December 2004

The Concept of Document Warehousing and Its Applications on Managing Enterprise Business Intelligence

Frank Tseng
National Kaohsiung First Univ. of Sci. & Tech.

Annie Chou
ROC Military Academy

The Concept of Document Warehousing and Its Applications on Managing Enterprise Business Intelligence*

Frank S.C. Tseng†
Dept. of Information Management
National Kaohsiung First Univ. of Sci. & Tech.
Kaohsiung, Taiwan, 811 ROC
imfrank@ccms.nkfust.edu.tw

Annie Y.H. Chou
Dept. of Computer and Information Science
ROC Military Academy
Kaohsiung, Taiwan, 830 ROC
yhchou@cc.cma.edu.tw

Abstract

During the past decade, data warehousing has been widely adopted in the business community. It provides multi-dimensional analyses on cumulated historical business data for helping contemporary administrative decision-makings. Nevertheless, it is believed there is only about 20% information can be extracted from data warehouses concerning numeric data only, the other 80% information is hidden in non-numeric data or even in documents. Therefore, many researchers now advocate it is time to conduct research works on document warehousing to capture complete business intelligence. Document warehouses, unlike traditional document management systems, include extensive semantic information about documents, cross-document feature relations, and document grouping or clustering to provide a more accurate and more efficient access to text-oriented business intelligence. In this paper, we discuss the basic concept of document warehousing and present its formal definitions. Then, we elaborate some useful applications to illustrate the importance of document warehousing. The work is essential for establishing an infrastructure to help combining text processing with numeric OLAP processing technologies. The combination of data warehousing and document warehousing will be one of the most important kernels of knowledge management and customer relationship management applications.

Keywords: Data Warehousing, Document Warehousing, Knowledge Management, OLAP.

1. Introduction

Data warehousing (Inmon 1993) and data mining techniques (Han and Kamber 2001) are gaining in popularity as organizations realize the benefits of being able to perform multi-dimensional analyses of cumulated historical business data to help contemporary administrative decision-making (Anahory and Murray 1997; Berry and Linoff 1997; Han and Kamber 2001; Kimball 1996). That inspires enterprises to eagerly delve useful business intelligence (BI) from both internal and external data. Business intelligence is supposed to provide decision makers with the tactical and strategic information they need for understanding, managing, and coordinating the operations and processes in organizations.

However, much of the efforts have only touched the tip of the information iceberg. While the techniques regarding data warehouses, multi-dimensional models, on-line analytical processing (OLAP), or even ad hoc reports have served enterprises well; they do not

* This research was partially supported by National Science Council, TAIWAN, ROC, under Contract No. NSC-91-2416-H-327-005.

† To whom all correspondence should be sent.

completely address the full scope of business intelligence. It is believed that (http://www.survey.com), for the business intelligence of an enterprise, there are only about 20% information can be extracted from formatted data stored in relational databases. The remaining 80% information is hidden in unstructured or semi-structured documents. This is because the most prevalent medium for expressing information and knowledge is text. For instances, market survey reports, project status reports, meeting records, customer complaints, e-mails, patent application sheets, and advertisements of competitors are all recorded in documents.

To alleviate such phenomenon, (Grigsby 2001; McCabe *et al.* 2000; Sullivan 2001) have advocated that documents should be properly *warehoused* according to some well-defined concepts for expanding the scope of business intelligence to include textual information. Hence, we think one of the next challenges in information community will be the study of topics about document warehousing and text mining to help enterprises on obtaining the complete business intelligence. Although research works regarding text mining have been conducted widely, for example, the gentle readers are referred to (Knight 1999; Lin et al. 1998; Loh et a. 2000; Tan 1999), however, the issues regarding document warehouse are rarely addressed.

Since there are usually many diverse concepts involved in a document, a document is also multi-dimensional in nature. Document warehouses, unlike traditional document management systems, include extensive semantic information about documents, cross-document feature relations, and document grouping or clustering to provide a more accurate and more efficient access to text-oriented business intelligence. To facilitate flexible and effective multi-dimensional on-line analytical document processing and browsing, a multi-dimensional query language for querying document warehouses is indispensable.

