Fan, Wai Pio 23 September 2011

A survey of collaboration filtering

Introduction

Today's information accessed on Internet is increasing dramatically. Huge number of information is unstructured. Internet users aren't easy to access their enquired information. Collaboration filtering (CF) is a type of techniques involving collaboration among multiple agents, viewpoints, data sources, etc. It is a process of filtering for information. Collaboration filtering techniques have been applied on many different areas including Internet. For example, it have been used in multiple sensors' data, financial data in multiple sources, user data in electronic commerce (E-commerce) and web 2.0 applications. Collaboration filtering is common use for user data on Internet. This survey is focused on how to apply collaboration filtering techniques on user data on Internet. Collaboration filtering techniques are making an information filter based on user data such as the interests of a user which collecting preferences or taste information from many users. In everyday life, people rely on recommendations from other people by spoken words, reference letters, news reports from news media, general surveys, travel guides, and so forth. The approaches of collaboration filtering is based on an assumption of those who happened in past tend to happen in the future. There are batch of collaboration filtering approaches such as item-based collaborative filtering invented by Amazon.com, Google's PageRank and Netfilx's movie recommendation. The target of this survey is to describe and compare variant of collaboration techniques.

Social component

The social component of collaborative filtering is that the result filters are constructed by making use of the interests of many users provided in past via different behavior on internet. These users' behaviors are social related and the collaborative filtering build a filter to make use of the relationship of many Internet users' behavior.

Challenges

The main difficulties of CF are data sparsity, scalability, synonymy, gray sleep, shilling attack etc. Online shopping companies such as eBay and Amazon are large. The challenge of CF is to build an E-commerce recommender system to provide fast and accurate recommendations will attract the interest of customers and produce benefits to the companies. Such commercial recommender systems are constructed on very large product sets and customers' data. These large data sets are extremely sparse and the performance of the recommender systems is challenged. There are several reasons that the data is sparse. First when a new customer or product has just arrived to the systems. There is not enough information, so it is difficult to find the past information to predict and produce recommendations. It is even more difficult for the new product that can't be recommended until customer rated or purchased it. There should enough records to let the recommender system be convergence.

The second difficulty of CF is scalability. The large E-commerce companies such as eBay and Amazon have huge numbers of existing customers and products. These numbers are growing fast.

The third difficulty of CF is synonymy. It is because there are huge numbers of the same or very similar items to have different names or entries. CF systems should be able to recognize those relations among those data and thus treat these products similarly.

The fourth difficulty of CF is gray sleep. Gray sheep refers to the users's data isn't consisted with each other. Their opinions even disagree others. This isn't benefit from collaborative filtering.

The last difficulty of CF is shilling attacks. A CF systems should prevent some users may provide information that give tons of positive recommendations for their own materials and negative recommendations for their competitors. It is challenge for CF systems to introduce rules that discourage this kind of phenomenon.

Algorithm

CF systems can be built by looking for the same rating pattern or purchase history between users. Use these past users' data to calculate a prediction for the future users. This method of CF is user-based collaborative filtering. A specific application of this is the user-based Nearest Neighbor algorithm. There is another algorithm called item-based collaborative filtering invented by Amazon.com. It is to solve a problem of users who bought x also bought y in an item-centric manner.

There are mainly three types of techniques: memory-based, model-based and hybrid. Memory-based technique uses user data to compute similarity between users or items. This is used for making recommendation and easy to implement and is effective. Typical examples of this type are neighborhood based CF and item-based/user-based top-N recommendations.

The second type of technique is model-based. Models are developed using data mining, machine learning algorithms to find patterns based on training data. These are used to make predictions for real data. There are many model based CF algorithms. These include Bayesian Networks, clustering models, latent semantic models such as singular value decomposition, probabilistic latent semantic analysis, Multiple Multiplicative Factor, Latent Dirichlet allocation and markov decision process based models.

The last type of technique is hybrid. It combine of memory-based and model-based algorithm to overcome the disadvantages of using these techniques sepearatly.

Fan, Wai Pio 23 September 2011

Data set

This survey will use the dataset of the Netfilx Prize. The Netflix Prize was launched in October 2006, it is an open competition for the best CF algorithm to predict user ratings for films, based on previous ratings. The Netflix prize challenge is featured with a large-scale industrial dataset (with 480,000 users and 17,770 movies), and a rigid performance metric of. Each training rating is a tuple of the form <user, movie, date of grade, grade>. The user and movie fields are integer IDs, while grades are from 1 to 5 (integral) stars.

In summary, the data used in the Netflix Prize looks as follows:

- Training set (99,072,112 ratings)
- Probe set (1,408,395 ratings)
- Qualifying set (2,817,131 ratings) consisting of:
- Test set (1,408,789 ratings), used to determine winners
- Quiz set (1,408,342 ratings), used to calculate leaderboard scores

Another dataset for UCI repository: Amazon Access Samples Data Set
Amazon's InfoSec is getting smarter about the way Access data is leveraged. This is an anonymized sample of access provisioned within the company.

Data Set Characteristics:	Time-Series, Domain-Theory	**Number of Instances:**	30000	**Area:**	Business
Attribute Characteristics:	N/A	**Number of Attributes:**	20000	**Date Donated**	2011-09-13
Associated Tasks:	Regression, Clustering, Causal-Discovery	**Missing Values?**	N/A	**Number of Web Hits:**	1121

Source: Dataset creator and donator: Ken Montanez email: kenmonta[at]cal.berkeley.edu institution: Information Security, Amazon Corp.

Project plan

This survey is trying to study the current collaboration filtering techniques including memory-based, model-based and hybrid. To investigate how these approaches to solve the data sparsity, scalability, synonymy, gray sleep, shilling attack etc. This survey is also trying to implement or make use existing libraries to evaluate the performance of different type of CF techniques. The data sets are Netflix prize's movie data and Amazon access sample data set.

Milestones:
1. Study collaboration filtering related problem and usage
2. Study collaboration filtering techniques
3. Find or implement collaboration algorithms
4. Use the datasets to evaluate those algorithms
5. Report