

# Text Technologies in the Mainstream: Text Analytics Solutions, Applications, and Trends

Seth Grimes

Alta Plana Corporation

301-270-0795 -- *<http://altaplana.com>*

INFORMS 2008

June 15, 2008

# Introduction

Seth Grimes —

Principal Consultant with Alta Plana Corporation.

Contributing Editor, *IntelligentEnterprise.com*.

Channel Expert (text analytics), *B-Eye-Network.com*.

Founding Chair, Text Analytics Summit, *textanalyticsnews.com*.

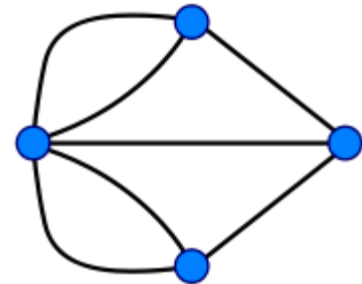
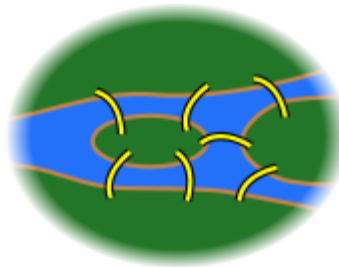
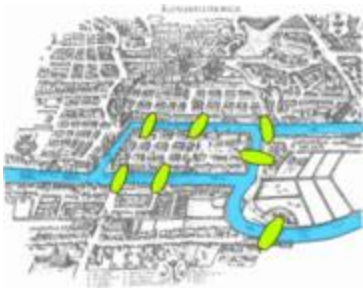
Instructor, The Data Warehousing Institute, *tdwi.org*.

# What is Analytics?



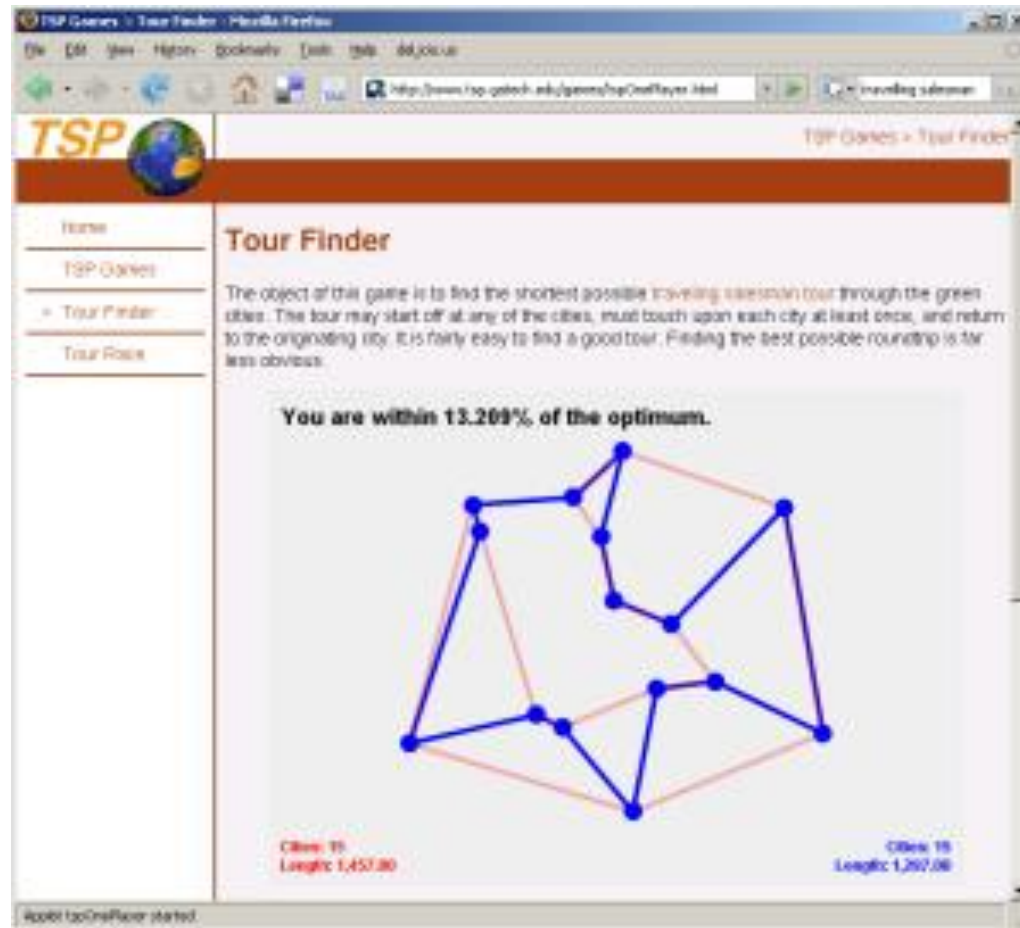
[http://www.tropicalisland.de/NYC\\_New\\_York\\_Brooklyn\\_Bridge\\_from\\_World\\_Trade\\_Center\\_b.jpg](http://www.tropicalisland.de/NYC_New_York_Brooklyn_Bridge_from_World_Trade_Center_b.jpg)

$$\begin{aligned} x(t) &= t \\ \longrightarrow y(t) &= \frac{1}{2} a (e^{t/a} + e^{-t/a}) \\ &= a \cosh(t/a) \end{aligned}$$



[http://en.wikipedia.org/wiki/Seven\\_Bridges\\_of\\_K%C3%B6nigsberg](http://en.wikipedia.org/wiki/Seven_Bridges_of_K%C3%B6nigsberg)

# What is Analytics?



# What is Analytics?

What do you do when you're working with this?

