International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459 (Online), An ISO 9001:2008 Certified Journal, Volume 3, Special Issue 1, January 2013)

International Conference on Information Systems and Computing (ICISC-2013), INDIA.

IMPACT OF VIDEO RECOMMENDATION SYSTEM AND FILTERING TECHNIQUE ON DISSEMINATION OF POLLUTED CONTENT

R. Shiva Ranjani[1] and T. Sheela[2]

[1] PG Student, ME Computer and Communication, Sri Sai Ram Engineering College, Chennai 44;
[2] Professor, Department of IT, Sri Sai Ram Engineering College, Chennai 44.

Shivaranjani1989@gmail.com.
Sheela_saiit@yahoo.com.

Abstract

A wide deployment of Internet User Created Contents (UCC) and Online Digital Video (ODV) enabled the instant growth of online Video and programs which can be designated by Users. Licensed broadcasting companies, a undersized quantity of video and satellite broadcasting operators are the only source for providing contents. So automatically it becomes more challenging, time consuming to find an attractive video using the User Profile Recommendation. We propose the Online Video Recommendation system under a Distributed Computing Environment to assure the customer's requirement. Open opportunistic allows users to upload, share the videos. So it has a possibility of introducing polluted content into the system by spammers and promoters. We address the issue of detecting video spammers and promoters using Supervised Classification Algorithms. It uses a manually labelling which provides high cost. Clustering Hybrid algorithm overcomes this problem by focusing on the features of given video, which increases the chance of generating more rules to classify the videos.

***Keywords*-- User profile recommendation, online video, recommendation, promoters, spammers, Clustering Hybrid Algorithm.**

I. INTRODUCTION

Video tag recommendation is to enrich the textual annotation of multimedia contents. When a new video is uploaded, the labels given by the classifiers are used to suggest the tags which are relevant to the video. We mined large volumes of Web pages and search the queries to discover a set of possible text entity categories and a set of associated relationships that map individual text entities to categories. Finally, we apply these relationships to synthesize a reliable set of categories which is most relevant to videos [8].

A social web site allows people to upload and share their video clips with the public at large or to invited guests. It provides an online service, platform, or site that focuses on facilitating the building of social networks or relations among people who share interests, activities, backgrounds, or real-life connections. Most social network services allow users to interact over the Internet, such as email and instant messaging [6].

As the communities built out of friends and family, they provide a more secure environment for online communication. Unfortunately, recent evidence shows that these trusted communities allow the spreading of malware and attacks [6] .In the existing, providers are very limited such as licensed broadcasting companies. Now users are given open opportunities for sharing and uploading the videos.

Web mining is technique to extract knowledge from Web data. Web becomes the fastest growing and most up to date source of information. The web has had tremendous success in building communities of users and information sources.

Such an opportunistic behaviour allows users to post a video as a response to a video topic. Some users' post unrelated videos as a response to a popular video topic, we call them as spammers. Other users try to gain the visibility of specific video by posting a large number of responses to boost the rank of video topic, we call them as promoters. Spread of such polluted content will automatically decrease user's patients and confidentiality towards such web sites since users cannot identify pollution before watching at least a segment of it.

Spammers and promoters are detected using decision tree classifier algorithm. It can be constructed faster, simple and easy to understand. It performs greedy search for generating the rules. So there is a chance of discarding some important rules. Later, they are detected using associative classifier algorithm. It performs a global search for generating more rules which becomes useless in classification [7].

In this paper, we are going to detect video spammers and promoters by adopting five steps. First, we are going to crawl a user data set from social websites.

Second, we created labelled collection with users “manually” classified as legitimate, spammers and promoters. Third, analyse attributes such as video, user and social to reflect the behaviour of our sampled users. Using these attributes which is based on users profile, user social behaviour and video posted by user, going to identify two types of polluters using lazy associative classifier.

Clustering Hybrid algorithm overcomes this problem by focusing on the features of given video, which increases the chance of generating more rules to classify the videos. It is responsible for an error rate reduction of approximately 10% when compared against eager associative and 20% when compared against decision tree classifier.

The rest of the paper is organized as follows. The next section discusses background. Section III describes related work. Section IV deals with classification algorithms. Clustering hybrid algorithm is described in the section V. Finally, section VI contains the conclusion.

