Web 2.0 Business Analytics: Data Quality through Data Quantity?

Alexander Löser
http://user.cs.tu-berlin.de/~aloeser

Database Systems and Information Management
Technische Universität Berlin
My Vision: Business Analytics on Text Data

Internet
Intranet
Forum, Wiki

Text Extraction
Semantic Integration
Load

Analysis Schema

Bankrupt Companies
- Company Locations
- Company Employees
- Company Customers

“list of aircrafts”

all about “Google”

Are we open Pandora’s Box?
Analytics over Web Data / User Generated Content?
TB of text data / GB of new data each day?
Structured Queries, Search Queries?
At “Google-Speed”?

© Chair for Database Systems and Information Management
Acknowledgements

Source: http://koala-vm-3.cis.cs.tu-berlin.de/?team=#team0809
Agenda

► Business Analytics and Web Search: Why now?

► Business Analytics over Text Data: Algebra and Execution Model

► What’s next? A bit of Dreaming.
OLAP Style Reporting

Measures by Product

<table>
<thead>
<tr>
<th>Product</th>
<th>Sales Amount</th>
<th>Total Product Cost</th>
<th>Margin</th>
<th>Margin %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bike</td>
<td>$80,861,515</td>
<td>$57,217,546</td>
<td>$13,544,169</td>
<td>16%</td>
</tr>
<tr>
<td>Mountain Bike</td>
<td>$32,324,691</td>
<td>$25,073,246</td>
<td>$7,251,445</td>
<td>18%</td>
</tr>
<tr>
<td>Road Bike</td>
<td>$35,417,136</td>
<td>$30,855,315</td>
<td>$4,861,821</td>
<td>12%</td>
</tr>
<tr>
<td>Touring Bike</td>
<td>$13,119,089</td>
<td>$11,689,085</td>
<td>$1,430,903</td>
<td>10%</td>
</tr>
</tbody>
</table>

KPI's for 2004

<table>
<thead>
<tr>
<th>KPIs</th>
<th>Value</th>
<th>Goal</th>
<th>Status</th>
<th>Trend</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orders</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Reseller Sales Amount by Date, Fiscal over Customer Geography and Geography for Calendar Year 2004

United States
- Alabama: $963,435.90
- Arizona: $3,564,715.24
- California: $1,020,454.90
- Colorado: $437,655.07
- Connecticut: $351,324.00 (min:max)
- Florida: $875,204.97
- Georgia: $351,324.00

Chart

- August 2003: $542,685
- October 2003: $1,062,047.06
- December 2003: $973,645.62
- February 2004: $1,197,256.74
- April 2004: $1,082,523.56
- June 2004: $938,258.09

Legend:
- Mountain Bikes
- Road Bikes
- Touring Bikes
To what extent will each of the following technologies be a Top 5 priority for you in 2009?

<table>
<thead>
<tr>
<th>Technology</th>
<th>Rank 2009</th>
<th>Rank 2008</th>
<th>Rank 2007</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business Intelligence Applications</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Enterprise Applications (ERP, SCM, and CRM)</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Server and Storage Technologies (Virtualization)</td>
<td>3</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Legacy Application Modernization</td>
<td>4</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Collaboration Technologies</td>
<td>5</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>Networking Voice, and Data Communications (VoIP)</td>
<td>6</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>Technical Infrastructure</td>
<td>7</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>Security Technologies</td>
<td>8</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Service-Oriented Technologies (SOA and SOBA)</td>
<td>9</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>Document Management</td>
<td>10</td>
<td>9</td>
<td>9</td>
</tr>
</tbody>
</table>

Source: 2009 Gartner Executive Programs CIO Survey, January 10, 2009
What are CIOs missing?

- Better/more information 22.9%
- Faster/quick retrieval 14.3%
- Accurate/updated data 11.4%
- Consistent platform 8.6%
- Better integration 8.6%
- Standardization 8.6%
- Other single mentions 40.0%

Estimated growth of content
- Published content: 3-4 Gb/day
- Professional web content: ~2 Gb/day
- User generated content: 5-10 Gb/day
- Private text content: ~2 Tb/day (200x more)

Source: Business Intelligence Survey, IDC, May, 2007
Search Engines: Why do we search?

Query Frequency

Source: Greg Pass, Abdur Chowdhury, Cayley Torgeson: A picture of search. Infoscale 2006: 1
Search Engines: Why do we search?

- Undirected (all about)
  [cheerleading in georgia]

- Selection & Projection
  [baby names]

- Aggregation (AVG, MIN, MAX, ...)
  [average emergency room nurse wages in virginia]

- Joins
  [accidents and emergencies 2005]

- Unions
  [a list of best ipods and mp3 players]

Source: AOL Crawl 2006, 20 Mio Queries (Google served AOL during that time)
Search Engines: Why do we search?

Navigational: Go to a web page I already have in mind.

