular dos dados em HTML ou LaTex. Apesar da similaridade e características relevantes, a mesma não fora citada no estudo de FABBRI *et al.* (2013) em 2013, entretanto, encontra-se citada em comparativo (MARSHALL e BRERETON, 2013a) do mesmo ano.

É importante citar que a ferramenta *CiteSpace* não é voltada diretamente para realização de mapeamento sistemático, mas sim visualização do conhecimento existente, em inglês conhecido como *knowledge domain visualization* ou KDViz (CHEN, 2004). Seu foco está em visualizar tendências no domínio de publicações utilizando análises estruturais e temporais acerca do conhecimento envolvido.

Apesar disso, essa ferramenta pode ser utilizada para a última etapa do mapeamento sistemático 2.5: extração dos dados e processo de mapeamento. Esse fato ocorre pois essa ferramenta possui um funcionamento diferenciado, além de mineração de dados em texto, ela inclui técnicas avançadas de extração do conhecimento.

Outra característica do *CiteSpace* é possuir a capacidade de focar nos relacionamentos existentes entre os documentos representados, analisando assim relaciona-

mentos de co-citação entre autores, redes de colaboração entre instituições e outras características intrínsecas dos relacionamentos entre os dados envolvidos. Devido a todas essas características, essa ferramenta possui um potencial a ser analisado em torno de suporte ao mapeamento sistemático.

Entretanto, pouco-se observa esse último tipo de abordagem nas ferramentas atuais, haja vista os resultados obtidos no estudo MARSHALL e BRERETON (2013a). Além disso, a partir desse estudo, pode-se entender que apesar de existirem ferramentas como as citadas, elas ainda se encontram em um estado primitivo de construção.

Basicamente, as ferramentas utilizam mineração de dados nos textos de resumos dos documentos apontados e tentam extrair informação que vise ajudar a classificar os documentos. Além de utilizarem técnicas de visualização de dados para melhor representação do conhecimento.

Portanto, a ferramenta *CiteSpace* se destaca nesse cenário por realizar um tratamento diferenciado ao analisar os relacionamentos implícitos existentes entre os documentos. Um dos principais pontos de apoio da teoria por trás da ferramenta *CiteSpace* trata-se dos nós *Pivots*:

> *"Pivot nodes have an essencial role in our method"* (CHEN, 2004).

Como explicado pelo autor:

> *"Pivot nodes are joints between different networks; they are either the common nodes shared by two networks or the gateway nodes that are connected by internetwork links".*

O autor continua por dizer acerca da importância do estudo dos *pivots*:

> *"This could be a particularly useful feature for the detection of significant articles that could be easily overlooked by falling below a single high-citation threshold".*

A forma de calcular esses *pivots* é descrita em um estudo posterior (CHEN, 2005), no qual o autor descreve que os *pivotal points* são calculados através da métrica de *bewteeness centrality* que é computada para cada nó, sendo sua implementação baseada no algoritmo de BRANDES (2001).

Como dito, certo pontos em uma rede implícita de ligações existentes entre documentos que se referenciam, podem ser úteis para auxiliar na tarefa de encontrar possíveis referências esquecidas por um autor de um documento que esteja nessa rede ou ser mapeado de alguma forma dentro dela.

É importante citar que apesar da semelhança com a área de sistemas de recomendação o presente trabalho não destina-se a atuar seguindo essa área, assim como

tampouco se enquadra nas técnicas gerais de recomendações para sistemas científicos como descritas em MANOUSELIS *et al.* (2011).

O estudo não entrará em pormenores nas características que o faz não ser qualificado como um sistema de recomendação seguindo as técnicas comuns da área, porém 2 características buscadas por esse trabalho já permitem demonstrar o porquê esse não utiliza técnicas de sistemas de recomendação. Uma delas é: buscar sugerir itens (que podem ser encarados como referências) que não necessariamente são similares em termos de descrição.

Esse fato é incompatível com uma das 3 classificações de categorias descritas por MANOUSELIS *et al.* (2011), nesse caso o *Content-Based Recommender Systems*, uma vez que não utiliza diretamente a descrição do item para recomendar, mas sim seus relacionamentos, e ainda que a utilizasse, essa técnica não possui o problema de somente indicar itens similares, como descrito em MANOUSELIS *et al.* (2011) para *Content-Based Recommender Systems*: *"The problem that only similar items are recommended"*.

A outra classificação de sistemas de recomendação que o presente trabalho poderia ser qualificado seria a: *Collaborative Recommender Systems*. Porém, nesse caso, fica ainda mais claro de se apresentar que isso não é possível, uma vez que o presente estudo não utiliza perfis de usuários ou itens para realizar sugestão.

A terceira classificação descrita por MANOUSELIS *et al.* (2011) é a técnica híbrida que utiliza uma mistura dos 2 métodos previamente citados, por fim, é fácil perceber que também não seria uma classificação válida para o presente estudo.

Por fim, destaca-se a existência da ferramenta chamada: PaperCube. De todas as citadas, essa é a que mais se aproxima do proposto por esse estudo tanto no contexto teórico quanto prático. Essa ferramenta é voltada para realização de um mapeamento, que apesar de não ser mencionado, pode ser sistemático ou não dependendo do intuito do usuário.

Em termos gerais, a ferramenta citada disponibiliza uma forma de mapeamento dos relacionamentos bibliográficos, através de múltiplos tipos de visões, entre elas a forma em grafo. Esse mapeamento é feito através de buscas iniciais por um artigo através de *strings* de busca. Após selecionar somente um artigo, é possível visualizar as relações diretas de citação e referências, mas não ambas ao mesmo tempo. Também é possível expandir esses relacionamentos para as referências seguintes, provendo assim uma expansão até no máximo 15 ligações entre o artigo inicial e um novo exibido.

Apesar de prover um meio interessante de se visualizar citações e referências, essa ferramenta, segundo seu estudo apresentado (BERGSTRÖM e ATKINSON, 2009), utiliza uma base estática. Isso significa dizer que a taxa de falha ao tentar encontrar indexação de um artigo cresce com o tempo. Além disso, essa ferramenta não provê

uma forma de expansão em múltiplos sentidos, dificultando assim uma análise mais ampla. Um outro aspecto de suas possíveis expansões é só poder expandir um único vértice de seu grafo ou todos de uma vez. Não é possível seguir múltiplos caminhos sem englobar a expansão completa dos relacionamentos.

Expansões bibliográficas, como as citadas, podem crescer de forma exponencial, tornando-se humanamente difícil de analisar as informações envolvidas. Apesar da ferramenta prover uma forma visual de se entender o cenário ao redor de uma pesquisa, dependendo da quantidade de referências, pode-se não ser uma forma ideal de enxergar um universo grande a ser analisado.

Portanto, percebe-se uma brecha até então ainda não explorada, a possibilidade de prover uma forma de expansão controlada e melhorada. Uma forma em que o pesquisador possa podar a expansão, possa guiá-la e até mesmo utilizar uma heurística que ajude-o a seguir por caminhos que otimizem suas buscas. Com isso o pesquisador poderia diminuir o conjunto de artigos a serem analisados otimizando assim seu esforço e tempo.

Além do citado, pode-se minimizar as falhas de indexação utilizando uma abordagem com uma base de dados atualizada. Aumentando-se assim a possibilidade de descobrir referências esquecidas, provendo uma forma de auxílio para pesquisas bibliográficas em andamento, recém finalizadas ou que estejam apenas iniciando.

Capítulo 3

Apoiando a pesquisa bibliográfica

O intuito do deste trabalho é prover uma forma de apoiar pesquisas bibliográficas. Para isso, destina-se realizar em uma rede gerada por um mapeamento sistemático ou KDViz, a detecção e indicação de referências que não foram incluídas nesses mapeamentos, porém possuem altas chances de terem sido esquecidas pelo pesquisador que realizou o estudo mapeado.

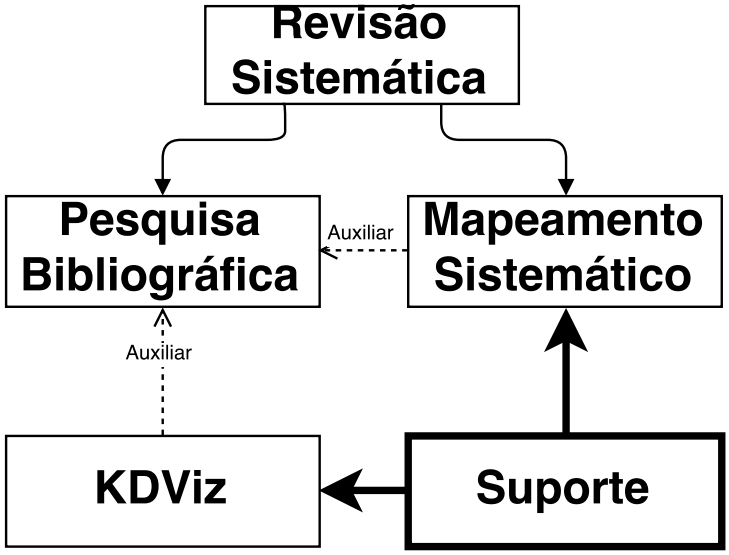

Figura 3.1: Ponto de suporte à revisão bibliográfica

Assim como mencionado na seção proposta (seção 1.3) há três aspectos que o presente trabalho considera relevantes para prover apoio a pesquisas bibliográficas: a relevância de cada referência conhecida para um propósito de pesquisa, a área à qual as referências envolvidas pertencem e a descoberta de referências relacionadas às iniciais (seja citando-as ou sendo referenciadas).

Uma vez que o conjunto universo de publicações possui uma quantidade gigantesca de elementos, haja vista que somente os dados de publicações indexadas por bases como *Scopus* (GOODMAN, 2005) seja na ordem de milhões, é necessário

que métodos que venham a expandir relacionamentos bibliográficos possuam uma heurística para podar de dados menos relevantes a fim de que o problema se torne humanamente e computacionalmente viável. Por isso a observância dos três aspectos se torna fundamental para expansão do conjunto de referências iniciais.

3.1 A relevância de cada referência envolvida

A relevância de cada referência pertencente a uma bibliografia, deve ser entendida como uma medida para mensurar o quanto importante a referência é para descrição do propósito da pesquisa ao qual a bibliografia pertence. Essa relevância pode ser inferida através de especialistas no assunto ao qual a bibliografia pertença.

Porém, um pesquisador ao iniciar um estudo, por não possuir domínio em determinados assuntos, pode sentir dificuldades em saber como conduzir suas buscas através dos documentos primariamente encontrados. Então ao se começar uma pesquisa em um assunto que seja previamente desconhecido é natural que haja dificuldade em identificar referências relevantes, principalmente senão houver auxílio de um especialista.

Somado a isso, ainda que um pesquisador seja um especialista em certas áreas, o mesmo pode não ser capaz de acompanhar toda a dinâmica de surgimento de novos estudos, acabando por não saber precisar a relevância de um estudo novo devido ao desconhecimento de seus impactos em outras publicações.

Devido a esses fatos, percebe-se a necessidade de haver um apoio à capacidade de classificação de estudos quanto a sua relevância ao redor de seus temas. A fim de otimizar o processo, o presente estudo busca uma forma automática de realizar esse auxílio, tentando prover uma alternativa à ausência de um especialista durante as buscas ou se houver um especialista, fornecer um método de apoio ao seu trabalho.

Nesse contexto, surge a possibilidade de utilização do algoritmo chamado HITS (*Hyperlink-Induced Topic Search* também conhecido como *Hubs and Authorities*) (KLEINBERG, 1999).

Os trabalhos encontrados mais próximos do que o presente trabalho busca apresentar, CHEN *et al.* (2009) e EGGERS *et al.* (2005) , utilizam HITS com o intuito de exibir dados de forma gráfica a fim de permitir que o pesquisador possa avaliar de forma visual tendências nos campos de estudo envolvidos através da entrada de dados temporais.

Porém, ambos consideram seus dados iniciais como o conjunto universo (indicado na figura 3.2 pelos nós existentes no círculo de referências iniciais) dos dados envolvidos. Isto é, nenhum outro dado que não esteja na base inicial será considerado. Portanto, não busca-se, através dos dados iniciais, novos relacionamentos que possam complementar a rede formada. Para esses estudos, os dados iniciais são

estáticos.

O fato citado abre margem para o presente estudo experimentar algoritmos que levem em consideração a expansão da base inicial de dados formada por referências existentes nos documentos que a constituem, como mostrado na figura 3.2.

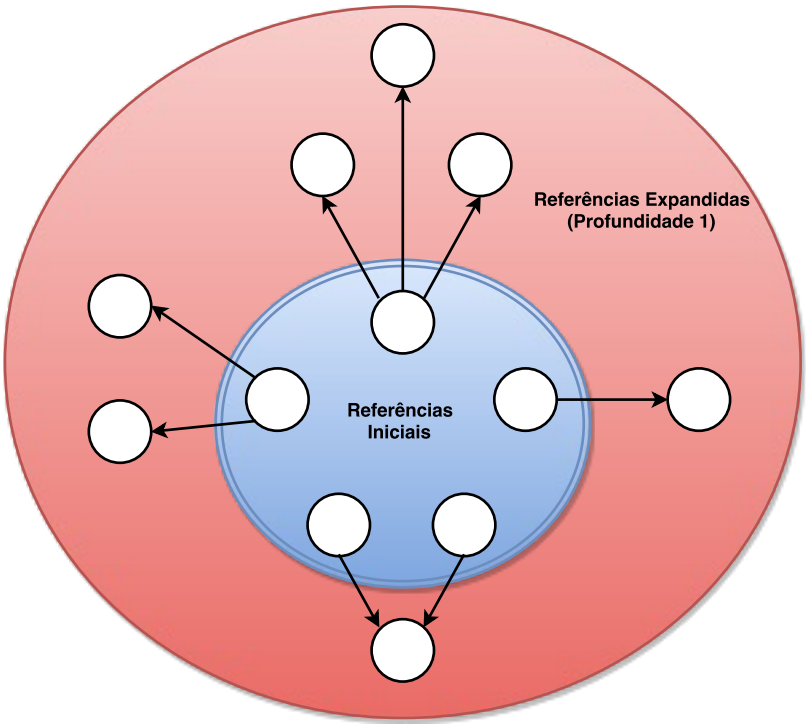

Figura 3.2: Expansão de uma rede inicial de referências para profundidade 1

Os conceitos de *hub* e de *authority*, para o algoritmo HITS acima citado, tratam-se de valores que indicam o quanto um nó em um grafo referencia outros pontos (*hubs*), assim como o quanto um nó é referenciado (*authority*) por outros nós. Basicamente, são medidas que demonstram a relevância dos nós em um grafo. O exemplo de um nó com tendência de valor alto para *hub* e outro para *authority* é apresentado abaixo:

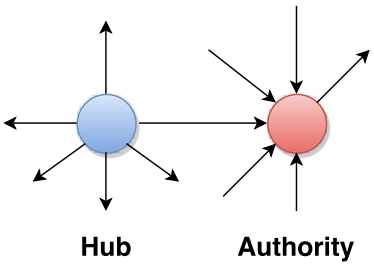

Figura 3.3: Exemplo de um possível *hub* e um possível *auhority*

Para justificar o *insight* para o uso da estratégia mencionada, partiu-se pri-

meiramente do mencionado acerca dos *pivots* em CHEN (2004): "*This could be a particularly useful feature for the detection of significant articles that could be easily overlooked by falling below a single high-citation threshold*", como também por sua importância e semelhança com HITS ao abordar os relacionamentos envolvidos.

Porém, apenas essa semelhança não basta como justificativa. Então a seguinte afirmação traz respaldo teórico para o uso de algoritmos, voltados para a finalidade citada, que sejam baseados nos relacionamentos existentes entre documentos (LIU e CHEN, 2013): "*the results confirm that topics from abstracts of citing papers have broader terms than topics from citation contexts formed by citing sentences*".

O que significa dizer que as relações envolvidas entre os documentos são fortes candidatas a especificar melhor o conteúdo abordado que os próprios conteúdos dos resumos dos documentos referênciados. Além disso, o autor de LIU e CHEN (2013) ainda adiciona : "*Sentences that cite specific references can provide a useful way to find the related work*".

3.2 Agrupamento de referências por áreas semelhantes

Ao se realizar um estudo, seu propósito de pesquisa é responsável por definir os aspectos das referências a serem buscadas pelo pesquisador, conforme descrito no capítulo de introdução (capítulo 1). O aspecto de largura de uma busca é a característica pela qual pode-se observar o quanto amplo ou específico é um estudo. São exemplos dessa diversidade: estudos interdisciplinares, que transitam entre diversas áreas do conhecimento realizando assim buscas de vasta largura. E artigos técnicos, que geralmente remetem-se a conteúdos bem específicos de uma determinada área, portanto necessitando de buscas de pouca largura.

Portanto, ao realizar pesquisas manuais, intuitivamente, o pesquisador estará controlando o quanto largo ou não será seu corpus envolvido. Porém as áreas de cada estudo nem sempre estão bem definidas. E tratando-se de uma quantidade grande de documentos a ser analisada, essa tarefa pode-se tornar complexa, acabando por misturar assuntos indevidos ou enveredar por buscas erradas.

Para auxiliar em tal aspecto, o presente estudo busca uma forma, também automática, de fornecer meios que permitam o pesquisador delinear da melhor forma as áreas envolvidas em seus estudos. Com isso, o pesquisador pode ser capaz de ter um apoio para controlar melhor que tipo de corpus final obterá.

Portanto fica claro a necessidade de se conseguir realizar agrupamentos com os documentos de uma busca a fim de conduzir melhor uma pesquisa. Para isso, o presente estudo sugere agrupar os documentos referenciados, através do método de

Louvain (BLONDEL *et al.*, 2008) a fim de separá-los por assuntos semelhantes.

Por fim, espera-se que sejam obtidos nos agrupamentos nós que possuam alto grau de similaridade, portanto, a referência de um possuirá alta probabilidade de ser indicada a ser referência de outro nó do mesmo agrupamento.

3.3 Expansão dos vértices conhecidos iniciais

O último ponto chave, trata-se da expansão dos vértices inicialmente conhecidos. Um conjunto de vértices iniciais, na melhor das hipóteses reflete por completo o tema abordado pelo documento do qual se extraiu tais vértices. Ou seja, no máximo, e ideal, representará uma bibliografia completa.

Porém, trata-se de algo estático em um determinado momento. Considerando o momento mencionado como presente, pode-se pensar em duas formas de enriquecer informações do grafo que representa esse cenário: seu passado e seu futuro. Ambos os casos analisando o cenário como observador atemporal.

Seu passado, pode ser obtido através da expansão dos vértices inicias no sentido de buscar vértices que sejam referenciados pelos vértices iniciais. Em outras palavras, buscar as referências que foram incluídas nos documentos usados como referências iniciais. Isso significa dizer que os novos vértices a serem adicionados refletem o passado dos vértices iniciais. Se esse entendimento for extrapolado para n passos, pode-se dizer que trata-se de uma forma de busca cuja a cronologia retoma o passado.

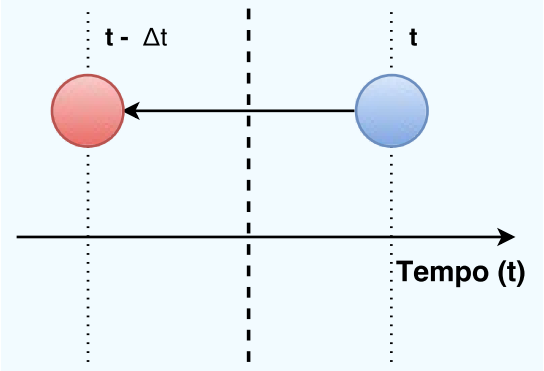

Figura 3.4: Expandindo o passado para obter relação de referência. Vértice conhecido em azul e expandido em vermelho

Seu futuro é representado pelo conjunto de vértices que referenciam os vértices iniciais. Em outras palavras, é representado pelos documentos que citam as referências iniciais. Ou seja, surgiram cronologicamente após o surgimento dos vértices iniciais. Se esse segundo entendimento for extrapolado também para n passos, pode-se percorrer e enriquecer o grafo com informações futuras ao momento inicial.

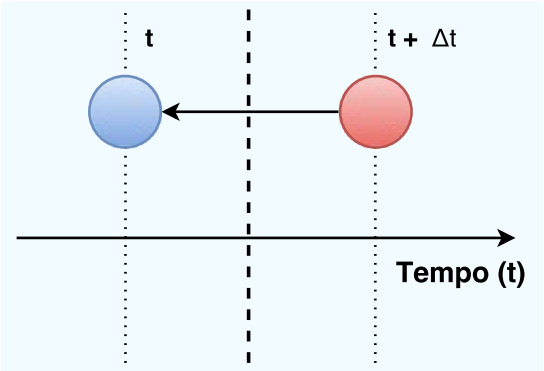

Figura 3.5: Expandindo o futuro para obter relação de citação. Vértice conhecido em azul e expandido em vermelho

Ambos os aspectos possuem relação com os dois principais aspectos do algoritmo mencionado para a análise da relevância de cada referência envolvida. *Hubs*, são vértices que realizam muitas referências, ou seja, no geral serão adicionados ao grafo mais facilmente através da busca para o futuro, enquanto os *Authorities*, são vértices que são muito citados, portanto, são criados de forma mais fácil através da busca no sentido passado.

Portanto, essa expansão enriquece o grafo com vértices que influenciarão tanto a sugestão final de novos estudos quanto a adição de vértices a cada etapa em que o algoritmo proposto por esse estudo for iterado. Havendo assim um ganho de informação com ambas as formas de expansão, cujo aproveitamento será de fato acrescido através da iteratividade, onde o produto final de uma iteração melhora o conjunto inicial da iteração seguinte.

3.4 Etapas de apoio

A figura 3.6 apresenta um diagrama de atividades que contempla os três aspectos enumerados nas seções 3.1, 3.2 e 3.3. O diagrama apresenta os passos propostos, pelo presente estudo, que, ao final de sua execução, apresentará um novo conjunto de elementos da bibliografia que são desconhecidos ou foram negligenciados pelo pesquisador.

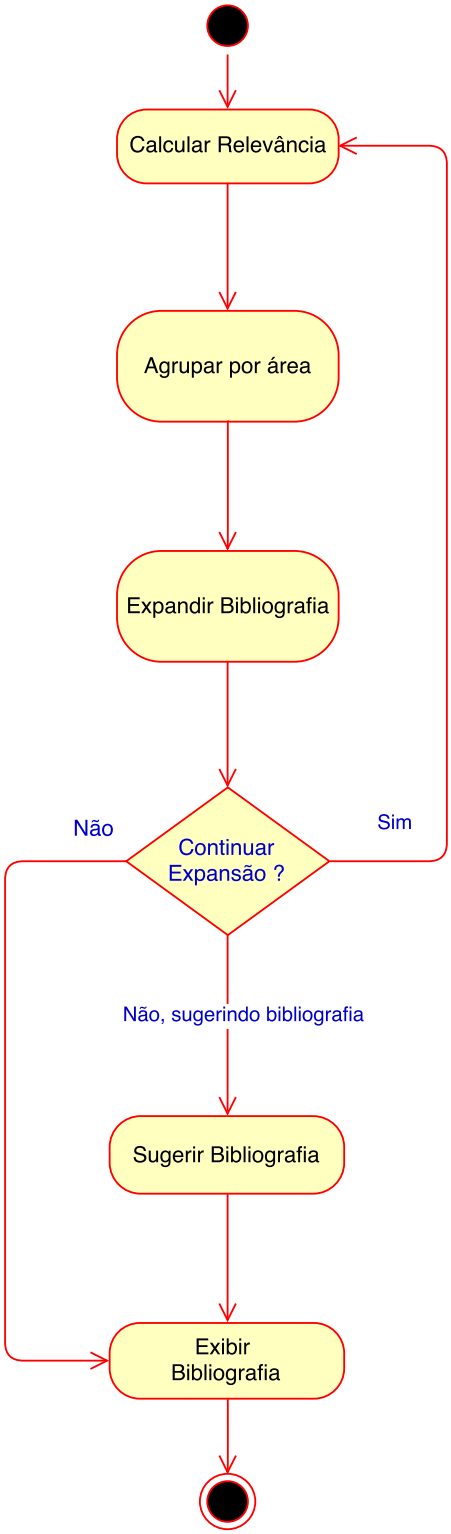

Figura 3.6: Diagrama de atividades das etapas de apoio

3.4.1 Cálculo de relevância

O passo de pontuação, será responsável por prover uma forma de quantificar a relevância mencionada na seção 3.1 para cada estudo a ser analisado. Para o de-

vido fim, pode-se utilizar não somente o algoritmo proposto, HITS, mas também outros algoritmos de cálculos de relevância para nós em um grafo ou outra forma de pontuação que leve em consideração os relacionamentos existentes entre os documentos envolvidos. Porém, o presente estudo, irá considerar o HITS como principal quantificador, conforme já justificado seu uso.

Essa é a primeira das três atividades a serem repetidas de forma iterativa conforme apresentado pelo diagrama de atividades da figura 3.6. A cada iteração, essa atividade irá recalcular as pontuações de todos os nós do grafo atual. Isso significa dizer que a cada expansão, todos os nós que já existiam terão suas pontuações recalculadas e possivelmente alteradas dependendo da nova estrutura formada após a adição de possíveis novos nós ao grafo pela iteração anterior.

A atividade de cálculo de relevância é de grande importância. Ela é responsável por gerar medidas que tornarão possível realizar tanto a atividade de expansão da bibliografia como a sugestão de novas bibliografias. O produto gerado por essa atividade irá se tornar parte das regras utilizadas para condução para as demais atividades citadas.

3.4.2 Agrupamento por área

Em seguida, para prover uma forma de executar o descrito pela seção 3.2, o passo de clusterização irá dividir as referências do grafo total em grupos através do método de *Louvain* (BLONDEL *et al.*, 2008). Esse método levará em consideração a estrutura do grafo existente, ou seja, os relacionamentos existentes no presente grafo. Cada comunidade encontrada, conforme já mencionado em proposta (seção 1.3), representará um conjunto de documentos pertencentes a um tema em comum.

Essa medida possibilita que a busca seja conduzida filtrando os temas envolvidos no propósito de pesquisa estabelecido. Com isso pode-se experimentar buscas com larguras mais equilibradas, nas quais cada área seja igualmente buscada ou pode-se escolher que apenas determinadas áreas sejam expandidas. Para esse passo, o presente estudo irá considerar como padrão a largura mais equilibrada possível a fim de não criar viés em suas indicações finais.

Além disso, utilizar largura mais equilibrada possível, também significa não pre-estabelecer o número de comunidades a serem formadas. Deixar que o algoritmo encontre uma convergência de acordo com seus critérios de maximização. Critérios esses que serão explicados na seção de algoritmo.

Essa é a segunda das três atividades a serem repetidas. A cada iteração, essa será responsável por encontrar todas as comunidades existentes no grafo atual. Da mesma forma que a atividade de pontuação, essa atividade poderá ser afetada pela inclusão de novos nós oriundos da iteração anterior. Ao serem adicionados novos

nós, possíveis novas comunidades podem ser encontradas e comunidades encontradas previamente podem ser repartidas.

3.4.3 Expansão da bibliografia

No final de cada iteração, se faz necessária a expansão da bibliografia para cobrir a necessidade apresentada na seção 3.3. Essa atividade ditará, através da quantidade de melhores referências pontuadas por comunidade, quantas serão as referências a serem consideradas para a expansão do conjunto bibliográfico existente. Isso proporciona uma outra forma de controlar o alcance da descoberta a ser realizada. Com isso, pode-se especificar o quão profundo deseja-se que a busca expanda cada assunto.

Além disso, há o sentido da busca. Trata-se de uma forma de se controlar qual será o sentido que alimentará o grafo com novas informações. Conforme explicado na seção 3.3, podem ser dois: sentido das referências, passado, e o sentido das citações, futuro. Vale enfatizar que as cronologias são relativas ao momento de criação de cada documento de onde se partiu a busca.

