# One Platform for Mining Structured and Unstructured Data: Dream or Reality?

Dina Bitton
SAP Labs
3421 Hillview ave
Palo Alto, CA 94304
(650)849-2695
Dina.bitton@sap.com

## 1. PANELISTS

Franz Faerber
SAP AG
Franz.faerber@sap.com

Laura Haas
IBM Almaden Research Center
laura@almaden.ibm.com

Jayavel Shanmugasundaram
Yahoo Labs
jaishan@yahoo-inc.com

## 2. INTRODUCTION

Although enterprises commonly utilize sophisticated data integration technology and business intelligence tools for analysis of structured data, analysis of unstructured data is a separate process and is often limited to capabilities supported by a search engine. Users have separate and vastly different interfaces for structured and unstructured data: Business Intelligence for structured data and Search for unstructured.

This panel will discuss these fundamental and controversial issues:

2.1 Information Integration: how do we "integrate" structured and unstructured data? Data warehousing versus federation; relational, multi-dimensional or other views?
2.2 Content mining: can we source, transform and analyze unstructured data alongside with structured data?
2.3 Can we retain the simplicity of the Search interface and yet provide some of the sophistication of structured data analysis tools?

*VLDB '06*, September 12–15, 2006, Seoul, Korea.
Copyright 2006 VLDB Endowment, ACM 1-59593-385-9/06/09

## 3. PREPARING UNSTRUCTURED DATA FOR REPORTING (Franz Faerber)

Most of enterprise information is stored in semi-structured or unstructured documents. Making this information available in a usable form is the goal of text analysis and text mining systems.

In this talk, we will compare the process of analyzing and mining structured data, with the process of analyzing and mining unstructured data. In particular, we will highlight the differences in the extraction step for both cases and discuss how it impacts the functionality that can be provided for analysis. In particular, we will highlight functionality related to the handling of master data.

The process of analyzing and mining structured data is well defined. It starts with and extraction process from an OLTP system, where all data that is relevant for reporting is copied to a data warehouse system. In the data warehouse system, the data is transformed and integrated.

Following this integration step, is a pre-aggregation step where the data is condensed to the granularity needed for reporting The data warehouse distinguishes between "master data" (for instance, products or customers) and operational data (for instance, sales orders) and will support different functionality on them.

In the third and last data preparation step, an index structure is generated (i.e. a star or snow-flake schema), which will allow fast reporting access to the data.

Once the data extracted from the OLTP system has gone through these 3 steps of 1) integration, 2) pre-aggregation and 3) indexing, it is then ready for OLAP. Usually, an OLAP query consists of two parts, the search on the master data (in order to select the portion of data the user wants to report on) and the real "aggregation" query.

A similar process can be defined for enabling reporting on unstructured data. The key feature here is the extraction of data (facts) and master data from documents in a "preprocessing" step. First documents are crawled (extraction step) from the document source system. The documents are first transformed (i.e. into html or xml). Then a linguistic analysis is performed. In an information extraction step, structured data is extracted from the documents. This structured data is then loaded into a data warehouse, where standard reporting functionality and master data search are supported. There are 3 types of master data, structured data, which are already available in the enterprise, master data which are extracted during an information extraction process, or the documents themselves, on which a standard text search can be performed.

The critical phase for high quality document reporting is the information extraction. Beside standard linguistic analysis such as language detection, tagging and phrase extraction, a named entity recognition system extracts entities such as persons, titles, products, locations and others. After the entity extraction step, a second extraction process takes place, where needed facts are extracted from documents. There are two different approaches to fact extraction. One uses “wrappers”, which extract data from semi-structured documents such as html tables. Wrappers can learn from examples presented by users. The second approach is a rule-based extraction system, where rules for extracting facts out of unstructured documents can be specified.

