Aplicações MapReduce – Recomendação baseada em conteúdo

PROFESSORA

Rosangela de Fátima Pereira – rpereira@larc.usp.br

DESCRIÇÃO

Esse documento descreve os passos necessários para a implementação e execução de uma aplicação de recomendação de produtos utilizando a biblioteca Java do Hadoop.

Recomendações são utilizadas para fazer sugestões sobre coisas que podem ser de interesse a alguém. Existem muitas formas de fazer recomendações:

Recomendação baseada em conteúdo: usa informação sobre o produto para identificar produtos similares. Por exemplo, pode-se utilizar por meio de categorias produtos similares e recomendá-los para os usuários.

Filtragem colaborativa: utiliza o comportamento do usuário para identificar similaridades entre produtos. Exemplo: se o mesmo usuário colocou uma nota alta para dois produtos, podemos inferir que há alguma similaridade entre esses dois produtos.

Nessa aplicação iremos utilizar uma base de dados da Amazon sobre produtos para fazer recomendações baseadas em conteúdo. Na base de dados, cada produto possui uma lista de itens similares calculado previamente pela Amazon. Nós utilizaremos esses dados para fazer as recomendações. A base de dados é extensa e por limitações de recursos computacionais, utilizaremos somente uma amostra dela. Porém você pode encontrar a base completa em: http://snap.stanford.edu/data/amazon-meta.html

A base de dados contém uma entrada para cada produto. Cada registro inclui um ID, título, categorização, itens similares a esses itens e informação sobre usuários que compraram e avaliaram os itens. Para fazer a captura desses dados, foi preciso necessário criar um novo formato de dados (Objeto InputFormat) para ler e estruturar os dados da base. Essa classe irá fazer a estruturação dos dados, emitindo os dados sobre cada produto Amazon como pares de chave-valor para a função map. Os dados sobre cada produto são representados como String e na classe UsuarioAmazon.java é incluído o código para gravar os dados de cada cliente.

Essa aplicação possui dois jobs. A tarefa map do primeiro MapReduce job recebe dados de cada produto como um par chave-valor. Quando a tarefa map recebe os dados do produto, ela emite o ID do cliente como sendo a chave e as informações do produto como o valor para cada cliente que comprou o produto. O Hadoop coleta todos os valores para a chave e invoca o reducer, que irá receber produtos que foram comprados pelo cliente. O reducer emitirá a lista de itens comprados por cada cliente, construindo um profile no campo valor e a quantidade de produtos comprados no campo chave. Para limitar o tamanho da base de dados, o reducer não apresentará na lista somente clientes que compraram mais que 5 produtos. O resultado desse job deverá ser uma lista com itens similares ao apresentado a seguir:

6 customerID=AST7UNIXVOAVA,review=ASIN=0783232063#title=Dead Men Don't Wear Plaid#salesrank=4766#group=DVD#rating=5#similar=6305308802|6305262225|0783230397|0783226799|B0001Z4P2I|,review=ASIN=0784011168#title=Jacob's Ladder#salesrank=2695#group=DVD#rating=5#similar=6305133131|B0001US62I|B00005V3Z4|B0002DB50E|078401213X|,review=ASIN=6305944288#title=Kronos#salesrank=4018#group=DVD#rating=3#similar=B00005R1O7|B00008G96N|B00006CXGG|B000063UR0|B00008975H|,review=ASIN=B00004W199#title=Carnival Of Souls#salesrank=25549#group=DVD#rating=5#similar=B0002V7O0Q|B00027JYLC|B0002DB50E|B00065GX64|B00004W3HE|,review=ASIN=B00005B1WS#title=House on Haunted Hill#salesrank=33305#group=DVD#rating=4#similar=B00009NHBC|B00000K3U3|B00005N5RQ|B00009NHB6|B00005AUK0|,review=ASIN=B00006FD9K#title=Godzilla King of the

Monsters#salesrank=13133#group=DVD#rating=4#similar=B00006FD9L|B00006FD9H|B00003IXDZ|B00006FD9G|B00006FD9I|

O segundo MapReduce job usa os dados gerados no primeiro job para fazer as recomendações para cada cliente. A tarefa map recebe dados sobre cada cliente como sendo o input e a recomendação é feita da seguinte forma:

a. Cada conjunto de dados de um roduto inclui itens similares a esse item. Dado um consumidor, primeiro a tarefa map cria uma lista de todos os itens similares para cada item que o cliente comprou.
b. Então a tarefa map remove qualquer item da lista de produtos similares que já foram comprados pelo mesmo cliente.
c. Então a tarefa map seleciona 10 itens para recomendar ao cliente

O resultado deverá conter registros similares ao apresentado a seguir:

AZ7LPCN9Y9SLJ [B00005JKFR, B00009NHC0, B000067FP3, B000063UR0, B00004RF9B, 6304698801, 6304698682, B00005NTNW, B00008CMT4, 0792842502]

AZI0O32W4ZYGH [B0006ZXTRK, B00008N6N5, 0061095508, 0671009443, B00004TQF7, 6304907613, 6305368171, B0000AOV4I, B00005S6K8, 6305696071]

AZOW89D0NXMUT [B00003CX95, B00003CWQU, B0001NBLVI, B00000JLWW, B0002J4ZWS, B00005JL78, 630587493X, 0767805712, B00005JL3T, B00005JN0T]

AZSN1TO0JI87B [B0000AOX0H, 0783230451, B0000AOX08, 0783232039, B0007PLLDI, B00004Y632, B00005OKQF, B00005LC4Q, 6305971099, B00006LPHA]

Na qual a chave é o ID do cliente e os valores são os IDs dos 10 produtos recomendados.

ATIVIDADES

1. Abra a IDE Eclipe.
2. Importe o projeto RecomendacaoProdutos a partir do caminho: /home/training/workspace
3. Após aberto o projeto, gere um jar da aplicação com o nome recomandacaoprodutos.jar
4. Abra o terminal dentro da VM da Cloudera
5. Verifique se você já possui o arquivo amazon-meta.txt no diretório entrada/amazon do HDFS
6. Caso o diretório e o arquivo ainda não exista, envie o arquivo /home/training/bases/amazon-meta.txt para o HDFS no diretório entrada/amazon
7. Execute o jar recomendacaoprodutos.jar passando como parâmetro a classe UsuariosFrequentesDriver e indicando o caminho de entrada e saída da aplicação.
8. Após terminar a execução do job, liste os arquivos do diretório de saída.
9. Verifique o conteúdo do diretório de saída.
10. Agora, execute novamente o jar recomendação produtos passando como parâmetro a classe RecomendacaoDriver e os diretórios de entrada e de saída. Nota: o caminho do arquivo de entrada deverá ser a saída da execução do job anterior.

11. Após terminar a execução do job, liste os arquivos do diretório de saída.

12. Verifique o conteúdo do diretório de saída.

Parabéns! Você concluiu as etapas de execução da aplicação de recomendação de produtos da Amazon.

BIBLIOGRAFIA:

PERERA, Srinath. **Hadoop MapReduce Cookbook**. Packt Publishing Ltd, 2013.