

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

VINICIUS J. SCHETTINO, Universidade Federal de Juiz de Fora

REGINA BRAGA, Universidade Federal de Juiz de Fora

MARCO ANTÔNIO P. ARAÚJO, Universidade Federal de Juiz de Fora

VICTOR STRÖELE, Universidade Federal de Juiz de Fora

JOSÉ MARIA NASAR DAVID, Universidade Federal de Juiz de Fora

A prática de *code review* está associada à qualidade de código e detecção precoce de defeitos de software. Intrinsecamente colaborativo, a eficiência do processo de revisão depende também do perfil histórico do revisor e da natureza do código em avaliação. No contexto de Desenvolvimento Distribuído de Software (DDS), 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 interação e troca de experiências 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 trabalho. Este estudo apresenta a arquitetura responsável pela extração dos dados e aplicação da ontologia proposta, avaliado por um cenário envolvendo projetos *open source* conhecidos de uma mesma organização.

CCS Concepts: •**Software and its engineering** → **Open source model**; **Programming teams**; *Software verification*; *Software design engineering*;

Additional Key Words and Phrases: code review, ontology, distributed software development

ACM Reference format:

Vinicius J. Schettino, Regina Braga, Marco Antônio P. Araújo, Victor Ströele, and José Maria Nasar David. 2019. Utilizando Ontologias para apoiar a recomendação de revisores de código entre projetos distintos. 1, 1, Article 4 (January 2019), 14 pages.

DOI: xxxxxxxx

1 INTRODUÇÃO

Code review é considerada uma das principais técnicas para diminuição de defeitos de software [8]. O autor de uma alteração na base de código de um projeto submete tal conteúdo ao crivo de um conjunto de pares, que irão revisar sua estrutura com base em uma lista de regras e convenções previamente definidas. Diferentes aspectos relacionados ao autor, ao revisor e ao processo de revisão estão diretamente relacionados à eficiência da prática. Autores relatam a diminuição da incidência de *anti-patterns* [25] de acordo com o nível de participação dos envolvidos e cobertura do código revisado [5, 27, 31]. Reputação [10] e experiência [26] do revisor também parecem impactar nos efeitos do *code review*

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

© 2019 Copyright held by the owner/author(s). Manuscript submitted to ACM

Intrínsecamente colaborativa, a atividade de *code review* é exercida com suporte de ferramentas computacionais específicas [3], principalmente no desenvolvimento distribuído. No contexto de workflows de trabalho descentralizados [18], a prática funciona como um *gateway* de qualidade que busca garantir que apenas alterações aderentes aos padrões de qualidade do projeto serão incorporadas 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.

Quando este processo é conduzido por indivíduos distribuídos em diferentes localidades e organizações, o mesmo é configurado como Desenvolvimento Distribuído de Software (DDS) [21]. Esta abordagem 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 geograficamente distribuída.

O Desenvolvimento Distribuído de Software afeta 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 determinado *patch*, relacionado ao vasto número de opções e pouca informação disponível sobre os aspectos técnicos e gerenciais (e.g. tempo disponível). A falta de contato co-localizado e a natureza distribuída deste tipo de desenvolvimento dificulta o processo de escolha do revisor, impactando negativamente a eficiência do processo [39].

Existem alguns trabalhos que demonstram métodos de recomendação de revisores [24, 44]. Também foram revisados outros trabalhos que apontam características de revisões, revisores e autores que possivelmente potencializam a colaboração [7, 25]. Além disso, o desafio de se encontrar potenciais colaboradores se estende para outras áreas, como recomendação de parceiros de pesquisa [13].

As principais abordagens encontradas na literatura se limitam a utilizar informações do repositório do projeto em questão para realizar a recomendação. 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 para contribuir em outros grupos. Os principais métodos existentes não são capazes de realizar esta recomendação entre projetos distintos.

