Making Sense of Unstructured Data

Thomas Hampp
OmniFind & UIMA Development Architect
September, 2007
Agenda

Why deal with unstructured data?
  Limitations of current Search & Business Intelligence (BI)

Why use a framework?
  The complexities of unstructured analytics

How does this help with business problems?
  A sample solution
Agenda

Why deal with unstructured data?
- Limitations of current Search & Business Intelligence (BI)

Why use a framework?
- The complexities of unstructured analytics

How does this help with business problems?
- A sample solution
Unstructured versus Structured Information: What does it mean?

Structured Information: Semantics of information captured in DB schema

<table>
<thead>
<tr>
<th>Name</th>
<th>Occupation</th>
<th>Organization</th>
<th>Age</th>
<th>Office Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jones</td>
<td>Engineer</td>
<td>IBM</td>
<td>29</td>
<td>San Francisco</td>
</tr>
<tr>
<td>Smith</td>
<td>Journalist</td>
<td>AP</td>
<td>32</td>
<td>Boston</td>
</tr>
<tr>
<td>Brown</td>
<td>CEO</td>
<td>Textract</td>
<td>42</td>
<td>New York</td>
</tr>
</tbody>
</table>

Unstructured Information: Semantics inherent in usage and context

The associated press reported today that, Ms Brown, CEO of Textract Corp, had been recently spotted at the Summit meeting in Zurich,............At, 42, Ms. Brown, is the youngest CEO at the Summit,....
Information is isolated in multiple silos - structured & unstructured

Customer Service  Marketing  Finance  Sales  Mfg  Legal  HR  R&D

Search and Analytics capabilities vary across content sources with inconsistent interfaces
Vast majority of content is unstructured, limiting its use

Lines between structured and unstructured content are blurring
Unstructured Data Challenges

Unstructured Data poses many challenges
   Governance
   Compliance
   Records Management/Retention Management
   Storage/Archival
   Search/Retrieval
   ...

Various products -from IBM and other companies- address these challenges

This presentation will focus on how to bring semantics & structure into unstructured data in a principled fashion to make better use of unstructured data in business processes
Search Today: Sometimes your word is used too often

Searching for “neon” finds signs and cars
Search Today: Sometimes your word isn’t used at all

Search for “Pat phone” finds nothing

Pat Lam's homepage » Pat's Phone Company
Pat’s Phone Company. Check out how to make cheap international calls from the UK with no accounts to open and no phone cards! ...
www.pat-lam.com/?page_id=52 - 7k - Cached - Similar pages

University News Briefs
Language Technologies Experts Get NSF Grant to Help PAT's Phone-Based Information Systems. Scientists at the Language Technologies Institute (LTI) in the ...
www.cmu.edu/cmnews/021121/021121_briefs.html - 22k - Cached - Similar pages

PDF Harmony: A Synchronization Framework For Tree-Structured Data
File Format: PDF/Adobe Acrobat - View as HTML
Business Intelligence Today

<table>
<thead>
<tr>
<th>Repair Date</th>
<th>Customer complaint</th>
<th>Technician Comment</th>
<th>Repair Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>7/28/04</td>
<td>CUST SMELLED BURNING UNDER THE HOOD</td>
<td>RUSTY CONNECTION TO FUEL PUMP RELAY. WIRING HARNESS CORRODED</td>
<td>$1,405</td>
</tr>
<tr>
<td>7/27/04</td>
<td>CUSTOMER STATES GET GAS ODOR INSIDE VEHICLE AFTER EVERY FILL UP:</td>
<td>VERIFIED CONCERN PERFORM FUEL SYSTEM LEAK TEST FOUND FUEL TANK LEAKING REPLACE FUEL TANK ASSEMBLY:</td>
<td></td>
</tr>
<tr>
<td>7/22/04</td>
<td>CAR JUST STALLED AND WOULD NOT START</td>
<td>FOUND SHORT IN HEADLIGHT WIRING</td>
<td></td>
</tr>
<tr>
<td>7/19/04</td>
<td>CAR WILL NOT START AFTER DRIVING IN THE SNOW. CUST. BRINGING IN CAR FOR SECOND TIME</td>
<td>NEED TO CHANGE HARNESS ROUTING</td>
<td></td>
</tr>
<tr>
<td>7/19/04</td>
<td>CUST REPORTS HEADLIGHTS FAIL</td>
<td>REPLACED FUEL PUMP AND FUEL FILTER. CORROSION WIRING HARNESS FOUND WHEN I REMOVED THE WIRED HARNESS COVER. RECOMMEND REPLACING WIRING HARNESS BUT WE DO NOT HAVE ACCESS TO PART. SHOULD RESCHEDULE</td>
<td></td>
</tr>
<tr>
<td>7/17/04</td>
<td>C/S THAT THE VEHICLE WOULD NOT START AFTER FUELING:</td>
<td>REPAIRED FAULTY HARNESS</td>
<td></td>
</tr>
<tr>
<td>7/16/04</td>
<td>CUSTS SITS WHILE DRIVING VEHICLE LOST ACCELERATION AND THE SEIGHT CAME AND NOW IT HAS NO ACCELERATION SEE HISTORY</td>
<td>TECH VERIFIED THE COMPLAINT I CHECKED THE FUSES AND FOUND THEM TO GOOD...</td>
<td>$1,162</td>
</tr>
<tr>
<td>7/15/04</td>
<td>C/S CAR STALLS</td>
<td>TECH REPLACED FUEL PUMP</td>
<td>$2,083</td>
</tr>
</tbody>
</table>

