

BISON: Providing Business Information Analysis as a Service

Hakan Hacigumus James Rhodes Scott Spangler Jeffrey Kreulen

IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120, USA

{hakan,jjrhodes,spangles,kreulen}@us.ibm.com

ABSTRACT

In this paper, we present the architecture of a Business Information Analysis provisioning system, BISON. The system is built based on a functioning business information analysis application, Business Insights Workbench that has been implemented at the IBM Almaden Research Center and used in numerous customer settings. The service provisioning system combines two prominent domains, namely structured/unstructured data analysis and service-oriented computing, in industrial solutions. We also discuss open research problems in the area.

1. INTRODUCTION

Today's highly competitive business environment challenges enterprises to push their limits to take advantage of all available information to improve business performance and stay competitive. This challenge becomes more difficult with the ever increasing amount of data from disparate sources. Although there is a glut of data generated by various sources, such as transactional systems, business support systems, partner systems, and external agencies etc., those data mostly are not in a form that could be directly used to support critical business decision making processes. Thus, the essential problem is transforming the data into information that provides insights into the business operations and the competitiveness measures. Typically, the data are available from heterogeneous resources in varied formats. It is well known that the amount of unstructured data in the text form far surpasses the amount of available data in the structured form. Therefore, one important step in exploiting the available resources is structuring the inherently unstructured data in meaningful ways. A well-established first step in gaining understanding is to segment examples into meaningful categories. This leads to the idea of taxonomies. The taxonomies are meaningful hierarchical categorizations of documents into topics reflecting the natural relationships between the documents and their business objectives. Improving the quality of these taxonomies and reducing the overall cost required to create them is therefore an important area of research. Supervised and unsupervised

text clustering are automated approaches to creating and maintaining document taxonomies. While there will be some commonality in some industries, these natural organizations will have significant diversity across domains and organizations. Clearly there is need for systems that automate these steps to enable knowledge workers to take full-advantage of available data sources and generate business reports in a timely manner.

We have developed such a system, called Business Insights Workbench (BIW), at the IBM Almaden Research Center. BIW has been used in numerous customer setups, to solve complex problems that require understanding and analysis of very large textual data sets to fulfill the business objectives. Some examples are problem ticket analysis, market trend analysis, patent data analysis, and web information mining.

Recent advances in networking and Internet technologies have fueled the emergence of new computing models for the industries. Possibly, one the most prominent new computing paradigm among those is the service-oriented computing. The service-oriented computing allows organizations to leverage IT solutions provided by the service providers, without having to develop them on their own. It alleviates the need for organizations to purchase expensive hardware and software, deal with software upgrades, and hire professionals for administrative and maintenance tasks. Instead, the new model allows third party service providers the capability to seamlessly host services for diverse organizations and to take over these tasks. By outsourcing, organizations can concentrate on their core competencies instead of sustaining a large investment on IT infrastructures and human resources.

We envisioned the union of these two technologies; namely the business information analysis and the service-oriented computing, to deliver even greater value for the customers. The resultant computing environment offers scores of new competencies such as better integration with intra- and inter-organization computing capabilities, enhanced and custom computing environments by composing available services from different providers,

