UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

SCAS-FUZZY: UMA ESTRATÉGIA SEMIAUTOMÁTICA PARA SELEÇÃO DE ESTUDOS PRIMÁRIOS EM ESTUDOS SECUNDÁRIOS

FÁBIO ROBERTO OCTAVIANO

ORIENTADORA: PROFª. DRª. SANDRA CAMARGO PINTO FERRAZ FABBRI

São Carlos - SP
Janeiro/2018

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

SCAS-FUZZY: UMA ESTRATÉGIA SEMIAUTOMÁTICA PARA SELEÇÃO DE ESTUDOS PRIMÁRIOS EM ESTUDOS SECUNDÁRIOS

FÁBIO ROBERTO OCTAVIANO

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Doutor em Ciência da Computação, área de concentração: Engenharia de Software
Orientadora: Profª. Drª. Sandra C. P. F. Fabbri

São Carlos - SP
Janeiro/2018

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia

Programa de Pós-Graduação em Ciência da Computação

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a defesa de tese de doutorado do(a) candidato(a) **Fábio Roberto Octaviano**, realizada em **23 de janeiro de 2018.**

Prof(a). Dr(a). Sandra Camargo P. F. Fabbri
(UFSCar)

****
Prof(a). Dr(a). Auri Marcelo Rizzo Vincenzi
(UFSCar)

****
Prof(a). Dr(a). Fabiano Cutigi Ferrari
(UFSCar)

****
Prof(a). Dr(a). Tayana Uchoa Conte
(UFAM)

****
Prof(a). Dr(a). Marcos Kalinowski
(PUC)

Certifico que a defesa realizou-se com a participação à distância dos membros **Tayana Uchoa Conte, Marcos Kalinowski, Fabiano Cutigi Ferrari e Auri Marcelo Rizzo Vincenzi** e depois das arguições e deliberações realizadas, os participantes à distância estão de acordo com o conteúdo do parecer da banca examinadora redigido neste relatório de defesa.

Prof(a). Dr(a). Sandra Camargo P. F. Fabbri
Presidente da Comissão Examinadora
(UFSCar)

AGRADECIMENTO

Primeiramente agradeço a Deus por sempre estar presente em minha vida, me abençoando, protegendo e dando forças para concluir este trabalho, e à Virgem Maria, por sua benção e intercessão materna.

À minha orientadora Prof. Sandra Fabbri, com quem tenho o imenso prazer de conviver há muitos anos e a quem muito admiro. Obrigado por acreditar em mim e permitir a realização deste trabalho.

Agradeço aos meus pais Clóvis e Wayni, que me deram a oportunidade de estudar e sempre me incentivaram na vida acadêmica, para que eu pudesse me tornar o melhor aluno possível, e agora pesquisador também. Aos meus irmãos Paulo e Carlos, os quais sempre foram ótimos exemplos de alunos e notas escolares.

À minha amada esposa Cassiana pela compreensão, apoio e carinho em todos os momentos e por sempre estar ao meu lado, e à minha filha Alana – minha Branca de Neve – por ser tão fofa e entender que não pude lhe dar a devida atenção em alguns momentos por conta deste trabalho. Vocês realmente são grandes presentes de Deus em minha vida, amo vocês!

Agradeço ao meu colega de laboratório Cleiton Silva por todo o suporte técnico e sempre estar à disposição nos momentos que precisei de seu auxílio. Ao meu grande amigo e colega de laboratório André Di Thommazo, por algumas explicações importantes, incentivo e amizade de muitos anos. Aos demais colegas e ex-colegas do LaPES: Elis, Kamilla, Guilherme “Baiano”, Deyse, Anderson, Juciara, Daniel, Arlindo, Bento, Abade e Rafael. Foram momentos incríveis, churrascos deliciosos e histórias engraçadas ao lado de vocês. Nossa amizade durará para sempre!

À minha colega Katia Felizardo, pela colaboração sobretudo no início deste trabalho.

Aos colegas professores e aos dirigentes do Instituto Federal de São Paulo (IFSP) que permitiram minha dedicação de maneira integral a esta pesquisa nos últimos anos do doutorado, o que contribuiu e muito para um trabalho de maior qualidade.

Aos funcionários e professores do Departamento de Computação da Universidade Federal de São Carlos (UFSCar).

Ao Programa Erasmus BE Mundus e ao Prof. Paolo Bottoni da Università La Sapienza di Roma, por terem me aceito para uma experiência incrível de mobilidade. *Grazie mille per tutto il vostro supporto*!

Enfim, meus sinceros agradecimentos a todos que conviveram comigo nesses anos todos e foram importantes, cada qual a seu modo, para a realização deste trabalho.

"Talvez não tenha conseguido fazer o melhor, mas lutei para que o melhor fosse feito. Não sou o que deveria ser, mas Graças a Deus, não sou o que era antes"

Marthin Luther King

RESUMO

Contexto: revisão sistemática e mapeamento sistemático são estudos secundários utilizados para localizar e agregar evidências relevantes da literatura sobre uma questão de pesquisa de interesse. Uma das atividades associadas aos estudos secundários é a seleção de estudos primários, que é uma tarefa manual e que pode demandar grande esforço dos pesquisadores. A qualidade da seleção de estudos primários afeta diretamente a qualidade geral dos estudos secundários. **Objetivo:** propor uma estratégia denominada SCAS-Fuzzy (*Score Citation semi-Automatic Selection using Fuzzy set)* para automatizar parte da atividade de seleção de estudos primários, minimizando o esforço consumido nessa atividade, mas mantendo a qualidade da seleção. **Metodologia:** inicialmente foi definida a estratégia semiautomática SCAS para seleção de estudos primários com base em dois parâmetros: o *score* e a ocorrência ou não de citações para um estudo. Ela foi avaliada por meio de um estudo de caso e de um experimento, que mostraram resultados promissores que motivaram a investigação de formas de melhorá-la. Criou-se então um coeficiente de citação, que passou a considerar a quantidade de citações recebidas por um estudo e o ano de publicação do mesmo. Além disso, utilizou-se lógica fuzzy para classificação dos estudos, agora com base no *score* e no novo coeficiente de citação. A estratégia melhorada, denominada SCAS-Fuzzy foi avaliada por meio de um estudo de caso. **Resultados:** o estudo de caso mostrou que, para as cinco revisões sistemáticas consideradas, a redução média de esforço aplicando a estratégia SCAS-Fuzzy foi de 39,1% e o percentual de erro foi de 0,3% para erros de exclusão e 3,3% para erros de inclusão, quando comparada à revisão manual, mostrando um nível de concordância substancial para com os revisores. Em comparação à estratégia original SCAS, os resultados são mais consistentes também. **Conclusão:** com os resultados é possível concluir que a estratégia SCAS-Fuzzy proporcionou resultados satisfatórios para redução de esforço da atividade de seleção inicial e com baixíssima quantidade de perde de evidências, mantendo a qualidade do estudo secundário, apresentando ainda resultados melhores como um todo em relação à estratégia SCAS inicialmente definida.

Palavras-chave: Seleção de estudos primários; estratégia de seleção; *Score Citation semi-Automatic Selection* (SCAS); revisão sistemática (RS); mapeamento sistemático (MS); engenharia de software baseada em evidência (ESBE); ferramenta StArt.

ABSTRACT

Context: Systematic review and systematic mapping are secondary studies used to identify and aggregate relevant literature evidence on a research question of interest. One of the activities associated with secondary studies is the selection of primary studies, which is a manual activity and may require great effort from the researchers. The quality of the selection of primary studies directly affects the overall quality of the secondary studies. **Objective:** To propose a strategy called SCAS-Fuzzy (Score Citation semi-Automatic Selection using Fuzzy set) to automate part of the activity of selection of primary studies, minimizing the effort required in this activity, but maintaining the quality of the selection. **Methodology:** it was proposed a semi-automatic strategy for the selection of primary studies based on two functionalities: the score and whether a study is cited or not, which was called SCAS. It was evaluated through a case study and an experiment, which showed promising results. Then, ways to improve it were investigated and a citation coefficient was created, which now considers the number of citations caught by a study and the year of its publication, besides using fuzzy logic for classification of studies, which is now based on their scores and citation coefficients. The improved strategy, called SCAS-Fuzzy, was evaluated through a case study. **Results:** the case study showed that, for the five systematic reviews considered, the general effort reduction applying the SCAS-Fuzzy strategy was 39.1% and the error percentage was 0.3% for automatically excluding studies and 3.3% for automatically including studies when compared to manual review, showing a substantial level of agreement with reviewers. **Conclusion:** based on the results it is possible to conclude that the SCAS-Fuzzy strategy provided satisfactory results to reduce the effort of the initial selection activity and with a very low amount of evidence loss, maintaining the quality of the secondary study, also presenting better results in general in relation to the original defined SCAS strategy.

Keywords: Primary study selection; selection strategy; Score Citation semi-Automatic Selection (SCAS); systematic literature review (SLR); systematic map (SM); evidence-based software engineering (EBSE); StArt tool.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

RS – Revisão Sistemática

MS – Mapeamento Sistemático

ES – Engenharia de Software

ESBE – Engenharia de Software Baseada em Evidências

SUMÁRIO

Capítulo 1

INTRODUÇÃO

Este capítulo apresenta o contexto do trabalho caracterizando as dificuldades na atividade de seleção inicial dos estudos primários, o que representa a principal motivação para propor a melhoria dessa atividade que compõe o processo dos estudos secundários. Além disso, descreve-se a metodologia adotada na condução deste trabalho e apresenta-se como o texto está organizado.

1.1 Contexto

Kitchenham, Dybå e Jørgensen (2004) afirmam que o objetivo da Engenharia de Software Baseada em Evidência (ESBE) deve ser o de fornecer meios pelos quais as melhores evidências de pesquisa possam ser integradas com a experiência prática e valores humanos no processo de tomada de decisões referente ao desenvolvimento e manutenção de software. Pode ser utilizada também para auxiliar a indústria na detecção e escolha das melhores tecnologias e métodos de desenvolvimento e manutenção, procedimentos de gerenciamento, entre outros. O termo ESBE é proveniente do termo Medicina Baseada em Evidência (MBE), adaptando a sua aplicação na área médica para a área de Engenharia de Software, mas mantendo suas principais características.

Na ESBE, as necessidades devem ser detectadas e transformadas em questões de pesquisa, buscando assim pelas melhores evidências para responder as questões formuladas. Jedlitschka e Ciolkowski (2004) afirmam que evidências têm por objetivo permitir que conclusões sobre um determinado tópico possam ser

generalizadas, com base em um conjunto de estudos relevantes sobre o tópico de pesquisa, incluindo estudos experimentais.

Dentre os métodos existentes para obtenção de evidências em ESBE estão os estudos secundários, que consistem em levantamentos bibliográficos feitos de maneira criteriosa e seguindo diretrizes preestabelecidas. Os estudos secundários utilizam estudos encontrados pelo pesquisador sobre um determinado tópico de pesquisa. Mafra e Travassos (2006) definem estudos primários como estudos que visam à caracterização de uma tecnologia ou método em uso dentro de um contexto específico. Os estudos primários normalmente fornecem resultados e avaliações referentes ao tópico de pesquisa desejado. A análise dos vários resultados e avaliações sobre uma determinada tecnologia ou método permite construir evidências sobre ele. Assim, em resumo, os estudos secundários normalmente fazem uso de estudos primários para evidenciar resultados sobre um tópico de pesquisa.

Os principais exemplos de estudos secundários são: Revisão Sistemática (RS) e Mapeamento Sistemático (MS). Em ambos, a seleção inicial de estudos relevantes é uma atividade essencial para a conclusão bem-sucedida de um estudo secundário e confiabilidade nos resultados obtidos.

A condução de estudos secundários, especialmente a de RSs, requer um grande rigor na pesquisa e muitas de suas atividades demandam bastante esforço do pesquisador (ZHANG, BABAR; TELL, 2011). Dentre as atividades que requerem mais esforço está a seleção de estudos, sobretudo quando o número de estudos identificados pelo pesquisador para determinados tópicos de pesquisa é muito elevado. Algumas tarefas exigidas na atividade de seleção podem ser subjetivas, como é o caso da aplicação de critérios de inclusão e exclusão de estudos. Outras tarefas podem ser suscetíveis a erros ou dependerem demasiadamente da experiência do pesquisador em relação ao tópico de pesquisa, como por exemplo, a determinação de palavras-chave referentes ao tópico de pesquisa abordado para identificação de estudos primários. Esses problemas são detalhados no Capítulo 2, à medida que os processos de RS e MS, e as atividades necessárias para suas conclusões, forem devidamente explicados.

Para tentar minimizar alguns desses problemas, estratégias de seleção bem formuladas e o suporte computacional podem ser instrumentos muito importantes. Dentre as estratégias de seleção inicial de estudos, há algumas propostas para

tentar melhorar e até mesmo semiautomatizar essa atividade. O objetivo é reduzir o esforço exigido para sua execução, mas ao mesmo tempo garantir que os estudos relevantes não sejam excluídos com a utilização da estratégia. Uma RS foi conduzida para identificar e descrever estratégias de seleção inicial de estudos propostas na área da Computação, e os resultados obtidos são mostrados em detalhes no Capítulo 3 deste trabalho, sendo possível deduzir que há uma lacuna na área, que ainda necessita de muita pesquisa. Há poucas estratégias para tentar aumentar a eficiência da atividade de seleção, sendo que as encontradas também não foram devidamente avaliadas.

As estratégias para tornar mais eficiente a atividade de seleção inicial precisam, normalmente, ser apoiadas por ferramentas computacionais que facilitem o uso por parte dos pesquisadores, sobretudo quando se referem a estratégias que visam à (semi) automatização dessa atividade. Essas ferramentas comumente fazem uso de recursos de processamento e mineração de texto para extrair informações dos estudos primários. Há algumas ferramentas que suportam todo o processo de RS, como é o caso da ferramenta StArt (FABBRI et al., 2016), e há outras que apoiam atividades específicas de uma RS. Entretanto, há poucas ferramentas identificadas que implementam estratégias para seleção de estudos primários, ou ao menos parte dessas estratégias.

Assim, esse é o contexto no qual esta pesquisa está inserida, sendo que a motivação e os objetivos que se esperam atingir são apresentados na próxima seção.

1.2 Motivação e Objetivos

Dado o contexto apresentado, pode-se sintetizar a motivação deste trabalho nos seguintes itens: (i) a importância dos estudos secundários para o levantamento de evidências sobre tópicos de pesquisa, (ii) a importância da atividade de seleção inicial de estudos no contexto de estudos secundários, (iii) as dificuldades existentes e esforço exigido para sua execução. Ressalta-se que, apesar dos estudos secundários serem aplicados nas diversas áreas de pesquisa, eles são especialmente importantes na área de engenharia de software, pois é através das

evidências produzidas pelos estudos secundários, que a ESBE se fortalece e pode dar apoio à comunidade da área. Assim, com base na motivação para esta tese os objetivos associados ao trabalho podem ser resumidos da seguinte forma:

- Explorar e combinar as funcionalidades de classificação de estudos com base em relevância de termos (*score*) e de número de citações dos estudos, com o intuito de elaborar uma estratégia semiautomática para apoiar a atividade de seleção inicial de estudos primários.
- Avaliar, por meio de estudos experimentais, a funcionalidade *score* existente na ferramenta StArt, que é um dos pilares da estratégia proposta neste trabalho, com o intuito de melhorá-la, se possível for, e torná-la mais eficaz no que diz respeito à identificação de estudos relevantes.
- Explorar e criar um coeficiente de citação de estudos, que pode ser baseado no número de citações e ano de publicação de estudos, uma vez que estudos mais antigos possuem chance maior de serem citados.
- Evoluir a ferramenta StArt de modo que a estratégia elaborada seja implementada e disponibilizada para uso da comunidade.

Assim, com base na motivação e nos objetivos descritos, a tese a ser defendida neste trabalho de doutorado pode ser redigida da seguinte forma:

"Considerando-se o processo de RS, é possível melhorar a eficiência da seleção inicial dos estudos, sem perda da qualidade, aplicando-se uma estratégia semiautomática baseada no *score,* associado às palavras-chave, e no coeficiente de citação, associado ao número de citações e ano de publicação, de um estudo".

1.3 Metodologia de Desenvolvimento do Trabalho

Para desenvolver esta pesquisa, fez-se um planejamento que considerasse:

1. Uma RS sobre as estratégias de seleção de estudos primários em estudos secundários existentes na literatura para levantamento do

estado da arte do tema e de possíveis alternativas que corroborassem a proposta da tese;

2. A proposição de uma estratégia semiautomática inicial para seleção de estudos primários em estudos secundários com base nas funcionalidades score e número de citações dos estudos;
3. Avaliação da estratégia inicial proposta por meio de estudos experimentais;
4. O refinamento da estratégia proposta com o uso de técnicas de inteligência computacional, buscando-se soluções que corroborassem a proposta da tese, com posterior avaliação em estudos experimentais;
5. A realização de estudo de caso para análise da estratégia de seleção semiautomática refinada e comparação de resultados com relação à estratégia de seleção inicialmente proposta.
6. Disponibilização da estratégia de seleção proposta em uma ferramenta de suporte computacional para que a mesma possa ser utilizada pela comunidade de pesquisadores que realizam estudos secundários.

Antes da proposição da estratégia inicial, foi conduzida uma RS seguindo as diretrizes propostas por Kitchenham e Charters (2007) para identificação das estratégias (automáticas ou não) existentes na literatura referentes à seleção de estudos primários em estudos secundários. Os resultados mostram que poucas estratégias foram encontradas e a maioria avaliada de forma bem superficial, como detalhado no Capítulo 3, o que permitiu a identificação da lacuna existente, fazendo com que a estratégia aqui apresentada agregasse conhecimento à área dos estudos secundários, caracterizando a originalidade deste trabalho.

Foi, então, criada uma estratégia que auxiliasse a atividade de seleção inicial dos estudos com base nas duas funcionalidades: *score* e número de citações dos estudos recuperados. A estratégia, denominada SCAS (*Score Citation semi-Automatic Selection*), permite a semiautomatização da atividade de seleção inicial de estudos nos processos de estudos secundários, dividindo os estudos em quadrantes com base em seus scores e números de citações, recomendando a inclusão automática dos estudos pertencentes ao quadrante 1, a exclusão automática dos estudos pertencentes ao quadrante 4, e a revisão manual dos estudos pertencentes

aos quadrantes 2 e 3, como apresentado detalhadamente na Seção 4.2 do Capítulo 4.

Para que a estratégia inicial proposta fosse avaliada, um estudo de caso contendo três RSs publicadas na literatura foi realizado com o intuito de analisar se as recomendações feitas pela estratégia SCAS estariam de acordo com a decisão tomada pelos revisores que conduziram as RSs. Os resultados foram animadores, como apresentado na Seção 4.3 do Capítulo 4, o que levou à publicação no periódico *Empirical Software Engineering* (OCTAVIANO et al, 2015). Posteriormente, foi realizado um experimento com alunos de pós-graduação (doutorandos) de algumas áreas de pesquisa da Universidade Federal de São Carlos durante uma disciplina de revisão sistemática oferecida pelo Programa de Pós-Graduação do Departamento de Computação. O experimento foi planejado de acordo com o paradigma GQM (Goal, Question, Metric) (BASILI; CALDIERA; ROMBACH, 1994). Os alunos aplicaram a estratégia SCAS em RSs conduzidas por eles e compararam as decisões sugeridas pela estratégia com as decisões que tomariam conduzindo as RSs manualmente. Os resultados foram bons, como apresentado na Seção 4.4 do Capítulo 4, o que levou à publicação na XX *International Conference on Evaluation and Assessment in Software Engineering* (EASE'16) (OCTAVIANO; SILVA; FABBRI, 2016).

Com a aplicação dos estudos experimentais, foram detectados pontos de possíveis melhorias na estratégia SCAS proposta inicialmente, como apresentado na Seção 5.2 do Capítulo 5. Dentre esses pontos, destaca-se a criação de um coeficiente de citação para os estudos, o qual passa a considerar o número de citações e o ano de publicação de um estudo, e a possibilidade do uso de inteligência artificial com a inclusão de recursos de lógica fuzzy e algoritmos genéticos na estratégia. Para tanto, foram necessárias a compreensão dos fundamentos dessas técnicas de inteligência computacional e a investigação de suas aplicações no tema desta tese. As melhorias realizadas levaram à proposição de uma estratégia semiautomática para seleção de estudos primários em estudos secundários denominada SCAS-Fuzzy, conforme apresentado na Seção 5.3 do Capítulo 5.

Um novo estudo de caso foi realizado para avaliação da nova estratégia SCAS-Fuzzy proposta considerando cinco RSs, sendo quatro publicadas na literatura e conduzidas por revisores experimentes e uma delas sendo a RS

conduzida para levantamento bibliográfico desta tese e apresentada no Capítulo 3. O estudo de caso tinha os objetivos de avaliar de a estratégia SCAS-Fuzzy apresentaria melhores resultados do que a estratégia SCAS inicialmente proposta, bem como de avaliar qual técnica para definição das funções de pertinência utilizados no sistema de inferência fuzzy definido seria a mais indicada. Os resultados mostraram que a estratégia SCAS-Fuzzy é mais precisa que a estratégia SCAS no que diz respeito à inclusão e exclusão automática de estudos, e que o uso de algoritmos genéticos na definição das funções de pertinência utilizadas no sistema fuzzy é positivo, conforme detalhado no Capítulo 6.

Para que a estratégia pudesse ser avaliada e, posteriormente, disponibilizada para uso da comunidade, torna-se fundamental sua implementação em uma ferramenta de apoio a estudos secundários. Assim, um módulo que implemente a estratégia SCAS-Fuzzy foi criado na ferramenta StArt, fazendo uso da funcionalidade pré-existente *score* e da nova funcionalidade implementada coeficiente de citação. Essas funcionalidades foram utilizadas como variáveis de entrada do sistema de inferência fuzzy, também incorporado à ferramenta, que gera a saída contendo a classificação de cada estudo (incluído automaticamente, excluído automaticamente ou revisão manual). Uma nova versão da ferramenta StArt contendo a estratégia semiautomática proposta está quase finalizada e será disponibilizada em breve para a comunidade. Para a realização dos estudos experimentais, foi utilizada uma versão alpha da ferramenta que inclui a estratégia proposta.

1.4 Organização do Trabalho

Esta tese está organizada em sete capítulos e dois apêndices. Este capítulo apresentou o contexto no qual a pesquisa está inserida, bem como a motivação, os objetivos e a metodologia de pesquisa seguida.

No Capítulo 2 é apresentada a Fundamentação Teórica dos temas que estão relacionados à pesquisa. Esses temas são ESBE, estudos secundários (revisões e mapeamentos sistemáticos) e os desafios referentes à atividade de seleção inicial de estudos, uma vez que essa atividade é o foco deste trabalho. Além disso, recursos

de mineração de texto e estudos experimentais também são apresentados no capítulo, visto que fazem parte do contexto e da avaliação da pesquisa.

No Capítulo 3 é apresentada a RS que foi conduzida para identificar e descrever as estratégias de seleção inicial de estudos propostas na literatura, bem como a maneira com que elas foram avaliadas, identificando lacunas de pesquisa na área.

No Capítulo 4 é apresentada a estratégica SCAS inicialmente proposta e seu funcionamento, bem como o estudo de caso e o experimento realizados para sua avaliação inicial.

No Capítulo 5 é apresentada a estratégica SCAS-Fuzzy com as melhorias realizadas em relação à estratégia SCAS proposta inicialmente, descrevendo seu funcionamento e recursos utilizados.

No Capítulo 6 é apresentado um estudo de caso para avaliação da estratégia SCAS-Fuzzy e sua comparação com a estratégia SCAS, discutindo-se os resultados obtidos.

No Capítulo 7 é apresentada a conclusão da tese, as contribuições da pesquisa e os trabalhos futuros.

Finalmente, no Apêndice A, a ferramenta StArt é brevemente descrita, visto que a estratégia definida será disponibilizada nela e, no Apêndice B, são apresentadas as *strings* de busca e o protocolo utilizados na RS conduzida e detalhada no Capítulo 3.

Capítulo 2

FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta uma visão geral dos principais temas relacionados com a tese proposta de modo a facilitar o entendimento da mesma.

2.1 Considerações Iniciais

Como apresentado no Capítulo 1, definiu-se uma estratégia semiautomática para a atividade de seleção inicial de estudos primários em estudos secundários, melhorando a eficiência da seleção inicial dos estudos sem perda da qualidade. Após a difusão do conceito de ESBE em 2004, o uso de estudos secundários por pesquisadores de Engenharia de Software cresceu consideravelmente e surgiram dificuldades em sua utilização. Muitas pesquisas foram realizadas para descrever essas dificuldades, especialmente no que se refere à atividade de seleção inicial de estudos. Nesse sentido, o uso de suporte computacional pode prover grande contribuição para a execução de parte ou de todo o processo de estudos secundários. Recursos de mineração de texto podem auxiliar a atividade de seleção por meio do processamento dos estudos coletados. Estudos experimentais são métodos avaliativos importantes e bem aceitos pela comunidade científica, podendo ser empregados para avaliar estratégias e ferramentas propostas. Assim, este capítulo tem por objetivo apresentar os principais conceitos relacionados com esses temas mencionados.

O capítulo está organizado da seguinte forma: na Seção 2.2 são apresentados conceitos sobre ESBE e estudos secundários; na Seção 2.3 são relatados alguns desafios apresentados por pesquisadores referentes à atividade de

identificação e seleção inicial de estudos; na Seção 2.4 são detalhados conceitos e métodos de mineração de texto; na Seção 2.5 são apresentados conceitos sobre estudos experimentais e, por fim, na Seção 2.6 são apresentadas as considerações finais deste capítulo.

2.2 Estudos Secundários em ESBE

Os estudos secundários são os meios mais utilizados por pesquisadores para obtenção de evidências na ESBE. Por meio deles é possível identificar e recuperar estudos relevantes referentes a tópicos de pesquisa na área de Engenharia de Software. Ao analisar os estudos recuperados, torna-se possível responder questões de pesquisa com o objetivo de caracterizar ou descrever uma área ou definir o estado da arte de algum tema de pesquisa.

Os principais exemplos de estudos secundários são: revisão sistemática e mapeamento sistemático, detalhados a seguir.

2.2.1 Revisão Sistemática

Kitchenham (2004) define RS como um meio capaz de identificar, avaliar e interpretar todas as pesquisas relevantes sobre uma determinada questão de pesquisa ou tópico de interesse, fazendo uso de uma metodologia confiável, rigorosa e que possa ser auditada. A autora propõe diretrizes para a execução de uma RS, extraídas de três documentos da área médica e adaptadas para resolver os problemas específicos de Engenharia de Software. Com base em um levantamento de como a comunidade executava as RSs, Kitchenham e Charts (2007) revisam as diretrizes inicialmente propostas e acrescentam novas visando a aumentar a qualidade na execução de RSs. Assim, as atividades necessárias para realização de uma RS são divididas em três fases: planejamento, execução e divulgação dos resultados. Felizardo et al (2017) abordam detalhadamente as fases do processo de uma RS, respondendo perguntas frequentemente realizadas por pesquisadores durante cada atividade de uma RS.

Na fase de planejamento, as atividades a serem realizadas são:

a) **Identificação da necessidade da RS:** o pesquisador deve realmente avaliar se é necessária a execução de uma nova RS e se já não há uma RS publicada sobre o tópico de interesse que satisfaça sua necessidade.

b) **Especificação das questões de pesquisa:** é a atividade principal da RS, pois as questões de pesquisa guiarão o pesquisador durante toda a RS. Para auxiliar na elaboração das questões, Petticrew e Roberts (2005) sugerem a utilização dos critérios PICOC (*Population, Intervention, Comparison, Outcome, Context*), que consiste em estruturar cada questão de pesquisa em função da população (pessoas, grupos ou áreas que serão afetadas pela intervenção), intervenção (metodologia, procedimento, tecnologia ou ferramenta investigadas), comparação ou controle (metodologia, procedimento, tecnologia ou ferramenta padrão ou bem conceituadas com as quais a intervenção deve ser comparada), resultados (melhorias esperadas em termos de confiabilidade, redução de custo, entre outros) e contexto (no qual a comparação ocorre ou os resultados têm efeito). Os critérios PICOC foram propostos para a área de Ciências Sociais e estende o método original da Medicina composto por população, intervenção e resultados (*population*, *intervention* e *outcome*, respectivamente). Staples e Niazi (2007) recomendam limitar o escopo da RS por meio de questões de pesquisa descritas de maneira clara e objetiva, pois elas influenciarão diretamente na seleção de estudos e na extração e análise dos dados.

c) **Criação de um protocolo da RS:** é de suma importância para a realização de uma RS. Nele são definidos todos os métodos e critérios que serão utilizados durante as atividades da RS. O objetivo do protocolo é a redução do viés de pesquisadores, como por exemplo, na seleção de estudos, por meio da utilização de critérios de seleção bem definidos, evitando ao máximo a subjetividade para realização dessa e de outras atividades. Os campos principais de um protocolo devem ser: a descrição da RS, as questões de pesquisa, a estratégia que será utilizada para buscar estudos primários (incluindo os locais de pesquisa), os critérios de inclusão e exclusão de estudos, o

procedimento para aplicação dos critérios de inclusão e exclusão, critérios de qualidade para avaliação dos estudos, o formulário de extração de dados que será preenchido para cada estudo relevante, a maneira como os dados extraídos serão sumarizados no final da RS, e a estratégia de divulgação dos resultados da RS. Opcionalmente pode-se descrever o cronograma para execução da RS.

Devido à importância do protocolo e com base em lições aprendidas ao executar RSs, Brereton et al. (2007) sugerem recomendações aos pesquisadores referentes à criação do protocolo: (i) estar atento à possibilidade de revisar as questões de pesquisa, à medida que o entendimento do problema aumente no decorrer da RS; (ii) fazer com que todos os pesquisadores envolvidos na RS participem da criação do protocolo; (iii) fazer um piloto considerando alguns estudos primários para averiguar se não há necessidade de modificações; e (iv) sugerem que, em muitos casos, um mapeamento superficial da área de pesquisa realizado antes de iniciar a RS pode auxiliar na criação do protocolo.

d) **Avaliação do protocolo:** o ideal é que o protocolo criado seja avaliado por pesquisadores mais experientes. Por exemplo, alunos de doutorados deveriam solicitar que seus orientadores fizessem a avaliação de seus protocolos.

Na fase de execução, as atividades a serem realizadas são:

e) **Identificação de estudos:** como o objetivo de uma RS é identificar o maior número possível de estudos relevantes sobre o tópico de pesquisa, uma estratégia de busca deve ser planejada, normalmente com o auxílio de pesquisadores experientes, e documentada. A estratégia deve basicamente dizer onde serão realizadas as buscas por estudos (sejam eles primários ou secundários, como por exemplo RSs já executadas sobre o mesmo tópico de pesquisa) e qual ou quais técnicas de busca serão utilizadas. Uma técnica amplamente utilizada na área de Engenharia de Software é a busca com *strings*, na qual *strings* de busca são criadas pelo pesquisador considerando um conjunto de palavras-chave e suas combinações lógicas, e executadas em bases de dados *online* que indexam trabalhos científicos

publicados tais como artigos de congressos e periódicos. É importante que o pesquisador considere todos os sinônimos conhecidos das palavras-chave referentes ao tópico de pesquisa para que o maior número possível de estudos seja identificado, considerando também diferentes idiomas quando aplicável. Dieste e Pádua (2007) comparam diferentes composições de *strings* utilizando combinações diversas de palavras-chave e mostram como os resultados retornados podem ser muito distintos dependendo da combinação escolhida. A documentação da estratégia de busca servirá para que leitores avaliem a robustez da pesquisa e consigam replicá-la da forma mais transparente possível, como por exemplo, as *strings* de busca utilizadas, em quais bases de dados elas foram executadas e quais os estudos retornados. Brereton et al. (2007) afirmam que o maior número possível de fontes de dados deve ser considerado e, no caso de busca apenas por artigos de conferências e revistas, tal decisão precisa ser justificada.

Outra técnica que tem ganho bastante força na identificação de estudos primários é o *snowballing*, que consiste em avaliar a lista de referências de estudos já conhecidos para identificar novos estudos relevantes para a pesquisa. Wohlin (2014) apresenta diretrizes para a aplicação da técnica *snowballing* de forma mais adequada. Ainda faz a definição dos termos *snowballing backward*, nome dado à busca e análise dos estudos (referências) citados pelos estudos primários já conhecidos, e *snowballing forward*, nome dado à busca e análise de estudos que citam os estudos primários já conhecidos. Wohlin (2016) recomenda realizar *snowballing* utilizando o texto completo e avaliando os trechos de texto que representam uma citação. Silva (2017) propõe critérios de priorização de estudos primários para *snowballing* com conjunto inicial formado por buscas com *strings* para melhorar a eficiência na aplicação da técnica, fornecendo suporte computacional para sua aplicação.

Kitchenham e Brereton (2013) sugerem o uso da técnica *snowballing* (*backward* e *forward*) como alternativa à busca por estudos por meio de *strings* de busca, podendo ser utilizadas complementarmente em

alguns casos. Mourão et al (2017) sugerem a aplicação de uma estratégia híbrida para identificação de estudos primários, aplicando *string* de busca em uma base de dados específica (Scopus) para identificar um conjunto inicial de estudos relevantes (chamado *seed*) e aplicação de *snowballing backward* e *forward* nesses estudos para identificação de novos estudos.

f) **Seleção inicial de estudos:** consiste na aplicação dos critérios de inclusão e de inclusão definidos no protocolo a partir da leitura do título, resumo (*abstract*) e palavras-chave de cada estudo identificado. Ao final dessa atividade, espera-se que os estudos claramente irrelevantes do conjunto inicial de estudos sejam excluídos da RS. Os remanescentes são estudos potencialmente relevantes e devem ser mais profundamente investigados.

g) **Avaliação da qualidade dos estudos selecionados:** além dos critérios de inclusão e exclusão de estudos, a qualidade dos estudos analisados deve ser considerada. Normalmente critérios de qualidade podem ser definidos pelo pesquisador e são utilizados para explicar diferenças em resultados apresentados por estudos, dar um peso maior a um estudo do conjunto em detrimento a outro, incluir ou excluir estudos de maneira mais criteriosa do que os critérios de inclusão e exclusão, e ainda detectar a necessidade de pesquisas adicionais sobre algum tema. O objetivo principal ao se utilizarem critérios de qualidade é minimizar o viés de resultados, isto é, reduzir a chance de que os resultados obtidos na RS sejam distantes dos resultados que deveriam ser realmente obtidos. Ao procurar avaliar a qualidade dos estudos primários por meio da aplicação de critérios de qualidade, o pesquisador busca aumentar a confiabilidade nos resultados que serão obtidos e também na generalização dos mesmos.

h) **Extração de dados:** um formulário de extração de dados deve ser criado no protocolo para que sejam coletadas todas as informações necessárias para responder as questões de pesquisa. O mesmo formulário deve ser preenchido pelo pesquisador para cada estudo relevante ao ler o seu texto completo. Após a leitura do texto completo, um estudo inicialmente considerado relevante pode ser considerado

agora como irrelevante, caso não se confirmem as expectativas do pesquisador quando leu o seu resumo. Em síntese, todos os estudos considerados relevantes após a leitura de seu texto completo devem ter o formulário de extração preenchido pelo pesquisador. Os campos de um formulário de extração podem permitir respostas descritivas, como por exemplo, a universidade ou grupo de pesquisa envolvido no estudo, respostas de uma lista pré-definida e que sejam exclusivas, como por exemplo se o estudo relata uma determinada técnica e cuja resposta dever ser sim ou não, e respostas de uma lista pré-definida, mas não exclusivas, como por exemplo qual ou quais técnicas o estudo relata, sendo que a lista foi pré-definida com quatro técnicas distintas. Kitchenham e Chartes (2007) afirmam que, se os critérios de qualidade forem utilizados para inclusão e exclusão de estudos, então um formulário à parte deve ser criado referente à avaliação de qualidade. No caso de os critérios de qualidade serem utilizados na síntese dos dados, então podem estar no mesmo formulário. Um piloto de extração deve ser executado considerando alguns estudos para validação do formulário de extração definido no protocolo, averiguando se modificações no mesmo são necessárias.

i) **Síntese dos resultados:** consiste na sumarização dos resultados coletados dos estudos incluídos na revisão. Essa sumarização pode ser qualitativa ou descritiva, mas também pode ser complementada com uma síntese quantitativa dos dados, por meio de técnicas estatísticas (metanálise). As questões de pesquisa devem ser respondidas. A sumarização qualitativa pode fazer uso de tabelas para exibir informações extraídas dos estudos, mostrando suas similaridades e diferenças, se apresentam resultados homogêneos ou heterogêneos, e as razões encontradas no caso de heterogeneidade, mas sempre olhando e tendo orientação das questões de pesquisa para não desviar do contexto da revisão. A sumarização quantitativa também pode incluir tabelas para auxiliar na leitura dos resultados. Quando possível, os resultados de diferentes estudos devem ser mostrados de forma comparativa. Dados provenientes de campos do formulário de extração com listas de respostas pré-definidas, como por

exemplo sim/não, podem ser calculados com base em alguma medida estatística como probabilidade (risco) ou razão de chances (*odds ratio*), entre outras. A utilização de gráficos é de grande valor para um melhor entendimento dos resultados e, nesse contexto, o gráfico de floresta (*forrest plot*) é indicado por conseguir exibir os meios e a variância das diferenças para cada estudo, porém outros tipos de gráfico podem ser utilizados.

Na fase de divulgação dos resultados, as atividades a serem realizadas são:

j) **Especificação da estratégia de divulgação:** consiste na definição e documentação via protocolo de como os pesquisadores pretendem comunicar os resultados da RS, que é uma atividade muito importante. No meio acadêmico, a publicação em congressos e revistas científicas é o mais adequado, por terem a garantia de uma revisão de especialistas, além de poder ser lida por um maior número de pessoas. Mas outras formas de divulgação podem ser consideradas, tais como relatórios técnicos, pôsteres, páginas Web e livros.

k) **Formatação do relatório principal da RS:** no caso de publicação em congressos ou periódicos, normalmente há um *template* a ser seguido. Em outros meios, o formato pode ser mais livre. Uma estrutura de conteúdo mínimo para divulgação dos resultados de uma RS é apresentada em (KITCHENHAM; CHARTES, 2007).

O processo de execução de uma RS não é estritamente sequencial, isto é, pode ser necessário que algumas atividades tenham de ser revisadas e melhoradas ao longo do processo, como por exemplo o protocolo. Fabbri et al. (2013) detalham a iteratividade necessária na execução de uma RS, mostrando os pontos passíveis de iteratividade dentro de cada fase da RS e também como as iteratividades podem ocorrer entre fases distintas. A iteratividade do processo é apresentada na Figura 2.1. Um exemplo de iteração dentro da mesma fase é quando, ao realizar um piloto para aprovação do protocolo, detecta-se a necessidade de adicionar novas palavras-chave no protocolo. Um exemplo de iteração entre fases distintas ocorre quando um pesquisador, durante a realização da seleção inicial de estudos, detecta novas palavras-chave que deveriam ser adicionadas à sua busca e, então, faz-se necessária a atualização do protocolo.

