

Chapter 4

Incorporating Text OLAP in Business Intelligence

Byung-Kwon Park
Dong-A University, Korea

Il-Yeol Song
Drexel University, USA

ABSTRACT

As the amount of data grows very fast inside and outside of an enterprise, it is getting important to seamlessly analyze both data types for total business intelligence. The data can be classified into two categories: structured and unstructured. For getting total business intelligence, it is important to seamlessly analyze both of them. Especially, as most of business data are unstructured text documents, including the Web pages in Internet, we need a Text OLAP solution to perform multidimensional analysis of text documents in the same way as structured relational data. We first survey the representative works selected for demonstrating how the technologies of text mining and information retrieval can be applied for multidimensional analysis of text documents, because they are major technologies handling text data. And then, we survey the representative works selected for demonstrating how we can associate and consolidate both unstructured text documents and structured relation data for obtaining total business intelligence. Finally, we present a future business intelligence platform architecture as well as related research topics. We expect the proposed total heterogeneous business intelligence architecture, which integrates information retrieval, text mining, and information extraction technologies all together, including relational OLAP technologies, would make a better platform toward total business intelligence.

DOI: 10.4018/978-1-61350-038-5.ch004

INTRODUCTION

Using a business intelligence solution, people get business insight from the vast amount of data they manage. In general, there are three categories of data: structured, semi-structured, and unstructured. Mostly, structured data are represented in a relational form; semi-structured in XML; and unstructured in text. It is known that only 20% of the data available are structured and are stored in relational databases, while about 80% are unstructured text and are stored in various forms of documents such as reports, news articles, e-mails, and largely web pages. Thus, in order to obtain complete business intelligence, incorporating and analyzing text data are essential.

Sullivan (2001) proposed to associate relational data warehouses with a text document warehouse for that purpose. Through analyzing the former, such information as where, when, and who did how many of what can be extracted. Through analyzing the latter, such information as why it was done can be figured out. For example, if we found that the sales volume of wide television set was cut down especially in an urban area, we can understand the reason by analyzing such documents as sales report, marketing report, product catalogs, and news articles at that time. Google Finance is another example of associating numerical data with web pages. The Google Finance web page contains a graph showing the stock price changes over time, and also hyperlinks from the extreme points of the graph to the web pages describing what happened at that time.

For multidimensional analysis, in general, online analytical processing (OLAP) technology is used. OLAP helps to perform multidimensional analysis of a vast amount of data from many perspectives. In this chapter, we call the multidimensional analysis on text documents using OLAP technology as *Text OLAP*. The major technologies handling text data are Text Mining (TM), Information Retrieval (IR), and Information Extraction (IE). A TM system mines, from a docu-

ment set, such information as top keywords, text summary, text classification, and text clustering. An IR system retrieves, from a document set, the documents containing the keywords given as a user query. An IE system extracts, from a document set, the structured information according to the schema given by a user. Integration of TM and IR technologies can contribute to the TEXT OLAP and IE technology to the linkage of unstructured text documents to the structured data in a relational database. We describe each approach with some examples, and discuss the future directions.

In this chapter, Section 2 focuses on analyzing text documents only (i.e. Text OLAP). Through Text OLAP over the reports on market trends, news articles, and web pages in Internet, business people can obtain important business information such as new competitors or competitive products coming out in market and consumer demand patterns changing. Section 3 focuses on linking structured relational data and unstructured text documents for the multidimensional analysis on the consolidated information. Section 4 focuses on the future research direction toward total business intelligence platform. Finally, we conclude in Section 5.

ANALYZING TEXT DOCUMENTS

Analyzing Based on Text Mining

Using TM technology, we can perform the operations of classifying, clustering, summarizing, or extracting top keywords from a set of text documents. Through those TM operations, we can perform multi-perspective analysis such as what documents belonging to each category, what clusters being formed, what being the summary in brief, and what being the top keywords from a set of text documents. In other words, we can perform a macro-level analysis and investigate the characteristics of a whole document set from various perspectives. For more information on

23 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:

www.igi-global.com/chapter/incorporating-text-olap-business-intelligence/58412

Related Content

Fintech Impact on EU Retail Savings and Investment

Manuel Gutierrez (2018). *Disruptive Technologies for Business Development and Strategic Advantage* (pp. 99-152).

www.irma-international.org/chapter/fintech-impact-on-eu-retail-savings-and-investment/206832

The Roles of Knowledge Management and Organizational Innovation in Global Business

Kijpokin Kasemsap (2016). *Business Intelligence: Concepts, Methodologies, Tools, and Applications* (pp. 1759-1783).

www.irma-international.org/chapter/the-roles-of-knowledge-management-and-organizational-innovation-in-global-business/142700

Business Intelligence in the Music Industry Value Chain: Ensuring Sustainability in a Turbulent Business Environment

Hanne Russ, Jean-Pierre Kuilboerand Noushin Ashrafi (2016). *Business Intelligence: Concepts, Methodologies, Tools, and Applications* (pp. 1557-1571).

www.irma-international.org/chapter/business-intelligence-in-the-music-industry-value-chain/142689

The Construction of Economic Intelligence as an Academic Discipline: A Theoretical Perspective

Florian Fougy (2014). *International Journal of Business Intelligence Research* (pp. 1-16).

www.irma-international.org/article/the-construction-of-economic-intelligence-as-an-academic-discipline/120049

Digitalization of Interlocking System to Optimize Logistics in Railway Transportation

Sipho Nzamaand Arnesh Telukdarie (2020). *International Journal of Business Analytics* (pp. 24-36).

www.irma-international.org/article/digitalization-of-interlocking-system-to-optimize-logistics-in-railway-transportation/246340