COVER SHEET FOR PROPOSAL TO THE NATIONAL SCIENCE FOUNDATION

PROGRAM ANNOUNCEMENT/SOLICITATION NO./CLOSING DATE: NSF 07-577 12/10/07  
FOR CONSIDERATION BY NSF ORGANIZATION UNIT(S): IIS - INFO INTEGRATION & INFORMATICS  
DATE RECEIVED NUMBER OF COPIES DIVISION ASSIGNED FUND CODE DUNS# (Data Universal Numbering System) FILE LOCATION  
002257350  

EMPLOYER IDENTIFICATION NUMBER (EIN) OR TAXPAYER IDENTIFICATION NUMBER (TIN)  
150532081  
SHOW PREVIOUS AWARD NO. IF THIS IS A RENEWAL OR AN ACCOMPLISHMENT-BASED RENEWAL  
IS THIS PROPOSAL BEING SUBMITTED TO ANOTHER FEDERAL AGENCY? YES ☐ NO ☐ IF YES, LIST ACRONYM(S)  

NAME OF ORGANIZATION TO WHICH AWARD SHOULD BE MADE  
Syracuse University  
AWARDEE ORGANIZATION CODE (IF KNOWN)  
0028829000  
NAME OF PERFORMING ORGANIZATION, IF DIFFERENT FROM ABOVE  
ADDRESS OF PERFORMING ORGANIZATION, IF DIFFERENT, INCLUDING 9 DIGIT ZIP CODE  
PERFORMING ORGANIZATION CODE (IF KNOWN)  

IS Awardee Organization (Check All That Apply) ☐ SMALL BUSINESS ☐ FOR-PROFIT ORGANIZATION ☐ MINORITY BUSINESS ☐ WOMAN-OWNED BUSINESS ☐ IF THIS IS A PRELIMINARY PROPOSAL THEN CHECK HERE  

TITLE OF PROPOSED PROJECT  
III-COR - Small: User-centered design for automated genre recognition  
REQUESTED AMOUNT PROPOSED DURATION (1-60 MONTHS) REQUESTED STARTING DATE SHOW RELATED PRELIMINARY PROPOSAL NO. IF APPLICABLE  
$449,023 24 months 07/01/08  
CHECK APPROPRIATE BOX(ES) IF THIS PROPOSAL INCLUDES ANY OF THE ITEMS LISTED BELOW  
☐ BEGINNING INVESTIGATOR (GPG I.G.2) ☐ HUMAN SUBJECTS (GPG II.D.6) Human Subjects Assurance Number ☐ INTERNATIONAL COOPERATIVE ACTIVITIES: COUNTRY/COUNTRIES INVOLVED (GPG II.C.2.j)  
☐ DISCLOSURE OF LOBBYING ACTIVITIES (GPG II.C) ☐ PROPRIETARY & PRIVILEGED INFORMATION (GPG II.D.1) ☐ SMALL GRANT FOR EXPLOR. RESEARCH (SGER) (GPG II.D.1) ☐ VERTEBRATE ANIMALS (GPG II.D.5) IACUC App. Date ☐ HIGH RESOLUTION GRAPHICS/OTHER GRAPHICS WHERE EXACT COLOR REPRESENTATION IS REQUIRED FOR PROPER INTERPRETATION (GPG I.G.1)  
PHS Animal Welfare Assurance Number  

PI/PD DEPARTMENT  
School of Information Studies  
PI/PD POSTAL ADDRESS  
Syracuse University Information Studies  
348 Hinds Hall  
Syracuse, NY 132444100  
United States  
PI/PD FAX NUMBER  
315-443-5806  
PI/PD NAME  
Kevin Crowston PhD 1991 315-443-1676 crowston@syr.edu  
Barbara Kwasnik PH.D. 1989 315-443-4547 bkwasnik@mailbox.syr.edu  
Nancy J McCracken PhD 1979 315-443-5484 njm@ecs.syr.edu  

Page 1 of 2
Project Summary: III-COR: Improving information-access for large digital collections through automated identification of document genre

The goal of this project is to improve information-access systems for large digital collections by incorporating metadata about documents’ genres/text types. The project starts from the observation that humans rarely have to read every word of a document to understand its meaning. Instead, they take a shortcut: they first identify the purpose of the document they are faced with, and then use the document in an appropriate way, that is, they start by assessing a document’s genre. For example, a grant proposal is used differently from a syllabus, a product brochure or a bank statement. Information-access systems would be more useful if they could similarly distinguish the purpose of documents. However, with the current state of the art, computerized systems generally cannot determine the social purposes of documents. Rather, a system can detect regularities of form, what some authors have labelled text type to distinguish it from genre. Text type and genre are closely related but still distinct, so an initial goal of this investigation is to discover a set of text types that can both be reliably automatically recognized and that correspond to genres useful to users’ information access and then to build a system that uses automatic recognition of these genres/text types to improve information access.

