From Knowledge Exchange To Knowledge Discovery

Edmund Yu, PhD

July 14, 2010

edmund@redcliffanalytics.com
esyu@syr.edu

From WOKE To PSF

- The original RFP states:
  - Although the “knowledge repository” approach of WOKE is suitable for collecting, organizing and making re-usable the accumulated knowledge of the workforce system, it is not easily adapted to a “conversational collaborative foundation of knowledge management” implied by the many social networking systems that have come about in the last few years.
  - PSF should incorporate business databases, information representation and retrieval techniques, and Web 2.0 functionalities into a single platform that registered users may use to quickly identify solutions or leads to solutions for their problems.

Information Retrieval

- Is the indexing and retrieval of textual documents.
- Searching for pages on the World Wide Web is the most recent and perhaps most widely used IR application
- Concerned firstly with retrieving *relevant* documents to a query.
- Concerned secondly with retrieving from *large* sets of documents *efficiently*.
- PSF uses Microsoft Search API, provided by Microsoft Search Server, to satisfy these requirements.

wwwcsif.cs.ucdavis.edu

Microsoft Search

- Search across site collections
- Scopes
- Content Sources
  - File Shares
  - Public Folders
  - External Websites (**Web Crawling**)
- Search Web Parts
- Federated Search
- Extensible
- No document limits

Web Crawling

- Start with a comprehensive set of root URL's from which to start the search
- Follow all links on these pages recursively to find additional pages. Which link is to be followed first depends on the selection policy: breadth first, depth first, pageranks, focused crawling, …
- Pages will be revisited based on the schedule
- Politeness policy is needed so that crawlers don't overload web servers (e.g. set a delay between HTTP requests)
- Web crawling can be parallelized to maximize download rate

Federated Search

Federated search consists of

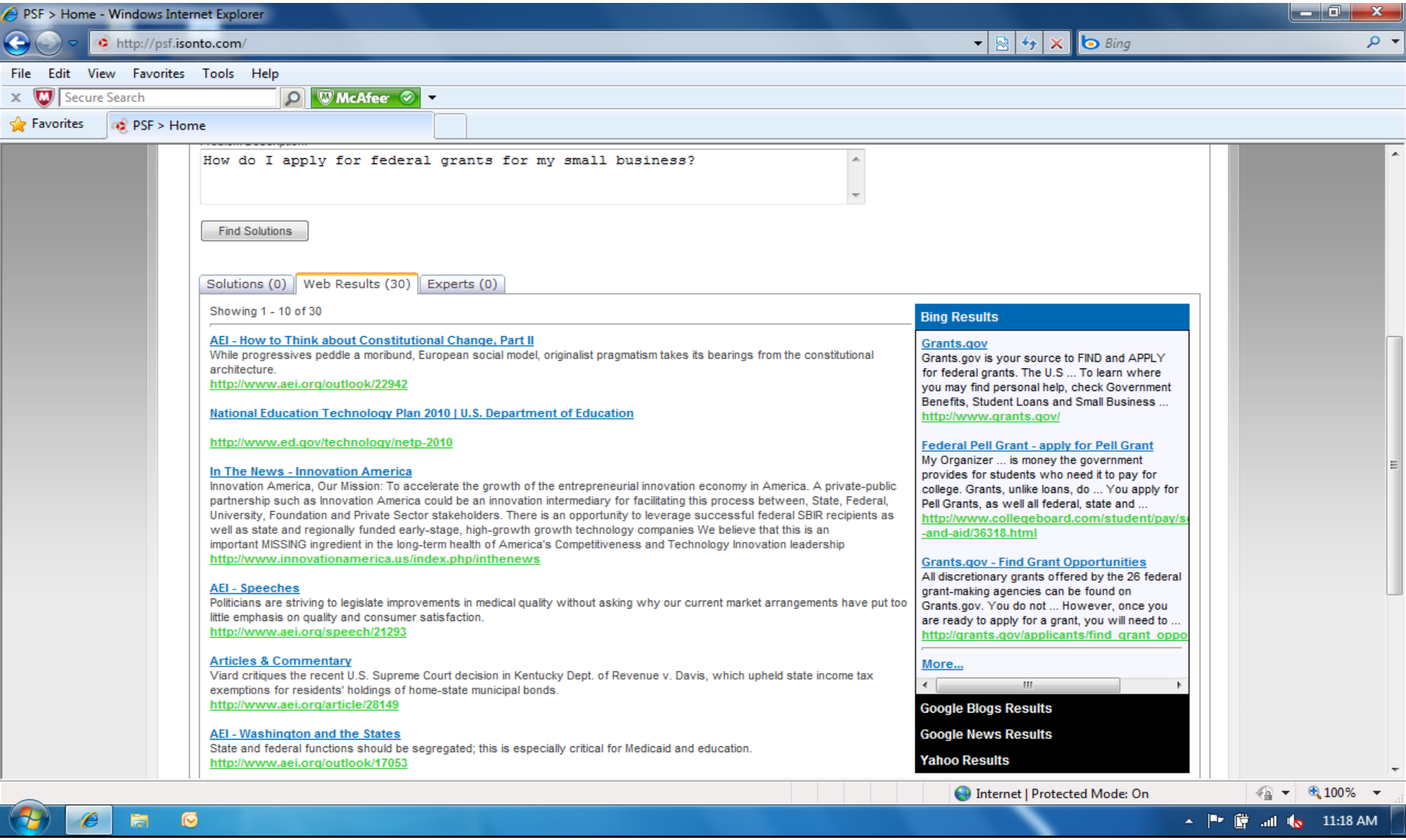
- transforming a query and broadcasting it to a group of disparate databases with the appropriate syntax,
- merging the results collected from the databases
- presenting them in a succinct and unified format with minimal duplication, and
- providing a means, performed either automatically or by the portal user, to sort the merged result set.

