KSU CIS Department Seminar

Information Extraction: Natural Language, Spatiotemporal Machine Learning, and Link Analysis Approaches

William H. Hsu

http://www.cis.ksu.edu/~bhsu

Laboratory for Knowledge Discovery in Databases (www.kddresearch.org)

Department of Computing and Information Sciences, Kansas State University

Sponsors

K-State National Agricultural Biosecurity Center (NABC)

U.S. Department of Defense, Department of Homeland Security & ONR

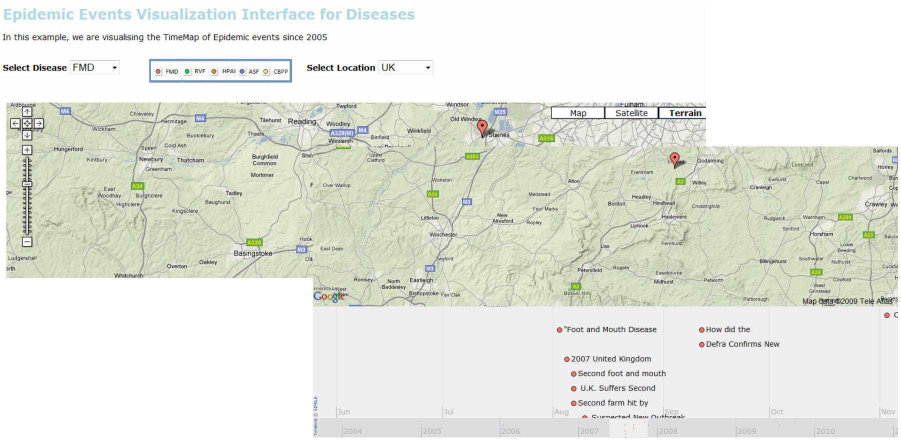
Joint Work With

Doina Caragea; Caterina Scoglio; Barry Erlick, Marty Vanier, KSU

Dan Roth, Chengxiang Zhai, and Jiawei Han, UIUC

Slides for this talk: http://bit.ly/4CQilt

Motivating Example: Summarizing News from the Web



http://fingolfin.user.cis.ksu.edu/timemap2gs

Based on

NLP Group NER Toolkit © 2005-2009 Stanford University

Simile © 2003-2009 Massachusetts Institute of Technology

Google Maps © 2007-2009 Tele Atlas, Inc. and Google, Inc.

Outline

- Three Information Extraction (IE) Tasks
  - Recognizing Textual Entailment (RTE)
  - Update Summarization
  - Question Answering (QA)
- Natural Language Learning/Reasoning Approaches
- Application: Spatiotemporal Event Extraction
- Data Mining: Link Prediction and Analysis
- Some Results from Link Mining, Text Extraction

Information Extraction Tasks: RTE, Summarization, QA

- Recognizing Textual Entailment (RTE)
  - Determine when meaning of text logically follows from that of another
  - Approaches: text categorization, semantic mapping, inference
  - Related to question answering: "true/false" questions
- Update Summarization
  - Produce brief synopsis of points in text where user has read others
  - Approaches: formal summarization, natural language (NL) synthesis
  - Related to question answering: collect relevant documents, digest
- Question Answering (QA)
  - Respond to query posed in natural language
  - Approaches: search, focused crawling, semantic mapping

OUTLINE

- Three Information Extraction (IE) Tasks
  - Recognizing Textual Entailment (RTE)
  - Update Summarization
  - Question Answering (QA)
- Natural Language Learning/Reasoning Approaches
- Application: Spatiotemporal Event Extraction
- Data Mining: Link Prediction and Analysis
- Some Results from Link Mining, Text Extraction

RECOGNIZING TEXTUAL ENTAILMENT [1]: EXAMPLES

SOURCE: A bus collision with a truck in Uganda has resulted in at least 30 fatalities and has left a further 21 injured.

TARGET: 30 die in a bus collision in Uganda. ✓ $S \models T$

SOURCE: Mrs. Bush's approval ratings have remained very high, above 80%, even as her husband's have recently dropped below 50%.

TARGET: 80% approve of Mr. Bush. ✘ $S \not\models T$

SOURCE: Take consumer products giant Procter and Gamble. Even with a $1.8 billion Research and Development budget, it still manages 500 active partnerships each year, many of them with small companies.

TARGET: 500 small companies are partners of Procter and Gamble. ✘ $S \not\models T$

Recognizing Textual Entailment [2]: Problem Definition

- **Given: Natural Language Input**
  - SOURCE sentence(s): usually complex text
  - TARGET sentence: usually simplified "gist" summary, proposition
- **Return**
  - True iff SOURCE logically entails TARGET (S ⊨ T)
  - Optional: Interpretation of SOURCE/TARGET
  - Optional: Chain of inferences
- **Possible Side Effects: Parsed Output**
  - Shallow parsing *aka* chunking: e.g., Named Entity Recognition (NER)
  - Full parsing: noun/verb phrases, Semantic Role Labeling (SRL)

Recognizing Textual Entailment [3]: Approaches

- **Algorithms**
  - Shallow parsing *aka* chunking: e.g., Named Entity Recognition (NER)
    - NER: people, places, organizations, quantities/dates, events
    - Part-of-speech (POS) tagging: e.g., verbs
  - Semantic Role Labeling: more in second problem (summarization)
- **Knowledge Representation**
  - Propositions
  - Limited first-order predicate calculus (shallow quantification)
- **Other Semantic Tasks**
  - Extracting terminology, relationships
  - Coreference resolution ("coref")