Este tipo de auxílio se faz importante dentro de organizações maiores. É o caso, por exemplo, da Apache Foundation [12], com dezenas de projetos com diferentes linguagens, arquiteturas e propósitos. É possível que desenvolvedores que trabalham em projetos específicos tenham conhecimentos aplicáveis em projetos distintos, apesar de sempre terem 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.

Assim, o objetivo deste estudo é estender a utilização de técnicas de recomendação de revisores para atender demandas entre projetos distintos. Esta extração é baseada em tópicos de conhecimento e na especialidade dos desenvolvedores. A abordagem 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 seja realizada a equivalência entre os tópicos de conhecimento de projetos de software. Ao invés de propor um novo método de recomendação, este trabalho apresenta uma forma de aumentar o alcance dos métodos existentes ao permitir que seja possível intercambiar potenciais revisores entre projetos distintos.

2 TRABALHOS RELACIONADOS

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 [32, 36].

Nos métodos tradicionais de recomendação de revisores de código, são utilizados dados de proximidade de código ou de semântica com as revisões anteriores [4, 23]. 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. [16] descrevem uma recomendação baseada em grafos de relacionamentos entre desenvolvedores e seus status de trabalho no projeto. Xia et al. [41] agrupam desenvolvedores utilizando métodos de proximidade para capturar relações implícitas e explícitas. Yu et al. [43, 44] 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. [42] utilizam social network analysis (SNA) baseadas em informações de atividade recente dos potenciais revisores no projeto. Nenhuma destas abordagens recomenda revisores para projetos que ele ainda não trabalhou, nem utiliza dados de projetos distintos para alimentar o histórico do desenvolvedor. Contudo, são abordagens passíveis de serem estendidas de acordo com a proposta do presente trabalho.

CoRRReCT [33] 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 *open source* feitas pelo desenvolvedor. A abordagem é similar ao proposto por Mo et al. [30]. Em relação a esta abordagem, a principal inovação da proposta pelo presente trabalho é o uso de ontologias para recomendar especialistas entre projetos distintos. A utilização de ontologias para recomendação é uma técnica já explorada em trabalhos anteriores [29] e indicada como sugestão para trabalhos futuros [1], e por isso é um potencial caminho para otimização dos resultados da recomendação.

Uma ontologia pode ser definida como um vocabulário ou linguagem comum para representar conhecimento [19]. Para computação, pode ser vista como uma especificação formal, explícita e não ambígua de um conceito compartilhado por diferentes agentes.

Sob o ponto de vista operacional, 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 conhecimento, antes não explícito, nas informações disponíveis. Estas máquinas inferem propriedades e relacionamentos baseadas em conhecimento diretamente representado [6]. Assim, no contexto deste trabalho as ontologias serão utilizadas para descrever a relação entre tópicos de conhecimento de projetos distintos, explorando desta forma o potencial de extração de conhecimento implícito trazido pelas ontologias para melhoria da recomendação em code review no contexto de DDS.

3 METODOLOGIA

A abordagem proposta consiste em alinhar os tópicos de conhecimento existentes entre projetos distintos, possibilitando assim o intercâmbio de revisores. Para este alinhamento, a ontologia proposta no presente trabalho foi alimentada com as tarefas dos projetos e cada uma de suas categorizações. Os dados foram processados para encontrar semelhanças entre

as categorias. Com base nas inferências e conhecimento explícito da ontologia, foi possível descrever as equivalências dos tópicos de conhecimento entre os projetos analisados. Por fim o alinhamento foi avaliado de acordo com a opinião de especialistas nos projetos selecionados para o cenário de uso, descrito nas próximas seções.

Considerando a prática de revisão de código, o uso de ferramentas computacionais se tornou prática comum neste contexto 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 revisões¹, também chamadas de *pull requests* no modelo de desenvolvimento “*pull based*” [17]. Esta plataforma foi escolhida devido ao número de projetos disponíveis e sua natureza aberta. As próximas subseções detalham o planejamento e execução de cada passo da metodologia proposta.

