

One Multi-source Heterogeneous Web Resources Aggregation Model

Wang Min, Wang Yongbin

School of Computer, Communication University of China, 100024, China

Abstract - To meet the diversification of the demand and terminal of web user's resource service, the paper brought a MSAM Model, which belongs to the multi-source heterogeneous web resources. First, it analyzed the current situation of web resource aggregation. Then, it raised a MSAM Model which is suitable for guiding the construction of web aggregation system, based on which an aggregation software system was given. Finally, through the case of the construction of the web music aggregation platform, it verified the rule of MSAM Model. Practice has proved that MSAM Model can realize a unified description of unstructured web resources, form a manageable web resource model, and provide a unified access mechanism of multi-source heterogeneous web resource.

Index Terms - multi-source heterogeneous, web resource, aggregation model, uniform Access

1. Introduction

On Web Sites such as Blog, BBS, WEB portal and social web, many existed approximate resources should be applied to user preference. Web resources aggregation mechanism is able to select, analyze and classify the resources from the Internet artificially or mechanically, then provide user with more useful and more targeted web resources[1-3,6-9]. Web resources aggregation is not simple resources accumulation, but the classification and selection of web resources after in-depth mining and analysis. The model usually selects, classifies, even dynamically restructures the contents through composite diagram. The output of web resources aggregation should consider the influence of user's characteristics.

Generally speaking, the current existing web resources is equipped with characteristics such as massive volume, multi source, isomerism, multiplex mode, uncertainty, dynamics and complex inline. These characteristics pose huge challenges to the web resources aggregation services. Therefore, it is a must to breach the current aggregation technologies, to meet the demand of application, to solve the integrated intelligence management problem of multi-source heterogeneous resource. It is also a must to refine an analyze model and a resource service mechanism of uncertain resource set which adapts to open environment[7-10]. For these purposes, the paper raised a MSAM Model, which belongs to the multi-source heterogeneous web resources.

2. The Status of Web Resources Aggregation

The purpose of web resource aggregation is to select, analyze, and classify the resources from the Internet artificially or mechanically. Finally, it provides users with excellent and useful web resources. According to the media-type, it can be divided into information aggregation, papers aggregation, audio aggregation, video aggregation and composite news aggregation. For example, in the field of

music, web resources aggregation is to integrate music on popular social media. In the websites, such as Blogs, forums, SNS websites, Twitter, P2P webs, Last.fm, iLike, the job is to analyze the feature of these websites, find resources and dynamically reorganize them into lists. In the field of music, this aggregation music platform already includes cool dog music, cool music, shrimp music, nine cool to listen to, a musical, multi-meter music, watercross radio, Baidu heart to listen, music TV and so on.

This paper summarized the existing researches on web resources aggregation, holding the view that it can be divided into link aggregation, social recommendation technology, interest adaptation technology, aggregation portal technology, resource pooling technology, and resources warehouse technology.

(1) Link aggregation

Link aggregation technologies include RSS, TAGs, search engine technology. These three technologies do not move resources from one location to another. Therefore they belong to Link Aggregation.

(2) Social recommendation technology

The social recommendation polymerization technology effectively improves the accuracy of resources recommendation, increasingly becoming an effective way to exploit valuable resources on the internet.

(3) Interest adaptation technology

Polymer filtration, based on users' interests, filtered to a new resource display page during polymerization.

(4) Portal display technology

Through the web resources aggregation portal system, it displays diversified resource services.

(5) Resources warehouse technology

Resources warehouse technology converts the information that needs to be shared into web resources that meet the regulatory requirements in advance and stores them in a public web resources warehouse. The real resources still exist in the original websites. Not only this technology ensures the rapid renewal of virtual view, it also contributes to the unified scheduling when web resources are requested.

(6) Resource pooling technology

This technology uses the resource pool to achieve resource load balancing after the polymerization and improves access performance. After polymerizing resources, it can achieve load balancing of resource access for multiple sources.

3. Related Concepts

Aggregation of web resources can be divided into two phases[4-7], including resource integration phase and customer service phase. Aggregation of web resources web

resources covers the whole process which is initiative to provide users with web resources. Resource integration is a prerequisite of user services, and customer service is the purpose of resource integration. Resource integration is completed by the information provider, and the service is targeted at Internet users. Integration emphasizes the integration of resources from different sources, while services stresses on user-oriented, user interests and meeting the diversified requirement of users.

