Utilizando Ontologias para apoiar a recomendação de revisores de código entre projetos distintos

Vinicius J. Schettino
Universidade Federal de Juiz de Fora
Juiz de Fora, Minas Gerais
vinicius.schettino@ice.ufjf.br

Regina Braga
Universidade Federal de Juiz de Fora
Juiz de Fora, Minas Gerais
regina.braga@ufjf.edu.br

Marco Antônio P. Araújo
Universidade Federal de Juiz de Fora
Juiz de Fora, Minas Gerais
marco.araujo@ufjf.edu.br

Victor Ströele
Universidade Federal de Juiz de Fora
Juiz de Fora, Minas Gerais
victor.stroele@ufjf.edu.br

ABSTRACT

A prática de *code review* está amplamente associada a qualidade de código e detecção precoce de defeitos de software. Intrinsecamente colaborativo, a eficiência do processo depende também do perfil histórico do revisor e da natureza do código em avaliação. Em contexto de desenvolvimento distribuído, consolidar todas as informações que ajudam a encontrar o revisor adequado se torna ainda mais difícil, e a importância da recomendação automatizada ganha mais força. Dentre outros desafios, a arquitetura e habilidades necessárias entre projetos são distintas, atrapalhando a troca de experiências e auxílios entre diferentes grupos de trabalho. Para auxiliar no processo de escolha do revisor, uma ontologia de alinhamento é proposta, de forma a relacionar os tópicos de conhecimento entre projetos distintos. Assim é possível encontrar especialistas que possam ser revisores adequados em diferentes grupos de tabalho. Este estudo apresenta a arquitetura responsável pela extração dos dados e aplicação da ontologia proposta, exemplificado por um caso de uso envolvendo projetos OpenSource conhecidos de uma mesma organização.

CCS CONCEPTS

•**Computer systems organization** →**Embedded systems;** *Redundancy;* Robotics; •**Networks** →Network reliability;

KEYWORDS

ACM proceedings, LaTeX, text tagging

ACM Reference format:
Vinicius J. Schettino, Regina Braga, Marco Antônio P. Araújo, and Victor Ströele. 2019. Utilizando Ontologias para apoiar a recomendação de revisores de código entre projetos distintos. In *Proceedings of Simpósio Brasileiro de Sistemas de Informação, Aracaju/SE - Brasil, Maio 2019 (SBSI'19),* 8 pages.
DOI: xxxxxxx

SBSI'19, Aracaju/SE - Brasil

DOI: xxxxxxx

1 INTRODUCTION

O *code review* é considerado uma das principais técnicas para diminuição de defeitos de software [8]. Nela, o autor de uma alteração na base de código de um projeto submete tal conteúdo ao crivo de um conjunto de pares técnicos, que irão revisar sua estrutura com base em um lista de regras e convenções previamente definida. Diferentes aspectos relacionados ao autor, ao revisor e ao processo de revisão em si estão diretamente relacionados à eficiência da prática. Autores relatam a diminuição da incidência de *anti-patterns* [21] de acordo com o nível de participação dos envolvidos e cobertura do código revisado [4, 23, 27]. Reputação [5, 10] e experiência [22] do revisor também parecem impactar nos efeitos do *code review*

Intrinsecamente colaborativa, a atividade de *code review* é exercida com suporte de ferramentas computacionais específicas [3], principalmente no desenvolvimento distribuído. Dentro de workflows de trabalho descentralizados [16], a prática funciona como um *gateway* de qualidade que busca garantir que apenas alterações aderentes aos padrões de qualidade do projeto serão incorporados à codebase principal. Esta etapa do desenvolvimento se torna uma oportunidade para disseminação de conhecimento, embate de ideias e discussão de melhores práticas entre profissionais de experiência e visões diferentes. Para tanto, percebe-se a necessidade de suporte computacional para essas atividades colaborativas.

Tais aspectos configuram o Desenvolvimento Distribuído de Software (DDS), onde equipes de desenvolvimento se encontram espalhadas por organizações e espaços geográficos distintos. Este novo ramo da Engenharia de Software vem modificando a relação entre empresas e sistemas, principalmente em relação às estratégias de negócios [2]. As próprias relações de negócios fomentam a distribuição das equipes, procurando diminuição dos custos e a incorporação de mão de obra qualificada que pode estar em qualquer lugar do planeta.

Estes desafios do Desenvolvimento Distribuído de Software afetam o *code review* de duas formas distintas. Primeiro, o processo de revisão pode se tornar lento e ineficiente quando a colaboração é afetada, devido aos baixos níveis de participação e cobertura. O mesmo vale para a disseminação do conhecimento, que fica prejudicada. Outro desafio que se consolida é a escolha do revisor adequado para aquele *patch*. Com um vasto número de opções e

pouca informação disponível sobre seus aspectos técnicos e gerenciais (e.g. tempo disponível) já que não há contato co-localizado entre eles, a natureza distribuída deste tipo de desenvolvimento dificulta o processo de escolha do revisor, impactando negativamente a eficiência do processo.

Existem alguns trabalhos congêneres que demonstram métodos de recomendação de revisores [20, 33, 37]. Esses trabalhos foram estudados e levados em consideração para escrita do presente texto. Também foram revisadas pesquisas que apontam características de revisões, revisores e autores que possivelmente potencializam a colaboração [5, 7, 21].

Neste contexto, porém, os os desafios à colaboração co-localizada são potencializados e as soluções tradicionais não são suficientes para fomentar este aspecto das atividades distribuídas [28]. Casey [11] mostra que, com a distribuição geográfica dos times, diversos outros desafios, antes considerados colaterais ou resolvidos, emergem de forma a ameaçar a colaboração entre os membros da equipe: barreiras culturais, temporais e geográficas; reengenharia dos processos de desenvolvimento; resistência em compartilhar informações e conhecimento com os pares distribuídos; entre outros desafios.