To develop the prototype system, the project includes a series of human-centered studies of automated classification of genres/text types for the development and testing of prototype, genre-enhanced information access tools. Specifically, a series of experiments will be conducted to determine a small set of genres that are useful for supporting an information-access task. On the NLP side, hierarchical clustering will be used to explore an already developed hierarchy of genres, using Self-Organizing Feature Maps (SOM) and the K-means method, to determine whether documents naturally fall into an appropriate number of clusters representing text types. Different classification algorithms will be used to cluster documents on text types. The goal of these interrelated studies is to find the set of useful genres that are matched to text types that can be reliably recognized using natural language processing (NLP) techniques. The project will then develop and test a prototype information-access system based on this design.

Expected intellectual contributions. The proposed study will make several contributions. Most fundamentally, it will provide a linkage between what until now have been two nearly disconnected bodies of research, namely the human use of genre and machine recognition of text type. On the one hand, research on automatic genre detection has used a variety of taxonomies of genre, but without explicitly examining their utility for supporting user information access. On the other hand, research on users and genre has not until now been linked explicitly to automatic classification. We expect the proposed research to provide a useful interchange between these two areas of research, thus illuminating each. More concretely, we will provide a small set of genres/text types that are useful and feasible, a set of algorithms for classifying documents into these text types and a prototype system that uses automatic classification to support information access.

Expected broader impacts. The project will have three kinds of broader impacts.

1. The project will tackle a significant problem facing users in their interactions with large digital collections, thus advancing knowledge and understanding within information science, and contributing to the development of the field;
2. The project will develop a prototype, genre-enhanced system, thus potentially contributing to society in general; and
3. The project includes students in the research, thus contributing to learning and education.

Keywords: Genre; text type; machine learning; information access
<table>
<thead>
<tr>
<th>Section</th>
<th>Total No. of Pages</th>
<th>Page No.* (Optional)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cover Sheet for Proposal to the National Science Foundation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Project Summary (not to exceed 1 page)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Table of Contents</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Project Description (Including Results from Prior NSF Support) (not to exceed 15 pages) (Exceed only if allowed by a specific program announcement/solicitation or if approved in advance by the appropriate NSF Assistant Director or designee)</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>References Cited</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Biographical Sketches (Not to exceed 2 pages each)</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Budget (Plus up to 3 pages of budget justification)</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Current and Pending Support</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Facilities, Equipment and Other Resources</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Special Information/Supplementary Documentation</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Appendix (List below.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Include only if allowed by a specific program announcement/ solicitation or if approved in advance by the appropriate NSF Assistant Director or designee)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appendix Items:</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Proposers may select any numbering mechanism for the proposal. The entire proposal however, must be paginated. Complete both columns only if the proposal is numbered consecutively.
III-COR: Improving information-access for large digital collections through automated identification of document genre

We seek to improve information-access systems for large digital collections by incorporating metadata about documents’ genres/text types. To this end, we propose a set of user and natural language processing (NLP) experiments leading to the design and testing of a genre-enabled information access system. The project is designed to address a significant problem in information access, namely that topic alone is not enough to define an information problem. Different users may require different solutions for seemingly similar topics because the situation (or context) of the user determines not only what topics are requested and what strategies are invoked in searching and evaluating output, but also what types of resources are considered relevant and useful. For example, “methods for learning mathematics” (a topic) will be construed differently by a student, by a parent and by a classroom teacher because of their different information-use situations. Indeed, even the same user may require different information at different times. A university professor looking for information about a branch of mathematics for the class that she teaches would most likely be interested in documents of educational genres (e.g., syllabi, assignments, class notes) but when working on a research paper in the area, the same professor would more likely appreciate scholarly works (e.g., conference papers, annotated biographies, NSF program announcements). The relevant documents for these two searches would be quite different, even though the topic and query keywords might be nearly the same. Although we know that it is important to understand the situation of the user, the actual representation of situations and then their implementation in a system remains a difficult problem. Research efforts to create user profiles, universal situation grammars, and so on suffer from limitations of scope to specific domains and lack of extensibility and flexibility.