Informational:

- Undirected: Learn anything about topic x.
- Locate: Find out if and where to obtain a resource.
- Aggregate/List: Intend a table as result.

Directed: Learn something specific about topic x.

Advice: Solve a problem/collect instructions.

Resource: Download/Obtain/Interact/Buy ... real world resource.

58%

Actual Query Intention: All about “Airbus”

SELECT documents FROM “The Web”
WHERE Company LIKE “AIRBUS”
AND document.text LIKE “Merger”, “Management Change”, “Customer” …
Simple Human Strategy

1. Read Top 100 News, e.g., BBC.com
2. Read BBC Articles
3. Filter out Information
4. Aggregate in PowerPoint
5. Send to your Manager
... is there someone reading the documents for me ....

Alternative 1: Ask your intern (12 €/hour)
... is there someone reading the documents for me ....

Alternative 1 : Ask your intern (12€/hour)

Alternative 2: Ask GOOLAP.INFO (free)

Foreign Ownership

Airbus and the Perils of Foreign Ownership Thursday, October 26, 2006  Crafty kids don't just share their dessert—they use it to make friends, build ...
Goolap.info: Answering Undirected Queries

Airbus

More Companies: East India Company White TransGas Development LLC

Companycompetitor (5)
Boeing
Emirates Airl...

Companylocation (1)
France
Toulouse

Companyproduct (2)
A380
Airbus A380

Companyreorganization (2)
announced
planned

47 spans
3 spans
6 spans
1 span
2 spans
1 span
1 span
1 span

Finance - December 9th

It make sense for the airline management teams and perhaps negotiated better pricing from Airbus and Boeing, enabling it to... and Emirates Airline. Deflation So we're in a recession? Duh...

ahmedismailbrahim.wordpress.com

Foreign Ownership

Airbus and the Perils of Foreign Ownership Thursday, October 26, 2006 Crafty kids don't just share their dessert—hey use it to make friends, build...

2006-10-26 | actionsbyt.wordpress.com
### Foreign Ownership

... aircraft manufacturing subsidiary of European-owned eads aerospace technology group, has been plagued with financial problems—problems so severe, the company recently announced the possibility of closing up to five airplane manufacturing complexes and laying off up to 10,000 employees. As can be imagined, the negative news has caused eads shares to cascade. Adding even more intense downward pressure on eads' share price was the ...

2006-10-25 | actionsbyt.wordpress.com
DEMO: GOOLAP.INFO
Many more Relevant Projects

- KnowItAll/ TextRunner (UW Seattle)
- IntelligenceInWikipedia (UW Seattle)
- DBpedia (U Leipzig & FU Berlin)
- SeerSuite (PennState)
- Cimple/ DBlife (U Wisconsin & Yahoo)
- Avatar/ System T (IBM Almaden)
- Libra (MS Research Beijing)
- SQoUT (Columbia U)
- WikipediaEntities (Yahoo Barcelona)
- ExpertFinding (U Amsterdam)
- Expertise Finding (U Twente)
- GOOLAP (TU Berlin)
- ... and more
- and G, Y, MS, SAP/BOBJ, IBM for products, locations, ...

Selected overviews in: ACM SIGMOD Record 37(4), Dec 2008
Business Analytics over Text Data
Data Flows and Operators

- Web Page Identification
- Fact Extraction
- Fact Verification
- Load/Index & Transfer
- Data Analysis

(Distributed) Text Query Processor

Existing OLAP & SQL Tools
**Data Flows and Operators**

- **Web Page Identification**
  - Copy pages that contain **relevant** textual facts
  - Without “owning” pages
Gather all pages

- Example: German Web ~ 1.4 Billion Pages (25 TB, text only)
- 80legs.com (web crawling service) charges $2.00 /1 Million pages
- $2800 to get a copy of the German Web
- $3500 for 25TB of storage
- +$0.03/hour (80legs) ... $0.10 (EC2) for distributed indexing and text scanning
- + monthly power bill + network costs + 1 ... 2 engineers

Efficiency Problem?

- Established Search Engines no longer have monopole on data
- Still too costly for individual researcher or small group
Generating Queries to Retrieve Promising Documents

a. Get document sample with “likely positive” examples. Label sample documents using information extraction system as “oracle.”
b. Generalize attribute values to types.
c. Shallow analysis on candidate sentences.
d. Filter words likely describing the relationship by lexico-syntactic patterns.
e. Identify discriminative words and phrases.
f. Generate keyword query
Generating Queries to Retrieve Promising Documents

