Information Extraction: Theory and Practice

Ronen Feldman
Computer Science Department
Bar-Ilan University, ISRAEL

feldman@cs.biu.ac.il

ClearForest
Leveraging Content
Creating Business Intelligence

Outline

- Introduction to Text Mining
- Information Extraction
  - Entity Extraction
  - Relationship Extraction
  - KE Approach
    - IE Languages
  - ML Approaches to IE
    - HMM
  - Anaphora resolution
  - Evaluation
- Link Detection
  - Introduction to LD
  - Architecture of LD systems
  - 9/11
  - Link Analysis Demos

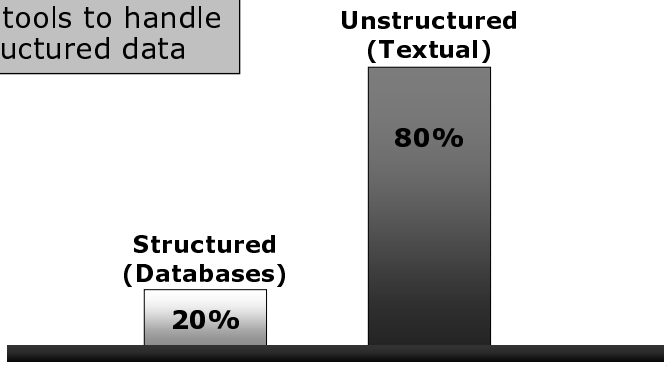
Background

- Rapid proliferation of information available in digital format
- People have **less time** to absorb **more information**

The Information Landscape

Problem

Lack of tools to handle unstructured data



TM != 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

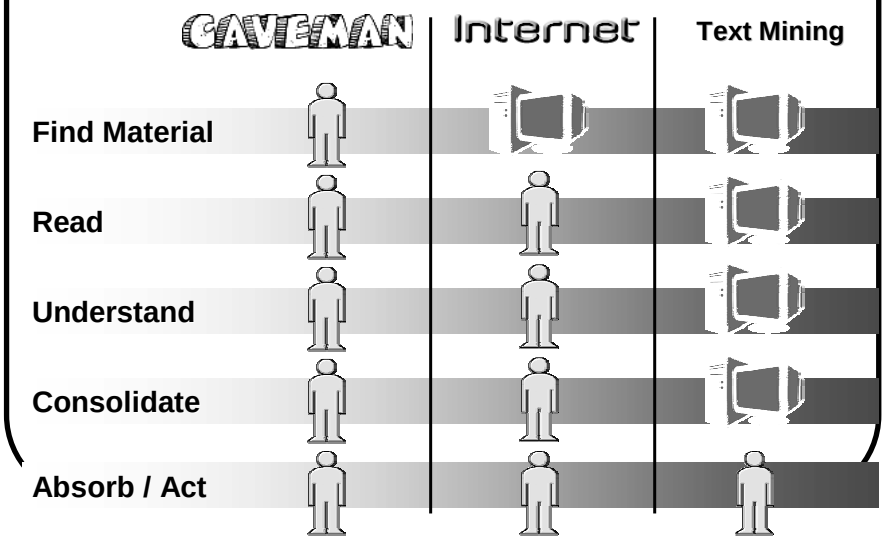
Patterns

Connections

Profiles

Trends

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.**

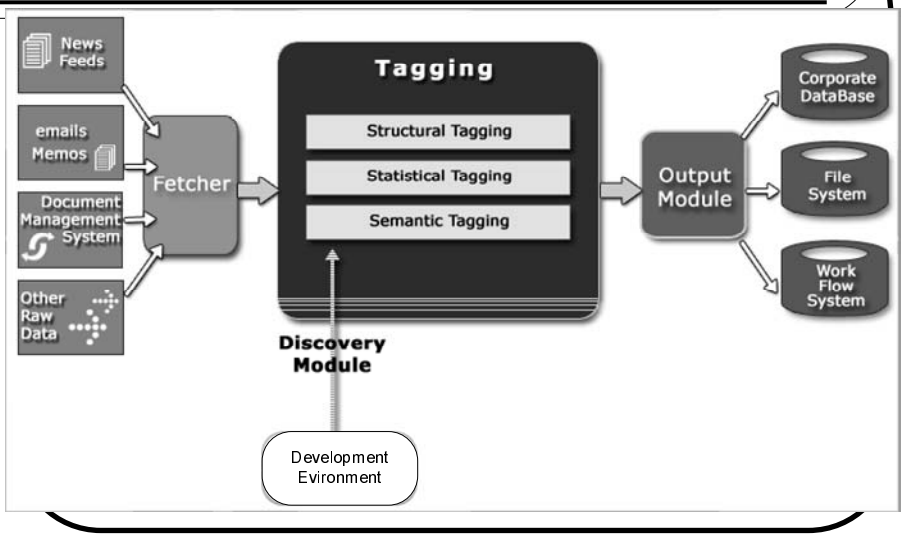
Document Types

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

Text Representations

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

The 100,000 foot Picture



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

…….

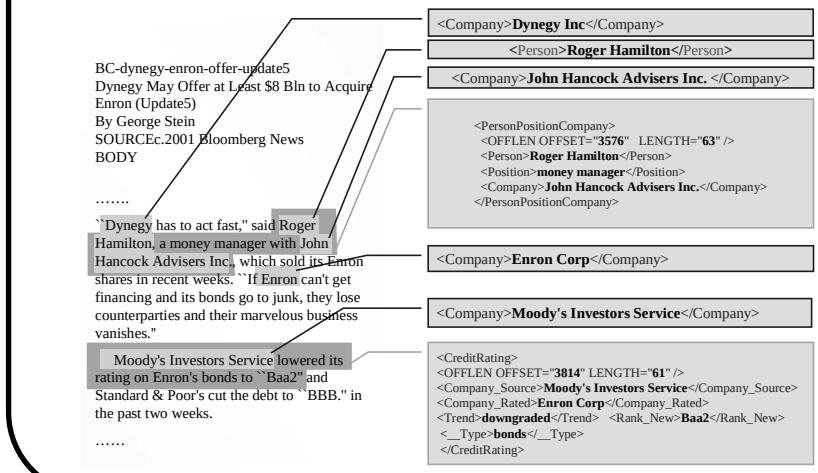
The Finsbury Park Mosque is the center of radical Muslim activism in England. Through its doors have passed at least three of the men now held on suspicion of terrorist activity in France, England and Belgium, as well as one Algerian man in prison in the United States.

``The mosque's chief cleric, Abu Hamza al-Masri 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 in 1999 for his alleged involvement in a Yemen bomb plot, but was set free after Yemen failed to produce enough evidence to have him extradited. ."

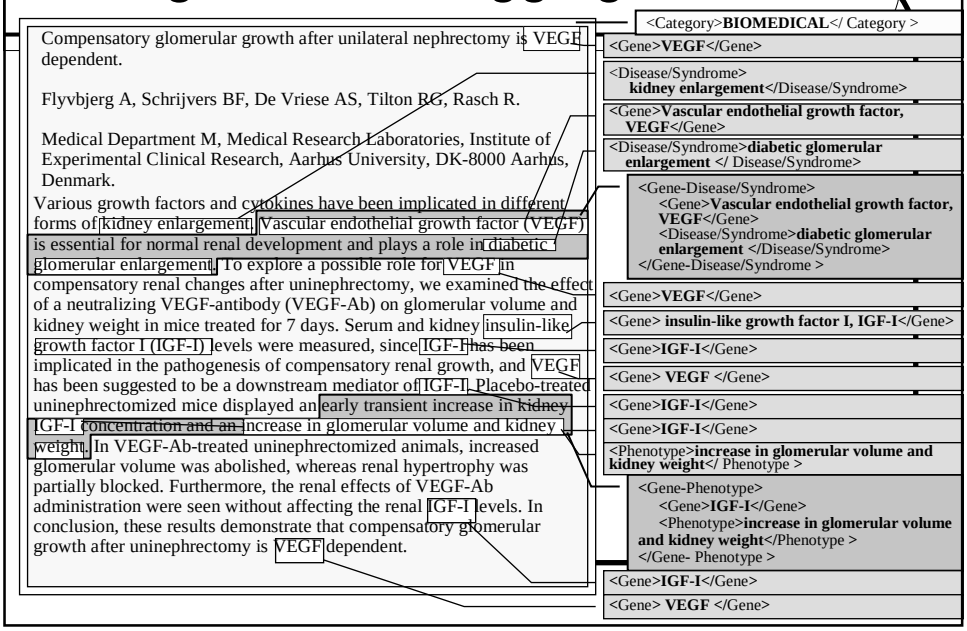
……

```
<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>
```

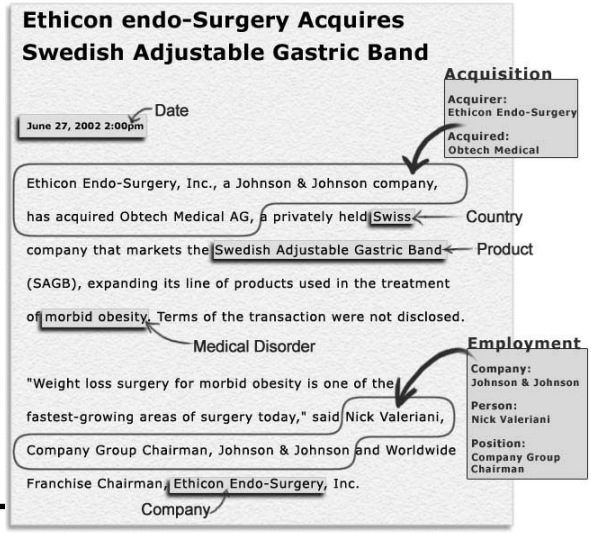
Tagging a Document



Intelligent Auto-Tagging



A Tagged News Document



Example

Phosphorylation Template

Kinase	**Protein**
Some kinase	*X*

"..BES1 is phosphorylated and appears to be destabilized by the glycogen synthase kinase-3 (GSK-3) BIN2..."

Kinase	**Protein**
glycogen synthase kinase-3 (GSK-3) BIN2	*BES1*

Phosphorylation

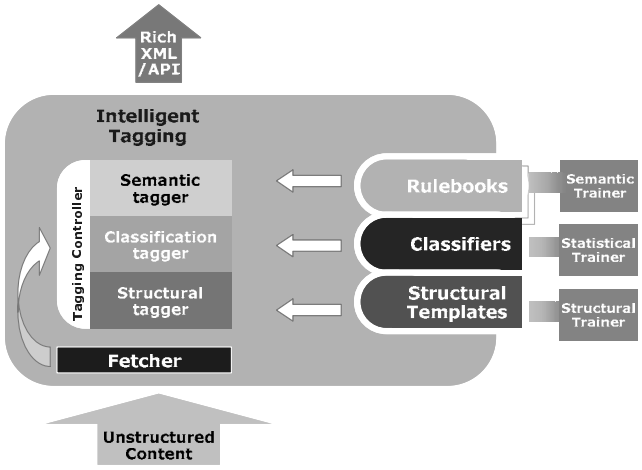
Leveraging Content Investment

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

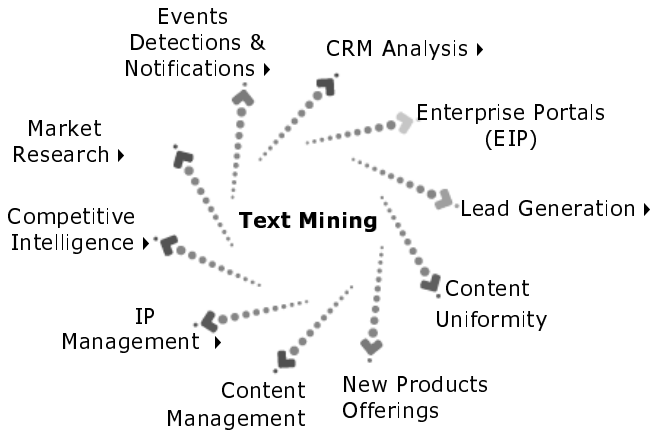
In any format
- Documents; PDFs; E-mails; articles; etc
- "Raw" or categorized
- Formal; informal; combination

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

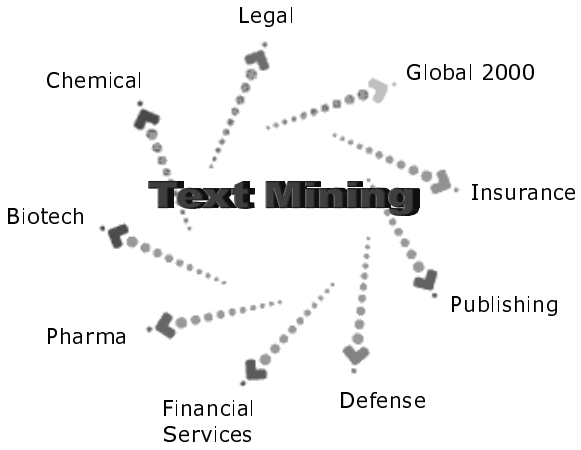
The Tagging Process (a Unified View)



Text Mining Horizontal Applications



Text Mining: Vertical Applications



Information Extraction

What is Information Extraction?

- IE does not indicate which documents need to be read by a user, it rather extracts pieces of information that are salient to the user's needs.
- Links between the extracted information and the original documents are maintained to allow the user to reference context.
- The kinds of information that systems extract vary in detail and reliability.
- Named entities such as persons and organizations can be extracted with reliability in the 90th percentile range, but do not provide attributes, facts, or events that those entities have or participate in.

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 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.

Example (Entities and Facts)

- Fletcher Maddox, former Dean of the UCSD Business School, announced the formation of La Jolla Genomatics together with his two sons. La Jolla Genomatics will release its product Geninfo in June 1999. Geninfo is a turnkey system to assist biotechnology researchers in keeping up with the voluminous literature in all aspects of their field.
- Dr. Maddox will be the firm's CEO. His son, Oliver, is the Chief Scientist and holds patents on many of the algorithms used in Geninfo. Oliver's brother, Ambrose, follows more in his father's footsteps and will be the CFO of L.J.G. headquartered in the Maddox family's hometown of La Jolla, CA.

Persons:	Organizations:	Locations:	Artifacts:	Dates:
Fletcher Maddox	UCSD Business School	La Jolla	Geninfo	June 1999
Dr. Maddox	La Jolla Genomatics	CA	Geninfo	
Oliver	La Jolla Genomatics			
Oliver	L.J.G.			
Ambrose				
Maddox				

PERSON	Employee_of	ORGANIZATION
Fletcher	Employee_of	UCSD Business
Maddox	Employee_of	School
Fletcher	Employee_of	La Jolla Genomatics
Maddox	Employee_of	La Jolla Genomatics
Oliver		La Jolla Genomatics
Ambrose		
ARTIFACT	Product_of	ORGANIZATION
Geninfo	Product_of	La Jolla Genomatics
LOCATION	Location_of	ORGANIZATION
La Jolla	Location_of	La Jolla Genomatics
CA	Location_of	La Jolla Genomatics

Events and Attributes

COMPANY-FORMATION_EVENT:

COMPANY:	La Jolla Genomatics
PRINCIPALS:	Fletcher Maddox Oliver Ambrose
DATE:	
CAPITAL:	

RELEASE-EVENT:

COMPANY:	La Jolla Genomatics
PRODUCT:	Geninfo
DATE:	June 1999
COST:	

NAME:	**Fletcher Maddox Maddox**
DESCRIPTOR:	**former Dean of the UCSD Business School his father the firm's CEO**
CATEGORY:	**PERSON**
NAME:	**Oliver**
DESCRIPTOR:	**His son Chief Scientist**
CATEGORY:	**PERSON**
NAME:	**Ambrose**
DESCRIPTOR:	**Oliver's brother the CFO of L.J.G.**
CATEGORY:	**PERSON**
NAME:	**UCSD Business School**
DESCRIPTOR:	
CATEGORY:	**ORGANIZATION**
NAME:	**La Jolla Genomatics L.J.G.**
DESCRIPTOR:	
CATEGORY:	**ORGANIZATION**
NAME:	**Geninfo**
DESCRIPTOR:	**its product**
CATEGORY:	**ARTIFACT**
NAME:	**La Jolla**
DESCRIPTOR:	**the Maddox family's hometown**
CATEGORY:	**LOCATION**
NAME:	**CA**
DESCRIPTOR:	
CATEGORY:	**LOCATION**

IE Accuracy by Information Type

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

Unstructured Text

POLICE ARE INVESTIGATING A ROBBERY THAT OCCURRED AT THE 7-ELEVEN STORE LOCATED AT 2545 LITTLE RIVER TURNPIKE IN THE LINCOLNIA AREA ABOUT 12:30 AM FRIDAY. A 24 YEAR OLD ALEXANDRIA AREA EMPLOYEE WAS APPROACHED BY TWO MEN WHO DEMANDED MONEY. SHE RELINQUISHED AN UNDISCLOSED AMOUNT OF CASH AND THE MEN LEFT. NO ONE WAS INJURED. THEY WERE DESCRIBED AS BLACK, IN THEIR MID TWENTIES, BOTH WERE FIVE FEET NINE INCHES TALL, WITH MEDIUM BUILDS, BLACK HAIR AND CLEAN SHAVEN. THEY WERE BOTH WEARING BLACK PANTS AND BLACK COATS. ANYONE WITH INFORMATION ABOUT THE INCIDENT OR THE SUSPECTS INVOLVED IS ASKED TO CALL POLICE AT (703) 555-5555.

