Uso de Clustering no tratamento de dados para melhoria de desempenho e escalabilidade em Sistemas de Recomendação

Alisson Mittaraquis
DCOMP-UFS
São Cristovão -SE
alisson_rpgmaster@hotmail.com

Maria Augusta S. N. Nunes
DCOMP-UFS
São Cristovão -SE
gutanunes@dcomp.ufs.br

Christian Nunes Aranha
Cortex intelligence
Rio de Janeiro -RJ
christian.aranha@cortex-intelligence.com

RESUMO
Clustering é a classificação, supervisionada ou não, de dados em grupos (clusters) baseada em padrões. Os problemas de agrupamento de dados vêm sendo abordados em muitos contextos e por pesquisadores de diversas áreas. Este artigo apresenta uma visão geral do método de agrupamento *Clustering* de uma perspectiva de reconhecimento de padrões em perfis de usuários do Twitter, com o objetivo de fornecer melhorias na aplicação das técnicas de recomendação como desempenho, e escalabilidade.

Termos Gerais
Algoritmos, Desempenho, Fatores Humanos.

Palavras-Chave
Clustering, Sistemas de Recomendação, Mineração de Textos, Twitter, TwiterTrending Topics Brasil.

1. INTRODUÇÃO
O universo da web oferece uma infinidade de músicas, filmes, produtos para a venda, conteúdo informativo, jogos, entre muitos outros. Encontrar algo que seja interessante, entre tantas opções, vem se tornando uma tarefa cada vez mais difícil e fatídica para os usuários da internet. Os Sistemas de Recomendação surgiram em resposta a este problema, por exemplo, um Sistema de Recomendação de um site recomenda produtos que possam atender as necessidades dos usuários, músicas que possam interessá-lo, textos que gostaria de ler, etc.

Mas, apesar do sucesso dos Sistemas de Recomendação, surgiram dois grandes desafios na área :

(i) o primeiro é relativo ao grande aumento de usuários e de informações disponíveis na web, para que recomendações em tempo real pudessem ser geradas, se fez necessário melhorar a escalabilidade dos algoritmos;

(ii) e o segundo, é melhorar a qualidade das recomendações realizadas, pois uma vez que um cliente recebe uma recomendação que ele não julgue interessante, ele provavelmente não voltará a usar o sistema.

Na tentativa de propor alguma solução para esses desafios propõe-se esse trabalho. Nesse trabalho propõe-se a utilização de técnicas de análise e classificação prévia dos dados antes da aplicação de alguma técnica de recomendação à base de dados. Dentre as técnicas de análise de dados mais utilizadas, o *Clustering* (clusterização) tem se destacado.

Recentemente, a técnica de clusterização tem sido aplicada em uma variedade de tópicos e áreas. O uso de técnicas de clusterização pode ser encontrado em Reconhecimento de Padrões, como descrito em [2], Compressão de Dados, como descrito em [3], Classificação de Dados, como descrito em [4], além de áreas como Psicologia, Geociências, Medicina, etc. Com base nessas características pode-se dizer que a clusterização é uma técnica-base e ferramenta para outras técnicas, assim como as ciências exatas matemática, física ou estatística.

Neste trabalho é proposto então o uso de técnicas de clusterização para tratar os dados anterior à aplicação das técnicas de recomendação. A base de dados escolhida como objeto de pesquisa foi a base do Twitter Trending Topics Brasil que é um sistema de mineração de textos aplicado aos perfis de usuários do twitter do Brasil.

O artigo está organizado da seguinte forma: a seção 2 descreve uma visão geral sobre Sistemas de Recomendação e suas principais técnicas; a seção 3 apresenta o Twitter Trending Topics Brasil e sua base de dados. A seção 4 descreve em detalhes a técnica de clusterização escolhida. A seção 5 discute os resultados obtidos e por fim a seção 6 apresenta a conclusão e os possíveis trabalhos futuros.

2. SISTEMAS DE RECOMENDAÇÃO
Sistemas de Recomendação começaram a aparecer nos anos 90. O objetivo era reduzir a sobrecarga de informação exibida aos usuários, selecionando um subconjunto de itens de um conjunto universal com base em preferências de outros usuários.

Sistemas de Recomendação são aplicativos que fornecem sugestões personalizadas para os usuários sobre os itens que possam interessá-los. Como descrito em [5], pode-se classificar as técnicas de recomendação em 5 categorias distintas. São elas:

- **Baseado em Conteúdo**: recomenda itens que são similares aos preferidos pelo usuário no passado. Itens (produtos, serviços ou pessoas) são definidos pelas suas características associadas.

- **Filtragem Colaborativa**: recomenda itens que pessoas com gostos semelhantes e preferências gostaram no passado. O perfil de usuário consiste de itens e suas respectivas avaliações feitas pelo usuário.

- **Demográfico**: recomenda itens, considerando as características demográficas. O perfil de usuário consiste em dados pessoais d demográficos do usuário.

- **Baseado em Conhecimento**: recomenda itens com base em inferências a partir de preferências e necessidades do usuário. O perfil de usuário consiste no conhecimento funcional estruturado e interpretado de acordo com uma máquina de inferência.

