8º FÓRUM ENSINO · PESQUISA EXTENSÃO · GESTÃO FEPEG
UNIVERSIDADE: SABERES E PRÁTICAS INOVADORAS
Trabalhos científicos · Apresentações artísticas e culturais · Debates · Minicursos e Palestras
REALIZAÇÃO: Unimontes
APOIO: FAPEMIG FADENOR
24 a 27 setembro
Campus Universitário Professor Darcy Ribeiro
www.fepeg.unimontes.br

Modelagem de um Sistema de Recomendação de Conteúdo com Base em Palavras-Chave Extraídas pelo Algoritmo TKG

Carlos Henrique Miranda Rodrigues

Introdução

A facilidade de acesso à informação propiciada pela internet tem se tornando cada vez maior. Através da internet as pessoas são capazes de compartilhar suas ideias, suas opiniões, encontrar pessoas, procurar produtos que elas estejam interessadas em comprar e diversas outras ações. Dentro desse contexto se encontram as redes sociais virtuais que são espaços nos quais os usuários podem acessar compartilhar fatos cotidianos com outros usuários, discutir sobre um determinado assunto, acompanhar notícias e se relacionar com outros usuários através da internet. Um serviço que tem se popularizado cada vez mais é o serviço de *microblog* Twitter, que permite que os seus usuários compartilhem mensagens e se comuniquem com outros usuários, desde que essas mensagens não ultrapassem 140 caracteres.

Este serviço vem cada vez mais se popularizando, visto que, cada vez mais pessoas e/ou empresas utilizam o *microblog* para informar aos demais usuários da rede sobre o que pensam ou o que creem ser interessante o suficiente para ser compartilhado com todos os demais usuários. Um problema que pode ser identificado nesse ponto é o fato de que quanto maior a quantidade de usuários deste tipo de serviço, maior será a quantidade mensagens que será gerada, o que pode gerar um aspecto um tanto quanto inconveniente aos seus usuários, pois a quantidade de informações geradas pelo serviço é muito maior do que o tempo disponível dos usuários para absorver todo este conteúdo. De acordo com o portal *Statistic Brain*[3], em uma pesquisa realizada em 1 de Janeiro de 2014, o número de usuários ativos do Twitter era de pouco mais de 645 milhões. Além disso, ainda de acordo com a pesquisa, estes 645 milhões de usuários demoram cerca de 5 dias para gerar um montante de 1 bilhão de *tweets* (mensagens do Twitter que podem ter no máximo 140 caracteres).

Uma forma de filtrar todo este volume de informação é através da aplicação de sistemas de recomendação que identificam itens que possam ser de interesse do usuário sem que para isso ele necessite navegar a procura da informação mais relevante. Estes sistemas utilizam de informações relacionadas aos usuários, que são previamente registradas, para tentar prever qual conteúdo possa estar mais de acordo com as características do usuário que possui.

Uma característica abordada no desenvolvimento de sistema de recomendação de conteúdo são as palavras-chave utilizadas pelos usuários. De modo, que o sistema de recomendação de posse das palavras-chave utilizadas pelo usuário busque por conteúdos relacionados às palavras-chave encontradas. Sendo assim, o conteúdo encontrado possui uma grande chance de ser relevante ao usuário, pois possui relação com as palavras-chave que o usuário tem utilizado.

Material e métodos

A. Estudo da aplicação do algoritmo TKG

Foi realizado um estudo visando à compreensão do funcionamento do algoritmo TKG (método proposto para a extração de palavras-chave de um conjunto de *tweets*).

O método TKG é um processo extração de palavras-chave de documentos utilizando como base a teoria dos grafos. Este processo é realizado em três etapas, sendo que cada uma dessas etapas é composta por duas fases. Na primeira etapa é realizado o pré-processamento de uma coleção de tweets, os quais são submetidos às fases de Análise Léxica e Remoção de Stopwords. Na segunda etapa são estabelecidos os vértices e as arestas do grafo textual que representa essa coleção. Nessa etapa, os vértices são definidos na fase chamada Atribuição de Vértices, enquanto as arestas são definidas na fase de Atribuição de Arestas. Na terceira etapa, por fim, é realizada a extração de palavras-chave a partir do grafo textual segundo as fases de Cálculo de Centralidade e Ordenação dos Vértices. Para isso, primeiramente, medidas de centralidade são calculadas para cada vértice. Então, esses vértices são ordenados formando um ranking, no qual as primeiras posições são aquelas que apresentam possíveis palavras-chave[1].

