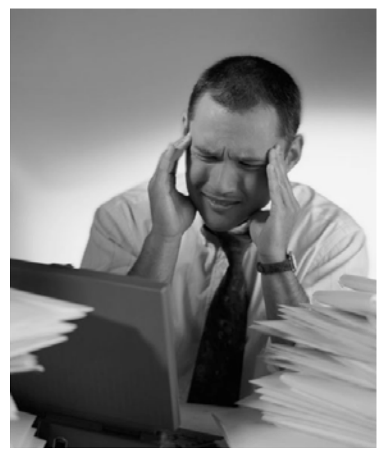
Applied Text Mining

Ronen Feldman

Information Systems Department
School of Business Administration
Hebrew University, Jerusalem, ISRAEL
Ronen.Feldman@huji.ac.il

Motivation

- Rapid proliferation of information available in digital format
- People have **less time** to absorb **more information**
- **Most information is free text, not in structured data**



Outline

- Intro to text mining
  - IR vs. IE
- Information extraction (IE)
  - IE Components
  - Case studies in IE
    - Whizbang!
    - CiteSeer and GoogleScholar
- Relation Extraction/Open IE
  - KnowItAll and SRES
- Blog Mining: Market Structure Surveillance
  - Visualization of extracted data

Text Mining != Search

Find *Documents* matching the Query

- Actual information buried inside documents
- Long lists of documents

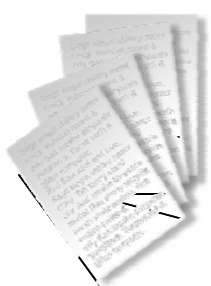
Display *Information* relevant to the Query

- Extract Information from within the documents
- Aggregate over entire collection

Text Mining

Input

Documents



Output

Patterns

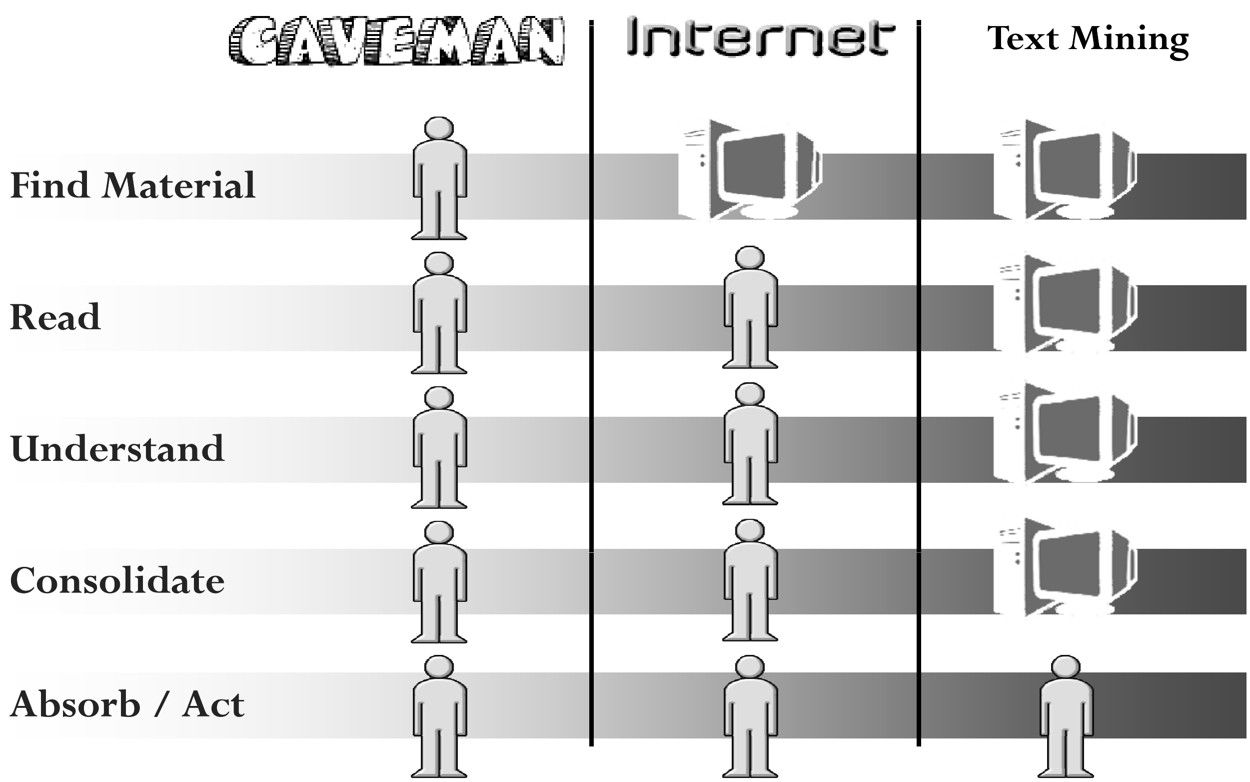
Connections

Profiles

Trends

Seeing the Forest for the Trees

Let Text Mining Do the Legwork for You



What Is Unique in Text Mining?

- **Feature extraction.**
- **Very large number of features that represent each of the documents.**
- **The need for background knowledge.**
- **Even patterns supported by small number of document may be significant.**
- **Huge number of patterns, hence need for visualization, interactive exploration.**

Text Sources

- Comments and notes
  - Physicians, Sales reps.
  - Customer response centers
  - Email
  - Word & PowerPoint documents
- The web
  - blogs
- Journal articles
  - Medline has 13 million abstracts
- Annotations in databases
  - e.g. GenBank, GO, EC, PDB

8

Document Types

- Structured documents
  - Output from CGI
- Semi-structured documents
  - Seminar announcements
  - Job listings
  - Ads
- Free format documents
  - News
  - Scientific papers
  - Blogs

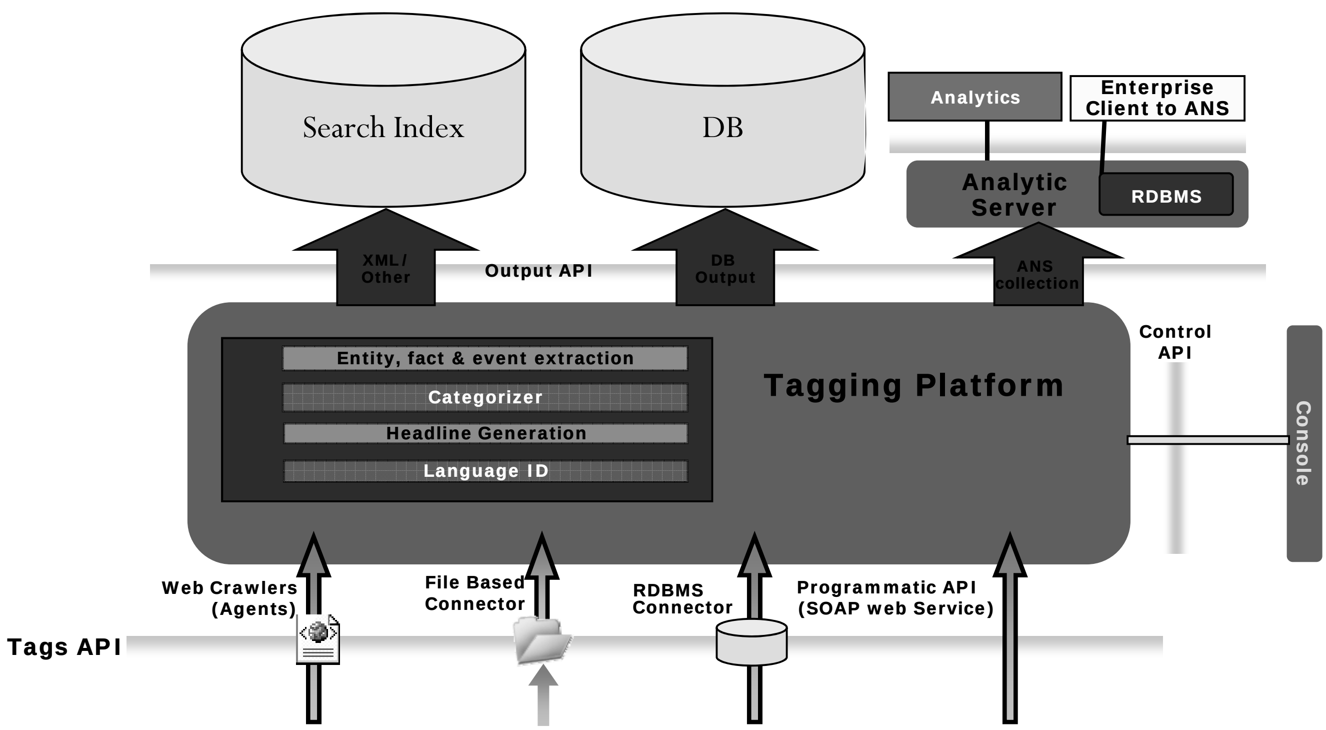
Text Representations

- Character Trigrams
- Words
- Linguistic Phrases
- Non-consecutive phrases
- Frames
- Scripts
- Role annotation
- Parse trees

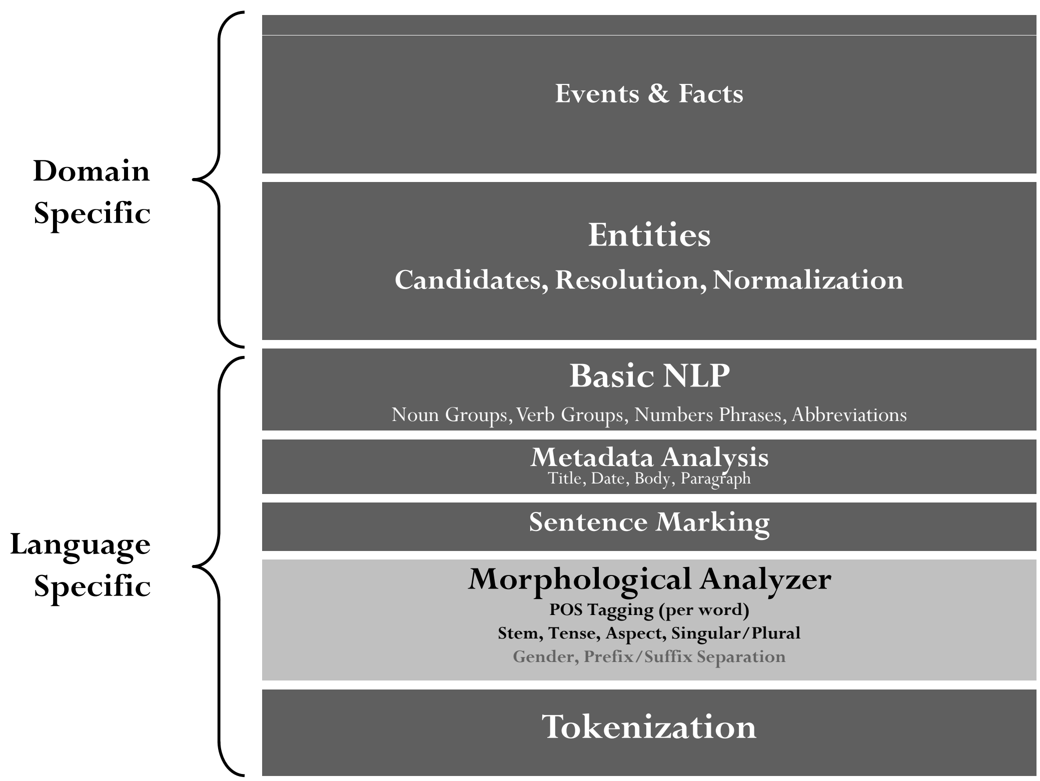
Text Mining: Key Questions

- What can text mining do?
  - What can be done now?
  - What will soon be possible?
- Different types of text mining
  - Information Retrieval (IR)
    - documents
  - Information Extraction (IE)
    - facts
- How well does it work?
  - Why text mining is hard
  - Why text mining is easy

