Towards Design Principles for Effective Context- and Perspective-Based Web Mining

Vijay K. Vaishnavi
Georgia State University
Computer Information Systems
Atlanta, Georgia 30302-4015
+1.404.413.7381
vvaishna@gsu.edu

Art Vandenberg
Georgia State University
Information Systems & Technology
Atlanta, Georgia 30302-3994
+1.404.413.4743
avandenberg@gsu.edu

Yanqing Zhang
Georgia State University
Computer Science
Atlanta, Georgia 30302-3994
+1.404.413.5733
yzhang@gsu.edu

Saravanaraj Duraisamy
Georgia State University
Computer Information Systems
Atlanta, Georgia 30302-4015
sduraisamy1@student.gsu.edu

ABSTRACT
A practical and scalable web mining solution is needed that can assist the user in processing existing web-based resources to discover specific, relevant information content. This is especially important for researcher communities where data deployed on the World Wide Web are characterized by autonomous, dynamically evolving, and conceptually diverse information sources. The paper describes a systematic design research study that is based on prototyping/evaluation and abstraction using existing and new techniques incorporated as plug and play components into a research workbench. The study investigates an approach, DISCOVERY, for using (1) context/perspective information and (2) social networks such as ODP or Wikipedia for designing practical and scalable human-web systems for finding web pages that are relevant and meet the needs and requirements of a user or a group of users. The paper also describes the current implementation of DISCOVERY and its initial use in finding web pages in a targeted web domain. The resulting system arguably meets the common needs and requirements of a group of people based on the information provided by the group in the form of a set of context web pages. The system is evaluated for a scenario in which assistance of the system is sought for a group of faculty members in finding NSF research grant opportunities that they should collaboratively respond to, utilizing the context provided by their recent publications.

Categories and Subject Descriptors
H.3.4 [**Systems and Software**]: Information networks
H.3.5 (**Online Information Services**]: Web-based services
K.4.3 [**Organizational Impacts**]: Computer–supported collaborative work

DESRIST 2009, May 6-8, 2009, Malvern, PA, U.S.A.

General Terms
Algorithms, Performance, Design, Experimentation, Human Factors, Theory, Verification

Keywords
Web discovery, Web mining, user context, user perspective, ontology, social networks

1. INTRODUCTION
Researchers in all fields of human endeavor including science and engineering recognize the potential and the challenges of the exponential growth of information in the World Wide Web (Etzioni 1996; Kobayashi et al. 2000; Dhyani et al. 2002; Mikroyannidis and Theodoulidis 2007). Taming the Web has spurred considerable research and commercial activity, such as (Brin and Page 1998; Flake et al. 2002; Chen 2003; Thelwall 2005). The available approaches can be broadly grouped into search engines (Brin and Page 1998; Clusty 2008), directories (Gardner and Shepherd 2004; Zeng et al. 2004), and web user adaptation and personalization systems (Gauch et al. 2004; Ferragina and Gulli 2005; Das et al. 2007; Mikroyannidis and Theodoulidis 2007). These approaches, however, have deficiencies in understanding the researcher's context for a web search and presenting results in manner tailored to the user perspective - a category or subcategory (such as "computational fluid dynamics" or "swim initiator neuron") representing the reference or point of view of the user.

This paper describes the initiation of a systematic design research study based on prototyping/evaluation and abstraction (Vaishnavi and Kuechler 2008) using existing and new techniques incorporated as plug and play components into a research workbench. The study investigates an approach called DISCOVERY for using (1) context/perspective information and (2) social networks such as ODP or Wikipedia (http://www.wikipedia.org/) for designing practical and scalable human-web systems for finding web pages that are relevant and meet the needs and requirements of a user or a group of users.

DISCOVERY uses novel granular machine learning methods for web data classification and clustering for the following research path (and its variations):

- Utilizing keywords and/or related context from the user
- Leveraging WordNet (WordNet 2006) and features of social networks (such as ODP or Wikipedia)
- Referencing user profiles (learned from prior interactions with the system)
- Semantically enhancing queries posed to a search engine,
- Visually presenting results in a structured manner (reflecting user needs and requirements) where user's selection among the presented results can be used to further learn the user profile.

The paper describes the current implementation of DISCOVERY and its initial use in finding web pages in a targeted web domain that arguably meets the common needs and requirements of a small group of people (based on a set of context web pages provided by the group). We conduct initial evaluation for a scenario in which the resulting system is used by a group of faculty members to identify possible NSF grant opportunities to which they can respond collaboratively (where their collaboration context is represented by a selection of their recent publications).

The paper is organized as follows: Section 2 outlines the DISCOVERY research approach and the implementation of an initial phase of the project. Section 3 describes the experimental approach used for evaluation and discusses the results. Section 4 discusses related literature in the context of the DISCOVERY approach. Section 5 provides discussion and outlines future work.

2. DISCOVERY APPROACH AND ITS IMPLEMENTATION

2.1 Overview

A system based on the DISCOVERY approach uses the following major subsystems (Figure 1):

a. Context/Perspective Extractor
b. Semantic Query Constructor
c. Results Categorizer/Visualizer
d. Context/Perspective Suggestion Agent (Profile Builder)