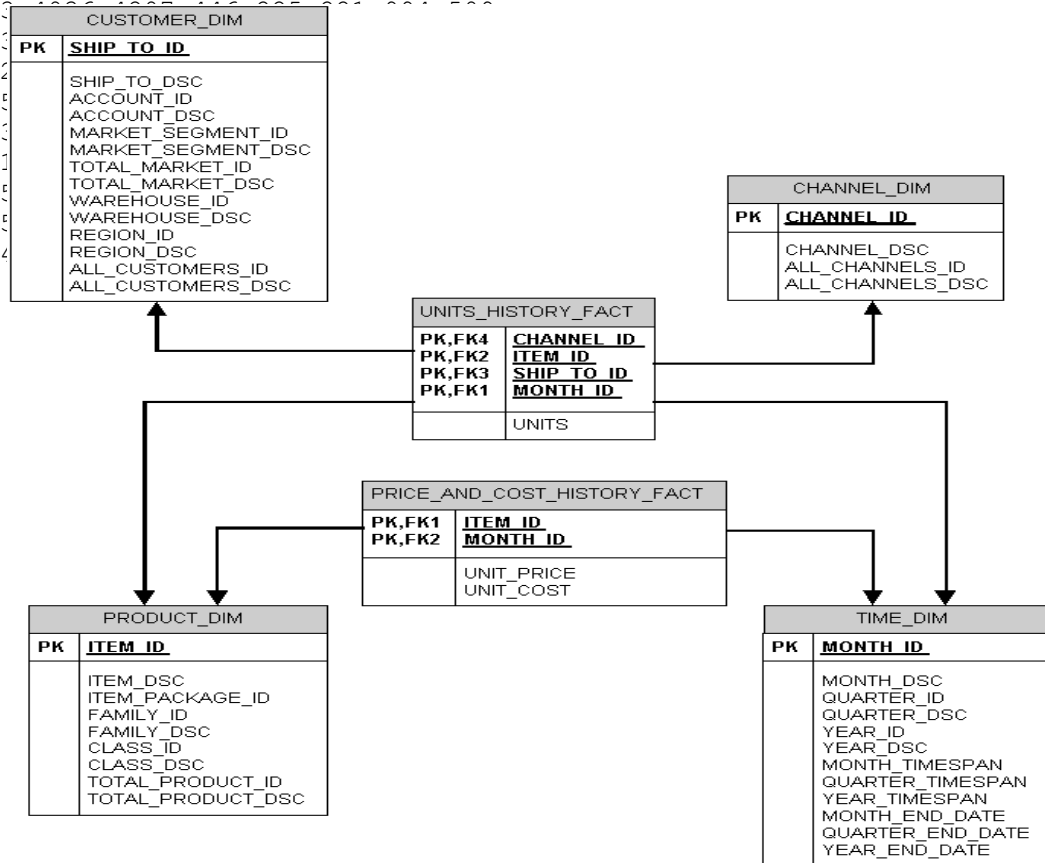
```
"SUMLEV","STATE","COUNTY","STNAME","CTYNAME","YEAR","POPESTIMATE",  
50,19,1,"Iowa","Adair County",1,8243,4036,4207,446,225,221,994,509  
50,19,1,"Iowa","Adair County",2,8243,4036,4207,446,225,221,994,509  
50,19,1,"Iowa","Adair County",3,8212,4020,4192,442,222,220,987,505  
50,19,1,"Iowa","Adair County",4,8095,3967,4128,432,208,224,935,488  
50,19,1,"Iowa","Adair County",5,8003,3924,4079,405,186,219,928,495  
50,19,1,"Iowa","Adair County",6,7961,3892,4069,384,183,201,907,472  
50,19,1,"Iowa","Adair County",7,7875,3855,4020,366,179,187,871,454  
50,19,1,"Iowa","Adair County",8,7795,3817,3978,343,162,181,841,439  
50,19,1,"Iowa","Adair County",9,7714,3777,3937,338,159,179,805,417
```

# Business Intelligence

## Traditional BI feeds off:

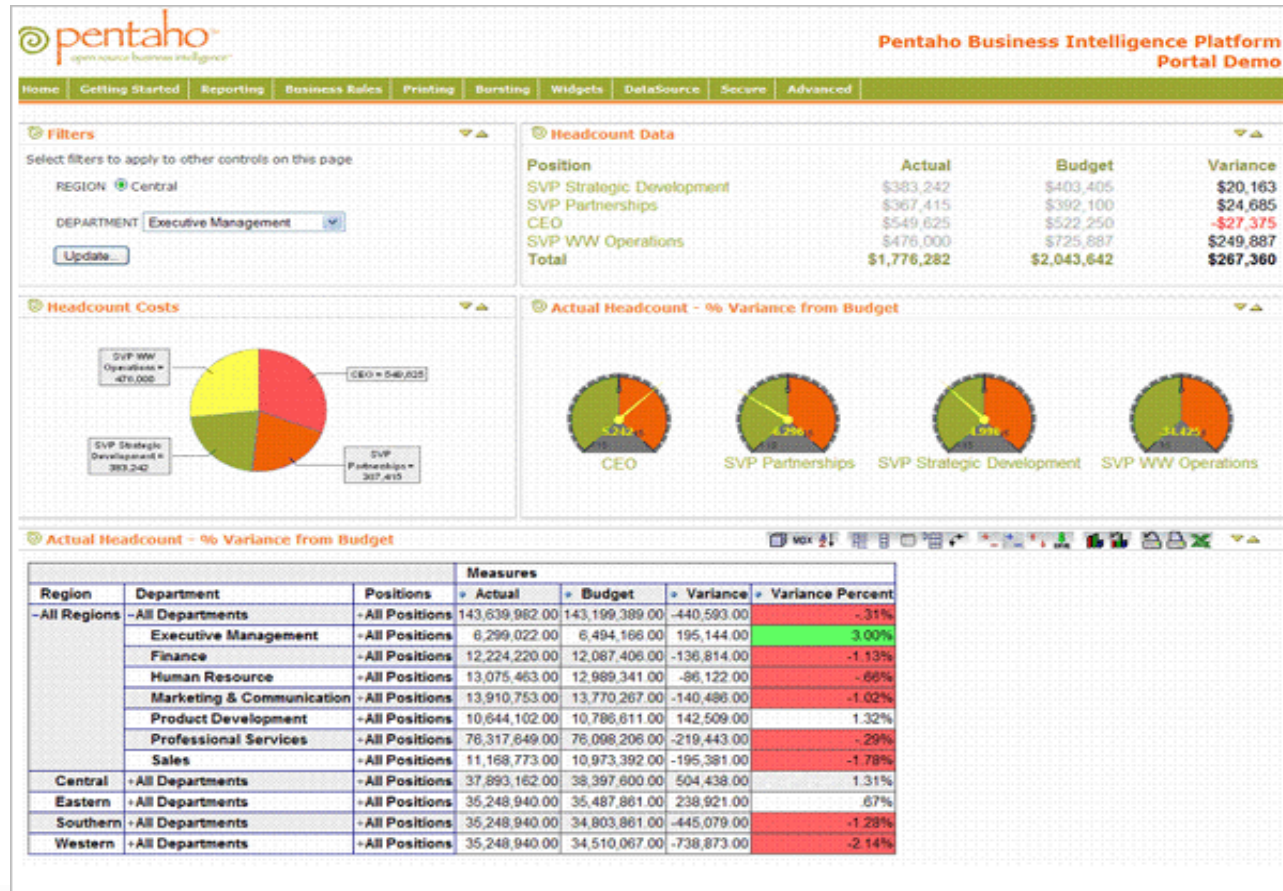
```
"SUMLEV", "STATE", "COUNTY", "STNAME", "CTYNAME", "YEAR", "POPESTIMATE",
50,19,1,"Iowa","Adair County",1,8243
50,19,1,"Iowa","Adair County",2,8243
50,19,1,"Iowa","Adair County",3,8212
50,19,1,"Iowa","Adair County",4,8095
50,19,1,"Iowa","Adair County",5,8003
50,19,1,"Iowa","Adair County",6,7961
50,19,1,"Iowa","Adair County",7,7875
50,19,1,"Iowa","Adair County",8,7795
50,19,1,"Iowa","Adair County",9,7714
```

