Social E-SECO

Integrando uma Rede Social em um Ecossistema de Software Científico

Jonathan Souza Muniz

JUIZ DE FORA

JULHO, 2017

Social E-SECO
Integrando uma Rede Social em um Ecossistema de Software Científico

Jonathan Souza Muniz

Universidade Federal de Juiz de Fora
Instituto de Ciências Exatas
Departamento de Ciência da Computação
Bacharelado em Ciência da Computação

Orientador: Regina Maria Maciel Braga
Coorientador: José Maria Nazar David

JUIZ DE FORA
JULHO, 2017

SOCIAL E-SECO

Integrando uma Rede Social em um Ecossistema de Software Científico

Jonathan Souza Muniz

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIAS EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Aprovada por:

Regina Maria Maciel Braga
Doutora em Engenharia de Sistemas e Computação

José Maria Nazar David
Doutor em Engenharia de Sistemas e Computação

Fernanda Cláudia Alves Campos
Doutora em Engenharia de Sistemas e Computação

Victor Ströele de Andrade Menezes
Doutor em Engenharia de Sistemas e Computação

JUIZ DE FORA

03 DE JULHO, 2017

Aos meus pais, por sempre acreditarem

À Bárbara, por todo seu amor e paciência

Resumo

A presente pesquisa baseia-se em construir uma rede social de pesquisadores e integra-la ao ecossistema de software científico E-SECO. O objetivo é encontrar relacionamentos entre pesquisadores e promover a colaboração direta entre os grupos de pesquisas a fim evoluir e complementar o resultado dos experimentos desenvolvidos com o suporte da plataforma. Os dados serão extraídos através de *APIs* públicas de plataformas conhecidas e com os mesmos, uma rede social de coautorias será construída, com o objetivo de proporcionar novas conexões para os pesquisadores no contexto do E-SECO.

Palavras-chave: Rede Social; Ecossistema de Software; Colaboração; E-SECOs.

Abstract

The present work aims to build a researchers social network and integrate it to the E-SECO scientific software ecosystem. The purpose is to find relations between researchers and promote direct collaboration between reaseach groups in order to evolve and complement the results of experiments developed with platform support. The data will be consumed from well known public APIs and used to build a coauthorship social network, aiming to promote new connexions between researchers in the E-SECO context.

Keywords: Social Network; Software Ecosystem; Collaboration; E-SECO.

Sumário

Lista de Figuras

Lista de Abreviações

API	Application Programming Interface
ECOS	Ecossistemas de Software
ECOSS	Ecossistemas de Software Científico
HTTP	Hypertext Transfer Protocol
LPS	Linha de Produtos de Software
MVC	Model View Controller
REST	Representational State Transfer
XML	Extensible Markup Language
JSON	Javascript Object Notation
NO-SQL	Not Only SQL
SQL	Structured Query Language

Agradecimentos

Aos meus pais, Jânice e José Braz, por todo o incentivo, por nunca terem desistido e sempre serem meu porto seguro. Sem vocês, esse trabalho não seria possível.

À Minha namorada Bárbara Fernandes por toda sua compreensão e paciência, além do auxílio nas revisões.

Aos meus amigos, em especial Anrafel Pereira e Magnus Ribeiro, que me ajudaram sempre desde o início e me mostraram qual prazeroso pode ser o trabalho em equipe.

Aos meus orientadores, Regina Braga e José Maria pela dedicação, conversas, conselhos, críticas e por estarem sempre de braços abertos, mesmo nos momentos nos quais eu hesitei. Vocês foram essenciais para minha formação acadêmica e estarão para sempre em minha memória.

Aos professores do Departamento de Ciência da Computação, em especial a professora Fernanda Campos, por ter me aberto as portas no início da minha graduação e estar em minha banca fechando esse ciclo e ao professor Victor Stroele, por seu trabalho ter sido minha principal referência e também estar presente em minha banca.

1 Introdução

1.1 Contextualização e Justificativa

A ciência moderna está cada vez mais buscando suporte da computação para conduzir experimentos, e o uso da mesma para apoiar trabalhos científicos é chamado *e-Science* (HINE, 2006). O desenvolvimento de software para o meio científico, em geral, é realizado de maneira informal. Normalmente, a maioria dos desenvolvedores não possui conhecimento na área de pesquisa. O cientista acaba assumindo a posição de desenvolvedor e é muito envolvido na elaboração do produto, porém, aprende sobre o desenvolvimento de maneira informal, o que pode tornar esse processo um tanto quanto caótico (SEGAL AND MORRIS, 2008).

Workflows científicos (HINE, 2006) são bastante usados nesse contexto, no entanto a elaboração dos mesmos requer conhecimento específico, interdisciplinar e habilidades de desenvolvimento dos cientistas, o que resulta em uma tarefa não trivial. O resultado são grandes barreiras e dificuldades no desenvolvimento e reuso de *workflows* de outros cientistas, o que geralmente leva a retrabalho. Torna-se então necessário a construção de ferramentas que apoiem os cientistas na tarefa de condução do experimento científico e que minimizem a necessidade do conhecimento específico e interdisciplinar, principalmente o do desenvolvimento de software, facilitando a construção de *workflows* e também o reuso dos mesmos por outros cientistas. O conceito de Linha de Produto de Software (LPS) foi utilizado para preencher essa necessidade, na tentativa de auxiliar os cientistas na elaboração do *workflow* (COSTA, 2013). Porém o processo de experimentação vai além desse passo.

(PEREIRA, 2014) evolui essa abordagem enriquecendo os artefatos presentes no núcleo de uma LPS através de elementos de colaboração (informações de percepção, contexto e mecanismo de comunicação). Com isso, é oferecida mais semântica para o domínio e facilita o trabalho dos cientistas gerando oportunidades de interação entre os mesmos, o que é uma das características principais de experimentos científicos complexos. Além

disso, aspectos como a grande quantidade de informação e a necessidade de suporte de mecanismos computacionais distribuídos requerem relacionamento intenso entre recursos e serviços e também entre os pesquisadores. Esses são elementos que constituem um Ecossistema de Software Científico.

Um Ecossistema de Software (SECO) pode ser considerado como um conjunto de atores trabalhando como uma unidade em um mercado comum de software e serviços, juntamente com o relacionamento entre eles (JANSEN et al., 2009). No contexto de *e-Science*, os atores são cientistas trabalhando em conjunto, interagindo em um domínio, utilizando software e serviços científicos que são relacionados através de *workflows*. Porém, como já mencionado para a construção de *workflows* científicos, os atores necessitam do suporte de uma plataforma capaz de preencher os requisitos de um Ecossistema de Software além de uma infraestrutura que promova o relacionamento entre os mesmos.

Nesse contexto surge o E-Science Software ECO system (E-SECO), um ecossistema baseado na web projetado para apoiar as atividades dos pesquisadores durante o ciclo de vida comum de um experimento científico. Essa é uma solução inovadora, aplicando conceitos de Ecossistema de Software no domínio do *e-Science*, visando apoiar o desenvolvimento colaborativo de experimentos científicos. E-SECO pode ser definido pelos seus relacionamentos com provedores de software científicos, institutos de pesquisas, pesquisadores, agência de desenvolvimento e quaisquer interessados nos resultados de uma pesquisa (SOUZA, 2015).

A utilização de plataformas de colaboração web estimulam colaboração de forma indireta entre os cientistas, uma vez que toda a comunicação deve passar antes por um servidor centralizado, o que pode trazer desvantagens em sua adoção (PHAM, 2006). O presente trabalho pretende promover a colaboração direta entre os grupos de pesquisa que utilizam o E-SECO através da análise e integração de uma rede social de pesquisadores.

Redes sociais são estruturas dinâmicas formadas por indivíduos ou organizações que são normalmente representados por nós conectados uns aos outros por tipos de relacionamentos entre os mesmos. Apesar de serem estruturas complexas, a sua análise permite detectar os tipos de conexão entre as pessoas dentro e fora de organizações.

No contexto de *e-Science*, os cientistas necessitam de compartilhar suas pesquisas

e se conectar com novos parceiros a fim de evoluir e complementar seus resultados. Por isso, a colaboração é essencial para realizar os experimentos (STROELE, 2013). A análise proposta das conexões entre os pesquisadores, como por exemplo através de coautoria e citações, pode proporcionar descobertas de novas oportunidades para os cientistas para colaboração em pesquisas, comunicação e reuso.

1.2 Objetivos da Pesquisa

O objetivo geral da presente pesquisa baseia-se em analisar uma rede social de pesquisadores e integra-la a plataforma E-SECO a fim de encontrar relacionamentos entre os pesquisadores e promover a colaboração direta entre os grupos de pesquisa e por fim evoluir e complementar o resultado dos experimentos desenvolvidos com o suporte da plataforma. Para tal, pretende-se extrair e analisar dados de pesquisadores disponíveis através de *APIs* públicas em plataformas acadêmicas como DBPL (DBLP, 1993) e Google Scholar (GOOGLE SCHOLAR, 2004). A partir da extração dos dados, uma rede social com relacionamentos de coautoria será gerada e uma representação visual disponibilizada a fim de proporcionar uma melhor análise do pesquisador com relação ao alcance de seus relacionamentos, encontrando novas conexões relevantes para os utilizadores no contexto do E-SECO. O relacionamento de coautoria é um dos mais importantes, devido ao fato de que os pesquisadores estejam estudando o mesmo assunto, então existe um forte interesse comum em um objeto de pesquisa. Os objetivos específicos são:

- Modelar e extrair os dados de pesquisadores de *APIs* públicas de plataformas de pesquisa;
- Consolidar os dados de tais plataformas de terceiros de forma que eles estejam unificados para o consumo no E-SECO;
- Montar uma página de perfil social de um pesquisador que esteja acessando o E-SECO, disponibilizando informações acadêmicas, métricas e publicações, bem como relacionamentos entre os maiores colaboradores;
- Disponibilizar uma materialização visual da rede social a fim de que o pesquisador

possa realizar uma melhor análise do alcance dos seus relacionamentos;

1.3 Metodologia e Organização

A pesquisa teve início com a revisão bibliográfica dos principais conceitos relacionados ao tema de pesquisa, com o intuito de verificar textos e publicações relacionadas com o assunto, bem como definições básicas da literatura a fim de conhecer melhor a respeito de como o assunto foi abordado em estudos anteriores e ter melhor domínio do problema. Tal revisão está contida no Capítulo 2 e apresenta conceitos essenciais sobre Software Científico, Colaboração, Ecossistemas de Software e Redes Sociais.

Nos Capítulos 3 e 4, a arquitetura E-SECO proposta por SOUZA, 2015 é brevemente apresentada. Em seguida, a abordagem Social E-SECO é introduzida como uma extensão ao E-SECO, visando mitigar o fato de uma plataforma web de colaboração oferecer somente suporte para colaborações indiretas e como a integração de uma rede social pode auxiliar em tal problema. Em seguida, serão descritas as fontes de dados utilizadas (DBLP e Google Scholar) e como tal extensão se integra à arquitetura do E-SECO. Posteriormente, detalhes da implementação são discutidos e como os mesmos se integram aos requisitos gerais propostos. Finalmente, um exemplo de uso ilustra a utilização da abordagem, explicando o funcionamento da arquitetura.

As considerações finais são apresentadas no Capítulo 5. É feita uma avaliação da adequação da solução final com os requisitos iniciais propostos e também são apontadas ramificações da presente pesquisa para trabalhos futuros.

2 Pressupostos e Revisão da Literatura

Conforme introduzido no capítulo anterior, o presente trabalho tem como principal objetivo, a modelagem e implementação de uma rede social de colaboração incorporando-a a um ecossistema para desenvolvimento de software científico. Esta rede de colaboração tem como finalidade a promoção da colaboração entre os utilizadores da plataforma.

Este capítulo apresenta uma revisão da literatura, englobando os principais conceitos envolvidos na abordagem. Assim, será realizada uma revisão dos conceitos relacionados a e-Science e os desafios para o desenvolvimento do software científico. Em seguida, serão discutidas, brevemente, as abordagens relacionadas a *e-Science*, como *workflows* científicos, e a consequente necessidade de colaboração entre os cientistas. Por último, para sustentar o principal objetivo desse trabalho, são apresentados conceitos de redes sociais, aprofundando-se em redes colaborativas científicas e como seu uso pode promover e apoiar a colaboração em um ecossistema de software científico.