(Jacsó 2004; http://en.wikipedia.org/wiki/Federated_search)

Federated Search – cont.

- Microsoft Search has **connecters** to
  - Main Search Engines: Bing/Live, Yahoo
  - News: Live.com News, Yahoo News, Google News, Wired, The Register (Sci/Tech News)
  - Media: Flickr (photo sharing), Yahoo Images, YouTube, PodScope (Podcast search engine), del.icio.us (social bookmarks)
  - Information Resources: MSDN, TechNet, Wikipedia, Encyclopedia Britannica
  - Blogs: Google Blog Search, Technorati
  - Social Networks: LinkedIn

Retrieval Functionalities of PSF



Improving Retrieval Functionalities

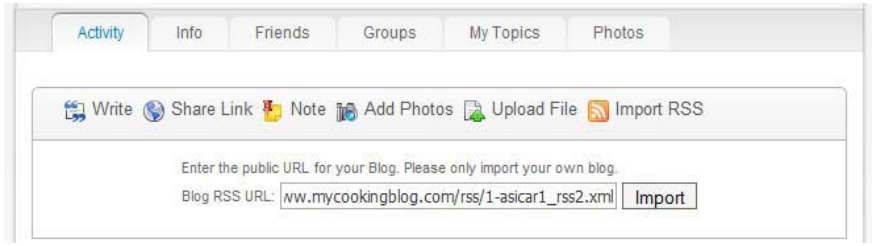
- Automatically classify the retrieved Web pages or documents (**Text Mining**), to make sure they match with the Problem Category.
- Given a specific problem description, gather the top X relevant Web pages or documents , and then identify the most relevant paragraph, i.e. the paragraph that contains the highest concentration of relevant terms. This paragraph has the highest likelihood of containing a solution *to that specific problem*.
- Given a specific problem description, gather the URLs of the top X search results. For those domain names that occur at least Y times, they are potentially information-rich, and should be crawled routinely to provide more useful information *for that specific problem.*

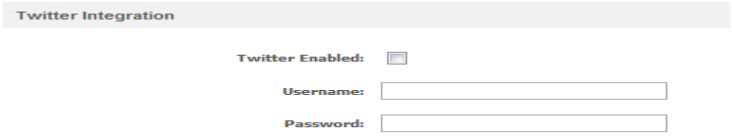
Web 2.0 Functionalities

- The term "Web 2.0" is commonly associated with web applications that facilitate interactive information sharing, interoperability, user-centered design, and collaboration on the WWW. A Web 2.0 site allows its users to interact with each other as contributors to the website's content, in contrast to websites where users are limited to the passive viewing of information that is provided to them. (http://en.wikipedia.org/wiki/Web_2.0)