3.1 Extração dos Dados

Os dados da análise foram extraídos do GitHub através da API² disponibilizada para desenvolvedores estenderem as funcionalidades da plataforma e integrar novos produtos que agreguem valor ao processo de desenvolvimento. A versão estável para integração foi a v3, que implementa o padrão arquitetural RESTful[14]. Existe a API recém lançada v4³, 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.

Os dados extraídos foram estruturados em um esquema relacional, mais especificamente em uma instância PostgreSQL 10.5. Este passo intermediário ocorreu para que os dados ficassem 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 baseada em grafo, ficando disponíveis para consultas posteriores sem degradar o desempenho 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 característica 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 na *thread* de discussão do *pullrequests*, representados pelos *discussioncomments* ou através de comentários no código em revisão, nos *reviewcomments*.

O algoritmo de extração dos dados foi projetado para garantir a integridade e disponibilidade das informações, diante das nuances do repositório e do domínio de dados. Os principais aspectos observados foram:

Inconsistência nos dados: Para evitar dados inconsistentes, toda a extração é feita no contexto de uma *transaction* relacional do SGBD, garantindo que todas as ações sejam executadas (ou canceladas) de forma atômica. As restrições de chave primária evitam 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 foi necessário tratar erros (como 400, 500 e 502) e evitar que a extração seja interrompida. Para isso, as