<table>
<thead>
<tr>
<th>Relation D(B,C)</th>
<th>Generated Keyword Phrases</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Company</strong>, <strong>Technology</strong></td>
<td>programming, leading, to support, based on, supports, to develop, combined, leader in</td>
</tr>
<tr>
<td><strong>Company</strong>, <strong>Location</strong></td>
<td>ltd, based in, located in, headquarters in, company of, business in, firm in</td>
</tr>
<tr>
<td><strong>Company</strong>, <strong>Employees</strong></td>
<td>employs, employs about, employs around, become</td>
</tr>
<tr>
<td><strong>Company</strong>, <strong>Product</strong></td>
<td>launched, launches, manufactured by, operating, released, to produce</td>
</tr>
</tbody>
</table>

Existing OLAP & SQL Tools

(Distributed) Text Query Processor

Load/Index & Transfer

Fact Extraction

Fact Verification

Web Page Identification

FOCUSSED SCAN
For years, Microsoft Corporation CEO Bill Gates was against open source. But today he appears to have changed his mind. "We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying...

---

**PEOPLE**

<table>
<thead>
<tr>
<th>Name</th>
<th>Title</th>
<th>Organization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bill Gates</td>
<td>CEO</td>
<td>Microsoft</td>
</tr>
<tr>
<td>Bill Veghte</td>
<td>VP</td>
<td>Microsoft</td>
</tr>
<tr>
<td>Richard Stallman</td>
<td>Founder</td>
<td>Free Soft..</td>
</tr>
</tbody>
</table>

---

Select Name
From PEOPLE
Where Organization = 'Microsoft'

---

Bill Gates
Bill Veghte

(from Cohen's IE tutorial, 2003)
**EXTRACT – Local or SaaS**

Out-of-the-box data
- Web Services for complex, atomic and named entities

Significant additional effort
- Infrastructures for extracting, managing and scalable storage of named entities
- Web Services for extracting named entities

High additional effort
- Screen scraper
EXTRACT – NLP Stack

Web Page Identification ➔ Fact Extraction ➔ Fact Verification ➔ Load/Index & Transfer ➔ Data Analysis

(Distributed) Text Query Processor

Existing OLAP & SQL Tools

Domain Specific

Events & Facts

Entities
Candidates, Resolution, Normalization

Basic NLP
Noun Groups, Verb Groups, Numbers Phrases, Abbreviations

Metadata Analysis
Title, Date, Body, Paragraph

Sentence Marking

Morphological Analyzer
POS Tagging (per word)
Stem, Tense, Aspect, Singular/Plural
Gender, Prefix/Suffix Separation

Language Specific

Tokenization

Source: Moty Ben-Dov, Ronen Feldman: Text Mining and Information Extraction. The Data Mining and Knowledge Discovery Handbook 2005: 801-831
public concept FamilyRelation {
    attributes:
        string Person;
        string FamilyRelation;
        string Person_Relative;
    }

wordclass wcThirdPersonPossessPronoun = "his" "her";
wordclass wcFamilyRelation = "father" "mother" "son" ... "nephew";

rule FamilyRelation {
    pattern:
        PersonAnaphora->pron //his
        wcFamilyRelation ->relation //nephew
        wcComma?
        PersonOrPersonDetails -> relative; //Inayatullah Khan Jadoon
    constraints:
        //pron is a relevant pronoun
        pron IS_A wcThirdPersonPossessPronoun;

        //make sure that the antecedent (here: "Amanullah Khan") isn’t empty
        ! pron.Antecedent.IsEmpty();
        //person is never a relative of himself!
        pron.Antecedent != relative.Person;
    actions:
        Add(Person<-pron.Antecedent,FamilyRelation<-relation
            Person_Relative<-relative.Person);
    }

EXTRACT – Rules/Lists

- High precision annotations 😊
- Extraction type system must be known *a priori* 😞
- Rules difficult to engineer (inter-dependencies!) 😞
- Cost intensive: Scales only to few relations 😞

Source: Open Calais 4.0 Web Service
Max is a student at TU Berlin.

Mappings between source and target view?
What are words that indicate the relationship?
Unsupervised, robust Extraction

MyDomain.Jobs

<table>
<thead>
<tr>
<th>Person</th>
<th>Role</th>
<th>Organization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max</td>
<td>student</td>
<td>TU Berlin</td>
</tr>
</tbody>
</table>

/* person_education.joy*/

SENTENCE SPANS

(NER.Person SPANS MyDomain.Jobs.Person) AS Slot1
IsFollowedBy (*, MyDomain.Jobs.Role, *) AS Between

Max is a student at TU Berlin.
Max is a student at TU Berlin.

<table>
<thead>
<tr>
<th>Person</th>
<th>Role</th>
<th>Organization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max</td>
<td>student</td>
<td>TU Berlin</td>
</tr>
</tbody>
</table>

### Sentence Extractor

**NER. Person** → **is a** → **LIST (MyDomain. Jobs. Role)** → **at** → **NER. Company**
Robust Extraction: Why is it working?