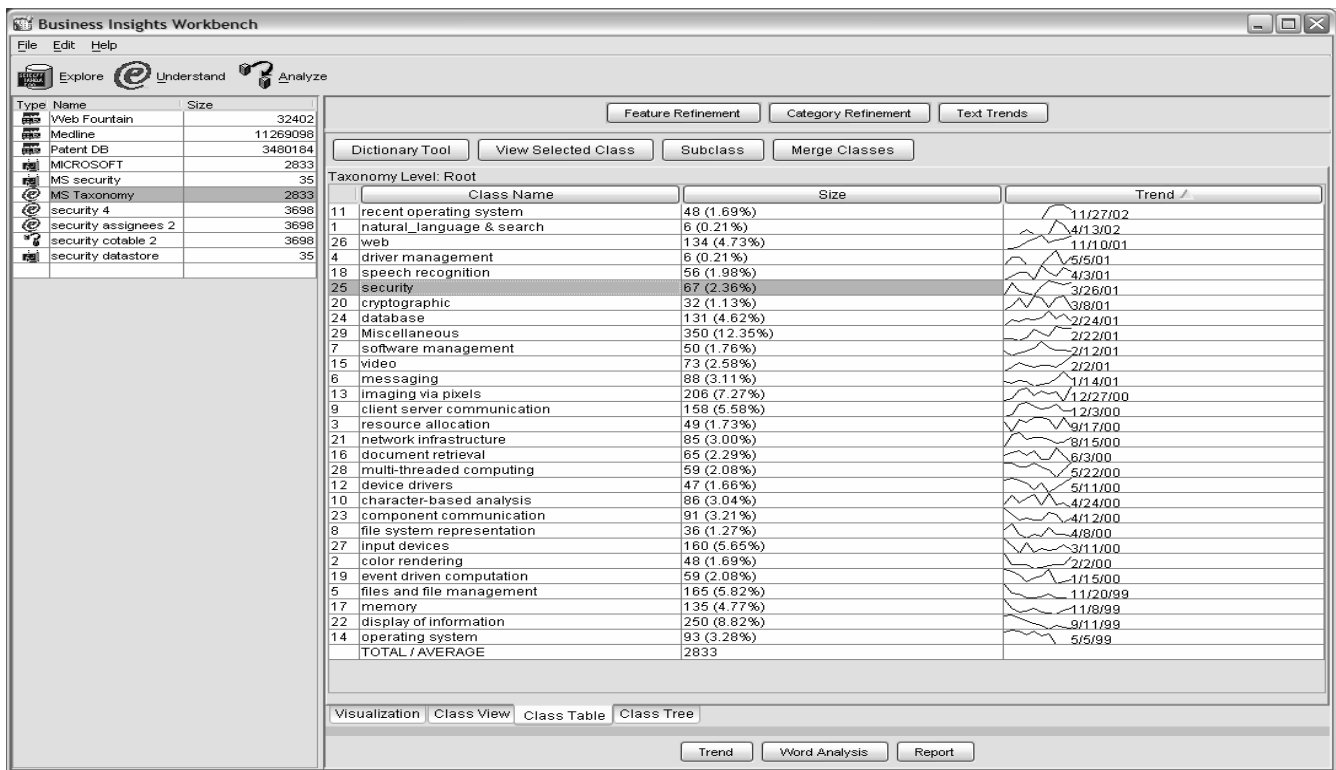


Figure 1. A Screenshot from BIW

larger customer base coverage, and solutions to clustering, scalability, extensibility, dynamic provisioning, and fault tolerance by leveraging the standard technologies and products.

The rest of the paper is organized as follows. In Section 2, we give an overview of Business Insights Workbench. Section 3 presents the BISON system. We discuss some open research problems in Section 4 and we conclude the paper in Section 5.

2. Business Insights Workbench (BIW)

In this section we give an overview of our business information analysis system, BIW. The details of the system can be found in [1] and [2]. BIW is a comprehensive data analysis application that allows a knowledge worker to learn from large collections of unstructured documents. BIW was designed in a way that applies domain expertise, through interactions with state-of-the-art unstructured data analysis algorithms and visualization, to provide a global understanding of a document collection. Most of the complexities inherent in text mining are hidden by using default behaviors, which can be modified as a user gains experience. The tool can be used to automatically categorize a large collection of text documents and then provide to a knowledge worker a broad spectrum of controls to refine the building of an arbitrarily complex hierarchical taxonomy. BIW has implemented numerous analytical, graphical, and reporting algorithms to allow a deep understanding of the concepts contained within a document collection. A sample screenshot from BIW is shown in Figure 1. The applicable tasks are grouped under

three categories, namely, Explore, Understand, and Analyze.

Explore operation performs the selection of the data of interest via queries or search from the data sources and summarizes structured values via metrics. It supports functions to create alternative ways of looking at a collection and includes data specific functions such as full text search, database drill down, data join, intersect, and subset selection.

Understand taxonomies uses text mining to extract higher level features from unstructured information. It provides tools to edit generated taxonomies visualize relationships among categories, build text models that can be applied to other data sets, and use nearest neighbor techniques to find related documents.

Analyze function examines the intersections between taxonomies and structured information. It is used to discover trends and correlations, visualize data categories over time, analyze relationships between categories in different taxonomies, and compare structured and unstructured information. Analyze function essentially combines the structured and unstructured data sources to provide a complete view.

3. Service-Oriented Approach

We have been developing a business information analysis service provisioning system, BISON, based on Business Insights Workbench that is presented above. The architecture diagram of BISON system is shown in Figure 2. We use J2EE standards as the implementation framework. Communications among the system entities, including the clients, are implemented as Web Services.