Definition 1. Resource aggregation process: the capacity of user-oriented resource integration and service under the condition of Internet $f(r)$.

$$f(r) = \frac{\text{interest}(user)}{\text{integration}(r)} \quad (1)$$

The $\text{integration}(r)$ and $\text{interest}(user)$ represent web resources integration and user interest adaptation

Definition 2. Resource integration phase: the process of collecting or searching, cleaning, identifying and normalizing resources $\text{integration}(r)$.

$$\text{integration}(r) = \text{normal}[\text{SimFind}(\text{clean}(\text{snap}(r)))] \quad (2)$$

$\text{Snap}(r)$, $\text{clean}(r)$, $\text{SimFind}(r)$, $\text{normal}[r]$, represent of the five resource operations mentioned above. The $\text{integration}(r)$ must return a normalized homogeneous set of resources $r\text{NormalSet}$.

Definition 3. Resource Service phase: One kind of construction of mining and analysis, restructuring ability in the same set of resources or resource configuration construct warehouse to use relevant information to create user interest perceived ability, on the basis of which to accomplish the process of dynamic resource recommendation $\text{interest}(user)$.

$$\begin{aligned} \text{interest}(user) &= (\text{resourcedw}(r\text{NormalSet}) | \text{findInterest}(user)) \\ &\triangleright \text{recommendFor}(user) \end{aligned} \quad (3)$$

$\text{resourcedw}(r\text{NormalSet})$ should represent the mining and analysis, restructuring operations of homogeneous set of resources. They $\text{findInterest}(user)$ represent the operation of perception of user interest $\triangleright \text{recommendFor}(user)$ represents the recommendation of the dynamic resource based on both of them.

4. The Web Resource Aggregation Model

Thus, the current web resource aggregation web service function is weak, decentralized service functions and service resources, service model and service means such as a single aggregate impact of web resources development bottlenecks. We proposed a web resources aggregation model named MSAM based on the existing web resources polymerization techniques.

Definition 4. Resource aggregation states: For any resources r ,

$$\forall r, \text{AggStatus}(r) \in \{\text{collected}, \text{converged}, \text{mined}, \text{serviced}, \text{accessed}\} \quad (4)$$

The order of constraints is $\text{collected} \rightarrow \dots \rightarrow \text{accessed}$; which represent collected state, converged state, mined state, serviced state and accessed state.

With respect to the stage of definition 1, collected , converged , mined respond to $\text{integration}(r)$; serviced , accessed respond to $\text{interest}(user)$. Therefore, the division of aggregation of definition 3 $\text{AggStatus}(r)$ meets the polymerization process proposed in definition 1 $f(r)$.

Definition 5. Resource aggregation model:

$$\text{MSAM}(user) = \sum_{i=1}^5 \text{layerFunction}_i(r) \quad (5)$$

$\text{layerFunction}_i(r) = \text{Transfer}(\text{status}_{i-1} - \text{status}_i)$ represents the hierarchical conversion from status_{i-1} to status_i .

According to the definition 3 of MSAM mentioned before, the Aggregation Software System based on MSAM should contain five levels – the gathering layer, the collective layer, the mining layer, the service and the access layer. The gathering layer is responsible for gathering and searching web resource from the Internet. The collective layer is responsible for the normalization of multi-source heterogeneous cyber source. The mining layer realizes the dynamic structure, the identification and the evolution management of the music set with data warehouse technology. The service layer is responsible for recommending resources to the users and adapting them to the users' interests. The access layer is responsible for providing access modes for a variety of terminals.

The five layers of MSAM are bottom-up and can gradually support each other. They finally build the framework of intelligent resources management model. The details are shown in the figure 1.

(1) The Gathering Layer:

This layer contains five components, including the RSS acquisition of cyber source and web search, the data analysis of cyber source, the cleaning of cyber source, the semantic search of cyber source and the source registration management of cyber source. The Gather Layer is responsible for gathering, storing, assembling and managing the cyber source which is structuring, semi-structuring or non-structuring. In such process, the Gathering Layer needs to clean the cyber source. In addition, on-demand searching should be realized here, as well as the related searching technique. The Gathering Layer can realize the gathering management of cyber source with extending the search scope and integration scope by semantic net.

(2) The Collective Layer:

This layer contains five components, including the standardization management of cyber source, the ontology semantic modeling of cyber source description, the similarity identification of cyber source, the normalization of cyber source and the management of resource pool. The Collective Layer is responsible for the standardization and normalization of the heterogeneous cyber source according to the unified

description. The Collective Layer is able to analyze the usability of multi-source heterogeneous cyber source from various aspects such as the information source, information itself and the user demands. In a unified feature space, the Collective Layer is able to realize the semantic modeling,

measure the similarity and normalize the semantics of the generous multi-source heterogeneous cyber source. In conclusion, the Collective Layer can exploit the statistical regularity of the multi-source heterogeneous cyber source and can solve the problem of semantic fusion.

