Feature weighting in content based recommendation system using social network analysis

Conference Paper · January 2008

DOI: 10.1145/1367497.1367646 · Source: DBLP

3 authors, including:

Niloy Ganguly
Indian Institute of Technology Kharagpur
287 PUBLICATIONS **3,111** CITATIONS

SEE PROFILE

Pabitra Mitra
Indian Institute of Technology Kharagpur
214 PUBLICATIONS **4,665** CITATIONS

SEE PROFILE

Feature Weighting in Content Based Recommendation System Using Social Network Analysis

Souvik Debnath
Indian Institute of Technology
Kharagpur, India - 721302
cs_souvik@yahoo.co.in

Niloy Ganguly
Indian Institute of Technology
Kharagpur, India - 721302
niloy@cse.iitkgp.ernet.in

Pabitra Mitra
Indian Institute of Technology
Kharagpur, India - 721302
pabitra@cse.iitkgp.ernet.in

ABSTRACT

We propose a hybridization of collaborative filtering and content based recommendation system. Attributes used for content based recommendations are assigned weights depending on their importance to users. The weight values are estimated from a set of linear regression equations obtained from a social network graph which captures human judgment about similarity of items.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*information filtering*

General Terms

Algorithms, Design, Experimentation

Keywords

Recommender System, Social Network, Feature Similarity

1. INTRODUCTION

Recommendation systems produce a ranked list of items on which a user might be interested, in the context of her current choice of an item. Recommendation systems are built for movies, books, communities, news, articles etc. There are two main approaches to build a recommendation system - collaborative filtering and content based [3]. Collaborative filtering computes similarity between two users based on their rating profile, and recommends items which are highly rated by similar users. However, quality of collaborative filtering suffers in case of sparse preference databases. Content based system on the other hand does not use any preference data and provides recommendation directly based on similarity of items. Similarity is computed based on item attributes using appropriate distance measures. We attempt to hybridize collaborative filtering and content based recommendation for circumventing the difficulties of these individual approaches. Item similarity measure used in content based recommendation is learned from a collaborative social network of users.

Some previous attempts at integrating collaborative filtering and content based approach include content boosted collaborative filtering [3], weighted, mixed, switching and feature combination of different types of recommender system [2]. But none of these talks about producing recommendation to a user without getting her preferences. We demonstrate the effectiveness of the proposed system for recommending movies in Internet Movie Database (IMDB) [1]. From the results it is seen that our recommendation is quite in agreement with IMDB recommendation.

WWW 2008, April 21–25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

2. FEATURE WEIGHTING IN CONTENT BASED RECOMMENDATION

In content based recommendation every item is represented by a feature vector or an attribute profile. The features hold numeric or nominal values representing certain aspects of the item like color, price etc. A variety of distance measures between the feature vectors may be used to compute the similarity of two items. The similarity values are then used to obtain a ranked list of recommended items. If one considers Euclidian or cosine similarity; implicitly equal importance is asserted on all features. However, human judgment of similarity between two items often gives different weights to different attributes. For example, while choosing a camera, price of a camera may be more important than the body color attribute. It may be stated that users base their judgments on some latent criteria which is a weighted linear combination of the differences in individual attribute. Accordingly, we define similarity S between objects O_i and O_j as

$$S(O_i, O_j) = \omega_1 f(A_{1i}, A_{1j}) + \omega_2 f(A_{2i}, A_{2j}) + \cdots + \omega_n f(A_{ni}, A_{nj}) \quad (1)$$

where ω_n is the weight given to the difference in value of attribute A_n between objects O_i and O_j, the difference given by $f(A_{ni}, A_{nj})$. The definition of f depends on the type of attribute (numeric, nominal, boolean). We normalize f's to have value in [0, 1]. In general the weights $\omega_1, \omega_2, \cdots, \omega_n$ are unknown. In the next section we describe a method of determining these weights from a social collaborative network.

We have used the above methodology for recommending movie in IMDB database. A set of 13 features are considered. The features along with their type, domain and distance measures are shown in Table 1. All these feature values can be obtained from the IMDB database.

3. DETERMINING FEATURE WEIGHTS

We estimate the feature weights from a social network graph of items. The underlying principle is to use exist-

Table 1: Features Used in Movie Recommendation

Feature	Type	Domain	Distance Measure
Release	Year	YYYY	$\frac{(300-\lvert Y_1-Y_2\rvert)}{300}$
Type	String	Movie,TV etc.	$T_1 = T_2?1:0$
Rating	Integer	(0-10)	$\frac{(10-\lvert R_1-R_2\rvert)}{10}$
Vote	Integer	(≥ 5)	$\frac{(V_{max}-\lvert V_1-V_2\rvert)}{V_{max}}$
Director	String	<Name>	$D_1 = D_2?1:0$
Writer	String	<Name>	$W_1 = W_2?1:0$
Genre	(String)*	Drama etc.	$\frac{\lvert G_1 \cap G_2\rvert}{G_{max}}$
Keyword	(String)*	College etc.	$\frac{\lvert K_1 \cap K_2\rvert}{K_{max}}$
Cast	(String)*	(<Name>)*	$\frac{\lvert C_1 \cap C_2\rvert}{C_{max}}$
Country	(String)*	France etc.	$\frac{\lvert C_1 \cap C_2\rvert}{C_{max}}$
Language	(String)*	English etc.	$\frac{\lvert L_1 \cap L_2\rvert}{L_{max}}$
Color	String	Color, B/W	$C_1 = C_2?1:0$
Company	String	<Name>	$C_1 = C_2?1:0$

ing recommendation by users to construct a social network graph with items as nodes. The graph represents human judgment of similarity between items aggregated over a large population of users. Optimal feature weights are considered to be those which induce a similarity measure between items best conforming to this social network graph.

