PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

MÉTODO PARA APOIO À CONSTRUÇÃO DE *STRINGS* DE BUSCA EM REVISÕES SISTEMÁTICAS POR MEIO DE MINERAÇÃO VISUAL DE TEXTO

GERMANO DUARTE MERGEL

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Ciência da Computação na Pontifícia Universidade Católica do Rio Grande do Sul.

Orientadora: Profa. Dra. Milene Selbach Silveira

Porto Alegre
2014

FICHA CATALOGRÁFICA

Dados Internacionais de Catalogação na Publicação (CIP)

M559i Mergel, Germano Duarte

Método para apoio à construção de *strings* de busca em revisões sistemáticas por meio de mineração visual de texto / Germano Duarte Mergel. – Porto Alegre, 2014.

103 f.

Dissertação (Mestrado) – Faculdade de Informática, PUCRS.
Orientador: Profª. Drª. Milene Selbach Silveira

1. Informática. 2. Engenharia de Software.
3. Mineração de Dados (Informática). I. Silveira, Milene Selbach.
II. Título.

CDD 005.1

Ficha Catalográfica elaborada pelo
Setor de Tratamento da Informação da BC-PUCRS

Pontifícia Universidade Católica do Rio Grande do Sul
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

TERMO DE APRESENTAÇÃO DE DISSERTAÇÃO DE MESTRADO

Dissertação intitulada "Método para Apoio à Construção de *Strings* de Busca em Revisões Sistemáticas por meio de Mineração Visual de Texto" apresentada por Germano Duarte Mergel como parte dos requisitos para obtenção do grau de Mestre em Ciência da Computação, aprovada em 31/07/2014 pela Comissão Examinadora:

Profa. Dra. Milene Selbach Silveira – Orientadora — PPGCC/PUCRS

Prof. Dr. Duncan Dubugras Alcoba Ruiz – PPGCC/PUCRS

Prof. Dr. Tiago Silva da Silva – UNIFESP

Homologada em 26/03/2015, conforme Ata No. 004 pela Comissão Coordenadora.

Prof. Dr. Luiz Gustavo Leão Fernandes
Coordenador.

PUCRS
Campus Central
Av. Ipiranga, 6681 – P32– sala 507 – CEP: 90619-900
Fone: (51) 3320-3611 – Fax (51) 3320-3621
E-mail: ppgcc@pucrs.br
www.pucrs.br/facin/pos

EPÍGRAFE

“[...] sei que algumas coisas nas quais acredito estão erradas; só não sei quais. Por isso ouço, leio, observo e discuto. Só assim chegarei um pouco mais perto da verdade.”

(Dr. Nigel Ashford, definindo humildade intelectual)

AGRADECIMENTOS

É incrível perceber o que aprendi nestes dois últimos anos. Ao longo de meu curso de Mestrado, aprendi a ser um aluno e pesquisador melhor, faminto por conhecimento. Agradeço a meus professores por isto, em especial à minha orientadora Profª Milene por sua confiança em mim e seu sempre presente otimismo. Não foram poucas as vezes que lhe ouvi responder: "*Não será fácil, mas acho que consegues!*". Agradeço, também, à Faculdade de Informática da PUCRS, por permitir que fosse, mais uma vez, seu aluno.

Aprendi a ser o pai de uma linda garotinha chamada Laís, que nasceu em Agosto de 2013, em um dia que terei para sempre na lembrança. Também aprendi que sempre poderei contar com o suporte, o carinho e a compreensão de minha amada esposa, Daniela, que sempre esteve presente e confiante. Aprendi, ainda, que o apoio incondicional de meus pais me trouxe até aqui e, sem ele, nada do que sou seria.

Com o comprometimento que me foi exigido, aprendi que não seria capaz de cursar uma pós-graduação sozinho. E sou grato em perceber que, ao longo de meu curso, nunca estive sozinho. Comigo matricularam-se minha família e minha orientadora, e lhes agradeço por tudo.

Obrigado por estes dois incríveis anos, que farão parte de minha vida para sempre!

MÉTODO PARA APOIO À CONSTRUÇÃO DE *STRINGS* DE BUSCA EM REVISÕES SISTEMÁTICAS POR MEIO DE MINERAÇÃO VISUAL DE TEXTO

RESUMO

Apesar do aumento na popularidade da aplicação de Revisões Sistemáticas da Literatura na Engenharia de Software, muitos pesquisadores ainda a apontam como um processo custoso e desafiador. Estudos levantados reportam problemas em diferentes atividades ao longo de seu processo, como na construção da *string* de busca da Revisão Sistemática e na seleção dos estudos primários.

Visando promover um auxílio à sua realização, métodos e ferramentas baseados em técnicas da área de Mineração Visual de Texto são propostas em estudos publicados da área, atuando em diversas etapas de uma Revisão Sistemática da Literatura. É percebida, porém, a ausência de métodos que auxiliem um pesquisador na construção da *string* de busca de sua Revisão Sistemática, na fase de planejamento da mesma.

Neste contexto, o presente trabalho visa qualificar o processo de construção da *string* de busca de uma Revisão Sistemática, propondo um método iterativo que, aplicando técnicas da Mineração Visual de Texto, apoia o pesquisador através da sugestão de termos relevantes de estudos selecionados. Os termos mais relevantes são extraídos de estudos selecionados e visualizados de forma a facilitar a decisão do pesquisador em incluí-los na *string* de busca utilizada, construindo e refinando a *string* de busca que será usada na Revisão Sistemática.

Uma ferramenta que implementa o método proposto foi desenvolvida, permitindo que testes com estes mesmos pesquisadores fossem realizados, e que uma análise sobre a viabilidade desta proposta fosse feita. Entrevistas realizadas com pesquisadores identificaram as dificuldades enfrentadas na realização de Revisões Sistemáticas e captaram suas opiniões a respeito da utilização do método proposto como solução.

Palavras-chave: Revisão Sistemática da Literatura, Mineração de Texto, Mineração Visual de Texto, Visualização de Informações, *string* de busca, método, ferramenta.

METHOD TO SUPPORT SEARCH STRING BUILDING IN SYSTEMATIC REVIEWS APPLYING VISUAL TEXT MINING

ABSTRACT

Despite the increased popularity of the adoption of Systematic Literature Reviews in Software Engineering, many researchers still indicate it as a costly and challenging process. Studies report problems in different activities throughout the review process, as in the construction of the Systematic Review search string and selection of primary studies.

Aiming to promote aid to its realization, tools based on methods and techniques from the Visual Text Mining area are presented in published studies, proposing assistance in various tasks of a Systematic Literature Review. However, it's perceived a lack of methods proposing to aid a researcher with the construction of the Systematic Review search string, on its planning phase.

In this context, this paper proposes an iterative method to assist the process of building the search string for a Systematic Review. Using Visual Text Mining techniques, it supports the researcher by suggesting terms for the search string. Relevant terms are extracted from studies selected by the researcher and shown in a visualization that facilitates the decision of the researcher to update the search string and include them, building and refining the search string that will be used in the Systematic Review.

A tool that implements the proposed method has been developed, allowing the execution of tests with researchers and an analysis of the feasibility of this proposal. Interviews with researchers identified the difficulties in performing Systematic Reviews and captured their opinions regarding the use of the proposed method, discussing its adoption.

Keywords: Systematic Literature Review, Text Mining, Visual Text Mining, Information Visualization, search string, method, tool.

LISTA DE FIGURAS

Figura 1 - Processo de Mineração de Dados (traduzido de Fayyad [Fay96]). 27

Figura 2 - Processo de Mineração de Texto (traduzido [Gar11]). 28

Figura 3 - Imagem que explora a capacidade visual humana na identificação de padrões [Nas06]. 35

Figura 4 - Exemplo ilustrativo de Treemap (criação do autor). 36

Figura 5 - Exemplo de um Grafo em Anel simples, com 8 nodos (criação do autor). 37

Figura 6 - Exemplo de grafo em anel com grupos e subgrupos (reproduzida de http://scaledinnovation.com/analytics/communities/communities.html). 38

Figura 7 - Exemplo de representação em Heatmap [criação do autor]. 39

Figura 8 - Processo de Revisão Sistemática, segundo Biolchini et al [Bio05]. 41

Figura 9 - Processo de Revisão Sistemática e técnicas de mineração por etapa (criação do autor, com base na literatura [Kit04]). 44

Figura 10 - Processo de Revisão Sistemática incluindo o método proposto, adaptado de Biolchini et al [Bio05].....45

Figura 11 - Fluxograma representando o método proposto incorporado às etapas de uma Revisão Sistemática da Literatura (criação do autor, com base na literatura [Kit04][Bio05]). 46

Figura 12 - Técnicas de Mineração de Texto aplicáveis em fases de uma Revisão Sistemática [Ana09]. 48

Figura 13 - Tela da ferramenta *ReVis* [Fel12]. 50

Figura 14 - Representação visual de termos-chave em um estudo [Chou11]. 51

Figura 15 - Análise do resultado sobre a experiência dos entrevistados em relação à Revisão Sistemática. 57

Figura 16 - Gráfico ilustrando as maiores dificuldades citadas pelos entrevistados, em relação ao número de entrevistados que a mencionaram. 59

Figura 17 - Tela da ferramenta *SLR.qub*. 61

Figura 18 - Área da string de busca da ferramenta *SLR.qub* em destaque. 61

Figura 19 - Edição do termo software da string de busca na ferramenta SLR.qub. 62

Figura 20 - Área dos termos sugeridos pela ferramenta. 62

Figura 21 - Área de opções da ferramenta SLR.qub. 63

Figura 22 - *Carousel* dos estudos resultantes da busca no *IEEExplore* e minerados pela ferramenta. 64

Figura 23 - Exemplo de marcação de estudos na ferramenta *SLR.qub*. Estudo 1 marcado como relevante, e estudo 3 como não-relevante. 64

Figura 24 - Fluxo de marcação dos estudos no carousel da ferramenta *SLR.qub*. 65

Figura 25 - Mapa de calor da ferramenta *SLR.qub*. 65

Figura 26 - Mapa de calor da ferramenta *SLR.qub* em detalhes. 66

Figura 27 - Página de resultados do site *IEEExplore*, com a área dos estudos resultantes destacada [Iee13]. 67

Figura 28 - Esquema representativo do fluxo de acionamento de um *bookmarklet* (criação do autor). 72

Figura 29 - Gráfico de marcações por documento. 77

Figura 30 - Marcação dos documentos por pesquisador. 79

LISTA DE TABELAS

Tabela 1 - Exemplo de valores de *tf(t,d)* calculados em cada uma das fórmula citadas. ..31

Tabela 2 - Valores de idf para os termos do exemplo, em uma coleção de três documentos. 32

Tabela 3 - Valores de tf-idf para os termos do exemplo, em uma coleção de três documentos. 33

Tabela 4 - Valores normalizados de *tf-idf* para os termos do exemplo. 33

Tabela 5 - Fases e etapas de uma Revisão Sistemática da Literatura (adaptado de Kitchenham [Kit04]). 40

Tabela 6 - Relação de estudos compreendendo ferramentas que auxiliam no processo de Revisão Sistemática, conforme resultado de um mapeamento sistemático (reproduzido de [Mar13]). 52

Tabela 7 - Relação de estudos e técnicas abordadas (adaptado de [Mar13]).....53

Tabela 8 - Síntese das ferramentas encontradas através do mapeamento sistemático [Mar13].....54

Tabela 9 - Relação de estudos sobre ferramentas de auxílio à Revisão Sistemática e as fases e etapas da revisão que são endereçadas (reproduzido de Marshall et al [Mar13]). 54

Tabela 10 - Roteiro de questões aplicadas na entrevista com pesquisadores. 55

Tabela 11 - Exemplificação da lista global de termos sugeridos em relação à marcação dos documentos como relevantes. 69

Tabela 12 - Valores de *tf-idf* em três documentos e na lista de termos sugeridos.....70

Tabela 13 - Síntese do resultado dos testes.....76

Tabela 14 - Visualização das marcações dos documentos nos testes realizados.....77

Tabela 15 - Strings construídas na execução dos testes, capturadas ao final dos mesmos. 79

Tabela 16 - Relação dos termos incluídos às strings de busca ao final dos testes. 80

Tabela 17 – Roteiro com questões aplicadas na entrevista pós-teste. 81

Tabela 18 – Síntese das respostas dos pesquisadores para as questões da entrevista pós-teste. 82

Tabela 19 - Recursos citados pelos pesquisadores como sendo mais úteis da ferramenta, em resposta à quarta questão (Q2.4) do questionário. 83

Tabela 20 - Experiência dos pesquisadores. 85

Tabela 21 - Estudo comparativo entre termos encontrados no corpo completo de um estudo [Bab09] e em seu *abstract*. 88

SUMÁRIO

1. INTRODUÇÃO 22

1.1 Metodologia de estudo e pesquisa 23

1.2 Estrutura do Trabalho 24

2. MINERAÇÃO VISUAL DE TEXTO (MVT) 25

2.1 Mineração de Dados 25

2.2 Mineração de Texto 27

2.2.1 Tarefas da Mineração de Texto 29

2.2.2 Técnicas da Mineração de Texto 30

2.3 Visualização de Informações 34

2.3.2 *Treemap* 36

2.3.3 Grafo em anel 36

2.3.4 Mapa de calor (*Heatmap*) 38

3. REVISÃO SISTEMÁTICA DA LITERATURA 40

3.1 A abordagem *quasi-gold standard (QGS)* 41

4. PROPOSTA DE MÉTODO DE APLICAÇÃO DE TÉCNICAS DE MVT NO AUXÍLIO À REVISÃO SISTEMÁTICA DA LITERATURA 43

4.1 Método proposto 45

4.2 Estudos relacionados 47

4.2.1 Identificando métodos de Mineração de Texto aplicáveis à Revisão Sistemática 47

4.2.2 Aplicação de técnicas de Mineração de Texto em auxílio à Revisão Sistemática 48

4.2.3 Abordagens de Mineração Visual de Textos no auxílio de uma Revisão Sistemática 49

4.2.4 Uma ferramenta de suporte à Revisão Sistemática 50

4.3 Levantamento das ferramentas utilizadas no auxílio à Revisão Sistemática da Literatura 51
4.4 Entrevistas com pesquisadores 55
4.4.1 Perfil dos entrevistados 56
4.4.2 Realização das entrevistas 56
4.4.3 Análise do resultado das entrevistas 56
5. ANÁLISE DO MÉTODO PROPOSTO 60
5.1 A ferramenta *SLR.qub* 60
5.1.1 Interface e utilização 60
5.1.2 Funcionamento e mineração 66
5.1.3 Fórmulas de relevância utilizadas 68
5.1.4 Formato 71
5.1.5 Controle de versionamento e disponibilidade 72
5.1.6 Limitações da implementação 73
5.2 Aplicação de testes com usuários 74
5.2.1 Descrição do teste 74
5.2.2 Perfil dos participantes 75
5.2.3 Seleção dos participantes 75
5.2.4 Apresentação dos resultados dos testes 75
5.3 Entrevistas pós-teste 81
5.3.1 Roteiro da entrevista 81
5.3.2 Apresentação do resultado das entrevistas 81
5.4 Discussão dos resultados 84
5.5 Limitações da pesquisa 87
6. CONSIDERAÇÕES FINAIS 90
6.1 Trabalhos futuros: suporte a outras bibliotecas digitais 92
6.2 Trabalhos futuros: utilização de algoritmos mais robustos 92
6.2 Trabalhos futuros: realização de testes com pesquisadores iniciantes em uma área 93

REFERÊNCIAS BIBLIOGRÁFICAS.....94
APÊNDICE A – ROTEIRO DE TESTE.....99
ANEXO A – TERMO DE CONSENTIMENTO.....102

1. INTRODUÇÃO

O volume de pesquisas empíricas na Engenharia de Software está em constante expansão. Com o número de trabalhos publicados diariamente, em muitas áreas tornou-se praticamente impossível para um pesquisador ler, realizar uma análise crítica e sintetizar o estado do conhecimento atual, muito menos atualizar esta base com certa periodicidade [Dyb08]. Como consequência deste aumento no número de estudos, revisões tornaram-se ferramentas essenciais para qualquer um que deseje manter-se atualizado na área. São as revisões que indicam, também, insuficiência nas bases de evidência pesquisadas até então, identificando a necessidade de novos estudos e oportunidades de pesquisa [Dyb08].

É necessário, porém, adotar uma abordagem sistemática para avaliar e agregar os resultados das revisões, com a finalidade de prover uma síntese objetiva e balanceada da evidência de pesquisa [Kit04]. Como solução, as Revisões Sistemáticas da Literatura avaliam e interpretam as pesquisas relevantes disponíveis para uma determinada questão de pesquisa, tópico da área ou fenômeno de interesse.

Classificadas como estudos secundários, as Revisões Sistemáticas da Literatura possuem um papel importante na pesquisa, uma vez que sintetizam o trabalho existente de maneira não tendenciosa [Kit04], utilizando-se de protocolos pré-estabelecidos para busca e identificação de estudos primários. Além disso, a estratégia adotada em uma revisão deve ser clara o bastante para permitir sua repetição por outros pesquisadores.

No entanto, Revisões Sistemáticas requerem um esforço consideravelmente maior do pesquisador, quando comparadas a revisões mais tradicionais [Kit04]. Talvez por isso, a maioria das Revisões Sistemáticas publicadas seja, atualmente, conduzida por pesquisadores mais experientes [Bab09]. Segundo estudos levantados [Dyb08] [Bab09] [Ria10], entre as maiores dificuldades da realização deste tipo de revisão, estão a construção da *string* de busca e a seleção de estudos primários. Estes são problemas comuns na realização de Revisões Sistemáticas, enfrentados tanto por pesquisadores iniciantes quanto por mais experientes [Ria10]. Assim sendo, tais problemas necessitam de atenção especial dos pesquisadores e poderiam, se endereçados, auxiliar a execução da Revisão Sistemática.

Hoje, é possível encontrar estudos que utilizam o conhecimento de outras áreas no auxílio à identificação e seleção de estudos primários na Revisão Sistemática da Literatura [Mar13]. São pesquisas que procuram aplicar técnicas de Mineração de Texto em estudos encontrados durante a fase de condução de uma Revisão Sistemática [Ana09] [Tho11]

[Fel12], estudam o impacto da Mineração Visual de Texto - uma área relacionada à Mineração de Dados - na seleção de estudos primários por um pesquisador [Mal07] [Fel11b], ou exploram a construção de uma ferramenta que apoia a realização de uma revisão [Cho11].

Como exemplo, Felizardo et al [Fel12] sugerem a utilização de técnicas de mineração e de visualização, a fim de facilitar a etapa de seleção de estudos em uma Revisão Sistemática. Sendo a maior parte do processo deste tipo de revisão realizada manualmente, a seleção de estudos primários pode tornar-se trabalhosa nos cenários em que muitos resultados são retornados na pesquisa. Dessa forma, justifica-se uma abordagem que utilize técnicas de mineração e de visualização que apoiem a descoberta de conhecimento [Kei02].

No intuito de auxiliar a realização de uma Revisão Sistemática da Literatura, esta pesquisa propõe um método que facilita e apoia o pesquisador na construção da *string* de busca de uma revisão do tipo, através do uso de técnicas de Mineração Visual de Texto.

Tendo em vista que o processo de Revisão Sistemática compreende uma busca na literatura por estudos usando uma *string* de busca pré-definida, foi desenvolvida uma ferramenta que utiliza a visualização e exploração dos estudos encontrados com o apoio de técnicas da área de Mineração Visual de Texto. Seu objetivo é o de associar algoritmos de Mineração de Texto e técnicas de Visualização de Informações no apoio à interação do usuário e, de maneira iterativa, auxiliá-lo na identificação e seleção de termos relevantes para construção e refinamento da *string* de busca de sua Revisão Sistemática da Literatura.

1.1 Metodologia de estudo e pesquisa

Para se alcançar os objetivos propostos neste trabalho foram executados alguns passos para entender e, desta forma, coordenar os conceitos da área de Mineração Visual de Texto e de Revisão Sistemática. Primeiramente foi feita uma revisão literária das áreas com o intuito de identificar as dificuldades dos pesquisadores na realização de Revisões Sistemáticas e as pesquisas existentes que procuram auxiliar o processo.

Após, foi elaborado um roteiro de entrevista com questões visando captar dificuldades enfrentadas pelos pesquisadores na realização de Revisões Sistemáticas e sua experiência com este tipo de revisão.

Para ajudar na análise da aplicabilidade do método proposto, foi desenvolvida uma ferramenta que o implementa, utilizando a visualização e exploração dos estudos

encontrados com o apoio de técnicas da área de Mineração Visual de Texto para extração e sugestão de termos para a *string* de busca. A ferramenta foi posta em prática em testes com os pesquisadores entrevistados. Com o intuito de obter suas opiniões em relação à aplicabilidade do método, foi elaborado um roteiro semiestruturado com questões abertas, utilizado em entrevistas realizadas logo após os testes.

Como resultado, foram capturadas as opiniões dos pesquisadores em relação ao método proposto, sendo relacionadas à experiência dos mesmos na realização de Revisões Sistemáticas e dificuldades enfrentadas com o objetivo de analisar a aplicabilidade do método proposto.

1.2 Estrutura do Trabalho

Os capítulos seguintes abordam a fundamentação teórica deste trabalho, introduzindo a área de Mineração Visual de Texto (seção 2), com tópicos da Mineração de Dados (2.1), Mineração de Texto (2.2) e Visualização de Informações (2.3), e a Revisão Sistemática da Literatura (seção 3).

Na seção 4, são apresentados o método proposto, estudos relacionados, e a entrevista realizada com pesquisadores, que visava captar a experiência de pesquisadores na realização de Revisões Sistemáticas e focar nas dificuldades por eles enfrentadas. É apresentada, na seção 5, a ferramenta implementada com o intuito de verificar a aplicabilidade do método proposto, através da análise do resultado de testes realizados com os pesquisadores entrevistados. Encerrando, na seção 6, são apresentadas as considerações finais sobre este trabalho, juntamente com sugestões para trabalhos futuros.

2. MINERAÇÃO VISUAL DE TEXTO (MVT)

A área de Mineração Visual de Texto é, na verdade, uma intersecção entre as áreas de Mineração de Texto e Visualização de Informações [Oli03]. Aplicações em Mineração Visual de Texto vão além da simples visualização de resultados de uma mineração, utilizando técnicas da Visualização de Informações durante o processo de Mineração de Texto, na representação visual e exploração dos dados.

Há de se considerar a diferença entre as áreas de Mineração Visual de Dados (bem como de texto) e a Mineração de Dados em imagens. Enquanto que a primeira área estuda métodos pelos quais uma visualização é gerada a partir de dados extraídos de uma base de dados ou documentos, como a Visualização de Modelos [Oli03], a Mineração de Dados sobre imagens aborda temas como o particionamento ou preparação de uma imagem para obtenção de informações sobre a mesma [Gon12].

Para melhor entender o conceito de Mineração Visual de Texto, faz-se necessário um estudo sobre Mineração de Dados, Mineração de Texto e Visualização de Informações, abordados nesta seção.

2.1 Mineração de Dados

A Mineração de Dados é definida como a área de pesquisa que estuda o processo de extração não-trivial de informações implícitas, previamente desconhecidas e potencialmente úteis, sobre bases de dados [Tan05]. Utilizando métodos de exploração e análise, por meios automáticos ou semiautomáticos, a Mineração é capaz de processar grandes quantidades de dados em busca da identificação de padrões significativos.

Tal processo compreende tarefas de dois tipos: preditivas e descritivas [Tan05]. As preditivas, em Mineração de Dados, têm por objetivo predizer o valor de um atributo específico, analisando um contexto de valores conhecidos. Já as tarefas descritivas têm por objetivo a descoberta de padrões implícitos, realizando análise de agrupamentos, associação e detecção de anomalias.

Uma diferente definição é encontrada em Fayyad [Fay96], na qual a atividade de mineração de dados é caracterizada como uma etapa do processo de Extração de Conhecimento (do inglês, *KDD - Knowledge Discovery in Databases*). Na concepção de Fayyad, *KDD* refere-se ao processo geral de descoberta de conhecimento útil em dados, e a mineração de dados é a atividade do processo na qual algoritmos específicos são aplicados para extração de padrões. Já a área de Mineração de Dados preocupa-se com o

desenvolvimento de métodos e técnicas “que deem sentido aos dados”, com o objetivo de endereçar um problema que a era da informação digital nos trouxe: a sobrecarga de dados. Segundo Fayyad [Fay96]:

> [...] Seja na ciência, marketing, economia, saúde, varejo, ou qualquer outro campo, a abordagem clássica para análise de dados depende fundamentalmente de um ou mais analistas tornarem-se intimamente familiares com os dados e servirem como de interface entre os dados e os usuários e produtos. [...] Na verdade, com o aumento drástico do volume de dados, este tipo de análise manual de dados torna-se completamente impraticável em muitos domínios. […] Acreditamos que esta tarefa não deva ser para humanos; portanto, o trabalho de análise precisa ser automatizado, ao menos em parte.

Neste trabalho, o termo “Mineração de Dados” referir-se-á à área de pesquisa ou ao processo de descoberta, variando seu significado segundo o contexto, sendo o mesmo válido para áreas relacionadas e referenciadas neste documento, como a Mineração de Texto.

O processo de Mineração de Dados, segundo Fayyad [Fay96], é interativo e iterativo, envolvendo diversas etapas. Nove delas são, por ele, consideradas básicas, e descritas como segue:

1. Compreensão do domínio da aplicação e definição do objetivo do processo de Mineração;
2. Seleção de uma base de dados ou subconjunto de dados sobre o qual o processo de descoberta será aplicado;
3. Limpeza e pré-processamento dos dados selecionados. Alguns exemplos de tarefas realizadas nesta etapa são remoção de dados ruidosos e estratégia para lidar com dados vazios;
4. Redução da dimensionalidade e projeção dos dados, procurando a melhor forma, dentro do objetivo proposto, para representar os dados;
5. Identificação de métodos de mineração que atendam ao objetivo proposto. Alguns dos métodos são: sumarização, classificação e *clustering*;
6. Escolha do algoritmo de mineração e método a ser utilizado na busca por padrões;
7. Aplicação da atividade de mineração propriamente dita, buscando por padrões de interesse;

8. Interpretação dos padrões minerados, com possível retorno a qualquer uma das etapas anteriores para maiores iterações e refinamentos. Esta etapa pode envolver visualização dos padrões extraídos;
9. Atuação sobre o conhecimento descoberto através do processo de Mineração, incorporando o resultado a um sistema ou simplesmente documentando o processo.

A figura 1 ilustra as etapas do processo de Mineração de Dados, como definidas por Fayyad [Fay96].

