International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 11 | Nov -2016 www.irjet.net p-ISSN: 2395-0072

A Recommendation System With Spam Reduction Based On Clustering With On Demand Service

R.Vignesh[1] and D.Kavitha[2]

[1]*Student,Department of Computer Science and Engineering, Valliammai Engineering College, TamilNadu, India.*

[2]*Professor, Department of Computer Science and Engineering, Valliammai Engineering College, TamilNadu, India.*

Abstract -*We have seen a numeral online video distribution websites, out of which YouTube is the most admired. It allows the users to upload a video under any particular category. This aspect provides a path for the malicious users to contaminate the system. We can find n number of spammers i.e. those who post unrelated or unwanted videos under a category. For example, if someone wants to make their video admired, the spammers will try to upload the scrupulous video under different categories. The users who search for their interest may get these spam videos as response. We suggest an efficient system to overcome this drawback by introducing a technique to reduce the spam contents. And also to bring forth the user interests with better results reducing network overhead. We also propose a system to avoid user interruption by providing a feature for the user or the video provider to opt for their category during registration with web server.*

***Key Words*:clustering, cloud storage, recommendation system, spam reduction, content filtering.**

1.INTRODUCTION

YouTube is a much admired video viewing website and we all find it so easy and user friendly comparing to other video viewing websites. Still it faces some kind issues with it. However these issues are not considered into account to rectify it though it is a colossal negative aspect of the entire system.

As it is a very well-liked website and we can find any kind of videos in that, and moreover it doesn't require any signup or login to view certain videos though it has age filters for some kind of videos. Still some malicious users can misuse the open facility of YouTube.

Secondly, it requires a massive storage capacity. It provides storage for each and every single user viewing videos in it. Even if n numbers of users are watching the same video, it provides separate storage for each and every single user viewing that video. Finally, the recommendation system of our existing system which provides recommendation based on video tags. Therefore, even malicious users can post unrelated videos with different video tag names which are considered to be immense drawback. However, it provides likes, dislikes, comments, etc., it is not taken into account. These videos are said to be spam videos and it cannot be eliminated. It is considered to be a big drawback in such a popular video viewing website.

In this paper, we propose different methodology to overcome the issues facing by the YouTube. A secure video viewing website is provided. Only authenticated users can able to view the videos as per their interests. To stick with the user interests more sharply we have provided a option to select the user's interest during their authentication. Hence we can reduce user distraction from the unrelated videos. And also it provides faster recommendation reducing the network overhead and reduces buffering of videos [4].

The system uses a big database called cloud for storage which is an on-demand service. Storage can be reduced or can enlarge as per the requirements. The system uses a clustering methodology for the efficient storage. If n numbers of users are watching a single video, instead of storing based on the individual users, the users are clustered based on their interests. This helps the users to obtain their favorites based on top list ranking based on content-filtering [2][9].

The very interesting and main role of the system is revealing of spam content [5]. The system provides a separate signup for users and video providers. The user/ the provider need to select their interest during the authentication process. The provider will be allowed toupload the video content in their selected category. Similarly, the user will be allowed to view the videos in their selected category.

If a particular user feels that the content is unrelated to their interest, then he/she is allowed to report the content. If the video has been reported by over a limit of users, the content is removed from the storage. Other than this, if the content uploaded by a particular provider is being reported continuously by the user, the particular provider itself will be blocked from uploading any more videos in the system [8].

2. PRELIMINARIES

Here we define the set of rules used in our proposed system.

2.1. Recommendation System

The user recommendation system is designed with the help of K-means algorithm. The K-means algorithm is efficient compared to the reputation algorithm. It is simple, robust, fast and easier to understand. It gives the best result even when the data are unique or well separated from each other.

K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The process follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters).

