ISSN (Online): 2349-7084
GLOBAL IMPACT FACTOR 0.238
DIIF 0.876

INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING IN RESEARCH TRENDS
VOLUME 2, ISSUE 9, SEPTEMBER 2015, PP 594-599

Survey on Collaborative Filtering and Content-Based Recommending

[1] Tekur Vijetha, [2] M.Sri lakshmi, [3]Dr.S.Prem Kumar
[1](M.Tech), CSE
[2]Assistant Professor, Department of Computer Science and Engineering
[3]Professor & HOD, Department of computer science and engineering,
G.Pullaiah College of Engineering and Technology, Kurnool, Andhra Pradesh, India.

Abstract:- Collaborative filtering (CF) is an important and popular technology for recommender systems. Recommender systems have been proven to be valuable means for web online users to cope with the information overload and have become one of the most powerful and popular tools in electronic commerce. Recommending and personalization are important approaches to combating information over-load.Machine Learning is an important part of systems for these tasks. Collaborative filtering has problems. Content-based methods address these problems (but have problems of their own).Integrating both is best.

Keywords: Collaborative Filtering (a.k.a. social filtering), Content-based Recommender Systems.

1 INTRODUCTION

Recommendation systems found their application in the field of e-commerce and internet where items suggest to a group of user on the basis of their requirement based on their area of interest. A recommendation system is an information filtering system that built a model from the characteristic of an item according to the rating or prediction, given by a user to an item. Recommendation system has an important component in social media sites (such as Amazon, IMDB, Movie Lens), social sites giants such as Amazon have been greatly gained from the capability of their recommenders in accurately delivering the correct item to the correct user [17]. Collaborative filtering (CF) is an important and popular technology for recommender system. CF methods are classified into user based CF and item-based CF. The basic idea of user-based CF approach is to find out a set of users who have similar favor patterns or interest to a given user and the basic idea of item-based CF approach is to find out a set of items having highest correlation with the given item. In reality, people may like to group items into categories, and for each category there is a corresponding group of people who like items in the category [18]. Cognitive psychologists find that objects (items) have different typicality degrees in categories in real life [19], [20], [21]. But these collaborative filtering methods have facing some problems.

II. SYSTEM ANALYSIS:

2.1. Recommender Systems

Systems for recommending items (e.g. books, movies, CD's, web pages, newsgroup messages) to users based on examples of their preferences. Many on-line stores provide recommendations (e.g. Amazon, CDNow).Recommenders have been shown to substantially increase sales at on-line stores. There are two basic approaches to recommending: Collaborative Filtering (a.k.a. social filtering), Content-based

Book Recommender

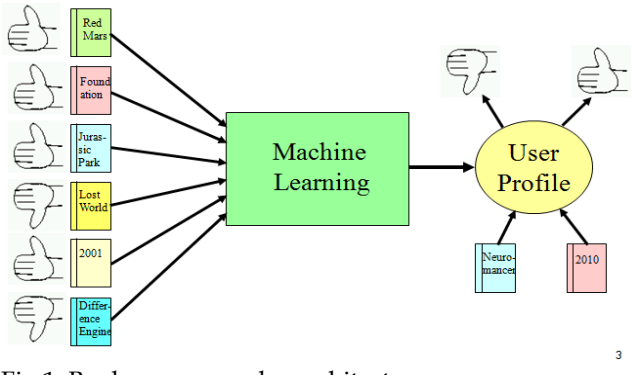

Fig 1. Book recommender architecture

ISSN (Online): 2349-7084

INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING IN RESEARCH TRENDS
VOLUME 2, ISSUE 9, SEPTEMBER 2015, PP 594-599

Personalization

- Recommenders are instances of personalization software.
- Personalization concerns adapting to the individual needs, interests, and preferences of each user.
- Includes:
  - Recommending
  - Filtering
  - Predicting (e.g. form or calendar appt. completion)

From a business perspective, it is viewed as part of Customer Relationship Management (CRM).