<table>
<thead>
<tr>
<th>CNT</th>
<th>SLOT1 Type</th>
<th>BETWEEN_TEXT</th>
<th>SLOT2 Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>44</td>
<td>Person</td>
<td>GxFBxJOBxTITLExxxxxxxxxxx ,</td>
<td>Organization</td>
</tr>
<tr>
<td>34</td>
<td>Person</td>
<td>of</td>
<td>Organization</td>
</tr>
<tr>
<td>31</td>
<td>Person</td>
<td>GxFBxJOBxTITLExxxxxxxxxxx of</td>
<td>Organization</td>
</tr>
<tr>
<td>12</td>
<td>Person</td>
<td>GxFBxJOBxTITLExxxxxxxxxxx of the</td>
<td>Organization</td>
</tr>
<tr>
<td>11</td>
<td>Person</td>
<td>s</td>
<td>Organization</td>
</tr>
<tr>
<td>9</td>
<td>Person</td>
<td>GxFBxJOBxTITLExxxxxxxxxxx .</td>
<td>Organization</td>
</tr>
<tr>
<td>6</td>
<td>Person</td>
<td>GxFBxJOBxTITLExxxxxxxxxxx</td>
<td>Organization</td>
</tr>
<tr>
<td>6</td>
<td>Person</td>
<td>GxFBxJOBxTITLExxxxxxxxxxx at</td>
<td>Organization</td>
</tr>
<tr>
<td>5</td>
<td>Person</td>
<td>GxFBxJOBxTITLExxxxxxxxxxx at</td>
<td>Organization</td>
</tr>
<tr>
<td>5</td>
<td>Person</td>
<td>from</td>
<td>Organization</td>
</tr>
<tr>
<td>4</td>
<td>Person</td>
<td>GxFBxJOBxTITLExxxxxxxxxxx</td>
<td>Organization</td>
</tr>
<tr>
<td>3</td>
<td>Person</td>
<td>Of</td>
<td>Organization</td>
</tr>
<tr>
<td>3</td>
<td>Person</td>
<td>GxFBxJOBxTITLExxxxxxxxxxx ,</td>
<td>Organization</td>
</tr>
<tr>
<td>3</td>
<td>Person</td>
<td>GxFBxJOBxTITLExxxxxxxxxxx at the</td>
<td>Organization</td>
</tr>
<tr>
<td>3</td>
<td>Person</td>
<td>vice chancellor and GxFBxJOBxTITLExxxxxxxxxxx ,</td>
<td>Organization</td>
</tr>
<tr>
<td>2</td>
<td>Person</td>
<td>Joins the GxFBxJOBxTITLExxxxxxxxxxx of the</td>
<td>Organization</td>
</tr>
<tr>
<td>2</td>
<td>Person</td>
<td>at</td>
<td>Organization</td>
</tr>
<tr>
<td>2</td>
<td>Person</td>
<td>for</td>
<td>Organization</td>
</tr>
<tr>
<td>2</td>
<td>Person</td>
<td>is GxFBxJOBxTITLExxxxxxxxxxx of</td>
<td>Organization</td>
</tr>
<tr>
<td>2</td>
<td>Person</td>
<td>is a GxFBxJOBxTITLExxxxxxxxxxx at</td>
<td>Organization</td>
</tr>
<tr>
<td>2</td>
<td>Person</td>
<td>on</td>
<td>Organization</td>
</tr>
<tr>
<td>2</td>
<td>Person</td>
<td>steps down as GxFBxJOBxTITLExxxxxxxxxxx of</td>
<td>Organization</td>
</tr>
<tr>
<td>2</td>
<td>Person</td>
<td>was a GxFBxJOBxTITLExxxxxxxxxxx of Church History at</td>
<td>Organization</td>
</tr>
<tr>
<td>2</td>
<td>Person</td>
<td>83, GxFBxJOBxTITLExxxxxxxxxxx of</td>
<td>Organization</td>
</tr>
<tr>
<td>2</td>
<td>Person</td>
<td>GxFBxJOBxTITLExxxxxxxxxxx, GxFERxORGANIZATIONxxxxxxxx ,</td>
<td>Organization</td>
</tr>
<tr>
<td>2</td>
<td>Person</td>
<td>GxFBxJOBxTITLExxxxxxxxxxx in the GxFERxORGANIZATIONxxxxxxxx ,</td>
<td>Organization</td>
</tr>
<tr>
<td>2</td>
<td>Person</td>
<td>GxFERxORGANIZATIONxxxxxxxx, Former GxFBxJOBxTITLExxxxxxxxxxx</td>
<td>Organization</td>
</tr>
<tr>
<td>2</td>
<td>Person</td>
<td>a GxFBxJOBxTITLExxxxxxxxxxx at</td>
<td>Organization</td>
</tr>
<tr>
<td>2</td>
<td>Person</td>
<td>a former GxFBxJOBxTITLExxxxxxxxxxx at the</td>
<td>Organization</td>
</tr>
<tr>
<td>2</td>
<td>Person</td>
<td>GxFBxJOBxTITLExxxxxxxxxxx of</td>
<td>Organization</td>
</tr>
<tr>
<td>2</td>
<td>Person</td>
<td>GxFBxJOBxTITLExxxxxxxxxxx of the</td>
<td>Organization</td>
</tr>
</tbody>
</table>
## Robust Extraction: Why is it working?