The major idea is to identify k centers, one for each cluster. These centers should be placed in a archness way because of unusual position causes different result. So, the improved choice is to place them as much as possible far away from each other. After thatit is to take each point belonging to a given data set and associate it to the nearest center. When no point is imminent, the first step is completed and an early group age is done. At this position we need to re-determine k new centroids as bar center of the clusters ensuing from the preceding step. After we have these k novel centroids, a new obligatory has to be done bordered by the same data set points and the adjacent new center. A ring has been created. As a consequence of this ring we may observe that the k centers modify their position step by step until no more modify are done or in other words centers do not shift any more. At last, this algorithm aspires at minimizing an objective function known as squared error function given by:

$$H(U) = \sum_{i=1}^{F} \sum_{j=1}^{F_i} (||y_i - u_j||)^2$$

where,

'$||y_i - u_j||$' is the Euclidean expanse linking y_i&u_j.

'F_i' is the amount of statistics points in i^{th} cluster.

'F' is the amount of statistics points in j^{th} cluster.

2.1.1. Algorithmic steps for k-means clustering

Consider Y = $\{y_1,y_2,y_3 \ldots y_n\}$ be the set of data points and V = $\{u_1,u_2,\ldots\ldots u_c\}$ be the set of centers.

1) At random go for 'F' cluster centers.
2) Determine the distance between each data point and cluster centers.
3) Allocate the data point to the cluster center whose distance from the cluster center is bare minimum of all the cluster centers.
4) Re-determine the new cluster center using:

$$U_i = (1/F_i) \sum_{j=1}^{F_i} y_i$$

where, 'F_i' represents the number of data points in i^{th} cluster.

5) Recalculate the remoteness linking each statistics point and new acquired cluster centers.
6) If no statistics point was reallocated then stop, or else do once more from step 3.

2.2. Spam Reduction

In this system the spam can be perceived and reduced with the help of Lazy Associative Algorithm. ALAC (active lazy associative classifier), which relies on an effective selective sampling strategy to deal with the high cost of labeling large amounts of examples.

LAC was extended to allow itself to select the subset of examples to be labeled, thus performing active classification. It does this sequentially, using the requested labeled examples to inform its decision of which example to opt for next. The anticipation is that by only call for the tag of most informative examples, ALAC can learn to detect spammers and promoters using significantly fewer labeled examples than would be required if the examples were randomly sampled. Next, we describe the sampling function used by ALAC as well its stop condition.

The classification results discussed in the previous section are obtained assuming that the correct identification of a user is equally important for users from all three classes. However, there might be scenarios in which a system administrator could prefer to correctly identify more users from one class at the possible expense of misclassifying more users from the other classes.

For instance, a system administrator, who is interested in sending automatic warning messages to all users classified as spammers, might prefer to act conservatively, avoiding sending messages to legitimate users by mistake, even if this comes at the cost of reducing the number of correctly identified spammers and/or promoters.

In contrast, another system administrator, who adopts the policy of manually inspecting each user flagged as polluter before sending a warning, might prefer to favor the accurate detection of spammers. In that issue, miscategorizing a few more genuine users has no immense penalty, and may be adequate, since these users will be cleared out through physical scrutiny.

2.2.1. Algorithmic steps for ALAC

Consider W be the set of all n training instances.
Consider G be the set of all m test instances.

1) For each t_i ε G do
2) Consider W_{ti} be the projection of W n features only from t_i.
3) Consider Z^l_{ti} be the set of all rules {Y ->D} mined from W_{ti}.
4) Sort Z^l_{ti} according to information gain
5) Select the first rule {Y -> Z} ε Z^l_{ti} , and predict class Z.

Decision tree classifiers carry out a greedy explore that may throw away imperative rules. Associative classifiers perform a worldwide search for rules; however it may generate a large number of rules. Lazy associative classifier conquer these problems by focusing on the features of the given trial instance.