Structured (Desired) Information

Crime	Address	Town	Time	Day
ABDUCTION	**8700 BLOCK OF LITTLE RIVER TURNPIKE,**	**ANNANDALE**	**11:30 PM**	**SUNDAY**
…	**…**	**…**	**…**	**…**
ROBBERY	**7- ELEVEN STORE LOCATED AT 2545 LITTLE RIVER TURNPIKE,**	**LINCOLNIA**	**12:45 AM**	**FRIDAY**
ROBBERY	**7- ELEVEN STORE LOCATED AT 5624 OX ROAD,**	**FAIRFAX**	**3:00 AM**	**FRIDAY**

MUC Conferences

Conference	Year	Topic
MUC 1	1987	Naval Operations
MUC 2	1989	Naval Operations
MUC 3	1991	Terrorist Activity
MUC 4	1992	Terrorist Activity
MUC 5	1993	Joint Venture and Micro Electronics
MUC 6	1995	Management Changes
MUC 7	1997	Spaces Vehicles and Missile Launches

The ACE Evaluation

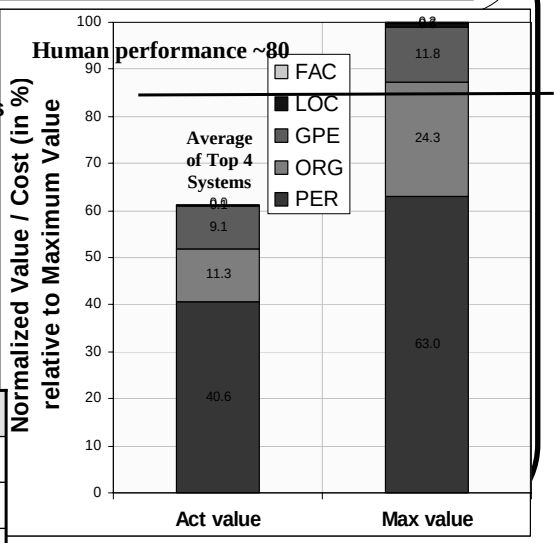
- The ACE program is dedicated to the challenge of extracting content from human language. This is a fundamental capability that the ACE program addresses with a basic research effort that is directed to master first the extraction of “entities”, then the extraction of “relations” among these entities, and finally the extraction of “events” that are causally related sets of relations.
- After two years of research on the entity detection and tracking task, top systems have achieved a capability that is useful by itself and that, in the context of the ACE EDT task, successfully captures and outputs well over 50 percent of the value at the entity level.
- Here value is defined to be the benefit derived by successfully extracting the entities, where each individual entity provides a value that is a function of the entity type (i.e., “person”, “organization”, etc.) and level (i.e., “named”, “unnamed”). Thus each entity contributes to the overall value through the incremental value that it provides.

ACE Entity Detection & Tracking Evaluation -- 2/2002

Goal: Extract entities. Each entity is assigned a value. This value is a function of its *Type* and *Level*. This value is gained when the entity is successfully detected. This value is lost when an entity is missed, spuriously detected, or mischaracterized.

Table of Entity Values

	PER	ORG	GPE	LOC	FAC
NAM	1	0.5	0.25	0.1	0.05
NOM	0.2	0.1	0.05	0.02	0.01
PRO	0.04	0.02	0.01	0.004	0.002



Miss 36%, False Alarm 22%, Type Error 6%

Applications of Information Extraction

- Routing of Information
- Infrastructure for IR and for Categorization (higher level features)
- Event Based Summarization.
- Automatic Creation of Databases and Knowledge Bases.

Where would IE be useful?

- Semi-Structured Text
- Generic documents like News articles.
- Most of the information in the document is centered around a set of easily identifiable entities.

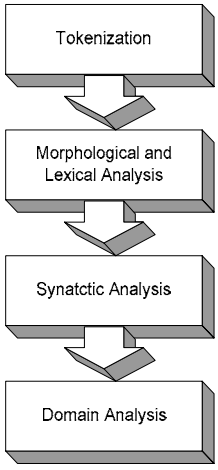
Approaches for Building IE Systems

- 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!

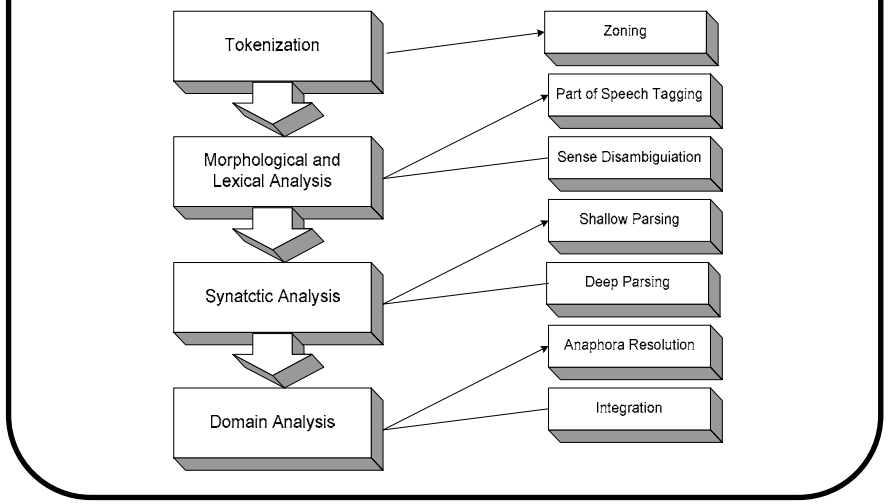
Approaches for Building IE Systems

- Automatically Trainable Systems
  - The techniques are based on pure statistics and almost no linguistic knowledge
  - 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.

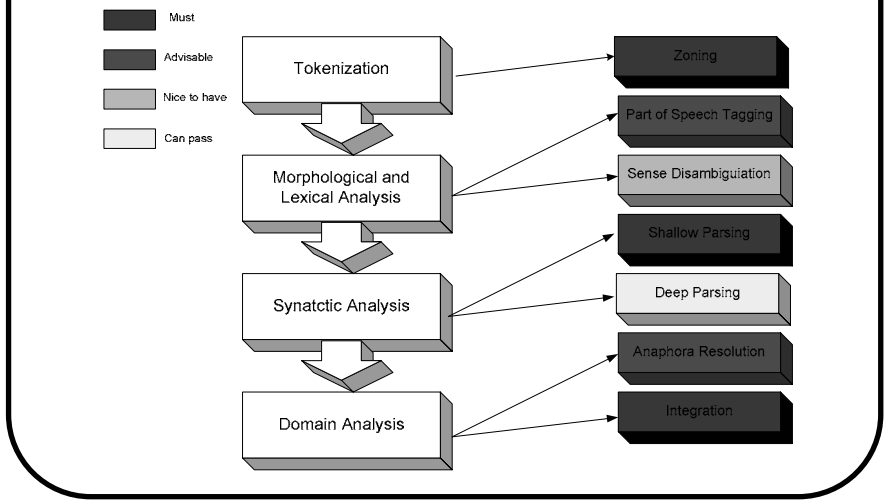
A Basic IE System



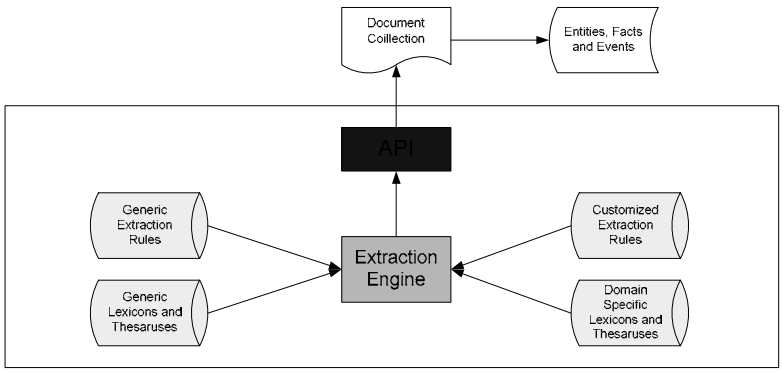
Architecture of a Full IE System



Components of IE System



The Extraction Engine



Why is IE Difficult?

- Different Languages
  - Morphology is very easy in English, much harder in German and Hebrew.
  - Identifying word and sentence boundaries is fairly easy in European language, much harder in Chinese and Japanese.
  - Some languages use orthography (like english) while others (like hebrew, arabic etc) do no have it.
- Different types of style
  - Scientific papers
  - Newspapers
  - memos
  - Emails
  - Speech transcripts
- Type of Document
  - Tables
  - Graphics
  - Small messages vs. Books

Morphological Analysis

- Easy
  - English, Japanese
  - Listing all inflections of a word is a real possibility
- Medium
  - French Spanish
  - A simple morphological component adds value.
- Difficult
  - German, Hebrew, Arabic
  - A sophisticated morphological component is a must!

Using Vocabularies

- “Size doesn’t matter”
  - Large lists tend to cause more mistakes
  - Examples:
    - Said as a person name (male)
    - Alberta as a name of a person (female)
- It might be better to have small domain specific dictionaries

Part of Speech Tagging

- POS can help to reduce ambiguity, and to deal with ALL CAPS text.
- However
  - It usually fails exactly when you need it
  - It is domain dependent, so to get the best results you need to retrain it on a relevant corpus.
  - It takes a lot of time to prepare a training corpus.

A simple POS Strategy

- Use a tag frequency table to determine the right POS.
  - This will lead to elimination of rare senses.
- The overhead is very small
- It improve accuracy by a small percentage.
- Compared to full POS it provide similar boost to accuracy.

Dealing with Proper Names

- The problem
  - Impossible to enumerate
  - New candidates are generated all the time
  - Hard to provide syntactic rules
- Types of proper names
  - People
  - Companies
  - Organizations
  - Products
  - Technologies
  - Locations (cities, states, countries, rivers, mountains)

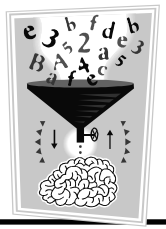
Comparing RB Systems with ML Based Systems

	Rule Based	**HMM**
Wall Street Journal		
MUC6	96.4	93
MUC7	93.7	90.4
Transcribed Speech		
HUB4	90.3	90.6

Building a RB Proper Name Extractor

- A Lexicon is always a good start
- The rules can be based on the lexicon and on:
  - The context (preceding/following verbs or nouns)
  - Regular expressions
    - Companies: Capital* [,] inc, Capital* corporation..
    - Locations: Capital* Lake, Capital* River
  - Capitalization
  - List structure
- After the creation of an initial set of rules
  - Run on the corpus
  - Analyze the results
  - Fix the rules and repeat…
- This Process can take around 2-3 weeks and result in performance of between 85-90% break even.
- Better performance can be achieved with more effort (2-3 months) and then performance can get to 95-98%

Introduction to HMMs for IE



Motivation

- We can view the named entity extraction as a classification problem, where we classify each word as belonging to one of the named entity classes or to the no-name class.
- One of the most popular techniques for dealing with classifying sequences is HMM.
- Example of using HMM for another NLP classification task is that of part of speech tagging (Church, 1988 ; Weischedel et. al., 1993).

What is HMM?

- HMM (Hidden Markov Model) is a finite state automaton with stochastic state transitions and symbol emissions (Rabiner 1989).
- The automaton models a probabilistic generative process.
- In this process a sequence of symbols is produced by starting in an initial state, transitioning to a new state, emitting a symbol selected by the state and repeating this transition/emission cycle until a designated final state is reached.

Disadvantage of HMM

- The main disadvantage of using an HMM for Information extraction is the need for a large amount of training data. i.e., a carefully tagged corpus.
- The corpus needs to be tagged with all the concepts whose definitions we want to learn.

Notational Conventions

- T = length of the sequence of observations (training set)
- N = number of states in the model
- q_t = the actual state at time *t*
- $S = \{S_1,...S_N\}$ (finite set of possible states)
- $V = \{O_1,...O_M\}$ (finite set of observation symbols)
- $\pi = \{\pi_i\} = \{P(q_1 = S_i)\}$ starting probabilities
- $A = \{aij\}=P(q_{t+1}= S_i \mid q_t = S_j)$ transition probabilities
- $B = \{b_i(O_t)\} = \{P(O_t \mid q_t = S_i)\}$ emission probabilities

The Classic Problems Related to HMMs

- Find *P*(O | λ): the probability of an observation sequence given the HMM model.
- Find the most likely state trajectory given λ and O.
- Adjust λ = (π, A, B) to maximize *P*(O | λ).

Calculating *P*(O | λ)

- The most obvious way to do that would be to enumerate every possible state sequence of length T (the length of the observation sequence). Let Q = Q1,...QT , then by assuming independence between the states we have
  - P(O|Q, λ) = $\prod_{i=1}^{T} P(O_i \mid q_i, \lambda) = \prod_{i=1}^{T} b_{q_i}(O_i)$
  - P(Q|λ) = $\pi_{q_1} \prod_{i=1}^{T-1} a_{q_i q_{i+1}}$
- By using Bayes theorem we have
  - P(O,Q|λ) = P(O|Q, λ) P(Q|λ)
- Finally The main problem with this is that we need to do $2TN^T$ multiplications, which is certainly not feasible even for a modest T like 10.

The forward-backward algorithm

- In order to solve that we use the forward-backward algorithm this is far more efficient. The forward part is based on the computation of terms called the alpha terms. We define the alpha values as follows,

$$\alpha_1(i) = \pi_i b_i(O_1)$$

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(O_{t+1})$$

$$P(O \mid \lambda) = \sum_{i=1}^{N} a_T(i)$$

- We can compute the alpha values inductively in a very efficient way.
- This calculation requires just N^2T multiplications.

The backward phase

- In a similar manner we can define a backward variable called beta that computes the probability of a partial observation sequence from t+1 to T. The beta variable will also be computed inductively but in a backward fashion.

$$\beta_T(i) = 1$$

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_i(O_{t+1}) \beta_{t+1}(j)$$

Solution for the second problem

- Our main goal is to find the "optimal" state sequence. We will do that by maximizing the probabilities of each state individually.
- We will start by defining a set of gamma variables the measure the probabilities that at time t we are at state S_i.

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{P(O \mid \lambda)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)}$$

- The denominator is used just to make gamma a true probability measure.
- Now we can find the best state at each time slot in a local fashion.

$$q_t = \arg\max_{1 \le i \le N} [\gamma_t(i)]$$

- The main problem with this approach is that the optimization is done locally, and not on the whole sequence of states. This can lead either to a local maximum or even to an invalid sequence. In order to solve that problem we use a well known dynamic programming algorithm called the Viterbi algorithm.

The Viterbi Algorithm

- **Intuition**
  - Compute the most likely sequence starting with the empty observation sequence; use this result to compute the most likely sequence with an output sequence of length one; recurse until you have the most likely sequence for the entire sequence of observations.
- **Algorithmic Details**
  - The delta variables compute the highest probability of a partial sequence up to time *t* that ends in state *Si*. The psi variables enables us to accumulate the best sequence as we move along the time slices.
- 1. Initialization:

$$\delta_1(i) = \pi_i b_i(O_1)$$
$$\psi_1(i) = 0$$

Viterbi (Cont).

- Recursion:

$$\delta_t(j) = \max_{1 \le i \le N} \left[\delta_{t-1}(i) a_{ij}\right] b_i(O_t)$$

$$\psi_t(j) = \arg\max_{1 \le i \le N} \left[\delta_{t-1}(i) a_{ij}\right]$$

- Termination:

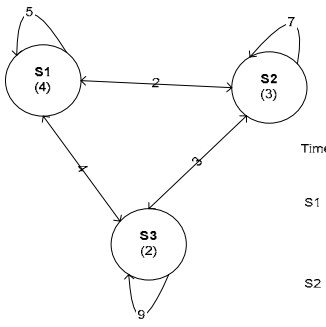
$$P^* = \max_{1 \le i \le N} \left[\delta_T(i)\right] \quad q_T^* = \arg\max_{1 \le i \le N} \left[\delta_T(i)\right]$$

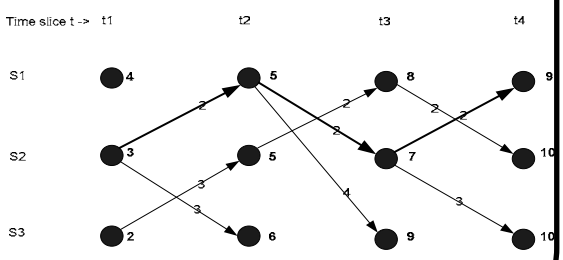
- Reconstruction:

$$q_t^* = \psi_{t+1}(q_{t+1}^*)$$

- For t = T-1,T-2,...,1. $q_1^* q_2^* \cdots q_T^*$ The resulting sequence, , solves Problem 2.

Viterbi (Example)

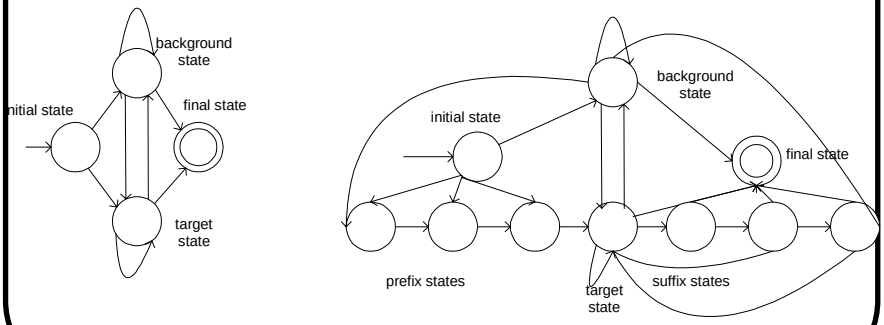




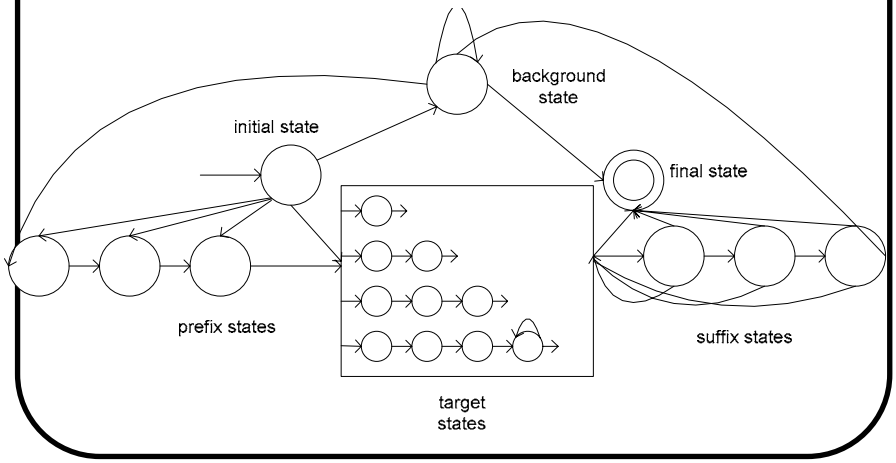
The Just Research HMM

- Each HMM extracts just one field of a given document. If more fields are needed, several HMMs need to be constructed.
- The HMM takes the entire document as one observation sequence.
- The HMM contains two classes of states, background states and target states. The background states emit words in which are not interested, while the target states emit words that constitute the information to be extracted.
- The state topology is designed by hand and only a few transitions are allowed between the states.