O processo de expansão será realizado de forma iterativa. Por isso é necessário que haja uma forma de controlar a quantidade de iterações a ser realizada. Para isso haverá uma medida que representará o alcance da profundidade na expansão do conjunto bibliográfico que o pesquisador possui inicialmente.

O alcance da profundidade descreve o número i de iterações na etapa de busca e servirá para regular o tamanho do crescimento da base de dados existente ao longo das iterações. Em outras palavras, seja um grafo que represente a bibliografia de um estudo, as citações e referências dessa bibliografia. O alcance da profundidade indica a distância máxima que deseja-se alcançar através das referências iniciais até referências que ainda não tenham sido expandidas (que não possuam mapeadas suas ligações com referências fora do conjunto inicial).

3.4.4 Sugestão de nova bibliografia

A atividade de sugestão, após n iterações mencionadas, irá prover, através do número de sugestões por área existente, quantas sugestões deseja-se obter. Serão selecionadas as n referências mais bem quantificadas pelo HITS ao longo de suas iterações. Essas sugestões irão representar o conjunto dos possíveis estudos que foram esquecidos ou negligenciados pelo estudo alvo envolvido e poderão ser sugeridos para uma leitura de seus resumos e em seguida, se assim for de fato relevante, sua leitura na íntegra.

3.5 Formalização do problema

Seja a bibliografia inicial de uma pesquisa em um momento t_0 representada por um grafo: $G^0(V^0, A^0)$, tal que:

- $V^0 = \{v_0, v_1, v_2, \ldots, v_m\}$ é o conjunto inicial de m documentos (vértices) da bibliografia.
- $A^0 = \{(v_x, v_y), (v_z, v_w), \ldots, (v_u, v_v)\}$ é o conjunto de referências (arestas) onde $\{v_x, v_y, v_z, v_w, \ldots, v_u, v_v\} \subseteq V^0$.

Cada aresta é representada por um par de vértices onde: o primeiro elemento, vértice de partida, representa um documento que faz uma referência e o segundo vértice simboliza o vértice incidente, o documento referenciado pelo primeiro documento. Logo, o grafo G^0 é um grafo direcionado.

Portanto, deseja-se encontrar um grafo $G^n = \{V^n, A^n\}$ tal que:

- $V^n = V^0 \cup V^p$ onde V^p é o conjunto de vértices adicionados ao grafo G^0
- $A^n = A^0 \cup A^p$ onde A^p é o conjunto de arestas adicionadas ao grafo G^0

É importante destacar que G^n precisa apresentar um subconjunto de vértices V^r, que são relevantes em relação a bibliografia inicial e ao propósito de pesquisa definido. Por exemplo, suponha que um pesquisador apresente um grafo de bibliografia inicial G^0 e esteja com um *propósito de pesquisa* de conhecer novas áreas afins com a que a sua pesquisa atual trabalha. Logo, $v_k \in V^r$ será relevante se for um vértice que apresente alguma correlação entre a área de pesquisa atual e uma área diferente das que apareciam em G^0. No presente exemplo, uma bibliografia G^0 com textos de bioinformatica tem como um subcojunto de vértices relevantes artigos de *big data*.

Onde V^r apresenta as seguintes propriedades:

1. $V^r \subset V^p$
2. $V^r \cap \ V^0 = \{\}$

A equação 3.1 formaliza o problema de maximizar o conjunto de referências relevantes V^r para um propósito de pesquisa definido e a sua bibliografia inicial G^0.

$$V^r = \text{argmax}(|V^r|) \tag{3.1}$$

Para o propósito do presente trabalho utilizaremos o operador profundidade máxima em $G^n = P[G^n]$ tal que $P[G^n]$ é dado pela maior excentricidade existente a partir do conjunto V^0. Entende-se como excentricidade o maior menor caminho

existente entre os vértices iniciais do grafo e todos os demais vértices. Conceito esse que pode ser entendido a partir do estudo DE FREITAS (2010). Assim como entende-se por menor caminho as arestas que são necessárias para ligar um vértice a outro de forma a utilizar a menor quantidade delas.

3.6 O algoritmo

Para execução das etapas citadas na seção etapas de apoio (seção 3.4), a presente seção irá demonstrar os algoritmos correspondentes a cada uma delas reunindo todos os sub passos nele contidos. O algoritmo completo é composto por duas etapas: expansão com heurística e sugestão.

A primeira etapa, conforme apresentada pelo algoritmo 1, realiza de forma iterativa e independente da segunda, três operações previamente descritas: pontuação, clusterização e expansão do conhecimento. Para realizar tal tarefa, o algoritmo recebe como entrada o grafo a ser iterado mais 5 parâmetros: profundidade, numero de melhores vértices pontuados por *cluster*, sentido da expansão, listagem de vértices a não expandir e verdadeiro ou falso para somente expandir pelos últimos vértices adicionados na iteração anterior.

Algoritmo 1 Expansão com heurística

procedimento EXPANSAOHEURISTICA($grafo$, $profundidade$, $numMelhoresVertices$, $sentidoBusca$, $verticesDesativados$, $somenteUltimosVertices$)
  $i \leftarrow 0$
  $prof \leftarrow profundidade$
  enquanto $i \leq prof$ **faça**
    CALCULARHITS($grafo$)
    EXECUTARLOUVAIN($grafo$)
    EXPANDIRGRAFO($grafo$, C, $numMelhoresVertices$, $sentidoBusca$, $somenteUltimosVertices$, $ultimosVertices$, $verticesDesativados$) ▷ C, comunidades encontradas
  fim enquanto
fim procedimento

A profundidade descreve a distância máxima que será gerada no grafo a partir dos nós iniciais até os vértices mais distantes usando como referência a quantidade de arestas entre eles. Em outras palavras, isso representará a excentricidade máxima que poderá ser percebida no grafo após todas as iterações. Pode-se também vê-la como o número máximo de iterações que serão realizadas na primeira etapa do algoritmo.

O número de vértices a serem expandidos, sendo representado pela variável: $numMelhoresVertices$ é responsável por controlar a quantidade máxima de vértices

expandido por *cluster* a cada iteração. Essa quantidade representará não necessariamente o número exato utilizado a cada iteração, pois deve-se observar que pode haver quantidade menor de vértices em certos *clusters* encontrados. Os vértices são ordenados por suas pontuações obtidas na operação inicial e em seguida são escolhidos os n vértices melhores pontuados para serem expandidos.

O parâmetro sentido da busca utiliza a variável: $sentidoBusca$. Tal parâmetro indica qual das três formas de expansão será utilizada. São elas: expansão das referências, expansão das citações ou ambas. A expansão das referências poderá agregar a cada vértice expandido novas arestas originarias em si e direcionadas a vértices novos ou preexistentes. E a expansão das citações poderá agregar a cada vértice expandido novas arestas direcionadas para o vértice expandido e que originam-se em nós novos ou preexistentes.

Há também a possibilidade da heurística começar recebendo uma listagem de vértices a serem desconsiderados na expansão. Isso possibilita a poda do grafo. Essa é informada pela listagem: $verticesDesativados$. Os vértices existentes nessa listagem serão removidos da listagem de vértices a ser utilizada na chamada para expandir grafo.

Outra possibilidade de influência na expansão do grafo trata-se da listagem de somente últimos vértices representada pelo parâmetro: $somenteUltimosVertices$. Esse repassa à heurística a decisão para considerar somente os vértices adicionados na última iteração ou todos os vértices no grafo. Ao começar a iteração, somente nessa etapa, independente da decisão passada por esse parâmetro, todos os vértices são considerados parte da última listagem de vértices adicionados.

A segunda etapa, conforme apresentada pelo algoritmo 2, realiza a sugestão de vértices que, de acordo com o propósito de pesquisa, podem ser relevantes. Para isso, são utilizados os resultados da última operação de pontuação realizada sobre o grafo total. Essa pontuação é ordenada de forma decrescente e exclui-se os vértices que pertenciam ao conjunto inicial. Escolhe-se os n primeiros vértices para serem sugeridos como candidatos a integrarem o conjunto inicial.

Algoritmo 2 Sugestão

```
1: procedimento SUGESTAO(grafo, s)
2:     lista ← heuristicSuggestion(grafo, s)        ▷ s, quantidade de sugestões
3:     print(lista)
4: fim procedimento
```

3.6.1 Cálculo de relevância usando Hyperlink-Induced Topic Search

O algoritmo HITS trata-se de um algoritmo criado inicialmente para avaliação da importância de páginas web através de seus relacionamentos. Tal algoritmo se baseia em dois principais conceitos: *hubs* e *authorities*. Os *hubs* são pontos de concentração de boas referências. Isso quer dizer que analisando as referências dos *hubs* pode-se chegar à prováveis documentos relevantes. Enquanto os *authorities* são pontos que possuem conteúdos relevantes (RAJENDRA e PAWAN, 2008).

Esse algoritmo, foi criado utilizando uma abordagem de grafo. Onde cada documento ou página é simbolizado por vértices e cada relacionamento entre esses vértices é chamado de aresta. O HITS computa dois valores para cada vértice, um valor de *hub* e outro de *authority*. Esse algoritmo é influenciado pela retroalimentação entre seus vértices. Isso quer dizer que a cada iteração as pontuações de *hubs* e *authorities* influenciam umas nas outras. Para realizar os cálculo dessas pontuações é utilizada uma variação da forma de calcular autovetores utilizada pelo algoritmo *PageRank*.

A implementação do HITS (HAGBERG *et al.*, 2013) utilizada para o presente trabalho inicializa as pontuações da seguinte forma: $\forall v$, authority(v)= $\frac{1}{numVertices}$ e hub(v) = $\frac{1}{numVertices}$, onde $numVertices$ é a quantidade de vértices existentes no grafo o qual deseja-se extrair a pontuação dos HITS.

Em seguida, de forma iterativa, o algoritmo repete as duas funções de atualização das pontuações dos vértices seguida de normalizações dessas pontuações. Sendo respectivamente: atualização dos *authorities*, normalização dos *authorities*, atualização dos *hubs* e suas normalizações. Ambas as normalizações são necessárias a fim de que as pontuações não aumentem indefinidamente e o algoritmo possa convergir, conforme demonstrado em KLEINBERG (1999).

Essa iteração continua até que o algoritmo consiga convergir para um resultado cujo o erro seja menor que o erro padrão tolerável ($1.0 * e^{-8}$) ou até que o algoritmo ultrapasse o número máximo de iterações especificada (por padrão, 100) terminando assim por não convergir.

O erro é calculado pela seguinte equação:

$$err = \sum_{i=1}^{n} abs(h[v_i] - hLast[v_i]) \tag{3.2}$$

onde abs() é uma função que retorna o valor absoluto de um número, h[v_i] retorna o valor de *hub* do vértice i e hLast[v_i] retorna o valor de *hub* para o vértice i na iteração anterior.

A atualização dos *authorities* é calculada pela seguinte equação:

$$authority(v) = \sum_{i=1}^{n} hub(v_i) \tag{3.3}$$

onde v é o vértice a ser atualizado, n é a quantidade total de vértices existentes no grafo que estão direcionados para v e i é cada um dos vértices que apontam para v. Isso significa dizer que a pontuação de *authority* de v será calculada através da soma das pontuações *hubs* dos vértices que apontam para v.

A normalização de cada pontuação dos *authorities* é feita através da seguinte equação:

$$authority(v) = authority(v) * s \tag{3.4}$$

na qual a variável s de normalização será recalculada a cada iteração através da equação a seguir, onde a é um conjunto de chaves e valores no qual as chaves são cada vértice de um grafo e seus valores são suas respectivas pontuações de *authorities*. A função values() retorna todos os valores desse conjunto em forma de lista e a função max() retorna o valor máximo de uma lista.

$$s = \frac{1.0}{max(a.values())} \tag{3.5}$$

De forma similar, o calculo das pontuações de *hubs* é realizado de acordo com a seguinte equação:

$$hub(v) = \sum_{i=1}^{n} authority(v_i) \tag{3.6}$$

onde v é o vértice a ser atualizado, n é a quantidade total de vértices para aos quais v está direcionado e i é cada um dos vértices apontados por v. Ou seja, a pontuação de *hubs* de v será calculada através da soma das pontuações *authorities* dos vértices apontados por v.

De forma análoga a fórmula 3.4 a normalização das pontuações de *hubs* é feita através da seguinte equação:

$$hub(v) = hub(v) * s \tag{3.7}$$

onde s é a variável de normalização dos valores de *hubs* e será recalculada a cada iteração através da equação abaixo, onde h é um conjunto de chaves e valores no qual as chaves são cada vértice de um grafo e seus valores são suas respectivas pontuações de *hubs*. As funções values() e max() assumem o mesmo comportamento explicado para 3.5.

$$s = \frac{1.0}{max(h.values())} \tag{3.8}$$

Após convergir, como última etapa, seus resultados serão normalizados (por padrão) de acordo com a seguinte equação para os valores de *authorities*:

$$authority(v) = authority(v) * s \tag{3.9}$$

onde s:

$$s = \frac{1}{\sum(a.values())} \tag{3.10}$$

e para *hubs*:

$$hub(v) = hub(v) * s \tag{3.11}$$

onde s:

$$s = \frac{1}{\sum(h.values())} \tag{3.12}$$

O algoritmo 3 ilustra o funcionamento da implementação de HITS utilizada pelo presente trabalho conforme demonstrada por KLEINBERG (1999).

3.6.2 Agrupamento por área usando o algoritmo Louvain

O algoritmo utilizado para clusterização do presente trabalho trata-se do método de Louvain descrito por BLONDEL *et al.* (2008) em *Fast unfolding of communities in large networks*. Esse foi desenvolvido para ser capaz de extrair comunidades de grandes grafos. Além disso, uma de suas características é possuir um bom desempenho em termos de tempo computacional gasto para sua execução. Para realizar tal tarefa, esse utiliza uma eurística cujo objetivo é maximiza a modularidade do grafo de entrada.

O termo modularidade no contexto de grafo, definido em NEWMAN e GIRVAN (2004),trata-se de uma métrica que busca mensurar a qualidade das comunidades encontradas em um grafo. Para isso o cálculo de modularidade utiliza dois conceitos principais: o peso total das arestas que apontam para um vérice i e o peso total das arestas no grafo completo.

A modularidade também pode ser entendida, ainda segundo NEWMAN e GIRVAN (2004), como uma medida que compara a quantidade de arestas em comunidades de um grafo com a fração de arestas para caso esse mesmo grafo houvesse tido suas arestas geradas de forma aleatória.

A fórmula abaixo demonstra como a modularidade de um grafo pode ser calculada:

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i * k_j}{2m} \right] \delta(c_i, c_j),$$

Algoritmo 3 *Hubs and Auhtorities*

```
1: procedimento CALCULARHITS(G)                          ▷ G recebe o Grafo
2:     para v in G faça
3:         v.hub = 1/numVertices
4:         ▷ v.hub é a pontuação hub de v e v.auth é pontuação de authority de v
5:     fim para
6:     i = 0                                              ▷ i é o número de iterações
7:     enquanto true faça
8:         hlast = G.hubs          ▷ Atribui à hlast as pontuações de hubs do grafo G
9:         h = (G.keys(), 0)                          ▷ Pontuações hubs dos vértices de G
10:        a = (G.keys(), 0)                    ▷ Pontuações authorities dos vértices de G
11:        para v in G faça                ▷ Atualizar todos os valores de authorities
12:            para q in v.incomingNeighbors faça
13:                v.auth += q.hub
14:            fim para
15:        fim para
16:        para v in G faça                          ▷ Atualizar todos os valores de hubs
17:            para q in v.outgoingNeighbors faça
18:                v.hub += q.auth
19:            fim para
20:        fim para
21:        s = 1.0/max(a.values())          ▷ Normalização dos valores de authorities
22:        para v in G faça                                ▷ para cada vértice do grafo G
23:            v.auth = v.auth * s
24:        fim para
25:        s = 1.0/max(h.values())                ▷ Normalização dos valores de hubs
26:        para v in G faça                                ▷ para cada vértice do grafo G
27:            v.hub = v.hub * s
28:        fim para
29:        err = 0
30:        para n in G.hubs faça   ▷ Calcula o erro através da pontuação de hubs
31:            err+ = abs(G.hubs[n] − hlast[n])
32:        fim para
33:        se err < 1.0 * e^−8 então                  ▷ Se o erro for menor que o padrão
34:            break
35:        fim se
36:        se i > 100 então                          ▷ Se ultrapassar máximo de passos
37:            O algoritmo HITS não convergiu com i + 1 iteracoes
38:        fim se
39:        i+ = 1
40:    fim enquanto
41:    s = 1.0/sum(a.values())
42:    para n in a faça
43:        a[n]∗ = s
44:    fim para
45:    s = 1.0/sum(h.values())
46:    para n in h faça
47:        h[n]∗ = s
48:    fim para
49:    return a,h
50: fim procedimento
```

onde $\delta(c_v, c_w) = 1$, se i e j forem designados para mesma comunidade. Caso contrário, o delta de Kronecker valerá 0. A_{ij} traduz a matriz de adjacência dos pesos das arestas entre i e j, $k_i = \sum_j [A_{i,j}]$ quantifica a soma total dos pesos incidentes em i, $\frac{k_j}{2m}$ representa a fração dos pesos das arestas de i para j caso houvessem sido geradas aleatoriamente e $m = \frac{1}{2}\sum_{i,j}[A_{ij}]$ é a soma total dos pesos das arestas. Portanto, percebe-se que $A_{ij} - \frac{k_i k_j}{2m}$ descreve a diferença entre entre os pesos das arestas no grafo real e caso o mesmo grafo tivesse tido suas arestas aleatoriamente geradas.

O método de *Louvain* busca maximizar Q através de sucessivas iterações re-agrupando vértices e recalculando o valor local de Q. Trata-se de um algoritmo guloso que possui duas fases iterativas: reorganização local dos vértices e alteração da função de ganho.

A primeira fase, reorganização local, irá começar com cada vértice pertencendo a uma comunidade. O algoritmo e irá iterar sobre todos os vértices mudando-os de comunidade, transferindo-os para comunidades de vértices vizinhos pertencentes a outras comunidades e irá recalcular o valor de Q para cada vizinho desse vértice. Se Q variar positivamente para algum de seus vizinhos, o vértice será movido por definitivo para comunidade testada, caso haja alteração positiva dos valores de Q, o vértice será mantido onde estava. Essa iteração ser manterá até que não haja mais mudança de comunidade para cada um dos vértices iterados.

A segunda fase irá considerar cada comunidade encontrada como um ponto único e irá representá-lo como um novo vértice. Todas as ligações dos vértices que foram agrupados com vértices que foram agrupados em outra comunidade, serão representadas entre os novos vértices que surgiram na segunda fase. Com isso a iteração volta para primeira fase e continua sua execução até que não seja possível agrupar novas comunidades. Quando assim acontecer, o grafo final representará as comunidades encontradas para os vértices do grafo inicial.

O algoritmo 4 descrito por AYNAUD e GUILLAUME (2010), demonstra o comportamento do método de *Louvain*.

Algoritmo 4 Método de Louvain

```
1: procedimento executarLouvain(G)
2:     enquanto true faça
3:         Coloque cada vértice de G em sua própria comunidade
4:         enquanto Vértices mudarem de comunidade faça
5:             para v in G faça
6:                 Coloque v em cada sua comunidade vizinha e verifique
   se houve ganho de modularidade para o vértice, se sim, mantenha-o
   lá, senão mantenha onde está
7:             fim para
8:         fim enquanto
9:         se  nova modularidade do grafo total for maior que a inicial
   então
10:            G = nova rede
11:        senão
12:            return
13:        fim se
14:    fim enquanto
15: fim procedimento
```

3.6.3 Expansão da bibliografia

A atividade de expansão bibliográfica utiliza um algoritmo próprio. Seu objetivo é agregar novos vértices ao grafo inicial que estejam relacionados de alguma forma, porém se desconhecia. Conforme já citado, o conjunto universo de publicações existentes é gigantesco. Sendo esse da ordem de milhões. Portanto trabalhar com todas as referências existentes se torna inviável do ponto de vista humano.

Por isso, torna-se necessário que o conhecimento a ser expandido surja de forma mais refinada. Necessita-se que a quantidade de publicações a ser trabalhada seja computacionalmente possível de ser utilizada, assim como o produto gerado por essa expansão seja humanamente útil para quem estiver utilizando-o.

Devido a esse contexto, o presente algoritmo busca uma forma de podar as informações externas ao conjunto inicial, buscando apresentar as informações mais relevantes. Para isso o algoritmo necessita dos produtos gerados pelas computações das atividades anteriormente citadas e de fazer acesso a uma base externa de dados a fim de poder agregar novas informações ao grafo conhecido.

Há dois principais pontos a serem considerados para a expansão do grafo modelado por esse estudo. O primeiro é o sentido e o segundo a profundidade. Ambos controlam a forma que o grafo é expandido pelo método proposto.

A expansão de um grafo bibliográfico pode ser feita em dois sentidos nas direções das arestas. São eles: sentido das referências e sentido das citações. Cada sentido desses representa um tipo de informação cronológica conforme já mencionado na seção 3.3. Enquanto as referências de um vértice representam o passado, as citações desse vértice representam o futuro. Ambos aspectos tendo como marco temporal o tempo de criação do vértice referenciado.

Um outro ponto importante a ser mencionado é a característica de profundidade. Ou a distância adicionada aos vértices iniciais até os vértices que estejam separados dos vértices iniciais pela maior quantidade de arestas possível, sendo esse conceito também conhecido por excentricidade conforme explicado na seção 3.5.

Essa característica foi uma medida adotada por esse estudo para controlar as iterações do método proposto. Vale enfatizar que apesar da profundidade do grafo poder aumentar a cada iteração, essa só aumenta em uma unidade nessa atividade. Pois na expansão somente são adicionados ao grafo os vértices imediatamente adjacentes aos vértices a serem expandidos.

A computação desse algoritmo começa recebendo o grafo inicial a ser expandido, uma lista de comunidades encontradas, o número n de melhores vértices, o sentido das expansões, um valor booleano para decidir se a expansão somente utilizará os últimos vértices adicionados, uma lista desses últimos vértices e uma lista de vértices desativados (podados explicitamente antes de iniciar a heurística). Esse grafo pode ser um grafo conexo (com ligação entre os vértices) ou até mesmo totalmente desconexo (sem arestas ligando quaisquer vértices).

Nessa atividade a ideia, conforme já mencionado, é utilizar os produtos das atividades anteriores para refinar essa expansão. Os dois produtos gerados anteriormente são: as pontuações de relevância do grafo e as comunidades a que cada vértice pertence.

O algoritmo nessa atividade, a fim de expandir somente os vértices mais relevantes, possui como critério de parada o fim da iteração sobre uma lista podada de vértices a serem expandidos. Esse irá a cada passagem pela atividade de expansão, escolher os n vértices melhores pontuados em cada comunidade encontrada para serem expandidos.

Após essa escolha, o algoritmo irá verificar se a expansão deve se restringir somente aos últimos vértices adicionados ao grafo. Caso o resultado dessa verificação seja positivo, a lista dos n vértices melhores pontuados será podada a fim de preservar somente os vértices que estiverem presentes em ambas. Caso contrário, quaisquer vértices das comunidades encontradas poderão estar na lista a ser expandida, desde que estejam entre os melhores pontuados.

Essas expansões serão realizadas de acordo com o sentido da expansão(referências e/ou citações). Para cada vértice da lista anteriormente podada, o algoritmo irá

recuperar vértices que possuam relação de: citação, recuperando todos os vértices que citem o vértice a expandir e/ou relação de referência, recuperando vértices que foram referenciados pelo vértice a ser expandido ou ambos.

O algoritmo 5 visa demonstrar melhor o funcionamento descrito:

Algoritmo 5 Expansão da Bibliografia

```
1: procedimento EXPANDIRGRAFO(G, C, numMelhoresVertices, sentidoExpansao, somenteUltimosVertices, ultimosVertices, verticesDesativados)
2:     tempUltimosVertices ← ∅
3:     para c in C faça                          ▷ Para cada comunidade C encontrada
4:         se somenteUltimosVertices então c = c ∩ ultimosVertices
5:         fim se
6:         c = c \ verticesDesativados           ▷ Remove da listagem, os vértices desativados
7:         c ← selecionaMelhoresVertices(c, numMelhoresVertices)   ▷ Ordena de forma decrescente de pontuação de authority os vértices de c e seleciona-se os primeiros numMelhoresVertices
8:         para v in c faça                      ▷ Para cada vértice da comunidade c
9:             se sentidoExpansao = sentidoReferencias OR sentidoExpansao = todos então
10:                novosVertices ∪ recuperarVertices(v, sentidoReferencias)
11:                adicionaVerticesAoGrafo(v, novosVertices)
12:            fim se
13:            se sentidoExpansao = sentidoCitacoes OR sentidoExpansao = todos então
14:                novosVertices ∪ recuperarVertices(v, sentidoCitacoes)
15:                adicionaVerticesAoGrafo(v, novosVertices)
16:            fim se
17:            tempUltimosVertices = tempUltimosVertices ∪ novosVertices
18:        fim para
19:    fim para
20:    ultimosVertices = tempUltimosVertices
21:    return G, ultimosVertices
22: fim procedimento
```

3.6.4 Sugestão de nova bibliografia

Para realizar a atividade de sugestão, o presente trabalho implementou um algoritmo *naive* deixando em aberto, para um trabalho futuro, a possibilidade de incrementar tal mecanismo. Nessa atividade, o foco é filtrar os dados gerados pelo fluxo de atividades anteriores. A característica escolhida para realização desse filtro foi a pontuação final de cada vértice.

Após a realização das três atividades anteriormente citadas, obtém-se um grafo possivelmente expandido, cujos vértices encontram-se pontuados de acordo com o

algoritmo escolhido para ser utilizado na primeira atividade (seção 3.6.1). Porém, vale salientar que na atividade atual são recalculadas as pontuações de todos os vértices no grafo final. Isso permite escolher um algoritmo de pontuação diferente do escolhido para cálculo da relevância da primeira atividade. Contudo, o estudo manteve como padrão, por motivos já justificados, o HITS.

Portanto, os parâmetros iniciais para execução do algoritmo dessa atividade são: o grafo final G, gerado pelas atividades anteriores e o número de sugestões desejadas. Inicialmente o algoritmo irá executar um novo cálculo das pontuações para todos os vértices de G. Em seguida, G terá suas pontuações ordenadas de forma decrescente e serão escolhidos o n vértices de melhor pontuação que não estejam no conjunto dos vértices iniciais. Como produto final, será gerada uma lista contendo os n vértices sugeridos.

Como dito anteriormente, trata-se de uma abordagem *naive*, pois além desse aspecto poderia-se utilizar outras informações a fim de tentar aprimorar os resultados obtidos. Mas devido ao amplo escopo desse trabalho, esses aprimoramentos serão listados entre os itens para trabalhos futuros.

O algoritmo 6 demonstra o funcionamento utilizado para fazer o filtro nessa atividade:

Algoritmo 6 Sugestão de Bibliografias

```
1: procedimento SUGERIRBIBLIOGRAFIAS(G, n)
2:     CALCULATEHITS(G)
3:     vertices = G.getNodes()          ▷ getNodes() retorna um dicionário de chaves-valores onde as chaves são os vértices e os valores são suas pontuações
4:     vertices = sorted(vertices, key, reverse = True)  ▷ sort ordena o dicionário (vertices) pelas chaves (key) em ordem decrescente(reverse=true)
5:     sugestoes = top(vertices, n) ▷ top retorna os n primeiros elementos de uma lista
6:     return sugestoes
7: fim procedimento
```

Capítulo 4

Desenvolvimento

O presente capítulo disserta acerca da cronologia de desenvolvimento do ferramental desse trabalho. Esse aborda: a escolha dos parâmetros utilizados, a arquitetura modelada e sua implementação.