Traditional BI does not allow to examine these text comments in a repair report database to make the information available for reports and alerts. People must manually read through all the technician comments.
Examples Of Other Tasks on Unstructured Data Which Are Hard To Solve Today

- Find **insurance accident reports** mentioning a Neon car sometime between June and August, referencing a license plate
- Find all **pictures** which show animals
- Find **SEC filings** for a specific company referencing unsecured debt over $300 million with an interest rate of more than 6%
- Determine if there are any **patent applications** for drugs with specific protein combinations
- Get an alert if your company name is mentioned Arabic **news videos**
- Find **blogs & web pages** in which the products and key managers or your company are talked about in a positive or negative way. Track how that evolves. Get alerts about important changes.
- Analyze **customer interaction records** (call center transcriptions, emails, comments) and find out what the must urgent complaints and problems and which customers are most like to switch to the competition
What is necessary to go beyond the today's state of the art

Distinguish between different meanings of the same term

  smart (car) vs. smart (adjective) vs. SMART (as an acronym for “Self-Monitoring Analysis Reporting Technology”)

  Provide an understanding of domain specific terms (auto parts, chemicals, etc.)

Understand concepts and identify facts (entity relationships)

  People, places and organizations or parts, problems, conditions and actions

  Is the person "located at" a place or are they "talking about" it?

  Did the condition "cause" the problem or "prevent" it?
Text Analysis Can “Understand” Text...

.. But it’s a complex multi-stage process...

A was driven by Neon, driven by Timothy Higgins.
The discovery of semantics in unstructured information & their explicit representation in structure through automated **Text (or multi-modal) Analysis**

The Associated Press reported today that, **Ms. Brown**, CEO of Textract Corp., had been recently spotted at the Summit meeting in **Zurich**. At 42, **Ms. Brown** is the youngest **CEO** at the Summit,…

**Analysis Engines** (pattern recognition techniques) discover Semantics & build in structure allowing us to process text as we would structured data.
Live Demo

Live Annotation of Web Content
News (LA Times)

Without annotations
With various general purpose annotation
Analytics Can Identify a Wide Variety Of Entities

Death toll rises among Afghan civilians, report says

Insurgents' attacks on nonmilitary targets are war crimes, U.S. human rights group charges.

By Shafiq Ahmad Saidi and Laura King, Special to The Times
April 17, 2007

KABUL, AFGHANISTAN — Insurgents in Afghanistan killed nearly 700 civilians last year, the largest annual number in more than five years of fighting since the toppling of the Taliban regime, New York-based Human Rights Watch said in a report issued Monday.

The group blamed NATO-led coalition forces for 230 civilian deaths in 2006, though estimates by officials and human rights groups put the number much higher.

Human Rights Watch said the attacks on civilian militiamen and other insurgents should be considered war crimes because they "intentionally targeted civilian objects that served no military purpose, including school bazaars."

Suicide bombings rise

Nearly 500 of the deaths at insurgents' hands came in suicide bombings, which are increasingly the Taliban's weapon of choice in the face of North Atlantic Treaty Organization firepower. Suicide attacks last year more than quadrupled from the previous year.