Although issues about document warehousing have been addressed in (Grigsby 2001; McCabe 2000; Sullivan 2001), there are still no formal definitions established up to now. In this work, we will first discuss the concept of document warehousing and formally define the related terms. Then, we elaborate some applications of document warehousing to sketch an attractive roadmap of using document warehouses.

As the Web applications proliferate tremendously, there will be a great deal of needs on rapid text processing and browsing. Document warehousing does not only provide an infrastructure for developing tools for business executives to systematically organize, understand, and properly categorize their documents to help strategic decision-makings, but also integrate all kinds of related documents being browsed instantly.

Document warehousing also provides an important platform for on-line analytical processing (OLAP) in text level for the interactive analysis of multi-dimensional documents of various granularities, which facilitates effective text mining, integrates documents into the business intelligence infrastructure, and provides the means to search for and target specific information the way we now do with numeric data. Furthermore, as the construction of data warehouses can be viewed as an important step for data mining, the construction of document warehouses can be regarded as an indispensable preprocessing step for text mining.

In this paper, we illustrate the general architecture of a document warehouse and its applications. The work is essential for establishing an infrastructure to help combining text processing with numeric OLAP processing technologies. Hopefully, the combination of data warehousing and document warehousing will be one of the most important kernels of knowledge management and customer relationship management applications.

We believe such infrastructure can help to extend numeric data analysis to be combined with

text processing technologies to make data warehousing and document warehousing being one of the most important kernel of knowledge management and customer relationship management applications. By combining document warehousing and data warehousing, documents can be integrated into the business intelligence infrastructure and provide the means to search for and target specific information the way we now do with numeric data.

Our paper is organized as follows. In Section 2, the important concepts of document warehousing are formally presented. Then, based on these definitions, we will propose a general architecture for constructing document warehouses in Section 3. Section 4 devotes to the dimensional modeling of document warehouses. Then, an application of document warehouses will be discussed in Section 5. Finally, we conclude and propose some future works in Section 6.

2. An Introduction to Document Warehousing

In the following, we give some definitions about *document*, *dimension*, *document tuple*, and *document cube* for document warehousing.

Definition 1: A *document* $T = \{t_1, t_2, \ldots, t_i\}$ is a logical unit of text characterized by a set of keywords $\{t_1, t_2, \ldots, t_i\}$.

To organize documents into structures, we need the concept of *dimension* defined as follows.

Definition 2: A *dimension D* is a tree structure of m levels, $m \geq 1$, which is used for representing the hierarchical relationships among a set of keywords. A node in a dimension D is called a *member*, and each internal node contains a special child called *summary member*, denoted '*', which is used for denoting the total concept of the other children of the internal node.

When drawing a dimension, we usually leave out a summary member, since it has the same meaning with its parent node. Besides, the keywords in a dimension are not limited to only those contained in document contents. Any property or metadata of a document file (e.g., those defined in Dublin Core Metadata Element Set (http://dublincore.org)) can also be regarded as a keyword in a dimension for constructing document cubes. Furthermore, if documents are organized into predefined categories, the category hierarchy to which a document belongs can also be regarded as a dimension.

According to the keyword sources, dimensions can be distinguished into the following types:

1. *Ordinary dimension*. A dimension contains keywords used for scanning the document contents.

2. *Metadata dimension*. A dimension contains keywords used for scanning document file properties or metadata. For example, in Dublin Core Metadata Element Set, there are title, creator, subject, description, publisher, contributor, date, type, format, identifier, source, language, relation, coverage, and rights, all can be regarded as metadata dimensions.

3. *Category dimension*. A dimension contains keywords correspond to the nodes in a category hierarchy, such as Wordnet (Miller 1990; http://www.globalwordnet.org), in which all considered documents should be multi-categorized.

To simplify our discussion, we mainly use ordinary dimensions, together with the metadata dimension *time* (i.e., date), in the following examples.

Definition 3: For a *dimension D*, the *ith-level member set*, denoted $D(i)$, is defined as $D(i) = \{a \mid a$ is a member in the i-th level of D, but a is not a summary member$\}$. Besides, we use

$D(0)$ to denote the union of all non-summary members in D, which is the union of all i-th level member sets in D. That is, $D(0) = \cup_{1 \le i \le h} D(i)$, where h is the height of D.

