Vol. (1) No. (1) March 2018

JK Research Journal in Mathematics and Computer Sciences

Hybrid Recommendation System Based on Thesaurus for Handling Cold Start

Zahid Maqbool[1] AamirMushtaq[2]

[1]Zahid Maqbool Department of Computer Science, GDC Dooru, Anantnag J&K- India
[2]Aamir Mushtaq Doctoral Student School of Basic Sciences IIT Mandi, HP-India

Abstract: Till now so many algorithms have been developed for providing recommendations. The Recommendation algorithms effectiveness is directly affected by the user's ratings. In general, more user ratings imply better recommendation we can serve. But usually, we have very sparse user ratings in the actual environment and hence directly affect the quality of recommendation.

In this paper, we will show how to handle cold start problem in news recommender system using synonyms of a topic. And finally we will show how ratings generated by synonyms gave better recommendations and at the same time, it effectively resolved the cold-start problem. We present an algorithm to solve the sparse rating problem and then use that algorithm to test on Reuter's data set and achieve precession of 63%, recall of 46%. The system was also able to rate items with zero ratings.

Keywords: *Cold-start; First Rater problem; Recommender System; Hybrid Recommender system; News Recommender System; Data mining.*

Introduction

Recommendation is a process by which a set of items are recommended to a user which he may consume (read, buyetc. depends on the item being recommended). Software which implements this process is called Recommender System (RS). Recommender systems play a vital role in many present-day systems like e-commerce, Blogs, News websites and many more. Recommendations are generated by processing large amount of information so that the recommended items best matchesusers' preferences and inturn improve conversion rate.

In general recommender systems are categorized into two broad categories Collaborative Filtering (CF) and Content-Based Filtering (CB) based on how they calculate rating. CF recommender systems are based on the ratings given by users who have similar taste and preferences while as CB recommender systems are based on the content/feature similarity of an item with respect to other items which user has consumed in the past. It is believed that CF gives usually better results as compared to CB but it suffers from Cold-Start (CS) problem. Cold-start problem occurs when the user is interacting with the system for the first time and has no previous ratings for any item. Success of CF recommender systems is wholly and solely

zahidcomp@gmail.com
*Zahid Maqbool

dependent on the amount of rating information available. CB doesn't get affected by CS but it suffers from overspecialization which means items that have high similarity score to those already rated will be recommended to the user. Apart from overspecialization it also gets affected by limited content analysis which means item details are not sufficient enough to find similarity with other items.

In addition to CF and CB recommender systems another type of recommendation system called Hybrid Recommendation system, which combines the advantages of CF and CB and reduces their limitations. Hybrid recommender systems are composed of basic recommender systems like Content-Based, Collaborative Filtering, and Demographic recommender systems. The main problem with basic recommender systems are Cold-Start (in CB, CF and demographic) and Stability/plasticity problems (where it is hard to change user's established profile i.e. user may want to check items which are different from its preference trend). [27, 31, 32]

Related Work

In collaborative filtering (CF), user ratings are used to generate recommendations for any user. It works by taking into consideration the taste of other users called neighbours. CF has been extensively studied and tested on movie domain [9, 25] but same methodology can be applied to a number of other domains as well, e.g. books [29], music [26, 31], jokes [8], news, newsgroups [13, 17], advertisements, and more. CF algorithms vary from basic "memory-based" methods [23, 25] to more advanced model-based methods in which we first train a model for example, a Bayesian network [6], a classifier [5, 18], a co-clustered matrix [27], or a truncated singular value decomposition matrix [5] based on some historical data, and then use this trained model to generate recommendations. So Many CF algorithms have been devised and tested, which includes machine learning (ML) methods [5, 22], graph-based methods [12], linear algebra based methods [5, 8], and probabilistic methods [11, 19, 20]. In addition to CF, many hybrid methods which combine Content-Based (CB) and CF techniques have also been proposed [9, 20, 33, 34], these systems are very productive when user ratings are sparse, for example in cold-start situations. Also, CF methods cannot be used at all if user has not rated any item, in such cases CB methods or hybrid CB/CF methods are required. We base our methodology for solving the cold-start problem by using a topic dictionary and use that for generating recommendations. This Topic Dictionary is created and updated periodically by using MALLET. We have devised a novel algorithm namely **MUPhi** for solving Cold-start problem using Hybrid System and after evaluation our algorithm showed significant improvement in handling Cold-start problem