¹<https://octoverse.github.com/>

²<https://developer.github.com/v3/>

³<https://developer.github.com/v4/>

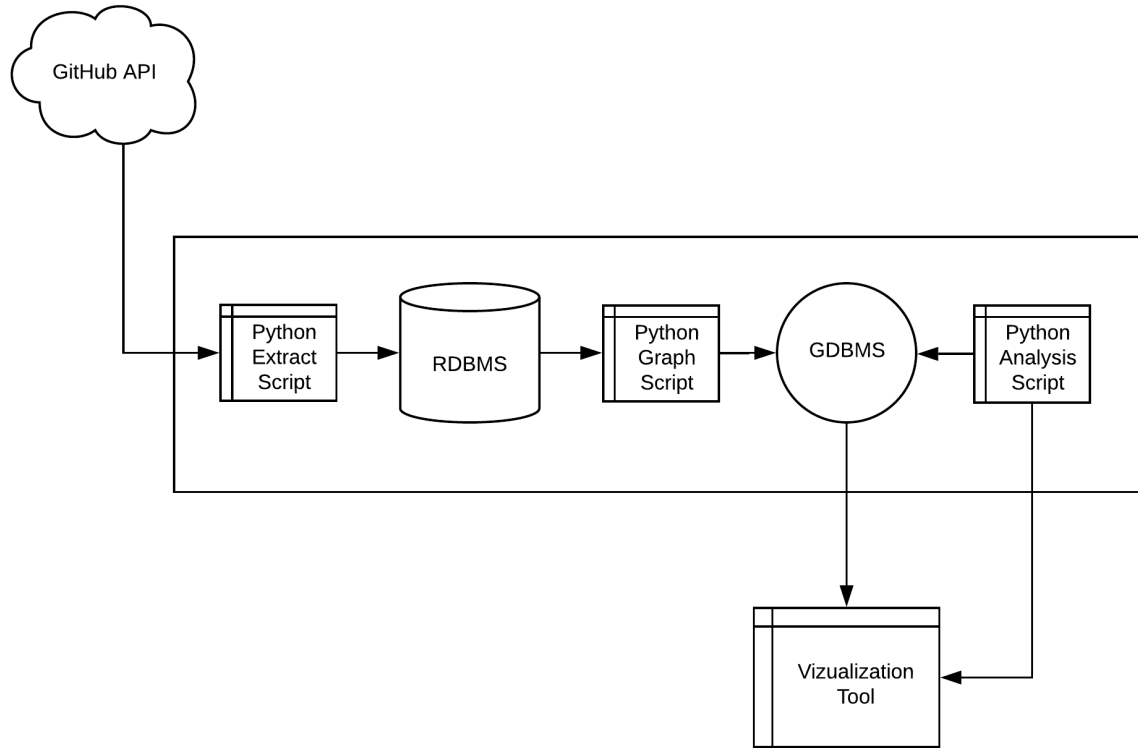


Fig. 1. Processo de extração dos dados

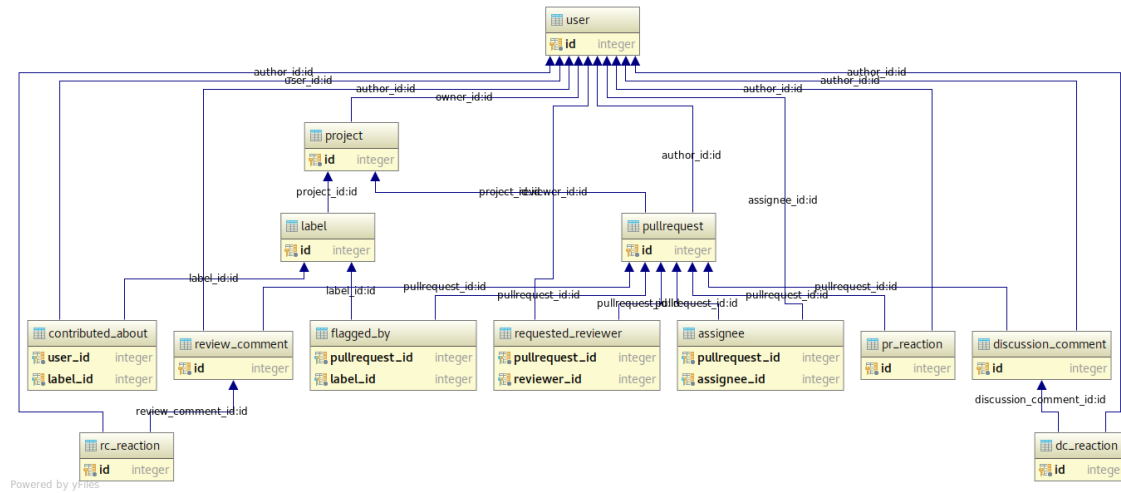


Fig. 2. Modelo Entidade Relacionamento

respostas são tratadas, e detectados erros temporários como estes, a requisição é agendada para se repetir em uma janela de tempo definida.

Ratelimit da API: . A API do GitHub limita o número de requisições 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 requisições, 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.

Dentre as possíveis melhorias futuras desta arquitetura, técnicas de paralelismo pode configurar uma diminuição do tempo de extração, através da utilização de chaves de acesso de API distintas. Esta melhoria faz importante especialmente para implementação de ferramentas automatizadas que tenham como requisito não funcional prioritário o tempo de execução.

3.2 Aspectos de Reprodutibilidade

Cada uma das tecnologias apresentadas foram encapsuladas através da tecnologia de virtualização Docker [28]. Esta abordagem é uma resposta às iniciativas de reprodutibilidade na ciência, buscando maior transparência, confiabilidade e possibilidade de extensão nos experimentos [15]. Os três componentes (banco de dados relacional, orientado a grafo e os scripts de extração, transformação e carga) foram encapsulados em contêineres distintos, como mostra a figura 3. Estes foram orquestrados através do *docker-compose* [28], que configura os parâmetros e acessos necessários para o funcionamento do sistema sem que o usuário precise realizar nenhuma ação extra.

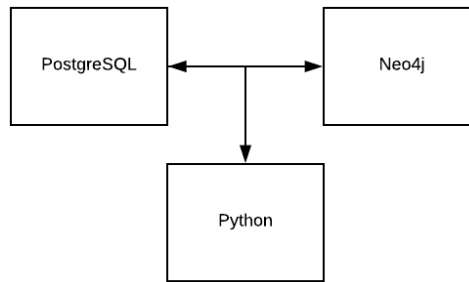


Fig. 3. Modelo de contêineres

Com um conjunto grande de dependências, tecnologias e minúcias que estão envolvidas neste tipo de experimento, o código fonte e a descrição ainda que detalhada dos dados não são suficiente para alcançar níveis adequados de reprodutibilidade [22]. Com auxílio dos containers Docker, foi 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].