We propose and seek to test an alternative approach, using genre/text type information as a non-topical characteristic of documents that signals their purpose. We note that humans rarely have to read every word of a document to understand its meaning. Instead, they start by identifying the purpose of each document they are faced with, and then use the document in an appropriate way, that is, they start by assessing a document’s genre. For example, a grant proposal is used differently from a syllabus, a product brochure or a bank statement. Each genre of document has a socially recognized purpose that informs its use. For example, because this proposal conforms to the grant-proposal genre, a reviewer can quickly determine the purpose of our communication and its likely relevance to their task (high for reviewers, but lower for researchers looking for the results of completed studies), locate relevant sections, evaluate the document’s contribution and use it to prepare reviews.

Information-access systems would be more useful if the systems could similarly distinguish the purpose of documents (that is, their genres). Explicit identification of genre seems likely to be particularly important for large digital collections because—unlike earlier collections of documents comprising a limited set of genres (e.g., a document database containing primarily journal articles)—these collections often contain documents with a diversity of genres (e.g., journal articles but also magazine articles, hot lists, memos, home pages, class syllabi, etc.). A user searching such a diverse document collection by topic will likely receive some documents of relevant genres along with many documents of irrelevant genres—a low precision result—even if all retrieved documents conform to search specifications regarding the topical content of the document. Indeed, one approach to developing a useful web search tool is to restrict it to documents to a particular genre, such as Indeed for job listings, CiteSeer and Google Scholar for academic articles and various Google specialized searches for blogs, books, business addresses, news articles, patents, product sales pages and source code (but excluding those specialized by media, e.g., searches for images or video). Yates and Sumner [87] argue that on the Web, genres help in both the production and consumption of documents because genre adds “fixity” in a medium that does not otherwise distinguish well between text types (say, a book and a post-it).
With the current state of the art, however, computerized systems have great difficulty determining the social purposes of documents. A system can more easily detect regularities of form, what some authors have labelled text type to distinguish it from genre (genre being defined primary by social purpose and only secondarily by form). There is a close association between genre and text type, purpose and form—indeed, some authors define genre as regularities of form and purpose—but a text type and a genre may not be identical. For example, memo is a text type because of its strong regularities of form, but may not be a genre in the strict sense because it can be used for many different communicative purposes. On the other hand, love letter is a genre since it has a socially recognized purpose, but can take many forms, and thus may not be regular enough to be a recognizable text type. Our study will add to the growing body of research on genre and text type by studying their relation as a possible tool in information retrieval, specifically, to improve information access to large digital collections.

Towards this end we propose a series of human-centred studies of automated classification of genres/text types for the development and testing of prototype, genre-enhanced information access tools. Specifically, we will carry out a series of experiments to determine which genres are simultaneously useful for supporting an information-access task and are matched to text types that can be reliably recognized using NLP techniques. We will then develop and test a prototype information-access system based on this design. That is, our goal at the most fundamental level is to identify genres/text types useful to humans in information-retrieval tasks and also identifiable automatically using their attributes as text types.

Results from prior work

The proposed research builds on prior work by two of the PIs, Crowston and Kwasnik, who have been funded by an NSF grant, IIS 04-14482, entitled “How can document-genre metadata improve information access for large digital collections?” In this study, we explored how genre metadata can best be utilized in information-access tasks. By “best” we meant initially improving users’ performance (e.g., time, accuracy or perceived usability) in information searching, filtering and evaluation tasks. Our overarching question was: would identification of document genres improve information access technologies in large digital collections such as digital libraries and the Web? The project has had several outputs that will be useful for the currently proposed study, specifically a taxonomy of genres and a coded corpus of web pages that will be used for experiments and for training machine-learning algorithms, as well as an experimental prototype information access system that will be used in the research proposed here.

