

Universidade Federal de Ouro Preto - UFOP
Instituto de Ciências Exatas e Biológicas - ICEB
Departamento de Computação - DECOM

COLETA E MINERAÇÃO DE DADOS DE REDES SOCIAIS

Aluno: Milton Stilpen Júnior
Matricula: 07.1.4171

Orientador: Álvaro Rodrigues Pereira Júnior

Ouro Preto
26 de setembro de 2011

Universidade Federal de Ouro Preto - UFOP
Instituto de Ciências Exatas e Biológicas - ICEB
Departamento de Computação - DECOM

COLETA E MINERAÇÃO DE DADOS DE REDES SOCIAIS

Proposta de monografia apresentada ao curso de Bacharelado em Ciência da Computação, Universidade Federal de Ouro Preto, como requisito parcial para a conclusão da disciplina Monografia I (BCC390).

Aluno: Milton Stilpen Júnior
Matricula: 07.1.4171

Orientador: Álvaro Rodrigues Pereira Júnior

Ouro Preto
26 de setembro de 2011

Resumo

As redes sociais sempre foram alvo de pesquisa, principalmente por parte dos sociólogos. Hoje, com o decorrente poder computacional o qual nos encontramos, a Web tornou-se capaz de conectar as diversas partes do globo, possibilitando a coleta e a análise de um rico universo de dados da massa de usuários que nela navegam. Assim, o estudo de redes sociais tem, dia após dia, se mostrado factível e valoroso para os pesquisadores da Ciência da Computação. Dessa forma, o que será proposto nesse trabalho de conclusão de curso é um *framework* para coleta e análise de dados de redes sociais.

Palavras-chave: Recuperação de Informação na Web. Redes Sociais. Engenharia de Software.

Sumário

1	Introdução	1
2	Justificativa	2
3	Objetivos	3
3.1	Objetivo geral	3
3.2	Objetivos específicos	4
4	Metodologia	6
5	Cronograma de atividades	7

Lista de Figuras

1	Arquitetura proposta	4
---	--------------------------------	---

Lista de Tabelas

1	Cronograma de Atividades.	7
---	-----------------------------------	---

1 Introdução

Encontra-se, na atualidade, uma explosão de informações graças à existência de diversas redes em escala mundial e de "gigantes" que trabalham os dados oriundos destas, como, a título e exemplo, a *Google* e o *Facebook*. Pode-se admitir, em razão do exposto, que pesquisar sobre determinado indivíduo no mundo contemporâneo é uma tarefa relativamente comum. Isso porque, a partir de um conhecimento prévio sobre a utilização de computadores e de acesso a internet, chega-se a várias informações sobre uma pessoa simplesmente através de seu nome. Seus gostos, interesses, cultura etc. E pode-se ir além, caso seja possível saber as suas interações com outras pessoas!

Nem sempre, no entanto, foi assim. Outrora, a dificuldade de se obter dados era assustadora se comparada com o agora. Obstáculos geográficos, pessoais, culturais, entre outros, atrapalhavam uma análise mais abrangente de uma rede social. A coleta de informações normalmente era feita por questionários e demandava tempo. Seu fim principal, contudo, não deixou de ser o mesmo: possibilitar o estudo aprofundado do comportamento humano. No caso das redes sociais online, o estudo dos dados de uso de um usuário podem trazer diversas oportunidades. Ferramentas de tomada de decisão de acordo com o que é dito nas redes, melhorias no design de interação de um sistema, ricos estudos de interação social [2] etc.

Logo, nós, cientistas da computação, através de algoritmos e estruturas de dados, temos a capacidade desenvolver uma ferramenta capaz de coletar do provedor web, e, a partir dessa, estruturar e analisar os dados coletados. É o que propomos nesse projeto. Desenvolver um *framework* extensível, com fraco acoplamento e de propósito geral para a coleta e análise de redes sociais.

2 Justificativa

Porque os dados de redes sociais são importantes? Consoante Nielsen Online[4], a cada cinco usuários ativos da internet, quatro visitam redes sociais. Americanos gastam mais tempo no Facebook do que em qualquer outro site dos Estados Unidos. Lá, aplicativos de redes sociais estão entre os três mais utilizados em smartphones. E podem-se ver esses dados repetirem no dia à dia. O mundo encontra-se interligado, através de várias redes sociais online. Dois terços da população online global visita ou participa de tais redes. O crescimento dessa área trouxe consigo o interesse dos cientistas, que começaram agora a pesquisar sobre o tema. O seu estudo em si ainda está na sua "infância"[1].