It runs off:



# Business Intelligence

Traditional BI produces:



# Business Intelligence

“The bulk of information value is perceived as coming from data in relational tables. The reason is that data that is structured is easy to mine and analyze.”

– *Prabhakar Raghavan, Yahoo Research, former CTO of enterprise-search vendor Verity (now part of Autonomy)*

That’s where BI operates, on data in a relational table that originated in transactional systems.

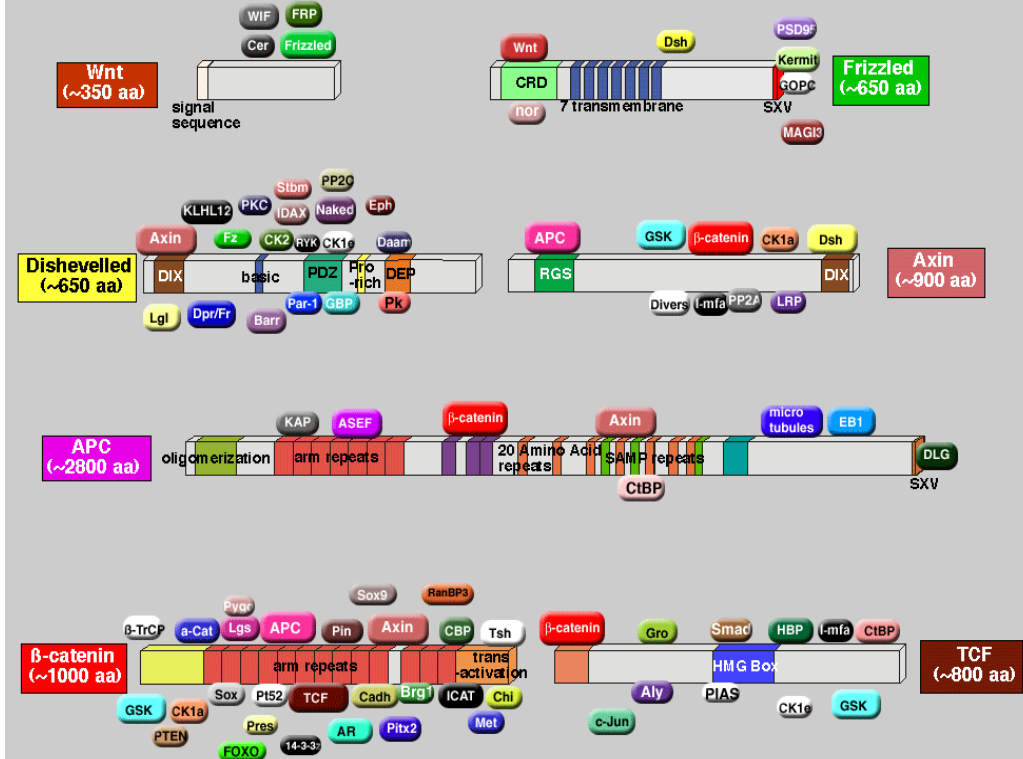
Yet it’s a truism that 80% of enterprise information is in “unstructured” form.



# The “Unstructured Data” Challenge

April 2006, Roel Nusse

These diagrams display interactions between proteins in Wnt signaling and the approximate sites of binding. The partners are hyper-linked to one literature reference in PubMed. From there, one can retrieve more literature.



[www.stanford.edu/~7ernusse/wntwindow.html](http://www.stanford.edu/~7ernusse/wntwindow.html)

Axin and Frat1 interact with dvl and GSK, bridging Dvl to GSK in Wnt-mediated regulation of LEF-1.

