A Three Layer System Architecture for Web-Based Unstructured Data Management

Zhiyong Peng #1, Hui Wang *2, Yuwei Peng #3, Bo Xu #, Zeqian Huang *

# Computer School of Wuhan University
Wuhan, Hubei, China 430072
1peng@whu.edu.cn

* State Key Lab. of Software Engineering
Wuhan, Hubei, China 430072
2wanghui_1022@mail.whu.edu.cn

Abstract—With the rapid growing of data on Web, we are facing three serious problems. Firstly, there are a huge number of data resources which are heterogeneous and dynamic. Secondly, most of data on Web are unstructured. Thirdly, there are various kinds of Web users who have different interests and requirements. In this paper, we proposed a new system architecture for unstructured data management on Web to solve these problems by integrating data spaces, database and meta search engine. The system architecture consists of three layers for data gathering on demand, dynamic management and personalized service, respectively. Data servicing layer allows Web users to create data spaces with advanced functions to manipulate and access Web data, eg, cross media query and automatic recommendation. Data managing layer models both Web data and their semantic relationships using our object deputy database named as TOTEM; it also supports schema evolution and dynamic classification. Data gathering layer extracts user’s interest from his or her data space; gathers the related data on Web and further analyzes their semantic relationships. Finally, we implemented a prototype system Tmusic based on this new system architecture to show its availability.

I. INTRODUCTION

Due to hardware progress and personal computer popularization, the INTERNET are developing rapidly. In the age of Web 2.0, a huge volume of data on Web has been generated, most of which are unstructured with self-contained content items, eg, image, audio, video, etc. Web with the mass and unstructured data becomes an important medium for people collecting and acquiring information in their daily life.

However, there are a huge number of data resources which are heterogeneous and dynamic. It is very hard for Web users to find their favorite data. In addition, most of data on Web are unstructured so that they are difficult to be managed by the traditional databases directly. Finally, since Web users have different interests and requirements, personalized services are necessary.

In order to solve these serious problems, we proposed a new system architecture for unstructured data management on Web, which consists of three layers for data gathering on demand, dynamic management and personalized services, respectively. That is, the meta search engine is employed to look for data according to Web users’ requirements. The data are analyzed so that their structures and semantic relationships can be extracted. The unstructured data with rich semantic relationships are managed by our object deputy database named as TOTEM, which can provide dynamic classification functions. Finally, Web users are allowed to create data spaces which are realized based on the flexible object views of TOTEM. Furthermore, the cross media query can be supported due to rich semantic relationships.

We have implemented a prototype system Tmusic to show the availability of this new system architecture. In Tmusic, Web users are allowed to create their own music spaces which are used to collect and manage their favorite music data. According to Web users’ interests extracted from their music space definitions, the meta search engine is employed to look for the related music data which are analyzed and then stored in TOTEM, where the music data are classified automatically so that they can be pushed into where they are needed. In their
personalized music spaces, Web users can enjoy music data in their favorite ways.

The remainder of this paper is organized as follows. Section 2 elaborates the new system architecture. Section 3 describes advanced functions of three layers in detail. Section 4 shows challenges and conclusions.

II. THREE LAYER SYSTEM ARCHITECTURE

To design an efficient Web-based unstructured data management system, it not only needs to manage unstructured Web data, but also needs to uniformly manage the data sources, data storage and data services. Therefore, the questions, how to gather Web data, how to uniformly manage and maintain Web data, and how to provide users with convenience Web data services, should be considered as a whole. Traditional methods well solved parts of these problems, and their drawbacks are obviously shown in uniformly Web data management.

- Web search engines are well used to help Web users to find their favorite data which are ranked according to how the data is related to the user’s requirement. Nevertheless, the results are usually very rough with a lot of noisy data. Web users have to choose what they really need. Therefore, although the results come from a large range of data sources, they can not match users’ needs exactly.
- Web databases are usually used to organize data for some specific purposes. They provide exact queries based on database functions. Web users can find the exact data through interfaces provided by Web browsers. However, Web databases only contain the data well organized for some specific purpose.
- Web data spaces are recently proposed to help Web users to collect and organize their favorite data. They have not strict data schema and can contain various types of data. However, they may cause a lot of data redundancy since each user has his or her own data copy.

It is better to integrate these three technologies together so that we proposed a new system architecture for Web-based unstructured data management which consists of three layers as shown in Fig. 1.

At the first layer, the meta search engine technology will be developed to invoke various search engines so that data can come from a large range of data resources. The heterogeneous data coming from different Web sources should be integrated and analyzed so that their semantic relationships can be found and extracted.

At the second layer, the unstructured data with rich semantic relationships will be stored and managed by our object deputy database TOTEM. In TOTEM, Web data will be defined by basic classes. Deputy classes will be created to maintain the relationships among different types of Web data. In addition, the data from Web can be sent to where they are needed by the dynamic classification function of TOTEM. Its schema evolution allows TOTEM to contain Web data of which schema are frequently changed.