4.1 O contexto

Inicialmente o trabalho surgiu com o objetivo de prover uma forma de auxiliar pesquisadores a realizarem pesquisas bibliográficas. Sua proposta inicial foi prover um suporte ao mapeamento sistemático ou ao *Knowledge Domain Visualization* (KDViz), porém que fosse capaz de expandir um conjunto bibliográfico inicial de forma automática. O resultado dessa expansão seria formado por referências que estivessem relacionadas ao conjunto inicial através de citações ou referências, resultado similar ao obtido de forma manual ao se realizar um procedimento de *snowballing* descrito por WOHLIN (2014).

Porém, durante sua implementação, o estudo foi sendo refinado. O primeiro refinamento surgiu com a ideia de realizar uma expansão um pouco melhorada. Inicialmente toda e qualquer expansão havia sido pensada como realizando a busca e recuperação de todas as informações associadas ao conjunto inicial de bibliografias a ser analisado. Porém, o presente estudo pretendia ir além de só realizar uma busca e recuperação da informação.

Então se almejou que a a expansão bibliográfica realizada por esse estudo fosse realizada através de um algoritmo que seguisse uma expansão tendenciosa a considerar as bibliografias mais relevantes em relação ao conjunto inicial. Com isso haveria uma redução das bibliografias recuperadas visando aumentar a precisão no acerto de referências relevantes. Então surgiram as três primeiras atividades das quatro atividades do algoritmo proposto pelo presente trabalho,

As três atividades que surgiram foram: pontuação, agrupamento e expansão. O objetivo nesse momento foi melhorar a forma de expansão para que a visualização

apresentasse as bibliografias mais relevantes. Para isso era necessário criar uma maneira de decidir uma medida que indicasse essa relevância. Então surgiu a atividade de pontuação. Essa atividade permitiria que as bibliografias envolvidas recebessem uma pontuação de acordo com sua relevância.

Dito isso, durante as pesquisas do presente trabalho entendeu-se que um possível caminho seria a utilização da estrutura de relacionamentos existentes entre bibliografias. Conforme mencionado na seção 3.1, certos aspectos contribuíram para esse entendimento, como a relevância de contextos de citações representarem bem o conteúdo do documento ao qual fazem citação.

A partir daí foi pesquisado um algoritmo que fosse voltado para tal análise. Essa análise seria feita em cima de um grafo que representaria as ligações existentes entre as diferentes bibliografias envolvidas. Então encontrou-se o HITS como sendo um forte candidato a integrar essa atividade desse estudo.

A atividade de agrupamento surgiu em consequência da necessidade de controlar quais áreas estariam envolvidas nessa expansão. Isso se faz necessário devido ao fato de estudos normalmente envolverem mais de um tipo de área do conhecimento, sejam elas áreas muito distintas ou apenas subáreas.

Para auxiliar na separação das áreas a que pertencem as bibliografias envolvidas na análise desse estudo, foi encontrado o algoritmo de *Louvain*. Esse algoritmo é direcionado a encontrar agrupamentos, também conhecidos como comunidades em grafos, que representem conjuntos cujos elementos possuam um relacionamento mais forte entre si do que com os elementos de outras comunidades encontradas.

Por fim, como última atividade que surgiu voltada para o refinamento da expansão das bibliografias iniciais, surge a atividade de expansão. Essa seria responsável por utilizar os produtos gerados pelas atividades anteriormente descritas com a finalidade de possuir critérios para efetuar uma decisão de quais bibliografias deveriam ser expandidas a fim de recuperar as mais relevantes de acordo com os critérios estabelecidos, conforme explicado na seção 3.3.

Com isso o intuito de refinar a expansão do conjunto inicial de bibliografias passou a ser coberto, produzindo assim uma visualização filtrada das bibliografias apresentadas ao final de expansões. Porém, o estudo ainda manteve a determinação de ir mais além. Sua ideia não passou a ser somente fornecer uma forma mais adequada de visualização do conhecimento ou até mesmo construir um mapeamento sistemático, mas também prover sugestões dentro do produto final gerado por suas expansões.

Portanto surge a última atividade, a atividade de sugestão. Para essa atividade pensou-se em algo que fosse simples, porém passível de futuras mudanças e aprimoramento. Então o algoritmo aqui utilizou parte dos produtos e ferramentas que dispôs nas etapas anteriores. A decisão aqui for optar pelas n melhores referências

que estivesse resultante de i expansões cujas suas pontuações obtidas nas atividades anteriores estivessem entre as n maiores pontuações. Concluindo assim uma forma de sugestão de bibliografias.

É claro que nessa última atividade muitos outros aspectos poderiam ter sido considerados. Porém, devido ao tempo para o desenvolvimento do trabalho, aprimoramentos para essa atividade ficaram para trabalhos futuros. Conforme serão descritos em um capítulo oportuno.

Vale aqui citar que o presente estudo demandou, além da parte teórica, a parte de implementação de uma ferramenta funcional para realização de experimentos desse trabalho, e também para ser livremente utilizada por pesquisadores. Somado a isso, houve uma preocupação no presente estudo para que se mantivesse ao máximo, a nível tanto de teoria como de implementação, tudo o mais escalonável e adaptável possível a fim de que futuramente esse trabalho continue sendo avançado.

A implementação utilizada foi a disponibilizada através de um módulo para linguagem Python chamado *Community*.Esse módulo foi implementado por Thomas Aynaud e encontra-se atualmente disponível em https://bitbucket.org/taynaud/python-louvain

4.2 Escolha de parâmetros

Conforme pode ser percebido no capítulo 3 sobre a proposta apresentada, alguns parâmetros surgiram a fim de controlar os aspectos mencionados na seção 1.2 sobre o problema envolvido. Esses parâmetros estão relacionados diretamente a arquitetura desenvolvida para suportar o desenvolvimento desse trabalho. Parâmetros variáveis: profundidade da busca, número de vértices candidatos a serem expandidos por *cluster* encontrado e o sentido da expansão.

Porém, outros parâmetros foram utilizados de forma fixa e também podem ser listados. Nessa lista encontram-se: o algorítimo de pontuação HITS utilizado para calcular a relevância de cada vértice, o qual utilizou a medida de autoridade já descrita e a expansão utilizando somente os últimos vértices adicionados, com a finalidade obter, a cada iteração, maior ganho de especificidade ou multidisciplinaridade, para busca em profundidade ou em largura respectivamente.

4.3 A arquitetura

A arquitetura estabelecida por esse trabalho e implementada para resolução do problema proposto e experimentação das características dissertadas foi composta de 4 componentes: componente de extração, componente de busca e recuperação de dados, componente de processamento de dados e componente de representação. Sendo

o componente principal o de processamento de dados, destinado a implementar a heurística descrita por esse estudo. Cada componente desse descreve uma parte da arquitetura que pode se implementada por diversos tipos de tecnologias e cada componente pode ser substituído por outro cujos padrões de entrada e saída sejam os mesmos.

A arquitetura geral desse sistema proposto pode ser entendida através diagrama mostrado pela imagem 4.1. Seus componentes implementados são descritos passo a passo pelas subseções seguintes de forma genérica sem se ater a uma tecnologia específica. Cada subseção explica como devem ser os componentes para construção de uma ferramenta que implemente a solução proposta pelo presente estudo.

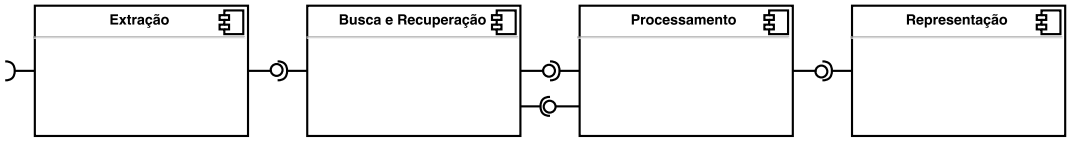

Figura 4.1: Arquitetura geral

4.3.1 Componente de extração

O componente de extração deve ser composto por um módulo capaz de ler um arquivo de entrada com uma listagem de referências a ser analisada. Essa listagem deve estar em um formato padrão que exiba os metadados de cada referência a fim de ser possível utilizar um *parser* que leia tais informações e as disponibilize de forma que possa ser processada. Esse estudo utilizou o padrão BIBTEX para representar suas referências.

Portanto, esse componente recebe como entrada um arquivo padrão e seu módulo de processar BIBTEX disponibiliza um objeto padronizado para acesso e leitura de todos os metadados das referências listadas. Sua representação é mostrada pela figura 4.2:

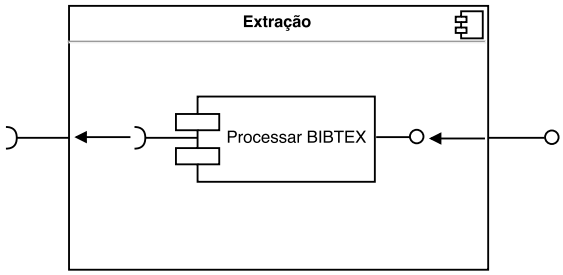

Figura 4.2: Componente voltado para processar listagens de referências em um formato padrão. Sua saída é um objeto com acesso padronizado para leitura das informações processadas

4.3.2 Componente de busca e recuperação

O componente de busca e recuperação deve ser composto por um módulo capaz de buscar em uma base de dados, seja ela local ou remota, informações acerca de referências, citações e detalhes associados a uma publicação. Para isso, esse componente recebe como entrada um objeto com acesso padronizado contendo informações de um ou mais artigos a serem buscados em formato de lista ou em formato de grafo. A saída disponibilizada por esse componente deverá ser um objeto grafo também padronizado para posterior acesso a suas informações.

O objeto de saída deve ser construído com base em 2 informações básicas: identificador de um artigo e ligações de referência e citação entre os artigos contidos no objeto. Esse também deve ser capaz de armazenar em sua estrutura interna dados associados a cada artigo representado pelos vértices do grafo gerado. A imagem 4.3 ilustra sua representação:

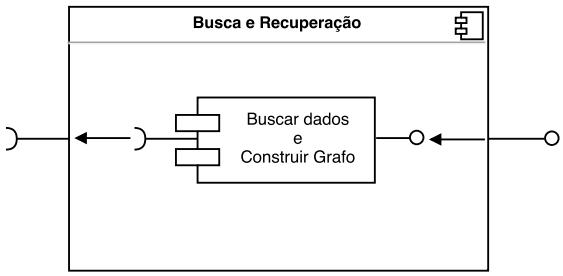

Figura 4.3: Componente voltado para processar listagens de referências em um formato padrão. Sua saída é um objeto com acesso padronizado para leitura das informações processadas

4.3.3 Componente de processamento

O componente de processamento é o cerne da implementação. Esse é o componente responsável por implementar a heurística proposta por esse estudo. Nele são processados os dados recebidos, através de um objeto com formato padronizado de acesso, e disponibiliza-se um objeto de saída para leitura e apresentação das informações geradas.

Esse componente é composto por 4 módulos: pontuação, clusterização, expansão e sugestão conforme mostra a figura 4.4 :

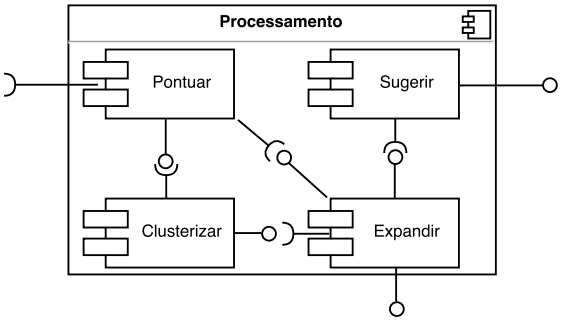

Figura 4.4: Componente voltado para processar listagens de referências em um formato padrão. Sua saída é um objeto com acesso padronizado para leitura das informações processadas

O módulo de pontuação é responsável por fazer a leitura inicial do objeto recebido pelo componente e processar os dados de acordo com o método proposto. Esse componente disponibiliza como saída um objeto com dados em formato de grafo, cujos vértices encontram-se pontuados.

O módulo de clusterização, recebe um objeto com os dados em formato de grafo e encontra *clusters* seguindo a o método escolhido para esse fim. Sua saída é o objeto grafo de entrada acrescido de atributos que armazenem as listas de *clusters* formados com seus respectivos vértices.

O módulo de expansão recebe como entrada um objeto grafo. O módulo é responsável por selecionar as informações a serem expandidas e as disponibilizar em formato de objeto grafo que contém em um atributo uma listagem de vértices desse grafo a serem expandidos. Ou ainda, esse módulo é responsável apenas por entregar o objeto de saída em formato de grafo já expandido.

O módulo de sugestão recebe como entrada um objeto grafo cujos vértices possuem pontuações acerca de suas relevâncias. Esse módulo é responsável por selecionar, seguindo seu método interno os vértices a integrarem uma listagem sugerida. Essa listagem passará a integrar um atributo do objeto grafo de saída.

Portanto, esse componente recebe como entrada um objeto grafo que pode passar por uma expansão iterativa necessitando se comunicar com o componente de busca e recuperação. Esse componente disponibiliza como saída um objeto grafo podendo conter uma listagem de vértices sugeridos.

4.3.4 Componente de representação

O componente de representação possui o módulo de gerar visualização que é responsável pela parte visual das informações. Esse recebe um objeto grafo com seus possíveis atributos adicionados, realiza sua leitura e o representa graficamente. A implementação desse componente pode se dar não necessariamente utilizando mode-

los de representação em grafos, mas qualquer modelo que seja capaz de representar os relacionamentos recebidos.

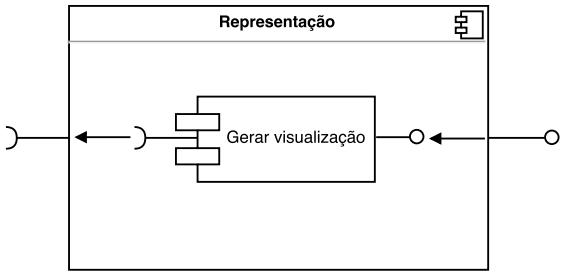

Figura 4.5: Componente voltado para processar listagens de referências em um formato padrão. Sua saída é um objeto com acesso padronizado para leitura das informações processadas

4.4 A implementação

A titulo de informações adicionais, a arquitetura descrita foi implementada como uma aplicação disponibilizada através de uma página para acesso via *browser*. Esse trabalho utilizou *framework* Django (`http://www.djangoproject.com`)) através da linguagem Python 2.7 para implementar a aplicação no lado do servidor. Módulos Python como: NetworkX (HAGBERG *et al.*, 2013) e *Community* (AYNAUD, 2009) também foram utilizados.

Para o lado do cliente, foi utilizado HTML (*HyperText Markup Language*) para descrever a estrutura de informações da tela da ferramenta. Somando-se a isso, foi utilizado JavaScript através de bibliotecas como Jquery e JSNetworkX (`http://jsnetworkx.org`), para controlar os comportamentos da página e gerar a visualização dos grafos (esses gerados utilizando SVG (Scalable Vector Graphics)). E por fim, a implementação também utilizou CSS (*Cascading Style Sheets*) para controlar o estilo da página.

A implementação dos componentes de: extração, busca e recuperação e processamento foi desenvolvida para o lado do servidor *web*. O componente de representação foi implementado para ser executado no lado do cliente através de um *browser* de internet. Maiores detalhes sobre como funciona essa implementação podem ser encontrados no apêndice B, tutorial da ferramenta. A imagem: 4.6 apresenta a implementação com a tecnologia citada:

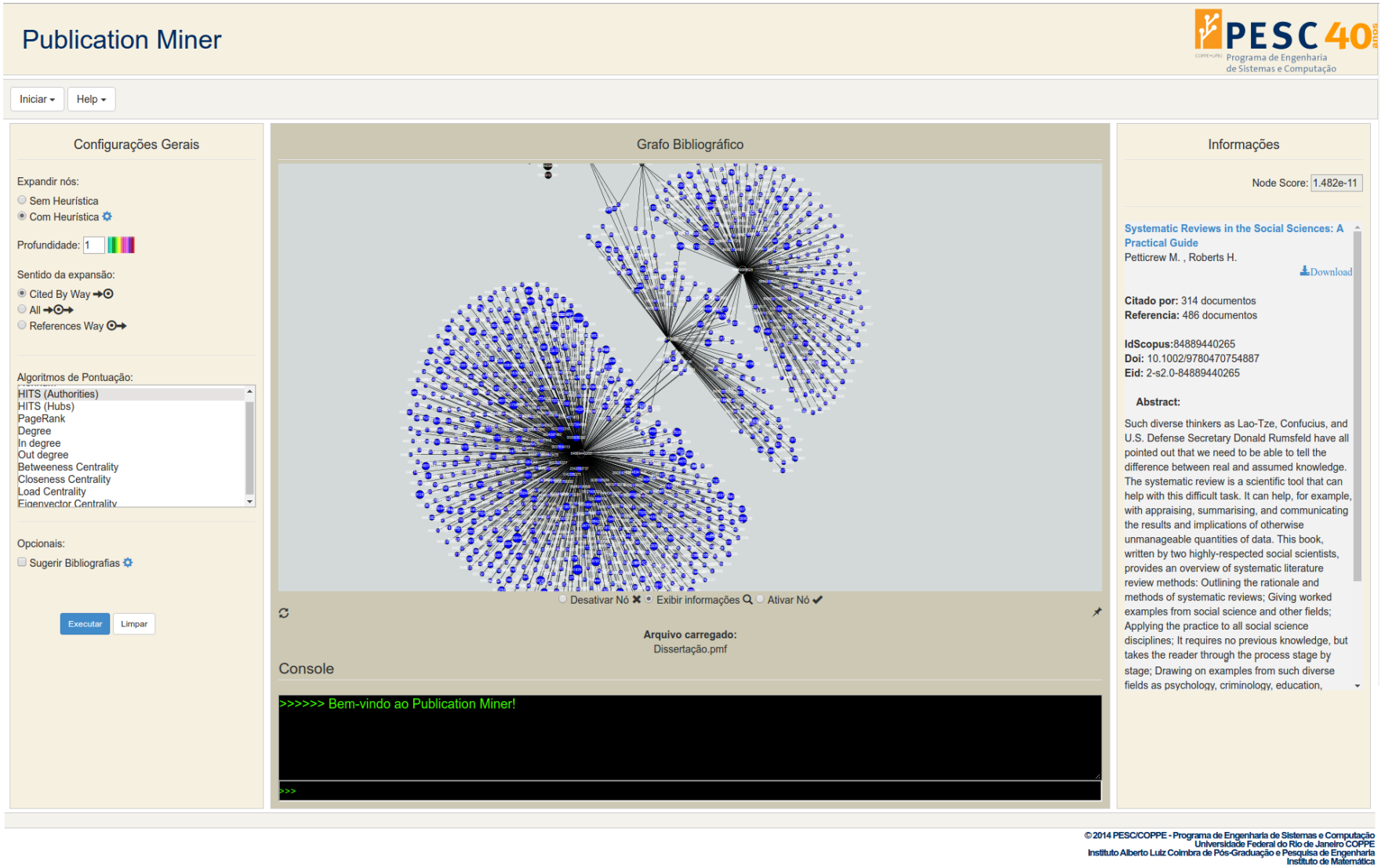

Figura 4.6: Tela utilizada pelo cliente através do *browser*.

Capítulo 5

Experimentos

Os experimentos são parte fundamental para um estudo ter suas ideias confrontadas com a realidade prática. Trata-se da forma em que a ciência dispõe para verificar se a teoria de fato pode ser aplicada ao mundo real e se a realidade se comporta conforme modelada.

Esse estudo abordou questões ligadas a uma tarefa prática: realizar pesquisas bibliográficas. Através dessa tarefa conseguiu identificar um problema a ser resolvido, encontrar bibliografias negligenciadas, auxiliando em mapeamentos sistemáticos, assim como identificou diversas questões periféricas a serem respondidas.

Dando continuidade, esse também dissertou sobre uma proposição para resolução desse problema de forma teórica e prática. Formalizou o problema e sua forma de tentar resolvê-lo. Porém, ainda restou submeter todo esse arcabouço teórico à experimentação. Portanto, esse capítulo destina-se a esse fim, verificar as proposições feitas.

De acordo com o descrito no capítulo 1.2, sobre o problema abordado por esse estudo, o presente capítulo visa experimentar as respostas dadas às indagações feitas acerca dos problemas relacionados à condução de avaliações sobre diversas questões envolvendo pesquisas bibliográficas. Para isso foi utilizada a ferramenta previamente descrita para realização de experimentos que permitissem uma avaliação dessas indagações.

A ideia principal foi utilizar dois modelos de avaliação, um voltado para analisar o impacto quantitativo causado na forma de realizar pesquisas bibliográficas, e outro voltado para analisar o impacto qualitativo. E para esse fim, foram realizadas comparações entre a forma auxiliada pela heurística e a forma não auxiliada nos dois modelos. Cada um desses modelos foi utilizado em um experimento.

Esse estudo propõe dois experimentos, cada um focado em um validar aspectos distintos dessa pesquisa. São eles: **uso coletivo por tópico** e o **uso por tema especializado**.

O **uso coletivo por tópico** foi uma abordagem quantitativa voltada para men-

surar os resultados obtidos pelo uso da heurística e compará-los com a ausência desse suporte. Essa utilizou um tema aleatório a ser pesquisado por um grupo de pessoas também aleatório.

O experimento: **uso por tema especializado** foi uma abordagem qualitativa cujos experimentos foram voltados para avaliar os conceitos principais desse estudo. Para isso foram utilizados temas nos quais os avaliadores fossem especialistas.

Esse capítulo foi organizado através desses dois experimentos supracitados contendo os seguintes tópicos cada um: conceitos a serem avaliados, os objetivos nos experimentos, voluntários e temas, a execução dos experimentos, avaliação dos voluntários e análise dos resultados. As tarefas a serem avaliadas individualmente estão relacionadas às propostas desse estudo para resolver o problema central ou responder à questões periféricas ao tema.

5.1 O experimento 1 - Uso coletivo por tópico

O presente experimento buscou realizar uma análise dos resultados obtidos pelos voluntários durante pesquisas bibliográficas de forma a extrair informações quantitativas. Suas tarefas foram voltadas para tentar mensurar o auxílio fornecido pela heurística descrita por esse estudo.

Esse experimento possuiu foco na análise coletiva dos voluntários envolvidos em tarefas iguais, direcionados ao mesmo tema central, porém divididos em 2 grupos, um utilizando a heurística proposta por esse estudo e outro não utilizando. Para uso da heurística, os voluntários não receberam parâmetros preestabelecidos. Esses foram mantidos o mais próximo à realidade que encontrariam ao dispor dessa ferramenta durante suas pesquisas bibliográficas reais.

Os conceitos a serem quantificados nesse experimento estão relacionados a perguntas de caráter objetivo. Os produtos finais de cada tarefa serviram de dados para a analise quantitativa final. Através dele o estudo pode ser avaliado de um ponto de vista macro, diferente da visão da segunda abordagem, cuja enfoque foi micro, focado em cada voluntário e direcionada as questões teóricas desse estudo.

5.1.1 Conceitos a serem avaliados

Os conceitos a serem avaliados por esse experimento envolvem características que podem ser percebidas em um ambiente real de pesquisa. São características que se apresentam de forma prática, surgindo ao longo de uma pesquisa e estando relacionadas a questionamentos básicos que podem surgir ao pesquisador se confrontar com uma área desconhecida.

Uma dessas característica, que pode ser percebida ao longo de uma pesquisa bibli-

ográfica, é a descoberta de documentos relacionados que são relevantes ao propósito central de um estudo. Deve-se entender esses documentos relacionados como sendo pertencentes a assuntos periféricos ao tema central. Um exemplo disso poderia surgir ao se tentar dissertar sobre processamento de linguagem natural. Documentos sobre aplicações de processamento de linguagem natural em língua portuguesa seriam periféricos ao tema central. Muitas vezes deseja-se descobrir meios de referenciar assuntos periféricos para certos estudos, porém se desconhece publicações para tal fim.

Outra característica que pode-se perceber é a existência de documentos que poderiam ser referenciados pelo o tema central, porém são muito específicos de um outro tema. Esses seriam documentos que abordam a fundo conceitos relacionados que são utilizados pelo tema central. Um exemplo disso seria ao pesquisar por processamento de linguagem natural se deparar com artigo de estatística sobre modelos de Markov. Apesar de ser um conceito utilizado pelo tema central, e candidato a ser referenciado, trata-se de um outro tema.

Uma terceira característica que pode ser percebida, de certa forma intuitiva, é descobrir quais são os artigos mais relevantes de certa área. No geral, quando se está realizando uma pesquisa acerca de um tema, deseja-se descobrir quais são as publicações mais influentes sobre sua área. Em outras palavras, quando deseja-se descobrir o estado da arte de uma certa área, é necessário dominar essa questão.

Os autores também são objetos de observação no cenário das publicações. Ao observá-los é possível perceber que determinados autores exercem maior autoridade sobre certos assuntos do que outros (WAGNER e LEYDESDORFF, 2005). Isso pode ser percebido através das citações, por exemplo. Nesse cenário pode-se perceber que certos autores possuem artigos muito citados em certas áreas. Saber quem são os autores mais influentes em um área de interesse pode agregar informação muito útil durante as pesquisas.

Palavras-chave também são instrumentos relevantes ao se realizarem busca. São mecanismos que ocorrem naturalmente ou as vezes propositalmente para tornar o estudo mais fácil de ser identificado. Então ao se conhecer as palavras-chave relacionadas ao tema no qual está realizando uma busca, pode-se de forma mais fácil encontrar artigos mais relevantes por saber como os documentos se relacionam.

Logo, esse experimento visa: (i) Avaliar o suporte provido pela heurística para o esclarecimento conceitos anteriores; (ii) Avaliar a experiência pessoal do voluntário com o uso da ferramenta.

A avaliação (ii) visa responder as seguintes questões:

Q_1: a implementação atendeu aos objetivos principais?

Q_2: Qual foi a qualidade dos estudos sugeridos?

Q_3: Qual a opinião do voluntário sobre a facilidade no uso da ferramenta?

Q_4: No geral como o voluntário classifica a experiência com a implementação da heurística apresentada?

As questões anteriores são perguntas a serem respondidas para avaliar os tópicos de forma quantitativa. A avaliação qualitativa será realizada no experimento 2.

5.1.2 Os objetivos dos experimentos

Essa subseção descreve os objetivos a serem alcançados por cada voluntário, avaliando às características descritas na subseção 5.1.1, durante as pesquisas bibliográficas. A seguir serão enumerados e descritos os 5 objetivos a serem alcançados.

O primeiro objetivo está ligado à **descoberta de documentos de temas relacionados ao tema central**, ou, em outras palavras, documentos periféricos, conforme exemplificado em 5.1.1. Para essa tarefa os voluntários listaram as publicações mais relevantes relacionadas ao tema que lhes foi informado.

O segundo objetivo avaliou **a capacidade de encontrar documentos que detalhassem pontos específicos do tema central**. Os voluntários deveriam atentar-se para o grau de detalhamento do conteúdo envolvido. Isto é, deveria listar documentos específicos de outras áreas, conforme exemplificado em 5.1.1, para análise dos resultados.

O terceiro objetivo, avaliou **a capacidade de encontrar os documentos mais relevantes do tema central** pelos voluntários. Os voluntários julgaram foram solicitados a separar os 5 artigos mais relevantes do tema central.

O quarto objetivo avaliou a **identificação dos autores mais influentes para o tema central**. O objetivo foi avaliar se os voluntários seriam capazes de localizar esses autores e se haveria uma convergência para um grupo em comum. Para isso, os voluntário tiveram que listar no máximo 5 autores que julgassem mais influentes para o tema central.

O quinto objetivo avaliou a **análise da capacidade de identificação das palavras-chave** de um estudo. O Objetivos se dividiu em 2 tarefas voltadas para questões distintas:

1. Realizar a identificação de palavras-chave que o voluntário identificou como relevantes ao tema central; 2. Identificar palavras-chave dos temas periféricos.