3. RELATED WORKS

3.1. Recommendation System

Generally, the design of recommender system that uses knowledge stored in the form of ontologisms. These recommender system uses a database to generate the recommendations. We may use a traditional two way recommender algorithm which is implemented by reducing the three way correlation to three-two way correlation & then applying a fusion method to re-associate these correlation[7].

Recommendations for users not on their similarity but on their trust relation to other users.Trust is meant to be anticipation of anmediator to be able to rely on some other representativessuggestions. The repayment of these reliance based algorithm includes strong personalization, no need to have a long rating history in the system[11].

In traditional recommender system there are categories: collaborative filtering and content based filtering. Collaborative filtering is a technique that automatically

predicts the interests of an active user by collecting rating information from other similar user or items[2]. These two filtering methods as drawback which may affect the functioning of the system. The collaborative filtering requires a large collection of user history data. The content based filtering lack the ability of understanding the interests and preferences[8].

3.2. Spam Reduction

There are set of spam ingredients namely spam model and spam metrics. Spam model is used to evaluate whether a content is a spam or not. Spam metric is nothing but if a content is said to be spam how the particular content is said to be spam is estimated. Spam model is of two types: synthetic model and trace-driven model. Synthetic model is based on set of assumptions. While the trace-driven model is based on real data observation. This is the mostly preferred model because they can produce more realistic results[5].

3.3. On-Demand Service

On-demand computing is an increasingly popular enterprise model in which computing resources are made available to the user as needed.The on-demand model was developed to prevail over the common challenge to an enterprise of being able to meet fluctuating demands efficiently. Because an enterprise's requirement on computing resources can vary drastically from one time to another, maintaining enough resources to meet peak requirements can be costly. On the other hand, if the enterprise cuts costs by only preserving minimal computing resources, there will not be an adequate amount of resources to meet peak requirements.

3.4. User Privacy in Cloud Computing

User privacy is also required in cloud. By using privacy the cloud or other users do not know the identity of the other user. The cloud can hold the user accounts for the data in cloud, and likewise, to provide services the cloud itself is accountable. The validity of the user who stores the data is also verified. There is also a need for law enforcement apart from the technical solutions to ensure security and privacy.

4.PROBLEM FORMULATION

4.1. System Model

1. Input design is the process of converting the user-oriented explanation of the input into the computer-based system. This blueprint is important to avoid errors in the data input process and show the correct direction to the management for getting information from the computerized system.

2. It is achieved by creating user-friendly screens for the data way in to handle large amount of data. The goal of designing input is to make data entry easier and to be free from faults. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides proof viewing facilities.

3. When the data is entered it will test out for its soundness. Data can be entered with the help of screens. Suitable messages are provided as when needed so that the user will not be in maize of immediate. Thus the objective of input design is to create an input layout that is easy to pursue.

5. CONSTRUCTION

In this section, we describe about the structure of the overall system module by which the well-organized system recommendation is provided for the users with the help of clustering and the lessening of spam.

5.1. Private storage Formation

We make a Private Storage Space for every Provider in our media storage Server. At the time of creating a Provider account the video storage space will be billed to the provider. That memory of the storage space is not a fixed, it can large-scale storage. From this set-up we can upload videos in private.

The Private Storage Formation (PSF) monitors the objective of consumer's private profile. The PSF supports management, scheduling, security, confidential control of the consumer's profile and the required resources.
The Media storage server is a web computing based storage for media contents which are broadcast over hundreds of broadcasting channels. Content Vendors (CV)

such as licensed broadcasting companies, small to medium operators, and content producers, store their own media contents on the media storage server. Service Agents (SAs) provide contents to consumers from the MSC, and generate statistical information, including a consumer's preference for contents based on the consumer's profile and analysis of their viewing history. MSC updates the user profiles at the Private Computing.