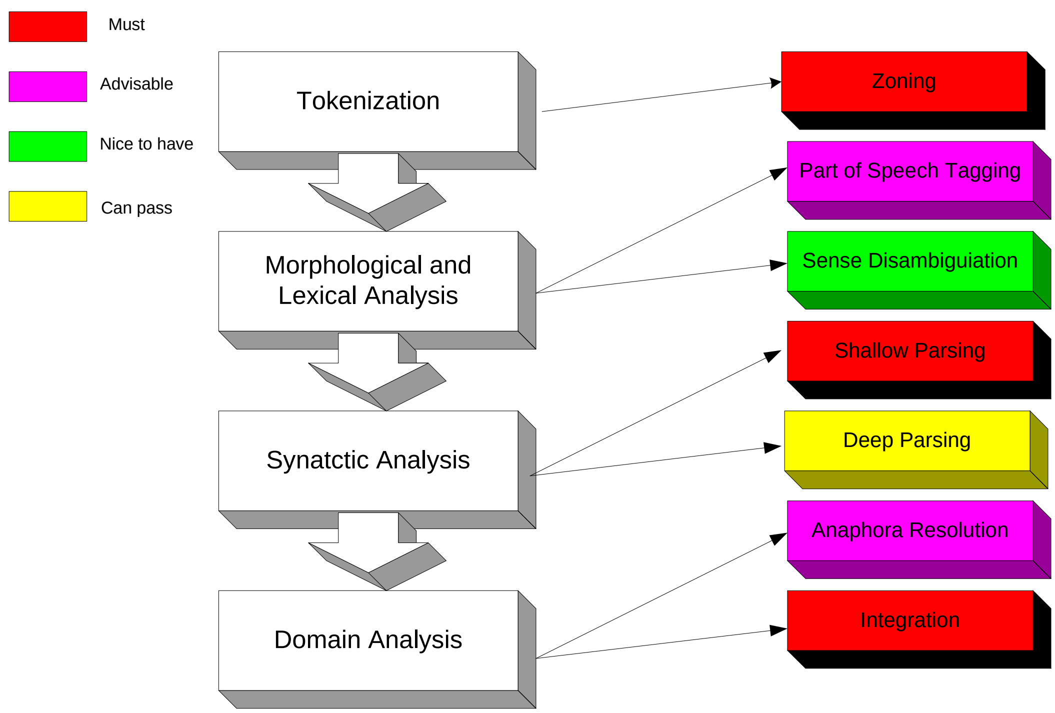
General Architecture



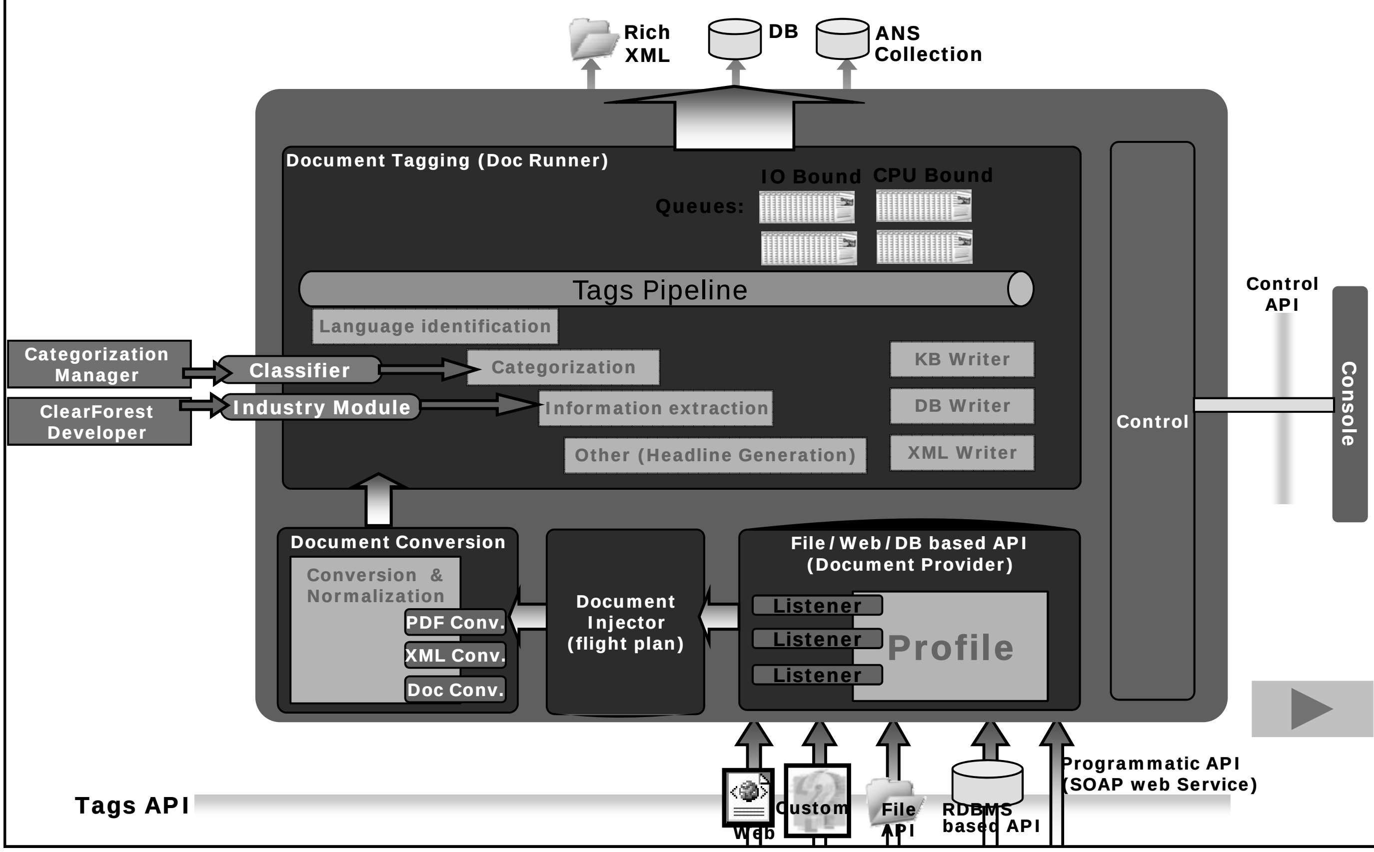
The Language Analysis Stack



Components of IE System



Tagging Architecture



Intelligent Auto-Tagging

(c) 2001, Chicago Tribune.
Visit the Chicago Tribune on the Internet at http://www.chicago.tribune.com/
Distributed by Knight Ridder/Tribune Information Services.
By Stephen J. Hedges and Cam Simpson

…….

The [redacted] is the center of radical Muslim activism in [redacted]. Through its doors have passed at least three of the men now held on suspicion of terrorist activity in [redacted] [redacted] and [redacted], as well as one Algerian man in prison in the [redacted].

``The mosque's chief cleric, [redacted]-[redacted] lost two hands fighting the Soviet Union in Afghanistan and he advocates the elimination of Western influence from Muslim countries. He was arrested in London [redacted] for his alleged involvement in a Yemen bomb plot, but was set free after Yemen failed to produce enough evidence to have him extradited. ."

……

<Facility>**Finsbury Park Mosque**</Facility>

<Country>**England**</Country>

<Country>**France** </Country>

<Country>**England**</Country>

<Country>**Belgium**</Country>

<Country>**United States**</Country>

<Person>**Abu Hamza al-Masri**</Person>

<PersonPositionOrganization>
<OFFLEN OFFSET="**3576**" LENGTH="**33**" />
<Person>**Abu Hamza al-Masri**</Person>
<Position>**chief cleric**</Position>
<Organization>**Finsbury Park Mosque**</Organization>
</PersonPositionOrganization>

<City>**London**</City>

<PersonArrest>
<OFFLEN OFFSET="**3814**" LENGTH="**61**" />
<Person>**Abu Hamza al-Masri**</Person>
<Location>**London**</Location>
<Date>**1999**</Date>
<Reason>**his alleged involvement in a Yemen bomb plot**</Reason>
</PersonArrest>

Business Tagging Example

SAP Acquires Virsa for Compliance Capabilities

By Renee Boucher Ferguson

April 3, 2006

Honing its software compliance skills, SAP announced April 3 the acquisition of Virsa Systems, a privately held company that develops risk management software.

Terms of the deal were not disclosed.

SAP has been strengthening its ties with Microsoft over the past year or so. The two software giants are working on a joint development project, Mendocino, which will integrate some MySAP ERP (enterprise resource planning) business processes with Microsoft Outlook. The first product is expected in 2007.

"Companies are looking to adopt an integrated view of governance, risk and compliance instead of the current reactive and fragmented approach," said Shai Agassi, president of the Product and Technology Group and executive board member of SAP, in a statement. "We welcome Virsa employees, partners and customers to the SAP family."

<Topic>BusinessNews</Topic>

<Company>SAP</Company>

<Company>Virsa Systems</Company>

<IndustryTerm>risk management software</IndustryTerm>

<Acquisition offset="494" length="130">
<Company_Acquirer>SAP</Company_Acquirer>
<Company_Acquired>Virsa Systems </Company_Acquired>
<Status>known</Status>
</Acquisition>

<Company>SAP</Company>

<Company>Microsoft</Company>

<Product>MySAP ERP</Product>

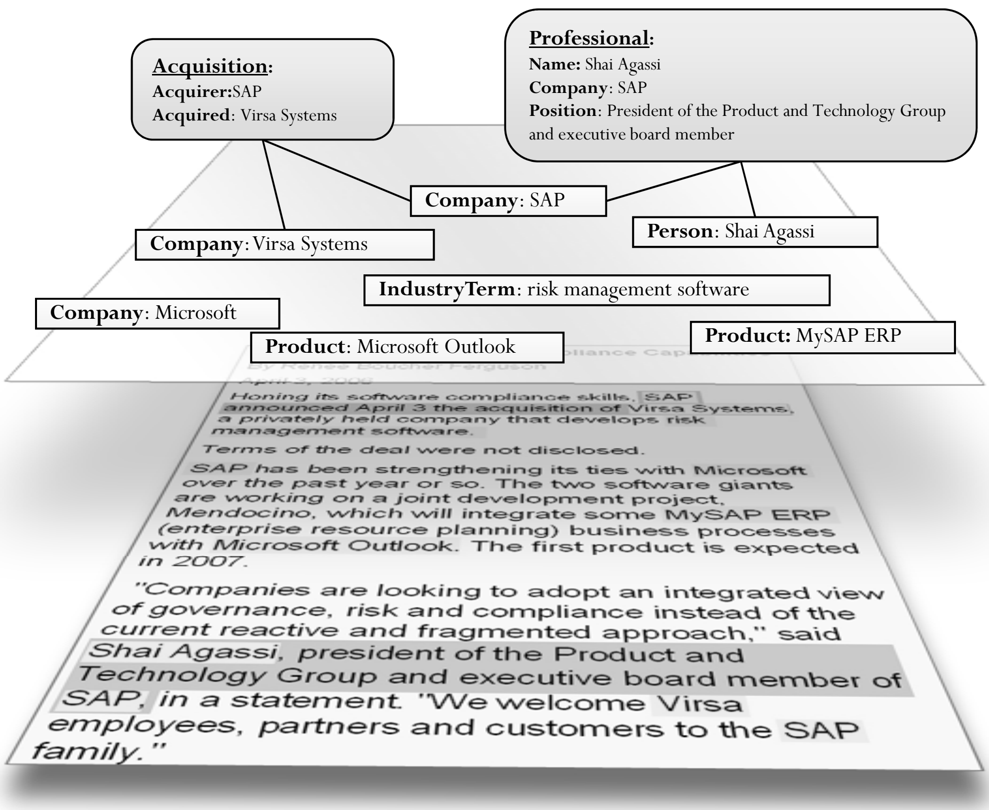
<Product>Microsoft Outlook</Product>

<Person>Shai Agassi</Person>

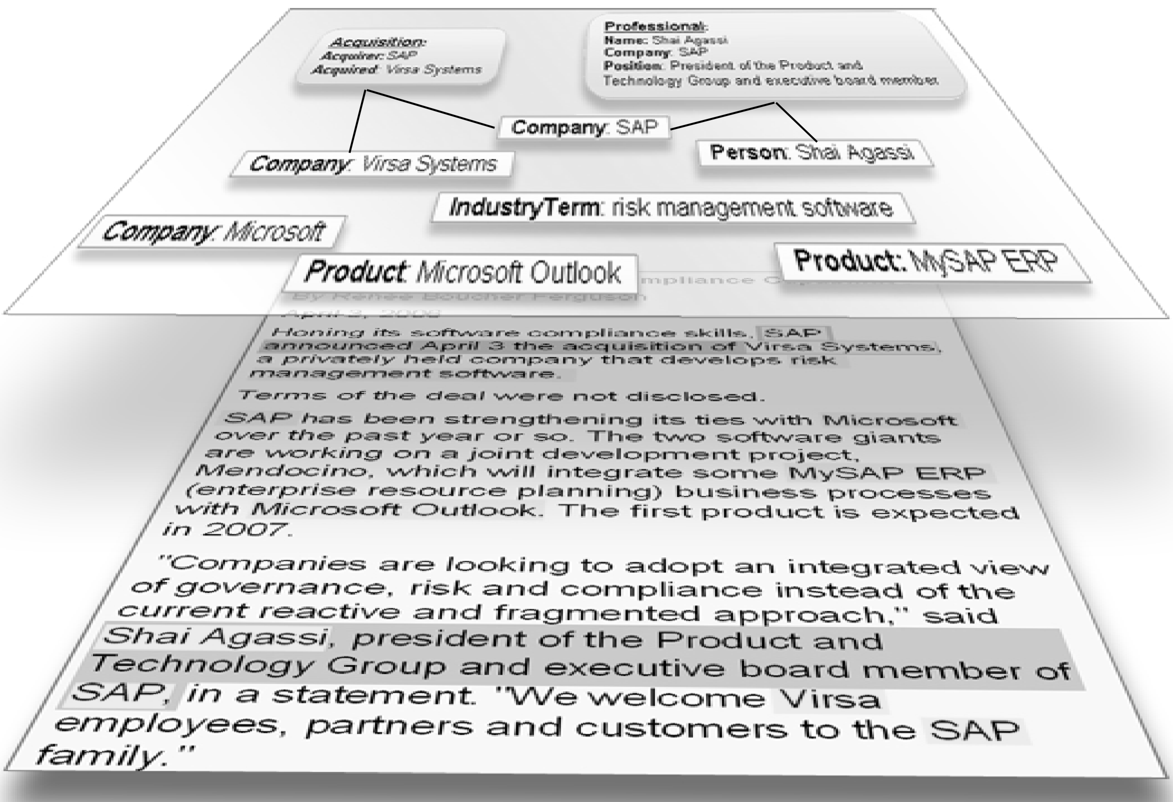
<Company>SAP</Company>

<PersonProfessional offset="2789" length="92">
<Person>Shai Agassi</Person>
<Position>president of the Product and Technology Group and executive board member</Position>
<Company>SAP</Company>
</PersonProfessional>

Business Tagging Example



Business Tagging Example



Leveraging Content Investment

Any type of content
- Unstructured textual content (current focus)
- Structured data; audio; video (future)

In any format
- Documents; PDFs; E-mails; articles; etc
- "Raw" or categorized
- Formal; informal; combination

From any source
- WWW; file systems; news feeds; etc.
- Single source or combined sources

Text mining is hard

- Language is complex
  - Synonyms and Orthonyms
    - *Bush, HEK*
  - Anaphora (and Sortal anaphoric noun phrases)
    - *It, they, the protein, both enzymes*
  - Notes are rarely grammatical
  - Complex structure
    - The first time I bought your product, I tried it on my dog, who became very unhappy and almost ate my cat, who my daughter dearly loves, and then when I tried it on her, she turned blue!

Text mining is hard

- Hand-built systems give poor coverage
  - Large vocabulary
    - Chemicals, genes, names
  - Zipf's law
    - *activate* is common;
      colocalize and *synergize*
      are not
    - Most words are very rare
  - Can't manually list all patterns
- Statistical methods need training data
  - Expensive to manually label data

22

Text mining is easy

- Lots of redundant data
- Some problems are easy
  - IR: bag of words works embarrassingly well
  - LSA (SVD) for grading tests
- Incomplete, inaccurate answers often useful
  - EDA
    - Suggest trends or linkages

Outline

- Intro to text mining
  - IR vs. IE
- Information extraction (IE)
  - IE Components
  - Case studies in IE
    - Whizbang!
    - CiteSeer and GoogleScholar
- Relation Extraction/Open IE
  - KnowItAll and SRES
- Blog Mining: Market Structure Surveillance
- Link Analysis

Information Extraction

Theory and Practice

Why Information Extraction?

xerox

Type	Public (NYSE: XRX)
Founded	Rochester, New York, USA (1906)
Headquarters	Norwalk, Connecticut, USA Offices in Rochester, New York
Key people	Anne M. Mulcahy, Chairman & CEO Ursula Burns, President Larry Zimmerman, CFO Gary R. Kabureck CAO Michael MacDonald, President, Marketing Operations
Industry	Document Services Computer Peripherals
Products	Digital Imaging Printers
Revenue	▲$17.2 billion USD (2007)
Employees	57,400 (2007)
Website	www.xerox.com

"Who is the CEO of Xerox?"