Neste trabalho, o aspecto da reprodutibilidade ganha contornos relevantes de acordo com os objetivos apresentados. Esta particularidade se deve ao fato do método proposto pode ser acoplado com diferentes métodos de recomendação de revisores, para ser comparado e colocado em prática em ferramentas de recomendação. Assim garantir altos índices de reprodutibilidade permite aumentar o número de acoplamentos testados e tecer avaliações mais abrangentes.

De acordo com Sinha et al. [38], a maturidade da reprodutibilidade de um experimento científico computacional pode ser medido 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 disponibilizado até a disponibilização integral 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) foram disponibilizados para uso futuro⁴. 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 tentativas 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 foram detalhados neste trabalho. Além disso, todas as análises são automatizadas e encapsuladas através de comandos executáveis dentro de um contêiner 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 *open source* e foi disponibilizado sob 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 fluxo de trabalho deste projeto está disponibilizado na infraestrutura Docker, a única restrição da máquina do pesquisador é ter o Docker instalado.

3.3 Escolha dos repositórios de dados

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

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

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

Popularidade e Escala: Os repositórios deveriam ser razoavelmente conhecidos e possuir um número de contribuidores que justificasse a necessidade de recomendação de revisores. Para isso foram escolhidos projetos integrantes do top-10 "projects with the most reviews" eleitos pelo próprio GitHub⁵

A tabela 1 ilustra os projetos escolhidos e informações de contexto e tamanho.

⁴https://github.com/vschettino/gitrev/blob/groupby_labels/dumps/complete_dump2018-08-05.tar.xz

⁵<https://octoverse.github.com/>

Nome	Linguagem	Watchers	Stars	Contribuidores	Pullrequests
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

Table 1. Descrição dos Respositórios Seleccionados

3.4 Equivalência de tópicos de conhecimento

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 identificar semelhanças de demandas entre projetos não é trivial. No GitHub, as *labels* dos *pullrequests* 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 especialistas que podem ser úteis em outros projetos.

Para atacar esta barreira, este trabalho propõe o uso de ontologias para alinhar as *labels* de diferentes projetos. Estas entidades são responsáveis por organizar e categorizar o fluxo de trabalho dos projetos no GitHub [17], indicando quais áreas de conhecimento, tecnologias e responsabilidades determinado *pullrequest* está associado. 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 especialistas em outros projetos, de acordo com a equivalência das *labels* envolvidas.

3.5 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* é originária de determinado projeto.

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

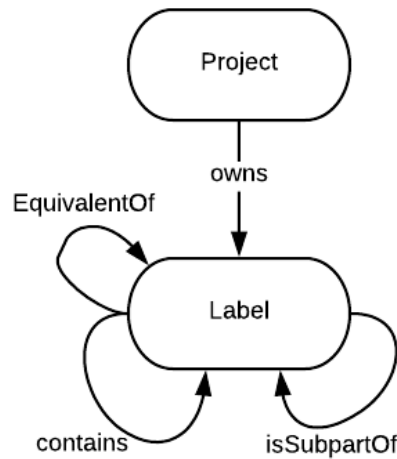


Fig. 4. Visualização da Ontologia Proposta

- 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 detalham como a ontologia proposta pode ser populada e cenário de utilização para averiguar a utilização prática dos métodos descritos.

Como descrito em seções anteriores, os dados são extraídos dos repositórios distribuídos 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 ontologia 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.

O software de extração em Python fornece mecanismos para o alinhamento proposto das Labels, para o então fornecimento da interface SPARQL que será utilizada pelos métodos de detecção e recomendação de especialistas entre projetos. Existem diversos métodos para buscar equivalência entre entidades oriundas de repositórios distintos. Algumas são baseadas na semântica ou na sintaxe dos descritores das entidades. Algumas tem como foco o *matching* seguro das informações, evitando ao máximo falsos positivos [40]. Em outros casos os autores buscam abordagens mais rebuscadas, especialmente para integração de fontes de dados heterogêneas [37].