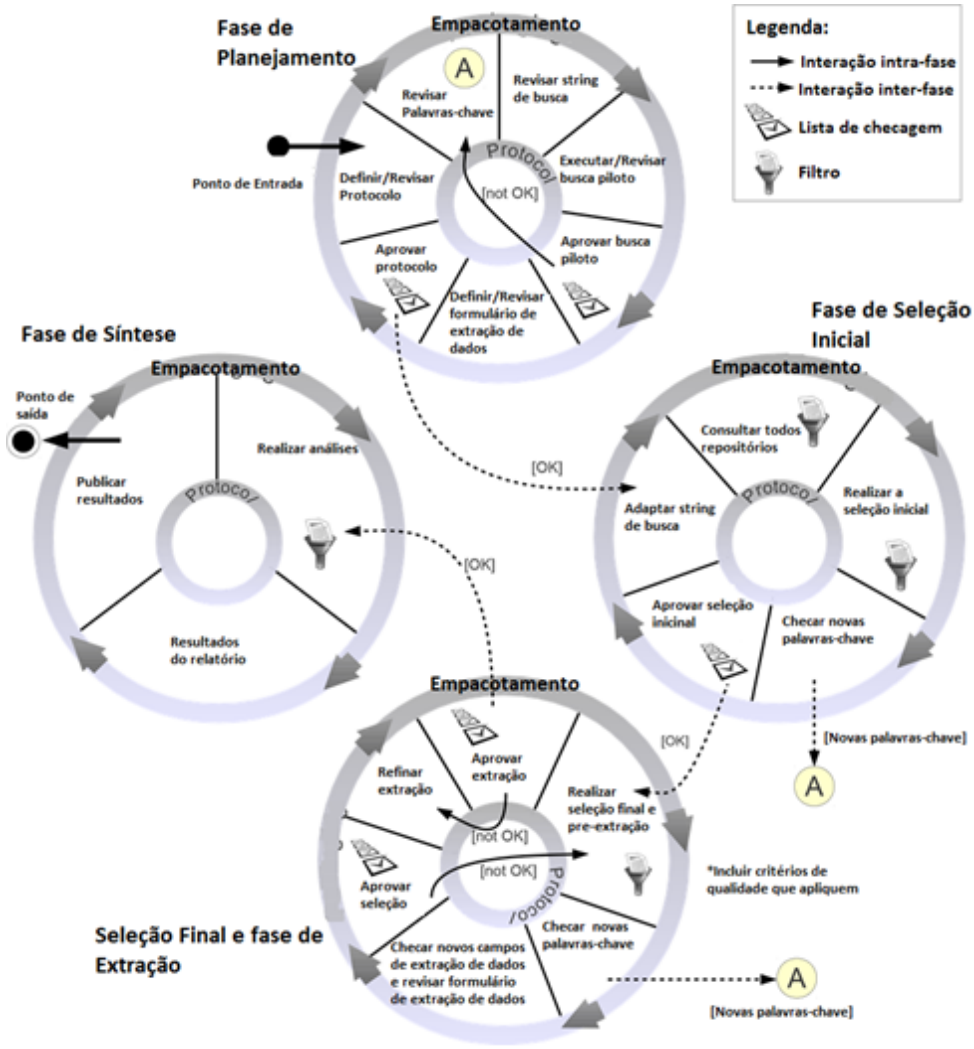


Figura 2.1. Visão geral das atividades de uma RS e os pontos possíveis de iteração (FABBRI et al., 2013)

Um tipo particular de RS ocorre quando um pesquisador opta por avaliar em sua pesquisa estudos secundários existentes sobre o tópico escolhido. Nesse caso, os estudos identificados que são utilizados para compor os resultados da revisão em questão são RSs ou MSs. A esse tipo de revisão dá-se o nome de revisão sistemática terciária. O procedimento para sua execução é o mesmo, porém os estudos primários utilizados na verdade são estudos secundários.

Kitchenham e Brereton (2013) avaliaram diversas RSs publicadas entre 2005 e a primeira metade de 2012 que discutiam técnicas para possíveis melhorias do processo de RS. O objetivo era o de analisar as diretrizes propostas em 2007 para a execução de RSs e detectar melhorias que pudessem ser adicionadas a elas. Como resultado, sugerem a remoção da diretriz de o pesquisador utilizar questões de pesquisa estruturadas para construção de *strings* de busca, pois concluíram que

elas tornam as *strings* de busca muito complexas e que, em muitos casos, muitas adaptações devem ser feitas para utilização em várias bases de dados. Outra recomendação que sugerem é o uso de uma busca manual limitada para auxiliar na construção de *strings* de busca e na avaliação do processo de busca, além da possibilidade de uso de *snowballing* para recuperação de estudos adicionais, conforme mencionado anteriormente.

2.2.2 Mapeamento Sistemático

MS é outro tipo de estudo secundário com origem na área médica e adaptado para a área de Engenharia de Software. Seu primeiro uso conhecido em Engenharia de Software foi apresentado por Bailey et al. (2007), porém sem uma explicação de como executá-lo. Petersen et al. (2008) propõem diretrizes para execução de MSs na área de Engenharia de Software, reportando também as diferenças entre uma RS e um MS e quando escolher entre um e outro, e ainda como podem ser combinados.

O objetivo principal de um MS é apresentar uma visão geral sobre uma área de pesquisa, identificando a quantidade e tipos de pesquisas existentes dentro dessa área e seus resultados. Exemplos de resultados obtidos com um MS podem ser a frequência de publicações ao longo dos anos para identificação de tendências, os locais de maior publicação de pesquisas na área, quais os métodos ou tecnologias mais empregadas, etc.

Petersen et al. (2008) apresentam a execução de um MS em cinco atividades principais: definição das questões de pesquisa, busca por estudos, seleção de estudos, leitura dos *abstracts* dos estudos e, por fim, a extração e mapeamento dos dados. Cada atividade deve produzir uma saída respectiva, conforme apresentado na Figura 2.2.

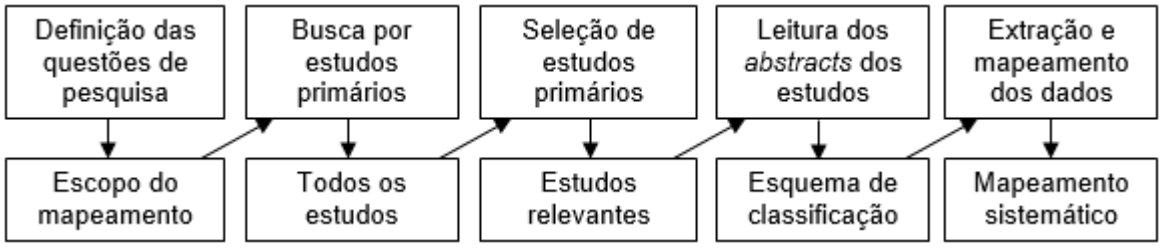


Figura 2.2. Atividades do mapeamento sistemático e saídas esperadas (Adaptado de PETERSEN *et al.*, 2008)

Algumas das atividades do MS são similares às da RS. As atividades necessárias são descritas a seguir.

a) **Definição das questões de pesquisa:** devem ser criadas questões de pesquisa com o objetivo de proporcionar uma visão geral sobre a área a ser mapeada. Exemplos de questões de pesquisa são: quais periódicos possuem publicações sobre projeto de software, quais são os métodos de desenvolvimentos mais utilizados na atualidade, que tipos de artigos estão sendo publicados ao longo do tempo, entre outras.

b) **Busca por estudos primários:** os autores sugerem a utilização da técnica de busca com *strings* em bases de dados científicas *online* ou a busca manual em anais de congressos e publicações em periódicos, com as *strings* de busca criadas por meio dos critérios PICOC.

c) **Seleção de estudos primários:** nessa atividade devem ser definidos critérios de inclusão e exclusão e aplicá-los em cada estudo, após leitura de seu título e *abstract*. O objetivo dessa atividade é selecionar os estudos que são potencialmente relevantes para responder as questões de pesquisa propostas.

d) **Leitura dos *abstract*s dos estudos:** nessa atividade deve ser feito o processo chamado pelos autores de *Keywording*. O processo deve ser realizado em duas etapas: (i) fazer a leitura do *abstract* e extrair palavras-chave e conceitos que reflitam a contribuição do estudo, contextualizando a pesquisa em andamento; e (ii) combinar as palavras-chave e conceitos extraídos dos *abstracts* para obter um entendimento em alto nível da natureza e contribuições da pesquisa, criando categorias de classificação nas quais os estudos serão distribuídos. Essa categorização é chamada de esquema de classificação, que é a saída esperada nessa atividade.

   Os autores alertam que, devido ao fato de que muitos *abstracts* da área de Engenharia de Software são pobres em conteúdo e mal estruturados, as seções de introdução e conclusão dos estudos podem ser avaliadas quando necessário.

e) **Extração e mapeamento dos dados:** nessa atividade, com base nos dados extraídos, os estudos primários são distribuídos no esquema de

classificação criado na atividade anterior. Com essa distribuição, é possível contar as frequências de publicação nas categorias definidas, permitindo identificar lacunas e áreas que precisam ser mais pesquisadas, que é a finalidade principal do mapeamento sistemático. Os resultados podem ser mostrados por meio de tabelas ou com recursos de visualização, como por exemplo o gráfico de bolhas.

2.2.3 Diferenças entre Revisão Sistemática e Mapeamento Sistemático

Kitchenham e Charters (2007) descrevem algumas diferenças básicas entre RS e MS. Entretanto, com base na experiência em executar ambos, Petersen et al. (2008) descrevem detalhadamente as principais diferenças existentes. São elas:

1. **Diferença de objetivos:** uma RS tem por objetivo estabelecer o estado da arte sobre o tópico de pesquisa, identificando, em muitos casos, as melhores práticas com base em evidências empíricas. Já um MS tem por objetivo a classificação, análise temática e identificação de locais e períodos de publicação de estudos, já que apenas uma análise superficial é feita em cada estudo;
2. **Diferença no processo:** a qualidade dos estudos não é avaliada no MS, mas é um item fundamental na RS, que busca caracterizar o estado da arte do tópico de pesquisa. Outra diferença está da extração dos dados, já que no MS é feita uma análise temática para definição de categorias, enquanto que na RS um nível mais profundo de extração é requerido por conta da síntese dos dados que deve ser realizada;
3. **Diferença em largura e profundidade:** em um MS mais estudos normalmente são considerados já que não são avaliados detalhadamente pelo pesquisador, ou seja, tem maior largura que uma RS. Já na RS, os estudos são selecionados com maior rigor, já que a avaliação de qualidade e os resultados apresentados em cada estudo são fundamentais, o que faz com que menos estudos sejam considerados relevantes para a sumarização dos resultados. Ou seja, uma RS tem uma avaliação mais profunda dos estudos;

4. **Diferença na classificação da área:** como muitos estudos primários são deficitários no que diz respeito ao rigor metodológico de pesquisa, pode ser que uma RS descarte estudos por questões de qualidade, o que pode comprometer a caracterização de uma área. Como a avaliação dos estudos primários em um MS é menos rigorosa, pode ser que em muitos casos ele seja mais recomendado para apresentar uma visão geral de uma área de pesquisa;
5. **Diferenças de validação:** Jørgensen e Shepperd (2007) apontam que muitos estudos são descritos incorretamente, isto é, seus autores relatam algo que na verdade não ocorre, como um experimento por exemplo, o que indica que um MS pode conter estudos categorizados incorretamente, uma vez que eles não são investigados com o mesmo nível de detalhamento e rigor que em uma RS. Embora o rigor tenha aumentado nos últimos anos em relação a isso, a chance desse problema ocorrer em um MS é bem maior do que em uma RS.
6. **Diferenças de acessibilidade industrial e relevância:** Petersen et al. (2008) relatam que, ao apresentar os resultados de RSs e de MSs para engenheiros de software na indústria, eles tiveram mais interesse nos dados dos MSs por julgarem muito complexos e difíceis de avaliar os dados das RSs. Os resultados detalhados das RSs são mais recomendados para grupos de pessoas específicos, normalmente com interesse muito grande no tópico pesquisado, isto é, especialistas no tema.

Embora existam diferenças, tanto Kitchenham e Charters (2007) quanto Petersen et al. (2008) defendem a ideia que os dois tipos de estudos secundários possam ser utilizados de forma complementar, normalmente executando um MS para um levantamento e categorização da área de pesquisa e, então, uma RS sobre tópicos específicos da área mapeada.

2.2.4 Ferramentas de Suporte a Estudos Secundários

Hassler et al. (2016) e Carver et al. (2013) reportam que há falta de suporte mais específico de ferramentas para a condução de estudos secundários, indicando que nenhuma ferramenta ainda está devidamente preparada para auxiliar

pesquisadores na realização de estudos secundários por completo. Entretanto, algumas ferramentas de suporte à execução de estudos secundários, em especial RSs, foram e estão sendo desenvolvidas nos últimos anos. Elas apoiam algumas atividades ou todo o processo de RS, dependendo da ferramenta.

Um mapeamento sistemático sobre ferramentas de suporte na área de Engenharia de Software é apresentado por Marshall e Brereton (2013). Os autores reportam dez ferramentas existentes: PEx, Revis, SLR-Tool, Hierarchical Cluster Explorer (HCE), Site Content Analyzer, UNITEX, SLuRp, SLRONT, StArt e DBPedia. Em relação à atividade de seleção de estudos, que é o foco desta pesquisa, PEx, Revis e SLRONT suportam apenas a atividade de seleção, enquanto SLR-Tool, SLuRp e StArt apoiam o processo todo de RS, que obviamente inclui a seleção de estudos. Marshall, Brereton e Kitchenham (2014) identificam uma quarta ferramenta que suporta todo o processo, que é a SLRTOOL e, nesse mesmo estudo, os autores compararam as quatro ferramentas (SLR-Tool, SLuRp, StArt e SLRTOOL) em termos de algumas características (economia, facilidade de instalação e configuração, suporte a atividades de RS e gerenciamento do processo). A pontuação geral de cada ferramenta foi calculada com base em pesos atribuídos a cada categoria mencionada. A classificação geral das ferramentas foi: SLuRp (65,4%), StArt (53,3%), SLR-Tool (53,2%) e SLRTOOL (45,1%).

2.3 Desafios Referentes à Identificação e Seleção de Estudos

A condução de estudos secundários, especialmente a RS, requer um grande rigor na pesquisa e muitas de suas atividades demandam bastante esforço do pesquisador (ZHANG; BABAR; TELL, 2011). O esforço será proporcionalmente maior quanto o número de estudos identificados pelo pesquisador. Zhang e Babar (2011) realizam uma investigação de como RSs são executadas na área de Engenharia de Software. Os autores concluem que os pesquisadores estão convencidos de que RSs são realmente valiosas e mais confiáveis do que revisões tradicionais, feitas de forma *ad-hoc*, porém o tempo necessário para concluir muitas de suas atividades é um desafio, sendo que 50% do esforço corresponde à etapa de

execução da RS (segunda fase). Dentre os pesquisadores entrevistados pelos autores, 43% dizem que a atividade da fase de execução da RS que consome mais tempo é a extração de dados, enquanto que 27% consideram a seleção de estudos primários a atividade mais desgastante.

A definição e aplicação de critérios de inclusão e exclusão de estudos é uma atividade que envolve subjetividade pois a interpretação de cada pesquisador pode ser diferente para alguns estudos analisados. Assim, diferentes níveis de experiência de pesquisadores em relação ao tema de pesquisa podem fazer com que haja discordância em relação à seleção de determinados estudos. Brereton et al. (2007) recomendam a participação de todos os envolvidos no processo desde o início para familiarizarem-se com o tema. A execução de um piloto de seleção, com a aplicação e discussão dos critérios por todos os pesquisadores envolvidos torna-se muito importante para minimizar o viés nessa atividade (BRERETON et al., 2007; FABBRI et al., 2013). Riaz et al. (2010) reportam que pesquisadores inexperientes gastam muito mais tempo do que os experientes em diversas atividades de uma RS, entre elas a aplicação de critérios de inclusão e exclusão de estudos. Assim, o bom entendimento e a aplicação desses critérios também pode ser considerado um desafio no que se refere à seleção de estudos primários.

Quando a RS envolve a utilização de critérios de qualidade como critérios de seleção de estudos, os seus resultados serão diretamente afetados com a aplicação correta ou não dos critérios de qualidade, o que pode ser outro desafio para os pesquisadores. Estudos inicialmente considerados relevantes podem ser excluídos se não satisfizerem as condições mínimas determinadas pelos critérios de qualidade definidos. Dybå, Dingsøyr e Hanssen (2007) relatam, por meio de suas experiências na condução de RSs, que a introdução de aspectos qualitativos em uma RS pode tornar a atividade de seleção de estudos mais complexa.

A pobreza ou a má formulação de muitos *abstracts* na área de Engenharia de Software também é um fator que pode interferir na seleção de estudos (BRERETON et al., 2007; DYBÅ, DINGSØYR e HANSSEN, 2007; KITCHENHAM et al., 2009; DYBÅ e DINGSØYR, 2008; RIAZ et al., 2010). A falta de informações relevantes pode fazer com que ou o pesquisador inclua desnecessariamente muitos estudos na seleção inicial, o que acarretaria num esforço ainda maior na segunda seleção com base no texto completo dos estudos, ou exclua indevidamente estudos que seriam relevantes, comprometendo o resultado final da RS. Brereton et al. (2007) defendem

que, pelo fato de os *abstracts* publicados na área de Tecnologia de Informação serem mal formulados, a seção de conclusão de cada estudo também deveria ser considerada, porém essa não é uma prática comum na seleção inicial de estudos. O esforço necessário para avaliar um estudo seria maior, além do tempo necessário para obtenção do texto completo do estudo para avaliação da seção de conclusão nele existente.

Em relação à identificação de estudos, um desafio é relacionado à terminologia utilizada pelos pesquisadores. Muitas vezes falta um padrão de terminologia, principalmente se a pesquisa for sobre um tópico muito recente e não houver termos já consolidados na área. Dieste, Grimán e Juristo (2009) relatam o quão difícil pode ser encontrar estudos primários sobre experimentos em Engenharia de Software devido à grande variedade de termos utilizados. Nesse sentido, o pesquisador precisa de muita atenção na atividade de identificação de estudos, considerando o maior número de termos possível na busca por estudos, como por exemplo na definição das palavras-chave da RS no protocolo e na criação das *strings* de busca que serão executadas nas bases de dados, que é a técnica mais frequentemente utilizada em Engenharia de Software. A criação e otimização de *strings* de busca é uma tarefa que demanda tempo e é bastante suscetível a erros (ZHANG; BABAR; TELL, 2011). Boell e Cezec-Kecmanovic (2011) ressaltam o quão difícil é criar uma *string* de busca ideal, uma vez que quanto mais ela for inclusiva (considerar mais termos), mais documentos irrelevantes recuperará, ao passo que quanto mais for restritiva (considerar menos termos), menos estudos relevantes serão recuperados.

A dificuldade de utilização das bases de dados é um fator que pode atrapalhar na identificação de estudos e deve ser considerado como um desafio. O pesquisador, muitas vezes, precisa adaptar suas *strings* de busca para cada base de dados desejada, pois não há um padrão de utilização de recursos dessas bases (KITCHENHAM; BRERETON, 2013). Algumas oferecem recursos mais sofisticados para execução das *strings* de busca, mas outras apenas recursos básicos, que ainda assim podem diferir entre bases distintas. Isso pode fazer com que o pesquisador não consiga recuperar estudos de determinadas bases por falta de conhecimento, ou ainda obter estudos indevidos por não conseguir adaptar as *strings* de busca para essas bases. Riaz et al. (2010) reportam como pesquisadores inexperientes tendem a ter problemas com a execução de *strings* de busca em bases de dados

online. Nesse sentido, a técnica *snowballing* pode ser uma alternativa para auxiliar na identificação de estudos primários.

Kuhrmann, Fernandéz e Daneva (2017) relatam as dificuldades em se aplicar as diretrizes propostas por Kitchenham e Charters (2007) e revisadas por Kitchenham e Brereton (2013) para a execução de estudos secundários na prática, sobretudo para pesquisadores inexperientes, em especial com relação às atividades de identificação e seleção de estudos primários. Os autores propõem, inclusive, algumas diretrizes auxiliares com base em suas experiências para colaborar com pesquisadores iniciantes na execução de tais atividades.

2.4 Mineração de Texto

Como mencionado anteriormente, algumas atividades dos estudos secundários, em especial da RS, requerem um grande esforço do pesquisador para serem concluídas, dentre elas a seleção inicial de estudos. Assim, o suporte computacional pode auxiliar na realização de muitas atividades dos estudos secundários. No que diz respeito à seleção inicial de estudos, foco principal deste trabalho, recursos de mineração de texto podem contribuir significativamente para auxiliar o pesquisador no processamento de títulos, *abstracts* e palavras-chave dos estudos primários e decidir por incluí-los ou não.

Feldman e Sanger (2007) definem mineração de texto como um processo de conhecimento intensivo em que o usuário interage com uma coleção de documentos por meio de um conjunto de ferramentas de análise. É originada das pesquisas em mineração de dados e seus sistemas de apoio possuem muitas características em comum, tais como rotinas de pré-processamento, algoritmos de descoberta de padrão e apresentação dos resultados por meio de ferramentas.

O objetivo da mineração de texto é extrair informações úteis de documentos de texto por meio da identificação e exploração de padrões nos mesmos. Como um documento é uma coleção de dados textuais não estruturados, a rotina de pré-processamento é fundamental no processo, visando a transformar os dados não estruturados em um formato intermediário estruturado. O pré-processamento se

baseia em técnicas de outras áreas, como por exemplo, recuperação de informação, linguística computacional e extração de informações (FELDMAN; SANGER, 2007).

Martins (2003) classifica as principais atividades relacionadas à atividade de mineração de texto em: (i) Categorização, que consiste em induzir uma classificação do documento de modo a determinar se pertence ou não a uma categoria pré-definida; (ii) Agrupamento (*Clustering*), que consiste em criar grupos finitos de documentos (*clusters*); e (iii) Sumarização, que consiste em reduzir o tamanho de um documento preservando seu significado.

Há algumas etapas que são comuns a todas as atividades de mineração de texto:

- **Coleta de documentos:** consiste em identificar documentos que sejam relevantes para o tema ou domínio sobre o qual se deseja extrair conhecimento.
- **Pré-processamento dos documentos:** consiste em extrair de textos escritos em linguagem natural uma representação estruturada, concisa e manipulável por algoritmos de agrupamento de textos. Para tanto, devem ser executadas atividades de tratamento e padronização dos textos, seleção dos termos mais significativos e, por fim, representação da coleção textual em um formato estruturado que preserve as principais características do texto (REZENDE; MARCACINI; MOURA, 2011).

  O tratamento e padronização normalmente se dá ao converter a coleção de documentos para textos sem formatação, pois os documentos podem estar todos em formatos bem distintos, o que dificulta o processamento.

  A seleção de termos representativos da coleção de documentos normalmente se dá ao aplicar alguma das técnicas a seguir (MARTINS, 2003):

  - **Eliminação de *stopwords*:** consiste na eliminação de termos sem sentido semântico para os documentos, tais como artigos, pronomes, advérbios e conjunções, com base numa lista pré-definida desses termos.
  - ***Stemming*:** consistem em uma normalização linguística, na qual as formas variantes de um termo são reduzidas ao seu radical

(*stem*), considerando termos derivados como sendo semelhantes.

- **TF-IDF:** consiste na aplicação do método de Luhn para seleção de termos. Utiliza as medidas TF (*Term Frequency*), que contabiliza a frequência absoluta de um termo no conjunto de documentos, e IDF (*Inverse Document Frequency*), que contabiliza o número de documentos em que termo aparece e o classifica de forma a diminuir o peso de termos que aparecem com maior frequência, como pode ser o caso de *stopwords*. Ao contabilizar a TF e ordenar o histograma resultante em ordem decrescente, obtém-se a Curva de Zipf, na qual o k-ésimo termo mais comum ocorre com frequência inversamente proporcional a k. Assim, os termos de alta frequência são julgados não relevantes por aparecerem na grande maioria dos textos e não trazer informações úteis, bem como os termos de baixa frequência por serem muito raros e, em geral, não possuírem caráter relevante. Desse modo, os termos intermediários são considerados os mais relevantes e são obtidos ao traçar subjetivamente pontos de corte superior e inferior da Curva de Zipf obtida (REZENDE; MARCACINI; MOURA, 2011).

Depois do pré-processamento dos documentos, tendo sido os termos mais representativos identificados, deve ocorrer a estruturação dos documentos por meio de uma representação de dados textuais. Para tanto, o modelo mais utilizado é o modelo espaço-vetorial, no qual cada documento é tratado como um vetor em um espaço multidimensional, e cada dimensão é um termo da coleção. Os textos são estruturados em um conjunto desordenado de termos no formato de uma matriz denominado "*bag of words*" (FELDMAN; SANGER, 2007). Essa matriz possui as dimensões de documento e termo, na qual o elemento di corresponde ao i-ésimo documento, tj representa o j-ésimo termo e aij é um valor que relaciona o i-ésimo documento com o j-ésimo termo, como pode ser observado na Figura 2.3. O valor do elemento aij pode indicar se um determinado termo está presente ou não em um dado

documento ou a importância ou distribuição do termo ao longo da coleção de documentos, como por exemplo, o valor de TF.

	t_N	t_N	...	t_M
d_1	a_{11}	a_{12}	...	a_{1M}
d_2	a_{21}	a_{22}	...	a_{2M}
⋮	⋮	⋮	⋱	⋮
d_N	a_{N1}	a_{N2}	...	a_{NM}

Figura 2.3. Exemplo da matriz documento-termo (REZENDE, MARCACINI; MOURA, 2011)

- **Extração de padrões:** consiste em aplicar técnicas de extração de conhecimento na coleção de documentos representada em um formato estruturado para descobrir padrões presentes nos documentos. Se os documentos estiverem estruturados no formato de *bag of words*, existem métodos específicos de Recuperação de Informação e Aprendizado de Máquina que podem ser utilizados para extração de padrões.
- **Avaliação e interpretação:** consiste em verificar se os resultados alcançados com a aplicação de mineração de texto são satisfatórios ou se etapas do processo devem ser refeitas. Para essa verificação, normalmente são utilizadas ferramentas de visualização e medidas estatísticas como *precision*/*recall* e F-*measure*. A avaliação e interpretação podem ser realizadas por um especialista do domínio ou usuário final (MARTINS, 2003).

No contexto deste trabalho, o processo de mineração de texto é um recurso importante para o processamento de estudos primários, pois os cálculos das funcionalidades *score* e coeficiente de citação, que são base para a estratégia proposta, estão diretamente vinculados a esse recurso. Estudos experimentais foram utilizados para avaliação do trabalho.

2.5 Estudos Experimentais em Engenharia de Software

Basili et al. (1996) sugerem que novos métodos, técnicas, linguagens e ferramentas propostos na área de Engenharia de Software não deveriam ser apresentados sem antes passarem por estudos experimentais, comparando-os com já existentes para evidenciar suas reais contribuições para a área. Os estudos experimentais em Engenharia de Software visam a caracterizar, avaliar, prever, controlar ou melhorar tanto os produtos, como também os processos, recursos, modelos ou teorias (TRAVASSOS; GUROV; AMARAL, 2002).

Há diferentes tipos de estudos experimentais, sendo que os três principais, de acordo com Wohlin et al. (2000), são: *survey*, estudo de caso e experimento, cabendo ao pesquisador decidir o mais apropriado com base no objeto de estudo que será avaliado.

O *survey* é uma investigação executada em retrospectiva. Pode ser utilizado, por exemplo, quando o objeto de estudo (uma ferramenta ou técnica) já foi usado por um determinado período. Os resultados de um *survey* podem ser generalizados para a população da qual a amostra foi selecionada, no entanto sua aplicação não oferece controle sobre a execução ou medição dos dados coletados. (WOHLIN et al., 2000). Normalmente, a coleta de dados é realizada por meio da aplicação de questionários à população, coletando muitas variáveis para análise. Pode ser de três tipos: (i) descritivo, quando o objetivo é determinar a distribuição características; (ii) explanatório, quando o objetivo é explicar a escolha de determinada técnica ou método; e (iii) exploratório, quando o objetivo é conduzi-lo preliminarmente para que uma investigação mais detalhada sobre um tópico de pesquisa seja realizada futuramente.

O estudo de caso é um estudo experimental utilizado para monitorar os projetos, atividades e atribuições (WOHLIN et al., 2000). Seu objetivo é observar um atributo específico, que pode ser entendido como um caso particular e, posteriormente, estabelecer relacionamentos entre atributos diferentes, que podem ser entendidos como princípios gerais (TRAVASSOS; GUROV; AMARAL, 2002). Tem por característica o baixo controle de sua execução por parte do pesquisador. No entanto, ao contrário do *survey*, o estudo de caso possui o controle sobre a medição das variáveis (TRAVASSOS; GUROV; AMARAL, 2002).

O experimento é o tipo mais indicado para confirmar as teorias e o conhecimento convencional, explorar os relacionamentos, avaliar a predição dos modelos ou validar medidas (TRAVASSOS, GUROV; AMARAL, 2002). Normalmente é realizado em laboratório e o pesquisador tem um elevado nível de controle sobre ele, tanto em relação ao processo como às variáveis envolvidas. Dessa forma, uma ou algumas variáveis são manipuladas enquanto as outras são mantidas fixas e, ao final, o resultado é medido (WOHLIN et al., 2000). Travassos e Barros (2003) afirmam que um experimento pode ser feito de quatro maneiras:

a) *In vitro*: executado e controlado em ambientes como laboratórios ou comunidades controladas, tais como universidades e grupos de pesquisa;
b) *In vivo*: executado com pessoas em seus próprios ambientes, ou seja, ocorre em circunstâncias reais;
c) *In virtuo*: executado em um ambiente virtual, composto por modelos numéricos, que são representações computacionais de elementos ou fenômenos do mundo real, permitindo a manipulação do modelo utilizado;
d) *In silico*: executado representando, além do objeto e do ambiente a serem estudados, o comportamento dos indivíduos envolvidos por meio de modelos computacionais, não permitindo manipulação.

Wohlin et al. (2000) mostram o processo de experimentação composto por cinco etapas: definição, planejamento, operação, análise e interpretação, e apresentação e empacotamento, conforme ilustrado na Figura 2.4.

Na etapa de definição deve ser caracterizado o problema e definidos os objetivos do experimento. Todos os aspectos importantes do experimento precisam ser definidos antes de seu planejamento e de sua execução. Wohlin et al. (2000) sugerem o uso do paradigma GQM (BASILI; CALDIERA; ROMBACH, 1994) para auxiliar nessa etapa.

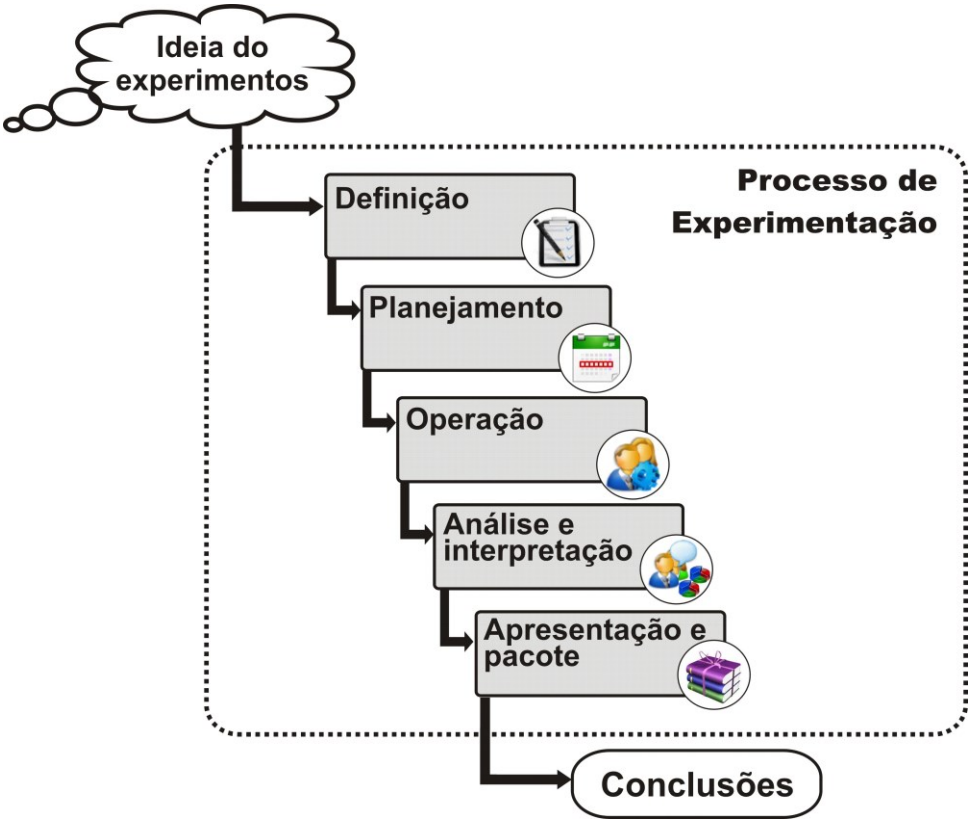


Figura 2.4. Processo de Experimentação (Adaptado de WOHLIN et al., 2000)

Na etapa de planejamento deve acontecer a definição e preparação de como ocorrerá o experimento, o que inclui a caracterização do ambiente no qual ele será executado, as hipóteses que serão investigadas, as variáveis independentes e dependentes utilizadas, o projeto experimental, a implementação do experimento e a avaliação da validade do experimento. A instrumentação do experimento deve ser definida também, o que inclui a definição de mecanismos de coleta (documento de consentimento, formulário de caracterização dos participantes, formulário de coleta de dados e formulários de *feedback*) e ferramentas necessárias para sua execução.

Na etapa de operação ocorre a preparação da execução do experimento (escolha e organização de participantes e materiais), a sua execução propriamente dita (participantes executam suas tarefas e ocorre a coleta de dados) e, por fim, a validação dos dados coletados (para averiguar se não é necessária uma nova coleta).

Na etapa de análise e interpretação, os dados coletados na etapa anterior devem ser analisados e interpretados. O uso de métodos estatísticos, quando aplicável, é recomendado para que as hipóteses sejam devidamente testadas.

Na última etapa, a de apresentação e empacotamento, deve ocorrer a apresentação e documentação dos resultados obtidos, com os objetivos principais de tornar o experimento replicável e de disponibilizar os resultados para a indústria e outros pesquisadores. Para tanto, a publicação de artigos em conferências e periódicos é um meio recomendado, além da elaboração de relatórios técnicos. O empacotamento é importante porque torna possível o armazenamento dos artefatos utilizados no experimento, impedindo a perda de informações importantes, o que auxilia no processo de replicação, que é uma das principais características de um experimento (TRAVASSOS; GUROV; AMARAL, 2002). Travassos, Gurov e Amaral (2002) propõem uma evolução no processo de experimentação inicialmente proposto no ano 2000 (WOHLIN et al., 2000), sugerindo que a etapa de apresentação e empacotamento seja executada em paralelo com as demais etapas, com o objetivo de evitar a perda de informações no decorrer do experimento.

2.6 Considerações Finais

Neste capítulo foram apresentados os temas relacionados diretamente com a estratégia apresentada neste trabalho – ESBE e estudos secundários, desafios relacionados às atividades de identificação e seleção de estudos e mineração de texto – assim como o tema de estudos experimentais, importante para avaliação da tese proposta.

Sobre EBSE e estudos secundários, foram caracterizados os tipos de estudos secundários mais utilizados em Engenharia de Software e diretrizes propostas paras suas utilizações, além das principais diferenças existentes entre eles.

Em relação aos desafios relacionados à identificação e seleção de estudos, foram reportados alguns dos desafios mais comuns encontrados por pesquisadores, o que podem indicar lacunas de pesquisa que podem ser mais bem exploradas. Uma RS sobre estratégias de pesquisa é apresentada no Capítulo 3.

Na seção sobre mineração de texto o intuito foi definir conceitos e apresentar o processo e algumas técnicas utilizadas, uma vez que tais recursos serão necessários para tornar viável esta proposta de tese.

Estudos experimentais foram apresentados a fim de elucidar como o processo de experimentação deve ser conduzido na área de Engenharia de Software e como eles são meios muito importantes para avaliação de técnicas, métodos, ferramentas e processos.

Todos os temas apresentados neste capítulo são passíveis de discussões mais extensas e completas. Contudo, optou-se por apresentar apenas os principais conceitos e descrever de maneira sucinta cada um deles.

Capítulo 3

REVISÃO SISTEMÁTICA SOBRE ESTRATÉGIAS DE SELEÇÃO DE ESTUDOS

Este capítulo apresenta a revisão sistemática realizada e que permitiu a identificação da lacuna de pesquisa explorada nesta tese.

3.1 Considerações Iniciais

Para auxiliar na elaboração deste trabalho, o seu autor, com suporte de sua orientadora, conduziu uma RS referente às estratégias de seleção de estudos primários encontradas na literatura, o que permitiu identificá-las, entender seus funcionamentos e detectar seus pontos positivos e negativos, visto que uma nova estratégia é proposta nesta tese.

O capítulo está organizado da seguinte forma: na Seção 3.2 é apresentado o método utilizado na RS; na Seção 3.3 são apresentados os resultados obtidos com a execução da RS, o que inclui as estratégias encontradas e uma discussão sobre seus pontos fortes e fracos; na Seção 3.4 são mencionadas as limitações da RS executada; e, por fim, na Seção 3.5 são apresentadas as considerações finais deste capítulo.

3.2 Método

A RS seguiu as diretrizes propostas por Kitchenham e Charters (2007) e apresentadas no Capítulo 2, Seção 2.2.1. Para auxiliar na execução de todas as fases da RS, foi utilizada a ferramenta StArt – State of the Art through Systematic Review (FABBRI et al., 2016), que é apresentada brevemente no Apêndice A deste trabalho. Para efeitos de replicação da RS, o protocolo utilizado está descrito no Apêndice B. Nas subseções a seguir são destacados alguns campos principais do protocolo.

3.2.1 Questões de Pesquisa

Foram definidas duas questões de pesquisa principais (QP) para a RS com o auxílio da aplicação dos critérios PICOC, como pode ser observado no protocolo do Apêndice B, que são:

a) QP1: Como funcionam as estratégias utilizadas para realizar a atividade de seleção inicial de estudos em RSs na área de computação?
b) QP2: Como funcionam as estratégias (semi) automáticas utilizadas para realizar a atividade de seleção inicial de estudos em RSs na área de computação?

3.2.2 Processo de Identificação de Estudos Primários

A busca por estudos primários foi realizada nas principais bases de dados *online* da área de Computação: SciVerse Scopus[1], ACM Digital Library[2], IEEE Xplore Digital Library[3] e Web of Science[4], conforme recomendação de Kitchenham e Brereton (2013), que sugerem o uso da IEEE Xplore e ACM Digital Library que são específicas da computação, além de ao menos outras duas de cunho mais geral,

[1] http://www.scopus.com

[2] http://dl.acm.org

[3] http://ieeexplore.ieee.org/Xplore/home.jps

[4] https://webofknowledge.com

dentre elas SciVerse Scopus e Web of Science. A pesquisa foi limitada à área da Computação, que é a área de interesse desta pesquisa.