Além desses cinco objetivos, o voluntário também respondeu à peguntas relacionadas à sua experiência pessoal com a ferramenta proposta. Também vale ressaltar que no primeiro e no segundo objetivo os voluntários não foram instruídos com a quantidade mínima ou máxima de documentos que deveriam selecionar e, portanto, deveriam usar a ferramenta e informar quais foram selecionados.

5.1.3 Voluntários e temas

Foram selecionados 18 alunos de pós-graduação do programa de engenharia de sistemas e computação do instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa de Engenharia (COPPE) da Universidade Federal do Rio de janeiro (UFRJ). Os alunos estão cursando mestrado, encontravam-se matriculados na disciplina de busca e recuperação da informação e possuíam experiência prévia para realização de pesquisas bibliográficas. Para a realização dos experimentos os alunos foram divididos em 2 grupos de mesma quantidade: G_1 Alunos usando a heurística; G_2 Alunos usando apenas o mecanismo de busca da Scopus. Ademais, o tema central utilizados, por todos os alunos, para realização dos experimentos foi: sumarização de texto.

5.1.4 A execução dos experimentos

Inicialmente todos os alunos foram informados acerca do tema e receberam os artigos *Approaches to text summarization - Questions and answers*(ALONSO *et al.*, 2004), para leitura completa, e *Automatic Summarization* (NENKOVA e MCKEOWN, 2011), para somente leitura do resumo e do capítulo 1. Todos os os voluntários dispuseram de 40 minutos para leitura desses artigos.

Em seguida, os voluntários do grupo G_1 receberam instruções de como utilizar a heurística, proposta por esse estudo, por meio de uma ferramenta que fornece suporte às tarefas do experimento. Esse grupo ficou livre para usar os parâmetros que julgasse mais apropriados. O grupo G_2 recebeu instruções de como realizar as mesmas tarefas, utilizando o motor de buscas *Scopus*.

Dito isso, os voluntários dispuseram de 1 hora e 30 minutos para realização de suas pesquisas bibliográficas. Apesar da limitação ao tempo de execução poder representar uma ameaça ao experimento, houve a necessidade de fixá-lo. Durante a execução de suas pesquisas, os voluntários não puderem fazer contato uns com os outros. Isso foi imposto a fim de evitar que a experiência de um grupo pudesse influenciar na do outro ou até mesmo que houvesse influência entre pessoas do mesmo grupo.

Após as instruções sobre o tema e as ferramentas de busca, os voluntários foram apresentados a um formulário de avaliação dos experimentos cujas as perguntas foram destinadas a cumprir os objetivos descritos na subseção 5.1.2, vale frisar que o quinto objetivo deu origem aos dois últimos tópicos. Esses tópicos foram:

i) Liste os artigos relacionados (periféricos) ao tema sumarização que você achou mais relevantes durante suas buscas.

ii) Liste os artigos que encontrou durante suas buscas por sumarização que iden-

tificou como detalhando um ponto específico de forma profunda. Esses artigos, apesar de serem referenciados por artigos de sumarização, são pertencentes a outra área e foram utilizados como suporte (seja matemático, computacional ou até mesmo para simples contextualização dos artigos de sumarização).

iii) Liste os 5 artigos mais relevantes para o tema sumarização encontrados em suas buscas.

iv) Liste, no máximo, os 5 autores mais influentes para o tema sumarização. (Citar conforme consta nas publicações encontradas em suas buscas)

v) Liste as palavras-chave que você identificou como relevantes para sumarização durante suas buscas.

vi) Liste as palavras-chave de assuntos relacionados (periféricos) ao tema sumarização encontrados em suas buscas.

O grupo G_1, que utilizou a ferramenta proposta por esse estudo, também precisou responder a perguntas relacionadas à sua experiência pessoal com o uso da ferramenta. Essas perguntas foram:

i) A implementação atendeu aos seus objetivos principais? Responder em um escada de 1 a 5, na qual 1 significa que discorda totalmente e 5 significa que concorda totalmente

ii) Qual foi a qualidade dos estudos sugeridos? Responder em uma escala de 1 a 5, na qual 1 significa considerou muito ruins e 5 que considerou excelentes

iii) Na sua opinião, como se caracteriza o uso da ferramenta? (Avaliar a facilidade). Responder em uma escala de 1 a 5, na qual 1 significa extremamente difícil e 5 significa extremamente fácil

iv) No geral, como você classifica sua experiência com a ferramenta utilizada? Responder em uma escala de 1 a 5, na qual 1 significa muito ruim e 5 significa excelente.

5.1.5 Análise dos resultados

Essa subseção realiza uma síntese dos dados gerados pela presente experimento. Esses resultados são apresentados integralmente no apêndice A. O intuito aqui é sumarizar os dados apresentados de forma a discuti-los posteriormente no capítulo de conclusões.

Os grupos G_1 e G_2 foram inicialmente submetidos ao mesmo conjunto de perguntas de 1 a 6. Além dessas 6 perguntas iniciais, o grupo G_1, que utilizou a heurística

implementada, foi submetido a 4 perguntas adicionais, de caráter pessoal, acerca da ferramenta proposta.

Um total de 18 respostas distintas foram contabilizadas para as perguntas de 1 a 6 e 9 respostas para as 4 perguntas de caráter pessoal. Portanto, foram sintetizadas um total de 108 respostas conceituais e 36 relacionadas a experiência pessoal com a implementação.

Para melhor visualização das respostas às perguntas de 1-3, devido à extensão dos títulos envolvidos, foi criada uma listagem, conforme apresentado na tabela 5.1. Seu intuito é servir de dicionário para tradução dos resultados apresentados em cada questão.

O intervalo de dados referente a cada resposta foi formado pela união de todos identificadores dos títulos dos artigos apresentados pelos 2 grupos de voluntários envolvidos. A cada artigo, foi atribuído um número identificador sequencial a ser representado em cada conjunto de respostas.

O mesmo princípio foi utilizado para os nomes dos autores citados e das palavras-chave nas respostadas dadas às questões 4, 5 e 6 respectivamente. Foram criadas listagens, por questão, baseadas na união das respostas dadas pelos dois grupos. Porém para essas questões não houve a necessidade de criação de um dicionário com identificadores sequenciais devido as palavras serem menores, o que não atrapalharia visualmente a análise posterior dos resultados.

Tabela 5.1: Dicionário dos Artigos

Id	**Título**
1	A compact forest for scalable inference over entailment and paraphrase rules.
2	A comparison of multiple approaches for the extractive summarization of Portuguese texts.
3	A comprehensive comparative evaluation of RST-based summarization methods.
4	A context-based word indexing model for document summarization.
5	A four dimension graph model for automatic text summarization.
6	A hybrid approach to automatic text summarization.
7	A Machine Learning Approach for Displaying Query Results in Search Engines.
8	A machine learning approach to sentence ordering for multidocument summarization and its evaluation.
9	A new approach for single text document summarization.
10	A new approach to hierarchical clustering and structuring of data with self-organizing maps.
11	A new evaluating method for Chinese text summarization not requiring model summary.
12	A new lexical chain algorithm used for automatic summarization.
13	A novel Chinese text summarization approach using sentence extraction based on kernel words recognition.
14	A query-oriented summarization system for XML elements.
15	A survey of paraphrasing and textual entailment methods.
16	A Survey of Text Summarization Extractive techniques.
17	A survey of text summarization techniques.
18	A system for query-specific document summarization.
19	Abstractive summarization of voice communications.
20	Advances in automatic text summarization.
21	Advantages of query biased summaries in information retrieval.
22	An empirical study of the textual similarity between source code and source code summaries.
23	An improved evolutionary algorithm for extractive text summarization.
24	Application and analysis of content-similarity-based automatic evaluation for summarization systems.

Tabela 5.1: Dicionário dos Artigos

Id	Título
25	Approaches to text summarization: Questions and answers.
26	Assessing sentence scoring techniques for extractive text summarization.
27	Assessing sentence similarity through lexical, syntactic and semantic analysis.
28	Automated multi-document summarization in neats.
29	Automated summarization evaluation based on clouds model.
30	Automated text summarization and the SUMMARIST system
31	Automatic abstractive summarization a systematic literature review.
32	Automatic Arabic text summarization: a survey.
33	Automatic condensation of electronic publications by sentence selection.
34	Automatic evaluation of information ordering: Kendall's Tau.
35	Automatic extractive multi-document summarization based on archetypal analysis.
36	Automatic multi document summarization approaches.
37	Automatic soccer video analysis and summarization.
38	Automatic summarising: The state of the art.
39	Automatic summarization method for Chinese document based on comprehensive background concept lattice.
40	Automatic summarization of online customer reviews.
41	Automatic text document summarization based on machine learning.
42	Automatic text structuring and summarization.
43	Automatic Text Summarization.
44	Challenges of automatic summarization.
45	Challenging issues of Automatic Summarization: Relevance detection and quality-based evaluation.
46	Chinese text automatic summarization based on affinity propagation cluster.
47	Clustering techniques and discrete particle swarm optimization algorithm for multi-document summarization
48	Clustering-based visual interfaces for presentation of web search results: An empirical investigation.
49	Coherent narrative summarization with a cognitive model.
50	Comparing summarisation techniques for informal online reviews.
51	Concept generalization and fusion for abstractive sentence generation.
52	Constructing literature abstracts by computer: techniques and prospects.
53	Discourse indicators for content selection in summarization.
54	DUC in context.
55	Efficient text summarization using lexical chains.
56	Efficient Voting-Based Extractive Automatic Text Summarization Using Prominent Feature Set.
57	Evaluation method of automatic summarization calculating the similarity of text based on hownet.
58	Evaluation of a sentence ranker for text summarization based on Roget's thesaurus.
59	Evaluation of automatic text summarization methods based on rhetorical structure theory.
60	Exploring events and distributed representations of text in multi-document summarization.
61	Extracting appraisal expressions.
62	Extractive single-document summarization based on genetic operators and guided local search.
63	Extractive text summarization using lexical association and graph based text analysis.
64	Fuzzy swarm based text summarization.
65	Gather customer concerns from online product reviews - A text summarization approach.
66	Generating Impact-Based Summaries for Scientific Literature.
67	Generic summaries for indexing in information retrieval
68	Generic text summarization using relevance measure and latent semantic analysis.
69	High quality information extraction and query-oriented summarization for automatic query-reply in social network.
70	Implementation and evaluation of evolutionary connectionist approaches to automated text summarization.
71	Improvement in quality of extractive text summaries using modified reciprocal ranking.
72	Improving web search ranking by incorporating summarization.
73	Inferring strategies for sentence ordering in multidocument news summarization.
74	Information navigation on the web by clustering and summarizing query results.
75	Kernel-based approach for automatic evaluation of natural language generation technologies: Application to automatic summarization.
76	Latent aspect rating analysis on review text data: A rating regression approach.

Tabela 5.1: Dicionário dos Artigos

Id	**Título**
77	Learning algorithms for keyphrase extraction
78	LexRank: Graph-based lexical centrality as salience in text summarization.
79	Multi-document summarization via group sparse learning.
80	Multi-video summarization based on Video-MMR.
81	Multidocument summarization: An added value to clustering in interactive retrieval
82	Multilingual summarization approaches.
83	Multimedia summarization for trending topics in microblogs.
84	Multiple documents summarization based on evolutionary optimization algorithm.
85	Multivariate Fuzzy C-Means algorithms with weighting.
86	Musical rhythmic pattern extraction using relevance of communities in networks.
87	New Methods in Automatic Extracting
88	News filtering and summarization on the Web.
89	NewsInEssence: summarizing online news topics.
90	On conceptual indexing for data summarization.
91	Optimizing text summarization based on fuzzy logic.
92	Panorama: Extending digital libraries with topical crawlers.
93	Paraphrase Extraction using fuzzy hierarchical clustering.
94	Popularity-based summarization of Chinese text: Implicit weight-based features for newspaper articles.
95	Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization.
96	Semantic inference at the lexical-syntactic level.
97	Semanticrank: Ranking keywords and sentences using semantic graphs.
98	Steds: Social Media Based Transportation Event Detection with Text Summarization.
99	Summarising customer online reviews using a new text mining approach.
100	Summarization evaluation without human models.
101	Summarization from medical documents: a survey.
102	Summarization of documentaries.
103	Summarization of films and documentaries based on subtitles and scripts.
104	Summarizing microblogs automatically.
105	Summarizing online customer reviews automatically based on topical structure.
106	Summarizing scientific articles: experiments with relevance and rhetorical status.
107	Summarizing text documents: sentence selection and evaluation metrics.
108	Supporting searching on small screen devices using summarisation.
109	SVM-Based Multi-Document Summarization Integrating Sentence Extraction with Bunsetsu Elimination.
110	Text mining techniques for patent analysis.
111	Text structuration leading to an automatic summary system: RAFI.
112	Text summarisation in progress: A literature review.
113	Text summarization by sentence segment extraction using machine learning algorithms.
114	Text summarization method applying vocabulary combination into sentence extraction.
115	The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases.
116	The effectiveness of automatic text summarization in mobile learning contexts.
117	Towards opinion summarization from online forums.
118	Tree view self-organisation of web content - Institute for Water Education.
119	User-based video abstraction using visual features.
120	Using Contextual Topic Model for a Query-Focused Multi-Document Summarizer.
121	Using lexical chains for text summarization.

Define-se o termo **concordância intragrupo** como sendo a convergência de 2 ou mais voluntários do mesmo grupo listando um mesmo item como parte de sua resposta. E **concordância intergrupos** como sendo a convergência de um ou mais voluntários de cada grupo listando um mesmo item como parte de suas repostas.

A síntese dos conjuntos de respostas dadas às 6 perguntas comuns aos grupos G_1

(alunos usando a heurística) e G_2 (alunos usando apenas o mecanismo de busca da Scopus) será apresentada pergunta a pergunta fornecendo os resultados obtidos por ambos os grupos a fim de facilitar as futuras conclusões. Seguem-se as informações geradas:

1) Liste os artigos relacionados (periféricos) ao tema sumarização que você achou mais relevantes durante suas buscas.

O conjunto união dos títulos listados pelos grupos G_1 e G_2 para resposta da presente questão foi um total de 65 artigos diferentes. Nesse universo houve um total de 5 concordâncias intragrupo G_1 . Essas surgiram com as seguintes quantidades de voluntários convergindo para a mesma resposta, seguidas pelo número de citações encontrado na literatura, a título de auxilo para posterior interpretação dos dados, conforme apresentado pela tabela 5.2:

Tabela 5.2: Tabela de concordância da questão 1 - intragrupo G_1

ID_ARTIGO	**Qnt Voluntários**	**Citações**
21	3	174
67	2	30
74	2	48
78	2	543
87	3	515

Para o grupo G_2, houve um total de 5 concordâncias intragrupo. Essas se apresentaram de acordo com a tabela 5.3:

Tabela 5.3: Tabela de concordância da questão 1 - intragrupo G_2

ID_ARTIGO	**Qnt Voluntários**	**Citações**
43	2	85
69	2	2
70	2	7
79	2	0
103	3	0

Com relação a concordância intergrupos, essa pode ser observada apenas 1 vez conforme apresentado pela tabela 5.4, a qual exibe também o número de citações do artigo no qual houve a concordância:

Tabela 5.4: Tabela de concordância da questão 1 - intergrupos G_1 e G_2

ID_ARTIGO	**Qnt Voluntários (G_1 e G_2)**	**Citações**
42	1 e 1	191

2) Liste os artigos que encontrou durante suas buscas por sumarização que identificou como detalhando um ponto específico de forma profunda. Esses artigos, apesar de serem referenciados por artigos de sumarização, são pertencentes a outra área e foram utilizados como suporte (seja matemático, computacional ou até mesmo para simples contextualização dos artigos de sumarização).

O conjunto união dos títulos listados pelos grupos G_1 e G_2 para resposta da presente questão foi um total de 51 artigos diferentes. Nesse universo houve um total de 1 concordância intragrupo G_1 . Essa surgiu com a seguinte quantidade de voluntários convergindo para a mesma resposta, seguida pelo número de citações encontrado na literatura conforme apresenta a tabela 5.5:

Tabela 5.5: Tabela de concordância da questão 2 - intragrupo G_1

ID_ARTIGO	**Qnt Voluntários**	**Citações**
21	2	174

Para o grupo G_2, houve um total de 1 concordância intragrupo. Essa se apresentou conforme mostrado na tabela 5.6 a seguir:

Tabela 5.6: Tabela de concordância da questão 2 - intragrupo G_2

ID_ARTIGO	**Qnt Voluntários**	**Citações**
103	4	0

Nenhuma concordância intergrupos foi observada.

3) Liste os 5 artigos mais relevantes para o tema sumarização encontrados em suas buscas.

O conjunto união dos títulos listados pelos grupos G_1 e G_2 para resposta da presente questão foi um total de 55 artigos diferentes. Nesse universo houve um total de 12 concordâncias intragrupo G_1 . Essas surgiram com as seguintes quantidades de voluntários convergindo para a mesma resposta, seguidas pelo número de citações encontrado na literatura conforme apresentado na tabela 5.7:

Tabela 5.7: Tabela de concordância da questão 3 - intragrupo G_1

ID_ARTIGO	**Qnt Voluntários**	**Citações**
16	2	73
21	4	174
38	2	142
42	3	191
52	2	150
67	2	30
74	3	48
77	2	365
81	2	30
87	3	515
89	2	56
106	3	206

Para o grupo G_2, houve um total de 6 concordâncias intragrupo Essas se apresentaram conforme apresentado pela tabela 5.8:

Tabela 5.8: Tabela de concordância da questão 3 - intragrupo G_2

ID_ARTIGO	**Qnt Voluntários**	**Citações**
6	2	4
56	2	0
60	2	0
70	2	7
112	3	41
116	3	7

Com relação a concordância intergrupos, essa pode ser observada 3 vezes. Essa concordância se deu conforme mostrado pela tabela 5.9:

Tabela 5.9: Tabela de concordância da questão 3 - intergrupos G_1 e G_2

ID_ARTIGO	**Qnt Voluntários (G_1 e G_2)**	**Citações**
16	2 e 1	73
42	3 e 1	191
112	1 e 3	41

4) Liste, no máximo, os 5 autores mais influentes para o tema sumarização. (Citar conforme consta nas publicações encontradas em suas buscas).

A essa resposta foi totalizado um total de 62 autores diferentes entre os citados por G_1 e G_2. Desse total houve 11 concordâncias intragrupo G_1. A seguir apresenta-se a tabela 5.10 a qual descreve os autores que fizeram parte do conjunto de concordâncias intragrupo G_1 seguidos pela quantidade de voluntários que concordaram desse grupo.

Além disso, as tabelas apresentadas a seguir também incluem uma pontuação chamada de índice H (HIRSCH, 2005), muito utilizada atualmente para representar o impacto individual de autores devido às suas publicações (THOMAZ *et al.*, 2011). Esse dado foi incluído a fim de auxiliar na discussão futura dos dados aqui apresentados.

Tabela 5.10: Tabela de concordância da questão 4 - intragrupo G_1

Autor	**Qnt Voluntários**	**Índice H**
Buckley, Chris	2	20
Edmundson, Harold P	3	4
Mitra, Mandar	2	15
Moens, M.	2	17
Radev, D.R.	3	24
Salton, Gerard	2	25
Sanderson Mark	2	20
Singhal, Amit	2	13
Sparck-Jones K.	3	10
Teufel, Simone	2	9
Tombros Anastasios	4	8

Para o grupo G_2 houve 3 concordâncias intragrupo a ser mostrado pela tabela 5.11.

Tabela 5.11: Tabela de concordância da questão 4 - intragrupo G_2

Autor	**Qnt Voluntários**	**Índice H**
Elena Lloret	5	4
Palomar, M.	2	11
Salim, N.	2	12

Para a presente questão foi encontrada apenas uma concordância intergrupos:

Tabela 5.12: Tabela de concordância da questão 4 - intergrupos G_1 e G_2

Autor	**Qnt Voluntários (G_1 e G_2)**	**Índice H**
Salton, Gerard	2 e 1	25

5) Liste as palavras-chave que você identificou como relevantes para sumarização durante suas buscas.

A presente questão obteve um total de 8 concordâncias intragrupo G_1. Essa são apresentadas a seguir pela tabela 5.13:

Tabela 5.13: Tabela de concordância da questão 5 - intragrupo G_1

Palavra-Chave	**Qnt Voluntários**
Document	3
Generic summarization	2
Information retrieval	2
Multidocument	2
Summarization	4
Summarizing	2
Text	2
Text summarizantion	3

O grupo G_2 obteve um total de 5 concordâncias intragrupo apresentadas pela tabela 5.14 a seguir:

Tabela 5.14: Tabela de concordância da questão 5 - intragrupo G_2

Palavra-Chave	**Qnt Voluntários**
Automatic summarization	3
Information retrieval	2
Multi-document summarization	2
Text processing	2
Text summarizantion	5

Foi encontrado um total de 7 concordâncias intragrupos G_1 e G_2. Essas são apresentadas pela tabela 5.15:

Tabela 5.15: Tabela de concordância da questão 5 - intergrupos G_1 e G_2

Palavra-Chave	**Qnt Voluntários (G_1 e G_2)**
Automatic summarization	1 e 3
Information retrieval	2 e 2
Query	1 e 1
Summarization	4 e 1
Summarization techniques	1 e 1
Text	2 e 1
Text summarizantion	3 e 5

6) Liste as palavras-chave de assuntos relacionados (periféricos) ao tema sumarização encontrados em suas buscas.

Essa questão obteve um total de 3 concordâncias intragrupo G_1, as quais são apresentadas a seguir na tabela 5.16

Tabela 5.16: Tabela de concordância da questão 6 - intragrupo G_1

Palavra-Chave	**Qnt Voluntários**
Abstract	2
Clustering	3
Summarization	2

O grupo G_2 obteve um total de 2 concordâncias intragrupo, as quais são apresentadas pela tabela 5.17:

Tabela 5.17: Tabela de concordância da questão 6 - intragrupo G_2

Palavra-Chave	**Qnt Voluntários**
Natural language processing systems	2
Text processing	2

Em relação a concordância intergrupos, foram obtidas 2 concordâncias conforme mostrado a seguir pela tabela 5.18

Tabela 5.18: Tabela de concordância da questão 6 - intergrupos G_1 e G_2

Palavra-Chave	**Qnt Voluntários**
Information retrieval	1 e 1
Text processing	1 e 2

A seguir apresentam-se os resultados das avaliações relacionadas à experiência pessoal dos voluntários do grupo G_1. Voltada a essas avaliações foi utilizada uma escala Likert, criada por Rensis Likert (LIKERT, 1932), para medir a opinião de cada avaliador com relação a uma afirmação ou pergunta apresentada.

As avaliações, utilizando a escala mencionada, foram destinadas a medir a satisfação dos voluntários com os resultados obtidos pela heurística e sua forma de implementação através da ferramenta apresentada. Seguem-se os resultados:

1) A implementação atendeu aos seus objetivos principais?

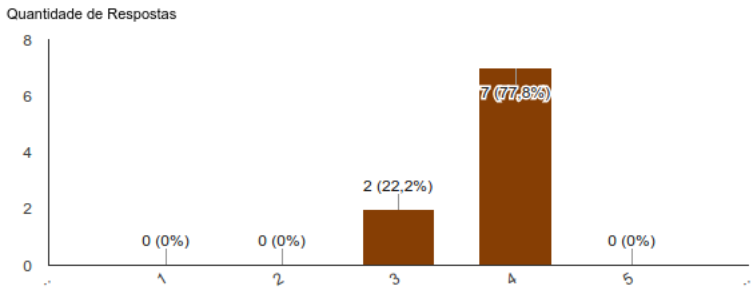

Figura 5.1: Escala de 1-5, onde 1 significa: discordo totalmente e 5 significa: concordo totalmente

2) Qual foi a qualidade dos estudos sugeridos?

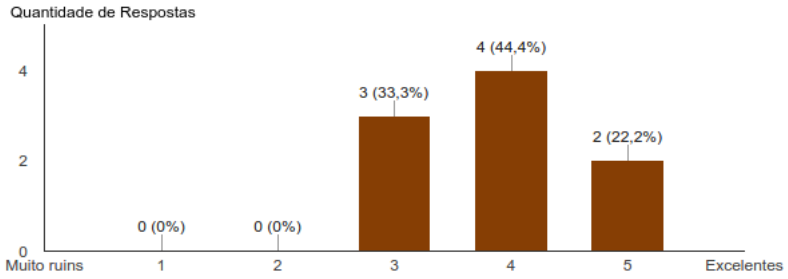

Figura 5.2: Escala de 1-5, onde 1 significa: muito ruim e 5 significa: excelente

3) Na sua opinião, como se caracteriza o uso da ferramenta? (Avaliar a facilidade)

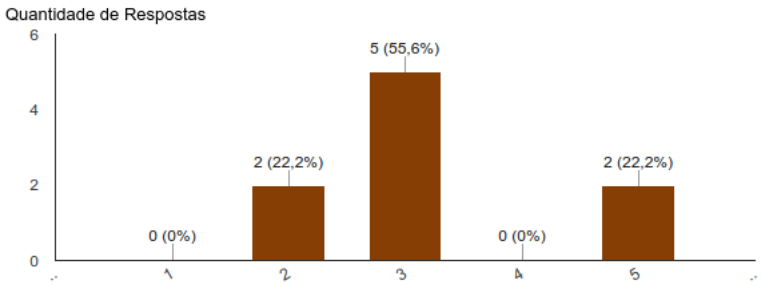

Figura 5.3: Escala de 1-5, onde 1 significa: extremamente difícil e 5 significa: extremamente fácil

4) No geral, como você classifica sua experiência com a ferramenta utilizada?

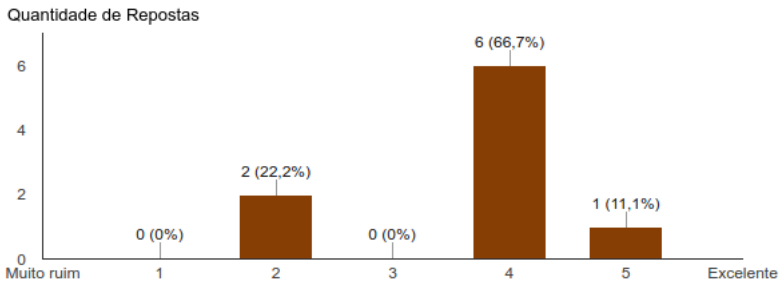

Figura 5.4: Escala de 1-5, onde 1 significa: muito ruim e 5 significa: excelente

5.2 O experimento 2 - Uso por tema especializado

O experimento chamado por esse estudo de: uso por tema especializado, foi direcionado a prover uma análise qualitativa às indagações teóricas do presente estudo assim como a avaliar a proposta para resolução de seu problema central. Esse possui foco na análise individual da pesquisa de cada voluntário envolvido ao ter seus temas de pesquisa auxiliados pela heurística proposta.

Os pontos a serem avaliados por esse experimento estão relacionados às questões como: um conjunto bibliográfico de um estudo encontra-se completo? Alguns anos após uma pesquisa ser concluída ela ainda encontra-se completa? Quão restrita a um certo tema deve ser uma uma pesquisa? Qual profundidade de conhecimento deseja-se alcançar para um certo tema? É possível prover um auxílio para um mapeamento sistemático? São essas as perguntas chaves que formam o conjunto de indagações que esse estudo se propõe a responder.

5.2.1 Conceitos a serem avaliados

As tarefas selecionadas para serem avaliadas possuem relação com os diversos conceitos estabelecidos pelo presente estudo assim como visam avaliar o propósito fim. Cada um desses desses conceitos e propósito fim é descrito na seção 1.2 que aborda o problema a ser resolvido. Juntos totalizam dois aspectos principais, aspecto temporal e o alcance da descoberta, a serem experimentados para alcançar o propósito fim, identificação de possíveis referências negligenciadas.

Esses aspectos podem ser representados através de quatro características a serem verificadas em uma pesquisa: a largura, a profundidade, se uma pesquisa encontra-se completa ao ser finalizada, se uma pesquisa encontrar-se-á completa alguns anos após ser finalizada. Durante os experimento, todas essas características direcionaram o foco do pesquisador para a capacidade de alcançar bibliografias, até então desconhecidas, levando-se em consideração cada tarefa a ser experimentada.