To illustrate the above definitions, we give an example as follows.

Example 1: Suppose there is a dimension, denoted R, for representing the regions of territory T as depicted in Figure 1. All nodes with label '*' are summary members. That is, the summary member in the second level has the same meaning with T, which represents {*South*, *North*}. Besides, the summary members under *South* and *North* have the same corresponding meaning with *South* and *North*, which denote {*SC1*, *SC2*, *SC3*} and {*NC1*, *NC2*, *NC3*}, respectively. By omitting all the summary members, Figure 1 is redrawn in Figure 2. According to the illustration of dimension R, we know that $R(1) = \{T\}$, $R(2) = \{South, North\}$, and $R(3) = \{SC1, SC2, SC3, NC1, NC2, NC3\}$, and $R(0) = \{T, South, North, SC1, SC2, SC3, NC1, NC2, NC3\}$. In Figure 3, another dimension, denoted P, representing the products of a company manufacturing consumer electronics is concisely depicted.

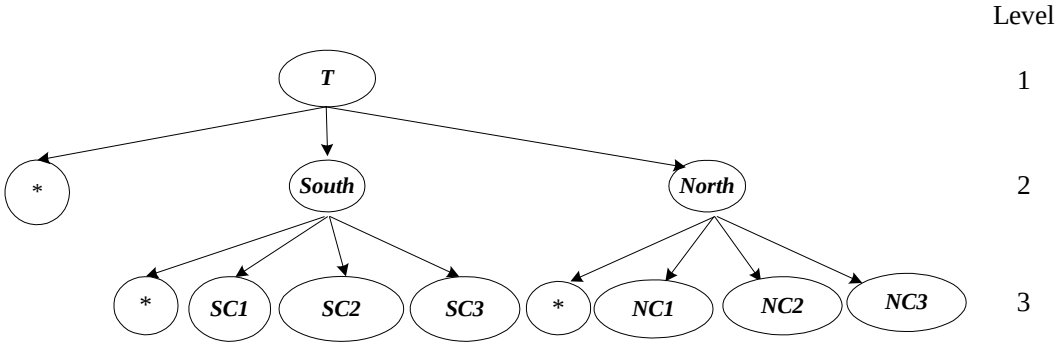

Figure 1: An illustration of dimension R.

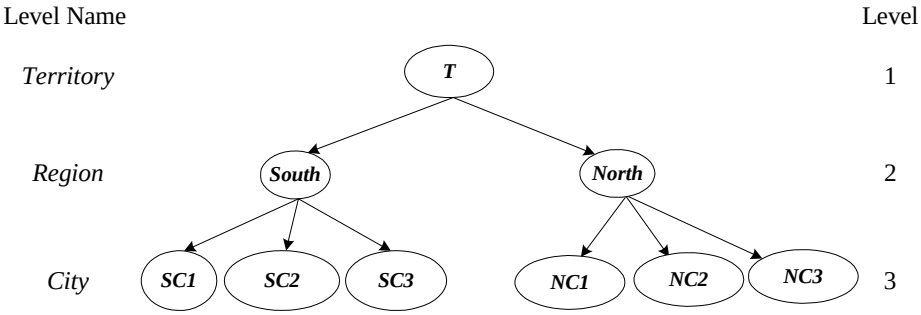

Figure 2: A concise illustration of dimension R.

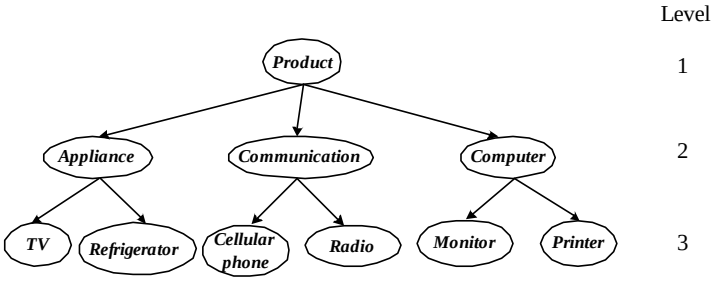

Figure 3: A concise illustration of dimension P.

For a dimension D, there are two basic operations called *drill-down* and *roll-up*, which are formally defined as follows.