Methodology

When a user is searching for some news item for the first time, its rating is zero. First Rater Problem (a.k.a Sparse Rating or Cold start) reduces the performance of a Recommendation System as items with zero rating; recommendation system cannot recommend any item to the user. So in order to address this issue, a Hybrid approach is proposed that uses both the techniques, content-based and collaborative.

Method

Our approach comprises of following components:

1. Topic Analyzer.
2. Thesaurus Generator
3. Topic Indexer.
4. Rater

Recommender Engine

1. **Topic Analyzer:** Topic Analyzer works by using a topic modelling tool, MALLET;we have chosen Latent Dirchlet allocation (LDA) as our topic model. We provide a set of documents and it returns a set of topics for each document according to the probability distribution of the topics and then we select top N' topics based on the probability score.

2. **Thesaurus Generator:** For each topic from Topic list generated by Topic Analyzer; find all its synonyms. Synonyms are generated using publicly available WordNet Lexicon.

3. **Topic Indexer:** Topic Indexer will index all the topics and their respective documents in a linked-list like structure.

4. **Rater:** Rater is used forrating the documents. Ratings will be used for generating recommendation using Collaborative Filtering approach.

Recommender Engine: Recommender Engine typically produces a list of recommendations either through collaborative or content-based filtering based on the presence of topic in Index. Collaborative filtering approaches to build a model from user's past behaviour's (i.e. User Profile) as well as similar decisions made by other users and use that model to predict news items that the user may have an interest in. If the user is reading a news item on a topic which is not present in the existing Index then content-based filtering approach will utilize a list of extracted topics and its synonyms for recommending new news items. These approaches are often combined called hybrid Recommender Systems.

Algorithm

The algorithm works in the following manner:

```
Algorithm 1    procedure My Procedure
    u ← current user
    d ← document read by user d
    ifuser(u) = Returning user then
    loop:
        Load User Profile UP.
        Apply Collaborative Filtering CF return recommended documents.
        Update UP
        gotoloop.
        close;
    else
        T = LDA(d)
    for Each topic t in T do
        ift belongs to T then
        returndocList
    Else TS = TS + synonym(t).
    for Each s in TS do  if s belongs to Ithen
    (t, d, p) :=> Indexer(I)
    returndocList
            UP <=: UP(d, 1)
    end
```

Description of Algorithm

STEP 1: First time a user logs in; it is identified whether it is a new user or returning user. If it is Returning user then its corresponding user profile is loaded and recommendations are generated using CF. If it is a new user then the document selected by the user is provided to topic modeller (Latent Dirchlet Allocation) which extracts most relevant topics from this document(say,T). Also, an entry is added in the user profile with rating score as 1 for the current document.

STEP 2: Topic T and document D is passed to Indexer for indexing purpose. Which means for each Topic t in T, there will be an entry in the index?

STEP 3: While generating are commendation for the current document, we search index for the topic to which this document belongs. If a match is found in the index then we recommend documents otherwise we first find synonyms of extracted topics of this document. After that index is searched for each synonym. If any match is found then alist of documents based on probability score is served as are commendation.

STEP 4: If a document is read by the user then User profile is accordingly updated by entering an entry for the current document and its score i.e. 1.

Above steps are diagrammatically shown in figures 1, 2, 3, and 4.

Implementation

We used Reuter Dataset for evaluation purpose, it contains news articles collected from various sources. This dataset comprises of 10,768 documents out of which we used 7,769 documents for training purpose and remaining 3,019 documents for testing purpose.

Working of the system can be diagrammatically shown as under:

Functioning

— Topic Analyzer extracts topics from the document and stores it in a Database (MongoDB).

