International Journal of Emerging Technology and Advanced Engineering

User Profile Based Personalized Research Paper Recommendation System Using Top-K Query
Sheetal Patil1, Prof. M. B. Ansari2
1ME-Student, 2Assistant Professor, Department of Computer Science and Technology, Shreyash College of Engineering and Technology, Aurangabad, India

Abstract- Researchers spent lots of time in searching published articles relevant to their project. Though having similar interest in projects researches perform individual and time overwhelming searches. But researchers are unable to control the results obtained from earlier search process, whereas they can share the results afterwards. We propose a research paper recommender system by enhancing existing search engines with recommendations based on preceding searches performed by others researchers that avert time absorbing searches. Top-k query algorithm retrieves best answers from a potentially large record set so that we find the most accurate records from the given record set that matches the filtering keywords.

Keywords- Recommendation System, Personalization, Profile, Top-k query, Steiner Tree

I. INTRODUCTION

The volatile expansion of the world-wide web in combination with promising popularity of e-commerce becomes responsible for large collection of data. To search any information user has to analyze the huge amount of data and then extract helpful information. This may results into exceed information overload problems [4]. So, there is need of robust high technology search engines which will help to find out accurate and precise information among large database.

Recommender systems are developed to bridge the gap between information collection and analysis. Goal of recommender system is to filter all of the available information and to provide most helpful and valuable information to the user. Now days, awareness in web-based recommendation system is rising widely because personalized services provided by such recommendation system are extremely used in internet shopping [5], [7]. Providing personalized of fitted information is required and essential for applying new marketing strategies as web personalization, customer relationship management (CRM), one to-one marketing [8]. Web recommendation system is a specific type of information filtering system that aims to predict the user preferred data. User can move towards the desired information from this model speedily and precisely.

Users search for research papers, based on domain, topics and keywords. So systems modern interests focuses on increasing accuracy of recommendation. Such recommender systems can be applied in variety of applications. One of them is the application which is rising rapidly is online publication of research papers. These days, many academic research papers are coming out from a group of conferences and journals. Academic researchers should go through all the conferences and journals which are correlated to their field of research and find out if there is any new article that may relate to their current works. Sometimes they search the articles from Google scholars with the keywords that might show interesting articles to them [21]. However, these methods require users to commit their time to search articles, which is labor-intensive, and also do not guarantee that they will find the exact articles related to their field of research.

As the numbers of published research papers are increasing day by day, older papers are getting digitized speedily. It is very difficult to find required paper from a large quantity of articles manually. To provide ease to process of finding research papers can be delivered through keyword-based searches or author profile. The search results of researchers with related interests can help direct a more efficient search, but the process of sharing search results is too difficult and time consuming to be feasible. A recommender system helps by automatically recommending papers depending on the preferences of other researchers having similar interests.

There are two main branches of recommender systems as collaborative filtering and content based filtering. Task of content based filtering is to establish relationships between items by examining intrinsic characteristics of the items. Collaborative filtering (CF) systems take advantage of collected information about users’ habits to recommend interesting items. The analysis of user behavior patterns, allows collaborative filtering systems to consider characteristics like author profile, domains, topics and keywords.
II. RELATED WORK

A. Recommendation System

Due to Recommendation systems, many people changed their way to search out products, information, and even other people [3]. The technology behind recommender systems has developed over the precedent 20 years into a rich collection of tools that facilitate the practitioner or researcher to build up effective recommenders. Goal of recommendation system is to deliver information according to user preferences. Recommendation system performs the filtering process to provide best results to user’s queries. Recommendation system can also be referred as ‘Information Filtering Technology’ which assists to discover the research papers which are needed by users speedily and precisely. Recommendation systems used to find research papers depending on domains, topics of research papers and keywords. ‘Extractor’ abstracts the data frequently required by users from huge database [10]. Extracted data get stored in pre-defined XML templates for further processing. Data collected in information extraction process is emerges as the document. Document contains environment for learning the rules of information extraction and extracting required information. Such document can be categorized into three types as 'Structured documents', 'Unstructured documents', 'Semi-structured documents'. As type of document varies, method for learning the rules also varies [9].

