Time-Ordered Collaborative Filtering for News Recommendation

Akshay Mane[1], Sneha Dixit[2], Mukul Patil, Supriyo Mandal[4],Jyoti Raghatvan[5],Nikhita Nerkar[6]

[1,2,3,4] *Students, Computer Department, RMD Sinhgad School of Engineering, Warje, Pune, Maharashtra*
[5,6] *Professor, Computer Department, RMD Sinhgad School of Engineering, Warje, Pune, Maharashtra*

Abstract - *Recommender systems play an important role in helping online users find relevant information by suggesting information of potential interest to them. Due to the potential value of social relations in recommender systems, recommendation systems have attracted increasing attention in recent years.*

Content based recommender systems have their roots in information retrieval and information filtering research. The content in these systems is usually described with keywords and the in formativeness of a keyword to a document is often measured by TF-IDF weight.

We present Time-Ordered Collaborative Filtering for News Recommendation, a novel semantic-based news recommendation system for users, which filters and recommends news to users based on their choices and preferences.

Key Words: **Collaborative filtering, News recommendation, Sparse matrix, Impact Analysis.**

1.INTRODUCTION

News recommender systems are filters and recommend news to users based on their choices and preferences. For getting the information about the interests of a specific user into news article recommendation, the interests are predicted from the data of user behaviors or the content of previous news articles which are read by a user.

In this paper, we focus on two algorithms Based Collaborative Filtering and Content-Based Filtering using TF-IDF. Data used for processing and recommendation would be extracted from popular online news papers like Times of India, Indian Express, Mumbai Mirror, Navbharat Times, Loksatta, Sakal, etc. after popularity analysis news filtering is done on the basis of two algorithms.

By calculating sparse matrix and impact analysis of news ranking of news is done and top news is recommended to the user.A large amount of data available on Word Wide Web but there is no such tools or technology to extract only required and relevant information from Web databases. Online-newspaper shows the valuable major headlines on the web page from that user wants to retrieve the news which is stored in the databases. News Web pages include content of news as well as noisy data like images, videos, advertisements and other links. Recommendation of top news done after user to user collaborative filtering.

Previous applications require lot of human involvement for news extraction and recommendation, which reduces the scalability. Later approaches focus on how to automatically recommendation framework. So this will reduce human involvement which increases the accuracy& scalability.

2.RELATED WORK

[1] This paper, mainly focuses on the memory based collaborative filtering methods. Aiming at the two important steps of collaborative filtering recommendation, we propose two new methods. One is to compute the similarity and the other is to predict the unknown value. This paper proposes a ratio-based method to calculate the similarity. Comparing the attribute values directly, calculating the similarity between users or between items. This method can be applied to memory-based collaborative filtering as well as to other aspects. Based on the proposed similarity calculation method, we present a new method to conduct the prediction for unknown values.

[2] This paper, present a visual analytics framework for event cueing using media data. As discourse develops over time, framework applies a time series intervention model which tests to see if the level of framing is different before or after a given date. If the model indicates that the times before and after are statistically significantly different, this cues an analyst to explore related datasets to help enhance their understanding of what (if any) events may have triggered these changes in discourse. Our framework consists of entity extraction and sentiment analysis as lenses for data exploration and uses two different models for intervention analysis. To demonstrate the usage of our framework, present a case study on exploring potential relationships between climate change framing and conflicts in Africa.

[3]This paper proposes a multi-stream SGD (MSGD) approach, for which the update process is theoretically convergent. On that basis, propose a CUDA (Compute Unified Device Architecture) parallelization MSGD (CUMSGD) approach. CUMSGD can obtain high parallelism and scalability on Graphic Processing Units (GPUs). On Tesla K20m and K40c GPUs, the experimental results show that CUMSGD outperforms prior works that accelerated MF on shared memory systems, e.g., DSGD, FPSGD, Hogwild!, and

CCD++. For large-scale CF problems, we propose multiple GPUs (multi-GPU) CUMSGD (MCUMSGD). The experimental results show that MCUMSGD can improve MSGD performance further. With a K20m GPU card, CUMSGD can be 5-10 times as fast compared with the state-of-the-art approaches on shared memory platform.