O aspecto temporal descreve o estado em que se encontra uma pesquisa ao longo

do tempo, desde sua finalização até os dias atuais. Esse apresenta-se na prática através de duas características: através da validação de um conjunto de bibliografias estar completo no momento da conclusão de uma pesquisa a qual essa bibliografia pertença e através da avaliação se o mesmo conjunto ainda encontra-se completo num momento futuro.

Já o alcance da descoberta descreve o comportamento ou objetivo de uma busca. Esse é definido pela mesma seção previamente citada como possuindo duas características: largura e a profundidade. Sendo a largura representando a margem de diferentes temas envolvidos em uma pesquisa e a profundidade o quão específico são os resultados encontrados. Ressalta-se que o termo "específico", conforme já explicado, está associado ao quanto mais próximo de conhecimento básico o documento encontrado aborda.

Por fim, esses conceitos culminam para definir uma forma de entender e resolver o propósito fim, encontrar referências possivelmente negligenciadas. As tarefas foram voltadas para o entendimento dessas características e da heurística de busca descrita através capítulo 3 sobre a proposta. Portanto, ambos os aspectos e propósito principal desse estudo foram testados com e sem auxílio do algoritmo proposto.

5.2.2 Os objetivos dos experimentos

Após separar-se os temas que seriam utilizados para os experimentos, deu-se início a execução. Começando pelos experimentos relacionados as duas características do alcance da descoberta, em seguida experimentando-se o aspecto temporal e por fim um experimento voltado para avaliação do suporte provido a um mapeamento sistemático. Testando o algoritmo não somente de forma pontual, mas ao longo de todo um processo de pesquisa bibliográfica.

A fim de avaliar a **profundidade**, foram realizadas pesquisas visando alcançar resultados mais específicos possíveis que pudessem ser citados pelo tema de cada voluntário. Nessa tarefa, o objetivo do voluntário era, na melhor das hipóteses, obter um resultado com somente referências que abordassem conceitos específicos.

Conforme já citado na subseção 5.1.1, esses resultados específicos seriam documentos que abordam a fundo conceitos que poderiam ser citados pelo tema central. Um exemplo disso seria ao pesquisar por processamento de linguagem natural se deparar com artigo de estatística sobre modelos de Markov. Esse artigo poderia ser citado pelo tema pesquisado, mas pertence a outra área.

Para tarefa voltada para avaliação da característica de **largura**, a pesquisa teve o objetivo de tentar localizar documentos periféricos ao tema central. Nesse caso, buscar documentos relevantes ao tema central, porém de áreas distintas a fim de embasar o estudo. Dessa forma, foi almejado alcançar um resultado com o máximo

possível de documentos multidisciplinares correlatos.

Conforme também já mencionado na subseção 5.1.1: um exemplo disso poderia surgir ao se tentar dissertar sobre processamento de linguagem natural. Documentos sobre aplicações de processamento de linguagem natural em língua portuguesa seriam periféricos ao tema central.

Relacionados a outro aspecto principal a ser entendido está o aspecto temporal. Esse, por sua vez, envolve questões sobre um conjunto de bibliografias estar completo ou não, conforme citado no início da presente sessão. Esse conceito possui duas características de temporalidade, o presente e o futuro. Representando respectivamente: o momento de finalização de um estudo e a análise da validade desse estudos anos depois.

A característica de um estudo estar completo ou não no momento de sua finalização, foi testada como objetivo validar ou invalidar um estudo de acordo com o estado atual do tema envolvido. O objetivo estabelecido para os voluntários nessa parte do experimento foi tentar encontrar algum trabalho novo que pudesse invalidar o estudo em análise. Apesar de não se possível aferir com precisão total, buscou-se entender se o resultado foi melhor ao se utilizar o algoritmo em comparação com a forma manual.

Por outro lado, há também a validade de estudos tempos depois de serem finalizados. De acordo com o dissertado anteriormente, um estudo pode ter estado completo no passado, ter considerado o máximo possível de conhecimento ao redor do seu tema, porém com o passar dos anos outros estudos o invalidaram. O objetivo ao se experimentar esse cenário foi de avaliar, o auxílio do algoritmo proposto atuando nessa tarefa de verificação. Apesar de similar ao objetivo anterior, para esse experimento o voluntário precisou atentar para a temporalidade, uma vez que o o foco são os estudos pós data de finalização.

Todas as quatro características citadas e que foram postas à prova nos experimentos estão relacionadas à capacidade do algoritmo, proposto por esse estudo, de sugerir bibliografias possivelmente negligenciadas. Cada uma representando um objetivo diferente. Sendo assim, pode-se testar o auxílio proposto por esse estudo através de conceitos intrínsecos e considerados como características básicas ao se realizar uma busca bibliográfica.

Além das tarefas supracitadas, foram realizados experimentos para avaliação do suporte a mapeamento sistemático que o algoritmo proposto poderia oferecer. Isso significa dizer, que o algoritmo foi utilizado, através da ferramenta implementada, durante todo um processo de pesquisa bibliográfica sistemática (descrita na sessão 2.4), a fim de avaliar o suporte prestado ao se tentar obter uma visão geral de uma área pesquisada (definição para mapeamento sistemático conforme o estudo PETERSEN *et al.* (2008)).

5.2.3 Voluntários e temas

Inicialmente, fez-se necessário a busca por voluntários que estivessem aptos à utilização da ferramenta criada para verificação das proposições feitas. Não bastava ser qualquer usuário, pois desejava-se avaliar um cenário que fosse o mais próximo possível à pesquisas realizadas em ambientes científicos.

Para isso foram selecionados alunos de mestrado e doutorado. Esses voluntários encontravam-se em diversos estágios, uns começando suas pesquisas para tese, outros ainda buscando temas, mas de forma geral todos com experiência em realizar pesquisas bibliográficas. Todos esses alunos oriundos do Programa de Engenharia de Sistemas e Computação (PESC) do instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa de Engenharia (COPPE) da Universidade Federal do Rio de janeiro (UFRJ)

No total 14 pessoas aceitaram o convite para serem voluntárias nos experimentos. Desse total, 5 mestrandos e 9 doutorandos. Dentre as áreas de especialidade desses voluntários estão: grafos, engenharia de software, sistemas de informação e inteligência artificial.

Os temas escolhidos foram áreas de pesquisa de cada voluntário envolvido nos experimentos. A razão disso foi, obviamente, fazer com que cada tarefa do experimento realizada fosse avaliada por um voluntário da área. Isso foi realizado também a título de motivar os voluntários a agirem com mais afinco em suas tarefas devido ao retorno pessoal. Para isso, foi requisitado que cada voluntário informasse seus temas de pesquisa atual.

Os temas envolvidos foram:

i) uso da função de ativação bi-hiperbólica no contexto de redes neurais, especialmente *auto-encoders*;

ii) *framework* de filtragem colaborativa utilizando a linguagem Julia;

iii) critérios de parada para testes de software;

iv) *active learning* para sistemas de recomendação;

v) complexidade e algoritmos aproximativos de problemas de ordenação;

vi) mineração de dados criminais;

vii) desafios no diagnóstico precoce do Alzheimer;

viii) representação de *design* de jogos;

ix) relação entre eventos em microblogs;

x) captura e explicitação de contexto em estudos experimentais;

xi) rastreamento de objetos;

xii) *crowd computing*

xiii) classificação de séries temporais;

xiv) visualização de informação.

5.2.4 A execução dos experimentos

Uma vez explicados os objetivos de cada experimento realizado, essa sessão descreverá como esses experimentos foram realizados. Cada tema possuiu um experimento a ele associado e cada experimento foi divididos entre cinco tarefas principais. Cada uma dessas tarefas relacionada aos objetivos descritos pela sessão 5.2.2, os objetivos dos experimentos. Para cada tarefa os pesquisadores tiveram que realizar buscas de acordo com o objetivo dessa e utilizando seu tema de pesquisa informado.

Ao final, todos os temas possuíram experimentos que executaram as cinco tarefas associadas aos objetivos citados anteriormente. Uma vez que os pesquisadores já haviam realizado suas pesquisas inicias acerca de seus temas, não foi necessário reproduzir novamente o cenário no qual não houvesse o auxílio de uma heurística, apenas utilizar o conhecimento prévio dos pesquisadores. A figura 5.5 sobre a sequência dos experimentos ilustra o mencionado:

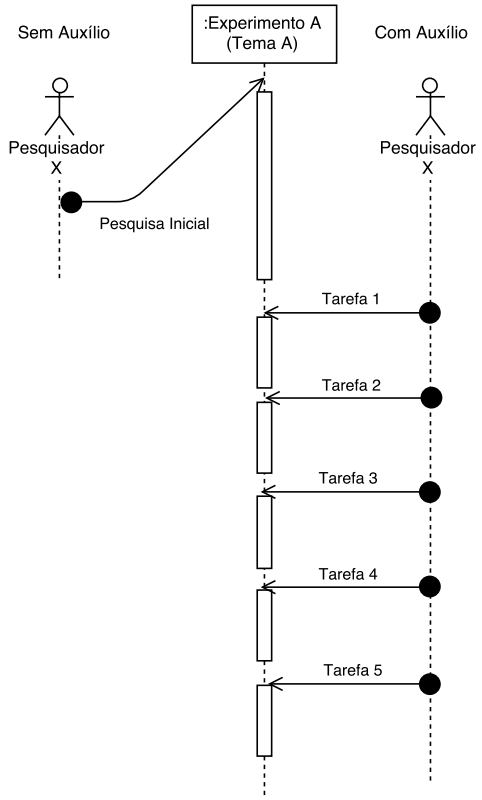

Figura 5.5: Sequência dos experimentos

Conforme descrito, foram executadas tarefas voltadas para cada objetivo por tema. Essas tarefas, por objetivo, foram executadas da seguinte maneira:

1) Avaliar a Profundidade

De acordo com os conceitos mencionados na sessão 1.2, problema sobre o qual esse trabalho se propõe a dissertar, a profundidade trata-se de uma característica do aspecto do alcance da descoberta a qual está ligada à característica dos resultados encontrados serem mais específicos ou menos específicos acerca de um determinado tema.

Para verificação de tal conceito, cada voluntário teve que conduzir buscas acerca de seu tema através do sentido das referências, uma vez que essa remete ao que já fora publicado antes do documento que as referencia. Isso conduz a dizer que a cada vez que se busca referências de outras referências, possivelmente os voluntários deveriam observar a heurística retornar documentos que descrevem princípios anteriores, consequentemente, cada vez mais básicos sobre o tema inicial.

Portanto, essa tarefa foi conduzida da seguinte forma: para cada tema selecionado essa tarefa demandou que os voluntários executassem buscas a fim de encontrar publicações que descrevessem os conceitos científicos mais primários do tema envolvido. Para isso lhes foi pedido que separassem as cinco publicações que julgassem

ser mais relevantes encontradas em suas pesquisas. Em seguida, os voluntários precisaram criar um arquivo contendo metadados dessas publicações em um formato chamado: Bibtex, a ser utilizado pela ferramenta desse estudo.

A partir desse conjunto, os voluntários seguindo seus temas, dispuseram de 25 minutos para conduzir a busca. Essa condução se deu de forma interativa com a ferramenta implementada para executar a heurística descrita por esse estudo. O número de iterações, que cada voluntário executou buscas, variou de acordo com seu julgamento conforme o surgimento dos resultados.

Para a execução dessa tarefa com a heurística, também foi passada a instrução de utilizar como parâmetro para quantidade de bibliografias por *cluster* a quantidade 2, no máximo, a fim de deixar a busca o mais profunda possível e com a largura menor possível. E como parâmetro para profundidade, no máximo a quantidade 9, além de utilizar a opção de continuar a expansão somente pelos últimos nós adicionados. Ambos os parâmetros foram definidos empiricamente devido ao crescimento do volume de dados a cada iteração do algoritmo. Além disso, também foi estabelecido como algoritmo para calcular as pontuações o HITS através da pontuação dos *authorities*.

2) Avaliar a Largura

A outra característica do aspecto do alcance da descoberta, a largura, representa o quanto multidisciplinar pode ser o resultado de uma busca. Em outras palavras quanto maior a largura, mais áreas periféricas ao tema central estarão sendo retornadas.

Para verificação dessa característica os voluntários conduziram as buscas de forma similar a anterior. Para cada tema, lhes foi pedido que tentassem buscar o máximo de publicações que contextualizassem seus temas centrais. Ou seja, a cada interação com o resultado da ferramenta, o usuário deveria observar o ganho de multidisciplinaridade que estivesse correlata ao tema central e fizesse sentido ser incluída em sua bibliografia.

Para isso, os voluntários precisaram conduzir buscas utilizando o sentido das citações, uma vez que ao recuperar os documentos que citam um inicial, deseja-se testar se os recuperados de fato remetem a uma largura maior. Isso significa que a cada vez que as citações são recuperadas, possivelmente os voluntários deveriam observar a heurística retornar documentos que descrevessem princípios mais abrangentes, porém ainda correlatos.

Além disso, os voluntários necessitaram novamente, do arquivo previamente utilizado na tarefa 1 contendo os metadados em formato Bibtex das 5 referências mais relevantes. Foi pedido aos voluntários que utilizassem os seguintes parâmetros através

da busca com heurística: no máximo profundidade 3 e para o número de bibliografias por *cluster*, no máximo, 10. Para essa tarefa também foi pedido que os voluntário utilizassem como algoritmo de pontuação o HITS através da pontuação dos *authorities*. Ambos dispuseram de 25 minutos para realizar essa tarefa.

A imagem 5.6 apresenta a sequência dos principais pontos que ocorrem durante a execução das tarefas 1 e 2. Essas se diferenciam pela variação dos parâmetros, os quais foram escolhidos de acordo com os conceitos a serem experimentados e pelo objetivo final de cada uma.

Ambas as tarefas 1 e 2 conservam a mesma cadeias de atividades que inclui: separar dentre as referências iniciais as cinco mais relevantes para serem utilizadas pela heurística, execução da heurística pelo número de vezes que o voluntário julgar necessário e análise final dos resultados.

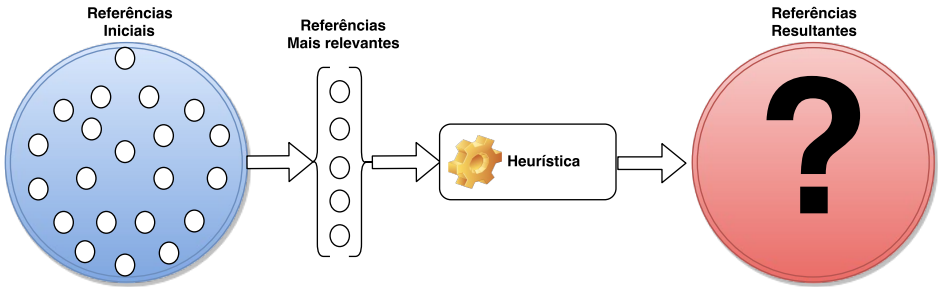

Figura 5.6: Execução das tarefas 1 e 2

As tarefas de 3 e 4 foram direcionada à experimentar o aspecto temporal de uma pesquisa descrito por esse estudo. Possuindo foco, respectivamente, no ponto finalização de uma pesquisa e no ponto futuro. Esses dois pontos são considerados por esse estudo como possíveis pontos de apoio para a utilização da heurística e por isso foram analisados. A figura 5.7 apresenta em que tempo esses dois pontos de uma possível utilização da heurística ocorre durante uma pesquisa real.

A tarefa 5 foi relacionada à experimentação do auxílio provido pela heurística no início e ao longo de um pesquisa. Essa, por sua vez, foi deixada por último devido a questões motivacionais. Apesar de, em uma pesquisa, essa ocorrer antes das duas anteriores, começou-se pelos temas atuais de pesquisa dos voluntários a fim de servir de estímulo. Além disso, o tempo de execução destinado a 3 e 4 foi menor, fazendo com que o resultado mais rápido também servisse de motivação para as tarefas seguintes.

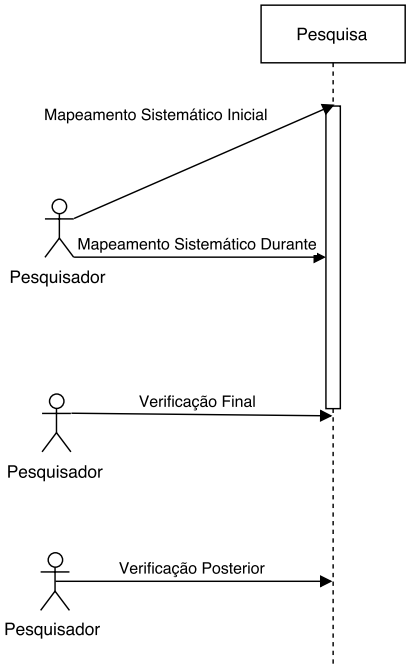

Figura 5.7: Pontos de apoio da heurística durante uma pesquisa

3) Avaliar a seguinte indagação: hoje a bibliografia encontra-se completa?

Essa questão trata de um problema correlacionado ao propósito principal do presente estudo. Esse tópico trata da validação de um conjunto de bibliografias de um estudo tido como finalizado. Isso se relaciona com o propósito fim desse estudo uma vez que se propõe a prover um auxílio à avaliação final do processo de pesquisa.

Para realização dessa tarefa, solicitou-se que os voluntários utilizassem conjuntos de bibliografias resultantes de um trabalho seu recentemente terminado a fim de criar os arquivos de metadados correspondentes para utilização através da ferramenta. Essa tarefa, diferentemente das anteriores, utilizou o conjunto completo, pois seu objetivo foi validar o estudo como um todo e não apenas experimentar aspectos pontuais.

Os voluntários após criarem seus arquivos de metadados a partir de suas referências foram instruídos a utilizar a ferramenta. Para isso, decidiram por conta própria a melhor forma de utilizá-la a fim de tentar achar referências que, caso não fossem de sua ciência, poderiam comprometer a validade de seus estudos.

As únicas instruções que foram passadas foram de limite máximo de tempo para realização dessa tarefas, que foi de 25 minutos e observações já acerca das características de largura e profundidade, como tentar manipulá-las através dos parâmetros de entrada. Esse último aspecto já havia sido explorado nas tarefas 1 e 2.

4) Avaliar a seguinte indagação: a bibliografia ainda encontra-se completa?

Para realização dessa tarefas, foi necessário a utilização de um estudo já finalizado a fim de tentar avaliar se o mesmo ainda continuava com sua característica de ser completo. Para isso, os voluntários precisaram utilizar um artigo que já fora publicado sobre o mesmo tema de suas pesquisas. Após essa seleção, os mesmos precisaram também criar uma arquivo com os metadados das referências do artigo selecionado para utilização com a ferramenta.

A execução dessa tarefa se deu de forma similar a tarefa de avaliar se a bibliografia encontrava-se completa ao término de uma pesquisa, porém o foco foi diferente, pois o voluntário precisou atentar não somente para as possíveis referências que esse estudo deveria sugerir se fosse escrito atualmente, mas também atentar-se para outros estudos que possam se concorrentes e tenham superado o mesmo.

Para a realização dessa tarefa, os voluntários dispuseram de 25 minutos e também ficaram livres para poder manipular os parâmetros de acordo com suas percepções de necessidade. Com isso ficaram com maior autonomia para conduzir a busca de forma a explorar os aspectos de largura e profundidade seguindo o que julgassem mais relevante durante a busca.

5) Avaliar a capacidade do algoritmo fornecer suporte ao mapeamento sistemático

Essa tarefa destinou-se à análise prática do auxílio provido pela heurística desse estudo às técnicas de mapeamento sistemático ou a visualização de domínio do conhecimento (KDViz) nas etapas inicias de uma pesquisa e ao longo dela. Portanto, tentando validar a capacidade do auxílio citado no último aspecto temporal restante.

Para realização dessa tarefa, uma vez que os voluntários já haviam realizado pesquisas iniciais e ao longo de seus estudos, foi pedido a cada um que escolhesse um outro tema que lhe agradasse a fim de utilizá-lo nessa tarefa. Para essa tarefas os voluntários não dispuseram de um conjunto inicial de bibliografias. Eles precisaram, primeiramente, utilizar uma ferramenta de busca para localizar 5 publicações que julgassem ser correlacionadas ao seus temas e a partir baixaram seus arquivos Bibtex.

Para essa etapa inicial os voluntários dispuseram de 10 minutos para localização de 5 bibliografias que julgassem correlacionadas. Foi indicado que não fosse realizado muito esforço para buscar as mais relevantes, uma vez que o teste dessa tarefa compreende exatamente em auxiliar nesse tipo de busca.

Feito isso, os voluntários dispuseram de 30 minutos para realizar suas pesquisas, utilizando como auxílio a ferramenta implementada. Foi pedido para que os voluntários tentassem alcançar um conjunto bibliográfico final que melhor refletisse uma revisão bibliográfica acerca de seus temas. Esse por sua vez deveria conter bibliografias que contextualizassem o estudo, apresentassem trabalhos correlatos, o estado da arte e qualquer ponto que o pesquisador julgasse relevante ser referenciado em um possível texto acerca do tema.

5.2.5 Avaliação dos voluntários

Os experimentos foram realizados de forma que os resultados dos 14 voluntários fossem avaliados por eles mesmos uma vez que já haviam atuado na área e possuíam conhecimento de cada tema envolvido. Dito isso, essa seção apresenta as avaliações dos resultados referentes a cada tarefa da seção 5.2.4 sobre a execução do experimento.

Para cada tarefa, os voluntários precisaram avaliar o suporte dado pela heurística para entendimento do conceito mencionado. Para responder às avaliações os voluntários utilizam uma escala Likert, voltada para medir o nível de concordância de uma pessoa a respeito de uma afirmação ou pergunta, conforme já explicado na subseção 5.1.5. Além disso, os mesmos foram convidados a responder a perguntas pessoais acerca da ferramenta que implementou a heurística utilizada.

Seguem os resultados das avaliações relacionadas as peguntas de aspecto conceitual, assim como uma breve análise dos resultados e do transcorrido ao longo dos experimentos:

1) Como você avalia o suporte dado a verificação da **característica de profundidade** ao tentar encontrar artigos que detalham a fundo conhecimentos específicos relacionados ao seu tema principal? Exemplo, se seu tema for processamento de linguagem natural, qual seria o suporte para encontrar artigos que falem somente sobre um método específico de inteligência artificial, mas que seria relevante de ser citado em seu texto.

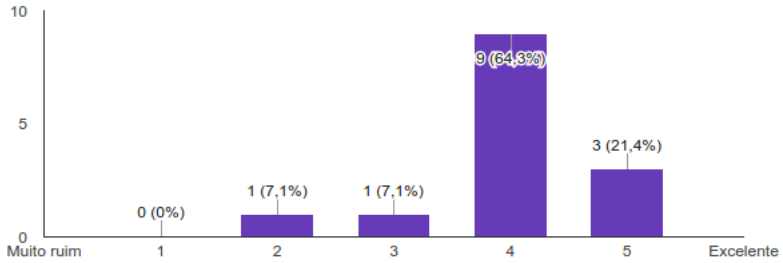

Figura 5.8: Escala de 1-5, onde 1 significa: muito ruim e 5 significa: excelente

Foi observado que de fato, através dos parâmetros estabelecidos. houve uma tendência a encontrar documentos desconhecidos que detalham conhecimentos específicos utilizados pelo tema do voluntário. Percebeu-se que o tema principal escolhido é um fator que influencia na quantidade de resultados encontrados. Quanto mais específico foi o tema do voluntário mais difícil foi de encontrar documentos que detalhassem ainda mais suas ideias. Um exemplo disso ocorreu com o tema: complexidade e algoritmos aproximativos de problemas de ordenação.

Com relação a estudo mais abrangentes, a tendência de encontrar documentos mais específicos foi mantida. Foram encontrados documentos de diferentes temas, mas que de alguma forma poderiam ser aplicados ao tema do voluntário. Porém, foi encontrado um caso em que o tema era bastante abrangente e apesar de ter seguido a tendência mencionada, também foram encontrados documentos periféricos seguindo a definição de largura. Segundo a voluntária e especialista no tema em questão, captura e explicitação de contexto em estudos experimentais, esse é um tema com trabalhos que possuem poucas referências especificando ainda mais esse tema.

Aparentemente a quantidade de artigos com a característica de profundidade influenciou nas notas mais baixas. O conjunto inicial de referências não pode ser expandido ao ponto de gerar uma quantidade satisfatória de estudos específicos. Esse fato aparentemente aconteceu devido a fatores como: base com dados incompletos e necessidade de mais expansões a fim de continuar observando a tendência a especificar os temas encontrados.

Ocorreram múltiplas situações em que os voluntários encontram artigos relevantes para o tema utilizado e consideraram ler por completo após o experimento.

2) Como você avalia o suporte dado a verificação da **característica de largura** ao tentar encontrar artigos relevantes que sejam relacionados ao seu tema? Em outras palavras como você avalia o suporte dado à capacidade de encontrar artigos multidisciplinares correlatos ao seu tema.

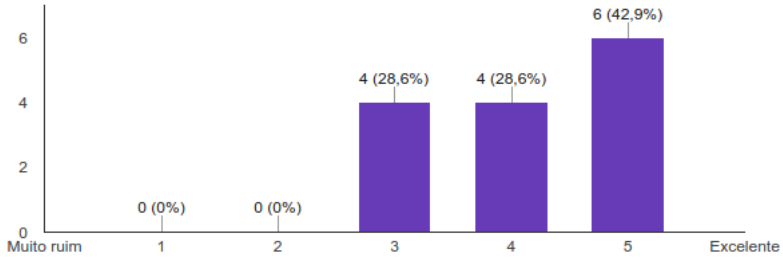

Figura 5.9: Escala de 1-5, onde 1 significa: muito ruim e 5 significa: excelente

Os resultados dessa tarefa indicaram, assim como na primeira tarefa, que o uso da heurística tendeu a alcançar o objetivo de encontrar referências periféricas. Por todos os voluntários foram encontrados artigos que se desconhecia e eram relevantes a serem utilizados pelo seu tema de pesquisa do experimento. Novamente a quantidade foi o fator influenciados nas notas segundo os voluntários.

3) Qual suporte dado para avaliar se hoje uma bibliografia recém finalizada encontra-se completa?

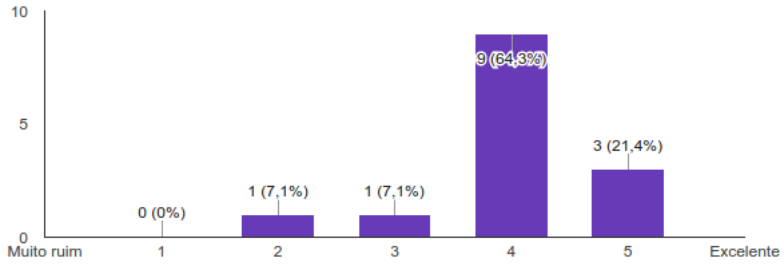

Figura 5.10: Escala de 1-5, onde 1 significa: muito ruim e 5 significa: excelente

Para essa tarefa a maioria dos voluntários tiveram a percepção de obter um suporte bom para verificação da característica experimentada. Porém, novamente fatores atrelados à base de dados utilizada, como a incompleta indexação de artigos, foi o motivo responsável por uma nota ruim e outra neutra.

Para os demais, características como poder carregar em uma ferramenta visual que pudesse expandir as ligações ao redor das referências de um documento a ser analisado contribuiriam muito para verificação desse aspecto.

4) Qual suporte dado para avaliar se uma bibliografia de um estudo já aceito ainda encontra-se completa?