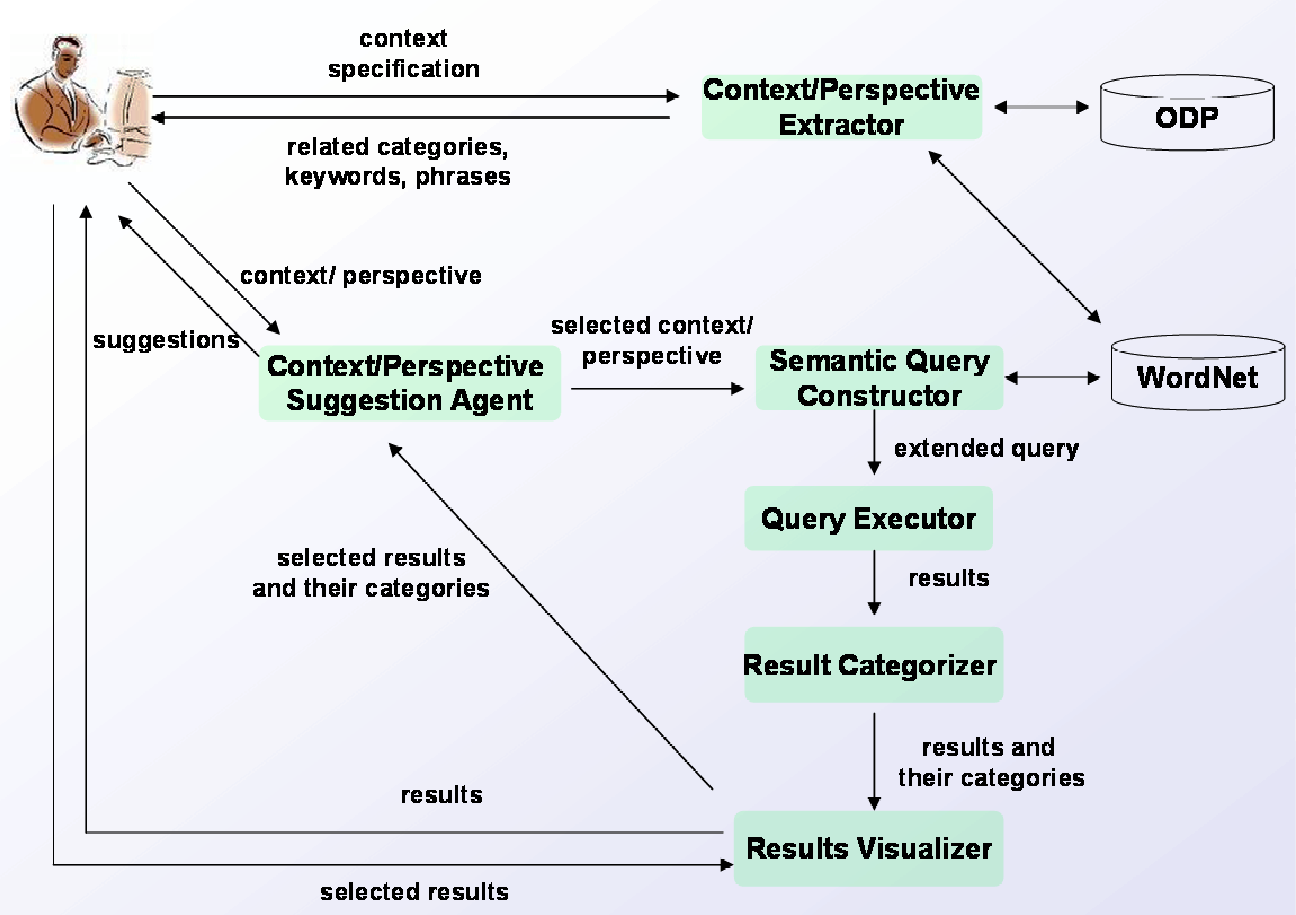

Figure 1: High Level System Flowchart for DISCOVERY

2.1.1 Context/Perspective Extractor

This subsystem extracts context/perspective details from the information provided by the user (keywords, relevant URLs and documents). It is highly likely that all the information provided by the user for a specific query will be around a single theme though we envision options for building customized, personal indices of searched pages.

The challenge is to identify the correct theme (perspective) and present it back to the user in the shortest time possible. To help identify the theme, ODP, as a human-edited hierarchy of categories (ontology) with descriptions, will be used.

A number of algorithms will be explored to improve the theme(s) identification. One approach will be to create a word-topic occurrence matrix and/or phrase-topic matrix. By topic is meant one of the ODP topic categories. The user-provided URLs/documents can be matched against the word/phrase-ODP topic matrices after Singular Value Decomposition (SVD) by identifying the cosine similarities (see Figure 2).

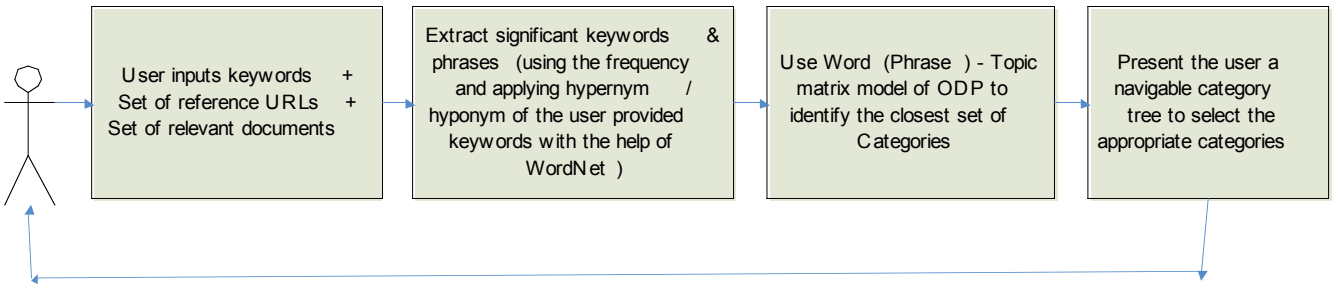

Figure 2: Identifying User Context/Perspective using Word/Phrase-ODP Topic Matrices

2.1.2 Semantic Query Constructor

Semantic query construction is an important part of the architecture for DISCOVERY. This process will be used to disambiguate query terms using WordNet (WordNet 2008). The industrial ontology group in Finland – http://www.cs.jyu.fi/ai/OntoGroup/InBCT_May_2004.html – has done some significant study about enhancing queries. WordNet's hypernyms can be used to provide a more generic representation of the query. Enhanced queries can also be constructed with the help of the most common keywords and phrases in the context information provided by the user and by using the keywords/phrases representative of the perspectives selected/provided by the user. Query refinement with lexicons and ontologies may be explored using a methodology called CONQUER (CONtext-aware QUERy processing) (Storey et al. 2008).

2.1.3 Results Categorizer/Visualizer

Clustering is an unsupervised learning process as opposed to classification. Clustering the documents against the perspectives selected by the user is an interesting problem. Some documents tend to be part of several perspectives (which may show the inter-relation between the perspectives). The main future research activity here is to identify the Best Matching Unit (BMU) between the documents and between documents and perspectives. Each perspective can be characterized with a set of keywords/phrases and their respective calculated distance to the retrieved results and presented to the user. This approach also paves the way to explore certain hidden information, based on the user's discretion. Self-Organizing Maps (SOM) can be used to cluster the result sets. The Kohonen SOM network is very effective for visualization of high-dimensional data (Zhao and Ram 2004). It compresses information while preserving the most important topological and geometric relationships of the primary data elements on the display. The main advantage is to gain insight into the (hidden) structure of data by observing the map, due to the topology preserving nature of SOM. (Bakus et al. 2002) describe the use of phrases for document clustering with SOM. (Amine et al. 2008) use SOM for concept based clustering of textual documents.