2.1 Software Científico

Com a evolução constante da tecnologia e a globalização do conhecimento, se percebe, nos últimos tempos, que os softwares científicos são essenciais para apoiar o desenvolvimento científico. Os mesmos não são somente instrumentos para a geração de resultados, mas são também cruciais para a maior parte das pesquisas recentes (MAXBILE, 2009). No entanto, o desenvolvimento de software científico é essencialmente diferente do convencional. SEGAL AND MORRIS (2008) apontam que, usualmente, quando um cientista necessita de um software para apoio à sua pesquisa, normalmente ele não tem completo domínio dos requisitos. Por esse motivo, se torna completamente inviável especificar os mesmos, uma vez que eles emergem juntamente com o software e com o consequente maior entendimento do progresso no domínio.

Além disso, outra característica bastante relevante é a necessidade de um profundo conhecimento do domínio no qual se pretende realizar o desenvolvimento. Sendo

assim, um especialista (cientista) deve estar muito envolvido no processo de desenvolvimento de software científico. Na maioria das vezes, este software será utilizado pelo próprio cientista ou por um grupo restrito de colegas próximos, e neste contexto, o cientista geralmente é o próprio desenvolvedor (SEGAL AND MORRIS, 2008). Isso pode ser um problema, pois a maioria dos especialistas aprende sobre o desenvolvimento de software de maneira informal, o que pode tornar todo o processo caótico, gerando um produto final de software sem qualidade, dificultando sua manutenção, reuso e interoperabilidade.

Segundo SOUZA (2011), algumas das principais características em um software científico são as seguintes: em geral, não é desenvolvido através de um processo padronizado; nem sempre a organização dispõe de mão de obra para o desenvolvimento e operação do software em termos de conhecimento de domínio da tecnologia, por isso, muitas vezes não se conta com equipe especializada em desenvolvimento; em grande parte, a reutilização de artefatos é realizada de maneira informal. A resolução do problema é a prioridade do projeto e muitas das vezes os requisitos não são claros, ou são levantados durante o processo e podem sofrer alteração. Por isso, a adaptação às mudanças é mais importante do que seguir um plano.

Com todos esses elementos postos em evidência, é perceptível que muitas práticas de Engenharia de Software podem ser aplicadas para contribuir para a geração de produtos científicos de maior qualidade. Sabe-se que a utilização de um processo formal é de suma importância para o sucesso do desenvolvimento de softwares em geral, ainda mais quando o mesmo tem natureza acadêmica, o que em geral implica em maior complexidade no problema a ser resolvido. Porém, quando existe a participação de especialistas em Engenharia de Software em colaboração com os cientistas, surgem conflitos entre eles.

Para se enfrentar todos esses obstáculos, se faz necessária a utilização de práticas com o objetivo de facilitar a comunicação entre os envolvidos, como por exemplo, a utilização de um modelo baseado em domínio, como propõe EVANS (2008). O autor sugere como elementos para uma modelagem efetiva, achar um terreno comum para os todos os envolvidos, ligar a modelagem com a implementação, cultivar uma linguagem comum entre os participantes baseada no modelo, desenvolver uma modelagem rica em conhecimento e disseminar a mesma entre a equipe, e acima de tudo, promover a experimentação.

Juntamente com o já apresentado, com os recentes avanços da internet, tecnologias de *grid* e de *cloud computing*, o número de recursos que um cientista pode aplicar em uma pesquisa foi estendido. Por consequência, a complexidade e o ciclo de vida de experimentos utilizando software científico também aumentaram. Com isso, recentemente, *workflows* se tornaram uma abordagem popular para modelar e organizar tais elementos (BELLOUM et al., 2011). O conceito de *workflow* pode ser entendido como um modelo que define o fluxo de processos ou tarefas coordenadas e encadeadas usando um plano sistemático. Segundo ZHAO et al. (2009), *workflows* científicos provaram ser uma tecnologia chave para engrandecer muitas disciplinas de pesquisa, promovendo o gerenciamento e automatização de processos de experimentos. Por isso, os mesmos estão se tornando recursos valiosos para a experimentação científica, permitindo que cientistas de todos os lugares do mundo possam colaborar em um experimento. Como resultado, cientistas podem promover a troca de conhecimento e aumentar a velocidade na condução de experimentos. Entretanto, apoiar a colaboração neste contexto não é uma atividade trivial. Um *Workflow* Científico pode ser definido como um encadeamento de atividades, sendo que cada uma delas é mapeada para uma aplicação, formando um fluxo coerente de informações e controles. *Workflows* científicos são diferentes de *workflows* aplicados a sistemas convencionais, pois possuem características extras que lhe adicionam uma maior complexidade, tais como: (i) fluxos com um grande número de passos; (ii) volatilidade, já que o *workflow* poderá ser alterado durante o processo de avaliação das hipóteses científicas; (iii) necessidade de parametrização para muitas tarefas (NARDI, 2009).

2.2 Colaboração

O desenvolvimento rápido de tecnologias fez com que o número de recursos que os cientistas possam utilizar em uma pesquisa aumentasse significativamente. Ou seja, os diferentes requisitos e diversidade de métodos de experimentação fazem com que os mesmos tenham uma coleção de ferramentas disponíveis para uso neste contexto. Somado a este fato, com a globalização da comunicação, hoje é possível realizar colaborações em grande escala, envolvendo cientistas geograficamente distribuídos. Com isso, experimentos científicos complexos envolvem recursos computacionais e dados distribuídos e uma intensa

colaboração entre os cientistas envolvidos (BELLOUM et al., 2011).

SHNEIDERMAN (2008) defende que estamos diante de um novo tipo de ciência, visto que a necessidade de colaboração justifica o foco na mesma. Segundo o autor, as principais diferenças entre a "nova ciência" e a tradicional, seriam a colaboração humana, a integração, experimentos intervencionais localizados no mundo real ao invés de uma ciência reducionista, com experimentos controlados conduzidos em laboratório. Dessa maneira, devemos expandir os métodos científicos tradicionais para lidar com questões complexas que a inovação tecnológica nos traz.

A colaboração começou a aparecer no meio científico nos séculos XVII e XVIII, e sua principal forma está na coautoria do trabalho de pesquisa e publicação. Apesar de que na ciência moderna as colaborações possam ir bem além disso, essa continua sendo uma das formas mais populares de interação entre pesquisadores (PHAM, 2006).

Na pesquisa moderna, a colaboração está se tornando cada vez mais relevante, podendo trazer grandes vantagens para os resultados das análises pois resultarão em avanços mais rápidos e maior qualidade na pesquisa (KRAUT et al., 1987), além de interações entre as comunidades acadêmicas. Entretanto, a colaboração traz muitos desafios, considerando a necessidade de recursos computacionais compartilhados para a condução de experimentos. Deve-se também considerar que tais colaborações vêm sendo feitas em escala global, o que leva ao fato dessas interações ocorrerem comumente através de portais colaborativos baseados na web. Dessa forma, um cientista consegue acessar os recursos da pesquisa de qualquer lugar através de um navegador web.

PHAM (2006) resume bem as características comuns das colaborações científicas que ocorrem nos dias atuais. Estas envolvem: o compartilhamento de recursos complexos e caros, além de um grande volume de dados; a necessidade de conhecimento multidisciplinar; o suporte à colaboração em escala global, não se limitando às fronteiras de uma universidade ou comunidade acadêmica; a competição entre os pesquisadores; e também, o suporte à comunicação informal.

As tecnologias que pretendem servir como base para colaboração científica devem levar em consideração e suportar as seguintes características: ser capaz de compartilhar os recursos, armazenar os dados de pesquisa, facilitar o compartilhamento do conheci-

mento e prover a segurança dos recursos pessoais de cada cientista, a fim de incentivar o envolvimento dos mesmos na plataforma. Portanto, deve-se suportar a experimentação científica de três ângulos: compartilhamento da informação, comunicação e coordenação (BELLOUM et al., 2011).

O modelo de colaboração 3C, apresentado originalmente por ELLIS et al. (1991) suporta de maneira orgânica tais necessidades. Para os autores, a colaboração pode ser analisada sobre três aspectos, tais como: comunicação, coordenação e cooperação. Conforme ilustrado na Figura 2.1, a coordenação intermedia o processo para o gerenciamento de pessoas, atividades e recursos, a cooperação é caracterizada pela atuação em conjunto no espaço compartilhado para a produção de objetos ou informações e a comunicação se realiza através da troca de mensagens entre os participantes, assegurando-se que a mesma seja entendida pelo receptor e este assuma compromissos relativos à mensagem. Apesar da separação das atividades para fins de análise, elas são realizadas de forma continua e interativamente durante o trabalho do grupo (FUKS et al., 2003).

Figura 2.1: O modelo de colaboração 3C (FUKS et al., 2003)

2.3 Ecossistemas de Software

Ecossistemas de Software (ECOS) é um campo de pesquisa que vem sendo explorado nos últimos anos e pode ser considerado como um conjunto de atores que colaboram e interagem com um mercado comum para software e serviços, juntamente com as relações entre esses atores. Tais relações são comumente sustentadas por uma plataforma tecnológica comum ou de mercado e operam através da troca de informações recursos e artefatos.

Um ecossistema é composto basicamente por um centralizador, uma plataforma e um conjunto de agentes de nicho. O centralizador atua no desenvolvimento da plataforma e no gerenciamento dos relacionamentos com as partes externas. Já os agentes de nicho são aqueles que influenciam no desenvolvimento do ecossistema. Tomando como exemplo o ecossistema do Android, o Google exerce o papel de centralizador, o Android é a plataforma e os agentes de nicho são os desenvolvedores de aplicativos e empresas como Samsung ou Motorola (SOUZA, 2015).

Um dos principais desafios relacionados ao desenvolvimento e modelagem de um ECOS está em sua caracterização (JANSEN et al., 2009). A saúde de um ECOS está intimamente ligada à gerência de relacionamentos das partes externas, e substancialmente como estes relacionamentos podem gerar valor para ambas as partes.

Existem dois cenários distintos que podem levar uma organização a caminhar em direção a tal abordagem. Um deles é quando a quantidade de funcionalidades a serem desenvolvidas para atender os clientes é muito maior do que ela poderia realizar sozinha. Outro fator é quando a aplicação possui grande demanda de customização em massa, de modo a estender sua plataforma com atores externos à organização (BOSCH, 2009).

2.4 Redes Sociais

Conforme apresentado na seção anterior, o suporte à colaboração aparece como principal forma de incentivar os avanços nas pesquisas científicas atuais, proporcionando maior agilidade e qualidade nos experimentos. Com o intuito de auxiliar os cientistas a compartilhar os seus recursos, armazenar os dados das pesquisas e facilitar a comunicação

entre os envolvidos, portais web foram criados. Dessa forma, um cientista pode ter acesso ao mesmo de qualquer lugar através de um navegador web.

Entretanto, conforme aponta PHAM (2006), o suporte que os portais colaborativos baseados na web oferecem para os cientistas é realizado de forma indireta como ilustra a Figura 2.2. Todas as colaborações têm de ser realizadas através de recursos que são administrados por terceiros. Em geral, isso pode trazer algumas limitações, tais como: (i) falta de suporte para colaborações entre grupos, que são comuns nas comunidades científicas e nas quais as pesquisas são, em geral, interdisciplinares; (ii) inflexibilidade para apoiar colaborações distribuídas em comunidades que não possuem tanto contato, uma vez que todas as atividades de colaboração acabam tendo de ser realizadas em um servidor centralizado; e, por último, (iii) como a solução depende de uma arquitetura baseada na web, com uma aplicação de servidor atendendo vários clientes, uma falha pode se tornar um gargalo para as comunicações.

A respeito do último problema apontado, com o avanço das tecnologias e arquiteturas de *cloud computing*, tem-se um caminho a seguir, uma vez que os servidores não são necessariamente centralizados, e é possível escala-los ou realizar um balanceamento das requisições de forma a otimizar a demanda. Os dois primeiros problemas, porém, possuem uma limitação de característica mais social, uma vez que as colaborações acontecendo pela web impedem a interação direta entre os cientistas. A utilização de redes sociais de colaboração em tais plataformas pode ser interessante para mitigar essa distância, uma vez que evidencia a conexão entre os integrantes da rede, o que denota um nível de conhecimento entre os cientistas. Sua análise, pode estimular os pesquisadores a buscarem novas interações diretas.

Existem muitos dados na web com características sociais, e muitos pesquisadores têm feito o uso desses recursos para desenvolvimento de seu trabalho científico. Na atualidade, a web é a principal interface pela qual os cientistas pesquisam sobre trabalhos relacionados aos seus, livros, artigos e pessoas que possam ajudar em sua pesquisa. Tal fenômeno vem possibilitando que a forma como os indivíduos estabelecem conexão, uns com os outros, seja estudada e como as mesmas evoluem com o tempo. Essas ligações entre as pessoas são representadas como uma rede social (STROELE, 2013).