Estes utilizam dados de expertise e reputação dos desenvolvedores para apontar os mais adequados para determinado conhecimento. São utilizados dados de proximidade de código ou de semântica com as revisões anteriores. Abordagens mais recentes de recomendação de revisores envolvem informações de colaboração, como por exemplo impacto do trabalho entre um revisor/autor no passado. Estas relações são geralmente modeladas como redes de colaboração e construídas através de análises semânticas de similaridade e frequência de colaboração entre determinados atores. Fu et al. [14] descreve uma recomendação baseada num grafo de relacionamento entre desenvolvedores e seu status de trabalho no projeto. Xia et al. [34] agrupa desenvolvedores utilizando métodos de proximidade para capturar relações implícitas e explícitas. Yu et al. [36, 37] estendem essa abordagem com métodos de aprendizado de máquina minerando os comentários para extrair metadados e enriquecer o modelo da rede colaborativa. Yang et al. [35] utilizam social network analysis (SNA) baseadas em informações de atividade recente dos potenciais revisores no projeto.

Contudo, é comum que tecnologias como linguagens, bibliotecas e frameworks sejam utilizados em projetos diferentes, fazendo com que a expertise de um desenvolvedor construída em um projeto o faça um potencial revisor adequado para contribuir em outros componentes. Os trabalhos em geral não focam neste tipo de recomendação entre projetos distintos.

Este tipo de auxílio se faz importante dentro de organizações maiores. É o caso da Apache Foundation, com dezenas de projetos com diferentes linguagens, arquiteturas e propósitos. É possível que desenvolvedores que trabalham em projetos específicos tenham conheciementos aplicáveis em projetos distintos, apesar de sempre ter contribuído em apenas um deles. A quantidade de projetos, desenvolvedores e de características intrínsecas de cada um destes é uma barreira para detecção e recomendação de especialistas entre repositórios distintos.

O uso de ferramentas computacionais para o processo de revisão de código se tornou prática comum nos últimos anos [3]. O GitHub é uma plataforma rica em repositórios de projetos de software. Muitos são de código aberto, disponíveis para mineração. São 24 milhões de usuários, 67 milhões de projetos e 47 milhões de revisões[1], também chamadas de *pull requests* no modelo de desenvolvimento "*pull based*" [15].

Tabela 1: Tabela do Template (para comparar)

Non-English or Math	Frequency	Comments
uai	1 in 1,000	At oprette
π	1 in 5	Common in math
\$	4 in 5	Used in business
Ψ_1^2	1 in 40,000	Unexplained usage

Esta característica permitiu a extração e análise automatizadas das informações sobre as revisões em projetos de código aberto, através de APIs disponibilizadas para este fim. Foram extraídas métricas apontadas como relevantes para nossos objetivos pela literatura relacionada.

Assim, o objetivo deste estudo é extender a utilização de técnicas de recomendação de revisores para atender demanda entre projetos distintos, baseado em tópicos de conhecimento e especialidade dos desenvolvedores. A ferramenta pode ser vista como um indicador de "projetos sugeridos"para desenvolvedores que buscam novos repositórios para contribuir. A proposta é que, através da mineração dos repositórios e do uso de uma ontologia para realizar a equivalência entre os tópicos de conhecimento de projetos distintos seja possível intercambiar potenciais revisores.

Ontologia é a definição de um vocabulário ou linguagem comum para representar conhecimentos [17]. Para computação, pode ser vista como uma especificação formal, explícita e não ambígua de um conceito compartilhado por diferentes agentes.

Uma ontologia é composta por: Definições de classes; Relações e Funções. Se tornaram comuns para representação de conhecimento na Web Semântica [6]. As especificações da ontologia permitem a utilização de máquinas de inferências para descoberta de conhecimentos, antes não explícitos, nas informações disponíveis. Estas máquinas inferem propriedades e relacionamentos baseadas em conhecimentos diretamente representados [6].

[ORGANIZAÇÃO DO TRABALHO AQUI]

2 TRABALHOS RELACIONADOS

Existem diversos trabalhos que visam recomendar revisores, alguns já citados em seções anteriores deste texto [14, 20, 33–37]. O foco desta seção é apresentar trabalhos relacionados a recomendação entre projetos distintos e o uso de ontologias para alinhamento de semânticas.

CoRReCT [29] é uma ferramenta que também utiliza o histórico dos potenciais revisores para predição. Contudo, as informações em questão tangem a história dele como desenvolvedor. O especialista é definido pelas tecnologias com as quais ele trabalhou no passado, definidas como tópicos de conhecimento. A principal inovação da proposta é utilizar o histórico de todas contribuições OpenSource feitas pelo desenvolvedor.

A abordagem é similar ao proposto por Mo et al. [26].

[1] https://octoverse.github.com/

A principal inovação da proposta aqui descrita é o uso de ontologias para recomendar especialistas entre projetos distintos. A utilização de ontolgias para recomendação é uma técnica já explorada em trabalhos anteriores [24, 25] e indicada como sugestão para trabalhos futuros [1], e por isso é um potencial caminho para otimização dos resultados da recomendação.

3 EXTRAÇÃO DOS DADOS

Os dados da análise são extraídos do GitHub através da API[2] disponibilizada para desenvolvedores do mundo todo estenderem as funcionalidades da plataforma e integrar novos produtos que agregem valor ao processo de desenvolvimento. A versão estável para integração foi a v3, que implementa o padrão arquitetural RESTful[12]. Existe a API em recém lançada v4[3], que utiliza da especificação GraphQL para exposição dos serviços. Esta não foi utilizada devido a documentação ainda recente e pela preferência ao padrão REST. A figura 1 mostra o pipeline de extração e estruturação dos dados.