2.1.4 Context/Perspective Suggestion Agent (Profile Building)

A significant part of DISCOVERY is building user-profiles by gathering perspective related information over a period of time. The objective of this module (a type of recommender system) is to learn and suggest perspectives for different types of topics. One approach could be to use collaborative filtering algorithms to identify the user perspectives proactively (Das et al. 2007). Once the system is put to use and data is gathered about user activities, further research can be conducted in this area of predicting user interests.

2.2 Current Implementation

The current implementation of a system for DISCOVERY (see Figure 1) implements the first two subsystems – Context/Perspective Extractor and Semantic Query Constructor. Figure 3 details on the flow of events in the current implementation.

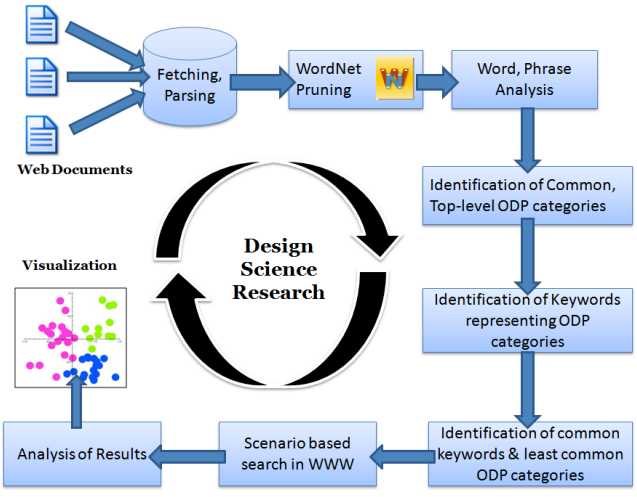

Figure 3: Experimental Prototype – Flow of Events

2.3 System Usage Scenarios

The following scenarios provide a sample of the ways a system implementing the DISCOVERY approach might be used.

Scenario A: A biology professor interested in neuroscience, and specifically in "molluscs," can use the tool as follows (Figure 4):

- **Context Profile:** The user can input keyword ("mollusc") and/or related documents (e.g. research publication URLs). Such related documents act as the user-context.
- The system interacts with the user presenting a set of perspectives drawn from ODP topic categories such as:
  top: recreation: outdoors: scuba diving: underwater life
  top: science: biology: flora & fauna: animalia: mollusca
  top: science: biology: zoology: malacology
  top: business: agriculture & forestry: aquaculture: equipment suppliers: fish farming.
- **Context/Perspective Selection:** The user may choose among suggested perspectives and/or may add his/her own perspective (e.g. "swim initiator neuron").
- The system disambiguates query terms using WordNet and then may semantically enhances the query using an ontology.
- The system executes the query and can classify, filter, and cluster the results based on the perspective(s) identified or suggested by the user.
- **Results Visualization:** The user will manipulate an easy-to-navigate interface with which to identify the result density for a particular perspective and the distance between the perspectives.
- User activities may be captured and recorded as part of an extended activity profile cover an extended period of time.

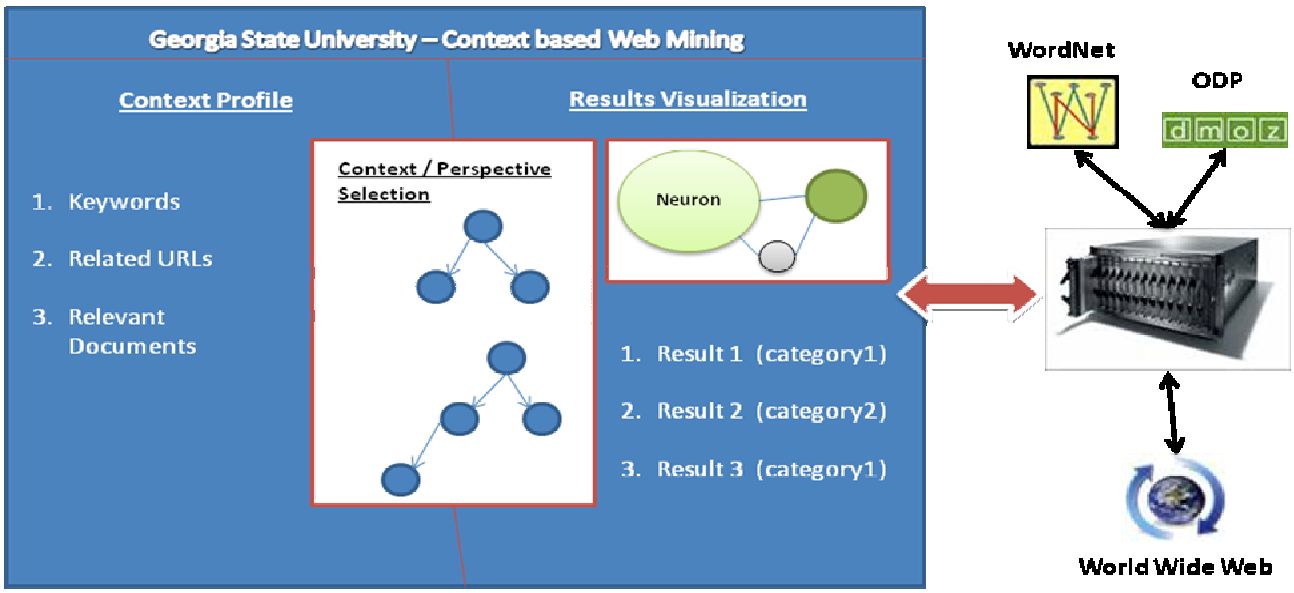

Figure 4: Abstract representation of the conceived system for Context/Perspective based Web Searching

Scenario B: It is desired that potential research collaborations may be automatically discovered based on research interests of a diverse set of faculty. One would like to have a convenient way to compare their research areas, find a potential topic of mutual interest, and then search National Science Foundation for possible granting research funding opportunities; Section 4 evaluates the prototype based on cases of this scenario. The scenario presents a practical application in which a university can suggest collaborative funding opportunities to its faculty based on the information it has on their publications. Some of these interdisciplinary collaborations may be between faculty who do not even know each other and/or are working in different fields.

The prototype may be used as follows:

- **Context Profile:** Sets of abstracts of recent publications are initial input. These abstracts become combined user-context
- The system will parse these abstracts and create a frequency list of keywords, bi-grams, and trigrams.
- The system disambiguates these phrases by processing them against WordNet (to eliminate proper nouns, such as author names) and then using Open Directory Project. The ODP step returns 1 to 5 (top level) categories related to each phrase.
- Phrases that return multiple ODP (top level) categories are deemed to be perhaps ambiguous. Those that return, say, all categories with the same top level (cf. "computers") are judged to be less ambiguous.
- ODP categories identified are used to query *NSF.GOV* domain.
- A user may then search further within these results, for instance searching for "*proposals*" to further refine the returned results.

