An Architecture for Unstructured Data Management

Yaohu Lin
School of Economics and Management
Beihang University
Beijing, P.R.China
linyaohu@buaa.edu.cn

Xuelian Lin
The Institute of Advanced Computing Technology
Beihang University
Beijing, P.R.China
linxl@buaa.edu.cn

***Abstract*—As the information age is coming, there is a vast amount of information available in the Internet. Most of data on Web are unstructured. But the significant data should be organized and stored in a suitable way for future purposes. One of the unsolved problems is the management of unstructured data. The unstructured data such as presentation, spreadsheet, text document, memo, images and web pages are difficult to manage while the data become a large scale and the users have different requirements and interests. In this paper, we proposed an architecture for unstructured data management by integrating source query, data collection and data management to solve these problems. The data collection layer extracts the data we care about, we use the existing tools to extract automatic and we can also add the data to the repository manually. The data management layer manage all the collection data by classifying the data, selecting nodes to store and managing centralized as index. The source query layer allows users to query and get the data diversity according the adaptive query service and recommendation service. Finally, we implemented a prototype system OCourse based on this system architecture to show it's feasible and efficient.**

Keywords-unstructured data; classification; storage

I. INTRODUCTION

The amount of information increases at a terrific rate due to hardware progress. A huge volume of data on Web has been generated with the personal computer popularization According to a prediction of IDC, volume of digital content grows to 2.7ZB (1ZB = 1 billion terabytes) in 2012, up 48% from 2011, rocketing toward 8ZB by 2015 [1]. According to Gartner Group statistics, 80% of today's data is unstructured data [2] which are from rich sources.

Internet with huge amount of data becomes an important means for people querying and getting information in their daily life. However, there are a huge number of data resources which are dynamic and heterogeneous. Web search engines are well used to help users to find their favorite data. But the results are often very rough with a lot of noisy data and the search engines does not validate the link. Therefore, it is very hard for Web users to find their favorite data. Traditional relational database for these complex types of unstructured data has been powerless.

In order to solve these serious problems, we proposed a system architecture for unstructured data management which consists of source query, data collection and data management. The unstructured data with rich semantic relationships are managed by an unstructured database name UDR (Unstructured Data Repository) which can provide dynamic classification functions, distributed file indexing and retrieval. Finally, Web users are allowed to retrieval the source diversity and a suitable node will be selected automatic according to the load of server to provide download service when they want to get the source files.

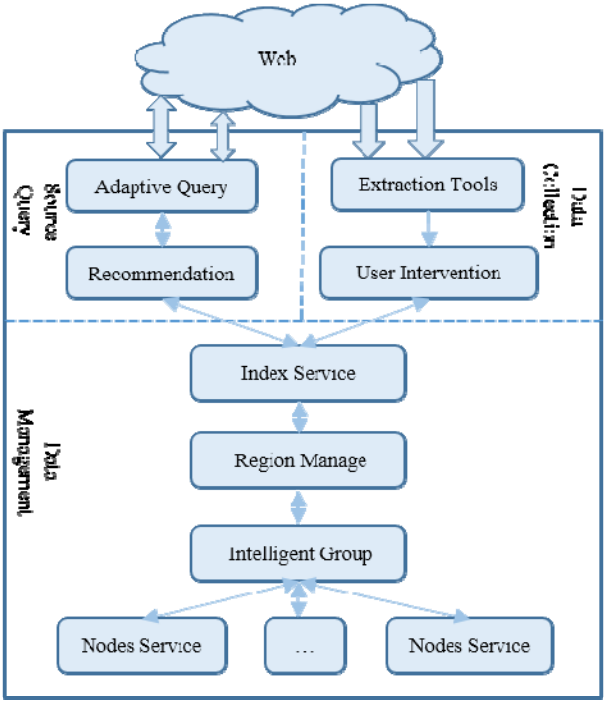

Figure 1. Unstructured data management architecture

We have implemented a prototype system OCourse to show the availability of this system architecture. In OCourse, Web users can retrieve the course ware and download all the resources which may contains presentation, text document, spreadsheet, video, audio, image, etc. they care about. The data collection tools will extract the course ware to UDR automatic and we can also add the resources to the repository manually. The UDR will select suitable nodes based on a node detection mechanism to store the resources, take a course ware contains the introduction of instructors, the description of course, have some video files and presentation files for example, the meta data of course will be saved as an

Published by Atlantis Press, Paris, France.

index and the video files will be stored in a high-performance node while the presentation files may be stored in another node.