Definition 4: For a *dimension D*, expanding an internal node to obtain all of its children is called *drill-down*, and shrinking a set of children to obtain their common parent is called *roll-up*.

Definition 5: For a document T with unique identifier id_T, a *document index of T* defined on n dimensions (D_1, D_2, …, D_n) is denoted $x = (id_T, K_T)$, where $K_T = (K_1, K_2, \ldots, K_i, \ldots, K_n)$ is an n-tuple of keyword sets, such that each K_i contains a set of keywords, and for all keywords $t_{ij} \in K_i$, t_{ij} occurred in T and $t_{ij} \in D_i(0)$, for all $1 \le i \le n$. In the following, the first and second components of a document index $x = (id_T, K_T)$ will be denoted x^1 and x^2 (i.e., $x^1 = id_T$ and $x^2 = K_T$), respectively.

Example 2: Suppose there is a complaint e-mail issued from a customer as shown in Figure 4. Then, a base document index of T defined on the above two dimensions (R, P) can be obtained as x = (A0001, ({*SC2*}, {*TV*})), where A0001 is the unique identifier of T.

The basic component of a document cube is called a *cell*, which is defined as follows.

Definition 6: A *cell* defined on n dimensions (D_1, D_2, …, D_n) is denoted $c = (t_c, X_c)$, where $t_c = (c_1, c_2, \ldots, c_i, \ldots, c_n)$, $c_i \in D_i(0) \cup \{\text{'*'}\}$, $1 \le i \le n$, and $X_c = \{x_1, x_2, \ldots, x_j, \ldots, x_m\}$ is a set of base document indices of the form $x_j = (id_{Tj}, (K_1, K_2, \ldots, K_n))$, where id_{Tj} is the unique identifier of some document T_j and $K_i \cap D_i(0) \neq \emptyset$, $1 \le i \le n$. The set of all such document unique identifiers id_{Tj} involved in the cell $c = (t_c, X_c)$ is denoted $ID(c) = \{x_j^1 \mid \forall\, x_j \in X_c\}$.

To whom it may concern:

We have bought a *TV* from your *SC2* branch last weekend. However, we found the screen is severely unstable. Please give us the phone number of your service center. Thank you for your kindly help.

Sincerely,

Frank S.C. Tseng

Figure 4: A complaint e-mail issued by a customer (A0001).

According to the above definitions, a *document cube* $DC = (S, (D_1, D_2, \ldots, D_n))$ is a cube composed of all cells $c_i = (t_{c_i}, X_{c_i})$ with $t_{c_i} \in \underset{1 \le j \le n}{\times} D_j(0)$ and $ID(c_i) \subseteq S$. A sample illustration of a document cube $DC = (S, (R, P, T))$ is shown in Figure 5.

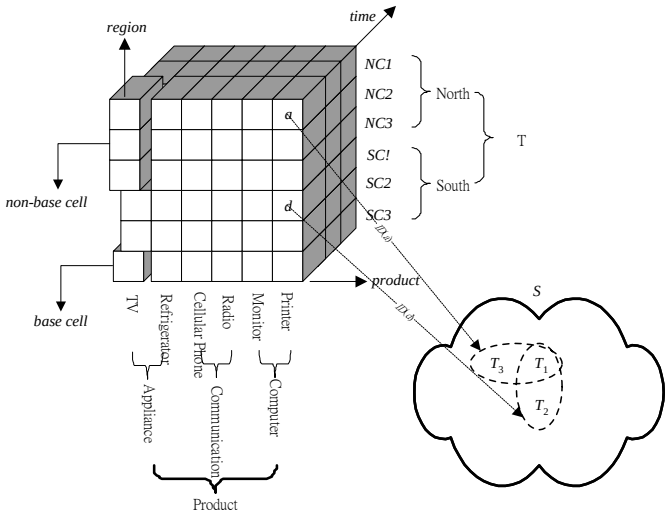

Figure 5: A sample illustration of a document cube.