[4]This paper proposed an enhanced measurement for computing QoS similarity between different users and between different services. The measurement takes into account the personalized deviation of Web services' QoS and users' QoS experiences, in order to improve the accuracy of similarity computation. Based on the enhanced similarity measurement, we proposed a location-aware CF-based Web service QoS prediction method for service recommendation. We conducted a set of comprehensive experiments employing a real-world Web service dataset, which demonstrated that the proposed Web service QoS prediction method significantly outperforms previous well-known methods.

[5] This paper describes an approach, to extract the major headlines as well as contents from the news pages, and these Web pages have been taken from the heterogeneous Indian news websites. Since experimental result can be used for real-time applications. the task is to extract the major headline & their contents from the news Web pages and find out the relative major headlines of the newspaper.

3. PROPOSED SYSTEM

The system flow is divided into 3 parts as shown below:

- User Module can register to the system. User module can log in to the system. It can also view the news.
- Database Module: In this the Module can fetch online news papers and store them.
- Recommendation Framework can preprocess the news data. Java Module can apply Collaborative Filtering and Content-Based Filtering on news data and can calculate sparse matrix of news and recommend top news.
  - Social news
  - Sport news
  - Entertainment news
  - Politics news
  - Business news
  - Crime news

At the time of user registration, the user creates their id and password and login to our system.Data used for processing and recommendation would be extracted from popular online news papers like Times of India, Indian Express, Mumbai Mirror, Navbharat Times, Loksatta, Sakal, etc. This can be done by feature-rich content extraction. This extraction mainly depends on Summary of Rich Site or RSS feeds and URL based content parsing of the data embedded in news papers.

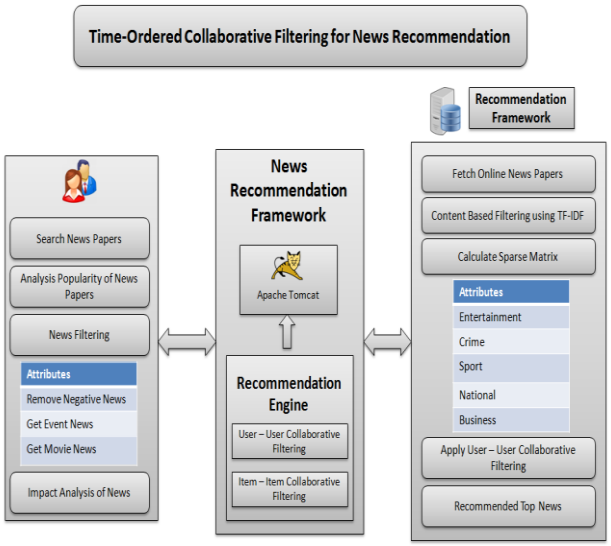

Fig-1: Diagram For Time-Ordered Collaborative Filtering for News Recommendation Architecture

Item data is divided into several attributes example. Consider U may be having attributes A1, A2, A3, A4…An.We have several items in the database may be U1, U2, U3...UN. Each item's attributes are compared to rest of the items in the database. and a score is calculated based on their similarities. Attributes are in many form it may be in text form or in the numeric form columns, which can not be directly matched. Hence an algorithm is used to match these attributes example if U1 has A1 as A, B, C and U2 have A1 as B, C then their matching score would be U1(A1)intersection U2(A1)/#of U1(A1).

Each user is presented as a vector of similarity ratings. For each item, we will be representing their relation with another item in numerical form this matrix will then be normalized so as each attribute value is in between 0-1. And this type of matrix is called as the sparse matrix. Attributes are such as Social, Sport, Entertainment, Politics, Business, Crime etc. After calculating sparse matrix the top-ranked news is recommended to a particular user according to his/her interest area.

4.PROPOSED ALGORITHM

User Collaborative Filtering

We are using User Based Collaborative Filtering algorithm Each user is presented as a vector of similarity ratings.

For example, the ith user will be denoted as Ri. Pearson Correlation Coefficient measures the extent to which two variables linearly relate with each other.

Pearson Correlation Coefficient between ui and uj can be calculated as . Where i denote the set of items rated by both ui and uj and S ∈ R n×n represents the user-user similarity matrix. Ri denotes the average rate of ui.

$$\mathbf{S}_{ij} = \frac{\sum_{k \in I}(R_{ik} - \bar{R}_i) \cdot (R_{jk} - \bar{R}_j)}{\sqrt{\sum_{k \in I}(R_{ik} - \bar{R}_i)^2}\sqrt{\sum_{k \in I}(R_{jk} - \bar{R}_j)^2}},$$

Steps of Algorithm:

Step 1. User u, set of attribute Au for which rating (u, a) is known.