"Female CEOs of public companies"

26

Applications of Information Extraction

- Routing of Information
- Infrastructure for IR and for Categorization
- Event Based Summarization.
- Automatic Creation of Databases
  - Company acquisitions
  - Sports scores
  - Terrorist activities
  - Job listings
  - Corporate titles and addresses

What is Information Extraction?

- IE extracts pieces of information that are salient to the user's needs.
  - Find named entities such as persons and organizations
  - Find find attributes of those entities or events they participate in
  - Contrast IR, which indicates which documents need to be read by a user
- Links between the extracted information and the original documents are maintained to allow the user to reference context.

Relevant IE Definitions

- **Entity:** an object of interest such as a person or organization.
- **Attribute:** a property of an entity such as its name, alias, descriptor, or type.
- **Fact:** a relationship held between two or more entities such as the position of a person in a company.
- **Event:** an activity involving several entities such as a terrorist act, airline crash, management change, new product introduction.

IE Accuracy by Information Type

Information Type	Accuracy
Entities	90-98%
Attributes	80%
Facts	60-70%
Events	50-60%

Information Extraction (IE)

JERUSALEM - A Muslim suicide bomber blew apart 18 people on a Jerusalem bus and wounded 10 in a mirror-image of an attack one week ago. The carnage could rob Israel's Prime Minister Shimon Peres of the May 29 election victory he needs to pursue Middle East peacemaking. Peres declared all-out war on Hamas but his tough talk did little to impress stunned residents of Jerusalem who said the election would turn on the issue of personal security.

IE – Extracted Information

MESSAGE: ID	TST-REU-0001
SECSOURCE: SOURCE	Reuters
SECSOURCE: DATE	March 3, 1996, 11:30
INCIDENT: DATE	March 3, 1996
INCIDENT: LOCATION	Jerusalem
INCIDENT: TYPE	Bombing
HUM TGT: NUMBER	"killed: 18"
	"wounded: 10"
PERP: ORGANIZATION	"Hamas"

32

IE - Method

- Extract raw text (html, pdf, ps, gif)
- Tokenize
- Detect term boundaries
  - We extracted *alpha 1 type XIII collagen* from …
  - Their house council recommended …
- Detect sentence boundaries
- Tag parts of speech (POS)
  - *John*/noun *saw*/verb *Mary*/noun.
- Tag named entities
  - Person, place, organization, gene, chemical
- Parse
- Determine co-reference
- Extract knowledge

33

Approaches for Building IE Systems

- Knowledge Engineering Approach
  - Rules are crafted by linguists in cooperation with domain experts.
  - Most of the work is done by inspecting a set of relevant documents.
  - Can take a lot of time to fine tune the rule set.
  - Best results were achieved with KB based IE systems.
  - Skilled/gifted developers are needed.
  - A strong development environment is a MUST!

Approaches for Building IE Systems

- Automatically Trainable Systems
  - The techniques are based on statistics and use almost no linguistic knowledge
    - Conditional Random Fields (CRFs)
  - They are language independent
  - The main input is an annotated corpus
  - Need a relatively small effort when building the rules, however creating the annotated corpus is extremely laborious.
  - Huge number of training examples is needed in order to achieve reasonable accuracy.
  - Hybrid approaches can utilize the user input in the development loop.

Conclusions

- What doesn't work
  - Anything requiring high precision and full automation
- What does work
  - Text mining with humans "in the loop"
    - Information retrieval
    - Message routing
    - Trend spotting
    - Fraud detection
- What will work
  - Using extracted info in statistical models
  - Speech to text

36

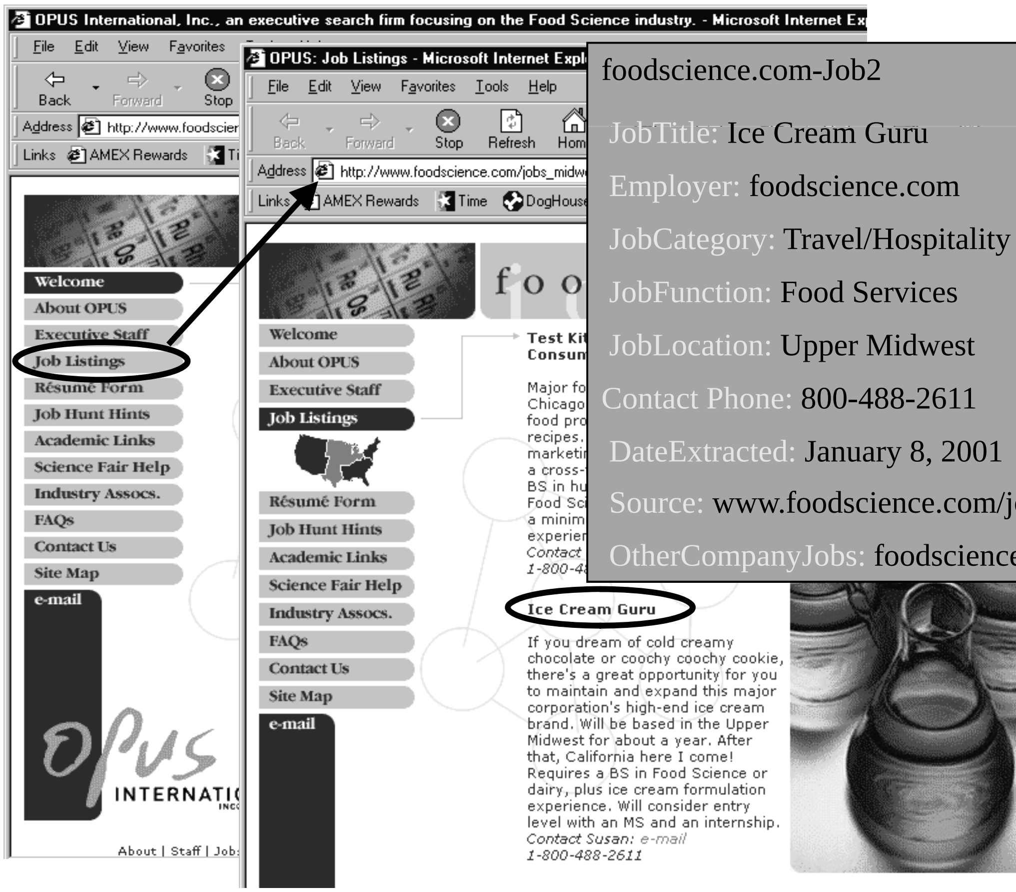
Case studies in Info. Extraction

- Whizbang!
- CiteSeer and GoogleScholar

Whizbang!

- A leading information extraction company
- Now closed.
- What did they do?
- What lessons can we draw?

Extracting Job Openings from the Web



foodscience.com-Job2

JobTitle: Ice Cream Guru

Employer: foodscience.com

JobCategory: Travel/Hospitality

JobFunction: Food Services

JobLocation: Upper Midwest

Contact Phone: 800-488-2611

DateExtracted: January 8, 2001

Source: www.foodscience.com/jobs_midwest.html

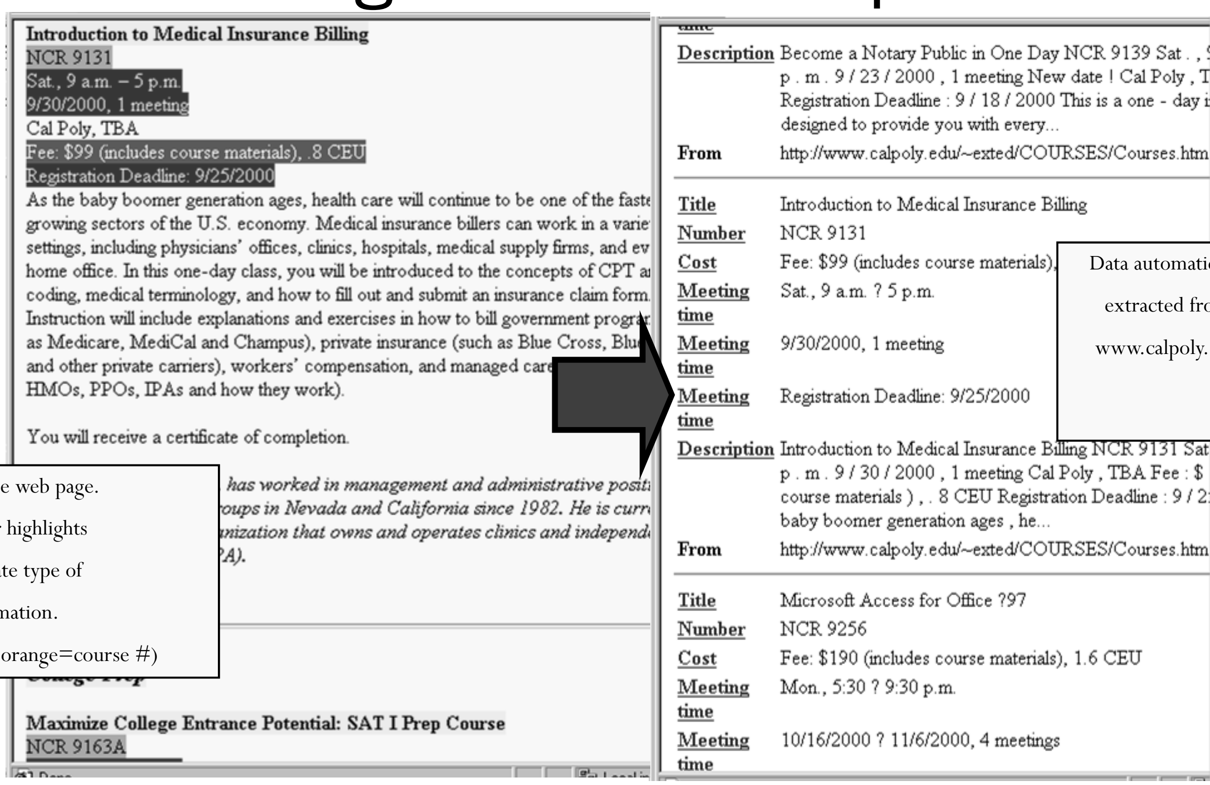
OtherCompanyJobs: foodscience.com-Job1

39

FlipDog


job search find employment careers @ FlipDog.com free! - Microsoft Internet Explorer
Address http://www.flipdog.com/home.html
Go
File Edit View Favorites Tools Help
Links
FlipDog.com
Home
Find Jobs
Your Account
Resource Center
• Support • Employers
Job Search at FlipDog.com: Employment & Career Management
647,514
Job Opportunities
from 53,641 Employers
Find a Job!
Post Your Resume
Employers
click here for
Products & Services
Pigskin Places
Health Care in NY 2,770
Health Care in MD 1,262
Sales in NY 3,751
Sales in MD 958
Computing in NY 8,050
Computing in MD 4,114
Jobs for Sports Fans
Head Football Coach
Football Coach
Asst. Football Coach
High School Football Coach
Univ. Asst. Football Coach
Job Seeker Newsletter
Enter your e-mail address:
Sign Me Up!
"Top 100 Web Sites"
PC Magazine, Nov. 2000
"Top 10 Career Web Site"
Media Metrix, Sept. 2000
"Top 10 Job Site"
Showcase Jobs
MRI
Management Recruiters of Charlotte North
We provide total staffing solutions in the areas of Human Resources, Compensation, Web-based HR self-service, and Customer Management Systems.
Learn More
Looking for a Vice President of Academic Affairs to oversee planning, operation and evaluation of the college's academic programs.
Learn More
powered by WhizBang!
Job Seekers: Find your dream job!
Check our 'Best Places to Find a Job' January report.
Open your FREE account and put your resume online.
Search 24x7 with our FREE automatic JobHunters™.
Research our database of over 50,000 employers.
Get expert advice at our new Resource Center.
Access salary surveys/calculators, relocation tools, networking opportunities, & training/testing tools.
Use FlipDog.com to search jobs right from your desktop! Download Snippets today!
Internet

Extracting Course Descriptions



Introduction to Medical Insurance Billing
NCR 9131
Sat., 9 a.m. – 5 p.m.
9/30/2000, 1 meeting
Cal Poly, TBA
Fee: $99 (includes course materials), .8 CEU
Registration Deadline: 9/25/2000
As the baby boomer generation ages, health care will continue to be one of the faste
growing sectors of the U.S. economy. Medical insurance billers can work in a varie
settings, including physicians' offices, clinics, hospitals, medical supply firms, and ev
home office. In this one-day class, you will be introduced to the concepts of CPT a
coding, medical terminology, and how to fill out and submit an insurance claim form.
Instruction will include explanations and exercises in how to bill government progra
as Medicare, MediCal and Champus), private insurance (such as Blue Cross, Blu
and other private carriers), workers' compensation, and managed care
HMOs, PPOs, IPAs and how they work).

You will receive a certificate of completion.

has worked in management and administrative positi
oups in Nevada and California since 1982. He is curr
nization that owns and operates clinics and independ
A).