3. A General Architecture of Document Warehouses

Designing a comprehensive architecture for document warehousing can be challenging because document warehousing covers a wide spectrum of concepts as we have shown in Section 2. Fortunately, there is already a general architecture being established for data warehousing in (Anahory and Murray 1997). Based on the architecture, we extend the constructs to include more features for documents warehousing. The proposed architecture is shown in Figure 6. The major components of a document warehouse are explained as follows.

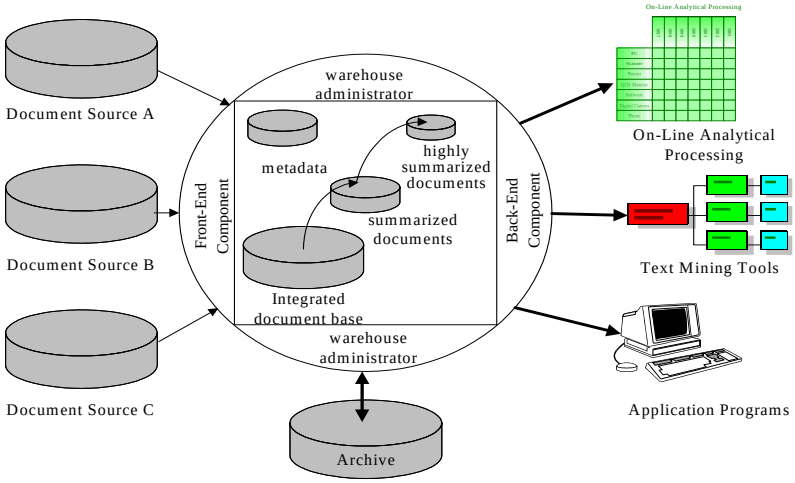

Figure 6: The proposed architecture of document warehouses.

3.1 Document Sources

The source of documents for a document warehouse is supplied from:

1. *Internal sources*: In an organization, there are documents in various formats spread throughout the organization on any kind of document repositories. The files may be in XML formats, MS Word formats, E-mail or even plain text.

2. *External sources*: Documents may also come from the Internet, including Web pages, FTP sites, commercially available document bases, private documents shared by private servers or document repositories associated with an organization's suppliers or customers.

3.2 Front-End Component

The front-end component performs all the necessary pre-processing of documents, such as text summarization (Goldstein 1999; Hahn and Mani 2000), text feature extraction (Feng and Croft 2001), document categorization (Appiani 2002), or other text mining procedures (Lin et al. 1998; Loh et al. 2000; Tan 1999), and then store the obtained features or patterns into the meta-data or store the summarized result as another summarized documents.

3.3 Warehouse Administrator

The warehouse administrator performs all the operations associated with the management of the documents in the warehouse. The operations include:

1. Enrich the metadata of all stored documents: Some of the document metadata (e.g., those defined in Dublin Core Metadata Element Set) may be missing and should be added manually by the warehouse administrator.

2. Perform necessary text mining operations or generate the summarization for documents either manually or by software tools (e.g., IBM Intelligent Miner for Text).

3. Create the dimensions and document indexes for constructing document cubes.

4. Archive documents and related data/metadata.

3.4 Back-End Components

The back-end component performs all the operations responsible for the management of user queries. It is typically composed of a set of document access tools, a multi-dimensional document query interface, document warehouse monitoring tools, and customized tools.

3.5 Highly Summarized documents

The part stores all the summarization derived from multiple documents, which belong to the same cluster or categorization. Some of such achievements have already been conducted (Chen and Huang 1999; Goldstein 2000).

The simplest format of a highly summarized document can be represented by a set of keywords appeared in the original document. Keywords of a document can be derived by computing the traditional *tf***idf* weights (Salton and Gill 1983; Salton 1988), pivoted cosine weights (Singhal et al. 1996), or one derived by any term-weighting scheme.

4. Data Modeling of Document Warehouses

The dimensional modeling technique (Inmon 1993; Kimball 1996) adopted widely in data warehouse modeling can be extended for document warehouses. Every dimensional model is composed of one central table with a composite key, called the *fact table*, which uses foreign keys to link to a set of *dimension tables*. This characteristic 'star-like' structure is also called a *star schema*. Such multi-dimensional data model for text permits the definition of any dimension of interest as defined in *Definition 2*. In Figure 7, we show a star schema for modeling document warehouses.