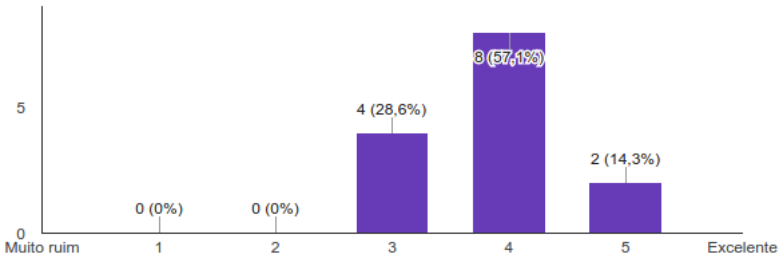

Figura 5.11: Escala de 1-5, onde 1 significa: muito ruim e 5 significa: excelente

Essa tarefa revelou mais um aspecto influenciado pelo tema do voluntário. Para voluntários de certas áreas do conhecimento, cujos artigos gerados não são em sua maioria superados, como no caso os que utilizam bases sólidas matemáticas, essa característica não poderia ser muito bem suportada. Isso ocorre pois o conhecimento de áreas assim é construído em cima do que já foi feito. De forma geral os resultados não são superados, mais sim melhorados.

Por esses motivos o voluntário da área de grafos citou que seu caso seria basicamente neutro. Já para os demais que apresentam posição neutra com relação a esse possível suporte, a falta de segurança da base de dados envolvida e o aspecto amplo do que poderia não tornar uma pesquisa completa foram os motivos que justificaram suas notas.

5) Qual suporte dado à realização de um mapeamento sistemático para entendimento de um assunto?

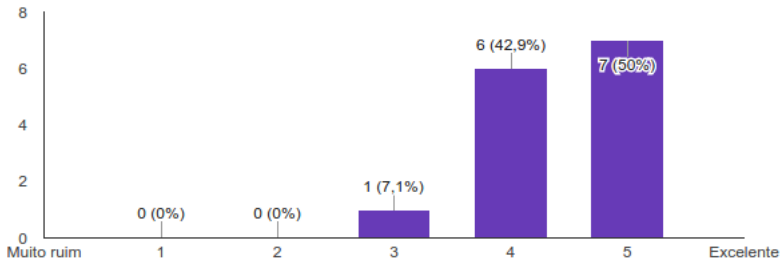

Figura 5.12: Escala de 1-5, onde 1 significa: muito ruim e 5 significa: excelente

Essa tarefa revelou que para praticamente todos os voluntários a heurística contribuiu para o entendimento de um assunto e descoberta de novos artigos relevantes. Percebeu-se que ao longo da execução dessa tarefa os voluntários selecionaram com entusiasmo artigos para serem lidos por completo posteriormente.

Um dos voluntários que se posicionou de forma neutra, justificou seu voto devido ao termo sistemático da pergunta. Seu entendimento foi que a ferramenta auxiliaria sim no mapeamento, mas a parte sistemática não seria feita por ela.

Outro aspectos apresentados por essa pesquisa e também relevantes são os resultados das experiências pessoais com a implementação da heurística apresentada. Esse estudo se prontificou a avaliar esse aspecto a fim de contribuir com uma ferramenta útil à comunidade científica e poder apresentar melhorias futuras condizentes com as necessidades reportadas pelos voluntários. Seguem suas avaliações:

6) Qual foi a qualidade dos artigos sugeridos ao longo das tarefas?

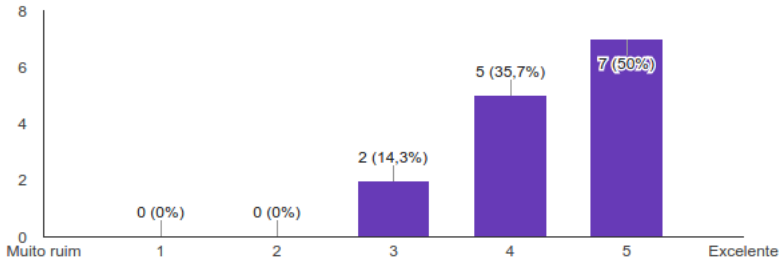

Figura 5.13: Escala de 1-5, onde 1 significa: muito ruim e 5 significa: excelente

Ao longo do experimento, os voluntários citaram que certos documentos retornados correspondiam de fato a autores consagrados. Outros disseram que encontraram artigos relevantes acerca do seus tema utilizado e para 2 voluntários os artigos encontrados no geral foram inconclusivos, precisariam ler mais a fundo para julgarem. Esses últimos se posicionaram de forma neutra.

7) Você foi capaz de encontrar referências relevantes que poderia ter incluído em seu trabalho utilizado na questão.

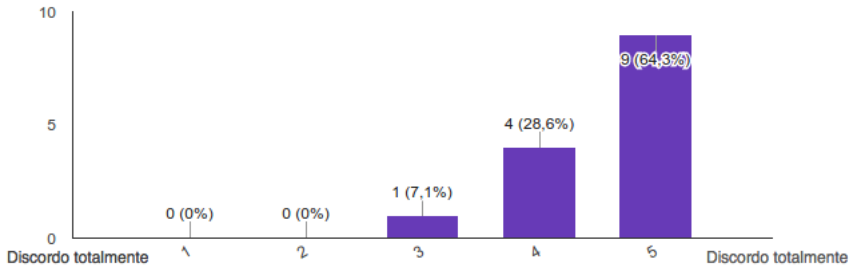

Figura 5.14: Escala de 1-5, onde 1 significa: discordo totalmente e 5 significa: concordo totalmente

Conforme já mencionado, os voluntários envolvidos por diversas vezes salvaram nomes, links e arquivos de artigos encontrados ao longo de suas buscas. Segundo os voluntários, aparentemente, ao observar seus resumos e nome de autores, seria importante guardá-los para posterior leitura, pois pareciam ser relevantes.

8) Na sua opinião, como se caracteriza o uso da ferramenta? (Avaliar a facilidade)

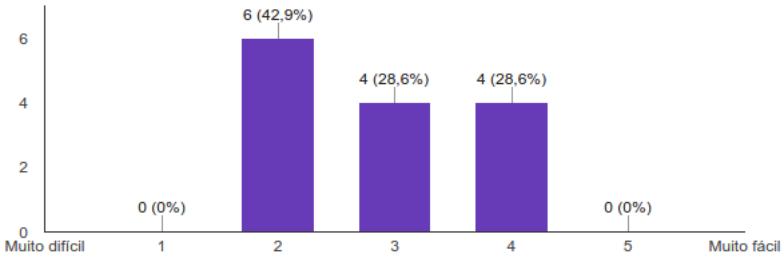

Figura 5.15: Escala de 1-5, onde 1 significa: muito difícil e 5 significa: muito fácil

As respostadas dadas a essa pergunta traduzem a relativa dificuldade encontrada pelos voluntários ao utilizarem a ferramenta. Relativa, pois os voluntários mencionaram que sua dificuldade estava mais diretamente relacionada aos termos descritos pela ferramenta. Após explicações verbais, os mesmos começaram a compreender seu uso, porém disseram que sem auxílio não seriam capazes de realizar os experimentos. Porém para avaliar disseram que levaram esse fato em consideração.

Diversas dicas de melhorias práticas foram mencionadas pelos voluntários. Foram elas: melhoria da disposição do botão de execução, exibição do ano dos artigos listados, mudança nos rótulos dos vértices do grafo, filtro para ano, remoção de vértices nulos devido a falta de indexação da base utilizada, deixar visualmente claro qual vértice está selecionado e mudança na representação da legenda de cores de cada nível do grafo.

9) No geral, como você classifica sua experiência com a ferramenta utilizada?

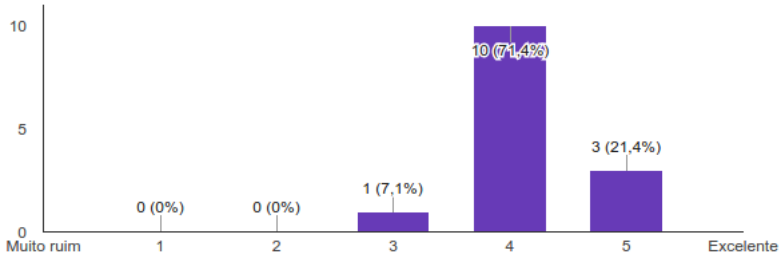

Figura 5.16: Escala de 1-5, onde 1 significa: muito ruim e 5 significa: excelente

Os voluntários apresentaram bastante entusiasmo com o uso da ferramenta. Diversos deles perguntaram se a mesma ficaria disponível para futuro uso de forma livre. As notas, em sua maioria, apresentara-se como boas ou excelentes. Apenas um voluntário se posicionou de forma neutra. Essa nota neutra foi justificada pelos argumento de que a ferramenta provê um suporte, mas a parte sistemática não é feita por ela mesma.

5.3 Base de Dados

Para realização de experimentos foi utilizada a base de dados chamada Scopus. Trata-se da maior base internacional de informações sobre literatura técnica e científica publicadas desde 1823. Seu conjunto total contempla mais de 54 milhões de registros todos possuindo resumos acerca de cada trabalho indexado. Desses, 33 milhões incluem dados também sobre referências.

Ressalta-se o fato dos pesquisadores, durante os experimentos, terem mencionado que utilizavam diversas bases de dados para consultas de artigos, porém abandonaram o uso das demais passando a utilizar somente a Scopus. Esses pesquisadores justificaram seu uso devido ao fato de na maioria das vezes o conteúdo existente nas demais sempre estar presente na Scopus e o inverso não ocorrer.

Para acesso aos dados, o portal da Scopus disponibiliza uma API (*Application Programming Interface*) via *web service*. As informações acerca dessa API podem ser encontradas através da página na *internet* no endereço: `http://dev.elsevier.com/api_docs.html`.

A API (*Application Programming Interface*) Scopus disponibiliza diversos tipos de consultas que retornam metadados acerca das informações buscadas. Entre essas consultas estão: busca por informações de autores, busca por informações de publicações e busca por instituições. Cada uma dessas buscas podem retornar resultados que variam de acordo com o tipo de visão utilizada pelo usuário desse serviço.

Os tipos de visão disponíveis aos usuários estão relacionadas a que tipo de privilégio os usuários possuem.

Basicamente, há dois tipos de privilégios para acesso às visões disponibilizadas. Eles se dividem entre: os que podem ser acessados publicamente e os que necessitam de cadastro prévio para obtenção de uma chave de acesso a ser utilizada a cada consulta à base. Para obter o segundo tipo de acesso é necessário estar conectado via alguma instituição de ensino conveniada.

Entre os tipos de consulta existentes há diversas visões que variam com o tipo de busca a ser feita. Para a busca por informações de artigos (chamada de *Abstract Retrieval*), por exemplo, há cinco tipos de visões dos metadados retornados. Sendo elas: *basic*, *meta*, *meta_abs*, *ref* e *full*. Cada uma possui suas peculiaridades pode ser consultada através do endereço: `http://api.elsevier.com/documentation/retrieval/AbstractRetrievalViews.htm`. As visões encontram-se listadas em ordem crescente de quantidade de informação disponibiliza. As duas últimas visões estão restritas a usuários que dispõem de chave de acesso.

Acerca das visões de acesso aos resultados de busca por autor (*Author Retrieval*), há também cinco tipos de visões. São elas: *basic*, *metrics*, *light*, *standard* e *enhanced*. Sendo as duas últimas visões restritas a usuários que dispõem de chada de acesso. Seus respectivos tipos de dados retornados encontram-se apresentados através do endereço: `http://api.elsevier.com/documentation/retrieval/AuthorRetrievalViews.htm`.

Por fim, há três tipos de visões dos metadados retornados através de buscas por instituições *Affiliation Retrieval*. São elas: *basic*, *light* e *standard*. Sendo todas visões de acesso público, não havendo necessidade de registro para utilização dos servições para esse tipo de busca. Os conteúdos retornados pelos diferentes tipos de visão encontram-sem apresentados no endereço: `http://api.elsevier.com/documentation/retrieval/AffiliationRetrievalViews.htm`.

Maiores informações acerca dos três tipos de busca descritos acima e seus metadados disponibilizados de acordo com o tipo de visão que cada um possui (*Abstract Retrieval Views*, *Author Retrieval Views* e *Affiliation Retrieval Views*) podem se encontrados através dos respectivos *links*: `http://api.elsevier.com/documentation/retrieval/AbstractRetrievalViews.htm`, `http://api.elsevier.com/documentation/retrieval/AuthorRetrievalViews.htm` e `http://api.elsevier.com/documentation/retrieval/AffiliationRetrievalViews.htm`.

É importante citar que a fim de minimizar erros de indexação, além da base Scopus também foi utilizada a base DBPL (*Digital Bibliography & Library Project*) para consulta de artigos cujo DOI(*Digital object identifier*) não houvesse sido informado no arquivo BIBTEX utilizado pela ferramenta, tampouco houvesse sido localizada sua indexação via título na base Scopus. Para isso artigos não encontrados direta-

mente na base Scopus foram pesquisados na DBLP para recuperação de seu DOI e posterior nova consulta na base Scopus.

Capítulo 6

Conclusão

6.1 Epílogo

A realização de pesquisas bibliográficas e a busca por métodos mais eficazes de realizá-las é certamente um tema de interesse geral no cenário científico. Isso se dá pelo fato de ser comum a todos a necessidade de se ter domínio sobre o estado de um assunto para poder dissertar sobre ele e propor novos trabalhos.

A ciência do presente evolui graças à ciência do passado. O presente é construído em cima da tentativa de evoluir ou superar um resultado previamente gerado. Ou até mesmo provar que um resultado passado estava errado. Porém não é característica da ciência construir trabalhos exatamente iguais a um trabalho já realizado. Esse comportamento em nada ajuda em sua evolução.

Porém não é sempre que a existência de fatos assim são propositais. Isso pode ocorrer devido a fatores que de alguma forma não permitiram que o pesquisador encontrasse referências sobre tais trabalhos iguais ao seu e acabasse por reinventar uma solução já apresentada.

A fim de minimizar os efeitos desses fatores, o presente estudo dissertou sobre: o problema apresentado, sua proposta para resolução, métodos existentes para realização de uma pesquisa seguindo um rigor metodológico, sobre os conceitos que estão envolvidos em sua proposta, a arquitetura utilizada para construção de uma ferramenta que implementasse sua proposta e por fim descreveu experimentos realizados a fim de obter resultados com o uso de sua proposta.

Esse capítulo faz uma revisão das questões principais a serem respondidas por esse estudo e disserta acerca de suas respostas. Em seguida são apresentadas informações advindas do processo de experimentação no qual contou com diversos voluntários das mais variadas áreas da computação: inteligência artificial, sistemas de informação, teoria dos grafos, engenharia de *software* e banco de dados.

Disto isso, também são enumerados os problemas encontrados ao longo desse es-

tudo. São descritos tanto os problemas de aspecto prático, através da implementação dos conceitos propostos, como também problemas relacionados a parte humana.

Por fim, são descritos os trabalhos futuros. Esses surgiram da necessidade percebida ao longo tanto da parte de desenvolvimento teórico, quanto da parte de implementação e experimentação. As ideias apresentadas foram resultantes de opiniões de voluntários, amigos e orientadores acerca desse estudo, porém que não caberiam no escopo desse trabalho.

6.2 Recapitulando os objetivos

O objetivo principal desse estudo foi prover uma forma de suporte ao mapeamento sistemático ou a uma forma de visualização do conhecimento. Seu intuito foi auxiliar a encontrar possíveis referências negligenciadas, esquecidas ou que não foram encontradas por um pesquisador em seu estudo. Em paralelo, também colaborar para minimizar o esforço durante pesquisas bibliográficas, tentando prover um método para auxiliar nesse processo.

Esse objetivo, por consequência, poderia tornar pesquisas bibliográficas menos onerosas do ponto de vista de tempo, ajudando o pesquisador a dedicar mais tempo ao seu objeto de estudo e a ter uma visão mais clara do estado atual de publicações relacionadas. Esse objetivo deu origem a 4 questionamentos iniciais que se pudessem ser respondidos de forma mais fácil auxiliariam o pesquisador em suas pesquisas bibliográficas. Esses são relacionados respectivamente aos aspectos temporais e de alcance da descoberta.

Aspectos temporais:

1) Atualmente a bibliografia proposta está completa?

Através do experimento 2, seção 5.2, observou-se que houve uma tendência dos voluntários a conseguirem responder mais claramente a essa questão através do ferramental teórico e prático desse estudo. Os voluntários, revelaram através de suas respostas, que teriam mais suporte para verificação dessa característica. Com isso seriam capazes de reduzir possíveis esquecimentos ou negligências bibliográficas ao término de seus estudos.

Apesar de não ter sido unanime, esse resultado aponta para a contribuição da redução da possibilidade de um pesquisador estar realizando o estudo previamente finalizado.

2) Ao se passarem inúmeros anos, essa bibliografia ainda estará "completa"?

Conforme já dissertado na seção 5.2 sobre o experimento 2, voltado a uma análise qualitativa, o resultado desse questionamento utilizando o suporte oferecido tendeu

a ser bom, segundo as respostas. Os voluntários se sentiram menos confortáveis para afirmar o suporte provido pela ferramenta se comparado a pergunta anterior, porém ainda assim apresentaram que seriam beneficiados ao utilizar o proposto por esse estudo.

Recapitula-se que houve um fato inesperado. O voluntário, cujo o tema era da área de teoria de grafos, visualizou esse questionamento como algo difícil de ser respondido através de seu tema. A explicação para isso se dá porque há áreas em que publicações são construídas com base em conhecimentos anteriores, porém sempre a nível de incrementar o já provado ou dissertado.

Aspectos do alcance da descoberta:

3) Quão restrita a um tema deve ser a pesquisa?

Para entender se a ferramenta implementada por esse estudo poderia prover um suporte para auxiliar a controlar a característica do quão restrita a um tema deve ser uma pesquisa, foram realizados experimentos a fim de testar o suporte que os voluntários teriam ao tentar encontrar artigos periféricos.

Os resultados desse estudo tenderam a apresentar uma boa avaliação para o suporte dado ao controle dessa característica. Os resultados dos voluntários do experimento 1, que utilizaram a ferramenta implementada por esse estudo, apresentaram artigos periféricos com quantidades mais altas de citações que os resultados dos voluntários desse mesmo experimento que só utilizaram o motor de busca da base de dados acordada.

Além dos resultados citados, os resultados dos voluntários do experimento 2 apresentaram, de forma qualitativa, que os voluntários foram capazes de perceber artigos que de fato poderiam ser referenciados em seu tema de pesquisa. Pode ser percebido que os voluntários, enquanto realizavam os experimentos acerca dessa característica, separavam os documentos encontrados para futura leitura integral de alguns artigos.

4) Quão profundo deve-se pesquisar em um certo tema?

A compreensão desse aspecto por parte do grupo de voluntários do experimento 1 não pareceu ser tão clara. Apesar do grupo que utilizou a heurística ter conseguido alcançar um resultado um pouco melhor que o grupo que não utilizou, o resultado quantitativo não foi tão relevante.

Porém, quando o mesmo aspecto foi analisado de forma qualitativa através do experimento 2, os voluntários conseguiram localizar artigos que os levaram a notar essa característica sendo acentuada.

Além desses 4 questionamentos iniciais, as avaliações quantitativas do experimento 1 também apresentaram que os resultados do grupo que utilizou a ferramenta superou o grupo sem ferramenta em:

i) bons resultados com relação aos artigos mais relevantes do tema utilizado. Os artigos encontrados com auxílio da ferramenta recebem quantidades de citações muito mais elevadas do que os encontrados sem auxílio;

ii) bons resultados com relação aos autores encontrados com o uso da heurística. Os autores encontrados possuem índice H mais elevados;

iii) maior quantidade de palavras-chave convergindo intragrupo.

Portanto, ficou claro o suporte recebido pelos voluntários, através da ferramenta, para buscas por referenciais teóricos, em largura, e para buscas de assuntos específicos, em profundidade. A percepção do quanto cada um notou esse suporte, variou de acordo com o tema utilizado. O mesmo ocorreu com o suporte para verificação do aspecto temporal. Houve uma tendência a respostas positivas sobre o provimento de suporte para responder a esse questionamento, porém o tema também foi um fator influenciador.

O resultado geral demonstra auxílio para controlar esses aspectos durante pesquisas bibliográficas. Todos os voluntários disseram realizar seus processos de pesquisa de forma manual, alguns sendo sistemáticos, outros não. Todavia nenhum voluntário mencionou utilizar alguma ferramenta com heurística para auxiliá-lo em suas buscas e o motivo foi apenas por desconhecimento.

6.3 Demais conclusões

Esse estudo fez uma análise vasta de diversas características percebidas ao longo de uma pesquisa. Essas características apresentaram-se relevantes quando o foco é encontrar referências negligenciadas, esquecidas ou apenas não encontradas pelo pesquisador.

Diversas conclusões paralelas aos aspectos abordados inicialmente por esse estudo puderam ser percebidas durante os experimentos. A primeira delas foi a clara influência do tema utilizado em cada experimento. Esse pode influenciar não somente no afinco do voluntário em realizar as pesquisas como também na dificuldade em termos de escassez para o aspecto do alcance da descoberta.

Foi percebido que temas cujos artigos referenciados geralmente abordam assuntos restritos de suas áreas, não são facilmente expandidos através de buscas chamadas por esse estudo de: busca em profundidade. Esse fato pode estar atrelado a razão das referências envolvidas já dissertarem sobre conhecimentos muito básicos de um certo tema, tornando difícil a tarefa de encontrar algo que especifique mais cada conceito envolvido ou por essas referencias tratarem temas muito genéricos.

Outra fato relevante foi a compreensão da existência de assuntos que são estudados por áreas distintas, mas que geralmente não se referenciam. Isso mostrou-se

como um problema, caso múltiplas áreas não estejam representadas através das referências inicias utilizadas pela heurística. Caso essas múltiplas áreas não sejam representadas, é possível que um mesmo assunto não referenciado inicialmente já tenha sido explorado por outra área de forma semelhante e será muito difícil para o pesquisador descobrir.

O termo sistemático, utilizado por esse estudo, levantou debate com um voluntário. Esse demonstrou resistência inicial com relação ao suporte provido a um estudo sistemático devido ao termo, em seu entendimento, estar associado a um objetivo específico.

Porém, ao longo de explicações ficou claro que o intuito desse estudo não é substituir um método, ou obrigar a heurística a ter base sistemática, mas sim apenas prover uma forma de um pesquisador que trabalha de forma sistemática a ser auxiliado. Portanto, focar em um objetivo, seja ele mais específico ou mais abrangente, seria trabalho do pesquisador. A ferramenta possibilita ambas escolhas.

Um resultado significativo relacionado a implementação do proposto por esse estudo foi a dificuldade em compreender a ferramenta. A primeira vista os voluntários disseram que não acharam intuitiva. Os termos utilizados foram muito específicos, e algumas partes do *layout* não foram intuitivas. Esses fatos geraram uma listagem de sugestões a serem apresentadas na seção sobre trabalhos futuros.

Os resultados gerais podem ser percebidos como satisfatórios, uma vez que tanto de forma quantitativa quanto de foma qualitativa o uso da heurística obteve melhores resultados que seu não uso. Os usuários perceberam suporte a tarefa de aprender sobre áreas desconhecidas ou melhorar seu conhecimento sobre algo já estudado.

6.4 Problemas encontrados

Durante a implementação da arquitetura proposta, foram encontrados alguns problemas não impeditivos, mas que dificultaram os experimentos e suas resoluções serão parte de trabalhos futuros. São eles:

i) acesso remoto a uma base de dados. Esse acaba por demandar um tempo computacional muito maior que utilizar uma base estática local. Porém sua vantagem é possuir dados atualizados;

ii) existência de problemas de indexação na base de dados utilizada, reduzindo assim a quantidade dos estudos retornados pela ferramenta;

iii) representação de uma quantidade grande de informação tornou o ambiente de visualização sobrecarregado devido a tecnologia utilizada, tornando assim a experiência do voluntário mais lenta.

Com relação aos problemas de caráter humano, pode-se citar a dificuldade em motivar certos grupos de voluntários a realizarem os experimentos com afinco, conforme o ocorrido no experimento 1. Esse foi realizado durante uma disciplina de curso de Pós-Graduação do PESC, no qual os alunos foram obrigados a participar com um tema que não necessariamente era de seu interesse. Esse fato pode possibilitar não levar ao máximo o potencial do experimento.

6.5 Trabalhos futuros

Muitas ideias surgiram ao longo do desenvolvimento do presente trabalho. Por se tratar de um tema com muito a se contribuir, conforme mencionado em MARSHALL e BRERETON (2013b), muitas dessas ideias não puderam ser implementadas devido ao tempo limite para conclusão dessa dissertação. Entretanto, essas ideias não implementadas não são menos importantes que as implementadas. Portanto, esse capítulo torna-se fundamental para o progresso do presente estudo.

A fim de aprimorar não somente a parte teórica, mas também a parte da ferramenta desenvolvida por esse trabalho, serão abordados alguns pontos que podem gerar maiores contribuições ao expandir-se o escopo atual.

Diversas foram as sugestões com relação ao *layout* da ferramenta que podem ser vistas como possíveis trabalhos futuros:

i) editar explicações das funcionalidades a fim de adequá-las ao entendimento de pesquisadores de áreas não só da computação;

ii) encontrar um posicionamento mais adequado para o botão de executar a heurística;

iii) alterar os rótulos de cada vértice no grafo. Um das sugestões foi apresentá-los como: número de citações/número de referências - ano;

iv) deixar claro qual vértice está atualmente selecionado;

v) modificar legenda dos níveis dos vértices no grafo;

vi) adicionar filtro de ano.

Conforme previamente mencionado na seção 3.6.4, o algoritmo utilizado para sugestão de referências trata-se de um algoritmo *naive*, fato que deixou em aberto a possibilidade de otimização. Para esse fim, o presente estudo identifica como relevante os seguintes pontos:

1. Utilização do fator de impacto do local em que a bibliografia envolvida foi publicada.

   Além da utilização da métrica extraída pelos cálculos de relevância, sugere-se que medidas como o fator de impacto sejam utilizadas. A sugestão seria utilizá-las em conjunto com o grau de relevância extraído dos relacionamentos do grafo. Nesse ponto, poderia-se atribuir uma nota final ponderada por tal fator. Um exemplo de medida que se sugere utilizar é a nota qualis que pode ser obtida através da plataforma Sucupira através da *internet* no endereço: https://qualis.capes.gov.br/.

2. Utilização do grau de relevância dos autores

   Além da métrica do fator de impacto dos locais, conforme supracitado, abre-se a oportunidade para utilização também dos fatores de impacto relacionados ao autor das publicações. Entre eles o: *h-index* (HIRSCH, 2005). Esse por sua vez poderia ser extraído de bases como *Scopus* (GOODMAN, 2005), utilizada para recuperação de dados desse trabalho e que disponibiliza forma de acesso padronizada a esse tipo de dado.

3. Utilizar o conteúdo dos resumos para ajudar a selecionar os mais relevantes.

   Uma sugestão um pouco mais rebuscada e computacionalmente mais custosa, seria a utilização dos conteúdos dos resumos de cada bibliografia envolvida. Há estudo realizado pelo Programa de Engenharia e Sistemas da COPPE, porém ainda não publicado, que demonstra uma taxa relevante de acerto ao se tentar filtrar entre uma série de resumos os artigos os que seriam mais relevantes a serem lidos por completo.

   Nesse ponto, a ideia seria utilizar esses resumos a fim de gerar mais uma métrica de avaliação dos artigos. Com isso, além da análise estrutural, ou seja dos relacionamentos existentes entre as bibliografias, o presente estudo seria capaz de analisar de forma automática o resumo do artigo e aprimorar a métrica existente para relevância.

Vale enfatizar que as otimizações citadas nos itens anteriores poderiam ser implementadas em dois pontos do método proposto por esse estudo. São eles: atividade de cálculo de relevância, no qual a otimização estaria atuando antes da expansão dos dados, provendo assim uma poda maior dos dados, ou na atividade de sugestão, provendo assim novas métricas para serem utilizadas em conjunto.

A fim de prover um melhor controle das áreas envolvidas ao longo da adição de nova bibliografias ao conjunto a ser analisado sugere-se, para melhor identificação visual, a utilização de *tag clouds* para representação das áreas envolvidas, conforme

já utilizada pelo estudo DE ALMEIDA (2012) para identificação de assuntos através de tópicos.