Maximize College Entrance Potential: SAT I Prep Course
NCR 9163A

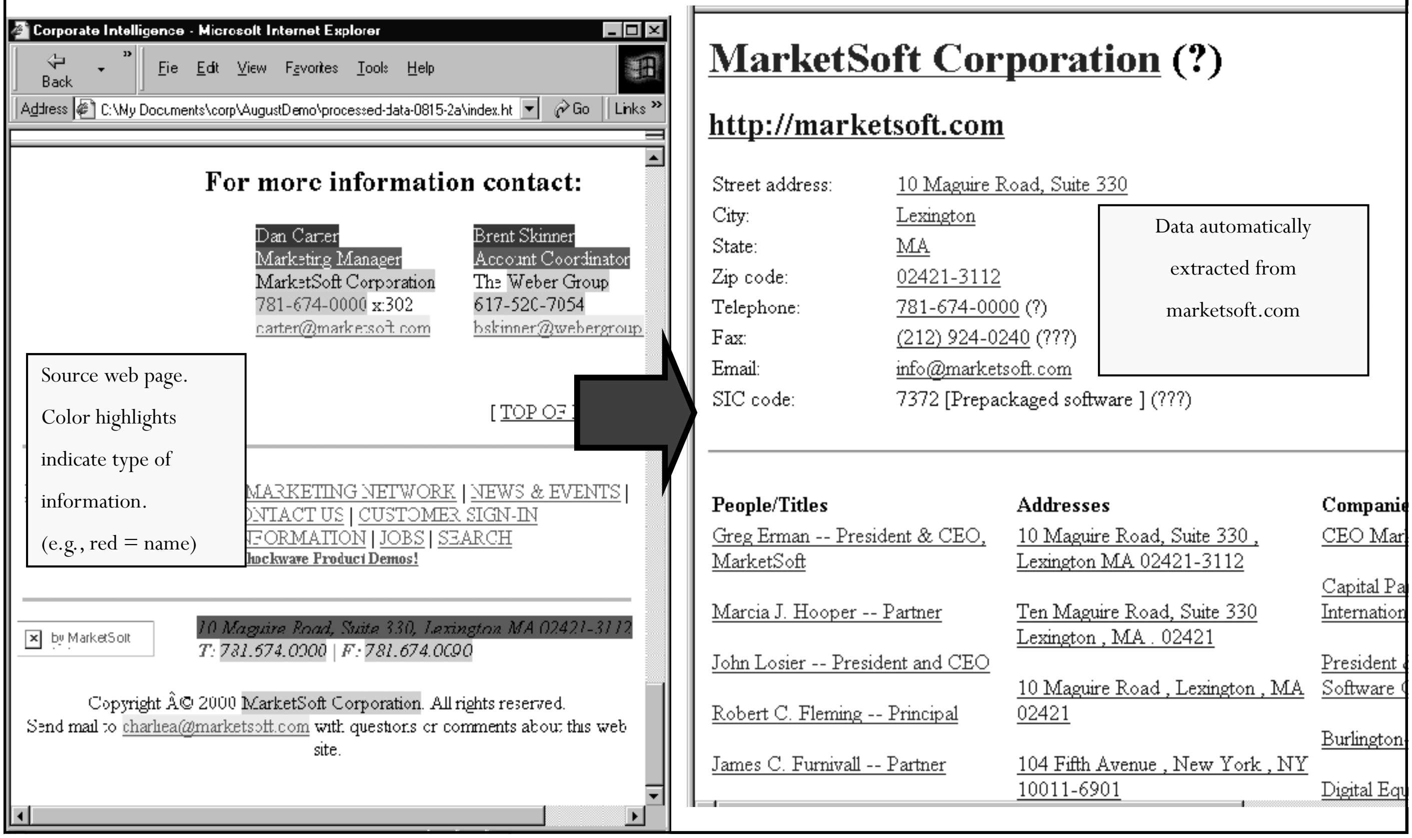
Description	Become a Notary Public in One Day NCR 9139 Sat . , p . m . 9 / 23 / 2000 , 1 meeting New date ! Cal Poly , T Registration Deadline : 9 / 18 / 2000 This is a one - day designed to provide you with every...
From	http://www.calpoly.edu/~exted/COURSES/Courses.htm
Title	Introduction to Medical Insurance Billing
Number	NCR 9131
Cost	Fee: $99 (includes course materials),
Meeting time	Sat., 9 a.m. ? 5 p.m.
Meeting time	9/30/2000, 1 meeting
Meeting time	Registration Deadline: 9/25/2000
Description	Introduction to Medical Insurance Billing NCR 9131 Sat p . m . 9 / 30 / 2000 , 1 meeting Cal Poly , TBA Fee : $ course materials) , . 8 CEU Registration Deadline : 9 / 2 baby boomer generation ages , he...
From	http://www.calpoly.edu/~exted/COURSES/Courses.htm
Title	Microsoft Access for Office ?97
Number	NCR 9256
Cost	Fee: $190 (includes course materials), 1.6 CEU
Meeting time	Mon., 5:30 ? 9:30 p.m.
Meeting time	10/16/2000 ? 11/6/2000, 4 meetings

Source web page.
Color highlights
indicate type of
information.
(e.g., orange=course #)

Data automatically
extracted from
www.calpoly.edu

41

Extracting Corporate Information



Why did Whizbang fail?

- People won't pay for info from the web
  - Technology rather than solution
- Too much cost for too little value
  - IE is inaccurate
  - High accuracy requires major human post-processing
  - Each application required major software development

data mining - ResearchIndex document query 10/26/2005 10:23 PM

CiteSeer Find: data mining [Documents] [Citations]

Searching for **PHRASE** **data mining**.
Restrict to: Header Title Order by: Expected citations Hubs Usage Date Try: Google (CiteSeer) Google (Web) Yahoo! MSN CSB DBLP
8330 documents found. **Only retrieving 1000 documents.** Retrieving documents... **Order: number of citations.**

A Tutorial on Support Vector Machines for Pattern Recognition - Burges (1998) (369 citations)
Conference on Knowledge Discovery &**Data Mining**. AAAI Press, Menlo Park, CA, 1995. B.
support vector machines for pattern recognition. **Data Mining** and Knowledge Discovery, 2(2)955-974, 1998. A
www.ai.mit.edu/courses/6.893/papers/tutorial_web_page.ps

Mining Generalized Association Rules - Srikant, Agrawal (1995) (253 citations)
Zurich, Swizerland, 1995 1 Introduction **Data mining**, also known as knowledge discovery in
www.almaden.ibm.com/cs/people/srikant/papers/vldb95_rj.ps

Dynamic Itemset Counting and Implication Rules for.. - Brin, Motwani, Ullman, .. (1997) (222 citations)
the results. 1 Introduction Within the area of **data mining**, the problem of deriving associations from
baskets. There are numerous applications of **data mining** which fit into this framework. The canonical
www-ai.cs.uni-dortmund.de/LEHRE/DATAWAREHOUSE98/Brin_etal_97a.ps.gz

Fast Subsequence Matching in Time-Series Databases - Faloutsos, Ranganathan.. (1994) (222 citations)
hypothesis testing and, in general, in `**data mining**' 1, 3, 4] and rule discovery. For the rest of
www.cse.cuhk.edu.hk/~unprog/csc5120/Papers/sigmod94.ps

An Optimal Algorithm for Approximate Nearest.. - Arya, Mount.. (1994) (210 citations)
applications, including knowledge discovery and **data mining** [FPSSU96]pattern recognition and
Uthurusamy. Advances in Knowledge Discovery and **Data Mining**. AAAI Press/Mit Press, 1996. Fre85] G. N.
www.cs.ust.hk/faculty/arya/pub/ANN.ps

Efficient and Effective Clustering Methods for Spatial Data Mining - Ng, Han (1994) (206 citations)
and Effective Clustering Methods for Spatial **Data Mining** Raymond T. Ng Department of Computer Science
V5A 1S6, Canada han@cs.sfu.ca Abstract Spatial **data mining** is the discovery of interesting relationships
ftp.fas.sfu.ca/pub/cs/han/kdd/vldb94.ps

data mining - Google Scholar 10/26/2005 10:26 PM

Google Scholar BETA

data mining [Search]

Advanced Scholar Search
Scholar Preferences
Scholar Help

Scholar Results **1** - **10** of about **402,000** for **data mining** [definition]. (**0.04** seconds)

An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants
E Bauer, R Kohavi, D **Mining**, SGI Visualization - Machine Learning, 1999 - kluweronline.com
... Our decision was to generate a Bootstrap sample from the original **data** S and continue up to a limit of 25 such samples at a given trial; such a limit was never ...
Cited by 453 - Web Search - metet.polsl.katowice.pl - robotics.stanford.edu - cs.utsa.edu - all 15 versions »

Data Mining: Concepts and Techniques
J Han, M Kamber, P Methods, H Methods, DB Methods, … - SIGMOD Record, 2002 - portal.acm.org
Page 1. **Data Mining**: Concepts and Techniques ... Any method used to extract patterns from a given **data** source is considered to be a **data mining** technique. ...
Cited by 1394 - Web Search - ir.iit.edu - cs.clemson.edu - ifsc.ualr.edu - all 18 versions »

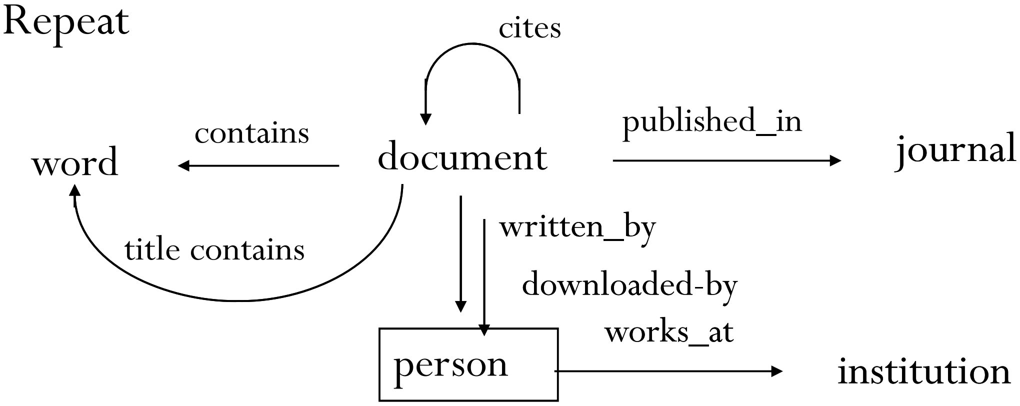
[BOOK] Advances in Knowledge Discovery and **Data Mining**
UM Fayyad, G Piatetsky-Shapiro, P Smyth, R … - 1996 - MIT Press
Cited by 1235 - Web Search - Library Search

From **Data Mining** to Knowledge Discovery: An Overview
UM Fayyad, G Piatetsky-Shapiro, P Smyth - … in knowledge discovery and **data mining** table of contents, 1996 - portal.acm.org
... From **data mining** to knowledge discovery: an overview. Source, Advances in knowledge discovery and **data mining** table of contents. Pages: 1 - 34. ...
Cited by 794 - Web Search - research.microsoft.com - galaxy.gmu.edu - ingentaconnect.com - all 7 versions »

[BOOK] The elements of statistical learning: **data mining**, inference, and prediction ·5
T Hastie, T Hastie, R Tibshirani, JH Friedman - 2001 - www-stat-class.stanford.edu
Page 1. Book Reviews 567 The Elements of Statistical Learning: **Data Mining**,

Building CiteSeer

- Pick seed URLs
- Spider the web
- Grab files
- Extract info
- Repeat



46

CiteSeer vs. GoogleScholar

- CiteSeer: A specialized search engine for computer science articles built by NEC
  - Searches the web for information
  - Run by academics
- GoogleScholar: a piece of Google
  - Uses proprietary data from publishers

Relation Extraction

"Oh dear."

IE for the Web

Challenges	Advantages
Difficult, ungrammatical sentences	"Semantically tractable" sentences
Unreliable information	Redundancy
Heterogeneous corpus	
Massive Number of Relations	**Open IE** [Banko, *et al.* 2007]

TextRunner Search

http://www.cs.washington.edu/research/textrunner/

TextRunner Search

[Banko et al., 2007]

Retrieved **2760** results for **What kills bacteria?**.

Grouping results by predicate. Group by: argument 1 | argument 2

kills - 42 results

strong antibiotics (103), Antibiotics (67), Benzoyl peroxide (50), ***175 more...*** **kills bacteria**
Ultraviolet disinfection devices (3), ozone (3), iodine (2), ***7 more...*** may **kill bacteria** and viruses
Levaquin (21) **kills** a variety of **bacteria**
INH (4), the medicine (4) **kills** the TB **bacteria**
many antibiotics (3), Antibiotics (2), the " bad " bacteria (2) also **kills** the " good " **bacteria**
Infact Doxy (4), only the Doxy (2) **kills** a whole bunch of various **bacteria**
Treatment (4), Penicillin treatment (2) will **kill** the syphilis bacterium
SILVER (3), our disinfectant solution (2) **kills** almost all known **bacteria**
boiling (2), boil-water alerts (2) will **kill bacteria** and parasites
a food (2), antibiotics (2) can **kill** all **bacteria**
Anti-bacterial cleaners (4) **kills** 99.9 % of **bacteria** Cleans appliances
Appropriate treatment (4) **kills** the Shigella **bacteria**
artemisinin (3) can **kill** other parasites and **bacteria**
the chlorine dioxide (3) **kills** the already formed **bacteria**
this mouthwash (3) **kills** germs and **bacteria**
those drugs (3) **killed** Andrew 's normal gut-protective **bacteria**
Antibiotics (3) **kill** gonorrhea **bacteria**
Proper cooking (3) **kills** food poisoning **bacteria**
that microwaves (2) can **kill** the anthrax **bacteria**
Hot dry vapor steam (2) **kills** mold , mildew , viruses , **bacteria**
One application (2) **kills bacteria** odors
Benzoyl peroxide (2) **kills** off **bacteria**
Iodine (2) will **kill** the lactic **bacteria**
the boiling (2) **kills** impurities and **bacteria**
The chlorine (2) **kills** iron **bacteria**
ozone (2) **kills** the acid producing **bacteria**
Ampicillin (2) **kills** susceptible **bacteria**
Any positively offset frequency (2) **kills** all **bacteria** , viruses and parasites

Find: Previous Next Highlight all Match case

Transferring data from turingc.cs.washington.edu...

does not kill - 1 result

Doxycycline (14), Freezing (11), Refri

to kill - 6 results

antibiotics (7), water (3), milk (3), ***2 n***

antibiotics (2) to **kill** extracellular **bac**

the ability (2) to **kill** a wide variety of

milk (2) to **kill** harmful **bacteria**

a second time (2) to **kill** any **bacteri**

macrophages (2) to **kill** the intracellu

helps kill - 2 results

Raw garlic (2), lime juice (2), uv germ

Benzoyl peroxide (3) helps **kill** skin **b**

does n't kill - 1 result

Freezing (6), Irradiation (4), antacids (2) does n't **kill bacteria**

kill not only - 1 result

Antibiotics (6), these drugs (3) **kill** not only harmful **bacteria**

x

Refrigeration and freezing do not kill bacteria, but slow th...