Wnt proteins transduce their signals through dishevelled (Dvl) proteins to inhibit glycogen synthase kinase 3beta (GSK), leading to the accumulation of cytosolic beta-catenin and activation of TCF/LEF-1 transcription factors. To understand the mechanism by which Dvl acts through GSK to regulate LEF-1, we investigated the roles of Axin and Frat1 in Wnt-mediated activation of LEF-1 in mammalian cells. We found that Dvl interacts with Axin and with Frat1, both of which interact with GSK. Similarly, the Frat1 homolog GBP binds Xenopus Dishevelled in an interaction that requires GSK. We also found that Dvl, Axin and GSK can form a ternary complex bridged by Axin, and that Frat1 can be recruited into this complex probably by Dvl. The observation that the Dvl-binding domain of either Frat1 or Axin was able to inhibit Wnt-1-induced LEF-1 activation suggests that the interactions between Dvl and Axin and between Dvl and Frat may be important for this signaling pathway. Furthermore, Wnt-1 appeared to promote the disintegration of the Frat1-Dvl-GSK-Axin complex, resulting in the dissociation of GSK from Axin. Thus, formation of the quaternary complex may be an important step in Wnt signaling, by which Dvl recruits Frat1, leading to Frat1-mediated dissociation of GSK from Axin.

[www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed&cmd=Retrieve&list\\_uids=10428961&dopt=Abstract](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed&cmd=Retrieve&list_uids=10428961&dopt=Abstract)

# Text (and Media) Technologies

What do people do with electronic documents?

1. Publish, Manage, and Archive.
2. Index and Search.
3. Categorize and Classify according to *metadata* & contents.
4. Information Extraction.

For textual documents, text analytics enhances #2 and enables #3 & #4.

Text analytics (a.k.a. text data mining) can be automated or interactive.

# The “Unstructured Data” Challenge

## Consider:

E-mail, news & blog articles, forum postings, and other social media.

Contact-center notes and transcripts; recorded conversation.

Surveys, feedback forms, warranty claims.

And every other sort of document imaginable.

These sources may contain “traditional” data.

The Dow fell 46.58, or 0.42 percent, to 11,002.14. The Standard & Poor's 500 index fell 1.44, or 0.11 percent, to 1,263.85, and the Nasdaq composite gained 6.84, or 0.32 percent, to 2,162.78.

# Search

Search is typically answer #1. Search involves –

Words & phrases: search terms & natural language.

Qualifiers: include/exclude, and/or, not, etc.

Search is not enough.

*Search helps you find things you already know about. It doesn't help you **discover** things you're unaware of.*

*Search results often lack **relevance**.*

*Search finds documents, not **knowledge**.*

*Search doesn't enable **unified analytics** that links data from textual and transactional sources.*

## Search++

Text analytics enables results that suit the information and the user, e.g., answers –

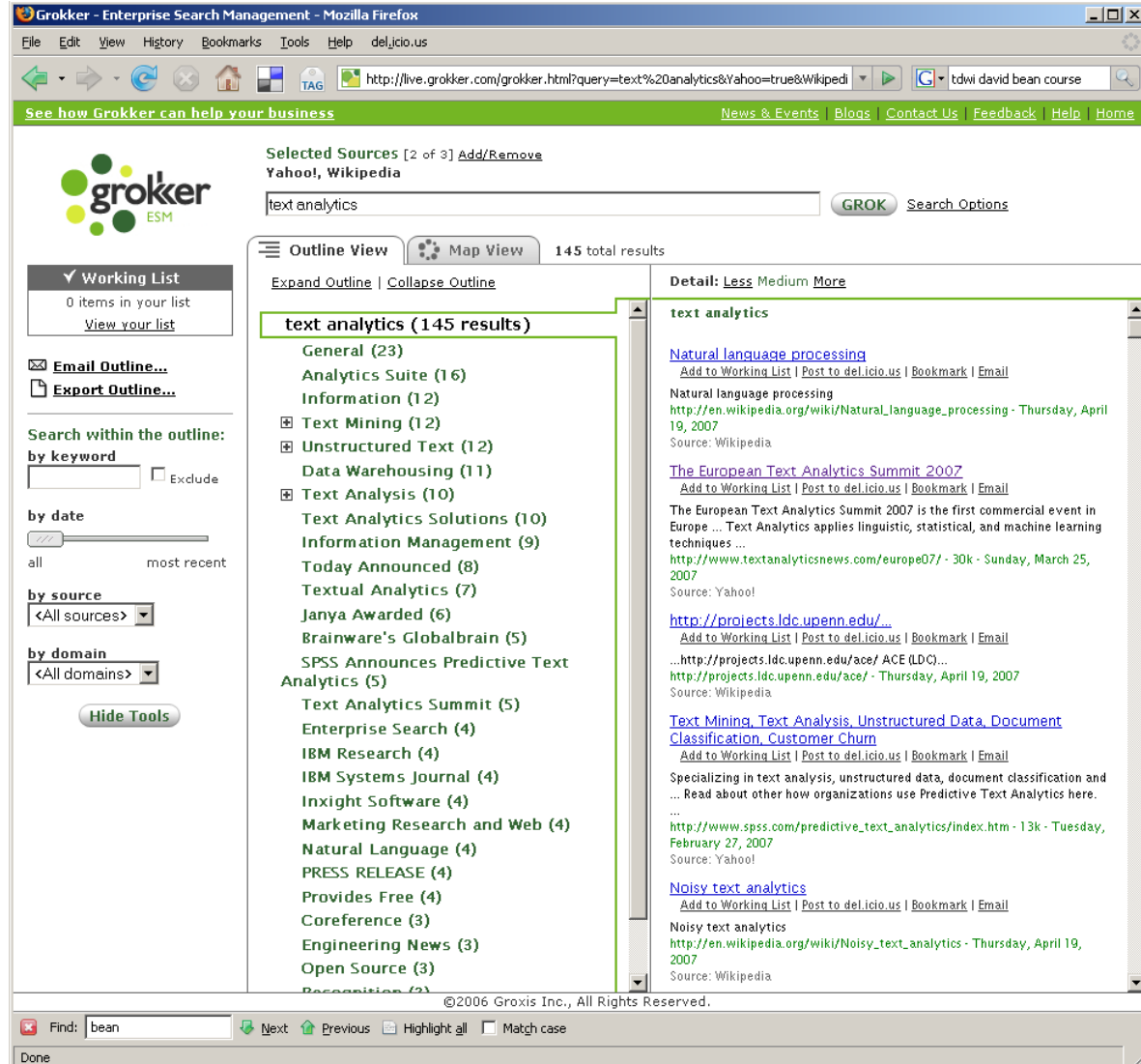


Now on to knowledge discovery, to discerning  
*interrelationships of presented facts...*

Alta Plana

Search can be pretty smart.

