Pesquisa Científica

Carlos Henrique Pacheco de Souza

Análise Exploratória de Dados em Redes de Compartilhamento de Conhecimento

Apresentado como requisito da disciplina de Monografia em Sistemas de Informação do DCC / UFMG.

Universidade Federal de Minas Gerais

Instituto de Ciências Exatas

Departamento de Ciência da Computação

Orientador: Clodoveu Augusto Davis Junior

Departamento de Ciência da Computação

Belo Horizonte

2017/1

Sumário

Introdução

O desenvolvimento de software apresenta grandes desafios que vão além do conhecimento técnico. Um problema muito comum ocorre durante a escolha de quais tecnologias e ferramentas utilizar. Diversos estudos são realizados na busca da combinação perfeita. Porém, alguns fatores importantes como os aspectos sociais e culturais são ignorados.

Nenhum sistema é criado sem a utilização de mão de obra especializada. Assim, o custo do projeto e consequentemente o seu sucesso, são influenciados pela demanda desses profissionais. Portanto, compreender como está a distribuição do conhecimento técnico nas diferentes regiões tem grande importância para essa tomada de decisão.

A grande utilização das redes sociais de compartilhamento de conhecimento, permite que seja realizada uma nova abordagem na análise de seus dados. Para isso, utilizaremos dados geográficos na visualização do comportamento de duas das maiores redes que foram selecionadas para esse trabalho devido a sua complementaridade proposta por Vasilescu, Filkov e Serebrenik (2013).

- GitHub é um serviço de hospedagem web para projetos que usam o controle de versionamento Git. Ao contrário do Git, que é uma ferramenta que opera estritamente na linha de comando, o GitHub fornece uma interface gráfica baseada na Web, bem como a integração com dispositivos móveis. Ele também fornece controle de acesso a vários recursos de colaboração, tais como rastreamento de bugs, solicitações de recursos, gerenciamento de tarefas e wikis por projeto. Em 2016 ele conta com aproximadamente 14 milhões de usuários e mais de 35 milhões de repositórios, tornando-se o maior servidor de código fonte no mundo(WIKIPEDIA, 2016a).

- Stack Overflow é um site que apresenta perguntas e respostas numa grande quantidade de tópicos de programação de computadores. Ele serve como plataforma para que os usuários façam perguntas e também as respondam. Mas além disso, podem através do registro e da participação ativa, que os usuários votem em questões e respostas mais ou menos úteis. O fechamento de perguntas é a principal diferenciação da ferramenta que previne perguntas com baixa qualidade. Em 2016 ele conta com aproximadamente 6 milhões de usuários e mais de 11 milhões de perguntas(WIKIPEDIA, 2016b).

Temos como objetivo principal, exibir a concentração relativa dos usuários por tópicos e ou períodos de interação nas redes. Bem como, identificar as regiões que produziram informações (códigos e ou respostas) de melhor qualidade nas redes. Para mensurar este fator, podemos utilizar o índice CPDScorer (HUANG et al.,). Dentre outras análises que também serão realizadas, podemos destacar a identificação dos lugares onde cada uma das redes sociais

possui maior presença relativa, além de realizar um comparativo entre as taxa de utilização e abandono das redes sociais por região.

Como objetivo secundário realizaremos um estudo sobre os lugares que mais introduzem novas tecnologias às ferramentas e os que demoram a adotá-las. Além de verificar, através de uma visão geográfica.

O presente trabalho traz consigo diversos desafios, desde a aplicação de conhecimentos obtidos em outras disciplinas como: Mineração de Dados, Banco de Dados Avançados(NoSQL), Processamento de Linguagem Natural, Algoritmo e Estrutura de Dados(Paralelismo e Programação Dinâmica) e Banco de Dados Geográficos. Além de integrar diversas tecnologias(Shell Script, RegEx, Python, Java, MongoDB e MySQL) e interfaces de programação de aplicativos(APIs - MapBox ou Google Maps).

1 Referencial Teórico

Com o surgimento de redes sociais como StackOverflow e GitHub, uma vasta quantidade de informações de desenvolvimento é criada diariamente. Tais dados de contexto pessoal e social tem um enorme potencial para apoiar na avaliação automática e eficaz da capacidade do desenvolvedor. O CPDScorer aborda uma modelagem e avaliação da capacidade de programação dos desenvolvedores através da mineração de informações heterogêneas, presentes nas comunidades do StackOverflow e GitHub. CPDScorer considera tanto a qualidade de resposta sobre tópicos de programação do StackOverflow e na qualidade do código fonte do projeto no GitHub. Um algoritmo de extração de termos da capacidade de programação também é projetado para rotular cada resposta e projeto com um termo de habilidade (HUANG et al.,).

A complementaridade das ferramentas foi constatada num estudo que mostra que "committers"ativos do GitHub perguntam menos e fornecem mais respostas no StackOverflow. Assim como, "askers"ativos do StackOverflow distribui seu trabalho de uma maneira menos uniforme que os outros (VASILESCU; FILKOV; SEREBRENIK, 2013). Bem como comportamentos semelhantes entre usuários das redes diferentes foram relatados por Badashian et al. (2014) ao verificar que desenvolvedores ativos contribuem para as principais atividades da plataforma (ou seja, realizar commits no GitHub e responder no StackOverflow), mas também se envolvem em outras atividades gerenciais (como gerenciamento de problemas no GitHub e voto de qualidade no StackOverflow).