Refrigeration does not kill bacteria and can not improve food quality.

Refrigeration and freezing do not kill bacteria, but s...

Refrigeration and freezing do not kill bacteria, but only slow their growth.

Refrigeration and freezing do not kill bacteria, but sl...

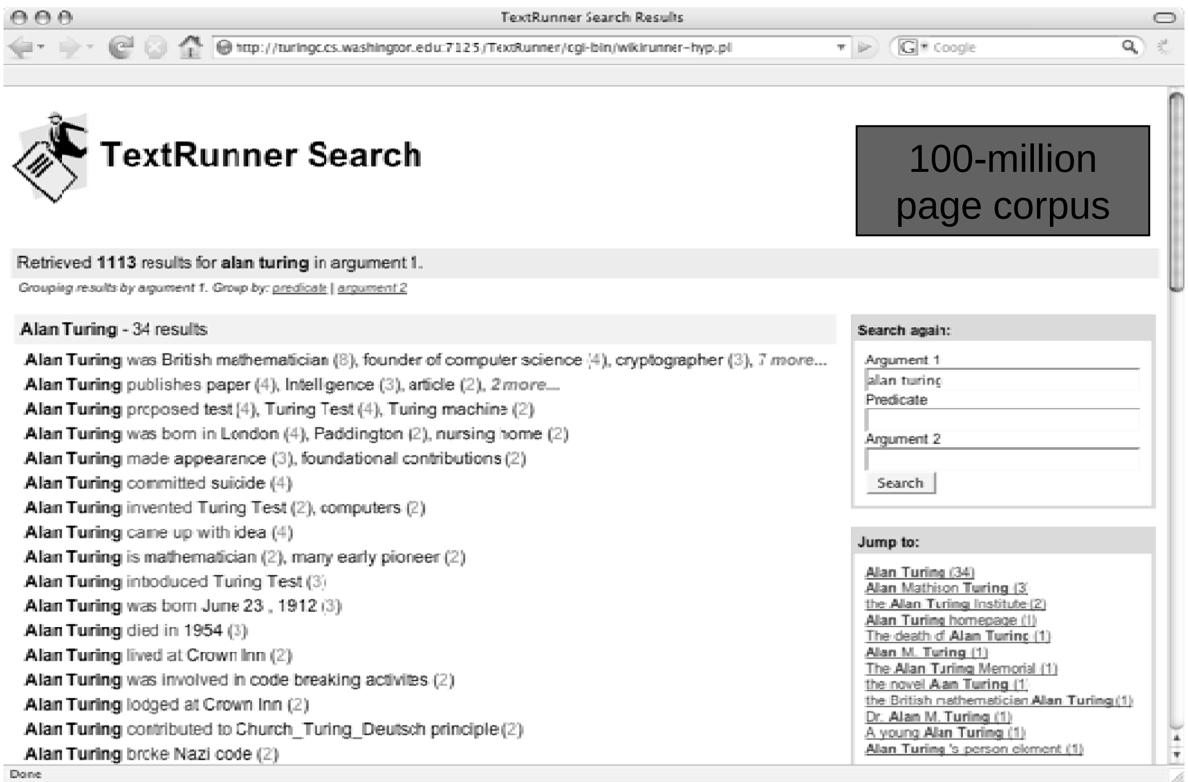
Refrigeration does not kill most bacteria.

Refrigeration and freezing do not kill bacteria, but slow their growth.

Remember: refrigeration does not kill bacteria; it only slows down their growth .

TextRunner

[Banko, Cafarella, Soderland, *et al.*, IJCAI '07]



100-million page corpus

53

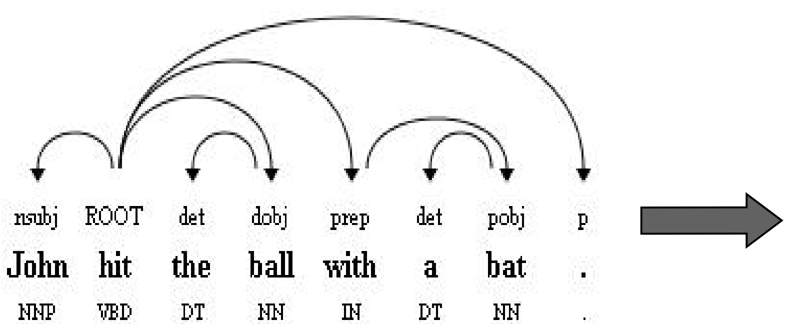
Open IE

- Relation-Independent Extraction
  - How are relations expressed, in general?
  - Unlexicalized
- Self-Supervised Training
  - Automatically label training examples
- Discover relations on the fly
  - Traditional IE: $(e_1, e_2) \in R$?
  - Open IE: ***What* is R?**

54

Training

- No parser at extraction time
- Use trusted parses to auto-label training examples
- Describe instances without parser-based features
  - Unlexicalized   PennTB ok



\+ (John, hit, ball)
\+ (John, hit with, bat)
\- (ball, with, bat)

Features

- Unlexicalized
  - Closed class words OK
- Parser-free
  - Part-of-speech tags, phrase chunk tags
  - ContainsPunct, StartsWithCapital, …
- Type-independent
  - Proper vs. common noun, no NE types

Relation Discovery

- Many ways to express one relation
- Resolver [Yates & Etzioni, HLT '07]

```
(Viacom, acquired, Dreamworks)
(Viacom, 's acquisition of, Dreamworks)
(Viacom, sold off, Dreamworks)

(Google, acquired, YouTube)
(Google Inc., 's acquisition of, YouTube)

(Adobe, acquired, Macromedia)
(Adobe, 's acquisition of, Macromedia)
```

$$P(R_1 = R_2) \sim \text{shared objects} * \text{strSim}(R_1,R_2)$$

57

IE vs. Open IE

	Traditional IE	Open IE
Input	Corpus + Relations + Training Data	Corpus + Relation-Independent Heuristics
Relations	Specified in Advance	Discovered Automatically
Features	Lexicalized, NE-Types	Unlexicalized, No NE types

58

Questions

- How does OIE fare when relation set is unknown?
- Is it even possible to learn relation-independent extraction patterns?
- How do OIE and Traditional IE compare when the relation is given?

59

Eval 1: Open Info. Extraction (OIE)

- OIE with Graphical Models (CRF) vs. Classifiers (Naïve Bayes)
- Apply to 500 sentences from Web IE training corpus [Bunescu & Mooney '07]

O-NB			O-CRF		
P	R	F1	P	R	F1
86.6	23.2	36.6	**88.3**	**45.2**	**59.8**

Category	Pattern	RF
Verb	E_1 Verb E_2 *X established Y*	37.8
Noun+Prep	E_1 NP Prep E_2 *the X* settlement with *Y*	22.8
Verb+Prep	E_1 Verb Prep E_2 X moved to Y	16.0
Infinitive	E_1 to Verb E_2 *X to acquire Y*	9.4
Modifier	E_1 Verb E_2 NP *X is Y winner*	5.2
Coordinate_n	E_1 (and|,|-|:) E_2 NP *X - Y deal*	1.8
Coordinate_v	E_1 (and|,) E_2 Verb *X , Y merge*	1.0
Appositive	E_1 NP (:|,)? E_2 *X hometown : Y*	0.8

Relation-Independent Patterns

- 95% could be grouped into 1 of 8 categories
- Dangerously simple
  - `Paramount , the` **`Viacom`** `- owned studio , bought` **`Dreamworks`**
  - **`Charlie Chaplin`** `, who died in 1977 , was born in` **`London`**
- Precise conditions
  - Difficult to specify by hand
  - Learnable by OIE model

62

Results

Category	O-NB			O-CRF		
	P	R	F1	P	R	F1
Verb	100.0	38.6	55.7	93.9	65.1	76.9
Noun+Prep	100.0	9.7	17.5	89.1	36.0	51.2
Verb+Prep	95.2	25.3	40.0	95.2	50.0	65.6
Infinitive	100.0	25.5	40.7	95.7	46.8	62.9
Other	0	0	0	0	0	0
All	86.6	23.2	36.6	**88.3**	**45.2**	**59.8**

63

Traditional IE with R1-CRF

- Trained from hand-labeled data *per relation*
- Lexicalized features, same graph structure
- Yes, many existing RE systems

  [*e.g.* Bunescu ACL '07, Culotta HLT '06]

  but want to isolate effects of

  - Relation-specific/independent features
  - Supervised vs. Self-supervised Training

  keeping underlying models equivalent

64

Eval 2: Targeted Extraction

- Web IE corpus from [Bunescu 2007]
  - Corporate-acquisitions (3042)
  - Birthplace (1853)
- Collected 2 more relations in same manner
  - Invented-Product (682)
  - Won-Award (354)
- Labeled examples by hand

Results

	R1-CRF			O-CRF	
Relation	**P**	**R**	**Train Ex**	**P**	**R**
Acquisition	67.6	69.2	3042	75.6	19.5
Birthplace	92.3	64.4	1853	90.6	31.1
InventorOf	81.3	50.8	682	88.0	17.5
WonAward	73.6	52.8	354	62.5	15.3
All	73.9	**58.4**	5931	**75.0**	18.4

Open IE can match precision of supervised IE ***without***

- Relation-specific training
- 100s or 1000s of examples *per relation*

66

Summary

- Open IE
  - High-precision extractions without cost of per-relation training
  - Essential when number of relations is large or unknown
- May prefer Traditional IE when
  - High recall is necessary
  - For a small set of relations
  - *And* can acquire labeled data
- Try it!

  http://www.cs.washington.edu/research/textrunner

67

Outline

- Intro to text mining
  - IR vs. IE
- Information extraction (IE)
  - IE Components
  - Case studies in IE
    - Whizbang!
    - CiteSeer and GoogleScholar
    - KDD Cup 2002
- Relation Learning / Open IE
  - KnowItAll and SRES
- Blog Mining: Market Structure Surveillance
- Link Analysis

Self-Supervised Relation Learning from the Web

KnowItAll (KIA)

- KnowItAll is a system developed at University of Washington by Oren Etzioni and colleagues (Etzioni, Cafarella et al. 2005).
- KnowItAll is an autonomous, domain-independent system that extracts facts from the Web. The primary focus of the system is on extracting entities (unary predicates), although KnowItAll is able to extract relations (N-ary predicates) as well.
- The input to KnowItAll is a set of entity classes to be extracted, such as "city", "scientist", "movie", etc., and the output is a list of entities extracted from the Web.

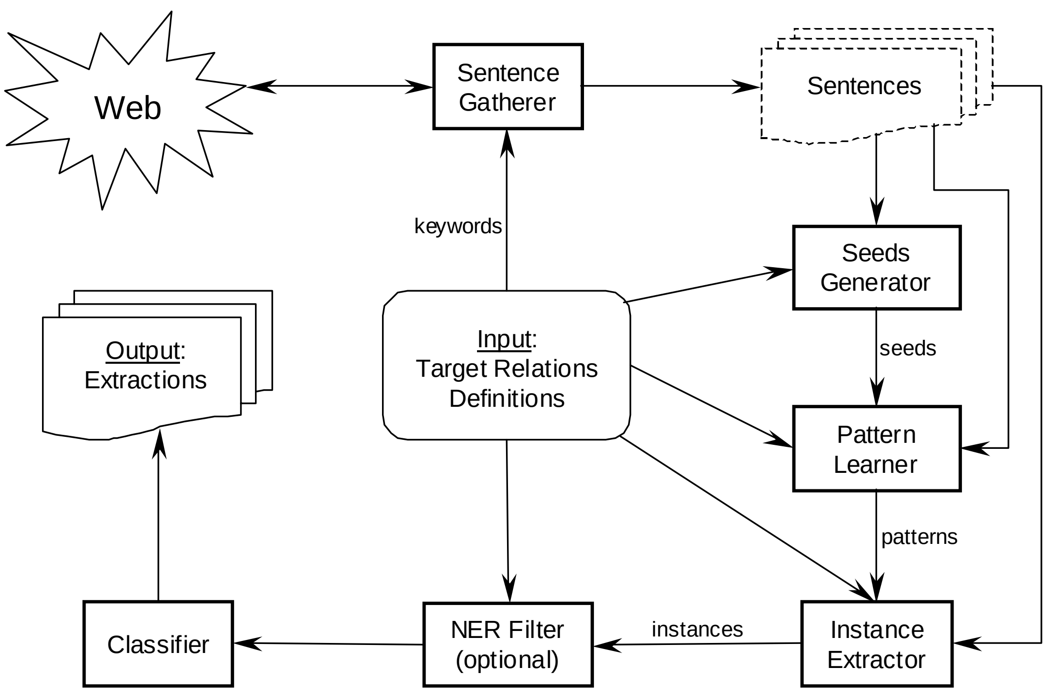
KnowItAll's Relation Learning

- The base version of KnowItAll uses only the generic hand written patterns. The patterns are based on a general Noun Phrase (NP) tagger.
- For example, here are the two patterns used by KnowItAll for extracting instances of the ***Acquisition*(*Company*, *Company*)** relation:
  - NP2 "was acquired by" NP1
  - NP1 "'s acquisition of" NP2
- And the following are the three patterns used by KnowItAll for extracting the ***MayorOf*(*City*, *Person*)** relation:
  - NP ", mayor of" <city>
  - <city> "'s mayor" NP
  - <city> "mayor" NP

SRES

- SRES (**Self-Supervised Relation Extraction System**) which learns to extract relations from the web in an unsupervised way.
- The system takes as input the name of the relation and the types of its arguments and returns as output a set of instances of the relation extracted from the given corpus.

SRES Architecture



Seeds for Acquisition

- Oracle – PeopleSoft
- Oracle – Siebel Systems
- PeopleSoft – J.D. Edwards
- Novell – SuSE
- Sun – StorageTek
- Microsoft – Groove Networks
- AOL – Netscape
- Microsoft – Vicinity
- San Francisco-based Vector Capital – Corel
- HP – Compaq

Positive Instances

- The positive set of a predicate consists of sentences that contain an instance of the predicate, with the actual instance's attributes changed to "<*AttrN*>", where *N* is the attribute index.
- For example, the sentence
  - *"The Antitrust Division of the U.S. Department of Justice evaluated the likely competitive effects of Oracle's proposed acquisition of PeopleSoft."*
- will be changed to
  - *"The Antitrust Division… …….effects of <Attr1>'s proposed acquisition of <Attr2>."*

Negative Instances II

- We generate the negative set from the sentences in the positive set by changing the assignment of one or both attributes to other suitable entities in the sentence.
- In the shallow parser based mode of operation, any suitable noun phrase can be assigned to an attribute.

