A. Rodrigues

Master in Informatics and Computing Engineering
Recomendação de Conteúdos: Aplicação de Agrupamento Distribuído a Conteúdos de TV
Alexandre José Monteiro Rodrigues
—

Abstract

With the explosive growth of multimedia content of the past few years, people are finding it increasingly more difficult to choose what is most relevant and most suited with their tastes. Recommender systems are software tools that can suggest items that users might like and play an important role because they help people choosing items to consume.

However, despite the popularity of these kind of systems, the development of an agnostic to the application domain is not trivial. This thesis considers the current applications of recommenders to the field of films and adapts it to the television programs domain, which is virtually unexplored in the literature. There are some challenges associated with the evolution of tastes over time and requirements of users of the system response time.

The study focus on the ways to improve the recommendation of content items (television programs and films) by combining various techniques and how the implementation can scale up with the increase of problem's size.

The study was conducted using the MinHash clustering technique that is a technique that links users to groups according to the similarity of the set of items they have seen before. This technique is complemented by Probabilistic Latent Semantic Indexing, which uses a mixture model to probabilistically represent the presence of sub-populations in a set of observations. The sub-populations are not identified a priori.

The system is fault tolerant, composed of several components that are on top of a distributed infrastructure. The architecture provides real-time recommendation requests and offline processing (using the MapReduce paradigm) of a set of observations that results in a set of partitions of users (communities) and in a probabilistic model which contains the affinity of content objects and users to a predefined number of latent classes, also considered as clusters. The solution also performs the observation accounting for each user-item interaction, keeping track of the observations by cluster. The statistics gathered are used in the calculation of a score for each item candidate for recommendation. The recommendation is a list of items, ordered by the calculated score.

The work is part of a context-aware services project of the company PT Inovação and the combination of techniques will be later applied to the MEO IPTV service dataset. Due to the bureaucratic process of access to confidential data, public datasets were used in the domain films, taking into account the nature of MEO's dataset.

The MinHash results are satisfactory. The parameters of this technique provide a good control of the number of clusters generated and the cover of the clusters in the universe of users.

The clusters are used in the recommendation calculation and we can conclude which kind of partition

scheme leads to better results for the dataset used. The responsiveness of the online components is also benchmarked and we can identify the points for improvement to be applied in production.

Resumo

Com o crescimento explosivo de conteúdos, as pessoas sentem cada vez mais dificuldade em escolher o que é mais relevante e o que mais se adequa aos seus gostos. Os sistemas de recomendação são ferramentas que sugerem itens que os utilizadores poderão gostar e desempenham um papel importante pois permitem ajudá-los a escolher com o mínimo de esforço os conteúdos ou itens a consumir.

No entanto, apesar da crescente popularidade destes sistemas, o desenvolvimento de uma aplicação agnóstica ao domínio não é trivial. Nesta dissertação considera-se a aplicação ao domínio dos filmes e reproduz-se a aplicação a um cenário de programas de televisão, que é praticamente inexplorado na literatura. Existem alguns desafios associados à evolução dos gostos dos utilizadores e aos requisitos de tempo de resposta do sistema.

O objectivo do trabalho é estudar como melhorar a recomendação de itens de conteúdo (programas de televisão e de filmes), combinando várias técnicas e assegurando a escalabilidade da aplicação.

O estudo foi realizado usando a técnica de clustering MinHash que associa os utilizadores a grupos de acordo com a semelhança do conjunto de itens que viram anteriormente. Esta técnica é complementada pela técnica Probabilistic Latent Semantic Indexing que recorre a um modelo mistura para modelar probabilisticamente a presença de sub-populações num conjunto de observações considerado, sem que as sub-populações estejam identificadas a-priori.

O sistema desenvolvido é tolerante à falha, composto por vários componentes que assentam numa infra-estrutura distribuída. A arquitectura contempla pedidos de recomendação em tempo real e um processamento offline (usando o paradigma MapReduce) de um conjunto de observações que resulta na partição de utilizadores por comunidades (clustering) e na modelação probabilística que determina a afinidade dos objectos de conteúdo e dos utilizadores a um número pré-definido de classes latentes, consideradas também como clusters. A contabilização de observações de cada item por cluster permite efectuar o cálculo de score a cada objecto candidato a recomendação e desta forma determinar a lista ordenada de objectos a recomendar.

O trabalho realizado está enquadrado num projecto de serviços baseados em contexto da PT Inovação e a combinação de técnicas estudadas será posteriormente aplicada a um conjunto de dados do serviço de IPTV (Internet Protocol Television) MEO. Devido ao processo burocrático de acesso a dados confidenciais, foram usados conjuntos de dados públicos no domínio de filmes, tendo em conta a natureza dos dados do conjunto da MEO.

Os resultados obtidos pela técnica MinHash são satisfatórios e a técnica permite controlar, através de parâmetros, o número de clusters pretendidos e a cobertura do universo de utilizadores. Os clusters são usados no cálculo de recomendação e determinaram-se quais os esquemas de partição que levam a melhores resultados para o conjunto de dados utilizado. A capacidade de resposta dos componentes online é também estudada, sendo identificados pontos de melhoria a serem aplicadas em produção.

Jury

- Chair: Luís Filipe Pinto de Almeida Teixeira
- External Examiner: Maria Benedita Campos Neves Malheiro
- Supervisor: João Correia Lopes
- Date: 13/7/2011