

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.



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 **connectors** 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

The screenshot shows a Windows Internet Explorer browser window with the address bar displaying <http://psf.isonto.com/>. The search bar contains the query "How do I apply for federal grants for my small business?". Below the search bar is a "Find Solutions" button. The results section shows "Solutions (0)", "Web Results (30)", and "Experts (0)". The first result is "AEI - How to Think about Constitutional Change, Part II" with a link to <http://www.aei.org/outlook/22942>. Other results include "National Education Technology Plan 2010 | U.S. Department of Education" and "In The News - Innovation America". A "Bing Results" sidebar is visible on the right, listing "Grants.gov", "Federal Pell Grant - apply for Pell Grant", and "Grants.gov - Find Grant Opportunities". The taskbar at the bottom shows the Windows logo, several application icons, and the system tray with the time 11:18 AM and 100% zoom level.

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 an 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

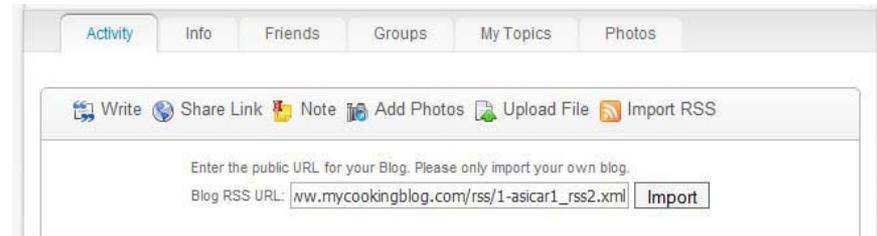
❖ You can enable Twitter Integration to have your status updates shared with your Twitter account

❖ Amazon S3 Integration

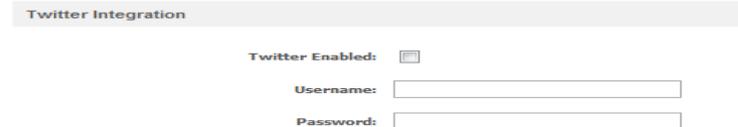
❖ You can enable Amazon S3 integration for storage

❖ Events...

❖ You can create events, send invitations & manage guests



The screenshot shows a web interface with a navigation bar containing tabs for Activity, Info, Friends, Groups, My Topics, and Photos. Below the navigation bar is a section with several icons and labels: Write, Share Link, Note, Add Photos, Upload File, and Import RSS. A text input field is labeled "Blog RSS URL:" and contains the text "www.mycookingblog.com/rss/1-asicar1_rss2.xml". To the right of the input field is an "Import" button. Above the input field, there is a small instruction: "Enter the public URL for your Blog. Please only import your own blog."



The screenshot shows a web interface with a section titled "Twitter Integration". Below the title, there is a checkbox labeled "Twitter Enabled:" which is currently unchecked. Below the checkbox are two input fields: "Username:" and "Password:".

PSF As A Web 2.0 Site – cont.

PSF > Community - Windows Internet Explorer
http://psf.isonto.com/Community.aspx

Secure Search McAfee

Home **Community** Groups Members Forums

Community Guest Logout

[All Items](#)
[Groups Created](#)
[Groups Joined](#)
[Forum Topics](#)
[Forum Replies](#)
[Shared Links](#)
[Notes](#)
[Friends Created](#)
[Videos](#)
[Photos](#)
[Status Messages](#)
[Journal Posts](#)
[Documents](#)
[Feeds](#)
[Groups Updated](#)
[Events Created](#)
[Events](#)

July 02

[Jian Qin](#) created the topic: [Function of post new topic](#)
a week ago [Comment](#)

[Wen Hsiao](#) is friends with [Jian Qin](#) and [Guest](#)
a week ago

[Guest](#) is friends with [Wen Hsiao](#)
a week ago

[Guest](#) Web crawling
Web crawling is time consuming.
a week ago [Comment](#)

Leave a comment...

[Guest](#)
M12 Capstone Conference
Time: 7:30PM Wednesday, July 14
Federal Reserve Bank of Chicago, Chicago, IL
a week ago [Your RSVP:Attending](#) [Invite Guests](#) [View Guest List](#) [Comment](#)

[Wen Hsiao](#) Break a leg!
3:05 PM July 03

Internet | Protected Mode: On 100% 8:48 AM

PSF As A Web 2.0 Site – cont.