II. BACKGROUND

In this section, we provide information about YouTube OSN, crawled YouTube data and the scope of proposed work.

2.1 YouTube

YouTube is a one of the video sharing online social website, it allows users to upload, view and share videos. Unregistered users can watch videos, while registered users can upload an unlimited number of videos. YouTube accepts the videos only in some formats. It also supports 3GP, by allowing videos to be uploaded from mobile phones.

Only YouTube will offer the users to see its videos even outside their website. Such videos consist of a piece of HTML which is used to embed it on any page on the Web. Video owner can be disable embedding as well as ranking and commenting. Such videos can also be shared through other websites, e-mail, mobile devices, and blogs. Using keywords, users are able to search for content and select the videos.

Unfortunately, YouTube is used to distribute malware. According to Secure Computing, attackers are using a fake video link on the site to initiate infection. We are going to detect such spam activities using clustering hybrid.

2.2 Crawled YouTube data

Process for creating a database is to give access to the collected data with browsing and searching interfaces. We were not only interested in collecting pages and videos from YouTube, but also a set of specific attributes such as titles, descriptions, ratings and comments.

Users provide a set of queries. The system uses these queries and search on YouTube. Set of meta data’s is extracted from results returned from YouTube. We define such data to be the information about the given video which are provided by the author for that video. The context capturing component captures various contextual information about the video items for which the meta data is already collected [2].

Each time such data is captured, a time stamp is recorded. This would include fields such as ratings and comments.

2.3 Scope of Proposed Work

A wide range of attacks are possible in online social network. Multiple types of attacks are executed such as 1) tag spam’s, 2) The bogus link will direct users to a website that forces the download of malicious videos, 3) Product advertisement, 4) Phishing attacks. Although purpose of each attack varies, main aim of spammers is to pollute the content of social network. First, our aim is to detect the users who spread the video pollution based on attributes that capture the feedback of users with respect to each other or to their contribution to the system. Second we use active learning approach to detect spammers.

III. RELATED WORK

Spread of content pollution has been observed in various applications, including Web search engines [9], Social networks [5]. A number of detection has been proposed [9], [1], [2], [4], [7], [5].

Content pollution can be detected based on text corpus. Text corpus is a structured set of texts. When taking notes while reading any form of text, information about each word is added to the corpus in the form of tag. In multimedia content, tag cannot be described well so recommendation based on tag may suggest unrelated video. To reduce the tag spam, we proposed clustering-based strategy to identify the spammers.

In web search engine, documents are indexed by textual search. It allows the web users to enter the term that are used to match the queries. Uploaders can provide textual information in several forms including title, description and a set of free form tags. Tags and title should have correct information about the content of video. We segment title and tag of each video independently and extract n gram. Construct lookup table by mapping n gram to a set of videos. Discard the n gram which corresponds to less or more videos.

Videos that are considered as similar based on recommended tags are identified and grouped together into sets labelled with textual categories.

We need to ensure the categories which use n gram is "consistent". For a category to be consistent associated classifier need to agree by having score above the threshold for the same video. It train the classifier based on content and uploader-supplied meta data.

IV. METHODOLOGIES

There are so many classifiers used in the existing. They are briefly explained below [7] and the comparative study is shown in **table 1**.

4.1 Decision Tree Classifier

At each internal node, the best split is chosen according to the information gain criterion. A DT is built using a greedy recursive splitting strategy. Decision tree can be considered as set of disjoint decision rules, with one rule per leaf. Such a greedy (local) search may discard important rules and expands only the current best rule. It can be constructed fast when compared to other methods and it is simple and easy to understand

4.2 *Eager Associative Classifier*

Associative classifier performs a global search and generates large number of rules and many rules may be useless during classification. Eager Associative Classifier mines all possible CARs with a given minimum support. During the testing phase, Associative classifier checks whether each CAR matches the test instance. The class associated with the first match is chosen. Eager Associative Classifier Steps: 1. Algorithm mines all frequent CARs. 2. Sort them in descending order of information gain. 3. For each instance, the first CAR matching is used to predict the class.