Machine Learning and Personalization

- Machine Learning can allow learning a user model or profile of a particular user based on:
  - Sample interaction
  - Rated examples
- This model or profile can then be used to:
  - Recommend items
  - Filter information

Predict behavior

2.2. Collaborative Filtering

Maintain a database of many users' ratings of a variety of items.For a given user, find other similar users whose ratings strongly correlate with the current user.Recommend items rated highly by these similar users, but not rated by the current user.Almost all existing commercial recommenders use this approach (e.g. Amazon).

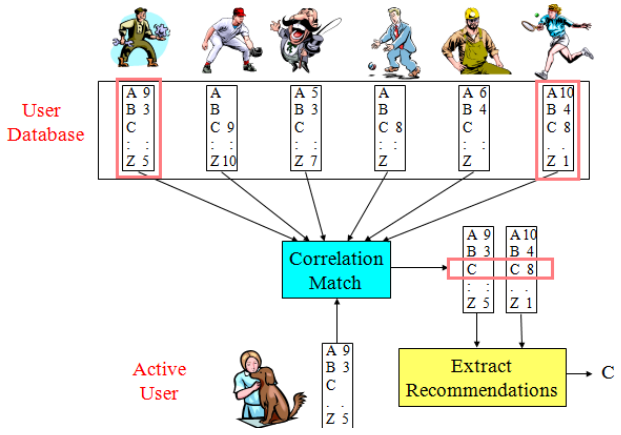

Fig 2. Collaborative Filtering

Weight all users with respect to similarity with the active user.Select a subset of the users (*neighbors*) to use as predictors.Normalize ratings and compute a prediction from a weighted combination of the selected neighbors' ratings.Present items with highest predicted ratings as recommendations.

- Typically use Pearson correlation coefficient between ratings for active user, *a*, and another user, *u*.

$$c_{a,u} = \frac{\text{covar}(r_a, r_u)}{\sigma_{r_a}\sigma_{r_u}}$$

r_a and r_u are the ratings vectors for the m items rated by **both** a and u

$r_{i,j}$ is user i's rating for item j

Covariance and Standard Deviation

Covariance:

$$\text{covar}(r_a, r_u) = \frac{\sum_{i=1}^{m}(r_{a,i} - \bar{r}_a)(r_{u,i} - \bar{r}_u)}{m}$$

$$\bar{r}_x = \frac{\sum_{i=1}^{m} r_{x,i}}{m}$$

Standard Deviation:

$$\sigma_{r_x} = \sqrt{\frac{\sum_{i=1}^{m}(r_{x,i} - \bar{r}_x)^2}{m}}$$

2.3. Neighbor Selection

For a given active user, a, select correlated users to serve as source of predictions.Standard approach is to use the most similar n users, u, based on similarity weights, wa,u Alternate approach is to include all users whose similarity weight is above a given threshold.

- For a given active user, *a*, select correlated users to serve as source of predictions.
- Standard approach is to use the most similar *n* users, *u*, based on similarity weights, $w_{a,u}$
- Alternate approach is to include all users whose similarity weight is above a given threshold.

Rating Prediction

- Predict a rating, pa,i, for each item i, for active user, a, by using the n selected neighbor users, u ∈ {1,2,…n}.
- To account for users different ratings levels, base predictions on differences from a user's average rating.

ISSN (Online): 2349-7084

INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING IN RESEARCH TRENDS
VOLUME 2, ISSUE 9, SEPTEMBER 2015, PP 594-599

- Weight users' ratings contribution by their similarity to the active user.

$$p_{a,i} = \bar{r}_a + \frac{\sum_{u=1}^{n} w_{a,u}(r_{u,i} - \bar{r}_u)}{\sum_{u=1}^{n} w_{a,u}}$$