Results Presentation: The results may be reviewed further by using the prototype tool to create a frequency list of keywords, bi-grams, and trigrams – which keywords, bi-grams and/or tri-grams can then serve as insight into potential topic areas, or be further employed in specific queries within the results set for *NSF.GOV* to discover very specific result pages.

2.4 Evaluation – Strategies for Objective and Subjective Contexts

Evaluation of the research approach and technologies used in implementation may be conducted in a multi-modal manner through experimentation and testing of individual techniques (objective evaluation of algorithms) as well as use of panels of experts and focus groups (subjective evaluations of utility and usability) in the context of the design science (Tremblay et al. 2008). Such evaluations conducted by using the Research Prototype as a design artifact then provide a mechanism (either objective or subjective) for more experimental validation and testing of various project hypotheses.

Commonly used metrics for reliable data classification evaluation include (1) accuracy, (2) sensitivity, (3) specificity, (4) G-Mean (the geometric mean of sensitivity and specificity) and (5) Area Under ROC Curve (AUC-ROC). Similar to AUC-ROC, Area Under Precision/Recall Curve (AUC-PR) can be used to indicate the detection ability of a classifier between precision and recall as a function of varying a decision threshold (Davis and Goadrich 2006). Evaluation metrics used for clustering include recall, precision, and F-Measure (an overall clustering metric) (Van Rijsbergen 1979; Larsen et al. 1999; Stein and Eissen 2002).

3. EXPERIMENTAL APPROACH AND PRELIMINARY RESULTS

We conceptualized and created a prototype (Figure 3) to help validate a potential hypothesis. *We were interested in whether the prototype could identify a common, collaborative theme across user-provided context information (publication abstracts) by leveraging some socially constructed information of the Web to discover a thematic concept. That identified theme could be used to construct a query to mine information from a targeted domain, such as nsf.gov.*

Consider that for a group of researchers it is desired that a common topic of research be found and used to locate possible National Science Foundation funding opportunities. We limited the scope of the prototype to analyzing context information of web documents (in this case URLs containing researchers' combined publication abstracts) and to leveraging information from the Open Directory Project (ODP). ODP (http://www.dmoz.org) is a socially constructed web resource of topic categories maintained by peer-reviewed volunteer experts. The prototype implementation included the parsing of web content to identify meaningful words and phrases (bigrams and trigrams) and using the WordNet library to prune the words that may not be relevant (such as proper names or acronyms). With the keywords and phrases, the prototype then leveraged ODP by submitting each phrase (for this paper the focus was on trigram phrases) as an ODP search string and parsing the suggested ODP topic categories. These topic categories were taken as themes for Google™ queries. Results were analyzed to compare the effectiveness of the prototype's ability to automatically discover relevant query themes from user context information. Refer to Figure 3 for the flow of events.

A trade-off in using WordNet is that perhaps certain relevant terms like "middleware" are eliminated as not relevant for the context information, but the overall efficiency in building the context improved (by, for instance, eliminating author names that tend to be frequently used phrases). The pruned words were analyzed for the presence of phrases (bigrams and trigrams) and their frequencies were counted.

The prototype was set to further analyze the top 10% of keywords and phrases by using Topic Categories of ODP. ODP's wealth of information is primarily organized under categories like Arts, Business, Computers, etc. The prototype created a phrase-ODP_category matrix. For each phrase a theme could be identified by selecting the ODP category that was dominantly occurring. Table 1 shows a simplified phrase-ODP_category matrix.

Table 1. Phrases with Suggested ODP Topic Categories

Trigram Phrase	Category
support vector machines	Computers: Artificial Intelligence: Support Vector Machines
support vector machines	Computers: Artificial Intelligence: Machine Learning
support vector machines	Computers: Artificial Intelligence: Machine Learning: Software
support vector machines	Computers: Artificial Intelligence: Neural Networks: People
support vector machines	Science: Biology: Bioinformatics: Publications
effective information integration	Computers: Data Communications: Software
effective information integration	Business: Business Services: Government Contracting: Contractors
effective information integration	Recreation: Scouting: Resources: Social Issues
metadata key component	Computers: Software: Master Data Management: Articles

To identify the dominant theme the ODP categories for each phrase were compared. A dominant category was one that occurred multiple times and shared a similar sequence of topics. For instance, for the phrase "support vector machines" in Table 1, the category "Computers: Artificial Intelligence: Machine Learning" is dominant, being shared twice, more than others. Effectively, the prototype used ODP Topic Categories as a socially constructed ontology that is searched for the lowest common ancestor in an overall topic digraph structure. Figure 5 is the snapshot of the prototype showing the URL input, the keyword and phrase parsing/analysis, and ODP returned categories with the top level "Computers" flagged.

Prototype trials analyzed the recent research publications of groups of researchers. After performing the logic explained above, the prototype identified candidate ODP Topic Categories representing the combined group research interest. In Tables 2 and 3 are presented the results of two trials. In each case, the results of the semantically enhanced queries are compared to queries that do not use the ODP discovered topic categories. In both cases, after an initial search was done (whether semantically enhanced by the prototype or no), there was a subsequent "search within results" for the additional keyword "proposal" (essentially a perspective refinement).

