The Music Resource Recognition for Internet

Wenqian Shang
School of Computer
Communication University of China
Beijing, China
e-mail: shangwenqian@cuc.edu.cn

***Abstract*—Along with the rapid development of the Internet, Music resources on the Internet grow very quickly. Its huge number makes the audience can not select what they need from the music files rely solely on traditional publicity and audition. According to the characteristics of the music files in the current Internet, we propose a machine learning-based platform for music recommendation. With data mining techniques, through behavior analysis of user actions, user preferences, it can achieve the automation of music resources push.**

Keywords-music resource recognition; machine learning; recommendation; data mining

I. INTRODUCTION

Along with the rapid development of the Internet, the explosive growth of all kinds of resources on the Internet, a lot of people's traditional habits change, especially in the music market, traditional methods such as music tape, CD and so on are gradual extinction, but the music market has not subsided but with the development of new media, networks, mobile Internet it has been more prosperity. In 2010, the overall market size of China's online music will reach 3.6 billion dollars (ISP total revenue), which is an increase of approximately 14.4% over 2009. The online music market income increased substantially, and the size of the online music market revenue reaches 44.5 million dollars, which is an increase of 64% from 2009. Wireless music market reaches 3.2 billion dollars (to the total revenue of the service provider), representing an increase of 9.8% over 2009 and more than 87.8% in the proportion of the overall size of the online music. It has become the backbone of supporting and promoting the development of the online music market force. In the year of 2010, the telecom operators earn revenue of 4.9 billion dollars in the wireless music market, which is an increase of 3.5% over the previous year. With the growth in the number of Internet music and fast-paced life, people can not select their favorite music as they do in the past by-audition, so how to recommend the user's favorite music to them quickly in current massive data will be a means to significantly improve the user's experience.

Music sites mostly classify the music files according to the style of music and the concert information of music. Users can find their favorite music files according to category, but this form can meet the need of users in the case of a small number of music files. But now what people face is music files from all over the world, the number will be far more than the number of music files traditional user can contact. Users now face the information explosion and the accelerated pace of life, there is no time for them to select the music files they like.

So in this paper, we propose a platform based on machine learning, gathering music resources from Internet, using data mining techniques to analyze user behavior and user preferences, in order to achieve the automation of music resources push.

II. THE ARCHITECTURE OF THE SYSTEM

The platform consists of the following modules:

A. The Discovery and Aggregation of Musci Resource

In this part, in order to get the music resource on the internet, we adopt the technology of focused crawler and web page structure analysis. With the help of these technologies, we can get the free music sources on the internet and at the same time, we can aggregate the music material and music resources provided by the third party.

B. Music Resource Downloads

After finding the music resources, we adopt the technology of audio stream data analysis to download the music resources to the local area, then make use of the resource more easily.

C. Storge and Management of Unstructured Music Data

As the music resources downloaded from the internet are unstructured, we could not store it directly. So in this part we use object-based technology to manage and store the unstructured data to the local area.

D. Retrieval and Access of Music Resource

Under the general framework of content aggregation, the search for digital content will be significantly simplify the user to get the music source by search engine interface provided by aggregating platform. Business search and access to music resources will be hold in a distributed environment.

E. Personality Push

The function of recommendation system based on server-side is limited by the web server's functionality, increasing the pressure of web server, there is a great threat to user privacy [1][2][3]. Therefore, we adopt the method of binding web log mining [4][5][6] with client-side data binding together. We can get what the user really like

through behavior analysis of user actions, user preferences analysis.