Our first goal in the prior study was to develop a better understanding of the use of genre in information access tasks and more specifically to develop a human-centred taxonomy of genres. We tried first to construct the taxonomy through a naturalistic field study. Because genres are situated in a community’s language and work processes, we felt it was important to learn about genres from people engaged in real tasks and in their own words. We solicited information about the genre of web pages from three groups of respondents (17–20 of each): K12 teachers, journalists and engineers, professionals for whom genre information might be useful—indeed necessary—for determining whether a given web page is relevant to their needs. We used think-aloud technique to capture the search goals and general strategy, followed by a debriefing. During the interview, for every page visited we asked: “What is your search goal? What type of web page would you call this? What is it about the page that makes you call it that? Was this page useful to you? How so (or why not)?” We recorded the URLs of the pages in the sequence in which it was visited, which we used to capture screenshots of each web page visited by the respondents. We collected nearly 1,000 screenshots of web pages visited by respondents, each accompanied by its original URL, along with transcribed audio recordings of the think aloud and debrief
sessions. Information supplied by respondents included participant-supplied genre terms, clues they used for identifying the genres and overall information on the task associated with the genre.

While our analysis yielded a repertoire of several hundred unique genre labels and a confusing number of genre attributes, it did not provide a suitable base for building the taxonomy we wanted. There were several challenges in analyzing this data. The first was the problem of identifying the genres themselves, identifying the genre unit, and eliciting unambiguous and non-contradictory genre labels. Respondents had difficulty distinguishing form and content and identifying the purpose of a page. When respondents described a web page, it was not always clear whether it was the page in toto, or some particular aspect or component of the page. Conversely several genre terms were sometimes applied to the same document. Frequently, different types of pages were labelled with the same term, even by the same person. Often a respondent would lack a term for a given genre and instead offer terms that were general, unspecific, idiosyncratic or vague. Another problem was eliciting genre attributes, since respondents could rarely be specific about how they knew that a page was of a particular genre. We did learn, however, that people do in fact use genres to distinguish between one site and another, but that in general, they have a difficult time naming the genres consistently, and they were not able to describe features in any way that revealed the kind of regularity we might have needed in order to build a useful faceted taxonomy.

Indeed, difficulties in precise recognition and naming of genre based on user input are reported in almost all studies dealing with the elicitation of genre information from participants. In order to move forward, researchers make accommodations and adjustments to genre inventories and coding [e.g., 76]; conduct content analysis of the genres themselves [e.g., 29, 41, 75]; or simply select a set of plausible genres on which to test their ideas [e.g., 60]. We therefore moved to the next phase of our study to expand the list of genre terms harvested in phase 1. We did this by working backwards from specific tasks to task-related genre. We constructed a set of 14 tasks that we believed might benefit from the use of genre in disambiguating search results (e.g., “Find a web page that names the person who invented the punch card”, for which distinguishing between product pages, manuals and encyclopedia articles could be helpful). We then searched for potentially useful web documents for these tasks (e.g., using the search string “punch card” for the task above) and retrieved approximately 200 pages for each, using the search engine clusty.com. For each page, we stored a spidereed copy of the page, the short summary presented by clusty.com in the search results and a keyword cluster assigned by clusty.com. Pages were then coded for genre. Initially we used the genre list from phase 1, but the codebook was evolved inductively based on the pages encountered. Inter-coder reliability and codebook refinement was achieved by talking to consensus on a sample of pages until 80% agreement was achieved. The result of this study is a taxonomy of 133 genre terms, currently arranged in a shallow hierarchy, and a corpus of 2564 web pages coded by these genres. This set of genres and the corpus will be key to carrying out our proposed research.

We are currently using the corpus in an experiment exploring the utility of genre meta-data, by comparing performance on the information seeking tasks when search results are clustered by keywords (i.e., using the cluster created by the clusty.com search engine) vs. by genre. Our current genre-based search interface prototype is shown in Figure 1. The studies included in this proposal are an extension of the prior work, as they add a focus on automatic recognition of text type and explore the relation between useful genres and feasible text types.

Overview of the proposed project

The goal of the proposed study is to develop a prototype, genre-enhanced information access system. The work will be carried out in two phases over two years, first, determining a set of genres/text types to be automatically recognized and then building and testing a prototype system.
In year 1 we will determine a set of genres that both are useful for supporting information access tasks and that correspond to text types that can be reliably detected using automated techniques. Our current genre hierarchy has far too many genres (133) to be feasible for automatic detection. We will therefore rerun the user studies testing a reduced taxonomy to determine a smaller set of useful genres. If time permits, we will also explore alternative user interfaces for presenting or making use of genre information.