Collaborative Filtering

Collaborative filtering [12] approach is widely used to design recommendation systems. This approach depends on gathering and examining a huge amount of information on use’s behaviors, activities or preferences and predicting what users will search for based on their similarity to other users. A key benefit of the collaborative filtering technique is that it does not rely on machine analyzable content and therefore it is proficient for accurately recommending complex items [11], [14]. When building a model from a user’s profile, a discrepancy is often made between explicit and implicit forms of data collection.

Content-based filtering

Content based filtering is centered on description of the item and a profile based on users choice. In a content-based recommendation system, keywords are used to express the items. A user profile is assembled to illustrate the type of item this user likes [13]. This technique has its origin in information retrieval and information filtering research.

B. Personalization

‘Collaborative Filtering’, 'Rule-Based Filtering', and 'Learning agent Filtering' are personalization techniques [2], [19] used for recommendation systems. Collaborative filtering uses other customer’s reviews, ratings, frequently used patterns and preferences to recommend related services. In Learning agent Filtering, learning agents are used to follow user's properties, habits, and personal preferences by examining log file such as website history, access location , frequency, time [15]. Rule-Base Filtering is an approach that system asks the question to users about demographic information and personally individual information, and then further recommends something corresponding to answer depending on the rule used [1].

C. User Profile

The significant information is correlated to the user’s preferences and user favorite data, which are usually referred to as the user profile [18]. User Profile can also identify as user modeling which is extremely dynamic field of research in information retrieval. Generally, User Profiles are assembled which consists of sets of weighted keywords, semantic networks, or weighted concepts, or association rules. User profiles are built from origin of information using a range of construction techniques based on machine learning or information retrieval [20].

III. PROPOSED SYSTEM

Proposed system recommends research papers based on keywords abstracted from research paper title, author profile, and abstract of research papers. User can search research papers according to domain, topic and keyword. Top-k query algorithm is used to retrieve best results from huge data set containing all articles. Top-K query scores every tuple in two ways as Direct Scoring and Indirect Scoring. Keyword extraction process is carried out to abstract keywords. Filtering process is carried out to refine list of research papers and author profiles.

A. Top-k query for Information Retrieval

Information retrieval systems use various approaches to rank query answers. Users are more concerned about the most important i.e., top-k query response in the potentially vast answer space. Different emerging applications demands for competent support for top-k queries. For example, in the context of the Web, the effectiveness and efficiency of meta search engines, which couples rankings from distinct search engines, are highly related to efficient rank aggregation methods [22].

210
Similar applications are present in the perspective of information retrieval and data mining. Most of these applications compute queries that involve joining and aggregating multiple inputs to provide users with the top-k results.

Aim of top-k query is to retrieve best answers from a potentially large record set. Using Top-k algorithm we find the most accurate records from the given record set that matches the filtering keyword, and arrange them according to their scores. An XML tree is also generated with each result set, which is termed as a Steiner Tree. The generated Steiner tree is a list of all the records arranged according to their scores, starting from the nearest result.

Top-K query algorithm uses scores documents against keywords. Here this algorithm is used to score “Tuple Units”. A tuple unit is a set of highly relevant tuples which contain query keywords. Using these tuple units we form tuple set. A single tuple set is formed when two tuples are directly associated with each other.

Top-K query scores every tuple according to two scoring methods, namely, Direct Scoring and Indirect Scoring. Direct Scoring is based on TF-IDF algorithm. TF-IDF stands for “Term Frequency, Inverse Document Frequency”. TF-IDF presents a way to score the importance of words (or “terms”) in a tuple based on how repeatedly they appear across multiple documents.

Scores are termed according to the following criteria:

1) If a word appears frequently in a tuple, it is termed as important and is scored high.

2) If a word appears in many tuples, it is termed as a unique identifier and is scored low.

