BI on Data and Content Together:
What will you do with the derived insights?

Mukesh Mohania
IBM Research -- India
Agenda

- Information Integration – Definition and architectures
- Existing solutions
- Context Oriented Information Integration
  - SCORE approach
  - EROCS approach
- Integrating Audio Streams with Structured Data
- Data Analytics and BI Applications
  - Improving Semantic Search
  - Preventing Customer Attrition
  - Preventing Information Leakage from Text Documents
  - Improving Cross/Up-Sale
- Conclusions
Market Insights: Information Management Challenges

60% + of CEOs: Need to do a better job capturing and understanding information rapidly in order to make swift business decisions.

Only 1/3rd of CFOs believe that the information is easy to use, tailored, cost effective or integrated.

85% of information is unstructured.

42% of transactions are still paper-based.

30-50% of application design time is spent on copy management.

30% of people’s time: searching for relevant information.

The average billion dollar company:
48 disparate financial systems
2.7 ERP systems

79% of companies have more than two repositories and 25% have more than 15

40% of IT budgets may be spent on integration.

Sources: IBM & Industry Studies, Customer Interviews, Forrester
Existing Solutions
II Architecture: A Data Warehousing Approach

Users

Analysis Tools  Report Writer  Spreadsheet  Data Mining Tools

MOLAP/ROLAP Server

MDDB
Multidimensional Database

DATA WAREHOUSE

Metadata

Integrator

Wrapper

Wrapper

Wrapper

IMS  RDBMS  Others

Data Mart

Data Mart

OLAP Server

Data Warehouse Management System

Front End Tools

Software Tools for Data Select, Transform, Integrate, Cleanse, and Copy

Operational Data Systems
II Architecture: Virtualization Layer Approach

- Leave the data in the sources.
- When a query comes in:
  - Determine the relevant sources to the query
  - Break down the query into sub-queries for the sources.
  - Get the answers from the sources, and combine them appropriately.
- Data is fresh. Approach scalable
- Issues:
  - Relating Sources & Mediator
  - Reformulating the query
  - Efficient planning & execution

Garlic [IBM], Hermes[UMD]; Tsimmis, InfoMaster[Stanford]; DISCO[INRIA]; Information Manifold [AT&T]; SIMS/Ariadne[USC]; Emerac/Havasu[ASU]
Structured and Unstructured Information Integration: A Brief Background on Existing Solutions

Existing solutions can be classified in terms of the query paradigm used:

- **Keyword Query Based Solutions (DB2 ESE, DbXplorer/BANKS)**
  - Relational data exposed to search engine as virtual text documents
  - Query both structured and unstructured information using keywords

- **SQL Query Based Solutions (SQL LIKE predicate, DB2 NetSearch Extender)**
  - Text data exposed to relational engine as virtual tables with text columns
  - Query both structured and unstructured information using SQL
    - Provide SQL primitives to search text in table columns using a set of keywords
Facts about business data

- Companies serve many millions of people and several business partners.
- Information about the business entities/customers is stored in several data sources.
- Two types of information (structured and unstructured including audio) available about business entities, but unfortunately stored in silos.
- Intelligent business decisions are made from integrated information.
- Unstructured data not used for Business Intelligence.

YOUR Goals:
- Break down information silos through the use of active, event-based, and context-oriented information integration technologies.
- Make unstructured content (including audio) accessible for analysis w.r.t structured data.
- Derive meaningful insights from unstructured text -- Provide holistic view of business Information based on search interface and/or a structured query.
- Provide a single 360 degree view of customer data (structured and unstructured) for intelligent business decisions.
- Perform off-line and (near) real-time analytics over structured and unstructured data to discover contextually relevant actionable business insights.
Facts about business data

• **Customer Data**
  - Structured data: Customer profiles, transactions, products, etc
  - Unstructured data:
    - Voice from call center, Emails
    - Agent notes
      - Call center summaries, Email resolution summary
    - Web interaction
      - Pages visited on website, forms, searches, dialog