3. CONTENT-BOOSTED COLLABORATIVE FILTERING

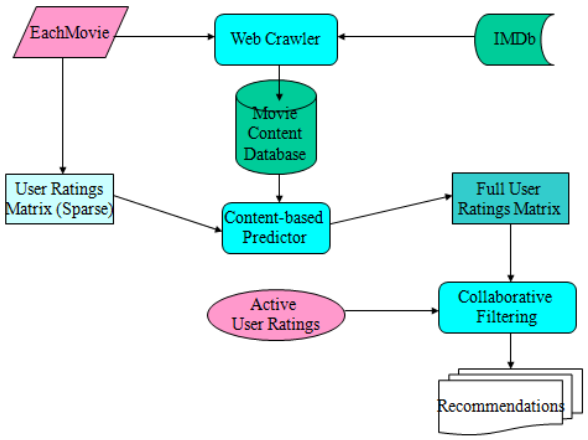

Fig 3. Content-boosted collaborative filtering architecture

Content-Boosted CF – I

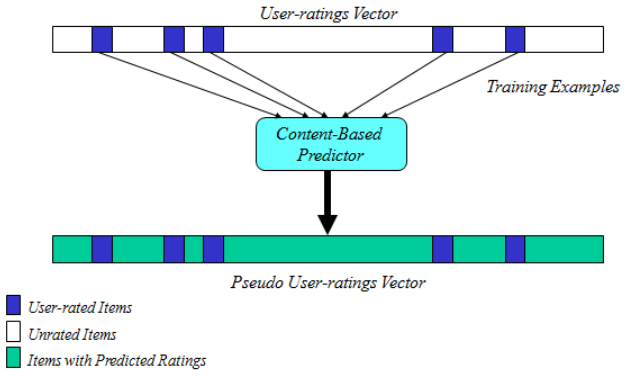

Fig 4.Content-Boosted CF – I

Problems with Collaborative Filtering

Cold Start: There needs to be enough other users already in the system to find a match.

Sparsity: If there are many items to be recommended, even if there are many users, the user/ratings matrix is sparse, and it is hard to find users that have rated the same items.

First Rater: Cannot recommend an item that has not been previously rated.

- New items
- Esoteric items

Popularity Bias: Cannot recommend items to someone with unique tastes.

- Tends to recommend popular items.

3.1. Content-Based Recommending

Recommendations are based on information on the content of items rather than on other users' opinions.

Uses a machine learning algorithm to induce a profile of the users preferences from examples based on a featural description of content.

Some previous applications:

Newsweeder (Lang, 1995)

Syskill and Webert (Pazzani et al., 1996)

Advantages of Content-Based Approach

- No need for data on other users.
  - No cold-start or sparsity problems.
- Able to recommend to users with unique tastes.
- Able to recommend new and unpopular items
  - No first-rater problem.
- Can provide explanations of recommended items by listing content-features that caused an item to be recommended.

Disadvantages of Content-Based Method

Requires content that can be encoded as meaningful features.Users' tastes must be represented as a learnable function of these content features.Unable to exploit quality judgments of other users.Unless these are somehow included in the content features.

4. LIBRA (LEARNING INTELLIGENT BOOK RECOMMENDING AGENT)

Content-based recommender for books using information about titles extracted from Amazon.

Uses information extraction from the web to organize text into fields:

- Author
- Title
- Editorial Reviews
- Customer Comments
- Subject terms
- Related authors
- Related titles

ISSN (Online): 2349-7084

INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING IN RESEARCH TRENDS
VOLUME 2, ISSUE 9, SEPTEMBER 2015, PP 594-599

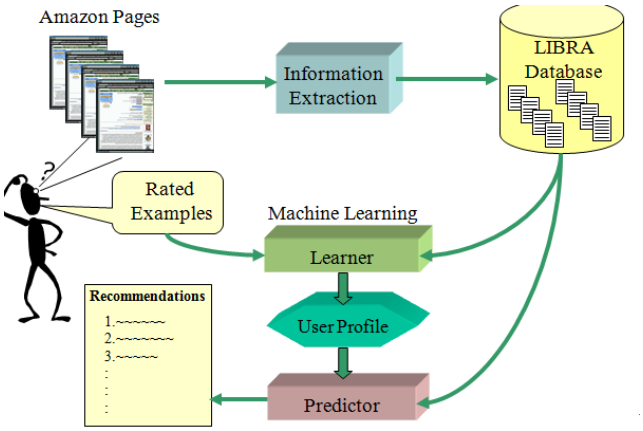

Fig5. LIBRA System Architecture

Sample Extracted Information

Title: <The Age of Spiritual Machines: When Computers Exceed Human Intelligence>
Author: <Ray Kurzweil>
Price: <11.96>
Publication Date: <January 2000>
ISBN: <0140282025>
Related Titles: <Title: <Robot: Mere Machine or Transcendent Mind>
Author: <Hans Moravec> >
…
Reviews: <Author: <Amazon.com Reviews> Text: <How much do we humans…> >
…
Comments: <Stars: <4> Author: <Stephen A. Haines> Text:<Kurzweil has …> >
…
Related Authors: <Hans P. Moravec> <K. Eric Drexler>…
Subjects: <Science/Mathematics> <Computers> <Artificial Intelligence> …

