A Document Authoring System for Credible Enterprise Reporting with Data Analysis from Data Warehouse

Masao Mori
IR Office, Kyushu University, Fukuoka, Japan
mori@ir.kyushu-u.ac.jp

Toshie Tanaka
IR Office, Kyushu University, Fukuoka, Japan
tanaka@ir.kyushu-u.ac.jp

Sachio Hirokawa
Research Institute for Information Technology, Kyushu University, Fukuoka, Japan
hirokawa@cc.kyushu-u.ac.jp

***Abstract*—In rapid progress of information technology, we are facing difficulties, "information explosion". From standpoint of using enormous quantity of data, there are many researches such as information retrieval and clustering information. On the other hand, in terms of creating credible enterprise reports, information explosion also becomes a big problem. If most of digital documents are unstructured, report writers may have significant difficulties with management and arrangement of digital documents. Actually in the case of university evaluations, report writers have been confronted with that difficulties. In addition, quantitative data from data warehouse is indispensable for enterprise reports. In this paper, we developed a document authoring system cooperating with data warehouse to settle these problems from viewpoint of reusing and reconstructing components of reports.**

***Keywords*-digital document; data warehouse; accreditation; knowledge management; web service**

I. INTRODUCTION

In recent years opportunities of enterprise reporting in companies, institutions and universities have been increasing rapidly. So that business intelligence and content management system for enterprise reporting are desirable, for instance, Priebe[1]. What is required to create credible enterprise reports? Morimoto et al.[2] asserts the following four processes of enterprise reporting from viewpoint of knowledge management: (1) collecting and accumulating documents, (2) searching and browsing documents, (3) extracting and identifying documents and (4) creating credible reports. In order to realize these processes completely, information must be structured. DITA[3] is one of the ideal architectures to extract information from documents effectively and to manage documents efficiently. However, not infrequently, non-structured information exceeds structured information, especially on-the-spot of university evaluations.

All Japaneses universities are obliged to be evaluated by certified organization, called *institutional certified evaluation and accreditation*. In addition, all Japaneses national universities must be evaluated for the purpose of information disclosure to government and nation, called *national university corporation evaluation*. They are called *university evaluations* which is undergone every six years. Universities must prepare self-assessment reports for university evaluations. Educational and research activities of university vary in many ways. In Kyushu university, one of national universities in Japan, though documents of committee and faculty council were stored, they had not yet been managed systematically. How to reuse these documents becomes a big problem. Authors of this paper have been supporting faculties and bureaus to create university evaluation reports. As in Figure1, the amount of document in evaluation report of Kyushu university 2009 was so large-scaled that it was hard even to fix formats of documents. In addition, many items and themes appear many times in both reports. So the writer must be thoughtful for consistency of both reports.

Fields	Schools	Contents of report		
		Sections	Viewpoints	Pages
Education	31	8	12	959
Research	20	5	5	311

Figure 1. An example of amounts of documents in the corporation evaluation report of educational and research activity of Kyushu university 2009

From our experience to support creation of evaluation reports, we have developed a document authoring system for enterprise report, especially for university evaluations, cooperating with data warehouse. In order to manage unstructured information efficiently, the proposing system provides users with a simple and uniform data structure for report components. Users can create enterprise reports by arranging report components in the tree structure of sections. Moreover, by reusing report components users can make sure of consistency of enterprise report.

Our system challenges the two targets as follows: (1) management of items and themes which appear frequently in various enterprise reports, (2) light-weight cooperation with data warehouse. This system is developed using Ruby on Rails and MySQL. Demonstration of our system can be seen on Youtube[1].

The paper is structured as follows: In Section 2 we review related works. In Section 3 we overview our system and introduce three main concepts, report components, report