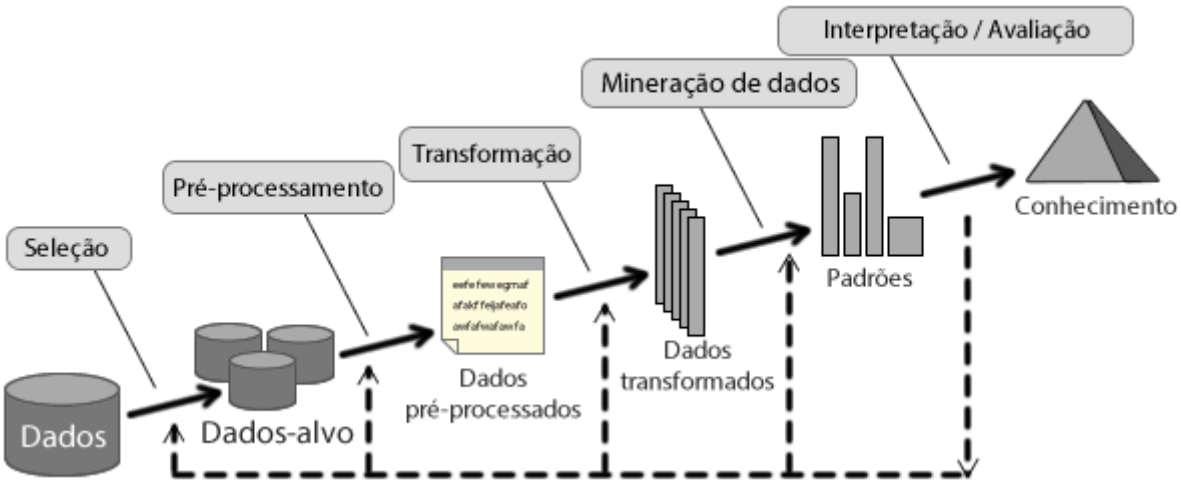

Figura 1 - Processo de Mineração de Dados (traduzido de Fayyad [Fay96]).

Han [Han06] sugere a caracterização das etapas do processo de mineração de dados como pertencentes a um dos seguintes estágios:

- **Pré-processamento:** preparação dos dados para realização da mineração, possivelmente envolvendo seleção e transformação dos dados;
- **Mineração de dados:** aplicação de um algoritmo de mineração e extração de informações;
- **Pós-processamento:** identificação de padrões encontrados na mineração e preparação para visualização das informações descobertas.

2.2 Mineração de Texto

Mineração de Texto é um processo de descoberta de novas e desconhecidas informações, extraídas automaticamente de um ou diferentes documentos de texto [Hea03] ou aplicado sobre coleções de textos, na descoberta de padrões e relações entre documentos [Gar11].

Por vezes tratado por Mineração de Dados Textuais, o processo de Mineração de Texto assemelha-se ao de Mineração de Dados na medida em que ambos utilizam técnicas

de mineração em grandes quantidades de dados, em busca de padrões significativos [Hea03]. Porém, na Mineração de Texto, os padrões são extraídos de textos em linguagem natural, ao invés de uma base de dados. Uma base de dados estruturada pode facilitar o processamento automático de programas, mas textos possuem significado semântico e requerem o emprego de métodos de mineração mais específicos, próprios da área de Mineração de Texto [Hea03].

Assim como o processo de Mineração de Dados, a Mineração de Texto é, também, um processo que compreende vários estágios [Gar11]. Entre eles estão o pré-processamento e transformação do texto, a seleção de propriedades, a descoberta de padrões e a interpretação e avaliação dos resultados. A figura 2 representa visualmente as etapas de um processo de mineração que utiliza recuperação de informações e processamento de linguagem natural para extrair informações de textos.

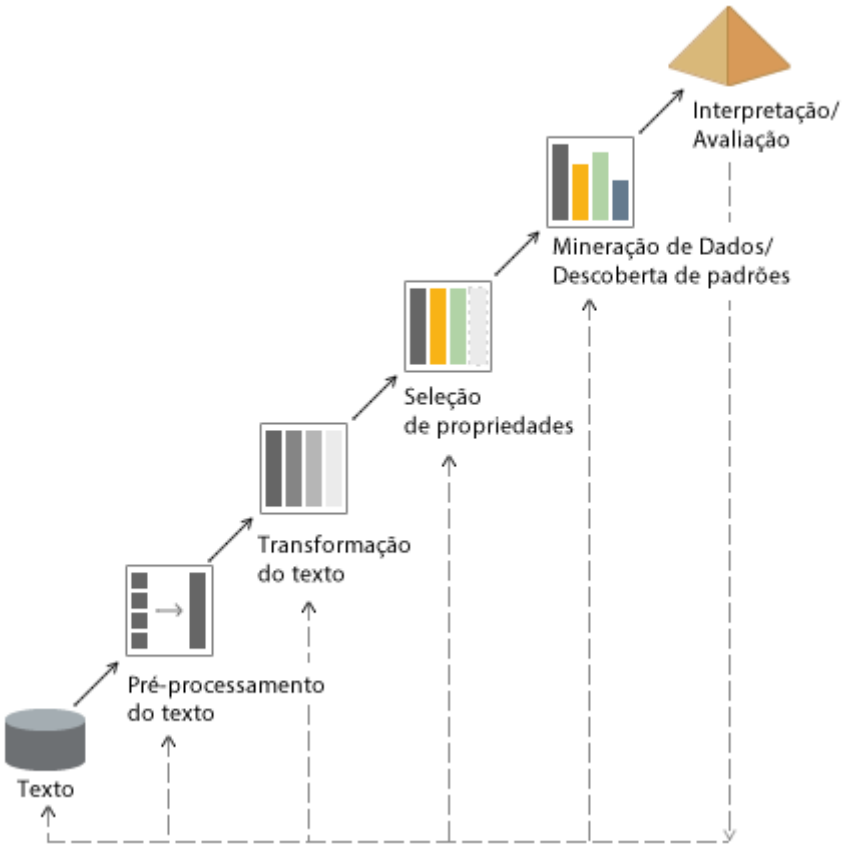

Figura 2 - Processo de Mineração de Texto (traduzido [Gar11]).

2.2.1 Tarefas da Mineração de Texto

Com o aumento no número de documentos eletrônicos disponíveis, a Mineração de Texto tem aumentado em popularidade e importância como campo de pesquisa da Mineração de Dados [Gar11]. Textos representam uma vasta e rica variedade de informações, mas conseguem codificar tais informações de maneira a dificultar sua interpretação automática. Métodos sofisticados devem ser, então, empregados na realização de tarefas de mineração de maneira eficiente [Hea99].

São várias as tarefas utilizadas na Mineração de Texto, mas algumas são consideradas mais relevantes, como o Reconhecimento Automático de Termos (*Automatic Term Recognition*), o Agrupamento de Documentos (*Document Clustering*), a Classificação de Documentos (*Document Classification*) e a Sumarização Automática de Documentos (*Automatic Summarization*) [Tho11]. Uma breve descrição de cada uma destas se faz necessária para melhor entender suas aplicações:

- **Reconhecimento Automático de Termos:** tarefa de mineração que procura identificar e extrair automaticamente termos técnicos de um documento. Estes termos tipicamente correspondem aos conceitos principais do estudo;
- **Agrupamento de Documentos:** objetiva agrupar coleções de documentos com base nos tópicos principais de seu conteúdo. Os grupos produzidos podem ser vistos como agrupamentos por tópico, abrangendo todos os documentos que compartilham um mesmo assunto. Identificar o tópico principal em um documento não é trivial para Mineração de Texto [Tho11], e abordagens atuais baseiam-se na frequência dos termos em um documento em comparação com sua frequência na coleção de todos os documentos;
- **Classificação de Documentos:** utiliza padrões encontrados em novos documentos analisados para identificá-los com uma categoria. Pode ser visto, na verdade, como uma função de mapeamento, que mapeia um item de dados em uma ou mais classes pré-definidas [Fay96];
- **Sumarização de Documentos:** tarefa na qual frases relevantes são extraídas de trechos do estudo com o intuito de criar um resumo do mesmo. Em uma definição mais abrangente, o método de sumarização é dito como a busca por uma descrição compacta de um subconjunto de dados [Fay96].

Os métodos aplicados na realização das tarefas mencionadas envolvem busca, extração e categorização de informações de documentos [Tho11]. Além dos métodos, técnicas que auxiliam na representação, indexação e classificação de documentos são de

suma importância na realização de uma Mineração de Texto, de acordo com o contexto em que elas são realizadas [Fay96]. Uma delas é a representação espacial dos documentos na forma de vetores (*Vector Space Model*), utilizando cálculo de relevância de termos por frequência, *tf-idf* (*term frequency-inverse document frequency*). Esta técnica, por ser utilizada na ferramenta proposta, é abordada na próxima subseção.

2.2.2 Técnicas da Mineração de Texto

Autores como Salton e Yang [Sal75] provaram que uma maneira interessante de representar documentos com o intuito de comparação e indexação é através da vetorização dos mesmo, utilizando a técnica denominada *Vector Space Model*. Com uma abordagem baseada em um espaço vetorial, mostraram ser possível utilizar a distância e o ângulo entre vetores na identificação de um vocabulário ideal para representação de uma coleção de documentos.

Na representação por *Vector Space Model*, considera-se uma coleção de documentos onde cada documento *Di* contém uma coleção de termos *Ti*, sendo os termos pesados conforme sua relevância naquele documento. É possível representar, portanto, o documento *Di* como um vetor de *t* dimensões, onde *t* termos aparecem em *Ti*.

Por definição, todo documento da coleção de documentos pode ser representado vetorialmente. Mas, para isso, cada termo deve ser representado numericamente, calculando seu peso ou relevância dentro de um documento. Assim, poderíamos reescrever o mesmo vetor de *Di* levando em consideração o peso de cada termo, como segue: *di = (wi1, wi2, wi3, …)*, onde *wi1* equivale ao peso *w* do termo *t1 no* documento *di* [Sal75].

Uma maneira muito comum na Mineração de Textos para se calcular o peso de cada termo de um documento é através da frequência com que o termo ocorre. Dessa forma, um determinado termo *t*, em um documento *d*, poderia ter sua frequência descrita como *f(t, d)* [Sal75]. Palavras muito frequentes em um documento têm um peso maior, o que não indicaria, necessariamente, que o termo é mais relevante. Considerando-se artigos e pronomes, por exemplo, são palavras frequentemente utilizadas sem que sejam relevantes na representação de um documento. Para resolver este problema, calcula-se um valor de relevância para os termos de um documento utilizando uma fórmula denominada *tf-idf* [Sal75].

Tf-idf (*term frequency - inverse document frequency*) é um cálculo estatístico que considera, além da ocorrência do termo em um documento, o número de documentos da coleção nos quais o termo é utilizado [Man08]. Dessa forma, termos muito utilizados em um

determinado documento, mas não presentes em outros documentos da coleção, recebem um maior peso e são percebidos como mais relevantes para aquele documento em particular. Para isso, o valor de *tf-idf* é calculado utilizando-se a seguinte fórmula [Man08]: $tf\text{-}idf(t,d) = tf\,(t,d) * idf\,(t)$, onde *d* é o documento no qual o termo *t* tem o valor de relevância *tf-idf*.

Para entender melhor a fórmula *tf-idf*, é preciso analisá-la em partes, sendo a primeira parte a frequência ou ocorrência de um termo em um documento. Denotada por *tf(t,d)*, onde *t* representa o termo e *d* o documento, a forma como a frequência é calculada pode variar de autor para autor. Algumas de suas principais fórmulas são conhecidas por natural (*natural term-frequency*), logarítmica (*logarithm term-frequency*) e booleana (*boolean term-frequency*) [Man08].

Na fórmula natural, contam-se as aparições de determinado termo em um documento. Logo, o valor de *tf* equivale à frequência daquele termo no documento. No cálculo logarítmico, o valor de *tf* é determinado pela seguinte fórmula:

$$log\ tf(t,d) = \begin{cases} 1 + \log\big(tf(t,d)\big)\ se\ \ tf(t,d) > 0 \\ 0\ sen\tilde{a}o \end{cases}$$

Dessa forma, termos cuja frequência é muito maior que outros têm seu valor amortizado, evitando grandes distorções na comparação entre valores. Na versão booleana, se um termo aparece ao menos uma vez no documento, seu valor para *tf* é 1. Caso não esteja presente no documento, seu *tf* será 0. A fórmula para a versão booleana é dada por:

$$booltf(t,d) = \begin{cases} 1\ se\ tf(t,d) > 0 \\ 0\ sen\tilde{a}o \end{cases}$$

Na tabela 1 encontra-se uma comparação entre as fórmulas citadas de *tf* para alguns termos em um único documento.

Tabela 1 - Exemplo de valores de *tf(t,d)* calculados em cada uma das fórmula citadas.

	Fórmulas		
Termos	natural	logarítmica	booleana
mineração	18	2,25	1
texto	32	2,50	1
carro	0	0	0

Tendo calculado a primeira parte da fórmula *tf-idf* (a frequência do termo *t* em um documento *d*) o valor da segunda parte (*idf*) pode ser calculado na seguinte fórmula [Man08]:

$$idf(t, D) = \log \frac{|D|}{|\{d \in D | t \in d\}|}$$

Nesta fórmula de *idf*, |*D*| refere-se à cardinalidade da coleção de documentos (número de documentos na coleção), e |*{d E D | t E d}*|, ao número de documentos pertencentes ao conjunto *D* onde o termo *t* está presente. Na tabela 2 está uma exemplificação de valores de *idf* para os termos da tabela 1, em um cenário hipotético com uma coleção de três documentos. Nela, a frequência dos termos para cada documento é encontrada nas colunas *D1*, *D2* e *D3*, bem como o número de documentos nos quais o termo está presente, exibido na coluna "*Total docs*". Por fim, o valor de *idf* para cada documento, na coleção de documentos, está na última coluna, denominada *idf*.

Tabela 2 - Valores de idf para os termos do exemplo, em uma coleção de três documentos.

	Freqüência			Total docs	
Termos	D1	D2	D3		idf
mineração	18	0	2	2	**0,18**
texto	32	2	5	3	**0**
carro	0	0	78	1	**0,48**

Percebe-se que a frequência do termo (*tf*) é calculada como atributo de um documento, utilizando dados locais (apenas daquele documento) para seu cálculo, enquanto que a frequência inversa nos documentos (*idf*) é um atributo global da coleção de documentos. Em outras palavras, dado um termo *t*, o valor de sua frequência (*tf*) variará (possivelmente) de documento para documento, mas sua frequência inversa em todos os documentos (*idf*) será um valor absoluto para a coleção de documentos [Man08].

Porém, uma variação no número de documentos observados na coleção (|*D*|+*1*, por exemplo) impactaria o valor encontrado em *idf* para cada os termos, obrigando o recálculo do mesmo, para cada termo em cada documento da coleção *D* [Man08].

Com os valores de *tf* e *idf* calculados, é possível, para o cenário hipotético de uma coleção de três documentos, calcular o valor final de *tf-idf* para cada termo em cada documento. Como exemplo, a fórmula escolhida para o cálculo de *tf* é a natural (*natural term-frequency*), como vista. Na tabela 3, é possível verificar os valores encontrados para *tf-idf* de cada termo (*"mineração"*, *"texto"* e *"carro"*), em cada um dos documentos (*D1*, *D2* e *D3*).

Tabela 3 - Valores de tf-idf para os termos do exemplo, em uma coleção de três documentos.

	Tf-idf		
Termos	D1	D2	D3
mineração	3,24	0	0,36
texto	0	0	0
carro	0	0	37,44

É recomendada, ainda, a normalização dos valores de frequência dos termos em um documento [Man08], uma vez que documentos com muitos termos podem ter um valor de frequência de termos (*tf*) maior que documentos com menos termos, favorecendo um valor maior no resultado do cálculo do *tf-idf*, sem, necessariamente, indicar uma maior relevância. Para isso, utiliza-se a maior frequência local (*tfmax*) - ou seja, o número de vezes que o termo mais utilizado aparece, entre os termos de um documento.

A fórmula da normalização da frequência de um termo é dada por:

$$ntf(t,d) = a + (1-a)\frac{tf(t,d)}{tfmax(d)}$$

Nesta equação *t* é um termo qualquer, *d* é um documento e *a* é um valor de balizamento, com o papel de amortecer a contribuição da segunda parte da fórmula. Este valor deve ser definido entre 0 e 1 [Man08]. Substituindo, então, *tf(t,d)* na fórmula de *tf-idf* por *ntf(t,d)*, obtêm-se a normalização dos valores encontrados para relevância, evitando distorções em documentos com muitos termos.

Com os valores calculados pelo *tf-idf* e normalizados pela fórmula supracitada, cada documento minerado da coleção gerará uma lista de termos mais e menos relevantes de seu conteúdo. Ainda no cenário hipotético, a tabela 4 traz os valores de *tf-idf* normalizados dos termos em relação aos documentos *D1*, *D2* e *D3*.

Tabela 4 - Valores normalizados de *tf-idf* para os termos do exemplo.

	Tf-idf normalizado		
	D1	D2	D3
Tfmax	32	15	120
Termos	-		
mineração	0,87	0,7	0,71
texto	1	0,74	0,71
carro	0,7	0,7	0,89
Para *a = 0.7*			

Como pode ser observado, os valores normalizados de *tf-idf* refletem melhor a relevância de um termo dentro de determinado documento. No caso hipotético, por meio da tabela 4 é possível perceber a proximidade entre o *tf-idf* normalizado do termo “mineração”

no documento *D1* (0,87) e do termo “carro” no documento *D3* (0,89). Na tabela contendo os valores não normalizados (tabela 3), é possível notar uma maior distância entre os respectivos valores (3,24 e 37,44).

A normalização, portanto, é uma prática interessante para o cálculo de relevância de termos através da técnica *tf-idf* [Man08]. Para fins desta pesquisa, o valor de relevância citado nas próximas seções será o normalizado, mesmo que a notação seja *tf(t,d)* (quando do valor de frequência em um documento) ou *tf(t,D)* (quando do valor de frequência global na coleção de documentos). Fica, portanto, implícita a utilização da normalização no cálculo da frequência dos termos.

2.3 Visualização de Informações

Considerada uma área emergente e multidisciplinar, a Visualização de Informações é definida por Nascimento e Ferreira [Nas06] como um campo de estudo da representação visual de informações abstratas, com os objetivos de (1) facilitar a comunicação de informações e (2) auxiliar na exploração e análise de dados.

Compartilhando deste ponto de vista, Stephen Few [Few10] afirma que a representação gráfica de informações abstratas tem dois propósitos: comunicar e analisar. Segundo ele, histórias interessantes vivem nos dados, e representá-las visualmente é “uma maneira poderosa de descobrir e entender estas histórias, e comunicá-las para outros”.

Para Nascimento e Ferreira [Nas06], as vantagens da visualização de informações estão na exploração da capacidade visual humana. Segundo eles, é possível condensar uma quantidade maior de dados em uma simples visualização, resumido no ditado popular “uma imagem vale mais do que mil palavras”. Além disso, o processo de visualização envolve o sentido humano com maior capacidade de captação de informações por unidade de tempo. “O sentido da visão é rápido e paralelo”, afirmam Nascimento e Ferreira.

Uma terceira vantagem estaria no reconhecimento instintivo de padrões pelo sistema visual humano. Somos treinados para identificar padrões e formas distintas rapidamente [Nas06], como exemplificado na figura 3. Nela, somos capazes de identificar de maneira rápida a forma de uma estrela em meio aos quadrados, e, ao lado, os padrões formados por aproximação.

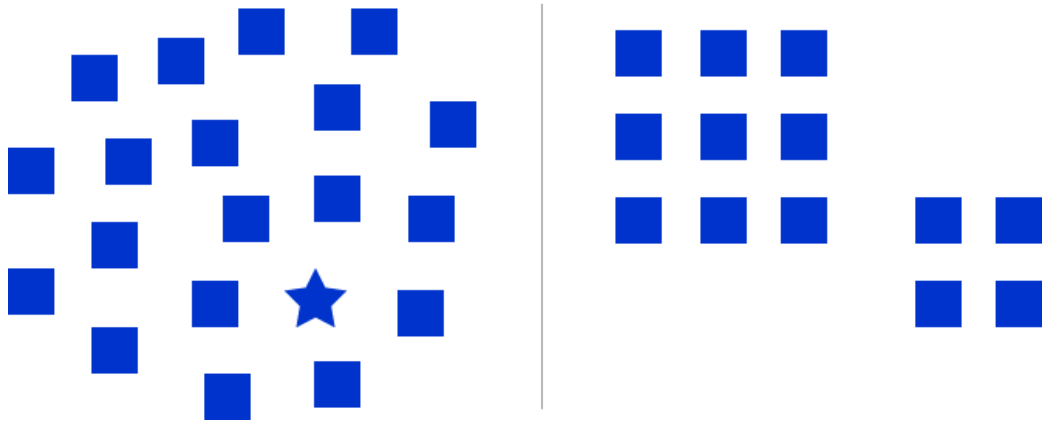

Figura 3 - Imagem que explora a capacidade visual humana na identificação de padrões [Nas06].

Seguindo a mesma linha, Stephen Few [Few10] afirma que um dos pontos mais fortes na visualização está na habilidade humana de processar informações visuais muito mais rapidamente que, por exemplo, verbalmente. “Processamento visual pré-atencional é aquela etapa que automaticamente ocorre no cérebro antes da consciência. Ela consiste de vários estágios, detectando atributos particulares da informação visual como comprimento, tamanho, tonalidade e intensidade da cor, ângulo, textura e forma, entre outras” [Few10].

Ainda que a visualização apresente tantas vantagens, o processo de construção visual não é uma tarefa simples. A tradução do abstrato para um modelo com propriedades visuais da física (como forma, tamanho, cor, etc.) só pode acontecer se entendermos a percepção visual e cognitiva do ser humano [Few10]. Afirma ele que, para tornar estas informações visualmente efetivas, devemos seguir princípios do design que são derivados do entendimento da percepção. A Visualização de Informações preocupa-se, então, com a forma como as representações visuais são construídas, compreendendo os princípios do design aplicados, bem como seus elementos e técnicas.

Segundo Steele et al [Ste10], uma boa visualização deve ser inusitada, informativa, eficiente e atraente. Com o intuito de criar uma visualização com tais características, algumas técnicas de devem ser adotadas. O tipo de dados que dispomos para construção visual unido à informação que pretendemos apresentar ou explorar são o que guiarão nossa escolha por técnicas específicas de visualização. *"Nosso questionamento, ainda durante a fase de preparação dos dados, deve guiar a construção do modelo visual. E responder as perguntas que surgem ao observar os dados não deveria ser nosso objetivo principal, mas sim encontrar a melhor maneira de fazê-lo*” [Few10].

Algumas das técnicas interessantes para visualização de proporções, relações e diferenças entre elementos representados visualmente são, respectivamente, o *Treemap*, o Grafo em anel e o Mapa de calor (*Heatmap*).

2.3.2 *Treemap*

Inventado como uma solução para visualizar uma estrutura de diretórios e arquivos [Shn92], a técnica de *treemapping* explora uma visualização recursiva que representa relações de hierarquia. Otimizado para que nenhum espaço seja deixado em branco, o *Treemap* utiliza formas retangulares que, somadas suas áreas, representam o universo visualizado.

O que as áreas dos retângulos representam é uma das propriedades. Uma ordem hierárquica é representada pela contenção de um retângulo por outro maior, como um subdiretório dentro de seu diretório. E uma terceira propriedade surge na utilização de cores na visualização. Enquanto que a tonalidade separa categorias, o brilho e a saturação podem ser explorados na representação de outro valor, como ilustrado na figura 4. Nela, um exemplo ilustrativo de *Treemap* apresenta, na tonalidade das cores, as quatro categorias principais, e no brilho das cores a indicação de, possivelmente, a relevância do que o retângulo representa dentro de uma categoria.

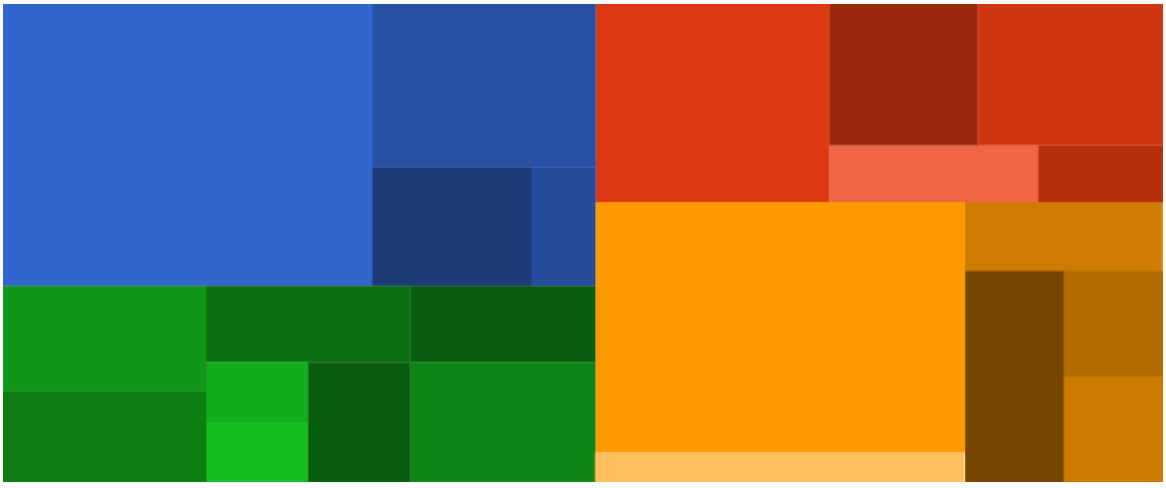

Figura 4 - Exemplo ilustrativo de Treemap (criação do autor).

2.3.3 Grafo em anel

Sugerido por Holten [Hol06] como uma forma de visualizar relações entre itens, o grafo em forma de anel preocupa-se em reduzir, em seu design, a desordem visual que uma visualização de relações entre muitos itens pode gerar. A abordagem é uma interessante forma de visualizar não só a existência de uma relação entre dois itens, mas

também de verificar a intensidade daquela relação, através da espessura das linhas internas [Hol06], como vistas na figura 5.

