Analyzing Social Media Data: Cubes, DAGs, Hierarchical Correlations

Umeshwar Dayal
Malu Castellanos, Chetan Gupta, Song Wang, Manolo Garcia-Solaco,
Meichun Hsu, Ming Hao, Riddhiman Ghosh
HP Labs
Palo Alto, California, USA

ER Conference, Florence, 17 October 2012
Mainstream business intelligence (BI) has traditionally focused on enterprise transactional data, and is batch oriented, often with long extract-transform-load latencies…
Emergence of an inflection point…

Sensors, mobile devices, real time events, web, and unstructured data have the promise of transforming the way we manage our customers, resources, environments, health…

• Analytics over big data ("Volume")
• Analytics over both stored and streaming data, and delivery of near real-time analytics wherever needed ("Velocity")
• Analytics over both structured and unstructured data, and enterprise-internal and open Web / social media data ("Variety")
Decisions at *any* time scale
Social media data: unstructured textual data
Complex New Social Landscape

• Over 500 Million users on Facebook and 170 Million on Twitter
• Facebook became bigger than Google in 2010 with 8.9% of all web traffic
• People spend over 700 billion minutes per month on Facebook
• Over 150 million active users currently accessing Facebook through their mobile devices
• Twitter generates 10% of global social media hits to websites
• Over 30 billion pieces of content (web links, news stories, blog posts, notes, photo albums, etc.) shared each month
• Over 50% of Twitter users follow brands!

Source: Randall Munrow/XKCD
Humans as sensors: tapping social media for insights in real time

**Listen**

What are consumers saying about products and services on blogs, reviews, twitter, Facebook,

**Understand**

Text analytics and natural-language processing tools used to determine sentiment, opinions, intentions, and link it to business data

**Act**

Modify business processes to react quickly to social media signals

- What are the main problems our customers are facing?
- What are people saying about our products?
- How will my operations be affected by recent news events?
Fusing Unstructured Data into the BI Pipeline via Information Extraction

- Extract structured information from text
  - Wrap information extraction algorithms inside operators
  - Plug operators into the ETL pipeline
  - Now, this data is available for querying, OLAP, reporting, analysis
  - Stream processing to meet decision latency requirements

What are the main problems that our customers are facing?
What are people saying about our products?
How will my operations be affected by recent news events?

contracts, call logs, reviews, reports, comments, email, etc
Exploiting Unstructured Data – Overall Pipeline

Two Phase Approach

- Offline Phase and domain specific
  - Specify entities, document categories
  - Model learning

- Online Phase
  - Classification and application of previously learned models
  - Analytics such as aggregation, querying, correlation

Specification → Model learning → Model application → Correlations, Cubes, etc.

Information Extraction → Analytics
Unstructured Textual Data

- Different kinds of textual data, e.g., news stories, reviews, forums, call records, …
- Extract concepts from textual data: named entities, events, topics, attributes, …
  - Each kind of data may require specialized information extraction techniques
- Extracted concepts typically form hierarchies
- Typically, there’s metadata associated with the date: author, time, location, …
- The metadata may also form hierarchies
- Data arrive at different rates: slow streams (stored data) versus fast streams
- Extract “measures” for analysis
  - Depend on the application, e.g., volume, sentiment scores, severity, popularity, …
- We want to perform multidimensional analysis on the measures along the metadata dimensions and the concepts
Analysis over Unstructured Textual Data

- Extract “measures” for analysis
  - Depend on the application, e.g., volume, sentiment polarity/score, severity, popularity, relevance, ...
- Associate measures with concepts and metadata
- We want to perform OLAP-style multidimensional analysis on the measures, e.g., aggregate sentiment for a product by time and geography
- We also want to do correlations, e.g., compare sentiments of “similar” concepts, detect events that “relate” to my business, ...

- We discuss two applications:
  - Sentiment Analysis: Use Sentiment Cube
  - Situational Awareness over Contracts: Use Contract Cube

© Copyright 2012 Hewlett-Packard Development Company, L.P.
The information contained herein is subject to change without notice.
HP Confidential
sentiment analysis
Customer Intelligence:
Sentiment Analysis

Analysis questions:

- **Summary** of the customers' opinions on a specific product.

- **What are the most severe problems of a product** according to the customers' opinion?

