A preliminary study on Tag-Based Recommender System

Hengyi Fu, MLIS 2013
Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign

Introduction

The purpose of recommender system is to build relationship between user's interest and item with an intermediary entity. An intermediary entity is needed because the direct relationship between user and item is unknown, assuming that the user has not yet tried the item. GroupLens introduces three common used explanations of recommendations: item-based, user-based, and feature-based, depending on the type of intermediary entity used to relate the user to the recommended item.

Method(con't)

Evaluation measure

Most online social networks use Top-N recommendation, so only Top-N items will be presented to the user. Typical evaluation measures for top-N recommendation are normalized discounted cumulative gain (NDCG) and precision.

In this project each dataset will be used for 10-fold cross-validation. For a certain user $u$, define:

$$ R(u) : \text{recommendation list of items with the length of N(N=10) for User } u \text{ generate by using training set}$$

$$ T(u) : \text{set of items user actually tag in testing set}$$

$$ \text{Precision} = \frac{1}{|T(u)|} \sum_{i \in T(u)} \mathbb{1}(R(u)_i \in T(u))$$

$$ \text{Recall} = \frac{|T(u)|}{|R(u)|} \sum_{i \in T(u)} \mathbb{1}(R(u)_i \in T(u))$$

Method(con't)

The term coverage is mainly associated with the percentage of the items for which the system is able to generate a recommendation.

$$ \text{Coverage} = \frac{\sum_{u \in U} |R(u)|}{|U||T(u)|}$$

Recommendation diversity means dissimilarity between each two items in recommendation list.

$$ \text{Diversity} = 1 - \frac{\sum_{u \in U} \sum_{i,j \in R(u)} \mathbb{1}(i \neq j) \mathbb{1}(i,j \text{ have common tag})}{|U||R(u)|}$$

Figure 4. Simple TagBased result

TagTFIDF:

$$ p_l(t) = \frac{n_l(t)}{\log(1+t^L)} \frac{1}{\sum_{i \in \text{tag(s) user uses}} p_l(i)}$$

Similar Tags:

For tag $l$, we use $N_l$ to denote the set of items which have tag $l$. $n_l(t)$ denotes the number of user who give tag $l$ to item $i$.So the similarity between $t$ and $l$ can be given by:

$$ \text{Sim}(t,l) = \sum_{i \in N_l} \frac{n_l(t)}{1+\log(1+n_l(t))} \sum_{i \in N_l} \frac{1}{|N_l|}$$

Result

Basic (Simple TagBased):

$$ \text{Precision} \quad \text{Recall} \quad \text{Coverage} \quad \text{Diversity} \quad \text{Popularity}$$

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Precision</th>
<th>Recall</th>
<th>Coverage</th>
<th>Diversity</th>
<th>Popularity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delicious</td>
<td>0.830</td>
<td>0.677</td>
<td>0.739</td>
<td>0.843</td>
<td>0.798</td>
</tr>
<tr>
<td>Last.fm</td>
<td>0.730</td>
<td>0.600</td>
<td>0.710</td>
<td>0.820</td>
<td>0.780</td>
</tr>
<tr>
<td>MovieLens</td>
<td>0.760</td>
<td>0.600</td>
<td>0.710</td>
<td>0.820</td>
<td>0.780</td>
</tr>
</tbody>
</table>

Figure 5. Similar Tags result

Limitations

- The three datasets used in this project are relatively small, which may hurt the performance of recommendation algorithms proposed
- Tag filter methods have not been applied in this project, which means the personal tags (e.g. "favorite", "2007") and synonym ("recommender system" and "recommendation system", "collaborative-filtering and collaborative_filtering") still exist in datasets

Why it's important?

Recommender systems can be used by E-commerce sites to suggest products to customers. The products can be recommended based on the top overall sellers on a site, based on the demographics of the customer, or based on analysis of the past buying behavior of the customer as a prediction for future buying behavior.

Recommend items to users to make user, content partner, websites happy!

References