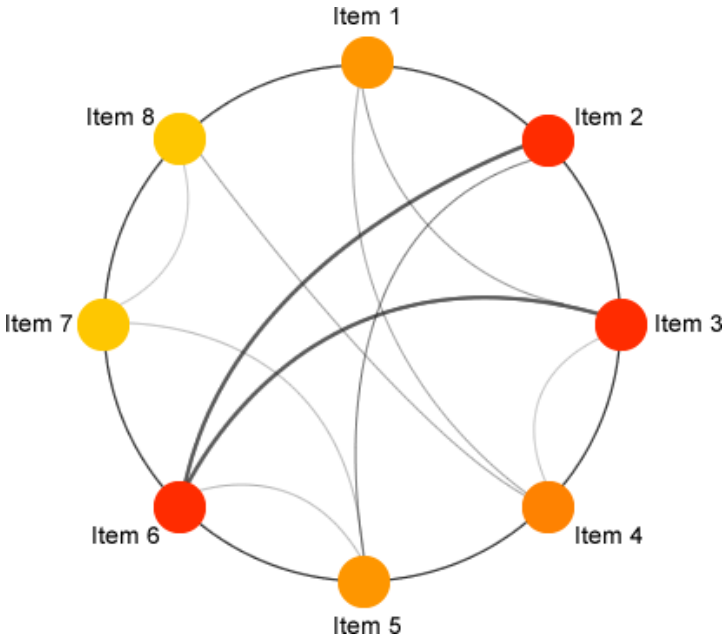

Figura 5 - Exemplo de um Grafo em Anel simples, com 8 nodos (criação do autor).

Colocados sob o círculo, os itens são representados por nodos. No exemplo, são encontrados na forma de círculos menores, mas não há regras para sua representação. A tonalidade das cores dos nodos pode ser utilizada para categorizar ou identificar os itens, ou ainda como a representação de uma propriedade dos itens, através da sua intensidade. No exemplo da figura 5, a tonalidade forte dos itens 2, 3 e 6 poderia indicar um valor maior de uma propriedade em comparação aos outros itens.

A parte principal do grafo em anel está nas relações entre os nodos, representadas por linhas curvilíneas. Cada linha liga dois nodos e sua espessura indica se a relação entre eles é forte ou fraca. O resultado “reduz a desordem visual quando lidando com um grande número de nodos adjacentes”, afirma Holten [Hol06].

A capacidade de representação de informações em um grafo em anel é expandida quando hierarquias e grupos são explorados na representação. Como exemplo, a figura 6 reproduz um grafo em anel onde não apenas as relações entre os nodos são mostradas, mas os itens pertencem a grupos e subgrupos. A utilização das diferentes tonalidades de cor caracteriza cada um dos grupos, e as formas em retângulos curvados separam de maneira clara os subgrupos.

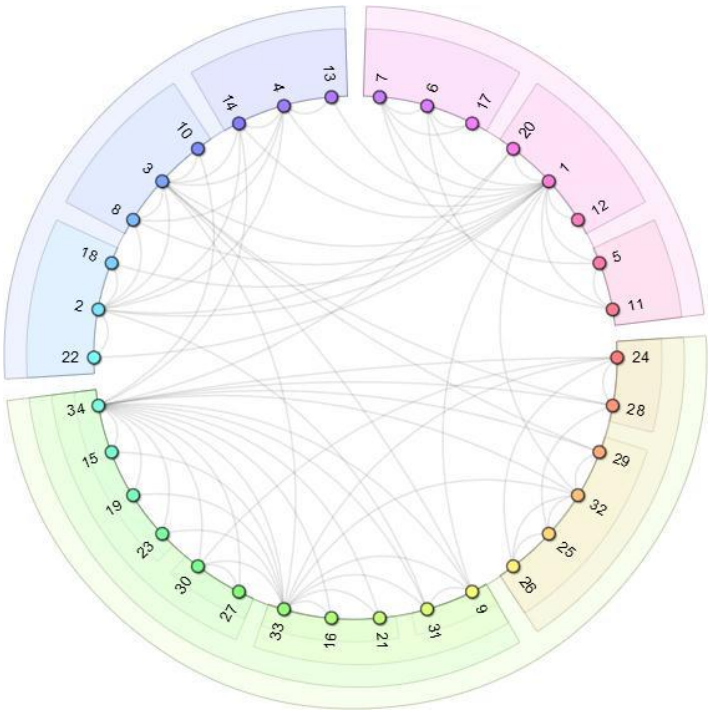

Figura 6 - Exemplo de grafo em anel com grupos e subgrupos (reproduzida de http://scaledinnovation.com/analytics/communities/communities.html).

A utilidade de um grafo que represente grupos é possibilitar a comparação e análise das relações transcendentes aos itens. É possível perceber se um grupo relaciona-se com outro, e complementar a exploração fazendo uma análise mais detalhada entre os nodos internos. A representação de hierarquia e grupos enriquece a visualização, e o grafo em anel facilita o acréscimo desta informação visual sem prejudicar as relações entre os itens [Hol06].

2.3.4 Mapa de calor (*Heatmap*)

Uma das melhores maneiras de comparar é expor todas informações em uma visualização [Yau11]. E é exatamente o objetivo do modelo visual conhecido por mapa de calor (*Heatmap*), que cria uma tabela comparativa entre diferentes itens e suas propriedades. Para facilitar a comparação, ao invés de números, um mapa de calor deve explorar uma graduação entre tonalidades distintas de cores, geralmente duas. A figura 7 ilustra uma comparação de cinco itens e suas propriedades através de um exemplo de mapa de calor.

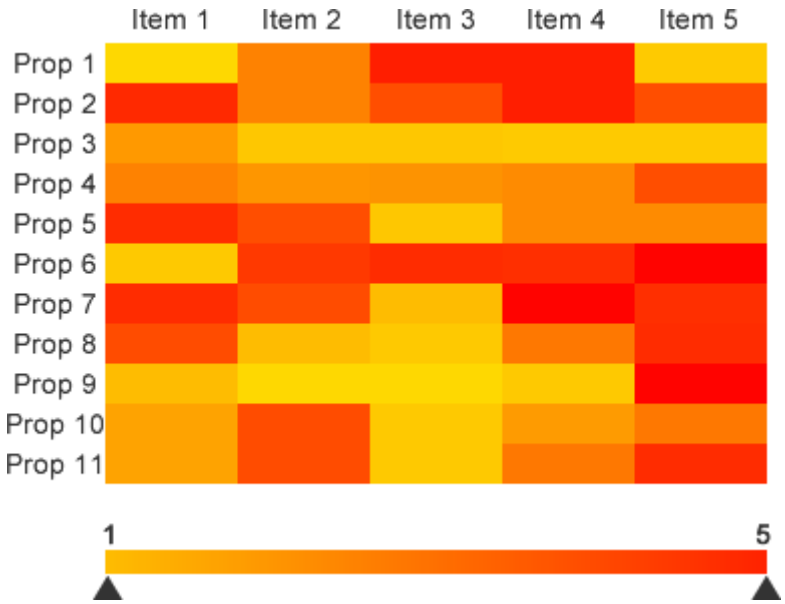

Figura 7 - Exemplo de representação em Heatmap [criação do autor].

A visualização em um mapa de calor facilita a comparação entre os itens, onde a diferença na tonalidade entre os tons de cores representa a distância entre os valores. No mapa de calor da figura 7, vemos como o item 5 mostra valores mais altos (em vermelho ou mais escuro) para a maior parte das propriedades, contrastando com as propriedades inferiores (em amarelo ou mais claro) do item 3. Mesmo não tendo acesso a valores exatos, para uma comparação rápida e eficiente de muitos itens e propriedades, o mapa de calor pode ser útil [Yau11].

3. REVISÃO SISTEMÁTICA DA LITERATURA

Definida como uma forma de identificação, avaliação e interpretação de trabalhos relevantes para uma determinada questão de pesquisa [Kit04], a Revisão Sistemática da Literatura provém mecanismos para identificar e agregar evidência de pesquisa na Engenharia de Software. É uma revisão que tem por objetivo prover uma completa e justa avaliação do estado da evidência relacionado a um tópico de interesse.

A revisão sistemática pode ser vista como uma metodologia específica de pesquisa, que obedece a uma bem definida e estrita sequência de passos metodológicos, em acordo com um protocolo aprioristicamente desenvolvido [Bio05]. Tal protocolo deve conter uma formulação da questão central, o foco da pesquisa, as bases de conhecimento utilizadas e critérios de filtragem na seleção dos trabalhos, entre outras definições.

Uma Revisão Sistemática envolve uma série de etapas, realizadas em três fases da revisão: planejamento, condução e síntese [Kit04]. As etapas associadas a cada uma das fases de uma revisão podem ser separadas como mostra a tabela 5.

Tabela 5 - Fases e etapas de uma Revisão Sistemática da Literatura (adaptado de Kitchenham [Kit04]).

Fase	Etapas
Planejamento	1. Identificação da necessidade de uma Revisão Sistemática; 2. Desenvolvimento do protocolo de revisão.
Condução	1. Identificação da pesquisa; 2. Seleção dos estudos primários; 3. Avaliação da qualidade dos estudos; 4. Extração dos dados; 5. Síntese dos dados.
Síntese	Criação do relatório da revisão

Embora as etapas listadas pareçam sequenciais, muitas delas envolvem iterações, como muitas atividades que são iniciadas durante o desenvolvimento do protocolo e refinadas durante a aplicação da revisão [Kit04].

De maneira similar, Biolchini et al [Bio05] definem o processo de Revisão Sistemática da Literatura como tendo, também, três fases: planejamento, execução e análise dos resultados. Durante o planejamento, os objetivos da pesquisa são listados e o protocolo é desenvolvido. Na execução da revisão, estudos primários são identificados, selecionados e avaliados de acordo com os critérios de inclusão e exclusão pré-definidos no protocolo

de revisão. Uma vez selecionados os estudos relevantes à pesquisa, é realizada a síntese das informações obtidas na revisão durante a fase de análise dos resultados.

No processo, propõe-se iterações entre as fases de uma Revisão Sistemática, abrangendo e tratando possíveis problemas encontrados. "*Antes de executar uma revisão sistemática, é necessário garantir que o plano de revisão seja factível*" [Bio05]. Assim, problemas relacionados a ferramentas de busca de bibliotecas digitais pesquisadas poderiam ser identificados após a execução da revisão, com o resultado da mesma, e remeterem o pesquisador a uma reexecução da revisão antes do início da análise dos resultados.

A figura 8 ilustra as três fases do processo, bem como o fluxo criado pelas iterações entre as fases de planejamento e execução, e entre a execução e análise dos resultados.

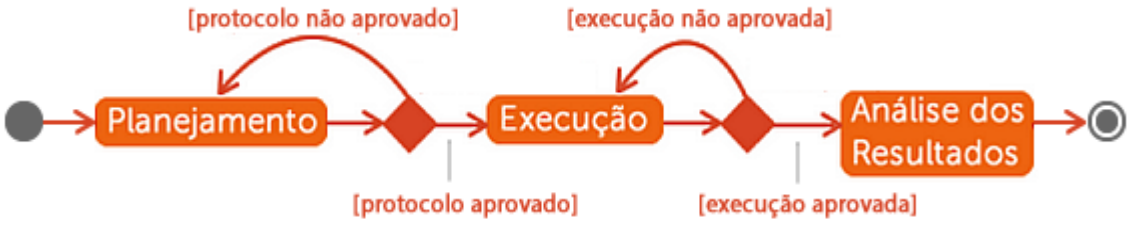

Figura 8 - Processo de Revisão Sistemática, segundo Biolchini et al [Bio05].

Há de se considerar, porém, que as dificuldades dos pesquisadores na realização de Revisões Sistemáticas relatadas na literatura, como na construção da *string* de busca [Bab09] e na seleção dos trabalhos encontrados [Dyb08] [Fel12], não são necessariamente endereçadas pela avaliação do plano de revisão e dos resultados da execução de uma Revisão Sistemática. Como exemplo, a construção da *string* de busca, no processo proposto, pertence ao planejamento de uma Revisão Sistemática, e seria avaliada antes do início da execução da revisão. Porém, é relatado que pesquisadores podem encontrar dificuldades na execução da revisão em razão da utilização de termos não-padronizados [Bab09] em diferentes ferramentas de busca, em um momento posterior à avaliação do plano, como proposta por Biolchini et al [Bio05].

3.1 A abordagem *quasi-gold standard (QGS)*

É na fase de planejamento de uma Revisão Sistemática, quando da elaboração do protocolo de revisão, que uma estratégia de busca deve ser definida [Kit07]. A estratégia de busca - que define os métodos de busca para encontrar os estudos relevantes - é

desenvolvida, por muitos pesquisadores de maneira subjetiva, utilizando seu conhecimento na área e explorando combinações de termos que busquem capturar os conceitos de interesse [Zha11].

Em Zhang [Zha11] é sugerida uma abordagem para o desenvolvimento da estratégia de busca - bem como a construção da *string* de busca - denominada *quasi-gold standard* (*QGS*). O conceito de *quasi-gold standard* define um conjunto de estudos conhecidos no domínio específico e reconhecidos pela comunidade. Este conjunto de estudos é, então, utilizado para extração de termos relevantes e validação da *string* de busca. Segundo o autor, os estudos do conjunto *QGS* são selecionados através de uma busca manual do pesquisador, já utilizando seus critérios de inclusão e exclusão definidos e observando as citações dos estudos da área.

Definido o conjunto de estudos do *QGS*, a abordagem para a construção da *string* de busca, segundo Zhang [Zha11], pode ser tanto de forma subjetiva como objetiva. Na sua forma subjetiva, o pesquisador lê todos os estudos do conjunto para encontrar termos relevantes e montar sua *string* de busca, reavaliando-a com o conjunto *QGS* - para validá-la, o pesquisador deve aplicar a *string* de busca em uma biblioteca digital e verificar se os estudos do conjunto estão entre os resultados. Na sua forma objetiva, o autor sugere que sejam utilizadas técnicas da área de Mineração de Texto para extração de termos relevantes do conjunto *QGS* de estudos. Uma maneira é através do cálculo de frequência e relevância dos termos [Zha11], abordado neste documento (seção 2.2.2) e utilizado no método proposto.

4. PROPOSTA DE MÉTODO DE APLICAÇÃO DE TÉCNICAS DE MVT NO AUXÍLIO À REVISÃO SISTEMÁTICA DA LITERATURA

Como visto nas seções anteriores, a Revisão Sistemática da Literatura tem ganhado popularidade desde sua introdução à área de Engenharia de Software [Kit04], mas ainda é tida como um processo custoso e desafiador por pesquisadores que a realizam [Dyb07] [Bab09] [Ria10]. As maiores dificuldades, segundo relatam pesquisadores, estão na etapa de construção da *string* de busca a ser utilizada [Bab09] e na seleção dos estudos primários retornados, quando a execução da Revisão Sistemática retorna um grande número de resultados [Fel12].

Hoje, já é possível encontrar estudos que utilizam o conhecimento de outras áreas no auxílio à identificação e seleção de estudos primários na Revisão Sistemática [Mal07] [Ana09] [Cho11]. São pesquisas que procuram aplicar técnicas de Mineração de Texto em estudos encontrados durante a fase de condução de uma Revisão Sistemática [Ana09] [Tho11], estudam o impacto da Mineração Visual de Texto na seleção de estudos primários por um pesquisador [Mal07] [Fel11b] [Fel12b], ou exploram a construção de uma ferramenta que apoie a realização de uma revisão [Cho11]. Durante a revisão literária dos trabalhos que utilizam técnicas de mineração no auxílio à Revisão Sistemática, pode-se perceber a falta de estudos que abordem o auxílio na construção da *string* de busca de uma revisão, tanto na aplicações de técnicas na forma de um experimento quanto na criação de uma ferramenta.

As técnicas encontradas durante a revisão bibliográfica realizada nos trabalhos supracitados, sugerem sua utilização dentro da fase de condução/execução de uma Revisão Sistemática, mas aplicam diferentes abordagens em etapas dentro desta fase. Um esquema adaptado das etapas de uma Revisão Sistemática [Kit04] pode ser encontrado na figura 9, e melhor representa algumas das técnicas de mineração em relação às etapas da condução de uma Revisão Sistemática.

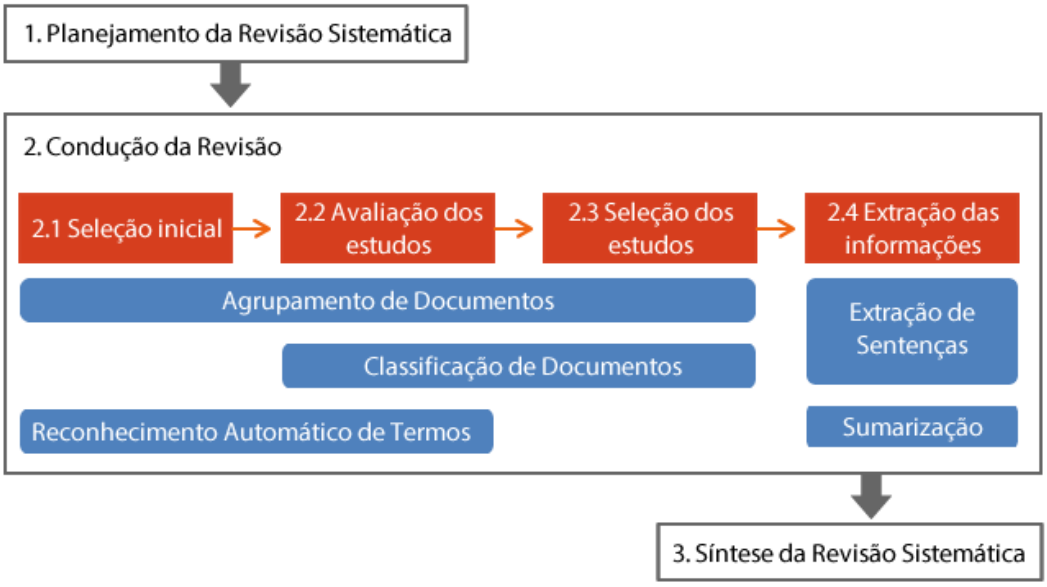

Figura 9 - Processo de Revisão Sistemática e técnicas de mineração por etapa (criação do autor, com base na literatura [Kit04]).

Como pode ser observado na figura 9, estas são técnicas que auxiliam o pesquisador na condução da Revisão Sistemática, e facilitam etapas como a seleção, avaliação e extração de informações dos estudos encontrados. Há de se perceber, porém, que são técnicas que agem exclusivamente na fase de condução da revisão, sem auxiliar o pesquisador a questionar ou propor alterações ao protocolo de sua Revisão Sistemática. Assim sendo, é possível inferir que tais técnicas procuram auxiliar a execução de uma Revisão Sistemática seguindo o resultado da fase de planejamento, sem sugerir alterações no protocolo de revisão do pesquisador. Se uma revisão é executada com uma *string* de busca ineficiente, retornando um grande número de trabalhos, tais técnicas auxiliariam na amortização do esforço de leitura e avaliação dos estudos primários encontrados sem questionar a reconstrução da string de busca ou propor alterações que melhorassem a relevância dos estudos retornados.

No processo de revisão proposto por Biolchini et al [Bio05], iterações ocorrem entre as três fases da Revisão Sistemática, com o objetivo de validar o resultado da fase anterior antes de seguir para a próxima. Um protocolo de Revisão Sistemática seria, então, validado anteriormente à sua execução, através de uma revisão do plano por um pesquisador mais experiente ou de um teste de execução. Não há, porém, neste processo [Bio05], uma iteração que, frente aos resultados obtidos na execução da revisão, leve o pesquisador a rever o plano de sua Revisão Sistemática. Tampouco encontrou-se, na literatura, ferramentas que auxiliem nesta tarefa, particularmente na construção da string de busca a ser utilizada. É, portanto, neste contexto, que surge o método proposto, abordado em maiores detalhes na próxima subseção.

4.1 Método proposto

Neste trabalho é proposto um método que utiliza técnicas de Mineração Visual de Texto no auxílio à construção da *string* de busca de uma Revisão Sistemática da Literatura, baseado no resultado da busca.

O método atuaria entre as fases de planejamento e execução da Revisão Sistemática, permitindo que o pesquisador construa e valide sua *string* de busca, indicando os estudos primários relevantes dentre os encontrados. Uma ferramenta construída para verificação do método utiliza técnicas de mineração para extrair termos dos estudos marcados como relevantes e sugerir melhorias para a *string* aplicada. A figura 10 ilustra a atuação da ferramenta no método proposto, inserida no processo de Revisão Sistemática iterativo proposto por Biolchini et al [Bio05].

Figura 10 - Processo de Revisão Sistemática incluindo o método proposto, adaptado de Biolchini et al [Bio05].

A ferramenta que implementa o método proposto, identificada por *q* na figura 10, age como um verificador da *string* de busca aplicada em relação aos estudos resultantes marcados como relevantes ou não-relevantes pelo pesquisador. Através do método, o pesquisador poderá incorporar melhorias sugeridas pela ferramenta à *string* de busca, retornando à fase de planejamento para atualizar seu protocolo de revisão. Caso não sejam identificadas melhorias, ou não seja da vontade do pesquisador incorporá-las à sua revisão, este poderá prosseguir com a execução da Revisão Sistemática da Literatura sem alterações no processo.

Um fluxograma representando o método proposto em maiores detalhes pode ser visto na figura 11, incorporado às etapas de uma Revisão Sistemática [Kit04]. Nele, é possível perceber a característica iterativa que o método pode assumir para a construção e refinamento de uma *string* de busca.

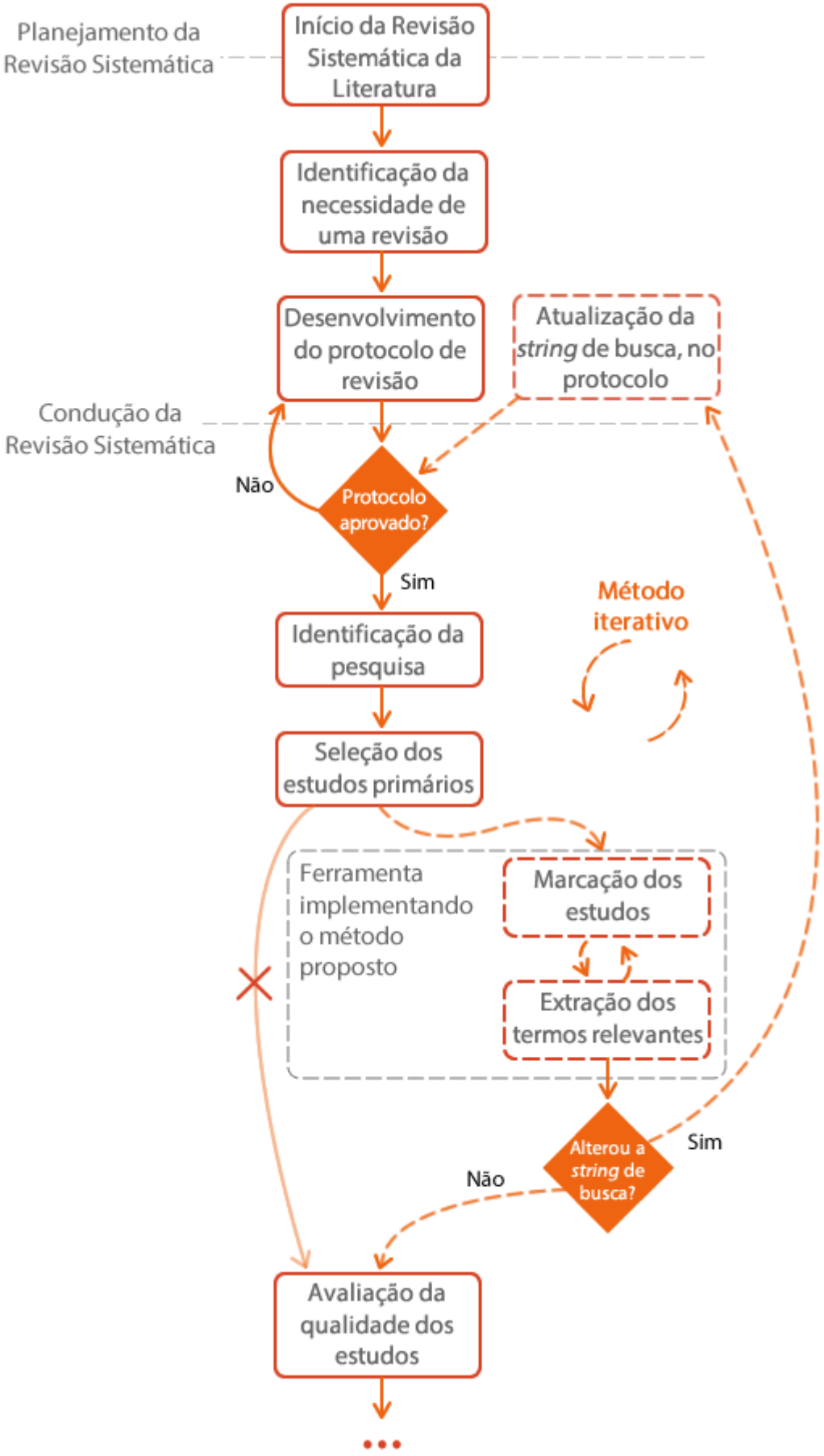

Figura 11 - Fluxograma representando o método proposto incorporado às etapas de uma Revisão Sistemática da Literatura (criação do autor, com base na literatura [Kit04][Bio05]).

4.2 Estudos relacionados

Esta seção tem por objetivo apresentar pesquisas identificadas como relacionadas a este trabalho, abordando temas similares e explorando a utilização de técnicas da Mineração de Texto e Visualização no auxílio à Revisão Sistemática.

4.2.1 Identificando métodos de Mineração de Texto aplicáveis à Revisão Sistemática

Ananiadou [Ana09] traz um extenso estudo sobre a utilização de soluções da Mineração de Texto no aprimoramento do processo de Revisão Sistemática. Na pesquisa desenvolvida em um projeto colaborativo, uma prova de conceito é conduzida sobre estudos primários da área de reabilitação de pessoas com problemas de saúde mental.

Para atingir o objetivo da pesquisa, a metodologia aplicada permitiu o desenvolvimento de protótipos de sistemas de mineração em iterações diretas com o usuário, no ambiente de trabalho do mesmo. Um levantamento rápido de requisitos e opiniões de pesquisadores que utilizariam os sistemas em um ambiente de desenvolvimento ágil são considerados os pontos fundamentais do trabalho realizado durante a pesquisa, nutrindo a colaboração necessária entre os envolvidos [Ana09].

Explorando a utilização das técnicas de mineração em diferentes fases de uma Revisão Sistemática, este trabalho analisa em que etapa cada técnica poderia ser aplicada. Durante a fase de busca em uma revisão, sugere-se a utilização da técnica denominada expansão da *query*, que busca encontrar outros estudos baseado na *string* de busca.