• **Customers communications contain important information**
  - What they are looking for
  - What problems they are having
  - Their satisfaction with products or services
  - Information they need
  - How the agent responded

• **Speech/text analytics can identify high level concepts and events from words or phrases in customer communication**
  - Concepts - sentiment, service quality, brand/competitor awareness, product interest, problems ..
  - Events - Threat to quit, product/service comparisons, complaint, request for product or information, etc.
Know your Customers Better

• To realize full potential of customer we have to answer certain questions about him/her
  • **Increase share of wallet**: What does a person need? What are his product affinities? What is his opinion? What hinders him from doing more business with us?
  • **Cross-sell/Up-sell**: What products sell best? (Cognos) Will he buy it? When is the right time to sell so that his likelihood of buying is high?
  • **Product Extension**: What features does he like? How can I improve his product experience?
  • **Reduce Churn**: Who is likely to churn? (SPSS) Why is he churning?
  • **Reduce cost to serve**: What is his problem? How can I solve it efficiently?

• Structured Data Analysis and Surveys are employed to answer these questions
  • Structured data analysis can answer only some of these questions
  • Unstructured data can answer more questions which cannot be answered by structured data.
SCORE Approach –

Associating text Documents with Structured Query Results
SCORE Overview

SELECT name, max(price) - min(price)
FROM stocks
GROUP BY name
ORDER BY 2
FETCH FIRST 3 ROWS ONLY

“Get the 3 companies with max price variation”

SELECT name, max(price) - min(price)
FROM stocks
GROUP BY name
ORDER BY 2
FETCH FIRST 3 ROWS ONLY

“Doctype:Patents”

“IBM” “ORCL” “MSFT”
“Database” “Software”
“Large-Cap”

“Doctype:Patents”

CIKM 2005 – Best Paper
Derive Context from a Broader Neighborhood

- Extend SCORE to derive the context from tables that refer to the rows in the query result (backward neighborhood)
EROCS Approach –

Associating Relevant Structured Data with Text Documents
Linkage Discovery (EROCS): Efficiently Linking Diverse Data

- Exploit partial information contained in a document to automatically identify and link relevant structured data

Main Idea
- View the structured data as a set of pre-defined “entities”
- Identify the entities from this set that best match the document, and also find embeddings of the identified entities in the document

Value Proposition
- Metadata Extraction (linking documents with structured data)
- Enhance semantic Search by exploiting this Metadata
- Enable BI across Structured and Unstructured Data
- Providing more metadata and richer text search in CM UIMA Annotator

VLDB 2006, SIGMOD 2007, PODS 2007, ICDE 2008 (Demo)
Example

Find the transaction that best matches the context

I am <Name> John Smith </Name>

.....

..... bought a
Company>Sony</Company>

<product> DVD player </product>

.....

from <Company>JK Electronics</Company> ......

Additional “sidebar” information available as a result of the annotation

<table>
<thead>
<tr>
<th>CustId</th>
<th>StoreId</th>
<th>Payment</th>
<th>Discount</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>A756K9</td>
<td>S8976</td>
<td>Card (AMEX)</td>
<td>Promo# 1236</td>
<td>NOREFND</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CustId</th>
<th>Name</th>
<th>Loyalty</th>
<th>Club</th>
<th>Addr</th>
</tr>
</thead>
<tbody>
<tr>
<td>A756K9</td>
<td>John W Smith</td>
<td>Platinum</td>
<td>IBM</td>
<td>Chicago, IL</td>
</tr>
</tbody>
</table>
Scenario for Linkage Discovery

• To link customer data residing in different operational silos (Call Center, CRM database, marketing databases)

• To link Voice-of customer data (emails, call records) stored in content repositories with CRM/DWH

• To link warranty claim data with product information or customer data

• To link medical reports (pathological, nurse notes, etc) with patient medical records, disease/symptoms database, etc

• To link the invoices (unstructured) for customers or products with sales statistics from data warehouses (structured).

