Text Mining and the Semantic Web

Tim Finin
University of Maryland
Baltimore County

Next Generation Data Mining Workshop
Baltimore, November 2002

http://umbc.edu/~finin-talks/ngdm02/
(1) Opening thoughts
(2) Some current research at UMBC
(4) Parting thoughts
“XML is Lisp's bastard nephew, with uglier syntax and no semantics. Yet XML is poised to enable the creation of a Web of data that dwarfs anything since the Library at Alexandria.”

“The web has made people smarter. We need to understand how to use it to make machines smarter, too.”

-- Michael I. Jordan, paraphrased from a talk at AAAI, July 2002 by Michael Jordan (UC Berkeley)
“The Semantic Web will globalize KR, just as the WWW globalize hypertext”

-- Tim Berners-Lee
“The multi-agent systems paradigm and the web both emerged around 1990. One has succeeded beyond imagination and the other has not yet made it out of the lab.”

-- Anonymous, 2001
IMHO

- The web is like a universal acid, eating through and consuming everything it touches.
  - Web principles and technologies are equally good for wireless/pervasive computing.
- The semantic web is our first serious attempt to provide semantics for XML sublanguages.
- The semantic web will provide mechanisms for people and machines (agents, programs, CGI scripts) to come together.
  - Solving the symbol grounding problem?
Some UMBC applications

(1) Semantic web and agents (ITTalks)
(2) Ontology mapping
(3) Cooperating personal agents
(4) Learning markup
(5) Information retrieval on the SW
(6) Modeling trust policies
(7) Semantic web & pervasive computing

Joint work with Anupam Joshi, Yun Peng, Scott Cost, Tim Oates, Jim Mayfield, Benjamin Grosof, Yelena Yesha, and many students.
(1) ITTALKS

- **ITTALKS** is a database driven web site of IT related talks at UMBC and other institutions. The database contains information on
  - Seminar events
  - People (speakers, hosts, users, ...)
  - Places (rooms, institutions, ...)
- Web pages with DAML markup are generated
- The DAML markup supports agent-based services relating to these talks.
- Users get talk announcements based on the interests, locations and schedules.

[http://ittalks.org/](http://ittalks.org/)
Electric Elves: Towards an Agent Facilitated Human Organization

Milind Tambe
University of Southern California
Information Science Institute

UMBC, ECS, LHS
2:00pm - 12:00pm, Tuesday, February 20, 2001

Abstract
Past few years have seen a revolution in the field of software agents, with agents now proliferating in human organizations, helping individuals in tasks such as information gathering, activity scheduling, managing email, etc. The "Electric-Elves" effort at USC/ISI is now taking the next step: dynamic teaming of all such different heterogeneous agents, as well as proxy agents for humans, to serve not just individuals, but to facilitate the functioning of entire organizations. The ultimate goal of our work is to build agent teams that assist in all organization activities, enabling organizations to act coherently, to robustly attain their mission goals and to react swiftly to crises. The results of this work could potentially be relevant to all organizations, including the military, corporations, and universities and research institutions. As a step towards this goal, we have had an agent team of about 15-20 agents, including 10 proxies (for 10 people) running 24/7 for the past four months at USC/ISI. The proxies communicate with us using different types of mobile wireless devices, and attempt to track our locations using wireless GPS transmissions. These agents assist us in several tasks: they track people's locations, reschedule meetings, decide presenters for research meetings (by auctioning research talk slots), and even order our lunch and dinner. In this talk, I will outline some of the lessons we have learned over the past several months in running this agent system. I will also outline our approach on some of the key research challenges, including agents' adjustable autonomy.

Biosketch
Dr. Milind Tambe is a project leader at the University of Southern California Information Sciences
Electric Elves: Towards an Agent Facilitated Human Organization

Past few years have seen a revolution in the field of software agents, with agents now proliferating in human organizations, helping individuals in tasks such as information gathering, activity scheduling, managing email, etc. The "Electric-Elves" effort at USC/ISI is now taking the next step: dynamic teaming of all such different heterogeneous agents, as well as proxy agents for humans, to serve not just individuals, but to facilitate the functioning of entire organizations. The ultimate goal of our work is to build agent teams that assist in all organization activities, enabling organizations to act coherently, to robustly attain their mission goals and to react swiftly to crises. The results of this work could potentially be relevant to all organizations, including the military, corporations, and universities and research institutions. As a step towards this goal, we have had an agent team of about 15-20 agents, including 10 proxies (for 10 people) running 24/7 for the past four months at USC/ISI. The proxies communicate with us using different types of mobile wireless devices, and attempt to track our locations using wireless GPS transmissions. These agents assist us in several tasks: they track people's locations, reschedule meetings, decide presenters for research meetings (by auctioning research talk slots), and even order our lunch and dinner. In this talk, I will outline some of the
(2) Ontology mapping

- Techniques for mapping between ontologies are thought to be important for the semantic web.
- Ontology mapping is a complex problem
- Even in our simple ITTALKS system we’ve found a need to map between limited ontologies.
What are talks about?