Na fase de filtragem [Ana09] (etapa de seleção de estudos em uma Revisão Sistemática), o ganho poderia vir na aplicação de técnicas como agrupamento de documentos, criando grupos de documentos por tópicos de interesse. Estes grupos corresponderiam, então, a um tópico que é compartilhado por todos os documentos que o contém. A visualização permitiria ao pesquisador enxergar a associação entre os documentos, selecionando tópicos e criando novos subgrupos. Outra técnica aplicável a esta fase da revisão seria a classificação de documentos. Através dela, os documentos minerados seriam automaticamente alocados em categorias existentes, facilitando a seleção dos estudos. Ananiadou sugere, também, que a técnica de classificação poderia ser aplicada sobre cada um dos grupos (*clusters*) criados pela técnica de agrupamento de documentos, estabelecendo a quais categorias os estudos pertencem.

Por fim, a fase de síntese de uma Revisão Sistemática poderia ser auxiliada, pelo uso de técnicas de sumarização de documentos [Ana09]. Tais técnicas permitiriam gerar

um resumo mais informativo de um documento, selecionando e extraindo frases relevantes do estudo para cada seção. Um resumo das técnicas sugeridas em Ananiadou [Ana09] pode ser visto na figura 12, bem como as fases onde seriam aplicáveis no auxílio de uma Revisão Sistemática.

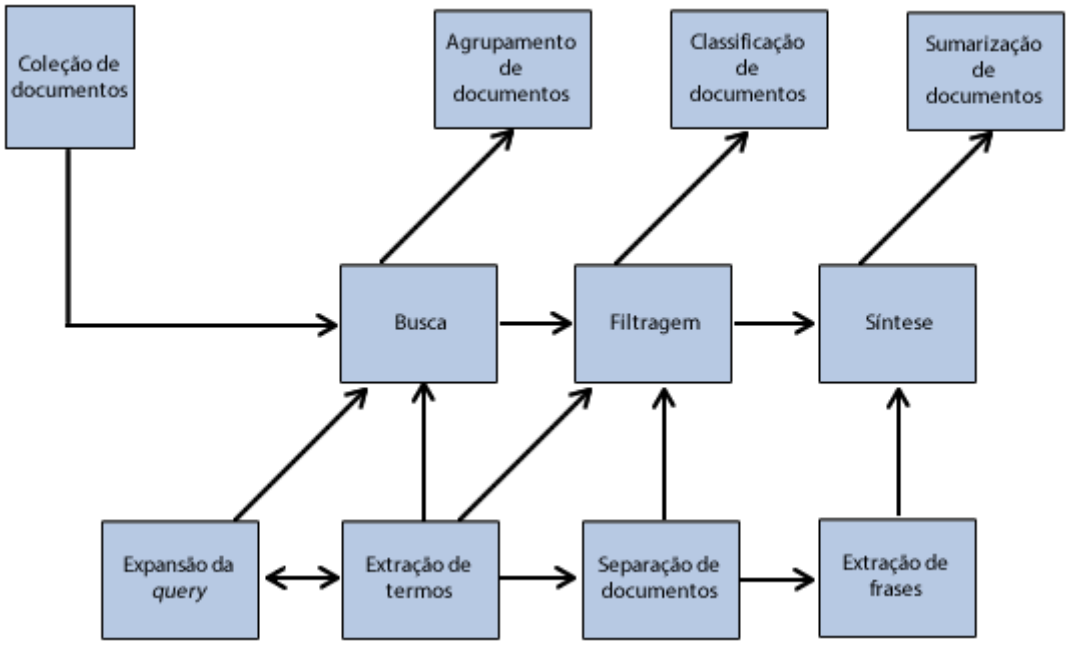

Figura 12 - Técnicas de Mineração de Texto aplicáveis em fases de uma Revisão Sistemática [Ana09].

4.2.2 Aplicação de técnicas de Mineração de Texto em auxílio à Revisão Sistemática

Em um estudo sobre aplicação de técnicas de Mineração de Texto em auxílio à Revisão Sistemática [Tho11], quatro técnicas de Mineração de Texto foram exploradas sobre o processo de Revisão Sistemática, visando a análise da contribuição destas técnicas no auxílio à identificação de estudos primários relevantes em Revisões Sistemáticas, considerando seus pontos fortes e limitações. As técnicas analisadas foram:

- Reconhecimento automático de termos;
- Agrupamento de documentos;
- Classificação;
- Sumarização de documentos.

Os resultados encontrados apontam, segundo o autor [Tho11], que as técnicas aplicadas têm potencial para auxiliar vários estágios do processo de revisão, mas que são relativamente desconhecidas na comunidade de Revisão Sistemática. Considera ele, ainda, que uma avaliação substancial e um maior desenvolvimento dos métodos de mineração em Revisão Sistemática são necessários antes que seu impacto possa ser inteiramente percebido.

4.2.3 Abordagens de Mineração Visual de Textos no auxílio de uma Revisão Sistemática

Em três trabalhos relacionados [Mal07] [Fel11b] [Fel12b], é abordada a utilização de ferramentas de Mineração Visual de Texto na Revisão Sistemática, comparando sua realização com o processo manual de revisão.

No primeiro estudo [Mal07], através da criação de uma hipótese, a qual sugere que a aplicação de técnicas de mineração possa melhorar significativamente o processo de Revisão Sistemática, os autores procuram validá-la, propondo uma comparação entre a realização de revisões com e sem o auxílio de uma ferramenta. A ferramenta utilizada chama-se *Project Explorer* (*PEx*), e é uma evolução do *Text Map Explorer*, desenvolvido pela Universidade de São Paulo (USP). Descrita como uma ferramenta potente e altamente flexível [Mal07], ela possui suporte para manuseio de textos e permite a realização de tarefas da Mineração Visual de Texto, como explorar uma coleção de documentos.

Já em outros dos estudos do grupo [Fel11b] [Fel12b], a hipótese é testada utilizando métricas de desempenho na realização de uma Revisão Sistemática em uma ferramenta desenvolvida pelos próprios autores, denominada ReVis. Tal ferramenta explora técnicas de visualização como grafo em anel (abordado na seção 2.3.3), mostrando as relações entre estudos primários e permitindo ao pesquisador encontrar os trabalhos mais referenciados com maior facilidade. Uma tela da ferramenta desenvolvida pode ser vista na figura 13.

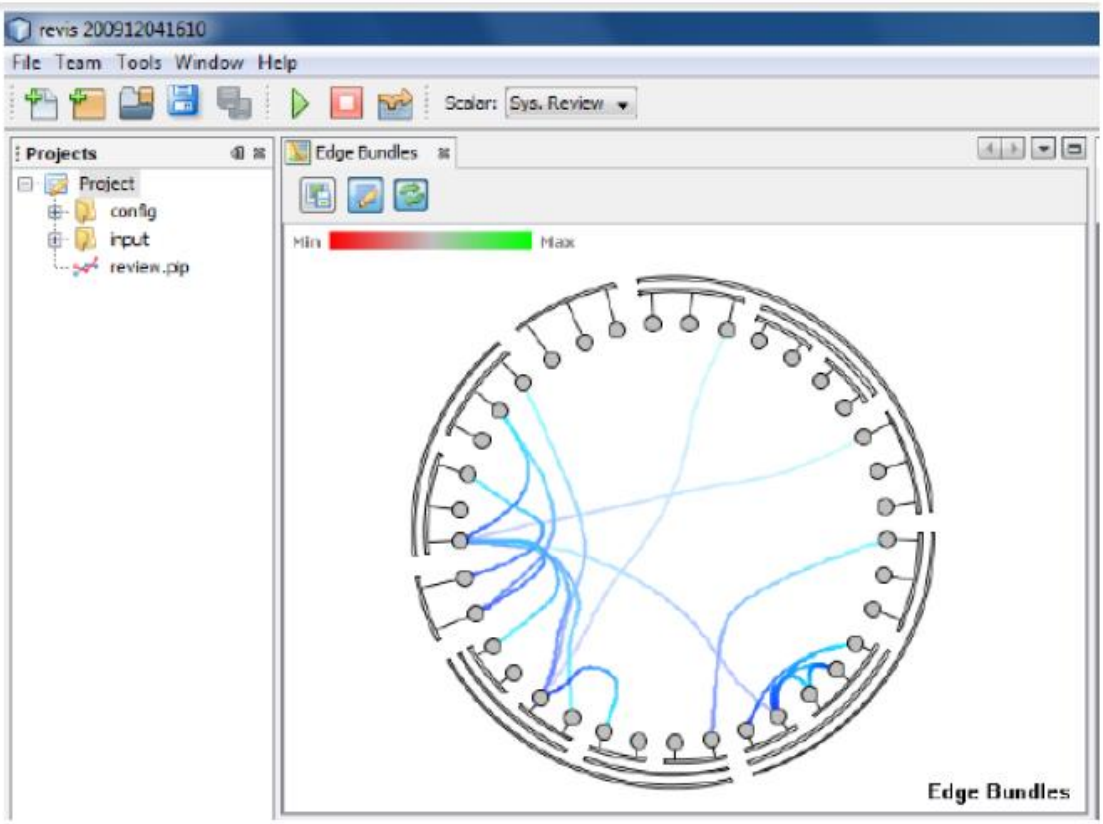

Figura 13 - Tela da ferramenta *ReVis* [Fel12].

Os resultados da pesquisa realizada [Fel12] apontam que a incorporação de técnicas de Mineração Visual de Texto na etapa de seleção de estudos primários da Revisão Sistemática reduziu o tempo gasto na atividade e aumentou o número de estudos corretamente incluídos, segundo o estudo conduzido.

4.2.4 Uma ferramenta de suporte à Revisão Sistemática

Em outro estudo [Cho11], é apresentada uma ferramenta que busca auxiliar a realização de uma Revisão Sistemática utilizando conceitos de Visualização de Informações, como *Radial Space Filling* (*RSF*) e *Treemapping*. Denominada pelos autores de *PaperVis*, a ferramenta aplica auxílios visuais na representação de atributos de estudos primários e suas relações, como referência. Uma de suas telas, mostrando termos-chave relacionados a um estudo, pode ser vista na figura 14. É possível perceber, nela, a utilização de técnicas de Visualização de Informações, como *Treemapping*, ainda que de forma radial, na representação dos termos.

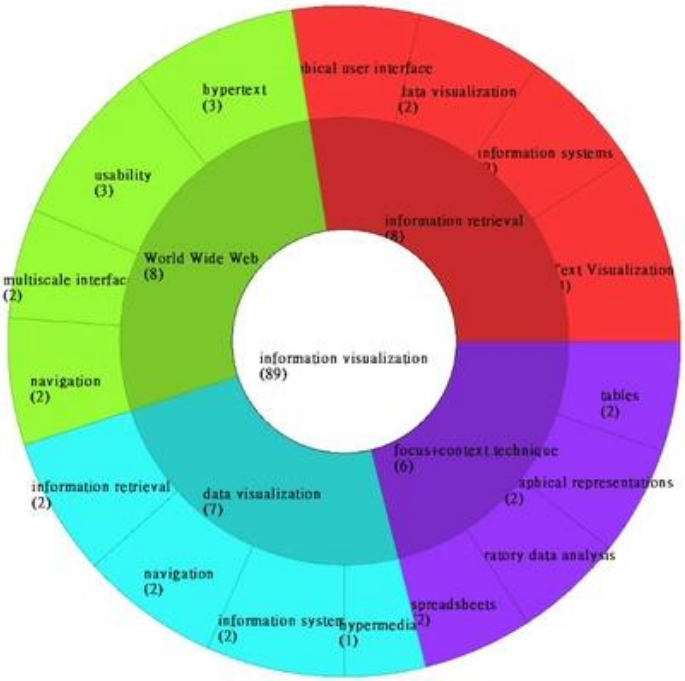

Figura 14 - Representação visual de termos-chave em um estudo [Chou11].

O objetivo da criação de uma ferramenta visual como o *PaperVis* é o de auxiliar tanto pesquisadores novatos como experientes na realização de uma Revisão Sistemática [Cho11]. Para Chou [Cho11], pesquisadores novatos são aqueles que ainda estão por conhecer a área de sua pesquisa, e que, portanto, encontrariam estudos de maior relevância com maior facilidade através de sua ferramenta. Já pesquisadores mais experientes, mais familiarizados com a área de pesquisa, poderiam explorar palavras-chave em determinadas categorias não muito abordadas pela comunidade acadêmica.

4.3 Levantamento das ferramentas utilizadas no auxílio à Revisão Sistemática da Literatura

Em um levantamento realizado por um estudo com o objetivo de identificar e classificar ferramentas que procuram automatizar parte ou todo o processo de Revisão Sistemática da Literatura [Mar13], é possível observar a carência de pesquisas que abordem o auxílio na etapa de planejamento da Revisão. Neste estudo, foi realizado um mapeamento sistemático - uma forma de revisão mais aberta que a Revisão Sistemática, com o objetivo de fornecer uma visão mais geral sobre determinada área ou tópico [Kit12] - e os resultados mostram-se interessantes para o desenvolvimento deste trabalho, justificando sua discussão na presente seção.

Para a realização do mapeamento sistemático, os autores do estudo [Mar13] definiram três questões de pesquisa, bem como critérios de inclusão e exclusão para

auxiliar na seleção dos estudos primários que abordassem ferramentas que auxiliem o processo de Revisão Sistemática. As questões de pesquisa definidas foram:

- QP1) Que ferramentas apoiam o processo de Revisão Sistemática da Literatura na Engenharia de Software foram reportadas?
- QP2) Que fases do processo de uma Revisão Sistemática são endereçados pelas ferramentas encontradas?
- QP3) Até que ponto as ferramentas foram avaliadas?

Ao final, foram selecionados 14 estudos abordando ferramentas conforme os critérios definidos [Mar13]. A tabela 6 traz uma relação dos estudos com o ano de publicação e sua referência. Alguns dos quais, por terem maior relação com o que foi desenvolvido no presente trabalho, foram abordados na seção 4.1.

Tabela 6 - Relação de estudos compreendendo ferramentas que auxiliam no processo de Revisão Sistemática, conforme resultado de um mapeamento sistemático (reproduzido de [Mar13]).

Estudo	Ano	Primeiro autor	Referência
ES01	2007	Malheiros	[MAL07]
ES02	2010	Felizardo	[FEL10]
ES03	2011	Felizardo	[FEL11b]
ES04	2012	Felizardo	[FEL12b]
ES05	2010	Fernández-Sáez	[FER10]
ES06	2007	Cruzes	[CRU07b]
ES07	2007	Cruzes	[CRU07a]
ES08	2012	Torres	[TOR12]
ES09	2012	Bowes	[BOW12]
ES10	2011	Felizardo	[FEL11a]
ES11	2012	Sun	[SUN12]
ES12	2011	Tomassetti	[TOM11]
ES13	2012	Hernandes	[HER12]
ES14	2012	Ghafari	[GHA12]

Ainda no mapeamento sistemático, os autores do levantamento [Mar13] realizaram uma análise das técnicas abordadas pelas ferramentas dos estudos encontrados. São técnicas das áreas de Visualização, Mineração de Texto, Mineração Visual de Texto, Ontologia e outras. A tabela 7 traz uma relação entre os estudos e as técnicas abordadas, bem como o total de estudos abordando determinada técnica.

Tabela 7 - Relação de estudos e técnicas abordadas (adaptado de [Mar13]).

Técnica abordada	Estudos	Total
Visualização	ES01; ES02; ES03; ES04; ES06 e ES10	6
Mineração de Texto	ES01; ES02; ES03; ES04; ES05; ES07; ES08 e ES12	8
Mineração Visual de Texto (MVT)	ES01; ES02; ES03 e ES04	4
Ferramentas que suportam todo o processo de Revisão Sistemática	ES05; ES09 e ES13	3
Ontologia	ES11	1
Ferramenta de busca	ES14	1

Segundo os autores [Mar13], as áreas que apresentaram um maior número de estudos sobre ferramentas abordando técnicas daquela área foram as de Mineração de Texto (8 estudos) e Visualização (6 estudos). Quanto à Mineração de Texto, as ferramentas mais citadas utilizando técnicas de sua área foram a ferramenta *Project Explorer* (*PEx*) [Mal07] e *ReVis* [Fel12b] - ambas referenciadas na seção 4.1. Em relação à utilização de técnicas da área de Visualização, além das ferramentas *Project Explorer* (*PEx*) e *ReVis* - já citadas e que também utilizam técnicas de Visualização - são citadas as ferramentas *HCE* (*Hierarchical Cluster Explorer*), uma ferramenta para identificação de padrões em *data-sets* multidimensionais [Cru07b], e *PEx-Graph* [Fel11b], uma extensão da ferramenta *PEx* [Mal07] que adiciona suporte a representações gráficas. Por abordarem técnicas de Mineração Visual de Texto, estes estudos encontram-se relacionados, também, nas áreas de Mineração de Texto e Visualização.

Outras ferramentas encontradas através do mapeamento sistemático realizado por Marshall et al [Mar13] objetivam apoiar o processo de Revisão Sistemática da Literatura como um todo (denominada *StArt*) [Her23], criar uma ontologia baseada em evidência para suporte de revisões do tipo (denominada *SLRONT*) [Sun12] ou apoiar a busca de estudos [Gha12].

Uma síntese das ferramentas encontradas [Mar13] pode ser visto na tabela 8, assim como os estudos que as abordam.

Tabela 8 - Síntese das ferramentas encontradas através do mapeamento sistemático [Mar13].

Ferramenta	Estudos	Total
Project Explorer (PEx)	ES01; ES02; ES10	3
ReVis	ES03; ES04	2
SLR-Tool	ES05	1
Hierarchical Cluster Explorer (HCE)	ES06	1
Site Content Analyzer	ES07	1
UNITEX	ES08	1
SLuRp	ES09	1
SLRONT	ES11	1
StArt	ES13	1
DBpedia	ES12	1
Sem nome	ES08; ES12	2

Ainda no mesmo estudo [Mar13], é feita uma análise das fases do processo de Revisão Sistemática da Literatura que são endereçados pelas ferramentas. Dos estudos encontrados (14), segundo Marshall et al, um total de 11 estudos apresentam ferramentas que endereçam etapas da fase de condução da Revisão Sistemática, e 3 estudos que apresentam ferramentas que endereçam o processo de Revisão Sistemática como um todo. A tabela 9 traz uma relação dos estudos e as fases e etapas de uma revisão deste tipo.

Tabela 9 - Relação de estudos sobre ferramentas de auxílio à Revisão Sistemática e as fases e etapas da revisão que são endereçadas (reproduzido de Marshall et al [Mar13]).

Fase	Etapa	Estudos	Total
Planejamento	Identificação da necessidade de uma Revisão	-	0
	Desenvolvimento do protocolo	-	0
Condução	Identificação da pesquisa	ES14	1
	Seleção de estudos	ES01; ES03; ES04; ES11; ES12	5
	Avaliação da qualidade dos estudos	-	0
	Extração de dados	ES02; ES08; ES011	3
	Síntese de dados	ES02; ES06; ES07; ES10	4
Síntese	Criação do relatório da Revisão	ES10	1
Todo processo de Revisão Sistemática da Literatura		ES05; ES09; ES13	3

Dos estudos que endereçam a fase de condução da Revisão Sistemática, segundo o estudo de Marshall et al [Mar13], a etapa de seleção dos estudos é a mais comumente endereçada. São cinco os estudos que endereçam esta etapa, dos quais três apresentam ferramentas que utilizam técnicas da Mineração Visual de Texto (MVT). As ferramentas são: *PEx* [Mal07], *PEx-Graph* [Fel11b] e *ReVis* [Fel12b].

Atualmente, a tarefa de construção da *string* de busca, parte da etapa de desenvolvimento do protocolo da Revisão Sistemática, não apresenta ferramentas que

auxiliem o pesquisador, segundo o mapeamento sistemático de Marshall et al [Mar13]. Das ferramentas que buscam auxiliar o processo de Revisão Sistemática como um todo, *SLR-Tool* [Fer10], *SLuRp* [Bow12] e *StArt* [Her12] são ferramentas que auxiliam na organização das atividades presentes nas fases de planejamento, condução e síntese da Revisão, salvando informações. Há, portanto, uma oportunidade observada para o desenvolvimento de um trabalho como o presente, explorando técnicas da Mineração Visual de Texto antes da fase de condução da Revisão Sistemática, no auxílio à construção da *string* de busca.

4.4 Entrevistas com pesquisadores

Tendo sido feita uma revisão literária sobre dificuldades dos pesquisadores na realização de Revisões Sistemáticas [Dyb07] [Bab09] [Ria10] [Fel12], percebeu-se a oportunidade de verificação das dificuldades apontadas na literatura em um universo de pesquisa mais próximo ao candidato. Com isso, pretende-se identificar dificuldades dos pesquisadores diretamente ligadas à construção da *string* de busca para fundamentar a necessidade de uma ferramenta que auxilie nesta tarefa da Revisão Sistemática da Literatura.

Para a realização das entrevistas, foi criado um roteiro visando captar a experiência de pesquisadores na realização de Revisões Sistemáticas e focando nas dificuldades encontradas. Além da identificação do entrevistado, o roteiro de entrevista continha seis questões abertas, como apresentado na Tabela 10.

Tabela 10 - Roteiro de questões aplicadas na entrevista com pesquisadores.

	Questões
Q1.1	Já aplicou uma Revisão Sistemática? Quantas vezes?
Q1.2	Já iniciou Revisões Sistemáticas que não terminou?
Q1.3	Já teve o resultado de uma Revisão Sistemática avaliado de alguma forma (por meio de banca, artigo, professor ou outro)?
Q1.4	Já teve uma Revisão Sistemática publicada?
Q1.5	Como aprendeu a aplicar Revisões Sistemáticas?
Q1.6	Na sua opinião, qual a maior dificuldade na realização de uma Revisão Sistemática? Em qual momento de sua aplicação?

O caráter aberto das questões trabalhadas na entrevista e o roteiro semiestruturado permitiram, ao entrevistador, aprofundar-se nas dificuldades experienciadas pelo

entrevistado, possibilitando uma exploração mais completa das informações desejadas [Hag03].

4.4.1 Perfil dos entrevistados

Em relação ao perfil dos entrevistados, todos pesquisadores selecionados são pós-graduandos que tenham realizado uma Revisão Sistemática da Literatura ao menos uma vez.

4.4.2 Realização das entrevistas

As entrevistas foram agendadas e realizadas pessoalmente, de forma presencial ou remota, por meio de ferramentas de comunicação como o *Skype*[1] e o *Google Hangout*[2], dentro da rede de contatos do autor deste estudo. Ao todo, foram sete os pesquisadores que participaram das entrevistas (identificados neste documento por *P1*, *P2*, *P3*, *P4*, *P5*, *P6* e *P7*), e o resultado é abordado na próxima subseção.

4.4.3 Análise do resultado das entrevistas

Em relação à experiência dos pesquisadores entrevistados, apenas dois afirmaram ter realizado mais de uma Revisão Sistemática (*P6* e *P7*), enquanto que a grande maioria aplicou a técnica apenas uma vez. Três entrevistados (*P1*, *P6* e *P7*) afirmaram terem iniciado uma revisão sem conseguirem terminar, por diferentes motivos, os quais foram explorados em maiores detalhes na continuação da entrevista. Em relação à avaliação e publicação de Revisões Sistemáticas, quatro entrevistados (*P3*, *P5*, *P6* e *P7*) afirmaram terem tido o resultado de suas revisões avaliados e dois (*P6* e *P7*) terem tido suas Revisões Sistemáticas publicadas. A figura 15 ilustra, através de um gráfico, a experiência dos entrevistados com Revisão Sistemática da Literatura.

---

[1] http://www.skype.com/pt-br/

[2] http://www.google.com/+/learnmore/hangouts/?hl=pt-BR

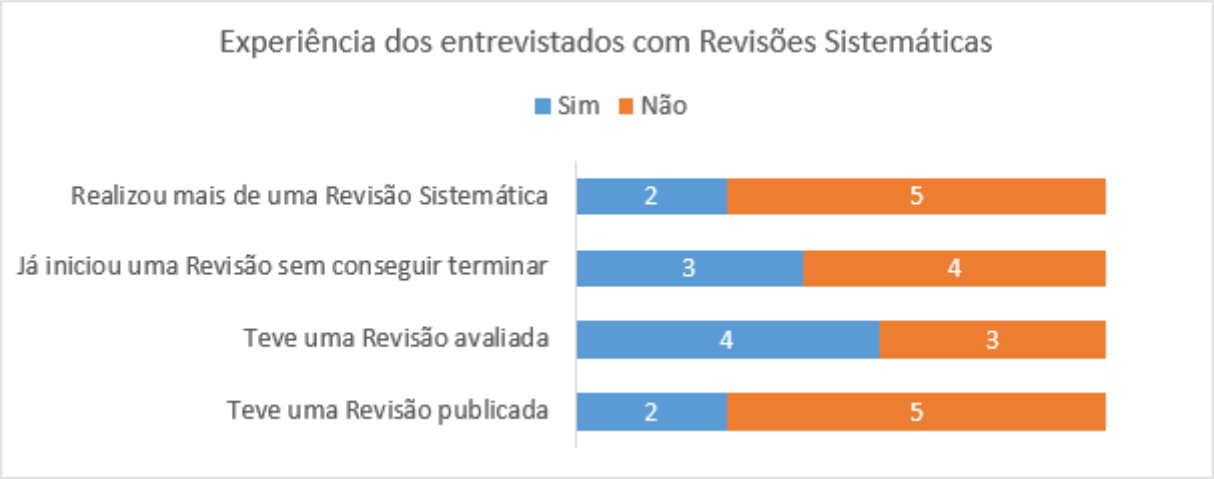

Figura 15 - Análise do resultado sobre a experiência dos entrevistados em relação à Revisão Sistemática.

De maneira geral, as entrevistas mostram o perfil de pesquisadores que tiveram pouco contato com a realização de Revisões Sistemáticas, por diferentes razões. Tais razões devem ser analisadas para compreender se, potencialmente, o método proposto pela presente pesquisa não poderia auxiliar nestes casos. Para tal, é feita uma análise das dificuldades que os pesquisadores enfrentaram na realização de suas Revisões Sistemáticas, através das questões abertas do questionário aplicado.

Analisando as respostas dadas, é possível identificar uma tendência na dificuldade dos pesquisadores na realização de Revisões Sistemáticas da Literatura, indo de encontro ao que é discutido em alguns dos trabalhos revisados durante a fundamentação teórica deste estudo [BAB09] [DYB08]. Em algumas entrevistas, são citados problemas especificamente na construção da *string* de busca. Para a questão *Q1.2*, indagado se já teve uma Revisão Sistemática iniciada sem conseguir conclui-la, determinado entrevistado (*P7*) afirmou que deixou de concluir uma Revisão Sistemática em função de uma *string* de busca mal construída, trazendo muitos resultados. "*[...] Talvez seja por ser uma Revisão em um domínio não conhecido, não pude definir a área e melhorar a string de busca*", considera o entrevistado.

