Contextualized Text OLAP Based on Information Retrieval

Lamia Oukid, LRDSI Laboratory, University of Blida 1, Blida, Algeria
Nadjia Benblidia, LRDSI Laboratory, University of Blida 1, Blida, Algeria
Fadila Bentayeb, ERIC Laboratory, University of Lyon 2, Lyon, France
Ounas Asfari, ERIC Laboratory, University of Lyon 2, Lyon, France
Omar Boussaid, ERIC Laboratory, University of Lyon 2, Lyon, France

ABSTRACT

Current data warehousing and On-Line Analytical Processing (OLAP) systems are not yet particularly appropriate for textual data analysis. It is therefore crucial to develop a new data model and an OLAP system to provide the necessary analyses for textual data. To achieve this objective, this paper proposes a new approach based on information retrieval (IR) techniques. Moreover, several contextual factors may significantly affect the information relevant to a decision-maker. Thus, the paper proposes to consider contextual factors in an OLAP system to provide relevant results. It provides a generalized approach for Text OLAP analysis which consists of two parts: The first one is a context-based text cube model, denoted CXT-Cube. It is characterized by several contextual dimensions. Hence, during the OLAP analysis process, CXT-Cube exploits the contextual information in order to better consider the semantics of textual data. Besides, the work associates to CXT-Cube a new text analysis measure based on an OLAP-adapted vector space model and a relevance propagation technique. The second part is an OLAP aggregation operator called ORank (OLAP-Rank) which allows to aggregate textual data in an OLAP environment while considering relevant contextual factors. To consider the user context, this paper proposes a query expansion method based on a decision-maker profile. Based on IR metrics, it evaluates the proposed aggregation operator in different cases using several data analysis queries. The evaluation shows that the precision of the system is significantly better than that of a Text OLAP system based on classical IR. This is due to the consideration of the contextual factors.

Keywords: Aggregation Operator, Context, Information Retrieval, Query Expansion, Text Cube, Text OLAP

1. INTRODUCTION

Data warehousing systems and OLAP technologies are effective to analyze data and support the decision-making, especially when data are numerical. However, most of enterprise data are presented in textual forms, such as: Reports, e-mails, etc. Unfortunately, standard tools are inadequate to analyze these textual data, in particular to extract their semantic contents. On the other hand, the aggregation of numerical data is performed by using standard aggregation functions such as sum, average, etc. However, these functions are not suitable to analyze textual data. To overcome this limitation, this paper proposes a new approach based on information retrieval (IR) techniques.
data due to their unstructured nature. Thus, it is important to propose new technologies to deal with textual data and aggregate them in an OLAP cube in order to perform effective analyses in the decision support process.

OLAP systems allow navigating through multidimensional cubes from one view to another in an interactive way. OLAP analysis allows expressing complex queries and viewing aggregated results relevant to the decision-making. In order to deal with textual data, we often use IR techniques which generally evaluate the data relevance to a query composed of simple keywords expressing the user needs. Most often, this relevance is based on term frequency in the document. The results generated by IR systems are therefore limited to extracting information from documents. Since in a Text OLAP system, we are interested in navigational analysis based on aggregation operators. For instance, if we consider a corpus of curricula vitae (CVs) for the selection of candidates during recruitment, the answer to a job offer is a large volume of documents that are hard to manage by the decision-maker. Applying OLAP analysis to these data allows decision-makers to navigate through OLAP cubes and to observe the data along several dimensions organized following different hierarchical levels. For instance, the decision-maker can observe the competencies in computer science for the year 2012 in France and then, by applying drill down operations, he can observe those for the year 2012 in Paris, Lyon, etc.

We propose to combine IR techniques with OLAP to better analyze textual data. This combination enables online analysis of large corpuses of documents and supports the navigation in text cubes for the decision-making. Therefore, IR techniques permit to extract relevant information from documents to construct textual analysis measures and to define aggregation operators for textual data. In fact, context, including user profile, task, location, time, etc., affects the information type relevant to a decision-maker and thus contextual information during the exploitation of data warehouses must be taken into account. Considering contextual factors in Text OLAP provides results that directly apply to the analysis context and generates a data model relevant to OLAP analysis. For instance, considering a contextual factor representing location allows us to express the fact that a city belongs to a country. While there have been a variety of context-aware applications, little work has been done on integrating context into data warehousing systems. Here, we consider contextual factors in Text OLAP.

In this paper, we propose a contextual text cube model called \textit{CXT-Cube}, associated with contextual dimensions. The fact is observed through a new measure for textual data analysis, based on an adapted vector space model. In order to calculate the weights of the document concepts, we propose a relevance propagation technique through a concept hierarchy. Also, we provide a new aggregation operator, denoted \textit{ORank}, to aggregate the documents during the analysis process. We propose a query expansion method based on the exploitation of the decision-maker context. Both the proposed textual analysis measure and the aggregation operator allow considering the contextual factors defined in our \textit{CXT-Cube} model. We have evaluated the results of the proposed \textit{ORank} operator by comparing between the results of three cases: Text aggregation using standard IR system; \textit{ORank} without considering the user context; \textit{ORank} considering the two defined contextual factors: \textit{Document context} and \textit{user context}. That allows quantifying the improvement provided by our system compared to the classical systems of text warehousing which do not consider the contextual factors and the relevance propagation technique.

The rest of this paper is organized as follows; section 2 overviews the related work and presents a comparative survey. Section 3 presents the context definition and its use in data warehousing systems. Section 4 presents our data model. We illustrate our OLAP aggregation approach which includes our query expansion method and our aggregation operator \textit{ORank} in section 5. Section 6 shows the evaluation and details the experimental study. Finally, we conclude this paper in section 7.
Influence of Domain and Model Properties on the Reliability Estimates’ Performance
www.igi-global.com/article/influence-domain-model-properties-reliability/37405?camid=4v1a

Business Intelligence Indicators: Types, Models and Implementation
Sandro Bimonte, Michel Schneider and Omar Boussaid (2016). *International Journal of Data Warehousing and Mining* (pp. 75-98).
www.igi-global.com/article/business-intelligence-indicators/171120?camid=4v1a