PSF > Forums - Windows Internet Explorer
http://psf.isonto.com/Forums.aspx

Thursday, July 08, 2010

Home Community Groups Members **Forums**

Forums Guest Logout

Unanswered Not Read My Topics Active Topics Forums My Profile My Settings Search Members

Groups

Forums	Topics	Replies	Last Post
PSF Development Team	0	0	
General Discussion Welcome to PSF Development Team	2	0	Function of post new topic jqin 07/02/2010 09:07 PM

Who is online:
There are currently 0 guests(s) online.
There are 1 of 5 member(s) online: **Guest**

Active Forums 4.2
NOT LICENSED FOR PRODUCTION USE
www.activemodules.com

Copyright 2009 - 2010 by The Institute for Work and the Economy
Privacy Statement Terms Of Use

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, Flickr, etc. mentioned earlier under **Federated Search**)

PSF As A Portal To Social Media

PSF > Home - Windows Internet Explorer

http://psf.isonto.com/

Secure Search McAfee

PSF > Home

How do I apply for federal grants for my small business?

Find Solutions

Solutions (0) Web Results (19)

Showing 1 - 10 of 19

[AEI - How to Think about Constitutional Change, Part II](#)
While progressives peddle a moribund, European social model, originalist pragmatism takes its bearings from the constitutional architecture.
<http://www.aei.org/outlook/22942>

[National Education Technology Plan 2010 | U.S. Department of Education](#)
<http://www.ed.gov/technology/netp-2010>

[AEI - Speeches](#)
Politicians are striving to legislate improvements in medical quality without asking why our current market arrangements have put too little emphasis on quality and consumer satisfaction.
<http://www.aei.org/speech/21293>

[Articles & Commentary](#)
Viard critiques the recent U.S. Supreme Court decision in Kentucky Dept. of Revenue v. Davis, which upheld state income tax exemptions for residents' holdings of home-state municipal bonds.
<http://www.aei.org/article/28149>

[AEI - Washington and the States](#)
State and federal functions should be segregated; this is especially critical for Medicaid and education.
<http://www.aei.org/outlook/17053>

[AEI - Speeches](#)
The full extent of the contemporary Court's dereliction at the structure front appears in sharpest relief against the purest structure court in American history: the Court of the Gilded Age.
<http://www.aei.org/speech/100014>

[AEI - Federalism after the Election](#)
How did the bitter presidential election of 2000 affect federalism?
<http://www.aei.org/outlook/12202>

[AEI - Speeches](#)

Bing Results

Google Blogs Results

[Has anyone ever applied for and received federal grant money ...](#)
Do you know who your "Congressman or Congresswoman is? Usually there is a Government Section Of your local Phone book. This usually has a list of Government Offices you can call for this information on "Federal Grant Money". You can look in the same "Government Section and look for "House of ... The Small Business Administration at <http://www.sba.gov> has a number of resources, which includes links to find information on funding sources, local resources, SCORE, etc. ...
<http://financialfred.com/has-anyone-ever-applied-for-and-received-federal-grant-money.php>

[Need a Grant Writer for Non Profit in US by nickmarks - Php ...](#)
Looking for someone to help me apply for government grants in US. I have a non profit that I run and am looking for someone to analyze my company and help me apply for some grants. Looking for someone who wants to change the world ... Need a grant writer with a proven track record by Critter1963 2009-11-03 21:29:24. Hi, I need a seasoned grant writer to help with a proposal for federal funds for a minority start up small business. I need to start this project right

Internet | Protected Mode: On 100%

8:01 AM

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* \mathbf{X} consisting of all possible text documents (or objects), and
 - ❖ A training set of labeled text documents for the *target classification function* $\mathbf{f}(\mathbf{x})$, which can take on any value from a finite set \mathbf{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 (\mathbf{x}, c) , where \mathbf{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/

Secure Search McAfee

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.

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.



1912 poster for Ziegfeld Follies

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.

Ads by Yahoo!

[Local Night Clubs](#)
Find night clubs & dance clubs in your zip code w/ Yellow Pages.
[yellowpages.com](#)

[Detroit, MI Jobs \(Hiring Immediately\)](#)
100s Of New Job Openings In Detroit.
[www.LocalJobRush.com](#)

[David Michael Cantor](#)
Local Experts. Professional Service Serving Tempe AZ 85282
[QuickLocal.com](#)

[berlin job postings \(Hiring\)](#)
Positions Available in Berlin Now Accepting Applications.
[Berlin.searchworklisti...](#)

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

Pre-Defined Categories, cont.