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, devem ser supervisionados por um especialista, que deve interferir quando achar necessário. Existem outros métodos que podem ser testados no futuro, estendendo o poder de alinhamento automático e auxiliando o processo com mais eficiência. Os principais métodos são:

- Direto, buscando *Labels* com descritores idênticos e fornecendo uma relação de *equivalentOf*.
- Parcial, buscando proximidade em detrimento de caracteres equivalentes (e.g. “lib/src” e “lib-src”).
- Radicais, identificando abreviações de labels equivalentes (e.g. “perf” e “performance”).
- Relação “todo-parte”, onde *Labels* são identificadas como “sublabels” de outras (e.g. “python-cripto” e “cripto”).

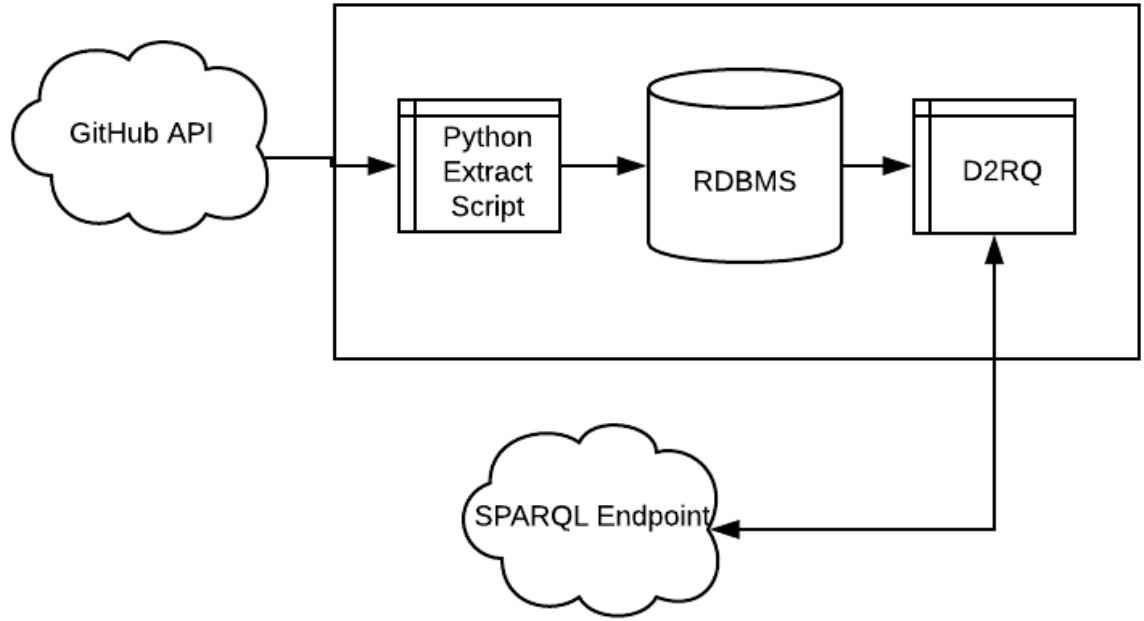


Fig. 5. Workflow de Extração

O D2RQ⁶ 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 é baseada na ontologia proposta, e pode ser estendida 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 [35]. Quando se tratam de grande volumes de dados, os bancos de dados relacionais proveem mais escalabilidade, e por isso abordagens híbridas são aconselhadas [20]. Como os projetos *open source* 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.

4 CENÁRIO DE UTILIZAÇÃO

Para demonstrar o funcionamento da estrutura proposta, utilizamos dados reais de um projeto de relevância no GitHub. Como o objetivo deste trabalho é apresentar a abordagem de intermediação de revisores entre projetos distintos, detalhes do método de recomendação utilizado não são apresentados. A abordagem em questão é independente do método de recomendação, dado que especialistas indicados podem ser categorizados de acordo com os tópicos de conhecimento que eles contribuíram, não sendo relevante o método utilizado para sua detecção. Portanto, o foco desta seção é avaliar

⁶<http://d2rq.org/>

a eficácia da equivalência entre as *labels* dos projetos distintos. Esta análise pode ser posteriormente utilizada para estender métodos de detecção ou recomendação de revisores.

