ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

COMPARISION BETWEEN THE CONTENT BASED FILTERING AND COLLABORATIVE FILTERING TECHNIQUES

Prof. Pragati Chowhan*, Miss.Harshada Deshmukh, Miss.Chaitali Kolhe
[*] Information Technology,S.G.B.A.U.Amravati, Maharashtra, India
Information Technology,S.G.B.A.U.Amravati, Maharashtra, India
Information Technology,S.G.B.A.U.Amravati, Maharashtra, India

ABSTRACT

On the Internet, content filtering (also known as information filtering) is the use of a program to screen and exclude from access or availability Web pages or e-mail that is deemed objectionable. Content filtering is used by corporations as part of Internet firewall computers and also by home computer owners, especially by parents to screen the content their children have access to from a computer. Content filtering usually works by specifying character strings that, if matched, indicate undesirable content that is to be screened out.

In this paper, we present the combination of the two filtering techniques including the content-based and collaborative filtering. Content-based filtering selects information based on semantic content, whereas collaborative filtering combines the opinions of other users to make a prediction for a target user. In this paper, we describe a new filtering approach that combines the content-based filter and collaborative filter to capitalize on their respective strengths, and thereby achieves a good performance .Finally, we comparing the two filtering techniques and in this two techniques find the result of which one is the best technique.

KEYWORDS: content filtering, Content-based filtering, Collaborative filtering

INTRODUCTION

The rapid growth of Internet leads to substantial information outburst in the cyberspace. The most used services of internet are flooded with anonymous content and source. There is no means to monitor the content of all the web pages on the WWW, and therefore, many web pages have questionable quality. A content filtering technique is a protection between the Internet and a user's computer, blocking access fromg potentially objectionable, offensive, or irrelevant subject. Corporations use it as part of Internet firewall and home computer users, especially parents' use content filtering to screen the content their children have access to from a computer. A content filter's blocking profile might be designed for children to restrict the access of inappropriate websites that contain information on gambling, illegal drugs, racist and sometimes even social networking.Content filtering effectively fix the security flaws in the corporate network.

Most information filtering methods fall into one of the following categories, content-based filtering (CBF) or collaborative filtering (CF) (Oard & Marchionini,1996). CBF selects the right information for users by comparing representations of searching information to representations of contents of user profiles that express the interests of users. For example, search engines recommend web pages with contents similar to user queries (Salton & McGill, 1983).

Collaborative filtering applies the speed of compters with the intelligence of human. CF is a technology wherein peer opinions are employed to predict the interests of others. The techniques of CF have been developed quickly not only in the research area but also in the commercial field.

RELATED WORKS

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

Anandampillai (2005) detailed about the Content Based Multicasting using JADE. The content based Multicasting is performed at the interior nodes of the IP multicast tree. Jean Sebastian (2006) discussed about the content filtering security.

Jeff Youmans (2007) detailed the Privacy and Content Filtering Rights and Wrongs. CIPA Filter's (2009) content filtering system use context sensitive pornography filtering algorithm for filtering web pages in a simplified and effective approach. Shane Hird (2009) showed that content filtering using heuristic algorithm can help alleviate the problems caused by legitimate bulk mail using other technical solutions,as mail is filtered based on the nature of the content. Emanuela Moreale (2003) proposed Agent-Based Approach to Mailing List Knowledge Management through the application of IE, IR and a novel information integration technique to the mailing lists. Liu Pei-yu, Zhang Li-wei, Zhu Zhen-fang (2009) described the research on E-mail filtering based on improved.

CONTENT FILTERING

The last approach is to allow access to the entire Internet but to examine the content retrieved before allowing it through to the user. These filtering products will look for certain .key words. In Web pages or for other characteristics that are supposed to indicate dubious content, such as graphics with large amounts of .flesh tones.. Content that fails to meet the acceptability tests will be blocked, regardless of whether it is a Rubens painting or a Penthouse centrefold.

Content filtering is appealing because it dynamically classifies incoming content as it arrives. Vendors do not have to manually examine large numbers of Web sites, and users do not have to constantly update lists of acceptable or unacceptable sites. The problem it faces is that accurately determining whether content should be allowed through is a very difficult computing task.

Content-based Filtering

Products that use content-based filtering techniques examine incoming content and outgoing requests to determine if they appear to be .unacceptable.. These products employ a variety of methods such as looking for key words, analysing images and looking for .known. characteristics of undesirable. Web pages.