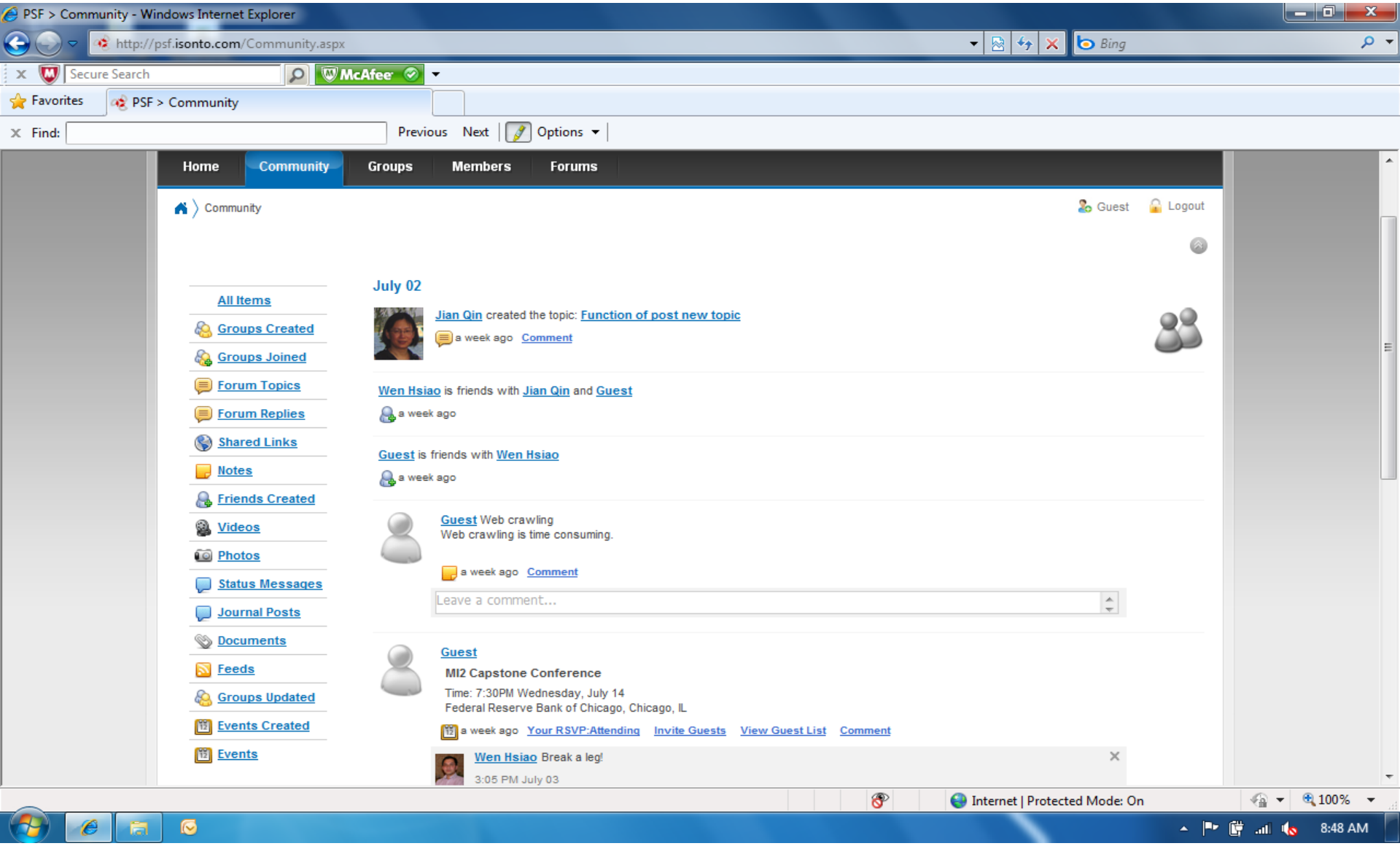
PSF As A Web 2.0 Site

- Is a online community builder. Members of the community can share information, exchange ideas, and network with each other.
- Additional features include:
  - RSS Integration
    - You can post RSS feeds or subscribe to them
  - Twitter Integration
    - You can enable Twitter Integration to have your status updates shared with your Twitter account
  - Amazon S3 Integration
    - You can enable Amazon S3 integration for storage
  - Events…
    - You can create events, send invitations & manage guests

