03-60-569: Semantic Web (2013 Fall)

Jianguo Lu

University of Windsor

September 4, 2013

Table of contents

The Web

Definition of the Web

The World Wide Web is a system of interlinked hypertext documents accessed via the Internet

The web relies on three mechanisms to make these resources available to audience:

- A uniform naming scheme for locating resources on the web (e.g., URIs).
- Protocols, for access to named resources over the web (e.g., HTTP).
- Hypertext, for easy navigation among resources (e.g., HTML).

What we use todays web for

- Browse (e.g., our course web site)
- Search (e.g., Google)
- Online social networking (facebook, twitter, ...)
- ...

Internet and the Web

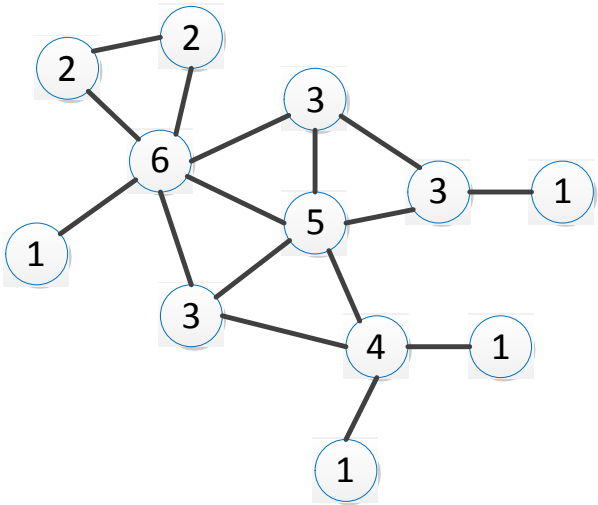
Internet:

- Global system of interconnected computer networks that use the standard Internet Protocol Suite (TCP/IP)
- Internet is a more general term
- Includes physical aspect of underlying networks and mechanisms such as email, FTP, HTTP