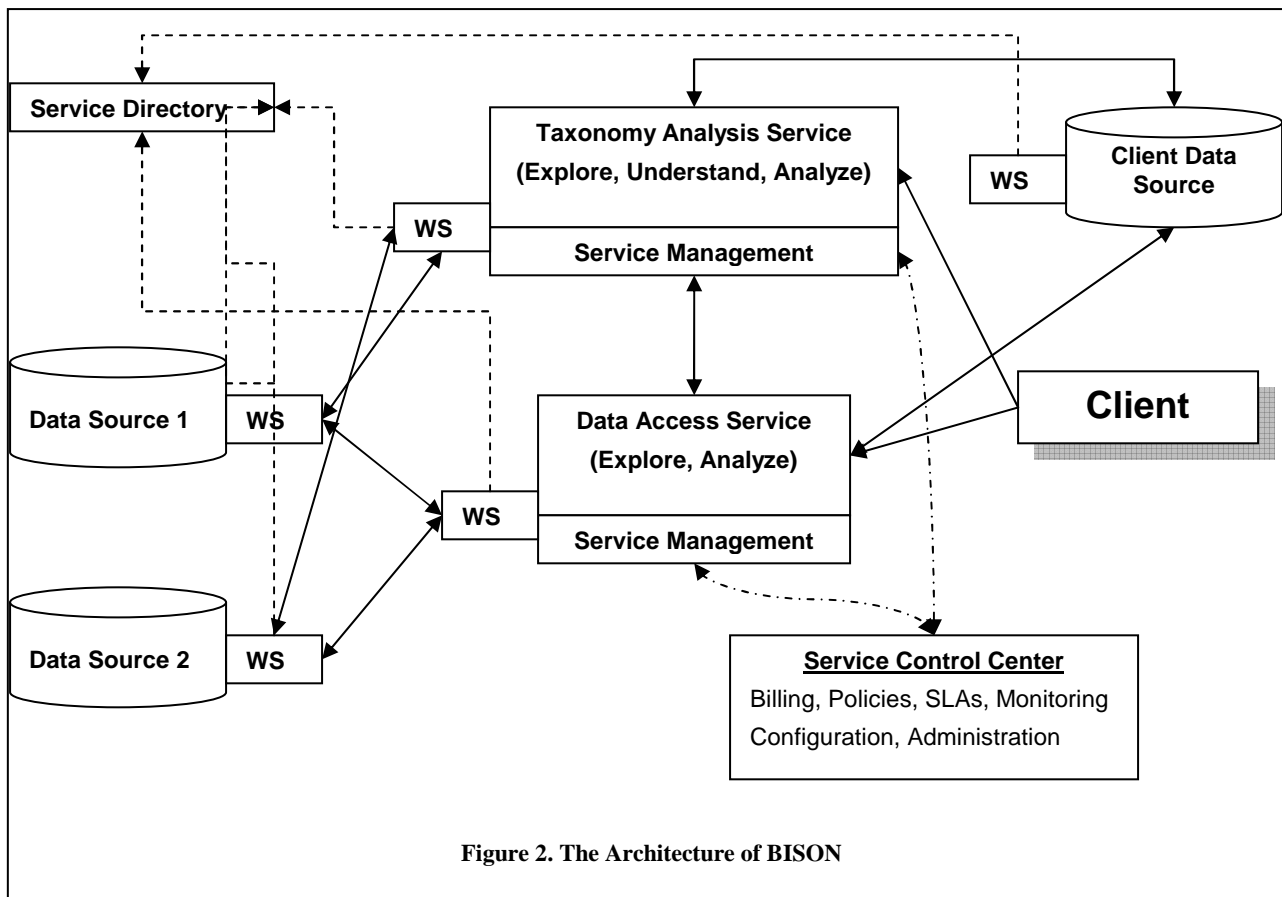


Figure 2. The Architecture of BISON

Each web service is defined by using WSDL (Web Services Definition Language) and the service discovery and registration mechanisms are driven by UDDI (Universal Description, Discovery and Integration).

Data Sources may provide structured and unstructured data and metadata information, such as full text indexes. Typical examples for text indexes are indexes created by crawling engines for increased full-text search performance. Data sources expose all the characteristics information, such as name of the source, schema of the database, and types of the data available, and data access mechanisms as web services. Data sources are registered the services directory so that they can be discovered by the other system entities, such as Data Access Service. The system architecture allows data sources dynamically join or leave the resource pool, or change the service characteristics thereby by providing flexibility and resilience.