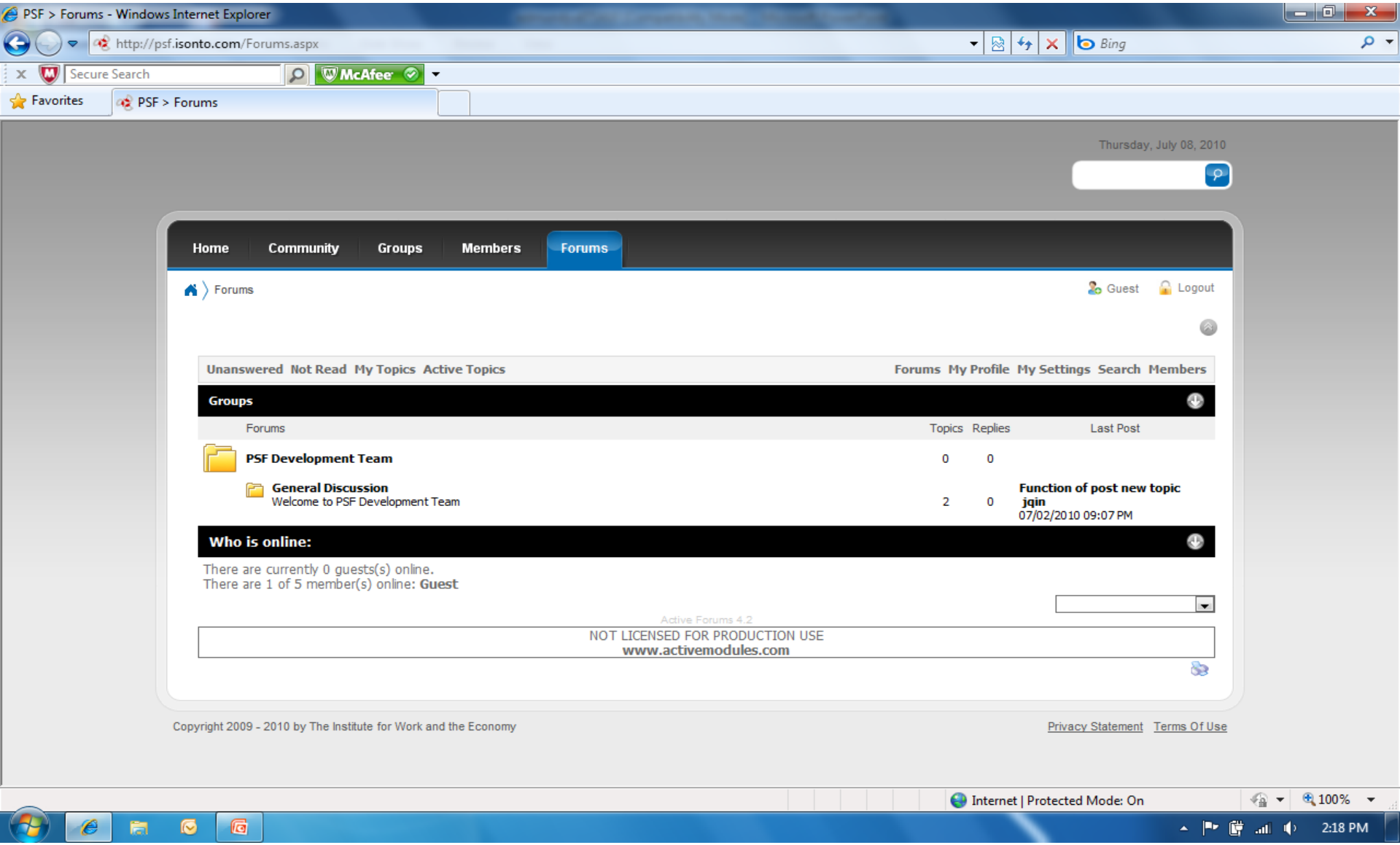




PSF As A Web 2.0 Site – cont.