Libra Content Information

- Libra uses this extracted information to form "bags of words" for the following slots:
  - Author
  - Title
  - Description (reviews and comments)
  - Subjects
  - Related Titles
  - Related Authors
- User rates selected titles on a 1 to 10 scale.
- Libra uses a naïve Bayesian text-categorization algorithm to learn a profile from these rated examples.
  - Rating 6–10: Positive
  - Rating 1–5: Negative
- The learned profile is used to rank all other books as recommendations based on the computed posterior probability that they are positive.
- User can also provide explicit positive/negative keywords, which are used as priors to bias the role of these features in categorization.

5. SYSTEM IMPLEMENTATION

Stopwords removed from all bags.A book's title and author are added to its own related title and related author slots.All probabilities are smoothed using Laplace estimation to account for small sample size.Lisp implementation is quite efficient:Training: 20 exs in 0.4 secs, 840 exs in 11.5 secs ,Test: 200 books per second

Explanations of Profiles and Recommendations

- Feature strength of word wk appearing in a slot sj :

$$\text{strength}(w_k, s_j) = \log \frac{P(w_k \mid \text{positive}, s_j)}{P(w_k \mid \text{negative}, s_j)}$$

6. EXPERIMENTAL DATA

- Amazon searches were used to find books in various genres.
- Titles that have at least one review or comment were kept.
- Data sets:
  - Literature fiction: 3,061 titles
  - Mystery: 7,285 titles
  - Science: 3,813 titles
  - Science Fiction: 3.813 titles

6.1. Rated Data

4 users rated random examples within a genre by reviewing the Amazon pages about the title:

- LIT1 936 titles
- LIT2 935 titles
- MYST 500 titles
- SCI 500 titles
- SF 500 titles

6.2. Method

- 10-fold cross-validation to generate learning

ISSN (Online): 2349-7084

INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING IN RESEARCH TRENDS
VOLUME 2, ISSUE 9, SEPTEMBER 2015, PP 594-599

curves.

- Measured several metrics on independent test data:
- Precision at top 3: % of the top 3 that are positive
- Rating of top 3: Average rating assigned to top 3
- Rank Correlation: Spearman's, rs, between system's and user's complete rankings.
- Test ablation of related author and related title slots (LIBRA-NR).
- Test influence of information generated by Amazon's collaborative approach.

6.3.Experimental Result Summary

- Precision at top 3 is fairly consistently in the 90's% after only 20 examples.
- Rating of top 3 is fairly consistently above 8 after only 20 examples.
- All results are always significantly better than random chance after only 5 examples.
- Rank correlation is generally above 0.3 (moderate) after only 10 examples.
- Rank correlation is generally above 0.6 (high) after 40 examples.