The paper is organized a follows: Section II describes the related work on all possible data management and Section III gives details explanation on research methodology. Section IV demonstrates the results on this research and Section V shows challenges and conclusions.

II. Related Works

A great amount of information is created since the World Wide Web is a growing database. There are three types of web pages which are structured, semi-structured and unstructured. Structured data can be processed automatically by machines. Semi-structured is typically can be handled by one or more patterns. Unstructured data are those in which the information is within simple text that no common pattern can be used to process. Unfortunately, most of web pages are unstructured and querying or accessing these data is not a simple task since it is represented in a human friendly format. There are many algorithms and solutions presented try to solve this problem, but unfortunately none of them can be considered as a perfect one.

A. Extraction and Classification

Data extraction is a process of retrieving and capturing data from one medium into another medium. The medium can be Web pages, databases, repository, documents or anything that consists of information. Many data extraction tools have been study and implement. RoadRunner [3] generated the HTML tag either semi-automatically or automatically. SRV [4] use Natural language processing or NLP techniques to build relationship between sentences elements and phrases. NoDoSE [5] and DEByE [8] are modeling based tools which a target structure is provided according to a set of modeling primitives that conform to an underlying data model. Brigham Young University Data Extraction Group [6] developed an ontology-based tool by relying directly on the data.

Data classification is to categorize data based or required needs. There are various classifications of wrapper. For example, Hsu and Dung [7] classified wrappers into four distinct categories, including hand-crafted wrappers that had heuristic-based and induction approaches. Laender [8] proposed the taxonomy such as languages for wrapper development that consists of HTML-aware tools, NLP-based tools, wrapper induction tools, modeling-based tools, and ontology-based tools.

B. Storage

Traditional relational database is very flexible but it is hard to represent unstructured data which has heterogeneous and complex structures. Multidatabase Systems [9] (MDBS) or Federated Database Systems [10] focus on the uniform interface of multiple database joint together and decrease the impact on operation of the existing database from integration perspective. MDBS provides users a unify data access environment from Local Databases (LDB) including database and file system. MDBS integrates the data pattern of heterogeneous database and forms a unify data pattern. Base on this pattern, users' queries are converted to the sub query of each sub database, the results are sent back to users after merging. Google Bigtable is a representative big data processing storage system. Bigtable uses many implementation strategy of database, but it does not support complete relational data model. It provides simple data models which users can control the data distribution and format. Bigtable treat all storage data as string without parsing. Finally, user can determine to put the data in the memory or hard disk by pattern parameters of Bigtable [11] . CouchDB is a document oriented database. User can use JavaScript to index and query as MapReduce style. CouchDB provides delta copy, incremental replication with bi-directional conflict detection and management. It can be accessed from any environment which allows HTTP request [12] .

C. Query

SQL (Structured Query Language) is a special-purpose programming language designed for managing data in relational database management systems. AQP（Adaptive Query Processing) [13] is a query engine on the basis of adjusting query plan according to the query amount and transmission speed. It can give query result step by step on occasion and adjust the incomplete results in time.

In [14], the authors used a content-based Mesh query engine to abstract the query as the concept of drift problem. This query can ignore some query condition in early phrase. Document [15] and [17] map relations into streams using a sequential data-stream-protocol-supported query language called CQL which can find the data-stream-related set just through an index. Document [16] finds the optimal result through the comparisons in the three series which it generates. Document [18] generates the query scheme and the evaluating system using a self-adapting algorithm, as for the massive data in XML, as to match the K values from the head efficiently.