4.1 Dimensions

As we have discussed in Section 2, dimensions can be distinguished into the following types:

1. *Ordinary dimension*. A document can be highly summarized by a set of keywords. Therefore, we can construct an ordinary dimension containing a set of keywords to allow users to pinpoint to the desired documents directly.

2. *Metadata dimension*. That is, those elements defined in Dublin Core Metadata Element Set: title, creator, subject, description, publisher, contributor, date, type, format, identifier, source, language, relation, coverage, and rights, all can be regarded as metadata dimensions. Some of the dimensions might be hierarchies or simply related data.

3. *Category dimension*. For example, a hierarchy such as Wordnet or its subset, or user-defined hierarchies can be employed as category dimensions. Notice that, there may be more than one category dimensions used to construct a document cube, since a document can be multi-categorized into different categories from various points of view.

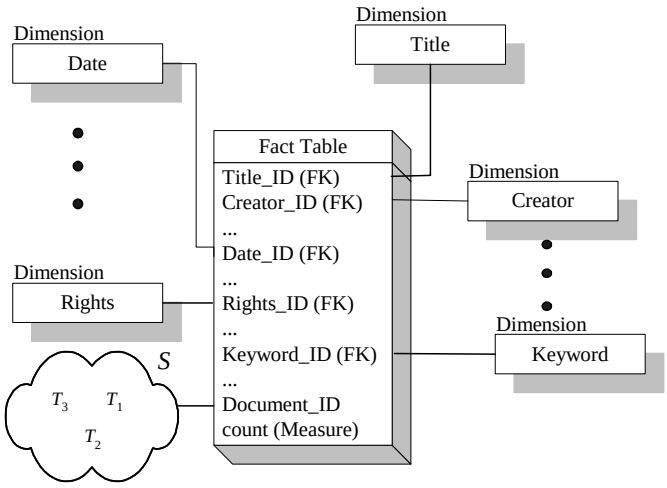

Figure 7: An example star schema of document warehouse.

4.2 The Fact Table

In general, the central fact table may be composed of the following attributes:

1. A composite key, which is composed of a set of foreign keys to the following dimensions:

   (a) Ordinary dimensions: For example, the dimension *Keyword* shown in Figure 7 is an ordinary dimension.

   (b) Metadata dimensions: For example, the dimensions *Title*, *Date*, *Creator*, …, and *Rights* as shown in Figure 7 are metadata dimensions.

   (c) Category dimensions. Notice that, there are no category dimensions shown in Figure 7.

2. Attributes used to derive the measures in a document cube. The document count (i.e., the attribute *count* of fact table in Figure 7) can be regarded as the default measure in a document cube. Another possible measure has been defined in (McCabe et al. 2000) as the weight of the *term frequency* of the corresponding keyword.

3. A Document_ID column stores the document pointer, which links to the file pathname as shown in Figure 7. That is, the set *S* in Figure 7 can be regarded as dimension containing all the document files and the Document_ID can be regarded as a measure.

5. An Application for Customer Relationship Management

In this section, we present an application of document warehouses. Suppose there is a company manufacturing appliances, communication equipments and computer peripherals, and it has established branches in the north and south regions. The objective is to warehouse customer complaint E-mails for customer relationship management.

After modeling the document cube, we obtain two metadata dimensions (Creator and Date) and three ordinary dimensions (Region, Product, and Time) as shown in Figure 8.

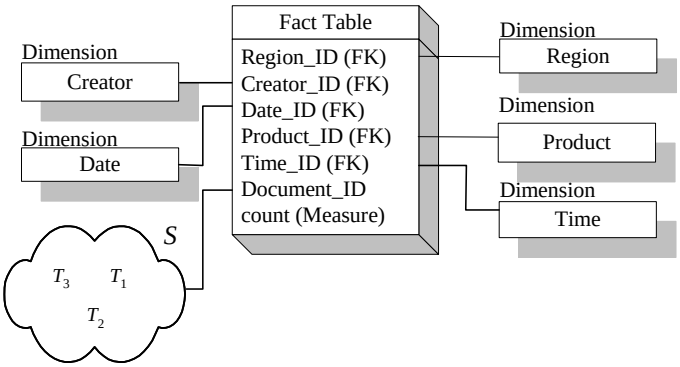

Figure 8: An example star schema for complaint e-mail management.

