ISSN No. 0976-5697

Volume 5, No. 7, September-October 2014

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

An Overview of Information Filtering Methods

Mu. Annalakshmi
Assistant Professor (Guest), Department of Computer Science
Government Arts College for Women (Autonomous)
Pudukkottai.

Abstract: The exponential increase of information in World Wide Web has made the retrieval of relevant information a challenging task. Information filtering and retrieval are the two major areas which have gained more attention in the recent years. This paper gives an account on the similarities and differences between information filtering and retrieval, the steps involved in filtering and the types of information filtering. It also throws light on the evaluation parameters for filtering.

Keywords: Information Retrieval, Information Filtering, Ontology, Filtering Types

I. INTRODUCTION

World Wide Web is a huge repository of information about various disciplines. Though the users have access to tremendously large data, they do not get the exact information they need form the several billions of web pages due to the lack of tools. Information filtering (IF) has recently emerged as a technique for effective delivery of the required and also relevant information. IF systems have the following features.

- Can be applied to unstructured or semi-structured data such as documents, e-mails, etc.
- Can handle large amounts of data.
- deal primarily with textual data
- are based on user profiles
- Remove irrelevant data from incoming streams of data.

Many of these features are also common to information retrieval (IR) systems. But there is a little difference. Information retrieval (IR) [1] is the activity of obtaining information resources relevant to an information need from a collection of information resources. Information filtering (IF) [1] is, a special type of information retrieval which removes redundant or unwanted information from an information stream using (semi)automated or computerized methods prior to presentation to a human user. Information retrieval focuses on all the relevant information but information filtering narrows down to the information required by the user based on his profile. Table 1 illustrates the difference between Information Retrieval and Information Filtering[2][5][8].

Table 1: Information Retrieval (IR) and Information Filtering (IF)

	Information Retrieval	***Information Filtering***
Representation of information needs	Queries	Profiles
Information Source	Static	Dynamic
Goal	Selection of relevant data items for query	Filtering out irrelevant data items or data collection, distribution based on profile
Scope of the System	Concerned only with relevance of data items	Also concerned with social issues such as user modeling and privacy
Frequency of use	Ad-hoc; one-time user queries	Repetitive use; long term users
Type of users	Not known to system	System keeps user profiles
User Profile	Not necessary	Essential
Information Seeking Behavior	Short Term	Long Term
User Query	Brief	Description or explanation of the information
User Interaction with the system	Single Information Seeking episode	Series of Information Seeking episodes

II INFORMATION FILTERING

IF Model

An IF system consists of four basic components [5] (see Figure 1).

a) Data Analyser component
b) Filtering component
c) User Model Component
d) Learning Component

The Data Analyser component gathers information from the users. It then analyses, stores them in appropriate format, which is used as input by the filtering component.

The User Model Component collects information about the user preferences either explicitly or implicitly and generates user profiles.

The Filtering Component is the main component of IF system which does the work of filtering the relevant information based on the user profile. The decision whether the information is relevant or irrelevant is made by comparing the user profile with the represented data items. The user determines the relevancy of the information provided by information filter and gives a feedback which is used by the learning component.

The Learning Component updates the user profiles based on the changing preferences and the feedback of the users without which filtering may return irrelevant results.

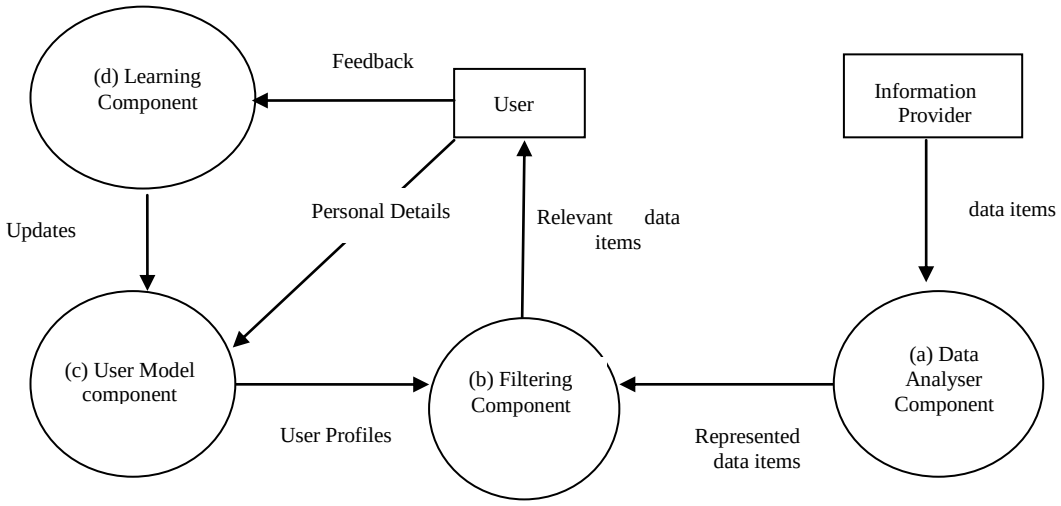

Figure 1. Information Filtering Model

Methods for creating user profiles

The IF systems use various methods to gather information about their users, for the purpose of creating user profiles, which play an important role in filtering of information. These methods can be classified as

- Explicit method based on user interrogation
- Implicit method based on inference from the user behaviour
- Mixed method

Explicit Method

The explicit method requires the use to provide the information about their preferences, priorities, area of interest and other relevant parameters to create their personal profile.