Step 2. Pick K other most similar item u1,....., uk :

$$\mathrm{DIST}(u, u') = \sum_{a_i \in A_u} |\mathrm{rating}(u, a_i) - \mathrm{rating}(u', a_i)|$$

Step 3. Score another user a by popularity with the "similar" ui :

$$\mathrm{SCORE}(a) = \sum_{i=1}^{K} rating(u_i, a)$$

Step 4. Recommended the top-scoring new user.

Content Based Filtering using TF-IDF:

Assuming that a clothing item's style is correlated to the item's brand, styles can be assigned to a large set of clothes by classifying the smaller subset of brands. For the classification of each brand as a member of one of the eight predefined styles, each brand was treated as a point in a 16-dimensional space and it is evaluated by the support vector machine. Eight of these dimensions measure the relevance of a style using the tf-idf method of text mining. b. Term frequency-inverse document frequency (tf-idf) is a statistical method to determine how important an individual word or phrase is to discerning the uniqueness of a block of text. For each brand, a block of text characterizing that brand was created by extracting the adjectives, adverbs, and nouns from the name and description of every item in the affiliate network that belongs to that brand. These text blocks were fed into the tf-IDF algorithm which determines the frequency of each word in the brand's text block relative to the frequency of that word in the global corpus of all text blocks. This is calculated into a tf-IDF statistic, which can be used to weigh the relative importance of that word in assigning a style to a brand. c. The two metrics involved in calculating the tf-idf statistic are term frequency and inverse document frequency. Term frequency is the frequency of a word in a document or the local importance of the individual text block. The inverse document frequency measures the global relevance of a word compared to the complete text corpus, which consists of the text blocks of all the evaluated brands. The tf-IDF statistic is the product of the term frequency and the inverse document frequency. • Term Frequency: occurrence of the word in document • Inverse Document Frequency: log 2(N/documents containing i) Where N is the total number of brand descriptions in the corpus. d. The tf-idf algorithm produces a value of 0 to 1 for each item-style association, meaning the item is likely to be associated with the given style.

5. CONCLUSIONS

Here we present a system Time-Ordered Collaborative Filtering for News Recommendation for users, which filters news among several and recommends news to users based on their choices and preferences. In this, we take advantage of user based collaborative filtering to recommend top scoring news to a user. The advantages of using this system is that it can be implemented in very low cost and provides better accuracy for recommendation of news.

REFERENCES

[1] Collaborative Filtering Service Recommendation Based on a Novel Similarity Computation Method" Xiaokun Wu, Bo Cheng, and Junliang Chen, IEEE 2015 Transactions on Services Computing.

[2] Yafeng Lu, Michael Steptoe, Sarah Burke, Hong Wang, Jiun-Yi Tsai, Hasan Davulcu, Douglas Montgomery, Steven R. Corman, Ross Maciejewski, Senior Member, IEEE"Exploring Evolving Media Discourse Through Event Cueing"IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, VOL. 22, NO. 1, JANUARY 2016.

[3] Hao Li, Kenli Li, Senior Member, IEEE, Jiyao An, Member, IEEE, Keqin Li, Fellow, IEEE"MSGD: A Novel Matrix Factorization Approach for Large-scale Collaborative Filtering Recommender Systems on GPUs" DOI 10.1109/TPDS.2017.2718515, IEEE Transactions on Parallel and Distributed Systems.

[4]]Jianxun Liu, Mingdong Tang, Member, IEEE, Zibin Zheng, Member, IEEE, Xiaoqing (Frank) Liu, Member, IEEE, Saixia Lyu "Location-Aware and Personalized Collaborative Filtering for Web Service Recommendation" Citation: DOI 10.1109/TSC.2015.2433251, IEEE Transactions on Services Computing.

[5] Yogesh W. Wanjari, Vivek D. Mohod, Dipali B. Gaikwad, Sachin N. Deshmukh"Automatic News Extraction System for Indian Online News Papers" IEEE 2014