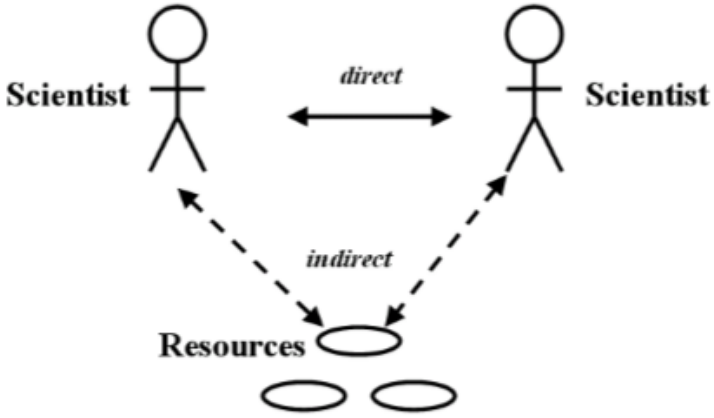

Figura 2.2: Suporte direto e indireto para colaborações (PHAM, 2006)

2.4.1 Rede Social

Uma rede social pode ser entendida como um conjunto de pessoas ou grupos que possuem conexões entre si, onde estes grupos podem ser definidos de diferentes formas, dependendo da situação ou do interesse que se tenha a analisar. Na linguagem de análise de redes sociais, as pessoas ou grupos são chamadas de atores e as ligações entre os mesmos de laços. Um ator pode ser uma única pessoa, um time ou uma companhia inteira e os laços podem representar amizades, colaboração ou mesmo uma característica em comum entre os atores representados (NEWMAN, 2001a).

Redes sociais é um assunto que vem ganhando bastante destaque nos dias atuais, porém é um tema que vem sendo abordado a bastante tempo pela ciência. De acordo com WASSERMAN AND FAUST. (1994), estudos precursores à respeito de redes sociais existem desde o final do século XIX, como por exemplo os de Émile Durkheim e Ferdinand Tönies, que começavam a enxergar a sociedade como grupos sociais que possuíam ligações pessoais ou impessoais. As redes sociais vêm sendo estudadas tanto do ponto de vista teórico quanto do empírico em diversas áreas de pesquisa. Através dos tempos, isso vem não só do inerente interesse em padrões, mas também das importantes implicações de sua estrutura, na qual se pode estudar por exemplo a disseminação de uma informação ou de uma doença (NEWMAN, 2001a).

Um famoso estudo empírico da estrutura de redes sociais foi o conduzido por Stanley Milgram, no qual ele escolheu aleatoriamente pessoas a partir de um catálogo telefônico do Nebraska. Posteriormente, o pesquisador pediu aos participantes do estudo

que enviassem uma carta a alguém em Boston, porém a instrução básica é que cartas só poderiam ser enviadas a pessoas que fossem diretamente conhecidas. Eventualmente, algumas dessas cartas chegaram ao seu destino e Milgram descobriu que o número médio de passos que as cartas tomavam antes de chegar era de cerca de seis. Esse resultado é usualmente usado como evidência para "hipótese do mundo pequeno", que afirma que a maioria dos pares de pessoas em uma população pode ser conectada por um pequeno caminho de conhecidos (NEWMAN, 2001a). Esse é um claro exemplo de usos de redes sociais para a análise da disseminação da informação, mas as mesmas também são usadas na epidemiologia, a fim de entender padrões de contato humano em doenças e também para inibir a disseminação de vírus, como o HIV em uma população, entre outros (WASSERMAN AND FAUST., 1994)

2.4.2 Redes Sociais Científicas

NEWMAN (2001a) descreve que não é simples realizar a análise de uma rede social, uma vez que não se tenha conhecimento da fonte de dados, a qualidade da coleta destes pode não ter sido ideal. Além disso, em uma rede que tenta analisar laços de amizade entre os atores, por exemplo, a definição de amizade pode variar de pessoa para pessoa, tornando os dados não confiáveis. Por isso, uma fonte de dados mais promissora é a rede de afiliações, na qual os atores estão conectados uns com os outros através de algum tipo de associação. No âmbito de pesquisa acadêmica e estudos científicos, a relação de coautoria é uma das associações mais importantes que se tem, devido ao fato de os pesquisadores estarem estudando o mesmo assunto (STROELE, 2013).

A ideia de construir uma rede de coautoria não é nova. O conceito, por exemplo do número Erdös, nomeado em homenagem ao famoso matemático húngaro, no qual se calcula o quanto um pesquisador está distante em termos de publicação de Erdös, mostra claramente a intenção de construção desse tipo de rede. Redes sociais científicas são redes sociais nas quais dois cientistas estão considerados conectados uma vez que eles foram coautores em um artigo. Em adição à distância entre os atores, existem muitos outros fatores interessantes que podem ser levados em consideração em tais redes, incluindo o número de coautores de cientistas e também o número de artigos que eles escrevem, e a

variação dos mesmos ao decorrer do tempo (NEWMAN, 2001b).

Existem dois tipos de redes sociais: homogêneas e heterogêneas. As primeiras, são as quais só existe um tipo de relacionamento, e o conhecimento flui através dessa ligação específica, por outro lado, em redes heterogêneas, também conhecidas como multi-relacionais, assume-se que os atores estejam compartilhando diferentes tipos de relação entre si.

2.4.3 Visualização de Redes Sociais

A utilização de imagens visuais é muito comum em vários ramos da ciência, e pode-se afirmar que é muito importante para o progresso de várias áreas. O historiador Alfred Crosby (1997) vai além disso e propõe que a visualização, juntamente com as medidas, são os fatores responsáveis pelo desenvolvimento explosivo de toda a ciência moderna. Esses dois fatores vêm sendo centrais no crescimento da análise de redes sociais. Avanços nas métricas têm sido rápidos e regulares nas pesquisas e, desde o começo, imagens visuais têm um papel chave nas pesquisas, provendo aos investigadores novas intuições sobre a estrutura das redes e facilitando as comunicações (FREEMAN, 2000).

Muitos estudos se desenvolvem somente para a resolução de problemas em análises de redes sociais, e sendo que muitas das vezes são extremamente complexos, não existindo foco na visualização das redes. Porém, a mesma é essencial e a representação de seus elementos permite que os pesquisadores tenham um melhor entendimento sobre o problema e o cenário a ser estudado. Podendo assim, simplificar a análise realizada pelos cientistas, tornando-se um passo importante na análise de redes sociais. As redes podem ser representadas por um grafo, nos quais os nós representam os atores e as arestas representam os laços entre os mesmos (PITAS, 2015).

3 E-SECO

O principal objetivo do presente trabalho é construir uma rede social de colaboração entre os pesquisadores que utilizam um ecossistema de software científico E-SECO (SOUZA, 2015) e integrá-la ao sistema. A mesma foi desenvolvida, extraindo e analisando-se dados de cientistas disponíveis na Web nas plataformas acadêmicas Google Scholar (GOOGLE SCHOLAR, 2004) e DBLP (LEY, 2009). O relacionamento de coautoria será utilizado para extrair as conexões entre os utilizadores. Com a incorporação no ecossistema, pretende-se promover uma maior visibilidade do alcance dos relacionamentos dos pesquisadores, bem como estimular a colaboração direta entre os utilizadores da plataforma que, segundo PHAM (2006), é uma das grandes deficiências de portais de colaboração baseados na web.

O presente capítulo irá inicialmente apresentar a motivação por trás da plataforma E-SECO, desde suas abordagens iniciais até a evolução em um ecossistema de software científico.

3.1 Abordagem

O desenvolvimento de software para o meio científico, em geral, conforme dito anteriormente, é realizado de maneira informal. Os desenvolvedores não possuem conhecimento na área de pesquisa, então o cientista acaba assumindo a posição de desenvolvedor, o que pode tornar todo processo caótico (SEGAL AND MORRIS, 2008).

Workflows científicos são bastante utilizados nesse contexto, no entanto a especificação e execução dos mesmos não é uma tarefa trivial e muitas vezes interdisciplinar, exigindo do cientista algum conhecimento em computação (HANNAY et al., 2009). O resultado são grandes barreiras e dificuldades no desenvolvimento e/ou reuso de *workflows* de outros cientistas, o que geralmente leva ao retrabalho. Torna-se então necessário a construção de ferramentas que apoiem o cientista e que minimizem a necessidade de conhecimento específico e multidisciplinar. O conceito de Linha de Produtos de Software

(LPS) vinha sendo utilizado nesse cenário (CASTRO et al., 2015)) na tentativa de minimizar esses problemas, obtendo ganhos na produtividade do processo de experimentação, além de aumentar a qualidade e reduzir o custo através do reuso sistemático e planejado de artefatos.

A abordagem proposta inicialmente através da arquitetura PL-Science (COSTA, 2013) utiliza o conceito de linha de produto de software científico a fim de auxiliar os cientistas na especificação de um *workflow* que mapeia todos os passos de uma aplicação científica de determinado experimento. Para a construção dos *workflows* e representar sua variabilidade, são utilizados modelos de *features*. Porém, devido à alta complexidade dos softwares científicos, tal modelo não é capaz de traduzir toda a variedade de características. O uso de ontologias é proposto a fim de trazer apoio semântico na construção dos produtos na LPSC, trazendo maior expressividade para a seleção de características desejáveis no desenvolvimento de *workflows* e/ou aplicações científicas no domínio trabalhado.

Como uma extensão do PL-Science, a abordagem Collaborative PL-Science foi proposta por (PEREIRA, 2014) a fim de explorar as lacunas deixadas por (COSTA, 2013) no domínio, como ausência de apoio para concepção de *workflows*, reuso das aplicações e principalmente dificuldades na cooperação e comunicação quando trabalhando com equipes de cientistas geograficamente dispersos. Introduzindo elementos de colaboração na arquitetura do PL-Science foi possível dar suporte à comunicação, cooperação e coordenação, além do desenvolvimento de uma memória de grupo para capturar os rastros deixados pelos cientistas durante o processo. Com isto, pesquisadores passam a ter a possibilidade de trabalhar de maneira colaborativa no desenvolvimento de *workflows* científicos a partir do núcleo de artefatos da LPSC. Como resultado, experimentos complexos que envolvem interações entre os cientistas, utilização e persistência de grandes volumes de dados, serviços e recursos computacionais distribuídos são apoiados.

Segundo BELLOUM et al. (2011), um experimento científico colaborativo passa por um ciclo de vida que se inicia na investigação do problema, seguido pela prototipação e execução do experimento até finalmente chegar a etapa de publicação das contribuições. Sob a perspectiva do ciclo de vida apresentado anteriormente, é possível afirmar que a abordagem Collaborative PL-Science está inserida na fase de prototipação do experi-

mento e, uma de suas limitações, se encontra no fato de a abordagem não proporcionar a execução do experimento e por consequência todas as fases posteriores do processo de experimentação colaborativo.

Com o intuito de atacar essa limitação apontada na abordagem Collaborative PL-Science, SOUZA (2015) estende mais uma vez a arquitetura proposta utilizando uma estratégia de desenvolvimento baseado em Ecossistemas de Software (ECOS) e apresenta a abordagem ECOS PL-Science, posteriormente renomeada para E-SECO. O objetivo principal é preencher a lacuna deixada por (PEREIRA et al., 2016), aumentando o escopo de suas funcionalidades com a realização de integrações com Sistemas Gerenciadores de *Workflow* Científico (SGWfC) de modo a possibilitar os cientistas a execução dos experimentos dentro da plataforma. Com isso, todas as etapas do ciclo de vida de um experimento científico proposto por BELLOUM et al. (2011) são contempladas, provendo recursos para suportar o processo, e estão elencados abaixo:

- Investigação do problema: Utilização de revisões sistemáticas da literatura, abrindo uma frente de pesquisa para explorar trabalhos e experimentos relacionados;
- Prototipação do experimento: Disponibilização aos cientistas recursos para utilizarem *workflows* e serviços web de plataformas renomadas. A LPS do PL-Science é então incorporada na etapa, aumentando o nível de reuso e qualidade no desenvolvimento dos experimentos;
- Execução do experimento: Integração de uma plataforma para transformação de *workflows* científicos com o objetivo de aumentar a flexibilidade para execução das tarefas;
- Disponibilização dos resultados: Todos os dados da execução do experimento e também de todo o processo de experimentação são armazenados no E-SECO, possibilitando que outros pesquisadores possam consultar e estender as pesquisas.