Recognizing Textual Entailment [4]: Challenges and Open Problems

- NER: Beyond Gazetteers (Dictionary) Approaches
- Coreference Resolution ("Coref")
  - Needed in multi-sentence tasks (RTE, QA, summarization)
  - Applications: anaphora (including pronoun resolution)
  - Inferential task
- Terminology Extraction: Finding New Named Entities, Verbs
- Relationship Extraction
  - Identity/equality: "exactly" / "only" (=)
  - Inequalities: "at least" (≥), "as many as" / "up to" (≤)
  - Relationships with sets: membership (∈), containment (⊆)
  - Terms of negation: "not", "never", "hardly", *etc.*

Recognizing Textual Entailment [5]: Applications

- Example: CNN, 2007 Foot-and-Mouth Disease (http://bit.ly/3gof6o)

  Tests have confirmed a second foot-and-mouth outbreak in southern England, the government announced, raising fears that the highly contagious animal virus is spreading.

  Chief Veterinary Officer Debby Reynolds said Tuesday that tests showed a herd of cattle had been infected.

  The animals were culled Monday evening after showing signs of the disease.

  Britain's Department for Environment, Food and Rural Affairs said Monday a herd of more than 50 cattle on a second farm within the two-mile (three-kilometer) protection zone in Surrey County, England, had shown signs of the highly contagious disease.

- Open Problems
  - Basic scientific, medical terminology: tests … confirmed
  - Anaphor resolution: the disease → [FMD]
  - Aggregates: herd of more than 50 cattle

Outline

- Three Information Extraction (IE) Tasks
  - Recognizing Textual Entailment (RTE)
  - Update Summarization
  - Question Answering (QA)
- Natural Language Learning/Reasoning Approaches
- Application: Spatiotemporal Event Extraction
- Data Mining: Link Prediction and Analysis
- Some Results from Link Mining, Text Extraction

Update Summarization [1]: Examples

SOURCE: A bus collision with a truck in Uganda has resulted in at least 30 fatalities and has left a further 21 injured.

TARGET: 30 people died and 21 people were injured in a bus collision in Uganda.

SOURCE: Mrs. Bush's approval ratings have remained above 80%, even as her husband's have recently dropped below 50%.

TARGET: President Bush's approval ratings have decreased to less than 50%.

SOURCE: Take consumer products giant Procter and Gamble. Even with a $1.8 billion R&D budget, it still manages 500 active partnerships each year, many of them with small companies.

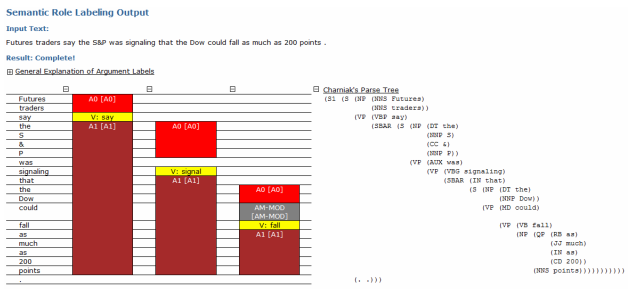
TARGET: Procter and Gamble has 500 partnerships per year.

UPDATE SUMMARIZATION [2]: PROBLEM DEFINITION

- **Given: Natural Language (NL) Input**
  - **SOURCE sentence(s): usually complex text**
  - **Previously digested text summaries (~ what user has previously read)**
- **Return**
  - **TARGET sentence: simple "gist" summary synthesized from SOURCE**
  - **Optional: Machine-readable interpretation of SOURCE**
  - **Optional: Rewriting, other transformations**
- **Possible Side Effects: Parsed Output**
  - **Chunking as for textual entailment**
  - **Semantic Role Labeling: may be needed more (for text generation)**

UPDATE SUMMARIZATION [3]: APPROACHES

- **Algorithms**
  - **SRL**



Semantic Role Labeling Demo
http://l2r.cs.uiuc.edu/~cogcomp/srl-demo.php
© 2009 University of Illinois

  - **Text generation**
- **Knowledge Representation: Parse Trees, Abstract Data Types**
- **Other Tasks**
  - **Filling in abstract data types (ADTs) *aka* frames, slot-filler structures**
  - **Natural language generation, content evaluation**

Update Summarization [4]: Challenges and Open Problems

- **Information Extraction (IE) Shared Tasks**
  - NER: as in RTE, needed to identify actors, label roles
  - Coreference resolution: needed to extract ADT representation
  - Terminology extraction: as in RTE, needed to expand set of entities
  - Relationship extraction: foundation of relational summarization
- **Relational Data Modeling and Summarization**
  - Summaries as tuples
  - "Who, what, when, where, why, how"
  - Example: disease, species, locale, quantity, date/time, expert, agency
  - Attributes may have *missing values*
- **Machine Learning and Inference: Imputation of Values**

Update Summarization [5]: Applications

- **Example: CNN, 2007 Foot-and-Mouth Disease (http://bit.ly/3gof6o)**

  Tests have confirmed a second foot-and-mouth outbreak in southern England, the government announced, raising fears that the highly contagious animal virus is spreading.

  Chief Veterinary Officer Debby Reynolds said Tuesday that tests showed a herd of cattle had been infected.

  The animals were culled Monday evening after showing signs of the disease.
- **Update Summarization**

  A second foot-and-mouth disease infection in a herd of cattle in southern England was responded to by culling on Monday evening and announced by Debby Reynolds on Tuesday.