In parallel, we will use the already coded pages from our corpus to train machine learning algorithms and experiment to determine which text types can be reliably detected. We anticipate that some of the genres in our current taxonomy will not correspond neatly to detectable text types and contrariwise, some easily-detectable text types will be of limited use in supporting information access, requiring some amount of back and forth to settle on a final set of text types/genres. The results of this year’s work will be a small set of genres that have been shown to be useful in supporting information access and a set of algorithms that can reliably detect documents of those genres/text types.

In Year 2, we will build and test a prototype information access system based on the genre/text types and user interface approaches identified in year 1. We will then use this system in a final round of user studies to determine its usefulness in supporting information access. In parallel, we will carry out follow-on studies that extend the results of year 1’s work.

Figure 1. Genre-based search interface experimental prototype, showing clustering of search results by genre on the left, and retrieved pages on the right. The experimental task is at the top.
Expected intellectual contributions of the proposed work

The proposed study will make several contributions. Most fundamentally, it will provide a linkage between what until now have been two nearly disconnected bodies of research, namely the human use of genre and machine use of text type. In particular, research on automatic genre detection has used a variety of taxonomies of genre, but without explicitly examining their utility for supporting user information access. On the other hand, research on users and genre has not until now been linked explicitly to automatic classification. We expect the proposed research to provide a useful interchange between these two areas of research, thus illuminating each. More concretely, we will provide a small set of genres/text types that are useful and feasible, a set of algorithms for classifying documents into these text types and a prototype information access system.

Theory: Genre and text type of digital documents

Before presenting the design of our research study in more detail, in this section we review prior work on genre, text type and digital documents to provide a theoretical grounding for the proposed work. We start with a discussion of genre of digital documents then review work on automatic identification of text type.

Genres and digital documents

Genres define a class of similar documents and such a categorization can be made on a variety of bases. Different approaches to defining genre have focused on different aspects of documents [e.g., 63, 69, 82] [see 6 for a review of the development of the concept]. In this proposal, we define genre following Campbell and Jamieson [8], who focused on the communicative purpose of the documents. They suggested that genres arise as responses to recurrent communicative situations: given a recognized need to communicate [i.e., a purpose, 4], individuals will express similar social motives, themes and topics in their communication, that is, they will communicate in a recognized genre. Miller [69] argued that the process of creating genre is one of typification, as similarities in situations are recognized. If the typification is useful, the use of the genre becomes routine. She further argued that genres must accomplish a recognized social function or purpose and thus, she criticized the environmental impact statement as an example of genre because it did not meet this criterion, but rather could have multiple conflicting motives. Swales [82] similarly viewed genres as “a system for accomplishing social purposes by verbal means.” He suggested analyzing documents in terms of their exhibited characteristic moves, defined as “a functional unit, used for some identifiable rhetorical purpose” [11]. For example, he analyzed the introduction, methods, results, discussion and conclusion sections of research articles as well as other research-process genres, such as abstracts, presentations, theses and grant proposals. Following Swales’s [82] approach, researchers have described other scientific genres such as research paper abstracts [67] and discussion sections [72], grant proposals [11], posters [64] and other visual presentations at scientific meetings [77], and French thesis defences [66], as well as journalistic genres such as news [62, 85], editorial letters [37] and magazine covers [44], genres from the commercial world, such as business letters of negotiation [74], letters of application for jobs [45], Chinese sales letters [89], and genres from many other areas of human endeavour, such as environmental impact statements [3], building reviews [7] and philanthropic direct mail [84]. For instance, this document is an example of the NSF grant proposal genre. It has a form familiar to many scientists and enforced by the NSF: a title page, narrative, budget, and so on. There are many document genres: some common, such as a report or a newsletter, and others restricted to specific domains, such as the course syllabus or problem set in higher education. Genre is applicable to electronic as well as physical documents. For example, in a study of Web documents, Crowston and Williams [29] were able to identify documents of many familiar genres and of a few genres new to the Web, such as the home page [32] and hotlist.
Genre classification

In order to use genre in a system, it is desirable to develop some sort of classification of them because a classification allows a compact and structured representation that can be manipulated and exploited to identify various levels of granularity and to establish useful relationships among the genres [57]. Though Askehave and Swales [2] note that classification has not been a priority of most genre researchers, there is still a substantial body of work on analyzing genre in printed documents and some work studying them on the Internet. A key goal of the project will be working out the best set of genres to use in an information access system, given the limitations imposed by the need to automatically recognize them.

