Integration Process for Multidimensional Textual Data Modeling

Rachid Aknouche, Ounas Asfari, Fadila Bentayeb and Omar Boussaid

ERIC Laboratory (Entrepts, Reprsentation et Ingnierie des Connaissances)
5 Avenue Pierre Mends France, 69676 Bron Cedex, France

Abstract. In this paper, we propose an original approach for text warehousing process. It is based on a decisional architecture which combines classical data warehousing tasks and information retrieval (IR) techniques. We first propose a new ETL process, named ETL-Text, for textual data integration and then, we present a new Text Warehouse Model, denoted TWM, which takes into account both the structure and the semantics of the textual data. TWM is associated with new dimensions types including: a metadata dimension and a semantic dimension. In addition, we propose a new analysis measure based on the modeling language widely used in IR area. Moreover, our approach is based on Wikipedia as external knowledge source to extract the semantics of the textual documents. To validate our approach, we develop a prototype composed of several processing modules that illustrate the different steps of the ETL-Text. Also, we use the 20 Newsgroups corpus to perform our experimentations.

1 Introduction

Research in data warehousing and On-Line Analytical Processing (OLAP) have produced new technologies for the design and implementation of information systems for decision support. To achieve the value of a data warehouse, incoming data must be transformed into an analysis-ready format. Currently, data warehousing systems often provide tools to assist in this process, especially when data are numerical [1]. Unfortunately, standard tools are inadequate for producing relevant analysis when data are complex, such as textual data. While recent studies confirm that 80% of an organization's information appear in a textual format [2]. Thus, it is important to propose new technologies to deal with the textual data and to extract and model their semantics. In fact, word sequences contain semantics information which are not accessible directly, this problem has been discussed in the literature, and several techniques have been proposed particularly in the data mining and information retrieval (IR) areas. However, in data warehousing and OLAP area, few works have focused on this issue. Moreover, it is necessary to quote the limits of the standard ETL (Extract-Transform-Load) on a textual data. The ETL process implies all the tasks that are necessary to feed and refresh a data warehouse. It is responsible for the extraction of data from several sources, their cleansing, customization and insertion into a data warehouse [3]. The current techniques for the ETL process provide conclusive results when the processed data are structured or

Aknouche R., Asfari O., Bentayeb F. and Boussaid O..
Integration Process for Multidimensional Textual Data Modeling.
DOI: 10.5220/0004602501190126
In *Proceedings of the 1st International Workshop in Software Evolution and Modernization* (SEM-2013), pages 119-126
ISBN: 978-989-8565-66-2

semi-structured, but the limitations of these techniques are obvious when integrating textual data in decisional systems.

In this paper, we propose a new approach for text warehousing process composed of three main components: (1) ETL-Text, (2) Text Warehouse Model TWM and (3) Text-OLAP. Here, we are interested on the first two phases; ETL-Text and TWM. The originality of our approach is that it extends the classical data warehousing process with new methodologies adapted for the text analysis. The proposed ETL-Text gathers process of extraction, transformation and loading appropriate to text warehouse. Moreover, TWM extends the constellation model to support representation of textual data in a multidimensional environment. Our solution to model textual data is based on IR techniques and information processing including language modeling. In addition, our approach provides a semantic layer which is represented in the ETL-Text by concepts extraction from an external knowledge source, and materialized in the TWM model by dimension named " *semantic dimension*". The semantic dimension is designed to assign topics to each documents in our data set. We depend on the online encyclopedia Wikipedia[1] as an external knowledge source.

The rest of this paper is organized as follows. In section 2 we present related work. Section 3 exposes our decisional architecture, the ETL-Text and the multidimensional model TWM. We give in section 4 our experimental study. Finally, we conclude this paper and we present research perspectives in section 5.

2 Related Work

Many research studies relate to textual documents integration in data warehouse environments, such as [4] which proposes a *meta-snowfalke* schema as multidimensional dynamic model for the text warehouses construction. They add a category index table to their snowflake schema. In [5], the authors propose a model named *DocCube* which offers a new kind of access to collections content based on topics, and help the user to improve and formulate his queries. [6] propose a document warehouse to textual data multidimensional analysis. They create dimension tables from keywords extracted from the documents. Inspired by the star schema, they use the ID and the number of relevant documents as measures in the *fact* table. In [7], the authors present a multidimensional IR (MIRE) that supports the integration of structured data and text. They use the inverted index as a IR technique in order to process and index the textual data. On the other hand, [8] generalize the concept of constellation (Kimball, 1996) to define a conceptual model on galaxy adapted to the data from XML documents. Their approach describes a multidimensional schema by the unique dimension concept, and thus the notion of fact is deleted.