Web: Associated with information stored on the Internet

- Refers to a broader class of networks, i.e. Web of English Literature

Both Internet and web are networks

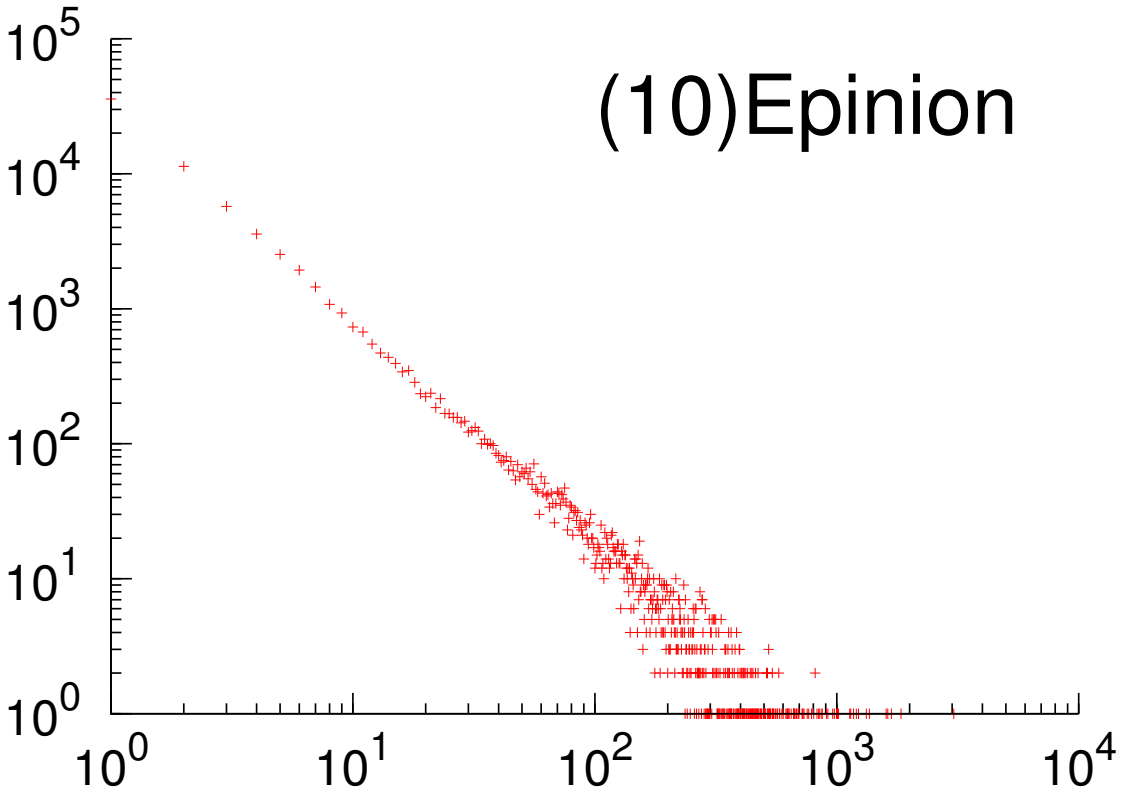


What are the challenging tasks for the Web

- Search Engine
  - **Crawling**: collect the web pages;
  - **Indexing:** There are tools that help us index documents, and provide relevant answers to queries. There are many matches. What matters are the top a few pages
  - **Ranking:** Important pages are returned first. How to decide the importance? Incoming links. Important link weights more. How to decide the importance?
  - **Detect similar pages** How do we detect similar pages (what do we mean by similar page)? When there are billions of pages, what is the algorithm?
- Online Social Networks. What are the network structure. E.g., the clustering coefficient, how many triangles are there in the network.
- How to handle the semantics
  - Latent Semantic Indexing
  - Ontology

Web characteristics

- Graph structure of the Web
- Power laws, small world, bow-tie structure
- Users and queries



Web crawling

- The first step to build a search engine
- Crawler architecture and policies, breadth first and depth first crawling;
- Crawling OSNs and Deep web

Near duplicate detection

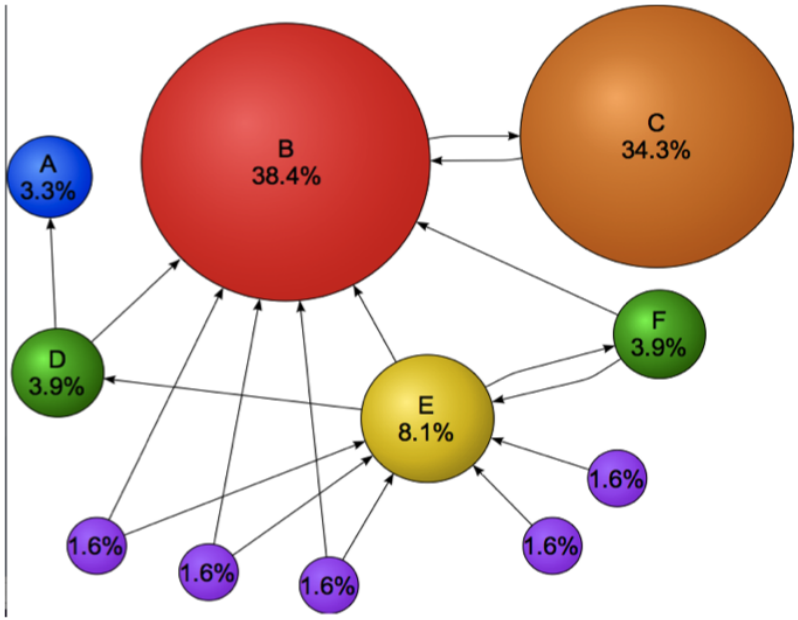
- Detect similar pages among vast number of documents
- Have to be efficient
- Use shingles to represent pages
- Jaccard similarity

Search Engines

- The process to build a search engine
- Use Lucene search engine API

Page Rank Algorithm

- The algorithm used by Google and other search engines;
- Have many engineering challenges



Problem with search engined: High recall, low precision

- Searching for *semantic web* in google matches estimated. 7,610,000 pages in 2009 14,000,000 pages in 2011
- Only first a few pages will be viewed
- Need to be intelligent
- Results are highly sensitive to vocabulary
  - If you search for *faculty member*, you may not be able to find *professor*.
- Results are single web pages
  - You will not be able to find *professors who are teaching semantic web in Canada*
  - Current search engines can not integrate related web pages in different places

Approach One: Latent Semantic Indexing

- Words used in the same context tend to have similar meanings
- Capture the semantic meanings that are otherwise latent in text
- Use SVD technique (Singular Value Decomposition) from linear algebra
- Ideas developed in the seventies, applied widely in information retrieval

Approach Two: Ontology

- Represent web content in a form that is more machine-processable.
- Use intelligent techniques to take advantage of these representations.