— User Profile of current user is supplied to Recommendation System in order to generate recommendations.

— If the user is returning user then Recommendation Algorithm will use User Profile information for generating recommendation using Collaborative Filtering.

— If the user is coming for the first time then there will be no User Profile for this user. So, RS will extract topics from the document which user is currently reading and then it scans the existing index for these topics and using Content-Based approach it will fetch documents from the existing index and will serve them as arecommendation.

— If no such index is found then system use WordNet lexicon for finding synonyms of the current topic and using those synonyms index will be scanned again and once a match is detected then recommendations are generated.

— With each User action, RS will keep updating User Profile based on the user's interaction with the system. Also, RS will incorporate ratings for the new topics as long as the user reads that article.

- It will find out log likelihood similarity and user neighborhood and return back recommended topic and documents associated with the topic to the user based on these calculations.
- Finally using CF. CB and Topic Synonyms in a systematic approach as discussed above, cold start problem is handled in a very effective manner.

Evaluation

To demonstrate that our recommendation method works, we tested devised system with Reuter's dataset. The dataset consists of 10,788 documents out of which 7769 were used for training purpose and rest used for testing purpose. Based on whether the user has read an article or not we used binary rating to represent it. Thus ratings will be 0 or 1. We have used a hybrid approach which helps to solve first ratter problems of both content-based filtering and collaborative filtering recommendation systems. The system works by exploiting the topic of the document as fine-grained approach for the recommendation. The system performs excellently for the test data and shows improvement over the baseline accuracy.

Results

For our experiment, we have taken 25 users. Entries of all these users are done in the table and then we have calculated recommendation for any of these users. Precision obtained for our system is 63.18. Recall obtained for our system is 35.29. F-Measure for our system is 45.28.For the system highest precision achieved is 100% and lowest precision achieved is 44.4 % for 25 users. Even for a new user with rating zero this system performs well and provides are commendation to that user as well.

The results are graphically shown by following:

Conclusion and Future Work

Recommender systems are an extremely potent tool utilized to assist the selection process easier for users. Not surprisingly, these systems, while used mainly in the e-commerce shopping world, can also be applied in myriad contexts as well. We developed a sophisticated model which will handle the Sparse Rating problem of collaborative filtering approach using a hybrid recommender system and enhance the quality of recommendations. This hybrid recommender system is designed and implemented to handle solve Cold start problem associated with collaborative filtering in a much-sophisticated way.

In Future work, we would like to extend it further by using demographic and time series information for providing more user-centric recommendations and also utilizing timestamp of news article while calculating finale relevance score. This will further improve the quality of recommended items and will further tackle new item problem as well.

References

1. C. C. Aggarwal, J. L. Wolf, K.-L.Wu, and P. S. Yu. Horting hatches an egg: a new graph-theoretic approach to collaborative filtering. In ACM KDD, pages 201212, 1999.

2. M. Balabanovic and Y. Shoham. Fab: content-based, collaborative recommendation. Communications of the ACM,40(3):6672, 1997.

3. J. Basilico and T. Hofmann.A joint framework for collaborative and content filtering.In ACM SIGIR, 2004.

4. C. Basu, H. Hirsh, and W. W. Cohen. Recommendation as classification: Using social and content-based information in recommendation. In AAAI/IAAI, pages 714720, 1998.

5. D. Billsus and M. J. Pazzani. Learning collaborative information filters. In ICML, pages 4654, 1998.[6] J. S. Breese, D. Heckerman, and C. Kadie.Empirical analysis of predictive algorithms for collaborative filtering.In UAI, pages 4352, 1998.

6. S.Sahebi, William W. Cohen. Community-based recommendations: a solution to the Cold-Start Problem WOODSTOCK97 El Paso, Texas USA

7. M. Deshpande and G. Karypis.Item-based top-n recommendation algo. ACM TOIS, 22(1):143177, Jan 2004.

8. K. Goldberg, T. Roeder, D. Gupta, and C. Perkins.Eigentaste:A constant time collaborative filtering algorithm. Information Retrieval, 4(2):133151, 2001.