Utilizamos o top-5 oficial da plataforma para projetos mais revisados⁷, selecionado aquele com mais estrelas (*stars*). Esta funcionalidade do GitHub pode ser utilizada como medida para o tamanho da comunidade passiva de um projeto[34]. Assim, este critério nos levou a escolher o **Node.js**⁸, uma popular engine JavaScript 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 **lnode**⁹ que mantém um plugin para debug do Node.js e o **build**¹⁰ que administra o pipeline de deploy e integração contínua da tecnologia.

Para o papel de especialista responsável pelo processo de *matching* foi escolhido um desenvolvedor de software com conhecimento do processo de desenvolvimento dos projetos em questão, bem como as regras de governança e workflow.

Os dados dos três projetos foram extraídos através do processo descrito anteriormente, e disponibilizados na instância relacional. O script de alinhamento é executado, fornecendo 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 manual lento e oneroso. Este aspecto potencializa a importância das técnicas automatizadas para suportar o processo.

Entre o projeto Node.js e o lnode, 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 “flasky test” foram alinhadas. Esta equivalência é relevante, pois identificar testes instáveis pode ser uma atribuição compartilhada dos projetos, podendo estar relacionada à escrita dos testes ou à infraestrutura de execução.

Já entre Node.js e build, houve alinhamento direto entre “tsc-review”, que indica que o pull request deve ser avaliado pelo Technical Steering Committee do projeto¹¹. 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 e por isso foram negados pelo operador. 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.

A tabela 2 demonstra os principais alinhamentos 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 <i>Labels</i> Alinhadas	Tipo de Relação
Direto	“tsc-review” e “tsc-review”	<i>equivalentOf</i>
Parcial	“flasky tests” e “flasky tests”	<i>equivalentOf</i>
Radicais	“install” e “installer”	<i>equivalentOf</i>

Table 2. Alinhamento automatizado

⁷<https://octoverse.github.com/>

⁸<https://github.com/nodejs/node>

⁹<https://github.com/nodejs/lnode>

¹⁰<https://github.com/nodejs/build>

¹¹<https://github.com/nodejs/TSC>

Com a construção deste cenário 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.

5 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 setenta 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 propoem métodos para recomendar revisores adequados, dado o contexto e o conteúdo de uma revisão. Os desafios do desenvolvimento global de software potencializam a necessidade desta recomendação: a quantidade de informações que influenciam na capacidade do revisor e a dificuldade de agregá-las de forma manual pode levar 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 [16, 41, 42]. 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 descrito um cenário de uso 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.

A abordagem proposta pode ser utilizada para estender métodos de recomendação de revisores de código, levando resultados relevantes para projetos distintos, baseada nas equivalências dos tópicos de conhecimento que podem existir entre projetos correlatos. Utilizando as *labels* que categorizam os *pullrequests* dos projetos, é possível identificar que potenciais revisores em determinadas linguagens, bibliotecas e tecnologias podem contribuir em outros projetos que compartilhem das mesmas necessidades.

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 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. Para aumentar a eficácia do processo de equivalência dos tópicos semânticos, pretendemos utilizar um Tesouro, como o WordNet¹². Através dos sinônimos cognitivos (*synsets*) pode ser possível encontrar equivalências mais adequadas e sem embasamento direto na sintaxe dos léxicos. Por fim, queremos

¹²<https://wordnet.princeton.edu/>

realizar avaliações aprofundadas da efetividade que desenvolvedores indicados entre projetos distintos no processo de code review.

