Future Research Challenges and Needed Resources for

The Web, Semantics and Data Mining

Tim Finin
UMBC, Baltimore MD
finin@umbc.edu

http://ebiquity.umbc.edu/resource/html/id/243

10/10/07
Overview

• Motivation
• The web’s new research challenges
• Needed resources
The Sum of Human Knowledge, 1907

WHEN IN DOUBT—“LOOK IT UP” IN

The

Encyclopaedia Brittanica

The Sum of Human Knowledge

29 volumes, 28,150 pages,
44,000,000 words of text.
Printed on thin, but strong
opaque India paper, each
volume but one inch in
thickness.

THE BOOK TO ASK QUESTIONS OF

FOR READING OR FOR STUDY
The Sum of Human Knowledge, 2007

I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions.
The Web

• The Web
• The 2.0 Web
• The social Web
• The semantic Web
• The game theoretic Web
• The next Web
The Web

• The Web is the most important new source of data and knowledge in our generation

• The Web is fundamentally unlike other information sources that preceded it
  – Today’s Web is not your father’s Internet

• We’ve been mining the Web since the beginning

• But it’s still evolving rapidly and unpredictably
  – Today’s Web won’t be your children’s Internet
The 2.0 Web

• The “current web” is characterized by
  – Mashups, dynamic data, tags, folksonomies
  – Sophisticated server-side programs
  – Sophisticated client-side programs (javascript)

• Challenges:
  – more of the information is bound up in client-side scripts -- how much should be evaluated?
  – More metadata and relational data
  – Taking advantage of new infrastructure like ping servers, feeds, rich APIs, etc.
The Social Web

• Estimates are that more than half of new Web content is being generated by users
  – Blogs, YouTube, wikis, forums, reviews, photo sharing sites, microblogging, public mailing lists

• Underlying social networks are of keen interest

• Challenges:
  – Very dynamic, subject to fads
  – Integrating and understanding the SNs
  – Modeling and recognizing communities, influence, sentiment
  – Serious privacy concerns
Once entered, the data is embedded into the blog post and Swoogle is pinged to index it.
The Semantic Web

• The W3C model has been evolving, from “KR on the Web” to a “Web of data”
  – Addresses data interoperability needs of Web 2.0
  – Nascent RDFa standard for integration of content & data
  – Slow but steady uptake industry (MS, Oracle, Adobe, …)

• Competing models like Google Base, Freebase, …

• **Challenges:** When and where and how much to do expensive reasoning, exploring interplay between KR and ML, trust and provenance, what’s the “right” paradigm, inventing SW infrastructure
• http://swoogle.umbc.edu/
• Running since summer 2004
• 2.3M RDF docs, 525M triples, 10K ontologies, 15K namespaces, 1.6M classes, 190K properties, 55M instances, 1000 registered users
The Game Theoretic Web

• At the 2007 Singularity Summit, Peter Norvig was asked if Google had been surprised by any emergent behavior on the Web
• His answer was quite interesting…
The Game Theoretic Web

“The other thing that I hadn't really thought about when we started this all is how game theoretic the whole thing is. At first we thought of ourselves as this observer of the Web. That the Web was out there and we made a copy of it and indexed it and if people wanted they could come and access that index. But it was just a reflection of the Web out there. And now we understand that we're co-evolving with the Web and that when we make a move it changes the Web and when the Web changes we change and going back and forth. And so all the search engine optimizers are watching what we do and we watch what they do and the Web is the interaction between us.”
Game Theoretic Web Challenges

• How do you do accurate data mining when the data providers are or may be trying to game the system?
• Current culprits: SEOs and spammers
• Increasingly: any web-savvy self promoter
  – E.g., most of us
“Honestly, Do you think people who make $10k/month from adsense make blogs manually? Come on, they need to make them as fast as possible. Save Time = More Money! It's Common SENSE! How much money do you think you will save if you can increase your work pace by a hundred times? Think about it…”

“Discover The Amazing Stealth Traffic Secrets Insiders Use To Drive Thousands Of Targeted Visitors To Any Site They Desire!”

“Holy Grail Of Advertising… “

“Easily Dominate Any Market, Any Search Engine, Any Keyword.”
(Some) Needed Resources

• Common datasets for Social Media and the Semantic Web
  – ICWSM is establishing a collection for social media

• Better abstract models for Social Media
  – E.g., SecondSpace model for the Blogosphere graph
  – E.g., understanding users’ intent in using social media

• Better models for trust and provenance

• Ideas to let us “Do no evil” by respecting privacy