- **Baseado em Utilidade**: recomenda itens a considerar a utilidade deles para os usuários.

Todas estas técnicas, para que funcionem adequadamente, elas precisam ser aplicadas a uma base de dados e dela, então, conseqüentemente, extrair as informações necessárias para realizar a recomendação. Note, que é interessante então que exista um tratamento nos dados da base (clusterização) para que esta busca seja feita somente nos dados que realmente contribuem para a recomendação.

3. TWITTER TRENDING TOPICS BRASIL

O Twitter Trending Topics Brasil é um serviço disponibilizado pelo CortexLabs da empresa Cortex Intelligence. A proposta do site é dar ao usuário do Twitter uma visão mais nacional do que rola nos tweets. A tecnologia praticada é o que chamamos de Twitter Mining [6]. O Twitter Mining é uma mineração de dados realizada numa base de dados que contem perfis de usuários do twitter assim como os posts que eles "twittam". O objetivo é classificar os twitters dos usuários em 13 temas centrais definidos pelo CortexLabs (Internet, Música, Globo, Esporte, Serviços, Jornalismo, Programas, Apresentadores, Blogs, Flashmobs, Humoristas, Pessoas e Política). Essa classificação cria um ranking dos usuários que possuem mais seguidores para cada categoria. Além disso, o Twitter Trending Topics Brasil tem também como objetivo fornecer uma lista alternativa à do Trending Topics oficial do Twitter, contendo associação entre as palavras mais twittadas por usuários brasileiros, e com isso trabalhar com uma abordagem brasileira do que está sendo mais twittado no Brasil.

Na base de dados cedida pela CortexLabs, estão contidas as informações de cada usuário, tais como nome, número de seguidores, número de usuários que está seguindo, descrição, número de identificação (id) e todas as postagens registradas.

O objetivo desse trabalho, é com base nessas informações, especificamente nas postagens de cada usuário e no número de seguidores, realizar uma pré-classificação, como descrito a seguir.

4. CLUSTERIZAÇÃO DA BASE DE DADOS

Como descrito anteriormente, uma cópia da base de dados, cedido pela CortexLabs, contendo informações de mais de trinta e dois mil usuários, está sendo utilizada nos experimentos de clusterização proposto por esse artigo.

O intuito do uso de clusterização é reduzir o espaço de busca quando forem aplicadas técnicas de recomendação e, então comparar os resultados obtidos.

O algoritmo ora proposto cria clusters compostos por grupos de usuários que têm preferências/características similares. As predições para um usuário são criadas por uma média de opiniões de os outros usuários desse cluster [5]. O agrupamento é realizado a partir de similaridades ou distâncias entre seus componentes (dissimilaridades). Os únicos pré-requisitos são medidas de similaridade ou dados sob os quais possam ser calculadas similaridades [7].

A técnica de clusterização adotada foi a Clusterização Conceitual Parcial baseada em Protótipos Exclusiva Particional. Ou seja, é conceitual por que a similaridade é obtida por conceitos e significados dos dados e não por medidas de distância numéricas, logo é subjetiva aos objetos e objetivos. Parcial por que nem todos os dados farão parte de um cluster. Os dados que não se enquadram em nenhuma das categorias são descartados (ruído). Baseada em Protótipos (centróides), pois são criados clusters iniciais com um elemento que contém em sua descrição as palavras-chaves mais relevantes a cerca do conceito que ele representa, exemplo apresentado na figura 1. Exclusivo, pois cada elemento só pode pertencer a um único cluster. E, Particional por que não existe hierarquia entre os clusters (sub-clusters).

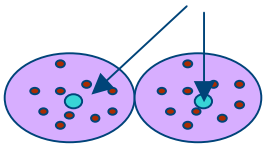

Figura 1. Dados agrupados a partir da similaridade com o Centróide.

Inicialmente era desejado que o algoritmo identificasse as palavras chaves e criasse os clusters de forma automática (classificação não supervisionada), mas a construção do algoritmo se mostrou consideravelmente complexa. Então, por questão de simplicidade, foi adotada a abordagem supervisionada baseada em centróides supracitada. Seguindo uma estrutura similar a do Twitter Trending Topics Brasil, a

base de dados é particionada em 9 clusters com os temas centrais: música, esporte, jogos, internet, seriados, profissões, religião, línguas e política. Cada cluster recebe um elemento centróide que tem em sua descrição cerca de vinte palavras-chave (definidas manualmente) relevantes ao tema. A partir daí os dados da base são confrontados com os centróides e então são adicionados aos devidos clusters.

5. RESULTADOS EXPERIMENTAIS

O algoritmo de clustering foi implementado em Java no ambiente de desenvolvimento Eclipse. O aplicativo foi executado no sistema operacional Windows XP 32 bits Service Pack 3, sobre a base de dados e gerou 9 clusters com os temas supracitados em aproximadamente 14,6 segundos. A máquina utilizada para os testes tem a seguinte configuração: processador com dois núcleos e freqüência total 3.2 GHz, 2 GigaBytes de memória RAM com freqüência 667 MHz.