- **Are there subgroups of people with similar opinions?**

- **How do customers' opinions change over time?**

- **Who are the key influencers?**

- **How do opinions spread through the social network?**

- **Does the opinion correlate with some other data that we have?**

- **Can we use opinions to predict some business outcome?**
Typical Product Review Site

The most helpful favorable review
28 of 38 people found the following review helpful:

🌟🌟🌟🌟🌟 So far I couldn't be happier
Ok, I confess, I just opened the box 10 minutes ago. But I already printed several documents and I couldn't be happier!

The cartridges are easy to install, the printer looks great and the product has a "smart" feeling to it.

Ohh, just one thing. For reasons that escape me, printers do not come with the USB cable these days. It's not just this model or just HP...

Read the full review
Published 12 months ago by Steven Burns

See more 5 star, 4 star reviews

26 of 29 people found the following review helpful:

🌟🌟🌟🌟🌟 Only good for very, very casual use
This machine is an ink guzzler. I can't print out my term papers without having to go through at least two cartridges. Unless you only expect to use it for very light use (print out the occasional e-mail, directions), I'd say look for something else, unless you want to keep a ton of cartridges in stock all the time.

Published 9 months ago by Dabby Cool

See more 3 star, 2 star, 1 star reviews

28 of 38 people found the following review helpful:

🌟🌟🌟🌟🌟 So far I couldn't be happier, September 3, 2007
Ok, I confess, I just opened the box 10 minutes ago. But I already printed several documents and I couldn't be happier!

The cartridges are easy to install, the printer looks great and the product has a "smart" feeling to it.

Ohh, just one thing. For reasons that escape me, printers do not come with the USB cable these days. It's not just this model or just HP. The Laser printer at the office came without it too.

So don't forget to get yourself a USB cable before you try to install it. I will probably upgrade to Vista within the next three months and this model works perfectly with Vista, which is a big plus compared to older models.

Anyway, I will update this review later, after a few weeks of printing and scanning.

Most Helpful First | Newest First
Sentiment analysis

I feel obligated to counter the bad reviews.
This printer is just fine.
I don't know what people are complaining about regarding the software but it installed seamlessly and is intuitive in its operation.
Even though I am dissatisfied with the paper tray altogether I am happy that I bought this wonderful printer.

Traditional Sentiment Analysis
Assign polarity (positive <-> negative) to complete review 😊👍
Attributed Sentiment Analysis

I feel obligated to counter the bad reviews.
This printer is just fine.
I don't know what people are complaining about regarding the software but it installed seamlessly and is intuitive in its operation. Even though I am dissatisfied with the paper tray altogether I am happy that I bought this wonderful printer.

Attributes

Sentiment Analysis
General exploration of text polarity (positive <-> negative)

Attributed Sentiment Analysis
Product feature- or attribute-based analysis
Extract Concepts (Attributes) and Sentiment Measures

I feel obligated to counter the bad reviews.
This printer is just fine.
I don’t know what people are complaining about regarding the software but it installed seamlessly and is intuitive in its operation.
Even though I am dissatisfied with the paper tray altogether I am happy that I bought this wonderful printer.

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>cartridge</td>
<td></td>
</tr>
<tr>
<td>paper tray</td>
<td>-1</td>
</tr>
<tr>
<td>price</td>
<td>0</td>
</tr>
<tr>
<td>printer</td>
<td>+1</td>
</tr>
<tr>
<td>scanner</td>
<td>0</td>
</tr>
<tr>
<td>software</td>
<td>+1</td>
</tr>
</tbody>
</table>
Sentiment Analysis Pipeline

- Source event
  - Filter
    - Sentence detection
      - Tokenization
        - POS
          - Attribute extraction
            - Attribute & sentiment annotation
              - Dynamic feedback from downstream state
                - Sentiment scoring
                  - Downstream ops (OLAP, correlation, prediction, etc)
                    - Rules, Lexicons
                      - Rules, Lexicons
                        - Sentiment extraction
                          - Rules, Lexicons
                            - Rules, Lexicons
                              - Source event
Multi-Dimensional Sentiment Analysis
Attributes/Concepts As A Dimension For OLAP
Sentiment Score as Measure
sentiment cube
Analysis Exploiting Hierarchies over Extracted Data

Sentiment Cube

- Aggregate sentiment scores via OLAP-like operations
- Non-traditional hierarchies
- Extended semantics for rollup operations
- New operations
Sentiment Cube

OLAP analysis over sentiment scores

- Sentiment Table schema:
  - `<doc-id; {metadata-dimensions}; concept-dimension; sentiment-score`

- OLAP like operations over metadata and concept dimensions
  - For example, get the average sentiment for a concept such as a “laptop” from location “X”

- Extend the scope of traditional operators to a subset of related concepts rather than just a single concept
  - Compare the sentiment score of a given HP laptop model with a given Apple model on “similar” features.