Data Access Service is primarily responsible for querying the data sources to retrieve the data of interest and performing database oriented tasks, such as joins. It also supports analyze functions over the data sources. The data access service is instrumented to facilitate service provisioning specific tasks. These tasks include, access control, metering, service monitoring for QoS and Service Level Agreements (SLAs). This information, along with the service management information from the other systems entities, is collected, monitored, acted upon by the Service Control Center.

Taxonomy Analysis Service provides all of the enhanced analysis capabilities defined over the taxonomies provided by BIW. These capabilities include explore, understand, and analyze functions defined in the previous section. The rationale behind the separation of data access service and the taxonomy analysis service is two fold. The first one is the service flexibility. There could be certain clients, who are only interested in taxonomy specific tasks without going through the data specific tasks provided by the data access service. The assumption here is that there are certain data sources that can feed the needed data into the taxonomy analysis service in an appropriate form. The second reason is the scalability. Depending on the customer need, the data analysis service or the taxonomy analysis service may require excessive system resources. One solution to that problem is cloning the demanding service over clusters while keeping the other services at the same configuration level to achieve the scalability and the lower level of redundancy. Similar to the data access service, the taxonomy analysis service is also instrumented for the service management tasks.

The Service Control Center oversees the service provisioning tasks in the system. It monitors and aggregates the system management data provided by the individual services. This aggregate information is used for billing, configuration management, dynamic provisioning, SLA management, policy enforcement. The service control center also provides the system administrators with the user

interfaces and the tools to perform the system administration tasks.

The client is a consumer of the services. Typically there are two different types of client in our system. First, the individual users those connect to the service directly by using a web browser. These clients use the web interfaces to navigate through the service functions and interact with the system. The second category of the clients is the service providers. They use our services to provide additional services to their end-users.

The Client Data Source is a temporary data source that is created to maintain the intermediate results between the service calls for the client.

4. Current Research and Technology Issues

Combining business information analysis applications and the service-oriented computing presents certain research and technology issues. We describe some those problems in this section.

4.1 Metering

Metering the service usage is a critical problem for any service offering environment. Metering is important for the cost analysis of the service delivery and generating the billing information for the customers. The metering is also important from the system management perspective as it provides valuable information for the performance of the system components and helps problem determination by identifying the system components that are negatively impact the overall system performance.

The real challenge is defining the metrics for metering. Although there are typical measures, such as CPU, bandwidth, and storage usage, we need to devise application specific metrics that would be meaningful for the information management functions.

4.2 Monitoring

Along with the overall service and system monitoring, we need application level monitoring techniques. Application level monitoring helps problem determination for the problem spots in the service delivery. It also enables application specific billing. Given that each application has its own characteristics, developing a common monitoring methodology that could be applied to all of the possible service components is a challenging problem.

4.3 Data Security and Privacy

The security of the data sources should be protected against the malicious users. This includes preventing unauthorized access to the data sources and preventing data disclosure to the parties who are not entitled to use the particular parts of the data. In addition to that, in certain cases the clients are concerned about revealing their identity for their particular interest on certain data sets. As a more challenging problem, if the client actually owns a data source and just uses the information analysis services, then the possible disclosure of confidential information to the service provider becomes an issue.

4.4 Business Integration

We provide business information analysis services in a service provisioning environment. In most of the cases this information analysis is a part of the higher level business processes at the client organizations. Hence, we need to develop methodologies that allow us to model our system processes and tie them into higher level business processes.

5. Conclusions

In this paper we have presented the architecture of a Business Information Analysis provisioning system, BISON. The system is built based on a functioning business information analysis application, Business Insights Workbench. The service provisioning system combines two very important domains, structured/unstructured data analysis and service-oriented computing, in industrial solutions. We have also described some open research problems that need to be solved for more effective service delivery and better integration with the systems in different settings.

6. REFERENCES

- [1] Scott Spangler, Jeffrey T. Kreulen: Interactive methods for taxonomy editing and validation. CIKM 2002: 665-668
- [2] William F. Cody, Jeffrey T. Kreulen, Vikas Krishna, W. Scott Spangler: The integration of business intelligence and knowledge management. IBM Systems Journal 41(4): 697-713 (2002)