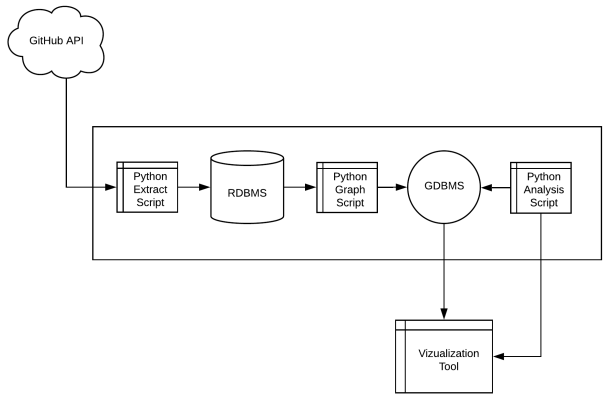

Figura 1: Processo de extração dos dados

Os dados extraídos são estruturados em um esquema relacional, mais especificamente em uma instância PostgreSQL 10.5. Este passo intermediário ocorre para que os dados fiquem disponíveis para outras análises, pesquisadores e ferramentas familiarizadas com o modelo relacional. Além disso, esta abordagem permite a extração de diversos metadados que não serão utilizados nas análises preliminares executadas na estrutura baseadas em grafo, ficando disponíveis para consultas posteriores sem degradar a performance do modelo em análise. A figura 2 mostra o modelo relacional proposto. As entidades *user* são normalizadas entre os diferentes projetos. Ou seja, o usuário é uma entidade única que se relaciona a diferentes projetos, sem duplicação de seu conteúdo. Esta caracterísitca facilita futuras análises de colaboração entre projetos distintos. As *labels* ligadas aos *pullrequests* são utilizadas como descritores semânticos do conteúdo da contribuição, como por exemplo *documentação*, *python* ou *performance*. As contribuições podem ser feitas através de comentários thread de discussão do *pullrequests*, representados pelos *discussioncomments* ou através de comentários no código em revisão, nos *reviewcomments*.

O algortimo de extração dos dados foi projetado com uma série de cuidados para garantir a integridade e disponibilidade das informações, diante das nuances do repostiório e do domínio de dados. Os principais aspectos observados são:

Inconsistência nos dados: Para evitar dados inconsistentes, toda a extração é feita dentro de uma transaction relacional do SGBD, garantindo que todas as ações serão executadas (ou canceladas) de forma atômica. As restrições de chave primária também são a primeira linha de defesa para evitar que informações inconsistentes sejam armazenadas (como por exemplo um comentário relacionado a um pullrequest que não existe).

Tratamento das exceções da API: . O repositório de dados pode se comportar de maneira anormal durante a extração dos projetos, por diversos motivos: indisponibilidade dos serviços, lentidão do cliente, problemas de conexão entre outros. Assim é necessário tratar erros (como 400, 500 e 502) e evitar que o dump seja interrompido. Para isso, as respostas são tratadas, e detectados erros temporários como estes, o request é agendado para se repetir em uma janela de tempo definida.

Ratelimit da API: . A API do GitHub limita o número de requests a 5000/hora. Por isso, é necessário que o cliente controle o número de requests para que não seja bloqueado. Assim, o componente de extração limita o seu número de requests, entrando em espera quando o número se aproxima ao limite imposto.

Unicidade de usuários: Usuários são entidades espalhadas entre os diferentes projetos. Para possibilitar análises globais e estatísticas realísticas de tamanhos de projetos, é necessário que não haja duplicação entre os usuários. O componente de extração também garante que não hajam usuários duplicados.

3.1 Aspectos de Reprodutibilidade

Cada uma das tecnologias apresentadas é encapsulada através da tecnologia de virtualização Docker. esta abordagem é uma resposta às iniciativas de reprodutibilidade na ciência, buscando maior transparência, confiabiliade e possibilidade de extensão nos experimentos [13]. Os três componentes (banco de dados relacional, orientado a grafo e os scripts de extração, transformação e carga) são encapsulados em contâiners distintos, como mostra a figura 3 Estes são orquestrados através do docker-compose, que configura os parâmetros e acessos necessários para o funcionamento do sistema sem que o usuário precise realizar nenhuma ação extra.

Com um conjunto grande de dependências, tecnologias e minúncias que estão envolvidas neste tipo de experimento, o código fonte e a descrição ainda que detalhada dos dados não é suficiente para alcançar níveis adequados de reprodutibilidade [19]. Com auxílio dos containers Docker, é possível criar instâncias executáveis dos experimentos que vão funcionar em diferentes computadores, arquiteturas e situações, sem necessidade de conhecimento técnico por parte do executor das tecnologias utilizadas [9].

De acordo com Sinha et al. [32], a maturidade da reprodutibilidade de um experimento científico computacional pode ser medido