We briefly describe these dimensions as follows:

1. *Ordinary dimension*. The dimensions Region and Product are as Figure 3 shows. The dimension Time is the purchase time described in the E-mail.

2. *Metadata dimension*. The dimension Creator stores the E-mail addresses of customers and dimension Date stores the date of receiving of E-mails.

3. *Category dimension*. There are no category dimensions shown in this example. However, the E-mail documents could be further categorized either manually or automatically categorized by software tools into hierarchical categories.

The fact table is composed of the following attributes:

1. A composite key, which is composed of a set of foreign keys to the fore-mentioned dimensions:
2. The attribute *count* is regarded as the default measure in this document cube.
3. A Document_ID column stores the E-mail file pathname.

After constructing the document cube, we can perform on-line analytical processing on the obtained document cube as illustrated in Figure 9. Notice that, each of the count shown in Figure 9 is actually a hyperlink, which links to a page containing the original E-mails.

2004/Quarter 1		Region	
Product		North	South
Appliances	TV	5	2
	Refrigerator	3	7
	DVD Player	7	9
Computer Peripherals	Laptop Computer	8	1
	Desktop Computer	2	3
Communication Equipments	Radio	6	1
	Cellular Phone	5	7

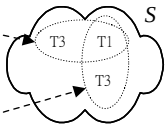

Figure 9: On-line analytical processing over the example document cube.

6. Conclusion and Future Directions

6.1 Conclusion

While data warehouses and the numeric-centric business intelligence technologies have served most of the enterprises well, they do not fully address the complete scope of business intelligence. In this paper, we advocate the importance of constructing document warehouses to support text-centric business intelligence, and propose an architecture for document warehousing. When documents are warehoused, users can perform *ad hoc* on-line analytical

processing (OLAP) over text in a document warehouse, just as the way users can perform OLAP over summarized data in a data warehouse.

The applications of document warehousing are versatile. In business, document warehousing can help administrators organize the meeting reports, gazettes, or even customer complaint e-mails, where the company personnel, products, and time may be regarded as the dimensions, such that documents related to some employees, or products in some time, at somewhere can be retrieved or browsed instantly.

When documents are warehoused, the task of version control will become very easy, since users can directly tracing the documents based on some criteria along the time dimension. Besides, document clustering can be achieved directly via visualizations. Users can also develop some document summarization tools to summarize a cluster of related documents. To sum up, data warehousing and document warehousing are not only one of the most infrastructure of knowledge management, but also the kernel of customer relationship management.

6.2 Future Works

In our future work, we will conduct more techniques for document warehousing. The preliminary components may include the following modules.

1. *Employ XML Schema to define document metadata*. We advocate using the Extensible Markup Language (XML) to be the intermediate media for document interchange.

2. *Incorporate automatic text summarization* (Goldstein et al. 1999; Goldstein et al. 2000; Knight 1999), *key feature extraction* (Feng and Croft 2001), *or even document classification and categorization* (Appiani et al. 2002) *techniques for document warehousing*. Develop related text summarization techniques to extract the most important 10% ~ 20% content for users to digest the documents more easily and propose how to bind a document summary with its corresponding documents for document warehousing.

Finally, since the construction of a document warehouse has to scan a large amount of documents, which is a task prone to time-consuming, the parallel architecture for such process will be further investigated in the future.

References

Anahory, S. and Murray D., *Data Warehousing in the Real World: A Practical Guide for Building Decision Support Systems*, Harlow, England: Addison-Wesley Longman, 1997.

Appiani, A., Cesarini, F., Colla, A., Diligenti, M., Gori, M., Marinai, S., and Soda, G. "Automatic document classification and indexing in high-volume applications," *International Journal on Document Analysis and Recognition*, Vol. 4, No. 2, 2002, pp. 69-83.

Berry, M.J.A., and G. Linoff, *Data Mining Techniques: For Marketing, Sales, and Customer Support.* New York: John Wiley & Sons, 1997.

Bertino, E. and B. Catania, "Integrating XML and Databases," *IEEE Internet Computing*, Vol. 5, No. 4, 2001, pp. 84-88.