  (Second since earlier report – hence "update".)
- **Compare: Recognizing Textual Entailment**

  A foot-and-mouth disease infection was reported the day after culling. (True.)

Outline

- Three Information Extraction (IE) Tasks
  - Recognizing Textual Entailment (RTE)
  - Update Summarization
  - Question Answering (QA)
- Natural Language Learning/Reasoning Approaches
- Application: Spatiotemporal Event Extraction
- Data Mining: Link Prediction and Analysis
- Some Results from Link Mining, Text Extraction

Question Answering [1]: Examples

SOURCE: A bus collision with a truck in Uganda has resulted in at least 30 fatalities and has left a further 21 injured.

QUERY [TARGET]: How many injuries [21] and how many fatalities [30] were reported in bus accidents in Uganda?

SOURCE: Mrs. Bush's approval ratings have remained above 80%, even as her husband's have recently dropped below 50%.

QUERY [TARGET]: What is President Bush's latest approval rating? [Less than 50%]

SOURCE: Take consumer products giant Procter and Gamble. Even with a $1.8 billion R&D budget, it still manages 500 active partnerships each year, many of them with small companies.

QUERY [TARGET]: How many active partnerships per year does Procter and Gamble have? [500]

Question Answering [2]: Problem Definition

- **Given: Natural Language Input**
  - SOURCE sentence: usually complex text
  - QUERY sentence
- **Return**
  - TARGET sentence: answers
    - from database query, OR
    - synthesized from data retrieved in response to query
  - Optional: other information retrieval (IR) functions
    - Data cubes (On-Line Analytical Processing): drill down, roll up
    - Visualization: thematic maps, hierarchies
    - Statistics and evidence in support of answer

Question Answering [3]: Approaches

- **Algorithms**
  - Simple ranking
    - Google *PageRank* / Kleinberg's HITS: hubs-authority score
    - Term frequency, inverse document frequency (TFIDF)
  - Entity search
  - Learning to rank
  - Query formation and semantics-preserving transformations
- **Knowledge Representation**
  - Queries and texts as documents
  - Propositional queries
- **Document Collections and Text Categorization**

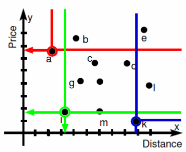
Question Answering [4]: Challenges and Open Problems

- Information Extraction (IE) Shared Tasks
  - NER: as in RTE and summarization, needed to produce NE phrases
  - Coreference resolution: needed to relate query to text
  - Terminology extraction: needed for synonymy, hypo/hypernymy
  - Relationship extraction: basis of query formation in relational model
- Relational Data Modeling and QA
  - Each relationship contains tuples
  - Queries on relational databases
  - Compare SQL SELECT … FROM … WHERE
  - Predicates: disease, species, locale, quantity, date, expert, agency
  - Translation of query needed
- Open Problem: Approximate/Tolerant (Skyline) Queries

Question Answering [5]: Applications

- Spatial Queries

===> What cities are within 250 miles of the capital of Italy?
I know that Italy's capital is Rome, Italy (source: START KB).
Using this information, I determined what cities are within 250 miles of Rome, Italy:
In Italy, the following cities are within 250 miles of Rome:
Naples, Italy is 118 miles (189.90298 kilometers) from Rome.
[Florence, Pisa, Bologna, Venice, Trieste, Verona]
Genoa, Italy is 249 miles (400.72745 kilometers) from Rome.
Source: START KB [http://start.csail.mit.edu]



http://bit.ly/UBq4q

University of Maryland

- Skyline Queries
  - Used in constrained decision support
  - Given: points p_1, p_2, …, p_N, each in d dimensions
  - Return: maximal (non-dominated) points – *i.e.*, Pareto front
  - QA: interpretation of NL queries (including skyline)

Outline

- Three Information Extraction (IE) Tasks
  - Recognizing Textual Entailment (RTE)
  - Update Summarization
  - Question Answering (QA)
- Natural Language Learning/Reasoning Approaches
- Application: Spatiotemporal Event Extraction
- Data Mining: Link Prediction and Analysis
- Some Results from Link Mining, Text Extraction

Machine Learning

- Notation and Definitions
  - Instance: $x = (x_1, x_2, \ldots, x_n)$, sometimes x_j, $1 \leqslant j \leqslant m$ with x_{ji}, $1 \leqslant i \leqslant n$
  - Instance space X such that $x \in X$
  - Data set: $D = \{x_1, x_2, \ldots, x_m\}$ where $x_j = (x_{j1}, x_{j2}, \ldots, x_{jn})$
- Clustering
  - Mapping from old $x = (x_1, x_2, \ldots, x_n)$ to new $x' = (x_1', x_2', \ldots, x_k')$, $k << n$
  - Attributes x_i' of new instance not necessarily named
  - Idea: project instance space X into lower dimension X'
  - Goal: keep groups of similar X together in X'
- Regression
  - Idea: given independent variable x, dependent variables $y = f(x)$, fit f
  - Goal: given new (previously unseen) x, approximate $f(x)$
- Classification
  - Similar to regression, except that f is boolean- or nominal-valued
  - "Curve fitting" figurative: approximator may be logical formula