Em resposta à questão *Q1.3*, sobre a avaliação de uma Revisão Sistemática realizada, um dos entrevistados (*P3*) destacou que teve o resultado de uma Revisão Sistemática realizada avaliado por banca de professores, e, em função do grande número de estudos resultantes, teve seu protocolo de revisão questionado.

Já na última questão da entrevista (*Q1.6*), abordando as dificuldades dos pesquisadores na realização da Revisão Sistemática da Literatura, foram cinco os

entrevistados que afirmaram terem encontrado dificuldades na etapa de planejamento da mesma, mais precisamente na construção da *string* de busca a ser utilizada.

Um dos entrevistados (*P1*) afirma que sua maior dificuldade na realização de Revisões Sistemáticas está na definição das palavras-chave da *string* de busca, e destaca a importância desta etapa: "*[...] Porque, a partir delas se consegue os resultados. Se for escolhido um grupo de palavras não-representativo para aquela área, não se conseguirá bons resultados. Logo, o difícil é encontrar as palavras certas para o que se procura [...]*". Ele afirma ainda que, na realização da Revisão Sistemática, teve que definir sua *string* de busca "*[...] por tentativa e erro, entrando diretamente na biblioteca digital, realizando a pesquisa e observando os resultados*". Segundo ele, já na etapa de condução da sua Revisão Sistemática, viu-se obrigado a redefinir as palavras-chave de sua *string*, retornando à etapa de planejamento.

Experiência similar foi relatada por outro entrevistado (*P6*). Afirmando que a construção da *string* de busca é, de fato, uma das maiores dificuldades observadas na realização de Revisões Sistemáticas, o entrevistado cita que perde-se muito tempo em função de *strings* de busca mal definidas: "*[...] Perdi umas três semanas de trabalho, certa vez, por causa de um termo equivocado na* string. [...] *Montei a* string*, retornei os artigos e vi que pouquíssimos artigos tinham relação com a Revisão proposta. Voltei às perguntas para, então, olhar para a* string *e refazê-la*".

Dois dos entrevistados (*P2* e *P3*) citaram outro problema, segundo eles, também decorrente da não utilização de uma *string* de busca apropriada, como indicado por um deles (*P2*) em sua resposta a última questão (*Q1.6*): "*A maior dificuldade que tive foi encontrar os termos certos. Não só encontrar, mas defini-los para que minha* string *de busca não gerasse um volume grande de artigos não-relevantes. Se a* string *de busca usa termos muito genéricos, encontrará estudos que não são do meu interesse [...]*", afirma. Este mesmo entrevistado (*P2*) indica ainda que teve dificuldades em relação à terminologia, durante a construção da *string*: "*[...] Outro problema que encontrei é o fato de a terminologia não ser padronizada. Cabe ao pesquisador conhecer a terminologia da área para realizar uma Revisão Sistemática. Sem conhecê-la, o pesquisador sofre para montar a* string*. Existem, também, termos que têm significados diferentes para diferentes pesquisadores, em diferentes estudos. [...]*", observa. A utilização de termos não-padronizados é identificada como sendo um dos problemas que levam a dificuldades na execução de Revisões Sistemáticas [Bab09], indo de encontro com a resposta do entrevistado (*P2*) supracitado.

Observando-se as respostas das entrevistas, não há dúvidas, entre os entrevistados, sobre a importância da estratégia de busca de uma Revisão Sistemática. Porém, nas palavras de um dos entrevistados (*P3*) para a mesma questão (*Q1.6*), é ressaltada também a importância da análise dos estudos encontrados: "*[...] Acho que o processo de busca é fundamental, mas não é só isso. A análise é, também, importante*".

Um gráfico com uma síntese das maiores dificuldades indicadas pelos entrevistados pode ser visto na figura 16, bem como o número de entrevistados que mencionaram aquela dificuldade.

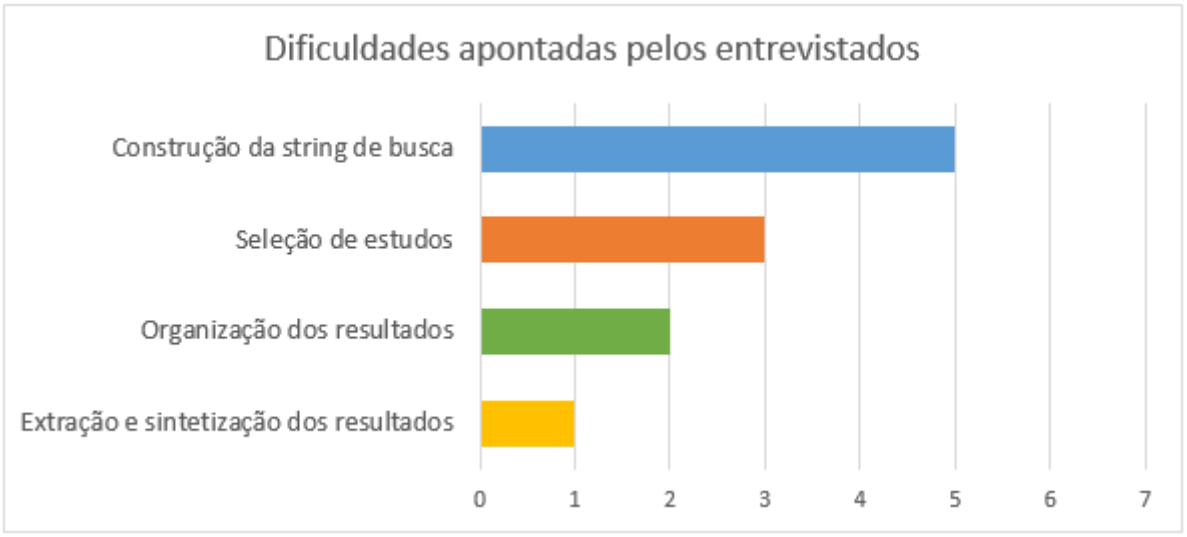

Figura 16 - Gráfico ilustrando as maiores dificuldades citadas pelos entrevistados, em relação ao número de entrevistados que a mencionaram.

5. ANÁLISE DO MÉTODO PROPOSTO

Esta seção aborda a análise do método proposto, descrevendo a ferramenta desenvolvida para viabilizar esta análise, denominada *SLR.qub* (*Systematic Literature Review - query builder*), bem como os testes realizados com pesquisadores e discussão dos resultados obtidos.

5.1 A ferramenta *SLR.qub*

A ferramenta *SLR.qub* é apresentada nas próximas subseções. Primeiramente sua interface e utilização são descritas (5.1.1). Seu funcionamento e técnicas de mineração aplicadas são apresentados na subseção 5.1.2, e as fórmulas de relevância por ela utilizadas estão na subseção 5.1.3. Na subseção 5.1.4 é abordado o formato em que foi desenvolvida, e detalhes de sua implementação (controle de versionamento) e disponibilidade estão na seção 5.1.5. No fechamento da seção (5.1.6), são mencionadas suas limitações conhecidas.

5.1.1 Interface e utilização

A ferramenta *SLR.qub* (*Systematic Literature Review - query builder*) atua minerando os *abstracts* de estudos resultantes da biblioteca digital *IEEExplore*[3], exibindo uma tela por sobre a página de resultados quando pronta. Nesta tela, é exibida uma representação visual de todos os documentos do resultado, em um *carousel* (recurso visual da área de *webdesign*). Um exemplo de tela da ferramenta *SLR.qub* pode ser visto na figura 17.

[3] Biblioteca digital da *IEEE (Institute of Electrical and Electronics Engineers)*.

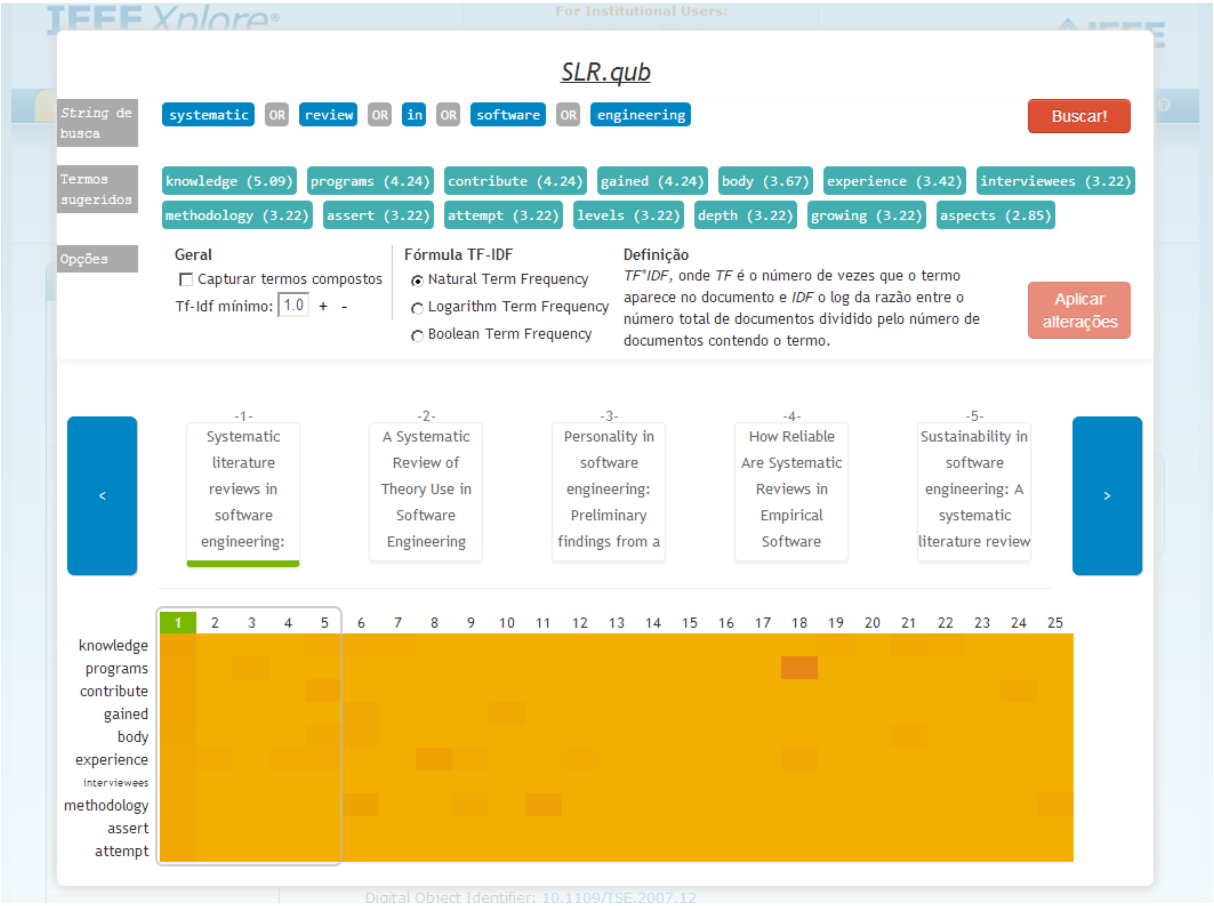

Figura 17 - Tela da ferramenta *SLR.qub*.

Além da coleção de estudos encontrados, a ferramenta exibe a *string* de busca utilizada, com seus termos destacados em azul e operadores lógicos em cinza, como destacado pela figura 18. Ao lado da *string* de busca, um botão com a legenda “Buscar!” indica a ação de efetuar uma nova busca no site *IEEExplore* utilizando a *string* de busca atual. Com a nova página de resultados aberta, porém, é necessário reexecutar o *bookmarklet*, clicando novamente nele.

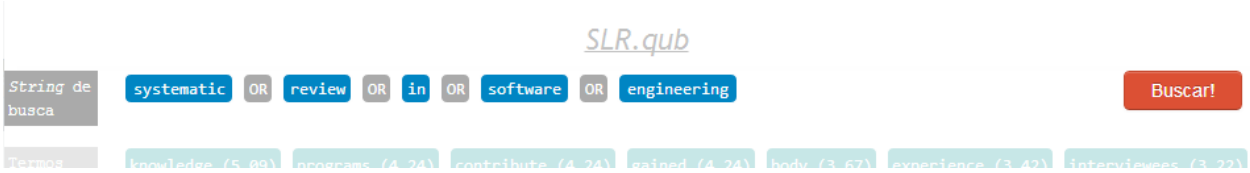

Figura 18 - Área da string de busca da ferramenta *SLR.qub* em destaque.

A área da *string* de busca da ferramenta permite que o usuário altere qualquer termo ou operador lógico utilizado, ou, ainda, adicione termos novos manualmente, mesmo que estes não tenham sido sugeridos pela ferramenta. Para isso, basta clicar em um dos termos ou operadores lógicos, ou espaços em branco ao longo da *string* de busca atual, habilitando a função de edição (como ilustrado na figura 19).

Figura 19 - Edição do termo software da string de busca na ferramenta SLR.qub.

Logo abaixo da *string* de busca, na ferramenta, encontra-se uma área reservada para listagem de termos sugeridos para composição da *string* de busca, baseado na indicação do pesquisador. Uma vez que o objetivo da ferramenta é o de obter termos sugeridos através da marcação de documentos relevantes e não-relevantes pelo pesquisador, esta área é dinâmica e varia de acordo com a interação do pesquisador no *carousel* de documentos, abordado na sequência. Os termos sugeridos nesta área, quando clicados, são adicionados à *string* de busca atual, sendo concatenados com um operador lógico OU (*OR*). Caberia ao usuário, então, editar a *string* de busca para melhor adequar o termo recém adicionado à *string* de busca. Um termo sugerido poderia ser adicionado, também, manualmente. Para tal, somente é necessário que o usuário edite a *string* diretamente e escreva o termo, escolhendo onde inclui-lo na *string*.

Como pode ser observado na figura 20, cada termo sugerido pela ferramenta traz um valor de *tf-idf*, calculado através das técnicas de mineração aplicadas. Este valor, para o usuário da ferramenta, indica a relevância de um termo em relação ao corpo de documentos marcados, e serve para efeitos de comparação entre os termos (para verificar, por exemplo, a distância entre um termo e outro). Na figura 20, é possível perceber que o primeiro termo sugerido (“*knowledge*”), possui um valor de relevância (*tf-idf*) de 5.09, enquanto que os próximos três termos (“*programs*”, “*contribute*” e “*gained*”) possuem o mesmo valor de relevância (4.24).

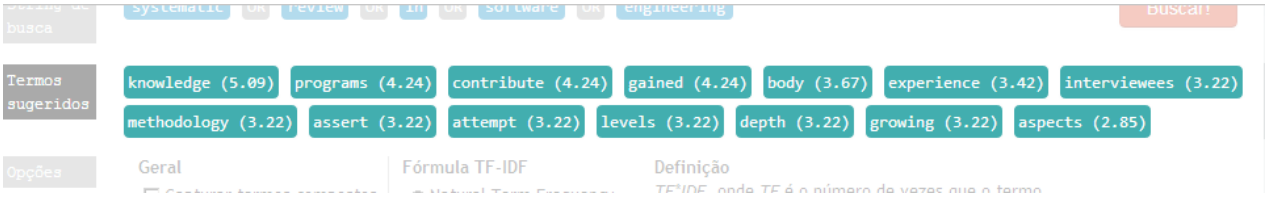

Figura 20 - Área dos termos sugeridos pela ferramenta.

Antes de abordar a marcação de documentos na ferramenta, faz-se necessária uma breve explanação sobre a área de opções da ferramenta, posicionada logo abaixo dos termos sugeridos. Cabe às opções ali presentes (figura 21) a função de possibilitar uma calibragem das técnicas de Mineração Visual de Texto utilizadas pela ferramenta.

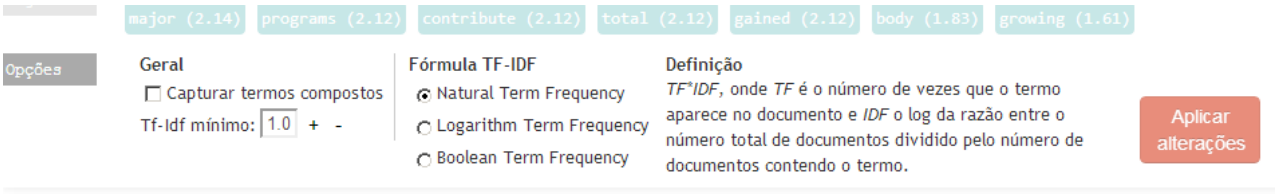

Figura 21 - Área de opções da ferramenta SLR.qub.

A primeira opção, dentro da seção de opções gerais (identificada pelo título *"Geral"*), é a de capturar termos compostos. Se marcada, a ferramenta passa a não utilizar o espaço para separar o texto dos documentos (*abstracts*), capturando termos compostos (como *"text mining"*). A próxima opção é o valor de *tf-idf* mínimo para os termos sugeridos; estando abaixo deste valor, o termo da lista de termos sugeridos será omitido, ajudando a filtragem dos valores de relevância dos termos.

Na outra seção de opções da área estão as fórmula para a frequência dos termos (*tf*), no cálculo do *tf-idf*. Ali, é possível selecionar qual das três fórmulas a ferramenta deve utilizar: natural (*natural term-frequency*), logarítmica (*logarithm term-frequency*) ou booleana (*boolean term-frequency*), como abordado na seção 2.2.2 (*Técnicas da Mineração de Texto*). Uma breve descrição de cada fórmula encontra-se ao lado, auxiliando o usuário a calibrar a ferramenta. Toda alteração nas opções da ferramenta só é aplicada após o usuário clicar no botão da área com legenda *"Aplicar alterações"*. Ao clicá-lo, a ferramenta mantém a *string* de busca e a marcação do usuário em relação aos documentos, atualizando apenas a lista de termos sugeridos com os novos valores de relevância.

A área de marcação dos estudos traz todos os documentos que foram minerados na visualização denominada *carousel*. Nela, o usuário tem a possibilidade de navegar na lista de maneira horizontal, para frente ou para trás, através de setas. A figura 22 mostra o *carousel* exibindo os cinco primeiros estudos resultantes da busca. Um estudo é representado na ferramenta por uma imagem de um documento e seu título. Acima, um índice auxilia o usuário a identificar o estudo na lista de resultados do *IEEExplore*, possibilitando que ele saiba qual estudo é referenciado e o consulte em paralelo.

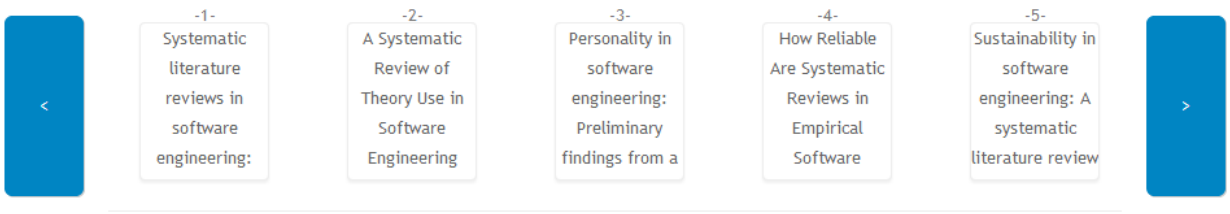

Figura 22 - *Carousel* dos estudos resultantes da busca no *IEEExplore* e minerados pela ferramenta.

Para marcar um estudo como relevante, o pesquisador deve clicar sobre a representação do estudo no *carousel*. No primeiro clique, a cor na parte inferior do documento torna-se verde, indicando que o estudo é considerado relevante para a Revisão Sistemática da Literatura em andamento. Como efeito desta interação, a área de termos sugeridos é atualizada para refletir a lista dos termos mais relevantes, após os cálculos supracitados ainda nesta seção.

Caso queira indicar que um documento não é relevante na revisão, o pesquisador deve clicar novamente sobre um estudo que fora marcado como relevante, no *carousel*. Logo após o clique, a cor passa de verde para vermelho, indicando que aquele documento é considerado não-relevante. A figura 23 mostra a marcação de um estudo como relevante (1) e outro como não-relevante (3).

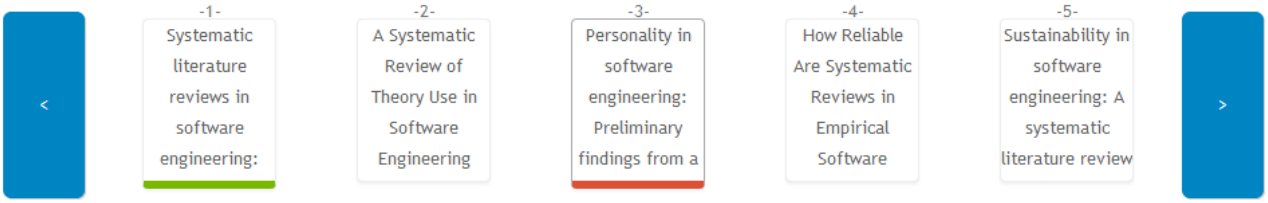

Figura 23 - Exemplo de marcação de estudos na ferramenta *SLR.qub*. Estudo 1 marcado como relevante, e estudo 3 como não-relevante.

A figura 24 ilustra o fluxo de marcação dos estudos. Inicialmente, o estudo encontra-se não-marcado. Quando clicado, muda seu estado para marcado como relevante (indicação em verde). Se clicado novamente, seu estado muda para marcado como não-relevante (indicação em vermelho). Caso seja clicado de novo, sua marcação é removida e ele retorna ao estado inicial. Neste estado (não marcado), um estudo é considerado não-definido, neutro para a análise realizada pela ferramenta, não interferindo nos resultados dos termos sugeridos.

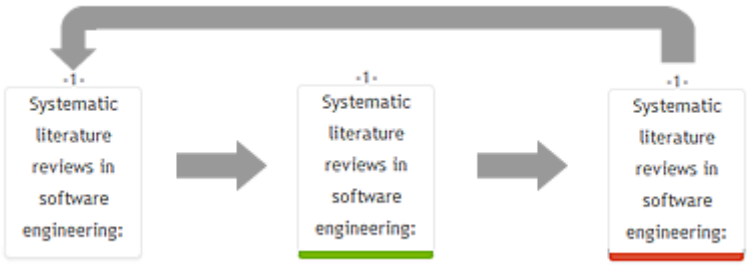

Figura 24 - Fluxo de marcação dos estudos no carousel da ferramenta *SLR.qub*.

Por fim, a ferramenta *SLR.qub* traz uma visualização baseada em mapa de calor (*heatmap*). O mapa de calor da ferramenta tem uma série de funções. A primeira é mostrar, dentre todos os estudos compreendidos, quais estão visíveis no *carousel* para o usuário, situando-o. É possível observar, na figura 25, a janela de cinco estudos que estavam visíveis no *carousel* no momento da captura de tela (*1*, *2*, *3*, *4* e *5*).

Figura 25 - Mapa de calor da ferramenta *SLR.qub*.

Sua segunda função é a de identificar, através das cores verde e vermelho, se um estudo está marcado como relevante, não-relevante, ou se não está marcado. Ainda na figura 25, vê-se a cor verde no estudo 1, identificando a marcação dele como relevante. Da mesma forma, o estudo 3 apresenta-se em vermelho, identificando a marcação dele como não-relevante. Outros estudos, como o 2, o 4, o 5 e ademais, apresentam-se como ainda não marcados, sem cor.

Na sua terceira função, o mapa de calor da ferramenta *SLR.qub* possibilita que o usuário clique em um estudo qualquer e navegue, no *carousel*, até aquele estudo. Como exemplo, na figura 25, clicando-se no estudo 16, tanto o *carousel* navegaria até este e exibiria os estudos 14, 15, 16, 17 e 18 (com o estudo clicado centralizado), como a janela identificada em cinza no mapa de calor estaria envolvendo os mesmos estudos, mostrando os estudos visíveis (primeira função do mapa de calor).

Sua quarta e última função - e a mais importante para o estudo em questão - está na apresentação, em cores, da relevância dos termos sugeridos em relação aos estudos. Como visto na seção 2.3.4, o mapa de calor (*heatmap*) é um modelo visual que expõe informações na forma de uma tabela para comparação através da graduação da tonalidade de cores [Yau11]. Na ferramenta *SLR.qub*, a tabela é construída através da relação entre a lista de termos sugeridos e os estudos compreendidos. Como observado em mais detalhes, na figura 26, a tonalidade mais avermelhada na célula indica uma relevância maior (*tf-idf(t,d)*) de um determinado termo (*t*) naquele estudo (documento *d*).

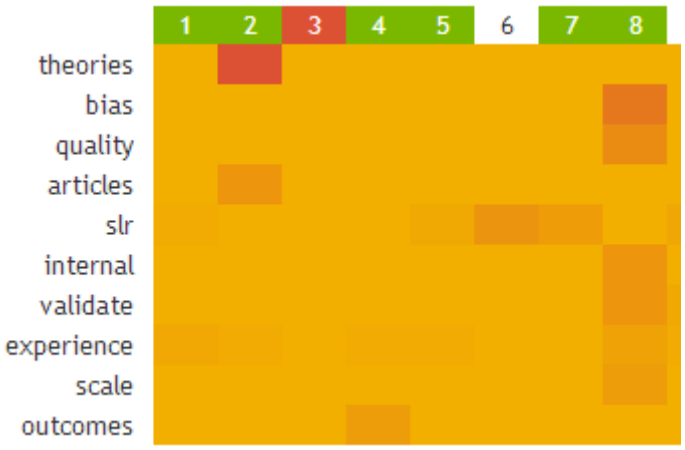

Figura 26 - Mapa de calor da ferramenta *SLR.qub* em detalhes.

É possível perceber que o termo “*theories*”, na figura 26, possui maior relevância no documento 2 que em qualquer outro, e que o termo “*bias*” possui maior relevância no documento 8. É nítida a diferença entre os dois valores quando se observa a tonalidade através do mapa de calor. A visualização desta forma auxilia o pesquisador de duas formas: na busca por estudos através de um termo que é sabido ser relevante para a pesquisa, ou na escolha dos termos que são sugeridos pela ferramenta. No primeiro caso, ainda na figura 26, se soubermos que o termo “*slr*” é relevante para nossa *string* de busca, veremos que o estudo 6 é potencialmente relevante para nós, pois apresenta uma tonalidade mais forte no mapa de calor para este termo. No segundo caso, se ainda não consideramos o termo “*slr*” como relevante para nossa *string* de busca, o mapa de calor o indica como relevante em um conjunto de documentos que marcamos como relevantes. Logo, existe uma evidência indicando que este é um termo a ser ao menos considerado.