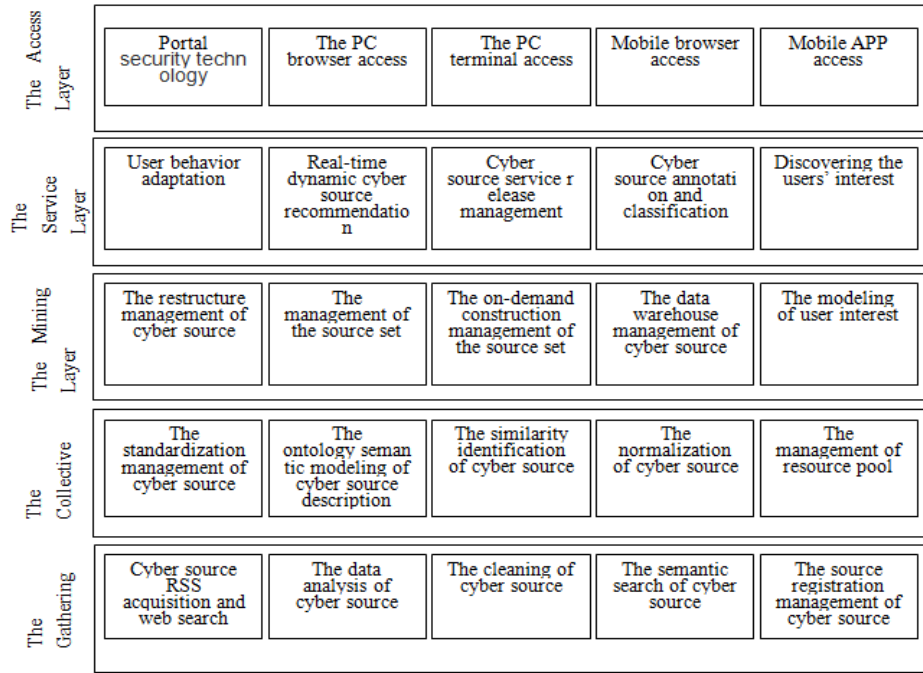


Fig1. MASM structure model

(3)The Mining Layer:

This layer contains five components, including the restructure management of cyber source, the management of source set, and the on-demand construction management of the source set, the data warehouse management of cyber source and the modeling of user interest. The Mining Layer is responsible for the restructure management, where knowledge discovery of multi-source heterogeneous cyber source is a problem. To solve the problem, resource warehouse technology is used, so that generous cyber source is able to be refined, integrated and restructured. In addition to the resource warehouse, the data mining is used to analyze the dynamic changes of the music set, realize key functions and provide technical support for the dynamic restructure of cyber source as well.

(4) The Service Layer:

This layer contains five components, including the user behavior adaptation, the real-time dynamic cyber source recommendation, the cyber source service release management, the cyber source annotation and classification and discovering the users' interest. The Service Layer is responsible for adapting to user's interest and providing on-demand services. It aims to set a demand driven real-time dynamic resource service system, which can catch the users' demands, behaviors and interests. In addition, the Service Layer will model the problem to form a group of resource service and manage the service release.

(5)The Access Layer:

This layer provides portal technology and access environment safety to support diversified terminal access including PC browser and terminal access, mobile browser access and mobile app access.

5. MSAM Framework Application

In essence, the aggregation of web resources can be divided into two stages, the resource integration phase and the resource service phase. When a user visits, Access Component identifies the user; and then identifies the user's interests through the Service Component, and returns the user's interest model; thirdly, the Mining Component conducts dynamic identification and resource set construction based on the user's interests; In the process, Converge Component should be made full use of to normalize heterogeneous resources. Again, the Collect Component search for resources to support the upper resource process. Finally, the user adaptive intelligence service is realized by the aggregation of resources.

MSAM model has been used successfully in the construction of some online music resource integration platform (musical aggregation web) MSAM model provides basic technical support for this kind of site. Main function of this musical aggregation web is making it accessible to converge and visit music resource uniformly. This system is mainly located in audio media automatically aggregate

system. On one hand, the system makes it possible to collect all kinds of music media information in real time and regularly; on the other hand, the providing of a variety of personalized services enables users browsing and retrieving for the web music they need.

Using the MSAM model, the musical aggregation platform is easier to realize:

(1) By aggregating dispersed music resources, the cost of obtaining resources has been reduced.

(2) Platform does not need to create its own music, on the contrary, just selects and edits resources from the Internet.

(3) For the procedure of aggregating essence music resources from the Internet, advantage in music resources is much higher than that of other single site.

(4) Only provides links to Internet music resource, and

the cost of hardware can be significantly reduced.

The main business structure of this system is shown in Figure 2 below.

The system is able to provide different types of users with uniform integrated music resources access, and form three types of integration models including integration of terminal users, web music resources and user interests, which is achieved by the use of MSAM model. In order to ensure the leading place of web music platform, it supports the diversity of the music source and the variety of music resource itself. Access to different types of music resources from different databases is available as well in this system. Meanwhile, for the system is open and extensible, the integration of existing online music and other service is achieved effectively.