Implicit Method

This method records the user's browsing behaviour and characteristics implicitly without the explicit involvement of the user. The parameters such as time spent by the user in the webpage, whether the user has saved the page or printed the page, etc are used for recording the user behaviour.

Mixed Method

It combines the best features of the above two methods namely implicit and explicit by implicitly inferring patterns form the user and also collecting data from them explicitly. They create an initial user profile and from then onwards they add knowledge to user profile from their browsing behaviour.

Information Filtering Types

Content-based filtering

This approach provides the users with the relevant information solely based on their profile. A content-based filter[3] provides information relevant to the user by analyzing the items rated by individual user and the content of the items and by calculating its similarity with the user's query.

Collaborative filtering

This method uses peer opinions to predict the interest of other users. The filtering is done by selecting and ranking items for a user based on the similarity of the user to other users who

liked similar items in the past. Similarity is calculated by using methods such as Pearson Correlation Coefficient. Collaborative filtering[5] is mainly used in recommender systems. But this method suffers from the problems of cold start and sparsity.

Hybrid Filtering

Hybrid filtering[4][5] combines the concepts of both content-based and collaborative filtering. It uses content-based filtering in the initial stage for creation of user profiles thereby avoiding the cold start and sparsity problems. It then uses collaborative filtering at the later stages to rate all types of data including multimedia items.

III EVALUATION OF IF APPLICATIONS

Two measures namely Precision and Recall are used for measuring the performance of IF applications.

Precision[1] is the fraction of documents retrieved that are relevant to the user's information need.

$$Precision = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|}$$

Recall[1] is the fraction of the documents that are relevant to the query that are successfully retrieved.

$$Recall = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{relevant\ documents\}|}$$

VI ONTOLOGY AND IF SYSTEMS

Ontology[6] is a specification of conceptualization. That is, ontology is a description of concepts and relationships that can exist for an agent or a community of agents. The information filtering systems may use ontology for improving the classic keyword-based search[9][10][11] expansion based on class hierarchies, relationships, etc. The information can be selected based on keywords, their synonyms, hyponyms and hypernyms. Hyponyms refer to the specialization of words whereas hypernyms refer to the generalization of words. These synonyms, hyponyms and homonyms can be referred form WordNet[7] – the lexical database of English words along with semantic relations. Ontology based IF Agents not only search for keywords but also evidence phrases (EP) such as synonyms, hyponyms and hypernyms of keywords generated form Wordnet. They also rate the relevance of websites by calculating the occurrence of EP, frequency of EP and the nearness factor.

IV CONCLUSION

This paper has just made a survey on information filtering, distinct features of information filtering and retrieval, a generic IF model, filtering types and the role of ontology in information filtering. It also focuses on the major parameters for the evaluation of information filtering and the methods for acquiring user preferences. This paper tries to pinpoint that ontology can help in improving information filtering.

V REFERENCES

[1] http://en.wikipedia.org/wiki

[2] Renganathan V, Babu AN,Sarbadhikari SN(2013), "A Tutorial on Information Filtering Concepts and Methods for Bio-medical Searching", J Health Med Informat 4: 131

[3] Peretz Shoval, Veronica Maidel, Bracha Shapira, "An Ontology content-based filtering Method, International Journal on Information Theories & Applications", Vol.15/2008.

[4] Uri Hanani, Bracha Shapira, Peretz Shoval, "Information Filtering: Overview of Issues, Research and Systems".

[5] Qing Li, Yuanzhu Peter Chen, Zhangxi Lin, "Filtering Techniques for selection of contents and products"

[6] http://tomgruber.org/writing/ontology-definition-2007.htm

[7] www.wordnet.princeton.edu

[8] Nicholas J.Belkin, W. Bruce Croft, "Information filtering and information retrieval: Two sides of the same coin?", Communications of the ACM New York, 1992.

[9] Andrew Burton Jones, Veda C.Storey, Vijayan Sugumaran, Sandeep Purao, A Heuristic-Based Methodology for Semantic Augmentation of User Queries on the Web

[10] Kwang Mong Sim, "Toward an Ontology-Enhanced Information Filtering Agent" in ACM SIGMOD Rec., Vol. 33, Mar 2004.

[11] Kwang Mong Sim and Pui Tak Wong, "Toward Agency and Ontology for Web-Based Information Retrieval", IEEE Transactions on Systems, MAN, and Cybernetics- Part C: Applications and Reviews, Vol 34, No.3, August 2004.

[12] Tsvi Kuflik, Bracha Shapira, Pertez Shoval, "Stereotype-Based versus Personal-Based Filtering Rules in Information Filtering Systems", Journal of the American Society for Information Science and Technology, Feb 1, 2003.