REFERENCES

- [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] Vipin Balachandran. 2013. Reducing human effort and improving quality in peer code reviews using automatic static analysis and reviewer recommendation. In *Software Engineering (ICSE), 2013 35th International Conference on*. IEEE, 931–940.
- [5] 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>
- [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] Tadeusz Chełkowski, Peter Gloor, and Dariusz Jemielniak. 2016. Inequalities in open source software development: analysis of contributor's commits in apache software foundation projects. *PloS one* 11, 4 (2016), e0152976.
- [13] Luciano Antonio Digiampietri, William Takahiro Maruyama, Caio Rafael do Nascimento Santiago, and Jamison Jose da Silva Lima. 2015. A Link Prediction System in Social Networks. In *Proceedings of the Annual Conference on Brazilian Symposium on Information Systems: Information Systems: A Computer Socio-Technical Perspective - Volume 1 (SBSI 2015)*. Brazilian Computer Society, Porto Alegre, Brazil, Brazil, Article 20, 8 pages. <http://dl.acm.org/citation.cfm?id=2814058.2814082>
- [14] 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.
- [15] 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.
- [16] 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.
- [17] 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.
- [18] 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.
- [19] 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.
- [20] Martin Hepp. 2008. Ontologies: State of the art, business potential, and grand challenges. In *Ontology Management*. Springer, 3–22.
- [21] James D. Herbsleb and Audris Mockus. 2003. An empirical study of speed and communication in globally distributed software development. *IEEE Transactions on software engineering* 29, 6 (2003), 481–494.
- [22] Darrel C Ince, Leslie Hatton, and John Graham-Cumming. 2012. The case for open computer programs. *Nature* 482, 7386 (2012), 485–488.
- [23] Jing Jiang, Jia-Huan He, and Xue-Yuan Chen. 2015. Coredevrec: Automatic core member recommendation for contribution evaluation. *Journal of Computer Science and Technology* 30, 5 (2015), 998–1016.
- [24] 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.
- [25] 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>
- [26] 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>
- [27] 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),

Manuscript submitted to ACM

- 37–44. <https://doi.org/10.1145/2661685.2661687>
- [28] Dirk Merkel. 2014. Docker: Lightweight Linux Containers for Consistent Development and Deployment. *Linux J.* 2014, 239, Article 2 (March 2014).
 - [29] 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.
 - [30] 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>
 - [31] 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.
 - [32] Mariano Pimentel, Hugo Fuks, and Carlos JP Lucena. 2008. Um processo de desenvolvimento de sistemas colaborativos baseado no Modelo 3C: RUP-3C-Groupware. *Anais do IV SBSI. SBSI 7* (2008).
 - [33] 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.
 - [34] 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.
 - [35] 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.
 - [36] Edeilson M Silva, Ricardo A Costa, Mario Godoy Neto, Robson YS Oliveira, and Silvio RL Meira. 2009. Promovendo melhorias na comunicação e colaboração em uma plataforma de gestão de conhecimento através de recomendações. *V Simpósio Brasileiro de Sistemas de Informação (SBSI)* 46 (2009).
 - [37] Giuseppe Silvestri, Jie Yang, Alessandro Bozzon, and Andrea Tagarelli. 2015. Linking Accounts across Social Networks: the Case of StackOverflow, Github and Twitter.. In *KDWeb*. 41–52.
 - [38] 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.
 - [39] P.a Thongtanunam, C.a Tantithamthavorn, R.G.b Kula, N.c Yoshida, H.a Iida, and K.-I.a Matsumoto. 2015. Who should review my code? A file location-based code-reviewer recommendation approach for Modern Code Review. *2015 IEEE 22nd International Conference on Software Analysis, Evolution, and Reengineering, SANER 2015 - Proceedings* (2015), 141–150. <https://doi.org/10.1109/SANER.2015.7081824>
 - [40] Bogdan Vasilescu, Vladimir Filkov, and Alexander Serebrenik. 2013. Stackoverflow and github: Associations between software development and crowdsourced knowledge. In *Social computing (SocialCom), 2013 international conference on*. IEEE, 188–195.
 - [41] 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>
 - [42] 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.
 - [43] 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>
 - [44] 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.