Examples

- *The Positive Instance*
  - *"The Antitrust Division of the U.S. Department of Justice evaluated the likely competitive effects of <Attr1>'s proposed acquisition of <Attr2>"*
- *Possible Negative Instances*
  - *<Attr1> of the <Attr2> evaluated the likely…*
  - *<Attr2> of the U.S. … …acquisition of <Attr1>*
  - *<Attr1> of the U.S. … …acquisition of <Attr2>*
  - *The Antitrust Division of the <Attr1> ….. acquisition of <Attr2>"*

Pattern Generation

- The patterns for a predicate *P* are generalizations of pairs of sentences from the positive set of *P*.
- The function *Generalize*(*S*1, *S*2) is applied to each pair of sentences *S*1 and *S*2 from the positive set of the predicate. The function generates a pattern that is the best (according to the objective function defined below) generalization of its two arguments.
- The following pseudo code shows the process of generating the patterns:

```
For each predicate P
    For each pair S1, S2 from PositiveSet(P)
        Let Pattern = Generalize(S1, S2).
        Add Pattern to PatternsSet(P).
```

Example

- *S1 = "Toward this end, <Arg1> in July acquired <Arg2>"*
- *S2 = "Earlier this year, <Arg1> acquired <Arg2>"*
- After the dynamical programming-based search, the following match will be found:

Toward		(cost 2)
	Earlier	(cost 2)
this	*this*	(cost 0)
end		(cost 2)
	year	(cost 2)
,	,	(cost 0)
<Arg1 >	*<Arg1 >*	(cost 0)
in July		(cost 4)
acquired	*acquired*	(cost 0)
<Arg2 >	*<Arg2 >*	(cost 0)

Generating the Pattern

- at total cost = 12. The match will be converted to the pattern
  - * * *this* * * , *<Arg1>* * *acquired* *<Arg2>*
- which will be normalized (after removing leading and trailing skips, and combining adjacent pairs of skips) into
  - *this* * , *<Arg1>* * *acquired* *<Arg2>*

Post-processing, filtering, and scoring of patterns

- In the first step of the post-processing we remove from each pattern all function words and punctuation marks that are surrounded by skips on both sides. Thus, the pattern from the example above will be converted to

, *<Arg1> * acquired <Arg2>*

- Note, that we do not remove elements that are adjacent to meaningful words or to slots, like the comma in the pattern above, because such anchored elements may be important.

Content Based Filtering

- Every pattern must contain at least one word relevant to its predicate. For each predicate, the list of relevant words is automatically generated from WordNet by following all links to depth at most 2 starting from the predicate keywords. For example, the pattern

  *<Arg1> * by <Arg2>*
- will be removed, while the pattern

  *<Arg1> * purchased <Arg2>*
- will be kept, because the word "*purchased*" can be reached from "*acquisition*" via synonym and derivation links.

Scoring the Patterns

- The filtered patterns are then scored by their performance on the positive and negative sets.
- We want the scoring formula to reflect the following heuristic: it needs to rise monotonically with the number of positive sentences it matches, but drop very fast with the number of negative sentences it matches.

$$Score(Pattern) = \frac{\left|S \in PositiveSet : Pattern \text{ matches } S\right|}{\left(\left|S \in NegativeSet : Pattern \text{ matches } S\right| + 1\right)^2}$$

Sample Patterns - Inventor

- X , .* inventor .* of Y
- X invented Y
- X , .* invented Y
- when X .* invented Y
- X ' s .* invention .* of Y
- inventor .* Y , X
- Y inventor X
- invention .* of Y .* by X
- after X .* invented Y
- X is .* inventor .* of Y
- inventor .* X , .* of Y
- inventor of Y , .* X ,
- X is .* invention of Y
- Y , .* invented .* by X
- Y was invented by X

Sample Patterns – CEO (Company/X,Person/Y)

- X ceo Y
- X ceo .*Y ,
- former X .* ceo Y
- X ceo .*Y .
- Y , .* ceo of .* X ,
- X chairman .* ceo Y
- Y , X .* ceo
- X ceo .*Y said
- X ' .* ceo Y
- Y , .* chief executive officer .* of X
- said X .* ceo Y
- Y , .* X ' .* ceo
- Y , .* ceo .* X corporation
- Y , .* X ceo
- X ' s .* ceo .*Y ,
- X chief executive officer Y
- Y , ceo .* X ,
- Y is .* chief executive officer .* of X

Shallow Parser mode

- In the first mode of operation (without the use of NER), the predicates may define attributes of two different types: P*roperName* and *CommonNP*.
- We assume that the values of the *ProperName* type are always heads of proper noun phrases. And the values of the *CommonNP* type are simple common noun phrases (with possible proper noun modifiers, e.g. "*the Kodak camera*").
- We use a Java-written shallow parser from the OpenNLP (http://opennlp.sourceforge.net/) package. Each sentence is tokenized, tagged with part-of-speech, and tagged with noun phrase boundaries. The pattern matching and extraction is straightforward.

Building a Classification Model

- The goal is to set the score of the extractions using the information on the instance, the extracting patterns and the matches. Assume, that extraction *E* was generated by pattern *P* from a match *M* of the pattern *P* at a sentence *S*. The following properties are used for scoring:
  1. Number of different sentences that produce *E* (with any pattern).
  2. Statistics on the pattern *P* generated during pattern learning – the number of positive sentences matched and the number of negative sentences matched.
  3. Information on whether the slots in the pattern *P* are anchored.
  4. The number of non-stop words the pattern *P* contains.
  5. Information on whether the sentence *S* contains proper noun phrases between the slots of the match *M* and outside the match *M*.
  6. The number of words between the slots of the match *M* that were matched to skips of the pattern *P*.

Experimental Evaluation

- We want to answer the following 4 questions:
  1. Can we train SRES's classifier once, and then use the results on all other relations?
  2. What boost will we get by introducing a simple NER into the classification scheme of SRES?
  3. How does SRES's performance compare with KnowItAll and KnowItAll-PL?
  4. What is the true recall of SRES?

Training

1. The patterns for a single model predicate are run over a small set of sentences (10,000 sentences in our experiment), producing a set of extractions (between 150-300 extractions in our experiments).
2. The extractions are manually labeled according to whether they are correct or no.
3. For each pattern match *Mk*, the value of the feature vector $\boldsymbol{fk} = (f1, \ldots f16)$ is calculated, and the label $Lk = \pm 1$ is set according to whether the extraction that the match produced is correct or no.
4. A regression model estimating the function $L(\boldsymbol{f})$ is built from the training data $\{(\boldsymbol{fk}, Lk)\}$. We used the BBR, but other models, such as SVM are of course possible.

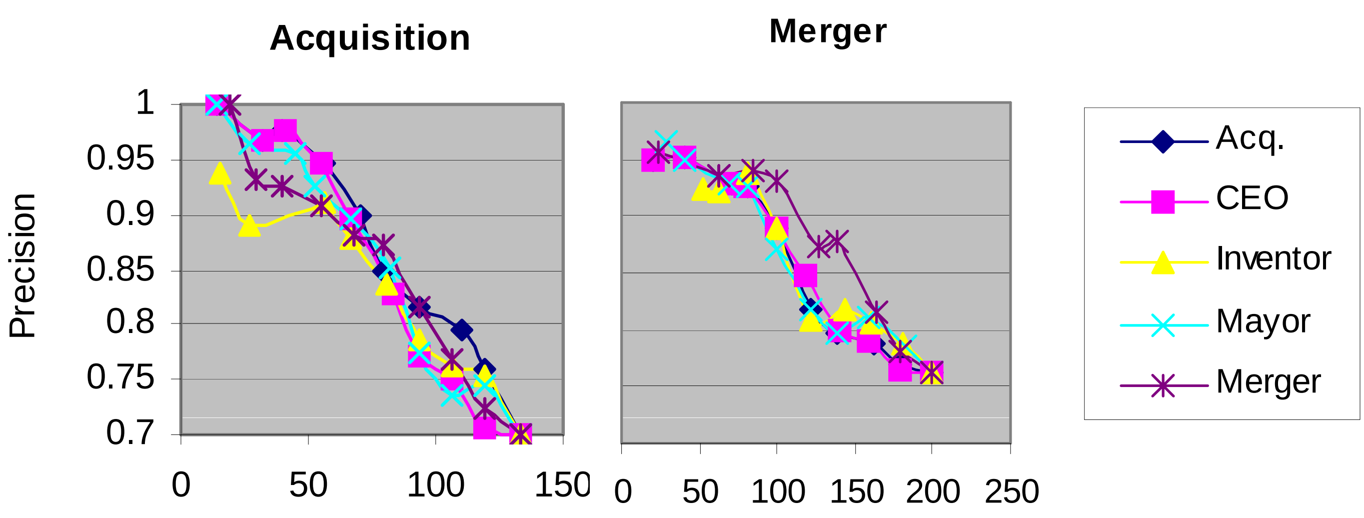
Testing

1. The patterns for all predicates are run over the sentences.
2. For each pattern match M, its score $L(f(M))$ is calculated by the trained regression model. Note that we do not threshold the value of L, instead using the raw probability value between zero and one.
3. The final score for each extraction is set to the maximal score of all matches that produced the extraction.

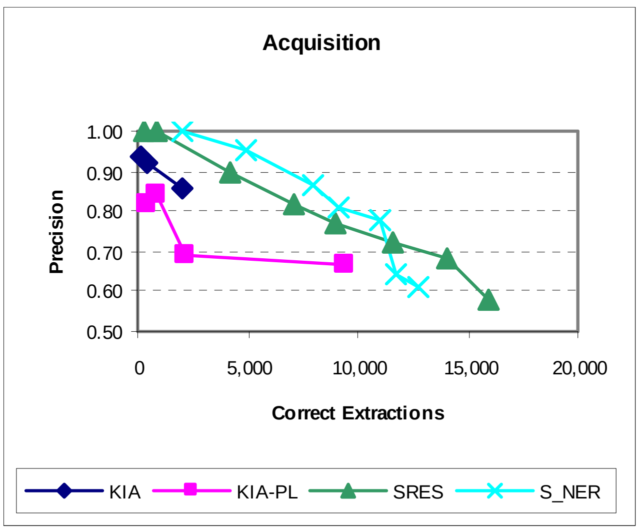
Sample Output

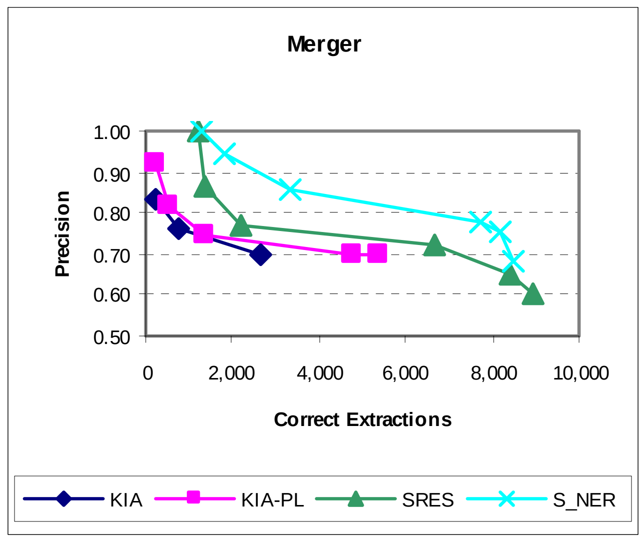
- <e> <arg1>HP</arg1> <arg2>Compaq</arg2>
  - <s><DOCUMENT>Additional information about the <X>HP</X> -<Y>Compaq</Y> merger is available at www.VotetheHPway.com .</DOCUMENT></s>
  - <s><DOCUMENT>The Packard Foundation, which holds around ten per cent of <X>HP</X> stock, has decided to vote against the proposed merger with <Y>Compaq</Y>.</DOCUMENT></s>
  - <s><DOCUMENT>Although the merger of <X>HP</X> and <Y>Compaq</Y> has been approved, there are no indications yet of the plans of HP regarding Digital GlobalSoft.</DOCUMENT></s>
  - <s><DOCUMENT>During the Proxy Working Group's subsequent discussion, the CIO informed the members that he believed that Deutsche Bank was one of <X>HP</X>'s advisers on the proposed merger with <Y>Compaq</Y>.</DOCUMENT></s>
  - <s><DOCUMENT>It was the first report combining both <X>HP</X> and <Y>Compaq</Y> results since their merger.</DOCUMENT></s>
  - <s><DOCUMENT>As executive vice president, merger integration, Jeff played a key role in integrating the operations, financials and cultures of <X>HP</X> and <Y>Compaq</Y> Computer Corporation following the 19 billion merger of the two companies.</DOCUMENT></s>

Cross-Classification Experiment

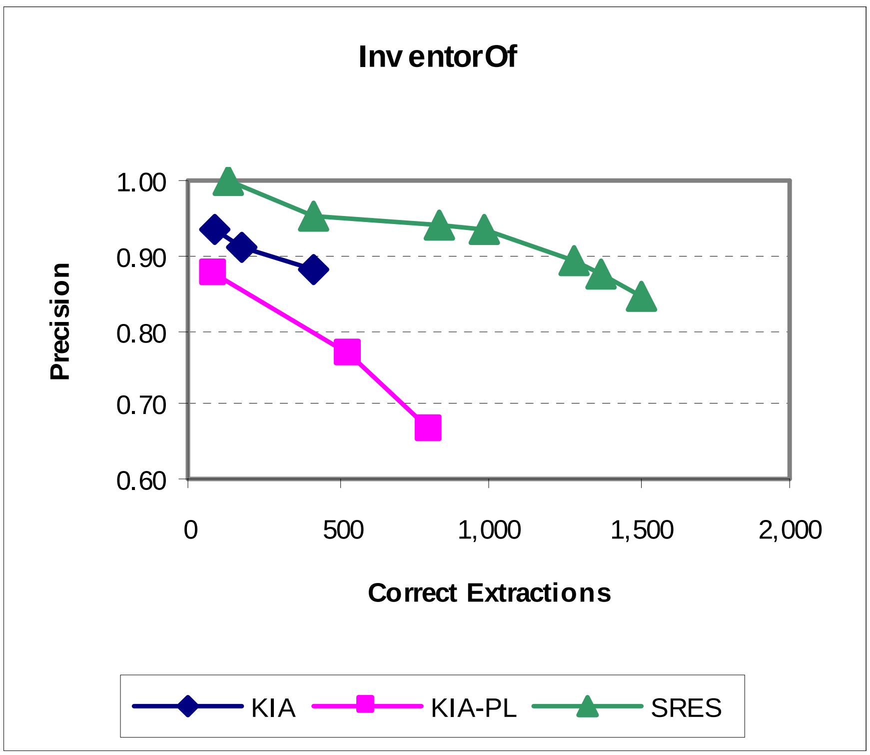


Results!