The system architecture can be shown as fig. 1:

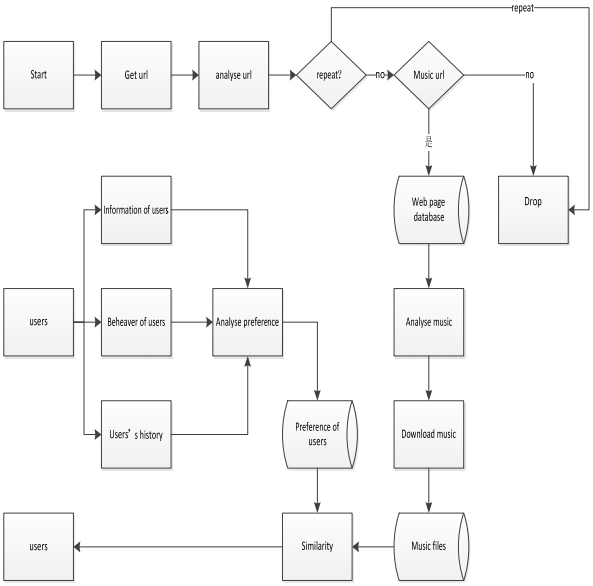

Figure 1. System Architecture.

III. THE KEY TECHNOLOGIES

A. *Discovery of Music Resource*

For automatic discovery of music resources on the Internet and effective integration with other related resources, the system is mainly consists of extracting the information from website, judging nature of website and learning feature of websites. This can be shown as fig. 2:

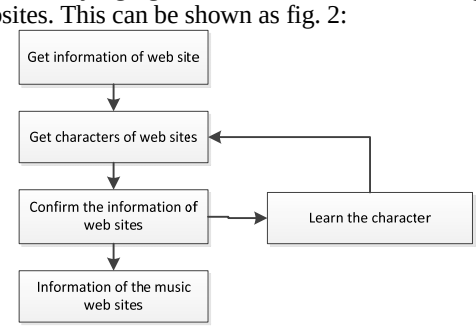

Figure 2. Discovery of Music Resources

Our system manages and balances searches by active search algorithm for Internet music resources. After getting website information, we analyze the pages using the technique of text extracting and structure analysis of pages to determine firstly whether the site is a music source website. The result is then recorded into the information library of music resources. In order to raise the accuracy of determination for websites, our system also studies the features of websites by intelligent learning.

B. *Activity Discovery Algorithm for Internet Audio & Video Resources*

With the rapid development of microprocessor technique, high-performance network technique and systematic software technique, cluster computing has gradually become a kind of computing resource with high cost performance in the area of large amount of calculation. All these tasks are managed and finished by multiple interconnected node computers. In the process of arranging the computing tasks, it has become a hot topic to realize the balance for load and the maximization of resource utilization rate. In comparison, traditional researches treat different resources separately, without considering the relationships among them or the characteristics of tasks and the difference of types and abilities among these tasks.

Analysis of Information Distribution Mechanism: Asynchronous communication among systems can dramatically raise the ability of task-processing; the message-sending process can make the system more reliable with its ability of fault recovering; the platform-crossing system makes it possible for the both sides of communication to have different physical platforms. As a result, it is necessary to analyze and determine the characteristics, working modes and exception handling of different physical platforms to design reliable and platform-crossing asynchronous information distribution mechanism for the preparation of correctly distributing searching tasks.

Research for self-adapting task managing technique: different node computers differ in system resources, while different searching tasks require different amount of calculation. In order to make sufficient use of the system resource in each node computers, namely, the maximization of resource-utilization-rate, it is needed to do researches on self-adapting task managing technique for the benefit of maximally matching system resources of all the idle node computers to the amount of calculation for searching. Meanwhile, it is needed to calculate current searching tasks and available idle resources according to the feedback from all the node computers to make advanced management to the tasks.

Research for dynamic load managing technique: Self-adapting task managing technique maximizes the utilizing rate of system resources. Under this situation, the load on each node computer differs a lot. Some of the nodes are doing infinite calculations, while some others are always idle. Thus, to further optimize the system, it is needed that the load on each node should be managed dynamically according to the type and calculating ability of system resources of each node while the amount of calculating task remains the same.

C. *Text Extraction and Structure Analysis of Pages*

Information on the Internet mainly appears as HTML pages. The pages contain large amount of structure less HTML data and some half-structured XML data. On the other hand, Internet information extraction has become the base of web information processing such as information searching and text abstraction. Thus, text extraction and

structure analysis of pages are made to be an important research topic in the area of Internet information search.