- Few human rules are required
- Scalable with size of training data (i.e. with training data from Freebase or DBPEDIA)
- Unsupervised extraction of complex relationships
- Filters promising patterns by relation-independent lexico-syntactic filters
Data Flows and Operators

- **Web Page Identification**
  - Find qualitative pages containing relevant textual facts
  - Without “owning” pages

- **Information Extraction**
  - Extract facts from text documents and web-pages
  - Open or Closed Type System
  - Learning/Rule/Human

- **Fact Verification, Cleansing and Fusion**
  - Drop false facts, solve ambiguities
  - Requires generally a holistic view of the data
RESOLVE (Text, Object)

Web Page Identification → Fact Extraction → Fact Verification → Load/Index & Transfer → Data Analysis

(Distributed) Text Query Processor

Existing OLAP & SQL Tools

<table>
<thead>
<tr>
<th>CompanyTechnology $403$</th>
<th>technology $404$</th>
<th>text_technology $404$</th>
<th>company $340$</th>
<th>text_company $340$</th>
</tr>
</thead>
<tbody>
<tr>
<td>272297</td>
<td>88097</td>
<td>Cell phones</td>
<td>11 b</td>
<td>272287</td>
</tr>
<tr>
<td>272306</td>
<td>90023</td>
<td>mobile phones</td>
<td>13 b</td>
<td>272287</td>
</tr>
</tbody>
</table>

Duplicates

<table>
<thead>
<tr>
<th>company $18$</th>
<th>text_company $18$</th>
</tr>
</thead>
<tbody>
<tr>
<td>909 Dell</td>
<td>4 b...</td>
</tr>
<tr>
<td>1417 Novell Inc.</td>
<td>11 b...</td>
</tr>
<tr>
<td>748 IBM</td>
<td>3 b...</td>
</tr>
<tr>
<td>3519 Canonical Ltd.</td>
<td>14 b...</td>
</tr>
<tr>
<td>3052 Dell Inc.</td>
<td>9 b...</td>
</tr>
<tr>
<td>435 Red Hat</td>
<td>7 b...</td>
</tr>
</tbody>
</table>

Dell to offer Ubuntu Linux on some computers

From Wikinews, the free news source you can write!

Wednesday, May 2, 2007

Dell Inc. (Dell) announced Tuesday that, due to public demand, they will be offering an alternative to Microsoft Windows on a selection of their desktop and laptop computers. Dell will partner with Canonical Ltd. to pre-install the recently released Ubuntu 7.04 distribution of the Linux operating system.
RESOLVE (Text, Object)

Web Page Identification → Fact Extraction → Fact Verification → Load/Index & Transfer → Data Analysis

(Distributed) Text Query Processor

Existing OLAP & SQL Tools

Entities | Facts
--- | ---
KnowItAll | 30 000 | 60 000
SUMO | 20 000 | 80 000
WordNet | 120 000 | 5 Mio.
Cyc | 300 000 | 8 Mio.
TextRunner | n/a | 8 Mio.
YAGO | 1.7 Mio. | 15 Mio.
DBpedia | 1.9 Mio. | 103 Mio.
Freebase | ??? | ???

Fabian M. Suchanek, Gjergji Kasneci, Gerhard Weikum: Yago: a core of semantic knowledge. WWW 2007
FUSE(Record1, Record2)

Web Page Identification — Fact Extraction — Fact Verification — Load/Index & Transfer — Data Analysis

(Distributed) Text Query Processor

Abstract text query processing by an algebra

- FULL SCAN / FOCUSED SCAN
- EXTRACT (Rule/List/Learning/OIE)
- RESOLVE/FUSE
- AVG/COUNT/JOIN/GROUP/ ... 
- ...many more

How do we implement our algebra at web scale?
MapReduce Programming Model

- For processing of huge amounts of data
- Massive parallelization of computing tasks
- Applicable to many real world applications
- MapReduce programs are easy to implement

MapReduce Engine

- Environment to run MapReduce programs
- Distributes computing tasks
- Errors are transparently handled
- Very scalable architecture
- Examples: Google MapReduce & Apache Hadoop
What is Hadoop?
- Free software framework for data intensive applications
- Enables distributed processing of vast amounts of data on cloud computing architectures
- Supports clouds with 1000+ nodes

Major Components
- Hadoop Distributed File System (HDFS)
- MapReduce Engine

Where can you get Hadoop?
MapReduce – Group AVG Example

Input Data

MAP(k,v)