B. Estudo dos métodos de requisição do Twitter API

Foi realizado um estudo visando identificar os métodos de requisição de informações do Twitter API para que fosse possível analisar quais os métodos poderiam ser utilizados no desenvolvimento do sistema de recomendação.

C. Especificação e modelagem do sistema de recomendação

Com base nas informações colhidas nas etapas anteriores foi proposta a modelagem de um sistema de recomendação de conteúdo com base nas palavras-chave extraídas pelo algoritmo TKG e os métodos de requisição do Twitter API para recuperação e busca de dados no Twitter.

Resultados

A. Descrição do Funcionamento do Sistema de Recomendação

O primeiro passo para o funcionamento do sistema é coletar as informações de acesso do usuário para que o sistema possa realizar a busca de informações dos *tweets* mais recentes do respectivo usuário. Em seguida, o sistema realiza a extração de palavras-chave, utilizando o algoritmo TKG. A próxima etapa é realizar uma busca no Twitter, utilizando as palavras-chave que foram obtidas no processo de extração, para tentar identificar outras postagens de outros usuários que se relacionem com as palavras-chave, usando novamente o Twitter API. A Figura 1 mostra um diagrama de fluxo de dados do sistema de recomendação proposto.

B. Especificação do Sistema de Recomendação

O sistema será chamado TwiCS, que é o acrônimo para Twitter Content Seeker. A Figura 2 exibe o diagrama de contexto do sistema de recomendação, no qual é possível visualizar as funcionalidades do sistema, bem como as interações do sistema com seus atores. Além disso, a Tabela 2 descreve cada uma das funções mostradas na Figura 2. A descrição dos atores que interagem com o sistema pode ser visualizada na Tabela 1.

Discussão

O sistema TwiCS é uma ferramenta que permite que um usuário do *microblog* Twitter encontre postagens de outros usuários que são similares as suas últimas postagens, com base em palavras-chave dos últimos *tweets* deste respectivo usuário. O objetivo desta ferramenta é tentar mostrar pessoas que, possivelmente, estejam discutindo sobre o mesmo assunto no Twitter, mesmo que elas não se conheçam ou não possuam nenhum vínculo no Twitter.

Como é citado na página oficial de ajuda do Twitter, em relação ao processo de recomendação de quais contas de usuários possam ser interessantes para um usuário específico se relacionar, as recomendações são feitas com "base nos tipos de contas que o usuário já está seguindo e em quem essas contas seguem"[2]. O sistema TwiCS, mencionado neste trabalho, propõe um abordagem um pouco diferente daquela citada na página oficial de ajuda do Twitter: uma análise com base nas palavras-chave dos últimos *tweets* do usuário, utilizando a própria API do Twitter para poder realizar as operações de recuperação de *tweets* e busca.

Uma vez que o sistema TwiCS somente utiliza de recomendação com base em informações disponibilizadas pelo próprio usuário através da interface de autenticação (API), ele não armazena nenhum dado em alguma base de dados. É um sistema que, para o escopo deste trabalho, não houve a necessidade de armazenamento de qualquer dado.

Conclusão/Conclusões/Considerações finais

A utilização do serviço de *microblog* Twitter proporciona ao seus usuários um grande volume de informações, como evidenciado no capítulo 1. No entanto, é praticamente impossível que um usuário consiga absorver todo este conteúdo de maneira rápida, por exemplo, se realizarmos um cálculo básico e dividirmos a quantidade de *tweets* gerados em um dia, 58 milhões, e dividirmos pelo número de horas no dia, 24, iremos obter o valor 2 milhões, um usuário teria que ler pouco menos de 2 milhões de *tweets* por hora para conseguir ler todas essas mensagens em um dia.