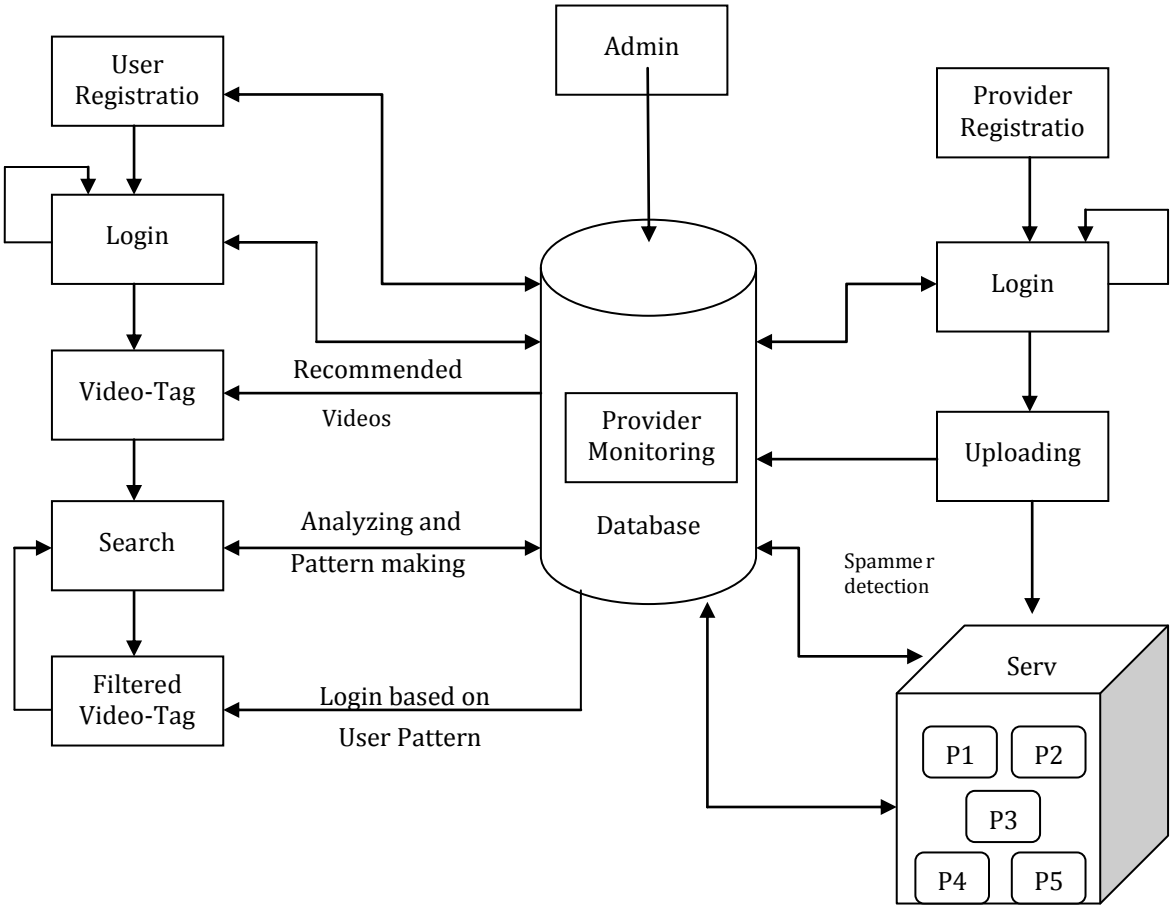

Fig.1. Overall System Architecture for the Recommendation System.

Videos uploaded from provider are stored to their own Storage space. Video files are uploaded to server based on the keywords. This is the private media storage space, they can assess and view there video files only private Storage Network is also a multi-purpose distribution platform that lithely, rapidly, and firmly distributes media content for reuse. You can use the private storage as origin storage for a content delivery network, embed content on websites, generate content-depot drop boxes, and use white lists to authorize right of entry via IP, or deploy child accounts for distribution groups.

The even supports leading de-duplication and compression technologies from Ocarina and others, ensuring the private Storage Network is an unsurpassed solution for digital content storage and distribution.

