TOWARDS A MULTI-USER DOCUMENT WAREHOUSE

Kaïs Khrouf[1], Maha Azabou[1], Jamel Feki[1] and Chantal Soulé-Dupuy[2]
[1]MIR@CL, University of Sfax, Airport road, Km 4, 3018 Sfax, P.O. Box 1088, Tunisia
[2]IRIT, University of Toulouse 1, Capitole 2 street of Doyen-Gabriel-Marty, 31042 Toulouse Cedex 9, France

Keywords: Document Warehouse, XML Documents, OLAP, Ontology.

Abstract: The development of the Internet generated the increase of data volume available and, the number of documents exchanged through this network. Currently, this data is more and more used by companies for economic, strategic, scientific and technical development. Therefore, it is essential to provide efficient tools for decision makers in order to help them analyzing document contents in a simple way; i.e., as they actually analyze factual or descriptive data. In this paper, we present a meta-model of document warehouse including three major components (Structural, Semantic and User components) and we describe an approach for multidimensional analyses of the warehoused documents.

1 INTRODUCTION

Data warehouses and OLAP systems (On-Line Analytical Processing) provide methods and tools for analyzing data that trace the enterprise activities. However, only few data of the business information system may be processed with the marketed OLAP systems; these systems mainly rely on factual operational data (in databases or data warehouses). About 80% of pertinent data are contained in documents and, therefore remains out of reach of OLAP systems due to the lack of available tools and processes dedicated to this data type. Besides, the emergence of new documentation standards such as XML, has offered the opportunity to identify and focus on parts of documents.

Thus, it is essential to provide decision makers with efficient tools that analyze information enclosed in text documents. This objective can be reached by using the concept of document warehouse which allows the storage of documents and their exploitation through the techniques of multidimensional analyses of information.

Within a document warehouse, documents are gathered into classes having similar characteristics, thus enabling users to better focus on (1) a document class which interests them (e.g., books, newspapers, proceedings), and/or (2) a document set dealing with a particular domain or theme. In order to reach our goal, we propose a meta-model for document warehouse, it is composed of three major components: *Structural*, *Semantic* and *Users*.

This paper is organized as follows. In section 2, we outline some works devoted to the document storage and the OLAP technology. Then, in section 3, we present our meta-model for warehousing documents. In section 4, we detail our proposed approach to realize multidimensional analyses on factual and textual data. Finally, in section 5, we give an overview of *DocWare* (*Doc*ument *Ware*house): our software prototype for the integration and the analysis of documents.

2 RELATED WORK

XML documents can be classified into two categories (Ravat et al., 2010):

- *Data-centric* XML documents are composed of short and precise elements and are similar to relational data. Mainly used by applications to exchange information (i.e. transactional data).
- *Document-centric* XML documents are text-rich documents and are the electronic version of traditional paper documents (i.e. scientific articles, internal reports, e-books).

For the OLAP manipulation of data-centric XML documents, several works have been proposed. In (Boussaid et al., 2006), the authors use a snowflake schema for an XML data warehouse built from breast cancer patient files. These proposals are

Khrouf K., Azabou M., Feki J. and Soulé-Dupuy C..
TOWARDS A MULTI-USER DOCUMENT WAREHOUSE.
DOI: 10.5220/0003937801490154
In *Proceedings of the 8th International Conference on Web Information Systems and Technologies* (WEBIST-2012), pages 149-154
ISBN: 978-989-8565-08-2

specific to complex data, but they cannot be adapted for analyses of textual data. The authors of (Hachaichi et al., 2009) propose an automatic approach to design multidimensional schemas starting from a DTD (Document Type Definition). However, their approach is purely syntactic since they do not tackle the semantic aspects of the DTD tags neither the semantics of the document contents.

Some other works have been focused on OLAP manipulations of document-centric XML documents. In (Tseng et al., 2006), the authors use the OLAP environment for document analyses. We note that these analyses are restricted to the count of documents (e.g., number of emails, articles, Web pages) by selected dimensions. The authors of (Ravat et al., 2008) define a function to aggregate textual data in order to obtain a summarized vision of the information extracted from documents. We note that this work does not take into account the heterogeneity of document structures.