Geocodificação é um conjunto de métodos capazes de transformar descrições em coordenadas geográficas. Endereços urbanos são uma das principais formas de expressão da localização geográfica em cidades. Muitos sistemas de informação incluem atributos para receber endereços e, assim, contam com uma referência espacial indireta. A obtenção de coordenadas a partir de endereços é um dos métodos de geocodificação mais importantes, mas é dificultada por variações comuns no endereço, como abreviações e omissão de componentes. No caso de endereços, existe uma expectativa de detalhamento hierárquico, com componentes que indicam o país, o estado e a cidade. O formato de apresentação desses componentes varia de país para país, e em muitas situações, alguns componentes são intencionalmente omitidos ou simplificados(MARTINS; JR; FONSECA, 2012).

Para contornar essa variabilidade na formação dos endereços, uma solução consiste na divisão do método em três passos ou estágios, sendo que cada estágio possui tarefas e interfaces de entrada e saída bem definidas. O primeiro estágio, chamado de "parsing", consiste na análise léxica que leve em conta as peculiaridades da estrutura de endereços do local ou país e posterior conversão da entrada textual contendo o endereço em uma estrutura de dados genérica. Essa estrutura de dados contém um número finito de atributos, que correspondem a

cada componente do endereço. O segundo estágio, chamado de "matching", recebe a estrutura de dados e realiza buscas em um banco de dados de referência, comparando valores por casamento exato ou aproximado de palavras e números, e definindo a melhor solução em caso de casamento parcial. Nesta fase, utiliza-se Levenshtein distance (NAVARRO, 2001) e Shift-And(WU; MANBER, 1992). O estágio seguinte, denominado "locating", consiste em recuperar as referências obtidas e extrair delas as coordenadas desejadas(JR; FONSECA, 2007).

2 Metodologia

O projeto pode ser dividido em 5 fases que são sequenciais e, devem seguir a ordem em que foram postadas.

1. Obter informações das bases de dados de ambas as redes sociais.
   Fazer o download de um backup do banco mais recente.

2. Mapear e extrair os dados pertinentes ao problema.
   Identificar a estrutura e elaborar consultas para relacionar e extrair apenas as informações que serão utilizadas.

3. Realizar um pré-processamento dos dados.
   Obter a localização geográfica dos usuários. Neste passo utilizaremos conceitos identificados no referencial teórico.

4. Desenvolver uma interface para visualização dos dados na forma geográfica.

5. Estudar a correlação dos conteúdos para algumas determinadas regiões.

3 Resultados Esperados

Pretendemos disponibilizar uma interface que permita, através de uma visão geográfica, analisar as diversas informações que as redes sociais podem oferecer. Além de tentar identificar algum padrão na correlação dos dados.

Outra oportunidade se faz no processo de geocodificação, onde podemos testar novas ideias para identificar as coordenadas a partir de endereços informados de modo livre. Uma vez que este tipo de endereço representa grande parte dos dados na internet hoje.

4 Etapas e Cronograma

Etapa \Mês	08/16	09/16	10/16	11/16	12/16	01/17	02/17	03/17	04/17	05/17	06/17
Identificar tema e orientador											
Entendimento e pesquisa de referencial teórico											
Elaborar proposta											
Obter informações das bases de dados											
Mapear e extrair os dados pertinentes ao problema											
Elaborar apresentação parcial											
Realizar um pré-processamento dos dados											
Desenvolver uma interface para visualização dos dados											
Estudar a correlação dos dados											
Elaborar apresentação final											
Relatório final											

Referências

BADASHIAN, A. S. et al. Involvement, contribution and influence in github and stack overflow. In: IBM CORP. *Proceedings of 24th Annual International Conference on Computer Science and Software Engineering*. [S.l.], 2014. p. 19–33.

HUANG, W. et al. Cpdscorer: Modeling and evaluating developer programming ability across software communities.

JR, C. A. D.; FONSECA, F. T. Assessing the certainty of locations produced by an address geocoding system. *Geoinformatica*, Springer, v. 11, n. 1, p. 103–129, 2007.

MARTINS, D.; JR, C. A. D.; FONSECA, F. T. Geocodificação de endereços urbanos com indicação de qualidade. *Proceedings XIII GEOINFO*, p. 36–41, 2012.

NAVARRO, G. A guided tour to approximate string matching. *ACM computing surveys (CSUR)*, ACM, v. 33, n. 1, p. 31–88, 2001.

VASILESCU, B.; FILKOV, V.; SEREBRENIK, A. Stackoverflow and github: Associations between software development and crowdsourced knowledge. In: IEEE. *Social Computing (SocialCom), 2013 International Conference on*. [S.l.], 2013. p. 188–195.

WIKIPEDIA. *GitHub — Wikipedia, The Free Encyclopedia*. 2016. [Online; accessed 9-September-2016]. Disponível em: ⟨https://en.wikipedia.org/w/index.php?title=GitHub&oldid=738492841⟩.

WIKIPEDIA. *Stack Overflow — Wikipedia, The Free Encyclopedia*. 2016. [Online; accessed 9-September-2016]. Disponível em: ⟨https://en.wikipedia.org/w/index.php?title=Stack_Overflow&oldid=736163294⟩.

WU, S.; MANBER, U. Fast text searching: allowing errors. *Communications of the ACM*, ACM, v. 35, n. 10, p. 83–91, 1992.