SIRQUEIRA et al. (2016) propõem o desenvolvimento de uma arquitetura para suporte à manutenção e evolução de experimentos científicos, fornecendo de maneira automatizada, informações estratégicas relacionadas a evolução e manutenção do experimento, de forma que os cientistas possam tomar decisões ou obter maior conhecimento

em relação ao mesmo. Para tal objetivo foi adicionado uma camada de proveniência de dados à arquitetura, a partir do registro da história da derivação dos dados, possibilitando a reprodutibilidade, interpretação dos resultados e diagnóstico de problemas. Essa arquitetura é denominada E-SECO ProVersion, e busca adicionar funcionalidades específicas para a gerência de configuração de experimentos científicos no contexto de ECOSC e laboratórios colaborativos

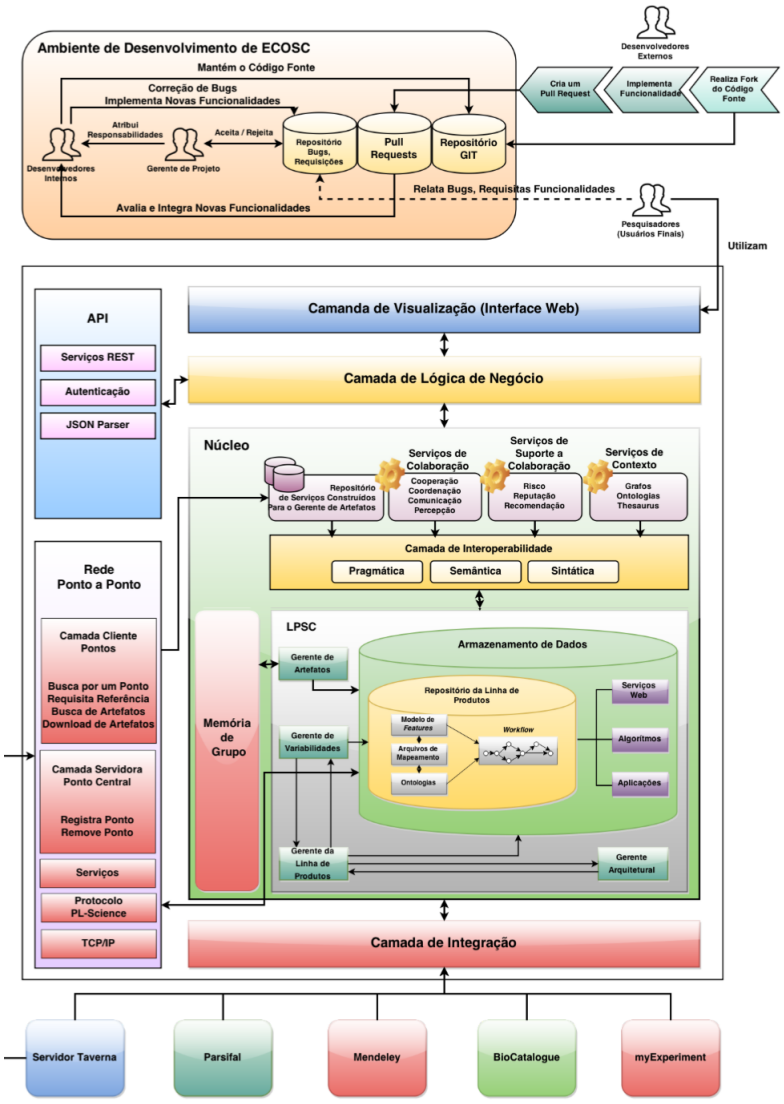

Figura 3.1: Arquitetura da abordagem E-SECO (SOUZA, 2015)

4 Social E-SECO

4.1 A abordagem Social E-SECO

Os problemas pesquisados pela ciência moderna estão cada vez mais complicados, e abordá-los requer conhecimento e experiência de uma ampla gama de disciplinas científicas. Os instrumentos solicitados para a execução de tais experimentos são também complexos e caros. Isso soma-se ao fato de que o volume dos dados gerados pelas pesquisas estão cada vez maiores de forma que não podem ser gerenciados por uma organização individual. Todos esses são fatores tornaram a colaboração distribuída global cada vez mais importante para a ciência moderna, ou "Ciência 2.0", como propõe (SHNEIDERMAN, 2008). Portais Web colaborativos são comumente utilizados como ambientes para interações entre pesquisadores distribuídos, que por sua vez estão sujeitos a um certo nível de administração centralizada e controle.

Conforme apresentado na seção anterior, a abordagem E-SECO proposta por SOUZA (2015) representa a extensão da ferramenta de experimentação científica colaborativa sugerida por PEREIRA (2014). Essencialmente, a solução é uma arquitetura através da qual o cientista se torna capaz de especificar e executar um experimento colaborativo, contemplando assim todas as fases de seu ciclo de vida. Entretanto, apesar de haver possibilidade de interação e compartilhamento de recursos entre os pesquisadores e os dados tanto da execução quanto do processo de um experimento estarem capturados na memória de grupo, as colaborações realizadas são indiretas. As mesmas têm de ser realizadas utilizando um servidor de terceiros. Segundo PHAM (2006) isso acarreta em diversas limitações, como falta de suporte para colaborações multidisciplinares e inflexibilidade para apoiar experimentos que são realizados entre comunidades que não são muito próximas geográfica e socialmente.

A abordagem Social E-SECO, proposta no presente trabalho, tem como objetivo atacar essa limitação. Sabe-se que as colaborações indiretas continuarão existindo na plataforma, uma vez que a mesma é inerente ao modelo proposto e, mesmo possuindo

suas limitações, é possivelmente a arquitetura que promove maior abrangência para que as colaborações sejam realizadas de maneira distribuída a qualquer momento em escala global. O que se propõe é que com a integração de uma rede social de colaboração científica na plataforma, o pesquisador tenha maior visibilidade sobre o alcance de seus relacionamentos, visto que as conexões do cientista estão evidenciadas. E, através dela, seja estimulado a buscar comunicações diretas com os colaboradores, bem como novas interações e parcerias com outros pesquisadores que estão conectados indiretamente a ele na rede. Com isso, a distância entre os cientistas que estão colaborando uns com os outros na plataforma será mitigada.

Assim, o propósito do presente trabalho é construir uma rede social científica de colaboração, na qual dois pesquisadores estarão ligados se possuírem publicações em coautoria. Consideramos que esse relacionamento é um dos mais importantes no domínio de pesquisa acadêmica, uma vez que evidencia que os autores estão estudando o mesmo assunto STROELE (2013). Além disso, tal relação se torna extremamente interessante para a abordagem proposta, uma vez que se dois pesquisadores publicaram um artigo em conjunto, é grande a probabilidade de que os dois se conheçam pessoalmente. Fato este que corrobora a ideia de mitigar o impacto das relações indiretas que a plataforma colaborativa web traz consigo.

Tal fato pode ser ilustrado, por exemplo, no caso de um pesquisador A ter como conexão direta com o pesquisador B (primeiro nível) e B tem uma conexão direta com o pesquisador C, que estuda o mesmo tema do que A. Tendo essa ligação evidenciada na rede colaborativa, o pesquisador A pode se interessar em entrar em contato com o pesquisador C a fim de trabalharem em conjunto e poderá utilizar o pesquisador B como intermediário, uma vez que a probabilidade de os dois se conhecerem é grande. Isso pode, de fato, estimular que colaborações diretas ocorram através da plataforma. Para construção da rede social colaborativa, utilizou-se de informações disponibilizadas por plataformas de apoio acadêmico bastante renomadas, no caso da presente pesquisa, o Google Scholar (GOOGLE SCHOLAR, 2004) e o DBLP *Computer Science Bibliography* (DBLP, 1993), que serão melhor apresentados em uma seção 3.3. Os dados das mesmas são consumidos por uma *API* e consolidados em um *middleware* e com isto é possível

modelar a rede social. Nada impede que dados de outras plataformas ou até mesmo uma base local sejam utilizados como fonte para rede. Para isso é necessário somente o trabalho de consolidar os dados de todas as origens no *middleware* que será apresentado como parte fundamental da arquitetura do Social E-SECO.

Os dados consolidados são disponibilizados de forma que sejam incorporados ao E-SECO em uma página de perfil do cientista logado na plataforma, a ser acessada a partir de um link na tela inicial do E-SECO, mostrada na Figura 4.1. Nesta tela, são disponibilizadas informações básicas sobre o cientista, bem como seus principais colaboradores, todos os seus artigos publicados, métricas acadêmicas e também uma visualização gráfica da rede social do pesquisador.

A proposta de busca de dados em plataformas externas torna a abordagem bastante versátil e pronta para utilizar, uma vez que possui acesso a uma considerável massa de dados sem a necessidade de armazenamento dos mesmos e também é simples incorporar uma nova plataforma como fonte de dados. Porém, como a informação é provida por terceiros, isso oferece um risco para a plataforma, uma vez que tais dados e plataformas estão em constante evolução e podem sofrer alterações de forma independente. Uma limitação da proposta está no fato de o cientista precisar de ter os dados disponíveis nas plataformas utilizadas na presente pesquisa para que possa ter acesso a sua rede social. No entanto trabalhos futuros podem atacar essa restrição construindo uma base de dados local para ser consumida pela aplicação.

4.1.1 Especificação de Requisitos

Antes de definir a estratégia de desenvolvimento, bem como escolher um modelo arquitetural que permita a evolução da plataforma, é necessário que se faça um levantamento dos requisitos a fim de viabilizar uma melhor análise e fundamentar o projeto arquitetural. O objetivo de tal trabalho é que durante o processo de desenvolvimento da solução, todos os requisitos estejam analisados e entendidos, evitando retrabalhos ou situações que sejam mais difíceis de se contornar.

A elicitação dos requisitos foi feita com base no estudo de trabalhos que tratam do processo de construção e visualização de redes sociais de colaboração científica

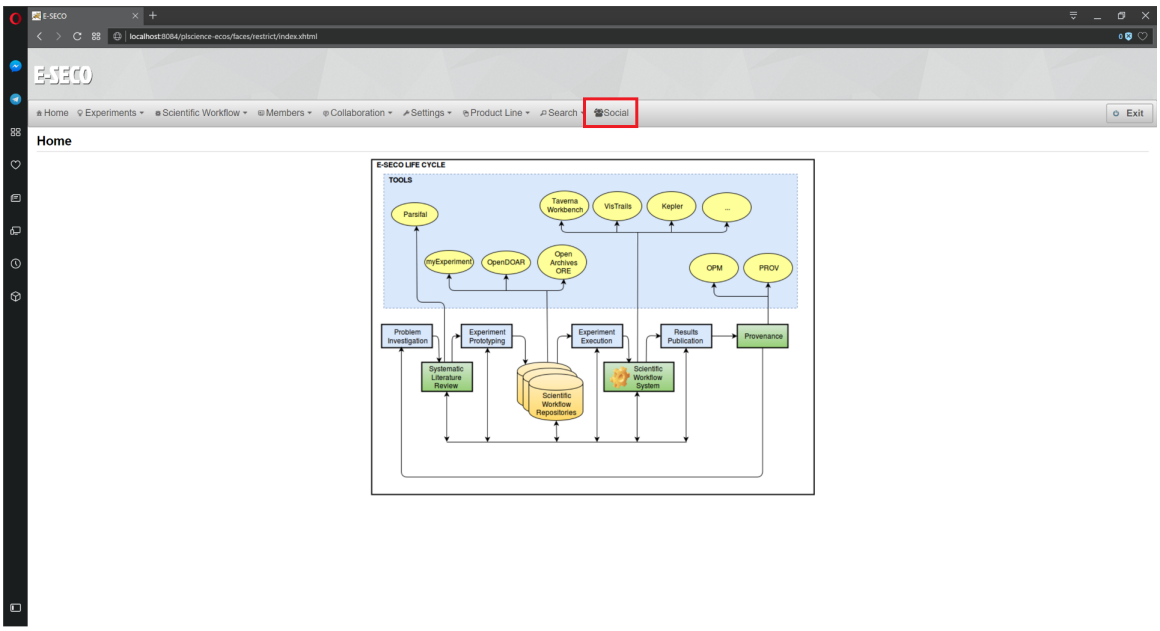

Figura 4.1: Página inicial do E-SECO com destaque para a opção SOCIAL no Menu

(NEWMAN, 2001a),(NEWMAN, 2001b), (STROELE, 2013) juntamente com os dados disponíveis nas plataformas que serviram como fonte de dados. Após entendimento e análise dos principais aspectos necessários e comuns ao desenvolvimento de uma rede social juntamente com as restrições levantadas com as plataformas acadêmicas consumidas, o escopo e funcionalidades a serem implementadas na proposta do Social E-SECO foram delimitadas. Nas seções 4.1.1 e 4.1.1 são apresentados respectivamente os requisitos funcionais e não funcionais levantados para o desenvolvimento da aplicação:

Requisitos Funcionais

- Permitir que as informações necessárias, que são: informações básicas sobre um autor, artigos de um autor, métricas acadêmicas e coautorias sejam consumidas de plataformas de terceiros (Google Scholar e DBLP).
- Oferecer a consolidação dos dados trazidos de tais plataformas de forma que os mesmos sejam unificados e disponibilizados para que seja construída a rede social.
- Permitir que o cientista logado na plataforma E-SECO possa acessar sua rede social de forma integrada à aplicação.
- Permitir a visualização de um perfil social que inclua os dados de publicações, coautorias e métricas acadêmicas do cientista.