In addition, several works in literature propose to deal the documents in multidimensional analysis, as in the works of [9], [10], [11], and [12]. They either extend the multidimensional models, which are used to process numeric data, to the documents analysis, or combine the IR techniques with OLAP for the multidimensional analysis of documents.

[1] http :// www.wikipedia.org/

Although these different proposed works, several difficulties still exist in the interpretation and the extraction of semantics from textual data. In fact, this data is extracted from various data sources and thus it requires different solutions to problems related to heterogeneity models, schemas and semantics. The proposed solution of this problem is firstly a new generation of data warehouse models, which allows to organize textual data in a multidimensional environment, and secondly, to define new OLAP operators for the textual analysis. From this point of view, the use of technologies issued from Information Retrieval IR area and Natural Language Processing NLP area constitute interesting tracks to process textual data in a multidimensional environment.

3 Text warehousing Architecture

In this section, we present a decisional architecture for text warehousing. It covers three main phases dedicated to the construction and the implementation of a data warehouse and use new methods to adapt them to textual documents. These phases are: (1) data integration ; (2) multidimensional modeling and (3) OLAP analysis. In this study, we are particularly interested in two major scientific challenges: (1) What are the appropriate techniques to perform the various tasks of textual ETL; (2) What modeling for textual data in a multidimensional environment.

As a solution to the first one, we propose an ETL process appropriated to textual data named ETL-Text. Beforehand, this process considers the select of meaningful terms before the loading task in the warehouse. In fact, the extraction task, in this process, is performed by data filtering operation based on Natural Language Processing NLP techniques to select from the original texts a set of terms relevant to the analysis. These relevant terms are then assigned to the transformation task to form so-called, in our approach, the *candidate-terms*. During this phase, we distinguish two transformations types: structural and semantic. The first focuses on the syntactic aspect of textual data, while the second considers the data semantics aspect based on an external knowledge source, Wikipedia in our approach. Moreover, to index documents, we depend on a modeling language notion widely used in Information Retrieval systems.

Otherwise, for the second challenge of the textual data modeling, we propose TWM as an extension of the classical multidimensional model. TWM depends on IR techniques to model textual data. It contains beside the classical dimensions, two another types; metadata dimension and semantic dimension in order to represent knowledge that describes documents and their semantics.

The main objective of our approach is to easily interact with an OLAP system and execute IR queries for text analysis. The proposed architecture consists of three phases: (1) ETL-Text phase which can extract, transform and load textual data in a warehouse; (2) Text warehouse modeling TWM phase; and (3) Text-Olap phase which allows to analyse textual data and to answer to decisional queries by using OLAP operator. In this paper, we will focus specifically on the first two phases.

3.1 The ETL-Text

The ETL-Text tasks can be summarized as follows:

1. Extract documents' textual entities and documents metadata: In this phase, the data, which is retrieved from various data sources, is prepared and filtered. Here, we distinguish three extraction types:
   - *Textual data extraction*: Firstly, we parse documents to extract textual entities (terms), this operation is called *Tokenization*. Then we filter irrelevant terms to the analysis (*stopwords*), such as: prepositions, pronouns, adjectives, adverbs, etc. They are grouped in a list named *stoplist*.
   - *Metadata extraction*: metadata is descriptive information accompanied the document to better qualify it. They do not need new techniques to identify and extract them from data sources. These operations can be performed by adapting those already known in a dedicated ETL to handle structured data. For example, metadata of emails such as: date, sender address, the recipient's address, subject, etc.
   - *Semantic descriptors extraction*: Semantic descriptors are the concepts organized into a knowledge source. In Wikipedia, for example, these concepts describe the articles subjects and they are represented by the page title. Each concept belongs to at least one category. Our goal, here, is to assign to each textual document in our collection its more related concepts and their categories.
2. Perform transformation on the extracted data: In this phase, we apply the morphsyntactic process on words that have minor difference in their forms and but same meaning or similarity, as in the case of the conjugated words. The method used to return these words to their canonical form or *Stem* is called *Stemming*, it consists of the endings words removal. Another solution, known as *lemmatization*, is to use a suffixes dictionary to extract the word root with its morphological. In our ETL-Text process, for this phase, we depend on the algorithm Porter [13] for English language and the algorithm proposed in CLEF[2] for French language.
3. Load the resulted data in the database.