Possible HMM Topologies



A more General HMM Architecture



Experimental Evaluation

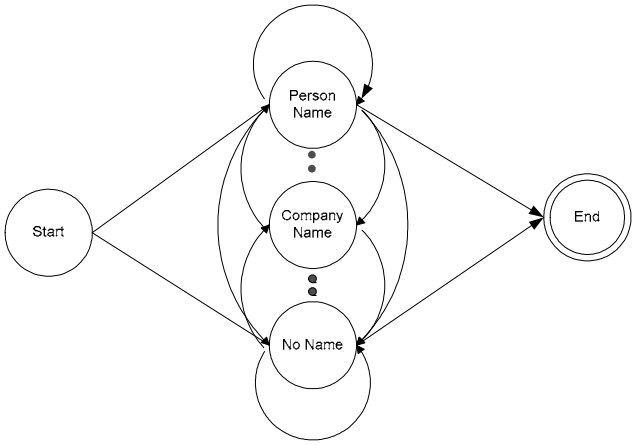
Acquiring Company	Acquired Company	Abbreviation of Acquired Company	Price of Acquisition	Status of Acquisition
30.9%	48.1%	40.1%	55.3%	46.7%

Speaker	Location	Start Time	End Time
71.1%	83.9%	99.1%	59.5%

BBN's Identifinder

- An ergodic bigram model.
- Each Named Class has a separate region in the HMM.
- The number of states in each NC region is equal to |V|. Each word has its own state.
- Rather then using plain words, extended words are used. An extended word is a pair <w,f>, where f is a feature of the word w.

BBN's HMM Architecture



Possible word Features

1. 2 digit number (01)
2. 4 digit number (1996)
3. alphanumeric string (A34-24)
4. digits and dashes (12-16-02)
5. digits and slashes (12/16/02)
6. digits and comma (1,000)
7. digits and period (2.34)
8. any other number (100)
9. All capital letters (CLF)
10. Capital letter and a period (M.)
11. First word of a sentence (The)
12. Initial letter of the word is capitalized (Albert)
13. word in lower case (country)
14. all other words and tokens (;)

Statistical Model

- The design of the formal model is done in levels.
- At the first level we have the most accurate model, which require the largest amount of training data.
- At the lower levels we have back-off models that are less accurate but also require much smaller amounts of training data.
- We always try to use the most accurate model possible given the amount of available training data.

Computing State Transition Probabilities

- When we want to analyze formally the probability of annotating a given word sequence with a set of name classes, we need to consider three different statistical models:
  - A model for generating a name class
  - A model to generate the first word in a name class
  - A model to generate all other words (but the first word) in a name class

Computing the Probabilities : Details

- The model to generate a name class depends on the previous name class and on the word that precedes the name class; this is the last word in the previous name class and we annotate it by w_{-1}. So formally this amounts to $P(NC \mid NC_{-1}, w_{-1})$.
- The model to generate the first word in a name class depends on the current name class and the previous name class and hence is $P(\langle w,f\rangle_{first} \mid NC, NC_{-1})$.
- The model to generate all other words within the same name class depends on the previoues word (within the same name class) and the current name class, so formally it is $P(\langle w,f\rangle \mid \langle w,f\rangle_{-1}, NC)$.

The Actual Computation

$$P(NC \mid NC_{-1}, w_{-1}) = \frac{c(NC, NC_{-1}, w_{-1})}{c(NC_{-1}, w_{-1})}$$

$$P(< w, f >_{first} \mid NC, NC_{-1}) = \frac{c(< w, f >_{first}, NC, NC_{-1})}{c(NC, NC_{-1})}$$

$$P(< w, f > \mid < w, f >_{-1}, NC) = \frac{c(< w, f >, < w, f >_{-1}, NC)}{c(< w, f >_{-1}, NC)}$$

c(<w,f>,<w,f>-1,NC), counts the number of times that we have the pair <w,f> after the pair $<w,f>_{-1}$ and they both are tagged by the name class NC.

Modeling Unknown Words

- The main technique is to create a new entity called UNKNOWN (marked _UNK_), and create statistics for that new entity. All words that were no seen before are mapped to _UNK_.
- split the collection into 2 even parts, and each time use one part for training and one part as a hold out set. The final statistics is the combination of the results from the two runs.
- The statistics needs to be collected for 3 different classes of cases: _UNK_ and then a known word (|V| cases), a known word and then _UNK_ and two consecutive _UNK_ words. This statistics is collected for each name class.

Name Class Back-off Models

- The full model take into account both the previous name class and the previous word ($P(NC| NC_{-1},w_{-1})$
- The first back-off model takes into account just the previous name class ($P((NC| NC_{-1}))$.
- The next back-off model would just estimate the probability of seeing the name class based on the distribution of the various name classes (P(NC)).
- Finally, we use a uniform distribution between all names classes (1/(N+1), where N is number of the possible name classes)

First Word Back-off Models

- The full model takes into account the current name class and the previous name class ($P(<w,f>_{first}| NC, NC_{-1})$).
- The first back-off model takes into account just the current name class ($P(<w,f>_{first}| NC)$).
- The next back-off model, breaks the <w,f> pair and just uses multiplication of two independent events given the current word class ($P(w|NC)P(f|NC)$)
- The next back-off model is a uniform distribution between all pairs of words and features (, where F# is the # of possible word features)

Rest of the Words Back-off Models

- The full model takes into account the current name class and the previous word (P(<w,f>|<w,f>$_{-1}$, NC)).
- The first back-off model takes into account just the current name class (P(<w,f>| NC)).
- The next back-off model, breaks the <w,f> pair and just uses multiplication of two independent events given the current word class (P(*w*|NC)P(*f*|NC))
- The next back-off model is a uniform distribution between all pairs of words and features (, where F# is the # of possible word features)

$$\frac{1}{|W| \cdot |F|}$$

Combining all the models

- The actual probability is a combination of the different models. Each model gets a different weight based on the amount of training available to that model.
- Lets assume we have 4 models (one full model, and 3 back-off models), and we are trying to estimate the probability of P(X|Y). Let P1 be probability of the event according to the full model, and P2, P3, P4 ate the back-off models respectively.
- The weights are computed based on a lambda parameter that is based on each model and it immediate back-off model. For instance λ_1 will adjust the wait between the full model and the first back-off model.

$$\lambda = \left(1 - \frac{c(Y)}{bc(Y)}\right) \frac{1}{1 + \frac{\#(Y)}{bc(Y)}}$$

Analysis

- Where c(Y) the count of event Y according to the full model, and bc(Y) is the count of event Y according to the back-off model. #(Y) is the number of unique outcomes of Y.
- Lambda has two desirable properties
  - If the full model and the back-off model both have the same support for event Y, then Lambda will be 0 and we will use just the full model.
  - If the possible outcomes of Y are distributed uniformly then the weight of lambda will be close to 0 since there is low confidence in the back-off model.

$$\lambda = \left(1 - \frac{c(Y)}{bc(Y)}\right) \frac{1}{1 + \frac{\#(Y)}{bc(Y)}}$$

Example

- We want to compute the probability of P("bank" | "river", "Not-A-Name"). Lets assume that river appears with 3 different words in the Not-A-Name" name class, and in total there are 9 different occurrences of river with any of the 3 words.

$$\lambda_1 = \left(1 - \frac{0}{9}\right)\left(\frac{1}{1 + \frac{3}{9}}\right) = 1 \cdot \frac{3}{4} = \frac{3}{4}$$

- so we will use the full model (P1) with 0.75, and the other back-off models with 0.25. We then compute λ2 which computes the weight of the first back-off model (P2) against the other back-off models, and finally λ3 which is the weight of the second back-off model (P3) against the last back-off model. So to sum up, the probability of P(X|Y) would be:
- P(X|Y) = λ1 * P1(X|Y) + (1 - λ1) * (λ2 * P2(X|Y) + (1 - λ2) * (λ3 * P3(X|Y) + (1 - λ3) * P4(X|Y)))

Using different modalities of text

- **Mixed Case:** Abu Sayyaf carried out an attack on a south western beach resort on May 27, seizing hostages including three Americans. They are still holding a missionary couple, Martin and Gracia Burnham, from Wichita, Kansas, and claim to have beheaded the third American, Guillermo Sobero, from Corona, California. Mr. Sobero's body has not been found.
- **Upper Case:** ABU SAYYAF CARRIED OUT AN ATTACK ON A SOUTH WESTERN BEACH RESORT ON MAY 27, SEIZING HOSTAGES INCLUDING THREE AMERICANS. THEY ARE STILL HOLDING A MISSIONARY COUPLE, MARTIN AND GRACIA BURNHAM, FROM WICHITA, KANSAS, AND CLAIM TO HAVE BEHEADED THE THIRD AMERICAN, GUILLERMO SOBERO, FROM CORONA, CALIFORNIA. MR SOBERO'S BODY HAS NOT BEEN FOUND.
- **SNOR:** ABU SAYYAF CARRIED OUT AN ATTACK ON A SOUTH WESTERN BEACH RESORT ON MAY TWENTY SEVEN SEIZING HOSTAGES INCLUDING THREE AMERICANS THEY ARE STILL HOLDING A MISSIONARY COUPLE MARTIN AND GRACIA BURNHAM FROM WICHITA KANSAS AND CLAIM TO HAVE BEHEADED THE THIRD AMERICAN GUILLERMO SOBERO FROM CORONA CALIFORNIA MR SOBEROS BODY HAS NOT BEEN FOUND.

Experimental Evaluation (MUC 7)

Modality	Language	Rule Based	HMM
Mixed case	English	96.4%	94.9%
Upper case	English	89%	93.6%
SNOR	English	74%	90.7%
Mixed case	Spanish	93%	90%

How much Data is needed to train an HMM?

Number of Tagged Words	English	Spanish
23,000	NA	88.6%
60,000	91.5%	89.7%
85,000	91.9%	NA
130,000	92.8%	90.5%
230,000	93.1%	91.2%
650,000	94.9%	NA

Limitations of the Model

- The context which is used for deciding on the type of each word is just the word the precedes the current word. In many cases, such a limited context may cause classification errors.
- As an example consider the following text fragment "The Turkish company, Birgen Air, was using the plane to fill a charter commitment to a German company,". The token that precedes Birgen is a comma, and hence we are missing the crucial clue **company** which is just one token before the comma.
- Due to the lack of this hint, the IndentiFinder system classified Birgen Air as a location rather than as a company. One way to solve this problem is to augment the model with another token when the previous token is a punctuation mark.

Example - input

<DOCUMENT>
<TYPE>NEWS</TYPE>
<ID> 4 Ryan daughters tied to cash (Thu Feb 13, 9:49 AM)</ID>
<TITLE> 4 Ryan daughters tied to cash </TITLE>
<DATE> Thu Feb 13, 9:49 AM </DATE>
<SOURCE> Yahoo-News </SOURCE>
<BODY> By Matt O'Connor, Tribune staff reporter. Tribune staff reporter Ray Gibson contributed to this report <p> Four of former Gov. George Ryan's daughters shared in almost 10,000 in secret payments from Sen. Phil Gramm's presidential campaign in the mid-1990s, according to testimony Wednesday in federal court. <p>
Alan Drazek, who said he was brought in at Ryan's request as a point man for the Gramm campaign in Illinois, testified he was told by Scott Fawell, Ryan's chief of staff, or Richard Juliano, Fawell's top aide, to cut the checks to the women. According to court records made public Wednesday, Ryan's daughter, Lynda Pignotti, now known as Lynda Fairman, was paid a combined 5,950 in 1995 by the Gramm campaign in four checks laundered through Drazek's business, American Management Resources.
<p>
<p>
In 1996, individual checks went to Ryan daughters Nancy Coghlan, who received 1,725, and Joanne Barrow and Julie R. Koehl, who each pocketed 1,000, the records showed.
<p>
<p>
A source said all four daughters had been given immunity from prosecution by federal authorities and testified before the grand jury investigating Fawell as part of the Operation Safe Road probe.
<p>
<p> Full story at Chicago Tribune <p>
<p>

<p> </BODY>
</DOCUMENT>

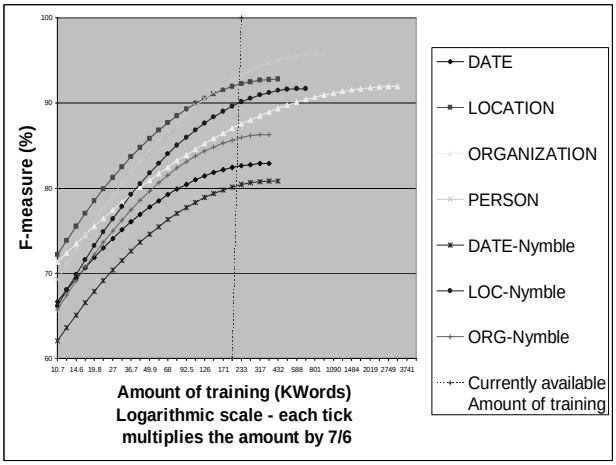
Example - Output

Full story at <LOCATION>Baltimore Sun</LOCATION>
By <PERSON>Matt O ' Connor</PERSON> , <ORGANIZATION>Tribune</ORGANIZATION> staff reporter . Tribune staff reporter <PERSON>Ray Gibson</PERSON> contributed to this report
Four of former Gov . <PERSON>George Ryan</PERSON> ' s daughters shared in almost <MONEY>10 , 000</MONEY> in secret payments from Sen . <PERSON>Phil Gramm</PERSON> ' s presidential campaign in the <DATE>mid - 1990 s</DATE> , according to testimony <DATE>Wednesday</DATE> in federal court .
<PERSON>Alan Drazek</PERSON> , who said he was brought in at <ORGANIZATION>Ryan</ORGANIZATION> ' s request as a point man for the Gramm campaign in <LOCATION>Illinois</LOCATION> , testified he was told by <PERSON>Scott Fawell</PERSON> , Ryan ' s chief of staff , or <PERSON>Richard Juliano</PERSON> , <PERSON>Fawell</PERSON> ' s top aide , to cut the checks to the women . According to court records made public <DATE>Wednesday</DATE> , Ryan ' s daughter , <PERSON>Lynda Pignotti</PERSON> , now known as <PERSON>Lynda Fairman</PERSON> , was paid a combined <PERCENT>5</PERCENT> , 950 in <DATE>1995</DATE> by the Gramm campaign in four checks laundered through Drazek ' s business , <ORGANIZATION>American Management Resources</ORGANIZATION> .
In <DATE>1996</DATE> , individual checks went to Ryan daughters <PERSON>Nancy Coghlan</PERSON> , who received <MONEY>1 , 725</MONEY> , and <PERSON>Joanne Barrow and Julie R . Koehl</PERSON> , who each <MONEY>pocketed 1 , 000</MONEY> , the records showed .
A source said all four daughters had been given immunity from prosecution by federal authorities and testified before the grand jury investigating Fawell as part of the Operation Safe Road probe .
Full story at <ORGANIZATION>Chicago Tribune</ORGANIZATION>

Training: ACE + MUC => MUC

r/p	muc7	ace+muc7
person	91.9/85.5	84.9/88.6
organization	91.1/93.7	83.1/95.9
date	90.9/76.6	59/89.5
time	76.4/77.6	68.6/92.5
location	90.7/91.3	77.7/91.7
money	97.6/82.1	86.6/82.1
percent	93.7/40.54	50/29.6