As an additional contribution to our previous work, the multidimensional approach we propose in this paper can be applied on data-centric XML documents as well as on document-centric XML documents. For that, we use our own meta-model (cf. Figure 1) for the document warehouse; this model is able to: (1) gather documents having identical or similar structures, and (2) provide the semantics of textual contents by using domain-ontologies. In this meta-model, we involved users and their profiles in order to permit the collaboration between these users and, the personalisation of their analysis.

3 DOCUMENT WAREHOUSE META-MODEL

3.1 Meta-model Description

The document warehouse constitutes a source of synthetic and homogeneous information. Accordingly, our proposed meta-model (cf. Figure 1) includes the following components:

- A set of documents (cf. Figure 1-a) to be integrated into the document warehouse.
- The hierarchical structures of all documents; it consists of two types of structures:

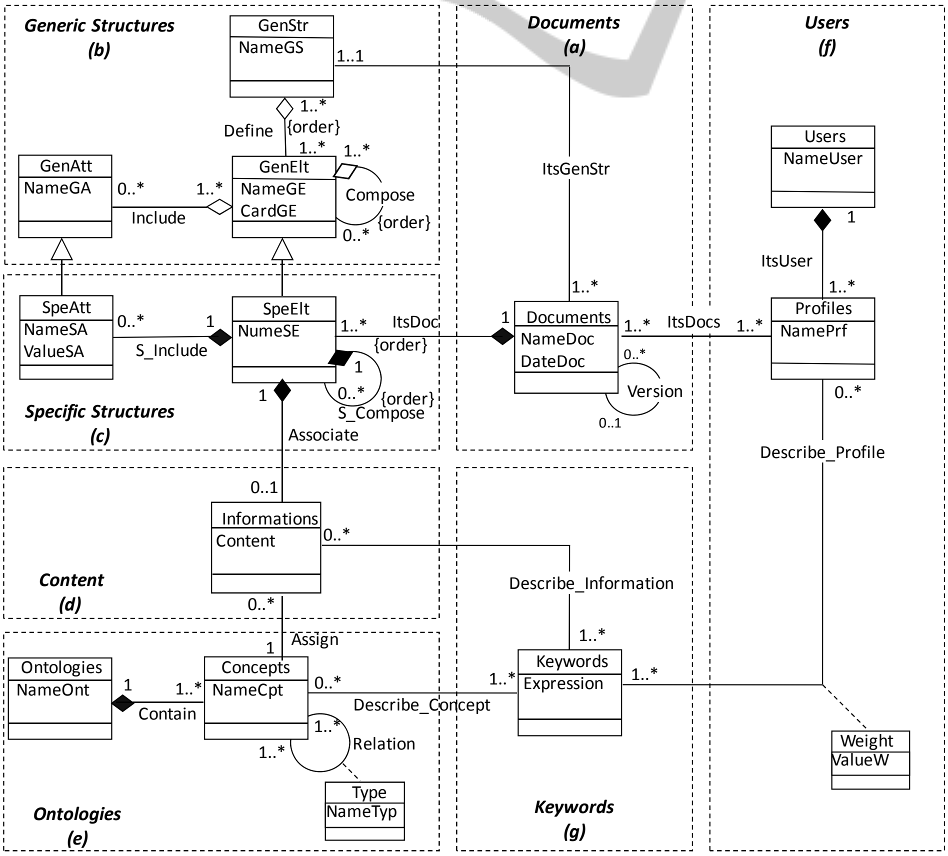

Figure 1: Meta-model for a document Warehouse.

1. The generic structures (cf. Figure 1-b): A generic structure is a common structure for a document set. This structure is defined by a set of generic elements (e.g., Title, Author), which can be composed of other generic elements. Each of these elements can also be described by generic attributes (e.g., Book-Id).

2. The specific structures (cf. Figure 1-c): It is associated to a single document and has to be compliant to one among the generic structures of the warehouse. This structure is defined by a set of specific elements that can include specific attributes.

- The content (cf. Figure 1-d). It is the content of the elements of the specific structures.
- The semantics layer (cf. Figure 1-e): It is defined using domain-ontologies. In our context, an ontology is composed of a set of concepts hierarchically organized where each concept is described by a set of keywords.
- The users (cf. Figure 1-f): They are able to define profiles; each profile is described by a set of keywords. These keywords can be weighted to indicate their importance compared with the profiles.
- Keywords (cf. Figure 1-g): They describe the textual content of documents, the ontology concepts of users' profiles.