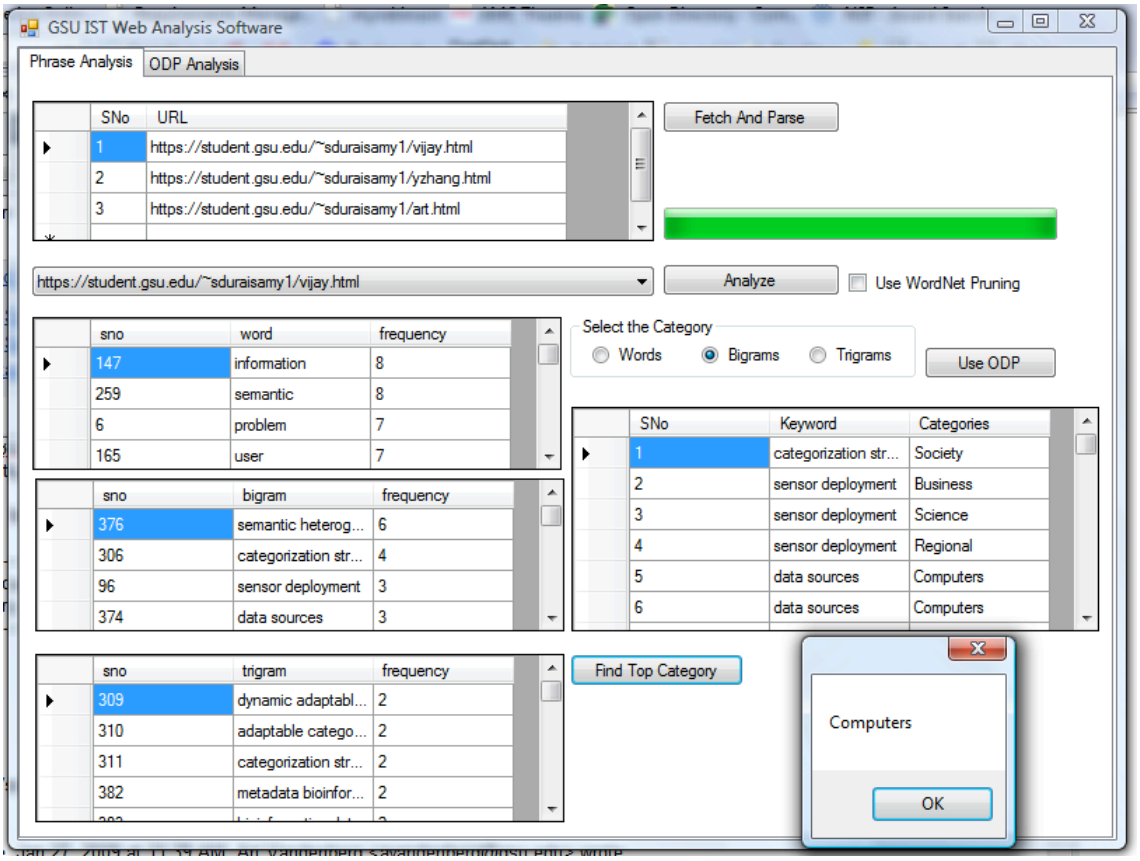

Figure 5: Prototype snapshot

The main challenge was finding a common theme of research collaboration to use as a query, a theme that equally well represented each individual researcher's interest. Finding common ground is important – especially if design principles or automated mechanisms can be found to improve effectiveness of otherwise individually conducted queries.

Tables 2 and 3 compare results of queries using only keywords (trigrams) selected from context documents and queries using the ODP Categories discovered by the prototype using the keywords (trigrams). Results suggest that the additional use of ODP Topic Categories can be more effective in reducing overall results and returning relevant results in the top 10 results.

Table 2. Case #1 – Three Researchers

All searches within domain= nsf.gov	**Google Search Results for Trigram or ODP based category**		***Search Within Results* for "PROPOSAL"**	
Case #1 Three Researchers	**# results**	**# relevant results in first 10 results**	**# results**	**# relevant results in first 10 "search within" results**
TRIGRAMS (context keywords only)	***Results Relevant to Individual***			
effective information integration	3,980	8	21,900 [sic]	8
component effective information	27,200	3	2,220	2
support vector machines	137	1	109	1
metadata key component	413	3	488 [sic]	8
ODP CATEGORY (via context keywords)	***Results Relevant to Group***			
Computers: Artificial Intelligence: Machine Learning (using Algorithm 1)	248	6	158	9

Table 3. Case #2 – Two Researchers

All searches within domain= nsf.gov	**Google Search Results for Trigram or ODP based category**		***Search Within Results* for "PROPOSAL"**	
Scenario #2: 2 Researchers	**# results**	**# relevant results in first 10 results**	**# results**	**# relevant results in first 10 "search within" results**
TRIGRAMS (context keywords only)	***Results Relevant to Individual***			
horizontal gene transfer	252	2	133	7
markov decision processes	197	0	122	0
intrusion detection systems	145	4	124	4
detection systems experience	512	1	612 [sic]	3
ODP CATEGORY (via context keywords)	***Results Relevant to Group***			
Computers: Security: Consultants (using Algorithm 1)	485	1	350	2
Computers: Artificial Intelligence or Computers: Security (using Algorithm 2)	685	5	397	9

Table 4. Results 1-10 (of 158) for Google Search on 1/30/2009

Search Term = Computers Artificial Intelligence Machine Learning site:nsf.gov proposal.

In the 10 top results, NSF Calls for Proposal account for 9, and of these 7 are NSF Information and Intelligent Systems (IIS) (to which the researchers in fact submitted a proposal)

nsf.gov - Computer & Information Science & Engineering (CISE ... / Jun 6, 2008 ... His general research interests lie in machine learning, artificial intelligence, and cognitive neuroscience, and his current research .../ www.nsf.gov/events/event_summ.jsp?cntn_id=110507&org=CISE - 44k - Cached - Similar pages
US NSF - CISE – IIS/ Novel advances in and integration across areas of artificial intelligence, such as machine learning, planning and problem solving, knowledge representation, .../ www.nsf.gov/cise/iis/ri_pgm.jsp - 37k - Cached - Similar pages
Artificial Intelligence & Cognitive Science (nsf03600)/ Mar 3, 2004 ... The Artificial Intelligence and Cognitive Science (AICS) program focuses on ... automated reasoning, machine learning, case-based reasoning, .../ www.nsf.gov/pubs/2003/nsf03600/nsf03600.htm - 48k - Cached - Similar pages
nsf.gov - Funding - Robust Intelligence - US National Science .../ Jul 23, 2008 ... Proposal and Award Policies and Procedures Guide ... and the brain increasingly draws on computer vision, robotics, and machine learning, .../ www.nsf.gov/funding/pgm_summ.jsp?pims_id=503305 - 35k - Cached - Similar pages
US NSF - CISE - IIS – About/ Jan 8, 2009 ... traditions of artificial intelligence, computer vision, human language research, robotics, machine learning, computational neuroscience, .../ www.nsf.gov/cise/iis/about.jsp - 36k - Cached - Similar pages
Computer and Information Science and Engineering/ Robotics and Machine Intelligence-- Supports research fundamental to the design ... topics as artificial intelligence, visualization, database management, .../ www.nsf.gov/nsf/nsfpubs/nsf9491/nsf9491d.htm - 23k - Cached - Similar pages
nsf.gov - Computer & Information Science & Engineering (CISE .../ Proposals and Awards. Proposal and Award Policies and Procedures Guide ... in theoretical computer science, machine learning, artificial intelligence, .../ www.nsf.gov/funding/pgm_summ.jsp?pims_id=503301&org=CISE - 45k - Cached - Similar pages
Chapter 2 - COMPUTER AND INFORMATION SCIENCE AND ENGINEERING/ Robotics and Machine Intelligence--Supports research essential to the design of ... the integration of such topics as artificial intelligence visualization, .../ www.nsf.gov/nsf/nsfpubs/guide/chap2.htm - 22k - Cached - Similar pages
Information and Data Management (nsf04500)/ Basic research in formal models of knowledge and information, machine learning and management of uncertainty is supported in the Artificial Intelligence and .../ www.nsf.gov/pubs/2004/nsf04500/nsf04500.htm - 64k - Cached - Similar pages
Human Language and Communication (nsf03613)/ The HLC program encourages innovative proposals involving computer processing of ... supervised and unsupervised machine learning, corpus-based approaches, .../ www.nsf.gov/pubs/2003/nsf03613/nsf03613.htm - 51k - Cached - Similar pages