9. N. Good, J. B. Schafer, J. A. Konstan, A. Borchers, B. M.Sarwar, J. L. Herlocker, and J. Riedl. Combining collaborative filtering with personal agents for better recommendations.In AAAI/IAAI, pages 439446, 1999.

10. Thomas Hess, Recommendation Engines Seminar Paper, February 1,2009

11. T. Hofmann and J. Puzicha. Latent class models for collaborative filtering. In IJCAI, pages 688693, 1999.

12. Z. Huang, H. Chen, and D. Zeng.Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. ACM TOIS, 22(1):116142, Jan 2004.

13. J. A. Konstan, B. N. Miller, D. Maltz, J. L. H. L. R. Gordon, and J. Riedl.GroupLens: applying collaborative filtering to Usenet news. Communications of the ACM, 40(3):7787, 1997.

14. B. Marlin. Collaborative filtering: A machine learning perspective. Masters thesis, University of Toronto, Computer Science Department.

15. S. McNee, S. Lam, J. Konstan, and J. Riedl.Interfaces for eliciting new user preferences in recommender systems.In UM, pages 178188, 2003.

16. P. Melville, R. Mooney, and R. Nagarajan.Content-boosted collaborative filtering.In AAAI, 2002.

17. B. N. Miller, J. T. Riedl, and J. A. Konstan. Experience with grouplens: MakingUsenet useful again. In USENIX annual technical conference, pages 219231, 1997.

18. K. Miyahara and M. J. Pazzani.Collaborative filtering with the simple Bayesian classifier.In PRICAI, pages 679689, 2000.

19. D. Pennock, E. Horvitz, S. Lawrence, and C. L. Giles. Collaborative filtering bypersonality diagnosis: A hybrid memory- and model-based approach. In UAI, pages473480, 2000.

20. A. Popescul, L. Ungar, D. Pennock, and S. Lawrence. Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments.In UAI, pages 437444, 2001.

21. A. Rashid, I. Albert, D. Cosley, S. Lam, S. Mcnee, J. Konstan, and J. Riedl. Gettingto know you: Learning new user preferences in recommender systems.In IUI, pages127134, 2002.

22. J. Rennie and N. Srebro. Fast maximum margin matrix factorization for collaborative prediction. In ICML, 2005.

23. P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. GroupLens: AnOpen Architecture for Collaborative Filtering of Netnews. In ACM CSCW , pages175186, 1994.

24. B. Sarwar, G. Karypis, J. Konstan, and J. Riedl.Application of dimensionalityreduction in recommender systemsa case study.In ACM WebKDD Workshop, 2000.

25. B. M. Sarwar, G. Karypis, J. A. Konstan, and J. Reidl.Item-based collaborativefiltering recommendation algorithms.In WWW, pages 285295, 2001.

26. U. Shardanand and P. Maes. Social information filtering:Algorithms for automatingword of mouth. In CHI, 1995.

27. L. Ungar and D. Foster. Clustering methods for collaborative filtering.InWorkshopon Recommendation Systems at AAAI, 1998.

28. M. R. W. Hill, L. Stead and G. Furnas. Recommending and evaluating choices ina virtual community of use. In ACM CHI, pages 194201, 1995.

29. U. Bashir, M. Nazir, & A. Yadav. A Mechanism for HandlingCold Startin Book Recommender system by sharing Student Profile,International Journal ofApplicationor Innovation in Engineering and Management (IJAIEM),pp318-322,vol 2,Issue12,December2013.

30. A. Gunawardana. Evaluating Recommender Systems Microsoft Research: A Survey of Accuracy Evaluation Metrics of RecommendationTasks

31. Umar Bashir Mir, MubbashirNazir and Anuj Yadav Fine-Tuned Content-BasedRecommender System Based on Apache Lucene, International Journal of EmergingTrends andTechnology in Computer Science (IJETTCS) pp18-20,vol 3,Issue 2,April 2014.