- Metadata dimensions are just “regular” dimensions, and the semantics of rollup and other OLAP operations carry over

- But …
Concept dimension is different from regular dimensions (1)

1. The hierarchy may contain root to leaf paths of different lengths.
   - Some OLAP systems insist on inserting dummy nodes to make all paths of equal length
   - Other systems support “ragged hierarchies”, but rollup operations are messy

2. Existence of “dangling tuples”: measure values associated with concepts that are not leaves.
   - Not supported by typical OLAP systems
   - Have to insert dummy children and additional fact tuples to associate the measure values
Concept dimension is different from regular dimensions (2)

3. No class labels: all concept instances are of the same class; identify levels by Level Number
   - Some OLAP systems support “parent-child hierarchies”, but operations are messy (have to resort to writing code).

4. Hierarchy may be a DAG, not a strict tree
   - Some OLAP systems support multiple parents, but to satisfy the “summarization” property for rollup, have to assign all of the child’s measure to one parent, or allocate it in some way across the multiple child-parent relationships (have to write even more messy code).
Cube Construction & Roll Up Operations

Differences from traditional roll ups for concept hierarchies

- The base cuboid is not the same as fact table.
- The results at the higher level cannot necessarily be obtained from lower level cuboids. For going up the hierarchy we need to check:
  - (i) If there are any leaf nodes at the higher level of abstraction (ii) the presence of dangling tuples.

Similarly, for roll-up by removing a dimension

<table>
<thead>
<tr>
<th>Location</th>
<th>Feature</th>
<th>Count</th>
<th>Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Austin</td>
<td>Screen</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Palo Alto</td>
<td>Screen</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>Austin</td>
<td>Keyboard</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>Dallas</td>
<td>Install</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>Palo Alto</td>
<td>Quality</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Palo Alto</td>
<td>Keyboard</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>San Jose</td>
<td>Keyboard</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Location</th>
<th>Feature</th>
<th>Count</th>
<th>Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Austin</td>
<td>Laptop</td>
<td>3</td>
<td>-1</td>
</tr>
<tr>
<td>Palo Alto</td>
<td>Laptop</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Dallas</td>
<td>Printer</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>Palo Alto</td>
<td>Printer</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>San Jose</td>
<td>Laptop</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Dallas</td>
<td>Laptop</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Location</th>
<th>Count</th>
<th>Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Austin</td>
<td>3</td>
<td>-1</td>
</tr>
<tr>
<td>Dallas</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Palo Alto</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>San Jose</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
For Level DAGs

- To obtain a correct answer for roll-up, we could:
  - Pre-compute all the paths from node u to the leaf nodes in the subtree rooted at u
  - From these paths obtain the set of all the individual nodes.

- Cumbersome for large hierarchies. Alternatively:
  - The correct answer for a roll-up is given by the subtree that has the maximum number of nodes.

- If a roll-up is performed by removing dimensions other than the concept dimension, then same as previous slide.
- If the roll-up is performed by removing the concept dimension, then compute the cuboid as $< \text{dim-levels, All} >$
New equivalence operators

Obtain new equivalence classes with new equivalence operators

- **Upward path equality**
  - Concept $b$ is considered to be equivalent to concept $a$, if $b$ is one of the $k^{th}$ ancestors of $a$.

- **Upward subtree equality**
  - Concept $b$ is considered to be equivalent to concept $a$, if $b$ exists in a subtree rooted at the $k^{th}$ ancestor of $a$.

- **Downward all-path equality**
  - Concept $b$ is considered to be equivalent to concept $a$, if $b$ is upto a $k^{th}$ descendant of $a$. 