- Topic hierarchies provide indexing terms
  - ACM CCS topic hierarchy
  - Open Directory
- Encoded as DAML ontologies
- These allow users to specify interests as well as browse the database of talks by topic
- Automatic classification of talks (based on title and abstract) and users (based on home page, CV, papers, etc.)
Ontomapper: Interactive topic ontology mapper

- Users create maps between ontologies with URIs to text describing classes & properties.
- Partially automates mapping, using hierarchical relations & user-specified

Used in XTalks to enable mappings between Alternative topic ontologies in DAML+OIL
Where will Semantic web markup come from?
- Created dynamically as web pages are generated from databases (e.g., ITTalks).
- Annotation editors (e.g., SMORE)
- NLP Information extraction apps (e.g., Aerotext)
- Learning markup for specific domains (e.g., ITTalks)

We’re experimenting with algorithms to learn to add DAML markup to text documents from a training corpus.

See poster later today
LTD DAML Markup Learning System

- We started with the Stalker algorithm developed at ISI for learning “web scraping” rules.
- The ontology structure constrains where objects are sought (e.g., Location has Room, Building, etc.)
- Rules are generated with a hill climbing algorithm.
- Many rules are generated so co-training techniques can be used to mark up plain documents.
- Results on learning to markup talk announcements are promising.
Problem: How do we do information retrieval over documents and queries which combine free text and semantic web markup?

- IR systems and KB systems use different models

One Solution: (1) index both the text and markup and then (2) use existing IR systems to find documents that match queries

Issues: (1) How do we index markup? (2) When and where do we do inferencing over the markup?

Applications: (1) Improved recall and precision for IR systems, (2) Retrieving documents for question answering.
OWLIR Framework

Event Descriptions

Text

Classification

Movie
Sport
Talk
Trip

Event Categories

Info Extraction

LMCO
AeroText + Java

Extract triples & reason

Jess

Text + triples

Agents

Expand Event Description

Text + triples

Convert triples to index terms

Index

Text

Convert triples to index terms

IR Engine

Text + triples

Query User Interface

Must
OK
Must not

Final Results

Inference on results

Results User Interface

Extract triples & reason

Jess

Text + DAML

Classifi cation

Query

User

Interface

Text

Movie
Sport
Talk
Trip
Parting Thoughts

- The Semantic Web opens up new opportunities to integrate knowledge-based reasoning and machine learning techniques.
- And to do so in an environment (the Web) in which it can have a major impact.
- Of special interest is the inclusion of distributed meta data of various kinds.
  - E.g., trust, provenance, significance, ...
- How will semantic web languages be used?
  - As markup on web documents?
  - As a web-based Knowledge interlingua?
  - As a part of an Agent Communication Language?
  - To express social meta data?
- Only experimentation will tell.
Classifying Talks

Topics

Ontology

uses

ACM CCS Ontology

uses

CMU Bow

statistical
text analysis
tools

e.g.: ACM CCS

e.g.: 5K ACM abstracts

Training corpus

ACM CCS classifier

topics

Now is the time for all good men to come to the aid of the country. Now is the time for

5K ACM abstracts
Mapping between topic ontologies

- **T1**
  - Topic ontology T1
  - Training corpus T1

- **T2**
  - Topic ontology T2
  - Training corpus T2

**CMU**

**Bow**

statistical
text analysis
tools

- T1 ↔ T2 mapper
  - \{(t2:bar, 0.8), (t2:qux, 0.7), \ldots\}

- t1:foo
(3) DAML and Agents

- Much multi-agent systems work is grounded in Agent Communication Languages (e.g., KQML, FIPA) and associated software infrastructure such as the DARPA Grid.
  - The paradigm has been peer-to-peer message oriented communication mediated by brokers and facilitators.
- The DAML program invites different paradigms which will require some changes in ACLs and their associates software systems.
  - Agents “publish” beliefs, requests, and other “speech acts” on web pages.
  - Agents “discover” what peers have published on the web.
- The software agent research community is very interested in the semantic web and DAML.
XPA is a configurable “personal agent” which accepts FIPA messages from XTalks and other instances of XPAs as well as applications, e.g. MS Outlook.
Results?

- We’ve done preliminary experiments measuring recall and precision over a small collection of 1500 event announcements and 12 queries.

- Comparing
  - Only free text
  - Free text + base triples but no inferencing
  - Free text + triples + inferred triples

- Results of this small experiment show better recall and precision with each step.
(6) Modeling trust policies

- There’s great interest in modeling and using trust in open systems such as the web and pervasive computing.
- Trust based systems are good when you have to deal with agents that are not well known authenticated users.
- Computing trust is usually
  - Based on observed properties of agents
  - Has a social component (e.g., reputation)
- We’ve developed a trust policy language (REI) that uses DAML+OIL as it’s encoding for (1) agent properties and (2) policy rules.