Many attempts to develop a categorization of genres have been top-down, that is, they analyzed a set of documents based on theoretical principles or according to a priori classifications. A key difference in these efforts is the number of genre categories distinguished. Many studies of Web pages have used fewer broader categories: for example, zu Eissen and Stein [90] used only 8 genres (help; article; discussion; shop; portrayal, non-private; portrayal, private; link collection; and download). At the other extreme, Görlach [40] offered a catalog of some 2000 genre (or text type) terms, which is intended to be an exhaustive list of the terms used in English. In between, Lee [59] categorized documents in the British National Corpus (BNC) into 70 genres or subgenres (with some document assigned more than one genre). However, he notes that the genre terms used were “meant to provide starting points, not a definitive taxonomy,” for example grouping textbooks and journal articles as academic texts that can be further distinguished by medium.

If the classification includes more than a handful of terms, it is useful to group together similar terms. For example, social science papers might be grouped with computer science implementation papers, biology research papers and so on. These genres share some similarities, such as a title, abstract and bibliography, but differ in other particulars, such as the expected section headings, types of arguments, etc. It is then necessary to deal with terms of different levels of generality [59]. Many organized lists of genres are structured as single hierarchies. Figure 2 shows a small section of the hierarchy of genres of Web documents identified by Crowston and Williams [29]. Advertisements and announcements are both examples of declaratory document genres; classified advertisements are a special kind of advertisement, and so on. Similarly, social science papers, computer science papers and biology papers might be seen as examples of a more general genre of research papers, which are in turn examples of papers or articles.

<table>
<thead>
<tr>
<th>&lt;declaratory document genres&gt;</th>
<th>advertisements</th>
<th>classified advertisements</th>
<th>Short paid announcements appearing in a periodical sorted according to the good or service being offered or requested</th>
</tr>
</thead>
<tbody>
<tr>
<td>announcements</td>
<td></td>
<td></td>
<td>Printed or published statements or notices that inform the reader of an event or other news</td>
</tr>
<tr>
<td>custom 404 page</td>
<td></td>
<td></td>
<td>A Web page announcing that the requested Web page could not be found on the server</td>
</tr>
<tr>
<td>news bulletins</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>press releases</td>
<td></td>
<td></td>
<td>Official or authoritative statements giving information for publication in newspapers or periodicals</td>
</tr>
</tbody>
</table>

**Figure 2.** A section of a hierarchy of document genres [from 29].
An advantage of a hierarchy is that it avoids the need to predetermine the level of detail needed in the classification. Depending on the circumstances, we can consider genres at any of these different levels and different levels might be more or less useful for different purposes. Of course, there is no guarantee that convenient and well-known terms will exist for all levels of the hierarchy. A second criticism of traditional hierarchies is that they rely on a single organizing principle, which may not be useful or appropriate for all cases. Harrell and Linkugel [42] note that there are multiple bases on which such a classification could be constructed. To overcome this problem, Kwasnik and Crowston [58] suggested using a facetted-analysis approach, following the example of previous genre-identification studies such as Päivärinta [71], Tyrväinen and Päivärinta [83] and Karjalainen et al. [52] who looked at the management of enterprise documents, and Kessler, Nunberg and Schuetze [54] who sought to identify a limited set of facets for communicative purposes. Crowston and Williams [29] based their classification of genres on the *Art and Architecture Thesaurus* [73], which is also a facetted classification overall.

A few researchers have attempted to identify genres bottom-up through user studies (as we did in the initial phase of our prior research). Dewe, Karlgren and Bretan [31] asked users to provide lists of genres found on the Web and received 67 responses. They noted though that users tended to conflate genre and topic. Nilan, Pomerantz and Paling [70] surveyed 242 Web users in person and via the Web about their purpose in searching the Web, the genre of document expected and the actual document found, and collected 1335 example pages; genres could be assigned for 1076, giving a total of 116 genres. They then grouped the genres, first, following the plan of the *Art and Architecture Thesaurus* and second, based bottom-up on the user responses. The second classification was much broader and shallower than the first, and some genres were classified in multiple locations, suggesting that there may be difference between genres in use and formal structured knowledge. Freund, Clarke and Toms [38] developed a list of 16 genres based on interviews with users and analysis of document repositories in a particular organization. They found that the 16 genres covered about 75% of documents randomly sampled from the repositories. These studies will be useful for the initial stage of the proposed research, selecting a small set of genres for supporting information access.