Esse método de identificação visual também seria de grande valia para o presente estudo. Uma vez que o pesquisador que utiliza o método aqui proposto poderia acompanhar o surgimento das áreas a cada iteração do algoritmo. Atualmente o algoritmo não dispõe de representação visual das comunidades que surgem, fundem-se ou desaparecem a cada computação.

Uma vez que a tarefa de controlar as áreas que são relevantes a um propósito de pesquisa está intimamente relacionada ao julgamento do pesquisador, é relevante a possível intervenção do mesmo de forma a orientar a expansão do grafo. A ideia mencionada ajudaria na tarefa de seleção das áreas a serem expandidas seria facilitada.

Esse estudo também demonstrou que a expansão dos dados em tempo de execução demanda um tempo excessivo. Devido a característica do acesso, via *web service*, à base utilizada, o tempo de recuperação das informações crescem proporcionalmente ao número de vértices expandidos no grafo representativo. A cada vez que surge um vértice novo, existe a necessidade de se fazer uma ou mais requisições para buscar as informações associadas a esse vértice.

Uma forma de reduzir esse tempo e maximizar a chance de encontrar indexação desses vértices, que também foi um problema relacionado à base, seria criar-se uma base local para acesso imediato. Primeiramente a ferramenta buscaria localmente a informação requerida, caso não a encontra-se faria o acesso remoto em busca de tais informações.

Uma outra ideia que não foi implementada por esse estudo, seria o teste de um novo componente de clusterização. Atualmente o componente de clusterização não provê parâmetros que tentem forçar o número de *clusters* a serem formados, ele decide por si só baseado em sua função de ganho. Porém a utilização de um componente que permitisse a especificação do número de *clusters*, talvez poderia ser capaz de controlar melhor o ganho em termos do conceito de largura apresentado por esse estudo.

Avaliar a variação do parâmetro para utilizar "somente últimos nós" pode também ser apontada como uma tarefa futura, uma vez que esse parâmetro foi mantido fixo nas avaliações realizadas pelo experimento 2. Talvez possa ser relevante não utilizar somente últimos nós numa busca com intuito de maximizar a quantidade dos resultados em largura.

O presente estudo também utilizou como parâmetro fixo o algoritmo de pontuação HITS. Porém outros algoritmos de pontuação também poderiam ser testados a fim de comparar o desempenho.

Por fim, um futuro experimento poderia ser realizado para comparar o uso da

ferramenta somente por especialistas em métodos sistemáticos com e sem o uso da ferramenta. Nesse experimento a ideia seria verificar a qualidade e quantidade dos estudos obtidos por cada especialista em um tempo estipulado.

Referências Bibliográficas

ALONSO, L., CASTELLÓN, I., FUENTES, M., et al., 2004, "Approaches to text summarization: Questions and answers", *Inteligencia Artificial*, v. 8, n. 22, pp. 79–102. Disponível em: <http://www.scopus.com/inward/record.url?eid=2-s2.0-2942675019&partnerID=40&md5=f3d1a2676e64ecd56b8ec0005043acea>. cited By 5.

ALVES, A. J., 2013, "A"revisão da bibliografia"em teses e dissertações: meus tipos inesquecíveis." *Cadernos de Pesquisa*, , n. 81, pp. 53–60.

AYNAUD, T., 2009, "Community", *URL http://perso.crans.org/aynaud/communities/*.

AYNAUD, T., GUILLAUME, J.-L., 2010, "Static community detection algorithms for evolving networks". In: *Modeling and optimization in mobile, ad hoc and wireless networks (WiOpt), 2010 proceedings of the 8th international symposium on*, pp. 513–519. IEEE.

BAILEY, J., BUDGEN, D., TURNER, M., et al., 2007, "Evidence relating to Object-Oriented software design: A survey." In: *ESEM*, v. 7, pp. 482–484. Citeseer.

BERGSTRÖM, P., ATKINSON, D. C., 2009, "Augmenting the exploration of digital libraries with web-based visualizations". In: *Digital Information Management, 2009. ICDIM 2009. Fourth International Conference on*, pp. 1–7. IEEE.

BLONDEL, V. D., GUILLAUME, J.-L., LAMBIOTTE, R., et al., 2008, "Fast unfolding of communities in large networks", *Journal of Statistical Mechanics: Theory and Experiment*, v. 2008, n. 10, pp. P10008.

BÖRNER, K., CHEN, C., BOYACK, K. W., 2003, "Visualizing knowledge domains", *Annual review of information science and technology*, v. 37, n. 1, pp. 179–255. ISSN: 1550-8382. doi: 10.1002/aris.1440370106. Disponível em: <http://dx.doi.org/10.1002/aris.1440370106>.

BOTELHO, L. L. R., CUNHA, C. C. D. A., MACEDO, M., 2011, "O método da revisão integrativa nos estudos organizacionais", *Gestão e Soc*, v. 5, n. 11, pp. 121–36.

BOWES, D., HALL, T., BEECHAM, S., 2012, "SLuRp: a tool to help large complex systematic literature reviews deliver valid and rigorous results". In: *Proceedings of the 2nd international workshop on Evidential assessment of software technologies*, pp. 33–36. ACM.

BRANDES, U., 2001, "A faster algorithm for betweenness centrality*", *Journal of Mathematical Sociology*, v. 25, n. 2, pp. 163–177.

CALDAS, M. A. E., 1986, *Estudos de revisão da literatura: fundamentação e estratégia metodológica*. Editora Hucitec com o apoio técnico e financeiro do MinC/Pro-Memória, Instituto Nacional do Livro.

CENDÓN, B. V., CAMPELLO, B. S., KREMER, J. M., 2000, *Fontes de informação para pesquisadores e profissionais*, v. 23. Editora Ufmg.

CHEN, C., 2004, "Searching for intellectual turning points: Progressive knowledge domain visualization", *Proceedings of the National Academy of Sciences*, v. 101, n. suppl 1, pp. 5303–5310.

CHEN, C., 2005, "The centrality of pivotal points in the evolution of scientific networks". In: *Proceedings of the 10th international conference on Intelligent user interfaces*, pp. 98–105. ACM.

CHEN, C., ZHANG, J., VOGELEY, M. S., 2009, "Visual analysis of scientific discoveries and knowledge diffusion". In: *Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI 2009)*.

CIENTÍFICO, D., 2004, "Revisão de Literatura e Desenvolvimento Científico: conceitos e estratégias para confecção", .

CORDEIRO, A. M., OLIVEIRA, G. M. D., RENTERÍA, J. M., et al., 2007, "Revisão sistemática: uma revisão narrativa", *Rev. Col. Bras. Cir*, v. 34, n. 6, pp. 428–431.

DE ALMEIDA, J. F., 2012, *BLOGMINER: REPRESENTAÇÃO TEMPORAL DE ASSUNTOS ATRAVÉS DE MODELAGEM DE TÓPICOS*. Tese de Doutorado, Universidade Federal do Rio de Janeiro.

DE FREITAS, L. Q., 2010, *Medidas de centralidade em grafos*. Tese de Doutorado, Universidade Federal do Rio de Janeiro.

DOS SANTOS, R. N. M., KOBASHI, N. Y., 2009, "Bibliometria, cientometria, infometria: conceitos e aplicações", *Tendências da Pesquisa brasileira em Ciência da Informação*, v. 2, n. 1.

DURELLI, V. H., FELIZARDO, K. R., DELAMARO, M. E., 2010, "Systematic mapping study on high-level language virtual machines". In: *Virtual Machines and Intermediate Languages*, p. 4. ACM.

EGGERS, S., HUANG, Z., CHEN, H., et al., 2005, "Mapping medical informatics research". In: *Medical Informatics*, Springer, pp. 35–62.

FABBRI, S., HERNANDES, E., DI THOMMAZO, A., et al., 2013, "Using information visualization and text mining to facilitate the conduction of systematic literature reviews". In: *Enterprise Information Systems*, Springer, pp. 243–256.

GLASS, G. V., 1976, "Primary, secondary, and meta-analysis of research", *Educational researcher*, pp. 3–8.

GOODMAN, D., 2005, "Web of Science (2004 version) and Scopus", *The Charleston Advisor*, v. 6, n. 3, pp. 5–5.

HAGBERG, A., SCHULT, D. A., SWART, P. J., 2013, "NetworkX", *URL http://networkx.github.io/index.html*.

HIRSCH, J. E., 2005, "An index to quantify an individual's scientific research output", *Proceedings of the National academy of Sciences of the United States of America*, v. 102, n. 46, pp. 16569–16572.

HULL, D., PETTIFER, S. R., KELL, D. B., 2008, "Defrosting the digital library: bibliographic tools for the next generation web", *PLoS computational biology*, v. 4, n. 10, pp. e1000204.

JÄSCHKE, R., HOTHO, A., SCHMITZ, C., et al., 2007, "Analysis of the publication sharing behaviour in BibSonomy". In: *Conceptual Structures: Knowledge Architectures for Smart Applications*, Springer, pp. 283–295.

KEELE, S., 2007, *Guidelines for performing systematic literature reviews in software engineering*. Relatório técnico, Technical report, EBSE Technical Report EBSE-2007-01.

KITCHENHAM, B., 2004, "Procedures for performing systematic reviews", *Keele, UK, Keele University*, v. 33, n. 2004, pp. 1–26.

KITCHENHAM, B., BRERETON, O. P., BUDGEN, D., et al., 2009, "Systematic literature reviews in software engineering–a systematic literature review", *Information and software technology*, v. 51, n. 1, pp. 7–15.

KLEINBERG, J. M., 1999, "Authoritative sources in a hyperlinked environment", *Journal of the ACM (JACM)*, v. 46, n. 5, pp. 604–632.

KLEINBERG, J. M., 2000. "Method and system for identifying authoritative information resources in an environment with content-based links between information resources". ago. 29. US Patent 6,112,202.

LIKERT, R., 1932, "A technique for the measurement of attitudes." *Archives of psychology*.

LIU, S., CHEN, C., 2013, "The differences between latent topics in abstracts and citation contexts of citing papers", *Journal of the American Society for Information Science and Technology*, v. 64, n. 3, pp. 627–639.

MAIA, R. T., 2008, "A importância da disciplina de metodologia científica no desenvolvimento de produções acadêmicas de qualidade no nível superior", *Revista Urutágua*, , n. 14.

MANOUSELIS, N., DRACHSLER, H., VUORIKARI, R., et al., 2011, "Recommender systems in technology enhanced learning". In: *Recommender systems handbook*, Springer, pp. 387–415.

MARSHALL, C., BRERETON, P., 2013a, "Tools to Support Systematic Literature Reviews in Software Engineering: A Mapping Study". In: *Empirical Software Engineering and Measurement, 2013 ACM/IEEE International Symposium on*, pp. 296–299. IEEE, a.

MARSHALL, C., BRERETON, P., 2013b, "Tools to Support Systematic Literature Reviews in Software Engineering: A Mapping Study". In: *Empirical Software Engineering and Measurement, 2013 ACM/IEEE International Symposium on*, pp. 296–299. IEEE, b.

MEADOWS, A. J., DE LEMOS LEMOS, A. A. B., 1999, *A comunicação científica*. Briquet de Lemos/livros.

NENKOVA, A., MCKEOWN, K., 2011, "Automatic summarization", *Foundations and Trends in Information Retrieval*, v. 5, n. 2-3, pp. 103–233. doi: 10.1561/1500000015. Disponível em: <http://www.scopus.com/inward/record.url?eid=2-s2.0-79960581921&

partnerID=40&md5=fa0f1f62485f643fe0309cbb1449a8fa>. cited By 66.

NEWMAN, M. E., GIRVAN, M., 2004, "Finding and evaluating community structure in networks", *Physical review E*, v. 69, n. 2, pp. 026113.

PETERSEN, K., FELDT, R., MUJTABA, S., et al., 2008, "Systematic mapping studies in software engineering". In: *12th International Conference on Evaluation and Assessment in Software Engineering*, v. 17. sn.

PETTICREW, M., ROBERTS, H., 2008, *Systematic reviews in the social sciences: A practical guide*. John Wiley & Sons.

RAJENDRA, A., PAWAN, L., 2008. "Building an intelligent web–theory and practice". .

THOMAZ, P. G., ASSAD, R. S., MOREIRA, L. F. P., 2011, "Uso do Fator de Impacto e do Índice H para avaliar pesquisadores e publicações", *Arq. bras. cardiol*, v. 96, n. 2, pp. 90–93.

VANHECKE, T. E., 2008, "Zotero", *Journal of the Medical Library Association: JMLA*, v. 96, n. 3, pp. 275.

WAGNER, C. S., LEYDESDORFF, L., 2005, "Network structure, self-organization, and the growth of international collaboration in science", *Research Policy*, v. 34, n. 10, pp. 1608 – 1618. ISSN: 0048-7333. doi: http://dx.doi.org/10.1016/j.respol.2005.08.002. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0048733305001745>.

WEBSTER, J., WATSON, R. T., 2002, "Analyzing the past to prepare for the future: Writing a literature review", *Management Information Systems Quarterly*, v. 26, n. 2, pp. 3.

WOHLIN, C., 2014, "Guidelines for snowballing in systematic literature studies and a replication in software engineering". In: *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*, p. 38. ACM.

Apêndice A

Resultados Integrais - Experimento 1

Esse apêndice apresenta, de forma integral, as respostas dos grupos G_1 e G_2 do experimento 1. Ou seja, são apresentadas as respostas às questões de 1 a 6 apresentadas no capítulo 5. As respostas de caráter pessoal, acerca da ferramenta, dadas pelo grupo G_1 encontram-se exclusivamente na seção 5.1.5, sobre análise dos resultados, uma vez que essa já se apresentava em sua integralidade.

A.1 Grupo com heurística (G_1)

As respostas do grupo que utilizou a heurística para responder às perguntas iniciais relacionadas aos conceitos descritos pela seção 5.1.1 são apresentadas a seguir. As respostas às perguntas de 1 a 3 seguem representadas através dos identificadores relacionados através da listagem geral de artigos 5.1.

Colunas e linhas representam respectivamente voluntários e artigos. Onde houver X marcado, entende-se que o voluntário citou tal artigo em sua resposta à questão em discussão.

1) Liste os artigos relacionados (periféricos) ao tema sumarização que você achou mais relevantes durante suas buscas.

Tabela A.1: Respostas da Questão 1

Id_Artigo	Voluntários								
	Vol 1	Vol 2	Vol 3	Vol 4	Vol 5	Vol 6	Vol 7	Vol 8	Vol 9
2									
3									
5									
6									
11									
14									

Tabela A.1: Respostas da Questão 1

Id_Artigo	Voluntários								
	Vol 1	Vol 2	Vol 3	Vol 4	Vol 5	Vol 6	Vol 7	Vol 8	Vol 9
15									
19									
21					X			X	X
22									
24									
26									
29									
31									
32									
33			X						
34									
37									
42				X					
43									
44									
45									
47									X
49									
50									
51									
54								X	
55									
56									
57									
58									
59									
63									
65		X							
66						X			
67					X			X	
69									
70									
72									
74							X	X	
75									
78	X	X							
79									
80									
81									X
83									
86									
87			X		X				X
88									
89							X		
93									
97									
98									
101						X			
102									
103									
106						X			
107				X					
109									

Tabela A.1: Respostas da Questão 1

Id_Artigo	Voluntários								
	Vol 1	Vol 2	Vol 3	Vol 4	Vol 5	Vol 6	Vol 7	Vol 8	Vol 9
110									
112									
114									
115			X						
116									
119									

2) Liste os artigos que encontrou durante suas buscas por sumarização que identificou como detalhando um ponto específico de forma profunda. Esses artigos, apesar de serem referenciados por artigos de sumarização, são pertencentes a outra área e foram utilizados como suporte (seja matemático, computacional ou até mesmo para simples contextualização dos artigos de sumarização).

Tabela A.2: Respostas da Questão 2

Id_Artigo	Voluntários								
	Vol 1	Vol 2	Vol 3	Vol 4	Vol 5	Vol 6	Vol 7	Vol 8	Vol 9
1									
4						X			
6									
7								X	
8									
10								X	
12									
13									
17						X			
18									X
19									
20				X					
21			X						X
23									
27									
30				X					
32									
37									
40							X		
47									X
48								X	
49									
55									
57									
61									
62					X				
63									
64									
69									
76									
77						X			
80									
81									X

Tabela A.2: Respostas da Questão 2

Id_Artigo	Voluntários								
	Vol 1	Vol 2	Vol 3	Vol 4	Vol 5	Vol 6	Vol 7	Vol 8	Vol 9
84					X				
85									
87									X
89			X						
90									
91									
92							X		
95			X						
96									
99					X				
100				X					
101									
102									
103									
108	X								
110									
113									
118		X							

3) Liste os 5 artigos mais relevantes para o tema sumarização encontrados em suas buscas.

Tabela A.3: Respostas da Questão 3

Id_Artigo	Voluntários								
	Vol 1	Vol 2	Vol 3	Vol 4	Vol 5	Vol 6	Vol 7	Vol 8	Vol 9
2									
3									
6									
9									
14									
16	X	X							
17									
18									X
20				X					
21			X		X			X	X
24									
25									
28				X					
32									
35									
36									
38	X					X			
39									
41									
42	X	X				X			
43									
45									
46									
47									X
49									
52					X		X		

Tabela A.3: Respostas da Questão 3

Id_Artigo	Voluntários								
	Vol 1	Vol 2	Vol 3	Vol 4	Vol 5	Vol 6	Vol 7	Vol 8	Vol 9
53									
54								X	
56									
59									
60									
67					X			X	
68				X					
70									
71									
72									
73							X		
74					X		X	X	
77	X					X			
81								X	X
82									
87			X		X				X
89			X				X		
94									
104		X							
105			X						
106	X	X				X			
107				X					
111						X			
112		X							
114									
116									
117									
120									
121				X					

As respostas às perguntas de 4 a 6 não possuem dicionário associado. O conjunto de respostas exibidos foi exatamente o mencionado por cada voluntário em suas respostas finais.

Para as respostas à questão 4, colunas e linhas representam respectivamente voluntários e autores listados, já para as repostas 5 e 6, representam voluntários e palavras-chave citadas.

4) Liste, no máximo, os 5 autores mais influentes para o tema sumarização. (Citar conforme consta nas publicações encontradas em suas buscas)

Tabela A.4: Respostas da Questão 4

Autores	Voluntários								
	Vol 1	Vol 2	Vol 3	Vol 4	Vol 5	Vol 6	Vol 7	Vol 8	Vol 9
Aliguliyev, R.M.									X
Androutsopoulos, I.									
Barzilay R								X	
Buckley, Chris		X		X					
Canhasi, E.									

Tabela A.4: Respostas da Questão 4

Autores	Voluntários								
	Vol 1	Vol 2	Vol 3	Vol 4	Vol 5	Vol 6	Vol 7	Vol 8	Vol 9
Chen, Y.									
Edmundson, Harold P					X			X	X
Ekin, A.									
Elena Lloret									
Erkan, G.		X							
Gallinari, P.									
Gomez-Hidalgo J.M					X				
Gupta, V.									
Hahn, U.									
Hirao, T.									
Hu, M.									
Jiang Peipei									
Kintsch, W.									
Kononenko, I.									
Lapata, M.									
Lee, L.									
Lehmam, Abderrafih						X			
Liu, B.									
Maña-López, M.J.									X
Manuel Palomar									
Marc Moens									
Marujo, L.									
Meena, Y.K.									
Mei, Qiaozhu						X			
Mitra, Mandar		X		X					
Moens, M.		X				X			
Naomie Salim									
Nenkova, A.									
Paice C.D					X				
Palomar, M.									
Panagiotis Stamatopoulosb									
Pang, B.									
Plaza, L.									
Radev, D.R.		X					X	X	
Roussinov							X		
Saggion, Horacio				X					
Sakai T.			X						
Salim, N.									
Salton, Gerard		X		X					
Sanderson Mark			X		X				
Silva, G.									
Simske, S.J.									
Singhal, Amit		X		X					
Sparck-Jones K.	X		X			X			
Stergos Afantenosa									
Teufel, Simone		X				X			
Thiago A. S. Pardo									
Tombros Anastasios			X		X			X	X
Torres-Moreno				X					
Tseng, Y.-H.									
van Dijk, T.A.									
Vangelis Karkaletsisa									

Tabela A.4: Respostas da Questão 4

Autores	Voluntários								
	Vol 1	Vol 2	Vol 3	Vol 4	Vol 5	Vol 6	Vol 7	Vol 8	Vol 9
Varadarajan, R									X
Wei Yongqing									
Wenjie Li									
Xu Mingying									
Zhang, Y.									

5) Liste as palavras-chave que você identificou como relevantes para sumarização durante suas buscas.

Tabela A.5: Respostas da Questão 5

Palavras-chave	Voluntários								
	Vol 1	Vol 2	Vol 3	Vol 4	Vol 5	Vol 6	Vol 7	Vol 8	Vol 9
Abstract					X				
Application									
Approaches									
Aproach									
Automated text summarization					X				
Automatic	X								
Automatic indicative summarization									
Automatic summarization				X					
Automatic text summarization									
Customer concern		X							
Distributed representations of text									
Document						X	X	X	
Document structure							X		
Document summarization									
Evaluation methods									
Event detection									
Extract					X				
Extraction approach									
Extraction techniques									
Extractive summarization									
Generic summarization					X				X
Indicative summarization									
Information								X	
Information retrieval				X					X
Keyphrase	X								
Knowledge								X	
Language generation				X					
Machine learning				X					
Method									
Multi-document summarization									
Multidocument							X		X
Natural language processing systems									
Product review		X							
Query			X						
Query focused summarization									X
Question answering				X					
Retrieval								X	
Search			X						
Semantics									

Tabela A.5: Respostas da Questão 5

Palavras-chave	Voluntários								
	Vol 1	Vol 2	Vol 3	Vol 4	Vol 5	Vol 6	Vol 7	Vol 8	Vol 9
Sentence extraction				X					
Software									
Summarization	X		X		X			X	
Summarization factor									
Summarization system				X					
Summarization techniques						X			
Summarizing						X	X		
System									
Techniques	X								
Text	X						X		
Text processing									
Text structuring									
Text summarizantion		X		X		X			
Text summarization relevant words									
Text summarization systems									
Ts				X					
Update summarization									
Web			X						

6) Liste as palavras-chave de assuntos relacionados (periféricos) ao tema sumarização encontrados em suas buscas.

Tabela A.6: Respostas da Questão 6

Palavras-chave	Voluntários								
	Vol 1	Vol 2	Vol 3	Vol 4	Vol 5	Vol 6	Vol 7	Vol 8	Vol 9
Abstract					X		X		
Artificial intelligence		X							
Automatic								X	
Automatic summarization									
Clustering			X					X	X
Community detection									
Comprehension			X						
Devices	X								
Documentaries									
Event-detection									
Extractive summarizations									
Films									
Fuzzy clustering									
Fuzzy logic									
Generic summarization									
Graph model									
Indexing model						X			
Information							X		
Information extraction									
Information retrieval									X
Information science									
Information summarization					X				
Keyphrase extraction						X			
Keyword extraction									
Machine learning									
Musical knowledge extraction									

Tabela A.6: Respostas da Questão 6

Palavras-chave	Voluntários								
	Vol 1	Vol 2	Vol 3	Vol 4	Vol 5	Vol 6	Vol 7	Vol 8	Vol 9
Musical rhythm									
Natural language									
Natural language processing systems									
Navigation			X						
Networks									
Paraphrasing									
Product review		X							
Query-reply									
Research								X	
Rhythmic pattern									
Screen	X								
Semantics									
Sentence extraction				X					
Sentence selection				X					
Similarity of text									
Small	X								
Speech			X						
Summarization	X				X				
Survey						X			
Svm									
Text evaluation				X					
Text extraction					X				
Text mining									
Text processing		X							
Text structuration						X			
Text summarization									
Video recording									
Video summaries									
Video summarization									
Videos									
Web search									X
Word frequency									

A.2 Grupo sem heurística (G_2)

As respostas do grupo que utilizou somente o mecanismo de busca da Scopus para responder às perguntas iniciais relacionadas aos conceitos descritos pela seção 5.1.1 são apresentadas a seguir. Conforme já mencionado, as respostas às perguntas de 1 a 3 seguem representadas através dos identificadores relacionados através da listagem geral de artigos 5.1.

Colunas e linhas representam respectivamente voluntários e artigos. Onde houver X marcado, entende-se que o voluntário citou tal artigo em sua resposta à questão em discussão.

1) Liste os artigos relacionados (periféricos) ao tema sumarização que você achou mais relevantes durante suas buscas.

Tabela A.7: Respostas da Questão 1

Id_Artigo	Voluntários								
	Vol 1	Vol 2	Vol 3	Vol 4	Vol 5	Vol 6	Vol 7	Vol 8	Vol 9
2						X			
3				X					
5							X		
6					X				
11				X					
14					X				
15			X						
19					X				
21									
22									X
24				X					
26			X						
29				X					
31		X							
32								X	
33									
34				X					
37				X					
42				X					
43		X			X				
44				X					
45				X					
47									
49								X	
50				X					
51						X			
54									
55						X			
56						X			
57				X					
58				X					
59				X					
63						X			
65									
66									
67									
69					X				X
70					X			X	
72					X				
74									
75							X		
78									
79			X					X	
80					X				
81									
83		X							
86	X								
87									
88			X						
89									
93							X		
97			X						
98									X

Tabela A.7: Respostas da Questão 1

Id_Artigo	Voluntários								
	Vol 1	Vol 2	Vol 3	Vol 4	Vol 5	Vol 6	Vol 7	Vol 8	Vol 9
101									
102					X				
103						X		X	X
106									
107									
109							X		
110				X					
112						X			
114					X				
115									
116					X				
119	X								

2) Liste os artigos que encontrou durante suas buscas por sumarização que identificou como detalhando um ponto específico de forma profunda. Esses artigos, apesar de serem referenciados por artigos de sumarização, são pertencentes a outra área e foram utilizados como suporte (seja matemático, computacional ou até mesmo para simples contextualização dos artigos de sumarização).

Tabela A.8: Respostas da Questão 2

Id_Artigo	Voluntários								
	Vol 1	Vol 2	Vol 3	Vol 4	Vol 5	Vol 6	Vol 7	Vol 8	Vol 9
1			X						
4									
6					X				
7									
8							X		
10									
12							X		
13							X		
17									
18									
19					X				
20									
21									
23							X		
27									X
30									
32								X	
37				X					
40									
47									
48									
49		X							
55						X			
57				X					
61			X						
62									
63						X			

Tabela A.8: Respostas da Questão 2

Id_Artigo	Voluntários								
	Vol 1	Vol 2	Vol 3	Vol 4	Vol 5	Vol 6	Vol 7	Vol 8	Vol 9
64							X		
69	X								
76			X						
77									
80					X				
81									
84									
85									X
87									
89									
90							X		
91			X						
92									
95									
96			X						
99									
100									
101		X							
102					X				
103	X	X				X		X	
108									
110				X					
113							X		
118									

3) Liste os 5 artigos mais relevantes para o tema sumarização encontrados em suas buscas.

Tabela A.9: Respostas da Questão 3

Id_Artigo	Voluntários								
	Vol 1	Vol 2	Vol 3	Vol 4	Vol 5	Vol 6	Vol 7	Vol 8	Vol 9
2						X			
3				X					
6		X			X				
9							X		
14					X				
16			X						
17			X						
18									
20									
21									
24				X					
25							X		
28									
32									X
35	X								
36			X						
38									
39								X	
41	X								
42				X					

Tabela A.9: Respostas da Questão 3

Id_Artigo	Voluntários								
	Vol 1	Vol 2	Vol 3	Vol 4	Vol 5	Vol 6	Vol 7	Vol 8	Vol 9
43									
45				X					
46		X							
47									
49								X	
52									
53			X						
54									
56	X				X				
59				X					
60	X								X
67									
68									
70		X						X	
71	X								
72					X				
73									
74									
77									
81									
82							X		
87									
89									
94		X							
104									
105									
106									
107									
111									
112			X			X	X		
114								X	
116		X			X			X	
117							X		
120									X
121									

Conforme previamente mencionado, as respostas às perguntas de 4 a 6 não possuem dicionário associado. O conjunto de respostas exibidos foi exatamente o mencionado por cada voluntário em suas respostas finais.