5.2. User Recommender systems

A content-based recommendation system recommends the most likely matched item then compares the recommendation list to a user's previous input data or compared to preference items.

A content-based recommendations system is based on information searching and generally uses a rating method which is used in the information searching. To measures for computing the user similarity, namely tag cloud-based cosine (TCC) and tag cloud similarity rank (TCSR). The Profile Filtering Agent (PFA) creates a personalized channel profile based on the accumulated viewed content list by using a content based filtering.

Users can recommend the videos to the user itself, at the time of user profile creation. The Recommended videos post to the client profile as video tag system. The video tag is generated based on the user Recommended.

Recommender system or recommendation systems (sometimes replacing "system" with a synonym such as platform or engine) are a subclass of information filtering system that seek to predict the 'rating' or 'preference' that a user would give to an item (such as music, books, or movies) or social element (e.g. people or groups) they had not yet measured, by means of a model built from the characteristics of an item (content-based approaches) or the user's social environment (collaborative filtering approaches). Recommender systems have become extremely common in recent years.

5.3. Content filtering and Reusability

A content-based recommendations system recommends the most likely matched item. To compares the recommendation list to a user's previous input data or compared to preference items. A content-based recommendations system is based on information searching and generally uses a rating method which is used in the information searching. The Profile Filtering Agent (PFA) creates a personalized channel profile based on the accumulated viewed content list by using a content based filtering.

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. Content filtering is used by corporations as part of Internet firewall computers and also by home computer owners, particularly by parents to monitor the content their children have right to use to from a computer.

Content filtering typically works by specifying temperament strings that, if coordinated, specify adverse content that is to be screened out. Content is classically monitored for pornographic content and sometimes also for aggression- or hate-oriented content. Detractor of content filtering programs point out that it is not difficult to unintentionally exclude desirable content.

Content sorting out and the yield that suggest this verify can be alienated into Web filtering, the screening of Web sites or pages, and video filtering, the screening of video for spam or other objectionable content.

A Web filter is a program that can screen an incoming Web page to determine whether some or all of it should not be exhibited to the user. The filter makes sure the source or content of a Web page in opposition to a set of policy provided by company or person who has installed the Web filter.

5.4. Spammer detections

Spammers may post an unrelated video as response to a popular one. We detect the spammers using customer suggestion private storage formation process. Lazy associative classification algorithms to automatically detect spammers.Categorizing them as spammers, promoters, and justifiable users. Using our trial gathering, we propose a categorization of content, character, and social feature that help differentiate each user class. We then scrutinize the practicability of via supervised classification algorithms to robotically identify spammers and promoters, and evaluate their usefulness in our test collection.

Analyzed a variety of video, individual and social attributes that reflect the behavior of our sampled users, aiming at drawing some insights into their relative inequitable

power in individual justifiable users, promoters, and spammers.

Fourth, using the same set of attributes, which are based on the user's summary, the user's social activities in the technique, and the videos uploaded by the customer as well as her intention (responded) videos, we explored the probability of applying supervised learning methods for identifying the two envisioned types of polluters.

We consider two state-of-the-art supervised classification algorithms, namely, support vector machines (SVMs) and lazy associative classification (LAC). We evaluated both algorithms over our test collection, finding that both techniques can effectively identify the majority of the promoters and spammers.

6. CONCLUSION& FUTURE WORK

Promoters and spammers can pollute video retrieval feature of online video SNs, compromising not only user satisfaction with the system, but also the usage of system resources and the effectiveness of content delivery mechanisms such as caching and content delivery networks. We here proposed an effective solution that can help system administrators to detect spammers and promoters in online video SNs. Relying on a sample of pre classified users and on a set of user behavior attributes, our supervised classification approaches are able to correctly detect the vast majority of the promoters and many spammers, misclassifying only a very small number of legitimate users.