3.2 Documents Enrichment based on Wikipedia

As mentioned above, we distinguish, in our ETL-Text process, two transformations types: structural, and semantic. Indeed the structural, previously detailed, does not capture the semantic relationships between terms. To consider this aspect in our approach, which is very useful for text analysis, we choice to use Wikipedia as an external knowledge source. In Wikipedia, each article describes one topic represented by one concept which we denote A_x. Also, it has a hierarchical categorization system where each concept belongs to at least one category C_y. Our goal, in this operation, is to determine the concepts A_x and their categories C_y more related to each document in our data set. To do that, we first index the Wikipedia corpus by Lucene engine[3], and then we use the stems S_i obtained for each document $d_j = (s_1, s_2, ...s_n)$, during the structural transformation phase, as query terms in Lucene engine in order to retrieve the relevant concepts to the documents. To compute the similarity between these query and concepts, we use a metric based on the Kullback-Leibler divergence (KL-divergence).

[2] http://www.clef-initiative.eu//

[3] http://lucene.apache.org/

However, the documents are already represented in a vector space of n dimensions by vectors d_j=(KL$_{1,1}$,KL$_{1,2}, ...$KL$_{1,n}$), where KL$_{i,x} \in [0,1]$ denotes the KL-divergence score of a document to the concepts A_x. These concepts are represented in the same vector space by A_x = (KL$'_{1,1}$,KL$'_{1,2}, \cdots,$ KL$'_{1,n}$), where KL$'_{x,y} \in [0,1]$ is also the KL-divergence concepts compared to their Wikipedia categories C_y. Finally, from the two matrixes we generate the corresponding documents to Wikipedia categories. This generation is performed by replacing the concepts of the first matrix by their respective categories in the second matrix. Also, the weights of these resulted categories take the KL-divergence score between the concepts and the document.

3.3 TWM: Multidimensional Text Warehouse Model

TWM extends the classical multidimensional model to consider the textual analysis processes. The constellation schema of the TWM model which is presented in Figure 1, adopts the graphic elements defined in [14]. This schema provides a uniform platform to integrate textual data and to represent their semantics. It organizes text warehouses information into a constellation set of *fact* and *dimension*. In our model, *Fact* represents a textual content of document to be analyzed according to different dimension types. A *dimension* represents the analysis axis according to which we want to observe this fact. Each dimension can be decomposed into several hierarchies constituting different granularity levels.

TWM Dimensions. The dimension can be defined as follows:

Definition 1. *Dimension: a dimension named Dim_j, is defined by $\{A_t^i, H^i\}$ where: A_t^i =$\{A_t^1, A_t^2 \cdots, A_t^n\}$ are dimensions attributes and H^i denote hierarchies defined by H^i =$\{H^1, H^2, \cdots, H^n\}$.*

In our TWM model, we distinguish three dimensions types:

1. *Classical Dimension:* It is characterized by simple attributes organized in a hierarchical way. Some of these attributes provide descriptive information. Others are used to specify how the observable data should be summarized. The classical dimension contains data from a single domain. For instance, ”*Stems*” dimension, ”*document*” dimension.
2. *Metadata Dimension:* This is informational elements describing a document. For example, we can include information explaining a scientific publication, as defined in the *Dublin Core Metadata Initiative* (*DCMI*)(title, author, language, conference, publisher, etc.). Another example is the information accompanying a tweet including the date, user, location, htag, etc.
3. *Semantic Dimension:* The semantic dimension is represented by a list of concepts which are organized in one hierarchy with several levels. They are related to a textual document and described its meaning. These concepts are obtained by applying the steps that are previously mentioned in documents enrichment phase of ETL-Text. This dimension permits to analyse documents through their semantic meaning in addition to simple keywords match. Indeed, semantic links between documents can be existed even if they have little or no common terms. Thus, the semantic dimension can be defined as:

Definition 2. *Semantic dimension: a semantic dimension denoted Dim_i^s is defined by $\{A_t^i, H_i^s\}$ where: $A_t^i = \{A_t^1, A_t^2 \cdots, A_t^n\}$ are the semantic dimension attributes and $H_i^s = \{A_x, C_y\}$ denote, respectively: the concepts of the external source knowledge defined by $A_x = (KL'_{1,1}, KL'_{1,2}, ...KL'_{1,n})$, where $KL'_{x,y} \in [0,1]$ is the KL-divergence of concepts compared to C_y; and C_y denotes a set of categories noted by $(c_1, c_2, \cdots, c_n)$ where c_i consists of a set of concepts A_x.*