Para as respostas à questão 4, colunas e linhas representam respectivamente voluntários e autores listados. Em relação às repostas a 5 e 6, colunas e linhas representam voluntários e palavras-chave citadas.

4) Liste, no máximo, os 5 autores mais influentes para o tema sumarização. (Citar conforme consta nas publicações encontradas em suas buscas)

Tabela A.10: Respostas da Questão 4

Autores	Voluntários								
	Vol 1	Vol 2	Vol 3	Vol 4	Vol 5	Vol 6	Vol 7	Vol 8	Vol 9
Aliguliyev, R.M.									
Androutsopoulos, I.			X						
Barzilay R									
Buckley, Chris									
Canhasi, E.	X								
Chen, Y.							X		
Edmundson, Harold P									
Ekin, A.				X					
Elena Lloret		X	X			X	X		X
Erkan, G.									
Gallinari, P.						X			
Gomez-Hidalgo J.M									
Gupta, V.			X						
Hahn, U.				X					
Hirao, T.							X		
Hu, M.					X				
Jiang Peipei								X	
Kintsch, W.					X				
Kononenko, I.	X								
Lapata, M.				X					
Lee, L.					X				
Lehmam, Abderrafih									
Liu, B.			X						
Maña-López, M.J.									
Manuel Palomar		X							
Marc Moens									
Marujo, L.	X								
Meena, Y.K.	X								
Mei, Qiaozhu									
Mitra, Mandar									
Moens, M.									
Naomie Salim		X							
Nenkova, A.			X						
Paice C.D									
Palomar, M.						X			X
Panagiotis Stamatopoulosb								X	
Pang, B.					X				
Plaza, L.						X			
Radev, D.R.									
Roussinov									
Saggion, Horacio									
Sakai T.									
Salim, N.						X			X
Salton, Gerard				X					
Sanderson Mark									
Silva, G.	X								
Simske, S.J.									X
Singhal, Amit									
Sparck-Jones K.									
Stergos Afantenosa								X	
Teufel, Simone									
Thiago A. S. Pardo		X							

Tabela A.10: Respostas da Questão 4

Autores	Voluntários								
	Vol 1	Vol 2	Vol 3	Vol 4	Vol 5	Vol 6	Vol 7	Vol 8	Vol 9
Tombros Anastasios									
Torres-Moreno									
Tseng, Y.-H.				X					
van Dijk, T.A.					X				
Vangelis Karkaletsisa								X	
Varadarajan, R									
Wei Yongqing								X	
Wenjie Li		X							
Xu Mingying								X	
Zhang, Y.							X		

5) Liste as palavras-chave que você identificou como relevantes para sumarização durante suas buscas.

Tabela A.11: Respostas da Questão 5

Palavras-chave	Voluntários								
	Vol 1	Vol 2	Vol 3	Vol 4	Vol 5	Vol 6	Vol 7	Vol 8	Vol 9
Abstract									
Application								X	
Approaches							X		
Aproach								X	
Automated text summarization									
Automatic									
Automatic indicative summarization						X			
Automatic summarization				X		X	X		
Automatic text summarization						X			
Customer concern									
Distributed representations of text	X								
Document									
Document structure									
Document summarization		X							
Evaluation methods				X					
Event detection	X								
Extract									
Extraction approach			X						
Extraction techniques			X						
Extractive summarization	X								
Generic summarization									
Indicative summarization						X			
Information									
Information retrieval		X			X				
Keyphrase									
Knowledge									
Language generation									
Machine learning									
Method								X	
Multi-document summarization	X	X							
Multidocument									
Natural language processing systems					X				
Product review									
Query									X

Tabela A.11: Respostas da Questão 5

Palavras-chave	Voluntários								
	Vol 1	Vol 2	Vol 3	Vol 4	Vol 5	Vol 6	Vol 7	Vol 8	Vol 9
Query focused summarization									
Question answering									
Retrieval									
Search									
Semantics					X				
Sentence extraction									
Software							X		
Summarization							X		
Summarization factor							X		
Summarization system									
Summarization techniques							X		
Summarizing									
System							X		
Techniques									
Text							X		
Text processing		X			X				
Text structuring				X					
Text summarization		X	X	X	X				X
Text summarization relevant words			X						
Text summarization systems			X						
Ts									
Update summarization									X
Web									

6) Liste as palavras-chave de assuntos relacionados (periféricos) ao tema sumarização encontrados em suas buscas.

Tabela A.12: Respostas da Questão 6

Palavras-chave	Voluntários								
	Vol 1	Vol 2	Vol 3	Vol 4	Vol 5	Vol 6	Vol 7	Vol 8	Vol 9
Abstract									
Artificial intelligence									
Automatic									
Automatic summarization						X			
Clustering									
Community detection	X								
Comprehension									
Devices									
Documentaries								X	
Event-detection									X
Extractive summarizations		X							
Films								X	
Fuzzy clustering							X		
Fuzzy logic		X							
Generic summarization		X							
Graph model							X		
Indexing model									
Information									
Information extraction									X
Information retrieval						X			
Information science					X				

Tabela A.12: Respostas da Questão 6

Palavras-chave	Voluntários								
	Vol 1	Vol 2	Vol 3	Vol 4	Vol 5	Vol 6	Vol 7	Vol 8	Vol 9
Information summarization									
Keyphrase extraction									
Keyword extraction			X						
Machine learning			X						
Musical knowledge extraction	X								
Musical rhythm	X								
Natural language							X		
Natural language processing systems		X				X			
Navigation									
Networks	X								
Paraphrasing			X						
Product review									
Query-reply									X
Research									
Rhythmic pattern	X								
Screen									
Semantics		X							
Sentence extraction									
Sentence selection									
Similarity of text				X					
Small									
Speech									
Summarization									
Survey									
Svm							X		
Text evaluation									
Text extraction									
Text mining				X					
Text processing					X	X			
Text structuration									
Text summarization						X			
Video recording					X				
Video summaries					X				
Video summarization					X				
Videos								X	
Web search									
Word frequency			X						

Apêndice B

Tutorial da Ferramenta

Esse apêndice foi escrito com a finalidade de prover um tutorial sobre a ferramenta implementada por esse estudo. Parte-se do pressuposto que o leitor já domina os conceitos dissertados nos capítulos integrantes desse estudo. Portanto, o conteúdo aqui descrito visa ser objetivo e se ater somente às funcionalidades, sem explicar novamente seus conceitos.

A título de esclarecimento, nessa ferramenta, os vértices descritos por esse estudo são chamados de nós.

B.1 Visão geral

Essa seção apresenta uma visão geral da ferramenta implementada. Nela são enumerados todos os campos existentes para posterior explicação de suas funcionalidades.

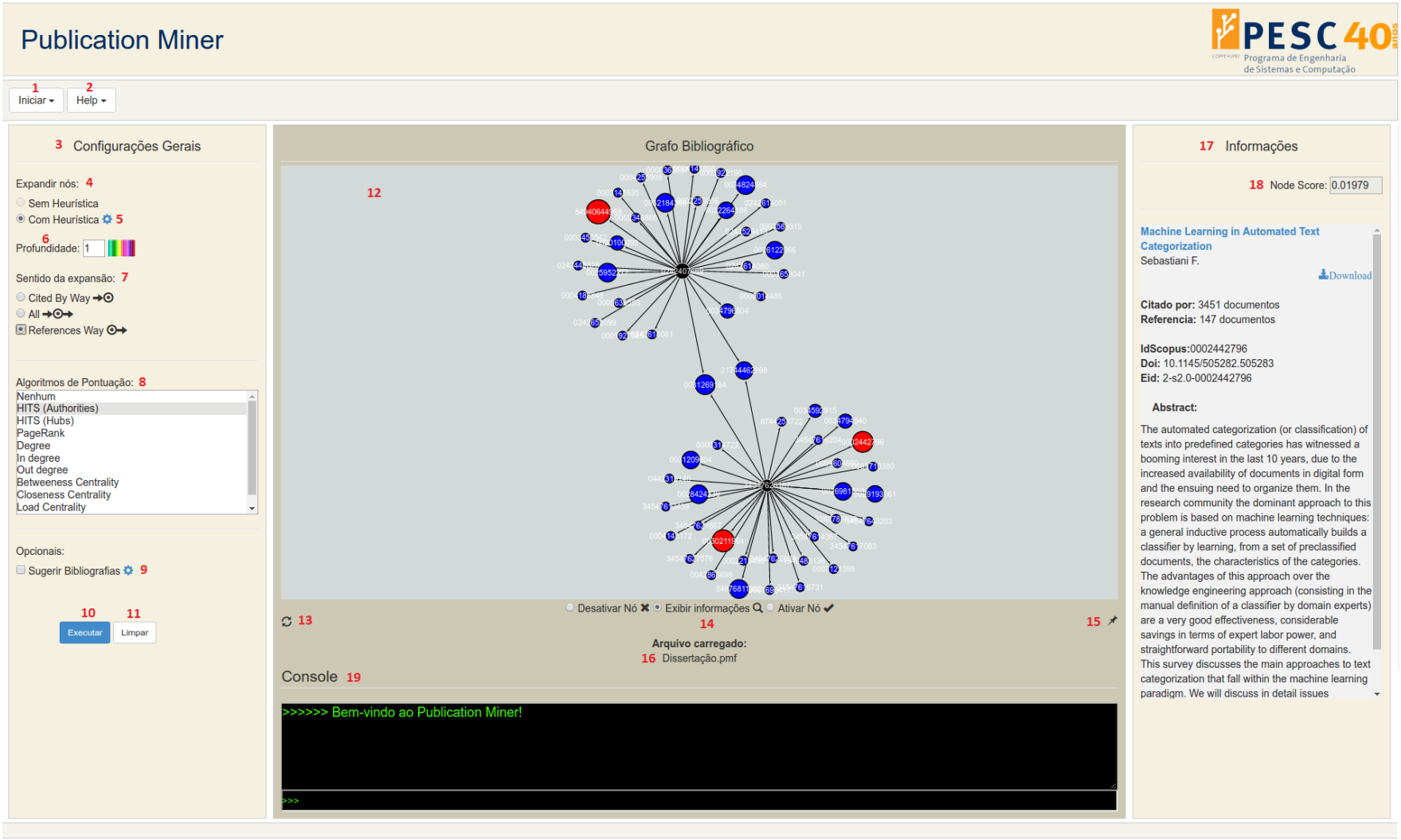

Figura B.1: Visão geral da ferramenta desenvolvida

1. Menu Iniciar
2. Menu Help
3. Configurações gerais
4. Expandir nós
5. Configurar heurística
6. Profundidade
7. Sentido da expansão
8. Algoritmos de pontuação
9. Sugerir bibliografias
10. Executar
11. Limpar
12. Exibir grafo
13. Recarregar grafo
14. Desativa/Ativar/Exibir Informações
15. Fixar nós
16. Arquivo carregado
17. Informações
18. Node score
19. Console

B.2 Funcionalidades

1. Menu Iniciar

   Esse menu possui acesso às funcionalidades básicas dessa ferramenta. Nele encontram-se os seguintes submenus conforme apresenta a figura B.2. Suas funcionalidade são listadas a seguir:

Figura B.2: Menu iniciar expandido

(a) Carregar um bib:

Essa funcionalidade realiza o carregamento de um arquivo com metadados, em formato BIBTEX, que representa uma ou mais referências. Cada referência é representada da seguinte forma:

@article{caled2016buzzword,

title={Buzzword detection in the scientific scenario},

author={Caled, Danielle and Beyssac, Pedro and Xexéo, Geraldo and Zimbrão, Geraldo},

journal={Pattern Recognition Letters},

volume={69},

pages={42–48},

year={2016},

publisher={Elsevier}

}

O arquivo a ser carregado por essa funcionalidade deve estar salvo é em formato texto com extensão ".bib". Seu conteúdo deve conter um ou mais registros como o anteriormente apresentado. Esses registros devem estar separados uns dos outros apenas por quebra de linha.

Esse tipo de dado pode ser exportado através de diversas ferramentas de consulta como Google Scholar. Basta buscar pela opção de exportar resultado para BIBTEX.

(b) Carregar um grafo:

Através desse recurso é possível carregar um grafo previamente salvo por essa ferramenta (formato ".pmf"). Trata-se de um arquivo em formato

JSON (*JavaScript Object Notation*).

(c) Salvar um grafo:

Essa opção serve para salvar o grafo atualmente carregado em formato ".pmf".

(d) Exportar grafo:

Exporta o grafo atualmente carregado para um formato padrão utilizado para lidar com grafos (Pajek) cuja extensão é ".net".

(e) Importar grafo:

Importa somente grafos gerados por essa ferramenta, pois a ferramenta depende de atributos específicos para poder carregar o grafo. Mas vale frisar que o formato de exportação deve ser lido normalmente pelas demais ferramentas compatíveis com esse formato.

(f) Configurações: Disponibiliza uma janela para que o usuário carregue um código chave que deve ser obtido seguindo as instruções apresentadas no ícone como obter chave.

(g) Sair: Através dessa opção a ferramenta é finalizada e o usuário é encaminhado para página do programa de mestrado que realizou a construção dessa ferramenta.

O Menu Help, apresentado pela figura B.3, possui as seguintes funcionalidades:

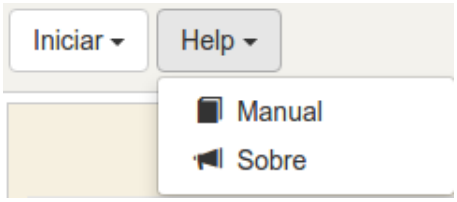

Figura B.3: Menu Help expandido

2. Menu Help

(a) Manual:

Exibe o manual simplificado de cada recurso;

(b) Sobre:

Exibe a descrição da ferramenta.

3. Configurações gerais

O quadro contendo as configurações gerais possui todos os campos direcionados a configurar os parâmetros para realização de buscas com a ferramenta. Nele são encontradas as seguintes definições:

4. Expandir nós

   Trata-se de um campo com duas opções que pesquisador pode optar. Entre essas duas opções de expansão dos vértices do grafo, também conhecidos como nós, estão:

   (a) Sem heurística: a expansão dos nós sem heurística baseia-se de uma expansão completa, isso quer dizer, expande todas as referências de suas referências ou citações de forma iterativa para construir o grafo de relacionamentos. Ou seja, se o conjunto inicial (profundidade 0) possui 3 referências, para o profundidade 1 o sistema irá adicionar ao grafo todas as referências dessas 3 referências. Para a profundidade 2 o sistema irá adicionar todas as referências das referências da profundidade 1, e assim sucessivamente. Deforma análoga o mesmo vale para o sentido das citações. Essa opção de expansão só depende dos parâmetros: profundidade e sentido da expansão.

   (b) Com heurística: esse é um tipo de expansão refinada que irá utilizar os parâmetro definidos em tela para tentar adicionar ao grafo somente nós que possuam maior relevância de acordo com a estrutura do grafo apresentada em tempo de execução. Essa expansão utilizará os parâmetros definidos pela janela apresentada pela imagem B.4 que pode ser aberta ao selecionar o ícone em formato de engrenagem, existente ao lado dessa opção. Além disso, essa heurística considera: profundidade, sentido da expansão e algorítimo de pontuação definidos.

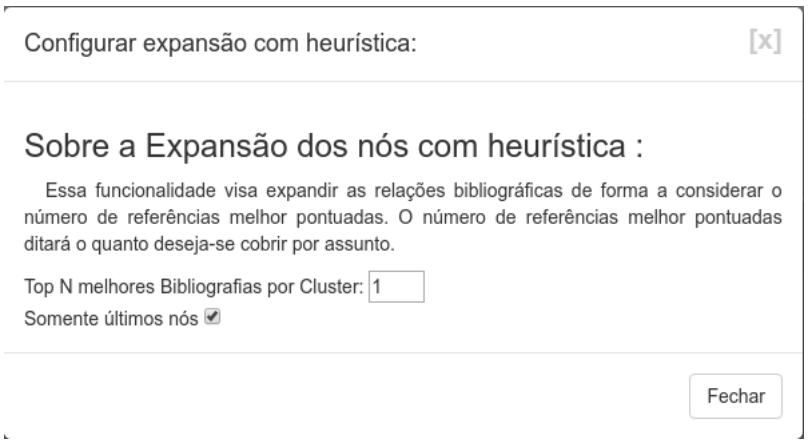

Figura B.4: Janela para definição de parâmetros da heurística

A diferença entre expandir os nós com heurística e sem heurística, é que a expansão com heurística utiliza a estrutura do grafo que está sendo construída

em tempo de execução para escolher por onde expandir de acordo com o caminho de melhor pontuação. Já a expansão sem heurística realiza expansão de todas as referências e citações possíveis.

5. Configurar heurística A opção de configurar heurística, acessada através da engrenagem indicada pela presente legenda possui os seguintes parâmetros a serem definidos:

   (a) Número de melhores bibliografias pontuadas a considerar a cada expansão. Indicará quantos possíveis nós podem ser expandidos.

   (b) O campo: somente últimos nós, estabelece se a cada iteração só serão expandidos os nós que estiverem entre os nós adicionados na iteração imediatamente anterior. Caso essa opção esteja desmarcada, o algoritmo poderá colocar como candidato a expansão qualquer nó do grafo total. Senão, esse tentará sempre maximizar seu ganho, seja por especificidade ou largura.

6. Profundidade

   O campo profundidade, reflete a profundidade atual do grafo. Em outras palavras, o maior caminho que há desde os nós iniciais aos últimos nós expandidos. Ou ainda, o número de iterações que ocorreram durante as expansões para que se pudesse visualizar o presente grafo (Dado que a cada iteração aumenta-se apenas em um a distancia máxima aos nós raízes).

7. Sentido da expansão

   Esse campo permite que o usuário possa escolher como será realizada a expansão do grafo, se irá expandir as referências, as citações ou ambos.

   Por exemplo: caso seja escolhido sentido das citações "*Cited by way*" significa que a medida em que o usuário aumentar a profundidade serão adicionados ao grafo nós que citam (ao apontam) para as referências que estão sendo expandidas. Caso seja escolhido o sentido das referências "*Reference's way*", então os nós adicionados serão apontados pelas referências que estão sendo expandidas. No caso de ser escolhido o sentido "*All*" então a ferramenta irá expandir nos dois sentidos citados.

8. Algoritmos de pontuação

   Através dessa opção o usuário pode escolher qual será o algoritmo usado para atribuir pontuação aos nós de acordo com a estrutura do grafo carregado. Essa é a pontuação a ser utilizada no caso de estar sendo realizada uma expansão com heurística e para realizar a tarefa de sugestão de bibliografias.

9. Sugerir bibliografias

   A opção sugerir bibliografia pode ser definida em qualquer momento. Caso um grafo já tenha sido expandido, pode-se apenas ativá-la e executá-la sem que se altere a profundidade. Pode-se também variar o tipo de algoritmo de pontuação a ser utilizado na sugestão.

   Há também um janela, indicada pela figura B.5, que se abre ao selecionar o ícone em formato de engrenagem existente ao lado desse campo. Essa irá definir quantas bibliografias se deseja que sejam sugeridas ao final de uma expansão.

Figura B.5: Janela para definiçaõ da quantidade de sugestões a serem indicadas.

10. Executar

    Botão responsável por ativar uma busca.

11. Limpar

    Botão responsável por redefinir configurações para parâmetros padrão.

12. Exibir grafo

    Área destinada a apresentar o grafo gerado pela ferramenta. Os nós podem ser arrastados e pode-se utilizar *zoom in* ou *zoom out* com o *scroll* do *mouse*. O tamanho de cada nó é calculado pela fórmula: $\log_2(n+1)+8$, onde n representa o número de citações do artigo representado pelo nó.

    As cores dos nós representam seus níveis e podem ser traduzidas pela seguinte escala (profundidade-cor): 0-preta, 1-azul, 2-verde, 3-amarelo, 4-laranja, 5-rosa, 6-violeta, 7-roxo, 8-azul escuro e 9-marrom.

13. Recarregar grafo

O botão em formato de recarregar localizado no canto esquerdo da divisão central, logo abaixo da visualização do grafo, serve para recarregar a exibição do grafo com sua máxima profundidade já alcançada.

14. Desativar/Ativar/Exibir Informações

    Cada nó pode ser selecionado para exibição de detalhes sobre ele. Esse são apresentados no campo: Informações. Basta selecioná-lo com a opção "exibir informações" marcada. Essa se localiza no centro da tela, abaixo da área de visualização do grafo, conforme indicado pela presente legenda. Nela, pode-se encontrar mais duas opções que servem para ativar e desativar nós do grafo. Isso significa que através desse recurso é possível controlar se deseja que o grafo deixe de expandir certos nós.

15. Fixar nós

    O botão localizado no canto direito da divisão central, logo abaixo da visualização do grafo (em formato de pin) serve para fixar os nós, fazendo com que eles não se movimentem na tela.

16. Arquivo carregado

    Esse campo indica qual foi o último arquivo carregado pela ferramenta.

17. Informações

    As informações sobre o atual *status* da ferramenta são mostradas no canto direito da tela. Nesse campo chamado de: Informações, são mostradas mensagens como: o que está sendo executado pela ferramenta, dados bibliográficos de um nó, a pontuação do nó que foi selecionado, entre outras possíveis mensagens que são carregadas nessa lateral.

    Dentre os dados bibliográficos de um nó estão informações como: *link* para baixar o arquivo contendo o texto integral indicado por esse nó, informações como quantidade de citação e referência, resumo do texto (*abstract*), DOI (*Digital object identifier*) e outros metadados.

18. Node score

    Campo que apresenta a pontuação do nó atualmente selecionado de acordo com o algoritmo de pontuação definido.

19. Console

    O terminal localizado no canto inferior da tela, destina-se ao pesquisador para programar em JS e intervir no grafo conforme desejar (Porém é necessário conhecer a estrutura interna do grafo).

Há também funções pré-definidas que podem ser utilizadas sem maiores dificuldades informando o idScopus do nó (esse pode ser obtido ao se detalhar um nós) ao qual deseja-se aplicar o resultado da função. São elas:

(a) tamanhoNoh(idScopus,tamanho);
Altera o tamanho de um nó informado.

(b) colorirNoh(idScopus,cor);
Colore um nó com a cor informada;

(c) tamanhoBordaNoh(idScopus,tamanho);
Altera o tamanho da borda de um nó.

(d) colorirBordaNoh(idScopus,cor);
Colore a borda de um nó.
Para executar as próximas funções sobre arestas, basta informar: idScopus do nó de origem, idScopus do nó de destino e cor. Porém se origem ou destino for omitido, será aplicado o resultado a todas as possíveis combinações com o parâmetro informado. Exemplo: ao se desejar colorir todas as arestas com origem em um determinado nó, basta somente informar o nó de origem. O mesmo pode ser feito para o de destino.

(e) colorirAresta(origem,destino, cor);
Colore uma aresta.

(f) colorirBordaAresta(origem, destino, cor);
Colore borda de uma aresta.

(g) tamanhoBordaAresta(origem, destino, tamanho);
Altera tamanho da borda de uma aresta.
Todos os argumentos devem ser passados como *string*, ou seja, entre aspas. Exemplo: tamanhoNoh(“100”,“10”)

B.3 Exemplo de uso

Essa seção descreve um exemplo de uso da ferramenta passo a passo para que o leitor tenha um noção geral do funcionamento dos itens anteriormente descritos. Para melhor acompanhamento, o leitor poderá utilizar a imagem B.1 da seção: visão geral e se guiar pelos números informados.

Ao começar utilizar a ferramenta, o primeiro passo será acessar o menu iniciar (1) e abrir o submenu configurações, caso esse não seja aberto automaticamente em seu primeiro acesso. Feito isso, o usuário deverá seguir as instruções para definir sua chave de acesso. Ao final, o uso de todas as funcionalidades estará disponível.

Para que a ferramenta carregue as informações iniciais em tela é necessário que se defina de que forma essas serão carregadas. Para isso, é possível utilizar 3 formas de entrada disponíveis através do menu iniciar (1): carregar bib, abrir grafo e importar grafo. Durante seu carregamento, mensagens sobre o andamento serão mostradas no quadro: informações (17). Informações sobre cada nó, ao ser detalhado, também serão mostradas em informações (17).

Após serem carregadas as informações inicias, seja utilizando qualquer uma das 3 maneiras previamente informadas, o usuário deverá visualizar na área de exibição do grafo (12) o grafo correspondente às informações carregadas. Nessa área é possível realizar *zoon in* e *zoom out*, detalhar um nó ao selecioná-lo com o *mouse* e desativar e ativar um nó selecionando as opções mostradas por (14).

Ainda direcionado à exibição do grafo, é possível através de (13) desenhar novamente o grafo ou através de (15), fixar os nós de forma que os mesmos não se movam. Outro ponto interessante a ser destacado é que o terminal indicado por (19) pode ser usado para intervir de forma visual no grafo através das diversas funções previamente já apresentadas ou utilizando *JavaScript*, caso o usuário domine a estrutura do objeto grafo em tela.

Caso seja desejado, em qualquer momento em que a ferramenta não estiver realizando carregamentos ou processamentos, é possível salvar o grafo atualmente carregado em tela, basta acessar (1) e escolher uma das 2 formas disponíveis: salvar grafo ou exportar grafo.

Uma vez carregado um grafo inicial, pode-se realizar diversas combinações de expansão. Para isso, basta variar as opções existentes em (3). Cada uma pode atender a um propósito distinto. Há dois tipos de expansão, conforme apresentado em (4): sem heurística e com heurística.

As expansões sem heurísticas são expansões completas, que recuperam todos os nós de acordo com o sentido e profundidades estabelecidos respectivamente por (6) e (7). Essa trata-se de uma expansão muito custosa em termos computacionais e de tempo. Caso o usuário não realize "podas" ao longo de suas expansões, essa poderá crescer de forma exponencial e será muito complicado para analisar as informações mostradas em tela devido a quantidade a ser analisada manualmente.

As expansões com heurística, conforme apresentado por esse estudo, possibilitam expansões mais controladas. Essas buscam seguir ampliando o grafo a cada iteração de acordo com a pontuação definida em (8). A o algoritmo de pontuação escolhido irá definir que tipo de pontuação será usada para selecionar os nós candidatos a expansão a cada iteração da heurística.

Para utilização dessa heurística, é necessário que se defina parâmetros estabelecidos em (5), profundidade (6), sentido da expansão(7) e algoritmo de pontuação(8), conforme já mencionado. Os parâmetros definidos em (5) são: número de melhores

bibliografias por *cluster* encontrado, que em outras palavras se refletirão como o quanto por área do conhecimento distinta será expandido e se é desejado sempre continuar uma expansão somente pelos últimos nós adicionados. A profundidade (6) irá influenciar conforme já explicado por esse estudo, assim como: (7) e (8), também explicados.

Além disso, pode-se utilizar a opção sugerir bibliografias (9) para que os n nós melhor pontuados sejam indicados como possíveis referências a serem utilizadas por uma pesquisa cujo conjunto bibliográfico inicial foi o carregado anteriormente. Essa opção permite que se escolha qual n será utilizado, para isso basta definir utilizando a janela aberta ao selecionar o ícone em formato de engrenagem ao lado de (9).

Por fim, uma vez definidas as configurações a serem utilizadas em um processamento, basta que o botão executar (10) seja pressionado. Feito isso, o usuário deverá aguardar para que o resultado seja mostrado na área de exibição do grafo (12). Enquanto aguarda, o usuário poderá acompanhar o andamento observando as mensagens exibidas em (17).

Vale frisar que um manual de referência rápida ou mais detalhes sobre a ferramenta podem ser obtidos acessando (2).