- Disponibilizar uma materialização da rede social científica colaborativa, a fim de que o cientista consiga ter uma visualização palpável do alcance de seus relacionamentos.

Requisitos não funcionais

A proposta do Social E-SECO representa uma extensão à abordagem de SOUZA (2015) e traz consigo algumas complexidades que derivam em alguns requisitos não funcionais, tais como: flexibilidade, extensibilidade, escalabilidade, disponibilidade e segurança que serão essenciais para o desenvolvimento e evolução da abordagem.

- Flexibilidade: A arquitetura deverá ser flexível, uma vez que ela depende de plataformas e serviços providos por terceiros que são mantidos e evoluídos de forma independente e podem sofrer alterações a qualquer momento.

- Extensibilidade: A arquitetura da solução deverá possibilitar que novas plataformas sejam incorporadas como fontes de dados para a rede de forma simples e que não comprometa a arquitetura. Para isso deverá ser desenvolvido um *middleware* que possibilite a conexão com qualquer serviço de terceiro, realize a consolidação dos dados e disponibilize os mesmos de forma unificada.

- Escalabilidade: A arquitetura do Social E-SECO deverá ser escalável, uma vez que com o aumento na utilização dos seus recursos, e abertura a novas integrações ocasionando aumento no número de requisições, a arquitetura deve suportar tal condição e não ser um gargalo comprometendo o desempenho da solução.

- Disponibilidade: Uma vez que os dados são obtidos através de serviços providos por terceiros, se faz necessária uma estratégia para que os mesmos estejam disponíveis e não sejam comprometidos mesmo que o provedor dos dados não esteja disponível no momento do acesso.

- Segurança: Como os dados consultados e disponibilizados através da solução tem cunho pessoal, o acesso aos mesmos deve ser controlado e protegido de forma a não expor dados confidenciais indevidos. As *APIs* desenvolvidas para consumo dos dados não deverão ser externalizadas e protegidas através de uma conexão privada com o servidor no qual se encontra o E-SECO.

4.2 Fontes de Dados

Para a obtenção dos dados utilizados na construção da rede social científica, decidiu-se pela obtenção através de informações disponibilizadas em plataformas acadêmicas, como citado anteriormente. Isso torna a solução flexível e facilmente implementável, pois não há a necessidade de realização de coleta de dados e nem mesmo o armazenamento de um grande volume de informações. As subseções 4.2.1 e 4.2.2 descrevem, brevemente, um histórico à respeito das plataformas utilizadas.

4.2.1 Google Scholar

O Google Scholar (GOOGLE SCHOLAR, 2004) é uma ferramenta gratuita de busca na web lançada em novembro de 2004 que indexa dados e metadados de literatura acadêmica de diversas áreas de pesquisa e em diferentes formatos. É uma das plataformas mais utilizadas por pesquisadores no intuito de localizar artigos, publicações e também outros cientistas. Ela provê uma maneira simples de realizar uma ampla pesquisa na literatura acadêmica. Centraliza, em um só lugar, artigos, teses, livros, resumos de diversas disciplinas e de diferentes editoras, repositórios, universidades e até mesmo outros websites.

Apesar de o Google Scholar não divulgar oficialmente, de acordo com a pesquisa de KHABSA et al. (2014), estima-se que cerca de 114 milhões de documentos acadêmicos em língua inglesa estão disponíveis na web e que destes, o Google Scholar possui pelo menos 100 milhões. Tal número evidencia a relevância e quantidade de dados que a plataforma disponibiliza.

4.2.2 DBLP

O DBLP *Computer Science Bibliography* (DBLP, 1993) é um projeto que teve seu início de forma despretensiosa em 1993 como um simples teste para a tecnologia web. Inicialmente, tabelas de conteúdos de importantes anais e jornais acadêmicos foram digitadas e marcadas utilizando HTML e colocadas à disposição em um servidor chamado "*Data Bases and Logic Programming*" (DBLP). Surpreendentemente esse pequeno servi-

dor web experimental se tornou bastante útil como catálogo para consultas acadêmicas e se tornou hoje em um popular serviço disponibilizado para a comunidade acadêmica de ciência da computação.

O DBLP é um projeto da University of Trier, da Alemanha e em 2009 possuía mais de 1,2 milhões de registros bibliográficos. A plataforma é uma popular ferramenta para a comunidade acadêmica de ciência da computação na qual é possível rastrear o trabalho de colegas e obter detalhes bibliográficos que podem ser usados na composição de listas de referências para novos artigos. Outro uso comum, porém, controverso do DBLP é o de ranquear e obter o perfil de pessoas, instituições, jornais e conferências (LEY, 2009).

4.3 Métricas

Algumas métricas acadêmicas, providas pelas plataformas citadas nas subseções anteriores são importantes para a construção da rede social e necessitam ser introduzidas. As mesmas estão listadas a seguir:

- Número de Citações: Representa o número de citações acadêmicas que um autor teve em seus artigos. Pode ser observada em granularidades diferentes: a nível de publicação, representa a quantidade de citações que um único artigo acadêmico possui; e a nível de autor, que representa o somatório de citações de todos os artigos de um autor. Essa métrica é importante para determinar a relevância (e reputação) de um autor e também de um artigo acadêmico;

- Número de Coautorias com um Autor: Representa o número de coautorias em publicações acadêmicas que um determinado autor possui com o outro. Essa métrica é importante para a concepção da rede de coautorias, uma vez que existe ligação entre dois autores se os mesmos forem coautores em uma publicação.

- *h-index*: É uma métrica proposta por (HIRSCH, 2005) cujo objetivo é medir tanto a produtividade de um autor quanto os impactos das publicações acadêmicas do mesmo. O índice é definido de forma que um autor, que possua um índice de valor h, deve ter h publicações citadas pelo menos h vezes em outras publicações;

- *i10-index*: É uma métrica criada pelo Google Scholar. Representa o número de publicações de um pesquisador que tenha pelo menos 10 citações. As suas principais vantagens é a facilidade de entendimento e cálculo, porém a grande desvantagem é sua utilização exclusiva do Google Scholar.

4.4 Arquitetura da Solução

Conforme já dito, o Social E-SECO representa uma extensão a proposta E-SECO e, para tal, sua arquitetura será acoplada na plataforma apresentada na Figura 4.2, a fim de incorporar a abordagem de redes sociais científicas à plataforma de ecossistema. A arquitetura está de acordo com a proposta original, utilizando o padrão MVC (*Model View Controller*) (MVC, 2003) em conjunto com o paradigma orientado à serviços (PERREY AND LYCETT, 2003). A extensão representada pelo Social E-SECO pode ser vista como um novo serviço plugado ao E-SECO. Essa arquitetura fornece um fraco acoplamento em relação ao resto do projeto, o que agrega extensibilidade e flexibilidade a solução. Esse é um objetivo que vem sendo buscado no desenvolvimento atual de software, a fim de evitar que a aplicação se torne monolítica, cuja manutenção, implantação e escalabilidade são comprometidos, uma vez que todo o sistema está fortemente acoplado.

A arquitetura proposta utiliza a camada de visualização e lógica de negócio do E-SECO, bem como da memória de grupo e seu banco de dados. A interação inicial é realizada a partir da interface web, contida na camada de visualização. A mesma está diretamente ligada com um controller na camada de lógica de negócio que é responsável, através de um client, de realizar uma requisição dos dados consolidados de um pesquisador a um *middleware*. Uma vez que os dados são retornados, o controlador gerencia o fluxo para a camada de visualização.

O *middleware* é responsável por receber uma requisição, na qual se deseja buscar os dados de um pesquisador e realizar tal pesquisa nas plataformas Google Scholar e DBLP e então retornar as informações consolidadas em um formato comum (JSON, 1999). Visto que um dos requisitos é a disponibilidade e como as informações são providas por serviços gerenciados por terceiros, tomou-se a decisão arquitetural de persistir os dados provenientes de tais plataformas em uma base de dados NoSQL que irá trabalhar como

um *document store*. Com isso, as informações carregadas serão "cacheadas" e estarão disponíveis mesmo que, eventualmente um dos serviços esteja fora do ar. Tal solução traz ainda como ganho mais velocidade para a comunicação, uma vez que evita a necessidade de realizar requisições HTTP para os servidores dos serviços e então serializar binariamente as respostas recebidas.

Os dados relativos aos pesquisadores armazenados na *document store* devem ser alvo de uma política de atualização, visto que podem se tornar obsoletos. A atualização será realizada assincronamente toda vez que houver uma nova requisição. Ao receber um pedido de busca de dados, o *middleware*, se possuir o mesmo em sua *document store* irá retorna-lo e então disparar uma tarefa assíncrona para atualizar sua *store*.

A Figura 4.2 ilustra o modelo arquitetural proposto, nela está representada uma abstração da arquitetura da abordagem E-SECO evidenciando a extensão do escopo da arquitetura apresentado no presente trabalho e o relacionamento dela com a arquitetura proposta originalmente para o ecossistema.

4.5 Implementação

Uma vez que o principal objetivo do presente trabalho é a construção de uma rede social utilizando dados providos por terceiros, a primeira etapa de desenvolvimento foi avaliar quais as interfaces disponíveis para consumir as informações desejadas. O DBLP oferece uma *Web API* de Consulta *REST*, na qual é possível consultar todas as informações de um autor, como publicações, coautorias e obtê-las através de um arquivo XML. O Google Scholar, porém, não oferece nenhuma *API* de consulta pública, logo para obtenção de seus dados decidiu-se desenvolver um *web crawler* que obtêm os dados de um determinado pesquisador. Essa foi a primeira e mais desafiadora etapa da implementação.

Para o desenvolvimento do *crawler* adotou-se a linguagem de programação *Python*, uma vez que a mesma facilita a conexão com uma página via requests HTTP. A partir do nome de um autor, faz-se a conexão com o Google Scholar e os dados do pesquisador, como índices, quantidade de citações, publicações e informações gerais podem ser obtidos.