As *strings* de busca executadas consideraram termos em inglês com o intuito de permitir que outras pessoas possam replicá-las, caso julguem interessante. Assim, a *string* de busca básica utilizada, e que precisou ser adaptada para cada base de dados, foi:

(selection OR screening) AND (studies OR papers OR articles OR evidence) AND ("systematic literature review" OR "systematic review" OR "systematic map")

As *strings* de busca executadas em são apresentadas no Apêndice B deste trabalho, com a finalidade de tornar a RS replicável por outros pesquisadores. Elas incluem os filtros adicionados por cada base em particular.

Os estudos primários recuperados foram avaliados e considerados satisfatórios pelos autores ao averiguarem que estudos relevantes conhecidos – (OCTAVIANO *et al.*, 2015), (OCTAVIANO; SILVA; FABBRI, 2016); (FELIZARDO et al, 2012) – foram recuperados. A exceção foi um único estudo que não foi encontrado nas bases de dados, mas era de publicação de um dos autores e precisou ser manualmente incluído no conjunto de estudos primários recuperados, como abordado na Seção 3.4. A distribuição dos estudos identificados é apresentada na Figura 3.1.

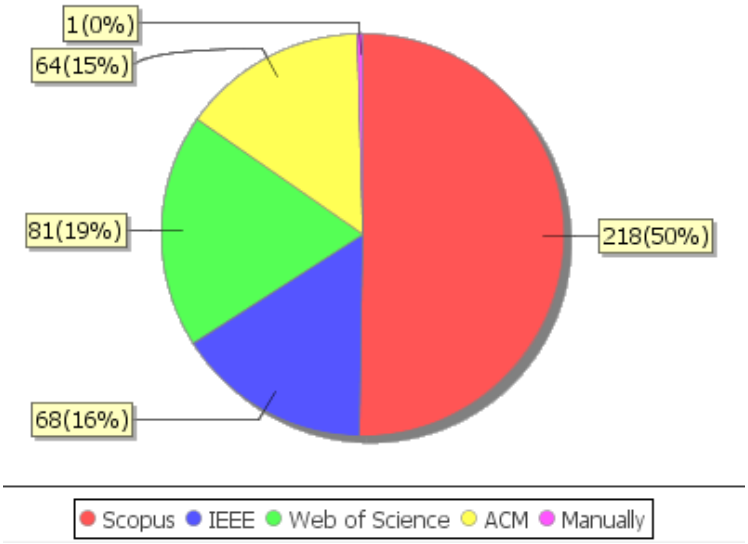


Figura 3.1. Distribuição dos estudos primários recuperados nas bases de dados

3.2.3 Critérios de Inclusão e Exclusão

Os critérios de inclusão e exclusão foram discutidos e definidos pelo autor deste trabalho e sua orientadora. Um piloto foi realizado considerando alguns estudos aleatoriamente a fim de averiguar se eram claros e suficientes. Após alguns ajustes, os critérios finais foram definidos. Dúvidas na aplicação dos mesmos durante a seleção de alguns estudos primários foram discutidas e uma decisão de incluí-lo ou não foi tomada em consenso.

A Tabela 3.1 apresenta os critérios de inclusão e exclusão finais definidos pelos autores para serem aplicados na atividade de seleção inicial de estudos primários.

Tabela 3.1. Critérios de Inclusão e Exclusão definidos para a RS

Tipo	Código	Descrição
Inclusão	CI1	O estudo relata ao menos uma estratégia para a seleção inicial de estudos
Exclusão	CE1	O estudo não relata uma estratégia para a seleção inicial de estudos
	CE2	O estudo não está escrito em inglês
	CE3	Não é possível encontrar o texto completo do estudo

Os autores definiram as seguintes regras para determinar se um estudo deveria ser incluído ou não na atividade de seleção inicial:

- O estudo deve ser incluído se o critério de inclusão for atribuído a ele;
- O estudo só deve ser excluído se ao menos um critério de exclusão for atribuído a ele;
- No caso de um estudo possuir tanto critérios de inclusão quanto de exclusão atribuídos a ele, como por exemplo, relatar uma estratégia de seleção, mas não estar em inglês, o mesmo deve ser excluído.

3.2.4 Formulário de Extração de Dados

O formulário de extração de dados também foi discutido e definido pelo autor deste trabalho e sua orientadora. Um piloto foi realizado considerando alguns

estudos relevantes previamente conhecidos pelos autores com o intuito de verificar se os campos do formulário de extração eram claros e suficientes. Mesmo com a execução do piloto, durante a atividade de extração de dados, foi detectada a necessidade de adicionar dois novos campos ao formulário de extração, conforme previsto em (FABBRI et al., 2013) e ilustrado na Figura 2.1. Dessa forma, os estudos primários que já haviam sido processados foram revisitados para analisar se não existiam dados para serem extraídos referentes aos novos campos. Os campos do formulário de extração final são apresentados na Tabela 3.2. A ferramenta StArt, utilizada na condução da RS, permite que os campos do formulário de extração sejam criados com opções de respostas pré-definidas no formato de alternativas (exibindo uma lista e permitindo a escolha de apenas uma resposta) ou de múltipla escolha (exibindo uma lista e permitindo a escolha de várias opções).

Tabela 3.2. Formulário de Extração de Dados utilizado na RS

Campo	Tipo do campo
Afiliação dos autores	Descritivo
O estudo relata uma estratégia para (semi) automatização da seleção inicial de estudos?	Sim ou Não
Descrição da(s) estratégia(s) para (semi) automatização da seleção inicial de estudos e recursos computacionais utilizados	Descritivo
O estudo relata uma ferramenta de suporte à seleção inicial de estudos?	Sim ou Não
Qual é o suporte provido pela ferramenta para a seleção inicial de estudos?	Descritivo
O estudo relata uma estratégia não automatizada para a seleção inicial de estudos?	Sim ou Não
Descrição da(s) estratégia(s) não automatizada(s) para a seleção inicial de estudos	Descritivo
O estudo apresenta uma avaliação da(s) estratégia(s) proposta(s)?	Sim ou Não
Como a(s) estratégia(s) foi(ram) avaliada(s)?	Estudo de Caso ou Experimento ou Outros
Descrição dos resultados obtidos após a avaliação da(s) estratégia(s)	Descritivo

3.2.5 Critérios de Qualidade

Foram definidos critérios de qualidade com a finalidade de classificar os estudos primários por relevância com base nas estratégias de seleção inicial encontradas e em suas avaliações. Assim, se um estudo apresenta uma estratégia, se ela foi implementada em uma ferramenta de suporte e se foi avaliada de alguma maneira para comprovar seu funcionamento, esse estudo tende a ser mais bem ranqueado do que outros que apresentam apenas uma estratégia sem avaliá-la ou sem implementá-la em alguma ferramenta de suporte.

Os critérios de qualidades definidos pelos autores são apresentados na Tabela 3.3.

Tabela 3.3. Critérios de Qualidade utilizados na RS

Código	Descrição
CQ1	A(s) estratégia(s) propostas foi(ram) apresentada(s) de maneira clara?
CQ2	A(s) estratégia(s) foi(ram) avaliada(s)?
CQ3	A(s) estratégia(s) foi(ram) implementada(s) em alguma ferramenta de suporte e está(ão) disponível(is) para utilização de pesquisadores?

As possíveis respostas para cada pergunta (que corresponde a um critério de qualidade) feita são: sim, não e parcialmente. Com o intuito de tornar possível o ranqueamento de estudos por critérios de qualidade, uma escala de pontuação foi elaborada: a cada resposta “sim” o estudo receberia 1 ponto; a cada resposta “parcialmente” o estudo receberia 0,5 pontos; e a cada resposta “não” o estudo não receberia pontos. Dessa forma, um estudo poderia ter no máximo 3 pontos e no mínimo nenhum ponto. Destaca-se que, para o CQ2, o estudo deveria receber 1 ponto se avaliado por meio de um estudo experimental, 0,5 pontos se apresentado somente um exemplo de uso e 0 pontos no caso de não haver avaliação ou demonstração de uso. É importante ressaltar ainda que, no caso um estudo não obter nenhum ponto, o mesmo seria excluído da RS por critério de qualidade.

3.3 Resultados

Após a execução das *strings* de busca, foram recuperados 432 estudos primários, dois quais 116 foram caracterizados como estudos duplicados, isto é, apareciam mais de uma vez por terem sido originados de bases de dados distintas. A ferramenta StArt tem um algoritmo que permite a detecção automática de muitos estudos duplicados com base em recursos de mineração de texto, o que auxiliou bastante nessa tarefa e poupou tempo dos autores, permitindo a avaliação manual de um número menor de estudos. Para os 316 estudos remanescentes, a atividade de seleção inicial foi realizada, com a leitura do título, *abstract* e palavras-chave dos estudos, e a aplicação dos critérios de inclusão e exclusão, sendo que 20 deles foram considerados potencialmente relevantes e incluídos para a atividade de leitura dos textos completos. Desses 20 estudos, não foi obtido acesso ao texto completo de dois deles, que foram excluídos por essa razão, restando então 18 estudos, dos quais apenas 13 foram considerados realmente relevantes para o contexto da pesquisa. O formulário de extração de dados foi preenchido para cada um desses estudos primários a fim de responder as questões de pesquisa formuladas.

A Tabela 3.4 apresenta a lista dos 13 estudos primários que foram considerados relevantes, ordenados decrescentemente por ano de publicação. Ela é composta pelo código que cada estudo primário relevante recebeu, composto pela letra E seguido de um número sequencial, pelo seu ano de publicação, a referência para sua identificação e a pontuação obtida com base nos critérios de qualidade, conforme explicado na Seção 3.2.5.

Os resultados obtidos na RS foram classificados em quatro categorias: estratégias semiautomáticas para seleção inicial de estudos, estratégias não automáticas para seleção inicial de estudos, comparações de recursos computacionais aplicados na seleção e ferramentas de suporte à seleção inicial de estudos. Os resultados de cada categoria são apresentados nas subseções a seguir.

Tabela 3.4. Lista dos estudos relevantes da RS

Código	Ano	Referência	Pontuação
E01	2016	(OCTAVIANO; SILVA; FABBRI, 2016)	3,0
E02	2015	(OCTAVIANO *et al.*, 2015)	3,0
E03	2014	(ABILIO *et al.*, 2014)	2,0
E04	2014	(FELIZARDO *et al.*, 2014)	2,5
E05	2014	(ALI; PETERSEN, 2014)	2,0
E06	2013	(FELIZARDO; SOUZA; MALDONADO, 2013)	2,5
E07	2012	(FELIZARDO *et al.*, 2012)	2,5
E08	2012	(FABBRI *et al.*, 2012)	1,5
E09	2011	(FELIZARDO *et al.*, 2011)	3,0
E10	2011	(TOMASSETTI *et al.*, 2011)	2,5
E11	2010	(WALLACE *et al.*, 2010a)	2,0
E12	2010	(WALLACE *et al.*, 2010b)	2,5
E13	2010	(FELIZARDO *et al.*, 2010)	2,0

3.3.1 Estratégias Semiautomáticas para Seleção de Estudos

No contexto deste trabalho, uma estratégia para seleção de estudos é considerada como semiautomática quando é utilizada para tomada de decisões automáticas em relação a incluir ou não estudos primários na RS. Não é considerada automática pelo fato de necessitar da intervenção humana ao menos na avaliação de parte dos estudos primários. Foram identificadas 3 estratégias, as quais são codificadas por ESA (estratégia semiautomática) seguido de uma sequência numérica.

a) ESA1

- **Descrição:** ESA1 é relatada no E10 (vide Tabela 3.4), o qual apresenta uma estratégia não nominada com base em recursos do DBPedia, que é um repositório de dados web para armazenamento de informações de Wikipedias em formato de dados estruturados (*linked data*). A estratégia começa com a definição de um conjunto inicial de estudos relevantes chamado I0, que pode ser definido com base em estudos previamente conhecidos pelos pesquisadores ou por meio de uma busca piloto para encontrar alguns estudos que sejam relevantes, se não for um tema que lhes seja familiar. Em seguida, os pesquisadores devem ler os títulos, *abstracts*, introduções e conclusões dos estudos de I0 e extrair o *bag of words* de cada estudo (chamado pelos autores de Modelo M). Vale ressaltar que no *bag of*

words a ordem dos termos não é levada em consideração, o que significa que um termo como “structured data” é considerado o mesmo que “data structured”, por exemplo.

Para os demais estudos que não pertencem a I0, termos (palavras-chave ou frases-chave) são extraídos deles e enviados ao DBPedia para encontrar novos termos que sejam sinônimos dos enviados. Então, os sinônimos são incorporados ao conjunto inicial de termos extraídos para representação do estudo, o que os autores chamam de processo de enriquecimento de termos (*linked data enrichment process*). Em seguida, os estudos enriquecidos são comparados a I0 por meio do classificador Naive Bayes, que é um algoritmo de mineração de texto muito conhecido para classificação de textos. Assim, se o percentual de similaridade entre o estudo enriquecido e I0 for maior que um limite preestabelecido, então o estudo é considerado relevante e o seu texto completo deve ser lido pelos pesquisadores, caso contrário não. Se o estudo for relevante, I0 deve ser atualizado com o novo estudo para efeito de novas comparações com os demais estudos que precisam ser processados.

Um protótipo Java foi desenvolvido para implementar a estratégia e um estudo de caso realizado para avalia-la, considerando uma RS publicada sobre o tópico estimativa de custo de software. Primeiramente, os autores utilizaram *bag of words* simples para representar os estudos, isto é, sem considerar o processo de enriquecimento mencionado anteriormente e compararam os resultados com os da RS original, que foi conduzida manualmente. Depois, repetiram o mesmo processo, porém utilizando *bag of words* enriquecidas para representar os estudos e comparar os resultados com os da RS original. Os resultados do estudo de caso mostraram que, sem perder nenhum estudo relevante, o processo sem enriquecimento foi 15% mais eficiente do que o original e o processo com enriquecimento foi 20% mais eficiente do que o original. Os números foram calculados com base no total de estudos que precisou ser processado para conseguir os mesmos resultados do que os da RS

original, já que os autores conheciam previamente os estudos primários relevantes.

- **Discussão:**

  Os pontos fortes identificados são:

  - A ideia de utilizar *bag of words* para representar os estudos e de enriquecimento dos termos extraídos é muito interessante no que diz respeito à comparação de estudos candidatos de uma RS.
  - A iteratividade do processo é positiva, pois estudos relevantes encontrados são sempre acrescentados ao conjunto inicial de estudos relevantes.

  Os pontos fracos identificados são:

  - Os autores não mencionam o que acontece com os estudos já excluídos uma vez que o conjunto de estudos relevantes é ampliado. Pode ser que, ao ampliar o conjunto inicial, um estudo anteriormente excluído passe a ser considerado relevante.
  - A avaliação foi realizada considerando apenas uma RS, e é difícil prever o que aconteceria ao utilizar a estratégia em várias RSs.
  - Não há uma ferramenta que permita a utilização da estratégia pela comunidade. Apenas um protótipo Java foi criado para realização do estudo de caso.
  - Embora os resultados mostrem uma redução de esforço, não está claro que 20% de redução de esforço seja bom, uma vez que os autores não mencionam o tempo gasto para envio e processamento dos termos no repositório DBPedia, que são tempos extras que um processamento manual não necessita, ainda que seja feito de forma automática.
  - Como a estratégia utiliza as seções de introdução e conclusão dos estudos, e não só os títulos e *abstracts*, os pesquisadores teriam de fazer o *download* do texto completo de todos os estudos primários inicialmente identificados para

processamento. Isso seria uma tarefa extremamente desgastante e com elevado consumo de tempo, que também não foi considerada na avaliação dos resultados. Normalmente, os estudos são exportados das bases de dados em algum formato, por exemplo Bibtex, mas possuem informações básicas, tais como título, autores, local e ano de publicação, *abstract*, entre outras.

b) ESA2

- **Descrição:** ESA2 é apresentada em E03 (vide Tabela 3.4), o qual relata uma estratégia com base em ranqueamento de estudos. O ranque é criado com base nos termos utilizados na *string* de busca e encontrados no título, *abstracts* e palavras-chave dos estudos recuperados. O Modelo Vetorial é utilizado para ranquear os estudos, que é um modelo algébrico para representação de documentos e consultas como vetores em um espaço t-dimensional, onde t é o número de termos distintos da coleção de termos. O peso de cada termo é calculado com base na frequência com que ele aparece no documento.

  Os autores propõem dois métodos: (i) os termos da *string* de busca são tratados igualmente, não importando se são ou não sinônimos, e desconsiderando termos duplicados e os conectores OR e AND. Utilizam o percentual de similaridade para comparar os termos extraídos de um estudo com a *string* de busca utilizada; e (ii) uma função de ranqueamento que simula a expressão booleana da *string* de busca é definida. Grupos de termos são criados com base nos conectores OR e AND da *string* de busca. Termos agrupados por OR são colocados no mesmo grupo, enquanto que o conectar AND indica a existência de um novo grupo. O percentual de similaridade é usado para calcular a similaridade de um estudo com a *string* de busca, sendo que o nível de similaridade de cada termo corresponde ao nível de similaridade máximo obtido pelo grupo. Há fórmulas matemáticas para os cálculos.

Um protótipo Java foi desenvolvido para implementar a estratégia e um estudo de caso realizado para avaliar os dois métodos propostos na estratégia considerando apenas uma RS sobre o tópico métricas contemporâneas para manutenção de software, que, inclusive, foi conduzida e publicada por alguns dos criadores da estratégia. A ferramenta JabRef foi utilizada para carregar arquivos BibTex com informações dos estudos primários originados de fontes diferentes e compilar todas as informações em um arquivo BibTex completo, que foi carregado no protótipo. Os resultados ao aplicar o método (i) mostram uma precisão de 28,6% e revocação de 60%, enquanto que os resultados ao aplicar o método (ii) mostram uma precisão de 50% e revocação de 80%. Um dado importante é que, para ambos os métodos, os estudos relevantes, previamente conhecidos pelos autores, aparecem entre os 20% de estudos mais bem ranqueados.

- **Discussão:**

  Os pontos fortes identificados são:

  - A ideia de ranquear estudos por suposta relevância é muito boa. O uso de frequência de palavras-chave no título e *abstract* para cálculo de relevância mostrou bons resultados no estudo de caso.
  - A ideia de agrupar termos (palavras-chave) ao analisar a *string* de busca é interessante. Termos agrupados pelo conector OR na *string* de busca pertencem ao mesmo grupo, enquanto que um conector AND, ao ser encontrado, define um novo grupo de termos.
  - A ideia de considerar pesos distintos para os termos é interessante.
  - Resultados obtidos por meio do método (ii) mostram que ele é promissor.

  Os pontos fracos identificados são:

  - A avaliação da estratégia foi realizada considerando apenas uma RS, e é difícil prever o que aconteceria ao utilizar a

estratégia em várias RSs. Para piorar, a única RS utilizada no estudo de caso foi conduzida por alguns dos próprios criadores da estratégia.

- Não há uma ferramenta que permita a utilização da estratégia pela comunidade. Apenas um protótipo Java foi criado para realização do estudo de caso.
- Método (i) não mostra resultados bons que justifiquem sua utilização na prática.
- A estratégia faz uso de uma terceira ferramenta (JabRef) para intermediar o processo de transformar os diferentes BibTex provenientes de bases de dados distintas em um único arquivo BibTex necessário para execução da estratégia.

c) ESA3

- **Descrição:** se refere à estratégia chamada SCAS, que é um dos frutos desta pesquisa e é apresentada com detalhes no Capítulo 4, além de ter sido apresentada e avaliada em E01 e E02 (vide Tabela 3.4).

3.3.2 Estratégias Não Automáticas para Seleção de Estudos

No contexto deste trabalho, uma estratégia para seleção de estudos é considerada como não automática se ela propõe melhorias na seleção de estudos, mas não toma nenhuma ação automática de incluir ou excluir estudos, apenas faz sugestões de possíveis ações com base em alguma classificação de estudos. Foram identificadas 3 estratégias, as quais são codificadas por ENA (estratégia não automática) seguido de uma sequência numérica.

a) ENA1

- **Descrição:** ENA1 é relatada em E07 (vide Tabela 3.4) e é baseada em técnicas de mineração de texto e visualização. Representação visuais de estudos são geradas após um processamento de informações de seus títulos, *abstracts*, palavras-chaves e referências por meio de recursos de mineração de texto. Uma ferramenta chamada Revis foi implementada para dar suporte à estratégia, a qual consiste, basicamente, em três visualizações que devem ser combinadas:

(i) Mapa de Documentos – estudos da RS são minerados e representados como bag of words, e, então, projetados de forma que estudos similares fiquem mais próximos uns dos outros, e estudos menos similares, mais distantes uns dos outros; (ii) Edge Bundles – mostra de forma gráfica os relacionamentos de citações existentes entre os estudos de uma RS (quais estudos são citados por outros); (iii) Rede de Citação – mostra graficamente os estudos e suas referências, visando a detectar referências em comum entre os estudos.

Quando um estudo é incluído, é possível verificar no Mapa de Documentos quais estudos têm conteúdo mais similares aos do estudo incluído. Assim, a ferramenta pode sugerir outros estudos similares ao incluído que possivelmente possam ser relevantes também. Por meio do Edge Bundles, é possível detectar quais estudos são os mais citados por seus pares, se tornando candidatos a serem incluídos também. Por fim, a Rede de Citação tenta auxiliar na identificação de possíveis novos estudos relevantes ao verificar as referências dos estudos já incluídos pelo pesquisador.

Um estudo de caso foi conduzido considerando quatro alunos de pós-graduação, divididos em dois grupos. Uma RS publicada foi escolhida como oráculo. O primeiro grupo conduziu a mesma RS de forma manual e o segundo grupo com o suporte da Revis. Os resultados mostram que os alunos que utilizaram a Revis foram mais rápidos na execução da RS e uma precisão melhor na inclusão e exclusão de estudos primários.

O estudo E06 apresentou um experimento que replicou o estudo de caso mencionado anteriormente. O experimento foi controlado e contou com a participação de quinze alunos de pós-graduação (mestrandos e doutorandos), os quais foram randomicamente divididos em dois grupos. Os resultados comprovaram que a estratégia, executada por meio da Revis, acelera a atividade de seleção e tem maior precisão do que a condução manual e, além disso, os doutorandos obtiveram melhores resultados do que os mestrandos, o que comprova como a

maior experiência do pesquisador tem influência nos resultados positivamente.

Outros estudos experimentais foram feitos para avaliar estratégias muito similares à apresentada, utilizando mineração de texto e visualização, aplicadas a MSs, mostrando bons resultados em relação à eficácia e eficiência. Elas são apresentadas, respectivamente, em E13, E04 e E12.

- **Discussão:**

  Os pontos fortes identificados são:

  - A estratégia apresentada gera várias visualizações interessantes que, quando combinados, parecem realmente auxiliar os pesquisadores na atividade de seleção.
  - Embora os estudos de caso e experimentos realizados fossem simples, os resultados são muito promissores.

  Os pontos fracos identificados são:

  - O esforço para preparar o arquivo de entrada utilizado para a ferramenta, o qual deve ter todas as referências dos estudos devidamente formatadas, é bem elevado. Esse tempo de preparação não é considerado na avaliação comparativa com o tempo de execução manual dos alunos, mas deveria. Não há nenhum processamento automático para realizar essa árdua tarefa.
  - Todos os estudos de caso e experimentos realizados utilizam a mesma RS como oráculo. A estratégia deveria ser avaliada com outras RSs para realmente comprovar seu valor e utilidade.

b) ENA2

- **Descrição:** ENA2 é relatada em E08 (vide Tabela 3.4), o qual também apresenta uma estratégia com base em recursos de mineração de texto e visualização. Apesar de o maior foco dessa estratégia estar na atividade de sumarização dos dados, ela também apoia a atividade de seleção inicial de estudos. É suportada pela ferramenta StArt. Os

pesquisadores devem carregar os textos completos dos estudos na ferramenta que, por meio de autômatos criados, extrai os principais campos das referências dos arquivos no formado PDF e insere-os nos campos correspondentes na própria ferramenta. Então, após gerada a visualização de referências das referências, é possível identificar estudos potencialmente relevantes, que não foram recuperados inicialmente, mas que são citados por estudos que foram incluídos na RS. Apenas um exemplo de uso é apresentado, nenhum estudo de caso ou experimento com RSs publicadas foi realizado.

- **Discussão:**

  Os pontos fortes identificados são:

  - A estratégia apresentada mostra um conjunto de visualizações que parecem interessantes para auxiliar na detecção de possíveis estudos primários não identificados na busca.
  - Nenhuma ferramenta adicional ou preparação de arquivos é necessária, tudo é processado pela ferramenta.

  Os pontos fracos identificados são:

  - Uma tarefa que exige bastante tempo é o carregamento de arquivos PDF de todos os estudos para a ferramenta. Seria de grande utilidade se fossem carregados automaticamente.
  - A estratégia não foi devidamente avaliada, há apenas um exemplo de como utilizá-la.

c) ENA3

- **Descrição:** ENA3 é relatada em E05 (vide Tabela 3.4), o qual apresenta uma estratégia com base nas decisões tomadas por dois ou mais pesquisadores em relação aos estudos primários. Cada estudo deve ser classificado como Relevante, Incerto ou Irrelevante. Os pesquisadores devem escolher aleatoriamente cinco estudos e discutir os critérios de inclusão e exclusão que aplicariam neles, tarefa essa chama pelos autores de *Think-Aloud Protocol*. Depois, os pesquisadores devem classificar um conjunto de estudos selecionados

aleatoriamente e calcular o nível de concordância entre eles por meio do coeficiente Cohen's Kappa (CARLETTA, 1996), fazendo ajustes se necessário para alinhamento de ideias. Então, os estudos são todos classificados como incluídos ou excluídos e, de acordo com as decisões dos pesquisadores, são categorizados em: A (Relevante/Relevante), B (Relevante/Incerto), C (Incerto/Incerto), D (Relevante/Irrelevante), E (Irrelevante/Incerto) ou F (Irrelevante/Irrelevante). Dessa forma, estudos pertencentes às categorias A e B são indicados para leitura dos textos completos e estudos pertencentes à categoria F devem ser excluídos. Estudos pertencentes às categorias D e E devem ser discutidos e recategorizados em A, C ou F. Estudos pertencentes à categoria C devem ter uma leitura adaptativa, isto é, ler a introdução, depois a conclusão, e outras seções se necessário até tomar uma decisão de inclui-los ou não.

Um estudo de caso foi conduzido pelos autores por meio de uma RS, e a estratégia foi avaliada considerando desde o caso mais restritivo (só incluir estudos da categoria A) até o mais abrangente (incluir estudos da categoria A até a categoria E), sempre calculando a precisão e esforço requerido para cada caso. Os autores concluem dizendo que é preciso evitar o caso mais restritivo devido à grande perda de estudos relevantes, e também o caso mais abrangente, no qual o esforço requerido é muito elevado em relação ao benefício de tentar recuperar todos os estudos relevantes.

- **Discussão:**

  Os pontos fortes identificados são:

  - Os estudos são divididos em categorias com base nas opiniões dos pesquisadores e há uma sugestão de decisão para cada categoria.
  - A ideia deveria ser mais explorada no sentido de combinar decisões humanas e computacionais em uma estratégia semiautomática.

Os pontos fracos identificados são:

- Dependendo do nível de experiência dos pesquisadores, muitos estudos poderiam ser classificados nas categorias que requerem a leitura do texto completo, ou até mesmo da leitura adaptativa, porque a estratégia diz que um estudo é categorizado quando ao menos um pesquisador toma uma decisão que o torne elegível a uma determinada categoria. Parece que uma simples discussão poderia economizar tempo dos pesquisadores em alguns casos.
- A avaliação da estratégia foi realizada considerando apenas uma RS, e é difícil prever qual seria o seu comportamento e se haveria redução de esforço significativa para outras RSs.

3.4 Limitações da RS

Uma limitação da RS conduzida é que nenhuma busca manual foi realizada, apenas buscas por meio de *strings* executadas em bases de dados, o que implica que pode haver estudos primários relevantes não recuperados. Para minimizar esse problema, estudos relevantes conhecidos pelos autores foram utilizados como controle para validação dos estudos retornados nas buscas. Apenas um estudo (E08) dentre os estudos relevantes previamente conhecidos não foi identificado por não estar indexado nas bases, mas o mesmo foi adicionado manualmente.

3.5 Considerações Finais

Este capítulo apresentou uma RS conduzida que deu bom embasamento da literatura para permitir o desenvolvimento desta tese.

Por meio da RS foram identificadas três estratégias semiautomáticas para seleção inicial de estudos e três estratégias para seleção não automática. As

estratégias foram brevemente descritas e alguns pontos positivos e negativos foram identificados em relação às estratégias semiautomáticas e também às não automáticas. Os resultados mostram que estratégias para diminuição do esforço na atividade de seleção, sobretudo com a sua (semi) automatização é uma lacuna a ser mais bem explorada em pesquisas.

O fato de a revisão da literatura ter sido conduzida por meio de um processo sistemático – a RS (KITCHENHAM; CHARTERS, 2007) – facilitou a atualização bibliográfica referente à pesquisa. Para tanto, bastou reaplicar as *strings* de busca registradas no protocolo sempre que desejado e analisar apenas os novos estudos retornados. Essa atividade foi bastante útil para garantir a originalidade desta pesquisa.

Capítulo 4

ESTRATÉGIA SEMIAUTOMÁTICA SCAS PARA SELEÇÃO DE ESTUDOS

Este capítulo descreve o funcionamento inicial da estratégia SCAS para seleção semiautomática de estudos primários, bem como apresenta os resultados iniciais obtidos com a aplicação da estratégia em revisões sistemáticas.

4.1 Considerações Iniciais

Conforme descrito anteriormente, a atividade de seleção pode exigir bastante esforço do pesquisador que conduz uma RS e estratégias que buscam agilizar essa atividade podem ser muito úteis. O capítulo anterior mostrou que poucas estratégias foram encontradas na literatura visando à aceleração da seleção de estudos, sobretudo no âmbito de sua (semi) automatização. Neste capítulo apresenta-se a estratégia SCAS, na sua versão inicial, cujos resultados experimentais motivaram sua evolução para a versão SCAS-Fuzzy, apresentada no Capítulo 5.

Este capítulo está organizado da seguinte forma: na Seção 4.2 é apresentada a descrição da estratégia semiautomática SCAS inicialmente definida; na Seção 4.3 é apresentado o estudo de caso realizado para avaliar a estratégia SCAS; na Seção 4.4 é apresentado o experimento realizado com alunos de pós-graduação para realizar uma nova avaliação da estratégia; e, por fim, na Seção 4.5 são apresentadas as considerações finais deste capítulo.

4.2 Descrição da Estratégia SCAS

A estratégia semiautomática para seleção inicial de estudos primários em estudos secundários proposta é chamada SCAS (*Score Citation semi-Automatic Selection*). A SCAS tem por base duas funcionalidades principais: *score* e número de citações de um estudo (OCTAVIANO; SILVA; FABBRI, 2016).

O *score* é um valor calculado e atribuído a cada estudo com base no número de vezes que os termos da *string* de busca que o localizou, que são as palavras-chave definidas pelo usuário no protocolo, são encontrados em partes específicas do texto (título, *abstract* e palavras-chave). Assim, cada vez que um termo da *string* de busca é encontrado em qualquer uma dessas três partes do texto, um valor é adicionado à contagem corrente do *score*, que inicialmente possui o valor zero. Depois de procurar por todos os termos da *string* de busca nas três partes específicas, o *score* final é calculado e atribuído ao estudo. Portanto, o resultado é totalmente dependente dos termos da *string* de busca definida.

Pesos diferentes são atribuídos quando um termo da *string* de busca é encontrado no título, *abstract* ou palavras-chave de um estudo. Ocorrências no título têm valor maior do que as ocorrências no *abstract*, que, da mesma forma, têm um valor mais elevado do que as ocorrências nas palavras-chave. As principais razões para atribuir um valor inferior a ocorrências nas palavras-chave são: (i) em alguns casos, as palavras-chave dos estudos não aparecem nos arquivos exportados das bases de dados de pesquisa (formato BibTex, por exemplo); (ii) como palavras-chave são em número limitado, pode ocorrer que o pesquisador não conseguiu definir todas as palavras relevantes para caracterizar o assunto sob investigação. Assim, nesses casos, atribuir um valor alto às ocorrências dos termos nas palavras-chave dos estudos poderia ter um impacto negativo na classificação de estudos por *score*. O valor mais alto é atribuído a ocorrências encontrados no título, uma vez que tendem a ser significativo para o contexto de uma RS.

A funcionalidade *score* está implementada na ferramenta StArt, a qual sugere valores pré-determinados para as ocorrências encontradas em cada parte do texto (título, *abstract* e palavras-chave) de um estudo primário. Esses valores (5 para ocorrências no título, 3 para ocorrências no *abstract* e 2 para ocorrências nas palavras-chave) foram inicialmente estabelecido com base no senso comum do

grupo de pesquisa do LaPES (Laboratório de Pesquisa em Engenharia de Software) que trabalha no desenvolvimento da ferramenta. Depois disso, esses valores foram avaliados muitas vezes, mostrando coerência nos estudos secundários realizados pelo grupo. No entanto, a ferramenta permite que os pesquisadores consigam alterar os pesos de acordo com as suas necessidades.

O autor deste trabalho realizou um estudo de caso para avaliar se a atual configuração (5, 3 e 2 pontos para ocorrências de termos da *string* no título, *abstract* e palavras-chave, respectivamente) era realmente a melhor configuração. A metodologia consistiu em fazer diversas simulações de valores distintos para ocorrências em cada uma das três partes e analisar os resultados considerando várias RSs compartilhadas por pesquisadores. Os resultados foram avaliados e chegou-se à conclusão que os valores atuais (5 para ocorrências no título, 3 para ocorrências no *abstract* e 2 para ocorrências nas palavras-chave) ainda são a melhor configuração e, por isso, foram mantidos.

A Figura 4.1 mostra a tela da ferramenta StArt que exibe os estudos primários carregados e classificados pelo pesquisador, e ordenados pela funcionalidade *score* (em destaque na figura). A ferramenta permite a exportação dos dados em formato de planilha eletrônica, o que facilita o processo de análise dos dados

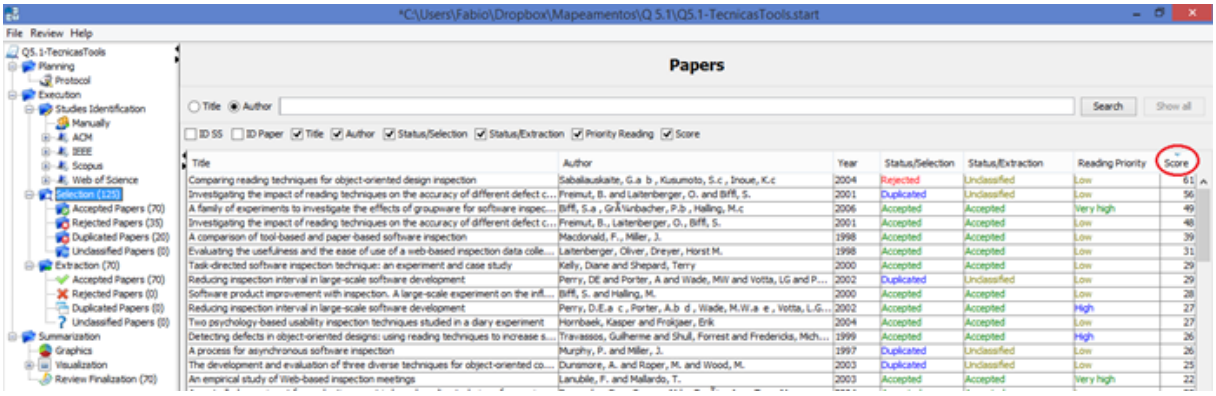


Figura 4.1. Estudos classificados por *score* na ferramenta StArt

A segunda funcionalidade utilizada na SCAS é o número de citações de cada estudo, que consiste em saber quantas vezes um estudo é citado por seus pares, isto é, pelos demais estudos recuperados pelo pesquisador para uma determinada RS. A ideia é que um estudo bastante citado por seus pares tenha a chance maior de ser um estudo considerado relevante para a RS.

O cálculo do número de citações de um estudo feito de maneira automática é um desafio bem grande. Primeiramente é preciso obter as referências dos estudos, que na grande maioria das vezes não é exportado pelas bases de dados. A ideia é evitar que o pesquisador tenha de gastar um tempo demasiado preparando as referências dos estudos (ainda que utilizando ferramentas de apoio como a JabRef, por exemplo), o que seria muito desmotivante na utilização da estratégia SCAS pelo pesquisador, e que inclusive foi apontado como ponto negativo em algumas estratégias de seleção apresentadas no Capítulo 3. A SciVerse Scopus é um exemplo de base de dados que exporta as referências completas dos estudos. Uma alternativa para os estudos que não possuem as suas referências nos arquivos exportados pelas bases de dados é carregar os textos completos (em formato PDF) na ferramenta e, por meio de recursos de mineração de texto, obter as referências de cada estudo. Essa funcionalidade já foi implementada na ferramenta StArt e está em processo de avaliação para melhorar a precisão do algoritmo.

Uma vez que as referências dos estudos tenham sido obtidas, o próximo passo é gerar uma matriz de citação dos estudos e, por meio dela, calcular quantas vezes um estudo é citado pelos demais, apresentando esse número na ferramenta.

Assim, tendo o *score* calculado para cada estudo, bem como o número de citações recebidas por ele em relação aos demais estudos, é possível executar a estratégia SCAS, que é composta por duas fases:

- **Fase 1 – Aplicando as funcionalidades *score* e número de citações**

Nessa fase, as funcionalidades *score* e citação são aplicadas isoladamente, a começar pelo *score*. Documentos relevantes, ou seja, estudos com *scores* altos, devem ser incluídos na RS, e documentos irrelevantes, ou seja, estudos com *scores* baixos, devem ser excluídos da RS. Para dizer se um *score* é alto ou baixo, um valor de corte tem de ser definido e, para isso, duas técnicas (regra dos 50% e árvore de decisão J48) são propostas.

Na regra dos 50%, os estudos primários são classificados decrescentemente por *score*, isto é, os estudos de *scores* mais altos são posicionados no topo da lista. O valor de corte é definido como sendo o *score* do estudo classificado no meio da lista. Por exemplo, se há vinte estudos a serem analisados, o *score* do décimo estudo é utilizado como o valor de corte. Se houver um número ímpar de estudos, o quociente deve ser truncado para zero casas decimais, de modo a obter o valor de

corte. Estudos com *score* acima do valor de corte (ou seja, estudos considerados com *scores* altos) são candidatos a serem incluídos na RS. Por outro lado, estudos com *scores* abaixo do valor de corte (ou seja, estudos considerados com *scores* baixos) devem ser excluídos da RS. É importante ressaltar que os estudos com *scores* idênticos ao valor de corte também devem ser incluídos. A regra dos 50% foi definida com base em observações feitas nas três RSs utilizadas no estudo de caso apresentado na Seção 4.3, no qual o número de estudos primários relevantes que seriam excluídos pode ser considerado baixo se comparado ao número de estudos relevantes classificados corretamente por meio da regra dos 50%.