**Automatic text type detection**

So far, we described manual identification and classification of genres that informs the proposed research. In this section, we describe the machine learning techniques that will be explored to automatically assign text type to web pages (using text type to refer to a classification of documents based on regularities of form vs. genre for purpose, though as noted, prior research has tended to mix the concepts). We have identified several challenges to automatic detection. The first challenges are to chose a set of classification techniques, accumulate and select a rich set of features characterizing attributes of the web pages and to use them together to explore the text types that result from those features. Another challenge arises from the large number of genres that we would like to identify where the number of training examples in the corpus may be insufficient to distinguish them. Finally, the large numbers of features themselves can pose practical problems to the efficiency of the machine learning algorithms. These challenges will be considered in turn.

**Techniques**

Researchers have applied numerous approaches to automatic text type identification, both for text documents and for web page documents. In addition to classifying documents by text type, studies that seek to classify web pages as to other types, such as thematic categories, provides relevant background into features that can be used to represent the content of web pages. In the previous studies of automatic text type identification several types of classifiers were used. In our work, we propose to use Support Vector Machines (SVMs), which have been shown to work well on text classification tasks with very
large feature sets [51, 79]. With a faceted classification, different techniques can be used recognize individual facets, as shown in the work of Kim and Ross [55]. Typically, automatic assignment studies evaluate their automatic systems by the measure of classification accuracy, although some systems report both precision and recall.

Features

In classifying documents, the development of features used to represent the characteristics of each document is one of the most important tasks. We have reviewed this work to identify the features that others have found useful in identifying text type. In classifying documents for text type or genre, almost all work has used some part of the content of the documents, some by using the well-known technique of using all the content words for features, known as the bag-of-words approach. One study looked at limiting this content, Stamatos, Fakotatis, and Kokkinakis [81], who classified documents in the Wall Street Journal according to 15 news text types and reported successful results using the 15 to 35 most frequent content words as features.

Many other researchers have focused on features relating to the form of the documents, such as Karlgren and Cutting [53], who used features such as the number of prepositions and adverbs, the average number of words per sentence and the number of first person pronouns, but results were good only on a small number (2–4) of text types. Kessler, Nunberg, and Schuetze [54] took a similar approach to classifying 6 text type in the Brown corpus and used 55 features of 3 different types: structural cues, including counts of passives, nominalizations and part-of-speech types; lexical cues, including particular titles, suffixes and temporal words, but not content words; and character level cues, such as punctuation and acronyms.

Yet other researchers have experimented with combinations of features types, establishing that content words are not sufficient by themselves. In work by Stamatos, Fakotatis, and Kokkinakis [80], Greek text was classified according to text type and author using style markers for features that included words, phrases and other syntactic levels from an NLP analysis of the text that included both the content and measures of the syntactic complexity of the text. The experiments showed that the NLP style markers outperformed the most frequent word method. Sets of features were compared in work by Dewdney, VanEss-Dykema, and MacMillan [30], where the sets consisted of traditional content words; presentation features using part-of-speech tags to identify tenses of verbs, frequencies of several closed class word sets, and character level markers, including frequencies of words and sentences, punctuation, and smilies. Combinations of word and presentation features outperformed either one separately.

The classification of web page documents into text types adds another layer of potential features from the form and functionality of the web page documents. In work by Rehm [75], features are discussed that include metadata such as the URL, http header, size and title; the HTML structure, including aspects of the hyperlinks, inline graphics and interactive elements; document-spanning features, including text type elements and position of document in the hypertext structure; linguistic and structural cues, including specific linguistic expressions at web-text type positions; and language issues, including the nature of the language and spelling. In their latest work on text type detection, Dong, Duffy, Shepherd and Watters [33] conducted a set of experiments classifying web pages using four types of text type: FAQ, news, e-shopping and personal home pages. Features were associated with content: text (words) visible in the web browser; form: names and attributes of HTML tags: title, head, font, bullet, div, style, tr, td; and functionality: names and attributes of HTML tags: applet, script, jsp, link, form, select, option, textarea, and input. Combinations of all three feature sets performed best at recall, and the combination of form and functionality performed best at precision, but all were in the 90-95% range. Finn and Kushmerick [36] examined three feature sets: a bag of words, a part-of-speech vector of ratios of different parts of
speech, and a vector of text statistics such as average sentence length and word length. They found that in most cases they had their best results when all three feature sets were used in combination.