Related technologies

- Explicit Metadata (XML)
- Ontologies
- Logic and Inference
- Web service
- Data integration
- Intelligent software agents

HTML


```
<h1> Agilitas Physiotherapy Centre
Welcome to the home page of the Ag
 Kelly Townsend (our lovely secretar
<h2> Consultation hours </h2>
Mon 11am – 7pm <br>
Tue 11am – 7pm <br>
Wed 3pm – 7pm <br>
Thu 11am – 7pm <br>
 Fri 11am – 3pm <p>
 But note that we do not offer cons
<a href=". . .">State Of Origin </a
```

Agilitas Physiotherapy Centre

Welcome to the home page of the Agilitas Physiotherapy Centre. Do you feel pain? Have you had an injury? Let our staff Lisa Davenport, Kelly Townsend (our lovely secretary) and Steve Matthews take care of your body and soul.

Consultation hours

Mon 11am - 7pm
Tue 11am - 7pm
Wed 3pm - 7pm
Thu 11am - 7pm
Fri 11am - 3pm

But note that we do not offer consultation during the weeks of the State Of Origin games.

Markups are for formatting or presentation. Web contents are for humans instead of programs

Problems with HTML

- Humans have no problem with this.
- Machines (software agents) do.
- How to distinguish therapists from the secretary;
- How to determine exact consultation hours;
- They would have to follow the link to the State Of Origin games to find when they take place.
-

One approach is to build intelligent program to extract the relevant information

A better representation-xml


```
- <company>
    <treatmentOffered>Physiotherapy</treatmentOffered>
    <companyName>Agilitas Physiotherapy Centre</companyName>
  - <staff>
      <therapist>Lisa Davenport</therapist>
      <therapist>Steve Matthews</therapist>
      <secretary>Kelly Townsend</secretary>
    </staff>
  </company>
```

Explicit Metadata

- This representation is far more easily processable by machines
- Metadata: data about data. Metadata capture part of the meaning of data

Ontology does not rely on text-based manipulation, but rather on machine-processable metadata

Meta data is not enough

- E.g., how can machine know that $<$ *secretary* $>$ and $<$ *therapist* $>$ are people? We need to define that $<$ *secretary* $>$ is a subclass of $<$ *people* $>$
- How can we specify the rule that $<$ *secretary* $>$ can not be a $<$ *therapist* $>$? We need to say that classes $<$ *secretary* $>$ and $<$ *therapist* $>$ are disjoint.

Topics

- Web as a graph, search basics
- Web analysis, degrees, diameter, scale-free network, near duplicate detection, page rank algorithm,
- Mining online social network
- Ontology–Introduction, RDF, RDFS, OWL, ontology engineering;

Marking scheme

- Final Exam: 40%
- Presentation: 10%
- Class participation: 10%
- One project–40%

Paper presentation 10% (before week 7)

- Papers and approximate presentation time are listed on course web site.
- When you are ready to present, let me know the time and the paper you selected.
- When your presentation slides is ready, send me the slides.
- Each presentation is 30 minutes long.

Class participation10%

Active in classes, paper presentation

Project

There are five 'small' tasks

- Build a crawler to collect web pages (8%) Say all the pages in the domain of uwindsor.ca
- Construct a search engine using Lucene (8%)
- Compute page rank (8%)
- Find near duplicate pages (8%)
- Compute clustering coefficient (8%)

- Lectures on these topics are finished within the first a few weeks, so that you can start the project soon;
- Naive implementation is straightforward. You can use any programming language you prefer.
- The submission deadline for the first two is November 10. The due date for the last two subtasks are on the last day of the class. You need to submit the source code as well as the ppt slides to be presented in the class.
- On the last day of the class, you will demonstrate and explain your project in front of the class.
- You can get high marks only if your program can process big data.
- Student presentations can address the details on these algorithms, especially the scalability techniques.

Tentative Schedule

- Week 1-4: Web as a graph, Crawling, Search Engine construction, paper presentations.
- Week 5-7: Link analysis (page rank algorithm), near duplicate detection, paper presentations.
- Week 8-10: Ontology (RDF, RDFS, OWL, Ontology query language)
- Week 11-12: Latent semantic indexing (Eigenvectors, singular value decomposition),

Exam

Web graph basics, Ontology, Web and OSN mining algorithms. Page rank, similar pages, singular value decomposition, latent semantic indexing.

- Test topics covered in lectures
- Test issues involved in projects
- Concepts, Logics, and Algorithms
- Student presentations will not be tested

Reading materials

MMD Anand Rajaraman and Jeff Ullman, Mining of massive datasets , 2012.

IR Christopher D. Manning, Prabhakar Raghavan and Hinrich Schutze,Introduction to Information Retrieval, Cambridge University Press. 2008. Chapters 18, 19, 20, 21 only.

SW Grigoris Antoniou, Frank van Harmelen,A Semantic Web Primer, MIT Press