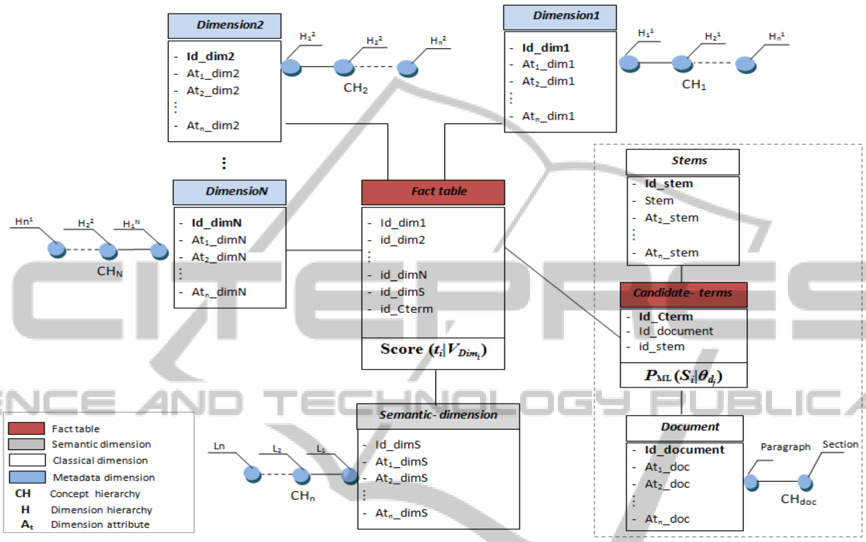

Fig. 1. Constellation Text Warehouse Model.

TWM Facts. In TWM model, the fact is defined as:

Definition 3. ***Fact:*** *a fact, denoted F, is defined by $F = \{V_{Dim_j}^i, M_l\}$ where: $V_{Dim_j}^i$ is a set of attributes values of dimension Dim_j. M_l is a set of fact measures $\{M_1, M_2, \cdots, M_n\}$.*

Definition 4. ***Candidate-terms:*** *A candidate-term $T_{(S_i,E_i)}$ of a word w_i in a text document d_j is a tuple $(S_i, P_{ML}(S_i|\theta_{d_j}))$ where S_i denotes the Stem of the word w_i in the vocabulary **V** and $P_{ML}(S_i|\theta_{d_j})$ is the maximum likelihood estimate MLE of stem S_i in the language model of document θ_{d_j}.*

The use of information processing, in particular the modeling language, has contributed to identify the facts relevant to the analysis. Indeed, they are modeled, in TWM, by two distinct and interconnected constellations. The first links the *"Stems"* dimension with the documents, whereas the second connects both *"Candidate-terms"* fact table with metadata and semantic dimensions as shown in Figure 3. The *Stems* dimension includes all terms resulted from the stemming process and thus constituting the lexicon of the document collection. The measure associated to *"Candidate-terms"* is defined by a probabilistic language model. It compute the weight of each value of "*Stems*" dimension $V_{Dim_j}^i$ in the language model of each document θ_d. Indeed, we consider each document d_j as a sample of the language and then we compute the probability of producing the values of dimensions V_{Dim_j} in this document. Also, we adapt the

KL-divergence metric to measure the facts of the second constellation. This measure generates a score between probability distributions for different dimensions values and the language model θ_q of IR query. The advantage of this proposed model that it aid to analyze and observe the textual documents through different dimensions values $V^i_{Dim_j}$ in considering IR queries composed of terms $q = (t_1 t_2 .. t_n)$.

4 Experiments

To validate our approach, we developed a platform for the storage and the analysis of textual data. The modules illustrating the ETL-Text steps have already been achieved in Java under Eclipse environment. These modules were tested on the 20 Newsgroups[4] corpus.

We processed the 20 Newsgroups corpus by applying different steps of our ETL-Text process. These steps include the data extraction, transforming and loading. To provide the candidate-terms we calculate the stems weighs by using the maximum likelihood estimation MLE. It permits to maximize the data likelihood in a document model, and compute the terms probability according to a multinational distribution defined as follows:

To evaluate the proposed ETL-Text process, we apply the precision/recall metric widely used in the performance evaluation of IR systems. To calculate these metrics we adopted on 20 Newsgroups corpus the *Lemur Toolkit* [5] which is a standard for conducting experiments in IR systems. By using the *Lemur's indexing* tool, an index of documents is created for each topic from 20 Newsgroups. To execute the experimental IR queries (vary from 01 to 07 terms) on all *candidate-terms* obtained during the ETL-Text process, we developed an information retrieval system. Then, we used TREC evaluation tool to calculate the precision and recall (the Perl script *ireval.pl* associated with the Lemur Toolkit to interpret the results of the *trec-eval*). As shown in Figure 4, the results obtained by applying ETL-Text process on 20 Newsgroups at the experimental queries show significantly improvement in the Precision/Recall metric compared to those on same corpus without applying the ETL-Text process. The improvement is precisely in the accuracy rate. It is obtained by using the three ETL-Text phase. However, at certain experimental queries, we notice that the ETL-Text process is not active, because these queries contain stop-words. In general, We can conclude that the ETL-Text does not cause any loss of textual data.