We describe below a linear regression framework for determining the optimal feature weights. Let the items under consideration be denoted by $O_1, O_2, \cdots, O_l$, they corresponds to the vertices of our social network. The edge weight between vertices O_i and O_j,

$E(O_i, O_j) = \#$ of users who are interested in both O_i, O_j.

$E(O_i, O_j)$, suitably normalized, may be considered as human judgment of similarity between O_i, O_j. Recall that feature vector (content based) similarity between O_i, O_j has been defined as $S(O_i, O_j)$ in Eq. (1). Equating $E(O_i, O_j)$ with $S(O_i, O_j)$ leads to the following set of regression equations. $\forall i, \forall j = 1..l \wedge i \neq j$,

$$\omega_0 + \omega_1 f(A_{1i}, A_{1j}) + \omega_2 f(A_{2i}, A_{2j}) + \cdots + \omega_n f(A_{ni}, A_{nj}) = E(O_i, O_j) \quad (2)$$

The values of $f(A_{1i}, A_{1j}), f(A_{2i}, A_{2j}), \cdots, f(A_{ni}, A_{nj})$ are known from the data as are the values of $E(O_i, O_j)$. Solving the above regression equations provide estimates for the values of $\omega_1, \omega_2, \cdots, \omega_n$. If there are l objects under consideration, it is possible to have ${}^lC_2$ regression equations of the above form. In the case of movie recommendation we have considered movies as nodes in the social network. The edge weight between two movies is the number of IMDB reviewers who have reviewed both the movies.

4. EXPERIMENTAL RESULTS

The movie database used in our recommendation system consists of 3×10^5 random movies downloaded from the IMDB. The movies voted by less than 5 people or the movies that have not been reviewed by a single person are filtered out. The data is then divided into three equal sets. Each movie is described by 13 features (Table 1).

4.1 Stability of Feature Weights

Our recommendation system is based on the presumption that feature weights are almost universal for different sets of users and movies. To test this presumption we consider different sets of regression equations and solve for the weights. We consider the following varieties of regression equations.

Table 2: Feature Weight Values

Feature	Mean	Variance
Type	0.18	0.0023
Writer	0.36	0.0048
Genre	0.04	0.0001
Keyword	0.03	0.0011
Cast	0.01	0.0003
Country	0.07	0.0013
Language	0.09	0.0004
Company	0.21	0.0110

I. Equations using only edge weights ≥ 1 (i.e. movie pairs having at least one co-reviewer)

II. Equations using only edge weights ≥ 2 . (Note that this gives a graph which is a sub-graph of the previous graph.)

For the above graphs we construct a set of equations for each of the three (partitioned) datasets having 10^5 movies. Thus we get six sets of regression equations which we solve using SPSS package. It is observed from the weight values obtained from each of the above six sets of regression equations that some of the features have stable weight values, while some features like Director, Rating, Vote, Year, Color have unstable or negative weight. We remove the features with unstable or negative weights from our regression equations and obtain the following set (Table 2) of stable weights for eight features. Also note, out of the 8, 3 features namely type, writer and company are particularly important. These features along with their weights are used to obtain the recommendations.

4.2 Performance of the Recommender System

The proposed algorithm is compared with pure content based method (considering equal weights for all features) and IMDB recommendations. Performance is measured using the classical *Recall* measure, considering IMDB recommendation as benchmark. The experiment has been done on 10 different movies. The proposed method achieves an average recall of 0.29. Where as, the pure content based method achieves a recall of 0.24 with IMDB. Thus the proposed method agrees well with IMDB recommendation and in this regard it outperforms pure content based method. This demonstrates the effectiveness of feature weighting.

5. CONCLUSION

A hybridization of content based and collaborative filtering based recommendation is proposed. The weights of different attributes of an item are computed from the collaborative social network using regression analysis. Further studies on other weight estimation techniques like sparse regression and isometric projection are being considered. Also more rigorous performance evaluation based on human judgment will be undertaken.

6. REFERENCES

[1] Internet Movie Database. http://www.imdb.com.

[2] Bruke, R. *Hybrid recommender systems: survey and experiments*, User Modeling and User Adapted Interaction 12 (2002) 331-370.

[3] P. Melville, R.J. Mooney, R. Nagarajan *Content-Boosted Collaborative Filtering for Improved Recommendations*, Proceedings of the 18th National Conference on Aritificial Intelligence (AAAI-2002), July 2002, Edmonton, Canada.