Eager associative classifiers search for CARs in a large search space. CARs that are important to some specific test instances may be missed. Eager classifiers generate CARs before the test instance is known. It often combines small disjuncts in order to generate more general predictions. This can reduce performance in highly disjunctive spaces, where single disjuncts may be important to classify specific instances.

4.3 *Lazy associative classifier*

Lazy learning methods postpone generalization model until a query is given. Lazy Associative Classifier induces CARs specific to each test instance. 1. It projects the training data only on features in the test instance. 2. From this projected training data, CARs are induced and ranked and best CAR is used. It is context-sensitive and focus the search for CARs in a much smaller search space, which is induced by the features of the test instance.

Lazy classifiers are often most appropriate when the search space is complex. Classifier is better when more CARs are generated.

LAC learns the classification function in two steps:

4.3.1 Demand-Driven Rule Extraction: The search space rule is huge and computational restrictions must be imposed during rule extraction. Let D and T be the sets of labelled training data and unlabeled testing data. Minimum support threshold σ_{min} is employed to select frequent rules. It is delayed until a set of users in T is given for classification. Each individual user d in T is used as a filter to remove irrelevant features and examples from D.

4.3.2 Prediction: some rules show stronger associations than others. Using a single rule to predict the class may produce error. Two key parameters of LAC are the maximum size of the rules and the minimum confidence

4.4 *Clustering Hybrid classifier*

Clustering hybrid classifier relies on an effective selective sampling strategy to deal with the high cost of labelling large amounts of examples. If the examples were randomly sampled, clustering hybrid classifier can learn to detect spammers and promoters using fewer labelled examples than would be required.

4.4.1 Sampling Function: Consider a set of unlabeled users U. select the users who carry almost the same information of all users in U. These informative users will compose the training data D. If a user u_i € U is inserted into D, then the number of useful rules for users in U that share feature values with u_i will increase. The users returned by sampling function are inserted into D.

Table 1
Comparative Study

	DECISION TREE CLASSIFIER	EAGER ASSOCIATIVE CLASSIFIER	LAZY ASSOCIATIVE CLASSIFIER
SEARCH	Local	Global	Global
CONSTRUCTION	Fast	Slow	Slow
PERFORMANCE	Poor	Good	Good
RULES	Generate one best rule per leaf	Generate more CARs	Generate more CARs but select one best CAR

In the next round, the new user is inserted into D so number of rules extracted from D for each user in U is changed and the sampling function is executed again.

Initially D is empty and no rule is extracted. The first user to be labelled is inserted into D which is selected from set of available users U. They form a cluster and used to extract the rules

4.4.2 Natural Stop Condition: The algorithm stops when all available users in U are less informative when compared to any users already inserted into D. this happens when ALAC selects a user which is already in D.

V. Proposed Work

Fig 1, shows an architectural diagram for our proposed work. In a user side, when a new user signs up, it asks about favourites so that it can recommend the videos when the user login again. Videos are tagged based on title, description and free tags. Based upon user query, it analyses and makes pattern matching in a server and recommends the related videos. In recommended search result, it recommends a video based upon video and user attributes.

In a provider side, when new provider signs up, it allocates private storage space for individual providers. In database it stores each provider's password, user name, mail id, contact number, etc. When provider uploads a video, it checks whether video topic is related to video. When a unrelated video is posted by spammers. It is detected using attributes and filtered by that corresponded provider.

5.1 Individuals Internet Terminals

User Behaviour Monitor (UBM) dynamically learns database user's access pattern, and automatically detects and alerts suspicious database access.UBM allows you to monitor user's behaviour via four types of Guarded Items:

5.1.1 Object Policies: monitors suspicious reads and writes on specific objects.

5.1.2 User Policies: monitors suspicious reads and writes by specific users.

5.1.3 Session Policies: monitors suspicious session behaviour.

5.1.4 User-Defined Rules: permits construction of your own rules.

In its Learn mode, UBM watches database usage and analyses usage events over time in order to compile a statistical behaviour profile. That profile enables you to derive your own definition of normal access behaviour. With a baseline, it becomes easier to identify, and prevent, malicious behaviour. In its Guard mode, UBM watches database usage on a near real-time basis and provides alerts, for specific-point-in-time access events.