Probabilistic and Frequentist Models

© 2008 C. Zhai
University of Illinois
http://sifaka.cs.uiuc.edu/ir/

Comparing News Articles

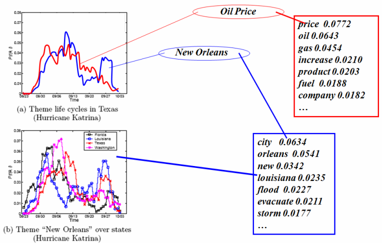
Iraq War (30 articles) vs. Afghan War (26 articles)

The common theme indicates that "United Nations" is involved in both wars

	Cluster 1	Cluster 2	Cluster 3
Common Theme	united 0.042 nations 0.04 …	killed 0.035 month 0.032 deaths 0.023 …	…
Iraq Theme	n 0.03 Weapons 0.024 Inspections 0.023 …	troops 0.016 hoon 0.015 sanches 0.012 …	…
Afghan Theme	Northern 0.04 alliance 0.04 kabul 0.03 taleban 0.025 aid 0.02 …	taleban 0.026 rumsfeld 0.02 hotel 0.012 front 0.011 …	…

Collection-specific themes indicate different roles of "United Nations" in the two wars

Theme Life Cycles ("Hurricane Katrina")



(a) Theme life cycles in Texas (Hurricane Katrina)

(b) Theme "New Orleans" over states (Hurricane Katrina)

The Database and Information Systems Laboratory at The University of Illinois at Urbana-Champaign
Large Scale Information Management

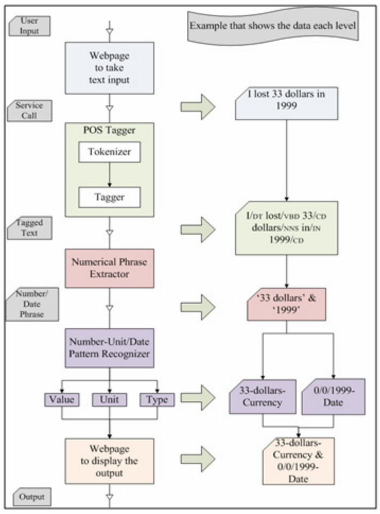
Current System: Tasks and Research Priorities

- Web document content extraction
  - Named entity recognition (NER)
  - Coreference, association
  - Relation extraction (*aka* link discovery)
- Geotagging: location extraction, map view
- Temporal tagging: date/time extraction, timeline view
- Semi-supervised document clustering
- Data integration: portal application
- Visual and text analytics
- Predictive epidemiological modeling interface

Information Extraction Pipeline

Project Data Flow Diagram:

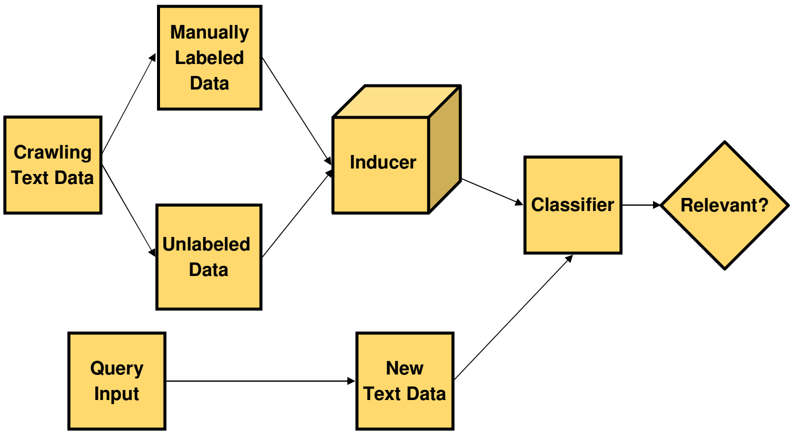
Numerical Entity Searcher



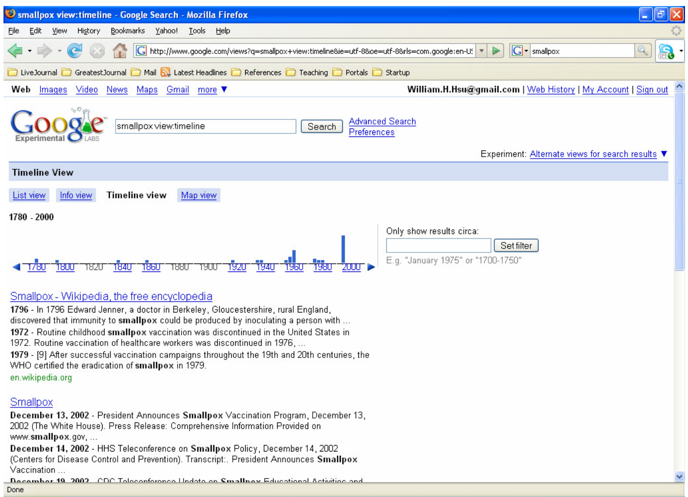
Outline

- Three Information Extraction (IE) Tasks
  - Recognizing Textual Entailment (RTE)
  - Update Summarization
  - Question Answering (QA)
- Natural Language Learning/Reasoning Approaches
- Application: Spatiotemporal Event Extraction
- Data Mining: Link Prediction and Analysis
- Some Results from Link Mining, Text Extraction