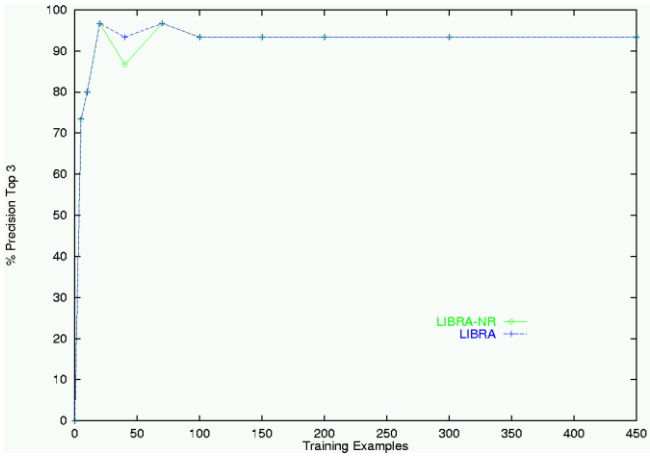

Fig 6.Precision at Top 3 for Science

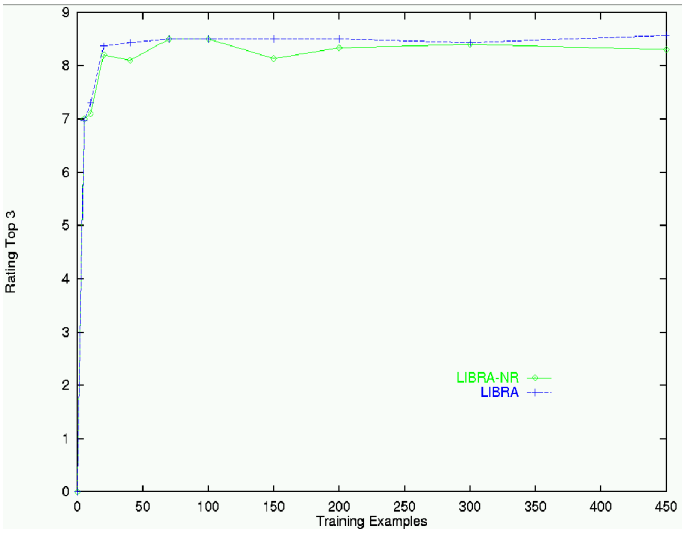

Fig7. Rating of Top 3 for Science

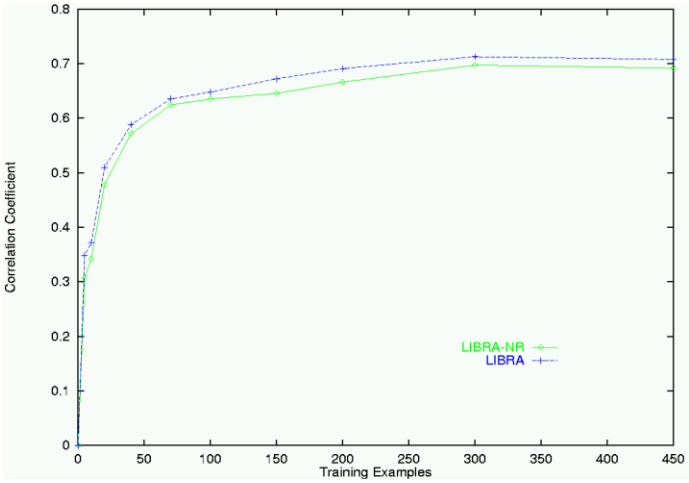

Fig 8.Rank Correlation for Science

7. CONCLUSSION

In this paper we investigate on Collaborative filtering (CF). Recommending and personalization are important approaches to combating information overload.Machine Learning is an important part of systems for these tasks. Collaborative filtering has problems. Content-based methods address these problems (but have problems of their own).Integrating both is best.

REFERENCES

[1] Z. Huang, H. Chen, and D. Zeng, "Applying Associative Retrieval Techniques to Alleviate the Sparsity Problem in Collaborative Filtering," ACM Trans. Information Systems, vol. 22, no. 1, pp. 116- 142, 2004.

[2] G. Adomavicius and A. Tuzhilin, "Toward the Next Generation of Recommender Systems: A Survey of the

State-of-the-Art and Possible Extensions," IEEE Trans. Knowledge and Data Eng., vol. 17, no. 6, pp. 734-749, June 2005.

[3] K.M. Galotti, Cognitive Psychology In and Out of the Laboratory, third ed. Wadsworth, 2004.

[4] G.L. Murphy, The Big Book of Concepts. MIT Press, 2002.

[5] L.W. Barsalou, Cognitive Psychology: An Overview for Cognitive Scientists. Lawrence Erlbaum Assoc., 1992.

[6] S. Schiffer and S. Steele, Cognition and Representation. Westview Press, 1988.

[7] D.L. Medin and E.E. Smith, "Concepts and Concept Formation," Ann. Rev. of Psychology, vol. 35, pp. 113-138, 1984.

[8] W. Vanpaemel, G. Storms, and B. Ons, "A Varying Abstraction Model for Categorization," Proc. Cognitive Science Conf. (CogSci'05), pp. 2277-2282, 2005.