[1] http://www.youtube.com/watch?v=okAT6aseks8

 ISBN: 978-1-61208-104-5

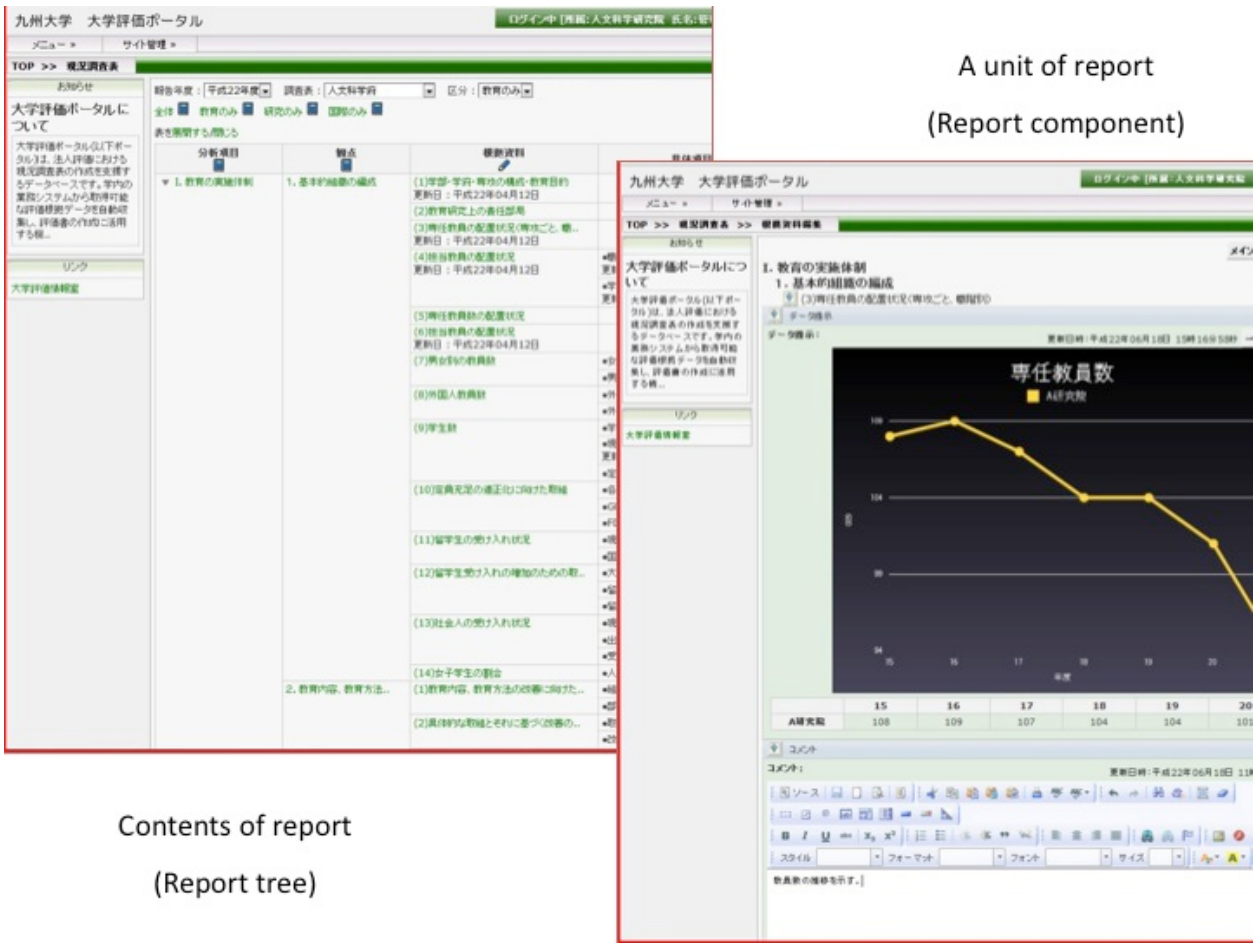

Figure 2. Views of the document authoring system

tree and data analysis queries. In Section 4 we present features of our approach comparing with related work. We conclude the paper with summary and future work.

II. Related work

We start to discuss related work by reviewing the assertion of processes for enterprise report in Morimoto et al.[2]: (1) collecting and accumulating documents, (2) searching and browsing documents, (3) extracting and identifying documents and (4) creating credible reports.

Considering document management, information retrieval is indispensable for accumulated digital documents. Beyer et al.[4] propose a method to discover patterns and rules of texts in structured documents in order to generate efficient search index. Linked Data[5] would be helpful to capture relationship between digital documents if they were structured. But we found that the prime consideration for documents in enterprise reporting, especially in university evaluation, is meta-data of digital documents and materials, such as their jurisdiction, creators and meanings of the documents. As a university is a complex organization consisting of many departments and bureaus with autonomy, meta-data of documents is indispensable for document management in the scene of university evaluations.

DITA[3][6] is a document architecture for extraction and management of documents. DITA enables users to extract and update information efficiently in large amounts of documents[7]. In order to adopt DITA and Linked Data, it is required to define an ontology for knowledge of enterprise. Since it is difficult to apply an ontology to present progressive enterprise processes and legacy systems, we decide to extract text from digital document by hand and to collect minimum concrete information (such as “Section 2 on page 23”) as meta-data about digital documents.