This slide and the next show dynamic, clustered search results from Grokker...



*live.grokker.com/grokker.html?query=text%20analytics&Yahoo=true&Wikipedia=true&numResults=250*

...with a zoomable display.

Clustering here utilizes statistical (text) data mining techniques to identifying cohesive groupings of retrieved documents.

The screenshot shows the Grokker Enterprise Search Management interface in Mozilla Firefox. The search query is 'text analytics' and it has returned 145 total results. The interface is in 'Map View' and displays a circular map of document clusters. A tooltip for a specific document is visible, showing the title 'Alias-i LingPipe 2.1 Released With Java Source for Text Analytics and Natural Language Processing', dated Mar 29, 2007, with a rank of 81 and source of Yahoo!. The map includes labels for various clusters such as 'Recognition', 'Unstructured Text', 'Text Mining', 'Information', 'Text Analysis', and 'Text Analytics Soluti...'. On the right side, a 'Detail' pane shows search results for 'Natural language processing' and 'The European Text Analytics Summit 2007'. The bottom of the browser window shows the search bar with 'bean' and navigation buttons like 'Next', 'Previous', 'Highlight all', and 'Match case'.

# Search++

Text analytics can do better.

Text analytics extracts and classifies by –

Entities: names, e-mail addresses, phone numbers

Concepts: abstractions of entities.

Facts and relationships.

Abstract attributes, e.g., “expensive,” “comfortable”

Opinions, sentiments: attitudinal data.

... and sometimes data objects.



# Text Analytics

Text Mining = Data Mining of textual sources.

Clustering and classification.

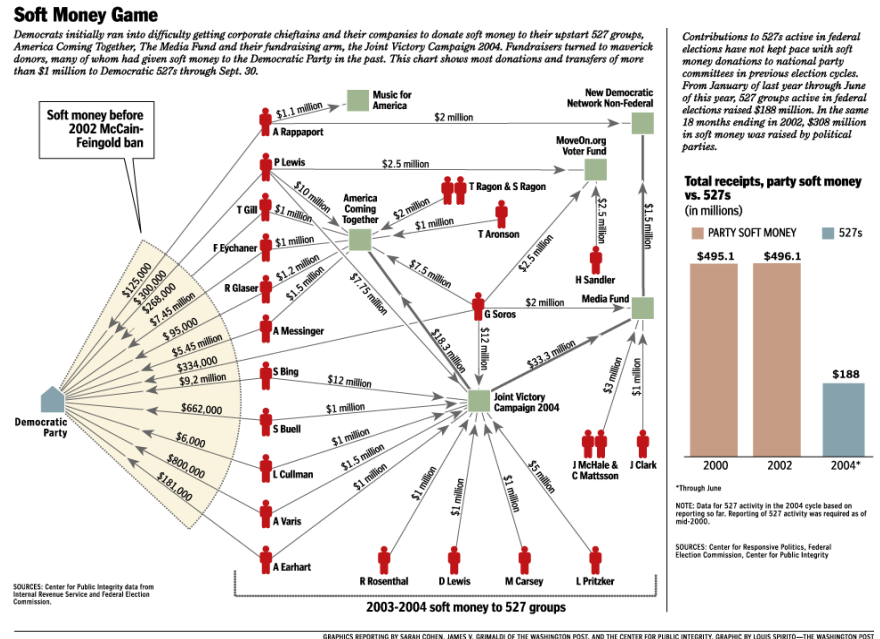
Link Analysis.

Prediction.

Association rules.

Regression.

Forecasting.



Text Mining = Knowledge Discovery in Text.

Search (Information Retrieval) is a first step.

Alta Plana

# Visualizing Interrelationships

The screenshot shows the Silobreaker website interface. At the top, there's a navigation bar with 'Home', 'Global Issues', 'Science & Technology', 'Business', and 'World'. Below that is the 'silobreaker' logo and a search bar with '360° Search', 'Network', 'Hot Spots', 'Trends', and 'My Page' tabs. A search bar contains 'SAP AG' and a 'Search Network' button. To the right, there are fields for 'User Name', 'Password', and a 'Sign In' button. Below the search bar, there's a 'Page Filter' section with a yellow background, showing 'SAP AG' selected. Underneath, there are sliders for 'Companies' (0 to 16), 'Organization' (0 to 1), 'Person' (0.2 to 1.1), 'City' (0 to 1), and 'Keyphrase' (0 to 11). A 'Visible: 30' and 'Left: 0' indicator is also present. The main content area is a network graph with 'SAP AG' at the center, connected to various entities like 'Salesforce.com Inc', 'Oracle Corporation', 'Cisco Systems Inc', 'BlackBerry', 'Wipro Ltd', 'Novell Inc', 'AT&T Inc', 'Business Objects SA', 'International Business Machines...', 'Microsoft Corporation', 'Gartner Inc', 'IDC International Data Corp', 'Hewlett-Packard Company', 'Database', 'Ecosystem', 'Berlin', 'Jim Shepherd', 'Vale Inco', 'Software as a Service', 'Acquisition', 'Radio-Frequency Identification', 'Environment', and 'Trademark'. On the right side, there are several panels: 'People' (Henning Kagermann, Jim Shepherd, Bill McDermott), 'Companies' (Business Objects SA, Oracle Corporation, International Business Machines Corp), 'Organizations' (Metropolitan Police Service, US Securities and Exchange Commission, European Union), 'Keyphrases' (Software as a Service, BlackBerry, Business Intelligence), 'Topics' (New Products & Services, Research & Development, Venture Capital, Mergers & Acquisitions), and 'Industries'. At the bottom, there are sections for 'Quotes' and 'Related Documents'. The 'Quotes' section contains a quote about SAP and hand-held terminals. The 'Related Documents' section lists various documents, news items, and reports related to SAP and Business Objects.