Intermediate (K,V)-Pairs

REDUCE(k,list(v))

Result

NewYork, US, 10
LosAngeles, US, 40
London, GB, 20
Berlin, DE, 60
Glasgow, GB, 10
Munich, DE, 30

...
Implementing the Algebra

► Relational Operators
  ▪ Selection & Projection → Map
  ▪ Joins → Merge / Cross
  ▪ Grouping / Aggregation → Reduce / Tree
  ▪ Sort (Order By) → Reduce / Tree

► Information Extraction Operators
  ▪ FOCUSED SCAN (Single page) → Map
  ▪ EXTRACT (Single page) → Map
  ▪ RESOLVE/FUSE (Single page) → Map

► Operators across Pages
  ▪ EXTRACT (multiple pages) → ???
  ▪ RESOLVE/FUSE (multiple pages) → ???
Challenge – Data Density & Completeness

Adaptive Crawling/Extraction/Resolve

No layoff for Boing?
Challenge – Data Source Dependency

Foreign Ownership

... aircraft manufacturing subsidiary of European-owned EADS aerospace technology group, has been plagued with financial problems—problems so severe, the company recently announced the possibility of closing up to five airplane manufacturing complexes and laying off up to 10,000 employees. As can be imagined, the negative news has caused EADS' share price to decline. Adding even more intense downward pressure on EADS' share price was the...

Who is leaderswedeserve and actionsbyt?
### Challenge – Data Source Dependency

Did leaders deserve to copy the fact from actionsbyt?
### Challenge – Data Accuracy

#### Is Emirate Airline really a competitor to Airbus?

#### Are Far East Russia and Joint Chief competitors?
### Challenge – Data Accuracy

#### US presidents

<table>
<thead>
<tr>
<th>Item Name</th>
<th>Image</th>
<th>Description</th>
<th>Date Of Birth</th>
<th>Full Name</th>
<th>Succeeded By</th>
<th>Date Of Death</th>
</tr>
</thead>
<tbody>
<tr>
<td>George Washing</td>
<td><img src="image1.png" alt="Image" /></td>
<td>On April 30, 1789, George Washington, standing on the balcony of Federal Hall on Wall Street in New York, took his oath of office as the first President of the United States. He successfully led the country through its greatest challenges and controversies.</td>
<td>February 22, 1732</td>
<td>No value found</td>
<td>John Adams</td>
<td>January 5, 1841</td>
</tr>
<tr>
<td>Thomas Jefferson</td>
<td><img src="image2.png" alt="Image" /></td>
<td>In the thick of party conflict in 1800, Thomas Jefferson wrote in a private letter, &quot;I have sworn upon the altar of God eternal hostility against every form of entangling alliances with foreign powers.&quot;</td>
<td>April 13, 1743</td>
<td>No value found</td>
<td>James Madison</td>
<td>July 4, 1826</td>
</tr>
<tr>
<td>Barack Obama</td>
<td><img src="image3.png" alt="Image" /></td>
<td>President Obama will sign an historic bill into law this afternoon, giving the Food and Drug Administration authority to regulate tobacco for the first time.</td>
<td>August 4, 1961</td>
<td>No value found</td>
<td>Kwame Raoul</td>
<td></td>
</tr>
<tr>
<td>Abraham Lincoln</td>
<td><img src="image4.png" alt="Image" /></td>
<td>Abraham Lincoln (February 12, 1809 – April 15, 1865) was the 16th President of the United States. He successfully led the country through its greatest challenges and controversies.</td>
<td>February 12, 1809</td>
<td>No value found</td>
<td>Abraham Lincoln</td>
<td></td>
</tr>
<tr>
<td>Theodore Roosevelt</td>
<td><img src="image5.png" alt="Image" /></td>
<td>Theodore Roosevelt was born on October 27, 1858, in a four-story brownstone at 20 East 20th Street, in the modern-day Gramercy section of New York City. He is known for his achievements in politics, education, and international relations.</td>
<td>October 27, 1858</td>
<td>No value found</td>
<td>Theodore Roosevelt</td>
<td></td>
</tr>
<tr>
<td>Harry S. Truman</td>
<td><img src="image6.png" alt="Image" /></td>
<td>The President Harry S. Truman Fellowship in National Security Science and Engineering, a distinguished postdoctoral three-year appointment at Sandia National Laboratories, honors those who have shown exceptional promise in the field.</td>
<td>May 8, 1884</td>
<td>No value found</td>
<td>Harry S. Truman</td>
<td></td>
</tr>
<tr>
<td>Calvin Coolidge</td>
<td><img src="image7.png" alt="Image" /></td>
<td>John Calvin Coolidge, Jr. was born in Plymouth, Windham County, Vermont. On July 4, 1872, the only U.S. President to be born on the Fourth of July, he went on to serve as the 30th President of the United States.</td>
<td>July 4, 1872</td>
<td>No value found</td>
<td>Calvin Coolidge</td>
<td></td>
</tr>
</tbody>
</table>