Results with our new algorithm



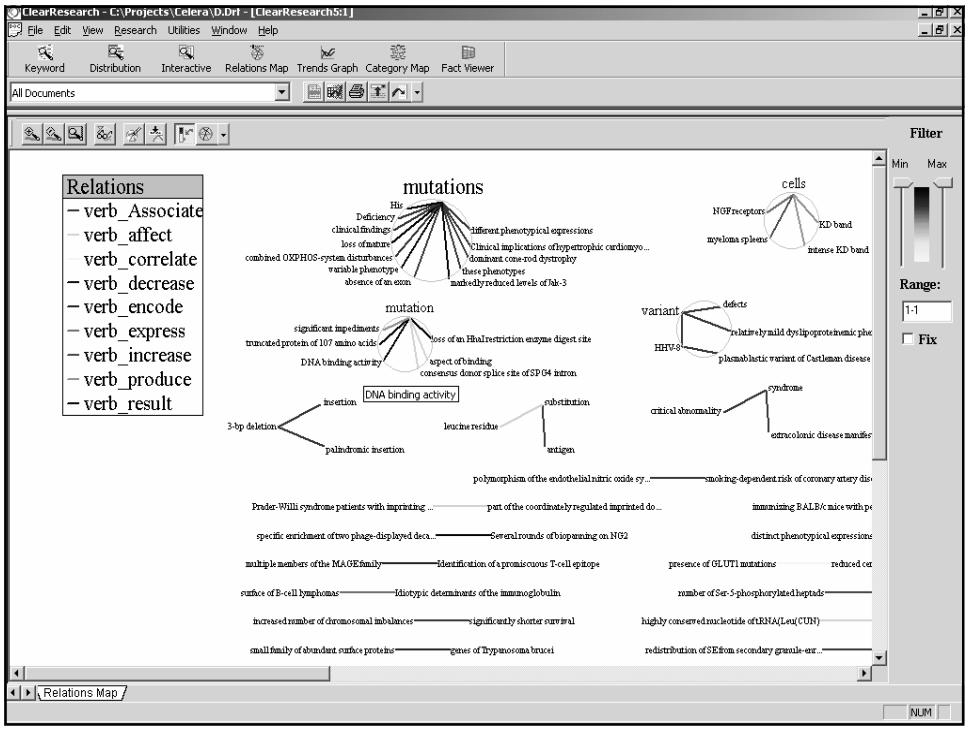
Some Useful Resources

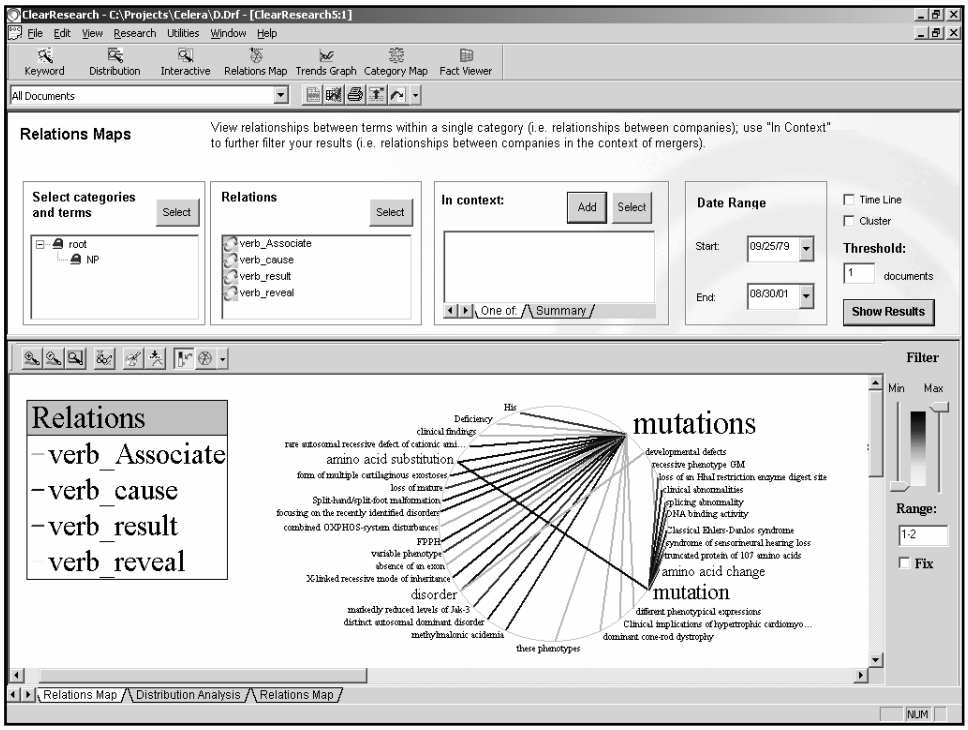
- Linguistic Data Consortium (LDC)
  - Lexicons
  - Annotated Corpora (Text and Speech)
- New Mexico State University
  - Gazetteers
  - Many lists of names
  - Lexicons for different languages
- Various Web Sources
  - CIA World Fact Book
  - Hoovers
  - SEC, Nasdaq (list of public company names)
  - US Census data
  - Private web sites (like Arabic, Persian, Pakistani names)

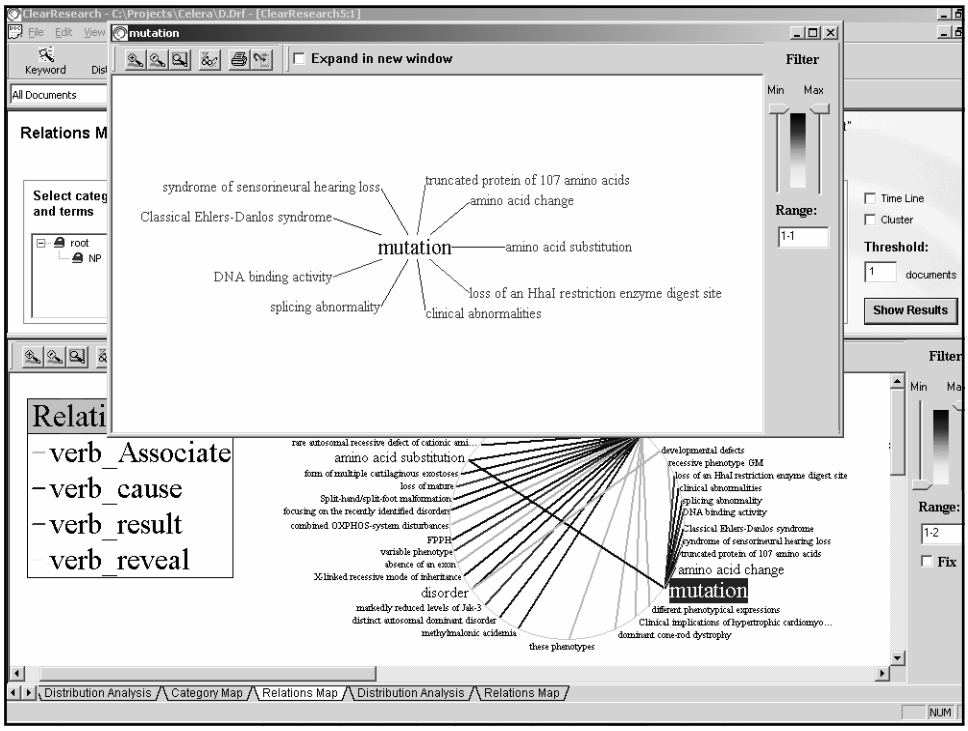
IE in BioTech

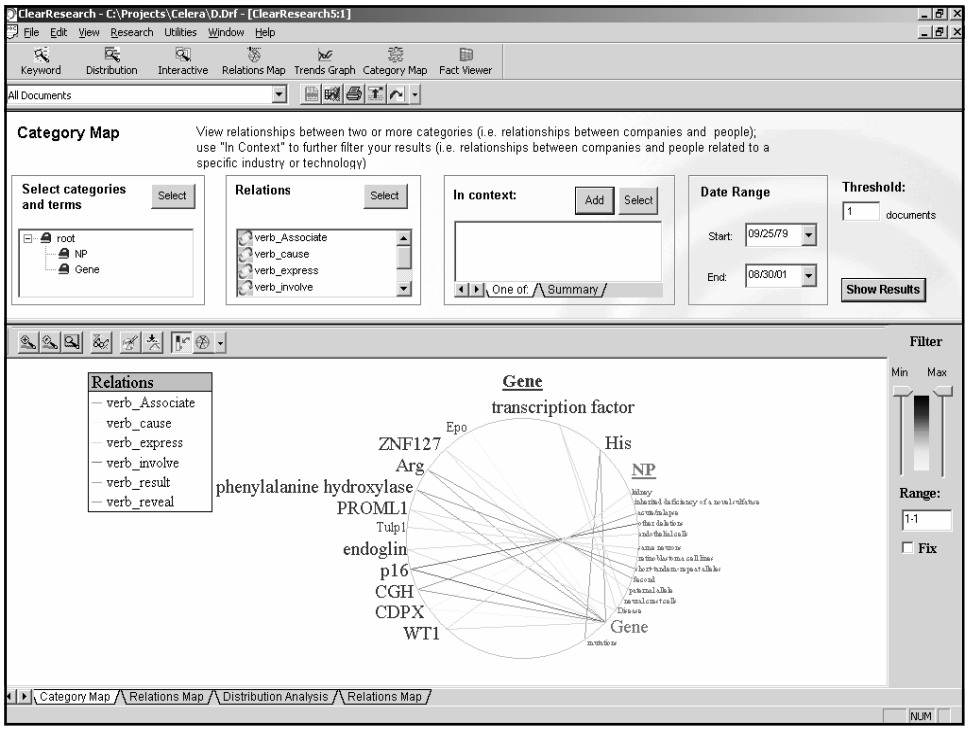
Two Levels of Processing

- Out-of-the-box Entity and Relationship Tagging
  - Identifies entities and categorizes them according to vocabularies (I.e. genes, tissues, …)
  - Identifies relationships based on verbs (and/or verb phrases)
  - Identifies co-occurrence relationships
- Advanced Tagging
  - Enables a complete discovery process (identify entities that are not in any vocabulary)
  - Add constraints, conditions and semantics for more granular relationship identification




ClearResearch - C:\Projects\Celera\D.Drf - [ClearResearch5:1]
File Edit View Research Utilities Window Help
Keyword Distribution Interactive Relations Map Trends Graph Category Map Fact Viewer
All Documents
Relations Maps
View relationships between terms within a single category (i.e. relationships between companies); use "In Context" to further filter your results (i.e. relationships between companies in the context of mergers).
Select categories and terms
Select
root
NP
Relations
verb_Associate
verb_cause
verb_result
verb_reveal
In context:
Add
Select
One of
Summary
Date Range
Start: 09/25/79
End: 08/30/01
Time Line
Cluster
Threshold:
documents
Show Results
Filter
Min
Max
Range:
Fix
Relations
verb_Associate
verb_cause
verb_result
verb_reveal
mutations
amino acid substitution
amino acid change
mutation
disorder
His
Deficiency
clinical findings
developmental defects
recessive phenotype GM
loss of an HhaI restriction enzyme digest site
clinical abnormalities
splicing abnormality
DNA binding activity
Classical Ehlers-Danlos syndrome
syndrome of sensorineural hearing loss
truncated protein of 107 amino acids
FPPH
variable phenotype
absence of an exon
X-linked recessive mode of inheritance
markedly reduced levels of Jak-3
distinct autosomal dominant disorder
methylmalonic acidemia
these phenotypes
different phenotypical expressions
dominant cone-rod dystrophy
Relations Map
Distribution Analysis
Relations Map
NUM


mutation
Expand in new window
Filter
Min
Max
Range:
1-1
syndrome of sensorineural hearing loss
truncated protein of 107 amino acids
amino acid change
Classical Ehlers-Danlos syndrome
mutation
amino acid substitution
DNA binding activity
loss of an HhaI restriction enzyme digest site
splicing abnormality
clinical abnormalities
Distribution Analysis
Category Map
Relations Map
Distribution Analysis
Relations Map
NUM



Advanced Tagging

- Enables a complete discovery process (identify entities that are not in any vocabulary)
- Add constraints, conditions and semantics for more granular relationship identification

Advanced Tagging

Enable a discovery process:

- Tag any tissue, **even if it does not exist in any vocabulary.**

Example:

Any noun-phrase that starts with "`kinase`" is probably a gene/protein

Advanced Tagging

- Enable a deeper, more accurate tagging process

Example:

"`intralysosomal accumulation of glycogen affects function of skeletal muscle`"

Naïve result: <glycogen> affects <function>

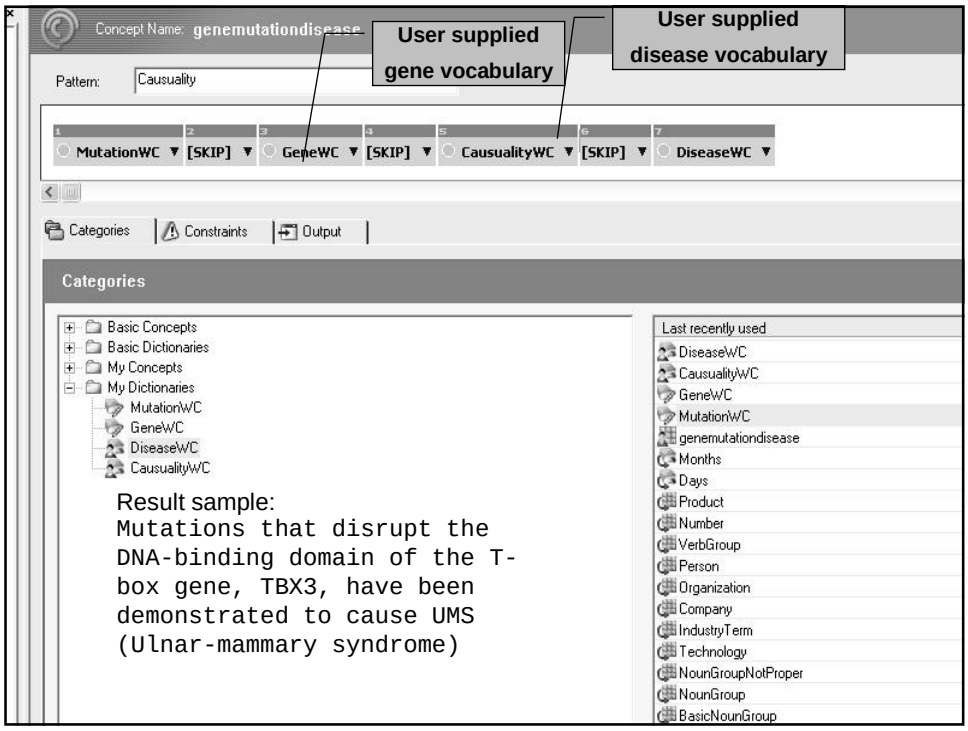
Correct result: <intralysosomal accumulation of glycogen> affects <function of skeletal muscle>

Advanced Tagging

- Enables constraints, conditions and semantics for more granular relationship identification

- Example:

Tag a <causality relationship> between <gene-mutation> and <disease>

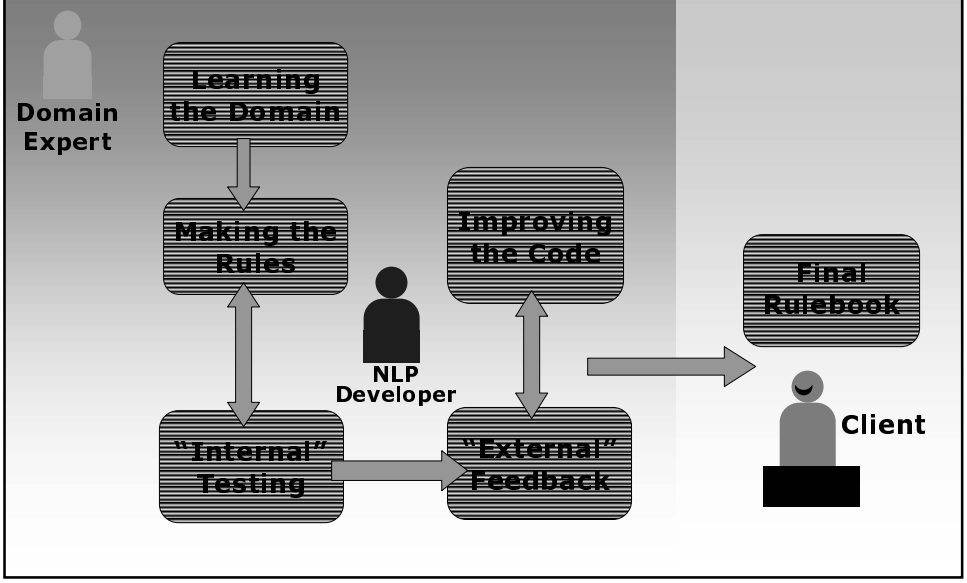


Result sample:

```
Mutations that disrupt the
DNA-binding domain of the T-
box gene, TBX3, have been
demonstrated to cause UMS
(Ulnar-mammary syndrome)
```

Developing IE Rules

The Process – Graphically



The DIAL Language

Declarative **I**nformation **A**nalysis **L**anguage

- Prolog (Logic Programming)- Based ("Rules")
- Unique Pattern-Matching Capabilities
- Enhanced by C++ - based Procedural Elements
- DIAL is implemented in C++
- DIAL code is compiled to C++ functions, from which an executive DLL is produced

DIAL Rules and Rulebooks

Rulebook

Formulation of real-world concepts as sets of DIAL predicates ("Program") that match and extract their occurrences in Texts.

Concepts

- Entities – "Basic" Object Types
- Relationships – Specific connection or occurrence involving two or more entities: (a "fact" or an "event").

Entities and Relationships- (examples)

- **Financial News Domain**
  - Entities: Company, Product, Person (Executive)
  - Relationships: Company Headquarter, Merger
- **Sports News Domain**
  - Entities: Athletes, Sports Teams
  - Relationships : Match Result
- **Genomics Domain**
  - Entities: Gene, Protein, Disease …
- **HTML Pages Parsing / Analysis**
  - Entities: URL (Links within the Page)

DIAL Predicates and Rules

Each Predicate has several Definitions (Rules)

Each Rule may have:

- A Series of Pattern Matching Elements
- A Set of Constraints that the Matched Patterns must obey
- A Set of Assignments of the Predicate's Parameters and/or actions concerning External Variables / Data Structures

Pattern Matching Elements

- An explicit token (String) : e.g. "*announces*"
- A wordclass – a predefined set of phrases that share a common **semantic** function.

  Example :

  ```
  wordclass wcResignation = resignation retirement departure ;
  ```

- (Another) Predicate Call
- Flow Control Operators: Cut, Local and Global Consumption (**@!**, **@>**, **@%**)

Constraints

Used for carrying out "on-the-fly" Boolean Checks on relevant Pattern Matching Elements

Example :

```
verify(InWC(P,@wcAnnounce))
```

Means that the *P* Pattern Matching Element must be a member of the wordclass *wcAnnounce*

A Full Rule – an Example

```
predicate toplevel CompanyName ( STRING
  Company ) ;
CompanyName(C) :-
CapWords -> d
wcCompanyExt->e
[ "." ]->f
  verify (! FirstIn
  (d,@wcCompNameNonStarters ) )
  @! @>
{ C = d+e+f ; } ;
```

Example of an Instance:

Crown Central Petroleum Corp.

Natural Language Processing – Challenges and Solutions

Why NLP is tricky – an Example

PersonLeftPosition (fired) event :

A company fired a person (an executive)

Example:

Banco Mercantil del Perú *yesterday* ***fired*** *the embezzler,* ***Pilar Gonzalez***

The “naïve” Approach

The Pattern:

Possible_Company … FOLLOWED BY
“fired” (or a similar word - wordclass) …
FOLLOWED BY
Possible_Person_Name

The Problem

A “*real*” sentence matching this very same rule:

Alabama Power*'s new gas-****fired*** *electric generating facility at* ***Plant Barry***.

The syntactical function of “*fired*” in the sentence must be a VERB !