A segunda técnica baseia-se na árvore de decisão J48, recurso disponível na ferramenta Weka[5]. A árvore de decisão J48 requer pelo menos duas variáveis de entrada (chamadas de atributos) e uma variável de saída (chamada de classe). O *score* e o número de citações dos estudos foram usados como atributos. O *score* foi normalizado, sendo o maior score definido como 1 e os demais valores calculados dividindo-se o *score* de cada estudo pelo maior *score*. O número de citações é definido como 0 (ou seja, o estudo não é citado) ou 1 (ou seja, o estudo é citado ao menos uma vez). Com base no *score* e no número de citações, o *status* de classe indica se um estudo deve ser incluído ou não. As decisões tomadas pelos pesquisadores que realizaram as três RSs utilizadas no estudo de caso foram usadas para treinar a árvore de decisão J48. Na verdade, 66% dos dados foram randomicamente usados para treinar a árvore de decisão, ou seja, foram fornecidos os *scores* e os números de citações desses estudos e suas classes correspondentes (decisão de incluir ou excluir o estudo, tomada pelo pesquisador), e 34% dos dados foram utilizados para executar a árvore de decisão, obtendo um percentual de classificação correta de quase 80%. A Figura 4.2 mostra a árvore de decisão J48 resultante gerada na Weka.

[5] https://www.cs.waikato.ac.nz/ml/weka/

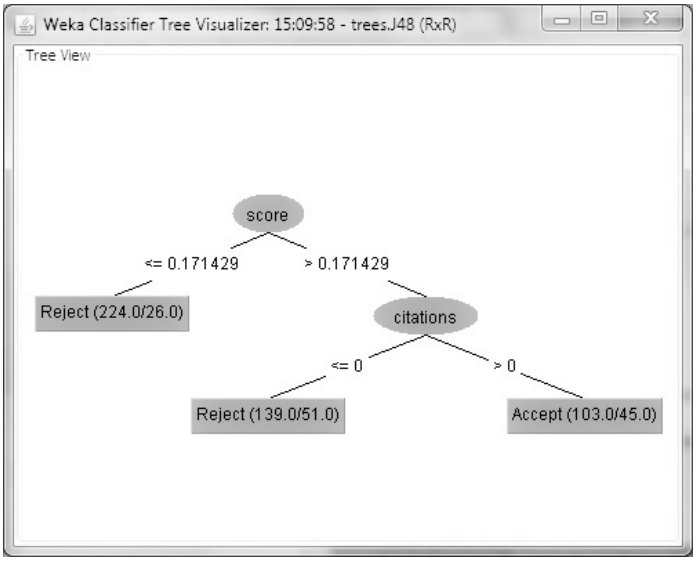


Figura 4.2. Árvore de decisão J48 gerada na Weka (OCTAVIANO; SILVA; FABBRI, 2015)

A árvore de decisão J48 executa dois níveis de "poda". O primeiro baseia-se somente no atributo *score*, sem considerar o atributo número de citações. O *score* sugerido como um valor de corte é 17,14% em relação ao valor do maior *score* calculado. O valor de corte deve ser arredondado para o valor inteiro imediatamente acima, no caso em que o valor de corte possua casas decimais. Por exemplo, se o valor de corte sugerido é de 27,35 ou 27,89, ambos devem ser arredondados para o score 28.

A segunda "poda" é baseada no atributo número de citações, e mostra as sugestões de quando os estudos deveriam ser incluídos ou excluídos (vide Figura 4.2), porém ela não é utilizada na SCAS. A razão disso é que o atributo número de citações não é aplicado aos estudos com *scores* baixos (que seriam todos rejeitados na primeira "poda"), como mostrado na Figura 4.2. Na SCAS, os estudos com *scores* baixos não são diretamente descartados, como sugere a árvore, mas sim avaliados conforme o número de citações, como será explicado na segunda fase da estratégia.

Um algoritmo para calcular o valor de corte foi implementado na ferramenta StArt com base nas duas técnicas apresentadas. Uma vez que os estudos são importados na ferramenta, as duas técnicas (regra 50% e árvore de decisão J48) são aplicadas para calcular o valor de corte. Primeiramente, o algoritmo classifica os estudos por *score* em ordem decrescente e escolhe o *score* do estudo classificado no meio da lista (50%) como um candidato. Depois, o algoritmo aplica as regras sugeridas pela árvore de decisão J48 e escolhe o valor de *score* que é a 17,14% do

score mais alto como um segundo candidato. Então, o algoritmo determina que o valor de corte é a menor valor entre os dois candidatos escolhidos. Por conseguinte, um maior número de estudos é considerado com *score* alto (estudos que provavelmente são relevantes).

O algoritmo utiliza o índice pré-definido de 17,14% que foi sugerido pela Weka com base nos dados das três RSs utilizadas no estudo de caso. Um pesquisador com pouca experiência ou que não tem dados de RSs anteriores realizados por ele, a recomendação é de utilizar o índice pré-definido. No entanto, se o pesquisador é experiente e não se sente confortável em utilizar o índice de 17,14% por alguma razão, ele pode usar dados coletados de RSs que tenha conduzido antes, ou até mesmo selecionar um conjunto de estudos pertencentes à RS que esteja em andamento, inserir os dados na Weka, como explicado anteriormente, para obter a sua própria árvore de decisão J48 e, em seguida, alterar o índice pré-definido na StArt, substituindo-o pelo novo índice obtido. O índice é um parâmetro que pode ser alterado a qualquer momento na ferramenta.

Vale a pena ressaltar que não é necessário inserir na Weka os dados (*score* e número e citações) dos estudos da RS que esteja sendo conduzida, uma vez que esses dados são utilizados apenas para treinar o algoritmo J48. Se o pesquisador não deseja modificar o índice pré-definido, não precisa utilizar a Weka.

A Tabela 4.1 apresenta quatro cenários possíveis e as ações factíveis em relação a como o pesquisador deve proceder para obter o valor de corte. Os cenários são derivados do nível de experiência do pesquisador em conduzir RSs e da disponibilidade de dados de RSs anteriores realizadas por ele.

O custo para configuração da SCAS só existirá se o pesquisador optar por utilizar a Weka para obter um novo índice. Assim, o tempo estimado pode ser calculado resumindo o tempo gasto nos passos seguintes:

a) Prepara-se o arquivo de entrada que será usado na Weka. Esse arquivo deve conter o *score* normalizado e o número de citações de cada estudo. Essa tarefa pode ser feita por meio da geração de um relatório existente na StArt que lista as informações dos estudos (incluindo o seu *score* e número de citações), e, em seguida, usando uma planilha para normalizar os *scores* (dividindo o *score* de cada estudo pelo maior *score* utilizando uma fórmula). Tempo estimado: cerca de 4 minutos;

b) Na Weka, deve-se escolher o modo de processamento de árvore de decisão J48, importar o arquivo de entrada gerado em (a), e obter o índice a ser usado na StArt. Tempo estimado: cerca de 5 minutos;

c) Na StArt, alterar o parâmetro referente ao índice pelo valor obtido em (b), e executar o algoritmo para obter o valor de corte. Tempo estimado: menos de 1 minuto.

Tabela 4.1. Ações possíveis para determinar o valor de corte com base em cenários distintos

			Ações Factíveis		
Cenário	**Nível de Experiência**	**Possui dados de RSs anteriores?**	Manter o índice 17.14% pré-definido na StArt e usar o algoritmo da StArt para calcular o valor de corte	Usar dados de RSs anteriores na Weka para obter um novo índice. Depois, alterar o parâmetro do índice na StArt e, finalmente, usar o algoritmo da StArt para calcular o valor de corte	Obter dados de um conjunto de estudos da RS atual. Depois, usar esses dados na Weka para obter um novo índice. Então, alterar o parâmetro do índice na StArt e, finalmente, usar o algoritmo da StArt para calcular o valor de corte
1	Pouca ou nenhuma	Não	X		
2	Pouca ou nenhuma	Sim	X	X	
3	Muita ou especialista	Não	X		X
4	Muita ou especialista	Sim	X	X	

Portanto, em resumo, o tempo de preparação da SCAS será em torno de 10 minutos, mas apenas se o pesquisador optar por usar a Weka. Se o pesquisador optar por manter o índice pré-definido na StArt, não há tempo de configuração. O tempo estimado para a ferramenta aplicar SCAS e dividir os estudos em quadrantes (explicado na Fase 2) é de apenas alguns minutos.

- **Fase 2 – Combinando as funcionalidades *score* e número de citações para sugerir o *status* de cada estudo primário**

Nessa fase, as funcionalidades aplicadas na primeira fase (*score* e número de citações) são combinadas, e os estudos primários são classificados em três categorias e quatro quadrantes: (i) Categoria 1 (inclusão "correta") - Quadrante 1: um estudo com *score* considerado alto (estudo posicionado acima do valor de corte) e que possui ao menos uma citação é provavelmente um estudo relevante, um candidato a ser incluído na RS; (ii) Categoria 2 (exclusão “correta") - Quadrante 4:

um estudo com *score* considerado baixo (estudo posicionada abaixo do valor de corte) e que não recebe citações é, provavelmente, um estudo irrelevante, um candidato a ser excluído da RS; e (iii) Categoria 3 (estudos para serem revistos) - Quadrantes 2 e 3: um estudo com *score* alto, mas sem citações, ou um estudo com *score* baixo, mas que recebe ao menos uma citação, devem ser revistos para determinar sua relevância ou irrelevância.

A Figura 4.3 ilustra os quadrantes de classificação. Uma vez que os estudos foram classificados, as ações que podem ser tomadas são as seguintes:

a) Estudos pertencentes ao Quadrante 1 - o pesquisador deveria aceitar a recomendação SCAS e classificar automaticamente esses estudos como incluídos, pelo fato de possuírem *score* alto e ao menos uma citação;

b) Estudos pertencentes aos Quadrantes 2 e 3 - o pesquisador deveria rever manualmente esses estudos, a fim de classificar cada um deles como incluído ou excluído. Nenhuma ação automática deveria ser tomada;

c) Os estudos pertencentes ao Quadrante 4 - o pesquisador deveria aceitar a recomendação SCAS e classificar automaticamente esses estudos como excluídos, pelo fato de possuírem *score* baixo e não possuírem citações.

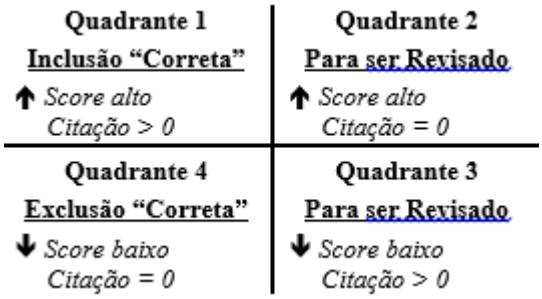


Figura 4.3. Combinando *score* e número de citações dos estudos para definir seus *status*

4.3 Estudo de Caso

Com o propósito de avaliar a estratégia SCAS, um estudo de caso contendo três exemplos foi realizado. Os exemplos (RS1, RS2 e RS3) são RSs publicadas na literatura e que variam em relação ao tópico de pesquisa e número de estudos primários considerados. Elas foram escolhidas porque foram conduzidas por pesquisadores experientes em realizar RSs e pelo fato dos pesquisadores terem disponibilizado todas as informações necessárias para aplicação da estratégia, tais como lista de estudos incluídos, lista de estudos excluídos e *strings* de busca utilizadas. É importante destacar que as RSs não foram conduzidas novamente, e sim os dados originais foram disponibilizados.

Os exemplos estão resumidos Tabelas 4.2, 4.3 e 4.4, que são organizadas em duas partes: (i) cabeçalho com informações das RSs (título, autores, temática e o número total de estudos primários incluídos e excluídos); e (ii) detalhes da RS: os estudos primários, seus *scores*, o número de citações que possuem e os seus *status* (incluído ou excluído pelos especialistas – os pesquisadores que conduziram as RSs).

Para a RS1 (vide Tabela 4.2), a regra dos 50% sugeriu que o valor de corte fosse o do *score* do estudo #48, que é 15. Segundo a técnica, todos os estudos com *score* 15 devem ser considerados também. Portanto, mais dois estudos foram classificados como estudos com *scores* altos, totalizando 50 estudos. No entanto, a árvore de decisão J48 sugeriu que o valor de corte fosse o *score* 14, estabelecendo o *score* do estudo #52 como o valor de corte. O pior caso das duas técnicas, ou seja, o resultado que considera mais artigos, deve ser escolhido. Em consequência, o *score* do estudo #52 foi escolhido como o valor de corte.

Os resultados da RS1 mostraram que 22 estudos foram classificados no quadrante 1 (ou seja, *scores* altos e possuem citações), ou seja, esses documentos são muito provavelmente relevantes para a RS1. Desse total, 19 foram incluídos (estudos 2, 3, 5, 6, 8, 11-13, 17, 19, 20, 25, 26, 28, 38, 40, 41, 44 e 48) e 3 foram excluídos (estudos 34, 46 e 51) pelos pesquisadores que realizaram a RS1 manualmente (vide coluna "Autores" na Tabela 4.2). O estudo #5 teve o maior número de citações; ele foi citado por 35 outros estudos. Os estudos 59-65, 52-57, 67, 69, 70, 73 75, 76, e 78-97 (38 no total), possuem *scores* baixos e não foram

citados - estes foram classificados no quadrante 4, ou seja, esses documentos muito provavelmente deveriam ser excluídos da RS1. Dentre eles, somente o estudo #62 não foi excluído pelos pesquisadores. Os estudos 1, 7 e 10 (quadrante 2 – *scores* altos, mas não citados) e os estudos 71, 72 e 77 (quadrante 3 – *scores* baixos, mas citados) são alguns exemplos de estudos que deveriam ser revisados manualmente.

Para a RS2 (vide Tabela 4.3), a regra dos 50% sugeriu que o valor de corte fosse o do estudo #18, com *score* 9. Segundo a técnica, todos os estudos com *score* 9 também devem ser considerados, assim um estudo adicional foi classificado como estudo com *score* alto, totalizando 19 estudos. Entretanto, a árvore de decisão J48 sugeriu que o valor de corte fosse o *score* 10, definindo o *score* do estudo #16 como o valor de corte. Como o pior caso deve ser escolhido, o *score* do estudo #18 foi definido como o valor de corte e um estudo adicional foi considerado com *score* alto.

Os resultados da RS2 mostraram que todos os estudos (1, 2, 6, 11 e 16) classificados como quadrante 1 (*scores* altos e que foram citados - provavelmente relevantes) foram incluídos pelos pesquisadores que realizaram manualmente a RS2 (vide coluna "Autores" na Tabela 4.3). Havia 15 estudos que foram classificados como quadrante 4 (*scores* baixos e que não foram citados - estudos provavelmente irrelevantes). Dentre eles, 7 foram incluídos (estudos 20, 22, 26, 32-34 e 37) e 8 foram excluídos pelos pesquisadores (estudos 21, 23-25, 27, 29, 35 e 36). Estudos 3-5, 7-10, 12-15, 17-19 (quadrante 2 – *scores* altos, mas sem citação) e os estudos 28, 30 e 31 (quadrante 3 – *scores* baixos, mas com citação) são exemplos de estudos que deveriam ser revisados manualmente.

Para a RS3, tanto a regra dos 50% quanto a árvore de decisão sugeriram o *score* do estudo #132 como sendo o valor de corte a ser adotado. Os resultados mostraram que 69 estudos foram classificados no quadrante 1 (*scores* altos e que foram citados - provavelmente relevantes) e, desse número, 33 foram incluídos e 36 foram excluídos pelos especialistas que conduziram a RS manualmente (vide coluna "Autores" na Tabela 4.4). Um total de 86 estudos foram classificados no quadrante 4 (*scores* baixos e que não foram citados - estudos provavelmente irrelevantes), sendo que 80 deles foram excluídos pelos especialistas. Um total de 63 estudos foram classificados no quadrante 2 (*scores* altos, mas sem citação), sendo os estudos 1, 2, 4 e 5 (vide Tabela 4.4) são exemplos de estudos que a SCAS sugere para serem revisados manualmente. O estudo 183 é um exemplo de estudo classificado no quadrante 3 (*scores* baixos, mas com citação) e que deveria ser revisado de forma

Tabela 4.2. Informações da RS1 utilizada como exemplo

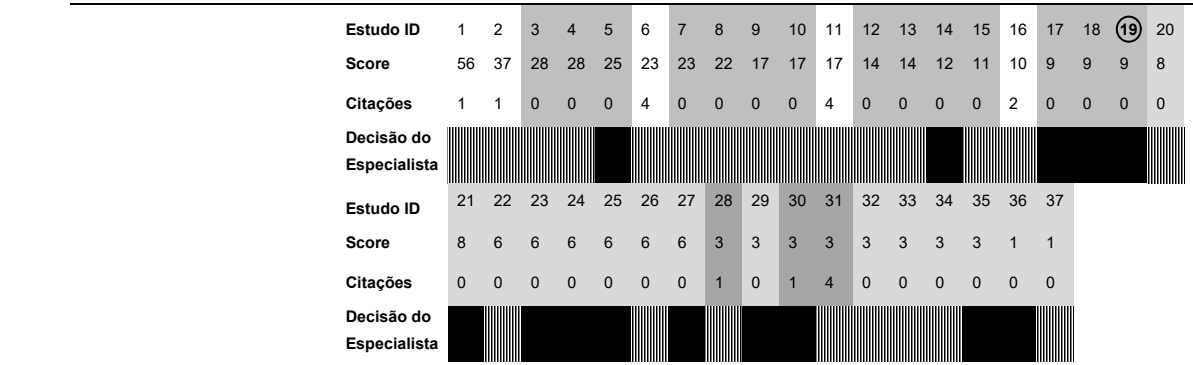
Estudo ID	Informações da RS1				
	Título	Autores	Referência	Temática	Total de Incluídos	Total de Excluídos
1	Experimenting with a multi-iteration systematic review in software engineering	Fabiano Cutigi Ferrari José Carlos Maldonado	ESELAW, 2008	Teste de software orientado a aspecto	34	63

Estudo ID	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	26	27	28	29	30	31	32	33	34	35
Score	78	69	65	56	49	46	44	43	38	38	37	35	35	35	32	32	32	32	31	31	31	29	29	28	26	26	25	25	24	23	22	22	22	22	22
Citações	0	1	1	0	35	6	0	18	0	0	9	3	1	0	0	0	7	0	1	1	0	0	0	0	4	9	0	8	0	0	0	0	0	1	0
Decisão do Especialista																																			
Estudo ID	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	**52**	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70
Score	20	20	20	18	18	17	17	16	16	16	16	16	15	15	15	14	14	13	13	13	12	12	12	11	11	10	10	9	9	9	9	9	9	9	9
Citações	0	0	2	0	10	6	0	0	3	0	2	0	1	0	0	6	0	0	0	0	0	0	2	0	0	0	0	0	0	0	1	0	3	0	0
Decisão do Especialista																																			
Estudo ID	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97								
Score	9	9	8	8	8	8	8	8	8	8	8	6	6	6	6	6	5	5	5	5	5	3	3	3	3	1	1								
Citações	7	6	0	3	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0								
Decisão do Especialista																																			

Legenda: Incluído | Excluído | Quadrante 1 | Quadrante 2 | Quadrante 3 | Quadrante 4 | Valor de corte

Tabela 4.3. Informações da RS2 utilizada como exemplo

Estudo ID	Informações da RS2					
	Título	**Autores**	**Referência**	**Temática**	**Total de Incluídos**	**Total de Excluídos**
2	Systematic literature reviews in software engineering – A systematic literature review	Barbara Kitchenham Pearl Brereton David Budgen Mark Turner John Bailey Stephen Linkman	IST, 51 (2009), 7–15	Engenharia de Software	23	14

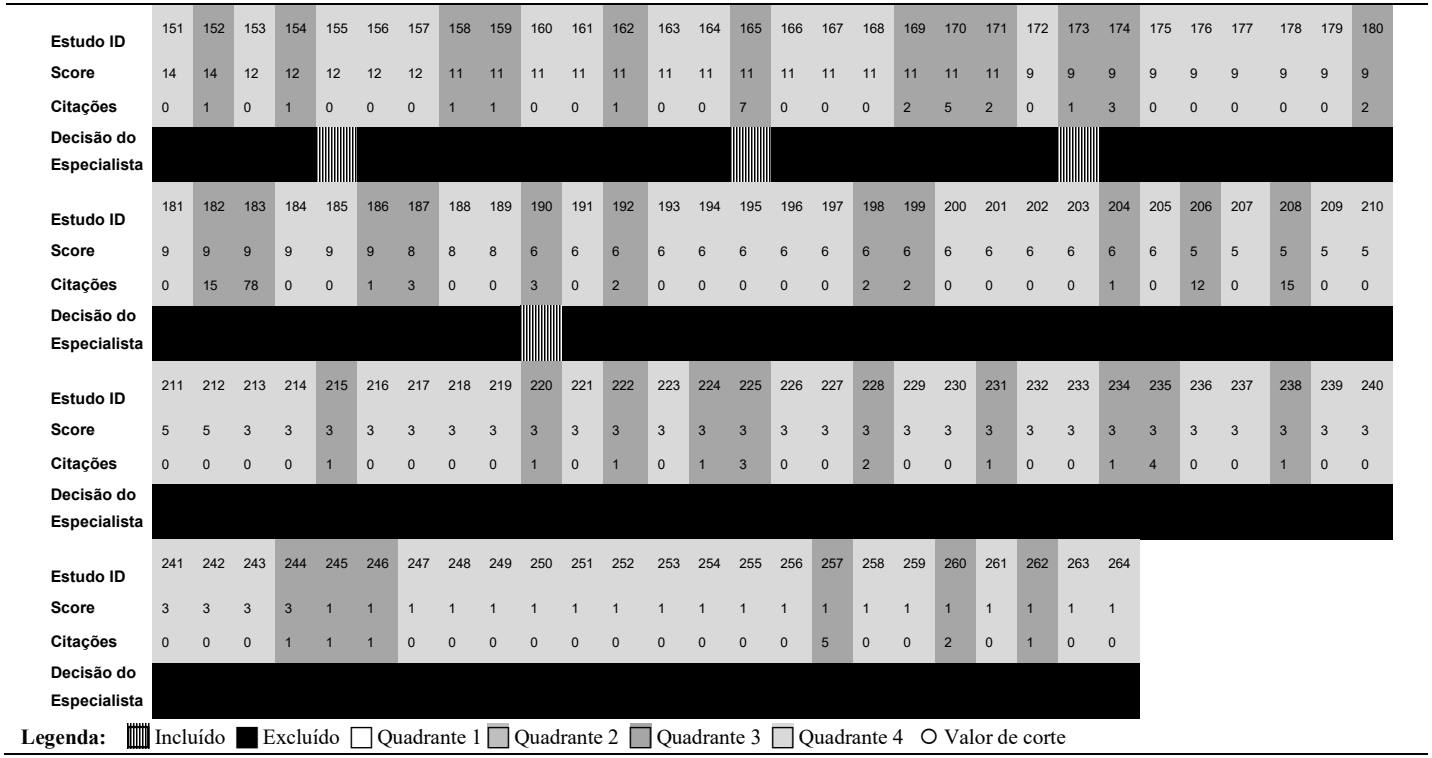


Legenda: Incluído Excluído Quadrante 1 Quadrante 2 Quadrante 3 Quadrante 4 Valor de corte

Tabela 4.4. Informações da RS3 utilizada como exemplo

Estudo ID	Informações da RS3					
	Título	**Autores**	**Referência**	**Temática**	**Total de Incluídos**	**Total de Excluídos**
3	Empirical studies of pair programming for CS/SE teaching in higher education: A systematic literature review	Norsaremah Salleh Emília Mendes John Grund	IEEE TSE 37(4), 509–522	Programação em pares	74	190

Estudo ID	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	26	27	28	29	30
Score	89	62	61	58	53	52	49	47	46	45	44	44	44	43	43	42	41	41	40	40	40	39	39	38	38	37	37	37	37	37
Citações	0	0	12	0	0	3	27	4	0	0	0	1	0	4	23	1	8	0	0	0	0	1	0	0	15	4	1	0	2	25
Decisão do Especialista																														
Estudo ID	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
Score	36	35	35	35	35	35	35	35	35	35	34	34	34	33	33	33	33	32	32	32	32	32	32	31	31	31	31	31	31	31
Citações	9	0	0	0	0	5	2	0	25	2	17	10	10	1	0	0	5	0	9	0	3	0	19	1	0	0	0	0	8	0
Decisão do Especialista																														
Estudo ID	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90
Score	30	30	30	29	29	29	29	29	28	28	28	28	28	28	26	26	26	26	26	26	26	26	26	25	25	25	24	24	23	23
Citações	10	42	0	0	0	0	0	3	16	4	6	1	1	6	1	3	3	0	1	29	0	0	0	3	0	0	0	0	3	1
Decisão do Especialista																														
Estudo ID	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120
Score	23	23	23	23	23	23	22	22	22	21	21	20	20	20	20	20	20	20	20	20	20	20	20	20	19	19	19	18	18	18
Citações	2	13	2	1	0	31	0	0	0	8	0	14	0	0	0	29	7	6	5	0	0	0	0	9	0	0	0	0	0	1
Decisão do Especialista																														
Estudo ID	121	122	123	124	125	126	127	128	129	130	131	**132**	133	134	135	136	137	138	139	140	141	142	143	144	145	146	147	148	149	150
Score	18	18	18	18	17	17	17	17	17	17	17	17	15	15	15	15	14	14	14	14	14	14	14	14	14	14	14	14	14	14
Citações	1	3	2	15	66	3	6	0	0	4	4	0	8	0	2	5	5	0	0	0	0	0	1	2	0	2	0	0	0	0
Decisão do Especialista																														


Estudo ID 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180
Score 14 14 12 12 12 12 12 11 11 11 11 11 11 11 11 11 11 11 11 11 11 9 9 9 9 9 9 9 9 9
Citações 0 1 0 1 0 0 0 1 1 0 0 1 0 0 7 0 0 0 2 5 2 0 1 3 0 0 0 0 0 2
Decisão do Especialista
Estudo ID 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210
Score 9 9 9 9 9 9 8 8 8 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 5 5 5 5 5
Citações 0 15 78 0 0 1 3 0 0 3 0 2 0 0 0 0 0 2 2 0 0 0 0 1 0 12 0 15 0 0
Decisão do Especialista
Estudo ID 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240
Score 5 5 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
Citações 0 0 0 0 1 0 0 0 0 1 0 1 0 1 3 0 0 2 0 0 1 0 0 1 4 0 0 1 0 0
Decisão do Especialista
Estudo ID 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264
Score 3 3 3 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
Citações 0 0 0 1 1 1 0 0 0 0 0 0 0 0 0 0 5 0 0 2 0 1 0 0
Decisão do Especialista
Legenda: Incluído Excluído Quadrante 1 Quadrante 2 Quadrante 3 Quadrante 4 Valor de corte

manual pelo pesquisador. Esse estudo, embora tenha *score* baixo, possui 78 citações.

A Tabela 4.5 compara as duas classificações de estudos primários pertencentes às RSs utilizadas no estudo de caso: (i) a classificação de cada estudo primário em um dos quadrantes por meio da estratégia SCAS; e (ii) a classificação de cada estudo primário (se incluído ou excluído) pelos especialistas que conduziram as RSs.

Tabela 4.5. Distribuição dos estudos primários das RSs nos quadrantes

RS	Total de estudos primários	Quadrante 1 Inclusão "Correta" ↑ *Score alto* Ao menos 1 *Citação*		Quadrante 2 Para ser revisado ↑ *Score alto* Sem Citação		Quadrante 3 Para ser revisado ↓ *Score baixo* Ao menos 1 *Citação*		Quadrante 4 Exclusão "Correta" ↓ Score baixo Sem Citação	
RS1	97	# Total	22	# Total	30	# Total	7	# Total	38
		# Incluídos	19	# Incluídos	10	# Incluídos	4	# Incluídos	1
		# Excluídos	3	# Excluídos	20	# Excluídos	3	# Excluídos	37
RS2	37	# Total	5	# Total	14	# Total	3	# Total	15
		# Incluídos	5	# Incluídos	9	# Incluídos	2	# Incluídos	7
		# Excluídos	0	# Excluídos	5	# Excluídos	1	# Excluídos	8
RS3	264	# Total	69	# Total	63	# Total	46	# Total	86
		# Incluídos	33	# Incluídos	29	# Incluídos	6	# Incluídos	6
		# Excluídos	36	# Excluídos	34	# Excluídos	40	# Excluídos	80

- **Discussão**

Para avaliar as sugestões apresentadas pela estratégia SCAS, foi assumido que as decisões tomadas pelos especialistas estivessem corretas, uma vez que as três RSs utilizadas no estudo de caso foram publicadas e conduzidas ou supervisionadas por pesquisadores experientes. Assim, as decisões tomadas pelos especialistas foram consideradas a base de comparação para avaliação das sugestões da estratégia SCAS.

Para a RS1, o esforço manual exigido na primeira fase da atividade de seleção, leitura de títulos e *abstracts*, foi reduzido em 61,85%; 60 estudos de um total de 97 (22 estudos pertencentes ao quadrante 1 e 38 pertencentes ao quadrante 4 – vide Tabela 4.5) poderiam ser classificados automaticamente utilizando as recomendações da estratégia SCAS. O percentual de erro foi de 4,12%; 4 estudos de um total de 97 (três pertencentes ao quadrante 1, e um pertencente ao quadrante 4) receberam classificação diferente pelos especialistas que realizaram a RS1. Os 3 estudos pertencentes ao quadrante 1 exigiram um esforço de leitura adicional, ou seja, 3 estudos "irrelevantes" precisaram ser lidos pelos pesquisadores seguindo as

recomendações da SCAS. O estudo pertencente ao quadrante 4 é uma decisão falso negativa, ou seja, é um estudo relevante que foi excluído pela SCAS.

Da mesma forma, os resultados da RS2 indicaram que a redução do esforço na atividade de seleção foi de 54,05%, e o percentual de erro foi de 18,91%; 20 estudos classificados nos quadrantes 1 e 4 poderiam ser classificados automaticamente. Vale destacar que, no caso da RS2, que é um estudo terciário, é esperado que a citação cruzada (citação entre os artigos da RS) seja baixa porque os estudos recuperados provavelmente não estejam relacionados uns aos outros.

Finalmente, os resultados da RS3 indicaram que a redução do esforço na atividade de seleção foi de 58,71% e o percentual de erro foi de 15,90%; 155 estudos pertencentes aos quadrantes 1 e 4 poderiam ser classificados automaticamente.

Com o objetivo de evidenciar que a SCAS fornece resultados melhores do que o uso das funcionalidades *score* e número de citações isoladamente, o percentual de erro foi calculado para cada um. Considerando-se que o recurso número de citações recomenda a inclusão de todos os estudos que tenham ao menos uma citação (FELIZARDO et al., 2011), um erro ocorre sempre que um estudo com citação é excluído pelo especialista ou um estudo sem citação é incluído pelo especialista. Considerando-se que a funcionalidade *score* recomenda a inclusão de todos os estudos com *score*s acima do valor de corte, um erro ocorre quando um estudo com *score* maior ou igual ao valor de corte é excluído pelo especialista ou um estudo com *score* inferior ao valor de corte é incluído pelo especialista. Portanto, com base nas Tabelas 4.2, 4.3 e 4.4, considerando-se o número de citações, os percentuais de erro foram: 16,49% para RS1, 45,95% para RS2 e 42,05% para RS3. Da mesma forma, considerando-se apenas o *score*, os percentuais de erro foram: 28,87% para RS1, 37,84% para RS2 e 31,06% para RS3. Assim, de acordo com a discussão anterior, a combinação dessas funcionalidades (*score* e número de citação) resultou em um percentual de erro menor do que a utilização dos recursos isoladamente.

Além disso, a fim de medir o nível de concordância entre SCAS e os especialistas, o coeficiente Cohen's kappa (CARLETTA, 1996) – também conhecido apenas por Kappa – foi calculado para cada RS utilizada no estudo de caso. Kappa é calculado por meio da equação k = (Pr (A) - Pr (e)) / (1-Pr (e)), em que Pr (A) é a concordância observada entre os avaliadores, e Pr (e) é a probabilidade hipotética

de concordância, utilizando os dados observados para calcular as probabilidades de cada observador aleatoriamente escolher cada categoria.

Se os avaliadores estão em completo acordo, então κ = 1. Se não houver acordo entre os avaliadores, então κ = 0. A Tabela 4.6 apresenta a interpretação dos valores Kappa sugerida por Landis e Koch (1977).

Tabela 4.6. Interpretação dos valores de Kappa (LANDIS; KOCH, 1977)

Valores de Kappa	Interpretação
<0	Sem concordância
0,00-0,19	Concordância pobre
0,20-0,39	Concordância distante
0,40-0,59	Concordância moderada
0,60-0,79	Concordância substancial
0,80-0,99	Concordância quase perfeita
1	Concordância perfeita

Com o objetivo de aplicar o coeficiente Kappa para comparar o nível de concordância entre os especialistas e a estratégia SCAS no estudo de caso, as seguintes premissas foram feitas: i) avaliadores – as decisões dos especialistas e as sugestões da SCAS; ii) categorias - incluídos e excluídos; e iii) dados observados - os estudos primários considerados pertencentes aos quadrantes 1 e 4, pois eles contêm os estudos que a SCAS recomenda incluir ou excluir automaticamente. Quadrantes 2 e 3 não foram considerados pois os pesquisadores devem rever manualmente os estudos que pertencem a eles.

A Tabela 4.7 apresenta os valores calculados para as RSs utilizadas no estudo de caso, bem como suas interpretações de acordo com a Tabela 4.6.

Tabela 4.7. Valores calculados e interpretação de Kappa para o estudo de caso

RS	Pr(a)	Pr(b)	Kappa	Interpretação
RS1	0,93	0,54	0,85	Concordância quase perfeita
RS2	0,65	0,45	0,36	Concordância distante
RS3	0,73	0,52	0,44	Concordância moderada
Geral	0,77	0,33	0,66	Concordância substancial

Com base nos dados apresentados na Tabela 4.7, é possível verificar que a RS1 alcançou um nível de concordância muito bom (concordância quase perfeita). Por outro lado, a mesma interpretação não se aplica para a RS2, principalmente porque se refere a um estudo terciário, que geralmente tem citação cruzada entre os estudos muito baixa. O Kappa geral, calculado com base nos dados das três RSs, mostrou um nível de concordância substancial, que é um bom resultado de acordo com a escala de interpretação apresentada na Tabela 4.6.

Por fim, considerando que a completude é fundamental para RSs, vale a ressalva que mesmo os seres humanos podem cometer erros e excluir um estudo relevante na triagem inicial. Vale ressaltar que uma das principais causas de erro está relacionada a títulos ruins ou *abstracts* mal escritos, nos quais um pesquisador pode não ser capaz de identificar palavras-chave relevantes ou um contexto relacionado às suas questões de pesquisa. Se os títulos e *abstracts* forem bem escritos e o pesquisador aplicar uma *string* de busca relativamente ampla, estudos relevantes provavelmente conterão as palavras-chave correspondentes à pesquisa de modo que esses documentos não pertencerão ao quadrante 4. Assim, uma decisão errada, provavelmente, será por conta do próprio pesquisador, uma vez que os estudos classificados no quadrante 1 serão aceitos automaticamente e os estudos selecionados para os quadrantes 2 ou 3 serão aceitos dependendo da decisão do pesquisador após revisão manual.

4.4 Experimento com Alunos de Pós-Graduação

Com o objetivo de se obter uma avaliação adicional da estratégia SCAS, foi realizado um experimento com estudantes de pós-graduação (todos doutorandos) de alguns departamentos da Universidade Federal de São Carlos durante uma disciplina específica de revisão sistemática da literatura.

4.4.1 Planejamento

Os principais objetivos do experimento são: (i) avaliar se a estratégia SCAS é mais eficiente do que a abordagem manual; (ii) avaliar o nível de concordância entre

SCAS e os revisores que realizaram a seleção inicial; e (iii) avaliar a eficácia da SCAS ao recomendar decisões para estudos com conflitos de decisão.

O experimento foi planejado usando o modelo Goal-Question-Metric (GQM) (BASILI et al, 1994), conforme apresentado na Figura 4.4. Quanto às métricas, que são as variáveis dependentes do experimento, é importante esclarecer o que exatamente elas pretendem medir:

- M1: Tempo gasto para gerar os quadrantes para os estudos usando a ferramenta StArt, ou seja, o tempo gasto para a aplicação de SCAS;
- M2: Tempo gasto pelos alunos para revisar manualmente os estudos pertencentes aos quadrantes 1 e 4. Consideramos apenas esses quadrantes porque são eles que a estratégia SCAS recomenda decisões automáticas;
- M3: Número de estudos incluídos pertencentes ao quadrante 1 de acordo com os membros do grupo, ou seja, quantos estudos os revisores incluíram uma vez que a recomendação da estratégia SCAS é para incluir esses estudos;
- M4: Número de estudos excluídos pertencentes ao quadrante 4 de acordo com os membros do grupo, ou seja, quantos estudos os revisores excluíram uma vez que a recomendação da estratégia SCAS é para excluir esses estudos;
- M5 / M6: Número de conflitos (divergências entre os revisores sobre a decisão de incluir ou não estudos) ocorridos na atividade de seleção inicial para estudos pertencentes aos quadrantes 1 e 4, respectivamente;
- M7: Número de estudos com conflitos de decisão pertencentes ao quadrante 1 que foram incluídos pelos revisores após discussão, ou seja, quantos conflitos foram resolvidos como incluídos uma vez que a recomendação SCAS é para incluir tais estudos;
- M8: Número de estudos com conflitos de decisão pertencentes ao quadrante 4 que foram excluídos pelos revisores após a discussão, ou seja, quantos conflitos foram resolvidos como excluídos uma vez que a recomendação SCAS é para excluir esses estudos.