**4 sources for "1982."**

- Date Of Death for Barack Obama
  - [Barack Obama in Google Squared](http://www.google.com/squaredlabs/us_presidents.html)
  - [Barack Hussein Obama Sr. - Freebase](http://www.freebase.com/view/wikipedia/images/en_id/20971417)
  - [date of death: 1982](http://en.wikipedia.org/wiki/Barack_Obama_Sr)
## Challenge – Data Accuracy

Below is a screenshot of a Google Squared search for US presidents. The table displays various details about each president, including their birth and death dates, but some dates are missing or incorrect. The challenge is to verify the accuracy of these dates and fill in the missing information.

<table>
<thead>
<tr>
<th>Item Name</th>
<th>Image</th>
<th>Description</th>
<th>Date Of Birth</th>
<th>Full Name</th>
<th>Succeeded By</th>
<th>Date Of Death</th>
</tr>
</thead>
<tbody>
<tr>
<td>George Washington</td>
<td></td>
<td>On April 30, 1789, George Washington, standing on the balcony of Federal Hall on Wall Street in New York, took his oath of office as the first President of the United States. He successfully led the country through its greatest challenges.</td>
<td>February 22, 1732</td>
<td>No value found</td>
<td>John Adams</td>
<td>January 5, 1943</td>
</tr>
<tr>
<td>Thomas Jefferson</td>
<td></td>
<td>In the thick of party conflict in 1800, Thomas Jefferson wrote in a private letter, “I have sworn upon the altar of God eternal hostility against every form...”</td>
<td>April 13, 1743</td>
<td>No value found</td>
<td>James Madison</td>
<td>July 4, 1826</td>
</tr>
<tr>
<td>Barack Obama</td>
<td></td>
<td>President Obama will sign an historic bill into law this afternoon, giving the Food and Drug Administration authority to regulate tobacco for the first time...</td>
<td>August 4, 1961</td>
<td>No value found</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Abraham Lincoln</td>
<td></td>
<td>Abraham Lincoln (February 12, 1809 – April 15, 1865) was the 16th President of the United States. He successfully led the country through its greatest challenges...</td>
<td>February 12, 1809</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Theodore Roosevelt</td>
<td></td>
<td>Theodore Roosevelt was born on October 27, 1858, in a four-story brownstone at 29 East 20th Street, in the modern-day Gramercy park section of New York City.</td>
<td>October 27, 1858</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Harry S. Truman</td>
<td></td>
<td>The President Harry S. Truman Fellowship in National Security Science and Engineering, a distinguished postdoctoral three-year appointment at Sandia National Laboratories...</td>
<td>May 8, 1884</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calvin Coolidge</td>
<td></td>
<td>John Calvin Coolidge, Jr. was born in Plymouth, Windham County, Vermont, on July 4, 1872, the only U.S. President to be born on the Fourth of July...</td>
<td>July 4, 1872</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

**4 sources for “1982.”**

**Date Of Death for Barack Obama**

- [Barack Obama in Google Squared. Date of Death: 1982](http://flickr.com/photos/sahikokun/3634831327)
- [Barack Hussein Obama. Sr. - Freebase](http://www.freebase.com/view/wiki/Barack_Hussein_Obama_
- [date of death: 1982](http://en.wikipedia.org/wiki/Barack_Obama_

---

© Chair for Database Systems and Information Management
Challenge – Data Accuracy
Challenge – Data Accuracy


4 sources for "1982."
Date Of Death for Barack Obama

- Barack Obama in Google Squared. Date of Death: 1982 on Flickr...
- Barack Hussein Obama Sr. - Freebase
  Creation Date: Edit. Import list: Detail view ...
  Date of death: 1936. Edit. Date of birth:
  1936. Edit. Import list: Detail view ...
  Out the Comet's Arm: Barack Obama's Moon
  I don't have a date for the Father's death but during 1982 there was a conjunction of Saturn-Pluto in Libra. This would have been in Obama's 8th ...
  date of death: 1982
Challenge – Robust Extraction

Learned Pattern:

```
<NN> <NN> <COLON> <DIGIT> <METRIC>
```

Population density: 97 inhabitants/km

Found Pattern:

```
<DASH> <NN> <DIGIT><COMMA><DIGIT> ... <METRIC>
```

-Density 2,2128 (5,512/sq mi)

Include Community & Web Data for deriving Semantics
Business Analytics and Web Search

When Harry met Sally
TAKE HOME MESSAGE

Beyond search – Web Scale Business Analytics
**TAKE HOME MESSAGE**

**Finance - December 9th**

Make sense for the airline management teams and perhaps negotiate better pricing from Airbus and Boeing enabling it to scale and Emirates Airline. Deflation So we're in a recession? Duh... 

cribe.blogspot.com

Emirates power in his own hands

Sarkozy's first official trip as president inside the country was to Toulouse, where he spoke to employees of Airbus threatened by mass redundancies. He later shared lunch with them in the canteen. Sarkozy promised to increase the share of the French state in the...