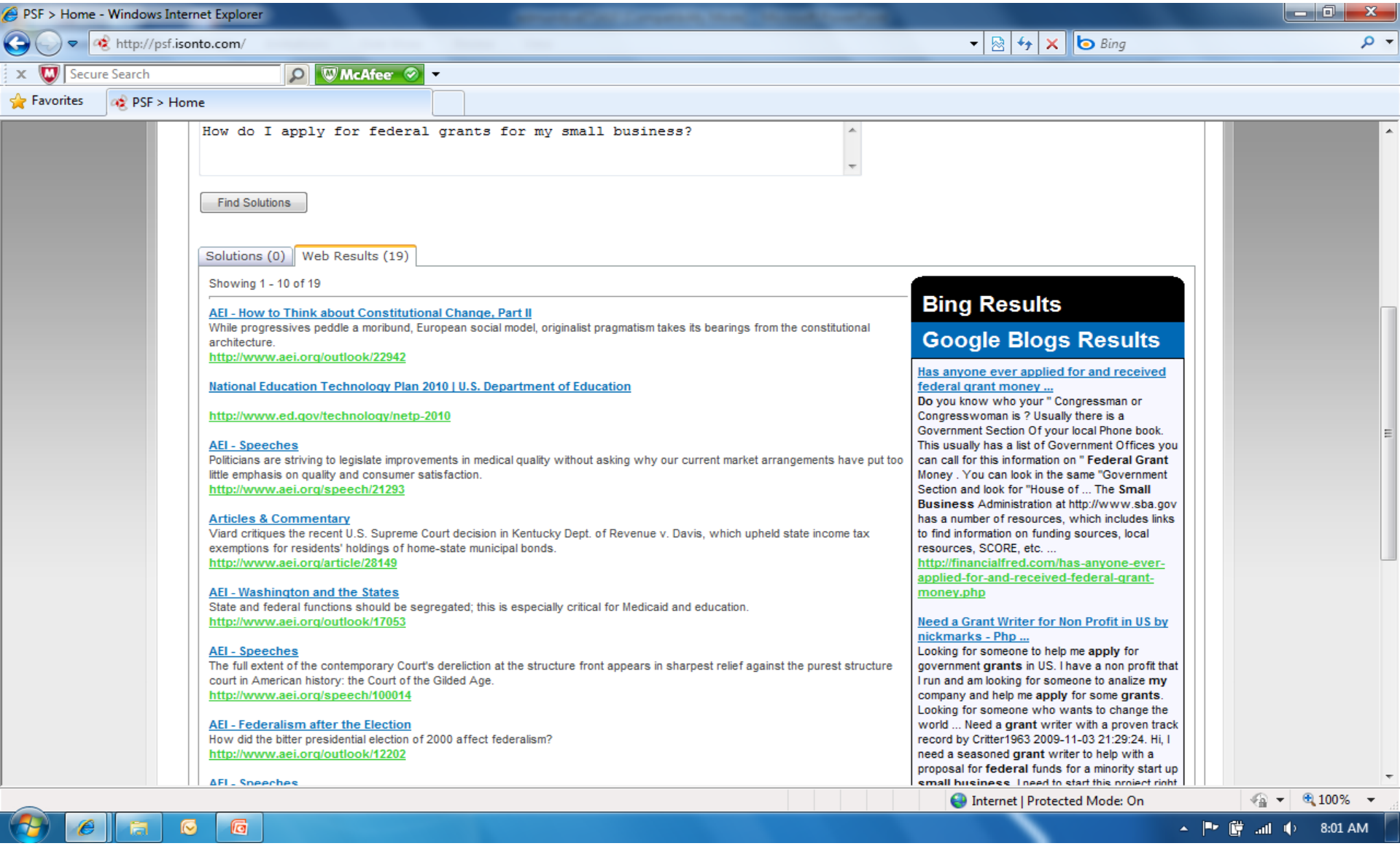
PSF As A Web 2.0 Site – cont.



PSF As A Portal To Social Media

- Social media describes the online tools and platforms that people use to share opinions, insights, experiences, and perspectives with each other. (http://en.wikipedia.org/wiki/Social_media)
- It builds on the ideological and technological foundations of Web 2.0
- It's a dynamic and growing area that includes blogs, tweets, wikis, forums, photo/video sharing sites, etc.
- PSF will function as a portal to those sites (e.g. Google Blogs, Twitter, YouTube, Wiki, Flicker, etc. mentioned earlier under **Federated Search**)