3.2 Meta-model Instantiation Example

Figure 2 depicts an instantiation example for the meta-model of Figure 1.

In this example, there are two users, the first has two profiles *Web* and *DB* (i.e., DataBase), and the second user has one profile *IS* (i.e., Information System). This last profile is described by two keywords *OLAP* and *OLTP*. The warehouse contains two documents *Doc1* and *Doc2*, the first is assigned to the *Web* profile and, the second is assigned to the *DB* and *IS* profiles. These two documents have a specific structure conform to the generic structure *Article*.

Doc1 is described by the Title *WWW*, has an author *Foulen*, it is published in the conference *ICEIS* during the year *2010*; the specific element 1417 constitutes the content of this article. *Doc2* (Title: *OLAP*) is written by *Dupond*, it is published in (*WebIST*, *2011*) and has a content (specific element 1838). The textual contents of documents were assigned to the corresponding concepts of ontologies. For example, the Title *OLAP* and the paragraph "Fact…" have been assigned to the *DW* concept.

3.3 Meta-model Instantiation Process

The integration of a document into the warehouse is

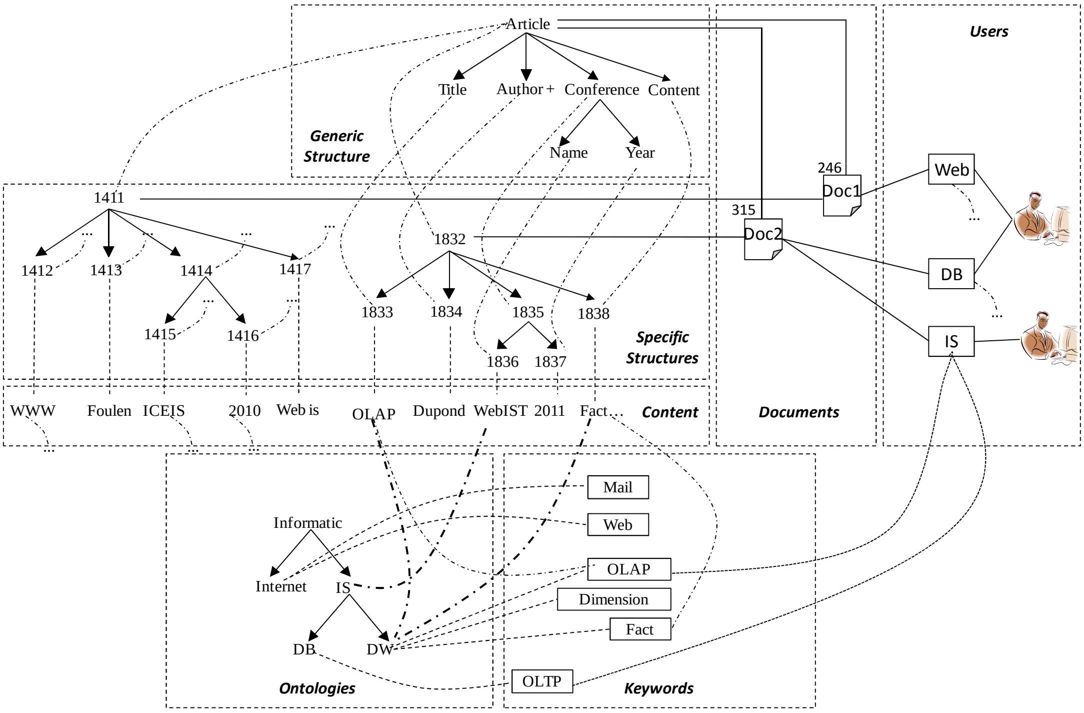

Figure 2: Example of instantiation of meta-model of Figure 1.

accomplished through the following steps:

1. Extraction of the specific structure of the document, i.e. the document tags and its hierarchical structure. It is performed using a parser as detailed in (Khrouf et al., 2004)

2. Selection of the most appropriate generic structure among the set of generic structures of the warehouse. This step is accomplished through a matching algorithm (Ben Messaoud et al., 2011).

3. Insertion of the document content (i.e., information and list of keywords) into the warehouse according to the structure selected in the previous step.

4. Allocation of concepts to the textual content: This step assigns to each leaf element of a document *d* the most significant concept from the available domain-ontologies (enhancing the warehouse semantics).