# Text Analytics

Typical steps in text analytics include –

Retrieve documents for analysis.

Apply statistical &/ linguistic &/ structural techniques to **identify, tag, and extract** entities, concepts, relationships, and events (features) within document sets.

Apply statistical pattern-matching & similarity techniques to **classify** documents and organize extracted features according to a specified or generated categorization / taxonomy.

– via a *pipeline* of statistical & linguistic steps.

# Text Analytics

Text analytics discerns linguistic and statistical structure inherent in the textual source materials. Let's look at some of the steps.

First, we'll do a lexical analysis of a text file, essentially a basic statistical analysis of the words and multi-word terms, looking at an article I wrote on sentiment analysis...

Keyword Density & Prominence Tool v1.5b - Mozilla Firefox

File Edit View History Bookmarks Tools Help del,icio.us

http://www.ranks.nl/cgi-bin/ranksnl/spider/spider.cgi?lang=

**RANKS.NL** KEYWORD DENSITY & PROMINENCE v1.5b Ranks Friends Log in New Report

Url tested : <http://altaplana.com/SentimentAnalysis.html> — More Domain / URL info —

Details  
 Comparison form  
 Header data  
 HTML  
 Totals, counts, special words  
 Page elements  
 Single word repeats

1423 total words in the file.  
 644 unique words in the file, short words included  
 5 possible StopWord(s) : an and the with www

word	repeats	density	Prominence	word	repeats	density	Prominence
sentiment	18 L,I	1.26%	46.93	for	17 L	1.19%	34.44
that	15	1.05%	55.22	text	15 L	1.05%	58.77
analytics	12 L	0.84%	52.83	from	10	0.70%	71.16
management	9 H	0.63%	50.37	analysis	9 L,I	0.63%	50.61
our	8	0.56%	20.36	are	8	0.56%	56.38
influence	7 H	0.49%	78.46	customer	7 H	0.49%	33.75
which	6	0.42%	63.18	understanding	6	0.42%	47.34
she	6	0.42%	68.22	notes	6	0.42%	51.18
have	6	0.42%	35.14	can	6	0.42%	55.43
been	6	0.42%	28.93	understand	5	0.35%	57.77
they	5	0.35%	54.28	sources	5	0.35%	87.31
not	5	0.35%	37.68	more	5	0.35%	42.90
mining	5	0.35%	55.84	mail	5	0.35%	63.50
extraction	5	0.35%	40.15	enterprise	5 H	0.35%	40.59
way	4	0.28%	23.61	time	4	0.28%	20.59
take	4	0.28%	14.78	surveys	4 L	0.28%	50.39
support	4	0.28%	21.75	results	4	0.28%	38.58
potential	4	0.28%	39.97	positive	4	0.28%	56.36
opinion	4	0.28%	71.71	networks	4 L	0.28%	75.03

Done

Keyword Density & Prominence Tool v1.5b - Mozilla Firefox

File Edit View History Bookmarks Tools Help del.icio.us

http://www.ranks.nl/cgi-bin/ranksnl/spider/spider.cgi?lang=

Google

Phrase repeats

Total 2 word phrases : 102 - Total Repeats : 246

phrase	repeats	density	Prominence
text analytics	9	1.26 %	58.87
of the	6	0.84 %	46.49
and the	4	0.56 %	48.45
e mail	4	0.56 %	62.86
from sources	4	0.56 %	88.12
influence networks	4 H	0.56 %	76.00
notes and	4	0.56 %	52.11
of text	4	0.56 %	52.37
to the	4	0.56 %	60.17
to understand	4	0.56 %	63.55
by the	3	0.42 %	34.65
call center	3	0.42 %	68.96
can be	3	0.42 %	81.68
customer experience	3 H	0.42 %	52.99
enterprise feedback	3 H	0.42 %	52.73
experience management	3 H	0.42 %	52.92
feedback management	3 H	0.42 %	52.66
in the	3	0.42 %	41.79
of opinion	3	0.42 %	69.97
real time	3	0.42 %	17.01
seek to	3	0.42 %	28.58
sentiment analysis	3 LI	0.42 %	69.52
sentiment extraction	3	0.42 %	37.29
the results	3	0.42 %	33.45
triggered by	3	0.42 %	26.00
a decision	2	0.28 %	20.41
a new	2	0.28 %	65.21
analytics can	2	0.28 %	97.15
analytics vendor	2	0.28 %	55.02
analyze attitudinal	2	0.28 %	96.66
and analyze	2	0.28 %	96.73
and other	2	0.28 %	37.70

Total 3 word phrases : 45 - Total Repeats : 93

phrase	repeats	density	Prominence
customer experience management	3 H	0.63 %	52.99
enterprise feedback management	3 H	0.63 %	52.73
of text analytics	3	0.63 %	46.78
analytics can be	2	0.42 %	97.15
analyze attitudinal information	2	0.42 %	96.66
and analyze attitudinal	2	0.42 %	96.73
and survey responses	2	0.42 %	95.54
applied to extract	2	0.42 %	96.94
articles blog postings	2	0.42 %	96.10
as articles blog	2	0.42 %	96.17
as varied as	2	0.42 %	96.31
attitudinal information from	2	0.42 %	96.59
be applied to	2	0.42 %	97.01
blog postings e	2	0.42 %	96.03
call center notes	2	0.42 %	95.75
can be applied	2	0.42 %	97.08
center notes and	2	0.42 %	95.68
ceo of text	2	0.42 %	55.24
cries for help	2	0.42 %	7.70
e mail call	2	0.42 %	95.89
experience management enterprise	2 H	0.42 %	62.65
extract and analyze	2	0.42 %	96.80
focus on applications	2	0.42 %	97.96
from linguamatics to	2	0.42 %	81.52
from sources as	2	0.42 %	96.45
information from sources	2	0.42 %	96.52
mail call center	2	0.42 %	95.82
management enterprise feedback	2 H	0.42 %	62.58
notes and survey	2	0.42 %	95.61
of opinion leadership	2	0.42 %	80.43
online consumer forums	2	0.42 %	55.90
postings e mail	2	0.42 %	95.96
real time two	2	0.42 %	18.58

Done

# Text Analytics

Those “tri-grams” are pretty good at describing the *Whatness* of the source text.