Acquisition
1.00
0.90
0.80
0.70
0.60
0.50
Precision
0
5,000
10,000
15,000
20,000
Correct Extractions
KIA
KIA-PL
SRES
S_NER


Merger
1.00
0.90
0.80
0.70
0.60
0.50
Precision
0
2,000
4,000
6,000
8,000
10,000
Correct Extractions
KIA
KIA-PL
SRES
S_NER

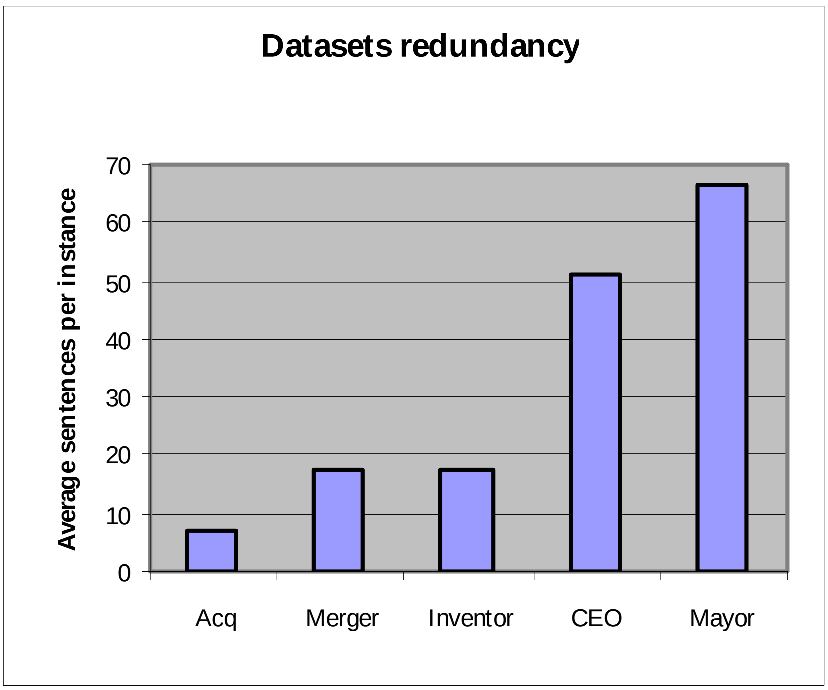
Inventor Results



When is SRES better than KIA?

- KnowItAll extraction works well when redundancy is high and most instances have a good chance of appearing in simple forms that KnowItAll is able to recognize.
- The additional machinery in SRES is necessary when redundancy is low.
- Specifically, SRES is more effective in identifying low-frequency instances, due to its more expressive rule representation, and its classifier that inhibits those rules from overgeneralizing.

The Redundancy of the Various Datasets



Outline

- Intro to text mining
  - IR vs. IE
- Information extraction (IE)
  - IE Components
  - Case studies in IE
    - Whizbang!
    - CiteSeer and GoogleScholar
- Relation Extraction/ Open IE
  - KnowItAll and SRES
- Blog Mining: Market Structure Surveillance

Market Structure Surveillance

Ronen
Feldman

Jacob
Goldenberg

Oded
Netzer

Research Objective

- Can we use the Web as a marketing research playground?
- Uncovering market structure from information consumers are posting on the web
- An example of the rapidly growing area of sentiment mining

OPINE

Ana-Maria Popescu, Bao Nguyen, Oren Etzioni

Home | Language: English

New York City hotels > **Renaissance New York Hotel Times Square**

Review Summary

Staff: excellent (7), great (3), very helpful (2), poor, fantastic, helpful, love, good, *view all* (17)

Location: great (4), best (3), good (2), fabulous, fantastic, ideal, superb, not great, love, *view all* (15)

Room: nice (5), great (2), not great (2), good (2), very nice (2), excellent, superb, lovely, average, *view all* (17)

Quality: best, fantastic, lovely, recommend, love, nice, fine, *view all* (7)

Food: very good (2), fantastic, lovely, not great, great, *view all* (6)

Bathroom beauty: beautiful

Bar: fabulous, great, *view all* (2)

Staff friendliness: friendly (4), very friendly (2), incredibly friendly, unfriendly, *view all* (8)

Room bed comfort: comfy (2), comfortable (2), extremely comfortable, *view all* (5)

Bathroom: great (2), elegant, very nice, nice, *view all* (5)

Room cleanness: **clean (2)**

User comments:

the rooms were clean and smelled great . Read more

The rooms were clean, spacious, soundproof and well-appointed . Read more

What are we going to do?

- Text mine consumer postings
- Use network analysis framework and other methods of analysis to reveal the underlying market structure

Market Structure Analysis

- Econometric models of brand choice data
- Large scale surveys
- Product similarities (multi-dimensional scaling)
- Often reveals what the structure is, but not why

Text Mining For Marketing Advantages

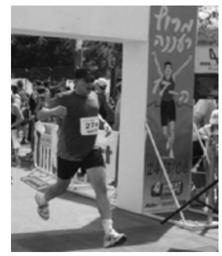
- Combines of observational and descriptive marketing research
- Non-invasive marketing research (no demand effect)
- Minimizes recall error
- Very rich data
- Permits both qualitative and quantitative marketing research
- Sample size is not an issue
- Real time data

Nielsen BuzzMetrics BUZZLOGIC Cymfony umbria

The Text Mining Process

- **Download:** html-pages are downloaded from a given forum site
- **Clean:** html-like tags and non-textual information like images, commercials, etc. are cleaned from the downloaded pages
- **Chunk:** the textual parts are divided into informative units like threads, messages, and sentences
- **Information Extraction:** products and product attributes are extracted from the messages
- **Extract comparisons between products:** either by using co-occurrence analysis or by using learned comparison patterns

Example Applications



- Three applications
  - Running shoes ("professionals" community)

RUNNER'S WORLD

  - Sedan cars (mature and common market)

edmunds.com where smart car buyers start®

Consumer Driven® edmunds CARSPACE

  - iPhone (innovation, pre-during-after launch)

iPhone Enthusiast iPhone Topsites

Product Co-occurrence Data

Message #1199 **Civic vs. Corolla** by mcmanus *Jul 21, 2007 (4:05 pm)*
Yes DrFill, the Honda car model is sporty, reliable, and economical vs the **Corolla** that is just reliable and economical. Ironically its Toyota that is supplying 1.8L turbo … Neon to his 16 year old brother. I drove it about 130 miles today. Boy does that put all this **Civic** vs. **Corolla** back in perspective! The Neon is very crudely designed and built, with no low …

Audi A6	Honda Civic	252
Audi A6	Toyota Corolla	101
Honda Civic	Audi 6	252
Honda Civic	Toyota Corolla	2762
Toyota Corolla	Audi A6	101
Toyota Corolla	Honda Civic	2762

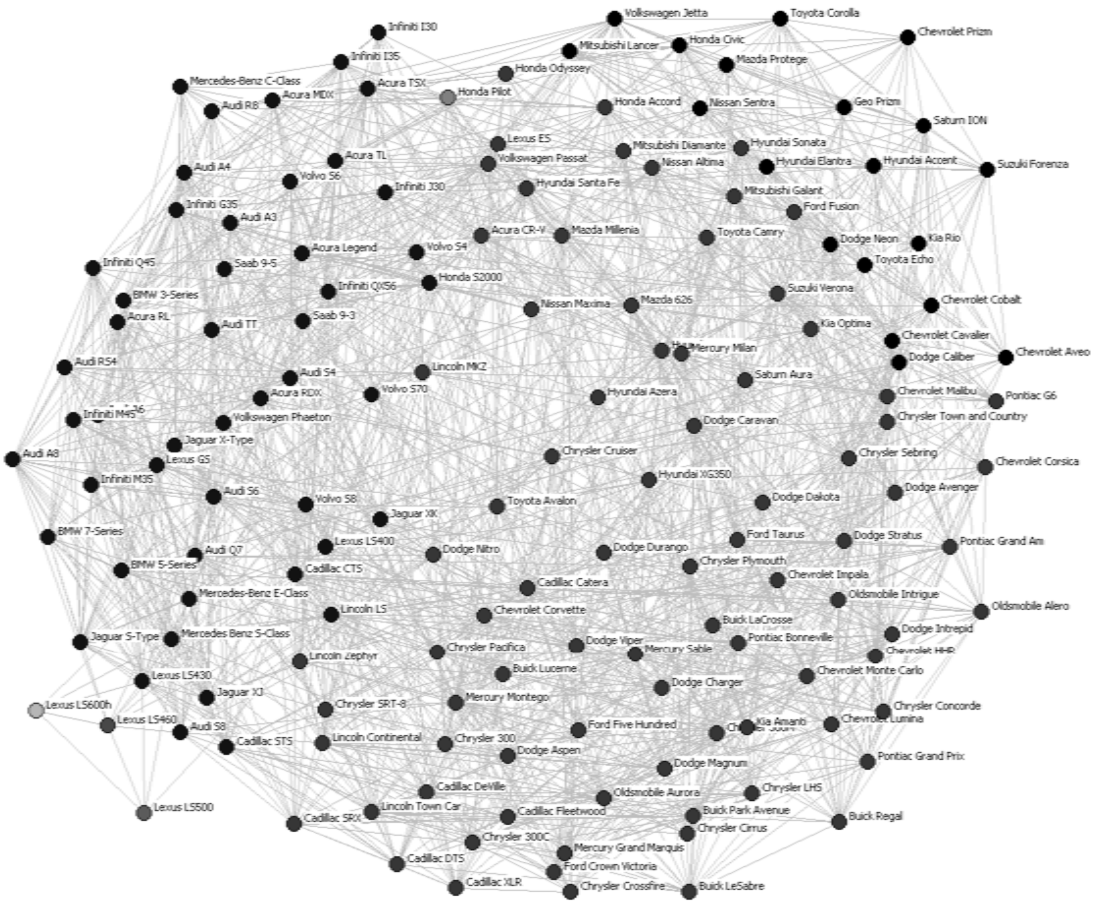
	Audi A6	Honda Civic	Toyota Corolla
Audi A6	---	252	101
Honda Civic	252	---	2762
Toyota Corolla	101	2762	---

$$lift(A,B) = \frac{N(A,B)}{N(A) \times N(B)}$$

Some Text Mining Difficulties

- We are interested in:
  - **Brand names** (Car companies, shoe companies)
  - **Model names** (Car models, shoe models)
  - Some **common terms** (mostly noun-phrases and adjectives)
- **Brand names** - are relatively easy
  - Need to deal with abbreviations and spelling mistakes
- **Models** - are more complex
  - Variations in writing styles
    - Honda Civic could be written as “Honda Civic”; “Civic”; “Honda Civic LS”; “Honda Civic LE”; “LE”; “H. Civic”; “Hondah Sivik”
    - Model numbers can be written as: 5, V, Five
      “Asics Speedstar (both I and II), I love the I and II's and can't wait for the III's”
    - Model can be referred to as numbers but numbers do not always refer to models (e.g., “1010 for New Balance 1010”, but $1010)

The Car Models Network



The Google Page-Rank of the Car Models

- Eigenvector centrality
  - Importance of a node in the network

$$x_i = \frac{1}{\lambda}\sum_{j=1}^{N} A_{i,j}x_j \quad \overrightarrow{x} = \frac{1}{\lambda}A\overrightarrow{x}$$

  - Used by Google for page ranking

Car Model	Eigenvector Centrality
Honda Accord	80.21
Toyota Camry	72.28
Hyundai Sonata	44.32
Nissan Altima	35.41
Ford Fusion	29.46
Acura TL	28.12
Honda Civic	23.64
Volkswagen Passat	22.10
Infiniti G35	16.60
Nissan Maxima	16.58
Toyota Avalon	15.21
Acura TSX	15.16
Chevrolet Malibu	12.95
Toyota Corolla	11.31
Chevrolet Impala	10.57

Predicting Sales Using Network Centrality

- **DV:**

  2004 cars sales data; Sales for 92 car models

- **IVs:**

  1) Eigenvector centrality

  2) Occurrence

Automotive News

R^2=0.354

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	49009.354	8092.532		6.056	.000
	occurance	3.980	.567	.595	7.017	.000

a. Dependent Variable: sales_2004

R^2=0.409

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	53029.018	7379.368		7.186	.000
	eigen	4066.596	515.209	.640	7.893	.000