Business and Economy in the Yahoo! Directory - Windows Internet Explorer

http://dir.yahoo.com/Business_and_Economy/

Secure Search McAfee

Business and Economy in the Yahoo! Directory

Business and Economy Email this page Suggest a Site Advanced Search

Directory > Business and Economy

Telephone System for Business
360TelecomEquipment.com Need a Telephone System? Free Quote From 4 Local Vendors.

SPONSOR RESULTS

Telephone Systems
Complete Phone Systems - Lowest Prices with Best...
www.aTelephoneSystem.c...

Business Phone Systems
Compare Free Price Quotes from Leading Phone System...
PhoneSystems.BuyerZone...

Business
WeAnswer provides complete call center services.
www.weanswer.com

AT&T Official Site
Phone, Internet, Wireless & More. Combine Services and Save...
att.com

Avaya IP Telephones
Explore Avaya Phones & Softphones Free True Cost of VOIP White Paper.

CATEGORIES (What's This?)

Commercial Categories

- [Business to Business](#) (256143) **NEW!**
- [Shopping and Services](#) (399023) **NEW!**

Additional Categories

- [Business and Finance Blogs@](#)
- [Business Libraries@](#)
- [Business Schools@](#)
- [Chats and Forums](#) (28)
- [Classifieds](#) (2592) **NEW!**
- [Cooperatives](#) (18)
- [Directories](#) (342)
- [Economics@](#)
- [Education@](#)
- [Employment and Work](#) (761) **NEW!**
- [Ethics and Responsibility](#) (73)
- [Finance and Investment](#) (1316) **NEW!**
- [Global Economy@](#)
- [History@](#)
- [Intellectual Property@](#)
- [Labor@](#)
- [Law@](#)
- [Marketing and Advertising](#) (188)
- [News and Media@](#)
- [Organizations](#) (11073)
- [Taxes@](#)
- [Trade](#) (249)
- [Transportation@](#)

POPULAR SITES

Internet | Protected Mode: On 100%

1:29 PM

Training Data – Global Economy

Global Economy in the Yahoo! Directory - Windows Internet Explorer

http://dir.yahoo.com/Social_Science/Economics/Global_Economy/

Secure Search McAfee

Favorites Global Economy in the Yahoo! Directory

SITE LISTINGS By Popularity | Alphabetical (What's This?) Sites 1 - 10 of 10 www.gwu.edu

- [Doing Business](#)
Offers an extensive database of indicators of the cost of doing business around the world by identifying regulations that enhance or constrain business investment, productivity, and growth. From the World Bank.
www.doingbusiness.org
- [Wikipedia: Globalization](#)
Hyperlinked overview of the term used to describe the changes in societies and the world economy.
en.wikipedia.org/wiki/Globalization
- [YaleGlobal Online Magazine](#)
Includes articles and resources discussing globalization issues. Also features books and reviews, academic papers, and related web sites.
yaleglobal.yale.edu
- [About.com: Globalization](#)
Introduction to the concept, pro and anti arguments, articles, and links to related topics regarding trade, human rights, and the environment.
globalization.about.com
- [Globaworks.com](#)
Features an overview of global economy. Includes topics in world economy, globalization, economies of scale, the third world, and NICs.
www.globaworks.com
- [Global 3.0](#)
Examines the current state of globalization. From American RadioWorks.
americanradioworks.publicradio.org/features/global30
- [GlobalCPR](#)
Features information on national income, globalization, international organizations, green economics, and international trade.
www.globalcpr.com
- [Econemisis](#)
Discusses globalization and other factors that impact the global economy. Includes the political and economic risks involved with international trade.
www.econemisis.com
- [Economic Earth](#)
Discusses areas of the global economy, ranging from globalization to economic sanctions.
www.economicearth.com
- [Global Econo](#)
Offers information about global economic trends, international trade, economic theory, and world finance.

www.gwu.edu

[A Global Economy](#)
Find Most Current Info & News About The Economy On Bing.
www.Bing.com

[Buy Textiles Apparel Global Economy](#)
Save up to 90% Used, New & eTexts In Stock for Fast, Free Shipping.
Textbooks.com

[See your message here...](#)

Internet | Protected Mode: On 100%

11:22 AM

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