Generally speaking, accumulating daily reports ensures enterprise reports, moreover it is advisable to study how to obtain meanings and attributes of documents[2]. If an enterprise report is required to be prompt, integration of document creation with OLAP is desirable[1]. In the case of university evaluation reports, frequency of reports is much lower than daily reports in companies. Actually evaluation report is usually conducted every year or every month at

 ISBN: 978-1-61208-104-5

most. A long-term vision rather than promptness is necessary for university management. One of important requests in university evaluations is to select documents efficiently and to organize them effectively rather than automatic reporting function. The proposing system provides users with an interactive interface to select documents and organize reports.

Integration of structured data in data warehouse and unstructured data in texts on news sites and blogs has been studied in [1][5][8][9]. Most of them are based on information retrieval and assume that ontology for structured data is given, whereas we assume that ontology is not given but the design of enterprise reports is given, like university evaluations. Our approach is different from those related work in terms of these assumptions.

III. Overview

A. Report Component and Report Tree

In this subsection, we will introduce the document authoring system for enterprise reporting (DASER for short). As we mentioned in the introduction, it is important to provide users with a uniform data structure in order to bundle essential information of materials and documents. A data structure, *report component*, is a unit in DASER, which consists of seven elements as follows:

1) id,
2) title (user input),
3) comment (user input),
4) data analysis query (user input),
5) data analysis,
6) attached documents (user input), and
7) meta-data.

Users may input data into attributes such as comment, data analysis query and attached documents. *Data analysis* is visualization of data obtained from data warehouse through "*data analysis query*" (DAQ for short). DAQ is URL of a CGI program in data warehouse. We will discuss DAQ in the next section. Meta-data is owner information and time-stamp. Each report component have visualizing function for CSV data obtained from DAQ.

The window on the right in Figure 2 is an example of report component. The graph is generated from CSV data which is obtained from data warehouse through DAQ. Note that the visualizing function does not depend on DAQ. One can visualize static CSV files located in other web server.

In DASER, we can define structure of enterprise report by giving a tree structure with report components as leaves and sections as internal nodes. This is called a *report tree*. Report tree is changeable corresponding to contents of every enterprise report, and it also can be changed depending on individual needs from users. Report tree can be construct with report component as leaf nodes, and with the root node and internal nodes. A root node and internal nodes have the same data structure as a report component and additional attributes as follows:

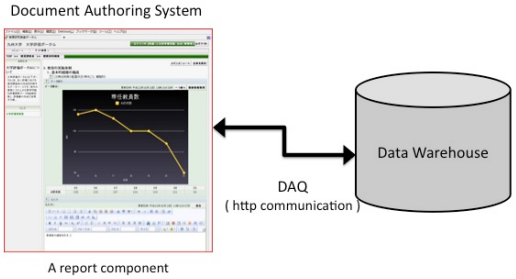

Figure 3. The report authoring system and data warehouse

8) a list of ids of children and
9) its parent's id.

Note that each report component does not depend on the definition of report tree.

B. Data Analysis Query

In this section we introduce data warehouse (DW for short) and its data analysis query. DAQ is WebAPI of DW. Data sources of DW is backup data of operational (business) systems. In the context of university evaluation, for example, they are information about students, teachers, teaching, research and finance of university. Flat files such as spreadsheets, are also data source of DW.

Administrator of DW provides users with programs in order to analyze data in DW. That is called *data analysis query*. As DAQs are implemented as CGI programs, one can access DW with DAQs over restful HTTP communication. DAQs return data in CSV format.

Let us consider the case of analyzing international students enrollment. One need to calculate numbers of students for every year and every department in order to show their changes. For example, the DAQ

```
http://dw.mydom/int_stdt.cgi?yr=5&dpt=eng
```

returns CSV data about changes of international students in the department of engineering (`dpt=eng`) for the past five years (`yr=5`). This DAQ is available for other departments and other year terms by changing parameters.

IV. Features

In the introduction, we mentioned that our challenges are: (1) management of items and themes which appear frequently in various enterprise reports, and (2) light-weight cooperation with data warehouse. In this section we will see achievement to the challenges.

A. Consistency and Credibility

Firstly we discuss how the proposing system contributes to consistency for enterprise reports.

Generally speaking, contents of an enterprise report form a tree structure. Leaf nodes are topics and themes and internal nodes are sections and chapters. So we define a

 ISBN: 978-1-61208-104-5

report component as a leaf node, which is a data structure with seven attributes, and chapters and section as internal nodes. When users create multiple reports in such as our case of two university evaluations, what user have to do is setting each report tree corresponding to a configuration of each report. Then DASER flexibly generates multiple reports. Even if some report components appear many times in different reports, DASER ensures consistency and credibility between different reports. Related work, such as [1][5][8][9], have not focused on the problem of multiple reports. This is one of unique features of our approach.