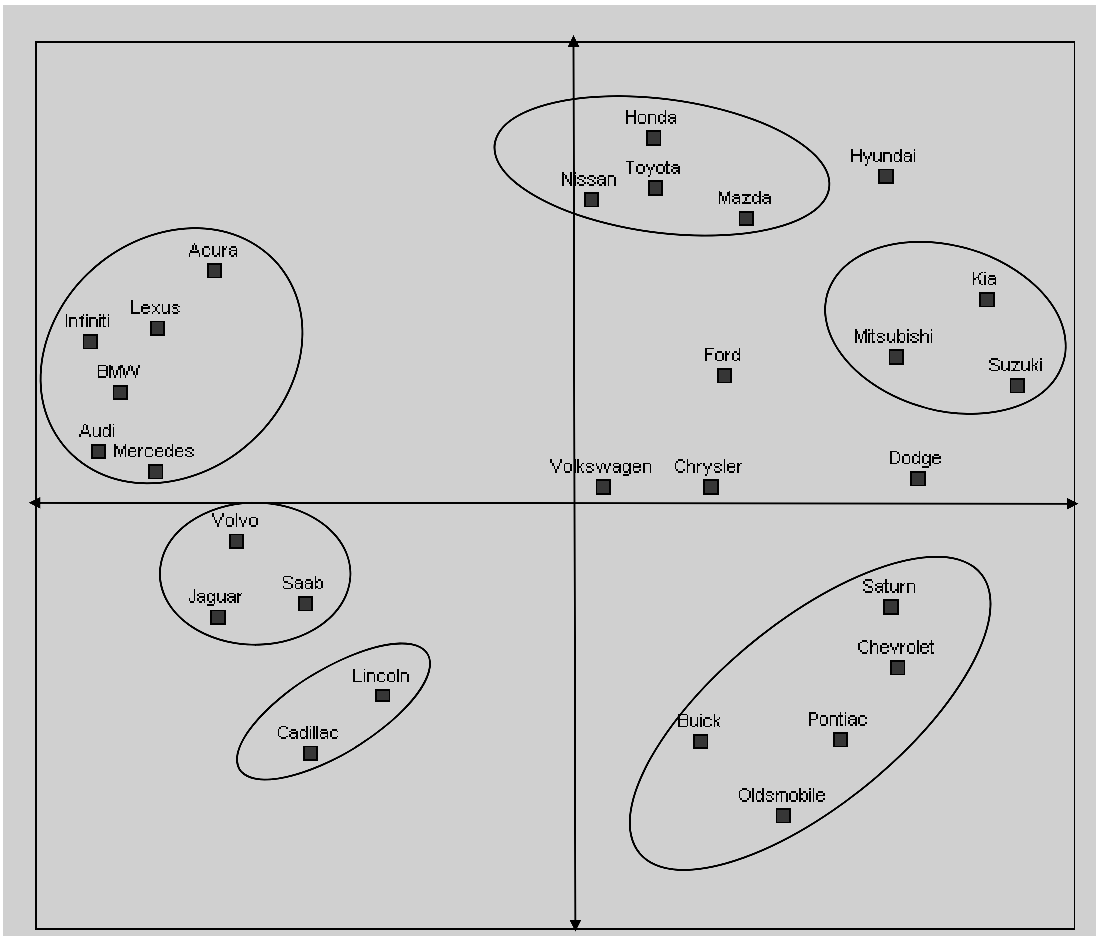
a. Dependent Variable: sales_2004

R^2=0.421

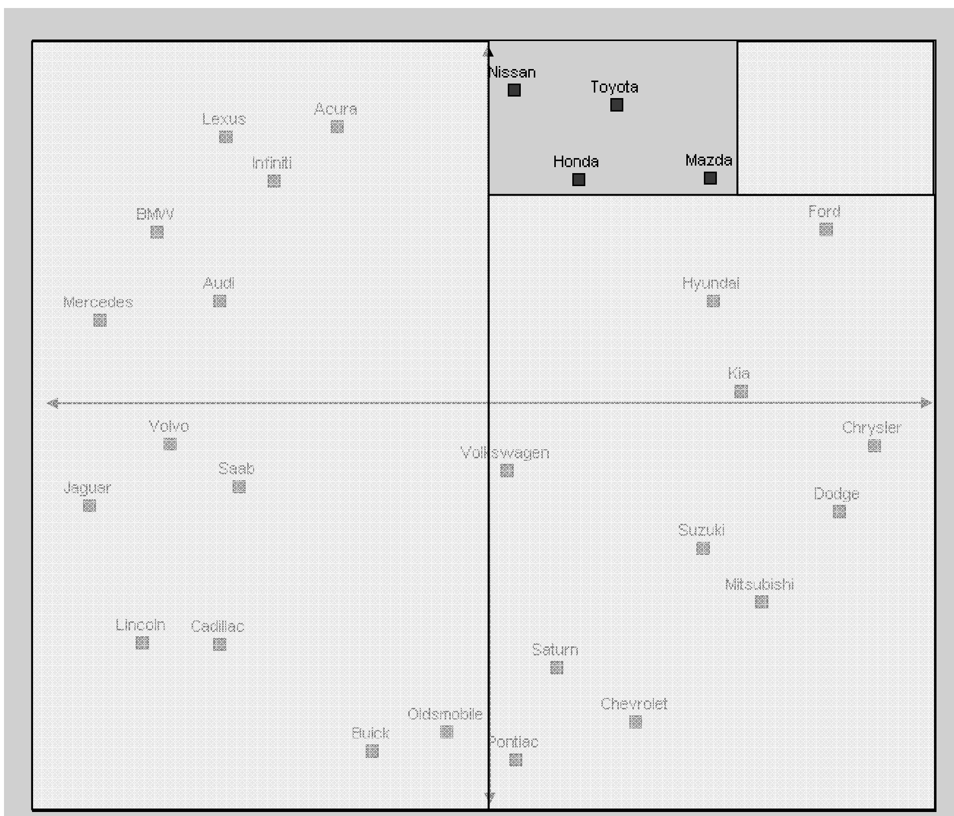
Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	57786.797	8177.648		7.066	.000
	occurance	-2.959	2.231	-.442	-1.326	.188
	eigen	6794.120	2119.945	1.069	3.205	.002

a. Dependent Variable: sales_2004

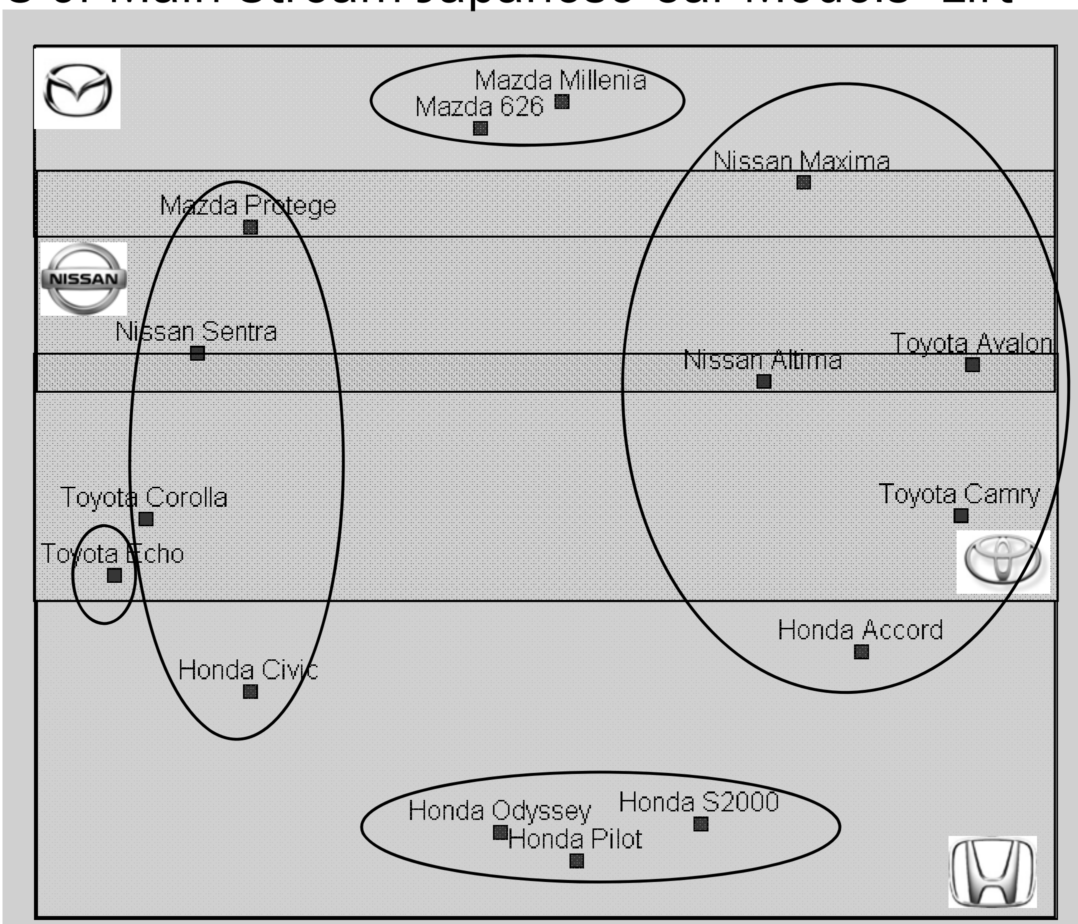
MDS of Brands Lift



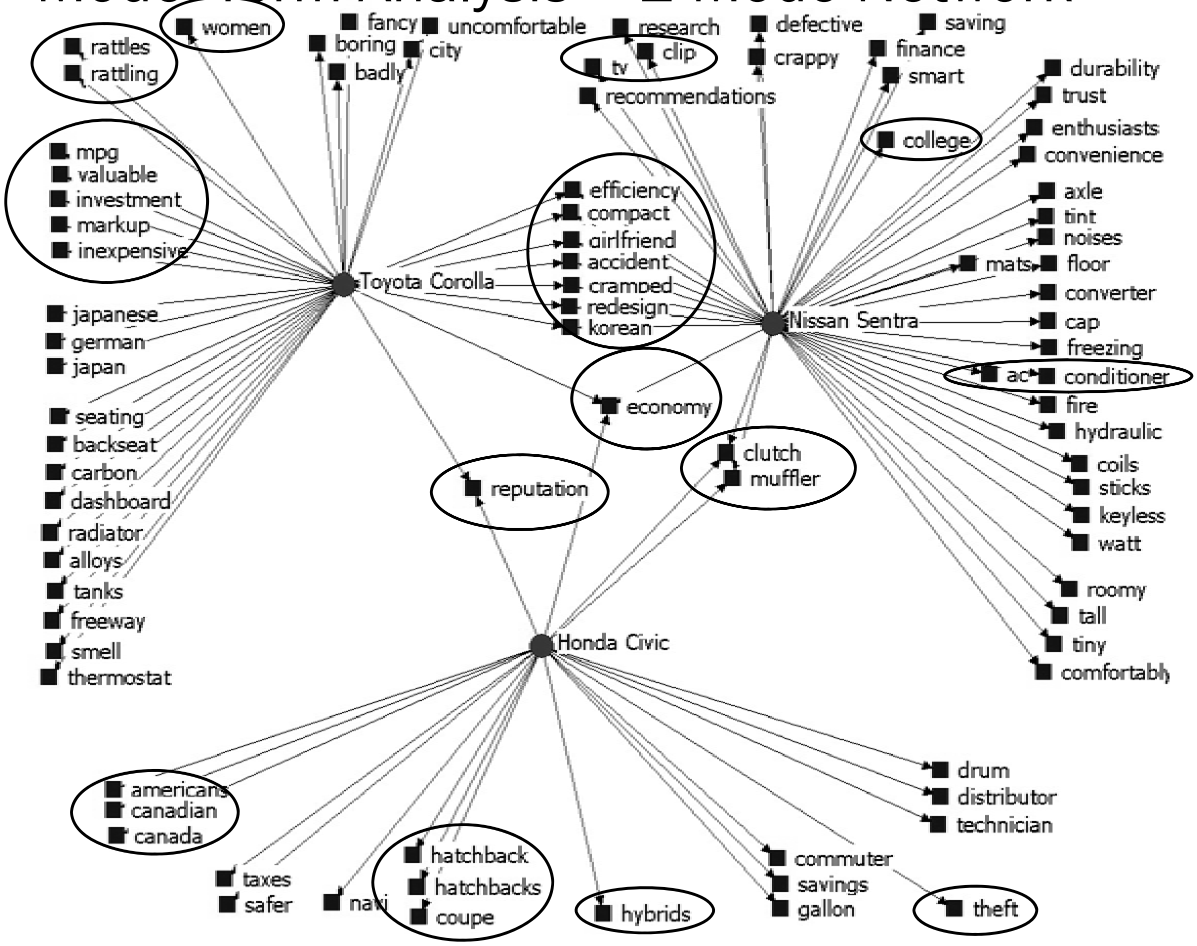
Digging in Deeper – Main Stream



MDS of Main Stream Japanese Car Models -Lift



Model-Term Analysis – 2 Mode Network



Most Stolen Cars Analysis

The **National Insurance Crime Bureau (**NICB®) has compiled a list of the 10 vehicles most frequently reported stolen in the U.S. in 2005

1) 1991 Honda Accord
2) 1995 Honda Civic
3) 1989 Toyota Camry
4) 1994 Dodge Caravan
5) 1994 Nissan Sentra
6) 1997 Ford F150 Series
7) 1990 Acura Integra
8) 1986 Toyota Pickup
9) 1993 Saturn SL
10) 2004 Dodge Ram Pickup

Top 10 cars mentioned with "stealing" phrases in our data ("Stolen", "Steal", "Theft")

1) Honda Accord (165)
2) Honda Civic (101)
3) Toyota Camry (71)
4) Nissan Maxima (69)
5) Acura TL (58)
6) Infinity G35 (44)
7) BMW 3-Series (40)
8) Hyundai Sonata (26)
9) Nissan Altima (25)
10) Volkswagen Passat (23)

Market Research Summary

- Text mining converts unstructured web data into useful information and knowledge
- Compute co-occurrence of
  - Pairs of brand names
  - Brands and attributes
- Visualize via clustering, MDS
- High face validity for using text mining for market structure analysis
  - Predicts sales, car thefts, ….
- Future Directions
  - Benchmarking against traditional market structure methods
  - Dynamics of the semantic network

The Text Mining Business

- Part of most big data mining systems
  - Fair Isaac. SAS, Oracle, SPSS …
- **AeroText** - Information extraction in multiple languages
- **Autonomy** - suite of text mining, clustering and categorization solutions for knowledge management
- **LanguageWare** - the IBM Tools for Text Mining.
- **Inxight** - text analytics, search, and visualization. (sold to Business Objects that was sold to SAP)
- **RapidMiner/YALE** - open-source data and text mining
- **Thomson Data Analyzer** - analysis of patent information, scientific publications and news.
- Lots more: Attensity, Endeca Technologies, Expert System S.p.A., Nstein Technologies. …
- Plus sentiment analysis: big boys plus Nielsen Buzzmetrics and many othres.

Summary

- Information Extraction
  - Not just information retrieval
  - Find named entities, relations, events
  - Hand-built vs. Learned models
    - CRFs widely used
- Open Information Extraction
  - Unsupervised relation extraction
    - Bootstrap pattern learning
- Sentiment analysis
- Visualize results
  - Link analysis, MDS, …
- Text mining is easy and hard

References

- See www.cis.upenn.edu/~ungar/KDD/text-mining.html