Porque criar um *framework* e não uma aplicação? O foco do estudo não é algo específico, haja vista o surgimento e desenvolvimento crescente de Interfaces de Programação de Aplicações; e o fato de a maioria das redes sociais poder ser representada por meio de grafos, sendo os vértices referentes aos usuários e suas arestas às conexões entre estes (seja por amizade, "seguidor" ou vídeos relacionados, entre outros), conhecidas como APIs.

A ideia de desenvolver um *framework* que encapsule várias ferramentas que possam modelar uma rede social a partir de sua API específica e que contenha utilitários como interpretadores, máquinas de pesquisa textual, agregador de valores, gráficos e mapas se mostra, portanto, significativamente interessante.

3 Objetivos

3.1 Objetivo geral

O objetivo, em linhas gerais, é chegar ao final da Monografia II (BCC 391) com um *framework* sólido bem como algumas aplicações de análise de algumas redes sociais já implementadas. Os componentes que o tal oferecerá vão da coleta até a visualização de gráficos e mapas, como é visto no diagrama abaixo¹. Por ser o primeiro desenho feito para a solução proposta, podem e devem acontecer alterações na mesma, sempre com a intenção de melhorar. Todos os módulos devem ser estendidos e configurados para a criação da aplicação. Na próxima seção serão explicados cada parte da solução proposta.