Os resultados da clusterização foram:

Cluster	Itens Clusterizados
Internet	2087
Línguas	1796
Música	1691
Profissões	1306
Esporte	891
Religião	378
Jogos	177
Séries	169
Política	136
Total	8631

Tabela 1. Resultado da clusterização da base de dados.

Ao total foram clusterizados 8.631 itens dos 32.816 contidos na base, ou seja, 26,30% dos itens da base foram adicionados a algum dos clusters durante o processo, enquanto que os 73,70% restantes foram considerados ruído e foram desconsiderados.

A percentagem de itens clusterizados foi relativamente baixa principalmente por que aproximadamente 12.600 usuários não digitaram nada na descrição de seus perfis (campo *description* com valor *null*) ou digitaram caracteres aparentemente sem sentido.

Se desconsiderarmos esses itens (de fato ruído) da base de dados, a percentagem de dados clusterizados em relação a itens válidos passa para aproximadamente 42,7%. O menor dos clusters formados (cluster Política) representa aproximadamente 0,41% de todos os dados da base, e o maior cluster (cluster Internet) representa aproximadamente 6,36% dos mesmos.

6. CONCLUSÃO E TRABALHOS FUTUROS

A clusterização se mostrou uma técnica bastante útil e apropriada para tratamento de dados e melhoria de escalabilidade, pois ao aplicarmos uma técnica de recomendação como, por exemplo, a filtragem colaborativa sobre os clusters gerados, teríamos uma sensível diminuição no espaço de busca. Novos usuários que se cadastrassem no sistema teriam também suas informações confrontadas com os centróides dos clusters e seriam então alocados em um destes. Além disto, mesmo que a base aumentasse largamente o número de dados, cada cluster representaria apenas uma pequena percentagem dela e a escalabilidade estaria mantida.

Contudo, os experimentos apontaram também uma dificuldade em classificar pessoas que não tivessem em sua descrição nenhuma das palavras-chave mantidas nos centróides, tendo em vista que mais da metade (≈ 57,3%) dos usuários ainda permaneceram sem classificação. Para amenizar o problema encontrado e aumentar a percentagem de usuários clusterizados, são propostas então como trabalho futuro algumas modificações no aplicativo de clusterização. São elas:

- Aumento do número de clusters, ou seja, adicionar mais temas centrais;
- Aumento do número e da qualidade (especificidade) das palavras-chaves contidas em cada centróide;
- Modificação do tipo da Clusterização Particional para Clusterização Hierárquica, introduzindo assim sub-clusters dos temos centrais. Exemplos: criar sub-clusters para os estilos musicais no cluster Música ou sub-clusters para categorias de esportes diferentes no cluster Esportes, etc.
- Utilização de clusterização automática (não supervisionada) para comparação de resultados.

Além disto, os resultados obtidos já demonstram algum potencial da técnica para ser utilizada em Sistemas de Recomendação de sites comerciais, por exemplo. Um site de vendas como o Submarino.com, que recomenda produtos apenas pelas compras e pesquisas já realizadas pelo usuário, poderia implementar um Sistema de Recomendação que efetuasse recomendações usando como base os resultados encontrados neste trabalho. Deste modo a submarino, por exemplo, poderia recomendar produtos para os usuários do Twitter (ex.: CDs e DVDs de música para usuários do cluster Música) ampliando conseqüentemente suas vendas ao alcançar novos nichos de consumidores.

7. REFERÊNCIAS

[1] SARWAR, B. M.; et al. Recommender Systems for Large-Scale E-Commerce: Scalable Neighborhood Formation

Using Clustering. 5th International Conference on Computer and Information Technology, 2002.

[2] C. Reina, U.M. Fayyad and P.S. Bradley. Initialization of iterative refinement clustering algorithms. In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98), pag. 194-198, 1998.

[3] K. Rose, E. Gurewitz, and G. C. Fox, "*Vector quantization by deterministic annealing*," IEEE Trans. Inform. Theory, vol. 38, nº 4, pag. 1249-1257, 1992.

[4] G. Fung and O. L. Mangasarian. Semi-supervised support vector machines for unlabeled data classification. Optimization Methods and Software, Abril 2001.

[5] Nunes, M. A. S. N. Recommender Systems based on Personality Traits: Could human psychological aspects influence the computer decision-making process?. 1. ed. Berlin: VDM Verlag Dr. Müller. v. 1. 140 p., 2009.

[6] Descrição do site do Twitter Trending Topics. Disponível em <www.twittertrandingtopics.com>. Acesso em 02 de junho de 2010.

[7] Kasznar, Istvan Karoly. Clustering – Agrupamento. Seção "Textos Quentes" do site da IBCI - Institutional Business Consultoria Internacional. Disponível em www.ibci.com.br. 2007.

[8] DO VALE, Marcos Neves. Agrupamento de Dados: Avaliação de Métodos e Desenvolvimento de Aplicativo para Análise de Agrupamento. Trabalho de conclusão de cursos – Pontifícia Universidade Católica do Rio de Janeiro, 2005.