III. System Design

With the development of Open Course Ware (OCW), more and more people from all over the world can get the excellent course ware. Course ware often includes course introductions, presentations, text documents, course videos, images, etc. We developed a course ware management system name as OCourse based on the framework shown in Fig. 1. The design of OCourse follows several principles: Firstly, the system needs to uniformly describe and can be easily realized in most Web applications. Secondly, data repository should manage Web data and their semantic relationships effectively, not only has flexibility of relational databases but also complex semantic modeling capability of object oriented database. Finally, cross queries based on Web data semantic relationships can be provided.

OCourse extracted course ware by extraction tools automatic and we can also add the course ware to the database manually. The UDR will update the course Meta information, such as course name, instructors, school, etc. At the same time, UDR will choose nodes to store these source.

Based on the Meta data analysis and semantic relationships analysis, various cross queries can be realized. The system consists of three parts for data collection, data management and source query.

A. *Data Collection*

There are a huge number of data resources on Web, extraction and classification become an important part. Various search engines have been developed to search for Web data through keywords. A single search engine is usually hard to meet all of Web users' requirement. And search engine often cannot distinguish links that do not work. We employ the search engine search the university or college website (*.edu.*) to invoke the existing extraction tools XWRAP [19] to gather the data on demand.

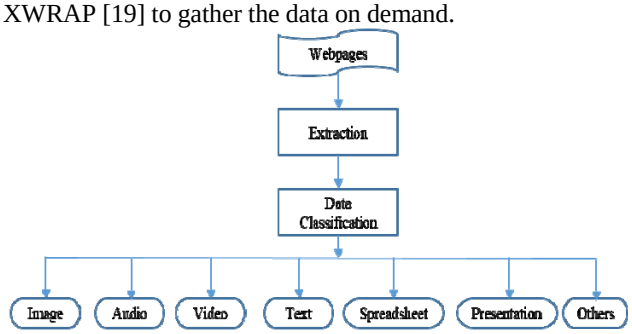

Figure 2. Data Collection

Webpages we concerned about are mainly the educational institution. Classification of data patterns is important for data extraction from website. Fig. 2 illustrates the progress of data collection. We extract the specific target and classify them into image, audio, video, text, spreadsheet, presentation and others.

B. *Data Management*

There are rich semantic relationships among course data. For instance, a course may have the introduction information including instructors, course number, level, etc. The study materials may contains course video, presentation and other electronic courseware.

The traditional relational database is very flexible to deal with many things but it's hard to represent complex semantic relationships. The object oriented database can get over this difficulty but it is not smart enough for dynamic characteristic of unstructured media data. We implement data management by integrating index service, region manage, intelligence group and nodes service together. The index service will serve the courses index information and use region management to master the details of clusters. The intelligence group provides a service to gather the nodes' physical information at regular intervals, and it knows the nodes exactly state. When the data come into UDR, in our architecture, the Meta data about course will be stored into the index server while the courseware files will be sent to the suitable node server to provide the download service. Courseware files were stored in a dynamic nodes according the internal algorithm we provided. All the nodes were divided into two groups: stream group and document group. The video or audio files are often larger than other file type. High-performance nodes are divided into stream group by intelligence group service because Web user may want to download the video or audio files or preview them.

C. *Source Query*

Based on the UDR, we provide Web users with cross queries and automatic recommendation. There are a lot of semantic relationships among media data in UDR which can be used to realize the source query. Various queries are provided by OCourse. One teacher may instruct several courses while a course with the same name may be lectured by several teachers. As a result, the related media data can be found through source query. For example, when a user query 'operating system' which is an instance of the deputy class Course. The Course class contains a set of query results, the related courseware will be found from the instance of deputy class Ware. Because the courseware files may be stored in different nodes, the nodes which files are selected to download should initialized the download service instance. Web user can get all the related resource they want.