A partir dessa aplicação inicial, desenvolveu-se o *middleware* citado na arquitetura da abordagem Social E-SECO, o mesmo basicamente é uma Web API RESTful que

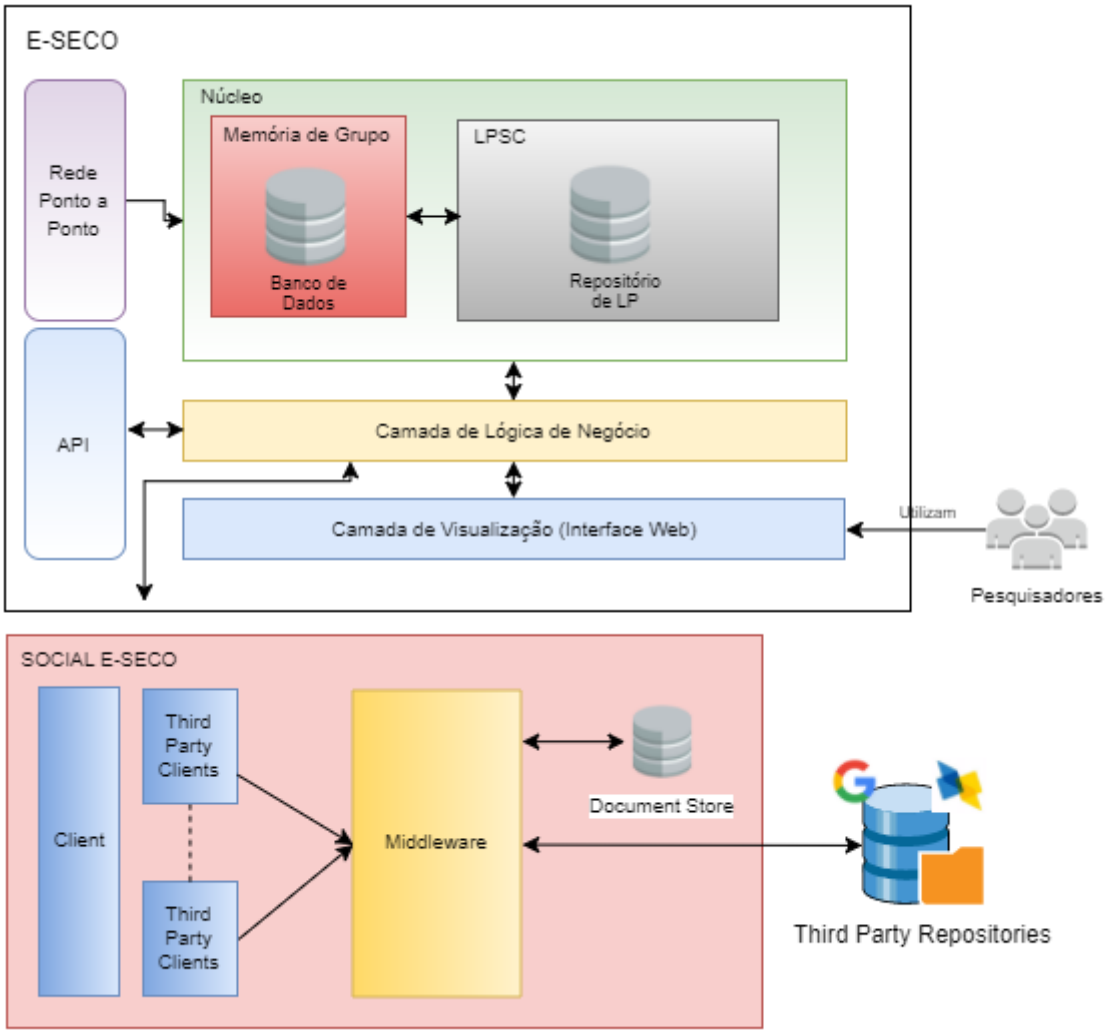

Figura 4.2: Arquitetura da Abordagem Social E-SECO

recebe requisições a respeito de um autor, processa as mesmas e devolve em um formato consolidado, no caso, JSON.

Para a obtenção dos dados DBLP, poderia se pensar em realizar a conexão diretamente, uma vez que a plataforma dispõe de uma API pública para consulta de seus dados, porém, o retorno das pesquisas é em XML. Tal fato iria, já de início, gerar uma discrepância em relação à consulta do Google Scholar, que retorna os dados em JSON, e geraria um cenário caótico para futuras integrações, uma vez que não se definiria um padrão. Devido a tais fatos, decidiu-se consolidar a *Web API Python* como *middleware* para a comunicação entre o E-SECO e os provedores de informação. De forma análoga a trabalhada com o Google Scholar, a API recebe uma requisição de um autor, realiza a consulta na *API* do DBLP, consolida e retorna os dados em JSON.

Um desafio na consulta dos dados em plataformas de terceiros se encontra na confiabilidade. Tais dados podem estar duplicados ou não corresponderem corretamente a realidade. No DBLP, por exemplo, ao se realizar a busca de um autor, informações duplicadas são retornadas. A fim de filtrar as informações, utilizou-se a biblioteca Python *DiffLib SequenceMatcher* (DiffLib, 2006), com a qual é possível "rankear" *string* por sua similaridade, então as mesmas são comparadas com o nome do pesquisador e a quantidade de publicações, a fim de encontrar os dados mais precisos e relevantes.

A tecnologia utilizada para o desenvolvimento da abordagem no E-SECO é a mesma proposta por SOUZA (2015), utiliza-se a linguagem Java, com modelagem MVC e a camada de visualização é desenvolvida com JSF 2.0 e PrimeFaces 4.0. Os dados requisitados por um usuário a partir da camada de visualização são recebidos e gerenciados por um controlador, que irá fazer a solicitação dos dados para o *middleware*. A fim de realizar tal requisição, desenvolveu-se em uma camada de integração dois clientes HTTP que tem o papel somente de fazer o *request* à *Web API Python* e serializar o JSON retornado em classes Java, que serão devolvidas ao controlador para serem exibidas na interface web.

Uma vez que um dos requisitos é a disponibilidade e os dados utilizados pela aplicação são providos por terceiros, decidiu-se pela criação de um banco de dados NoSQL para "cachear" os dados consumidos. Utilizou-se o MongoDB (MongoDB, 2009) para tal, um famoso banco de documentos, que tem como vantagem ser independente de esquema, tornando-o bastante flexível. Com isso é possível armazenar os resultados sem uma estrutura rígida, o que se torna bastante adequado para as intenções da abordagem. Uma vez que as informações guardadas podem se tornar defasadas, foi necessária definir uma política de atualização dos dados, decidiu-se a implementação de uma atualização assíncrona. Toda vez que os dados de um autor são requisitados, existem duas possibilidades: se eles estiverem na base MongoDB, são imediatamente retornados e é disparada uma tarefa assíncrona para realizar a atualização das informações. Se eles não estiverem na base, os mesmos são consultados no Google Scholar e DBLP, retornados e persistidos na base. Tal abordagem demonstrou também como vantagem um aumento na velocidade de resposta das requisições.

Para a implementação da visualização da rede na forma de um grafo, utilizou-se a biblioteca gráfica JavaScript Vis.JS (Vis.JS, 2015). A mesma "renderiza" os dados de coautoria que são retornados pelo middleware e monta um grafo da rede de colaboração do autor.

Uma *cloud tag* foi implementada contendo as palavras mais utilizadas pelo pesquisador no título de suas publicações. Para tal, incorporou-se a biblioteca Java WordCounter (Stoyanr, 2013) ao ecossistema. Pode-se, em trabalhos futuros, utilizar-se de tais palavras-chave para encontrar relacionamentos entre pesquisadores que se interessem pelo mesmo assunto.

Um fluxo padrão de uma requisição ao Social E-SECO pode ser visualizado na Figura 4.3. Inicialmente, um utilizador acessa o E-SECO através do login. Suas informações de acesso são gerenciadas pelo controlador, validando-as no repositório de dados. Após autenticação, o pesquisador tem acesso a todas as funcionalidades do E-SECO, uma vez clicando em "Social", o *controller* repassa as credenciais do pesquisador para um *client* que irá requisitar as informações do usuário ao *middleware*. Uma vez recebidas as informações de um usuário, o *middleware* irá consultar a *document store* a respeito da disponibilidade dos dados. Em caso positivo, os mesmos serão retornados, em caso negativo, o *middleware* realizará consultas em plataformas de terceiros, persistirá os dados na *document store* e os retornará. As informações consolidadas serão disponibilizadas para o usuário na camada de visualização.

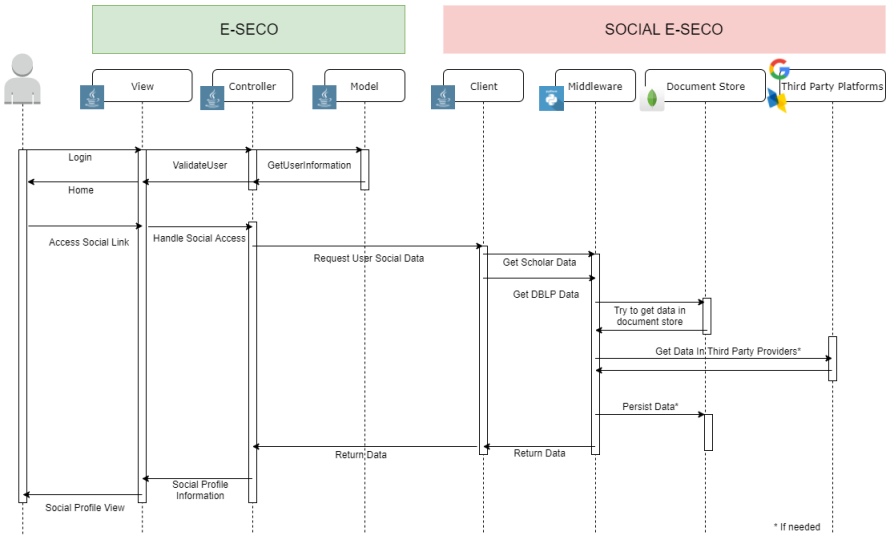

Figura 4.3: Diagrama de Sequencia de uma requisição

4.6 Modelo de Rede de Colaboração

A utilização de representações visuais é muito comum em vários ramos da ciência e é um dos fatores responsáveis pelo desenvolvimento explosivo de toda ciência moderna (FREEMAN, 2000). Para a pesquisa em redes sociais, as visualizações em forma de imagens também são de extrema importância, e desde o início dos estudos desempenham um papel chave provendo aos pesquisadores novos esclarecimentos sobre a estrutura das redes e ajudando-os a compartilhar suas descobertas.

Porém, apesar de muito esforço ter sido despendido no estudo da análise de redes sociais, muitas das vezes, tais estudos não focam na visualização das redes sociais, isso acontece, pois, o tópico é muitas vezes extensivo e bastante complexo. Contudo, sabe-se que a visualização é um passo muito importante na análise de uma rede social, pois permite um melhor entendimento do domínio/ problema representado e, em um cenário científico, provê filtros que auxiliam a simplificar a análise.

A representação mais comum de uma rede social é na forma de um grafo, na qual os nós representam os atores e os vértices representam os relacionamentos entre os mesmos. No presente trabalho, a rede social construída representa um relacionamento de coautoria entre os pesquisadores. Para tal, cada cientista é representado em um nó do grafo, associado ao seu nome, e estão conectados se existe alguma publicação acadêmica em conjunto. Além disso, decidiu-se modelar o peso de tais contribuições, as arestas irão ser mais grossas ou mais finas, de acordo com a quantidade de publicações nas quais os autores foram coautores.

Uma das grandes dificuldades na modelagem da visualização de uma rede social é a complexidade e tamanho dos dados, muitas das vezes as arestas podem acabar estando misturadas e comprometer a visualização. Neste trabalho, decidiu-se fornecer ao pesquisador utilizador da plataforma a escolha de representação de sua rede em até três níveis de indireção, e somente as relações indiretas mais relevantes (com maior número de coautoria) serão representadas visualmente. Isso auxiliará na abstração da dificuldade da representação e estará provendo ao usuário as informações mais importantes. A Figura 4.4 demonstra a modelagem proposta.

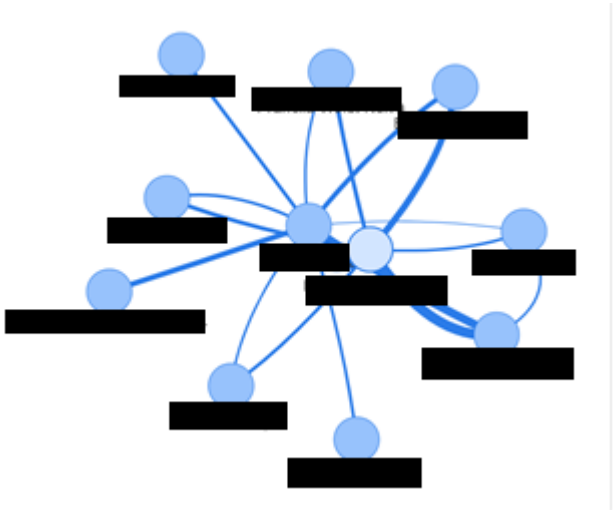

Figura 4.4: Modelagem para materialização da rede social de colaboração

4.7 Exemplo de Uso do E-Seco Social

Na presente seção, será realizado um estudo de caso, exemplificando passo a passo a execução da abordagem proposta a fim de validar o fluxo de funcionamento da arquitetura.

1. O cientista irá realizar o acesso na plataforma E-SECO através da página de Login como já era realizado anteriormente na abordagem proposta por Souza, 2015. Após inserir suas credenciais de acesso, o pesquisador será redirecionado para a página inicial do E-SECO, na qual terá uma opção na Barra de Menu para acessar o ambiente Social E-SECO, conforme ilustra a Figura 4.5.