Semisupervised Annotation: Mostly Unlabeled Data



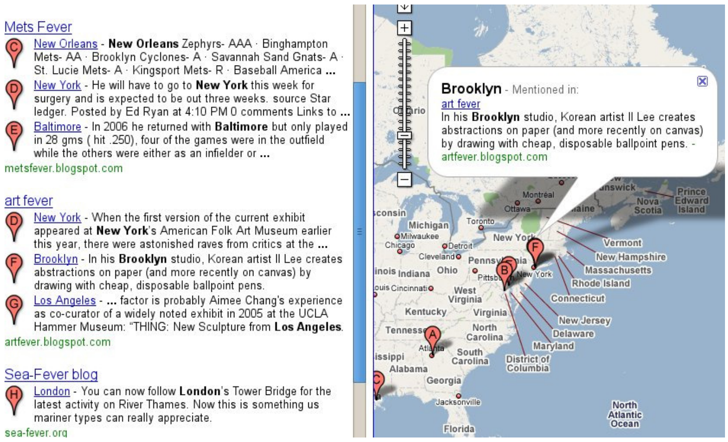
Automatic Timeline Generation Task



Search phrase: "smallpox"

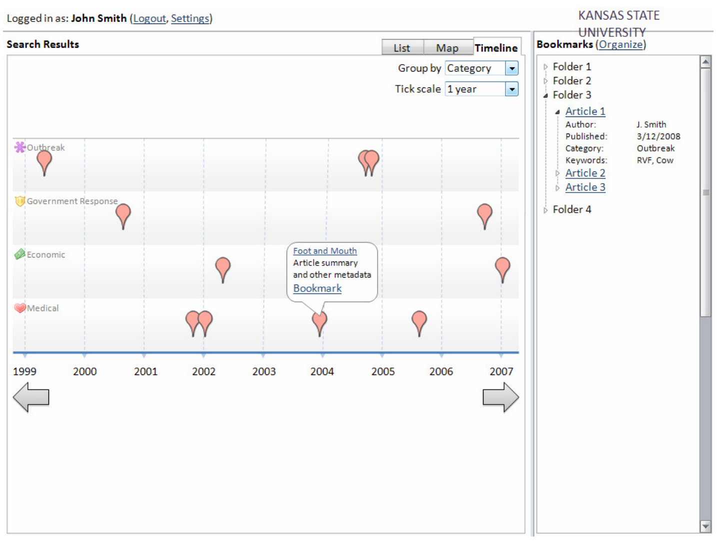
Spatial Annotation Task: Disambiguation and Classification

Current off-the-shelf applications fall into ambiguity problems



© 2008 W. Elshamy

Joint Work With Elder Research



© 2008 J.R. Lawhorne

ELDER RESEARCH INC. DATA MINING & PREDICTIVE ANALYTICS

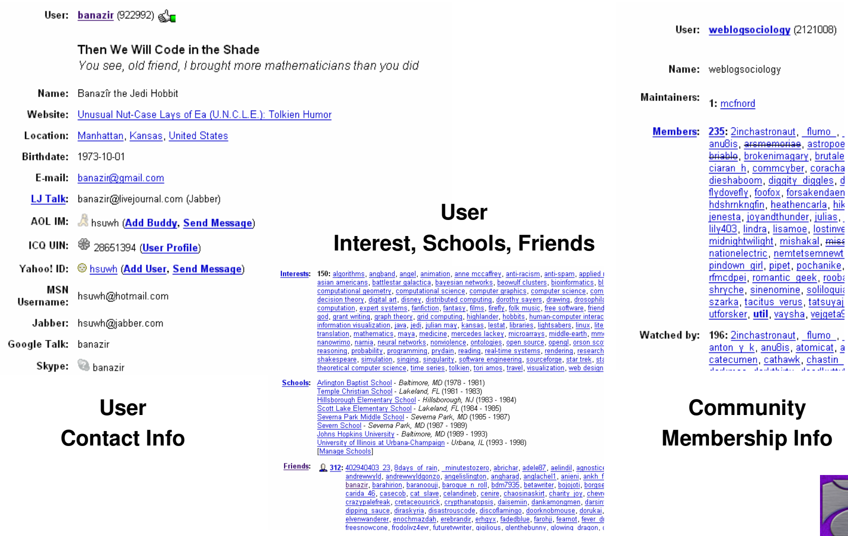
Outline

- Three Information Extraction (IE) Tasks
  - Recognizing Textual Entailment (RTE)
  - Update Summarization
  - Question Answering (QA)
- Natural Language Learning/Reasoning Approaches
- Application: Spatiotemporal Event Extraction
- Data Mining: Link Prediction and Analysis
- Some Results from Link Mining, Text Extraction

Link Mining in Social Networks

- Problem Definition
  - Given: records of users of weblog or social network service
  - Discover
    - ⇨ Features of entities: users, communities
    - ⇨ Relationships: friendship, membership, moderatorship
    - ⇨ Explanations and predictions for relationships
- Goals
  - Boost precision and recall of link existence prediction
  - Find relevant features
- Significance: Recommendations (Friendship, Membership)
- Data Set: Crawled from *LiveJournal* Blog Service

K-State Test Bed: *LJMiner* Corpus



LiveJournal Topology: Definitions

Start	End	Link Denotes
User	User	Trust or friendship
User	Community	Readership or subscribership
Community	User	Membership, posting access, maintainer
Community	Community	Obsolete

Types of links in the blog service LiveJournal.