Shallow Parsing in IE

- Only Core Constituents are extracted
- No attempt is made at full parses
- Relevant prepositional attachments are extracted.
  - *I saw the man with a telescope*
- *Only adverbials related to location and time are processed, others are ignored.*
- *Quantifiers, modals, and propositional attitudes are ignored, or treated in a simplified way.*

Why not Full Parsing?

- Full Parsing for IE was tried in:
  - SRI Tacitus system (MUC 3)
  - NYU Proteus (Muc-6)
- Main Issues:
  - Slow (combinatorial explosion of possible parses)
  - Erroneous
  - A simple predicate-argument structure is needed.

Coreference

- The general problem is related to co referential relations between expressions
  - Whole – part relationship
  - Containment relationship (set/subset)
- A simplest version is to find which noun phrases refer to the same entity
- An even more restricted version is to limit it just to proper names.
- Example:
  - The President, George Bush, George W. Bush, or even “W”, all refer to the same entity.

Easy and hard in Coreference

- “Mohamed Atta, a suspected leader of the hijackers, had spent time in Belle Glade, Fla., where a crop-dusting business is located. **Atta** and other Middle Eastern men came to South Florida Crop Care nearly every weekend for two months.
- “Will Lee, the firm's general manager, said the **men** asked repeated questions about the crop-dusting business. **He** said the questions seemed "odd," but **he** didn't find the **men** suspicious until after the Sept. 11 attack.”

The “Simple” Coreference

- Proper Names
  - IBM, “International Business Machines”, Big Blue
  - Osama Bin Ladin, Bin Ladin, Usama Bin Laden. (note the variations)
- Definite Noun Phrases
  - The Giant Computer Manufacturer, The Company, The owner of over 600,000 patents
- Pronouns
  - It, he , she, we…..

Coreference Example

Granite Systems provides Service Resource Management (SRM) software for communication service providers with wireless, wireline, optical and packet technology networks. Utilizing Granite' Xng System, carriers can manage inventories of network resources and capacity, define network configurations, order and schedule resources and provide a database of record to other operational support systems (OSSs). An award-winning company, including an Inc. 500 company in 2000 and 2001, Granite Systems enables clients including AT&T Wireless, KPN Belgium, COLT Telecom, ATG and Verizon to eliminate resource redundancy, improve network reliability and speed service deployment. Founded in 1993, the company is headquartered in Manchester, NH with offices in Denver, CO; Miami, FL; London, U.K.; Nice, France; Paris, France; Madrid, Spain; Rome, Italy; Copenhagen, Denmark; and Singapore.

A KE Approach to Corefernce

- Mark each noun phrase with the following:
  - Type (company, person, location)
  - Singular vs. plural
  - Gender (male, female, neutral)
  - Syntactic (name, pronoun, definite/indefinite)
- For each candidate
  - Find accessible antecedents
    - Each antecedent has a different scope
      - Proper names's scope is the whole document
      - Definite clauses's scope is the preceding paragraphs
      - Pronouns might be just the previous sentence, or the same paragraph.
  - Filter by consistency check
  - Order by dynamic syntactic preference

Filtering Antecedents

- George Bush will not match "she", or "it"
- George Bush can not be an antecedent of "The company" or "they"
- Using a sort Hierarchy we can use background information to be smarter
  - Example: "The big automaker is planning to get out the car business. The company feels that it can never longer make a profit making cars."

AutoSlog (Riloff, 1993)

- Creates Extraction Patterns from annotated texts (NPs).
- Uses Sentence analyzer (CIRCUS, Lehnert, 1991) to identify clause boundaries and syntactic constituents (subject, verb, direct object, prepositional phrase)
- It then uses heuristic templates to generate extraction patterns

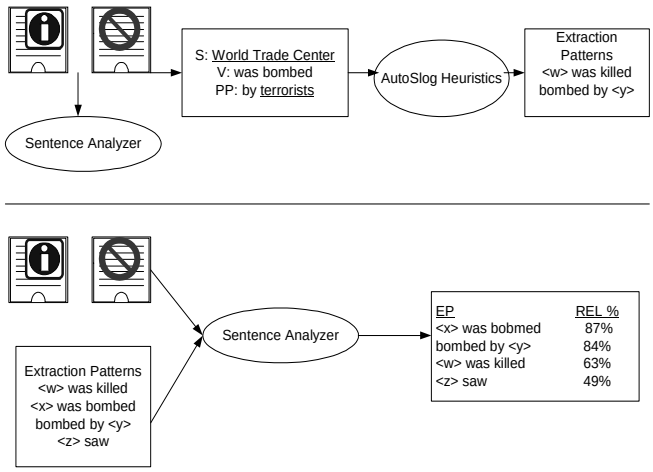
Example Templates

Template	Example
<subj> passive-verb	<victim> was murdered
<subj> aux noun	<victim> was victim
Active-verb <dobj>	Bombed <target>
Noun prep <np>	Bomb against <target>
Active-verb prep <np>	Killed with <instrument>
Passive-verb prep <np>	Was aimed at <target>

AutoSlog-TS (Riloff, 1996)

- It took 8 hours to annotate 160 documents, and hence probably a week to annotate 1000 documents.
- This bottleneck is a major problem for using IE in new domains.
- Hence there is a need for a system that can generate IE patterns from un-annotated documents.
- AutoSlog-TS is such a system.

AutoSlog TS



Top 24 Extraction Patterns

<subj> exploded	Murder of <np>	Assassination of <np>
<subj> was killed	<subj> was kidnapped	Attack on <np>
<subj> was injured	Exploded in <np>	Death of <np>
<subj> took place	Caused <dobj>	Claimed <dobj>
<subj> was wounded	<subj> occured	<subj> was loctated
Took place on <np>	Responsibility for <np>	Occurred on <np>
Was wounded in <np>	Destroyed <dobj>	<subj> was murdered
One of <np>	<subj> kidnapped	Exploded on <np>

Evaluation

- Data Set: 1500 docs from MUC-4 (772 relevant)
- AutoSlog generated 1237 patterns which were manually filtered to 450 in 5 hours.
- AutoSlog-TS generated 32,345 patterns, after discarding singleton patterns, 11,225 were left.
- Rank(EP) = $\frac{rel - freq}{freq} \log_2 freq$
- The user reviewed the the top 1970 patterns and selected 210 of them in 85 minutes.
- AutoSlog achieved better recall, while AutoSlog-TS achieved better precision.

Learning Dictionaries by Bootstrapping (Riloff and Jones, 1999)

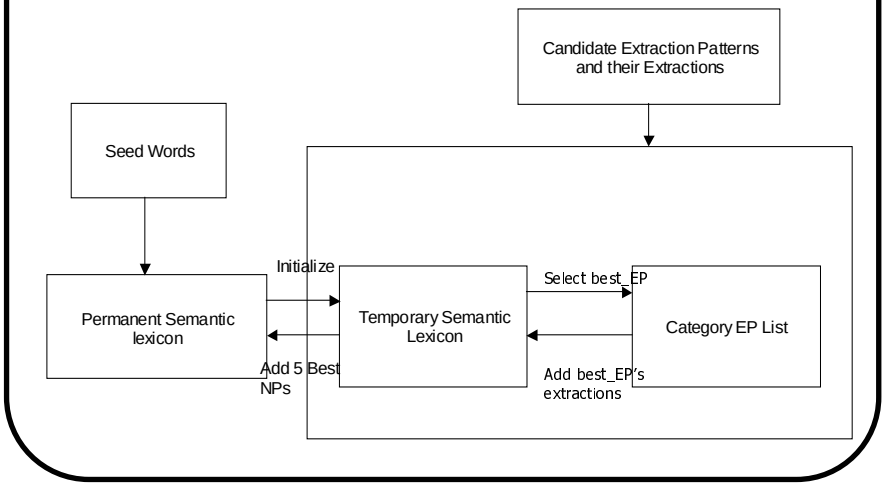
- Learn dictionary entries (semantic lexicon) and extraction patterns simultaneously.
- Use untagged text as a training source for learning.
- Start with a set of seed lexicon entries and using mutual bootstrapping learn extraction patterns and more lexicon entries.

Mutual Bootstrapping Algorithm

- Using AutoSlog generate all possible extraction patterns.
- Apply patterns to the corpus and save results to EPData
- SemLex = Seed Words
- Cat_EPList = {}

1. Score all extraction patterns in EPData
2. Best_EP = highest scoring pattern
3. Add Best_EP to Cat_EPList
4. Add Best_EP's extractions to SemLex
5. Goto 1

Meta Bootstrapping Process



Sample Extraction Patterns

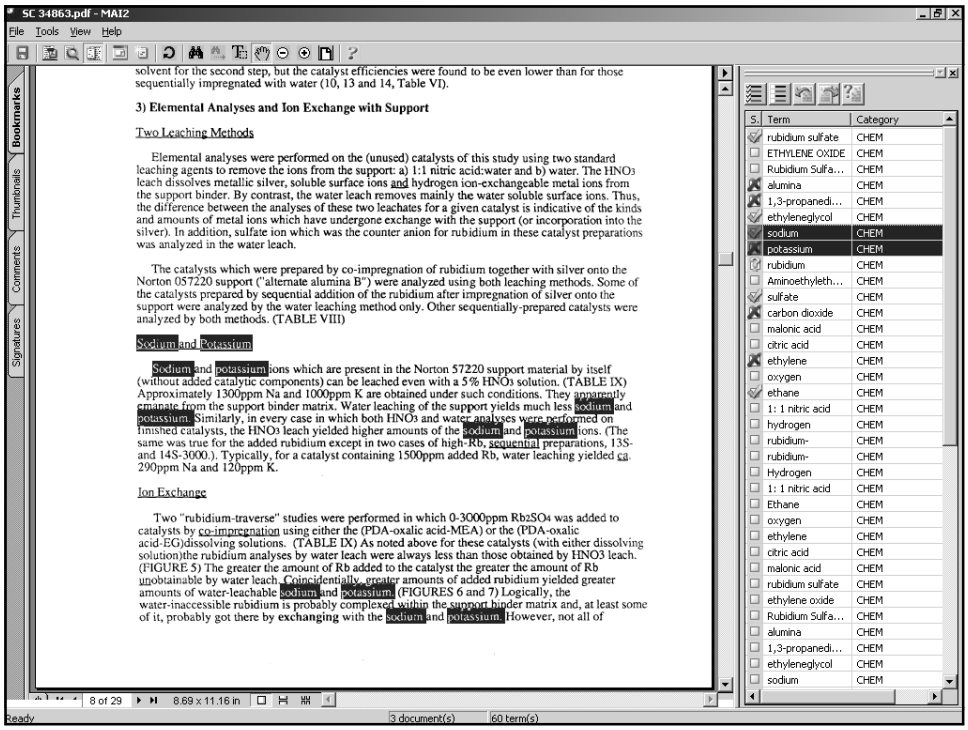
www location	www company	terrorism location
offices in <x>	owned by <x>	living in <x>
facilities in <x>	<x> employed	traveled to <x>
operations in <x>	<x> is distributor	become in <x>
operates in <x>	<x> positioning	sought in <x>
seminars in <x>	motivated <x>	presidents of <x>
activities in <x>	sold to <x>	parts of <x>
consulting in <x>	devoted to <x>	to enter <x>
outlets in <x>	<x> thrive	ministers of <x>
customers in <x>	message to <x>	part in <x>
distributors in <x>	<x> request information	taken in <x>
services in <x>	<x> has positions	returned to <x>
expanded into <x>	offices of <x>	process in <x>

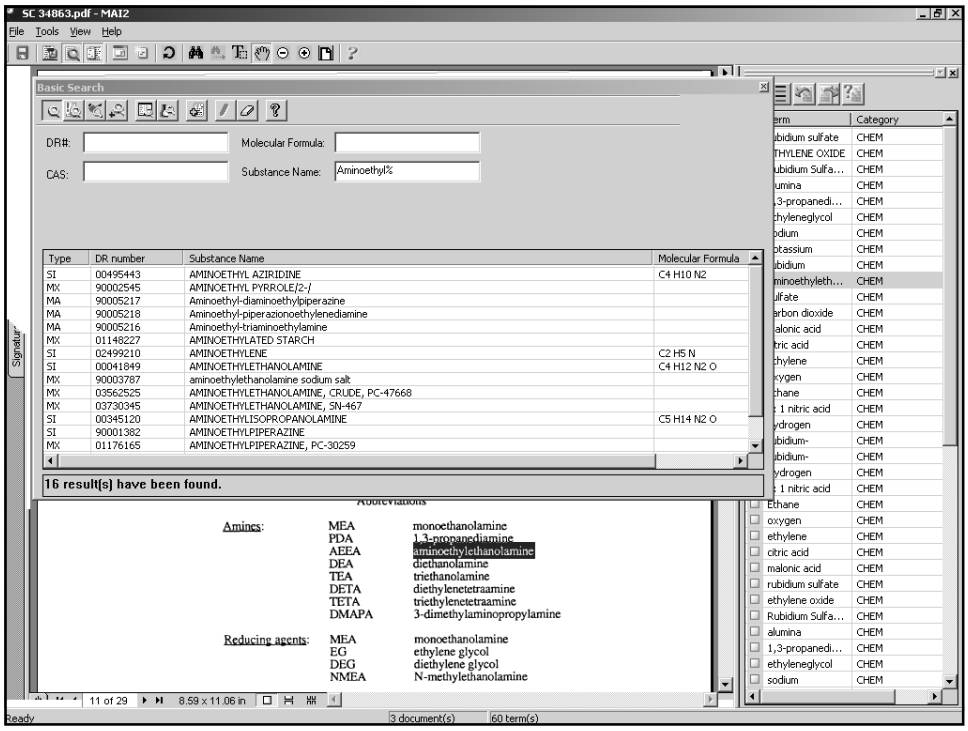
Experimental Evaluation

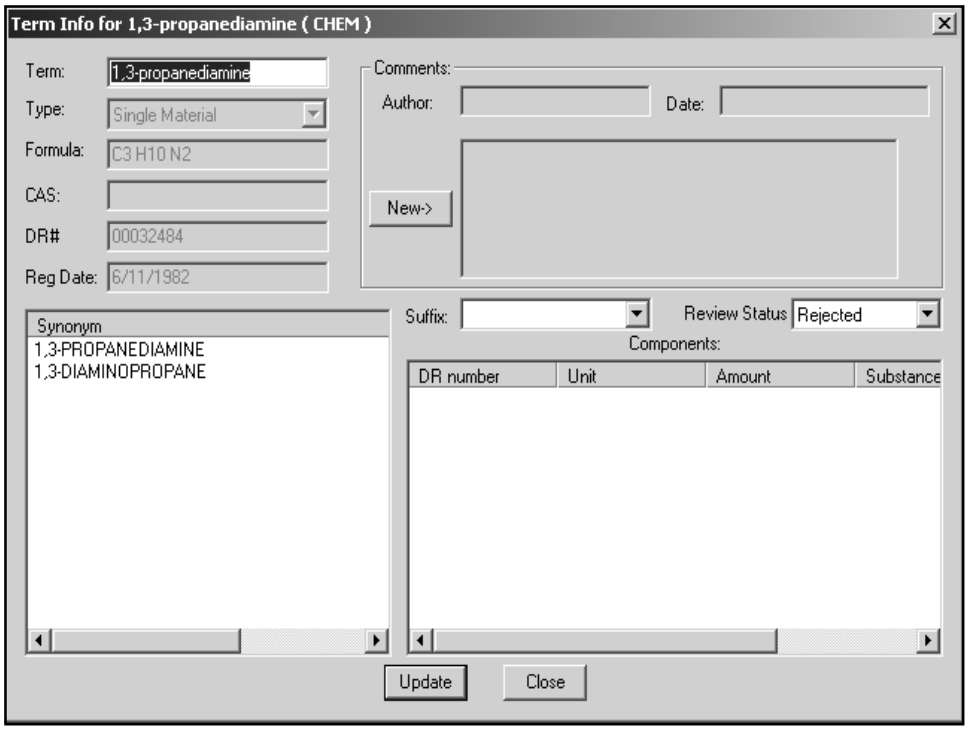
	Iter 1	Iter 10	Iter 20	Iter 30	Iter 40	Iter 50
Web Company	5/5 (1)	25/32 (.78)	52/65 (.80)	72/113 (.64)	86/163 (.53)	95/206 (.46)
Web Location	5/5 (1)	46/50 (.92)	88/100 (.88)	129/150 (.86)	163/200 (.82)	191/250 (.76)
Web Title	0/1 (0)	22/31 (.71)	63/81 (.78)	86/131 (.66)	101/181 (.56)	107/231 (.46)
Terr. Location	5/5 (1)	32/50 (.64)	66/100 (.66)	100/150 (.67)	127/200 (.64)	158/250 (.63)
Terr. Weapon	4/4 (1)	31/44 (.70)	68/94 (.72)	85/144 (.59)	101/194 (.52)	124/244 (.51)