[2]https://developer.github.com/v3/
[3]https://developer.github.com/v4/

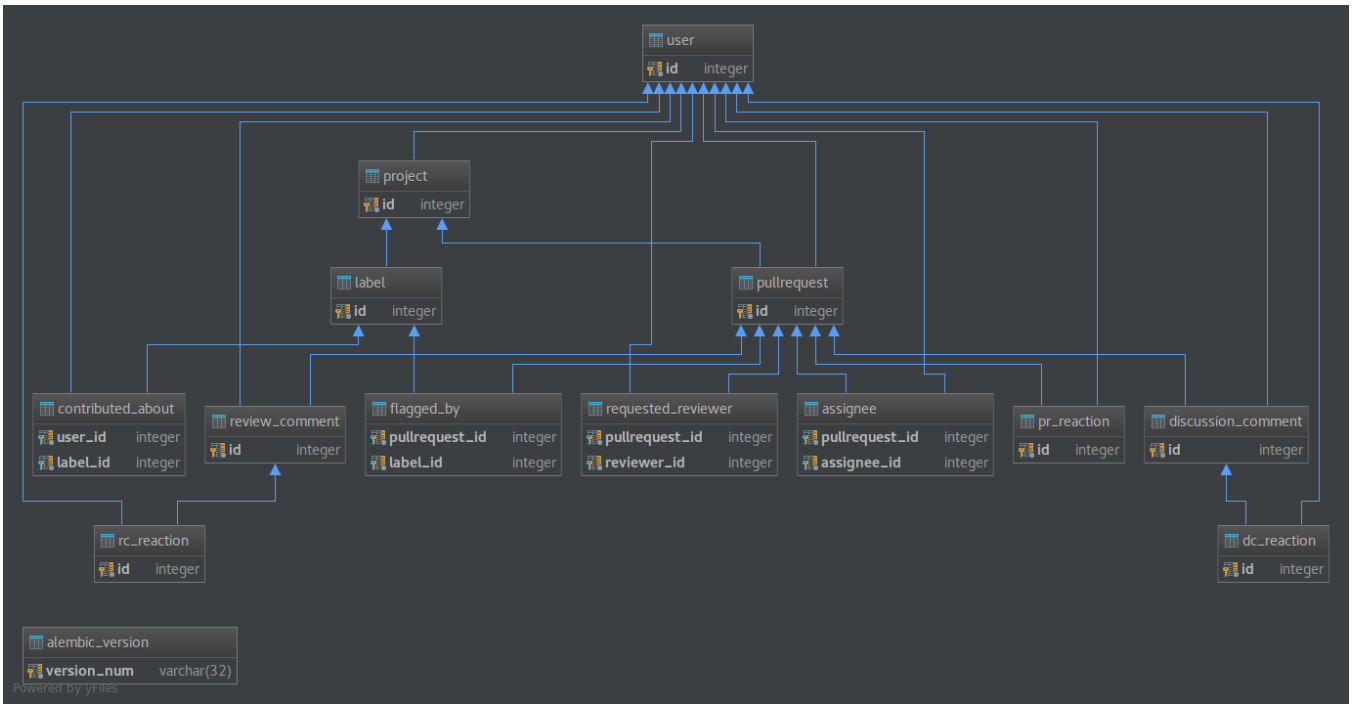

Figura 2: Modelo Entidade Relacionamento

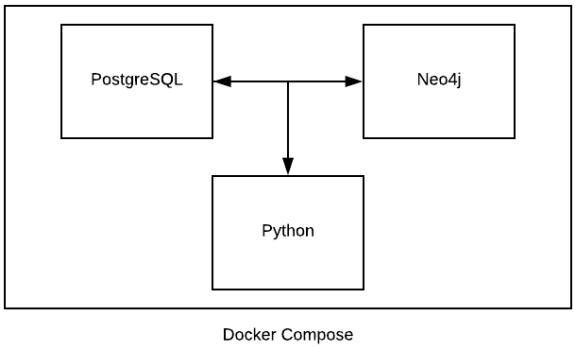

Figura 3: Modelo de contâiners

de acordo o nível dos seguintes aspectos que foram trabalhados em sua disponibilização:

Dados - Evidências Primárias/Secundárias: Análise da disponibilidade dos dados utilizados para futuros pesquisadores. Pode variar de simplesmente não dispobilizado até a disponibilização íntegra das informações, valendo-se de meios para uma oferta não repudiável e com garantias de integridade e veracidade. Neste trabalho, todos os dados (em suas diferentes agregações) são disponibilizados para uso futuro. Além disso os mecanismos de extração são automatizados, permitindo buscar outros projetos ou atualizar os dados dos já utilizados.

Modelo e Parâmetros: Verificação dos modelos e parâmetros utilizados no experimento. São essenciais para a discussão e reprodução do experimento, além de qualquer tentativa de otimização. Varia de não disponibilizado até a disponibilização plena com documentação adequada, interface robusta que trate valores fora do domínio (como nulos) e também prática, facilitando os testes com parâmetros e dados distintos. Os modelos de dados e de execução são detalhados neste trabalho. Além disso, todas as análises são automatizadas e encapusladas através de comandos executáveis dentro de um contâiner Docker.

Código Fonte: Disponibilidade do código fonte utilizado nas análises. Pode variar de não disponível (proprietário) até open source com direito de extensão e modificação. Neste projeto todo o código fonte é opensource e disponibilizado sob a permissiva licença MIT.

Sistema computacional requerido: Detalhamento das informações de hardware e software necessários, como por exemplo memória, arquitetura, processador, versão e plataforma. O nível ideal é a disponibilização do experimento em ambiente virtualizado em nuvem, em arquitetura que permita tanto a reprodução "as is"quanto extensão e modificação de parâmetros e dados de entrada. Como todo o pipeline deste projeto está disponibilizado na infraestrutura Docker, a única restrição da máquina do pesquisador é ter o Docker instalado. Informações do hardware utilizado nos experimentos serão disponibilizadas nas próximas seções.

3.2 Escolha dos repositórios de dados

Para avaliar as diferentes propriedades das redes colaborativas propostas, é necessário definir um conjunto de dados inseridos no contexto de estudo. As principais características buscadas foram:

Repositórios de GSD:. Como a recomendação de revisores têm maior relevância em desenvolvimento global de software, os repositórios escolhidos devem ser representantes deste contexto de desenvolvimento.

Acesso aberto: Os repositórios devem estar disponíveis para uso em pesquisa e para futuras reproduções e extensões do estudo. Além disso há a restrição de estarem no GitHub para acesso via API.

Popularidade e Escala: Os repostórios devem ser razoavelmente conhecidos e possuir um número de contribuidores que justifique a necessidade de recomendação de revisores. Para isso foram escolhidos projetos integrantes do top-10 "projects with the most reviews"eleitos pelo próprio GitHub[4]

A tabela 2 ilustra os projetos escolhidos e informações de contexto e tamanho.

4 RECOMENDANDO ESPECIALISTAS ENTRE PROJETOS

Especialistas de um projeto podem ser detectados de diversas formas. Analisando a atividade recente, por exemplo, é possível encontrar os desenvolvedores responsáveis pelas principais contribuições. Estas contribuições podem ser revisões, modificações no código, documentação e tradução dos artefatos.