IV. RESULTS

The user interface of OCourse is illustrated as Fig. 3 that allows web users to search interested course. User can select the resource types he or she cares about. OCourse give the recommendation according to the download times and the specific resource type.

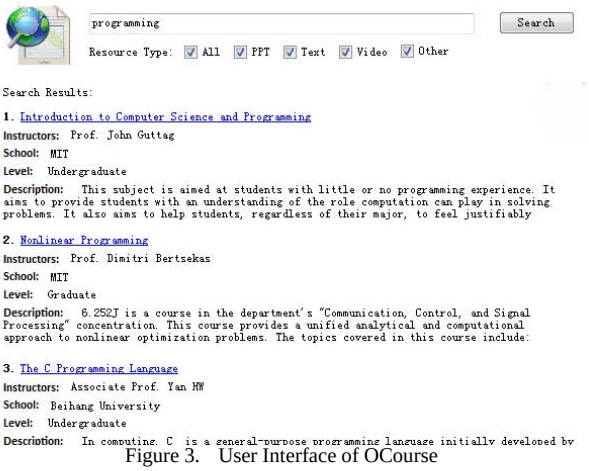

Figure 3. User Interface of OCourse

The Meta data of course are extracted and save into the index server. Fig. 4 shows part of the index information of the course.

idCourse	courseName	instructors	school	level	description	eval	hotSpot
58	Compliers	Wang ting	National ...	1	602 staff room	15	25
59	Compliers	Zhang Li	Beihang ...	1	specialized core course of undergrate...	15	30
60	Introduction to Anthropo...	Prof. James H...	MIT	1	This class introduces students to the ...	20	24
61	Principles of Microecon...	Prof. Jonathan...	MIT	1	14.01 Principles of Microeconomics is...	24	22
62	Introduction to Compute...	Prof. John Gut...	MIT	1	This subject is aimed at students with ...	21	30
63	Introduction to Compute...	Prof. Eric Grim...	MIT	1	This subject is aimed at students with ...	23	22
64	Introduction to Electrical...	Prof. Leslie Ka...	MIT	1	This course provides an integrated int...	14	23
65	Dynamic Programming a...	Prof. Dimitri Be...	MIT	2	This course covers the basic models ...	26	14
66	Nonlinear Programming	Prof. Dimitri Be...	MIT	2	6.252J is a course in the department's...	24	15
67	Introduction to Algorithms	Prof. Erik Dem...	MIT	1	This course provides an introduction t...	11	16

Figure 4. Index information

Based on the information of index server, when user clicks the specific course of the search result, related resources will be listed as Fig. 5. From the download link listed, courseware files can be retrieved.

Resource List

NO.	File Name	File Types	File Size	Download
1	1.avi	avi	6252225.0	Link
2	1-2_Lecture1-2.pdf	pdf	5429984.0	Link
3	10_Maximum_Flow.pdf	pdf	4984310.0	Link
4	11_string_matching.pdf	pdf	3209161.0	Link
5	12_NP-Completeness.pdf	pdf	5166672.0	Link
6	13_Approximation_Algorithms.pdf	pdf	1766679.0	Link
7	1_Lecture1.pdf	pdf	2272331.0	Link
8	3_Sorting Algorithms.pdf	pdf	4438515.0	Link
9	4_Dynamic_Programming.pdf	pdf	2775446.0	Link
10	5_Greedy_Algorithms.pdf	pdf	3047667.0	Link
11	6_Amortized Analysis.pdf	pdf	1829026.0	Link

Figure 5. Resources of specific course

User can get the information of instructors, level, description and related resource files from OCourse. It validates the effectiveness of download link and makes sure the resource can be download normally.