To do so, we calculate the weight of each concept C_i relative to each leaf element E_j of the document *d*. This weight is calculated using Formula 1.

$$Weight\,(C_i^{E_j}) = \frac{\left|C_i^{E_j}\right|}{\left|C_i^{d}\right|} * Weight\,(C_i^{O_K}) \qquad \forall\ j \qquad E_j \in d \qquad (1)$$

Where:

- $\left|C_i^{Ej}\right|$ denotes the number of occurrences of concept C_i (or one of its terms) in element E_j,
- $\left|C_i^{d}\right|$ represents the number of occurrences of C_i (or one of its terms) in document *d*, and
- $Weight(C_i^{O_k})$ is the weight of the concept C_i in the domain-ontology O_k. This weight is calculated by assigning more importance to leaf elements.

The experiments realized on a set of 140 documents downloaded from Wikipedia have shown that the allocation of concepts to textual contents was clearly improved with weighted-concepts.

5. Allocation of documents to user profiles: When the user inserts a document into the warehouse, we calculate the relevance of this document compared with the user profiles. Each user who connects to the warehouse system will be informed of the documents inserted (by other users) that are relevant to its profiles. Therefore, he/she can access the content of these documents in order to accept/reject them compared to their profile. Thereafter, the user can visualize or manipulate only those documents belonging to its profile(s).

4 MULTIDIMENSIONAL ANALYSES

The proposed multidimensional process, to analyze the document warehouse, is decomposed into two phases: (1) Document mart construction, and (2) views generation and result visualization.

- **Document mart construction**

Let us remember that a generic structure gathers a set of documents having identical or similar structures. The decision makers can focus on a generic structure to perform his/her analyses.

The first step consists in selecting the analysis context through the choice of the generic structure on which analyses will be applied. Then, he/she should select the multidimensional schema components; i.e., one *fact* and a set of related *dimensions*. The fact represents the subject to be analyzed (e.g., *Number of articles*) and the dimensions represent the context of recording and analyzing the fact (e.g. *Author*, *Year*, *Conference*). In addition, the decision-maker indicates the order of dimensions and the aggregate function (i.e., *Count*, *Sum*, *Max*…) to be applied on the fact measures. An example of multidimensional analysis is: *Number of published articles* by *Author*, *Year* for the *Database* Concept.

- **Views Generation and result visualization**

For every multidimensional schema component selected, the system generates a view containing the three attributes *Doc*, *Anc* and *Inf* as follows.

```
CREATE VIEW Dim_n (Doc, Anc, Inf) AS
SELECT
e.S_Compose…  S_Compose.ItsDoc.NumDoc,  (1)
e.S_Compose…  S_Compose.NumSE,          (2)
e.Content                               (3)
FROM SpeElt e                           (4)
WHERE e.S_Compose…  S_Compose.ItsDoc.
ItsGenStr.NameGS ='NameGS';             (5)
And e.S_Compose…  S_Compose.ItsDoc
In(SELECT
  THE(SELECT p.ItsDocs
  FROM Profiles p
  WHERE P.ItsUser.NameUser ='NameUser'))(6)
--If the dimension is a generic element
And e.Inherit.NameGE = 'NameGE'         (7)
--If the dimension is a concept
AND e.Associate.Affect.NameCpt='NameCpt;(8)
```

where:

(1) The document identifier
(2) The identifier of specific elements that inherit from the first common ancestor of all analysis elements.
(3) The content of the specific element.
(4) The table name among those of the meta-model.

(5) Selection of documents belonging to the generic structure chosen by the user (i.e., decision-maker).
(6) Selection of documents belonging to the user profiles.
(7) Indicates the selected name of the generic element when a dimension is based on a generic element.
(8) Indicates the chosen name of the concept when a dimension is based on a concept.

Note that the fact view is generated in the same way like are dimensions.

After generating the fact-view and its dimensions views, it is necessary to link these views on their first two attributes *Doc* and *Anc*; thus we generate a new view called *VJoint*. From this view, a last view describing the document mart is generated by grouping (*Group by*) all the dimensions and applying the aggregation operation chosen by the decision-maker. The content of this view will be visualized in a multidimensional table (cf. Figure 4).