![Diagram](image)
Rich set of queries can be expressed with these operators embedded in SQL

- Compute the average sentiment for concept “printer” from all attributes up to a certain depth: use downward all-path equality
  
  \[
  \text{SELECT } \text{avg}(\text{measure}) \text{ FROM Table } T1 \text{ WHERE } \text{“printer” } \equiv_{a}^{k} T1.\text{concept}
  \]

- Compare the sentiment of a particular concept with those of its ancestors up to a certain distance: use upward path equality
  
  \[
  \text{WHERE “printer”}= T1.\text{concept AND T2.}\text{concept } \equiv_{p}^{k} T1.\text{concept}
  \]

- Compare the sentiment of a particular concept with those of concepts that share a common ancestor and are up to a specified depth: use upward subtree equality
  
  \[
  \text{WHERE “printer” } = T1.\text{concept AND T2.}\text{concept } \equiv_{s}^{k} T1.\text{concept AND level(T2) } \leq \text{level (T1)+1}
  \]

- The user does not need to know the concept hierarchy
beyond sentiment analysis
Intention Capture & Understanding

Provide analysis and reports on customer intentions and future plans from multiple sources:

- Web forums and discussion groups (Disney Mom’s Panel, Yahoo Answers, …)
- Mostly Q&A type forums with explicit or implicit intentions

- Automatically extract intention phrases and other information when available

We **plan to attend** the MNSSHP on **Oct 22**. There are **5 adults** plus **2 kids (5 and 2)** If the grandparents don’t care about rides, will they need to buy the special tickets or can they just stay with their regular admission and not partake in the activities.
situational awareness for contracts
Contract Situational Awareness

- An enterprise must be able to respond to events that might affect existing business partnerships
  - Political instability or natural disasters in a country: what contracts do we have with suppliers or distributors based in this country?
  - Significant fluctuations in currency values: what contracts do we have that are denominated in this currency?
  - Changes in commercial law: how does the change in commercial law affect our risk in each contract?
  - Mergers and acquisitions: what contracts do we have with the parties involved in the merger?
- Correlate information in stored repositories or slow streams (e.g., contracts) with information derived from fast moving streams (e.g., RSS news feeds)
Example

A typhoon in the Pacific region where an enterprise has its main suppliers.

- Key capabilities:
  - Extraction
  - Correlation
  - Near real-time
Analysis Exploiting Hierarchies over Extracted Data

Situational awareness over contracts

- Two streams of unstructured data
- Traditional hierarchies
- Correlate two streams by “approximate joins”
Situational Awareness Data Flow
Example

A typhoon in the Pacific region where an enterprise has its main suppliers.
Correlation

- We formalize the situational awareness problem as one of correlating two streams, $S1$ and $S2$ (of unstructured data) arriving at different rates $r1$ and $r2$
- Performing correlation via an equi-join is overly restrictive.
- We need “approximate joins”
  - For example, while analyzing contracts, a natural disaster in a higher granularity location (e.g., a region) can affect contracts of suppliers located in cities in that region
  - $\text{Select } S1.a, S2.b \text{ From } S1, S2 \text{ Where } d_c(S1.c,S2.c) < k \text{ And } S1.c=”Pacific”, \text{ And } t_1 < S1.timestamp < t_2, \text{ and } t_3 < S2.timestamp < t_4$
- Introduce a new “concept-distance” operator
- Need to figure out equivalence
Implementing hierarchical similarity joins on streams using HNTs

- Find correlation between streams of unstructured data
- Use of Hierarchical Neighborhood Trees (HNTs)
  - Data structures to compute similarity between categorical variables (i.e., extracted entities) in two streams
  - Hierarchical similarity-based join
  - Scale based neighborhood relationship
    - Measures the level in an HNT of the closest common ancestor between two entities
  - A set of HNTs forms a cube
    - Used to compute the similarity of contracts and interesting news articles
- Slow stream contracts are inserted in the contract cube
  - As a news item streams in, its neighbors (i.e., contracts that the news item affect) are found using the contract cube
Operations Using HNTs

News article tuple $i$

... || (Pacific, typhoon, Sept 8) || ...

News article tuple $i$

... || (Tetra, fraud, integrated circuits) || ...

Contract Cube

Geography

Organization

Products

Philippines

Tetra

motherboard

Acme

IC

$d=1$

$d=2$

$d=3$

... || (Acme, Philippines, peso, motherboard) || ...

Contract tuple $k$

$$S_{i,k} = \prod_{j=1}^{d} s_{i,k,j}$$
conclusion
Live Analytics Platform
Deep analytics and data management over massive multi-source streaming and stored data, across many time scales
Summary

- Integrate structured and unstructured, stored and streaming data into a common processing pipeline
- Use a combination of information extraction, multi-dimensional (OLAP-style) analysis over hierarchies, and downstream analytics (e.g., correlation)
- Defined extended semantics and operations on hierarchies
- Many challenges remain:
  - More accurate extraction algorithms
  - Additional semantics of concept hierarchies
  - Additional analytics tasks: concept and influence propagation, prediction
  - Optimization
Thank You!