At the third layer, data spaces will be realized by the flexible object views of TOTEM. They can avoid data redundancy and allow Web users to organize data in their favorite ways. Cross media queries in data spaces can be supported by path expression based on the rich semantic relationships. Semantical related multi types of Web data can be retrieved in a single query.

The main difference between our three layer system architecture and other existing system can be summarized in three aspects. Firstly, our architecture considers data gathering, management and servicing as a whole, and it can be easily realized in most Web applications. Secondly, our object deputy data base, which has not only flexibility of relational databases but also complex semantic modeling capability of object oriented database, can effectively manage Web data and their semantic relationships. Thirdly, cross media queries based on Web data semantic relationships can be provided.

III. UNSTRUCTURED DATA MANAGEMENT IMPLEMENTATION

Using this new system architecture, we developed a Web-based music data management system named as TMUSIC. TMUSIC allows Web users to create their own music spaces of which definitions imply their interests. According to Web users’ requirements, various search engines such as Baidu, Soso and etc. are employed to search for their favorite music, lyrics, MTVs and etc, which are typical unstructured data. The music data are analyzed and stored into TOTEM which can represent their rich semantic relationships and automatically recommend them into music spaces where they are needed by its dynamic classification function. Based on the rich semantic relationships, various cross media queries can be realized. The system consists of three layers for unstructured data gathering on demand, dynamic management and personalized service, respectively.

A. Gathering On Demand

There are a huge number of data resources which are heterogeneous and dynamic. Various search engines have been developed to search for Web data by keywords. A single search engine is usually hard to meet all of Web users’ requirement. We employ the meta search engine to invoke different search engines for the music data gathering on demand.

As we know, the structured query in Web databases and keyword query in search engines can’t provide us with the uniform information. In TMUSIC, a retrieval program is developed for meta search engine to search data in both Web databases and search engines in an integrated way. Web
data retrieved by this program consists of meta data used to describe data related to the media and the media data themselves. The meta data for music will be found in Web databases, such as 9sky.com. The media data will be gathered by search engines. For example, music can be obtained from music.baidu.com and music.soso.com; pictures can be found in image.baidu.com; and MTV can be fetched from youku.com. Web users’ interest can be extracted from their music spaces definition in the form of attribute/value and further clustered to generate users’ common interests with low redundancy. According to Web users’ common interests, the query forms will be constructed by the retrieval program and submitted automatically to acquire query results.

The retrieved result pages will be analyzed since there are rich semantic relationships among media data. For instance, in a page for introduction of an album, the retrieved image may be the album’s picture. In TMUSIC, the semantic analysis function is implemented together with the media data extraction. The extracted media data will be filtered out according to Web users’ common interests.

B. Dynamic Management

There are rich semantic relationships among various media data. For instance, a singer’s song can be related to its corresponding MTV because they share the same melody. The unstructured media data with rich semantic relationships will be stored and managed by a database so that they can be well organized for exact queries by Web users. The traditional relational database is very flexible but it is hard to represent complex semantic relationships. The object oriented database can overcome this difficulty but it is not flexible enough for dynamic characteristic of unstructured media data. The object deputy model proposed by us[1], [2] has not only flexibility of relational data model but also complex semantic modeling capability of the object oriented data model. Based on the object deputy model, we developed an object deputy database system named as TOTEM which can provide flexible object views, dynamic classification and cross class queries. These advanced functions are very suitable for dynamic management of the unstructured media data.

We use the object deputy model to define the music data. In this model, each entity of the real world is represented by an object. Deputy object can selectively inherit source object’s attributes or methods, the additional attributes or methods can be defined on demand as well. Switching operations are defined on an inherited attributes or methods, any access to them will be switched to the corresponding attributes or methods in the source object. Schema of a deputy object is defined by a deputy class. The object deputy algebra is used to derive four kinds of deputy classes, which are selection, union, join and group deputy classes and are used to represent the rich semantic relationships between objects and deputy objects, such as specialization, generalization, aggregation and grouping. There is a bilateral link between a deputy object and its source object, which is utilized for update propagation to support dynamic classification. TOTEM has been successfully used in many Web applications, eg. geographic information system[3] and biological information management[4].

The schema hierarchy in TOTEM for the Tmusic is defined as shown in 2. The media data include music, lyrics, pictures of albums, and MTVs of which schema are defined as classes: Music, Lyrics, Albums, MTV. Various deputy classes are defined to represent their semantic relationships. For example, the join deputy class Music_Lyr is defined for aggregation of Music and Lyrics, which mean songs and their lyrics are linked together. The select deputy class Avril_Lyr is specialization of Music_Lyr, which shows Avril’s songs can be managed separately. The union deputy class A BJ_Lyr is the generalization of BonJovi_Lyr and Avril_Lyr, which we can put BonJovi and Avril’s songs together to make a new music space. The grouping deputy class Album is used to group Music_Lyr, which can help people to search music by album.