In the latest such attack, a bomber blew himself up Monday outside a police post in the normally calm northern city of Kunduz, killing 11 police officers and injuring nearly three dozen people. A day earlier, a similar attack in the eastern city of Khowst killed seven police officers and a civilian.

Human Rights Watch noted that even militants' strikes aimed at NATO and Afghan forces, such as those on Sunday and Monday, were also likely to hurt or kill civilians.

"Insurgent attacks are often so indiscriminate that Afghan civilians end up as the main victims," Joanne MacInerney, the organization's terrorism and counterterrorism director, said in Kabul, the Afghan capital.

The report included graphic testimony from civilians injured in suicide bombings, with many expressing disbelief that they would be targeted by fellow Afghans. One was a 3-year-old girl in Kabul whose abdomen was ripped open in a March 2006 suicide bombing.
Analytics Bridge the Unstructured & Structured Worlds

Unstructured Information

- High-Value
- Most Current
- Fastest Growing (80% of Corporate Data)

...BUT...

- Buried in Huge Volumes (Noise)
- Implicit Semantics
- Inefficient Search

Structured Information

- Explicit Semantics
- Efficient Search
- Focused Content

...BUT...

- Slow Growing
- Narrow Coverage
- Less Current/Relevant

Discover Relevant Semantics → Build into Structure

- Docs, Emails, Phone Calls, Reports
- Topics, Entities, Relationships
- People, Places, Org, Times, Events
- Customer Opinions, Products, Problems
- Threats, Chemicals, Drugs, Drug Interactions...

Text and Multi-Modal Analytics
Agenda

Why deal with unstructured data?
   Limitations of current Search & Business Intelligence (BI)

Why use a framework?
   The complexities of unstructured analytics

How does this help with business problems?
   A sample solution
UIMA to Tame Combination Complexity of Text Analytics: A new standard for content processing and analysis

- Text Analysis is a complex, multi-step process
- It varies a lot between languages and domains (e.g. automotive vs. medical)
- No single vendor will be able to provide all optimal components for all steps in languages and domains
- UIMA is an open, standardized, plug-and-play framework that allows to integrated technology from different vendors
- Provides an SDK for building and composing text analytics
- Enables development of new and re-use of existing components for analysis

Text Analysis Modules – aka “Annotators”
IBM has submitted the Unstructured Information Management Architecture (UIMA) specification to the Organization for the Advancement of Structured Information Standards (OASIS)

Members of the Technical Committee

EMC, IBM, NStein, TEMIS, Thomson, …
CMU, Univ. Manchester, Univ. Tokyo, Mayo Clinic, …
SAIC, Pacific Northwest National Lab, Army Intelligence and Information Warfare Directorate, Los Alamos National Laboratory

In addition, the UIMA source code has been contributed to the Apache Software Foundation and an Apache incubator project has been established to foster collaborative, consensus-based development of new software based on UIMA
Over 200 IBM researchers across 6 WW labs with UIM as a Strategic Focus. Many expert, small groups independently building analysis technologies. Needed a common architecture and framework to facilitate reuse and technology transfer.
Live Demo
Marvel Overview
UIMA in IBM

In active use since 2002

**IBM Internal Component Repository**
- 80+ Analysis Components and
- 23+ UIMA-based systems/solutions
  - E.g., Deep and Shallow Parsing, Categorization, Summarization, Semantic Class Detection, POS, English/Chinese/Japanese NE, Classifier Trainers, Machine Translation, Video and Speech Analytics, BioInformatics…

**Products**
- Lotus Workplace, WebSphere Portal Server, **OmniFind**, (more to come)

**Some Research Projects Building on UIMA**
- Open-Domain Question Answering (ARDA/AQUAINT)
- Life Sciences/BiolInformatics (Joint Program with Mayo Clinic, Sloan Kettering Cancer Center)
- Search and Categorization (IBM websites)
- Machine Translation (DARPA)
- Multi-Lingual/Multi-Modal Search (DARPA/TALES)
- Automatic Content Extraction (DARPA/ACE)
- Knowledge Integration and Knowledgebase Population (ARDA(DTO)/NIMD)
- Video Analysis (ARDA(DTO)/VACE, Marvel)
- Standard Analysis Component Plug-in Architecture in Streaming Analytic project
- Customer Relationship Management
UIMA Highlights Outside IBM