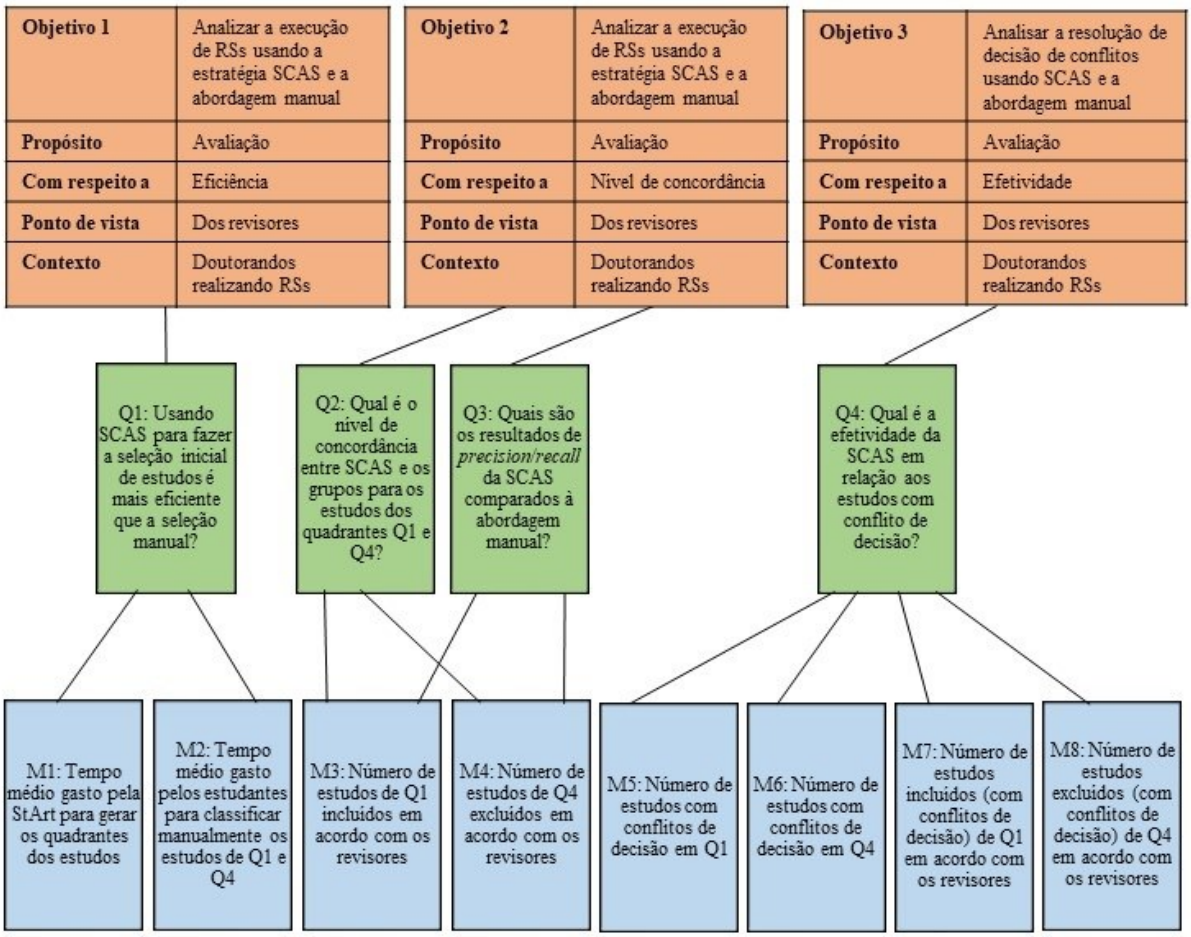


Figura 4.4. Planejamento do experimento usando GQM

4.4.2 Hipóteses

As hipóteses nulas do experimento são:

- H0,1: executar a atividade de seleção inicial usando SCAS não é mais eficiente do que a atividade de seleção inicial executada manualmente;
- H0,2: executar a atividade de seleção inicial usando o SCAS não fornece resultados semelhantes aos da atividade de seleção inicial manual;
- H0,3: SCAS não ajuda a resolver conflitos entre revisores sobre a decisão de incluir ou não estudos.

4.4.3 População

A população foi composta de 21 estudantes de alunos de pós-graduação (todos doutorandos), durante uma disciplina específica de revisão sistemática. Alguns doutorandos já conheciam o processo de RS e, para os demais, o mesmo foi apresentado durante a disciplina. Havia alunos de áreas de pesquisa distintas: doze da informática (engenharia de software), cinco da engenharia de produção e quatro da educação. Eles foram divididos em grupos de acordo com suas áreas de pesquisa, a fim de executar RSs para assuntos de interesse. Os estudantes de engenharia de software foram divididos em três grupos distintos de acordo com a subárea específica de pesquisa a que pertenciam.

É importante mencionar que a disciplina de revisão sistemática oferecida foi multidisciplinar, oferecido a alguns departamentos universitários e este é o principal motivo para que os participantes não estejam relacionados apenas com a disciplina de engenharia de software. Além disso, seria bom analisar se a estratégia SCAS proporcionaria resultados satisfatórios para outras disciplinas além da engenharia de software.

4.4.4 Operação

Durante a disciplina de RS, antes do experimento, todos os alunos foram treinados em como realizar um processo de RS e em como usar a ferramenta StArt para dar suporte a todo o processo de RS. A disciplina teve duração de 64 horas e 24 delas foram utilizadas para aprendizagem do processo de RS e da ferramenta StArt. Mesmo os alunos que já conheciam o processo de RS participaram da mesma fase de treinamento. Depois disso, os participantes foram divididos em cinco grupos com base em suas disciplinas de interesse, conforme apresentado na Tabela 4.8.

Tabela 4.8. Definição dos grupos para o experimento

Nº do Grupo	Membros	Área de Pesquisa
1	5	Engenharia de Produção
2	4	Educação
3	5	Engenharia de Software
4	4	Engenharia de Software
5	3	Engenharia de Software

O experimento foi realizado em dois dias:

Primeiro dia:

a) Os objetivos do experimento foram apresentados aos participantes;
b) Cada grupo escolheu um tema distinto de interesse relacionado à sua área de pesquisa;
c) Cada grupo criou um protocolo de RS e recuperou os estudos primários das bases de dados *online* que escolheram;
d) Cada grupo aplicou a estratégia SCAS para os estudos primários recuperados e obteve os quadrantes para os estudos, tomando nota do tempo gasto para a aplicação da SCAS;
e) Cada membro do grupo realizou manualmente a atividade de seleção inicial para os estudos pertencentes aos quadrantes 1 e 4, lendo os títulos e *abstracts*, e decidindo por incluir ou não esses estudos, tomando nota do tempo gasto para executar essa atividade. É importante ressaltar que os participantes foram convidados a analisar cuidadosamente esses estudos sem levar em consideração a classificação anterior fornecida pela estratégia SCAS;
f) Cada membro do grupo analisou e comparou as recomendações da estratégia SCAS para os quadrantes 1 e 4 com as decisões que ele tomou para os estudos pertencentes a esses quadrantes;
g) Cada membro do grupo respondeu um questionário informando seu nome, área de pesquisa e número do grupo; o tempo gasto para aplicar a estratégia SCAS e para realizar a seleção manual dos estudos pertencentes aos quadrantes 1 e 4; o número de estudos incluídos e excluídos pertencentes ao quadrante 1; o número de estudos incluídos e excluídos pertencentes ao quadrante 4.

Segundo dia:

h) Os membros do grupo se encontraram e identificaram os conflitos existentes em relação às decisões tomadas para os estudos pertencentes aos quadrantes 1 e 4;

i) Após a discussão, os membros do grupo tomaram uma decisão final sobre os estudos que possuíam conflitos de decisão;

j) Cada grupo analisou e comparou as recomendações da estratégia SCAS para os quadrantes 1 e 4 com as decisões finais tomadas pelo grupo para os estudos pertencentes a esses quadrantes;

k) Um membro de cada grupo respondeu um questionário informando a área de pesquisa e identificação do grupo; o número de estudos incluídos e excluídos pertencentes ao quadrante 1 após a reunião de consenso; o número de estudos incluídos e excluídos pertencentes ao quadrante 4 após a reunião de consenso; o número de estudos com conflitos de decisão pertencentes ao quadrante 1 e quantos deles foram incluídos e excluídos; o número de estudos com conflitos de decisão pertencentes ao quadrante 4 e quantos deles foram incluídos e excluídos.

4.4.5 Resultados e Análises

A última tarefa de cada grupo durante o experimento foi responder um questionário sobre os resultados obtidos. A Tabela 4.9 apresenta as respostas dadas por cada grupo após a realização da atividade de seleção inicial e a reunião de consenso. Ele mostra as decisões tomadas para estudos pertencentes a quadrantes 1 (Q1) e 4 (Q4), ou seja, quantos estudos primários foram incluídos e excluídos nos quadrantes 1 e 4, respectivamente. Não foram considerados os estudos pertencentes aos quadrantes 2 e 3 porque eles deveriam ser revisados manualmente usando a estratégia SCAS ou não.

As decisões tomadas pelos participantes foram a base de comparação para avaliar as recomendações da SCAS. Eles escolheram os tópicos de pesquisa com que eram familiarizados e tinham mais experiência. Além disso, todas as decisões finais relativas aos estudos foram tomadas após uma reunião de consenso, onde os estudos com conflito foram discutidos para se tomar uma decisão final.

Tabela 4.9. Decisões tomadas para os estudos dos quadrantes 1 e 4

	Nº Grupo	1	2	3	4	5
	Área de Pesquisa	Engenharia de Produção	Educação	Engenharia de Software	Engenharia de Software	Engenharia de Software
	Total de estudos	256	260	269	252	154
	Tempo médio gasto para SCAS	4	6	4	3	2
	Tempo médio gasto por revisores	82	105	85	115	88
Q1	Total de estudos	13	15	35	40	10
	Total de incluídos	9	10	23	24	8
	Total de excluídos	4	5	12	16	2
Q4	Total de estudos	57	44	23	19	10
	Total de incluídos	2	3	2	0	1
	Total de excluídos	55	40	21	19	9

Em relação à eficiência, considerando a hipótese H0,1, ao compararmos o tempo gasto para adotar as recomendações SCAS com o tempo gasto para realizar a revisão manual de estudos dos quadrantes 1 e 4, é possível observar (Tabela 4.9) que a aplicação da estratégia SCAS é muito mais rápida. É importante mencionar que o tempo gasto para aplicar a SCAS pode variar dependendo da configuração do computador onde a ferramenta StArt foi instalada, não está apenas relacionado ao número de estudos processados. Em média, a estratégia SCAS levou cerca de quatro minutos para executar, enquanto os revisores levaram cerca de 95 minutos para completar a seleção inicial com base em títulos e *abstracts* dos estudos primários. Isso significa que a estratégia SCAS é mais eficiente do que a revisão manual, como esperado, o que não confirma a hipótese H0,1.

Considerando o número de estudos a serem avaliados na seleção inicial, o grupo 1 teve uma redução de esforço de 27,34% (70 estudos classificados automaticamente de um total de 256 estudos recuperados). Da mesma forma, o grupo 2 obteve uma redução de esforço de 22,7%, o grupo 3 obteve de 21,56%, o grupo 4 obteve de 23,41% e o grupo 5 obteve de 13%.

O percentual de erro para o grupo 1 foi de 2,34%: 6 estudos de um total de 256, sendo 4 pertencentes ao quadrante 1 e 2 pertencentes ao quadrante 4, receberam classificação diferente por revisores do grupo 1. Os quatro estudos pertencentes ao quadrante 1 são decisões falso positivas e exigiriam esforço de

leitura adicional, ou seja, quatro estudos "irrelevantes" que precisariam ser lidos pelos revisores seguindo a recomendação da estratégia SCAS de aceita-los. Os dois estudos pertencentes ao quadrante 4 são decisões falso negativas, ou seja, são estudos relevantes que foram excluídos pela estratégia SCAS. Da mesma forma, o percentual de erro para o grupo 2 foi de 3,08% (5 falsos positivos e 3 falsos negativos), para o grupo 3 foi de 5,2% (12 falsos positivos e 2 falsos negativos), para o grupo 4 foi de 6,35% (16 falsos positivos e nenhum falso negativo), e para o grupo 5 foi de 1,95% (2 falsos positivos e um falso negativo). Portanto, a perda de evidências (decisões falso negativas) é muito baixa em comparação ao tempo economizado usando a estratégia SCAS. Alguns estudos irrelevantes foram incluídos para a leitura de texto completo, mas provavelmente seriam excluídos logo depois de ler a introdução ou a conclusão de seus textos completos.

A Tabela 4.10 sintetiza a análise realizada para os dados apresentados na Tabela 4.9, em que a coluna de tempo economizado refere-se ao tempo gasto pelos revisores menos o tempo gasto ao aplicar a estratégia SCAS; a coluna de redução de esforço refere-se ao número de estudos que não precisariam ser lidos na RS; a coluna de percentual de erro refere-se ao número de estudos incorretamente classificados pela estratégia SCAS de acordo com as decisões dos revisores; e a coluna de evidências perdidas mostra o número de decisões falso negativas.

Tabela 4.10. Resumo da análise feita para os grupos

Grupo	Tempo Economizado	Redução de Esforço	Percentual de Erro	Evidências Perdidas
1	78 minutos	27,34%	2,34%	2
2	99 minutos	22,70%	3,08%	3
3	88 minutos	21,56%	5,20%	2
4	112 minutos	23,41%	6,35%	0
5	86 minutos	13,00%	1,95%	1

Além disso, para medirmos o nível de concordância entre a estratégia SCAS e os grupos, calculou-se o coeficiente Kappa (CARLETTA, 1996). As seguintes associações foram feitas: i) avaliadores - as decisões do grupo e as recomendações SCAS; ii) categorias - incluídas e excluídas; e iii) dados observados - os estudos primários considerados pertencentes ao quadrante 1 e ao quadrante 4, pois contêm

os estudos que SCAS recomenda incluir ou excluir automaticamente. Novamente os quadrantes 2 e 3 não foram considerados como pesquisadores devem revisar manualmente os estudos que pertencem a eles, como explicado anteriormente.

A Tabela 4.11 apresenta os valores calculados para cada grupo (revisão manual e SCAS) e suas interpretações. É possível verificar que os grupos 1, 2 e 5 alcançaram uma concordância substancial ao se comparar as decisões do grupo com as recomendações da estratégia SCAS. No entanto, a mesma interpretação não vale para os grupos 3 e 4 que alcançaram um nível de concordância moderado. O Kappa geral, calculado com base nos dados dos cinco grupos, mostra um nível substancial de concordância, o que indica ser um bom resultado.

Tabela 4.11. Valores e interpretações do Kappa para os grupos

Grupo	Kappa	Interpretação
1	0,70	Concordância substancial
2	0,62	Concordância substancial
3	0,53	Concordância moderada
4	0,49	Concordância moderada
5	0,70	Concordância substancial
Geral	0,62	Concordância substancial

Uma análise adicional foi realizada ao serem calculadas as métricas de precisão e revocação (*precision/recall*) para cada RS conduzida pelos grupos. Lembrando que, na recuperação de informações com classificação binária, a precisão é a fração de instâncias recuperadas que são relevantes, enquanto que a revocação é a fração de instâncias relevantes que são recuperadas. No experimento em questão, como o objetivo é avaliar os quadrantes 1 e 4 para os quais a SCAS recomenda decisões automáticas, a precisão e a revocação foram calculadas para as RSs considerando apenas os estudos pertencentes a esses quadrantes, conforme apresentado na Tabela 4.12, que também inclui a precisão e revocação gerais considerando dados dos cinco grupos.

Tabela 4.12. Precisão e revocação referentes às RSs conduzidas

Grupo	Precisão	Revocação
1	69,23%	81,82%
2	66,67%	76,92%
3	65,71%	92%
4	60%	100%
5	80%	88,89%
Geral	65,49%	90,24%

Assim, considerando a hipótese H0,2, com base no percentual de erro médio baixo (vide Tabela 4.10), no valor Kappa calculado que mostra o nível geral de concordância entre a SCAS e os revisores como substancial (vide Tabela 4.11) e no bom resultado de revocação geral (vide Tabela 4.12), podemos deduzir que a estratégia SCAS fornece resultados semelhantes a uma atividade de seleção inicial totalmente manual, o que não confirma a hipótese H0,2.

Os conflitos de decisão que ocorreram para os quadrantes 1 e 4 também foram analisados para verificar se a estratégia SCAS seria útil em caso de divergência de opiniões entre os revisores. A Tabela 4.13 mostra os conflitos de decisão relatados por cada grupo e as decisões correspondentes que tomaram depois da reunião de consenso. Considerando-se apenas os conflitos, o índice de acerto (quando as recomendações da estratégia SCAS e as decisões dos revisores são as mesmas) para o grupo 1 foi de 80%, ou seja, 16 dos 20 conflitos (3 para o quadrante 1 e 13 para o quadrante 4) foram resolvidos de acordo com as recomendações da SCAS. O índice de acerto para o grupo 2 foi de 50%, para o grupo 3 foi de 60%, para o grupo 4 foi de 53,85% e para o grupo 5 foi de 25%. Os resultados, com exceção do grupo 1, não são bons o suficiente para provar que a estratégia SCAS é útil na resolução de conflitos de decisões. No entanto, quando analisamos as Tabelas 4.9 e 4.13 em conjunto, é possível notar que todos os falsos negativos, que são o pior caso porque implicam em perda de evidência, referem-se a estudos em conflito entre os membros do grupo. Isso significa que a estratégia SCAS excluiu alguns estudos que não tinham concordância total entre os revisores. Assim, podemos imaginar, a partir de uma análise superficial, que esses estudos provavelmente não são os estudos mais relevantes das RSs por causa do

desentendimento dos revisores e porque obtiveram *scores* baixos uma vez que foram classificados no quadrante 4.

Tabela 4.13. Conflitos relatados para os quadrantes 1 e 4 e suas resoluções

	Grupo	**1**	**2**	**3**	**4**	**5**
Q1	Total de estudos	13	15	35	40	10
	Total de conflitos	5	7	18	23	3
	Total de conflitos resolvidos como inclusão	3	0	12	11	1
	Total de conflitos resolvidos como exclusão	2	7	6	12	2
Q4	Total de estudos	57	44	23	19	10
	Total de conflitos	15	13	2	3	1
	Total de conflitos resolvidos como inclusão	2	3	2	0	1
	Total de conflitos resolvidos como exclusão	13	10	0	3	0

Assim, considerando a hipótese H0,3, com base nos dados da Tabela 4.13, a estratégia SCAS estava correta em 58,89% das sugestões feitas para as resoluções de conflito, o que não é suficiente para provar que é útil na resolução de conflitos de decisão.

4.4.6 Ameaças à Validade

Com base nas ameaças à validade mencionadas em (WOHLIN et al, 2000), é possível destacar as seguintes ameaças:

- **Validade interna e de construção:** os tópicos de pesquisa e o nível de experiência dos participantes são ameaças, uma vez que possivelmente começaram seus papéis como pesquisadores em diferentes períodos, são oriundos de áreas distintas e também possuem níveis de experiência distintos na realização de RSs. Para minimizar essas ameaças, foram selecionados estudantes de pós-graduação (doutorandos), os quais devem realizar pesquisas de modo habitual, imaginando, assim, estarem em um nível de maturidade aceitável como pesquisadores. Além disso, eles participaram de um curso de 64 horas sobre todo o processo de RS para padronizar o nível

de conhecimento antes da execução das RSs e escolheram tópicos de pesquisa com os quais eram familiarizados e capazes de avaliar estudos.

- **Validade da conclusão:** a execução de uma RS pode ser um processo subjetivo, sendo influenciada pelos perfis dos participantes, pelo nível de experiência e pelo nível de compreensão que eles adquiriram na fase de treinamento. Além disso, os participantes realizaram a classificação manual de estudos tendo conhecimento prévio da classificação feita pela estratégia SCAS. Tentando minimizar essa ameaça, as decisões tomadas pelos membros do grupo foram comparadas entre eles em uma reunião de consenso. Os percentuais de erro entre as recomendações da estratégia SCAS e as decisões tomadas pelos grupos foram comparadas e o nível de concordância foi avaliado por meio do coeficiente Kappa, mas somente após os grupos tomarem as decisões finais relativas aos estudos.
- **Validade externa:** é plausível dizer que os resultados obtidos podem ser diferentes em outro conjunto de participantes. Tentando minimizar essa ameaça, foi selecionada uma população de pesquisadores composta por doutorandos, imaginando-se ter um nível aceitável de maturidade para a realização de pesquisas. Além disso, a generalização de nossos resultados está sujeita a certas limitações, principalmente porque apenas três tópicos de engenharia de software foram analisados, bem como apenas outras duas áreas de pesquisa (engenharia de produção e educação) além da engenharia de software participaram do experimento.

4.5 Considerações Finais

Este capítulo apresentou a estratégia semiautomática SCAS para auxiliar na atividade de seleção de estudos primários. Ela foi definida com base na funcionalidade *score* calculada para cada estudo primário em conjunto com a existência ou não de citações para o estudo primário. Com tal combinação, é

possível classificar os estudos primários em quadrantes e obter as recomendações sugeridas pela estratégia de incluir automaticamente os estudos pertencentes ao quadrante 1, excluir automaticamente os estudos pertencentes ao quadrante 4 e fazer a revisão manual tradicional para os estudos pertencentes aos quadrantes 2 e 3.

A estratégia foi submetida a avaliações por meio de um estudo de caso e de um experimento, e, em ambos, os resultados foram positivos. Isso levou à continuidade da pesquisa para melhorar a estratégia inicial definida e, com base em sua aplicação nos estudos experimentais, foi possível detectar possíveis pontos de melhorias. Isso levou à definição de uma nova estratégia melhorada chamada SCAS-Fuzzy, a qual é apresentada no próximo capítulo.

Capítulo 5

ESTRATÉGIA SEMIAUTOMÁTICA SCAS-FUZZY PARA SELEÇÃO DE ESTUDOS

Este capítulo descreve o funcionamento da estratégia SCAS-Fuzzy, que é uma evolução da estratégia SCAS para seleção semiautomática de estudos primários. Apresenta ainda os resultados obtidos por meio de um estudo de caso com a aplicação da estratégia melhorada em revisões sistemáticas e comparações com a estratégia original.

5.1 Considerações Iniciais

A estratégia SCAS mostrou bons resultados para apoiar a seleção de atividade de estudos primários, conforme relatado no capítulo anterior. No entanto, alguns pontos de melhorias foram detectados, especialmente relacionados ao ponto de corte para determinação de *scores* altos e baixos e à criação de um coeficiente de citação com base no número de citações e no ano de publicação dos estudos primários, e não apenas considerar se um estudo é citado ou não. Tais melhorias geraram uma evolução da estratégia inicial, a qual foi chamada de estratégia SCAS-Fuzzy, uma vez que utiliza lógica fuzzy (ZADEH, 1965) para lidar com a imprecisão com relação à relevância das funcionalidades *score* e coeficiente de citação. Na sequência são apresentadas as melhorias realizadas na estratégia SCAS original e o funcionamento da estratégia melhorada SCAS-Fuzzy.

Este capítulo está organizado da seguinte forma: na Seção 5.2 são apresentadas as melhorias realizadas na estratégia SCAS-Fuzzy em relação à estratégia SCAS; na Seção 5.3 é apresentada a descrição da estratégia SCAS-Fuzzy definida; e, por fim, na Seção 5.4 são apresentadas as considerações finais deste capítulo.

5.2 Melhorias Realizadas

Esta seção apresenta as duas melhorias que fizeram a estratégia SCAS evoluir para a versão SCAS-Fuzzy. Essas melhorias correspondem à maneira como a citação de um estudo é tratada e à utilização da lógica fuzzy para decidir sobre a aceitação de um estudo.

5.2.1 Coeficiente de Citação

A primeira melhoria realizada em relação à estratégia SCAS original foi a criação de um coeficiente de citação para os estudos. Ele leva em consideração quantas vezes um estudo é citado pelos demais estudos de uma RS e seu ano de publicação. A estratégia original só considera se um estudo é citado ou não, não importa se ele possua uma ou várias citações, ou se tenha sido publicado recentemente ou não. A ideia é que quanto mais velho for um estudo, maior será a chance de ser citado por outros estudos. Assim, entende-se que um estudo publicado em 2015 que é citado somente uma vez deva possivelmente possuir um coeficiente de citação maior do que um estudo publicado em 2010 que também é citado apenas uma vez. No entanto, ao considerarmos estudos com diferentes números de citações e anos de publicação distintos, é difícil estabelecer o nível de importância desses estudos no contexto de uma RS.

A técnica para se criar um coeficiente de citação foi baseada no número de citações e no ano de publicação dos estudos. É resultante de uma equação que multiplica dois termos principais, um baseado em citações e outro baseado no ano de publicação. A seguinte equação foi estabelecida para se calcular o coeficiente de citação para os estudos pertencentes a uma RS, na qual CIT é o número de citações

de um estudo, ANO é o ano de publicação de um estudo, AMR é um valor constante que representa o ano mais recente de publicação dos estudos da RS em execução, e índice é um valor constante utilizado para penalizar (dar menos peso) ao ano de publicação em relação ao número de citações de um estudo:

$$COEF\_CIT = (1 + CIT) * (1 – índice * (AMR – ANO))$$

A equação para calcular o coeficiente de citação foi inicialmente estabelecida considerando os dados das RSs utilizadas no estudo de caso que permitiram avaliar a estratégia SCAS original e que foi apresentado na Seção 4.3. Os estudos pertencentes a essas RSs foram inicialmente ranqueados por um novo valor obtido pela multiplicação de seus *scores* por seus coeficientes de citação calculados. A funcionalidade *score* não pôde ser ignorada para classificar os estudos porque é o principal pilar da estratégia SCAS para classificar os estudos nos quadrantes de inclusão e exclusão (OCTAVIANO et al, 2015). Após o ranqueamento, índices multiplicadores distintos foram testados na fórmula de cálculo do coeficiente de citação a fim de se conseguir valores resultantes diversos, uma vez que as demais variáveis da fórmula do coeficiente de citação (CIT, ANO e AMR) são provenientes dos estudos. A ideia foi avaliar qual índice proporcionaria melhores resultados para o coeficiente de citação combinado com a funcionalidade *score*. Assim, para cada índice, os estudos foram classificados pelo seu novo valor resultante em ordem decrescente. Os índices testados foram: 0, 0,001, 0,01, 0,1, 0,2, 0,3, 0,5, 0,7, 0,8 e 1. Depois disso, as posições de classificação dos estudos incluídos pelos especialistas foram somadas para cada índice distinto. A Figura 5.1 mostra um exemplo de ranqueamento de alguns estudos primários da RS1 utilizada no estudo de caso inicial, testando o índice 0,05 para cálculo do coeficiente de citação. As posições destacadas em azul são dos estudos incluídos e as mesmas foram somadas para se obter a soma final de posições do índice 0,05 para a RS1.

RS1	Índice: 0,05						
ID Paper	Score	Year	Citações	Coef_Cit	Valor	Status	Posição
207	49	2004	35	30,6	1499,400	A	1
297	43	2003	18	15,2	653,600	R	2
286	37	2006	9	9,5	351,500	A	3
261	46	2004	6	6,0	273,700	A	4
291	26	2006	9	9,5	247,000	R	5
260	32	2005	7	7,2	230,400	R	6
283	18	2006	10	10,5	188,100	A	7
298	25	2002	8	6,8	168,750	R	8
212	69	2006	1	1,9	131,100	A	9
225	35	2005	3	3,6	126,000	R	10
217	65	2006	1	1,9	123,500	A	11
257	26	2005	4	4,5	117,000	R	12
243	17	2005	6	6,3	107,100	R	13
211	14	2006	6	6,7	93,100	R	14
252	78	2006	0	1,0	74,100	A	15
254	35	2007	1	2,0	70,000	R	16
285	9	2005	7	7,2	64,800	A	17
242	16	2006	3	3,8	60,800	A	18
255	31	2006	1	1,9	58,900	R	19
292	9	2005	6	6,3	56,700	R	20
245	56	2007	0	1,0	56,000	A	21

Figura 5.1. Exemplo de ranqueamento para escolha do melhor índice

Da mesma forma, a soma das posições dos estudos incluídos foi calculada para cada índice para todas as RSs do estudo de caso inicial. O menor valor de soma indicaria, para cada RS, qual o melhor índice a ser usado uma vez que os estudos mais relevantes se encontrariam nas posições de ranqueamento mais baixas. Após análise de todos os valores resultantes das somas para cada índice e para cada RS, o índice 0,05 foi escolhido para ser utilizado na equação do coeficiente de citação pois mostrou os melhores resultados.

Um exemplo de cálculo do coeficiente de citação pode ser visto na Figura 5.1. Para um estudo (vide estudo #207) citado 35 vezes pelos outros estudos da mesma RS, publicado em 2004, e pertencente a uma RS cujo estudo mais recente recuperado é de 2007, o coeficiente de citação para este estudo seria:

$$COEF\_CIT = (1 + 35) * (1 - 0,05 * (2007 - 2004)) = 30,6.$$

Para esse mesmo estudo, o valor para ranqueamento, calculado ao multiplicarmos o seu *score* por seu coeficiente de citação, seria igual a 1499,4 (49 multiplicado por 30,6), conforme pode ser observado na Figura 5.1.

5.2.2 Uso de Lógica Fuzzy

A lógica fuzzy foi introduzida por Zadeh (1965) e propõe, ao invés de simplesmente usar o verdadeiro ou o falso, o uso de uma variação de valores entre uma completa afirmação e uma absoluta negação. Na teoria de conjuntos clássica, existem apenas duas possibilidades de pertinência para um elemento em relação a um conjunto como um todo: o elemento pertence ou não a um conjunto (ARTERO, 2009). Na lógica fuzzy, a falta de precisão é expressa de maneira quantitativa, na qual os valores pertencem ao intervalo fechado real entre 0 e 1 por meio de uma função de pertinência (LUGER, 2014). O processo de conversão de um número real em sua representação fuzzy é chamado de "fuzzyficação".

Outro conceito importante na lógica fuzzy está relacionado às regras que utilizam variáveis linguísticas na execução do processo de suporte à decisão. As variáveis linguísticas são identificadas por nomes, têm um conteúdo variável e assumem valores linguísticos que são nomes de conjuntos difusos (ARTERO, 2009). As variáveis linguísticas de entrada são chamadas de antecedentes enquanto as de saída são chamadas consequentes. No contexto deste trabalho, as variáveis linguísticas são o *score* e o coeficiente de citação. Elas podem assumir os valores nebulosos "baixos", "médios" e "altos", que serão representados por uma função de pertinência. Uma função de pertinência é uma curva que define como cada ponto no espaço de entrada é mapeado para um valor de pertinência (ou grau de pertinência) com uma variação que cobre o intervalo entre 0 e 1, operando no domínio de todos os valores possíveis dentro do intervalo (ZADEH, 1965).

O uso da lógica fuzzy para a funcionalidade *score* visa a evitar uma situação de corte abrupta como acontecia na estratégia SCAS original. Por exemplo, se o valor de corte fosse o *score* 17, o *score* 17 seria considerado alto, porém o score 16 já seria considerado baixo. Aplicando a lógica fuzzy, é possível definir uma função de pertinência para a variável linguística *score*, o que torna a classificação de um *score* mais imprecisa. Nesse contexto, por exemplo, um *score* pode ter um valor de pertinência 0,6 de ser alto e um valor de pertinência 0,4 de ser médio, sendo que a

sua classificação de acordo com a estratégia SCAS original seria simplesmente o valor alto ou o valor baixo.

Do mesmo modo, o mesmo conceito foi utilizado para o coeficiente de citação. Por meio de lógica fuzzy, é possível definir uma função de pertinência para ele, tornando sua classificação imprecisa. Assim, por exemplo, um coeficiente de citação pode ter um valor de pertinência 0,3 de ser baixo e um valor de pertinência 0,7 de ser médio. A estratégia SCAS original considera apenas se um estudo é citado ou não, independentemente do total de citações e sem considerar o seu ano de publicação.

Além da definição das variáveis linguísticas (*score* e coeficiente de citação), seus possíveis valores (baixo, médio ou alto) e suas funções de pertinência, é necessário definir também as regras de inferência fuzzy, as quais são responsáveis por tomar decisões com base nos valores (conjuntos nebulosos) das variáveis linguísticas. Elas devem representar o valor de saída como um conjunto nebuloso também. No contexto deste trabalho, os possíveis valores de saída (decisão tomada para um estudo) são "inclusão automática", "exclusão automática" ou "revisão manual", também definidos por meio de funções de pertinência de saída.

Depois da aplicação das regras de inferência, pode ser interessante que o sistema fuzzy produza uma saída numérica ao invés de termos linguísticos. Este processo que mapeia saídas difusas para valores numéricos é chamado de "defuzzyficação" (KLIR; YUAN, 1995). Três dos principais métodos para o processo de defuzzificação são o critério máximo (MAX), a média do máximo (MdM) e o centro de gravidade (CdG). O método CdG foi adotado por ser o método de defuzzificação mais comum. Todos os detalhes sobre as definições de funções de pertinência para o *score*, coeficiente de citação e variáveis de saída, além das regras de inferência fuzzy para apoiar a tomada de decisão são apresentadas a seguir.

5.3 Descrição da Estratégia

A estratégia SCAS-Fuzzy possui duas fases. A primeira fase baseia-se em duas características: i) o *score*, que apoia a análise dos estudos primários com base em seu conteúdo, como mencionado anteriormente; e ii) o coeficiente de citação,

que é calculado considerando quantas vezes um estudo é citado por outros estudos pertencentes à mesma RS e seu ano de publicação. Nessa fase, os *scores* e os coeficientes de citação dos estudos devem ser calculados e normalizados. A segunda fase é baseada em um sistema de inferência fuzzy predefinido, que combina as características do *score* e do coeficiente de citação e classifica os estudos em três categorias: inclusão automática, exclusão automática e revisão manual. O sistema de inferência fuzzy utiliza variáveis linguísticas, seus valores, funções de pertinência e regras de inferência pré-definidos, como é apresentado melhor na sequência.

Como mencionado anteriormente, as variáveis linguísticas definidas para a estratégia são o *score* e o coeficiente de citação. Os valores nebulosos definidos para essas variáveis são: baixo, médio ou alto. Suas funções de pertinência foram inicialmente definidas com base na análise de dados de duas de três RSs utilizadas no estudo de caso inicial relatado na Seção 4.3. A RS terciária usada naquele estudo de caso foi descartada porque a citação cruzada entre os estudos é muito baixa em RSs terciárias, uma vez que seus estudos primários são na verdade estudos secundários (RSs). Depois de se calcular e normalizar os *scores* e os coeficientes de citação de cada RS, os estudos foram primeiramente classificados por seus *scores* normalizados. Então, com base na observação de decisões tomadas pelos especialistas, foi possível definir as funções de pertinência iniciais para a variável *score*. Em seguida, os estudos foram classificados por seus coeficientes de citação normalizados e, também com base na observação de decisões tomadas pelos especialistas, foi possível definir as funções de pertinência iniciais para a variável coeficiente de citação. As funções iniciais de pertinência foram melhoradas considerando dados de outras duas RSs usadas em um novo estudo de caso, que é apresentado no Capítulo 6. O mesmo processo de classificação e observação foi usado para as novas RSs a fim de se obter uma calibragem ainda melhor dessas funções de pertinência.

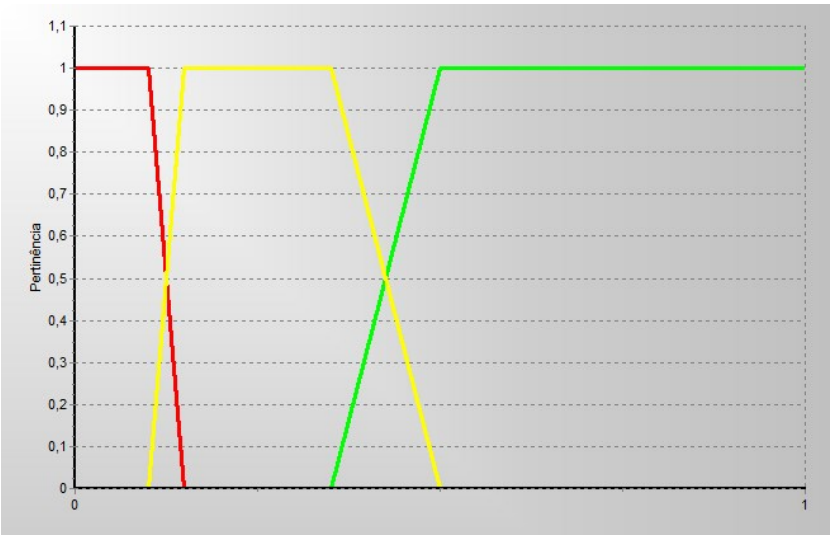
Os valores das variáveis de saída foram definidos de acordo com as três categorias esperadas para um estudo: inclusão automática, exclusão automática e revisão manual. Assim, foi possível definir as regras de inferência, que usa duas variáveis linguísticas de entrada (antecedentes) e uma de saída (consequente). As regras de inferência definidas são apresentadas na Tabela 5.1.

Tabela 5.1. Regras de inferência da estratégia SCAS-Fuzzy

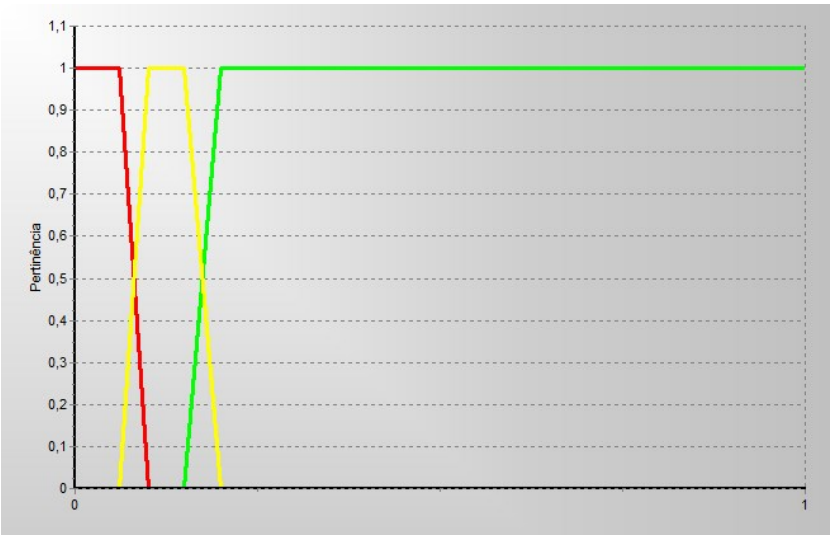
Regra		Antecedente		Consequente
1	Se	*Score* = “alto” E citação = “alta”	Então	“Inclusão automática”
2	Se	*Score* = “alto” E citação = “média”	Então	“Inclusão automática”
3	Se	*Score* = “alto” E citação = “baixa”	Então	“Revisão manual”
4	Se	*Score* = “médio” E citação = “alta”	Então	“Inclusão automática”
5	Se	*Score* = “médio” E citação = “média”	Então	“Revisão manual”
6	Se	*Score* = “médio” E citação = “baixa”	Então	“Revisão manual”
7	Se	*Score* = “baixo” E citação = “alta”	Então	“Revisão manual”
8	Se	*Score* = “baixo” E citação = “média”	Então	“Exclusão automática”
9	Se	*Score* = “baixo” E citação = “baixa”	Então	“Exclusão automática”

As funções de pertinência para a variável de saída, assim como as demais, foram inicialmente definidas com base na observação. Funções iniciais foram definidas e, depois do processamento das variáveis de entrada para todos os estudos pertencentes às RSs escolhidas, foi possível calibrá-las observando-se o valor numérico de saída gerado pela defuzzyficação e as classificações sugeridas para os estudos.