Shallow parsing and statistical analysis can be enough, for instance, to support classification.

It can help you get at meaning, for instance, by studying co-occurrence of terms.

But statistical pattern matching alone – the bag of words approach in a vector-space model – may fall short.

# The Need for Linguistics

Consider –

The Dow *fell* 46.58, or 0.42 percent, to 11,002.14. The Standard & Poor's 500 index fell 1.44, or 0.11 percent, to 1,263.85, and the Nasdaq composite *gained* 6.84, or 0.32 percent, to 2,162.78.

The Dow *gained* 46.58, or 0.42 percent, to 11,002.14. The Standard & Poor's 500 index fell 1.44, or 0.11 percent, to 1,263.85, and the Nasdaq composite *fell* 6.84, or 0.32 percent, to 2,162.78.

Example from Luca Scagliarini, Expert System.

Let's try syntactic analysis of a bit of text...



Connexor - Technology - Machineese - Demo - Machineese Syntax - demo - Mozilla Firefox

File Edit View History Bookmarks Tools Help del.icio.us

http://www.connexor.eu/technology/machineese/demo/syntax/ Google

**connexor**  
natural knowledge

Sitemap

Home Company Solutions Technology Partners Contact

Technology > Machineese > Demo > Machineese Syntax - demo

Machineese

- Machineese Metadata
- Machineese Syntax
- Machineese Semantics
- Machineese Phrase Tagger
- Demo

## Machineese Syntax

Machineese Syntax is a syntactic parser that returns base forms and compound structure, produces part-of-speech classes, inflectional tags, noun phrase markers and syntactic dependencies. Syntactic dependencies show functional relations between words and phrases in sentences.

What's the best price for new laptop that I'll use for business trips and around the office?

English text Apply Syntax

This demo is intended for evaluation purposes only.

Done

The screenshot shows a Mozilla Firefox browser window displaying the Connexor website. The page title is "Connexor - Technology - Machinese - Demo - Machinese Syntax - demo - Mozilla Firefox". The address bar shows the URL "http://www.connexor.eu/technology/machinese/demo/syntax/". The Connexor logo is visible at the top left, with the tagline "natural knowledge". A navigation menu includes "Home", "Company", "Solutions", "Technology", "Partners", and "Contact". The main content area is titled "Analysis of Machinese Syntax for English:" and displays a complex syntax tree for the sentence: "What is the price of the best laptop for new use that I will use that for around office business the". The tree starts with a red "root" node, which branches into "main:" and "comp:". "main:" branches into "subj:" (What) and "s:". "s:" branches into "comp:" (is) and "price". "price" branches into "attr:" (best) and "det:" (the). "det:" branches into "for". "for" branches into "pcomp:" (laptop) and "mod:". "pcomp:" branches into "atr:" (new) and "mod:". "mod:" branches into "ch:" (use) and "ha:". "ch:" branches into "subj:" (I) and "ll:". "ll:" branches into "that". "that" branches into "mp:". "mp:" branches into "for". "for" branches into "cc:". "cc:" branches into "around". "around" branches into "pcomp:". "pcomp:" branches into "atr:" (office) and "det:" (the). "atr:" branches into "business".

**Note:** The Connexor Machinese demos are intended for evaluation purposes only.

Connexor Oy, Helsinki Business and Science Park, Finland. info@connexor.com  
© Connexor Oy. Powered by [ToimiSait](#)

Applet Dtree started

Connexor - Technology - Machine - Demo - Machine Phrase Tagger - demo - Mozilla Firefox

File Edit View History Bookmarks Tools Help delicio.us

http://www.connexor.eu/technology/machine/demo/tagger/

Google

**Connexor**  
natural knowledge

Sitemap

Home Company Solutions Technology Partners Contact

Technology > Machine > Demo > Machine Phrase Tagger - demo

Machine  
Machine Metadata  
Machine Syntax  
Machine Semantics  
Machine Phrase Tagger  
Demo

## English Machine Phrase Tagger 4.6 analysis:

Text	Baseform	Phrase syntax and part-of-speech
What	what	nominal head, pro-nominal
's	be	main verb, indicative present
the	the	premodifier, determiner
best	good	premodifier, superlative adjective, noun phrase begins
price	price	nominal head, noun, noun phrase continues
for	for	postmodifier, preposition, noun phrase continues
new	new	premodifier, adjective, noun phrase continues
laptop	lap top	nominal head, noun, noun phrase ends
that	that	nominal head, pro-nominal
I	I	nominal head, pro-nominal
'll	will	auxiliary verb, indicative present
use	use	main verb, infinitive
for	for	preposed marker, preposition
business	business	premodifier, noun, noun phrase begins
trips	trip	nominal head, plural noun, noun phrase ends
..d	..d	..d

Connexor Oy, Helsinki Business and Science Park, Finland, info@connexor.com  
© Connexor Oy. Powered by TomiSait

Done

# Information Extraction

Let's see tagging in action. We'll use GATE, an open-source tool...

The screenshot displays the GATE 4.0 build 2752 interface. The main window shows a document titled "GATE document\_00020" with the following text:

Sentiment Analysis: A Focus on Applications  
 by **Seth Grimes**  
 Published: February 19, 2008  
 Text analytics can be applied to extract and analyze attitudinal information from sources as varied as articles, blog postings, e-mail, call-center notes and survey responses.