Mutual Friends: $\{ v \mid (v, u) \in E \wedge (u, v) \in E \}$

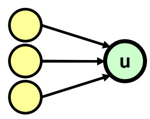
Also Friend Of: $\{ v \mid (v, u) \in E \wedge (u, v) \notin E \}$

LJCrawler and *LJClipper*

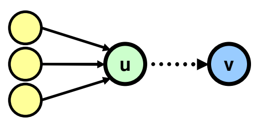
- Three Parts
  - Client, Injector, Parser
  - Ancillary: Multi-threading, distribution, storage
  - *LJClipper*, *LJStats*
- What Makes *LJCrawler* Different?
  - Distributed implementation of focused crawler
  - Offline data synthesis: LJClipper
- Runtime Efficiency
  - 200 users/sec maximum, 5 users/sec allowed
  - ~2.3 million pages crawled

Graph Features [1]: Node, Pair, Link-Dependent

Node-dependent feature:
Indegree of *u*



Pair-dependent feature:
Common interests of *u* and *v*
Alternate distance from *u* to *v*
(degrees of separation)



Link-dependent feature:
Duration of friendship between *u* and *v*
"How does *u* know *v*?"

Laboratory for Knowledge Discovery in Databases
Departmental Seminar Wednesday, 30 Sep 2009
Computing & Information Sciences Kansas State University

Graph Features [2]: Node and Pair Features in *LJMiner*

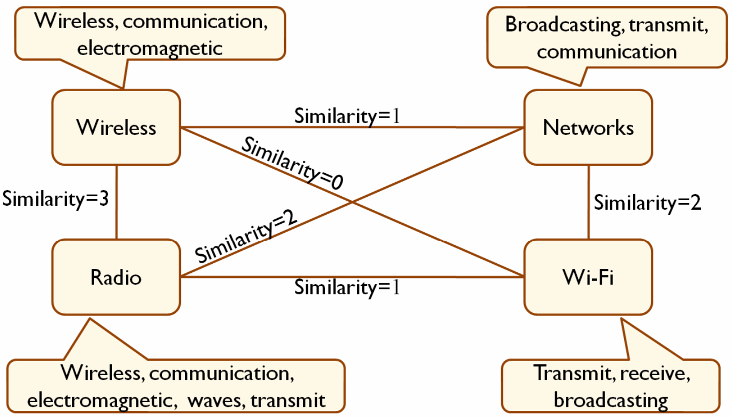
Graph Features

1. Indegree of u: popularity of the user
2. Indegree of v: popularity of the candidate
3. Outdegree of u: number of other friends besides the candidate; saturation of friends list
4. Outdegree of v: number of existing friends of the candidate besides the user; correlates loosely with likelihood of a reciprocal link
5. Number of mutual friends w such that $u \rightarrow w \wedge w \rightarrow v$
6. "Forward deleted distance": minimum alternative distance from u to v in the graph without the edge (u, v)
7. Backward distance from v to u in the graph

Interest-Related Features

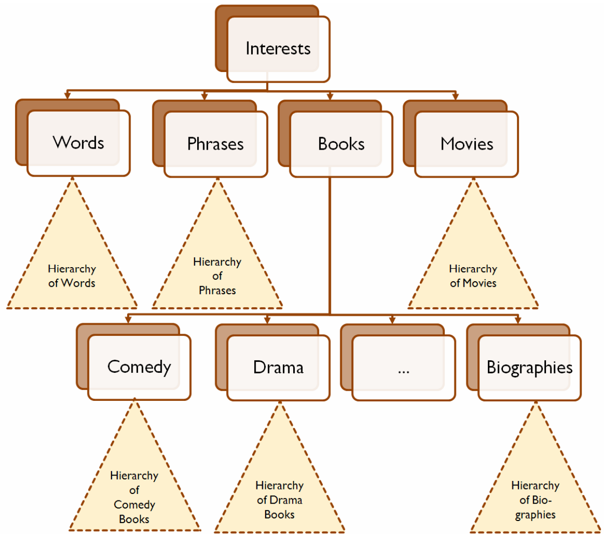
8. Number of mutual interests between u and v
9. Number of interests listed by u
10. Number of interests listed by v
11. Ratio of the number of mutual interests to the number listed by u
12. Ratio of the number of mutual interests to the number listed by v

Similarity Measures for Ontology Extraction

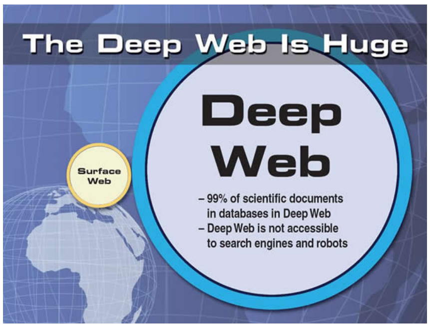


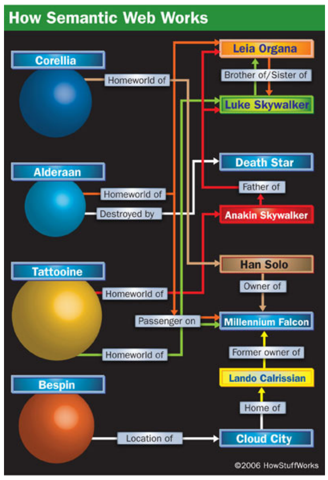
Similarity Metric

ONTOLOGY EXTRACTION BY HIERARCHICAL AGGLOMERATIVE CLUSTERING



DEEP WEB & SEMANTIC WEB





Warnick, W. L. (2006). Global Discovery: Increasing the Pace of Knowledge Diffusion to Increase the Pace of Science.
http://www.osti.gov/speeches/fy2006/aaas/

Wikipedia: "aka *Deepnet, invisible Web, hidden Web* ... refers to World Wide Web content not part of surface Web indexed by search engines"