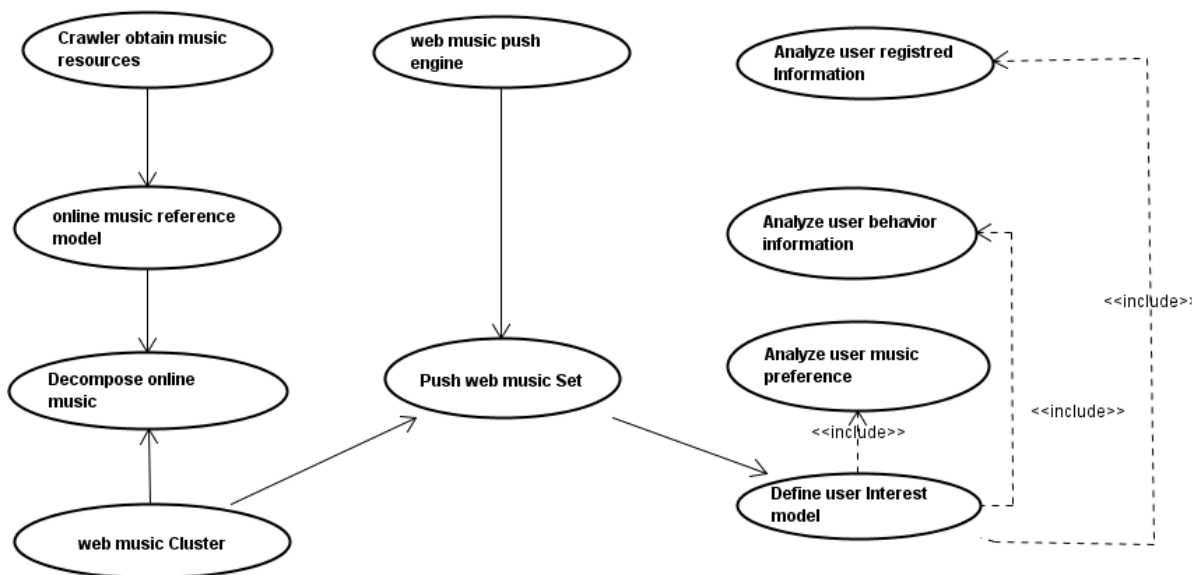


Fig.2. Business structure of web music platform

6. Conclusions

Based on current web resources aggregation technologies, the research raised a MSAM Model. The model distinguishes the duty of web resources aggregation system from five-layer angles. MSAM Model could be used as a solution to help select, analyze and classify the contents of massive music resources on the Internet. Finally it provides users with more useful and more targeted web resources. With MSAM Model, web resources aggregation is not single resource accumulation, but a good realization of in-depth mining and analysis, and user services of massive resources.

Reference

[1] Fabian Abel, Ivana Marenzi, et al. Sharing distributed resources in learn web2.0. EC- TEL 2009, LNCS 5794: 154- 159.
 [2] Ning Xiaomin, Jin Hai, Wu Hao. RSS: A framework enabling ranked search on the semantic web. Information Processing & Management, 2008, 44 (2):893 - 909.

[3] Stanford Center for Biomedical Informatics Research. protege owl api programmer's guide[OL]. [2008-12-23]. <http://protege.stanford.edu/plugins/owl/api/guide.html>.
 [4] Zhang Chengzhi, Su Xinning. Modelling and method of information resources integration. New Technology of Library and Information Service, 2005, (10).
 [5] Gan Wen-yan, Li De-yi, and Wang Jian-min. A hierarchical clustering method based on data fields. Acta Electronica Sinica, 2006, 34(2): 258-262.
 [6] Zhang Ming-wei, Liu Ying, Zhang Bin, et al. Concept based data clustering model. Journal of Software, 2009, 20(9):2387-2396.
 [7] Alicja, Piotr. Quality Assessment of k-NN Multi-label Classification for Music Data. Lecture Notes in Computer Science. 2006, 42(03): 389-398.
 [8] Chang Liaoyu, Yu Xiaoqing, Wan Wanggen, Yao Jincao. Research on Fast Music Classification Based on SVM in Com pressed Domain. International Conference on Audio, Language and Image Processing, 2008, 638-642.
 [9] Owyang J. The future of the social web: In five eras. [2010- 02- 27]. <http://www.web-strategist.com/blog/2009/04/27/future-of-the-social-web/>.
 [10] Sigvat Eide, Terje Kristensen, et al. A model for dynamic content based e-learning systems. Proceedings of the 2008 Euro-American Conference on Telematics and Information Systems, 2008: 2.