Dream or Reality? The extraction process produces good results for domains with high quality extraction rules (or wrappers). Nevertheless you can’t trust this information as you would trust true structured data", because even with high quality extraction system, the error rate is quite high. Reporting system have to take into account this uncertainty. Applying standard reporting technology to structured data derived from unstructured data may therefore produce unreliable reports.

## 4. WHAT CLASS OF APPLICATIONS DEMANDS BOTH STRUCTURED AND UNSTRUCTURED INFORMATION? (Laura Haas)

The database community has been integrating structured and unstructured information since the mid-‘90’s if not earlier. [1,2]. More recently, text analytics has emerged as a way to extract structure from text. So clearly, a single platform for analysis and mining for both types of data is conceivable and achievable in the near future. Hence the main question we need to ask is, "To what end? What class of applications demands both structured and unstructured information?" Transactional systems cannot run on unstructured data as the query semantics are too weak (e.g., Payroll cannot and will not run on the combination). Therefore the class of applications has to be those that can tolerate noise in the results. This might include enterprise search applications such as intranet search, email search, etc, but, will likely be driven by such needs as compliance (= governance + litigation + supervision) and customer call centers and/or self-service CRM. These applications will require a broad range of sources, including email, traditional databases, content management systems storing contracts and records of complaints, and other sources such as blog entries.

The next question is "What approach should we take to these problems?"

Here, we believe the most promising answer is to convert the relevant and useful parts of your unstructured data into structured data and then expose this combined information to applications. There are several important research problems to be addressed here, namely: managing the overall task of information extraction and resolution [3,7,4], addressing the problem of uncertainty which is inherent in the extraction process [6], and determining the right representation for this combination of extracted information and structured data. A key question is, "How is this information manifested to the user?" Today's business intelligence applications are restrictive and part of the reason is the lack of accessibility -- i.e., users can’t find or retrieve the information that they need, even though they have access privileges to it. To enable a larger class of users to obtain the aforementioned information, the access paradigm has to be one of simple Search and this leads to a large class of research problems such as "How do you resolve the ambiguity in keyword search ?" , an active research area [5].

## 5. SEARCHING STRUCTURED AND UNSTRUCTURED DATA (Jai Shanmugasundaram)

A unified platform for the mining of structured and unstructured data is a possible dream, an inevitable reality. I'll illustrate my position with two scenarios:

5.1 Exploratory search over databases. Databases are widely used for structured data, but there is an increasing need to support flexible search over such system for a variety of applications such as catalog management, deep web search and enterprise search. Consequently, databases are slowly morphing into a unified search/query system.

5.2 Structured search using search engines. Search engines are widely used tool for querying unstructured data, but there is a growing interest in incorporating structured information behind the "simple" search interface. I'd describe a few such examples including entity extraction and "live" advertisements. Consequently, search engines are (not so slowly) morphing into a unified search/query system.

## 6. REFERENCES

[1] Luis Gravano, ed. Special Issue on Text and Databases, IEEE Data Engineering Bulletin. Vol 24 No 4, December 2001.

[2] A. Tomasic, L. Raschid, and P. Valduriez. Scaling Heterogeneous Databases and the Design of DISCO. In Proc. ICDCS, 1996

[3] V. Chakaravarthy, H. Gupta, P. Roy, M. Mohania. Efficiently Linking Text Documents with Relevant Structured Information. To appear in Proc. VLDB, Korea, 2006.

[4] Unstructured Information Managment Architecture, UIMA, http://jedi.watson.ibm.com/uima/index.htm

[5] S. Agrawal, S. Chaudhri and G. Das, DBXPlorer: A keyword-based Search Over Relational Databases

[6] T.S. Jayram, R. Krishnamurthy, S. Raghavan, S.Vaithyanathan and H.Zhu Avatar Information Extraction System, IEEE Data Engineering Bulletin, 2006

[7] A. Doan, R. Ramakrishnan, S. Vaithyanathan, Managing Information Extraction, SIGMOD Tutorial, 2006