Wilson, T. V. (2006). How Semantic Web Works.
http://computer.howstuffworks.com/semantic-web.htm

Outline

- Three Information Extraction (IE) Tasks
  - Recognizing Textual Entailment (RTE)
  - Update Summarization
  - Question Answering (QA)
- Natural Language Learning/Reasoning Approaches
- Application: Spatiotemporal Event Extraction
- Data Mining: Link Prediction and Analysis
- Some Results from Link Mining, Text Extraction

Network Statistics: Graph Distance

Distance *d*	Frequency (= *d*)	Cumulative (≤ *d*)
1	6204	6204
2	107307	113511
3	69896	183407
4	59926	243333
5	3400	246733
6	255	246988
7	16	247004
8	1	247005
9	0	0
10	0	0
∞	9731	256735

1000 nodes

Distance *d*	Frequency (= *d*)	Cumulative (≤ *d*)
1	19410	19410
2	370568	389978
3	403075	793053
4	520373	1313426
5	123747	1437173
6	18453	1455626
7	2657	1458283
8	339	1458622
9	29	1458651
10	0	1458651
∞	174534	1633185

4000 nodes

Hsu, W. H., King, A. L., Paradesi, M., Pydimarri, T., & Weninger, T. (2006).
AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs .
http://bit.ly/LQmqR

Link Prediction as Classification: Early Results

- 941-node graph (Hsu *et al.*, 2006): ***LJCrawler*** **v1 output**
- 1000-4000 node graphs: ***LJCrawler*** **v2 output**

Inducer	All	NoDist	BkDist	Dist	Interest
J48	*98.2*	94.8	95.8	97.6	88.5
OneR	95.8	92.0	95.8	95.8	88.5
Logistic	91.6	90.9	88.3	88.9	88.4

Percent accuracy for predicting all classes using the 941-node graph.

Inducer	All	NoDist	BkDist	Dist	Interest
J48	*89.5*	65.7	67.7	83.0	5.4
OneR	67.7	41.1	67.7	67.7	4.5
Logistic	38.3	33.3	0	4.5	4.5

Precision (true positives to all positives) using the 941-node graph.

Inducer	Accuracy	Precision	Recall
J48	99.9	97.5	96.1
OneR	99.6	91.7	91.8

Percent accuracy, precision and recall using a 1000-node graph (10-fold CV).

Inducer	Accuracy	Precision	Recall
J48	99.8	95.8	92.0
OneR	99.7	91.1	89.9

Percent accuracy, precision and recall using a 2000-node graph (10-fold CV).

Inducer	Accuracy	Precision	Recall
J48	99.8	94.5	88.3
OneR	99.7	88.2	84.3

Percent accuracy, precision and recall using a 4000-node graph (10-fold CV).

Hsu *et al.* (2006)   http://bit.ly/LQmqR

Hsu, W. H., Lancaster, J. P., Paradesi, M. S. R., & Weninger, T. (2007).
First International Conference on Weblogs and Social Media (ICWSM).
http://bit.ly/34NwTE

Ontology Extraction: Most Recent Results [1]

(Predicting Friendships)

- Graph-based and interest-based numerical features

Exp#	*Ontology*	*SVM*	*Logistic*	*J48*	*Random Forest*	*OneR*
4	(graph only)	0.92 +/- 0.03	0.91 +/- 0.04	0.94 +/- 0.03	0.97 +/- 0.03	0.86 +/- 0.09
11		0.92 +/- 0.03	0.91 +/- 0.04	0.94 +/- 0.02	0.98 +/- 0.01	0.86 +/- 0.09
12(a)	O1	0.94 +/- 0.04	0.94 +/- 0.02	0.93 +/- 0.05	0.97 +/- 0.02	0.88 +/- 0.04
12(b)	O2	**0.95 +/- 0.03**	0.94 +/- 0.03	0.94 +/- 0.03	0.98 +/- 0.01	**0.91 +/- 0.04**
13(a)	Sub-O1	0.90 +/- 0.04	0.91 +/- 0.04	0.94 +/- 0.03	0.97 +/- 0.03	0.86 +/- 0.06
13(b)	Sub-O2	0.93 +/- 0.04	0.92 +/- 0.04	0.93 +/- 0.05	0.98 +/- 0.01	0.91 +/- 0.08

Table reports AUC values
BLUE-BOLD highlights significant improvements compared to the baseline
RED highlights improvements compared to the baselines that are not significant

Bahirwani, V., Caragea, D., Aljandal, W. & Hsu, W. H. (2008).
Second ACM SIGKDD Workshop on Social Network Mining and Analysis (SNA-KDD).
http://bit.ly/32UnGs

Ontology Extraction: More Recent Results [2]

(Predicting Friendships)

Features Used				*SVM*	*Logistic*	*J48*	*Random Forest*	*OneR*
Nom. Interest based	*Num. Interest based*	*Graph based*	*Ontology 2*					
✓								
✓			✓					
	✓			**0.66**	0.64	0.59	0.61	0.58
	✓		✓	**0.76**	0.73	0.69	0.73	0.64
		✓		0.92	0.91	0.94	**0.97**	0.86
✓		✓						
✓		✓	✓					
	✓	✓		0.92	0.91	0.94	**0.98**	0.86
	✓	✓	✓	0.95	0.94	0.94	**0.98**	0.91

Bahirwani *et al.* (2008).
http://bit.ly/32UnGs

Ontology Extraction: More Recent Results [3]