Contudo, o intercâmbio destes especialistas entre projetos distintos é um desafio. Isso porque desenvolvedores podem ser especialistas em determinadas atividades ou tecnologias, e indentificar semelhanças de demandas entre projetos não é trivial. No GitHub, as labels dos pull requests tem como objetivo esta caracterização da modificação que está sendo feita, e por isso ajuda a identificar o tipo de conhecimento que o autor e o revisor demonstram ter influência. Ainda assim, as labels são particulares de cada projeto e podem ser editadas por seus mantenedores, o que reforça o desafio de buscar especialsitas que podem ser úteis em outros projetos.

Para atacar esta barreira, este trabalho propõe o uso de ontologias para alinhar as tags de diferentes projetos. O objetivo é que desenvolvedores conceituados dos projetos alvos do alinhamento tenham uma infraestrutura para documentar e expressar a relação entre as diversas labels dos projetos. Assim, é possível que especialistas encontrados em determinado projeto possam ser recomendados como especialsitas em outros projetos, de acordo com a equivalência das labels envolvidas.

4.0.1 Ontologia Proposta. A ontologia proposta pode ser visualizada na figura 4. Apenas duas entidades são criadas, *Project* e *Label*. Entre elas uma Object Property *owns* que denota que determinada label originalmente é originária de determinado projeto.

As diversas Object Properties que relacionam as Labels denotam relacionamento semântico entre elas. Enquanto *contains* siginifica que a *Label* de destino representa parte do conteúdo denotado pela entenidade de origem, *isSubPartOf* traduz o sentido inverso

[4]https://octoverse.github.com/

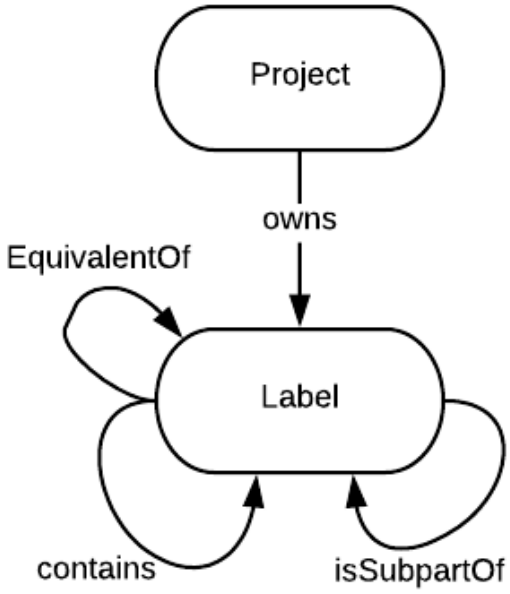

Figura 4: Visualização da Ontologia Proposta

e pode ser inferida pelos *reasoners*. Já *equivalentOf* denota uma correspondência completa entre duas entidades.

Através desta ontologia, é possível expressar os diferentes tipos de equivalência entre labels de projetos distintos. Assim, é possível inferir especialsitas de um projeto que podem contribuir em outro, dentro de suas especialidades já identificadas. Este alinhamento é o principal diferencial da abordagem baseada em ontologias, pois:

- Oferece uma interface direta para consulta das *Labels* já alinhadas, sem necessidade de processamento custoso em tempo real.
- Permite a utilização de Object Properties para descrever os relacionamentos de forma completa, onde o SGBD se limitaria a relacionamentos triviais como *"de..para"*.
- A estrutura é facilmente extensível para criar novos tipos de relacionamento entre as *labels* sem mudanças na estrutura relacional.

As próximas subseções mostram como a ontologia proposta pode ser populada e um estudo de caso para averiguar a utilização prática dos métodos descritos.

4.0.2 Populando a Ontologia. Como descrito em seções anteriores, os dados são extraídos dos repositórios distribuidos e consolidados em uma base de dados relacional. Este é o ponto central de onde outras transformações são executadas, como a criação da rede colaborativa. Para popular a ontolgia proposta, é necessário acessar este repositório central. A figura 5 mostra o processo de extração para disponibilizar uma interface SPARQL para prover as informações necessárias para os serviços de recomendação.

Nome	Linguagem Principal	Watchers	Stars	Contribuidores	Pull Requests
Node.js	JavaScript	2.868	52.709	2.088	8.440
TensorFlow	C++/Python	8.293	108.249	1.624	8.562
Kubernetes	Go	2.675	40.610	1.776	41.059
Symfony	PHP	1.297	18.381	1.692	17.475

Tabela 2: Descrição dos Respositórios Selecionados

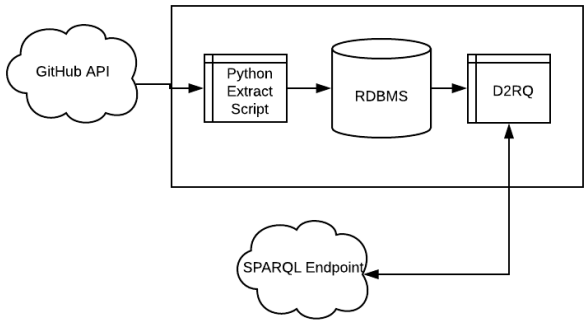

Figura 5: Workflow de Extração

O script de extração em Python fornece mecanismos para o alinhamento proposto das Labels, para o então fornecimento da interface SPARQL que será utilizada pelos metodos de detecção e recomendação de especialistas entre projetos. Assim como descrito na seção **??** que trata da integração de diversas bases, existem diversos métodos para buscar equivalência entre entidades oriundas de repositórios distintos. Algumas são baseadas na semântica ou na sintaxe dos descritores das entidades.