PSF As A Portal To Social Media



Towards Knowledge Discovery

- Refers to systems that create new knowledge through the implementation of intelligent algorithms such as data mining, and through the inference of data relationships (Fayyad et. al. 1996).
- Discovering Knowledge from Text Databases (reports, journal articles, Web pages, blogs, wiki's, tweets,…) → **Text Mining (TM)**
- The phrase "text mining" is generally used to denote any system that analyzes large quantities of natural language texts and detects lexical or linguistic usage patterns in an attempt to extract probably useful (although only probably correct) information
- PSF + Text Mining = Automated PSF (**APSF**)

Two Main TM Tasks - TC

- Text Categorization (TC) is the problem of automatically assigning predefined categories to natural language texts
- Or, more formally
  - Given an ***instance space*** **X** consisting of all possible text documents (or objects), and
  - A training set of labeled text documents for the ***target classification function*** **f(x)**, which can take on any value from a finite set **V** (e.g. interesting, not_interesting, about_genetics, not_about_genetics)
  - Learn from this training set to predict the target value for subsequent text documents
  - The text documents are usually represented in the form (**x**, *c*), where x is a vector of feature-values, such as the number of occurrences or the presence of the words, phrases, or other semantic entities in the documents, and *c* is the class label

Pre-Defined Categories

Yahoo! Directory - Windows Internet Explorer

http://dir.yahoo.com/

Yahoo! Directory

Advanced Search | Suggest a Site | Email This Page

Arts & Humanities
Photography, History, Literature...

Business & Economy
B2B, Finance, Shopping, Jobs...

Computers & Internet
Hardware, Software, Web, Games...

Education
Colleges, K-12, Distance Learning...

Entertainment
Movies, TV Shows, Music, Humor...

Government
Elections, Military, Law, Taxes...

Health
Diseases, Drugs, Fitness, Nutrition...

News & Media

The Spark: Glorifying the American Girl

By Dave Sikula
Tue, July 6, 2010, 12:01 am PDT

Opening a Broadway show on July 8 in the pre-air-conditioned year of 1907 was an unorthodox move, but no one ever accused Florenz Ziegfeld, Jr. of being conventional.