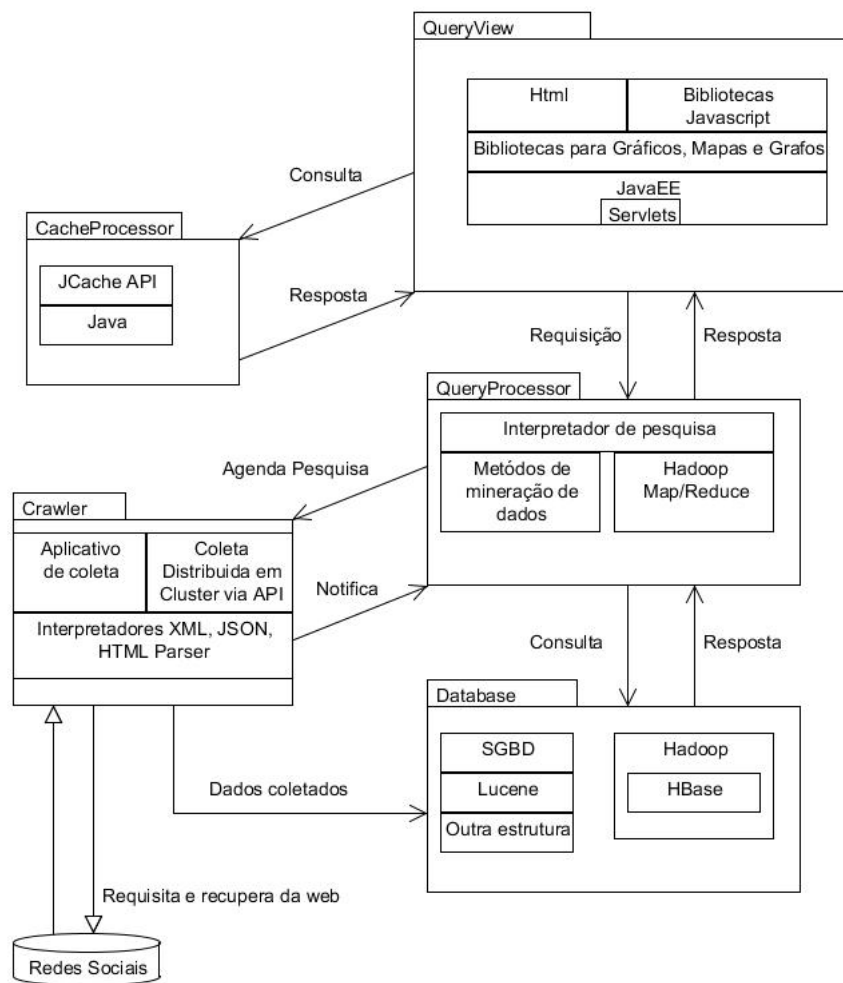


Figura 1: Arquitetura proposta

3.2 Objetivos específicos

- **Database:** O módulo de base de dados, nomeado Database, terá encapsulado ferramentas para armazenamento de dados estruturados e não estruturados. Pode-se citar a utilização de, por exemplo, SGBDs e *Apache HBase*, para dados estruturados, distribuídos ou não, e *Apache Lucene* para dados não estruturados.
- **Crawler:** O módulo de coleta, nomeado Crawler, será responsável por prover classes genéricas para coleta de usuários e suas informações, como mensagens públicas, geolocalização, possível lista de amigos ou seguidores através de APIs[5], primeiramente. Conterá parsers de HTML, interpretadores XML, JSON etc. Seu fluxo de coleta será endereçado diretamente para o módulo Database, que foi configurado pelo desenvolvedor do aplicativo. Podem existir diversas formas de coleta como, por exemplo, coleta distribuída[3] e coletas por aplicativo. Também serão disponíveis métodos de inserção manual de informações.
- **QueryProcessor:** O módulo de processamento de requisições de dados, nomeado QueryProcessor, terá encapsulado ferramentas de mineração de dados e agregadores de valores. O ideal é que esse componente se encontre em um ambiente

distribuído, como *Apache Hadoop*, para paralelizar seus pesados processos de mineração em cima da extensa base.

- QueryCache: Levando em consideração que o processamento pode demorar minutos, e até horas, o módulo de cache, denominado QueryCache, existe para diminuir o tempo de resposta de pesquisas feitas pelo usuário. Sua funcionalidade, a princípio, é armazenar resultado de consultas mais frequentes.
- QueryView: O módulo responsável pela camada View de uma aplicação, intitulada QueryView, conterá ferramentas de desenho de gráficos e mapas, de acordo com as funções requisitadas ao QueryProcessor. Extendida pelo desenvolvedor, deve gerar a interface gráfica web do aplicativo desenvolvido.

4 Metodologia

A Metodologia utilizada nesse projeto para o desenvolvimento do *framework* é muito usual em Engenharia de Software, o Scrum. Esse processo é iterativo e incremental, e será executado de maneira adaptada. Será desenvolvido primeiramente um *backlog* e, a partir deste, ocorrerá a divisão de tarefas para o planejamento de *sprints* a longo prazo. Todo *sprint* tem um entregável ao final, que será uma parte do *framework* testada e com uma parte da aplicação de exemplo, validando o módulo. Os *sprints* irão variar de 2 a 3 semanas.

Porém, como um primeiro *sprint*, é necessário estudar outros *frameworks* e bibliotecas que serão encapsulados para prover as funcionalidades desejadas no *framework* em questão. Além de implementar o modelo da própria arquitetura, em questão de código fonte. Assim, para os próximos *sprints*, podemos ter entregáveis de dois módulos por vez e, para o último, o módulo QueryView seria entregue sozinho.

Dessa maneira, ao final desse *backlog*, já teremos um *core* do *framework* com uma aplicação desenvolvida em cima deste. A aplicação será da rede social *Twitter*, pelo fato de possuir a API mais utilizada atualmente, ou seja, possuir uma API robusta como guia.

5 Cronograma de atividades

Na Tabela 1, o planejamento para o desenvolvimento do *framework*.

Atividades	Ago	Set	Out	Nov	Dez
Definição do tema	X				
Estudo sobre coletas em redes sociais	X	X			
Construção da arquitetura proposta			X		
Implementação e testes de uma aplicação e escrita do relatório final			X	X	X
Apresentação do Trabalho					X

Tabela 1: Cronograma de Atividades.

Referências

- [1] Fabrício Benevenuto, Jussara M. Almeida, and Altigran S. Silva. Explorando redes sociais online: Da coleta e análise de grandes bases de dados às aplicações, 2011.
- [2] Fabrício Benevenuto, Tiago Rodrigues, Meeyoung Cha, and Virgílio Almeida. Characterizing user behavior in online social networks. *IMC '09 Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, 09, 2009.
- [3] Duen Horng Chau, Shashank Pandit, Samuel Wang, and Christos Faloutsos. Parallel crawling for online social networks. *WWW 2007 / Poster Paper*, 2007.
- [4] Nielsen Online. State of the media: The social media report q3 2011. Technical report, Nielsen/McKinsey Company, 2011.
- [5] Matthew A. Russel. *Mining the Social Web*. O'Reilly, 1 edition, 2011.