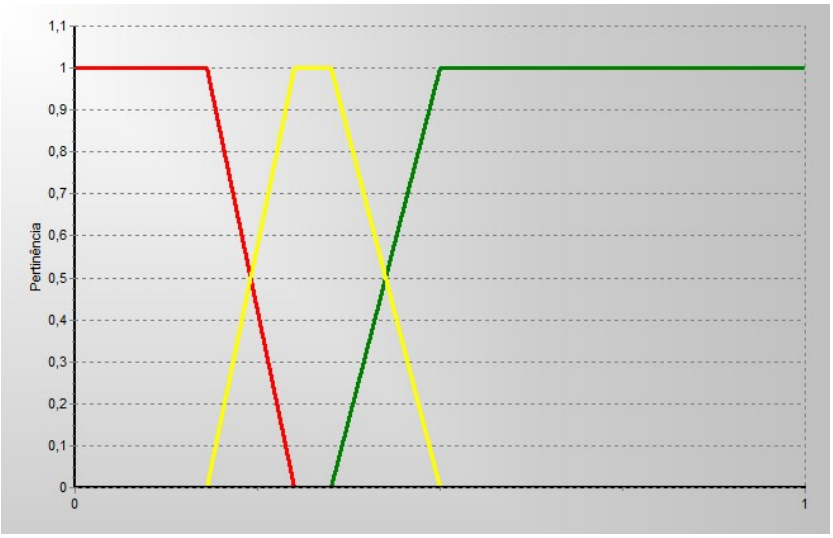
As primeiras funções de pertinência definidas para as variáveis linguísticas após a calibração são mostradas na Figura 5.2. O eixo vertical indica a pertinência (valor entre 0 e 1) de cada variável de entrada ou saída analisada, Quanto mais próximo de um, maior a pertinência de o resultado ser o correto. O eixo horizontal indica o valor normalizado em função do maior valor encontrado para a variável analisada, sendo um número entre 0 e 1. Por exemplo, na Figura 5.2.a, que corresponde à primeira função de pertinência, é possível observar que quanto mais próximo de zero o valor normalizado do score (eixo horizontal), maior a pertinência de o artigo ser excluído da RS (linha vermelha). Se observarmos a linha verde, é possível notar que próximo aos 50% do score normalizado, a pertinência de um artigo ser incluído é bastante elevada. A linha amarela destaca a região de scores que seriam considerados médios pelo fato de não se ter uma decisão clara sobre incluir ou excluir um artigo da RS, sugerindo a revisão manual.



(a)



(b)



(c)

Figura 5.2. Funções de pertinência para as variáveis linguísticas *score* (a), coeficiente de citação (b) e saída (c)

Como as funções de pertinência iniciais foram definidas com base na observação de dados, optou-se então pela utilização de outro recurso de inteligência computacional para verificar se as funções de pertinência iniciais eram boas ou se algumas melhorias poderiam ser feitas a elas: os algoritmos genéticos (HOLLAND, 1975).

- **Uso de algoritmos genéticos**

Os algoritmos genéticos são normalmente utilizados na busca de solução para problemas sem nenhum algoritmo conhecido (ARTERO, 2009). São baseados em uma metáfora biológica, inspirados na teoria da genética de Mendel e na Teoria da Evolução de Darwin. O aprendizado, para os algoritmos genéticos, é similar a uma competição em uma população de soluções evolutivas, na qual seus indivíduos são candidatos a resolver um problema (LUGER, 2014). Eles fazem uso de diversos conceitos dessas teorias, dentre os quais podemos destacar: a reprodução, a mutação e a avaliação da capacidade de sobrevivência de indivíduos de uma determinada população. A ideia é que as características dos indivíduos considerados melhores em uma determinada geração sejam transferidas para a geração seguinte, tal qual na Teoria da Evolução de Darwin, por meio de operações análogas à transferência de genes durante a reprodução sexual.

Para buscar soluções para um determinado problema, um conjunto de valores aleatórios iniciais deve ser gerado com o intuito de avaliar se algum desses valores corresponde à solução do problema. Esses valores gerados são chamados de cromossomos ou indivíduos. Não encontrando a solução entre os indivíduos dessa geração inicial, alguns indivíduos devem ser descartados, normalmente os que obtiverem piores resultados, e outros são combinados para gerar uma nova geração de indivíduos. Essa combinação pode ser feita por meio de cruzamentos de genes e seus indivíduos também serão avaliados.

Os principais conceitos de algoritmos genéticos são: (i) genes, que representam algum parâmetro do problema, e cuja representação pode ser feita por meio de valores inteiros, reais ou conjunto de caracteres; (ii) cromossomos, que representam cada indivíduo da população por meio de uma cadeia de genes; (iii) indivíduos, que correspondem a cada cromossomo e representam uma solução possível para o problema abordado; (iv) população, que é composta por um grupo de indivíduos que competem por sobrevivência e reprodução, a fim de manterem

suas características nas gerações; e (v) função de aptidão (*fitness*), que é usada para medir a capacidade de cada indivíduo no que diz respeito à sua sobrevivência.

A Figura 5.3 mostra os passos necessários para a utilização de algoritmos genéticos.

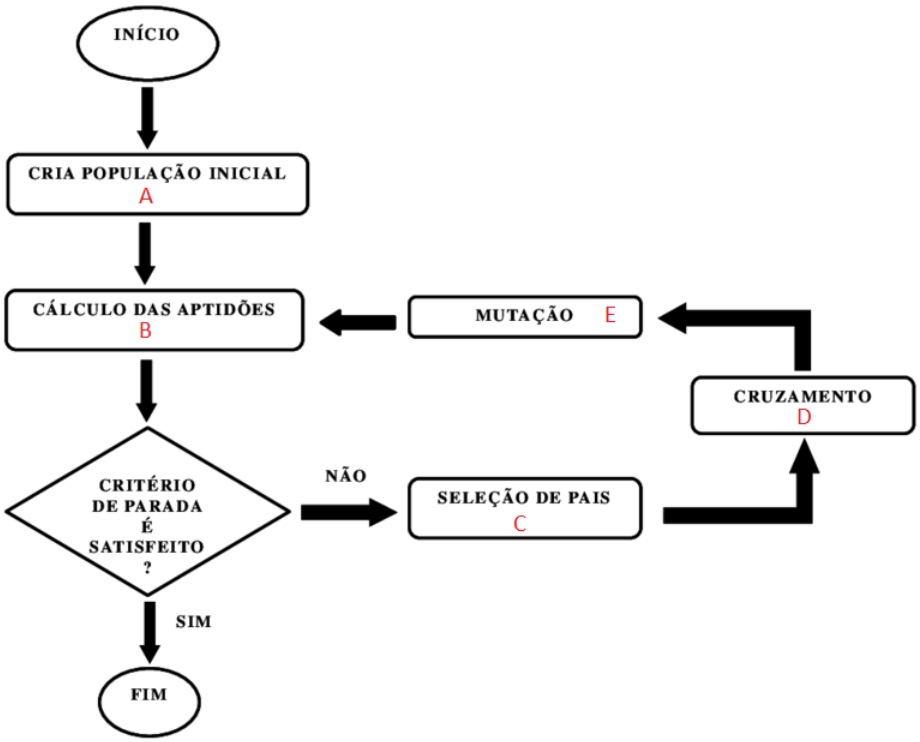


Figura 5.3. Passos para a utilização de algoritmos genéticos

O primeiro passo de um algoritmo genético é a geração da população inicial (Figura 5.3-A), a qual é composta por vários indivíduos (representados por seus respectivos cromossomos). Normalmente a população inicial é gerada aleatoriamente. O passo seguinte é avaliar se há um indivíduo nessa população que solucione o problema em questão. Tal avaliação é feita calculando-se a aptidão de cada indivíduo em resolver o problema (Figura 5.3-B). Se um indivíduo for avaliado com aptidão suficiente para resolver o problema, então a condição de parada será satisfeita e o algoritmo será finalizado, apresentando tal indivíduo como solução do problema. Caso contrário, é necessário realizar a fase de seleção (Figura 5.3-C), na qual são selecionados os indivíduos que permanecerão na próxima geração e os que darão origem a novos indivíduos por meio de cruzamentos (Figura 5.3-D) e/ou mutações (Figura 5.3-E).

Para realização da seleção, deve-se escolher alguma dentre as principais técnicas existentes: (i) seleção aleatória, na qual escolhe-se aleatoriamente alguns indivíduos para fazerem parte da nova geração; (ii) seleção por torneio, na qual a cada dois indivíduos selecionados aleatoriamente, calcula-se suas funções de aptidão e o de melhor resultado permanece para a nova geração; e (iii) seleção por roleta, na qual, para todos os indivíduos da geração, calcula-se seus valores de aptidão que correspondem proporcionalmente aos setores de uma roleta. Quanto maior a aptidão de um indivíduo, maior será a probabilidade de ele ser selecionado para a próxima geração.

Uma vez que os indivíduos sobreviventes são definidos, novos indivíduos devem ser gerados para completar a população da nova geração. Esses novos indivíduos serão definidos por meio de cruzamentos e/ou mutações. Em cruzamentos, dois indivíduos trocam material genético de forma aleatória e um novo indivíduo é composto por parte genética de um dos pais e parte genética do outro pai. Já em mutações, o cromossomo é aleatoriamente modificado em algumas de suas partes genéticas, originando assim novos indivíduos com características distintas. Normalmente utiliza-se uma taxa de mutação baixa (em torno de 1%) e uma taxa de cruzamento alta (acima dos 60%) para a criação de uma nova geração (ARTERO, 2009).

No contexto deste trabalho, um cromossomo é definido por 36 valores numéricos que representam os pontos das funções de pertinência das variáveis linguísticas *score*, coeficiente de citação e saída. Todas as funções serão representadas como trapézios, cujas representações necessitam de quatro pontos. Os primeiros doze pontos representam três funções de pertinência para os possíveis valores do *score* (baixo, médio e alto). Os doze pontos seguintes representam as três funções de pertinência para os possíveis valores do coeficiente de citação (baixo, médio e alto). Finalmente, os últimos doze pontos representam as três funções de pertinência para os possíveis valores de saída (exclusão automática, revisão manual e inclusão automática).

Com base na representação das funções por meio de trapézios, o indivíduo inicial foi representado pelo cromossomo contendo todos os pontos dos valores das funções de pertinência definidos por meio da observação de dados. Os pontos para representar as funções de pertinência do *score* são: 0, 0, 0,1 e 0,18 (*score* baixo),

0,10, 0,18, 0,40 e 0,49 (*score* médio) e 0,40, 0,49, 1 e 1 (*score* alto). Os pontos para representar as funções de pertinência do coeficiente de citação são: 0, 0, 0,09 e 0,12 (coeficiente de citação baixo), 0,09, 0, 12, 0,18 e 0,20 (coeficiente de citação médio) e 0,18, 0,20, 1 e 1 (coeficiente de citação alto). Os pontos para representar as funções de pertinência da saída esperada são: 0, 0, 0,16 e 0,30 (exclusão automática), 0,16, 0,30, 0,35 e 0,52 (revisão manual) e 0,35, 0,52, 1 e 1 (inclusão automática). A Figura 5.4 mostra um exemplo da representação de um indivíduo por meio de seu cromossomo contendo os pontos que definem os trapézios que representam as funções de pertinência das variáveis linguísticas *score*, coeficiente de citação e a saída.

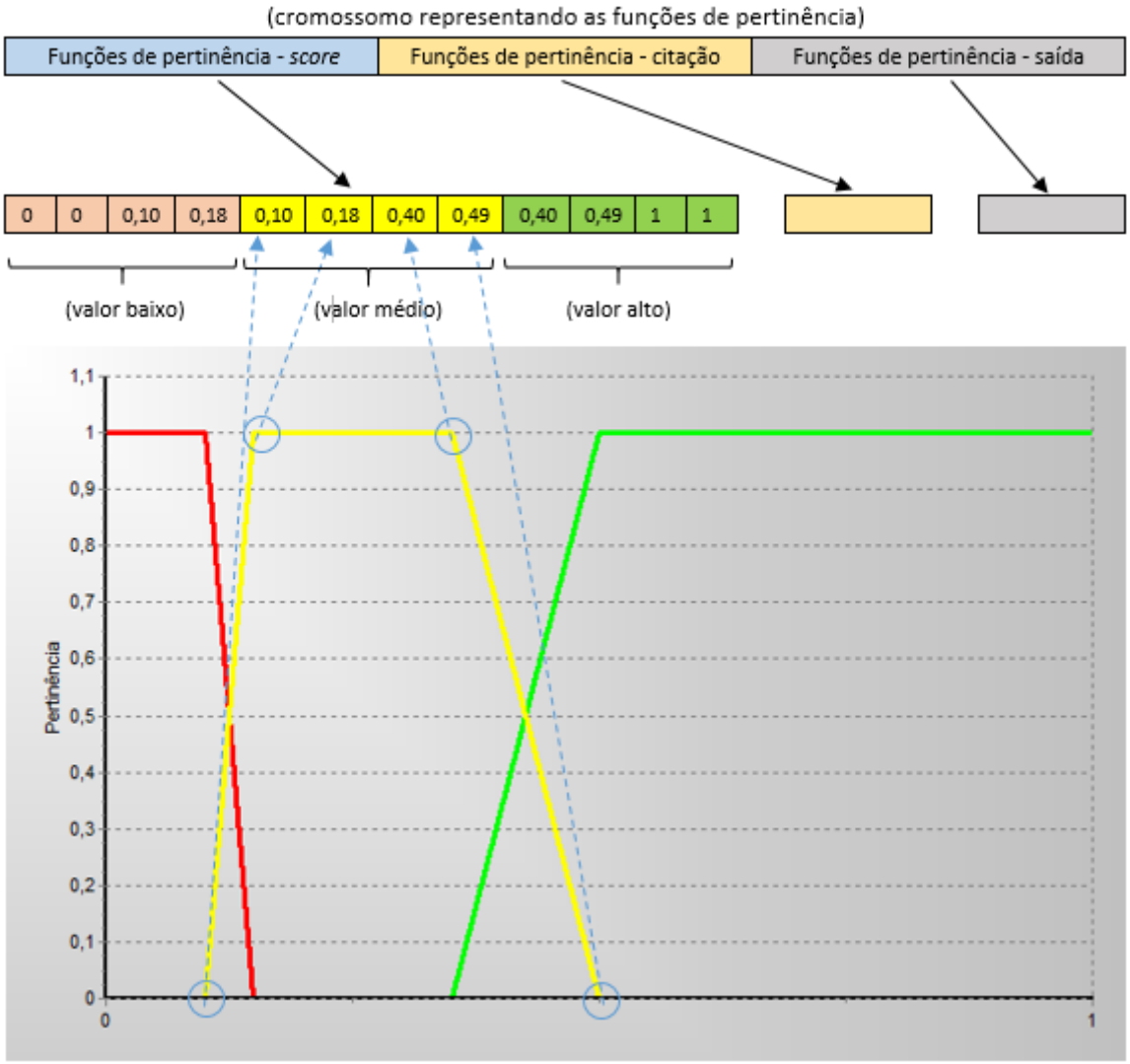


Figura 5.4. Representação de um indivíduo no contexto deste trabalho

Após a criação do cromossomo inicial, o próximo passo foi a geração da população inicial. Para termos uma população bem densa, foi definido que a população seria composta por 1000 indivíduos. Assim, os outros 999 cromossomos foram gerados aleatoriamente com base no cromossomo inicial, alterando partes dele. O passo seguinte foi a realização da fase de seleção, na qual os indivíduos que permaneceriam para a próxima geração foram selecionados. O algoritmo da seleção por roleta foi usado, selecionando 40% da população original, com base na função de aptidão (*fitness*). A função de aptidão, no contexto deste trabalho, avalia a eficácia de cada indivíduo na recomendação de decisões em comparação com as decisões dos especialistas para os estudos pertencentes às RSs escolhidas. Para cada indivíduo da população, todos os estudos foram processados considerando as funções de pertinência definidas pelo seu cromossomo, e o percentual de acerto foi calculado comparando-se as decisões sugeridas pela estratégia SCAS-Fuzzy com as decisões dos especialistas que conduziram manualmente as RSs. O percentual de acerto mostra quão perto cada indivíduo está de encontrar a "melhor solução", que seria recomendar todas as decisões de acordo com as decisões dos especialistas. Assim, 400 indivíduos (40% da população) foram selecionados para a próxima geração com base nos melhores resultados. A isso dá-se o nome de elitismo.

Depois da aplicação do elitismo, os operadores de mutação e cruzamento deveriam ser aplicados para gerar os 600 indivíduos restantes de população. No entanto, no contexto deste trabalho, apenas o operador de mutação pôde ser aplicado, uma vez que uma posição aleatória de um cromossomo é modificada com um novo valor, também gerado aleatoriamente. O cuidado tomado foi que o novo valor gerado seria o ponto de um trapézio e a ordem dos valores dos pontos deveria sempre ser respeitada, isto é, o valor do ponto 1 menor ou igual ao do ponto 2, o valor do ponto 2 menor ou igual ao do ponto 3, e o valor do ponto 3 menor ou igual ao valor do ponto 4. Já o operador de cruzamento não pôde ser aplicado porque, ao aplicá-lo, dois cromossomos são selecionados aleatoriamente para serem cruzados, gerando novos indivíduos, o que causou muitos problemas à definição lógica dos trapézios, como por exemplo invertendo-se os pontos 2 e 3 de cromossomos distintos, o que desfigura a construção de um trapézio resultando em indivíduos totalmente ilógicos para o contexto das funções de pertinência.

Uma vez determinada a nova geração composta por 40% de indivíduos oriundos da geração anterior por elitismo e 60% de novos indivíduos gerados a partir de mutações, o cálculo das aptidões foi realizado novamente em busca da solução. É um processo iterativo. Foi estabelecido como critério de parada a determinação do melhor indivíduo (cromossomo), isto é, aquele que possuir o melhor valor de aptidão (fitness), após um máximo de 100 iterações. Esse limite foi estabelecido para evitar que a aplicação dos algoritmos genéticos executasse infinitamente no caso de não encontrar nenhum indivíduo que satisfizesse o critério de parada de concordar totalmente com as decisões dos especialistas. No final das 100 iterações, o melhor indivíduo foi escolhido. A Figura 5.5 exemplifica as alterações sugeridas nas funções de pertinência para as variáveis linguísticas *score*, coeficiente de citação e saída, comparando suas definições iniciais com base na observação (vide Figura 5.5-a, Figura 5.5-c e Figura 5.5-e) com as suas definições após a aplicação de algoritmos genéticos (vide Figura 5.5-b, Figura 5.5-d e Figura 5.5-f).

Assim, as funções de pertinência das variáveis linguísticas foram definidas por meio de duas técnicas, uma com base na observação de dados e outra com base no uso de algoritmos genéticos. Ambas foram testadas em um estudo de caso apresentado no Capítulo 6.

Em resumo, os seguintes passos são necessários para aplicação da estratégia SCAS-Fuzzy em uma RS:

1. Cálculo dos *scores* dos estudos;
2. Normalização dos *scores*, estabelecendo o *score* mais alto como sendo o valor 1, enquanto que todos os demais valores são calculados dividindo-se o *score* de cada estudo pelo maior *score* encontrado na RS;
3. Cálculo dos coeficientes de citação por meio da equação apresentada na Seção 5.2.1;
4. Normalização dos coeficientes de citação estabelecendo o coeficiente mais alto como sendo o valor 1, enquanto que todos os outros valores são calculados dividindo-se o coeficiente de citação de cada estudo pelo maior coeficiente de citação encontrado na RS;
5. Utilização dos *scores* e coeficientes de citação normalizados como entradas no sistema de inferência fuzzy criado, obtendo a saída para

cada estudo, que é a recomendação de inclusão automática, exclusão automática ou revisão manual. A saída é calculada com base nas funções de pertinência e regras de inferência descritas anteriormente.

6. Aplicação das decisões de seleção automáticas sugeridas pela estratégia no que diz respeito à inclusão e exclusão de estudos, e seleção manual dos demais estudos (classificados como revisão manual), fazendo a leitura de seus títulos e *abstracts*.

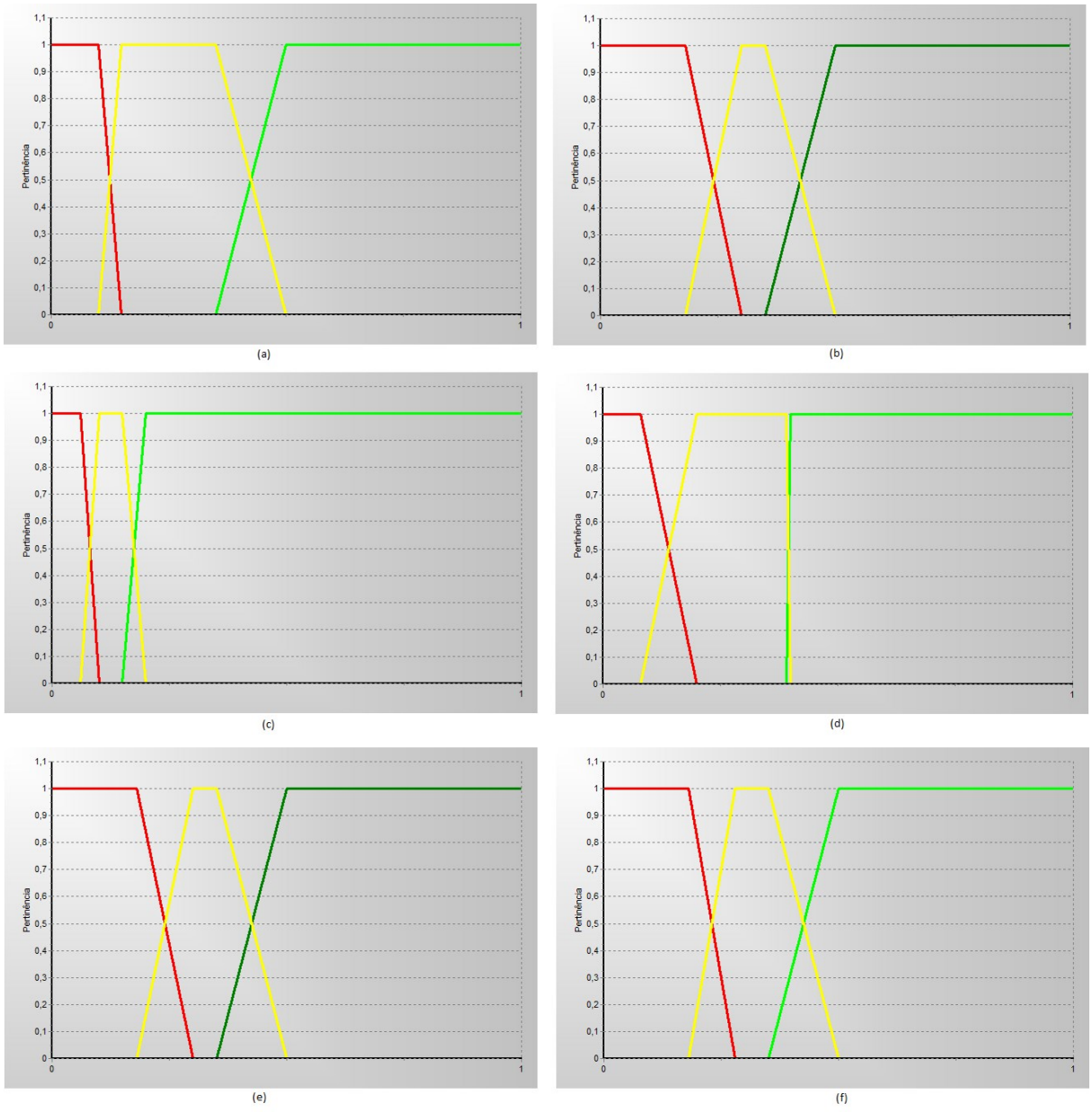


Figura 5.5. Definições das funções de pertinência antes e depois da aplicação de algoritmos genéticos

A Figura 5.6 ilustra etapas para aplicar a estratégia SCAS-Fuzzy.

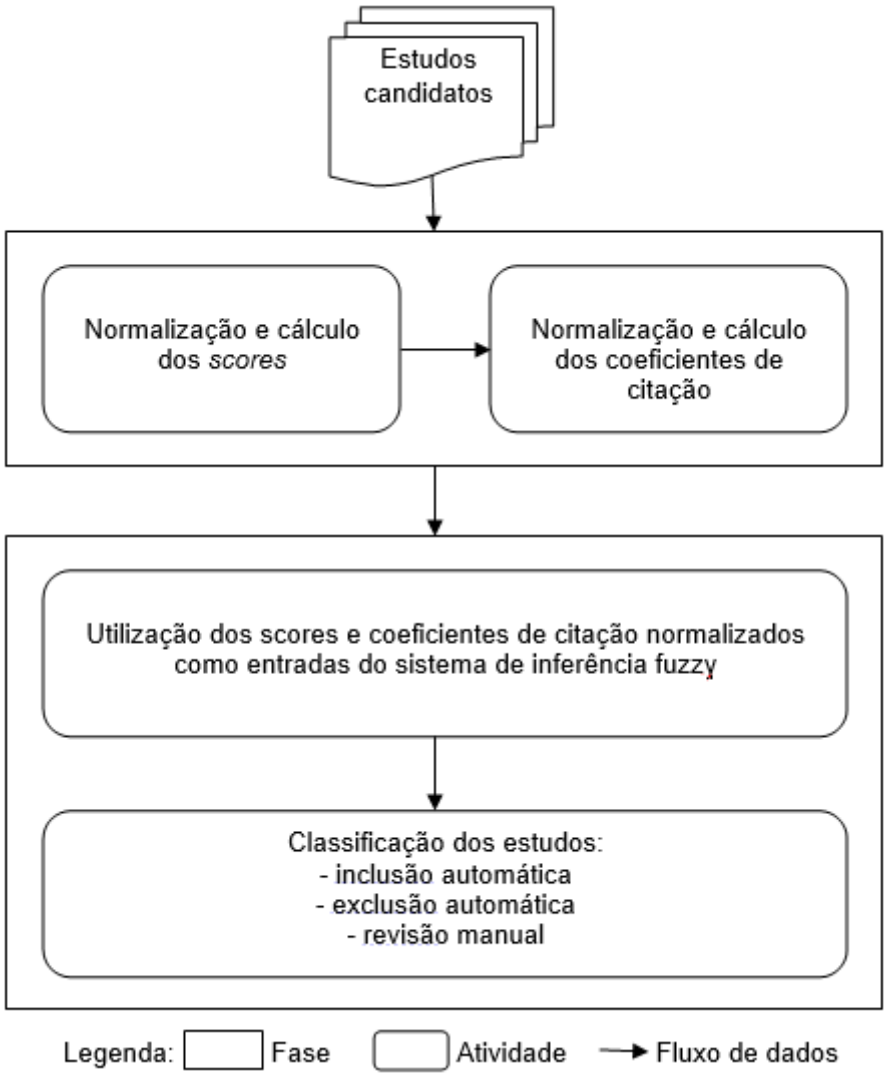


Figura 5.6. Passos para execução da estratégia SCAS-Fuzzy

A estratégia SCAS-Fuzzy foi implementada na ferramenta StArt (FABBRI et al, 2016). O suporte computacional torna mais fácil e rápido o uso da estratégia e ajudou muito para a realização do estudo de caso apresentado no próximo capítulo.

5.4 Considerações Finais

Este capítulo apresentou a estratégia semiautomática SCAS-Fuzzy definida a partir da estratégia SCAS inicial, após a detecção de pontos de melhoria e

incorporação dos mesmos à estratégia. São eles: a definição de um coeficiente de citação com base no número de citações de um estudo primário e seu ano de publicação, não mais considerando apenas se um estudo é citado ou não, e o uso de lógica fuzzy para auxiliar na classificação de um estudo primário, considerando seu *score* e coeficiente de citação. O sistema fuzzy utilizado precisa de duas variáveis linguísticas de entrada (*score* e coeficiente de citação), cujos possíveis valores nebulosos são alto, médio e baixo, para então gerar uma variável linguística de saída que recomenda se um estudo primário deve ser incluído ou excluído automaticamente, ou se deve ser revisado manualmente. Os valores dessas variáveis são determinados por funções de pertinência, que foram definidas de duas maneiras: inicialmente com base em observação e, depois, com o uso de algoritmos genéticos.

Com a definição da estratégia SCAS-Fuzzy, foi preciso avaliá-la. Isso foi feito por meio de um novo estudo de caso, a fim de verificar se os resultados seriam positivos e também fazer uma comparação com a estratégia SCAS inicialmente proposta. O estudo de caso é apresentado no capítulo a seguir, bem como uma discussão de seus resultados.

Capítulo 6

ESTUDO DE CASO PARA AVALIAÇÃO DA ESTRATÉGIA SCAS-FUZZY

Este capítulo descreve o estudo de caso para avaliar os resultados obtidos com o uso da estratégia SCAS-Fuzzy, a fim de compará-los com os resultados da estratégia SCAS original e com a abordagem manual realizada pelos pesquisadores que conduziram as revisões sistemáticas usadas nesse estudo de caso.

6.1 Considerações Iniciais

Para mostrar o uso da estratégia SCAS-Fuzzy e a sua comparação com a estratégia SCAS original e a abordagem totalmente manual, é apresentado neste capítulo um estudo de caso contendo cinco exemplos. Os exemplos incluem quatro RSs (RS1, RS2, RS3 e RS4) manualmente conduzidas e publicadas na literatura, e uma RS (RS5) ainda não publicada e que foi conduzida pelo autor deste trabalho e sua orientadora, a qual foi apresentada no Capítulo 3. Essas RS variam em tópico de pesquisa e em número de estudos primários considerados (de dezenas a centenas de estudos). Elas foram escolhidas por dois motivos principais: (i) foram conduzidas e verificadas pelos revisores com experiência na realização de RSs; e (ii) continham todas as informações necessárias para aplicar a estratégia original e a melhorada. Vale ressaltar novamente que as RSs não foram refeitas, uma vez que os revisores forneceram os dados originais. A Tabela 6.1 apresenta informações sobre as RSs mencionadas. Três RSs (RS1, RS2 e RS5) foram usadas para calibrar as funções de pertinência utilizadas no sistema de inferência fuzzy.

Tabela 6.1. Informações das RSs utilizadas no estudo de caso

RS	Título	Autores	Referência	Temática	Total de estudos	Total de incluídos
RS1	Experimenting with a multi-iteration systematic review in software engineering	F. Ferrari; J. Maldonado	ESELAW, 2008	Teste de software orientado a aspecto	97	34
RS2	Empirical studies of pair programming for CS/SE teaching in higher education: A systematic literature review	N. Salleh; E. Mendes; J. Grund	IEEE TSE 37(4), 509–522	Programação em par	264	74
RS3	A systematic literature review on the description of software architectures for systems of systems	M. Guessi; V. Neto; T. Bianchi	SAC, 2015	Arquitetura de software	215	104
RS4	Comparing local and global software effort estimation models – reflections on a systematic review	S. MacDonell; M. Shepperd	ESEM, 2007	Estimativa de custo	140	10
RS5	Estratégias de apoio à seleção de estudos em revisões sistemáticas na área da computação	F. Octaviano; S. Fabbri	-	Revisões sistemáticas da literatura	289	23

6.2 Método e Resultados

O estudo de caso foi proposto para tentar demonstrar, como objetivo principal, que a estratégia SCAS aprimorada com o uso de lógica fuzzy (independentemente de como as funções de pertinência são definidas) produz melhores resultados do que a estratégia SCAS original. Além disso, um objetivo secundário é descobrir qual das estratégias SCAS-Fuzzy (com as funções de pertinência definidas com base em observação ou usando algoritmos genéticos) fornece melhores resultados e pode ser sugerida como sendo a estratégia SCAS a ser utilizada pela comunidade. Dessa forma, para cada RS utilizada no estudo de caso, as três estratégias propostas (SCAS original, SCAS-Fuzzy com base em observação e SCAS-Fuzzy usando algoritmos genéticos) foram executadas e as recomendações para cada estudo analisadas e comparadas com as decisões tomadas pelos especialistas autores das RSs em questão.

A Tabela 6.2 apresenta dados da RS1, ou seja, estudos que são necessários para aplicar as estratégias SCAS-Fuzzy e a SCAS original, bem como as recomendações de decisões feitas por elas. Para cada estudo, podemos ver seu

ano de publicação, seu *score* calculado e seu *score* normalizado, o número de citações que obteve, seu coeficiente de citação e seu coeficiente de citação normalizado, representados, respectivamente, pelos cabeçalhos de linha "Ano", "Score", "Score Norm.", "Citação", “Coef. Citação” e “Coef. Cit. Norm.”. Além disso, podemos observar a decisão tomada pelo especialista referente ao estudo (cabeçalho de linha "Especialista"), a decisão recomendada pela estratégia SCAS original (cabeçalho de linha "SCAS"), a decisão recomendada pela estratégia SCAS-Fuzzy com funções de pertinência definidas com base em observação (cabeçalho de linha "SCAS-M1") e a decisão recomendada pela estratégia SCAS-Fuzzy com funções de pertinência definidas com base em algoritmos genéticos (cabeçalho de linha "SCAS-M2"). Os valores possíveis para a linha "Especialista" são I (incluído) ou E (excluído), e para as linhas "SCAS", "SCAS-M1" e "SCAS-M2" são I (inclusão automática), M (revisão manual) e E (exclusão automática). Os *scores* foram calculados com base na frequência das palavras-chave definidas no protocolo da RS e que foram encontradas nos títulos, *abstracts* e palavras-chave nos estudos recuperados pela mesma RS (OCTAVIANO et al, 2015). A citação representa quantas vezes um estudo foi citado pelos demais na mesma RS. O coeficiente de citação foi calculado de acordo com a equação apresentada na Seção 5.2.1 do Capítulo 5. Por exemplo, para o estudo #3, considerando o ano mais recente da RS como sendo 2007, o coeficiente de citação é calculado como Coef_Cit = (1 + 1) * (1 - 0,05 * (2007 - 2006)) = 1,9. Os *scores* e coeficientes de citação normalizados dos estudos são obtidos após feito o ranqueamento de estudos por *score* ou coeficiente de citação e dividindo seus valores pelo maior *score* ou maior coeficiente de citação, respectivamente, como mencionado anteriormente. Por exemplo, para o estudo #3, o *score* normalizado é 71 dividido por 84, que é igual a 0,85, e o coeficiente de citação normalizado é 1,9 dividido por 30,6 que é igual a 0,06. A recomendação da SCAS original é baseada no quadrante de que um estudo é classificado, sendo "I" quando classificado no quadrante 1, "E" quando classificado no quadrante 4 e "M" quando classificado nos quadrantes 2 ou 3 (OCTAVIANO et al, 2015). Uma recomendação SCAS-Fuzzy é obtida após a entrada dos *scores* e do coeficiente de citação normalizados no sistema de inferência fuzzy, observando o resultado gerado, que é a própria recomendação para o estudo.

Para compararmos as decisões tomadas pelos especialistas com a estratégia SCAS original e as novas estratégias SCAS-Fuzzy ("SCAS-M1" ou "SCAS-M2"),

devemos considerar apenas estudos classificados como "I" (inclusão automática) ou "E" (exclusão automática) porque eles seriam classificados automaticamente pelas estratégias sem a revisão manual do pesquisador. Assim, comparando as recomendações da estratégia SCAS original com as decisões dos especialistas, observamos que cinco estudos (#11, #13, #45, #54 e #58) foram incorretamente classificados pela SCAS (eles são destacados na Tabela 6.2 com fundo na cor cinza) uma vez eles seriam incluídos pela SCAS, mas na verdade eles foram excluídos pelo especialista. Esses cinco estudos são considerados falsos positivos, pois seriam incluídos na atividade de seleção inicial, mas possivelmente excluídos mais tarde durante a leitura dos textos completos. Não foram detectados falsos negativos, ou seja, nenhum estudo seria incorretamente excluído pela SCAS causando perda de evidência.

Ao compararmos as recomendações da estratégia SCAS-Fuzzy baseada em observação ("SCAS-M1") com as decisões do especialista, percebemos que três estudos (#11, #13 e #54) foram incorretamente classificados como “I” por “SCAS-M1” e excluídos pelo especialista (falsos positivos) e nenhum foi classificado como “E” por “SCAS-M1” e incluído pelo especialista (falso negativo), o que significa que SCAS-Fuzzy classificaria corretamente todos os estudos pertencentes à área de decisão de exclusão automática. No entanto, vale destacar que o número de estudos classificados automaticamente por "SCAS-M1" (33 de 97 estudos - 34,02%) é inferior ao número de estudos classificados automaticamente por SCAS (53 dos 97 estudos - 54,6%).

Ao compararmos as recomendações da estratégia SCAS-Fuzzy usando algoritmos genéticos com as decisões do especialista, percebemos que nenhum estudo foi incorretamente classificado como “I” por “SCAS-M2” e excluído pelo perito (falso positivo) nem classificado como “E” por “SCAS-M2” e incluído pelo perito (falso negativo), o que significa que “SCAS-M2” classificaria corretamente todos os estudos pertencentes às áreas de decisão automática. No entanto, o número de estudos classificados automaticamente por "SCAS-M2" (25 de 97 estudos - 25,7%) é inferior ao número de estudos classificados automaticamente pela SCAS (53 dos 97 estudos - 54,6%) e pela "SCAS- M1” (33 de 97 estudos - 34,02%).