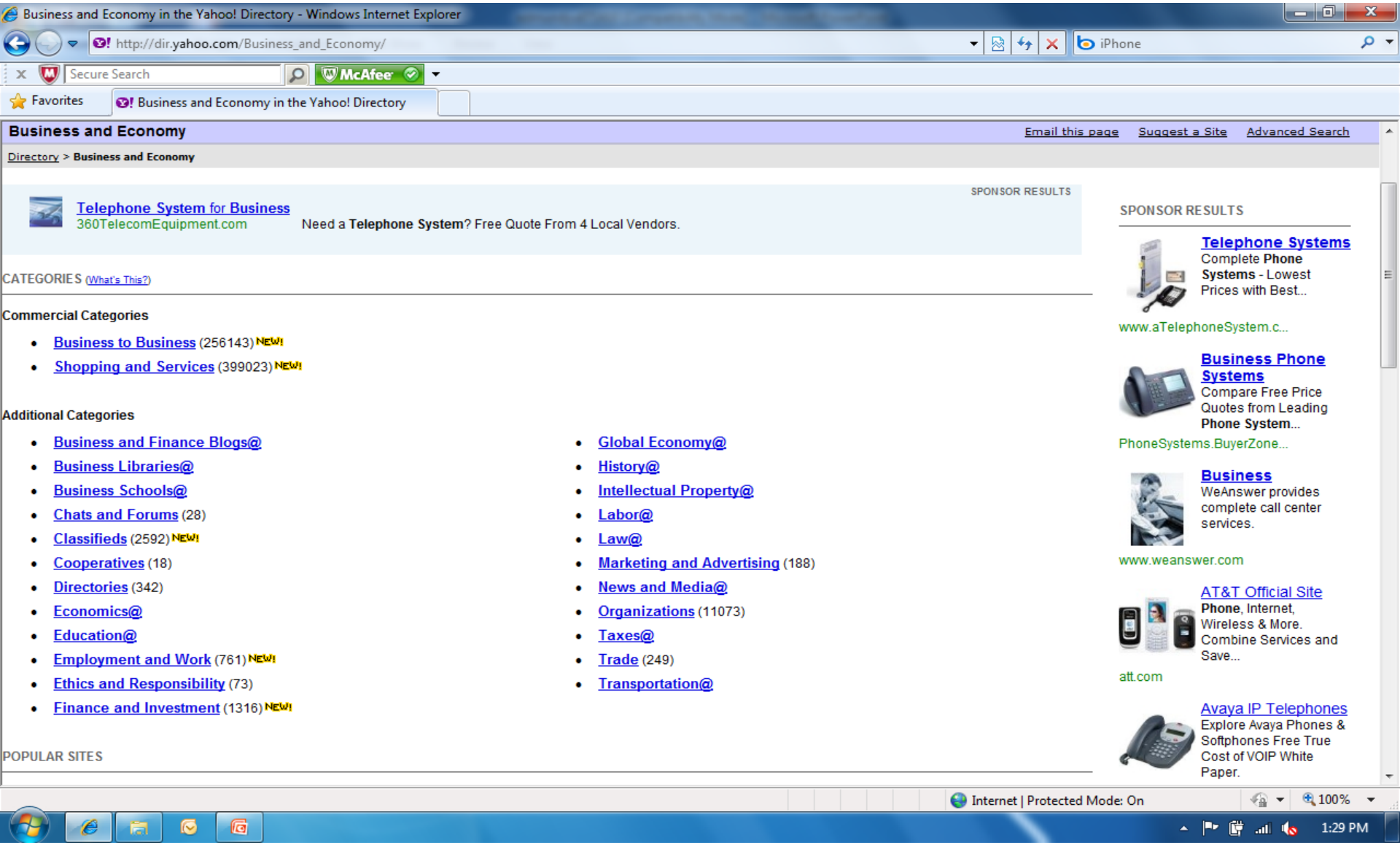
1912 poster for Ziegfeld Follies

Ziegfeld (pronounced "ZEEG-feld," if you please, not "Zig-field") began his career as a small-time showman, presenting Eugen Sandow, "The World's Strongest Man," at the 1893 Chicago World's Fair. Sandow scandalously performed nearly nude or in flesh-colored tights and soon became a leading attraction. Ziegfeld parlayed that success by importing French singer Anna Held to star in a smash-hit show -- and then marrying her.