5 IMPLEMENTATION

To validate our proposals, we developed a software prototype called *DocWare* (acronym of *Doc*ument *Ware*house) dedicated to the integration and the analysis of documentary information. The multidimensional analyses can be performed on:

- Data-centric XML documents (cf. Example 1): We use only the generic structure components.
- Document-centric XML documents (cf. Example 2): We combine the elements of the generic structure and the ontology concepts.

Example 1: Data-centric XML documents

Assuming we want to analyze the *Sum of quantity* in *Transactions* by *Product*, *Customer* and *Year*.

To perform this analysis, the system displays a list of all existing generic structures in the warehouse. Among these structures, the decision-maker selects the corresponding generic structure, namely *Transactions* that will be visualized by a tree as displayed in Figure 3.

After that, the decision-maker specifies the role (i.e., dimension or fact) of some elements to build the mart by using contextual menus. The elements chosen by the decision-maker are highlighted by using different shapes and colors for the dimensions and fact. In our example, the dimensions are the *Name of Product* (called *nom* in Figure 3), the *Customer* (*client* in Figure 3) and the *Year* (*annee* in Figure 3). The measure is the sum of quantities (*qt* in Figure 3).

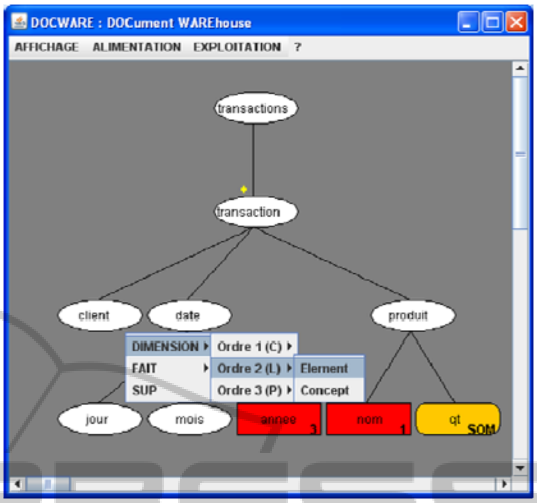

Figure 3: The analysis components for Example 1.

To visualize the result, the system creates views according to the approach described in Section 4 and displays the following multidimensional table.

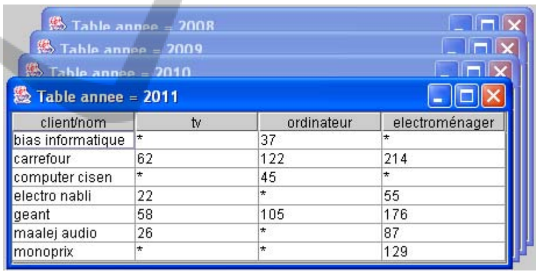
Table annee = 2008
Table annee = 2009
Table annee = 2010
Table annee = 2011

client/nom	tv	ordinateur	electroménager
bias informatique	*	37	*
carrefour	62	122	214
computer cisen	*	45	*
electro nabli	22	*	55
geant	58	105	176
maalej audio	26	*	87
monoprix	*	*	129

Figure 4: Multidimensional table for Example 1.

Example 2: Document-centric XML documents

Now, let us analyze the scientific articles dealing with *Data Warehouse Domain* as well as by *Author* and *Published Year*.

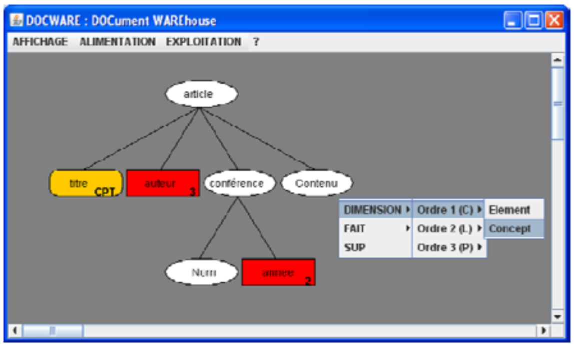

Figure 5: Choice of analysis components for Example 2.

The decision-maker assigns a *Data Warehouse* Concept to generic element *Content* (*contenu* in Figure 5). This is the first dimension. Then, he/she selects the generic element *Year* as the second dimension (*année* in Figure 5), the generic element *Author* (*auteur* in Figure 5) as third dimension. The measure is the count of titles (*titre* in Figure 5).