5.1.2 Funcionamento e mineração

Como mencionado na subseção anterior, a utilização da ferramenta se dá por sobre uma página de resultados de uma busca no *IEEExplore*. Para acioná-la, é preciso que tal

página esteja aberta, ou o *bookmarklet* avisará que não pôde carregar a ferramenta. Com uma página de resultados aberta, como ilustrado na figura 27, ativa-se o *bookmarklet* criado ao clicar nele. Na figura 27, a área de resultados de uma busca no *IEEExplore* encontra-se destacada.

Figura 27 - Página de resultados do site *IEEExplore*, com a área dos estudos resultantes destacada [Iee13].

Ativado o *bookmarklet* no navegador *web*, cada documento é identificado pelo algoritmo minerador da ferramenta, e salvo em um objeto *javascript* para mineração. Como

preparação do texto para mineração, *stop words*[4] - palavras que aparecem com muita frequência nos documentos, como artigos e pronomes - são removidas do *abstract* de cada documento, além de uma filtragem por pontuação, quebrando o texto em uma lista de termos. Com essa separação, os termos resultantes são, em sua maioria, substantivos compostos (como “text mining”). Isto facilita a identificação dos termos relevantes do abstract, já que não estarão sendo considerados, a priori, artigos, pronomes e outras classes de palavras não relevantes. Por padrão, a ferramenta ainda separa os termos do texto por espaço, resultando em uma lista de substantivos simples (como “text” e “mining”). Nas opções da ferramenta, é possível, ainda, retirar a seleção de quebra por espaços, com a finalidade de obter substantivos compostos.

Como a frequência dos termos é uma métrica-chave na aplicação do cálculo de relevância, espera-se que, com as medidas supracitadas de preparação do documento, sejam identificados termos pertinentes em um *abstract*. Para a aplicação do cálculo de relevância, uma função passa por todos os termos dos *abstracts*, contando com que frequência aparece cada termo e salvando os resultados em um *hashmap*, indexando pelo próprio termo. Com as frequências encontradas, pode-se encontrar o número de documentos que contenham determinado termo ao menos uma vez. Esta medida nos fornece a frequência do termo nos documentos da coleção (*document frequency*, ou *df*), e é parte da fórmula do cálculo de relevância (*tf-idf*). Foi criada, ainda, uma função que normaliza os valores de frequência de termos encontrados (*tf*), assim como a abordada na seção 2.2.2.

Calculado os valores de *tf-idf* para cada termo em cada *abstract* - e armazenados estes valores para consultas futuras - é possível identificar os termos mais relevantes de cada abstract minerado, dependendo apenas da marcação do usuário para que estes termos sejam sugeridos.

5.1.3 Fórmulas de relevância utilizadas

Na ferramenta, um documento pode ser marcado como relevante, não-relevante ou deixado sem marcação pelo usuário. Quando um documento é marcado como relevante, seus termos mais relevantes são adicionados à lista de termos sugeridos, ordenados por seus valores de relevância. Se um termo já encontra-se presente na lista de termos sugeridos, seus valores de relevância (de todos os documentos marcados como relevantes

[4] A lista de *stop words* utilizada foi encontrada em um projeto aberto denominado *stop-words*, hospedado no endereço https://code.google.com/p/stop-words/, e é reproduzida no anexo A deste documento.

onde ele é utilizado) são somados. Após, todos valores de relevância dos termos desta lista são divididos pelo número de documentos marcados como relevantes, resultando na média aritmética para o valor de relevância (*tf-idf*) dos termos. A tabela 11 ilustra este cálculo, exemplificando um cenário hipotético de dois documentos (*D1* e *D2*), cada qual com uma lista de termos ordenados pela relevância (coluna *tf-idf*).

Tabela 11 - Exemplificação da lista global de termos sugeridos em relação à marcação dos documentos como relevantes.

Valor de *tf-idf* nos documentos				Lista global de termos sugeridos			
D1		D2		Apenas D1		D1 e D2	
Termo	***Tf-idf***	**Termo**	***Tf-idf***	**Termo**	***Tf-idf***	**Termo**	***Tf-idf***
vermelho	5.5	vermelho	5.2	vermelho	**5.5**	vermelho	**5.35**
azul	4.2	amarelo	5	azul	**4.2**	amarelo	**4.55**
amarelo	4.1	verde	4	amarelo	**4.1**	azul	**2.1**
		laranja	2			verde	**2**
						laranja	**1**

Quando apenas o documento *D1* é marcado como relevante pelo pesquisador (coluna *Apenas D1* da tabela 11), a lista global de termos sugeridos limita-se aos termos relevantes de *D1*. Porém, quando ambos documentos *D1* e *D2* são selecionados, a lista resultante traz termos sugeridos com valores de relevância equivalentes à média entre as listas dos documentos. Como exemplo, o valor de *tf-idf* do termo "*vermelho*" é a média entre os valores 5.5 (*D1*) e 5.2 (*D2*), resultando em 5.35. Com o cálculo desta média, percebe-se, também, novos termos adicionados à lista global ("*verde*" e "*laranja*"), bem como a alteração na ordem de termos presentes na lista (no caso, "*amarelo*" passando para segundo da lista e "*verde*" caindo para terceiro). É possível, portanto, representar o valor global de um termo *t* como *tf-idf(t,D)*, onde *D* é o universo de documentos minerados.

Da forma como foi desenvolvida a ferramenta, a cardinalidade do conjunto de documentos marcados como relevantes é o denominador utilizado no cálculo da média dos valores de relevância dos termos globais sugeridos. Assim, podemos definir os valores globais de relevância para um determinado termo *t* como na seguinte fórmula:

$$tf - idf(t, D) = \sum_{di \in R} tf - idf(t, di)$$

Nesta fórmula, *R* é o conjunto de documentos marcados como relevantes e |*R*| a cardinalidade deste conjunto.

Utilizando uma lógica similar, a marcação de um documento como não-relevante, pelo usuário da ferramenta, implica na redução do valor de relevância dos termos da lista global de termos sugeridos. Neste caso, porém, não é acionado o cálculo da média, pois o conjunto de documentos marcados como relevantes não foi alterado. Utilizando o mesmo cenário hipotético anterior com a adição de um terceiro documento (*D3*), a tabela 12 exemplifica o cálculo do *tf-idf* para este caso. Quando marcado como não-relevante, os termos mais significativos de *D3* impactam negativamente os valores de relevância da lista de termos sugeridos, como pode ser observado.

Tabela 12 - Valores de *tf-idf* em três documentos e na lista de termos sugeridos.

Valor de *tf-idf* nos documentos						Lista de termos sugeridos			
D1		D2		D3		D1 e D2		**D1, D2 e ¬D3**	
Termo	***Tf-idf***	**Termo**	***Tf-idf***	**Termo**	***Tf-idf***	**Termo**	***Tf-idf***	**Termo**	***Tf-idf***
vermelho	5.5	vermelho	5.2	vermelho	3	vermelho	5.35	vermelho	**3.85**
azul	4.2	amarelo	5	amarelo	2.5	amarelo	4.55	amarelo	**3.3**
amarelo	4.1	verde	4	verde	2	azul	2.1	azul	**2.1**
		laranja	2			verde	2	laranja	**1**
						laranja	1	verde	**0**

É possível adicionar a redução do valor de *tf-idf* global dos termos à fórmula supracitada. Sendo *NR* o conjunto de documentos marcados pelo usuário como não-relevantes e |*NR*| sua cardinalidade, a seguinte fórmula nos daria a soma de todos os valores de *tf-idf* de um termo *t* nos documentos de *NR*:

$$\sum_{dj \in NR} tf - idf(t, dj)$$

Adicionando esta fórmula como parte do cálculo do valor de *tf-idf* global de um termo *t*, tem-se a seguinte fórmula:

$$tf - idf(t, D) = ((\sum_{di \in R} tf - idf(t, di)) - (\sum_{dj \in NR} tf - idf(t, dj)))/|R|$$

Nesta fórmula, *R* é o conjunto de documentos relevantes e *NR* o conjunto de documentos não-relevantes, marcados pelo usuário, e |*R*| e |*NR*| suas respectivas cardinalidades. *D* é o conjunto de todos os documentos e *t* o termo. Esta é a fórmula final para o cálculo de relevância global de um termo, utilizada no algoritmo da ferramenta *SLR.qub* e aplicada a cada marcação de documentos.

É baseando-se em uma interação simples do usuário que a ferramenta *SLR.qub* realiza cálculos que indicam termos que poderiam ser adicionados à *string* de busca de

uma Revisão Sistemática da Literatura, aplicando técnicas de Mineração Visual de Texto em tempo real e exibindo um modelo visual dos resultados que procura auxiliar e facilitar a decisão do pesquisador que a utiliza.

5.1.4 Formato

A ferramenta foi desenvolvida na forma de um *bookmarklet*, ou seja, um código escrito em linguagem *Javascript* que se comporta como uma página guardada nos favoritos de um navegador. Ao acioná-lo, porém, o navegador não abre uma página, mas executa um *script* que atua por sobre a página aberta. Os ganhos com esta abordagem vão desde a possibilidade de executar a ferramenta sobre a página que se encontra aberta no navegador, para leitura de seu conteúdo, à simplicidade com que é acionada, com um clique.

Os *bookmarklets* são utilizados como ferramentas que agem sobre o documento *HTML* e seus elementos, alterando a maneira como a página é exibida, personalizando sua visualização, depurando páginas em desenvolvimento ou extraindo informações de seu conteúdo. Para ser executado, o *bookmarklet* utiliza o protocolo *javascript:* [Kan98]. Em razão disso, é suportado por todos os navegadores que implementam tal protocolo, como o *Microsoft Internet Explorer*[5], *Mozilla Firefox*[6] e *Google Chrome*[7], mesmo em suas versões mais antigas (como o *IE6.0*).

Na figura 28 tem-se uma representação visual do fluxo de execução do *bookmarklet* criado: primeiramente, temos um navegador da *web* com uma página aberta (1), exibindo os favoritos e, entre eles, o *bookmarklet da ferramenta*. Clicando no *bookmarklet* (2), um código em *Javascript* é executado, através do protocolo *javascript:* implementado e suportado pelo navegador (3). Este código, por sua vez, invoca a inclusão e execução de um arquivo com código *Javascript* de um repositório acessível (4), que atua sobre a página aberta no navegador (5), finalizando o fluxo.

[5] http://windows.microsoft.com/pt-br/internet-explorer/download-ie
[6] http://www.mozilla.org/pt-BR/firefox/new/
[7] http://www.google.com/intl/pt-BR/chrome/

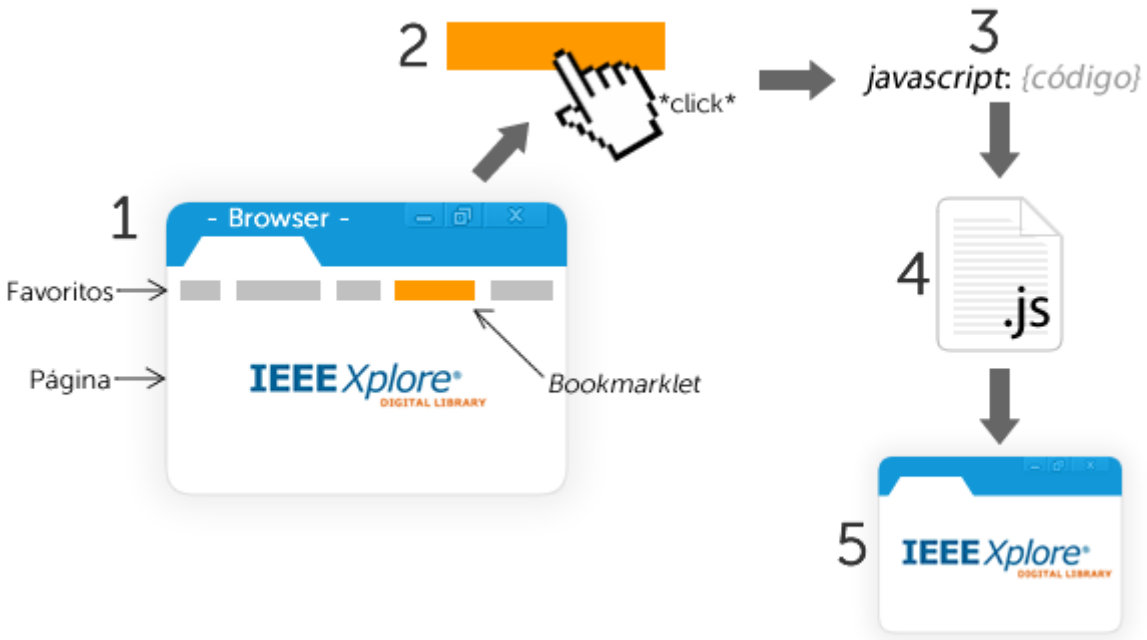

Figura 28 - Esquema representativo do fluxo de acionamento de um *bookmarklet* (criação do autor).

Como o *bookmarklet* utiliza um espaço limitado de caracteres - no máximo 508 caracteres no *Internet Explorer versão 6.0* - uma boa prática no seu desenvolvimento é a de limitá-lo para, ao invés de conter todo o código que será executado, invocar e executar o código de um arquivo remoto contendo *Javascript*. Para que isso seja possível, o *script* invocado deve ser um arquivo acessível na internet. Aplicando essa boa prática, é possível, também, continuar o desenvolvimento e aprimoramento do script invocado, mantendo-o constantemente atualizado; como o acionamento do favorito invoca um arquivo remoto, a versão mais atualizada do script sempre será executada pelo *bookmarklet*, evitando que o mesmo tenha que ser atualizado manualmente [Kan98].

5.1.5 Controle de versionamento e disponibilidade

Com o intuito de manter a ferramenta sempre atualizada e disponível para a comunidade acadêmica, foi selecionado um sistema de controle de versionamento e distribuição para seu desenvolvimento, denominado *Git*[8].

O controle de versionamento em *Git* possui ênfase na velocidade de desenvolvimento e facilidade de uso. Seu modelo distribuído permite que operações possam ser realizadas localmente, e não seja necessário comunicar um servidor central a

[8] http://git-scm.com/

cada alteração. Se comparado ao controle de versionamento *Subversion* (*svn*)[9], para algumas operações comuns como *commit*, *diff* e *update* (comandos), a vantagem da utilização do *Git* é considerável[10].

A utilização de um sistema distribuído de controle de versão permite que cada desenvolvedor tenha uma cópia de todos os arquivos do projeto enquanto trabalha, mantendo múltiplos *backups* destes arquivos. Além disso, o próprio *Git* é distribuído com uma licença de código-aberto, permitindo que qualquer desenvolvedor possa contribuir com sua solução.

Os projetos iniciados com o controle distribuído de versionamento *Git* permitem a contribuição de diferentes desenvolvedores, desde que tenham acesso ao seu repositório. Para facilitar a visibilidade e acessibilidade dos projetos em questão, um serviço de hospedagem online e portal de projetos foi criado, com o nome de *GitHub*[11]. Nele, um usuário cadastrado pode ter seus projetos iniciados e compartilhados com outros usuários. Se públicos, os projetos podem ter a contribuição de um desenvolvedor da comunidade, bastando apenas que este inicie um repositório local (utilizando *Git*) com os arquivos do projeto e realize *commits* (comando para incluir alterações em um projeto), enviando suas alterações para o projeto principal.

No centro da plataforma de hospedagem de projetos e repositórios *GitHub* está sua comunidade ativa, uma verdadeira rede social de desenvolvedores e empreendedores. Fazendo parte dela, é possível receber notícias e acompanhar contribuições de projetos, além de fazer comentários sobre alterações e sugestões de implementação. A ferramenta desenvolvida encontra-se hospedada no *GitHub*, sob o título de *SLR.qub*, e permite a participação de outros membros da comunidade.

5.1.6 Limitações da implementação

Apesar de ferramentas da área de Mineração de Textos geralmente utilizarem algoritmos robustos sobre grandes volumes de dados (textos) [Hea99], a ferramenta construída atua minerando o texto contido no *abstract* de cada um dos estudos resultantes da primeira página de resultados de uma busca na biblioteca digital *IEEExplore.* Por padrão - caso não seja alterado nas opções do site da biblioteca digital - são retornados 25 estudos na primeira página de resultados.

---

[9] http://subversion.apache.org/

[10] http://git-scm.com/documentation

[11] https://github.com/

Mesmo com esta limitação, optou-se pelo desenvolvimento de uma ferramenta que possibilitasse a mineração dos resultados em tempo real e que não impedisse ou deixasse lenta a utilização da mesma pelos pesquisadores participantes do teste neste trabalho.

A opção pela criação de uma ferramenta que funciona sobre resultados da biblioteca digital *IEEExplore* deu-se pelo fato de os *abstracts* dos estudos resultantes, neste site, encontrarem-se presentes e expostos como parte da página (ou seja, como parte do documento *HTML*) e permitirem a mineração sem maior complexidade. Não há impedimentos, porém, na adaptação da ferramenta para diferentes bibliotecas digitais, bastando para tal modificar a forma com que lê a estrutura da página e extrai o conteúdo dos *abstracts*. Como a ferramenta, criada em *Javascript*, escaneia a página aberta com os resultados é necessária a criação de uma rotina de leitura dos estudos para adicionar suporte a bibliotecas digitais que não mostrem os *abstracts* da mesma maneira; o restante da execução da ferramenta permaneceria sem alteração. Uma vez que o projeto da ferramenta encontra-se disponível, é facilitado o acesso de desenvolvedores que queiram contribuir para tal propósito.

5.2 Aplicação de testes com usuários

Com o objetivo de observar a aplicabilidade do método proposto - que usa técnicas de Mineração Visual de Texto no auxílio à construção da *string* de busca de uma Revisão Sistemática, através da utilização da ferramenta *SLR.qub* - foi elaborada uma etapa de testes com usuários. Esta seção aborda os testes realizados e apresenta seus resultados. Uma discussão geral abordando os resultados deste teste pode ser encontrada na seção 5.4.

5.2.1 Descrição do teste

Para realização do teste, os pesquisadores selecionados foram situados no início de uma Revisão Sistemática da Literatura, na etapa de desenvolvimento do protocolo da revisão, quando da construção da sua *string* de busca. O cenário foi descrito antes do início da execução de cada teste, como consta na seção 3 do apêndice A deste documento.

A Revisão Sistemática proposta no teste tinha como objetivo entender qual o estado-da-arte deste modelo de revisão literária na área de Engenharia de Software e quais dificuldades os pesquisadores têm enfrentado na sua adoção. Para tal, foram definidas questões de pesquisa e critérios para inclusão e exclusão de estudos, também presentes na seção 3 do apêndice A.

Realizada a etapa de preparação para o teste, descrita na seção 4.A do apêndice A deste documento, os participantes iniciaram a execução do teste na página de resultados da biblioteca digital *IEEExplore*, acionando a ferramenta através de seu *bookmarklet* (como mencionado na seção 5.1.2). Com a ferramenta aberta, a tarefa dos participantes era a de marcarem estudos resultantes da busca utilizando uma *string* inicial (“*systematic review in software engineering”*) que achassem ser relevantes ou não-relevantes para a Revisão Sistemática proposta no teste. Os critérios para o término do teste, foram: (1) indicação do participante de haver chegado a uma *string* de busca relevante para a revisão proposta ou (2) haverem transcorrido 25 minutos do início da sua execução.

Durante o teste, foram coletadas métricas consideradas interessantes para análise dos resultados. Estas métricas compreendem todo tipo de interação do usuário com a ferramenta, como cliques nos elementos da interface, marcação de estudos ou edição da *string* de busca. Junto à ação do usuário, foi coletado, também, o tempo transcorrido do início do teste e o endereço *IP* dos participantes, para facilitação da análise dos resultados.

5.2.2 Perfil dos participantes

Os participantes do teste deveriam ser pesquisadores familiarizados com a área de Engenharia de Software que já houvessem realizado ao menos uma Revisão Sistemática da Literatura. Ter como pré-requisito para o teste a experiência do pesquisador era crucial para o entendimento e realização da tarefa proposta.

5.2.3 Seleção dos participantes

Optou-se pela realização dos testes com os mesmos participantes das entrevistas, uma vez que já era conhecido o seu nível de experiência em relação à Revisão Sistemática, além de ter sua opinião quanto às dificuldades enfrentadas pelos pesquisadores capturada, conforme abordado na seção 4.4.

5.2.4 Apresentação dos resultados dos testes

Ao todo, foram sete os testes realizados com pesquisadores. A tabela 13 traz uma síntese dos testes realizados, tendo na coluna *P* a identificação do pesquisador (de *P1* a *P7*), na coluna *I* o número total de interações do pesquisador com a ferramenta - onde interação é como encontra-se dito na descrição do teste (seção 5.2.1) - e na coluna *D* a duração, em minutos, de cada teste. Em relação ao resultado dos testes, nas colunas *Relevantes* e *Não relevantes*, são exibidas as identificações dos documentos minerados na primeira página de resultados do *IEEExplore* e marcados pelo pesquisador (de *D1* a *D25*).

O número total de documentos marcados como relevantes e não relevantes é visto em, respectivamente, em *Rs* e *NRs*.

Tabela 13 - Síntese do resultado dos testes.

P	*I*	*D*	Relevantes	Não relevantes	*Rs*	*NRs*
P1	201	0:25	*D1, D2, D3, D7, D8 e D9*	*D4, D12 e D23*	6	3
P2	141	0:25	*D1, D3, D4, D6, D7, D12 e D16*	*D2*	7	1
P3	156	0:25	*D1, D5, D23 e D25*	*D9 e D21*	4	2
P4	322	0:22	*D1, D2, D5, D6, D14, D18, D21 e D23*	*D3, D4, D7, D8, D9, D10, D11, D12, D13, D15, D16, D17, D19, D20 e D22*	8	15
P5	71	0:25	*D1, D4, D6, D12, D15, D21 e D23*	*D3, D8, D9 e D22*	7	4
P6	163	0:25	*D1, D2, D3, D4, D5, D6, D7, D8, D9, D10, D11, D12, D13, D14, D15, D16, D17 e D23*	*D18, D19, D20, D21, D22, D24 e D25*	17	8
P7	279	0:25	*D1, D2, D5, D6, D10, D12, D13 e D25*	*D3, D4, D9, D11, D15, D16, D17, D18, D19, D20, D21, D22, D23 e D24*	8	14

Para uma visualização mais clara de quantas vezes cada documento foi marcado, as marcações dos documentos são apresentadas na tabela 14, de maneira estendida. Nela, os pesquisadores que participaram do teste estão identificados de *P1* a *P7*, na parte esquerda da tabela. Os 25 estudos presentes na primeira página de busca de resultados do *IEEExplore* estão identificados por *1* a *25* no topo da tabela. No corpo da tabela, portanto, encontram-se as marcações de cada pesquisador, onde *r* identifica um documento marcado como relevante, *nr* um documento não-relevante e em branco um documento que não foi marcado pelo pesquisador. A tabela reflete o estado dos documentos marcados ao final dos testes.

As colunas *Rs1* e *NRs1*, na parte de baixo da tabela 14, são referentes ao número de vezes que o documento em questão foi marcado como relevante e não-relevante, respectivamente. Já *Rs2* e *NRs2*, no extremo direito da tabela, identificam o número de documentos marcados como relevantes e não-relevantes, respectivamente, por cada pesquisador. No somatório do resultado dos testes, contabilizando-se todas as marcações efetuadas, houve 57 marcações de documentos como relevantes, 47 como não-relevantes e 71 deixados em branco. Em média, foram marcados 8 documentos como relevantes e 7 como não-relevantes por teste.

Tabela 14 - Visualização das marcações dos documentos nos testes realizados.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	Rs2	NRs2
P1	r	r	r	nr			r	r	r			nr											nr			6	3
P2	r	nr	r	r		r	r					r				r										7	1
P3	r				r				nr												nr		r		r	4	2
P4	r	r	nr	nr	r	r	nr	nr	nr	nr	nr	nr	nr	r	nr	nr	nr	r	nr	nr	r	nr	r			8	15
P5	r		nr	r		r		nr	nr			r			r						r	nr	r			7	4
P6	r	r	r	r	r	r	r	r	r	r	r	r	r	r	r	r	r	nr	nr	nr	nr	nr	nr	nr	nr	17	8
P7	r	r	nr	nr	r	r			nr	r	nr	r	r		nr	nr	nr	nr	nr	nr	nr	nr	nr	nr	r	8	14
Rs1	7	4	3	3	4	5	3	2	2	2	1	4	2	2	2	2	1	1	0	0	2	0	3	0	2	57	-
NRs 1	0	1	3	3	0	0	1	2	4	1	2	2	1	0	2	2	2	2	3	3	3	4	3	2	1	-	47

A frequência de seleção de documentos como relevantes e não-relevantes fica mais clara em um gráfico como o da figura 29, no qual percebemos quantas vezes cada documento foi marcado como relevante, não-relevante ou deixado sem marcação.

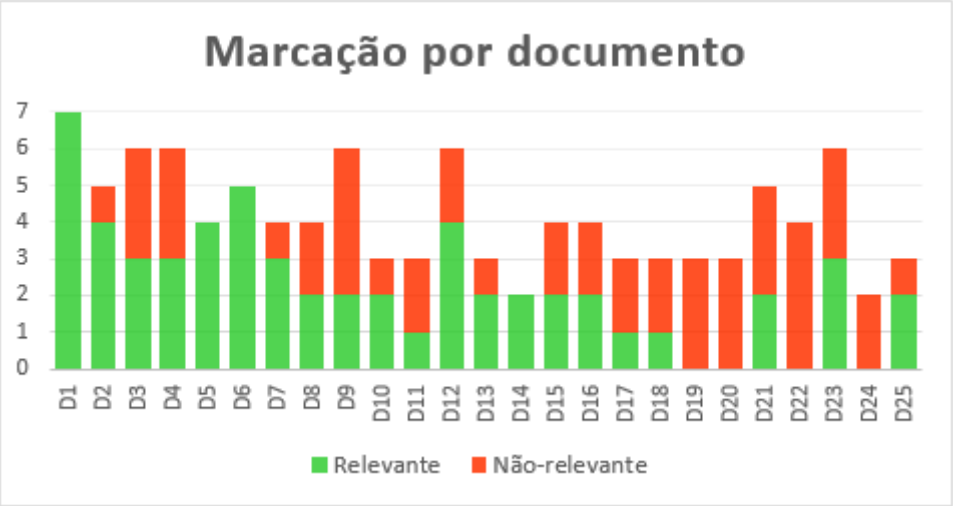

Figura 29 - Gráfico de marcações por documento.