Tabela 6.2. Comparação das decisões tomadas para os estudos da RS1

Estudo #	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
Ano	2006	2006	2006	2007	2007	2004	2006	2006	2003	2004	2006	2005	2006	2005	2006	2007	2007	2006	2005	2005	2007	2007	2007	2007	2006
Score	84	83	71	59	56	55	55	55	52	49	48	45	44	43	41	40	39	38	38	38	35	35	35	34	34
Score Norm.	1.0	0.99	0.85	0.7	0.67	0.65	0.65	0.65	0.62	0.58	0.57	0.54	0.52	0.51	0.49	0.48	0.46	0.45	0.45	0.45	0.42	0.42	0.42	0.4	0.4
Citação	1	0	1	0	0	6	10	9	18	35	2	0	2	3	0	0	0	0	7	7	1	0	0	0	1
Coef. Citação	1.9	0.95	1.9	1.0	1.0	5.95	10.4	9.5	15.2	30.6	2.85	0.9	2.85	3.6	0.95	1.0	1.0	0.95	7.2	7.2	2.0	1.0	1.0	1.0	1.9
Coef. Cit. Norm	0.06	0.03	0.06	0.03	0.03	0.19	0.34	0.31	0.5	1.0	0.09	0.03	0.09	0.12	0.03	0.03	0.03	0.03	0.24	0.24	0.07	0.03	0.03	0.03	0.06
Especialista	I	E	I	I	E	I	I	I	I	I	E	I	E	I	E	I	E	E	I	I	I	E	E	E	I
SCAS	I	M	I	M	M	I	I	I	I	I	**I**	M	**I**	I	M	M	M	M	I	I	I	M	M	M	I
SCAS-M1	M	M	M	M	M	I	I	I	I	I	**I**	M	**I**	I	M	M	M	M	I	I	I	M	M	M	M
SCAS-M2	M	M	M	M	M	I	I	I	I	I	M	M	M	I	M	M	M	M	M	M	M	M	M	M	M
Estudo #	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
Ano	2006	2007	2007	2005	2007	2004	2007	2002	2006	2006	2007	2006	2006	2007	2006	2006	2005	2005	2006	2006	2005	2007	2006	2007	2004
Score	34	32	32	32	32	31	31	31	29	29	29	29	28	28	27	27	26	26	25	25	25	24	23	22	22
Score Norm.	0.4	0.38	0.38	0.38	0.38	0.37	0.37	0.37	0.35	0.35	0.35	0.35	0.33	0.33	0.32	0.32	0.31	0.31	0.3	0.3	0.3	0.29	0.27	0.26	0.26
Citação	0	0	0	2	0	1	0	8	0	0	0	9	0	0	0	0	0	4	0	1	6	0	0	0	0
Coef. Citação	0.95	1	1	2.7	1	1.7	1	6.75	0.95	0.95	1	9.5	0.95	1	0.95	0.95	0.9	4.5	0.95	1.9	6.3	1	0.95	1	0.85
Coef. Cit. Norm	0.03	0.03	0.03	0.09	0.03	0.06	0.03	0.22	0.03	0.03	0.03	0.31	0.03	0.03	0.03	0.03	0.03	0.15	0.03	0.06	0.21	0.03	0.03	0.03	0.03
Especialista	I	E	I	I	E	I	E	I	I	E	I	I	I	E	I	E	I	I	E	E	I	E	E	E	E
SCAS	M	M	M	I	M	I	M	I	M	M	M	I	M	M	M	M	M	I	M	**I**	I	M	M	M	M
SCAS-M1	M	M	M	I	M	M	M	I	M	M	M	I	M	M	M	M	M	M	M	M	I	M	M	M	M
SCAS-M2	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M
Estudo #	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75
Ano	2006	2005	2007	2006	2004	2007	2007	2005	2003	2005	2005	2007	2006	2006	2007	2004	2000	2007	2007	2007	2005	2006	2005	2007	2007
Score	22	22	22	20	20	20	18	18	17	17	17	17	16	16	16	15	15	15	14	13	13	13	11	11	11
Score Norm.	0.26	0.26	0.26	0.24	0.24	0.24	0.21	0.21	0.2	0.2	0.2	0.2	0.19	0.19	0.19	0.18	0.18	0.18	0.17	0.15	0.15	0.15	0.13	0.13	0.13
Citação	3	0	0	6	1	0	0	3	0	6	1	0	0	0	0	0	0	0	0	0	5	0	3	0	0
Coef. Citação	3.8	0.9	1	6.65	1.7	1	1	3.6	0.8	6.3	1.8	1	0.95	0.95	1	0.85	0.65	1	1	1	5.4	0.95	3.6	1	1
Coef. Cit. Norm	0.12	0.03	0.03	0.22	0.06	0.03	0.03	0.12	0.03	0.21	0.06	0.03	0.03	0.03	0.03	0.03	0.02	0.03	0.03	0.03	0.18	0.03	0.12	0.03	0.03
Especialista	I	E	E	E	I	E	E	E	E	I	I	E	E	E	I	E	E	E	E	E	I	E	E	E	E
SCAS	I	M	M	**I**	I	M	M	**I**	M	I	I	M	M	M	M	M	M	M	E	E	M	E	M	E	E
SCAS-M1	M	M	M	**I**	M	M	M	M	M	I	M	M	M	M	M	M	M	M	M	M	I	M	M	M	M
SCAS-M2	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M
Estudo #	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97			
Ano	2007	2007	2007	2006	2007	2000	2007	2006	2007	2007	2006	2006	2002	2006	2007	2003	2006	2007	2003	2006	2007	2007			
Score	11	11	11	10	10	9	9	9	9	8	8	8	8	8	8	8	6	6	6	3	3	3			
Score Norm.	0.13	0.13	0.13	0.12	0.12	0.11	0.11	0.11	0.11	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.07	0.07	0.07	0.04	0.04	0.04			
Citação	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
Coef. Citação	1	1	1	0.95	1	0.65	1	0.95	1	1	0.95	0.95	0.75	0.95	1	0.8	0.95	1	0.8	0.95	1	1			
Coef. Cit. Norm	0.03	0.03	0.03	0.03	0.03	0.02	0.03	0.03	0.03	0.03	0.03	0.03	0.02	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03			
Especialista	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E			
SCAS	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E			
SCAS-M1	M	M	M	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E			
SCAS-M2	M	M	M	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E			

Da mesma forma que para a RS1, as estratégias foram aplicadas para as outras quatro RSs. Destaca-se que os dados da RS2, RS3, RS4 e RS5 não foram apresentados no formato de tabela como feito para a RS1 na Tabela 6.2 pelo fato de elas possuírem um número grande de estudos considerados, o que tornaria a visualização menos compreensível. Os resultados da aplicação das estratégias nas cinco RSs são apresentados na Tabela 6.3. Ela mostra a comparação dos resultados entre as recomendações feitas pelos três tipos de estratégias SCAS em avaliação (a original, a usando Fuzzy com funções de pertinência definidas com base em observação e a usando Fuzzy com funções de pertinência definidas com base em algoritmos genéticos) e as decisões tomadas pelos revisores da RS (especialistas). O objetivo é determinar qual das estratégias fornece resultados mais semelhantes às decisões tomadas pelos especialistas. Para tanto, para cada RS, calculou-se o percentual de erro para as decisões de inclusão automática e exclusão automática das três estratégias quando comparadas às decisões dos especialistas. No contexto deste estudo de caso, um erro é calculado quando uma estratégia recomenda automaticamente incluir ou excluir um estudo e a decisão tomada pelo especialista é diferente da recomendação. Além disso, foi calculado o percentual de automatização de decisões para cada RS, o que representa a redução de esforço da atividade de seleção inicial.

Com base nos dados apresentados na Tabela 6.3, em relação à RS1, a estratégia SCAS-Fuzzy usando algoritmos genéticos não teve erros, o que significa que todas as decisões automáticas de incluir ou excluir estudos foram corretas (mesmas decisões que as tomadas pelo especialista). No entanto, a porcentagem de automatização dessa estratégia foi de 25,8% (25 dos 97 estudos), que foi a menor redução de esforço entre as estratégias consideradas. A estratégia SCAS original obteve uma redução de esforço de 54,6% (53 em 97 estudos), mas teve cinco falsos positivos (percentual de erro de 5,2% para inclusão automática), o que significa que a leitura dos textos completos seria equivocadamente realizada para cinco estudos adicionais. A estratégia SCAS-Fuzzy baseada em observação obteve resultados intermediários com três falsos positivos (percentual de erro de 3,1%) e uma redução do esforço de 44% (33 dos 97 estudos). Todos não tinham falsos negativos, o que significa que não houve perda de evidências.

Tabela 6.3. Comparação das decisões dos especialistas com as recomendações das estratégias SCAS propostas

		SCAS Original		SCAS-Fuzzy (baseada em observação)		SCAS-Fuzzy (algoritmos genéticos)	
	Estudos	**Total**	**%**	**Total**	**%**	**Total**	**%**
RS1	Total de “I”	27	27,8	16	16,5	6	6,2
	Erros de “I”	5	5,2	3	3,1	0	0
	Total de “E”	26	26,8	17	17,5	19	19,6
	Erros de “E”	0	0,0	0	0,0	0	0,0
	Total de “M”	44	45,4	64	66,0	72	74,2
RS2	Total de “I”	69	26,1	26	9,8	12	4,5
	Erros de “I”	36	13,6	10	3,8	5	1,9
	Total de “E”	86	32,6	89	33,7	106	40,2
	Erros de “E”	6	2,3	2	0,8	2	0,8
	Total de “M”	109	41,3	149	56,4	146	55,3
RS 3	Total de “I”	30	14,0	186	86,5	58	27
	Erros de “I”	20	9,3	89	41,4	20	9,3
	Total de “E”	39	18,1	1	0,5	19	8,8
	Erros de “E”	2	0,9	1	0,5	1	0,5
	Total de “M”	146	67,9	28	13,0	138	64,2
RS4	Total de “I”	25	17,9	33	23,6	7	5,0
	Erros de “I”	16	11,4	23	16,4	3	2,1
	Total de “E”	62	44,3	90	64,3	99	70,7
	Erros de “E”	0	0,0	0	0,0	0	0,0
	Total de “M”	53	37,9	17	12,1	34	24,3
RS5	Total de “I”	50	17,3	22	7,6	9	3,1
	Erros de “I”	40	13,8	15	5,2	5	1,7
	Total de “E”	85	29,4	48	16,6	58	20,1
	Erros de “E”	1	0,3	0	0,0	0	0,0
	Total de “M”	154	53,3	219	75,8	222	76,8
Geral	Total de “I”	201	20,0	283	28,2	92	9,2
	Erros de “I”	117	11,6	140	13,9	33	3,3
	Total de “E”	298	29,7	245	24,4	301	30,0
	Erros de “E”	9	0,9	2	0,2	3	0,3
	Total de “M”	506	50,3	477	47,5	612	60,9

No contexto da RS2, as estratégias SCAS-Fuzzy apresentaram dois falsos negativos, mas foram melhores do que a estratégia SCAS original que apresentou

seis falsos negativos. A redução de esforço foi semelhante para ambas as estratégias SCAS-Fuzzy (44,7% usando algoritmos genéticos e 43,6% para a baseada em observação), mas foram menores do que a estratégia SCAS original, que foi de 58,7%. A estratégia SCAS-Fuzzy usando algoritmos genéticos obteve os melhores resultados ao considerarmos os falsos positivos, pois apresentou apenas cinco erros contra dez da estratégia SCAS-Fuzzy com base em observação e 36 erros da estratégia SCAS original.

Considerando a RS3, houve uma discrepância na redução do esforço, já que a estratégia SCAS-Fuzzy baseada em observação classificou automaticamente 87% dos estudos, enquanto que a SCAS-Fuzzy usando algoritmos genéticos classificou automaticamente 35,8% e a SCAS original classificou automaticamente 32,1% dos estudos. Além disso, houve outra discrepância ao considerarmos os falsos positivos, pois a estratégia mais eficiente teve 41,4% (89 estudos) classificados incorretamente como incluídos contra 9,3% (20 estudos) de cada uma das outras duas estratégias. Ambas as estratégias SCAS-Fuzzy tiveram um falso negativo e a estratégia SCAS original teve dois falsos negativos, o que implicaria na perda de mais evidências.

Na RS4, nenhum falso negativo foi encontrado para as três estratégias. A estratégia SCAS-Fuzzy usando algoritmos genéticos teve o menor número de falsos positivos (três estudos), enquanto que a SCAS original teve 16 falsos positivos e a SCAS-Fuzzy com base em observação teve 23 falsos positivos. Considerando a redução de esforço, a estratégia SCAS-Fuzzy baseada em observação foi a estratégia mais eficiente pois classificou automaticamente 88,9% dos estudos (123 de 140 estudos), enquanto que a estratégia SCAS-Fuzzy usando algoritmos genéticos classificou automaticamente 75,7% dos estudos e a estratégia SCAS original classificou automaticamente 62,1% dos estudos.

Levando em consideração a RS5, não foram encontrados falsos negativos para ambas as estratégias SCAS-Fuzzy e foi encontrado um falso negativo para a estratégia SCAS original. Foram encontrados cinco falsos positivos para a estratégia SCAS-Fuzzy usando algoritmos genéticos (percentual de erro de 1,7%), 15 para a SCAS-Fuzzy baseada em observação (percentual de erro de 5,2%) e 44 para a SCAS original (percentual de erro de 40%). A maior redução do esforço pôde ser observada para a estratégia SCAS original (46,7%), seguida pela SCAS-Fuzzy baseada em observação (24,2%) e, por fim, mas com valor próximo à anterior, pela SCAS-Fuzzy com algoritmos genéticos (23,2%).

Observando-se os resultados gerais, ao considerarmos todas as RSs, a estratégia SCAS original teve uma redução de esforço de 49,7%, classificando 29,7% dos estudos para serem excluídos automaticamente, com um percentual de erro (falsos negativos) de 0,9%, e 20% de estudos para serem automaticamente incluídos, com um erro percentual (falsos positivos) de 11,6%. A estratégia SCAS-Fuzzy baseada em observação teve uma redução de esforço de 52,5%, classificando 24,4% dos estudos para serem excluídos automaticamente, com um percentual de erro de 0,2%, e 28,2% dos estudos para serem automaticamente incluídos, com um percentual de erro de 13,9%. Finalmente, a estratégia SCAS-Fuzzy usando algoritmos genéticos teve uma redução de esforço de 39,1%, classificando 30% dos estudos para serem excluídos automaticamente, com um percentual de erro de 0,3%, e 9,2% dos estudos para serem automaticamente incluídos, com um percentual de erro de 3,3%.

Além dos cálculos de porcentuais de erro e redução de esforço, o coeficiente Kappa foi calculado para medir o nível de concordância entre as estratégias distintas e os especialistas. A mesma tabela de interpretação dos valores Kappa (Tabela 4.6) sugerido em (LANDIS; KOCH, 1977) foi utilizada nesse estudo de caso, tal qual no estudo de caso que avaliou a estratégia SCAS original. A Tabela 6.4 apresenta os valores calculados para as RSs utilizadas no estudo de caso e suas interpretações conforme a Tabela 4.6.

Analisando os resultados apresentados na Tabela 6.4, é possível identificar que a estratégia SCAS-Fuzzy usando algoritmos genéticos sempre obteve um nível de concordância maior ou ao menos igual ao das outras estratégias (vide coluna Interpretação). Ao olharmos para o valor Kappa geral, a estratégia SCAS-Fuzzy usando algoritmos genéticos também obteve os melhores resultados (nível de concordância substancial contra o nível de concordância moderado dos outros dois). Em relação às estratégias SCAS original e SCAS-Fuzzy baseada em observação, a primeira apresentou melhores resultados do que a segunda para as RS3 e RS4, a segunda apresentou melhores resultados do que a primeira para a RS2, e ambas obtiveram resultados semelhantes para as RS1, RS5 e Kappa geral.

Tabela 6.4. Valores e interpretações do Kappa para as estratégias propostas

RS	Estratégia	Kappa	Interpretação
RS1	SCAS Original	0,81	Concordância quase perfeita
	SCAS-Fuzzy (observação)	0,82	Concordância quase perfeita
	SCAS-Fuzzy (algoritmos genéticos)	1,00	Concordância perfeita
RS2	SCAS Original	0,43	Concordância moderada
	SCAS-Fuzzy (observação)	0,67	Concordância substancial
	SCAS-Fuzzy (algoritmos genéticos)	0,63	Concordância substancial
RS3	SCAS Original	0,30	Concordância distante
	SCAS-Fuzzy (observação)	0,01	Concordância pobre
	SCAS-Fuzzy (algoritmos genéticos)	0,45	Concordância moderada
RS4	SCAS Original	0,44	Concordância moderada
	SCAS-Fuzzy (observação)	0,39	Concordância distante
	SCAS-Fuzzy (algoritmos genéticos)	0,71	Concordância substancial
RS5	SCAS Original	0,22	Concordância distante
	SCAS-Fuzzy (observação)	0,39	Concordância distante
	SCAS-Fuzzy (algoritmos genéticos)	0,58	Concordância moderada
Geral	SCAS Original	0,42	Concordância moderada
	SCAS-Fuzzy (observação)	0,48	Concordância moderada
	SCAS-Fuzzy (algoritmos genéticos)	0,71	Concordância substancial

Para uma análise mais profunda, a precisão e a revocação (*precision/recall*) foram calculadas para cada estratégia aplicada às RSs do estudo de caso. Como o objetivo era avaliar as classificações automáticas sugeridas pelas estratégias, a precisão e a revocação foram calculadas considerando apenas os estudos classificados automaticamente como incluídos ou excluídos. Assim, nesse contexto, a precisão significa a porcentagem dos estudos incluídos automaticamente que são realmente relevantes (foram aceitos pelos especialistas), enquanto a revocação significa a porcentagem de estudos relevantes que seriam corretamente classificados como incluídos considerando todas as decisões automáticas tomadas. A Tabela 6.5 apresenta a precisão e a revocação calculadas para as estratégias para cada RS, incluindo também a precisão e a revocação gerais.

Tabela 6.5. Cálculo de precisão e revocação das estratégias propostas

RS	Estratégia	Precisão	Revocação
RS1	SCAS Original	81,5%	100%
	SCAS-Fuzzy (observação)	81,3%	100%
	SCAS-Fuzzy (algoritmos genéticos)	100%	100%
RS2	SCAS Original	47,8%	84,6%
	SCAS-Fuzzy (observação)	61,5%	88,9%
	SCAS-Fuzzy (algoritmos genéticos)	58,3%	77,8%
RS3	SCAS Original	33,3%	83,3%
	SCAS-Fuzzy (observação)	52,2%	100%
	SCAS-Fuzzy (algoritmos genéticos)	65,5%	97,4%
RS4	SCAS Original	36,0%	100%
	SCAS-Fuzzy (observação)	30,3%	100%
	SCAS-Fuzzy (algoritmos genéticos)	57,1%	100%
RS5	SCAS Original	20,0%	90,9%
	SCAS-Fuzzy (observação)	31,8%	100%
	SCAS-Fuzzy (algoritmos genéticos)	44,4%	100%
Geral	SCAS Original	41,8%	90,3%
	SCAS-Fuzzy (observação)	50,5%	98,6%
	SCAS-Fuzzy (algoritmos genéticos)	64,1%	95,2%

Conforme apresentado na Tabela 6.5, a estratégia SCAS-Fuzzy usando algoritmos genéticos teve a melhor precisão para todas as RSs, exceto a RS2, na qual a estratégia SCAS-Fuzzy baseada em observação teve uma precisão ligeiramente melhor. No entanto, considerando a revocação, a estratégia SCAS-Fuzzy baseada em observação obteve resultados melhores ou pelo menos semelhantes às demais estratégias para todas as RSs. É possível deduzir que, ao aplicar a estratégia SCAS original, obteríamos 90,3% dos estudos relevantes com uma precisão de 41,8%, o que significa que a maioria dos estudos classificados como incluídos seria excluída mais tarde. Ao aplicar a estratégia SCAS-Fuzzy baseada em observação, obteríamos 98,6% dos estudos relevantes com uma precisão de 50,5%, o que significa que quase metade dos estudos classificados como incluídos seriam excluídos mais tarde. Finalmente, ao aplicar a estratégia SCAS-Fuzzy usando algoritmos genéticos, obteríamos 95,2% dos estudos relevantes com uma precisão de 64,1%, o que significa que quase a terça parte dos estudos classificados como incluídos seriam excluídos mais tarde.

6.3 Ameaças à Validade

Considerando as ameaças à validade mencionadas em (WOHLIN et al., 2000), é possível destacar:

- **Validade interna e de construção:** o nível de experiência dos revisores das RSs são ameaças potenciais. Com o objetivo de minimizá-los, procurou-se selecionar quatro RSs (RS1-RS4) já publicadas na literatura, tendo todos os dados necessários para aplicar a estratégia SCAS sem ter de refazê-las, e uma RS adicional (RS5) que foi conduzida pelo autor desta tese e sua orientadora, ambos com experiência de anos no assunto. Além disso, os pesquisadores que realizaram ou supervisionaram essas RSs são pesquisadores experientes que têm algumas publicações sobre o tema de revisão sistemática da literatura;
- **Validade de conclusão:** o método escolhido para comparar os resultados após a aplicação das estratégias é uma ameaça potencial. Tentando minimizá-lo, comparou-se o percentual de erro entre as decisões sugeridas pela estratégia SCAS original e as estratégias SCAS-Fuzzy com as decisões tomadas pelos pesquisadores das RSs, avaliou-se o nível de concordância por meio do coeficiente Kappa e calculou-se a precisão e a revocação dos estudos automaticamente incluídos ou excluídos. Apenas três RSs (RS1, RS2 e RS5) foram usadas para calibrar as funções de pertinência utilizadas no sistema de inferência fuzzy;
- **Validade externa:** a generalização dos resultados está sujeita a certas limitações, principalmente porque somente foram analisadas cinco RSs. Assim, mais RSs devem ser considerados em um próximo estudo, a fim de alcançar resultados ainda mais conclusivos.

6.4 Discussão

Muitos cálculos foram apresentados na Seção 6.2 referente ao estudo de caso: percentuais de erros, redução de esforço, nível de concordância entre as recomendações das estratégias e as decisões dos especialistas, precisão e revocação. Com esses cálculos, espera-se atingir dois objetivos. O primeiro deles é demonstrar que a estratégia SCAS melhorada com recursos de lógica fuzzy (independentemente de como as funções de pertinência são definidas) fornece melhores resultados do que a estratégia SCAS original. O segundo objetivo, uma vez alcançado o primeiro objetivo, é descobrir qual estratégia SCAS-Fuzzy (com as funções de pertinência definidas com base em observação ou usando algoritmos genéticos) fornece melhores resultados. É importante ressaltar que os resultados apresentados no estudo de caso (vide Tabelas 6.3, 6.4 e 6.5) devem ser analisados em conjunto, não considerando apenas cálculos isolados.

Quando analisamos os resultados gerais, as duas estratégias SCAS-Fuzzy obtiveram melhores resultados do que a estratégia SCAS original, considerando a precisão e a revocação (vide Tabela 6.5) e o nível de concordância medido por meio do Kappa (vide Tabela 6.4). Em relação à redução do esforço, a estratégia SCAS original está em uma posição intermediária, pois apresentou melhores resultados do que a estratégia SCAS-Fuzzy usando algoritmos genéticos, mas piores resultados do que a estratégia SCAS-Fuzzy baseada em observação. Ambas as estratégias SCAS-Fuzzy apresentaram melhores resultados em relação a percentuais de erro para as recomendações de exclusão automática do que a estratégia SCAS original, com menor percentual de perda de evidência (o que é muito importante no contexto de uma RS). A estratégia SCAS original está em uma posição intermediária ao considerarmos os índices de acerto de inclusão automática, pois apresentou melhores resultados do que a estratégia SCAS-Fuzzy baseado em observação, mas piores resultados do que a estratégia SCAS-Fuzzy usando algoritmos genéticos (vide Tabela 6.3).

Também é importante analisar os resultados para cada RS isoladamente, a fim de evitar possíveis discrepâncias. Por exemplo, uma estratégia poderia obter os melhores resultados para quatro das cinco RSs, mas em uma RS em particular, por algum motivo, poderia obter resultados muito ruins, o que talvez levasse a uma

interpretação errônea dos resultados gerais. Vale destacar que a RS3 e a RS4 não foram utilizadas para calibrar as funções de pertinência utilizadas no sistema de inferência fuzzy utilizado na estratégia SCAS-Fuzzy, como ocorreu em relação à RS1, RS2 e RS5. Assim, em relação à precisão e à revocação (vide Tabela 6.5), a estratégia SCAS original apresentou resultados piores do que as estratégias SCAS-Fuzzy para as RS3 e RS5. Para a RS1, todas as estratégias tiveram a mesma revocação (100%), porém a precisão da SCAS-Fuzzy usando algoritmos genéticos foi a melhor (100%), enquanto as outras duas foram muito similares (81,5% para a SCAS original e 81,3% para a SCAS baseada em observação). Para a RS2, a estratégia SCAS com base em observação apresentou a melhor precisão e recuperação (61,5% e 88,9%, respectivamente), enquanto que a SCAS original apresentou melhor revocação (84,6%) e pior precisão (47,8%) do que SCAS-Fuzzy usando algoritmos genéticos (77,8 % e 58,3%, respectivamente). Para a RS4, todas as estratégias tiveram a mesma revocação (100%), mas a precisão da SCAS-Fuzzy usando algoritmos genéticos foi a melhor (57,1%), seguido pela SCAS original (36%) e, finalmente, pela SCAS-Fuzzy baseada em observação (30,3%).

Em relação ao nível de concordância entre recomendações das estratégias e decisões dos especialistas medidos para cada RS, a estratégia SCAS original nunca apresentou os melhores resultados. Obteve o pior valor Kappa para as RS2 e RS5, e o Kappa intermediário para a RS1, RS3 e RS4. A estratégia SCAS-Fuzzy usando algoritmos genéticos obteve os melhores resultados para todas as RSs (vide Tabela 6.4).

Em relação à redução de esforço, a estratégia SCAS original apresentou os melhores resultados para a RS1, RS2 e RS5 (54,6%, 58,7% e 46,7%, respectivamente), mas os piores resultados para a RS3 e RS4 (32,1% e 62,1%, respectivamente). A estratégia SCAS-Fuzzy baseada em observação obteve os melhores resultados para a RS3 e RS4 (87% e 88,9%, respectivamente), resultados intermediários para a RS1 e RS5 (44% e 24,2%, respectivamente) e os piores resultados para a RS2 (43,6%). A estratégia SCAS-Fuzzy usando algoritmos genéticos apresentou resultados intermediários para a RS2, RS3 e RS4 (44,7%, 35,8% e 75,7%, respectivamente) e os piores resultados para a RS1 e RS5 (25,8% e 23,2%, respectivamente). No entanto, devemos considerar os resultados da redução de esforço combinados com os percentuais de erros, pois não seria muito útil classificar automaticamente muitos estudos, mas com um alto número de erros.

Nesse sentido, a estratégia SCAS-Fuzzy usando algoritmos genéticos obteve os melhores ou os mesmos resultados (nos piores casos) do que as outras estratégias para todas as RSs, tendo apenas três falsos negativos (o mesmo número de SCAS-Fuzzy baseada em observação) contra nove da SCAS original, e os melhores ou os mesmos resultados (no pior caso - RS3) para falsos positivos. As estratégias SCAS-Fuzzy usando algoritmos genéticos e SCAS original classificaram incorretamente 13 estudos como exclusão automática (percentual de erro de 11,8%), enquanto a estratégia SCAS-Fuzzy baseada em observação não teve nenhum falso negativo, porém ressaltando que recomendou apenas um estudo a ser definido como exclusão automática.

Portanto, considerando os resultados gerais e individuais das RSs, analisando-os em um conjunto, é possível concluir que a estratégia SCAS original, na maioria dos casos, apresenta resultados piores ou similares (em poucos casos) do que as estratégias SCAS-Fuzzy. Isso faz com que o primeiro objetivo – que é tentar demonstrar que a estratégia SCAS-Fuzzy obtém melhores resultados do que a estratégia SCAS original – seja alcançado.

Para avaliar o segundo objetivo – descobrir qual dentre as estratégias SCAS-Fuzzy consegue melhores resultados – as duas estratégias SCAS que utilizam lógica fuzzy foram comparadas. Ao considerarmos os resultados globais, a estratégia SCAS-Fuzzy usando algoritmos genéticos obteve um nível de concordância (Kappa) maior, uma melhor precisão de decisões automáticas e um menor número de falsos positivos, enquanto que a estratégia SCAS-Fuzzy baseada em observação obteve uma maior redução de esforço, um menor número de falsos negativos e uma melhor revocação. No entanto, ao considerarmos as RSs de forma isolada, a estratégia SCAS-Fuzzy usando algoritmos genéticos apresentou melhores níveis de concordância e taxas de precisão do que a SCAS-Fuzzy baseada em observação para quase todas as RSs (exceto RS2), o mesmo percentual de erro de exclusões automáticas para todas as RSs e o menor percentual de erro de inclusão automática para todas as RSs. Além disso, embora a redução do esforço fosse menor para a estratégia SCAS-Fuzzy baseada em observação, a estratégia SCAS-Fuzzy usando algoritmos genéticos apresentou o maior índice de estudos classificados como exclusão automática e com um percentual de erro muito baixo para todas as RSs, enquanto a estratégia SCAS-Fuzzy baseada em observação obteve os índices mais altos de estudos classificados como inclusão automática,

mas com maior percentual de erro, o que resultará em esforço adicional em uma próxima atividade de leitura de texto completo (ou leitura adaptativa). Assim, considerando tudo isso, é possível concluir que a estratégia SCAS-Fuzzy usando algoritmos genéticos apresenta, em geral, melhores resultados do que a estratégia SCAS-Fuzzy baseada em observação, atingindo assim o segundo objetivo proposto.

6.5 Considerações Finais

Este capítulo apresentou o estudo de caso realizado para avaliação da estratégia semiautomática SCAS-Fuzzy para seleção de estudos primários. O estudo de caso considerou dados de 5 RSs e avaliou, primeiramente, a eficácia e eficiência da estratégia em relação à revisão manual feita pelos especialistas que conduziram as RSs consideradas. Além disso, permitiu a comparação de resultados com a estratégia SCAS inicialmente definida e a avaliação de qual técnica seria mais indicada para a definição de funções de pertinência das variáveis linguísticas utilizadas no sistema de inferência fuzzy criado para aplicação da estratégia, se a com base em observação ou a com base no uso de algoritmos genéticos.

Os resultados do estudo de caso e a discussão feita neste capítulo sobre os mesmos foram muito importantes para as conclusões apresentadas no próximo capítulo.

Capítulo 7

CONCLUSÃO

Este capítulo apresenta as conclusões, contribuições e limitações deste trabalho, bem como lições aprendidas, oportunidades de pesquisa identificadas e as publicações obtidas até o momento.

7.1 Conclusões

Normalmente, uma RS é uma tarefa trabalhosa que envolve um grande conjunto de estudos primários que precisam ser analisados. Os pesquisadores geralmente passam muito tempo realizando a atividade de seleção inicial, lendo títulos e *abstracts* de centenas ou até mesmo milhares de estudos primários. Nesse contexto, este trabalho apresentou **estratégias semiautomáticas – SCAS e SCAS-Fuzzy – para seleção de estudos primários em estudos secundários** e os estudos experimentais realizados para corroborar, preliminarmente, a tese de que **considerando-se o processo de RS, é possível melhorar a eficiência da seleção inicial dos estudos, sem perda da qualidade, aplicando-se uma estratégia semiautomática baseada no *score,* associado às palavras-chave, e no coeficiente de citação, associado ao número de citações e ano de publicação, de um estudo**, sendo que a estratégia SCAS-Fuzzy apresenta melhores resultados que a estratégia SCAS.

Inicialmente, foi proposta uma estratégia semiautomática para seleção de estudos no processo de RS chamada SCAS, totalmente baseada em duas funcionalidades: *score* e número de citações dos estudos. A estratégia propunha a

classificações dos estudos de uma RS em quatro quadrantes, sendo que os estudos com *scores* altos e ao menos uma citação pertenceriam ao quadrante 1 e deveriam ser incluídos automaticamente na atividade de seleção inicial, e os estudos com *scores* baixos e sem citação pertenceriam ao quadrante 4 e deveriam ser excluídos automaticamente na seleção inicial. Os estudos com *scores* altos e sem citação (quadrante 2) e os estudos com *scores* baixos e ao menos uma citação (quadrante 3) deveriam ser analisados manualmente. A estratégia foi avaliada primeiramente por meio de um estudo de caso e, em seguida, por meio de um experimento, e ambos apresentaram resultados promissores. No entanto, alguns pontos de melhorias foram identificados e, assim, uma nova estratégia aprimorada foi proposta para apoiar a seleção inicial semiautomática dos estudos primários.

As principais melhorias realizadas foram a definição de um coeficiente de citação, que se baseia no número de citações e no ano de publicação dos estudos, e o uso de lógica fuzzy para classificar os *scores* e coeficientes de citação dos estudos como sendo altos, médios ou baixos, ao invés de se utilizar valores de corte abruptos, como acontecia na estratégia inicial. Além disso, com o uso de lógica fuzzy, duas maneiras de se definir as funções de pertinência das variáveis linguísticas *score* e coeficiente de citação foram propostas: com base em observação e com o uso de algoritmos genéticos. Com tais melhorias, uma nova estratégia – chamada SCAS-Fuzzy – foi definida e avaliada por meio de um estudo de caso considerando-se dados de cinco RSs. O objetivo era determinar se a estratégia SCAS-Fuzzy proporcionaria melhores resultados do que a SCAS original e, se isso fosse verdadeiro, avaliar qual dos métodos de definição das funções de pertinência produziria melhores resultados: o baseado em observação ou o baseado no uso de algoritmos genéticos.

Os resultados gerais do estudo de caso mostraram que as duas estratégias SCAS-Fuzzy (com funções de pertinência definidas com base no uso de algoritmos genéticos e com base em observação) obtiveram melhores resultados do que a estratégia SCAS original para precisão e revocação, nível de concordância medido por meio do Kappa e percentual de erro das recomendações de exclusão automática (o que implica em menor porcentagem de perda de evidência). A estratégia SCAS original está em uma posição intermediária para redução de esforço (perdendo para a SCAS-Fuzzy baseada em observação) e percentual de erro das recomendações de inclusão automática (perdendo para SCAS-Fuzzy usando algoritmos genéticos).

Comparando apenas as duas estratégias SCAS-Fuzzy, a SCAS-Fuzzy usando algoritmos genéticos obteve um nível de concordância mais alto, uma melhor precisão para decisões automáticas e um menor número de falsos positivos, enquanto que a SCAS-Fuzzy baseada em observação obteve uma redução de esforço maior, um número menor de falsos negativos e uma melhor revocação.

Analisando os resultados de cada RS de forma isolada, a estratégia SCAS original nunca obteve os melhores resultados para precisão e revocação nem para nível de concordância. Em relação à redução do esforço, a SCAS original obteve os melhores resultados para três das cinco RSs, e a SCAS-Fuzzy baseada em observação obteve os melhores resultados para as outras duas RSs. Ao olharmos para a redução de esforço combinada com os percentuais de erro, a estratégia SCAS-Fuzzy usando algoritmos genéticos obteve os melhores resultados para todas as RSs. Além disso, é possível notar que SCAS-Fuzzy usando algoritmos genéticos obteve um índice elevado de estudos classificados como exclusão automática e com percentual de erro muito baixo para todas as RSs, enquanto que a SCAS-Fuzzy baseada em observação apresentou um alto índice de estudos classificados como inclusão automática, mas com maior porcentual de erro.

Assim, o estudo de caso mostrou que, considerando a redução média de esforço de 39,1% ao se aplicar a estratégia SCAS-Fuzzy, os percentuais de erro de 0,3% para erros de exclusão e de 3,3% para erros de inclusão quando comparada à revisão manual, além de um nível de concordância substancial para com os revisores, a estratégia SCAS-Fuzzy proporcionou resultados satisfatórios para redução de esforço da atividade de seleção inicial e com baixíssima quantidade de perde de evidências, mantendo a qualidade do estudo secundário.

Além disso, foi possível concluir que a estratégia SCAS-Fuzzy proporcionou melhores resultados do que a estratégia SCAS original e que, ainda, a estratégia SCAS-Fuzzy com funções de pertinência das variáveis linguísticas do sistema fuzzy definidas com o uso de algoritmos genéticos fornece, em geral, melhores resultados do que a estratégia SCAS-Fuzzy baseada em observação. Logo, a estratégia SCAS-Fuzzy deve ser adotada para a seleção semiautomática de estudos primários em estudos secundários, e as funções de pertinência utilizadas no sistema de inferência fuzzy, o qual é parte da estratégia, devem ser definidas com base no uso da técnica de inteligência computacional de algoritmos genéticos.

Vale ressaltar que a estratégia independe do meio de identificação de estudos escolhidos para que possa ser aplicada. Seja a identificação feita por meio de *strings* de busca ou *snowballing*, seja utilizando uma combinação de ambas as técnicas, ou até mesmo qualquer outra técnica, a estratégia SCAS-Fuzzy pode ser aplicada. A única exigência é que as informações dos estudos (títulos, *abstracts*, palavras-chave e referências) sejam disponibilizadas por meio de arquivos formatados, como por exemplo BibTex, para que a ferramenta StArt consiga entendê-los e calcular seus *scores* e coeficientes de citação. Nesse sentido, o uso de *strings* de busca facilita o processo para utilização da estratégia, pois as principais bases exportam as informações dos estudos identificados.

Procurou-se, também, investigar o porquê da ocorrência dos falsos negativos nas RSs avaliadas, mas não se chegou à conclusão sobre a razão principal de alguns estudos com *scores* baixos, ainda que poucos, serem aceitos pelos revisores. Acredita-se que a subjetividade inerente à atividade de seleção seja a razão que leva o revisor a aceitar um estudo, o que impede o tratamento de ocorrências de falsos negativos de forma automática, ao menos no atual estágio da pesquisa.

7.2 Contribuições da Tese

As contribuições da tese são detalhadas na sequência:

- Conduzir uma RS sobre estratégias de seleção de estudos primários em estudos secundários, descrevendo seus funcionamentos e resultados, ferramentas que as suportam e os métodos pelas quais foram avaliadas.
- Definição da estratégia semiautomática SCAS para seleção inicial de estudos primários no contexto de revisões sistemáticas da literatura. A estratégia permite a inclusão e exclusão automáticas de parte dos estudos primários recuperados, reduzindo o esforço e sem perda significativa de qualidade.
- Avaliação da estratégia inicial proposta por meio de um estudo de caso com RSs publicadas na literatura e de um experimento com alunos de pós-graduação. Ambos mostraram resultados satisfatórios e foram de

grande importância para publicação da estratégia em periódico e evento relevantes da área.

- Refinamento da estratégia inicial e o uso de inteligência computacional para resultar em uma estratégia semiautomática para seleção de estudos primários melhorada chamada SCAS-Fuzzy, a qual faz uso de um sistema de inferência fuzzy para classificação dos estudos, cujas funções de pertinência foram definidas com base em observação e, posteriormente, no uso de algoritmos genéticos.
- Apresentação de um estudo de caso para comprovar que a estratégia SCAS-Fuzzy produz resultados ainda melhores que a estratégia SCAS original e que as funções de pertinência das variáveis linguísticas utilizadas resultam em classificações mais adequadas quando definidas por meio de algoritmos genéticos.
- Implementação, inicialmente, da estratégia SCAS, e, posteriormente, da estratégia SCAS-Fuzzy na ferramenta StArt, o que tornou possível a execução dos estudos experimentais apresentados nesta pesquisa, bem como permite a utilização da estratégia SCAS-Fuzzy pela comunidade.

7.3 Limitações do Trabalho

- Os estudos experimentais apresentados nos capítulos anteriores foram conduzidos ou pelo próprio autor deste trabalho ou sob sua supervisão. A estratégia aqui proposta não foi utilizada por um pesquisador externo, que não tivesse envolvimento com sua definição. Dessa forma, ainda não se consegue caracterizar a real dificuldade em aplicar a estratégia SCAS-Fuzzy.
- Os resultados obtidos nos estudos experimentais não podem ser generalizados para outros contextos. Dessa forma, não se pode afirmar que os valores obtidos nesses estudos se repitam em outros casos. Certamente, para caracterizar valores de referência em relação à

redução de esforço e percentual de erro associado à aplicação da estratégia, muitas outras RSs teriam que fazer uso da mesma.

- As regras de inferência do sistema fuzzy e as funções de pertinência foram calibradas com dados de apenas três RSs. Assim, para que calibragem dessas funções seja mais acurada, é necessária a inclusão de mais dados de RSs no sistema fuzzy.
- A implementação da estratégia SCAS-Fuzzy na ferramenta StArt, que possibilitou a realização dos estudos experimentais ainda necessita de melhorias para se tornar mais estável, de forma a ser utilizada pela comunidade.
- Como o intuito da RS conduzida no contexto deste trabalho era apenas o de identificar as estratégias existentes e os métodos de avaliação utilizados para elas, não foram feitas meta-análises das informações obtidas por meio da RS.