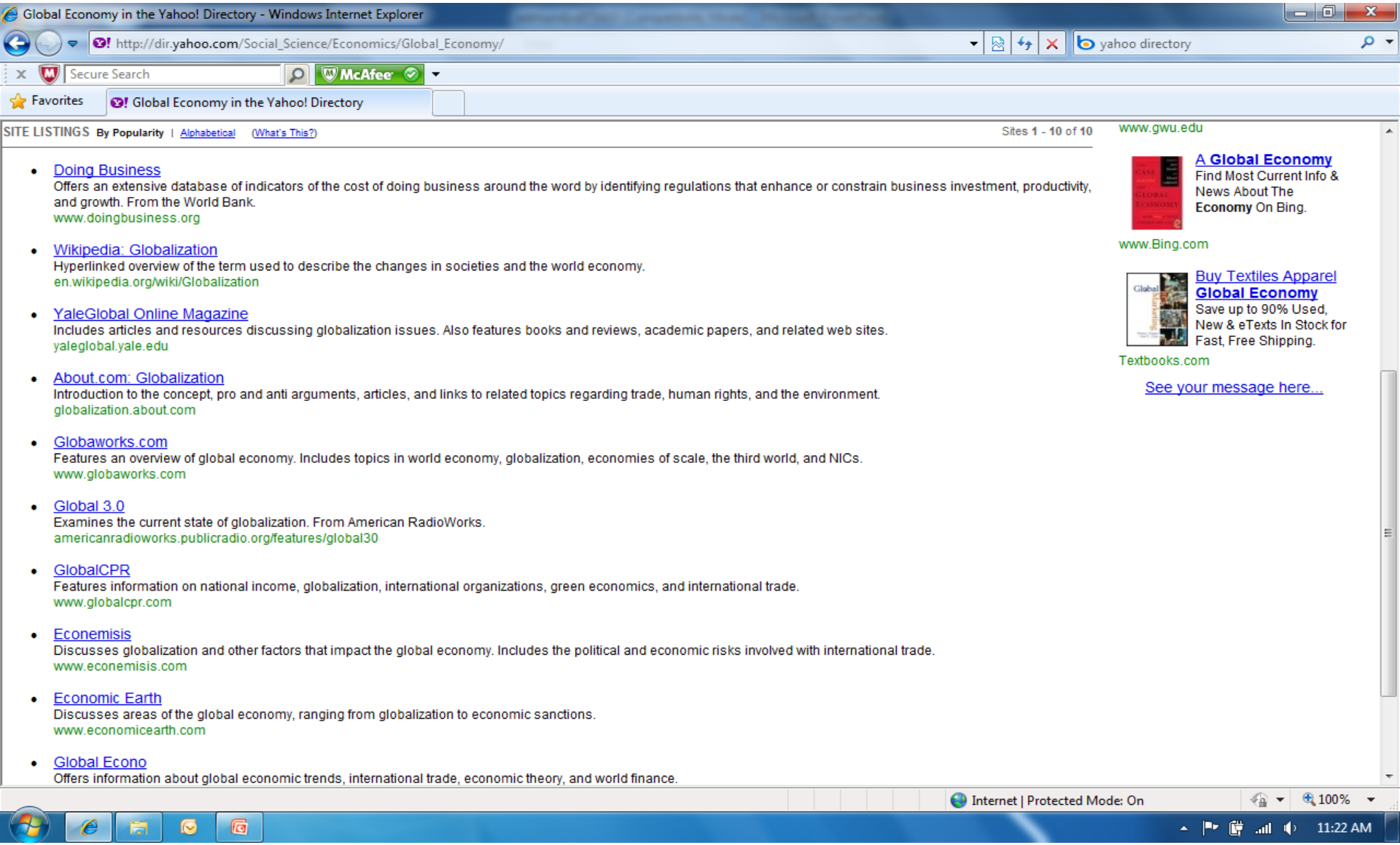
In 1906, Ziegfeld, who always had an eye for feminine beauty, decided he would emulate Paris' famous Folies Bergère and present his own "Follies" in New York, dedicated to "Glorifying the American Girl." He surrounded top comedians with lavish sets and costumes, songs by such writers as George Gershwin, Irving Berlin, and Jerome Kern -- and the tallest, most beautiful women Ziegfeld could find. The Follies became another smash and continued annually until 1927 -- when even Ziegfeld couldn't outrun the Great Depression. But in those years, such entertainment legends as Nora Bayes, Sophie Tucker, Fanny Brice, Ed Wynn, W.C. Fields, Will Rogers, and Eddie Cantor appeared in the Follies -- and Ziegfeld even managed to break Broadway's color barrier by starring Bert Williams, the first black performer to appear on an American stage with an otherwise white cast.

Internet | Protected Mode: On 100% 1:14 PM

Pre-Defined Categories, cont.



Training Data – Global Economy



Two Main TM Tasks - IE

An end-to-end information extraction (IE) system should consist of:

- Named Entity Recognition and Classification (NERC) component, the purpose of which is to recognize and classify named entities such as persons, organizations, companies, locations, products, dates, monetary amounts...
- Relationship Extraction component, the purpose of which is to identify the relationship between a pair of entities extracted by NERC. Common relations of interest include:
  - [Person] employee_of [Company]
  - [Product] product_of [Company]
  - [Location] location_of [Company]

Two Main TM Tasks – IE, cont.

- Co-reference Resolution component. Co-reference is the linguistic phenomenon whereby two or more linguistic expressions may represent or indicate the same entity. These include:
  - variant forms of name expression (Steve Jobs … Jobs)
  - definite noun phrases and their antecedents (Jobs … the co-founder and CEO of Apple …)
  - pronouns and their antecedents (Jobs … he).
- Event Extraction component. In event extraction, an event is defined as an activity or occurrence of interest such as a terrorist act, an airline crash, a rocket launch, or a product release
  - A [Product_Release] event might involve a [Company: Apple], a [Product : iPhone 4.0], a [Date: June 24, 2010] and [Cost: $299]

Current IE Capabilities

- By using the API provided by the currently available IE systems, we will be able to identify:
  - Entities located in web pages, news, blogs, or tweets
  - Structured information about an entity (entity profiles)
  - How entities are related (entity web)
  - Popular entities on the entity web broken down by various categories
  - How entities or combinations of entities are related to documents, images and video found on the web
  - Media recommendations for entities, or combinations of entities, based on web pages, news, blogs, or tweets
- They can help answer questions about those entities

Envisioned Automated PSF (APSF)

Discovered (Potential) Solutions