• To link the customer contracts, survey, complaints (unstructured), with customer profile or customer sales history (structured).

• To link software contracts (unstructured) with proactive expiration warnings against a structured database.

• To link complaints (products with most customer complaints) with the structured database for quantities of those products present in distribution centers.
1. **Entity/sentiment extraction from text.**

2. **Integration of key annotation technologies in common stack; Tooling to configure, test, deploy extraction annotators**
Data Analytics and BI Applications
CallAssist: Integrating Real-time Audio with Databases

CallAssist
- A novel system for linking audio conversations with relevant structured data automatically in real-time,
- Suggests informative queries to narrow down the context

Value Proposition
**Cost Reduction:** the end-to-end time taken to support a customer will be reduced.
- Reduced call escalation rates
- Lower average time for call resolution
- Reduction in agent training costs
  -- Agents need not be have in-depth understanding of business processes
  -- System will prompt agent with informative queries that can narrow context

**Potential Sales:** By suggesting opportunities for up-selling/cross-selling products and services to customer.

**Potential Applications**
- Automate Q&A for Call Center
- Automated Solution Provider
- Dynamic Learning/Self-help Program
- Integrated Trouble Ticket cum Diagnosis Generator
- Extend Customer Data Integration tools to support real-time audio

VLDB 2007 (Demo), SIGMOD 2008
<table>
<thead>
<tr>
<th>Tuple Rank</th>
<th>Model</th>
<th>Location</th>
<th>Promotion</th>
<th>Type</th>
<th>Rate</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>pontiac grand</td>
<td>Stamford Connecticut</td>
<td>NA</td>
<td>mid size</td>
<td>20</td>
<td>4-door 2 adult</td>
</tr>
<tr>
<td>2</td>
<td>pontiac grand</td>
<td>Stamford Connecticut</td>
<td>AAA</td>
<td>mid size</td>
<td>25</td>
<td>4-door 2 adult</td>
</tr>
<tr>
<td>3</td>
<td>pontiac grand</td>
<td>Dallas Fort Worth</td>
<td>Club</td>
<td>mid size</td>
<td>20</td>
<td>4-door 2 adult</td>
</tr>
<tr>
<td>4</td>
<td>Chevy Aveo or Similar</td>
<td>Dallas Fort Worth</td>
<td>AAA, COSTCO</td>
<td>mid size</td>
<td>18</td>
<td>4-door, 4 people, 2 bags, unlimited mileage</td>
</tr>
<tr>
<td>5</td>
<td>Chevy Aveo or Similar</td>
<td>Dallas Love field</td>
<td>AAA, COSTCO</td>
<td>mid size</td>
<td>18</td>
<td>4-door, 4 people, 2 bags, unlimited mileage</td>
</tr>
</tbody>
</table>

**Suggested Questions**

*What promotion you prefer? Expected Gain=2.32*

*What rate you prefer? Expected Gain=2.32*

*What promotion you prefer? Expected Gain=1.92*

**Suggested Promotions**

<table>
<thead>
<tr>
<th>Promotion From</th>
<th>To</th>
<th>Maximum Discount(%)</th>
<th>Promo Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>pontiac grand</td>
<td>Chevy Aveo</td>
<td>10.0</td>
<td>Upsell</td>
</tr>
<tr>
<td>pontiac grand</td>
<td>Chevy Impala</td>
<td>30.0</td>
<td>Upsell</td>
</tr>
<tr>
<td>pontiac grand</td>
<td>Kia sportage</td>
<td>30.0</td>
<td>Upsell</td>
</tr>
<tr>
<td>mid size</td>
<td>Standard</td>
<td>20.0</td>
<td>Upsell</td>
</tr>
<tr>
<td>mid size</td>
<td>minivan</td>
<td>10.0</td>
<td>Upsell</td>
</tr>
<tr>
<td>mid size</td>
<td>Fullsize SUV</td>
<td>40.0</td>
<td>Upsell</td>
</tr>
</tbody>
</table>
I have an account in your bank in TX (# 0214-452). I am currently facing problems in accessing my net-banking account. Whenever I try to login, I get a message "account locked". I cannot go to the branch to reset my passwords as I am currently traveling and outside the US. Can you please reset my password to my old one? 