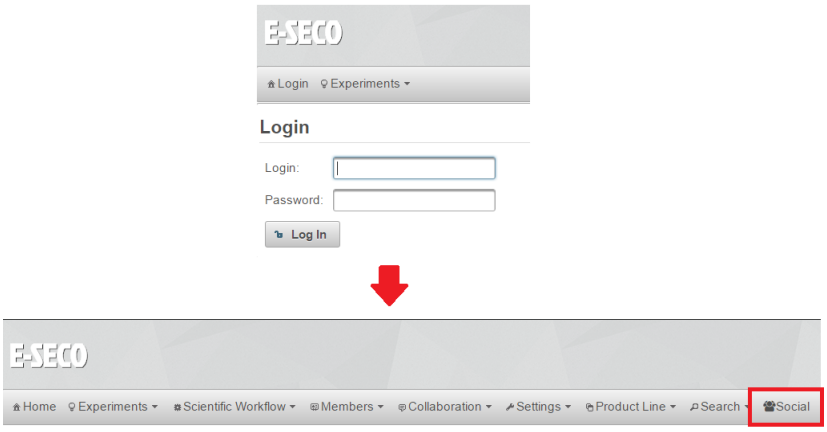

Figura 4.5: Login e Página Inicial do E-SECO

2. Ao requisitar acesso ao ambiente, o sistema irá realizar requisições GET para o mid-

dleware apresentado na arquitetura, que por sua vez, realizará as buscas e retornará os resultados em formato JSON. O retorno do Google Scholar está apresentado na Figura 4.6, na qual se percebe as informações de um autor, informações gerais e específicas de uma publicação. O retorno do DBLP está apresentado na Figura 4.7, na qual percebe-se as informações de um autor, suas publicações e seus coautores.

Figura 4.6: Retorno do Google Scholar. a) Informações do autor; b) Informações gerais de publicações, c) Informações específicas de uma publicação

3. Após retorno da requisição, todos os dados são apresentados para o pesquisador em sua página de perfil, ilustrada na Figura 4.8. Cada container será detalhado em seguida.

   (a) O primeiro container contém informações básicas sobre o pesquisador, como sua foto, nome, e-mail, quantidade de citações, quantidade de coautorias e principais interesses.

   (b) O segundo container possui todas as publicações do pesquisador, ordenadas por ordem de maior quantidade de citações. É possível clicar na lupa para obter maiores informações sobre o artigo e também realizar o download do mesmo,

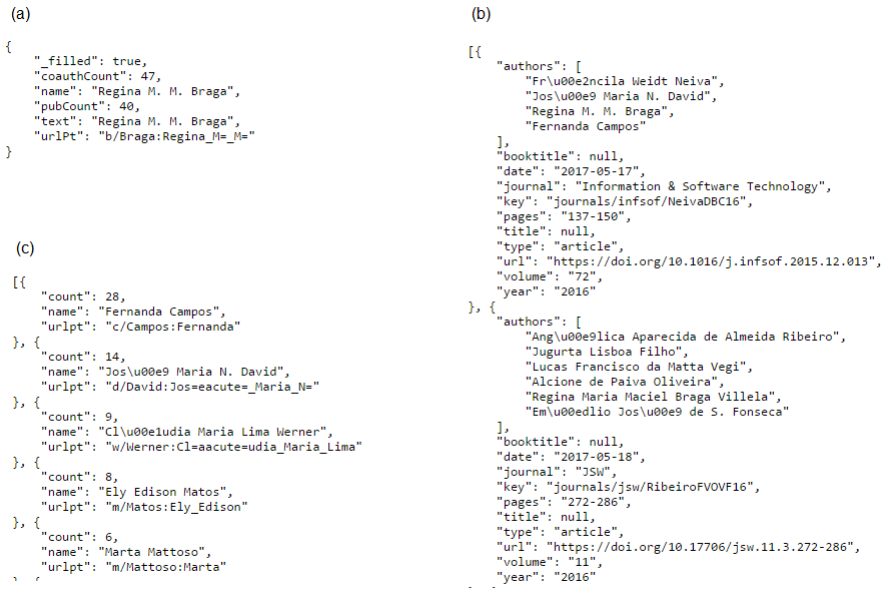

Figura 4.7: Retorno do DBLP. a) Informações do Autor; b) Informações gerais de publicações; c) Informações de coautores

quando disponível, como mostra a Figura 4.9.

(c) O terceiro container contém uma Cloud Tag com as palavras mais utilizadas nos títulos das publicações do autor. Futuramente, podem ser úteis para encontrar relações entre autores pesquisando assuntos relacionados.

(d) O quarto container possui as métricas apresentadas na seção 4.3. Separadas em duas categorias: Geral e Desde 2012.

(e) O quinto container possui a lista de coautores do pesquisador ordenados por maior número de contribuições, que está representada entre parênteses após o nome do autor colaborador.

(f) O sexto contêiner contém a materialização da rede de colaboração. Ela apresenta os pesquisadores representados como nós de um grafo. O relacionamento entre eles é representado pelas arestas, que possuem espessuras diferentes, representando a quantidade de coautorias entre os dois atores envolvidos. O pesquisador poderá escolher em até 2, o nível da materialização, a fim de analisar o alcance de seus relacionamentos.

4. Clicando em uma publicação, o pesquisador tem acesso a maiores informações sobre a mesma (Figura 4.9), como: título, autores, data, número de páginas, resumo,

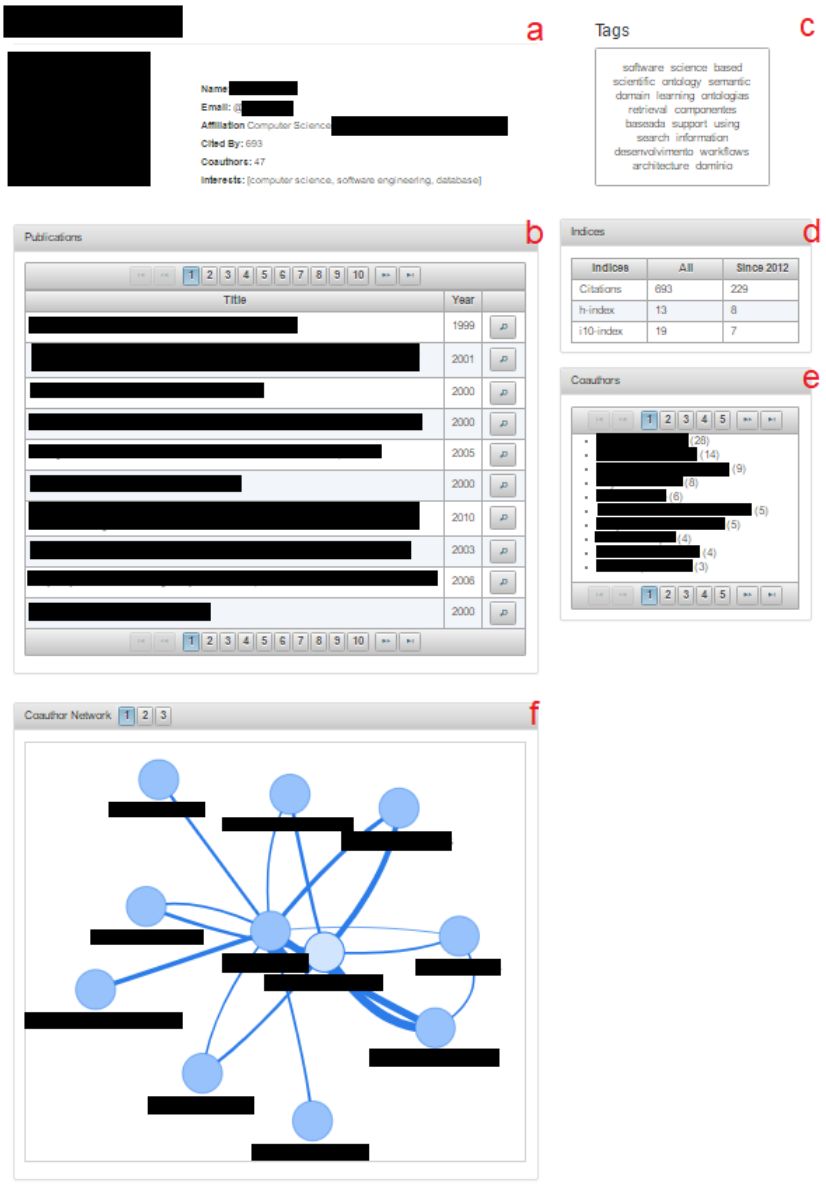

Figura 4.8: Página Inicial do E-SECO

quantidade de citações e também, se disponível, download da publicação.

5. Alterando o nível da materialização da rede social, o pesquisador aumentará ou reduzirá a abrangência da visualização de sua rede de coautoria, como está representado na Figura 4.10.

4.8 Evolução da Abordagem

A abordagem proposta no presente trabalho é um passo inicial na integração de redes sociais no E-SECO. Foi incluída uma rede de coautoria, que apesar de ser um

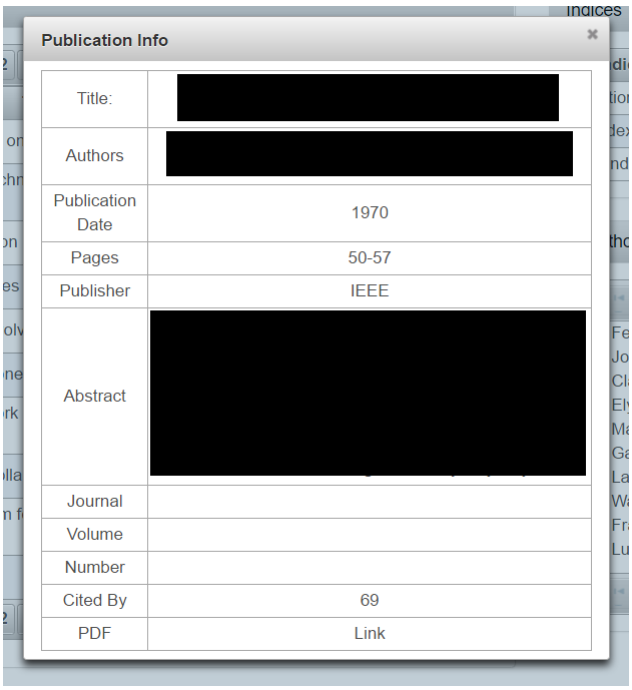

Figura 4.9: Informações detalhadas de uma publicação

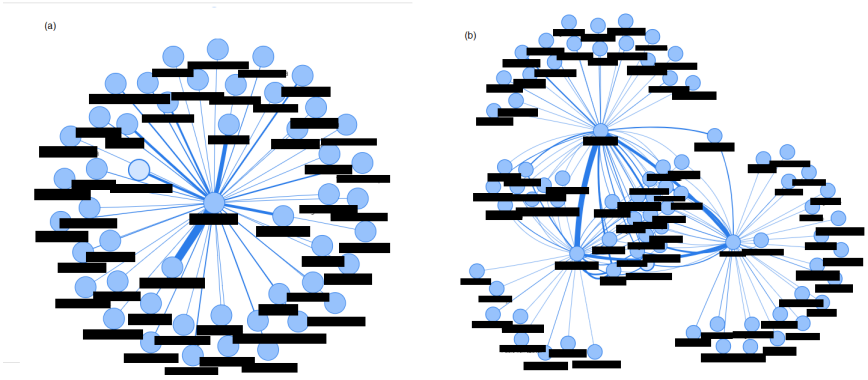

Figura 4.10: Materialização da rede em diferentes níveis. a) Um nível de associação b) Dois níveis de associação

dos relacionamentos acadêmicos mais importantes, não é o único. A abordagem deve ser estendida, passando a rede social de homogênea para heterogênea, enriquecendo-a e possibilitando que os dados representem de melhor forma o mundo real.

Além disso, trabalhos futuros podem ter a abordagem Social E-SECO como base. Pode-se aplicar diversas técnicas de grafos e redes sociais a fim de extrair mais informações relevantes. Entre elas, pode-se citar: a análise dos dados utilizando algoritmos de menor caminho em grafos, a fim descobrir novos relacionamentos entre pesquisadores; a análise do fluxo da rede, com o intuito de elucidar como a informação flui através da mesma, e levantar quais os atores estão mais fortemente relacionados. Com isso, é possível aplicar

conceitos de clusterização, e estudar a rede em diferentes níveis de granularidade, entender como pesquisadores se relacionam dentro de um mesmo departamento, como universidades se comunicam umas com as outras, como cientistas ou grupo de cientistas se relacionam com os de outros países, entre várias possibilidades. Trabalhos futuros podem também incluir uma funcionalidade de pesquisa de autor, com a qual o utilizador pode descobrir como alcançar um outro pesquisador e os melhores caminhos para tal.