Thus, our proposed approach poses a promising alternative to simply considering all users as legitimate or to randomly selecting users for manual inspection. Moreover, given that the cost of the labeling process may be too high for practical purposes, we also propose an active learning approach, which was able to produce results very close to the completely supervised solutions, but with a greatly reduced amount of labeled data evolve. We intend to explore other refinements to the proposed approach such as to use different classification methods, perhaps combining multiple strategies.
We believe that better classification effectiveness may require exploring other features which include temporal aspects of user behavior and also features obtained from other SNs established among YouTube users. Additionally, we intend to explore a better combination of features to improve classification results. Finally, we also plan to extend our general approach to detect malicious and opportunistic users in other online SN sites and contexts.

As the future work, user summary have been achieved from statement information, but users for eternity make no comment after viewing their fascinated video, which directs to errors during clustering. For future work, we plan to switch the data scarcity of user summary. One more important point that should be deliberate is scheming a dispersed recommendation cache to get better recommending hit rate. The cache can also diminishworking out pressures caused by the quantity of simultaneous rule reordering and implementation.

REFERENCES

[1]MagdaliniEirinaki, Malamati D. Louta "A Trust-Aware system for personalized user recommendations in social Networks", Members, IEEE, and IraklisVarlamis, Members, IEEE

[2] H. Ma, H. Yang, M. R. Lyu, and I. King, "Sorec: Social recommendation using probabilistic matrix factorization," in Proc. 17th ACM Conf. Inf.Knowl. Manage. 2008, pp. 931–940.

[3] X.Wu, Y. Zhang, J. Guo, and J. Li, "Web video recommendation and long tail discovering," in Proc. IEEE ICME, 2008, pp. 369–372.

[4] Z.-D. Zhao and M.-S.Shang, "User-based collaborative-filtering recommendation algorithms on Hadoop," in Proc. WKDD, 2010, pp. 478–481.

[5]Paul Heymann, Georgia Koutrika and Hector Garcia-Molina "Fighting spam on social websites: A survey of approaches and future challenges" Standford University.

[6]Chi-Cheng Tsai,Ching-I Chung,Yi-Ting Huang, Chia-HsingShen, Yu-Chieh Wu and Jie-Chi Yang "VCSR: Video Content Summarization for Recommendation" Graduate Institute of Network Learning Technology.

[7]AlexandrosNanopoulos"Item Recommendation in Collaborative Tagging Systems". IEEE transactions on systems, Man and Cybernetics.

[8] Bo Shao, Dingding Wang, Tao Li and MitsunoriOgihara "Music Recommendation based on acoustic features and user access patterns" IEEE transactions on Audio, speech and language processing.

[9] Alan Mislove, MassimilianoMarcon, Krishna P. Gummadi, Peter Druschel, Bobby Bhattacharjee Measurement and Analysis of Online Social Networks" Max Planck Institute for Software Systems.

[10]Guozhu Dong, Jiawei Han, Joyce M.W. Lam, Jian Pei, Ke Wang, and Wei Zou "Mining Constrained radients in Large Databases" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING.

[11] F. E. Walter, S. Battiston, and F. Schweitzer, "Personalized and dynamic trust in social networks,"in Proc. 3rd ACM Conf. Recommender Syst., 2009, pp. 197–204.

[12]Getting videos onto your websites "websitehelpers.com/video/".

[13] Identifying Video Spammers in Online Social Networks"web.cse.lehigh.edu/2008/.../benevenuto_2008_spam_video.

[14] Search Engine Click Spam Detection Based on Bipartitewww.thuir.cn/group/~YQLiu/publications/wsdm2014.

[15] Recommender system - Wikipedia, the free encyclopedia."en.wikipedia.org/wiki/Recommender_system"

[16]SimRate:ImproveCollaborative Recommendation Based"link.springer.com/chapter/10.1007%2F978-3-642-17313-4_17".

.