Last month, I looked at **Sentiment Analysis: Opportunities and Challenges**, promising a follow-on focus on applications. It's the breadth of opportunities – promising ways text analytics can be applied to extract and analyze attitudinal information from sources as varied as articles, blog postings, e-mail, call-center notes and survey responses – and the difficulty of the technical challenges that make existing and emerging applications so interesting.

We will explore three applications – influence networks, assessment of marketing response and customer experience management/enterprise feedback management – via mini-case studies.

Influence Networks

Aafia Chaudhry, a physician who calls herself "a passionate enthusiast in the science of opinion leadership in healthcare systems," is president of **81 qd**, a New York company that consults on pharmaceutical life-cycle management. Chaudhry

On the right, the "Original markups" panel shows a list of HTML tags with checkboxes: a (checked), blockquote, body, br, div, em, h3, html, p, span.

At the bottom, a table lists annotations:

Type	Set	Start	End	Features
a	Original markups	48	59	{href=/channels/index.php?filter_channel=1394, c
a	Original markups	266	266	{href=http://www.clarabridge.com/, isEmptyAndS
a	Original markups	290	338	{href=http://www.b-eye-network.com/view/6744, t
a	Original markups	1072	1076	{href=http://www.81qd.com/, target=_blank}
a	Original markups	1199	1211	{href=http://www.linguamatics.com/, target=_blan
a	Original markups	1728	1738	{href=http://www.lexalytics.com/index.php, target=
a	Original markups	3919	3937	{href=http://www.andersonanalytics.com/, target=

15 Annotations (1 selected)

Document Editor Initialisation Parameters

The screenshot shows the GATE 4.0 build 2752 interface. On the left is a tree view with categories: Applications (ANNIE\_0002B), Language Resources, GATE document\_00020, Corpus for GATE document\_00020, Processing Resources (ANNIE OrthoMatcher\_00036, ANNIE NE Transducer\_00035, ANNIE POS Tagger\_00034, ANNIE Sentence Splitter\_00031, ANNIE Gazetteer\_00030, ANNIE English Tokeniser\_00020), and Data stores.

The main area is titled 'Messages' and shows 'GATE document\_00020' and 'ANNIE\_0002B'. It contains two tables:

- Loaded Processing resources:** (Empty table)
- Selected Processing resources:**

!	Name	Type
●	Document Reset PR_0002C	Document Reset PR
●	ANNIE English Tokeniser_00020	ANNIE English Tokeniser
●	ANNIE Gazetteer_00030	ANNIE Gazetteer
●	ANNIE Sentence Splitter_00031	ANNIE Sentence Splitter
●	ANNIE POS Tagger_00034	ANNIE POS Tagger
●	ANNIE NE Transducer_00035	ANNIE NE Transducer
●	ANNIE OrthoMatcher_00036	ANNIE OrthoMatcher

Below the tables is a 'Corpus:' dropdown menu set to 'Corpus for GATE document\_00020'. A message states: 'The corpus and document parameters are not available as they are automatically set by the controller!'.

A table below shows 'No selected processing resource' with columns: Name, Type, Required, Value.

At the bottom right is a 'Run' button. The status bar at the bottom shows 'Serial Application editor Initialisation Parameters' and 'ANNIE\_0002B run in 0.766 seconds'.

Messages GATE document\_00020 ANNIE\_0002B

Annotation Sets Annotations Co-reference Editor Text

- via mini-case studies.  
Influence Networks

Aafia Chaudhry, a physician who calls herself "a passionate enthusiast in the science of opinion leadership in healthcare systems," is president of 81qd, a New York company that consults on pharmaceutical life-cycle management. She applies text-analytics software from Linguamatics to perform targeted influence-mapping studies. She seeks to understand the correlation between sentiment, mined from sources that include event and interview transcripts, presentations, media releases and PubMed biomedical literature about clients' scientific and promotional messaging about therapies. She has concentrated on sources where large volumes of readily mineable information are available; she is exploring adding blogs to the mix.

Jeff Catlin, CEO of text-analytics vendor Lexalytics, describes similar work at Cisco, which he characterizes as his company's best success story. Cisco "used the sentiment engine to determine which executives have the highest correlation to positively moving the stock price when they deliver positive news. They found that certain executives had a positive influence on the markets, while others actually had a negative influence because of the tone of their delivery."

Aafia Chaudhry's 81qd clients are "looking to develop relationships with key opinion leaders," and text-mining along with peer-to-peer network analysis facilitate the task.

Type	Set	Start	End	text
a	Original markups	290	338	{href=http://www.b-eye-network.com/view...
JobTitle		1059	1068	{rule=JobTitle1}
a	Original markups	1072	1076	{href=http://www.81qd.com/, target=_blank}
a	Original markups	1199	1211	{href=http://www.linguamatics.com/, target=...
Person		1686	1697	{gender=male, rule=PersonFinal, rule1=P...
JobTitle		1699	1702	{rule=JobTitle1}
a	Original markups	1728	1738	{href=http://www.lexalytics.com/index.php...

67 Annotations (1 selected)

Document Editor Initialisation Parameters

ANNIE\_0002B run in 0.766 seconds

# Information Extraction

For content analysis, key in on extracting information to databases.

Entities and concepts (features) are like dimensions in a standard BI model. Both classes of object are hierarchically organized and have attributes.

We can have both discovered and predetermined classifications (taxonomies) of text features.

Once you've done information extraction, you can mine the data and create predictive models.



# Applications

Text analytics has applications in –

Intelligence & law enforcement.

Life sciences.

Media & publishing including social-media analysis and contextual advertizing.

Competitive intelligence.

Voice of the Customer: CRM, product management & marketing.

Legal, tax & regulatory (LTR) including compliance.

Recruiting.

Questions?

Discussion?

Thanks!

Seth Grimes

Alta Plana Corporation

301-270-0795 – *http://altaplana.com*

*Alta Plana*