Todos os métodos apresentados servem para facilitar o processo de *matching* e evitar o trabalho operacional humano, mas para evitar a existência de falsos positivos, deve ser supervisionados por um especialista, que deve interferir quando achar necessário. Existem outros métodos que podem ser testados no futuro, extendendo o poder de alinhamento automático e auxiliando o processo com mais eficiência. Os principais métodos são:

- Direto, buscando *Labels* com descritores idênticos e fornencendo uma relação de *equivalentOf*.
- Parcial, buscando proximidade em detrimento de caracteres equivalentes
- Radicais, identificando abreviações de labels equivalentes
- Relação "todo-parte", onde *Labels* são identificadas como "sublabels" de outras

A tabela 3 mostra exemplos de alinhamentos que podem ser identificados automaticamente. Todas as saídas são sugestões de alinhamento para o especialista, que deve aprová-las para utilização posterior dos métodos de detecção e recomendação.

Método	Exemplos de Labels Alinhadas	Tipo de Relação
Direto	"python"e "python"	*equivalentOf*
Parcial	"lib_src"e "lib-src"	*equivalentOf*
Radicais	"perf"e "performance"	*equivalentOf*
"Todo-Parte"	"python-crypto"e "crypto"	*contains / isSubPartOf*

Tabela 3: Alinhamento automatizado

O D2RQ[5] é uma plataforma que provê uma interface de acesso às bases de dados relacionais como triplas RDF, acessíveis através de uma interface que interpreta consultas SPARQL. A conversão entre os modelos é feita nativamente para tipos mais comuns e baseada na ontologia proposta, e pode ser extendida através do desenvolvimento de classes que implementem o comportamento esperado. A conversão pode ser feita em tempo real ou criando um arquivo TTL (Turtle Triple Language) que solidifica todas as informações necessárias para posterior consulta.

Apesar do poder que as máquinas de inferência possuem, o desempenho desta abordagem ainda é um fator limitante de sua adoção geral na web semântica [31]. Quando se tratam de grande volumes de dados, os bancos de dados relacionais proveem mais escalabilidade, e por isso abordagens híbridas são aconselhadas [18]. Como os projetos opensource encontrados no GitHub possuem centenas de milhares de registros, testes preliminares apontaram pela inviabilidade da inferência através destas tecnologias. Assim, o D2RQ se mostrou como alternativa plausível que permite a especificação de serviços e modelos SPARQL aproveitando o poder da ontologia como linguagem de especificação, permitindo a utilização da infraestrutura relacional que dispõe de maior escalabilidade e desempenho.

5 AVALIAÇÃO DA ONTOLOGIA PROPOSTA

Para uma avaliação preliminar da estrutura proposta, foi realizado um estudo de caso com dados reais de um projeto de relevância no GitHub. Buscando por repositórios com quantidades razoáveis de dados, utilizamos o top-5 oficial da plataforma para projetos mais revisados[6], selecionado aquele com mais estrelas. Esta funcionalidade do GitHub pode ser utilizada como medida para o tamanho da comunidade passiva de um projeto[30]. Assim, este critério nos levou a escolher o **Node.js**[7], uma popular engine JavaScript

[5] http://d2rq.org/
[6] https://octoverse.github.com/
[7] https://github.com/nodejs/node

especialmente utilizada para aplicações *server-side* de alta vazão (I/O).

A organização responsável pelo projeto conta com outros 148 repositórios, a maioria deles bem menores e de alguma forma relacionados ao projeto principal. Para este estudo de caso foram escolhidos repositórios: o **llnode**[8] que mantém um plugin para debug do Node.js e o **build**[9] que administra o pipeline de deploy e integração contínua da tecnologia.

Os dados dos três projetos foram extraídos através do processo descrito anteriormente, e disponibilizados na na instância relacional. O script de alinhamento é executado, fornencendo sugestão de equivalência entre as diversas labels. Os resultados são descritos abaixo. Juntos, estes projetos reúnem mais de 200 labels, o que torna o alinhamento completamente manual lento e oneroso. Este aspecto potencializa a importância das técnicas automatizadas para suportar o processo.

Entre o projeto Node.js e o nodejs/llnode, diversas labels foram alinhadas de maneira direta: "V8 changes" e "feature request" por exemplo. A segunda é comum a diversos projetos e está mais relacionada ao workflow de desenvolvimento do que efetivamente um tópico de conhecimento, e pode ser negada pelo operador caso ele considere que este não seja um bom fator para recomendação de especialistas para projetos distintos. Utilizando o alinhamento parcial, as tags "flasky tests" e "flaky test" foram alinhadas. Esta equivalência é relevante pois identificar testes instáveis pode ser uma atribuição compartilhada dos projetos, pois pode estar relacionada à escrita dos testes ou à infraestrutura de execução.

Já entre Node.js e nodejs/build, houve alinhamento direto entre "tsc-review", que indica que o pull request deve ser avaliado pelo Technical Steering Committee do projeto. Outros alinhamentos como "enhancement" e "bug" foram identificados, e provavelmente não servem para detecção de especialistas entre projetos pois são comuns às atividades de software e não necessariamente a um tópico de conhecimento. Já a comparação por radicais apontou equivalência entre as labels "install" e "installer", que parecem ter funções parecidas nos projetos mas foram grafadas de forma distinta. Neste aspecto o alinhamento parece relevante, pois os principais revisores em cada uma destas labels são os mesmos em cada um dos projetos.

Com este estudo de caso foi possível observar que algumas equivalências automatizadas competem a tópicos muito generalistas, onde um especialista de determinado projeto não parece ter competência explícita para o ser em outro. Contudo, diversas equivalências parecem ser relevantes para os projetos selecionados, e por isso podem expressar relacionamentos que apoiam o processo de alinhamento manual. Estudos de caso posteriores podem ajudar a investigar se os especialistas têm performance equiparável quando contribuindo em outros projetos.

