Improving Recommender Systems by Reducing Hubness

Dontlu Vamsi Krishna (15111016)  
vamsi@iitk.ac.in

Dhekane Eeshan Gunesh (13248)  
eeshangd@iitk.ac.in

Project Guide - Prof. Piyush Rai (piyush@cse.iitk.ac.in)

Abstract

In this project, we propose to develop a Recommender System that can overcome the problem of Hubs, which is usually faced by most of them. Since hubs are mainly generated due to high dimensionality of data and the measures used to find the similarities between the data points, we try to implement different approaches that can reduce generation of Hubs.

Motivation

Recommender Systems or Recommendation Systems (RS) are a subclass of information filtering system that seek to predict the ‘rating’ or ‘preference’ that a user would give to an item\(^5\). Collaborative Filtering technique does this job by studying the item preferences of large number of users. This requires similarity measures between the data points, which is estimated using $k$–Nearest Neighbours ($k$–NN) algorithm. The data used by RS is usually high-dimensional. In high dimensional feature space some of the data-set objects emerge as Hubs\(^2\). These are the objects which appear frequently in the $k$ nearest neighbours of other objects. It is observed that the presence of hubs may affect the performance of $k$ nearest neighbour search\(^4\). One of the reasons is that if a hub object is present in the training set, it might be a neighbour of many of the other training objects. This can make the prediction of the class of a test object biased towards the labels of the hubs. This also leads to results which are not very useful. Thus, in order to make RS more effective, there is a need to overcome the hub problem.

Data Set

We will be using publicly available movie rating data-sets.

- **MovieLens 100K**: 100,000 ratings from 943 users on 1,682 items.
- **NetFlix Prize Data Set**: 100,480,507 ratings from 480,189 users on 17,770 items.

In addition to these, we might use other generated data sets\(^4\).
Methodology

We will start with the study of the concept of hubs, their origin and the effects on RS. Based on this, we will implement the techniques that would reduce the hub problem and test the results against the mentioned data sets. Once the results are obtained, we will implement the working methods in our Recommender System to test the improvement in its performance.

Time-line

<table>
<thead>
<tr>
<th>Work Description</th>
<th>Tentative Dates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Literature Survey</td>
<td>10/02/2016 – 29/02/2016</td>
</tr>
<tr>
<td>Code Development</td>
<td>01/03/2016 – 20/03/2016</td>
</tr>
<tr>
<td>Testing and Experimentation</td>
<td>21/03/2016 – 31/03/2016</td>
</tr>
<tr>
<td>Integrating</td>
<td>01/04/2016 – 10/04/2016</td>
</tr>
</tbody>
</table>

References