C. Personalized Service

Based on our object deputy database, we can provide Web users with advanced personalized services, which are data spaces[5], cross media query and automatic recommendation.

1) Data Spaces: The data space is a data set consists of various types of media data along with their semantic relationships. Our system allows Web users to create their data spaces by defining deputy classes of which instances are deputy objects of their favorite media data. For example, when a Web user is interested in both Bon Jovi and Avril’s music, he or she can define a union deputy class in the following way.

CREATE UNIONDEPUTYCLASS A BJ_Lyr(comment text, rank text)

AS (SELECT * FROM BonJovi_Lyr
UNION SELECT * FROM Avril_Lyr);

If the user also prefers Avril’s albums, he or she can define the following select deputy class Avril_Alb in his or her data space.

CREATE SELECTDEPUTYCLASS Avril_Lyr(comment text, rank text)

AS (SELECT * FROM Album WHERE artist = ‘Avril’)  
In the data space, deputy objects inherit attributes of media data and can have additional attributes such as comment and rank which are used to store the user’s comments and ranks of the music. The user’s interests are actually defined as the semantic constraints of the deputy class. The Web user can classify their music in his or her data space by defining the deputy class of the existing deputy class.

2) Cross Media Query: As shown in 2, there are a lot of semantic relationships among media data in TOTEM which can be used to realize the cross media query. This query is a unique feature of TMUSIC. As a result, the related media data can be found and used together in a cross media way. For example, when a user is listening to the song ‘Girlfriend’ which is an instance of the deputy class Avril_Lyr; the related media data, such as the lyric, album picture and corresponding MTV will be very helpful for enjoying the music.
these media data are stored as instances of the different basic classes, cross class queries are necessary for finding the related media data. TOTEM provides a bidirectional path expression which can be used to define cross media queries. In order to obtain the corresponding MTV of the song ‘Girlfriend’, the cross media query can be defined as follows.

```
SELECT (Avril_Lyr → Music_Lyr → Music → Music_MTV → MTV).mtv FROM Avril_Lyr WHERE title = 'Girlfriend' AND artist = 'Avril';
```

The bidirectional path expression “(Avril_Lyr → Music_Lyr → Music → Music_MTV → MTV).mtv” is defined according to the object deputy database schema. When the cross media query is executed, the bidirectional path expression will guide the navigation which starts from the deputy class Avril_Lyr to its source class Music_Lyr and then to the base class Music. Since Music and MTV are joined together by the deputy class Music_MTV, the MTV of Avril’s music “Girlfriend” can be obtained by following bilateral links from Music to Music_MTV and then from Music_MTV to MTV. The related album picture and lyric can be obtained through the similar cross media queries. As a result, the cross media query example is shown in 3.

3) **Automatic Recommendation:** The meta search engine is employed to search for the media data on Web. When the new media data are inserted into TOTEM, the update propagation will be triggered for dynamic classification so that their deputy objects will be automatically generated and inserted into the deputy classes through the update propagation. Thus, the music ‘Girlfriend’ will appear automatically in the data spaces of some Web users who are interested in the music. In this way, TMUSIC can recommend various new media including music, lyrics, pictures of albums and MTVs to the users’ data spaces automatically according to their interests.

**IV. CHALLENGES AND CONCLUSIONS**

In this paper, we proposed a three layer architecture and implemented a Web-based music system Tmusic to prove its availability. However, there are still the following challenges.

**A. Data quality**

The meta search engine in the data gathering layer only retrieves Web data according to the user’s requirement. It lacks the ability to judge the data quality. Data quality problems in Web data management have many different situations. For instance, different data sources may have different data granularity. Errors or mistakes may happen in the attributes or values of Web data. It is necessary to gather a high accurate Web data.

**B. Consistency between Web data and users’ requirements**

Because the storage of a database is limited, it can only store the data which satisfies Web users’ requirements. However, both Web data sources and users’ requirements are changed frequently. It is vital important to maintain the consistency of data in the database with the users’ requirement and keep data in the database to be the update of Web data resources.

**C. Information sharing among data spaces**

Data spaces support Web users to manipulate Web data in their personalized way. However, it still has not a better method to share information among data spaces so that Web users can collaborate with each other. Therefore, how to share Web data among data spaces becomes an urgent and interesting issue.

**ACKNOWLEDGMENT**

This research is supported by National Basic Research Program of China (2007CB310806) and the Major Research plan of the National Natural Science Foundation of China (90718027).

**REFERENCES**