B. Light-weight cooperation and its effectiveness

DASER is connected to DW only through DAQs by restful http communication which is one of web service techniques. We could successfully develop DASER and DW separately. In other words, DW can offer the CSV data to other service besides DASER, and DASER can refer to static CSV files from other data source besides DW.

Sharing data warehouse inside of intranet has been a trend for a decade [10]. Our approach is to develop an integration of qualitative data and quantitative data for enterprise reporting, whereas we must develop data warehouse for not only reporting but also sharing information inside of our university. This situation is different from [8][1].

C. Flood of unstructured and valuable XML data

In order to accomplish information disclosure, enterprise documents are always accumulated. This issue is for not only big organizations such as big universities, for but also any small organizations such as elementary schools.

Unfortunately, in many universities and schools in Japan, most of their digital documents, like word processor files and spread sheets, are unstructured data. That is why we must assume nonexistence of ontology for our approach. When user creates an enterprise report on our system, she/he is supposed to set up report components and report trees. Giving report components and report trees would lead to the ontology for the enterprise report. That is one of unique feature of our system.

V. Conclusion and future work

In this paper we developed a document authoring system for enterprise reporting cooperating with data warehouse. And we realized a light-weight cooperation between our system by using the technique of restful http communication.

Two problems still remain. First problem is flexibility of report component. Under current configuration of DASER, user cannot variously set the contents of report component to the context of each enterprise reporting. Second problem is flexibility of composing results of DAQ. Cross tabulation of two or more results of DAQs is impossible. From our researches like [11][12], it is considerable to apply the method of web mash-ups to the second problem.

References

[1] T. Priebe and G. Pernul, “Ontology-based integration of olap and information retrieval,” in *Proceedings of the 14th International Workshop on Database and Expert Systems Applications*, 2003, p. 610.

[2] Y. Morimoto, H. Mase, and H. Tsuji, “Perspectives on reuse process support systems for document-type knowledge,” in *Human-Computer Interaction, Part IV, HCII 2007*, ser. Lecture Note in Computer Science 4553, 2007, pp. 682–691.

[3] O. Standard, *DITA Version 1.1 Architectural Specification*, OASIS, 2007.

[4] K. Beyer, V. Ercegovac, and R. K. et al., “Towards a scalable enterprise content analytics platform,” *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 2009.

[5] C. Bizer, T. Health, K. Idehen, and T. Berners-Lee, “Linked data on the web (ldow2008),” in *Proceeding of the 17th international conference on World Wide Web*, 2008, pp. 1265–1266.

[6] M. Priestley, “Dita xml: A reuse by reference architecture for technical documentation,” in *Proceedings of SIGDOC’01*, October 2001, pp. 152–156, santa Fe, New Mexico, USA.

[7] O. Dìaz, F. I. Anfurrutia, and J. Kortabitarte, “Using dita for documenting software product lines,” in *Proceedings of the 9th ACM symposium on Document engineering*. Association for Computing Machinery, 2009, pp. 231–240.

[8] A. Ferrández and J. Peral, “The benefits of the interaction between data warehouses and question answering,” in *Proceedings of the 2010 EDBT/ICDT Workshops*, ser. ACM International Conference Proceeding Series, vol. 426, 2010.

[9] B. Riger, A. Kleber, and E. von Maur, “Metadatabased integration of qualitative and quantitative information resources approaching knowledge management,” in *ECIS 2000 Proceedings*, 2000.

[10] R. Kimball and R. Merz, *The Data Webhouse Toolkit: Building the Web-Enabled Data Warehouse*. Wiley, 2000.

[11] M. Mori, T. Nakatoh, and S. Hirokawa, “Links anc cycles of web databases,” in *The 4th Italian Workshop on Semantic Web Applications and Perspectives*, 2007, pp. 21–30.

[12] ——, “Functional composition of web databases,” in *Proceedings of International Conference Asian Digital Libraries 2006*, ser. Lecture Note in Computer Science 4312. Springer Verlag, 2006, pp. 439–448.

[13] S. Mushhad, M. Gilani, J. Ahmed, and M. A. Abbas, “Electronic document management: A paperless university model,” in *Proceedings of 2009 2nd IEEE International Conference on Computer Science and Information Technology*, 2009, pp. 440–444.

 ISBN: 978-1-61208-104-5