Bertino, E. and E. Ferrari, "XML and Database Integration," *IEEE Internet Computing*, Vol. 5, No. 6, 2001, pp. 75-76.

Champion, M, "Native XML *vs*. XML-Enabled: the Difference Makes a Difference," http://www.softwareag.com/xml/library/champion_nativexml.htm, Software AG: The

XML Company.

Chen, H.H. and Huang, S. J., "A summarization system for Chinese News from multiple sources," *Proceedings of 4th International Workshop on Information Retrieval with Asia Language*, 1999, pp. 1-7.

Dublin Core Metadata Initiative, http://dublincore.org/

Feng, F. F. and Croft, W. B. "Probabilistic techniques for phrase extraction," *Information Processing and Management*, Volume: 37, Issue: 2, Mar. 2001, Page(s): 199-220.

Goldstein, J., Kantrowitz, M., Mittal, V., and Carbonell, J. "Summarizing text documents: Sentence selection and evaluation metrics," *Proc. of SIGIR*, 1999, Page(s): 121-128.

Goldstein, J., Mittal, V. O., Carbonell, J. G., Callan, J. P., "Creating and Evaluating Multi-Document Sentence Extract Summaries," *Proceedings of the 9th International Conference on Information and Knowledge Management*, 2000, pp. 165-172.

Grigsby, M., "The Internet Document Warehouse: Content Management for the Back Office," Technical Report, IMERGE Consulting, Inc., 2001. http://www.imergeportal.com/publishedarticles.asp

Hackathorn, R., Data Warehousing Energizes Your Enterprise, *Datamation*, Feb. 1995, Vol. 1, pp. 38-42.

Hahn, U. and Mani, I. "The challenges of automatic summarization," *IEEE Computer*, Volume: 33, Issue: 11, Nov. 2000, pp. 29-36.

Han, J. and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, 2001.

http://www.survey.com, "Development Snapshot: Warehouse Data of the Future," *Application Development Trends*, Feb. 2000.

Inmon, W.H., *Building the Data Warehouse*, New York, NY: John Wiley and Sons, 1993.

Kimball, R, *The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses*, John Wiley & Sons, Inc., 1996.

Knight, K., "Mining Online Text," *Communications of the ACM*, Vol. 42, No. 11, 1999.

Lin, S.-H., C.-S. Shih, M.C. Chen, J.-M. Ho, M.-T. Ko, and Y.-M. Huang, "Extracting Classification Knowledge of Internet Documents with Mining Term Associations: A Semantic Approach," *Proc. SIGIR*, 1998.

Loh, S., L.K. Wives, and J.P. de Oliverira, "Concept-Based Knowledge Discovery in Texts Extracted from the Web," *SIGKDD Explorations*, Vol. 2, No. 1, Jun. 2000.

McCabe, M.C., J. Lee, A. Chowdhury, D. Grossman, and O. Frieder, "On the Design and Evaluation of a Multi-dimensional Approach to Information Retrieval," *Proc. 23th Annual International ACM SIGIR Conference*, 2000, pp. 363-365.

Miller, G.A., "Wordnet: An Online Lexical Database," *Int'l J. Lexicography*, Vol. 3, No. 4, 1990, pp. 235-312.

Salton, G.., "*Automatic Text Processing*," Addison-Wesley Publishing Company, 1988.

Salton, G.., Gill, M., "*Introduction to Modern Information Retrieval*," McGraw-Hill, 1983.

Singhal, A., C. Buckley, and M. Mitra, "Pivoted Document Length Normalization," *Proc. 19th Annual International ACM SIGIR Conference*, 1996, pp. 21-29.

Sullivan, D. *Document Warehousing and Text Mining: Techniques for Improving Business Operations, Marketing and Sales*, John Wiley & Son, Inc., 2001.

Tan, Ah-Hwee, "Text Mining: The State of The Art and The Challenges," *Proc. PAKDD 99---Workshop on Knowledge Discovery from Advanced Databases*, Beijing, 1999, pp. 50-70.

The Global Wordnet Association, http://www.globalwordnet.org/

Acknowledgement: This research was partially supported by the National Science Council, ROC, under contract No. NSC 92-2416-H-327-005.