Neste gráfico, é possível observar que o documento *D1* foi marcado por todos os participantes do teste como relevante. *D1* é o primeiro estudo resultante da busca no *IEEExplore*, e aborda uma discussão sobre as dificuldades dos pesquisadores na realização de Revisões Sistemáticas da Literatura. Este assunto está diretamente ligado aos critérios da Revisão Sistemática proposta na descrição do teste e foi marcado acertadamente como relevante por todos os participantes.

Há, também, documentos como o *D5* e *D6*, que não tiveram nenhuma marcação como não-relevantes. E, da mesma forma, documentos como o *D19*, *D20*, *D22* e *D24*, que não foram considerados relevantes em nenhum dos testes. Em relação ao documento *D5*, este é um artigo que tem, entre os autores, Bárbara Kitchenham, autora dos principais

estudos sobre Revisão Sistemática da Literatura na área de Engenharia de Software [Kit04] [Kit07] [Kit12]. Através da leitura da transcrição da execução dos testes, é possível perceber que três dos quatro participantes que marcaram o estudo como relevante procuraram pelo estudo nos resultados da página do *IEEExplore* (na outra aba aberta no navegador), e o marcaram como relevante ao perceberem o nome da autora.

Cinco dos sete participantes marcaram o documento *D6* como relevante. Após o teste, questionado sobre sua seleção do documento em questão, um dos participantes (*P6*) referiu-se à conferência na qual fora publicado. A *EASE*[12] (*Evaluation and Assessment in Software Engineering*) é uma conferência de compartilhamento de conhecimento relacionado a estudos empíricos na área de Engenharia de Software e, entre os tópicos aceitos na conferência, está a Revisão Sistemática da Literatura.

Quanto aos estudos que não foram considerados relevantes em nenhum dos testes (*D19, D20, D22* e *D24*), um deles é relacionado à área de teste de *software* (*D19*), outros dois abordam certificações em gerência na Engenharia de Software (*D20* e *D22*) e o último é uma meta-análise sobre a relação entre personalidade de desenvolvedores de *software* (*D24*). Portanto, não foram marcados como relevantes acertadamente na avaliação dos pesquisadores.

É possível, ainda na análise da tabela de síntese dos resultados do teste (tabela 13), observar a seleção de relevância dos documentos em relação aos pesquisadores, como exposto no gráfico da figura 30. Nele, observamos quais pesquisadores marcaram mais documentos, e quantos foram marcados como relevantes e não-relevantes. A discussão desta observação encontra-se na seção 5.4.

[12] http://ease2014.org/index.html

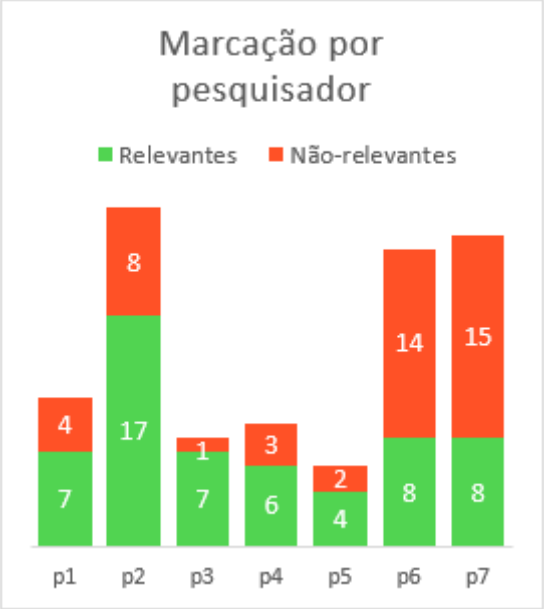

Figura 30 - Marcação dos documentos por pesquisador.

Como resultado de cada teste, a *string* de busca encontrada pelos pesquisadores foi capturada. A tabela 15 traz uma relação das *strings* encontradas, identificadas pela coluna *S* (de *S1* a *S7*). Nela, é possível verificar qual pesquisador (de *P1* a *P7*, na coluna *P*) é responsável por qual *string* de busca encontrada.

Tabela 15 - Strings construídas na execução dos testes, capturadas ao final dos mesmos.

P	S	*String* final encontrada no teste
P1	*S1*	*systematic review in software AND engineering AND lessons OR groups OR slr OR method AND statistical*
P2	*S2*	*("systematic review" AND "software engineering") AND (motivation OR replication)*
P3	*S3*	*(systematic review in software engineering OR "systematic reviews") AND (knowledge OR outcomes OR contribute OR "search strings" OR "search approach" OR "reviews findings")*
P4	*S4*	*systematic review in software engineering*
P5	*S5*	*((systematic and review) or slr or (tertiary and study)) and ((software and engineering) or se)*
P6	*S6*	*"systematic review" OR software OR engineering OR methodology OR replication*
P7	*S7*	*(systematic review OR systematic literature review OR slr OR tertiary study) AND (software engineering)*

Fazendo uma rápida leitura sobre as *strings* de busca resultantes da tarefa proposta no teste (visualizadas na tabela 15), observa-se a incidência de termos incluídos pelos pesquisadores. Além dos termos presentes na Revisão Sistemática proposta pelo teste

(descrito na seção 5.2.1), como “*systematic*”, “*review*”, “*software*” e “*engineering*”, foram incluídos, pelos participantes, termos como “*slr*” (incluído por 3 participantes), e “*method*”, “*replication*”, “*study*” e “*tertiary*” (incluídos por dois participantes). O termo “*search*” foi utilizado duas vezes pelo mesmo participante (*P3*), em sua *string* de busca (*S3*). A tabela 16 apresenta a utilização de todos os termos incluídos em relação às *strings* construídas pelos participantes.

Tabela 16 - Relação dos termos incluídos às strings de busca ao final dos testes.

	termo	freq total	S1	S2	S3	S4	S5	S6	S7
t1	*review*	10							
t2	*systematic*	9							
t3	*engineering*	7							
t4	*software*	7							
t5	*slr*	3							
t6	*method*	2							
t7	*replication*	2							
t8	*search*	2							
t9	*study*	2							
t10	*tertiary*	2							
t11	*approach*	1							
t12	*contribute*	1							
t13	*findings*	1							
t14	*groups*	1							
t15	*knowledge*	1							
t16	*lessons*	1							
t17	*literature*	1							
t18	*methodology*	1							
t19	*motivate*	1							
t20	*outcomes*	1							
t21	*search*	1							
t22	*statistical*	1							
t23	*string*	1							

5.3 Entrevistas pós-teste

Com o objetivo de captar a opinião dos pesquisadores em relação à ferramenta e ao método proposto, foram realizadas entrevistas logo após a realização dos testes. Esta seção descreve tanto o roteiro utilizado como base para as entrevistas como apresenta os resultados obtidos a partir delas.

5.3.1 Roteiro da entrevista

Para a realização das entrevistas pós-teste, foi criado um roteiro com 6 perguntas, separadas em dois grupos: (1) questões relacionadas à execução e resultados do teste e (2) questões abordando à experiência dos pesquisadores em relação ao método utilizado pela ferramenta. As questões encontram-se na tabela 17.

Tabela 17 – Roteiro com questões aplicadas na entrevista pós-teste.

Execução e resultados do teste	
Q2.1	Você acredita ter encontrado uma *string* de busca relevante para a Revisão Sistemática proposta? Qual é ela?
Q2.2	O quanto você acredita que a ferramenta lhe ajudou na construção da *string* de busca?
Q2.3	Como você costuma realizar a construção da *string* de busca de uma Revisão Sistemática?
Q2.4	Qual recurso da ferramenta você achou mais útil?
Experiência dos pesquisadores e o método	
Q2.5	Você acredita que a ferramenta ajudaria pesquisadores iniciando uma Revisão Sistemática em uma área desconhecida a encontrar termos mais relevantes? Por que?
Q2.6	Na sua opinião, pesquisadores experientes poderiam também se beneficiar do uso de uma ferramenta como a utilizada? Por que?

5.3.2 Apresentação do resultado das entrevistas

Uma síntese das respostas dos participantes do teste pode ser vista na tabela 18. Nela, os pesquisadores estão identificados por *P1* a *P7*, e as questões de *Q2.1* a *Q2.6*. Cabe ressaltar que as respostas dos pesquisadores foram, em sua maioria, subjetivas e, em alguns casos, não permitem uma análise unicamente quantitativa. Por isso, as questões

são abordadas nesta seção em maiores detalhes, incluindo todas as observações feitas pelos pesquisadores.

Tabela 18 – Síntese das respostas dos pesquisadores para as questões da entrevista pós-teste.

	P1	P2	P3	P4	P5	P6	P7
Q2.1	não	não	sim, mas teria que testá-la	sim	não	não	não muito
Q2.2	bastante	me apontou termos que não incluí	bastante, na escolha dos termos	ajudou 100%	pouco	apenas na forma de visualizar os estudos	ajudou apenas com um termo
Q2.3	referência de estudo/autor conhecido	começo com termos que conheço	referência de estudo/autor conhecido	começo com termos que conheço			referência de estudo/autor conhecido
Q2.4	*carousel* e mapa de calor	relevância dos termos sugeridos e edição da *string*	opções e mapa de calor	mapa de calor e termos sugeridos	edição da *string*	mapa de calor	termos sugeridos e mapa de calor
Q2.5	sim	sim	sim	sim	sim	sim	sim
Q2.6	sim, mas de outra forma	sim, auudando a confirmar os termos	sim, mas depende da abordagem	talvez não	sim, mas para se atualizar em realação aos termos	sim	sim, para lembrar termos

Analisando as respostas para a primeira questão (*Q2.1*), é possível observar, na tabela 18, que cinco dos sete pesquisadores indicam não terem encontrado uma *string* de busca relevante para a Revisão Sistemática proposta no teste. Três destes cinco pesquisadores (*P5*, *P6* e *P7*) indicaram a falta de termos realmente relevantes nas sugestões da ferramenta.

Os outros dois pesquisadores que indicaram não terem encontrado uma *string* de busca relevante para a Revisão Sistemática proposta (*P1* e *P2*) indicaram falta de tempo para trabalhar na marcação dos estudos, seleção dos termos sugeridos e edição da *string* de busca. Observando a resposta de um dos pesquisadores (*P1*), ele pondera se, com mais tempo, teria conseguido obter uma *string* relevante:

> *"Acredito que, se tivesse mais tempo para trabalhar, conseguiria uma boa* string*. Com o uso da ferramenta, consegui encontrar termos que não tinha nem pensado. Um era a sigla* SLR*, e nem tinha pensado em colocar na* string*. Acho que, com o tempo, pode-se montar uma* string *boa com os termos sugeridos. Como o tempo foi curto, a* string *não ficou otimizada."*

Em relação à maneira como os pesquisadores costumam realizar a construção da *string* de busca de suas Revisões Sistemáticas da Literatura (abordado na questão *Q2.3*), quatro pesquisadores indicaram iniciar a construção por termos conhecidos por eles (*P2*, *P4*, *P5* e *P6*). Através de uma busca simples com estes termos, eles capturam sinônimos dos termos pela leitura de estudos resultantes, atualizando a *string* de busca. Um dos pesquisadores do grupo (*P6*) afirma realizar a construção dessa forma: "*Primeiro pego as palavras-chave que conheço. Jogo na busca, reviso os artigos e atualizo a* string *de busca. Repito o procedimento até encontrar uma* string *de busca interessante [...]*".

Os outros três pesquisadores (*P1*, *P3* e *P7*) afirmaram utilizarem estudos mais conhecidos (de autores conhecidos) ou mais referenciados (com mais citações) da área onde aplicarão a Revisão Sistemática. "*[...] Depois, procuro extrair os termos. Seria mais ou menos parecido com a ferramenta, mas em um conjunto de estudos pré-filtrado*", observou um dos pesquisadores (*P3*).

Quando questionados sobre qual recurso da ferramenta acharam mais útil (*Q2.4*), cinco dos sete pesquisadores (*P1*, *P3*, *P4, P6* e *P7*) citaram o mapa de calor (*heatmap*). Em segundo lugar no número de citações está a sugestão de termos mais relevantes (recurso-chave da ferramenta e do método), tendo sido citado por três pesquisadores (*P2*, *P4* e *P7*).

Tabela 19 - Recursos citados pelos pesquisadores como sendo mais úteis da ferramenta, em resposta à quarta questão (Q2.4) do questionário.

Recurso	Citações
Mapa de calor (*heatmap*)	5
Termos sugeridos	3
Edição da *string*	2
Carousel dos documentos	1
Opções	1

Um dos pesquisadores que citaram o mapa de calor (*P5*) destaca sua utilidade na busca de termos relevantes: "*[...] O mapa de calor me ajudou a ver, quando encontrado um termo interessante, quais estudos continham aquele termo, para que eu não deixasse de selecionar um estudo relevante*".

Os pesquisadores foram unânimes nas suas respostas quando questionados se a ferramenta poderia ajudar pesquisadores realizando uma Revisão Sistemática da Literatura em uma área desconhecida (*Q2.5*), na busca de termos relevantes para a construção da *string* de busca. Todos acreditam que a ferramenta ajudaria pesquisadores com pouca experiência em determinada área. Um dos pesquisadores (*P3*) afirma isso da seguinte

forma: “*Sim ajudaria, pois auxiliaria justamente na questão dos termos. Se não conhecemos uma determinada área, é difícil encontrar alguns termos, e a ferramenta auxiliaria neste sentido, descobrindo palavras-chave para a busca*”.

Já quanto a sua utilização por pesquisadores experientes em uma determinada área, a maioria dos pesquisadores afirmou que a ferramenta auxiliaria na construção da *string* de busca da Revisão Sistemática, mas com algumas ressalvas. Os pesquisadores *P2*, *P3*, *P5* e *P7* afirmam que é mais provável que tal ferramenta seja utilizada por pesquisadores experientes em casos onde já se tenha ao menos um esboço da *string* de busca, na validação de seus termos. Um dos pesquisadores (*P5*) ainda indicou que, sendo experiente em uma área na qual fosse realizar uma Revisão Sistemática, utilizaria a ferramenta para se atualizar quanto aos termos da área. Segundo ele: “*[...] a ferramenta poderia ser usada para se atualizar em relação aos termos conhecidos, por exemplo. O processo da revisão não mudaria e, por mais que fosse experiente em uma área, utilizaria a ferramenta sim*”.

5.4 Discussão dos resultados

A experiência dos pesquisadores em relação à realização de Revisões Sistemáticas pode ser vista na tabela 20. Percebe-se que os pesquisadores com maior experiência (*P6* e *P7*) - que já realizaram mais de uma Revisão Sistemática e tiveram alguma revisão do tipo avaliada e publicada - mencionam dificuldades na construção da *string* de busca, como destacado na mesma tabela. Tal constatação vai de acordo com autores que afirmam que problemas com a construção da *string* de busca e seleção de estudos primários são comuns na realização de Revisões Sistemáticas, sendo enfrentados tanto por pesquisadores iniciantes quanto por mais experientes [Ria10].

Tabela 20 - Experiência dos pesquisadores.

P	Realizou apenas uma Revisão?	Iniciou uma Revisão sem terminar?	Teve uma Revisão avaliada?	Teve uma Revisão publicada?	Mencionou dificuldades na construção da string?
P1	sim	sim	não	não	sim
P2	sim	não	não	não	sim
P3	sim	não	sim	não	sim
P4	sim	não	não	sim	não
P5	sim	não	sim	não	não
P6	não	sim	sim	sim	sim
P7	não	sim	sim	sim	sim

Os pesquisadores que mostraram ter maior experiência com Revisões Sistemáticas da Literatura também estão entre os que marcaram mais estudos como relevantes e não-relevantes, como evidenciado pela figura 30, na seção 5.2.4. O conhecimento adquirido pela realização de revisões deste tipo facilitou a marcação de estudos que sabiam serem referências na área, sem a necessidade de leitura de seus *abstracts* ou de uma análise mais detalhada. Um destes pesquisadores mais experientes (*P6*), quando questionado se já conhecia os estudos que marcara como relevantes, afirmou conhecer a maioria deles. Porém, foi ressaltado pelo mesmo pesquisador que, embora tenha selecionado estudos que sabia serem relevantes, não percebeu termos relevantes sendo sugeridos pela ferramenta: "*[...] o que estou sentindo falta é de algumas palavras que me ajudariam a construir a* string*. O que tenho aqui são palavras que realmente aparecem nos artigos selecionados - e sei que estes artigos são bons - mas que não incluiria na* string *de busca da revisão. [...] Os estudos que estou marcando são estudos que consideraria como relevantes, vinculados com o tema e que me ajudariam a responder as questões de pesquisa*".

Por se tratar de um pesquisador mais experiente, é compreensível que *P6* tenha certa expectativa em relação aos termos que são sugeridos, tendo em mente palavras que sabe serem relevantes para a revisão proposta. Este pesquisador está entre os participantes que sentiram falta de termos relevantes nas sugestões da ferramenta, grupo no qual também encontra-se outro dos pesquisadores mais experientes (*P7*).

Como mencionado pelos pesquisadores *P2*, *P3, P5* e *P7* quando da utilização da ferramenta por pesquisadores mais experientes, é possível perceber que o pesquisador *P6* procurou validar os termos que já conhecia, o que é considerada uma utilização válida do método proposto. Porém, para que o método funcione para validação, deve ser utilizado de

maneira iterativa, executando novas buscas à medida que a *string* de busca é modificada. Assim, refina-se a lista de estudos resultantes ao ponto de encontrar um conjunto de estudos conhecidos pelo pesquisador, confirmando a relevância da *string* de busca construída.

Enquanto o pesquisador *P6* indicou iniciar a construção de sua *string* de busca utilizando termos sabidamente relevantes, o pesquisador *P7* afirmou partir de um conjunto de estudos consolidados da área (de autores conhecidos e mais citados). "*Eu, então, observo seus* abstracts*, seu título, introdução e conclusão. Comparando a minhas questões de pesquisa, procuro construir minha* string *de busca*", afirma. Outros pesquisadores (*P3* e *P5*) também indicaram utilizar a mesma prática para construção de suas *strings* de busca. Por mais experiente que o pesquisador seja em determinada área, partir de um conjunto de estudos consolidados indica ser uma boa prática [Zha11a] [Zha11b]. Tal abordagem é permitida pelo método proposto, onde a marcação dos estudos consolidados como relevantes geraria uma lista de termos relevantes sugeridos para construção da *string* de busca.

Em um guia de procedimentos para realização de Revisões Sistemáticas [Kit07], é recomendada a construção de questões de pesquisa estruturadas de forma a auxiliar a extração de termos para a *string* de busca. Esta recomendação, porém, foi recentemente revista e retirada por seus autores [Kit13], alegando que ela "*[...] leva a* strings *de busca muito complexas*". Em seu lugar, é recomendado utilizar o *quasi-gold standard* (*QGS*) [Zha11] (abordado na seção 3.1) em um conjunto limitado de estudos selecionados manualmente no auxílio à construção da *string* de busca. Esta abordagem vai ao encontro da prática dos participantes supracitados, partindo de um conjunto conhecido de estudos (possivelmente o mesmo conjunto do *QGS*) para extração de termos relevantes e construção da *string* de busca.

O resultado dos testes indica que, embora a maioria dos pesquisadores tenha indicado não terem encontrado uma *string* de busca relevante ao término dos testes, alguns termos relevantes para a Revisão Sistemática proposta foram sugeridos e incluídos, aparecendo nas *strings* finais. Um dos termos mais incluídos durante os testes (três vezes) foi o termo "*slr*". Sendo o termo "*slr*" o acrônimo em inglês para Revisão Sistemática da Literatura (*Systematic Literature Review*), sua inclusão na *string* de busca da revisão proposta no teste é um acerto.

Na literatura, *strings* de busca utilizadas por autores que realizaram um mapeamento sistemático de estudos que abordam tanto a utilização de técnicas da Mineração Visual de

Texto no auxílio à condução da Revisão Sistemática [Fel12a] como um levantamento de ferramentas que auxiliam este tipo de revisão [Mar13] incluem o acrônimo "*SLR*". Portanto, tendo o termo "*slr*" sido incluído por três dos sete pesquisadores (*P1*, *P5* e *P7*), é possível considerar que a ferramenta alcançou o objetivo de sugerir um termo relevante e auxiliar na construção da *string* de busca, ao menos nestes casos.

Segundo comentários dos próprios pesquisadores, quando da entrevista pós-teste, nota-se o destaque dado à consideração do termo "*slr*" através da ferramenta: "*Com o uso da ferramenta, consegui encontrar termos que não tinha nem pensado. Um era a sigla SLR, e nem tinha cogitado colocar na* string *[...]*", afirma um deles (*P1*). "*Dos termos sugeridos, só aproveitei o acrônimo* SLR", observa outro (*P7*). Este ainda complementa: "*[...] o (termo)* slr*, que não lembrei, foi sugerido pela ferramenta, mas poderia ter me passado batido. Como foi sugerido, lembrei do termo e adicionei à* string". Tais comentários constatam que o método proposto funcionou como esperado na sugestão de, ao menos, um termo relevante para construção da *string* de busca.

Se utilizado de maneira iterativa - fazendo novas pesquisas a cada atualização da *string* de busca - e com mais tempo para análise dos estudos por parte dos usuários, o método tem o potencial de auxiliar mais na sugestão de termos relevantes.

5.5 Limitações da pesquisa

Uma das principais limitações observadas na realização deste trabalho é técnica, e refere-se à tarefa de mineração de textos. Dado que a ferramenta construída atua na forma de um *script*, lendo a página de resultados da biblioteca digital *IEEExplore*, a obtenção do corpo de texto de cada estudo mostrou-se tecnicamente inviável. Isso deve-se ao fato de a página de resultados do *IEEExplore* trazer, no seu conteúdo, somente o *abstract* de cada estudo, e não o estudo completo. Assim sendo, as técnicas de mineração de texto aplicadas pela ferramenta limitam-se somente ao texto presente no *abstract* de cada estudo analisado.

Com tal limitação, a relevância dos termos é calculada em cima de um corpo de texto que, na média e por observação do autor, mostrou-se em torno de 48 a 54 palavras. Com estudos recentes questionando a qualidade de *abstracts* de estudos da área de Engenharia de Software [Bre07] e indicando que os *abstracts* das publicações contém, em geral, menos da metade das informações contidas em um estudo [Ana09], esta limitação introduz um viés ao resultado da pesquisa.

Porém, através de uma comparação entre os termos mais frequentes encontrados em um estudo primário e seu respectivo *abstract*, é possível observar uma semelhança na lista de termos relevantes. Utilizando-se as mesma técnicas de mineração propostas sobre o primeiro estudo encontrado buscando-se pela *string "Systematic Review in Software Engineering"* no *IEEExplore* [Bab09] - como proposto nos testes aplicados neste trabalho - obteve-se o resultado encontrado na tabela 21. No estudo completo, são contabilizados 979 palavras, e, no *abstract* do mesmo estudo, palavras. Na tabela 21, estão representadas as frequências dos termos encontrados através da mineração, e valores normalizados de frequência, obtidos através da divisão da frequência de um termo pelo termo de maior frequência do mesmo texto (documento ou *abstract*).

Tabela 21 - Estudo comparativo entre termos encontrados no corpo completo de um estudo [Bab09] e em seu *abstract*.

	Documento				Abstract		
	Termo	#	norm		Termo	#	norm
TD1	*slr*	96	1.00	TA1	*review*	9	1.00
TD2	*review*	89	0.93	TA2	*systematic*	8	0.89
TD3	*software*	74	0.77	TA3	*software*	6	0.67
TD4	*systematic*	61	0.64	TA4	*engineering*	6	0.67
TD5	*engineering*	58	0.60	TA5	*knowledge*	4	0.44
TD6	*researches*	55	0.57	TA6	*experience*	3	0.33
TD7	*interview*	51	0.53	TA7	*reporting*	2	0.22
TD8	*studies*	45	0.47	TA8	*studies*	2	0.22
TD9	*reporting*	44	0.46	TA9	*gained*	2	0.22
TD10	*data*	37	0.39	TA10	*methodology*	2	0.22
TD11	*se*	36	0.38	TA11	*based*	2	0.22
TD12	*experience*	34	0.35	TA12	*body*	2	0.22
TD13	*guidelines*	34	0.35	TA13	*researches*	2	0.22
TD14	*interviewees*	29	0.30	TA14	*programs*	2	0.22
TD15	*questions*	25	0.26	TA15	*contribute*	2	0.22

É possível observar, na tabela 21, que o estudo em análise possui termos semelhantes na comparação entre a mineração sobre seu texto completo e sobre seu *abstract*. São encontrados, em ambas as listas, termos como *review* (TD2 e TA1), *software* (TD3 e TA3), *engineering* (TD5 e TA4) e *researches* (TD6 e TA13), entre outros. Existe, portanto, uma notada representatividade nos termos utilizados no estudo completo se comparados aos termos mais relevantes em seu *abstract*, apesar de questionável qualidade [Bre07] e conteúdo incompleto [Ana09].

Há de se observar, ainda, a limitação existente em relação ao universo de documentos (no caso da aplicação deste trabalho, *abstracts*) analisados. Tendo em vista que a ferramenta lê apenas a página atual de resultados na busca efetuada, o número de

documentos minerados será aquele configurado como limite de resultados pela ferramenta de busca da biblioteca digital utilizada (no caso, *IEEExplore*). Por padrão, o limite é de 25 resultados. Um universo de 100 ou mais documentos forneceria um resultado de mineração mais refinado.

Outra limitação do presente trabalho é o fato da ferramenta desenvolvida estar limitada à biblioteca digital *IEEExplore*, como citado anteriormente. Mesmo a busca nesta biblioteca ser uma recomendação para a realização de Revisões Sistemáticas [Kit13], o fato de aplicar o método em apenas uma biblioteca digital (como faz a ferramenta *SLR.qub*) limita o conjunto de estudos que podem ser encontrados. Uma nova ferramenta, que concatenasse os estudos encontrados em todas as bibliotecas digitais antes da aplicação das técnicas de mineração sugeridas pelo método poderia obter um resultado mais interessante. Tal abordagem seria permitida pelo método, se observado que estudos não estejam duplicados no conjunto minerado.

De certa maneira, as limitações técnicas supracitadas oferecem oportunidades de aplicação de algoritmos de mineração de texto mais robustos. A seção 6.1 aborda sugestões de trabalhos futuros que exploram tais oportunidades.

6. CONSIDERAÇÕES FINAIS

Com o aumento no volume de pesquisas empíricas na Engenharia de Software, a Revisão Sistemática da Literatura é uma maneira de manter-se atualizado ou identificar a necessidade de novos estudos e oportunidades de pesquisa em uma área. As Revisões Sistemáticas avaliam e interpretam pesquisas relevantes disponíveis para uma determinada questão de pesquisa, tópico da área ou fenômeno de interesse.