Original Software Development Kit and Source Code on alphaWorks
Source Code contributed to open source community Q1 2006: http://uima-framework.sourceforge.net/
10.000+ Downloads as of April 2008

IBM Sponsored “UIMA Innovation Grants”
>100 Faculty Applications in 2006
From Universities specializing in language, video analysis &/or knowledge management

Mayo Clinic – an early adopter

Government Related
UIMA Working Group driven by DARPA and IBM advanced framework
Initial group of academics & researchers to evaluate, provide feedback and advance framework
Stanford, Carnegie Mellon, Columbia, UMASS, BBN, MITRE, SAIC (Object Sciences)

DARPA/ITPO GALE Project (Speech-to-text, Translation, Distillation)
Three Major Contractors + ~25 Universities adopting UIMA as integrating platform

National Labs and DHS (Threat Assessment)

Universities
Many including: Stanford, UMASS, Columbia, CMU, University of Sheffield (GATE Interoperability)
Taught in classes at CMU’s Language and Technology Institute
UIMA Component Repository at CMU: http://uima.lti.cs.cmu.edu/resources.html
Live Demo: Crosslingual Multimodal Search (TALES)
Live Demo: Semantic Search

myEmailSearch

99.9% Emails Indexed

Looking for John's phone number? Try "john phone". Or the Word document from Mark? Try "mark doc". You may also try "IBM uni", "mike address", "receipt date", and more.

© 2006 IBM
Agenda

Why deal with unstructured data?
  Limitations of current Search & Business Intelligence (BI)

Why use a framework?
  The complexities of unstructured analytics

How does this help with business problems?
  A sample solution
Online Media Analysis: Provide insight into what people are saying. Manage reputation, public image and brand awareness

**Benefits**

- Monitor a wide variety of sites for voiced opinions of customers and opinion makers
- Detect trends and swings in opinion more quickly and be able to react earlier
- Drill down to find the root of public image
- Discover new topic related to company/brand

**Functionality**

- Access blogs, news sites, feeds, competitor sites, customer information sites etc.
- See overall tone, news types, hot topics, brands, products, most vocal sources
- See development of each topic over time
- Drill down on each discovered topics
Text analysis annotators can identify positive and negative statements in texts.
Other annotators can extract the names car brands, models, countries, topics and other entities.
Delivering it all in an user interface that combines the familiar ease of use of search with the power and insight of analytics.
Search with the full power of IBM OmniFind

- Use overview of active drill-up and drill down conditions to navigate
- Show overview based on Search results
- Show Hot Terms computed dynamically for the search results
- Show automatically detected concepts specific to application domain
- Click on any item to drill down and refine results
- Detects tonality in documents and marks them
## Save Investigations

<table>
<thead>
<tr>
<th>Query</th>
<th>Additional query terms</th>
<th>Tonality</th>
<th>Media Relevance</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>fuel economy</td>
<td></td>
<td>positive (81.0%)</td>
<td>Other a… (67.7%)</td>
<td>Feb 21, 2007</td>
</tr>
<tr>
<td></td>
<td></td>
<td>neutral (8.8%)</td>
<td>Fuel econ… (12.9%)</td>
<td>10:39 AM</td>
</tr>
<tr>
<td>hybrid car</td>
<td></td>
<td>positive (81.4%)</td>
<td>Other a… (58.4%)</td>
<td>Feb 21, 2007</td>
</tr>
<tr>
<td></td>
<td></td>
<td>neutral (8.6%)</td>
<td>Hybrid car (3.6%)</td>
<td>10:35 AM</td>
</tr>
<tr>
<td>hybrid car</td>
<td>CarBrand: Toyota</td>
<td>positive (88.6%)</td>
<td>Other a… (88.8%)</td>
<td>Feb 21, 2007</td>
</tr>
<tr>
<td></td>
<td></td>
<td>neutral (4.4%)</td>
<td>Hybrid car (1.2%)</td>
<td>10:38 AM</td>
</tr>
</tbody>
</table>
Summary

Unstructured analytics can bring additional semantics & structure to unstructured data
This kind of analytics are best implemented in an open, standardized, principled and flexible framework like UIMA
In the world of search the additional structure can be used to drive new search paradigms
In the world of business intelligence the additional structure can be used to feed improved mining and reporting
There are many business problems that can benefit from this emerging technology