Key word Filtering

Products using key word filtering scan Internet content as it is being loaded into a user computer and look for words that are included in a black list. A page is blocked if it contains any of the words in the block list. Filtering products also often check requests before they are sent out to prevent users from using search engines to find sites that may contain .undesirable. content but not included in the products black lists.

Key word filtering can be very efficient and so is suitable for older, less powerful, personal computers.

There are several problems with key word filtering technologies:

- They only check text, and cannot block objectionable pictures that are not accompanied by (in)appropriate text. This could be a particular problem for pornographic content, as Russian or Japanese sexually explicit photographs look much the same as Australian or US pornography but may not come with any helpful English key words.
- They have to be able to distinguish the .acceptability. of a word from its context. Early key word scanning products had a reputation for being simplistic, blocking words regardless of how they were being used, and unnecessarily blocking access to desirable content as a result. The classic example is the term breast cancer, which would be picked up by a key word filter looking for the word breast, resulting in blocking the entire site.

Phrase Filtering

Phrase filtering is a more sophisticated extension of keyword filtering. Phrase filtering does not consider words in isolation, but as part of a phrase. This allows for more fine-grained classification, as it would allow one to consider the phrases huge breasts, and breast cancer in their respective contexts. While this approach might be expected to do better than keyword filtering alone, it still has many of the associated problems (such as deciding how many objectionable phrases are required before a page is blocked, and being useless for non-English sites), and, in addition, has the added difficulty of having to enumerate all the different phrases that are considered objectionable.

Profile filtering

Several companies have introduced products that filter Internet content based on the characteristics of the received content. Vendors tend to be circumspect on how these products work, but some of the features they look for include the ratio of pictures-to-text and links to other known undesirable sites.

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

Profile analysis can be computationally intensive and result in an unacceptable slow down in perceived Internet access times. Subsequent attempts to retrieve similar content from the same site will be blocked. Content-based filtering is often used in conjunction with other methods, such as URL filtering, to evaluate content that is not already on a black list.

Image analysis filtering

Some filtering products examine images as they are delivered to a user. This approach tries to determine if incoming content contains images of naked bodies, often looking for large amounts of skin tones and on the analysis of images themselves. It is computationally intensive and a difficult task, and computers will invariably experience difficulty in distinguishing between art and pornography, between a Rubens painting and a Penthouse centrefold, or even between pornography and pictures of the family at the beach.

Pages that are found to contain undesirable images are then added to the black list and will not be available in the future.

COLLABORATIVE FILTERING

Collaborative filtering (CF) is a technique used by some recommender systems.[1] Collaborative filtering has two senses, a narrow one and a more general one.[2] In general, collaborative filtering is the process of filtering for information or patterns using techniques involving collaboration among multiple agents, viewpoints, data sources, etc.[2] Applications of collaborative filtering typically involve very large data sets. Collaborative filtering methods have been applied to many different kinds of data including: sensing and monitoring data, such as in mineral exploration, environmental sensing over large areas or multiple sensors; financial data, such as financial service institutions that integrate many financial sources; or in electronic commerce and web applications where the focus is on user data, etc. The remainder of this discussion focuses on collaborative filtering for user data.

Collaborative prediction

Prediction for an item is then calculated by performing a weighted average of deviations from the neighbor"s mean. Here we use the top N rule to select the nearest N neighbors based on the similarities of users. The general formula (Resnick et al., 1994) for a prediction on item i by user k is:

$$P_{k,i} = \overline{R}_k + \frac{\sum_{u=1}^{n} (R_{u,i} - \overline{R}_u) \times sim(k,u)}{\sum_{u=1}^{n} |sim(k,u)|}$$

CONTENT-BASED FILTERING VS. COLLABORATIVE FILTERING

The first approximations to information filtering were based on content [Foltz and Dumais 1992]. These systems select which items to recommend based on their content.

Therefore, the user profile is a representation of the content in which the user is interested. This kind of filtering is especially effective when retrieving text documents, where each document is represented by a set of keywords. However, these systems have several limitations [Shardanand and Maes 1995].

First, the items should be analyzable by a machine. This is difficult when retrieving multimedia information where machine perception of the content (colors, textures, etc.) differs greatly from user perception. Although the assignment of attributes by a person (annotated multimedia content) solves this problem, at least in part, content-based filtering is insufficient to deal with much of the information available today.