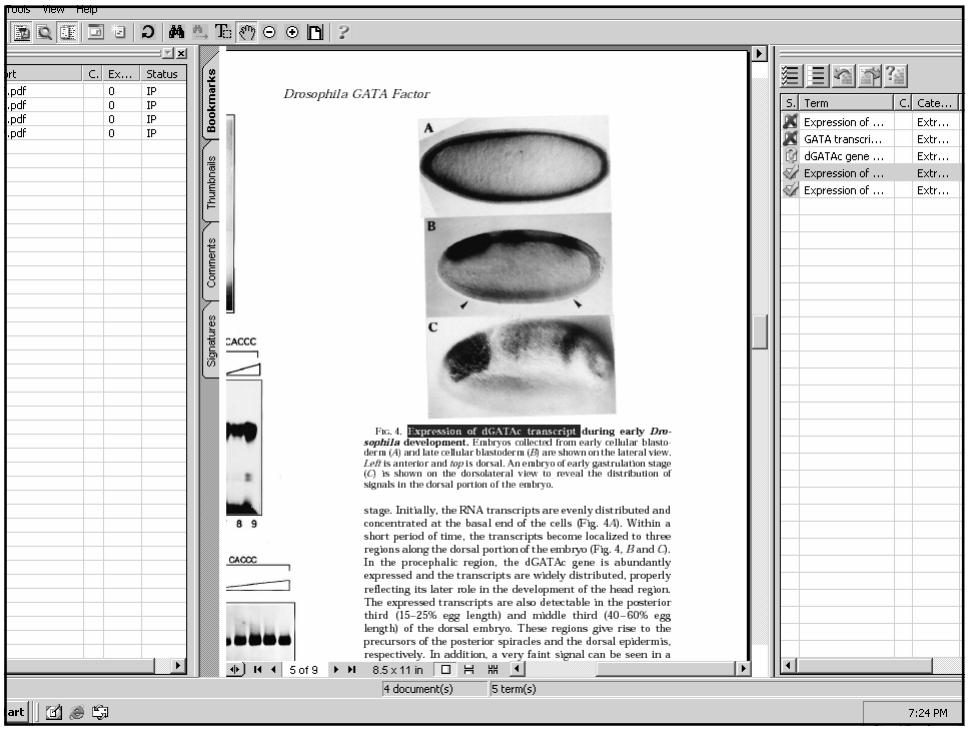
Semi Automatic Approach

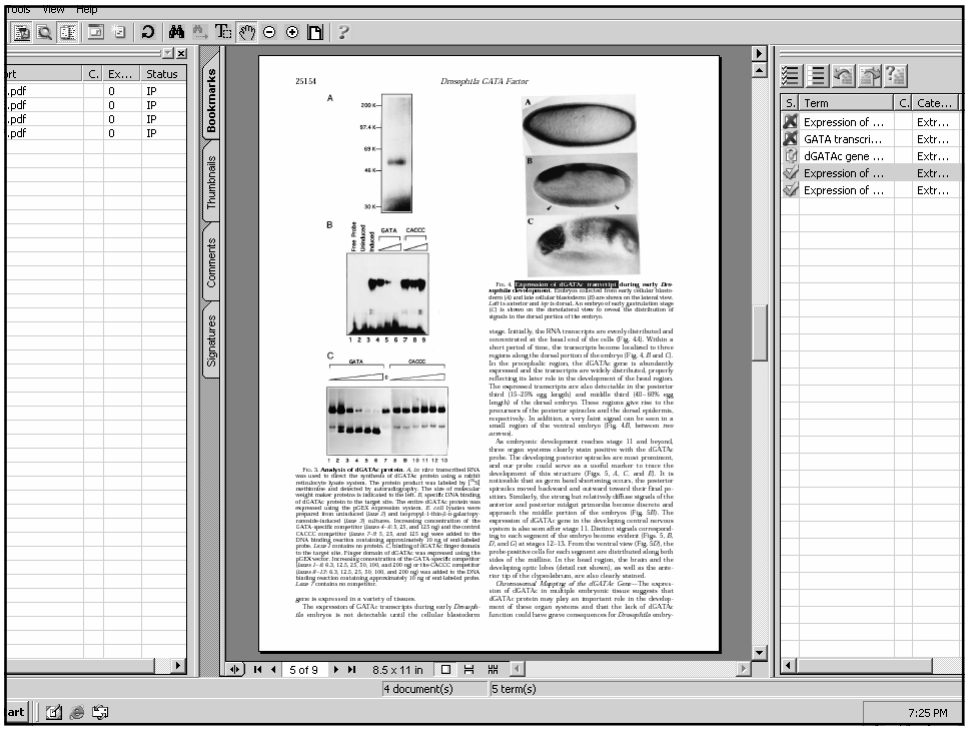
Smart Tagging

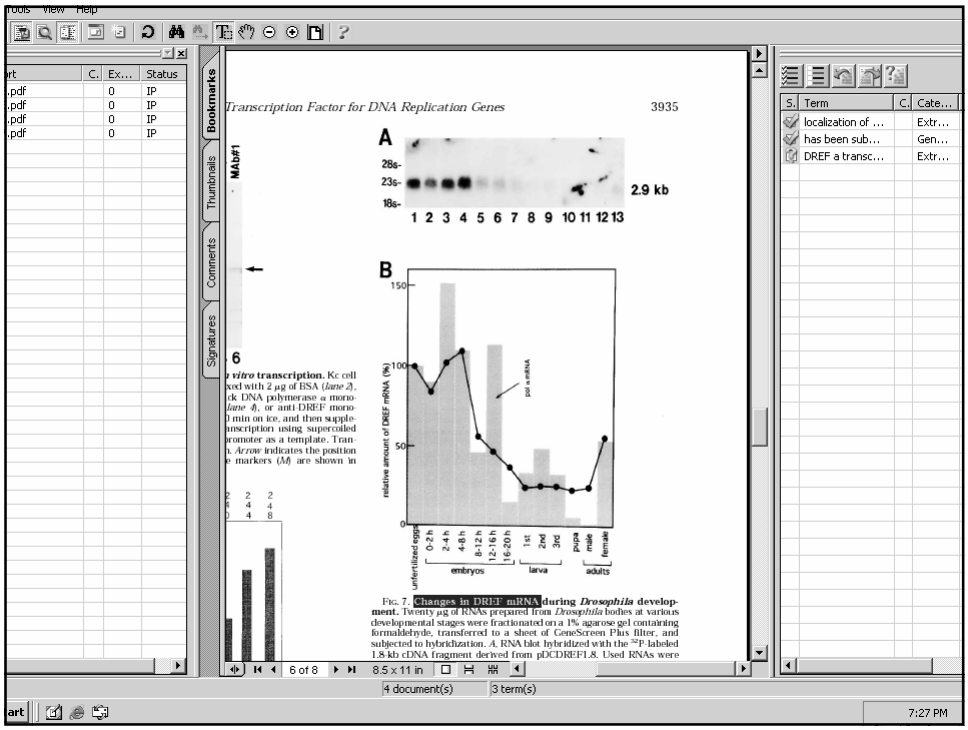

SC 34863.pdf - MAI2
File Tools View Help
Bookmarks
Thumbnails
Comments
Signatures
solvent for the second step, but the catalyst efficiencies were found to be even lower than for those sequentially impregnated with water (10, 13 and 14, Table VI).
3) Elemental Analyses and Ion Exchange with Support
Two Leaching Methods
Elemental analyses were performed on the (unused) catalysts of this study using two standard leaching agents to remove the ions from the support: a) 1:1 nitric acid:water and b) water. The HNO3 leach dissolves metallic silver, soluble surface ions and hydrogen ion-exchangeable metal ions from the support binder. By contrast, the water leach removes mainly the water soluble surface ions. Thus, the difference between the analyses of these two leachates for a given catalyst is indicative of the kinds and amounts of metal ions which have undergone exchange with the support (or incorporation into the silver). In addition, sulfate ion which was the counter anion for rubidium in these catalyst preparations was analyzed in the water leach.
The catalysts which were prepared by co-impregnation of rubidium together with silver onto the Norton 057220 support ("alternate alumina B") were analyzed using both leaching methods. Some of the catalysts prepared by sequential addition of the rubidium after impregnation of silver onto the support were analyzed by the water leaching method only. Other sequentially-prepared catalysts were analyzed by both methods. (TABLE VIII)
Sodium and Potassium
Sodium and potassium ions which are present in the Norton 57220 support material by itself (without added catalytic components) can be leached even with a 5% HNO3 solution. (TABLE IX) Approximately 1300ppm Na and 1000ppm K are obtained under such conditions. They apparently emanate from the support binder matrix. Water leaching of the support yields much less sodium and potassium. Similarly, in every case in which both HNO3 and water analyses were performed on finished catalysts, the HNO3 leach yielded higher amounts of the sodium and potassium ions. (The same was true for the added rubidium except in two cases of high-Rb, sequential preparations, 13S- and 14S-3000.). Typically, for a catalyst containing 1500ppm added Rb, water leaching yielded ca. 290ppm Na and 120ppm K.
Ion Exchange
Two "rubidium-traverse" studies were performed in which 0-3000ppm Rb2SO4 was added to catalysts by co-impregnation using either the (PDA-oxalic acid-MEA) or the (PDA-oxalic acid-EG)dissolving solutions. (TABLE IX) As noted above for these catalysts (with either dissolving solution)the rubidium analyses by water leach were always less than those obtained by HNO3 leach. (FIGURE 5) The greater the amount of Rb added to the catalyst the greater the amount of Rb unobtainable by water leach. Coincidentially, greater amounts of added rubidium yielded greater amounts of water-leachable sodium and potassium. (FIGURES 6 and 7) Logically, the water-inaccessible rubidium is probably complexed within the support binder matrix and, at least some of it, probably got there by exchanging with the sodium and potassium. However, not all of
8 of 29
8.69 x 11.16 in
S. Term Category
rubidium sulfate CHEM
ETHYLENE OXIDE CHEM
Rubidium Sulfa... CHEM
alumina CHEM
1,3-propanedi... CHEM
ethyleneglycol CHEM
sodium CHEM
potassium CHEM
rubidium CHEM
Aminoethyleth... CHEM
sulfate CHEM
carbon dioxide CHEM
malonic acid CHEM
citric acid CHEM
ethylene CHEM
oxygen CHEM
ethane CHEM
1: 1 nitric acid CHEM
hydrogen CHEM
rubidium- CHEM
rubidium- CHEM
Hydrogen CHEM
1: 1 nitric acid CHEM
Ethane CHEM
oxygen CHEM
ethylene CHEM
citric acid CHEM
malonic acid CHEM
rubidium sulfate CHEM
ethylene oxide CHEM
Rubidium Sulfa... CHEM
alumina CHEM
1,3-propanedi... CHEM
ethyleneglycol CHEM
sodium CHEM
Ready
3 document(s)
60 term(s)

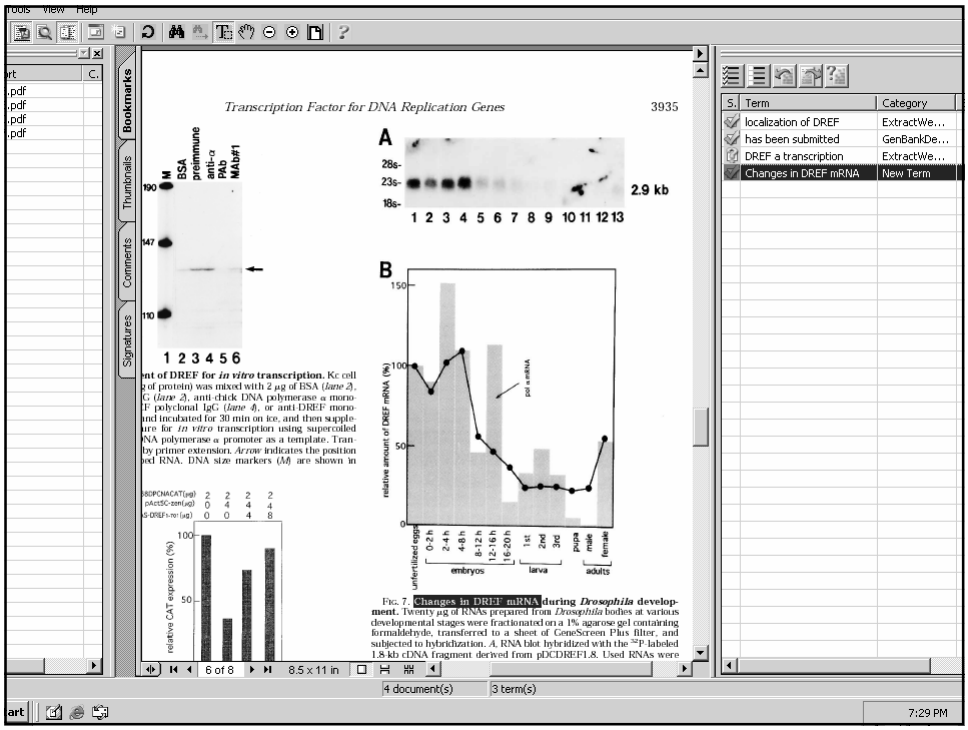

SC 34863.pdf - MAI2
File Tools View Help
Basic Search
DR#:
Molecular Formula:
CAS:
Substance Name: Aminoethyl%
Type DR number Substance Name Molecular Formula
SI 00495443 AMINOETHYL AZIRIDINE C4 H10 N2
MX 90002545 AMINOETHYL PYRROLE/2-/
MA 90005217 Aminoethyl-diaminoethylpiperazine
MA 90005218 Aminoethyl-piperazinoethylenediamine
MA 90005216 Aminoethyl-triaminoethylamine
MX 01148227 AMINOETHYLATED STARCH
SI 02499210 AMINOETHYLENE C2 H5 N
SI 00041849 AMINOETHYLETHANOLAMINE C4 H12 N2 O
MX 90003787 aminoethylethanolamine sodium salt
MX 03562525 AMINOETHYLETHANOLAMINE, CRUDE, PC-47668
MX 03730345 AMINOETHYLETHANOLAMINE, SN-467
SI 00345120 AMINOETHYLISOPROPANOLAMINE C5 H14 N2 O
SI 90001382 AMINOETHYLPIPERAZINE
MX 01176165 AMINOETHYLPIPERAZINE, PC-30259
16 result(s) have been found.
Abbreviations
Amines: MEA monoethanolamine
PDA 1,3-propanediamine
AEEA aminoethylethanolamine
DEA diethanolamine
TEA triethanolamine
DETA diethylenetetraamine
TETA triethylenetetraamine
DMAPA 3-dimethylaminopropylamine
Reducing agents: MEA monoethanolamine
EG ethylene glycol
DEG diethylene glycol
NMEA N-methylethanolamine
11 of 29
8.59 x 11.06 in
Ready
3 document(s)
60 term(s)


Term Info for 1,3-propanediamine (CHEM)
Term:
1,3-propanediamine
Type:
Single Material
Formula:
C3 H10 N2
CAS:
DR#
00032484
Reg Date:
6/11/1982
Comments:
Author:
Date:
New->
Synonym
1,3-PROPANEDIAMINE
1,3-DIAMINOPROPANE
Suffix:
Review Status
Rejected
Components:
DR number
Unit
Amount
Substance
Update
Close


Drosophila GATA Factor
Bookmarks
Thumbnails
Comments
Signatures
Term
Expression of ...
GATA transcri...
dGATAc gene ...
5 of 9
8.5 x 11 in
4 document(s)
5 term(s)
7:24 PM


Tools View Help
Bookmarks
Thumbnails
Comments
Signatures
Term
Expression of ...
GATA transcri...
dGATAc gene ...
Expression of ...
Expression of ...
5 of 9
8.5 x 11 in
4 document(s)
5 term(s)
7:25 PM

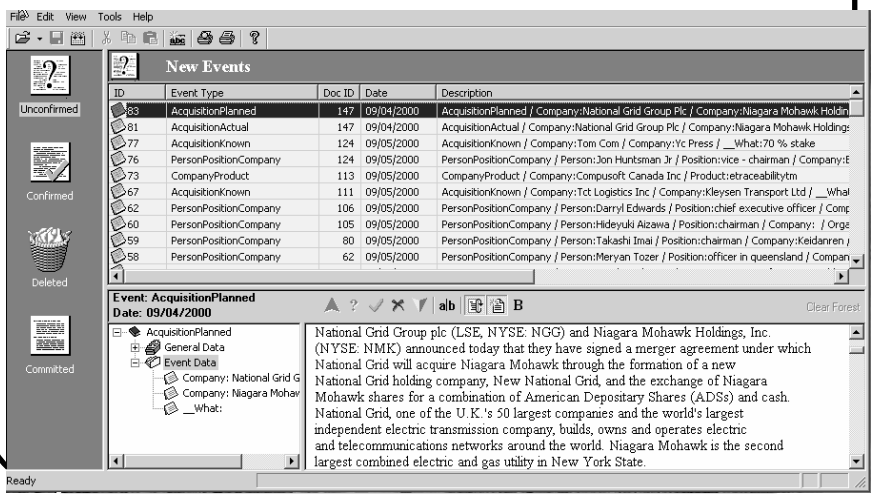

Tools View Help
Bookmarks
Thumbnails
Comments
Signatures
Term
localization of ...
has been sub...
DREF a transc...
6 of 8
8.5 x 11 in
4 document(s)
3 term(s)
7:27 PM



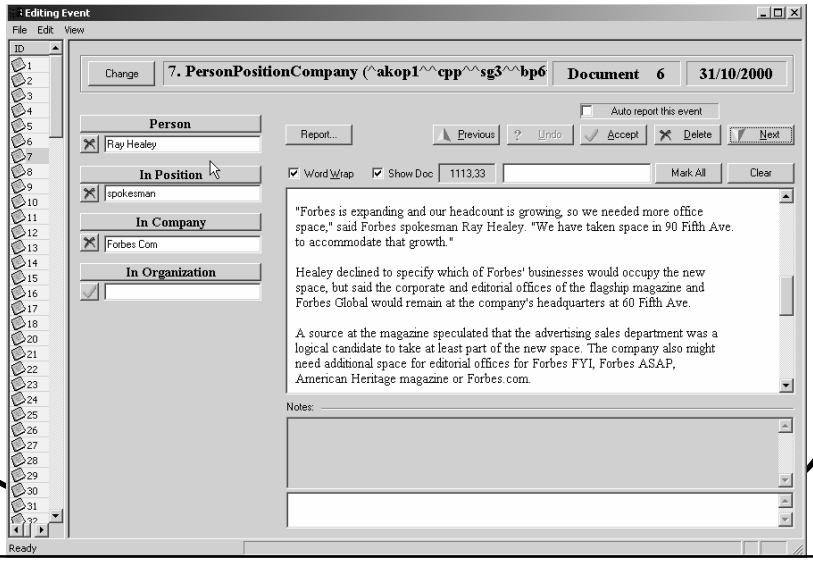
Auditing Environments

Fixing Recall and Precision Problems

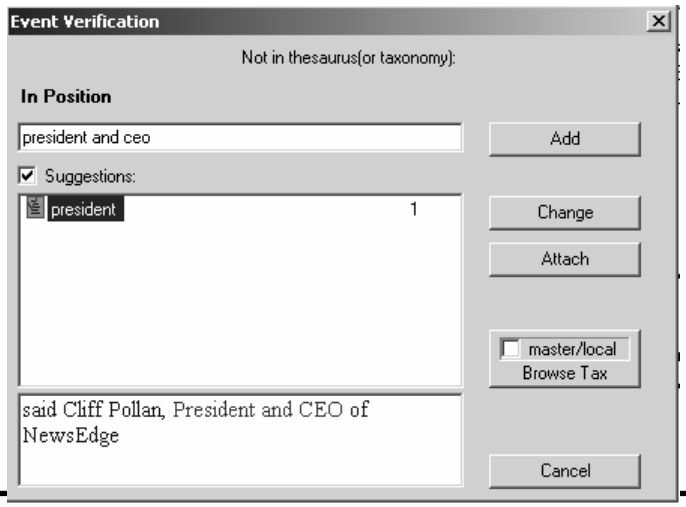
Auditing Events



Auditing a PersonPositionCompany Event



Updating the Taxonomy and the Thesaurus with New Entities

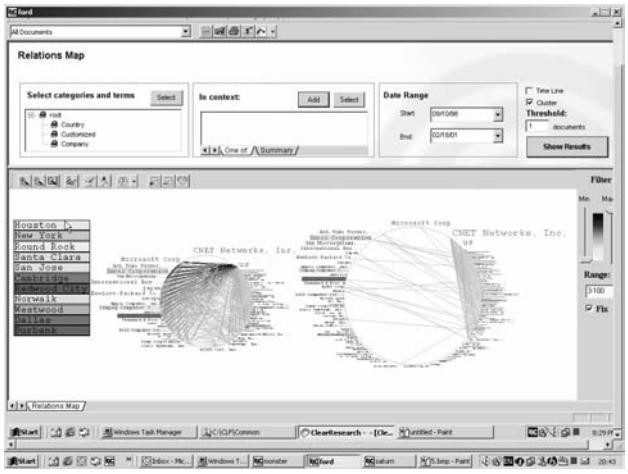


Using Background Information

- Often the analyst would like to use background information about entities, which is not available in the documents.
- This background information enables the analyst to perform filtering, clustering and to automatically color entities in various colors.
  - Examples:
    - CIA World FactBook
    - Hoovers
    - SIC Codes
- A database gateway can create virtual nodes in the taxonomy based on properties stored in the database.