To assign a concept to a generic element, the system displays a list of all existing ontologies in the warehouse to permit the decision-maker to choose the appropriate ontology (cf. Figure 6).

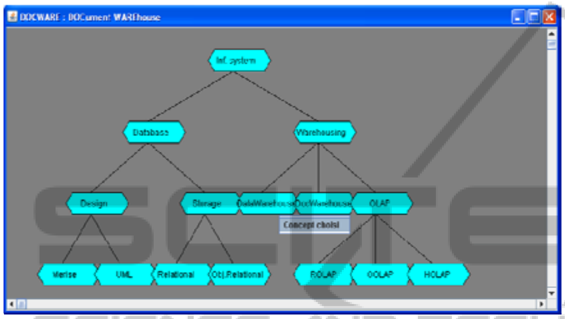

Figure 6: Choice of ontology for Example 2.

The result of this analysis is displayed using a multidimensional table as shown in Figure 7, where the first column represents *Years* and the second column gives the *Number of articles* for the *Author* in the sheet heading.

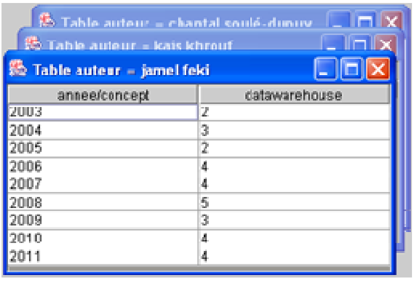

Table auteur – chantal soulé-dupuy

Table auteur – kais khrouf

Table auteur – jamel feki

annee/concept	datawarehouse
2003	2
2004	3
2005	2
2006	4
2007	4
2008	5
2009	3
2010	4
2011	4

Figure 7: Multidimensional table for Example 2

6 CONCLUSIONS

The document warehouse constitutes a solution to exploit and analyze textual data extracted from documents.

The meta-model proposed in this paper is suitable for (1) storing heterogeneous documents according to their structures and semantics, and (2) applying the techniques of multidimensional analysis, i.e. analyzing data according to several dimensions through graphical language interfaces that offers a high simplicity for users.

The perspectives that we intend to lead in order to extend this work concern the following areas: (1) determining semantic structures of the documents integrated in the warehouse and gathering theses structures into semantic classes, (2) defining the user profiles from the parts of domain-ontologies (instead of simple keywords), and (3) involving the user in the construction process of document mart (recommendation of potential analysis components based on the user profiles and the relevant domain-ontologies).

REFERENCES

Ben Messaoud, I., Feki, J., Khrouf, K., Zurfluh, G., Unification of XML Document Structures for DOCW. In *International Conference on Enterprise Information Systems (ICEIS'11)*, p. 85-94, Beijing, China, 2011.

Boussaid, O., Ben Messaoud, R., Choquet, R., Anthoard, S., 2006. X-Warehousing: An XML-Based Approach for Warehousing ComplexData. In *East European Conference. on Advances in Databases and Information Systems (ADBIS'06)*, p. 39–54. Thessaloniki, Hellas.

Hachaichi, Y., Feki, J., Ben-Abdallah, H., 2009. Designing Data Marts from XML and Relational Data Sources. In *Design and Advanced Engineering Applications: Methods for Complex Construction, Advances in Data Warehousing and Mining Series*, GI Global, p 55-80, Bellatreche Edition.

Khrouf K., Soulé-Dupuy C., 2004. A Textual Warehouse Approach: a Web Data Repository, In *Intelligent Agents for Data Mining and Information Retrieval*, (Chapter VII), p. 101-124, Idea Group Publishing.

Ravat, F., Teste, O., Tournier, R., Zurfluh, G. 2008. Top_Keyword: an Aggregation Function for Textual Document OLAP. In *International Conference on Data Warehousing and Knowledge Discovery (DaWaK'08)*, p. 55-64, Turin, Italiy.

Ravat, F., Teste, O., Tournier, R., Zurluh, G., 2010. Finding an application-appropriate model for XML data warehouses. In *Information Systems*, Vol. 35, issue 6, p. 662-687, Elsevier.

Tseng, F., S., C., Chou, A., Y., H., 2006. The concept of document warehousing for multi-dimensional modeling of textual-based business intelligence. In *Decision Support Systems*, Vol. 42, p. 727– 744, Elsevier.