Mail # 1

I am indeed privilege to get your response to quickly. However, I have not got any service out of your net banking on the given dates. I was following up the wrongful debits to my account on account of the car loan which you have refunded now. That being the case, there is no justification for you to charge me the extra $40.

Mail # 2

Find Emails from Privilege Customers complaining about net-banking

Regular (non-linked) Search

LD Search

Linkage Discovery

CRM
Improved BI spanning both on Content and Data

Show me the top 4 pain points of my most privileged customers from North region who have reduced their balance by more than 50% in the last quarter.

- Traditional BI systems cannot answer this type of hybrid query requiring manual analysis
- OmniFind Analytics Edition with Linkage Discovery can.

ICDE 2008 (Demo)
Banking Industry Insights Due to Linkage Discovery

• What are the 10 categories that receive the maximum (unhappy) complaints from the most profitable (Band 5) customers?

• Find the City with the most complaints on products sold by Branch XYZ.

• Among the customers who have been complaining about Service charges, how many hold more than one product with the bank? How many of them have been delinquent in the past?

• List customers who have complained about “out-station cheque service charges” by their region/demographics/profitability.

• Analyze text correspondence, identify pain points of the most profitable customers to reduce loyal customer churn.
Customer Relationship Management: Churn Prediction

- Attrition Prevention
  - Use SCORE to deduce common features of the set of customers who have cancelled their credit card
  - Prioritize customer retention campaign for remaining customers exhibiting these characteristics

“List all customers cancelled their credit cards in last 3 months”

“CardType: Gold”,
“Category: High Interest”
“City: Mumbai”,
“No_of_Complaints: >2”,
“Sentiments: Unhappy”,
“Band: 5”,

“CardType: Silver”,
“Category: Late Payment”
“City: New Delhi”,
“No_of_Complaints: >3”,
“Sentiments: Unhappy”,
“Band: 4”,

RDBMS/WII

Cognos

Content Repository

Linkage Discovery

Other customers having the same context
Example: Complaint Routing

Complaints → Linkage Discovery → Rule Based Router

Transactions data

Identified transaction with relevant context *(Payment details, promotion utilized, customer profile, products)*

Linkage Discovery can also help proactively identify the *minimal* additional information needed from the customer to identify the relevant transaction.

PODS 2006 (Entity Completion)
Sanitization involves removing sensitive information from a document.

Problem Statement

- Given a document D and a parameter K, delete a minimum number of terms so that the remaining document T is K-safe.

- K-safety: A set of terms T is K-safe, if for any entity e, at least K other entities contain \( T \cap C(e) \) in their context.
Problem Scenario 1: Document Sanitization Based on Access Control Policies

- Dynamically sanitize a document for a specified user, based on his/her access privileges defined on a structured database.
  - Sensitive information hidden from the user in the database should also be removed from the document.

![Diagram showing document sanitization process]

- Input Document: John Ralphson of IRL is suffering from cancer .........
- Document Sanitizer
- Sanitized Document: John Ralphson of IRL is suffering from cancer

- DB + Access Control Policies
- Doctor
Problem Scenario 2: Document Sanitization for Securing Entities

- Database contains a set of entities
- Each entity $e$ has context $C(e)$: a set of terms associated with $e$.
- Sanitization $\rightarrow$ Hide information from a document so that the entity mentioned in the document cannot be identified.
- Identification: Happens by searching the database and matching terms.
Conclusions

- Information Integration has become widely popular – However a lot needs to be done

- Context Oriented Information Integration
  - SCORE: Automatically finds relevant unstructured data for a SQL Query
  - EROCS: Finds links between structured and unstructured data

- Data Integration Applications
Thank You!

Work done at IRL. This is a joint work with Himanshu, Prasan and Mukesh.