5 Conclusions

In this paper, we proposed an original approach for building a text warehouse. It uses both natural language processing techniques and information retrieval methods to process and integrate textual data in a warehouse. Our contribution is that firstly, we proposed a new ETL process appropriate for textual data called ETL-Text. Secondly, we proposed a multidimensional model for a text warehouse called TWM (Text Warehouse Model). TWM is associated with new dimensions types including: a metadata

[4] http ://people.csail.mit.edu/jrennie/20Newsgroups/

[5] http//www.lemurproject.com

dimension and a semantic dimension. Also, it has a new analysis measure adapted for text analysis based on the modeling language notion. The documents semantics are extracted by using Wikipedia as an external knowledge source. To validate our approach, we have developed a prototype composed of several processing modules that illustrate the different phases of the ETL-Text. These modules are tested on the 20 Newsgroups corpus. In perspective, we plan to define a new aggregation operators adapted to OLAP analysis on textual data.

References

1. Bentayeb, F., Maiz, N., Mahboubi, H., Favre, C., Loudcher, S., Harbi, N., Boussaid, O., Darmont, J.: Innovative Approaches for efficiently Warehousing Complex Data from the Web. In: Business Intelligence Applications and the Web : Models, Systems and Technologies. IGI BOOK (2012) 26–52
2. Lai, K.K., Yu, L., Wang, S.: Multi-agent web text mining on the grid for enterprise decision support. In: Proceedings of the international conference on Advanced Web and Network Technologies, and Applications. APWeb'06, Berlin (2006) 540–544
3. Vassiliadis, P., Simitsis, A., Skiadopoulos, S.: Conceptual modeling for etl processes. In: Proceedings of the 5th ACM international workshop on Data Warehousing and OLAP. DOLAP '02, New York, NY, USA, ACM (2002) 14–21
4. Bleyberg, M., Ganesh, K.: Dynamic multi-dimensional models for text warehouses. In: Systems, Man, and Cybernetics, 2000 IEEE International Conference on. Volume 3. (2000) 2045–2050 vol.3
5. Mothe, J., Chrisment, C., Dousset, B., Alaux, J.: Doccube: multi-dimensional visualisation and exploration of large document sets. Journal of the American Society for Information Science and Technology, JASIST, Special 54 (2003) 650659
6. Tseng, F.S.C., Chou, A.Y.H.: The concept of document warehousing for multi-dimensional modeling of textual-based business intelligence. Decis. Support Syst. 42(2) (November 2006)
7. McCabe, M.C., Lee, J., Chowdhury, A., Grossman, D., Frieder, O.: On the design and evaluation of a multi-dimensional approach to information retrieval. In: Proceedings of the 23rd annual international ACM SIGIR, New York, NY, USA (2000) 363–365
8. Ravat, F., Teste, O., Tournier, R., Zurlfluh, G.: A conceptual model for multidimensional analysis of documents. In: Proceedings of the 26th international conference on Conceptual modeling. ER'07, Berlin (2007) 550–565
9. Lin, C.X., Ding, B., Han, J., Zhu, F., Zhao, B.: Text cube: Computing ir measures for multi-dimensional text database analysis. In: In ICDM. (2008) 905–910
10. Zhang, D., Zhai, C., Han, J., Srivastava, A., Oza, N.: Topic modeling for olap on multidimensional text databases: topic cube and its applications. Stat. Anal. Data Min. 2(56) (December 2009)
11. Pérez, J.M., Berlanga, R., Aramburu, M.J., Pedersen, T.B.: A relevance-extended multi-dimensional model for a data warehouse contextualized with documents. DOLAP '05, New York, NY, USA, ACM (2005) 19–28
12. Keith, S., Kaser, O., Lemire, D.: Analyzing large collections of electronic text using olap. CoRR abs/cs/0605127 (2006)
13. Porter, M. F.: Readings in information retrieval. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1997) 313–316
14. Golfarelli, M., Maio, D., Rizzi, S.: Conceptual design of data warehouses from e/r schemes. (1998) 334–343