6 CONCLUSÃO

O *code review* é uma prática estabelecida com efeitos conhecidos e confirmados por vários autores através dos anos. Desde o processo tradicional que nasceu nos anos 70"baseado em um workflow rígido, a técnica se desenvolveu constantemente para o modelo leve e ágil que valoriza mais a troca de conhecimento e a relação entre pessoas.

Estudos recentes mostram o impacto da escolha do revisor correto na eficiência e impacto do processo na qualidade do código e na detecção precoce de defeitos. Isso fez com que diversos autores propossem métodos para recomendar revisores adequados, dado o contexto e o conteúdo de umaa revisão. Os desafios do desenvolvimento global de software potencializam a necessidade desta recomendação, visto a quantidade de informações que influenciam na capacidade do revisor e a dificuldade de agregá-las de forma manual, levando a escolhas tardias ou fundamentalmente ruins.

Trabalhos mais recentes apresentam propostas e métodos de avaliação baseados em informações de colaboração, como atividade, eficiência de revisões passadas e capacidade de interação. Estes resultados motivaram a proposta de novos métodos de recomendação baseados em redes colaborativas e voltados para o desenvolvimento global de software, com suas peculiaridades.

Para expansão dos métodos já avaliados, especialistas podem ser recomendados para projetos distintos de suas contribuições originais. Para isso, este trabalho propõe a utilização de ontologias para alinhamento dos tópicos de semânticos internos dos repositórios, de forma a inferir que se um desenvolvedor domina determinado conhecimento de um projeto, poderá auxiliar em outro projeto com demandas equivalentes. Para testar esta hipótese, foi realizado um estudo de caso com o alinhamento de projetos relevantes da comunidade opensource. Foram identificadas diversas equivalências, algumas relevantes e outras que podem ser negadas pelo o operador do processo, um especialista realiza a curadoria do alinhamento e julga as sugestões geradas automaticamente.

Como trabalhos futuros, existem diversas aplicações de ontologia nos métodos propostos. É possível sua utilização para integração e alinhamento dos dados de diferentes bases que contém informações sobre desenvolvedores e sua reputação, como o StackOverflow (já explicitado neste trabalho) e ferramentas como o Twitter. Além disso, testes aprofundados e avaliação dos métodos de recomendação conhecidos podem apontar novas aplicações de ontologia com o objetivo de melhorar a qualidade das saídas e a predição de revisores cada vez mais adequados.

[8]https://github.com/nodejs/llnode
[9]https://github.com/nodejs/build

REFERÊNCIAS

[1] Gediminas Adomavicius and Alexander Tuzhilin. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge & Data Engineering* 6 (2005), 734–749.

[2] Jorge Luis Nicolas Audy and Rafael Prikladnicki. 2007. *Desenvolvimento distribuído de software.* Elsevier.

[3] Alberto Bacchelli and Christian Bird. 2013. Expectations, Outcomes, and Challenges of Modern Code Review. In *Proceedings of the 2013 International Conference on Software Engineering (ICSE '13)*. IEEE Press, 712–721.

[4] G. Bavota and B. Russo. 2015. Four eyes are better than two: On the impact of code reviews on software quality. *2015 IEEE 31st International Conference on Software Maintenance and Evolution, ICSME 2015 - Proceedings* (2015), 81–90. https://doi.org/10.1109/ICSM.2015.7332454

[5] O. Baysal, O. Kononenko, R. Holmes, and M.W. Godfrey. 2013. The influence of non-technical factors on code review. *Proceedings - Working Conference on Reverse Engineering, WCRE* (2013), 122–131. https://doi.org/10.1109/WCRE.2013.6671287

[6] Tim Berners-Lee, James Hendler, and Ora Lassila. 2001. The semantic web. *Scientific american* 284, 5 (2001), 34–43.

[7] C. Bird, T. Carnahan, and M. Greiler. 2015. Lessons learned from building and deploying a code review analytics platform. *IEEE International Working Conference on Mining Software Repositories* 2015 (2015), 191–201. https://doi.org/10.1109/MSR.2015.25

[8] Barry Boehm and Victor R. Basili. 2001. Software Defect Reduction Top 10 List. *Computer* 34, 1 (12 2001), 135–137. https://doi.org/10.1109/2.962984

[9] Carl Boettiger. 2015. An introduction to Docker for reproducible research. *ACM SIGOPS Operating Systems Review* 49, 1 (2015), 71–79.

[10] A. Bosu and J.C. Carver. 2014. Impact of developer reputation on code review outcomes in OSS projects: An empirical investigation. *International Symposium on Empirical Software Engineering and Measurement* (2014). https://doi.org/10.1145/2652524.2652544

[11] Valentine Casey. 2010. Virtual software team project management. *Journal of the Brazilian Computer Society* 16, 2 (2010), 83–96.

[12] Roy T Fielding and Richard N Taylor. 2002. Principled design of the modern Web architecture. *ACM Transactions on Internet Technology (TOIT)* 2, 2 (2002), 115–150.

[13] Juliana Freire, Philippe Bonnet, and Dennis Shasha. 2012. Computational reproducibility: state-of-the-art, challenges, and database research opportunities. In *Proceedings of the 2012 ACM SIGMOD international conference on management of data*. ACM, 593–596.

[14] Chenbo Fu, Mingming Zhou, Qi Xuan, and Hong-Xiang Hu. 2017. Expert recommendation in oss projects based on knowledge embedding. *International Workshop on Complex Systems and Networks (IWCSN)*, 149–155.

[15] Georgios Gousios, Martin Pinzger, and Arie van Deursen. 2014. An exploratory study of the pull-based software development model. In *Proceedings of the 36th International Conference on Software Engineering*. ACM, 345–355.

[16] Georgios Gousios, Margaret-Anne Storey, and Alberto Bacchelli. 2016. Work practices and challenges in pull-based development: The contributor's perspective. In *Software Engineering (ICSE), 2016 IEEE/ACM 38th International Conference on*. IEEE, 285–296.