Our paper combines page structure analysis and machine learning techniques to meet our needs. Techniques related to DOM are used to analyze the structure of pages, to extract URLs and to process them by web crawlers. Likewise, information of music resources is processed by the audio processing model, while text information is analyzed in detail. The process is shown as follows:

Automatic word segmentation for Chinese characters is needed for the extraction of information in Chinese. Here we combine NLP and Machine Learning techniques to optimize the speed and quality of our word segmentation algorithm.

This sub-system gets and stores music resources on the Internet automatically. It is supposed to be able to support common packaged formats such as wav, mp3 and wma, eliminating encoded or copy-prevent protected audio streams. These encoded data should be processed by other methods, like paying.

This model contains mainly five sub-nodes. The first is the pre-analyzing node for the propagation mode of programs, which analyzes the propagation mode of the file being downloaded according to the information provided by the program source locating library. The second node is the transfer protocol analyzing node, while the third is the file downloading node. The fourth one is the Transcoding and restoring node, which transcodes the downloaded files and restores them to the server. The last is the program preview node. Downloaded files can be previewed through this node.

D. Mass Structureless data Processing

Audio and video information is anywhere on the internet, and as a result we should deal with it to meet the needs of users. Internet audio and video programs and their abstract, or their website information are mass resource in the form of multimedia. Since the video and audio information on the Internet, such as pictures, music, video programs can rarely be associated to be structured. In our system, we decide to process these data using structureless data managing technology based on object to realize the management of obtained multi-media information. Using the theory of OO, Object is the abstraction of entities in the real world. It is assembled by the data describing the inner status and the operation to them. According to OO theory, we can abstract the information entities as numbers, which provides the combination of data and service with an available structure.

In the number of objects, the data stream means the properties of these number objects. Inside the data stream, the combination of multiple types of number resources and different types of metadata is packed together. According to the needs, any amount or type of data can be involved in the same number object. When the need changes, only the adjustment to the inner data or metadata is needed, which makes number objects more flexible.

IV. CONCLUSION

Our platform has made advance to current music recommendation system. In this system we have introduced into our system the techniques of machine learning to deal with huge amount of Internet music files with efficiency. Our system has in a large scale to satisfy the needs of our users. Since there is now rare good solution to the protection of music copyright, and that it is becoming a trend to make the copyright of music safe. Now copyright protection is becoming the key point in the advancement of our following modules including the crawling module and the recommendation one. We will take all these into consideration in our further work on our video and audio recommendation system.

V. ACKNOWLEDGMENT

This paper is partly supported by "The technology research of cinema management system and collaborative network service platform" (2012BAH02F04).

REFERENCES

[1] O. R. Zainne, M. Xin, and J. Han, "Discovering web access patterns and trends by applying OLAP and data mining technology on web logs," Proc. Of the IEEE International Forum on Research and Technology Advances in Digital Libraries. Los Alamitos: IEEE CS Press, pp. 19–29, 1998.

[2] G. Paliouras, C. Papatheodorou, and V. Karkaletsis et al, "Clustering the users of large web sites into communities," Proc. Of the 17th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers, pp.68–73, 1892.

[3] A. Nanopoulos and Y. Manolopoulos, "Mining patterns from graph traversals," Data and Knowledge Engineering, 37(3): 243-266, 2001.

[4] S. Berkovsky, Y. Eytani and T. Kuflik, et al, "Privacy enhanced collaborative filtering," Workshop on Privacy Enhanced Collaborative Filtering Edinbrugh Uk, 2005.

[5] J. Canny, "Collaborative filtering with privacy," Proc. Of the Annual ACM Symposium on Applied Communicating, Oa-Kland, pp. 45-57, 2002.

[6] R. Ramakrishnan, B. J. Keller and B. J. Mirza, et al, "Provacy risks in recommenders system," IEEE Internet Computing, 11(2): 54-62, 2001.