Therefore, common words like “the” and “for”, which appear in many tuples, will be scaled down. Words that appear frequently in a single tuple will be scaled up.

The second type of scoring method is Indirect Scoring. Indirect Scoring scores a tuple unit based on the keyword that is indirectly present in the tuple unit. The keyword is scrutinized against each tuple unit for indirect similarity.

**B. Keyword Extraction**

Proposed user profile based research paper recommendation system consists of components as UserInterface, Extractor, Filter, KeyFinder, Profile Manager, and Database Manager as shown in figure 1. UserInterface [16] phase contacts with users, takes input from users and generates appropriate outputs for user requests. In proposed user profile based research paper recommendation system, UserInterface is assigned to collect title of research paper or author name to be searched and recommend research paper according to the domain or topics.

Extractor extracts keywords from title of paper, abstract, author profile, and topics. If some article does not contain any keyword, it could be extracted using proper keywords in titles of papers via KeyFinder. Filter component carried out filtering process to refine the collected information which all gets stored in Database. UserProfile [17] contains personal information of authors, title of research paper, abstract, keywords along which can be display using XML files. Profile manager contains all filtered data of authors of research papers.

Filter performs filtering process to provide proper records of authors and research papers to users via UserProfile which contains preference information with XML forms, and provides list of refining research paper titles and author profiles to users.

UserProfile has been stored the user favorite information by topic. When user chooses some record with UserInterface, they replicate user’s preference information by weight value of upgrade.

The process of keyword extraction starts from preprocessing research papers according to domains and topics. The process can be done by extracting title of research papers, author names, and keywords. To get keywords, we can make assumptions as:

1) Keywords section exists between introduction and topic.

2) Keywords section begins with ’Keywords’, ends ’n’ or ’”’.  

3) Keywords section exists in the first page of the paper.

In case where we are unable to extract keywords from published research papers, we trace most appropriately representative words instead of keywords and use them to recommend research papers required by users. Paper used the basic assumption that whatever words are frequently expressed in paper can be treated as topic and keywords.
When the paper doesn't include keywords, we choose the suitable words among the words which are composed of topic. Keywords are abstracted from research paper title, author name. All the articles and words having low priority for example “a”, “is”, “that”, “the”, “for”, etc., are ignored.

We can easily calculate degree of reflection for user posts by using following equation:

$$\text{Reflex} = \frac{\text{# Keyword Title}}{\text{# Title Term}}$$

Where, title term includes all noun and compound noun.

C. User Based Personalization

UserProfile is composed hierarchical architecture as Figure 2.

![Fig 2: Architecture of UserProfile](image)

Where, $WD_{ij}$ = weight of domain i to Topic j, $WT_{jk}$ = weight of topic j to keyword k.

Generally, each user profile stored in XML format which further forms hierarchical architecture as shown in fig.4. Preferred information can be easily indicated by domain and topic of the users. It can be traced easily preference information by Domain and Topic of the users, accommodate simply with a variation of signal information by forming hierarchical architecture. The update of User Profile is made for every click of the research paper collected by topic. Whenever a research paper is selected, it makes increasing of a frequency of Domain, Topic, and Keyword then, recalculates a rate of each occurrence and reflects to UserProfile.

If user searches for domain, then all the topics related to that particular domain and their respective keywords against the domain will be searched. Similarly, when user searches for topic under a particular domain then all the keywords under that topic will be searched.

IV. CONCLUSION

While writing academic research paper, researchers have to spend much more time to find out the latest research papers according to topics. So, robust recommendation system is required to search latest research papers to save time and efforts of researchers. Proposed system provides rapid and effective search of research papers according to the domain, topic and keywords. Top-k query algorithm is used to retrieve best research papers from a potentially large record set. An XML tree is generated with each result set, which is termed as a Steiner Tree. The generated Steiner tree is a list of all the records arranged according to their scores, starting from the nearest result.

Future work will be implemented about grouping of research papers related to specific subject and active recognition of research trends continuously.

REFERENCES