Another major problem with content-based filtering is its inability to evaluate the quality of an item. For example, it cannot distinguish a good article from a bad one if both articles use similar words. In fact, the quality of an item is a highly subjective feature that depends on the tastes, ideas, culture, etc., of each person and that would be difficult for a machine to analyze.

Finally, content-based filtering does not have a way of finding serendipitous items that are interesting for the user, that is, really good items that are not apparently related to the user profile. Collaborative filtering systems [Shardanand and Maes 1995] are less sensitive to these problems since they are not based on the content of items but rather on the opinions of other users. The system will recommend items that have received high ratings by other users with similar tastes or interests. In these techniques, the items are actually rated by people. Thus, the system does not need to analyze content (and, therefore, it is valid for any type of item including nonannotated multimedia content), and the quality or subjective evaluation of the items is also considered. In collaborative filtering–based systems, the user profile is the set of ratings given to

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

different items. These ratings can be captured explicitly, that is, by asking the user,or implicitly by observing his/her interaction with the system. Generally, the rating is represented as a unary value (showing only the relevant items), binary (allowing to distinguish between good and bad items) or, more commonly, as a numerical value on a finite scale.The user ratings are stored in a table known as the rating matrix. This table is processed in order to generate the recommendations. Depending on how the data of the rating matrix are processed, two types of algorithms, memory-based and model-based,can be differentiated.Memory-based algorithms use the whole table to compute their prediction. Generally,they use similarity measures to select users (or items) that are similar to the active user. Then, the prediction is calculated from the ratings of these neighbors. (This is why they are also called neighbor-based.) Most of these algorithms can be classified as user-based algorithms or item-based algorithms depending on whether the process of getting neighbors is focused on finding similar users [Resnick et al. 1994; Shardanand 1994] or items [Sarwar et al. 2001].

FUTURE WORK

The areas of both content-based and collaborative filtering, coupled with the test-bed supporting real users, is reach with future work possibilities. Both content-based and collaborative systems can provide such a service, but individually they both face shortcomings. The design of the adapting population of collection agents takes advantage of these overlaps to dynamically converge on topics of interest, both automatically identifying communities of interest and providing the possibility of significant resource savings when increasing the numbers of users and documents. Initial experiments validate our profile construction methods, and show anecdotally that the emergent properties we postulated for collection agents are indeed being exhibited, namely agents specializing to topics and serving multiple users where appropriate.

CONCLUSION

Collaborative Filtering combining the strength of human intelligence in understanding information content with the speed of computers in information processing.Unfortunatelly collaborative filletring techniques alone can be ineffective when users have not rated at item,for new users of the filtering system,or for users who do not generally benefits of the opinions of the others. Content-based filtering techniques can be combined with collaborative filtering technique. A unique approach to integrating content-based and collaborative filtering. CBF can directly select information based on a user own profile contents without the opinions of other users, while CF can recommend information according to other opinions.

REFERENCES

[1] “A Content Filtering With MDaemon 6.0”, Alt-N Technologies, 2002.
[2] BASILICO, J. AND HOFMANN, T. 2004. Unifying collaborative and content-based filtering. In Proceedings of the 21st International Conference on Machine Learning (ICML’04). ACM, New York, NY, 9.
[3] BENNETT, J. AND LANNING, S. 2007. The netflix prixe. In Proceedings of KDD Cup and Workshop (KDDCup’07).ACM, 4.
[4] BILLSUS, D. AND PAZZANI, M. J. 1998. Learning collaborative information filters. In Proceedings of the 15th International Conference on Machine Learning. Morgan Kaufmann, San Francisco, CA, 46–54.
[5] HUANG, Z., ZENG, D., AND CHEN, H. 2007. A comparison of collaborative-filtering recommendation algorithms for e-commerce. IEEE Intell. Syst. 22, 5, 68–78.
[6] Marko Balabanovic and Yoav Shoham.Content-based,collaborative recommendation. Communications of the ACM 40(3) March 1997.
[7] Basu, C., & Cohen, W. W. (1998). Using social and content-based information in recommendation. In Proc. of the AAAI-98.
[8] Hauver, D. B., & French, J. C. (2001). Flycasting: Using collaborative filtering to generate a play list for online radio. In Proc. of Web Delivery of Music.