Entretanto existem algumas técnicas utilizadas pelo Twitter para tentar filtrar o conteúdo que os usuários possam estar mais interessados. Essas técnicas envolvem recomendação de usuários com base nos seus seguidores e nos seguidores de seus seguidores. No entanto, existe a possibilidade de que um usuário siga outro usuário, porém eles estejam discutindo sobre temas completamente diferentes nos últimos 2 meses, por exemplo. A estratégia analisada neste trabalho tem por objetivo mostrar pessoas que possam estar discutindo sobre assuntos semelhantes, de modo que elas possam compartilhar conhecimento e experiências em relação ao assunto que elas estão discutindo.

Referências

[1] ABILHOA, W, D. **Um Método para Extração de Palavras-chave de Documentos Representados em Grafos**. 2013. 82f. Dissertação (Mestrado em Engenharia Elétrica) – Universidade Presbiteriana Mackenzie, São Paulo, 2013.

[2] **Help Center Twitter**. 2014. Disponível em < https://support.twitter.com/articles/227220-twitter-s-suggestions-for-who-to-follow>. Acesso em: 20 abr.2014.

[3] **Twitter Statistics**. 2014. Disponível em: <http://www.statisticbrain.com/twitter-statistics/>. Acesso em: 09 abr. 2014.

Tabela 1. Atores do sistema

Número de ordem	Ator	Definição
1	Usuário	Usuário que possua uma conta no Twitter
2	Twitter API	Interface de acesso às funcionalidades do Twitter. Esta interface fornece um conjunto de métodos que possibilitam o acesso a informações do microblog.
3	Twitter *Search* API	Interface de acesso às funcionalidades relativas a busca do Twitter.

Tabela 2. Funções do sistema

Número de ordem	Caso de Uso	Definição
1	Fazer Login	Processo de realização de login de acesso do usuário ao sistema. Este processo irá utilizar do Caso de Uso Autenticar Usuário.
2	Autenticar Usuário	Autenticação de usuários com base em dados de acesso ao Twitter
3	Buscar *Tweets* do Usuário Logado	Com base nos dados do usuário logado, o sistema realizará uma requisição à API do Twitter para obter os 50 *tweets* mais recentes do respectivo usuário
4	Extrair Palavras-chave	De posse dos 50 *tweets* do usuário logado o sistema deverá extrair as palavras-chave desse conjunto de *tweets.*
5	Buscar Palavras-chave	Uma vez que o sistema tenha as palavras-chave, ele deve realizar uma busca através do *Twitter Search API*, utilizando as palavras-chave como parâmetros de busca.
6	Visualizar recomendações	Apresentação dos resultados da busca realizada através do *Twitter Search API*, utilizando as palavras-chave encontradas pelo sistema.

Figura 1. Diagrama de fluxo de dados do sistema

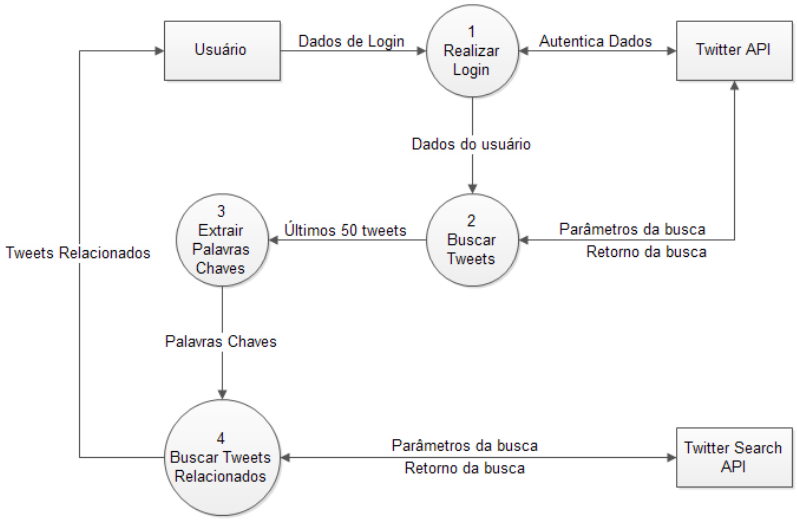

Figura 2. Diagrama de contexto do sistema