É possível também realizar uma análise temporal da rede, e levantar como os relacionamentos foram construídos ou até mesmo destruídos com o passar do tempo. Além disso, pode-se levar em consideração essa informação temporal para compor o peso dos relacionamentos, uma vez que coautorias que tenham mais tempo, mesmo em maior quantidade, podem não ser relevantes.

5 Considerações Finais

Este capítulo tem como objetivo analisar as contribuições deste trabalho em frente aos requisitos inicialmente propostos, bem como suas limitações e trabalhos futuros.

5.1 Contribuições

Analisando os requisitos funcionais e não funcionais propostos no capítulo anterior considera-se que o presente trabalho conseguiu atingi-los em sua maioria. Quanto aos requisitos funcionais, foi possível extrair dados de um pesquisador de plataformas de terceiros, consolida-los em um *middleware* e disponibilizar tais informações para o utilizador da plataforma E-SECO em uma página de perfil social. Apesar de inicialmente ter enfrentado dificuldades para a obtenção dos dados do Google Scholar, a decisão de construção de um *web crawler* se mostrou conveniente. A despeito de demandar um maior custo quanto ao tempo de desenvolvimento, a solução final atendeu completamente às necessidades. Com a materialização da rede social, disponibilizou-se para o cientista de forma simples e gráfica os relacionamentos que o mesmo possui em relação a autoria e o alcance de seus relacionamentos. Com isso, espera-se que as colaborações diretas sejam estimuladas através da análise da rede, encontrando pesquisadores que estudam temas relacionados.

Em relação aos requisitos não funcionais, conseguiu-se atingir um grande nível de flexibilidade e extensibilidade através do desenvolvimento do *middleware* proposto, uma vez que ele centraliza a comunicação com serviços externos e deixa isso transparente para o E-SECO. Com isso, se torna simples plugar novas plataformas e estender a abordagem proposta. A maior ameaça ao presente trabalho se encontrava na disponibilidade, uma vez que tal requisito poderia comprometer a experiência do usuário tanto em velocidade quanto em visualização dos dados. A solução de persistir as informações recuperadas em um *document store* se mostrou bastante conveniente, uma vez que a velocidade de recuperação dos dados aumentou consideravelmente e, como as informações do pesqui-

sador estavam "cacheadas" nesse banco, diminuiu-se o acoplamento que se tinha com os provedores de dados.

A maior contribuição do presente trabalho está na possibilidade de colaboração direta entre os pesquisadores. Com a visualização da rede se torna possível descobrir possibilidades de colaborações que não se tinha conhecimento, porém ficaram evidenciadas com a plataforma. Alguns trabalhos da literatura já exploraram o conceito de redes sociais aplicadas ao *e-Science*, porém a integração a um ECOS, realizada no presente trabalho evidenciou todos os ganhos que uma análise e visualização de redes sociais pode trazer no âmbito de colaborações.

Além disto, o presente trabalho evidencia a extensibilidade desejada da abordagem E-SECO, uma vez que se mostrou simples e orgânico o desenvolvimento e integração de uma nova *feature* à plataforma.

5.2 Limitações

Durante o processo de desenvolvimento da plataforma, bem como durante a levantamento dos requisitos, ficou claro que a parte mais frágil da arquitetura proposta estava na comunicação de provedores de informação gerenciados por terceiros. Logo, as maiores limitações do sistema se encontram em torno desse fato.

Uma das limitações está na necessidade de o pesquisador ter suas informações disponíveis tanto no Google Scholar quanto no DBLP para que a plataforma possa exibir seus dados. Outra limitação se encontra na disponibilidade dos dados, uma vez que mesmo com a persistência das informações em um *document store*, nem sempre pode-se garantir que os dados para uma determinada consulta serão retornados. Tais limitações abrem brecha para trabalhos futuros na plataforma que serão discutidos na seção 5.3. Outra limitação que deve ser mencionada se encontra na consulta de informações, uma vez que as fontes de de dados podem trazer informações duplicadas. Utilizou-se algoritmos de comparações de *strings* para tentar mitigar esse fato e realizar um filtro.

5.3 Trabalhos Futuros

Como levantado na seção anterior, uma das limitações da abordagem se encontra na disponibilização dos dados, quanto a isso pode-se citar como trabalhos futuros:

- Integração de novas plataformas a fim de garantir uma maior quantidade e qualidade dos dados disponíveis;
- Construção de uma base de dados local a fim de se ter controle sobre a disponibilidade das informações. Pode-se utilizar a abordagem proposta de guardar as informações que foram requisitadas, juntamente com a construção incremental de uma base a ser cadastrada pelo utilizador via plataforma E-SECO;

Além disso, uma outra ramificação de trabalhos futuros se apresenta a partir das redes socais. O presente trabalho representa uma etapa inicial de integração da rede, a mesma pode, e deve ser ampliada. É possível citar como trabalhos futuros:

- Análise dos relacionamentos de um cientista a fim de sugerir automaticamente novas possíveis colaborações. Para isso, pode-se utilizar a rede de autorias em conjunto com as tags de um pesquisador para encontrar caminhos mais curtos entre pesquisadores que estudem assuntos relacionados.
- Construção e análise de *clusters* dentro da rede em diferentes níveis de granularidade. Com isso, pode-se analisar como os pesquisadores se relacionam dentro de um departamento, dentro de uma universidade ou até mesmo analisar os relacionamentos entre as organizações.
- Análise do fluxo da rede, com o objetivo de descobrir os relacionamentos mais fortes e relevantes entre os pesquisadores e *clusters*.
- Extensão da rede construída para uma rede heterogênea, agregando mais fatores além da coautoria e tornando-a mais rica e correspondente a realidade.
- Análise temporal dos relacionamentos, com isso pode-se entender como os relacionamentos foram construídos e relacionados ao longo do tempo.

Referências Bibliográficas

Belloum, A.; Inda, M. A.; Vasunin, D.; Korkhov, V.; Zhao, Z.; Rauwerda, H.; Breit, T. M.; Bubak, M. ; Hertzberger, L. O. Collaborative e-science experiments and scientific workflows. **IEEE Internet Computing**, v.15, n.4, p. 39–47, 2011.

Bosch, J. **From software product lines to software ecosystems**. In: Proceedings of the 13th international software product line conference, p. 111–119. Carnegie Mellon University, 2009.

Costa, G. C. B.; Braga, R.; David, J. M. N. ; Campos, F. A scientific software product line for the bioinformatics domain. **Journal of biomedical informatics**, v.56, p. 239–264, 2015.

Costa, G. C. B. **Uma abordagem para linha de produtos de software científico baseada em ontologia e workflow**. 2013. Dissertação de mestrado - Universidade Federal de Juiz de Fora.

Dblp. http://dblp.uni-trier.de. Acesso em: 16-06-2017.

Difflib. https://docs.python.org/2/library/difflib.html. Acesso em: 16-06-2017.

Ellis, C. A.; Gibbs, S. J. ; Rein, G. Groupware: some issues and experiences. **Communications of the ACM**, v.34, n.1, p. 39–58, 1991.

Evans, E. **Domain-driven design: tackling complexity in the heart of software**. Addison-Wesley Professional, 2004.

Freeman, L. C. Visualizing social networks. **Journal of social structure**, v.1, n.1, p. 4, 2000.

Fuks, H.; Raposo, A. B.; Gerosa, M. A. ; Lucena, C. J. P. Do modelo de colaboração 3c à engenharia de groupware. **Simpósio Brasileiro de Sistemas Multimídia e Web–Webmidia**, p. 0–8, 2003.

Google scholar. https://scholar.google.com. Acesso em: 16-06-2017.

Hannay, J. E.; MacLeod, C.; Singer, J.; Langtangen, H. P.; Pfahl, D. ; Wilson, G. **How do scientists develop and use scientific software?** In: Software Engineering for Computational Science and Engineering, 2009. SECSE'09. ICSE Workshop on, p. 1–8. Ieee, 2009.

Hine, C. **New infrastructures for knowledge production: Understanding e-science**. IGI Global, 2006.

Hirsch, J. E. An index to quantify an individual's scientific research output. **Proceedings of the National academy of Sciences of the United States of America**, p. 16569–16572, 2005.

Jansen, S.; Finkelstein, A. ; Brinkkemper, S. **A sense of community: A research agenda for software ecosystems**. In: Software Engineering-Companion Volume, 2009. ICSE-Companion 2009. 31st International Conference on, p. 187–190. IEEE, 2009.

Json. http://www.json.org. Acesso em: 16-06-2017.

Khabsa, M.; Giles, C. L. The number of scholarly documents on the public web. **PloS one**, v.9, n.5, p. e93949, 2014.

Kraut, R. E.; Galegher, J. ; Egido, C. Relationships and tasks in scientific research collaboration. **Human–Computer Interaction**, v.3, n.1, p. 31–58, 1987.

Ley, M. Dblp: some lessons learned. **Proceedings of the VLDB Endowment**, v.2, p. 1493–1500, 2009.

Maxville, V. **Preparing scientists for scalable software development**. In: Proceedings of the 2009 ICSE Workshop on Software Engineering for Computational Science and Engineering, p. 80–85. IEEE Computer Society, 2009.

Patterns, D.; Pattern, C. **Model-view-controller**, 2003.

Mongodb. https://www.mongodb.com. Acesso em: 16-06-2017.

Nardi, A. R. **Uma arquitetura de baixo acoplamento para execução de padrões de controle de fluxo em grades**. 2009. Tese de Doutorado - Universidade de São Paulo.

Newman, M. E. Scientific collaboration networks. i. network construction and fundamental results. **Physical review E**, v.64, n.1, p. 016131, 2001.

Newman, M. E. The structure of scientific collaboration networks. **Proceedings of the National Academy of Sciences**, v.98, n.2, p. 404–409, 2001.

Perrey, R.; Lycett, M. **Service-oriented architecture**. In: Applications and the Internet Workshops, 2003. Proceedings. 2003 Symposium on, p. 116–119. IEEE, 2003.

Pereira, A. F. **Collaborative pl-science: Utilizando elementos de colaboração em uma linha de produtos de software científico**. Juiz de Fora, MG, Brasil, Julho, 2014 2014. Dissertação de mestrado - Universidade Federal de Juiz de Fora.

Pereira, A. F.; Braga, R.; Campos, F. ; others. **An architecture to enhance collaboration in scientific software product line**. In: 2016 49th Hawaii International Conference on System Sciences (HICSS), p. 338–347. IEEE, 2016.

Pham, T. V. **A Collaborative e-Science architecture for distributed scientific communities**. 2006. Tese de Doutorado - The University of Leeds.

Pitas, I. **Graph-based social media analysis**. Chapman and Hall/CRC, 2015.

Segal, J.; Morris, C. Developing scientific software. **IEEE Computer Society, IEEE Software**, v.25, n.4, p. 18–20, July-Aug 2008.

Shneiderman, B. Copernican challenges face those who suggest that collaboration, not computation are the driving energy for socio-technical systems that characterize web 2.0. **Science**, v.319, n.5868, p. 1349–1350, 2008.

Sirqueira[1], T. F.; Dalpra[1], H. L.; Braga[1], R.; Araújo[1], M. A.; David[1], J. M. N. ; Campos[1], F. E-seco proversion: Manutenção e evolução de experimentos científicos. **Anais do XXXVI Congresso da Sociedade Brasileira de Computação: Computaçãoeinterdisciplinaridade.**, 2016.

Souza, N. B. O. d.; others. **Caracterização de software científico: um estudo de caso em modelagem computacional**. Dissertação de Mestrado - .

Souza, V. F.; others. **Ecos pl-science: Uma arquitetura para ecossistemas de software cient´ ıfico apoiada por uma rede ponto a ponto**. Dissertação de Mestrado - .

StröEle, V.; ZimbrãO, G. ; Souza, J. M. Group and link analysis of multi-relational scientific social networks. **Journal of Systems and Software**, v.86, n.7, p. 1819–1830, 2013.

Stoyanr wordcounter. https://github.com/stoyanr/Wordcounter. Acesso em: 16-06-2017.

Vis.js. http://visjs.org. Acesso em: 16-06-2017.

Wasserman, S.; Faust, K. **Social network analysis: Methods and applications**, volume 8. Cambridge university press, 1994.

Zhao, Z.; Belloum, A. ; Bubak, M. Special section on workflow systems and applications in e-science. **Future Generation Computer Systems**, v.25, n.5, p. 525–527, 2009.