Using the CIA World Fact Book

Using Hoovers



Link Detection

The Importance of Analysis

"Analysis is the key to the successful use of information; it transforms raw data into intelligence. It is the fourth of five stages in the intelligence process: collection, evaluation, collation, analysis, dissemination. Without the ability to perform effective and useful analysis, the intelligence process is reduced to a simple storage and retrieval system for effectively unrelated data"

The Intelligence Analysts Training Manual
Scotland Yard, London

What is Link Analysis? (Law Enforcement context)

Link Analysis is the graphic portrayal of investigative data, done is a manner to facilitate the understanding of large amounts of data, and particularly to allow investigators to develop possible relationships between individuals that otherwise would be hidden by the mass of data obtained.

Coady, 1985

Generalized Definition

Link Analysis is the graphic portrayal of extracted/derived data, done is a manner to facilitate the understanding of large amount of data, and particularly to allow analysts to develop possible relationships between entities that otherwise would be hidden by the mass of data obtained.

Characteristics of Criminal Networks

- Huge size (many nodes, relatively small numbers of links)
- Incompleteness
  - Not all “real” nodes and links will be observed.
- Fuzzy Boundaries
  - Organizations are often interlaced
- Dynamic
  - Each link strength has distribution over time

Notions of Centrality(node)

- Degree
  - # of other nodes to which it is **directly** linked.
- Betweeness
  - # of geodesics (shortest paths between 2 other nodes) which pass through it.
- Closeness
  - Minimizing the maximum of the minimal path length to other nodes in the network. (the central nodes have a minimal radius).
- Center of gravity
- Point Strength
  - Increase in the number of maximal connected subcomponents upon removal of the node.
- Business
  - How busy is each node in transmitting information

Applications of Centrality

- Targeting
  - Betweeness
  - Business
- Identification of Network Vulnerability
  - Betweeness
  - Point Strength
  - Business

Equivalence

- Substitutability
  - Objects a, b of category C are structurally equivalent if, for any relation M and any object x of C, aMx iff bMx and xMa iff xMb.
- Stochastic Equivalence
  - Given a stochastic multigraph X, actors i and j are stochastically equivalent iff the probability of any event concerning X is unchanged by interchanging actors i and j.
  - Two network nodes are stochastically equivalent if the probabilities of them being related to any other individual are the same.
- Role Equivalence
  - In a network X, a and b are role equivalent if there exist an automorphism f of X which maps a into b and b into a, and which is link preserving (f(a) = b, f(b) = a; f(c) is linked to f(d) iff c is linked to d).
  - Two nodes can be role equivalent even if they are in two disconnected components. Obviously they are not stochastically equivalent.

Applications of Equivalence

- Assessment of network vulnerability
  - Whether a target individual has a substitute or not.
- Detecting Aliases
  - No link joining them directly
  - Many paths of length 2 connecting them
- Role equivalence can be used for template matching. Building models and seeking a match within the link graph.

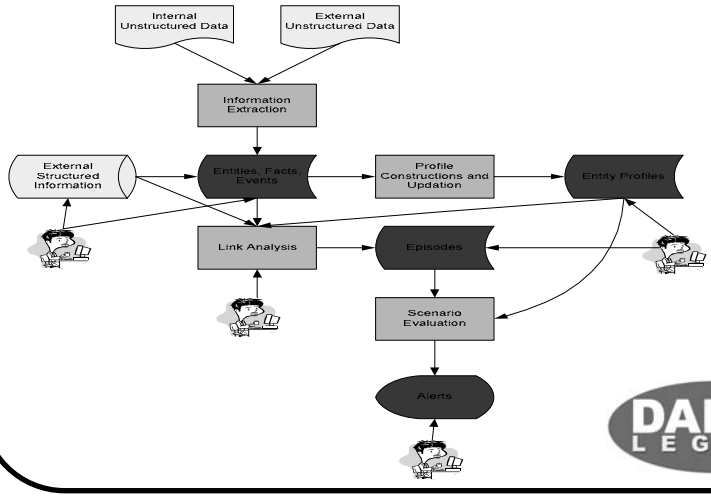
Weak Ties

- Weak Ties are the ties which add most to the efficiency of communication within the network. They usually connect cliques to the outside world.
- Detecting activity of the weak ties usually signals a critical operation.
- Eliminating weak ties may hurt the network the most.

How to derive relationships?

- Wayne Baker and Robert Faulkner looked at archival data to derive relationship data.
  - The data they used to analyze illegal price-fixing networks were mostly court documents and sworn testimony. This data included accounts of observed interpersonal relationships from various witnesses.
- Bonnie Erickson (Erickson, 1981) talk about trusted prior contacts for the effective functioning of a secret society.
  - The 19 hijackers appeared to have come from a network that had formed while they were completing terrorist training in Afghanistan.
  - Many were school mates, some had lived together for years, and others were related by kinship ties.
  - Deep trusted ties, that were not easily visible to outsiders, wove this terror network together.

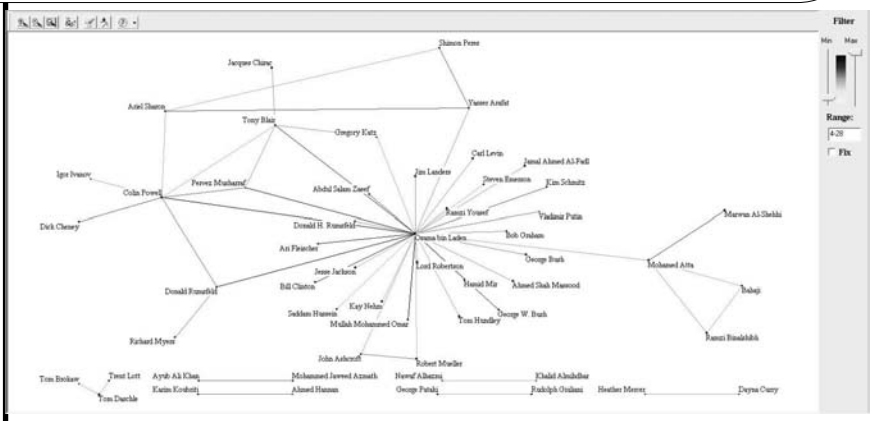
A Complete Link Detection System



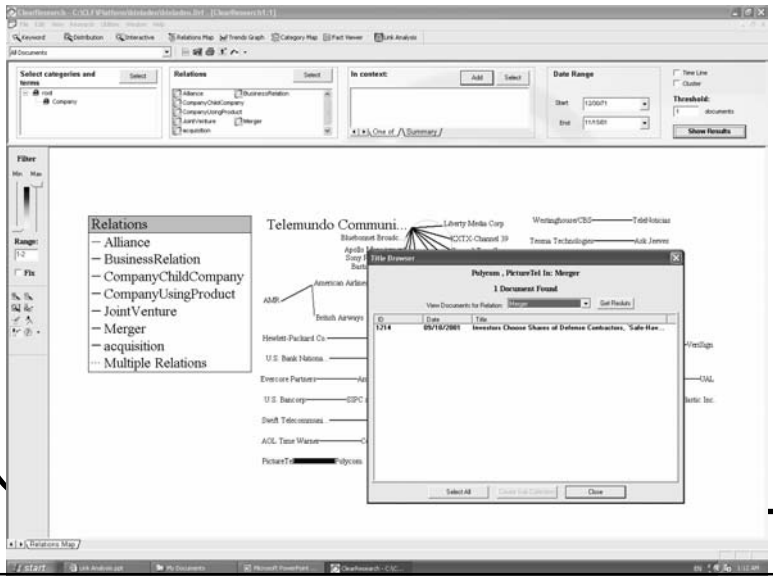
Types of Link Detection Questions:

- Who is Central in the organization?
- Which 3 individuals' removal or incapacitation would sever this drug-supply network?
- What role or roles does specific individual appear to be playing in a given organization?
- Which communication channels without a terrorist organization are worth monitoring?
- What significant changes have taken place in the supply operation of a given organization since this time last year?

Spring Graph of People Co Occurrence Graph



Visualizing Relationship Between Companies

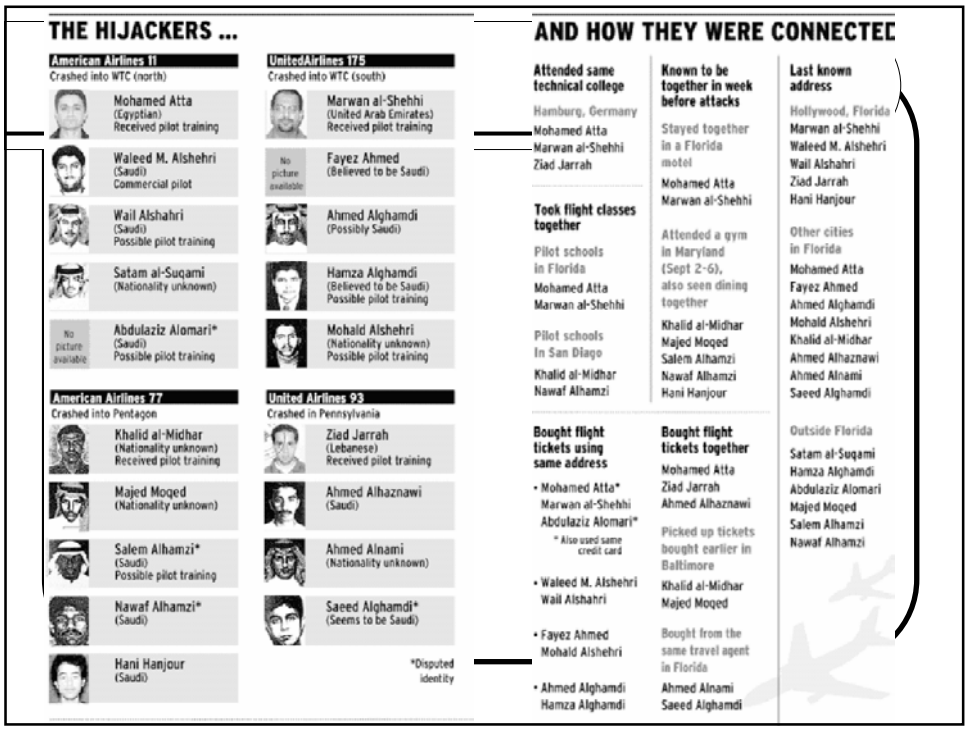


Link Analysis in the 9/11 Context (Krebs 2001)

Analysis (Krebs 2001)

- many on the same flight were more than two steps away from each other. A strategy for keeping cell members distant from each other, and from other cells, minimizes damage to the network if a cell member is captured or otherwise compromised.
- "Those who were trained to fly didn't know the others. One group of people did not know the other group."
  - Usama bin Laden

	Average Path Length
Contacts	4.75
Contacts + Shortcuts	2.79



Trusted Prior Contacts

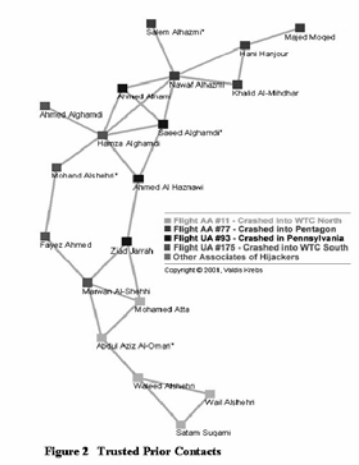


Figure 2 Trusted Prior Contacts

So How was the “work” done?

- Special Meetings were held to connect distant parts of the network.
- These meetings created transitory shortcuts (watts, 1999).
- One of the known meetings was held in Las Vegas.
- As a results, 6 shortcuts were added to the network and it reduced the average path length by 40%.
- All pilots formed a small clique!

Measuring Network Parameters

Degrees		Betweeness		Closeness	
Mohammad Atta	0.361	Mohammad Atta	0.588	Mohammad Atta	0.587
Marwan Al-Shehhi	0.295	Essid Sami Ben Kemais	0.252	Marwan Al-Shehhi	0.466
Hani Hanjour	0.213	Zacarias Moussaoui	0.232	Hani Hanjour	0.445
Essid Sami Ben Kemais	0.18	Nawaf Alhazmi	0.154	Nawaf Alhazmi	0.442
Nawaf Alhazmi	0.18	Hani Hanjour	0.126	Ramzi Bin Al-Shibh	0.436
Ramzi Bin Al-Shibh	0.164	Djamal Beghal	0.105	Zacarias Moussaoui	0.436
Ziad Jarrah	0.164	Marwan Al-Shehhi	0.088	Essid Sami Ben Kemais	0.433

Ranking the Terrorists

Degrees		Betweenness		Closeness	
0.417	Mohamed Atta	0.334	Nawaf Alhazmi	0.571	Mohamed Atta
0.389	Marwan Al-Shehhi	0.318	Mohamed Atta	0.537	Nawaf Alhazmi
0.278	Hani Hanjour	0.227	Hani Hanjour	0.507	Hani Hanjour
0.278	Nawaf Alhazmi	0.158	Marwan Al-Shehhi	0.500	Marwan Al-Shehhi
0.278	Ziad Jarrah	0.116	Saeed Alghamdi*	0.480	Ziad Jarrah
0.222	Ramzi Bin al-Shibh	0.081	Hamza Alghamdi	0.429	Mustafa al-Hisawi
0.194	Said Bahaji	0.080	Waleed Alshehri	0.429	Salem Alhazmi*
0.167	Hamza Alghamdi	0.076	Ziad Jarrah	0.424	Lotfi Raissi
0.167	Saeed Alghamdi*	0.064	Mustafa al-Hisawi	0.424	Saeed Alghamdi*
0.139	Lotfi Raissi	0.049	Abdul Aziz Al-Omari*	0.419	Abdul Aziz Al-Omari*
0.139	Zakariya Essabar	0.033	Satam Suqami	0.414	Hamza Alghamdi
0.111	Agus Budiman	0.031	Fayez Ahmed	0.414	Ramzi Bin al-Shibh
0.111	Khalid Al-Mihdhar	0.030	Ahmed Al Haznawi	0.409	Said Bahaji
0.111	Mounir El Motassadeq	0.026	Nabil al-Marabh	0.404	Ahmed Al Haznawi
0.111	Mustafa al-Hisawi	0.016	Raed Hijazi	0.400	Zakariya Essabar
0.111	Nabil al-Marabh	0.015	Lotfi Raissi	0.396	Agus Budiman
0.111	Rayed Abdullah	0.012	Mohand Alshehri*	0.396	Khalid Al-Mihdhar
0.111	Satam Suqami	0.011	Khalid Al-Mihdhar	0.391	Ahmed Alnami
0.111	Waleed Alshehri	0.010	Ramzi Bin al-Shibh	0.391	Mounir El Motassadeq
0.083	Abdul Aziz Al-Omari*	0.007	Salem Alhazmi*	0.387	Fayez Ahmed
0.083	Abdussattar Shaikh	0.004	Ahmed Alghamdi	0.387	Mamoun Darkazanli
0.083	Ahmed Al Haznawi	0.004	Said Bahaji	0.371	Zacarias Moussaoui
0.083	Ahmed Alnami	0.002	Rayed Abdullah	0.367	Ahmed Khalil Al-Ani
0.083	Fayez Ahmed	0.000	Abdussattar Shaikh	0.360	Abdussattar Shaikh
0.083	Mamoun Darkazanli	0.000	Agus Budiman	0.360	Osama Awadallah
0.083	Osama Awadallah	0.000	Ahmed Alnami	0.353	Mohamed Abdi
0.083	Raed Hijazi	0.000	Ahmed Khalil Al-Ani	0.350	Rayed Abdullah
0.083	Salem Alhazmi*	0.000	Bandar Alhazmi	0.343	Bandar Alhazmi
0.056	Ahmed Alghamdi	0.000	Faisal Al Salmi	0.343	Faisal Al Salmi
0.056	Bandar Alhazmi	0.000	Majed Moqed	0.343	Mohand Alshehri*
0.056	Faisal Al Salmi	0.000	Mamoun Darkazanli	0.340	Majed Moqed
0.056	Mohand Alshehri*	0.000	Mohamed Abdi	0.340	Waleed Alshehri
0.056	Wail Alshehri	0.000	Mounir El Motassadeq	0.330	Nabil al-Marabh
0.056	Zacarias Moussaoui	0.000	Osama Awadallah	0.327	Raed Hijazi
0.028	Ahmed Khalil Al-Ani	0.000	Wail Alshehri	0.319	Ahmed Alghamdi
0.028	Majed Moqed	0.000	Zacarias Moussaoui	0.298	Satam Suqami
0.028	Mohamed Abdi	0.000	Zakariya Essabar	0.271	Wail Alshehri
0.128	MEAN	0.046	MEAN	0.393	MEAN
0.306	CENTRALIZATION	0.296	CENTRALIZATION	0.372	CENTRALIZATION