Additionally noted in Table 3, the ODP Category results are shown for Algorithm 1 and Algorithm 2. Where Algorithm 1 considered the dominant ODP category as the one that occurred most frequently, Algorithm 2 was a refinement that recognized the condition where several ODP categories might be equally dominant. The revised algorithm used both ODP Topic categories in a search.

We were interested in whether the prototype could identify a common, collaborative theme across user-provided context and scenario information (publication abstracts) by leveraging socially constructed information of the Web. Could that identified theme could be used to construct a query that mined information from a targeted domain, such as nsf.gov? **Our several trials of the prototype suggest that we can accomplish reasonable search results using context to enhance queries. Further, enhancing the query by using ODP's socially constructed categories can deliver even further improvement.**

Table 4 is an example of the result set from an ODP Topic Category enhanced search. These results correspond to the last row in Table 2's ***Search Within Results*** **for "PROPOSAL"** in which 158 results were returned and 9 of the first 10 results were, in fact, NSF calls for proposal. It is noted in particular, that 7 of these 9 were the NSF Information and Intelligent Systems (IIS) program and that the researchers had, indeed, submitted a collaborative proposal to that program recently.

Finally, it is observed that the prototype can post-process the results sets such that all keyword and phrases are identified providing additional insight into phrases. Potentially these phrases can then be processed to discover ODP Topic Categories or further search refinement.

3.1 NSF.GOV as Dataset for Further Experimental Validation

We note that our use of NSF.GOV as a target search domain resulted from our interest identifying a search trial that would have some compelling relevance – that is, finding funding sources well matched to research interests. However, we argue that our design science artifact has provided an interesting value add that we suspected, but are now more confident about and encouraged to pursue further.

That is, we see NSF.GOV as a dataset for conducting rigorous experimental validation of our context- and perspective-based web mining. Consider that NSF.GOV has Abstracts of Recent Awards associated with most of its program areas. It appears to that these are "answers" that can be used to test our query-enhancing prototype.

NSF Awards abstracts may have multiple researchers associated as Principal or Co-Principal Investigators (PIs). We identify several recent publications for each of these PIs and use these publication abstracts as context documents representing the varied research interests of faculty. Our trials treated a several groups of such NSF Award PIs as a candidate collaboration group. The prototype extracted keywords (trigrams for instance) from this combined, group context, and optionally enhanced the keywords using ODP's socially constructed categories. Searches against NSF.GOV suggested that we might indeed discover the very programs to which these PIs were likely to apply, or indeed, from which these PIs did find funding. (See discussion re Table 4 results preceding.)

We foresee design and conduct of a strong experimental validation by recognizing that NSF.GOV's Abstracts of Recent Awards offers a unique pool of data and the potential to generate additional interesting design science activities.

4. RELATED WORK IN THE CONTEXT OF RESEARCH APPROACH AND ITS IMPLEMENTATION

4.1 Information Retrieval

Information retrieval is a highly competitive field, highly demanding, and with continuing research opportunities. Significant research has been conducted in the areas of leveraging context. (Zhu and Dreher 2007) create a special browser that uses ODP to categorize results and disambiguate queries through interaction with users, though user context is not addressed as a pre-process to conducting a search. Data mining methods were used to mine Web search logs (context) to better present search results (Wang and Zhai 2007).

(Zakos and Verma 2005) describe a novel technique to dynamically generate a context-based measure of document term significance during retrieval that supplements term frequency measure. Such context matching significantly improves retrieval results, and would suggest that such a technique, when coupled with user supplied context documents (such as we propose) can especially improve results for a specific user request. (Stefanidis et al. 2007) describe user personalization model that depends on context, such as user location. They argue that relaxing context constraint (such as replacing a hierarchical attribute by one at a higher level) may improve performance results. We foresee that DISCOVERY approach further leveraging a similar effect by interacting with the user so that they themselves can select among suggested context constraints, directly evaluate results, and refine their personalization model.

(Li et al. 2007) discuss relation-based searches based on the Semantic Web, implementing an instance of semantic search using RDF relation tuples to provide improved results. However, they acknowledge their dependence on next generation Semantic Web and the broad implementation of RDF. We believe that DISOVERY approach employing user provided context information can also provide improvements by relating the context data to the search results – and not requiring any a priori annotation (such as expected with RDF.)

Investigations can be made to suggest the perspectives and result-sets for a query belonging to a particular domain. Well-established concepts like "collaborative filtering" can be applied (Das et al. 2007; Harpale and Yang 2008). (Wang et al. 2007) describe the use of topical n-grams, rather than relying only on a bag-of-words approach to discover topics and topical phrases. DISCOVERY approach makes similar use of n-grams and ordered phrases, applying it additionally at the pre-query stage when interacting with the user to establish context for a proposed search, as well as subsequently when presenting results and offering discovered topics as options for user perspective.

(Smucker et al. 2008) consider how human question answering systems may improve retrieval results when including an interactive component; since human interaction can be variable, better solutions will combine benefits of precise information retrieval and flexible question answering systems. We suggest that

our approach not only seeks to engage the user, but also leverages supplied context information to prompt or automatically suggest user responses – thereby offsetting the downside (i.e. variable interaction) of reluctant users. (Lin and Smucker 2008) further explore their hypothesis that a content-similarity browsing tool can compensate for poor retrieval results. They evaluate this concept with PubMed, which provides related-links function that serves to prompt or encourage the user in their browsing. DISCOVERY approach extends this concept in several ways, by using user supplied context as a pre-matching element that guides an automated search, as well as a post-search mechanism when browsing results and potentially leveraging user-selected content (from initial context specification, or when selecting among results) as a content-similar "reference set" presented in a browsing mode.