Apesar do aumento na popularidade da aplicação de Revisões Sistemáticas na Engenharia de Software, muitos pesquisadores ainda a apontam como um processo que demanda muito tempo e que apresenta desafios. Por se tratar de uma metodologia relativamente nova na área da Engenharia de Software, estudos que abordam lições aprendidas na realização de Revisões Sistemáticas reportam problemas em diferentes atividades ao longo de seu processo. Uma das atividades mencionadas é a construção da *string* de busca da Revisão Sistemática, para a qual é reportada a falta de padronização dos termos utilizados, a dificuldade de definição de termos relevantes para um tópico e *strings* de busca resultantes muito complexas. Uma *string* de busca mal construída pode resultar em poucos ou muitos estudos sendo retornados, ou ainda impedir que estudos relevantes para a pesquisa sejam encontrados.

Entrevistas realizadas na presente pesquisa captaram a experiência de pesquisadores quanto à realização de Revisões Sistemáticas, listando as dificuldades por eles enfrentadas quando da aplicação deste tipo de revisão. Durante estas entrevistas, problemas na tarefa de construção da *string* de busca foram mencionados pela maioria dos entrevistados. De fato, dificuldades na construção da *string* de busca são enfrentadas por pesquisadores independentemente da sua experiência na realização de Revisões Sistemáticas da Literatura. São raras, porém, as propostas de métodos ou ferramentas nas atuais publicações da área que enderecem esta tarefa, na fase de planejamento da Revisão Sistemática. Portanto, foi vista como uma oportunidade, a proposta de um método que auxiliasse nesta tarefa.

Desta forma, com o objetivo de auxiliar na tarefa de construção da *string* de busca de uma Revisão Sistemática da Literatura, o presente trabalho propôs um método iterativo que apoia o pesquisador através da sugestão de termos relevantes de estudos selecionados manualmente. Através de técnicas de Mineração Visual de Texto, os termos mais relevantes dos estudos selecionados pelo pesquisador são extraídos e apresentados por meio de uma visualização em mapa de calor que possibilita identificar sua frequência

em todos estudos resultantes da busca. O método permite, então, que a *string* de busca seja alterada para incluir termos sugeridos, e que uma nova busca seja feita, repetindo o processo quantas vezes for necessário para a construção e refinamento da *string* de busca.

Para a análise dessa proposta, foi desenvolvida uma ferramenta que implementa o método descrito e foram realizados testes com os mesmos pesquisadores, anteriormente entrevistados. Ao final de cada teste, o pesquisador foi submetido a uma nova entrevista, com o intuito de capturar sua opinião quanto ao método e à ferramenta, relacionando com suas dificuldades e experiência na realização de Revisões Sistemáticas.

Em resposta às questões da entrevista realizada após os testes, todos os entrevistados afirmaram acreditar que a ferramenta poderia auxiliar pesquisadores a realizar uma Revisão Sistemática da Literatura em uma área desconhecia, sugerindo termos relevantes que não sejam de conhecimento do pesquisador. A maior parte dos entrevistados afirmou, também, acreditar no auxílio da ferramenta a pesquisadores que já tenham experiência na área onde realizam uma Revisão Sistemática, na validação dos termos da *string* de busca que este já conheça ou na sua atualização em relação aos termos utilizados.

O resultado consolidado das respostas dos entrevistados no pós-teste aponta a aceitação por uma ferramenta baseada no método proposto, indicando também que os entrevistados a utilizariam na realização de futuras Revisões Sistemáticas da Literatura. É notado, porém, que a maioria dos pesquisadores que utilizaram a ferramenta afirmou não ter encontrado uma *string* de busca relevante para a Revisão Sistemática que lhe foi proposta no teste. Os motivos apontados pelos pesquisadores foram o tempo limitado do teste e a falta de termos sugeridos que fossem realmente relevantes para a revisão proposta, o que aponta para a necessidade, como próximo passo desta pesquisa, de um aprofundamento da análise realizada. Seria preciso refinar a ferramenta utilizada para endereçar suas limitações, submetendo-a para utilização por uma amostra maior e mais diversificada de pesquisadores - com diferentes experiências em relação na área de Revisão Sistemática – e, principalmente, por um tempo mais prolongado. Assim seria possível capturar mais informações e alimentar discussões mais aprofundadas e futuros trabalhos nesta promissora linha de pesquisa.

6.1 Trabalhos futuros: suporte a outras bibliotecas digitais

Como mencionado, uma das limitações do presente trabalho está na implementação de uma ferramenta compatível apenas com o site da biblioteca digital *IEEExplore*, limitando o conjunto de estudos que podem ser encontrados na aplicação do método. Sugiro, portanto, que este trabalho tenha continuidade no desenvolvimento de ferramentas que implementem o método proposto, suportando a leitura e mineração de resultados em diferentes bibliotecas digitais.

Para isso, a primeira parte do algoritmo – que varre a página de resultados da biblioteca digital *IEEExplore* – deve ser alterada para incluir a leitura de outras páginas de resultados. Dessa forma, o *bookmarklet* seria capaz de, quando acionado, perceber qual biblioteca digital está aberta e executar o *script* que faça a leitura e extração de dados. Com tal suporte, a mineração seria executada sobre um conjunto maior de estudos, possivelmente fazendo com que as sugestões de termos relevantes mostrem-se mais apuradas.

No entanto, o desenvolvimento de uma ferramenta que implemente o método proposto e suporte múltiplas bibliotecas digitais deve observar dois pontos importantes: (1) evitar a duplicação de estudos que possam ser encontrados em mais de uma biblioteca digital, sob o risco de impactar o resultado da mineração, e (2) a forma como a estrutura lógica da *string* de busca – operadores lógicos e hierarquia entre os termos – é entendida pelas bibliotecas suportadas, uma vez que cada biblioteca digital utiliza a sua maneira de estruturar a *string* de busca.

Como sugestão, a ferramenta resultante de tal implementação pode ser implementada como uma contribuição à ferramenta apresentada neste trabalho, através do projeto *Qub*, disponível no portal *GitHub*.

6.2 Trabalhos futuros: utilização de algoritmos mais robustos

Outra oportunidade de exploração de trabalhos futuros é a aplicação das técnicas utilizadas neste estudo sobre todo o texto dos estudos primários, e todos os estudos encontrados em uma busca. Para tal, algoritmos que consigam capturar textos dos arquivos onde se encontram os estudos - em formato *PDF* (*Portable Document Format*) - e realizar o cálculo de *tf-idf* em um corpo maior de análise.

Mesmo demandando um maior poder de processamento, seria interessante comparar o resultado na sugestão dos termos mais relevantes com o que foi encontrado na

presente pesquisa, com a mineração dos *abstracts*. É possível que, através de algoritmos que minerem o texto contido em todos os estudos, a lista de termos sugeridos seja refinada e apresente termos mais relevantes.

6.2 Trabalhos futuros: realização de testes com pesquisadores iniciantes em uma área

O teste com usuários realizado no presente trabalho propôs a construção da *string* de busca de uma Revisão Sistemática de Revisões Sistemáticas. Como pré-requisito e para enriquecimento da discussão dos resultados, foram selecionados participantes com alguma familiaridade com o conceito de Revisão Sistemática. Assim sendo, os pesquisadores que participaram do teste tinham experiência no tema proposto na revisão do teste.

Uma sugestão para continuidade deste estudo está na aplicação de testes similares, mas que proponham uma Revisão Sistemática em uma outra área (como exemplo, Mineração de Texto), e que sejam selecionados, para o teste, pesquisadores que não tenham familiaridade com o tema proposto. Com isso, objetiva-se observar a aplicação do método proposto no presente trabalho por pesquisadores iniciantes em uma área.

Sem conhecer muitos termos relevantes da área, o método, se funcionar como proposto, deverá sugerir termos e auxiliar a construção de uma *string* de busca relevante. Tal estudo, se realizado, poderá ter seus resultados discutidos e comparados aos obtidos no presente trabalho.

REFERÊNCIAS BIBLIOGRÁFICAS

[Ana09] S. Ananiadou, B. Rea, N. Okazaki, R. Procter, J. Thomas. “Supporting systematic reviews using text mining”. *Social Science Computer Review (SSCR)*, vol. 27, n. 4, Abril 2009, pp. 509-523.

[Bab09] M. A. Babar, H. Zhang. “Systematic literature reviews in software engineering: Preliminary results from interviews with researchers”. *In: Proceedings of the Empirical Software Engineering and Measurement (ESEM)*, 2009, pp. 346-355.

[Bio05] J. Biolchini, P.G. Mian, A.C.C. Natali, G.H. Travassos. “Systematic review in software engineering”. *Technical Report ES, System Engineering and Computer Science Department COPPE/UFRJ*, 2005, 679 p.

[Bow12] D. Bowes, T. Hall, S. Beecham. “SLuRp: a tool to help large complex systematic literature reviews deliver valid and rigorous results”. *In: Proceedings of the Empirical Software Engineering and Measurement (ESEM)*, 2012, pp. 33-36.

[Bre07] P. Brereton, B. Kitchenham, D. Budgen, M. Turner, M. Khalil. “Lessons from applying the systematic literature review process within the software engineering domain”. *Journal of Systems and Software*, vol. 80, n. 4, Abril 2007, pp. 571-583.

[Cho11] J. K. Chou, C. K. Yang. “PaperVis: Literature review made easy”. *In: Proceedings of the Eurographics / IEEE - VGTC conference on Visualization (EuroVis)*, 2011, pp. 721-730.

[Cru07a] D. Cruzes, M. Mendonça, V. Basili, F. Shull, M. Jino. “Automated information extraction from empirical software engineering literature: is that possible?” *In: Proceedings of the Empirical Software Engineering and Measurement (ESEM*), 2007, pp. 491-493.

[Cru07b] D. Cruzes, M. G. Mendonça, V. Basili, F. Shull, M. Jino. “Using context distance measurement to analyze results across studies”. *In: Proceedings of the Empirical Software Engineering and Measurement (ESEM)*, 2007, pp. 235-244.

[Dyb07] T. Dybå, T. Dingsøyr, G. K. Hanssen. “Applying systematic reviews to diverse study types: An experience report”. *In: Proceedings of the Empirical Software Engineering and Measurement (ESEM)*, 2007, pp. 225-234.

[Dyb08] T. Dybå, T. Dingsøyr. “Strength of evidence in systematic reviews in software engineering.” *In: Proceedings of the Empirical Software Engineering and Measurement (ESEM)*, 2008, pp. 178-187.

[Fay96] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth. "From Data Mining to Knowledge Discovery: An Overview". *AI Magazine*, vol. 17, n. 3, Fevereiro 1996, pp. 1-37.

[Fel10] K. R. Felizardo, E. Y. Nakagawa, D. Feitosa, R. Minghim, J. C. Maldonado. “An approach based on visual text mining to support categorization and classification in the systematic mapping”. *In: Proceedings of the Evaluation and Assessment in Software Engineering (EASE)*, 2010, pp. 1-10.

[Fel11a] K. R. Felizardo, M. Riaz, M. Sulayman, E. Mendes, S. G. MacDonell, J. C. Maldonado. “Analysing the use of graphs to represent the results of Systematic Reviews in Software Engineering”. *In: Proceedings of the Brazilian Symposium on Software Engineering (SBES)*, 2011, pp. 174-183.

[Fel11b] K. R. Felizardo. “Using visual text mining to support the study selection activity in systematic literature reviews”. *In: Proceedings of the Empirical Software Engineering and Measurement (ESEM)*, 2011, pp. 77-86.

[Fel12a] K. R. Felizardo, S. G. MacDonell, E. Mendes, J. C. Maldonado. “A systematic mapping on the use of visual data mining to support the conduct of systematic literature reviews”. *Journal of Software*, vol. 7, n. 2, Fevereiro 2012, pp. 450-461.

[Fel12b] K. R. Felizardo, G. F. Andery, F. V. Paulovich, R. Minghim, J. C. Maldonado. “A visual analysis approach to validate the selection review of primary studies in systematic reviews”. *Information and Software Technology*, vol. 54, n. 10, Outubro 2012, pp. 1079-1091.

[Fer10] A. M. Fernández-Sáez, M. G. Bocco, F. P. Romero. “SLR-tool - a tool for performing Systematic literature reviews”. *In: Proceedings of the International Conference on Software and Data Technologies (ICSOFT)*, 2010, pp. 144.

[Few10] S. Few. “Data visualization for human perception”. Capturado em: https://www.interaction-design.org/encyclopedia/data_visualization_for_human_perception.html, Outubro 2013.

[Gar11] R. Garg, Heena. “Study of text based mining”. *In: Proceedings of the International Conference on Advances in Computing and Artificial Intelligence (ACAI)*, 2011, pp. 5-8.

[Gha12] M. Ghafari, M. Saleh, T. Ebrahimi. “A Federated Search Approach to Facilitate Systematic Literature Review in Software Engineering”. *International Journal of Software Engineering and Applications (IJSEA)*, vol. 3, n. 2, Março 2012, pp. 13-24.

[Gon12] L. F. P. Gonzalez. “Uma abordagem para mineração de dados e visualização de resultados em imagens batimétricas”. Capturado em: http://repositorio.pucrs.br/dspace/handle/10923/1574, Outubro 2012.

[Han06] J. Han, M. Kamber, P. Jian. “Data Mining: Concepts and Techniques”. *Morgan Kaufmann*, 3ª Ed., 2011, 744 p.

[Hea03] M. Hearst. “What is text mining?”. Capturado em: http://www.jaist.ac.jp/~bao/MOT-Ishikawa/FurtherReadingNo1.pdf, Fevereiro 2011.

[Hea99] M. Hearst. “Untangling text data mining”. *In: Proceedings of the Association for Computational Linguistics on Computational Linguistics (ACL)*, 1999, pp. 3-10.

[Her12] E. Hernandes, A. Zamboni, S. Fabbri, A. D. Thommazo. “Using GQM and TAM to evaluate StArt-a tool that supports Systematic Review”. *CLEI Electronic Journal*, vol. 15, n. 1, Abril 2012, pp. 1-15.

[Iee13] IEEExplore. “IEEE Xplore - Search Results”. Capturado em: http://ieeexplore.ieee.org/search/searchresult.jsp, Janeiro 2013.

[Kei02] D. A. Keim. “Information visualization and visual data mining”. *Visualization and Computer Graphics*, vol. 8, n. 1, Janeiro 2002, pp. 1-8.

[Kit04] B. Kitchenham. “Procedures for performing systematic reviews”. *Keele University*, 2004, 33 p.

[Kit07] B. Kitchenham, S. Charters. “Guidelines for performing systematic literature reviews in software engineering”. *EBSE Technical Report, Keele University and Durham University Joint Report*, 2007, 65 p.

[Kit12] B. Kitchenham, P. Brereton, D. Budgen. “Mapping study completeness and reliability-a case study”. *In: Proceedings of the Evaluation and Assessment in Software Engineering (EASE)*, 2012, pp. 126-135.

[Kit13] B. Kitchenham, P. Brereton. “A systematic review of systematic review process research in software engineering”. *Information and Software Technology*, vol. 55, n. 12, Dezembro 2013, pp. 2049-2075.

[Mal07] V. Malheiros, E. Hohn, R. Pinho, M. Mendonça, J. C. Maldonado. “A visual text mining approach for systematic reviews”. *In: Proceedings of the Empirical Software Engineering and Measurement (ESEM*), 2007, pp. 245-254.

[Man08] C. D. Manning, P. Raghavan, H. Schütze. “Introduction to information retrieval”. *Cambridge University Press*, 2008, 506 p.

[Mar13] C. Marshall, P. Brereton, B. Kitchenham. “Tools to Support Systematic Literature Reviews in Software Engineering: A Mapping Study”. *In: Proceedings of the Empirical Software Engineering and Measurement (ESEM)*, 2013, pp. 296-299.

[Nas06] H. A. do Nascimento, C. B. R. Ferreira. “Visualização de Informações: uma abordagem prática” *In: Proceedings of the XXIV Congresso da Sociedade Brasileira de Computação (SBC)*, 2005, pp. 1262-1312.

[Oli03] M. C. de O. Ferreira, H. Levkowitz. “From visual data exploration to visual data mining: a survey”. *Visualization and Computer Graphics*, vol. 9, n.3, Jul-Set 2003, pp. 378-394.

[Ria10] M. Riaz, M. Sulayman, N. Salleh, E. Mendes. “Experiences conducting systematic reviews from novices’ perspective”. *In: Proceedings of the Evaluation and Assessment in Software Engineering (EASE)*, 2010, pp. 44-53.

[Shn92] B. Shneiderman. “Tree visualization with tree-maps: 2-d space-filling approach”. *ACM Transactions on Graphics (TOG)*, vol. 11, n. 1, Janeiro 1992, pp. 92-99.

[Ste10] J. Steele, N. Iliinsky. “Beautiful Visualization: Looking at Data through the Eyes of Experts”. *O’Reilly Media*, 2010, 416 p.

[Sun12] Y. Sun, Y. Yang, H. Zhang, W. Zhang, Q Wang. “Towards evidence-based ontology for supporting systematic literature review”. *In: Proceedings of the Evaluation and Assessment in Software Engineering (EASE)*, 2012, pp. 171-175.

[Tan05] P. Tan, M. Steinbach, V. Kumar. “Introduction to data mining”. *Addison-Wesley*, 2005, 769 p.

[Tho11] J. Thomas, J. McNaught, S. Ananiadou. “Applications of text mining within systematic reviews”. *Research Synthesis Methods*, vol. 2, n. 1, Março 2011, pp. 1-14.

[Tom11] F. Tomassetti, G. Rizzo, A. Vetro, L. Ardito, M. Torchiano, M. Morisio. “Linked Data approach for selection process automation in Systematic Reviews”. *In: Proceedings of the Evaluation and Assessment in Software Engineering (EASE)*, 2011, pp. 31-35.

[Tor12] J. A. S. Torres, D. S. Cruzes, L. do N. Salvador. “Automatic Results Identification in Software Engineering Papers. Is it Possible?”. *In: Computational Science and Its Applications (ICCSA)*, 2012, pp. 108-112.

[Yau11] N. Yau. “Visualize This: The FlowingData Guide to Design, Visualization, and Statistics”. *John Wiley & Sons*, 2011, 384 p.

[Zha11a] H. Zhang, M. A. Babar, X. Bai, J. Li, L. Huang. “An empirical assessment of a systematic search process for systematic reviews”. *In: Proceedings of the Evaluation and Assessment in Software Engineering (EASE)*, 2011, pp. 56-65.

[Zha11b] H. Zhang, M. A. Babar, P. Tell. “Identifying relevant studies in software engineering”. *Information and Software Technology*, vol. 53, n. 6, Junho 2011, pp. 625-637.

APÊNDICE A – ROTEIRO DE TESTE

Apoio à avaliação de critérios de interação humano-computador de sistemas computacionais

Protocolo de Pesquisa Registro CEP 11/05667

Faculdade de Informática/PUCRS

Avenida Ipiranga, 6681 – Prédio 32 - 90619-900 – Porto Alegre – RS

Tel.: (51) 3320-3558

1. Definição do teste

Este teste tem por objetivo observar a aplicação de uma ferramenta que utiliza técnicas de Mineração Visual de Texto em auxílio à construção de uma *string* de busca de uma Revisão Sistemática.

Para tal, o pesquisador realizará o início de uma Revisão Sistemática, na fase de planejamento, construindo uma *string* de busca para a revisão sugerida. O ambiente do teste é o site da biblioteca digital *IEEE* (*IEEExplore*) e ocorrerá com a utilização da ferramenta criado.

Seguindo questões de pesquisa e critérios de seleção pré-definidos pelo autor do teste, o pesquisador selecionará estudos primários que considerar relevante, encontrando termos sugeridos pela ferramenta para sua *string* de busca e reaplicando a busca na ferramenta de busca do site do *IEEE*, de maneira contínua. Quando considerar ter montado uma *string* de busca relevante para a Revisão Sistemática proposta, ou findos 25 minutos, o teste será considerado concluído.

2. Seleção dos participantes

Os participantes do teste são pesquisadores familiarizados com a área de Engenharia de Software que já tenham realizado ao menos uma Revisão Sistemática, e que participaram de uma entrevista prévia sobre as dificuldades dos pesquisadores na realização de Revisões Sistemáticas.

3. Definição do cenário de teste

Como pesquisador, você deseja realizar uma Revisão Sistemática sobre Revisões Sistemáticas na Engenharia de Software para entender qual o estado-da-arte deste modelo de revisão literária na área em questão e quais dificuldades os pesquisadores têm enfrentado na sua adoção.

Para isso, define-se as seguintes questões de pesquisa e critérios de seleção de estudos.

Questões:

QP1) Que estudos reportam experiências ou investigam o uso da Revisão Sistemática na Engenharia de Software entre os anos de 2005 e 2012?

QP2) Que problemas foram observados por pesquisadores da área de Engenharia de Software na realização de Revisões Sistemáticas?

Critérios:
C1) Publicações entre 2005 e 2012;

C2) Menção explícita à Engenharia de Software (*Software Engineering*, ou termos similares) no título ou no *abstract* do estudo;

C3) Menção explícita à Revisão Sistemática (*Systematic Review*, ou termos similares) no título ou no *abstract* do estudo.

Tendo conhecimento sobre a área, sabe-se que a seguinte *string* é um bom início para sua pesquisa: ***systematic review in software engineering***.

Com as questões, os critérios e a *string* inicial, você tem tudo que precisa para iniciar o teste. Para fins de objetividade do teste, limitaremos a busca ao site da *IEEExplore* apenas. A *string* inicial é simples e a ferramenta permitirá a análise de termos sugeridos segundo a seleção dos estudos primários, além do uso de operadores lógicos, na construção da *string* de busca final.

Sua tarefa será a de, com a *string* inicial do enunciado, utilizar o protótipo de ferramenta desenvolvido para encontrar uma *string* de busca para a Revisão Sistemática proposta que considere satisfatória.

4. Definição de procedimentos e métricas

A) Preparação:

1. Abra o navegador *Google Chrome;*
2. Abra a seguinte página: https://raw.github.com/gmergel/Qub/master/bookmark.code;
3. Copie o código exibido na página;
4. Clique com o botão direito na barra de favoritos;
5. Adicione um novo favorito;
6. Dê-o o nome de "*Qub*";
7. Cole o código copiado (item 3) no campo *URL* do novo favorito;
8. Abra a biblioteca virtual IEEExplore: http://ieeexplore.ieee.org;
9. Digite a *string* de busca inicial no campo de pesquisa: *systematic review in software engineering* (e dê enter/clique em buscar);
10. Uma vez aberta a página com resultados, clique com o botão direito do mouse na aba aberta no navegador e clique em *Duplicar* (abrindo uma nova aba com a mesma página);
11. Na primeira aba, clique no favorito *Qub* (criado na preparação do teste) e a ferramenta abrirá;
12. Tudo pronto para o início do teste!

B) Execução do teste:

- Utilize a definição do cenário do teste para selecionar estudos considerados relevantes e estudos considerados não-relevantes para sua pesquisa
- Adicione termos sugeridos a sua *string* de busca como desejar

- Edite sua *string* de busca como desejar, adicionando termos não sugeridos e operadores lógicos (e parêntesis) como preferir;
- Altere opções e aplique suas alterações como desejar;
- Utilize o botão *Buscar* para encontrar novos resultados no *IEEExplore* utilizando a nova *string* de busca - lembre-se apenas de reativar a ferramenta clicando no favorito novamente quando os resultados tiverem carregado.

Critérios para fim do teste:

- Ter chegado a uma *string* que considere satisfatória como *string* de busca para sua Revisão Sistemática;
- Terminados 25 minutos de execução do teste.

ANEXO A – Termo de Consentimento

Apoio à avaliação de critérios de interação humano-computador de sistemas computacionais

Protocolo de Pesquisa Registro CEP 11/05667

Faculdade de Informática/PUCRS

Avenida Ipiranga, 6681 – Prédio 32 - 90619-900 – Porto Alegre – RS

Tel: (51) 3320-3558

Primeiramente, obrigado pela disponibilidade em participar deste teste! Seu objetivo é investigar questões relacionadas à funcionalidade e auxílio de técnicas de Mineração Visual de Textos na realização de Revisões Sistemáticas. Para isto, como participante deste teste, o convidarei a utilizar a ferramenta por mim criada sob minha observação. Aspectos relevantes de sua interação com a ferramenta serão capturados na forma de métricas, além de serem registrados em papel.

Caso o teste seja realizado de maneira remota, pedirei que compartilhe sua tela através da ferramenta que estiver em uso no momento (Google Hangout ou Skype), de forma que possa observar a execução do teste. Em nenhum momento do teste a tela será capturada. Observa-se que será capturado o endereço IP do participante como parte das métricas, mas apenas para fins de facilitação da análise dos resultados.

Lembro que o presente teste não tem por objetivo, de maneira alguma, avaliar o participante, mas sim a utilização das técnicas supracitadas. O uso que se faz dos registros efetuados durante o teste é estritamente limitado a atividades de pesquisa e desenvolvimento, garantindo-se para tanto que:

1. O anonimato dos participantes será preservado em todo e qualquer documento divulgado em foros científicos (tais como conferências, periódicos, livros e assemelhados) ou pedagógicos (tais como apostilas de cursos, slides de apresentações, e assemelhados);
2. Todo participante terá acesso a cópias destes documentos após a publicação dos mesmos;
3. Todo participante que se sentir constrangido ou incomodado durante uma situação de teste pode interromper o teste e estará fazendo um favor à equipe se registrar por escrito as razões ou sensações que o levaram a esta atitude. A equipe fica obrigada a descartar o teste para fins da avaliação a que se destinaria;
4. Os participantes que forem menores de idade terão, obrigatoriamente, que apresentar o consentimento de seu responsável, para participação no estudo, o qual será declarado ciente do estudo a ser realizado através de sua assinatura no presente Termo de Compromisso;
5. Todo participante tem direito de expressar por escrito, na data do teste, qualquer restrição ou condição adicional que lhe pareça aplicar-se aos itens acima enumerados (1, 2, 3 e 4). A equipe se compromete a observá-las com rigor e entende que, na ausência de tal manifestação, o

participante concorda que rejam o comportamento ético da equipe somente as condições impressas no presente documento;

6. A equipe tem direito de utilizar os dados dos testes, mantidas as condições acima mencionadas, para quaisquer fins acadêmicos, pedagógicos e/ou de desenvolvimento contemplados por seus membros.

Se de acordo e apto a realizar o teste, por favor, peço que responda este email com "*Eu, <seu nome completo>, concordo com o termo referido e aceito participar no teste de livre e espontânea vontade.*"

Mais uma vez, obrigado!

Germano Mergel