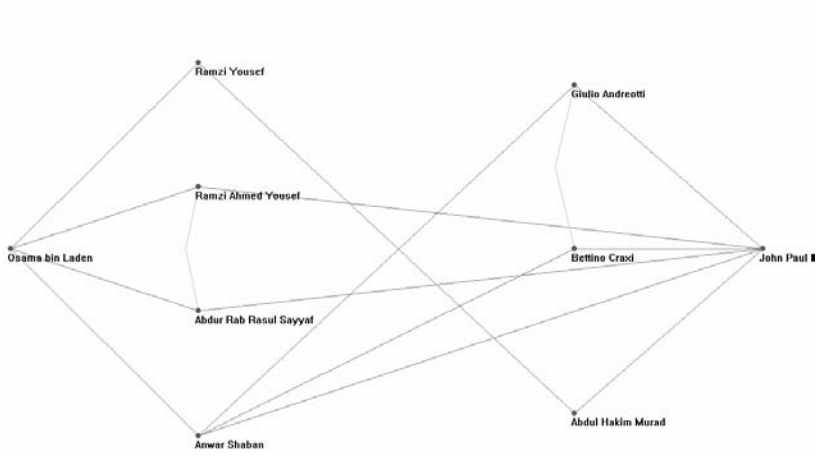
* suspected to have false identification

Building Networks

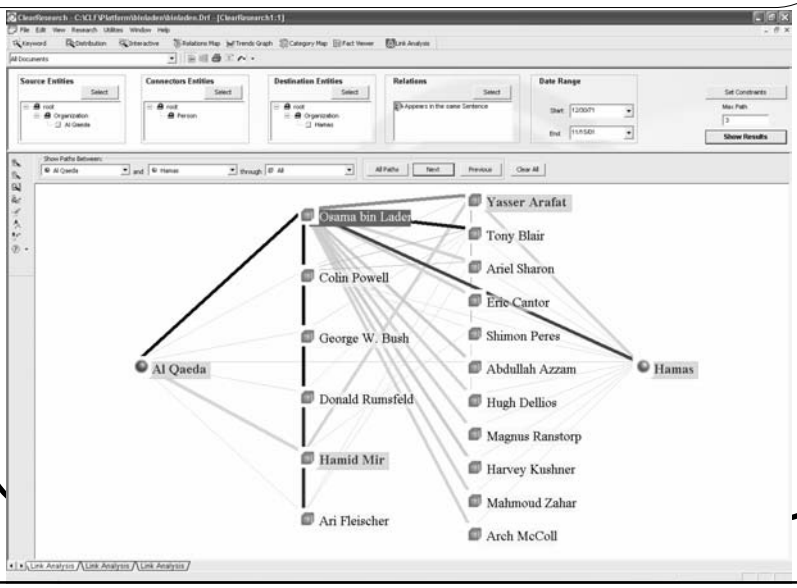
Table 4. Networks to Map

Relationship / Network	Data Sources
1. Trust	Prior contacts in family, neighborhood, school, military, club or organization. Public and court records. Data may only be available in suspect's native country.
2. Task	Logs and records of phone calls, electronic mail, chat rooms, instant messages, web site visits. Travel records. Human intelligence – observation of meetings and attendance at common events.
3. Money & Resources	Bank account and money transfer records. Pattern and location of credit card use. Prior court records. Human intelligence – observation of visits to alternate banking resources such as Hawala.
4. Strategy & Goals	Web sites. Videos and encrypted disks delivered by courier. Travel records. Human intelligence – observation of meetings and attendance at common events

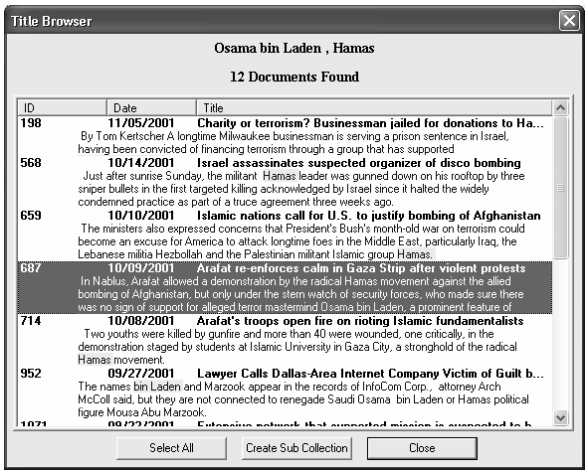
Example



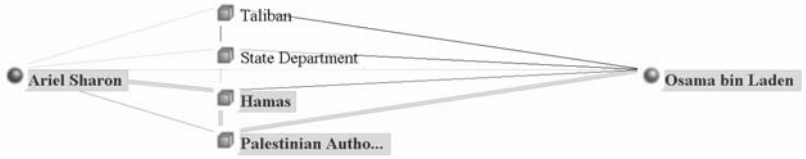
Another Example



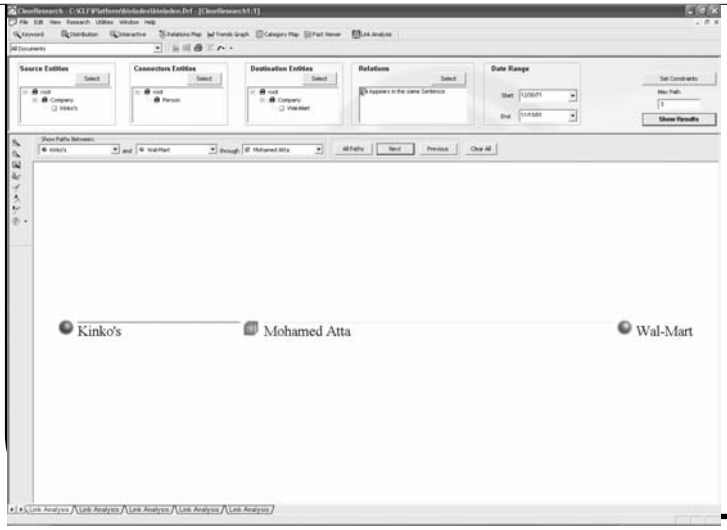
Drilling Down



Person-Organization-Person



Which Person Connects Wal-Mart and Kinko's?



Conclusions

- Information Extraction is a mature technology that can create a solid foundation for real text mining.
- Entity extraction can be performed with very high accuracy (~95% breakeven).
- Entity extraction can be implemented either by using rule based approaches or by using machine learning approaches (HMM or TBR)
- Link Detection is a very effective method to uncover indirect relationships between entities.

Future Work

- Machine learning approaches for relationship extraction
- Hybrid approaches (rules based + machine learning) for relationship extraction
- Visual authoring environments for rule based methods
- Visual environment for defining complex link detection queries.

References I

- Anand T. and Kahn G., 1993. Opportunity Explorer: Navigating Large Databases Using Knowledge Discovery Templates. In Proceedings of the 1993 workshop on Knowledge Discovery in Databases.
- Appelt, Douglas E., Jerry R. Hobbs, John Bear, David Israel, and Mabry Tyson, 1993. ``FASTUS: A Finite-State Processor for Information Extraction from Real-World Text'', Proceedings. IJCAI-93, Chambery, France, August 1993.
- Appelt, Douglas E., Jerry R. Hobbs, John Bear, David Israel, Megumi Kameyama, and Mabry Tyson, 1993a. ``The SRI MUC-5 JV-FASTUS Information Extraction System'', Proceedings, Fifth Message Understanding Conference (MUC-5), Baltimore, Maryland, August 1993.
- Bookstein A., Klein S.T. and Raita T., 1995. Clumping Properties of Content-Bearing Words. In Proceedings of SIGIR'95.
- Brachman R. J., Selfridge P. G., Terveen L. G., Altman B., Borgida A., Halper F., Kirk T., Lazar A., McGuinness D. L., and Resnick L. A., 1993. Integrated Support for Data Archaeology. International Journal of Intelligent and Cooperative Information Systems 2(2):159-185.
- Brill E., 1995. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging, Computational Linguistics, 21(4):543-565
- Church K.W. and Hanks P., 1990. Word Association Norms, Mutual Information, and Lexicography, Computational Linguistics, 16(1):22-29
- Cohen W. and Singer Y., 1996. Context Sensitive Learning Methods for Text categorization. In Proceedings of SIGIR'96.
- Cohen. W., "Compiling Prior Knowledge into an Explicit Bias". Working notes of the 1992 AAAI spring symposium on knowledge assimilation. Stanford, CA, March 1992.
- Daille B., Gaussier E. and Lange J.M., 1994. Towards Automatic Extraction of Monolingual and Bilingual Terminology, In Proceedings of the International Conference on Computational Linguistics, COLING'94, pages 515-521.
- Dumais, S. T., Platt, J., Heckerman, D. and Sahami, M. Inductive learning algorithms and representations for text categorization. Proceedings of the Seventh International Conference on Information and Knowledge Management (CIKM'98), 148-155, 1998.
- Dunning T., 1993. Accurate Methods for the Statistics of Surprise and Coincidence, Computational Linguistics, 19(1).

References II

- Feldman R. and Dagan I., 1995. KDT – Knowledge Discovery in Texts. In Proceedings of the First International Conference on Knowledge Discovery, KDD-95.
- Feldman R., and Hirsh H., 1996. Exploiting Background Information in Knowledge Discovery from Text. Journal of Intelligent Information Systems. 1996.
- Feldman R., Aumann Y., Amir A., Klösgen W. and Zilberstien A., 1997. Maximal Association Rules: a New Tool for Mining for Keyword co-occurrences in Document Collections, In Proceedings of the 3rd International Conference on Knowledge Discovery, KDD-97, Newport Beach, CA.
- Feldman R., Rosenfeld B., Stoppi J., Liberzon Y. and Schler, J., 2000. "A Framework for Specifying Explicit Bias for Revision of Approximate Information Extraction Rules". KDD 2000: 189-199.
- Fisher D., Soderland S., McCarthy J., Feng F. and Lehnert W., "Description of the UMass Systems as Used for MUC-6," in Proceedings of the 6th Message Understanding Conference, November, 1995, pp. 127-140.
- Frantzi T.K., 1997. Incorporating Context Information for the Extraction of Terms. In Preceedings of ACL-EACL'97.
- Frawley W. J., Piatetsky-Shapiro G., and Matheus C. J., 1991. Knowledge Discovery in Databases: an Overview. In G. Piatetsky-Shapiro and W. J. Frawley, editors, Knowledge Discovery in Databases, pages 1-27, MIT Press.
- Grishman R., The role of syntax in Information Extraction, In: Advances in Text Processing: Tipster Program Phase II, Morgan Kaufmann, 1996.
- Hahn, U.; and Schnattinger, K. 1997. Deep Knowledge Discovery from Natural Language Texts. In Proceedings of the 3rd International Conference of Knowledge Discovery and Data Mining, 175-178.
- Hayes Ph. 1992. Intelligent High-Volume Processing Using Shallow, Domain-Specific Techniques. Text-BasedIntelligent Systems: Current Research and Practice in Information Extraction and Retrieval. New Jersey, P.227-242.
- Hayes, P.J. and Weinstein, S.P. CONSTRUE: A System for Content-Based Indexing of a Database of News Stories. Second Annual Conference on Innovative Applications of Artificial Intelligence, 1990.
- Hobbs, J. (1986), Resolving Pronoun References, In B. J. Grosz, K. Sparck Jones, & B. L. Webber (Eds.), Readings in Natural Language Processing (pp. 339-352), Los Altos, CA: Morgan Kaufmann Publishers, Inc.
- Hobbs, Jerry R., Douglas E. Appelt, John Bear, David Israel, Megumi Kameyama, and Mabry Tyson, ``FASTUS: A System for Extracting Information from Text'', Proceedings, Human Language Technology, Princeton, New Jersey, March 1992, pp. 133-137.
- Hobbs, Jerry R., Mark Stickel, Douglas Appelt, and Paul Martin, 1993, ``Interpretation as Abduction'', Artificial Intelligence, Vol. 63, Nos. 1-2, pp. 69-142. Also published as SRI International Artificial Intelligence Center Technical Note 499, December 1990.

References III

- Hull D., 1996. Stemming algorithms - a case study for detailed evaluation. Journal of the American Society for Information Science. 47(1):70-84.
- Ingria, R. & Stallard, D., A computational mechanism for pronominal reference, Proceedings, 27th Annual Meeting of the Association for Computational Linguistics, Vancouver, 1989.
- Cowie J., and Lehnert W., "Information Extraction," Communications of the Association of Computing Machinery, vol. 39 (1), pp. 80-91.
- Joachims, T. Text categorization with support vector machines: Learning with many relevant features. Proceedings of European Conference on Machine Learning (ECML'98), 1998
- Justeson J. S. and Katz S. M., 1995. Technical Terminology: Some linguistic properties and an algorithm for identification in text. In Natural Language Engineering, 1(1):9-27.
- Klösgen W., 1992. Problems for Knowledge Discovery in Databases and Their Treatment in the Statistics Interpreter EXPLORA. International Journal for Intelligent Systems, 7(7):649-673.
- Lappin, S. & McCord, M., A syntactic filter on pronominal anaphora for slot grammar, Proceedings, 28th Annual Meeting of the Association for Computational Linguistics, University of Pittsburgh, Association for Computational Linguistics, 1990.
- Larkey, L. S. and Croft, W. B. 1996. Combining classifiers in text categorization. In Proceedings of SIGIR-96, 19th ACMInternational Conference on Research and Development in Information Retrieval (Zurich, CH, 1996), pp. 289–297.
- Lehnert, Wendy, Claire Cardie, David Fisher, Ellen Riloff, and Robert Williams, 1991. ``Description of the CIRCUS System as Used for MUC-3'', Proceedings, Third Message Understanding Conference (MUC-3), San Diego, California, pp. 223-233.
- Lent B., Agrawal R. and Srikant R., 1997. Discovering Trends in Text Databases. In Proceedings of the 3rd International Conference on Knowledge Discovery, KDD-97, Newport Beach, CA.
- Lewis, D. D. 1995a. Evaluating and optmizing autonomous text classification systems. In Proceedings of SIGIR-95, 18th ACMInternational Conference on Research and Development in Information Retrieval (Seattle, US, 1995), pp. 246–254.
- Lewis, D. D. and Hayes, P. J. 1994. Guest editorial for the special issue on text categorization. ACMT ransactions on Information Systems 12, 3, 231.
- Lewis, D. D. and Ringuette, M. 1994. A comparison of two learning algorithms for text categorization. In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval (Las Vegas, US, 1994), pp. 81–93.
- Lewis, D. D., Schapire, R. E., Callan, J. P., and Papka, R. 1996. Training algorithms for linear text classifiers. In Proceedings of SIGIR-96, 19th ACMInternational Conference on Research and Development in Information Retrieval (Z ̈urich, CH, 1996), pp. 298–306.

References IV

- Lin D. 1995. University of Manitoba: Description of the PIE System as Used for MUC-6 . In Proceedings of the Sixth Conference on Message Understanding (MUC-6), Columbia, Maryland.
- Piatetsky-Shapiro G. and Frawley W. (Eds.), 1991. Knowledge Discovery in Databases, AAAI Press, Menlo Park, CA.
- Rajman M. and Besançon R., 1997. Text Mining: Natural Language Techniques and Text Mining Applications. In Proceedings of the seventh IFIP 2.6 Working Conference on Database Semantics (DS-7), Chapam & Hall IFIP Proceedings serie. Leysin, Switzerland, Oct 7-10, 1997.
- Riloff Ellen and Lehnert Wendy, Information Extraction as a Basis for High-Precision Text Classification, ACM Transactions on Information Systems (special issue on text categorization) or also Umass-TE-24, 1994.
- Sebastiani F. Machine learning in automated text categorization, ACM Computing Surveys, Vol. 34, number 1, pages 1-47, 2002.
- Sheila Tejada, Craig A. Knoblock and Steven Minton, Learning Object Identification Rules For Information Integration, Information Systems Vol. 26, No. 8, 2001, pp. 607-633.
- Soderland S., Fisher D., Aseltine J., and Lehnert W., "Issues in Inductive Learning of Domain-Specific Text Extraction Rules," Proceedings of the Workshop on New Approaches to Learning for Natural Language Processing at the Fourteenth International Joint Conference on Artificial Intelligence, 1995.
- Srikant R. and Agrawal R., 1995. Mining generalized association rules. In Proceedings of the 21st VLDB Conference, Zurich, Swizerland.
- Sundheim, Beth, ed., 1993. Proceedings, Fifth Message Understanding Conference (MUC-5), Baltimore, Maryland, August 1993. Distributed by Morgan Kaufmann Publishers, Inc., San Mateo, California.
- Tipster Text Program (Phase I), 1993. Proceedings, Advanced Research Projects Agency, September 1993.
- Vapnik, V., Estimation of Dependencies Based on Data [in Russian], Nauka, Moscow, 1979. (English translation: Springer Verlag, 1982.)
- Vapnik, V., The Nature of Statistical Learning Theory, Springer-Verlag, 1995.
- Yang, Y. and Chute, C. G. 1994. An example-based mapping method for text categorization and retrieval. ACMT ransactions on Information Systems 12, 3, 252–277.
- Yang, Y. and Liu, X. 1999. A re-examination of text categorization methods. In Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval (Berkeley, US, 1999), pp. 42–49.