5.2 Private Storage Space

The web space can serve two basic purposes. In the first place, it allows you to upload file information (HTML files, image files, etc.) on the World Wide Web where it will be available at a global scale. Second, this resource enables you to store various files that are not visible to website visitors but play an important role for the proper functioning of your website. Some of the popular 'invisible' files taking up web space on the server where your website is located are PHP files, database files and CGI program files. PHP files are stored on the server with a .php extension and are used for various important on-site activities such as order form processing for online stores, poll results management, etc. Database, in turn, store data such as product codes, customer details, etc., which is retrieved by PHP scrips and CGI programs. CGI programs serve for processing data inputs from online forms, which require that the collected information be stored on the website's server. Other web space occupying files worth mentioning include externally linked CSS files and JavaScript files.

Server provides a private space for each user. It can provide each user to block spam activities individually.

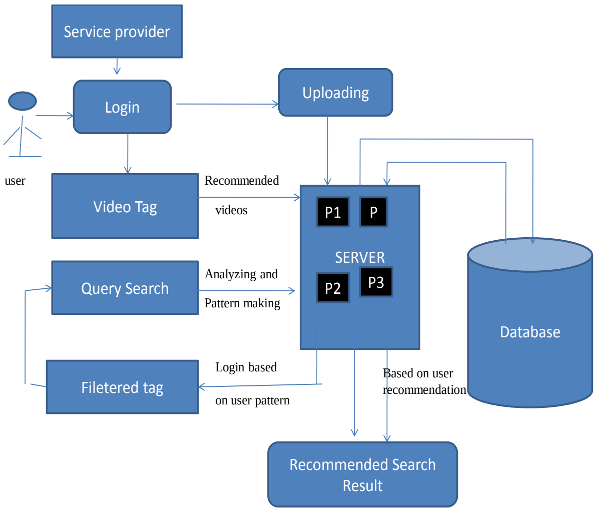

Fig 1. Architectural diagram

VI. Conclusion

Promoters and spammers can pollute video retrieval features of online video SNs, usage of system resources is wasted, and effectiveness is reduced and not satisfying user satisfaction. We propose an effective solution which automatically detects spammers and promoters in online video SNs.

Supervised classification approaches are able to detect the promoters and many spammers but misclassifying small number of legitimate users. Thus, our proposed approach deals with clustering of highly informative users using active learning approach. This reduces manually labelling where the cost is too high.

REFERENCES

[1] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, “Detecting spammers on twitter,” in Proc. Annu. CEAS Conf., 2010, pp. 1–9.

[2] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, and M. Gonalves, “Detecting spammers and content promoters in online video social networks,” in Proc. Int. ACM Conf. Res. Develop. Inf. Retrieval (SIGIR), 2009, pp. 620–627.

[3] R. Agrawal, T. Imielinski, and A. Swami, “Mining association rules between sets of items in large databases,” in Proc. Int. Conf. Management Data, 1993, pp. 207–216.

[4] C. Costa, V. Soares, J. Almeida, and V. Almeida, “Fighting pollution dissemination in peer-to-peer networks,” in Proc. ACM SAC, 2007,pp. 1586–1590.

[5] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Zhao, “Detecting and characterizing social spam campaigns,” in Proc. ACM SIGCOMM Conf. Internet Meas., 2010, pp. 3547.

[6] P. Heymann, G. Koutrika, and H. Garcia-Molina, “Fighting spam on social web sites: A survey of approaches and future challenges,” IEEE Internet Comput., vol. 11, no. 6 , Nov./Dec. 2007, pp. 36–45.

[7] A. Veloso, W. Meira, and M. J. Zaki, “Lazy associative classification,” in Proc. Int. Conf. SDM, 2006, pp. 645–654.

[8] George Toderici, Hrishikesh Aradhye, Marius Pasca, Luciano Sbaiz, Jay Yagnik, “Finding Meaning on YouTube: Tag Recommendation and Category Discovery”, Google Inc.1600, Amphitheatre Parkway Mountain View, California 94043.

[9] D. Fetterly, M. Manasse, and M. Najork, “Spam, damn spam, and statistics: Using statistical analysis to locate spam web pages,” in Proc. Int. WebDB, 2004, pp. 1–6.