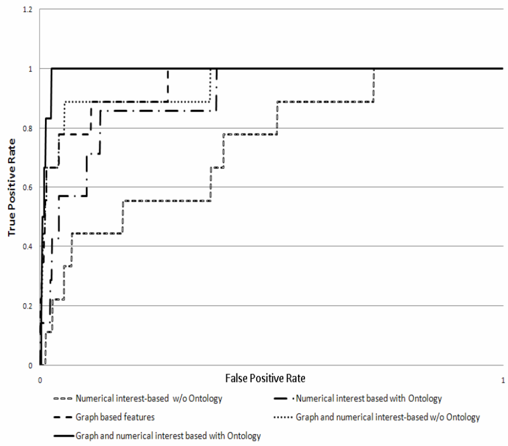


Figure 3: ROC curves for SVM when using different sets of attributes to predict friends

Bahirwani *et al.* (2008).
http://bit.ly/32UnGs

Ontology Extraction: Most Recent Results

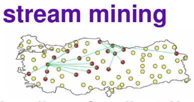
Area Under the ROC Curve (ROC-AUC)
Support Vector Machines, Logistic Regression, Random Forest, Decision Trees
Average of 5 Replications

	Features	SVM	LR	RF	J48
10% links known	Graph only	0.69±0.01	0.67±0.01	0.70±0.04	0.61±0.08
	Graph, without O	0.68±0.01	0.68±0.01	0.69±0.05	0.57±0.09
	Graph, O (best level)	**0.70±0.00**	**0.69±0.01**	**0.74±0.04**	**0.64±0.06**
		(42,35,37,42,34)	(42,28,17,21,17)	(9,13,38,26,27)	(2,3,5,22,6)
25% links known	Graph only	0.71±0.01	0.67±0.01	0.72±0.02	0.67±0.05
	Graph, without O	0.74±0.01	0.72±0.01	0.71±0.03	0.65±0.04
	Graph, O (best level)	**0.76±0.01**	**0.74±0.01**	**0.79±0.02**	**0.71±0.05**
		(42,36,42,41,23)	(42,40,42,29,32)	(42,36,19,31,27)	(6,22,2,5,6)
50% links known	Graph Only	0.82±0.01	0.79±0.01	0.80±0.01	0.77±0.03
	Graph, without O	0.85±0.01	0.83±0.01	0.82±0.02	0.76±0.02
	Graph, O (best level)	**0.86±0.01**	**0.85±0.01**	**0.86±0.02**	**0.78±0.02**
		(42,42,42,27,23)	(42,23,21,29,42)	(42,36,26,18,27)	(6,28,2,26,27)

Caragea, D., Bahirwani, V., Aljandal, W., & Hsu, W. H. (2009).
Eighth Symposium on Abstraction, Reformulation and Approximation (SARA).
http://bit.ly/32UnGs

Continuing Work

- Natural Language Learning and Information Extraction
  - Multi-lingual NER
  - Extracting domain lexicons and ontologies using coreference
  - Maximum entropy methods for event extraction
  - From topic detection to update tracking: stream mining
  - Spatial disambiguation
  - Skyline QA



Roy Chowdhury, Scoglio, & Hsu (2009)
Epidemics 2, to appear.

- Link Mining
  - Ontology-aware link annotation: towards causal explanations
  - Spatiotemporal fluents
- Predictive Epidemiology
  - Parameter estimation
  - Graphical models of probability: continuous-time Bayes nets
- Other Topics
  - Information trust: using constrained conditional models
  - Vertical portals (e.g., http://dblife.cs.wisc.edu)

References

- Natural Language Learning and Information Extraction
  - RTE: PASCAL http://bit.ly/2VZn62
  - Update Summarization: NIST TAC 2009 http://bit.ly/sx9ws
  - Question Answering: NIST TAC 2008 http://bit.ly/lkRFH
  - IR: Manning *et al.* (2008), Zhai (2009)
- Link Mining
  - Barabási & Crandall (2003)
  - Han & Kamber (2006), Chapter 9
- Predictive Epidemiology
  - Sørensen *et al.* (1999)
  - Barthelemy *et al.* (2004), Colizza *et al.* (2007)
- Machine Learning and Data Mining
  - Han & Kamber, 2e (2006)
  - Witten & Frank, 2e (2005)
  - Mitchell (1997)
  - See also: KDD Group Bibliography (work in progress)

Acknowledgements

- Collaborators
  - Doina Caragea – Department of Computing and Information Sciences
  - Caterina Scoglio – Department of Electrical and Computer Engineering
- K-State Lab for Knowledge Discovery in Databases
  - Grad RAs: Elshamy, Kallumadi, Roy Chowdhury, Volkova
  - Other grad students: Greene, Hart; Bujuru, Karanam, Mohammed, Peddola, Singireddy, Tirumalae
  - Alumni: Aljandal, Bahirwani, Paradesi, Weninger, Xia
  - Undergrads: Berggren, Drouhard, Fowles, Henke, Jones
- Predictive Epidemiology, Social Networks, Graph Algorithms
  - Marty Vanier, Barry J. Erlick (Nat'l. Ag. Biosecurity Center)
  - Z. Buckner, J. Dimeo, J. R. Lawhorne (Elder Research, Inc.)
- Machine Learning
  - Dan Roth, Jiawei Han, Kevin Chang, ChengXiang Zhai (University of Illinois at Urbana-Champaign)
  - AnHai Doan (University of Wisconsin – Madison)