[17] Nicola Guarino. 1998. *Formal ontology in information systems: Proceedings of the first international conference (FOIS'98), June 6-8, Trento, Italy*. Vol. 46. IOS press.

[18] Martin Hepp. 2008. Ontologies: State of the art, business potential, and grand challenges. In *Ontology Management*. Springer, 3–22.

[19] Darrel C Ince, Leslie Hatton, and John Graham-Cumming. 2012. The case for open computer programs. *Nature* 482, 7386 (2012), 485–488.

[20] Jing Jiang, Yun Yang, Jiahuan He, Xavier Blanc, and Li Zhang. 2017. Who should comment on this pull request? Analyzing attributes for more accurate commenter recommendation in pull-based development. *Information and Software Technology* 84 (2017), 48–62.

[21] C. F. Kemerer and M. C. Paulk. 2009. The Impact of Design and Code Reviews on Software Quality: An Empirical Study Based on PSP Data. *IEEE Transactions on Software Engineering* 35, 4 (07 2009), 534–550. https://doi.org/10.1109/TSE.2009.27

[22] O.a Kononenko, O.b Baysal, L.c Guerrouj, Y.b Cao, and M.W.a Godfrey. 2015. Investigating code review quality: Do people and participation matter? *2015 IEEE 31st International Conference on Software Maintenance and Evolution, ICSME 2015 - Proceedings* (2015), 111–120. https://doi.org/10.1109/ICSM.2015.7332457

[23] A. Meneely, A.C.R. Tejeda, B. Spates, S. Trudeau, D. Neuberger, K. Whitlock, C. Ketant, and K. Davis. 2014. An empirical investigation of socio-technical code review metrics and security vulnerabilities. *6th International Workshop on Social Software Engineering, SSE 2014 - Proceedings* (2014), 37–44. https://doi.org/10.1145/2661685.2661687

[24] Stuart E Middleton, David C De Roure, and Nigel R Shadbolt. 2001. Capturing knowledge of user preferences: ontologies in recommender systems. In *Proceedings of the 1st international conference on Knowledge capture*. ACM, 100–107.

[25] Stuart E Middleton, Nigel R Shadbolt, and David C De Roure. 2004. Ontological user profiling in recommender systems. *ACM Transactions on Information Systems (TOIS)* 22, 1 (2004), 54–88.

[26] Wenkai Mo, Beijun Shen, Yuming He, and Hao Zhong. 2015. GEMiner: Mining Social and Programming Behaviors to Identify Experts in Github. In *Proceedings of the 7th Asia-Pacific Symposium on Internetware (Internetware '15)*. ACM, New York, NY, USA, 93–101. https://doi.org/10.1145/2875913.2875924

[27] Rodrigo Morales, Shane McIntosh, and Foutse Khomh. 2015. Do code review practices impact design quality? A case study of the Qt, VTK, and ITK projects. *2015 IEEE 22nd International Conference on Software Analysis, Evolution and Reengineering (SANER)* 00 (2015), 171–180.

[28] Ana Maria Nicolaci-da Costa and Mariano Pimentel. 2011. Sistemas colaborativos para uma nova sociedade e um novo ser humano. *Sistemas colaborativos. PIMENTEL, M.; FUKS, H.(Orgs.). Rio de Janeiro: Elsevier* (2011).

[29] Mohammad Masudur Rahman, Chanchal K Roy, and Jason A Collins. 2016. CoRReCT: code reviewer recommendation in GitHub based on cross-project and technology experience. In *Software Engineering Companion (ICSE-C), IEEE/ACM International Conference on*. IEEE, 222–231.

[30] Jyoti Sheoran, Kelly Blincoe, Eirini Kalliamvakou, Daniela Damian, and Jordan Ell. 2014. Understanding watchers on GitHub. In *Proceedings of the 11th Working Conference on Mining Software Repositories*. ACM, 336–339.

[31] Pavel Shvaiko and Jérôme Euzenat. 2013. Ontology matching: state of the art and future challenges. *IEEE Transactions on knowledge and data engineering* 25, 1 (2013), 158–176.

[32] Rajesh Sinha and Prem Sewak Sudhish. 2016. A principled approach to reproducible research: a comparative review towards scientific integrity in computational research. In *Ethics in Engineering, Science and Technology (ETHICS), 2016 IEEE International Symposium on*. IEEE, 1–9.

[33] X.a Xia, D.b Lo, X.a b Wang, and X.a Yang. 2015. Who should review this change?: Putting text and file location analyses together for more accurate recommendations, In IEEE 23rd International Conference on Software Analysis, Evolution, and Reengineering (SANER 2016). *2015 IEEE 31st International Conference on Software Maintenance and Evolution, ICSME 2015 - Proceedings* (2015), 261–270. https://doi.org/10.1109/ICSM.2015.7332472

[34] Z. Xia, H. Sun, J. Jiang, X. Wang, and X. Liu. 2017. A hybrid approach to code reviewer recommendation with collaborative filtering. In *2017 6th International Workshop on Software Mining (SoftwareMining)*. 24–31. https://doi.org/10.1109/SOFTWAREMINING.2017.8100850

[35] Xin Yang, Norihiro Yoshida, Raula Gaikovina Kula, and Hajimu Iida. 2016. Peer review social network (PeRSoN) in open source projects. *IEICE TRANSACTIONS on Information and Systems* 99, 3 (2016), 661–670.

[36] Y. Yu, H. Wang, G. Yin, and C. X. Ling. 2014. Reviewer Recommender of Pull-Requests in GitHub. In *2014 IEEE International Conference on Software Maintenance and Evolution*. 609–612. https://doi.org/10.1109/ICSME.2014.107

[37] Yue Yu, Huaimin Wang, Gang Yin, and Charles X Ling. 2014. Who should review this pull-request: Reviewer recommendation to expedite crowd collaboration. In *Software Engineering Conference (APSEC), 2014 21st Asia-Pacific*, Vol. 1. IEEE, 335–342.