V. CONCLUSIONS AND CHALLENGES

This paper presents an architecture for unstructured data management and implemented a prototype system OCourse to prove it's feasible and efficient. However, there are still the following challenges. Firstly, when the data coming into the repository, the service can select suitable node which has a good performance in all respects to store the file, but due to the environment is dynamic and change all the time, we cannot ensure the specific node has a good performance all the time. Performance problems in distributed system have many different solutions, a suitable copy management can be used in this situation. Secondly, both Web data sources and users' requirements are changed frequently. It is important to maintain the consistency of data in the database. Although new storage nodes can be easily add to the UDR in our architecture, the capacity of the repository is limited comparing to the data in Web.

REFERENCES

[1] IDC, "TOP 10 PREDICTIONS," IDC, pp. 1-26, 2011

[2] Diane Berry, Coveo, "Unstructured data: Challenge or asset," ZDNet, http://www.zdnet.com/news/unstructured-data-challenge-or-asset/6356681, 2012

[3] Valter Crescenzi, Giansalvatore Mecca, Paolo Merialdo, "RoadRunner: Towards Automatic Data Extraction from Large Web Sites," VLDB Conference, pp. 624-624, 2001

[4] Freitag, Dayne, "Information extraction from HTML: Application of a general machine learning approach," AAAI, pp. 517-523, 1998

[5] B. Adelberg, "NoDoSE: A Tool For Semi-Automatically Extracting Structured And Semi-Structured Data From Text Documents," SIGMOD Record, vol. 27(2), pp. 283-294, 1998

[6] T. Chartrand, "Ontology-Based Extraction of Rdf Data From The World Wide Web," Brigham Young University, 2003

[7] Chun-Nan Hsu, Ming-Tzung Dung, "Generatingfinite-statetransducers for semi-structured dataextraction from the Web," Information Systems, vol. 23(8), pp. 521–538, 1998

[8] Alberto H.F. Laender, Berthier Ribeiro-Neto, Altigran S. da Silva, "DEByE – DataExtraction By Example," Data & Knowledge Engineering, vol. 40(2), pp. 121-154, 2002

[9] Dayal, Umeshwar, Hwang, Hai-Yann, "View Definition and Generalization for Database Integration in a Multidatabase System," IEEE Transactions on Software Engineering, vol. 10(6), pp. 628-645, 2009

[10] W.-S. Li, V. S. Batra, V. Raman, W. Han, and I. Narang, "QoS-based data access and placement for federated systems," Proceedings of the 31st international conference on Very large data bases, pp. 1358-1362, 2005

[11] Fay Chang, Jeffrey Dean, Sanjay G., et al, "Bigtable:A Distributed Storage System for Structured Data," TOCS, pp. 1-26, 2008

[12] Apache, "CouchDB," http://couchdb.apache.org/, 2008

[13] Anastasios Gounaris,Norman W.Paton,Rizos Sakellariou,Alvaro A.A.Femandes, "Adaptive Query Processing and the Grid: Opportunitites and Challenges," .Information Sciences, vol. 177 (17) , pp. 3574-3591 , 2007

[14] Rimma V. Nehme, Elke A. Rundensteiner, Elisa Bertino, "Self-tuning query mesh for adaptive multi-route query processing," Extending Database Technology, pp. 803-814, 2009

[15] A. Arasu, S. Babu, and J. Widom, "The CQL continuous query language: Semantic foundations and query execution," VLDB, vol. 15(2), pp. 121-142, 2006

[16] S. Babu and P. Bizarro, "Adaptive query processing in the looking glass," Second Biennial Conference on Innovative Data Systems Research, pp. 238-249, 2005

[17] Sai Wu, Quang Hieu Vu, Jianzhong Li, Kian-lee Tan, "Adaptive Multi-join Query Processing in PDBMS," Data Engineering, pp. 1239-1242, 2009

[18] Marian, A., Amer-Yahia, S., Koudas, N.,Srivastava, D, "Adaptive Processing of Top-k Queries in XML," Data Engineering, pp. 162-173, 2005

[19] Georgia Institute of Technology, "XWRAP," http://www.cc.gatech.edu/projects/disl/XWRAP/, 2000