7.4 Publicações

Nesta seção são apresentadas as publicações do autor desta tese durante o período de doutorado. As publicações estão organizadas em três subseções: artigos completos publicados em periódicos, artigos completos publicados em anais de congressos (*full papers*) e capítulos de livros.

7.4.1 Publicações em Periódicos

1) FABBRI, S. C. P. F.; FELIZARDO, K. R.; FERRARI, F. C.; HERNANDES E. C. M.; OCTAVIANO F. R.; NAKAGAWA, E. Y.; MALDONADO, J. C. Externalising tacit knowledge of the systematic review process. **IET Software**, v. 7, n. 6, p. 298-307, dezembro 2013. Doi:10.1049/iet-sen.2013.0029.

2) OCTAVIANO, F. R.; FELIZARDO, K. R.; MALDONADO, J. C.; FABBRI, S.C.P.F. Semi-automatic selection of primary studies in systematic

literature reviews: is it reasonable? **J. Empirical Software Engineering**, v. 20, n. 6, p. 1898–1917, dezembro 2015. Doi:10.1007/s10664-014-9342-8.

7.4.2 Publicações em Anais de Congresso (*Full Papers*)

3) OCTAVIANO, F.; SILVA, C.; FABBRI, S. Using the SCAS strategy to perform the initial selection of studies in systematic reviews: an experimental study. In: 20º INTERNATIONAL CONFERENCE ON EVALUATION AND ASSESSMENT IN SOFTWARE ENGINEERING – EASE, Limerick, IR, 2016. **Proceedings…** New York: ACM 2016. Paper 25. Doi:10.1145/2915970.2916000

4) FABBRI, S.; OCTAVIANO, F.; SILVA, C.; HERNANDES, E.; DI THOMMAZO, A.; BELGAMO, A. Improvements in the StArt tool to better support the systematic review process. In: 20th INTERNATIONAL CONFERENCE ON EVALUATION AND ASSESSMENT IN SOFTWARE ENGINEERING – EASE, Limerick, IR, 2016. **Proceedings…** New York: ACM 2016. Paper 21. Doi:10.1145/2915970.2916013.

7.4.3 Publicações de Capítulos de Livros

5) FABBRI, S.; OCTAVIANO, F.; HERNANDES, E. Protocolo da revisão sistemática. In: FELIZARDO, K.; NAKAGAWA, E.; FABBRI, S.; FERRARI, F. **Revisão sistemática da literatura em engenharia de software – teoria e prática**. 1ª ed. Elsevier, 2017, 144 p.

7.5 Oportunidades Futuras

Apresentam-se algumas ideias que deverão ser desenvolvidas no grupo de pesquisa como continuidade desta tese, a saber:

- Continuar avaliando a estratégia SCAS-Fuzzy por meio de novos estudos experimentais visando, entre outros pontos, a avaliar o esforço

necessário para o processamento de falsos positivos e investigar profundamente as principais causas de falsos negativos.

- Obter *feedback* sobre a estratégia SCAS-Fuzzy dos usuários da StArt uma vez que eles utilizem a estratégia em suas RSs. Isso é possível por meio da StArt *Online Community*, disponível na página oficial da ferramenta StArt: http://lapes.dc.ufscar.br/tools/start_tool.
- Possível recalibragem das funções de pertinência das variáveis linguísticas *score* e coeficiente de citação à medida que dados de novas RSs sejam coletados e colocados como entrada no sistema de inferência fuzzy criado para esta pesquisa.
- Investigar a inclusão da técnica *snowballing* para complementar a busca por estudos relevantes. Uma ideia seria avaliar a aplicação de *snowballing backward* sugerida por Silva (2017) nos estudos automaticamente aceitos pela estratégia SCAS-Fuzzy, uma vez que a mesma já está implementada na ferramenta StArt.

REFERÊNCIAS

ABILIO, R. et al. Systematic literature review supported by information retrieval techniques: A case study. In: 40º LATIN AMERICAN COMPUTING CONFERENCE, Montevideo, UR, 2014. **Proceedings**… Montevideo: IEEE Computer Society 2014. p. 1-11.

ALI, N. B.; PETERSEN, K. Evaluating strategies for study selection in systematic literature studies. In: 18º INTERNATIONAL SYMPOSIUM ON EMPIRICAL SOFTWARE ENGINEERING AND MEASUREMENT – ESEM, Torino, IT, 2014. **Proceedings…** Nova Iorque: ACM 2014. Doi:10.1145/2652524.2652557

ARTERO, A. O. **Inteligência Artificial: Teoria e Prática**. Livraria Física, 2009.

BAILEY, J. et al. Searchengine overlaps: do they agree or disagree?. In: *2º* INTERNATIONAL WORKSHOP ON REALISING EVIDENCE-BASED SOFTWARE ENGINEERING - REBSE, Minneapolis, USA, 2007. **Proceedings…** IEEE 2007. p. 1–6. Doi: 0.1109/ICSECOMPANION.2007.6

BASILI, V. R.; CALDIERA, G.; ROMBACH, H. D. **Goal Question Metric Approach**. Encyclopedia of Software Engineering, London, UK, 1994.

BASILI, V. et al. The empirical investigation of Perspective-Based Reading. **J. Empirical Software Engineering**, v. 6, n. 2, p. 133-164, 1996.

BOELL, S.; CEZEC-KECMANOVIC, D. Are systematic reviews better, less biased and of higher quality?. In: 19ª EUROPEAN CONFERENCE ON INFORMATION SYSTEMS – ECIS, Helsinki, FI, 2011. **Proceedings…** Helsinki, FI 2011: Paper 223.

BRERETON, P. et al. Lessons from applying the systematic literature review process within the software engineering domain. **Journal of Systems and Software**, v. 80, p. 571–583, abril 2007.

CARVER, J. C. et al. Identifying barriers to the systematic literature review process. In: 7º International Symposium on Empirical Software Engineering and Measurement – ESEM, Baltimore, US, 2013. **Proceedings…** Washington: IEEE 2013. p. 203–212. Doi:10.1109/ESEM.2013.28

CARLETTA, J. Assessing agreement on classification tasks: The kappa statistic. **Computational Linguistics**, v. 22, n. 2, p. 249-254, 1996.

DIESTE, O.; GRIMÁN, A.; JURISTO, N. Developing search strategies for detecting relevant experiments. **J. Empirical Software Engineering**, v. 14, n. 5, p. 513-539, outubro 2009. doi:10.1007/s10664-008-9091-7

DISTE, O.; PADUA, A. G. Developing search strategies for detecting relevant experiments for systematic reviews. In: 1º INTERNATIONAL SYMPOSIUM ON EMPIRICAL SOFTWARE ENGINEERING AND MEASUREMENT – ESEM, Madri, SP, 2007. **Proceedings…** IEEE Computer Society 2007. p. 215-224.

DYBÅ, T.; DINGSØYR, T. Empirical studies of agile software development: A systematic review. **Information and Software Technology**, v. 50, n. 9, p. 833-859, agosto 2008.

DYBÅ, T.; DINGSØYR, T.; HANSSEN, G. K. Applying systematic reviews to diverse study types: An experience report. In: 1º INTERNATIONAL SYMPOSIUM ON EMPIRICAL SOFTWARE ENGINEERING AND MEASUREMENT – ESEM, Madri, SP. **Proceedings…** IEEE Computer Society 2007. p. 225-234. Doi:10.1109/ESEM.2007.59

FABBRI, S. et al. Externalising tacit knowledge of the systematic review process. **IET Software**, v. 7, n. 6, p. 298-307, dezembro 2013. doi:10.1049/iet-sen.2013.0029

FABBRI, S. et al. Improvements in the StArt tool to better support the systematic review process. In: 20th INTERNATIONAL CONFERENCE ON EVALUATION AND ASSESSMENT IN SOFTWARE ENGINEERING – EASE, Limerick, IR, 2016. **Proceedings…** New York: ACM 2016. Paper 21. Doi:10.1145/2915970.2916013

FABBRI, S. et al. Using information visualization and text mining to facilitate the conduction of systematic literature reviews. In: 14ª INTERNATIONAL CONFERENCE ON ENTERPRISE INFORMATION – ICEIS, Wroclaw, PL, 2012. **Proceedings…** Springer 2012. p. 243-256.

FABBRI, S.; OCTAVIANO, F.; HERNANDES, E. Protocolo da revisão sistemática. In: FELIZARDO, K.; NAKAGAWA, E.; FABBRI, S.; FERRARI, F. **Revisão sistemática da literatura em engenharia de software – teoria e prática**. 1ª ed. Elsevier, 2017, 144 p.

FELDMAN, R.; SANGER, J. **The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data**. Cambridge: Cambridge University Press, 2007.
FELIZARDO, K. R. et al. A visual analysis approach to validate the selection review of primary studies in systematic reviews. **J. Information and Software Technology**, v. 54, n. 10, p. 1079-1091, outubro 2012.

FELIZARDO, K. R. et al. An approach based on Visual Text Mining to Support Categorization and Classification in the Systematic Mapping. In: 14º INTERNATIONAL CONFERENCE ON EVALUATION AND ASSESSMENT IN SOFTWARE ENGINEERING – EASE, UK, 2010. **Proceedings…** Swindon: BCS Learning & Development Ltd 2010. p. 34-43.

FELIZARDO, K. R. et al. A visual analysis approach to update systematic reviews. In: 18º International Conference on Evaluation and Assessment in Software Engineering, London, UK, 2014. **Proceedings...** Nova Iorque: ACM 2014. Paper 4.

FELIZARDO, K. R. et al. Using Visual Text Mining to Support the Study Selection Activity in Systematic Literature Reviews. In: 5º INTERNATIONAL SYMPOSIUM ON EMPIRICAL SOFTWARE ENGINEERING AND MEASUREMENT – ESEM, Banff, CN, 2011. **Proceedings…** IEEE Computer Society 2011. p. 77-86. Doi:10.1109/ESEM.2011.16

FELIZARDO, K. R.; SOUZA, S. R.; MALDONADO, J. C. The use of visual text mining to support the study selection activity in systematic literature reviews: A replication study. In: 3º INTERNATIONAL WORKSHOP ON REPLICATION IN EMPIRICAL SOFTWARE ENGINEERING RESEARCH - RESER, Baltimore, USA, 2013. **Proceedings…** New York: ACM 2013. p. 91-100.

HASSLER, E. et al. Identification of slr tool needs – results of a community workshop. **Information Software Technology**, v. 70, p. 122–129, fevereiro 2016. Doi:10.1016/j.infsof.2015.10.011

HOLLAND, J. H. **Adaptation in natural and artificial systems**. Ann Arbor, MI: The University of Michigan Press, 1975.

JEDLITSCHKA, A.; CIOLKOWSKI, M. Towards evidence in software engineering. In: INTERNATIONAL SYMPOSIUM ON EMPIRICAL SOFTWARE ENGINEERING – ISESE, 2004. **Proceedings…** Washigton: IEEE Computer Society 2004. p. 261-270.

JøRGENSEN, M.; SHEPPERD, M. A systematic review of software development cost estimation studies. **IEEE Transactions on Software Engineering**, v. 33, n. 1, p. 33-53, janeiro 2007. doi:10.1109/TSE.2007.256943

KITCHENHAM, B. A. **Procedures for Performing Systematic Reviews**. Keele University, Keele, UK, 2004.

KITCHENHAM, B. A.; DYBÅ, T.; JøRGENSEN, M. Evidence-based software engineering. In: 26ª INTERNATIONAL CONFERENCE ON SOFTWARE ENGINEERING – ICSE, Edinburgh, UK, 2004. **Proceedings…** Washington, US: IEEE Computer Society 2004. p. 273-281.

KITCHENHAM, B.; BRERETON, P. A systematic review of systematic review process research in software. **Information and Software Technology**, v. 55, p. 2049-2075, dezembro 2013.

KITCHENHAM, B.; CHARTERS, S. **Guidelines for performing Systematic Literature Reviews in Software Engineering**. Keele University, Keele, UK, 2007.

KITCHENHAM, B. et al. Systematic literature reviews in software engineering – A systematic literature review. **Information and Software Technology**, v. 51, n. 1, p. 7-15, janeiro 2009.

KLIR, G. J.; YUAN, B. **Fuzzy Sets and Fuzzy Logic: Theory and Applications**. 1ª ed. Prentice Hall, 1995.
LANDIS, J.; KOCH, G. The measurement of observer agreement for categorical data. **J. Biometrics**, v. 33, n. 1, p. 159-174, março 1977.

KUHRMANN, M.; FERNÁNDEZ, D. M.; DANEVA, M. On the pragmatic design of literature studies in software engineering: an experience-based guideline **J. Empirical Software Engineering**, v. 22, n. 6, p. 2852–2891, dezembro 2017. Doi: org.ez31.periodicos.capes.gov.br/10.1007/s10664-016-9492-y

LUGER, G. F. **Inteligência Artificial**. 6ª ed, São Paulo: Pearson Education do Brasil, 614 p., 2013.

MAFRA, S. N., & TRAVASSOS, G. H. **Estudos Primários e Secundários Apoiando a Busca por Evidência em Engenharia de Software**. Programa de Engenharia de Sistemas e Computação (PESC), COPPE/UFRJ, Rio de Janeiro, 2006.

MARSHALL, C.; BRERETON, P. Tools to Support Systematic Literature Reviews in Software Engineering: A Mapping Study. In: 7º INTERNATIONAL SYMPOSIUM ON EMPIRICAL SOFTWARE ENGINEERING AND MEASUREMENT - ESEM.

Baltimore, USA, 2013. **Proceedings…** IEEE 2013. p. 296-299. Doi:10.1109/ESEM.2013.32

MARSHALL, C.; BRERETON, P.; KITCHENHAM, B. Tools to Support Systematic Reviews in Software Engineering: A Feature Analysis. In: 18º INTERNATIONAL CONFERENCE ON EVALUATION AND ASSESSMENT IN SOFTWARE ENGINEERING - EASE. Torino, IT, 2014. **Proceedings…** Nova Iorque: ACM 2014. p. 139-148. Doi 10.1145/2601248.2601270

MARTINS, C. A. **Uma abordagem para pré-processamento de dados textuais em algoritmos de aprendizado**. 2003. Tese (Tese em Ciência da Computação) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo – USP, São Carlos, 2003.

MOURÃO, E. et al. Investigating the use of a hybrid search strategy for systematic reviews. In: 11º INTERNATIONAL SYMPOSIUM ON EMPIRICAL SOFTWARE ENGINEERING AND MEASUREMENT – ESEM, Toronto, CA, 2017. **Proceeding…** 2017.

OCTAVIANO, F. R. et al. Semi-automatic selection of primary studies in systematic literature reviews: is it reasonable? **J. Empirical Software Engineering**, v. 20, n. 6, p. 1898–1917, dezembro 2015. doi:10.1007/s10664-014-9342-8

OCTAVIANO, F.; SILVA, C.; FABBRI, S. Using the SCAS strategy to perform the initial selection of studies in systematic reviews: an experimental study. In: 20º INTERNATIONAL CONFERENCE ON EVALUATION AND ASSESSMENT IN SOFTWARE ENGINEERING – EASE, Limerick, IR, 2016. **Proceedings…** New York: ACM 2016. Paper 25. Doi:10.1145/2915970.2916000

PETERSEN, K. et al. Systematic mapping studies in software engineering. 12º INTERNATIONAL CONFERENCE ON EVALUATION AND ASSESSMENT IN SOFTWARE ENGINEERING – EASE, Bari, IT, 2008. **Proceedings…** Electronic Workshops in Computing (eWiC) 2008. p. 68-77.

PETTICREW, M.; ROBERTS, H. How to appraise the studies: An introduction to assessing study quality. In: _____. **Systematic reviews in the social sciences: A practical guide**. Malden: Blackwell Publishing, 2005. p. 125–163.

REZENDE, S. O.; MARCACINI, R. M.; MOURA, M. F. O uso da mineração de textos para extração e organização não supervisionada de conhecimento. **Revista de Sistemas de Informação da FSMA**, v. 7, p. 7-21, 2011.

RIAZ, M. et al. Experiences conducting systematic reviews from Novices' perspective. In: 14th INTERNATIONAL CONFERENCE ON EVALUATION AND ASSESSMENT IN SOFTWARE ENGINEERING – EASE, Keele, UK, 2010. **Proceedings…** British Computer Society 2010. p. 1-10.

SILVA, C. **Critérios para priorização de estudos primários identificados por *snowballing* com conjunto inicial gerado por *string* de busca**. 2017. Dissertação (Dissertação em Ciência da Computação) – Universidade Federal de São Carlos – UFSCar, São Carlos, 2017.

TOMASSETTI, F. et al. Linked data approach for selection process automation in systematic reviews. In: 15º ANNUAL CONFERENCE ON EVALUATION ASSESSMENT IN SOFTWARE ENGINEERING – EASE, Durham, UK, 2011. **Proceedings…** IET 2011. p. 31-35. Doi:10.1049/ic.2011.0004

TRAVASSOS, G. H.; BARROS, M. O. Contributions of in virtuo and in silico experiments for the future of empirical studies in software engineering. In: 2nd WORKSHOP ON EMPIRICAL SOFTWARE ENGINEERING THE FUTURE OF EMPIRICAL STUDIES IN SOFTWARE ENGINEERING, 2003. **Proceedings…** 2003. p. 117-130.

TRAVASSOS, G. H.; GUROV, D.; AMARAL, E. A. **Introdução à Engenharia de Software Experimental**. COPPE / Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, 2002.

WALLACE, B. et al. Active learning for biomedical citation screening. In: 3rd INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING – WKDD, Phuket, TH, 2010. **Proceedings…** Washington: ACM 2010. p. 173-181.

WALLACE, B. et al. Semi-automated screening of biomedical citations for systematic reviews. **BMC Bioinformatics**, janeiro 2010. doi:10.1186/1471-2105-11-55

WOHLIN, C. et al. **Experimentation in Software Engineering**. London: Springer, 2000.

WOHLIN, C. Guidelines for snowballing in systematic literature studies and a replication in software engineering In: 18th INTERNATIONAL CONFERENCE ON EVALUATION AND ASSESSMENT IN SOFTWARE ENGINEERING – EASE, London, UK, 2014. **Proceedings…** ACM 2014. p. 1-10.

WOHLIN C. Second-generation systematic literature studies using snowballing. In: 20th INTERNATIONAL CONFERENCE ON EVALUATION AND ASSESSMENT IN

SOFTWARE ENGINEERING – EASE, Limerick, IR, 2016. **Proceedings…** New York: ACM 2016. p. 1-6.

ZADEH, L. A. Fuzzy sets. **Information and Control**, v.8, n. 3, p. 338-353, 1965.

ZHANG, H.; BABAR, M. A. On Searching Relevant Studies in Software Engineering. In: 14th INTERNATIONAL CONFERENCE ON EVALUATION AND ASSESSMENT IN SOFTWARE ENGINEERING – EASE, Keele, UK, 2010. **Proceedings…** British Computer Society 2010. p. 111-120.

ZHANG, H.; BABAR, M. A.; TELL, P. Identifying relevant studies in software engineering. **Information and Software Technology**, v. 53, p. 625-637, junho 2011. doi:10.1016/j.infsof.2010.12.010

Apêndice A

A FERRAMENTA START

A ferramenta StArt – State of the Art through Systematic Review (FABBRI et al, 2016) – foi desenvolvida para fornecer suporte computacional para o maior número possível de atividades de uma revisão sistemática, desde o preenchimento do protocolo na fase de planejamento, passando pelas atividades de seleção inicial e extração de dados na fase de execução, até a fase de sumarização dos dados.

1. Fase de Planejamento:

Em relação ao planejamento, o protocolo trazido pela ferramenta possui os mesmos campos propostos por Kitchenham (2004) para auxiliar os pesquisadores na condução da revisão sistemática e também na repetitividade do processo. Permite o registro de informações tais como objetivo, questões de pesquisa, métodos utilizados para seleção, palavras-chave utilizadas nas *strings* de busca, critérios de inclusão e exclusão, bases de dados utilizadas, formulário de extração, descrição de critérios de qualidade e de como a sumarização será feita. A Figura A.1 mostra alguns campos do protocolo existente na ferramenta.

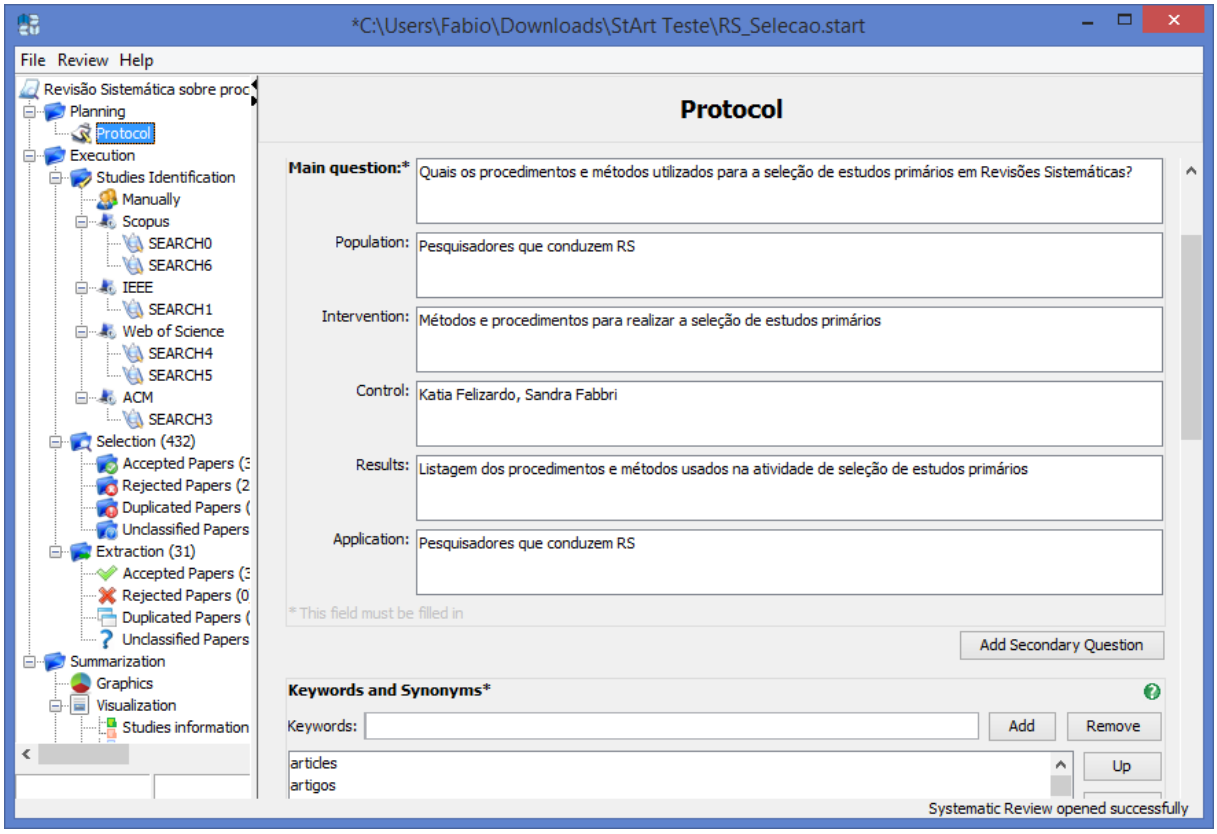


Figura A.1. Exemplo de protocolo na ferramenta StArt

2. Fase de Execução:

Em relação à execução, a ferramenta permite que os estudos primários recuperados sejam carregados por meio de arquivos exportados de várias bases de dados no formato BibTex, RIS, MedLine ou Cochrane. Muitos estudos duplicados são detectados automaticamente pela ferramenta por meio de técnicas de mineração de texto.

Além disso, a StArt traz um importante recurso para auxiliar os pesquisadores na seleção inicial de estudos primários: o *score*. Esse recurso considera a frequência com que as *keywords* definidas no protocolo aparecem no título, *abstract* e palavras-chave dos estudos primários recuperados, fazendo um ranqueamento de estudos. Quanto maior o *score* de um estudo, maior é a sua chance de ser relevante ao tema de pesquisa (OCTAVIANO et al, 2015). A Figura A.2 mostra um exemplo de tela da StArt no qual os estudos são classificados por *score* (destacado em vermelho).

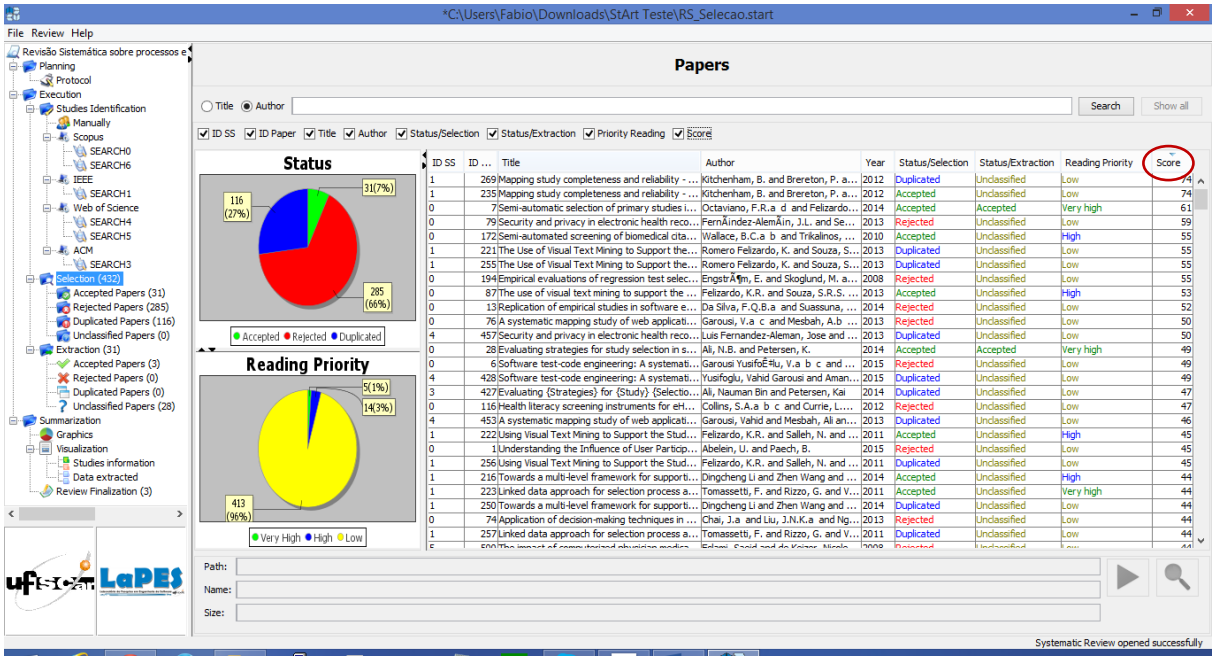


Figura A.2. Estudos classificados por *score* na atividade de seleção inicial

Outros recursos adicionais que foram implementados são o número de citações entre os estudos primários (*cross-citation*) e o coeficiente de citação que são calculados pela ferramenta. Esses recursos permitem que a estratégia SCAS-Fuzzy seja disponibilizada para uso na StArt, uma vez que utiliza a combinação dos *scores* e dos coeficientes de citação dos estudos.

Outro recurso existente que pode auxiliar o pesquisador na seleção é o percentual de similaridade de estudos, em que, por meio de mineração de texto, a ferramenta apresenta estudos similares em conteúdo em relação a um estudo selecionado pelo pesquisador. Assim, ao detectar um estudo relevante, é possível navegar por outros estudos com conteúdo similares, que supostamente tendem a ser relevantes também.

Na etapa de extração, a ferramenta apresenta os campos do formulário de extração definido no protocolo para que as mesmas informações sejam extraídas de todos os estudos.

3. Fase de Sumarização dos Dados:

Em relação à sumarização, a StArt apresenta funcionalidades para facilitar a sumarização dos dados, tais como o uso de recursos de visualização e a geração de relatórios no formato Excel de acordo com as necessidades do pesquisador. É

possível gerar visualizações nos formatos de grafo, árvore e grafo-radial, e por diversos campos, tais como ano de publicação, autores, locais de eventos, campos do formulário de extração, entre outros. A Figura A.3 mostra um exemplo de visualização no formato grafo-radial por ano de publicação apenas dos estudos incluídos na seleção inicial de uma RS.

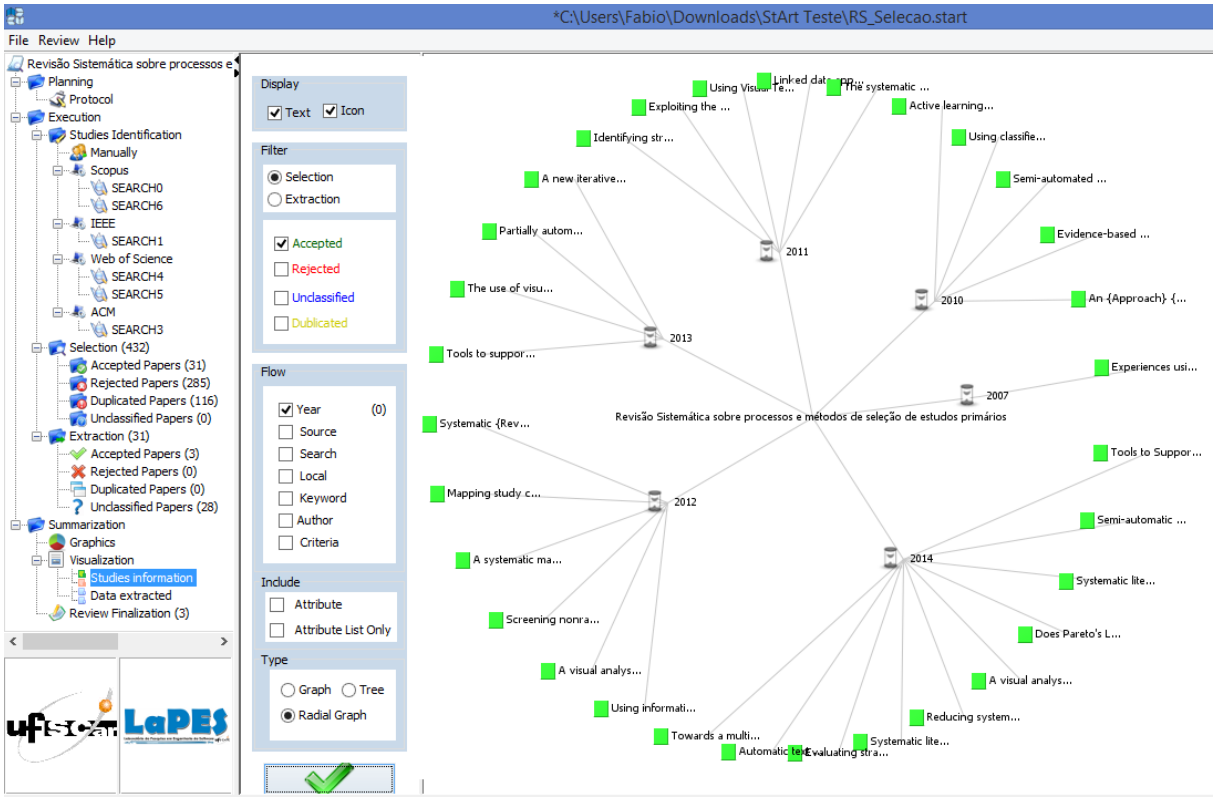


Figura A.3. Exemplo de visualização por ano de publicação na ferramenta StArt

Apêndice B

PROTOCOLO E *STRINGS* DE BUSCA DA REVISÃO SISTEMÁTICA

Este apêndice mostra as *strings* de busca e o protocolo utilizados na RS sobre estratégias de seleção apresentada no Capítulo 3, com o efeito de permitir que ela seja repetível por outros pesquisadores, caso necessário.

As *strings* de busca foram adequadas às particularidades de cada base de dados, e os filtros adicionais utilizados nas bases também são exibidos.

• Scopus:

TITLE-ABS-KEY((selection OR screening) AND (studies OR papers OR articles OR evidence) AND ("systematic literature review" OR "systematic review" OR "systematic map") AND (LIMIT-TO(SUBJAREA,"COMP"))

• IEEE Xplore Digital Library:

(selection OR screening) AND (studies OR papers OR articles OR evidence) AND ("systematic literature review" OR "systematic review" OR "systematic map")

Filtro adicional aplicado na base de dados: computer science

• Web of Science:

((selection OR screening) AND (studies OR papers OR articles OR evidence) AND ("systematic literature review" OR "systematic review" OR "systematic map"))

Filtros adicionais aplicados na base de dados: (computer science interdisciplinar applications or computer science artificial intelligence or computer science information systems)

• ACM Digital Library:

((Title:selection or Title:screening) AND (Title:studies or Title:papers or Title:articles or Title:evidence) AND (Title:"systematic literature review" or Title:"systematic review" or Title:"systematic map")) OR ((Abstract:selection or Abstract:screening) AND (Abstract:studies or Abstract:papers or Abstract:articles or Abstract:evidence) AND (Abstract:"systematic literature review" or Abstract:"systematic review" or Abstract:" systematic map")) for: (((Title:selection or Title:screening) AND (Title:studies or Title:papers or Title:articles or Title:evidence) AND (Title:"systematic literature review" or Title:"systematic review" or Title: "systematic map")) OR ((Abstract:selection or Abstract:screening) AND (Abstract:studies or Abstract:papers or Abstract:articles or Abstract:evidence) AND (Abstract:"systematic literature review" or Abstract:"systematic review" or Abstract:" systematic map")))

A Tabela B.1 apresenta o protocolo utilizado na RS sobre estratégias de seleção na área de Computação.

Tabela B.1. Protocolo da RS apresentada no Capítulo 3

Informações Gerais	
Título	Estratégias de seleção de estudos primários na área de computação
Pesquisadores	Fábio Octaviano; Sandra Fabbri
Descrição	Uma revisão sistemática para identificar e entender as estratégias para realizar a atividade de seleção inicial de estudos em RSs na área de computação.
Objetivos	Descrever as estratégias (automatizadas ou não) propostas na literatura para realizar a atividade de seleção inicial de estudos em RSs na área de computação, identificando seus pontos fortes e fracos, além de um levantamento das ferramentas que suportam as estratégias e recursos computacionais utilizados nas estratégias.
Data da Última Atualização	Outubro/2017
Questões de Pesquisa Principais	
Questão de Pesquisa 1	Como funcionam as estratégias utilizadas para realizar a atividade de seleção inicial de estudos em RSs na área de computação?
População	Estudos que apresentam estratégias para realizar a atividade de seleção inicial de estudos em RSs.
Intervenção	Estratégias para realizar a atividade de seleção inicial de estudos em RSs.
Comparação	N/A
Resultados	Descrição das estratégias encontradas e os métodos de avaliação utilizados.
Contexto	Condução de RSs.
Questão de Pesquisa 2	Como funcionam as estratégias (semi) automáticas utilizadas para realizar a atividade de seleção inicial de estudos em RSs na área de computação?
População	Estudos que apresentam estratégias (semi) automáticas para realizar a atividade de seleção inicial de estudos em RSs.
Intervenção	Estratégias (semi) automáticas para realizar a atividade de seleção inicial de estudos em RSs.
Comparação	N/A
Resultados	Descrição das estratégias (semi) automáticas encontradas e os métodos de avaliação utilizados.
Contexto	Condução de RSs.
Método para recuperação de estudos	
Palavras-chave	articles; evidence; papers; screening; selection; studies; systematic literature review; systematic map; systematic review;

Critérios de seleção das Fontes de busca	- Indexar estudos da área de computação; - Exportar as principais informações (metadados) de estudos como arquivos no formato BibTex; - Exportar os textos completos dos estudos como arquivos no formato PDF.
Linguagens dos estudos	Inglês
Bases de dados	Scopus; IEEE; Web of Science; ACM.
Métodos de busca nas bases de dados	Primeiramente, construir as *strings* de busca combinando as palavras-chave identificadas. Em seguida, executar as strings de busca nas bases de dados selecionadas. Depois, exportar os metadados dos estudos identificados como arquivos no formato BibTex para carrega-los na ferramenta StArt.
Critérios de Inclusão e Exclusão dos estudos	Critérios de Inclusão: • CI1: O estudo relata ao menos uma estratégia para a seleção inicial de estudos. Critérios de Exclusão: • CE1: O estudo não relata uma estratégia para a seleção inicial de estudos. • CE2: O estudo não está escrito em inglês. • CE3: Não é possível encontrar o texto completo do estudo.
Tipos de Estudos	Todos os tipos de estudos.
Seleção Inicial de Estudos	Pesquisadores lerão o título, *abstract* e palavras-chave dos estudos recuperados e os classificarão como incluídos ou excluídos, aplicando os critérios de inclusão e exclusão definidos. Um estudo deve ser incluído se ele satisfizer ao menos um critério de inclusão. Um estudo deve ser excluído se ele satisfizer ao menos um critério de exclusão. No caso de um estudo satisfizer tanto critérios de inclusão quanto de exclusão, o mesmo deve ser excluído.
Informações de extração de dados e de qualidade	
Avaliação da qualidade dos estudos	Critérios de Qualidade: • CQ1: A(s) estratégia(s) propostas foi(ram) apresentada(s) de maneira clara? • CQ2: A(s) estratégia(s) foi(ram) avaliada(s)? • CQ3: A(s) estratégia(s) foi(ram) implementada(s) em alguma ferramenta de suporte e está(ão) disponível(is) para utilização de pesquisadores? As respostas possíveis para cada questão são: Sim (1 ponto), Não (0 pontos), ou Parcialmente (0,5 pontos). Com base nas respostas, um *score* de qualidade deve ser calculado para cada estudo e os que possuírem o score de qualidade menor do que 0,5 pontos devem ser excluídos.
Campos do Formulário de extração de dados	• Afiliação dos autores. • O estudo relata uma estratégia para (semi) automatização da seleção inicial de estudos? (Respostas permitidas: sim ou não). • Descrição da(s) estratégia(s) para (semi) automatização da seleção inicial de estudos (se aplicável) e recursos computacionais utilizados. • O estudo relata uma ferramenta de suporte à seleção inicial de estudos? (Respostas permitidas: sim ou não). • Qual é o suporte provido pela ferramenta para a seleção inicial de estudos? (se aplicável). • O estudo relata uma estratégia não automatizada para a seleção inicial de estudos? (Respostas permitidas: sim ou não). • Descrição da(s) estratégia(s) não automatizada(s) para a seleção inicial de estudos (se aplicável). • O estudo apresenta uma avaliação da(s) estratégia(s) proposta(s)? (Respostas permitidas: sim ou não). • Como a(s) estratégia(s) foi(ram) avaliada(s)? (Respostas permitidas: Estudo de Caso, Experimento, Outros, N/A) • Descrição dos resultados obtidos após a avaliação da(s) estratégia(s).
Sumarização e Publicação	
Sumarização de resultados	Todas as estratégias identificadas devem ser descritas, incluindo os métodos de avaliação utilizados para avaliá-las. Além disso, seus pontos fortes e fracos devem ser discutidos. As ferramentas que suportam as estratégias e os recursos computacionais utilizados nas estratégias devem ser descritos.

Estratégia de publicação	Preparar um artigo científico e submetê-lo a uma revista ou conferência importante, com o intuito de disponibilizar os resultados para a comunidade.