2007-06-02 | ahmedismailibrahim.wordpress.com

**Foreign Ownership**

Airbus and the Perils of Foreign Ownership Thursday, October 26, 2006 Crafty kids don't just share their dessert—they use it to make friends, build...

2006-10-26 | actionsbyt.wordpress.com
**TAKE HOME MESSAGE**

<table>
<thead>
<tr>
<th>Company</th>
<th>companyproduct</th>
<th>companylayoffs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Airbus</strong></td>
<td>A380 Aircraft</td>
<td>(2)announced 10,000</td>
</tr>
<tr>
<td></td>
<td>Airbus A380 Aircraft</td>
<td>(1)see more »</td>
</tr>
<tr>
<td><strong>Lockheed Martin</strong></td>
<td>Joint Strike Aircraft</td>
<td>(2)see more »</td>
</tr>
<tr>
<td></td>
<td>F-35 Joint Strike Aircraft</td>
<td>(1)see more »</td>
</tr>
<tr>
<td><strong>Boeing</strong></td>
<td>Chinook Aircraft</td>
<td>(3)see more »</td>
</tr>
<tr>
<td></td>
<td>787 Dreamliner Aircraft</td>
<td>(2)see more »</td>
</tr>
<tr>
<td></td>
<td>KC-45 Aircraft</td>
<td>(1)see more »</td>
</tr>
<tr>
<td><strong>British Airways</strong></td>
<td>Airbus A380 Aircraft</td>
<td>(1)see more »</td>
</tr>
<tr>
<td></td>
<td>F-10s Aircraft</td>
<td>(1)see more »</td>
</tr>
<tr>
<td><strong>Bilderberg Group</strong></td>
<td>Mig-15s Aircraft</td>
<td>(1)see more »</td>
</tr>
<tr>
<td><strong>Far East Russia</strong></td>
<td>Mig-15s Aircraft</td>
<td>(1)see more »</td>
</tr>
<tr>
<td><strong>Northrop Grumman</strong></td>
<td>KC-45 Aircraft</td>
<td>(1)see more »</td>
</tr>
<tr>
<td><strong>FedEx</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

**Foreign Ownership**

...aircraft manufacturing subsidiary of European-owned eads aerospace technology group, has been plagued with financial problems—problems so severe, the company recently announced the possibility of closing up to five airplane manufacturing complexes and laying off up to 10,000 employees. As can be imagined, the negative news has caused eads shares to cascade. Adding even more intense downward pressure on eads’ share price was the ...

2005-10-28 | actionsbyt.wordpress.com
Queries integrating two fact types (Query domain: Fraud detection)
Q2.1: persons working in companies, person convictions (data source: www.cnn.com)
Q2.2: persons working in companies, person travel location (data source: entire web)
Q2.3: person gender and person arrestments (data source: entire web)
Q2.4: persons working in companies, person age (data source: www.washingtonpost.com)

Queries integrating three fact types (Query domain: Human resources and nature preservation)
Q3.1: person working in company and position, person age, company technology (data source: www.cnn.com)
Q3.2: company location, number of company employees, company products (data source: www.ft.com)
Q3.4: natural disaster location, company location, company technology (data source: www.cnn.com)

Queries integrating four fact types (Query domain: Mergers & acquisitions)
Q4.1: company location, company employees, company technology, company bankruptcy

Queries integrating five fact types (Query domain: General information about a person)
Q5.1: person attributes, person convictions, person career, person travel, person arrestment

We ask these queries already and create a market. Academia, Open Source, Search Engine and BI Tool Vendors already started to focus on this market. Let become part of this movement.
Come and visit us in Berlin …

- Capital of Germany
- 3.5m Inhabitants

- Major Science Location in Germany
  - 130,000 students at 27 universities
  - 70 Research Institutes (50,000 researchers)
    - Max-Planck, Helmholtz, Leibnitz, Fraunhofer, Hertz Institutes
  - Technology Park Adlershof (15,000 researchers, 800 Tech Companies)

- TU Berlin is a leading science university in Germany
  - 8 Nobel Prize Winners
  - ~300 tenured professors
  - ~2000 Postgraduate Researchers
  - ~2000 Student assistants
  - Affiliated Labs: T-Labs, DCAITI
  - 27,000 students at TU Berlin

- School of Electrical Engineering and Computer Science @ TU Berlin
  - ~2,500 Students, ~50 tenured/chaired professors

… as Master/PhD. Student, Colloquium Speaker, Post-Doc, Visiting Scientist!