4.2 Ontology

As early as 1995, the importance of ontology and ontology structure for efficient information integration was discussed in (Farquhar et al. 1995). The significance of ontology in Enterprise modeling has been examined in (Uschold et al. 1998). With the Web being a large dataset involving information about virtually every aspect of human endeavor, it may not be practically possible to construct and follow a well-defined ontology. Experiments have been conducted to enable automated ontology learning from domain text using Natural Language Processing and Machine Learning techniques (Navigli et al. 2003). Significant efforts have been made to organize the information available in the Web using directory structure (e.g. ODP), a form of knowledge network (Contractor and Monge 2002). ODP, as of now, has categorized 4,576,062 sites into 590,000 categories and is considered to be the largest and most comprehensive human-directory on the web. Social bookmarking and Folksonomies have also gathered momentum in classifying and tagging the publicly available Web information (Mathes 2004) but they need to handle semantic heterogeneity. (Wang and Taylor 2007) describe an algorithm to extract concept forests from a document with the assistance of natural language ontology, the WordNet database. (Wei and Croft 2007) investigate retrieval performance using manually built topic models derived from a handcrafted directory resource (socially constructed ODP).

These social network efforts, however, remain challenged by the exponential growth and scale of the Web – Google reported processing 1 trillion unique web links as of July 2008 (Google 2008). DISCOVERY approach differentiates itself by seeking automated discovery of ODP (or other social network initiatives) topic models and using these models as extant ontology resources. Unlike traditional ontologies, ODP and other social networks such as Wikipedia are perhaps better suited to the evolving nature and inherent ontological drift (Vaishnavi and Kuechler 2005) of Web data.

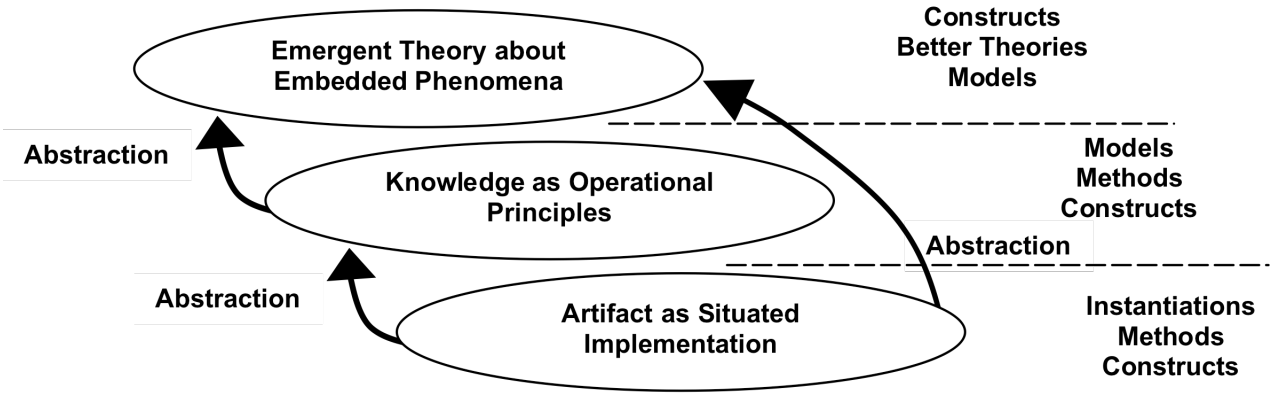

Figure 6. Outputs of Design Science Research (Purao 2002)

5. CONTRIBUTION AND FUTURE RESEARCH

The work presented in this paper follows the design science research paradigm (Montgomery and Runger 1999; Hevner et al. 2004). Building research artifacts and evaluating the same to test for feasibility, effectiveness, and efficiency, and abstracting the knowledge gained (see Figure 6) in terms of design principles and theories (Kuechler and Vaishnavi 2008) are among the important research activities in design science research (Vaishnavi and Kuechler 2008). We start with a general approach, DISCOVERY, and carry out its situated implementation for carrying out explorations of classes of web mining tasks. Experimentation and evaluation of the resulting artifact aims to find general design principles and mid-range web mining theories (Kuechler and Vaishnavi 2008).

The results reported in the paper point to the potential of the DISCOVERY approach, following design science methodology, to advance knowledge. Future work will complete the research study based on the DISCOVERY approach. It will also develop a rigorous experimental framework for evaluating a system based on the approach (using NSF.GOV as a source of experimental data), and generalizing the system.

REFERENCES

[1] Amine, A., Elberrichi. Z., Bellatreche, L., Simonet, M. and Malki, M., "Concept-based clustering of textual documents using SOM*," Proc. IEEE/ACS International Conference on Computer Systems and Applications (AICCSA 2008)*, 2008.

[2] Bakus, J., Hussin, M. F. and Kamel, M., "A SOM based document clustering using phrases," *Proc. 9th International Conference on Neural information Processing (ICONIP'OZ)*, Vol. 5, 2002.

[3] Brin, S. and Page, L., "The Anatomy of a Large-Scale Hypertext Web Search Engine," *Proc. 7th WWW Conference*, 1998.

[4] Chen, H., "Web Retrieval and Mining," *Decision Support Systems* (35) pp. 1-5, 2003.

[5] Clusty, http://clusty.com/, last accessed on December 16, 2008.

[6] Contractor, N. S. and Monge, P. R., "Managing knowledge networks." *Management Communication Quarterly* 16(2), pp. 249-258, 2002.

[7] Das, A., Datar, M., Garg, A. and Rajaram, S., "Google News Personalization: Scalable online collaborative filtering," *Proc. WWW 2007 / Track: Industrial Practice and Experience*, 2007.

[8] Davis, J. and Goadrich, M., "The relationship between Precision-Recall and ROC curves," *Proc. 23rd ICML*, pp. 233-240, 2006.

[9] Dhyani, D., Ng, W K. and Bhowmick, S. R., "A Survey of Web Metrics," *ACM Computing Surveys* (34:4), pp. 469-503, 2002.

[10] Etzioni, O., "The World Wide Web: Quagmire or Gold Mine," *Communications of the ACM*, pp. 65-68, 1996.

[11] Farquhar, A., Fikes, R., Pratt, W. and Rice, J., "Collaborative Ontology Construction for Information Integration," Technical Report, Stanford University, 1995.

[12] Ferragina, P. and Gulli, A., "A Personalized Search Engine Based on Web-snippet Hierarchical Clustering," *Proc. 14th International World Wide Web Conference*, Chiba, Japan, pp. 801-810, 2005.