32. M.Nazir and U. Bashir. Actionable Web Analytics: Customer SegmentationApproach Based on Behavioral Patterns in E-Commerce Industry,InternationalJournal ofEmerging Trends and Technology in Computer Science (IJETTCS) ,pp 227-229, Vol2, Issue6,November-December 2013.

33. M. Nazir, U. Bashir, & A. Yadav. A Mechanism for MiningTop Impacted Segments in Web Analytics based on the Variants used for AB Testing,International Journal of Application or Innovation in Engineering and Management (IJAIEM),pp66-71 vol 3, Issue 2,February 2014.

34. M. Nazir, U. Bashir, & A. Yadav.A Novel approach for evaluating effectiveness of Recommendation Algorithms" IPASJ International Journal ofComputer Science (IIJCS), pp 29-33 vol 2, Issue 4, April 2014.2, Issue 4, April 2014.

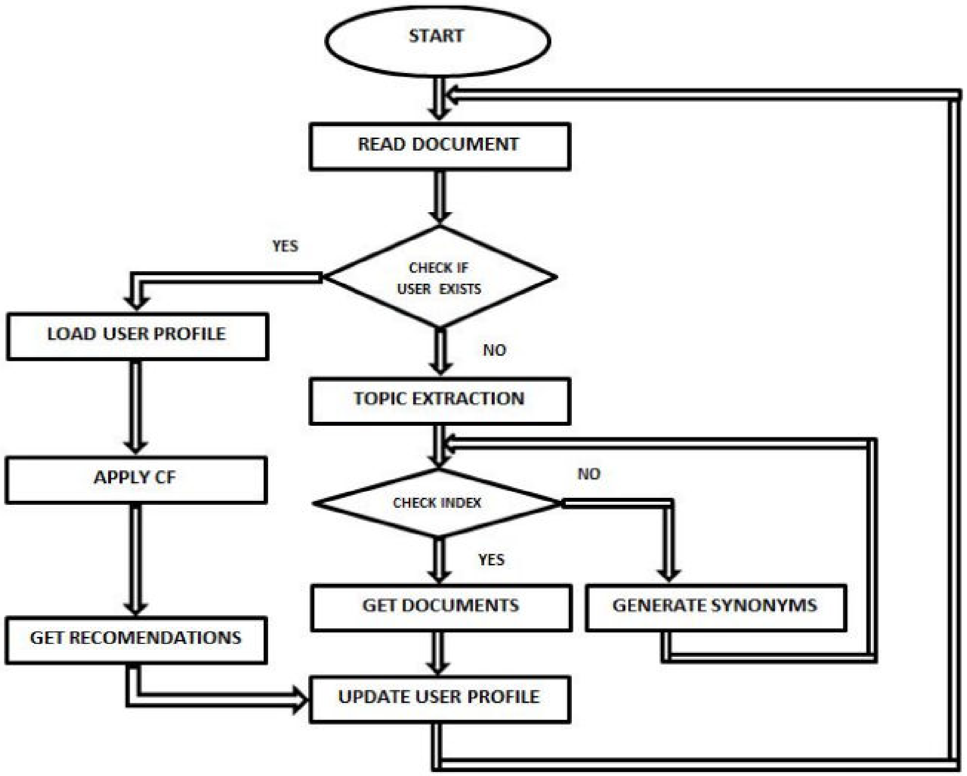


Fig. 1. Overall System Flow Chart

Fig. 2. Dictionary Building

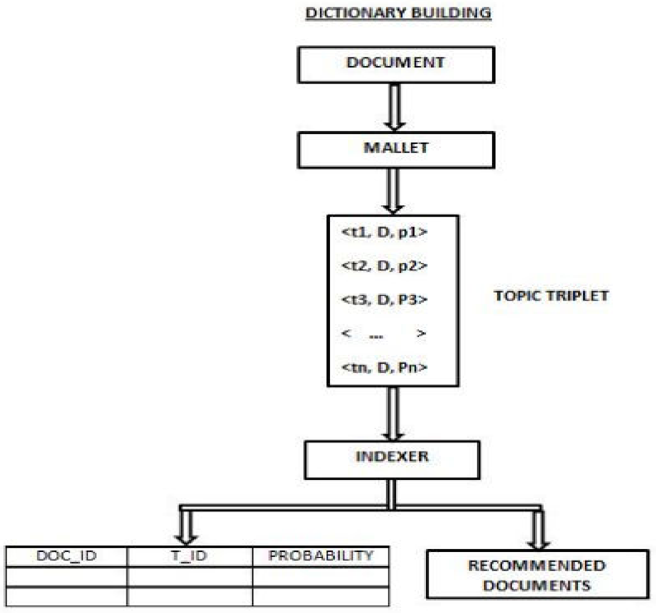


Fig. 3. Content Filtering

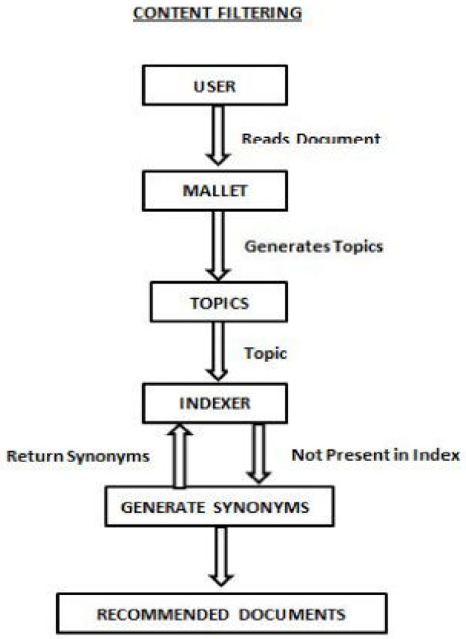


Fig. 4. Collaborative Filtering

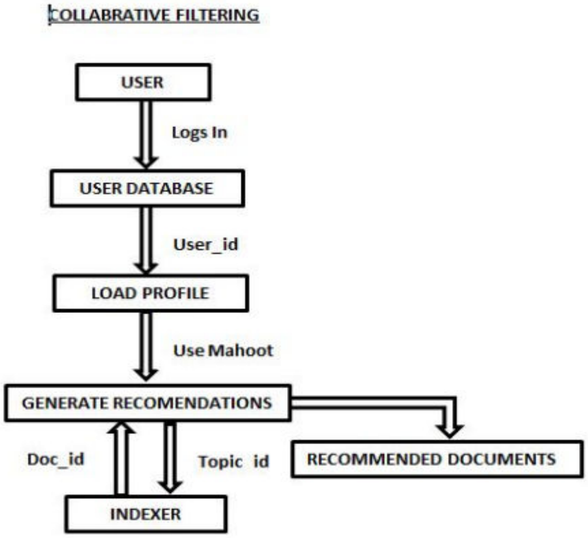
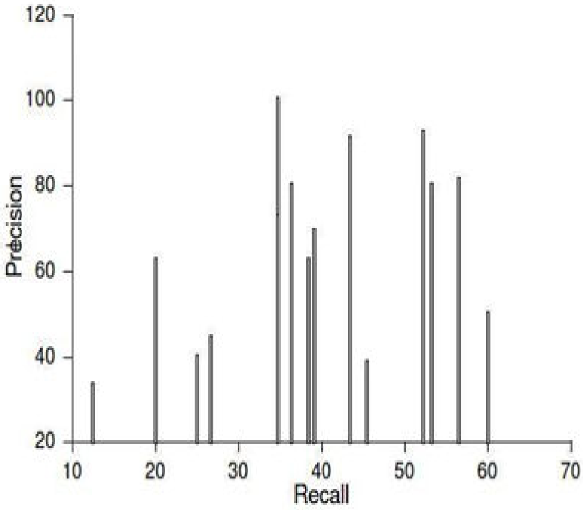


Fig. 5. Precision Vs Recall


Precision
Recall