[13] Flake, G W., Lawrence, S., Giles, C L. and Coetzee, F M., "Self-Organization and Identification of Web Communities," *IEEE Computer* (35:3), pp. 66-71, 2002.

[14] Gardner, D. and Shepherd, G.M., "A gateway to the Future of Neuroinformatics," *Neuroinformatics* (2:3), pp. 271-274, 2004.

[15] Gauch, S., Chafee, J. and Pretschner, A., "Ontology-based personalized search and browsing," *Web Intelligence and Agent Systems* (1:3-4), pp. 219-234, 2004.

[16] Google, "We knew the web was big…", The Official Google Blog, 2008, http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html, last accessed Dec 14, 2008.

[17] Harpale, A. S. and Yang, Y., "Personalized Active Learning for Collaborative Filtering," *Proc. SIGIR'08*, Singapore, July 20–24, 2008.

[18] Hevner, A., March, S., Park, J. and Ram, S. "Design Science in Information Systems Research," *MIS Quarterly* (28:1), pp. 75-105, 2004.

[19] Kobayashi, M. and Takeda, K., "Information Retrieval on the Web," *ACM Computing Surveys* (32:2), pp. 144-173, 2000.

[20] Kuechler, B. and Vaishnavi, V., "On Theory Development in Design Science Research: Anatomy of a Research Project," *European Journal of Information Systems*, Vol. 17, No. 5, pp. 489-504, 2008.

[21] Larsen, B. and Aone, A. "Fast and Effective Text Mining Using Linear-time Document Clustering," *Proc. Fifth ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining*, pp. 16-22, 1999.

[22] Li Y., Wang Y. and Huang X., "A Relation-Based Search Engine in Semantic Web," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 2, pp. 273-282, Feb. 2007.

[23] Lin, J. and Smucker M. D., "How Do Users Find Things with PubMed? Towards Automatic Utility Evaluation with User Simulations," Technical Report LAMP-TR-148/HCIL-2008-07, University of Maryland, College Park, February 2008.

[24] Mathes, A., "Folksonomies - Cooperative Classification and Communication through Shared Metadata," University of Illinois Urbana-Champaign, December 2004.

[25] Montgomery, D.C. and Runger, G.C. *Applied Statistics and Probability for Engineers*, John Wiley & Sons, Inc., 1999.

[26] Mikroyannidis, A. and Theodoulidis, B., "Heraclitus: A Framework for Semantic Web Adaptation," *IEEE INTERNET COMPUTING*, pp. 45-52, 2007.

[27] Navigli, R., Velardi, P. and Gangemi, A., "Ontology Learning and Its Application to Automated Terminology Translation," *IEEE Intelligent Systems*, Vol. 18, No. 1, pp. 22-31, 2003.

[28] Purao, S., "Design Research in the Technology of Information Systems: Truth or Dare." Georgia State University Department of CIS Working Paper 2002.

[29] Smucker M D., Allan J. and Dachev B., "Human Question Answering Performance using an Interactive Information Retrieval System," Technical Report IR-655, Center for Intelligent Information Retrieval (CIIR), Department of Computer Science, University of Massachusetts Amherst, January 2008.

[30] Stefanidis, K., Pitoura, E. and Vassiliadis, P., "On Relaxing Contextual Preference Queries," *Proc. International Conference on Mobile Data Management*, 2007.

[31] Stein, B. and Eissen, S.M.Z. "Document Categorization with Major CLUST," *Proc. 12th Annual Workshop On Information Technologies And Systems (WITS'02)*, Barcelona, Spain, 2002.

[32] Storey, V. C., Burton-Jones, A., Sugumaran, V. and Purao, S., "CONQUER: A Methodology for Context-Aware Query Processing on the World Wide Web," *Information Systems Research*, Vol. 19, No. 1, pp. 3-25, March 2008.

[33] Thelwall, M., "Scientific Web Intelligence: Finding Relationships in University Webs," *Communications of the ACM*, pp. 93-96, 2005.

[34] Tremblay, M. C., Hevner, A. R. and Berndt, D. J., "The Use of Focus Groups in Design Science Research," *Proc. Third International Conference on Design Science Research in Information Systems and Technology*, V. Vaishnavi & R. Baskerville, Eds., Atlanta, Georgia, 2008.

[35] Uschold, M., King, M., Moralee, S. and Zorgios, Y., "The Enterprise Ontology," *Knowledge Engineering Review,* pp. 31-89, 1998.

[36] Vaishnavi, V. K. and Kuechler, W L., *Design Science Research Methods and Patterns: Improving and Innovating Information and Communication Technology*, Auerbach Publications, Taylor & Francis Group," New York, NY, 2008.

[37] Van Rijsbergen, C. "Information Retrieval," Butterworth, London, 1979.

[38] Wang, X., McCallum, A. and Wei, X., "Topical N-grams: Phrase and Topic Discovery, with an Application to Information Retrieval," *Proc. Seventh IEEE International Conference on Data Mining*, 2007.

[39] Wang, J. Z. and Taylor, W., "Concept Forest: A New Ontology-assisted Text Document Similarity Measurement Method - J Wang, W Taylor," *Proc. IEEE/WIC/ACM International Conference on Web Intelligence*, 2007.

[40] Wang X. and Zhai, C., "Learn from Web Search Logs to Organize Search Results," *Proc. 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'07)*, pp. 87-94, 2007.

[41] Wei, X. and Croft, W. B., "Investigating Retrieval Performance with Manually-Built Topic Models," 2007.

[42] WordNet, Cognitive Science Laboratory, Princeton University, Princeton, New Jersey, 2006. http://wordnet.princeton.edu/, last accessed on December 12, 2008.

[43] Zakos, J. and Verma, B., "A Novel Context Matching Based Technique for Web Document Retrieval," *Proc. Eighth International Conference on Document Analysis and Recognition*, 2005.

[44] Zeng, H J., He, Q C., Chen, Z., Ma, W Y. and Ma, J W., "Learning to Cluster Web Search Results," *Proc. 27th annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM Press, Sheffield, United Kingdom, pp. 210-217, 2004.

[45] Zhao, H. and Ram, S. "Clustering Schema Elements for Semantic Integration of Heterogeneous Data Sources," *Journal of Database Management* (15:4), pp. 88-106, 2004.

[46] Zhu, D. and Dreher, H., "Determining and Satisfying Search Users Real Needs via Socially Constructed Search Concept Classification," *Proc. Inaugural IEEE International Conference on Digital Ecosystems and Technologies (IEEE DEST 2007)*, 2007.