Feature selection

In many of the works cited above, the number of features defined for a problem has been quite large, ranging into the hundreds of thousands. Even for SVM, these large feature sets can be challenging to both the performance accuracy of a classifier and to the practical performance in the time needed to train classifiers. In many types of classification, feature selection can be beneficial to both of these problems by reducing the number of features actually used by the classifier. Typically, in feature sets that contain thousands of binary features, statistical measures are used to compare the contribution of different features and to select a smaller number. These binary features arise when, for example, content words are represented as features that can either be present or not. In the text categorization area, a study of statistical feature selection by Yang and Pederson [86] found that among feature selection measures of Document Frequency Thresholding, Information Gain, Mutual Information, the $\chi^2$ statistic (CHI) and Term Strength, the Information Gain and CHI were most effective in reducing the number of features without sacrificing classification accuracy. [86] Although many studies in automatic text type classification did not report on using feature selection, in [30], information gain was used to reduce the feature set. Also, Dong, Watters, Duffy and Shepherd compared statistical feature selection measures in [34] and found that information gain performed best. On the other hand, for features that are multi-valued nominals, such as a closed set of possible terms, or for numeric features, such as the average number of sentences, wrapper methods of feature selection, while they may take longer to compute, may yield better results in feature selection. Wrapper methods are suitable for smaller numbers of features because the method trains a classifier on different subsets of features and compares the performance [56]. The wrapper method does not try all combinations of subsets of features; rather it uses a search algorithm to find the best subset. For example, in our work in [68], we used a combination of genetic search and bidirectional search to find the best feature set.

Target classification of text types

A third challenge is picking the set of text types to be recognized. In the studies of automatic genre recognition cited above, the number of genre tended to have a small number of categories (from 4 to a maximum of 32, with a median of 10). These numbers are small compared to the hundreds found in user studies, including the user study from our previous NSF grant, and general Web surveys. Experimenters have attempted to detect fewer text types because increasing the number to be detected reduces precision and demands a larger corpus for training. As Boese [5] noted, however, definitions of broader categories have be “softer” to include more documents and so very broad categories may be less useful to users or for the information access systems discussed above. Another concern is that the design of classification interacts with the features chosen. Grouping together documents with similar purposes but dissimilar structural features (i.e., genres as opposed to text types) may be useful for users, but create problems for automatic classification.

In this investigation, we do not have an a priori set of text types. Rather we will use clustering as a data exploration method on the features to establish what text types are identified by the features. As noted above, an important part of this investigation is to discover a subset of the text type that can both be reliably automatically recognized and that can be useful to the user’s information search. The use of clustering for this type of data exploration is well-known, for example, different types of clustering and in what ways they group the data is described by Jain et al [50]. Specifically, we propose to use hierarchical clustering to explore any naturally occurring hierarchy of text types, to cluster by Self-Organizing Feature Maps (SOM) which are often used to reduce high dimensional data sets, and to cluster by the often-used
K-means method, for various small values of k in the neighbourhood of 15, to explore whether the text types naturally fall into an appropriate number of clusters [50].

Training data

A final problem related to the available training corpora is that the number of training examples for each genre type is a skewed distribution across the genre. There may be only a few training examples for some genre, which would be too few for the classification algorithm to learn how to recognize that genre. One way that this sparse data problem can be solved is to use hierarchical classification to reduce the number of genre distinguished by any one classifier. This technique [35], trains one classifier to identify just the small number of genre at the topmost level of the hierarchy. Then additional classifiers are trained to recognize subgenres.

Another technique for improving classifiers that have only a small amount of training data is to train the classifier with unlabelled examples in addition to the web page examples that have been labelled with text type [51]. In this technique, a similarity measure is defined for web pages based on the feature representations, and similar unlabeled documents are also used to train the classifier. This technique can be characterized as a semi-supervised learning technique and has successfully been used in sentiment categorization of a collection of movie reviews [39]. Another technique for using unlabeled data to help compensate for a small number of training examples is that of co-training. In this technique, two different types of classifiers are used: one is trained on the labelled data and used to label the unlabeled data, while the second classifier is then trained on both the labelled and unlabelled data (with labels from the first classifier). This technique has also been successful in sentiment classification, in experiments involving identifying emotions in dialog [65].

Methods and procedures

Now that we have presented the literature on which we will build, in this section we present the details of the methods and procedures to be employed to answer our research questions, namely, using genre/text type meta-data to improve information access systems.

Year 1

The goal of year 1 is to determine a set of genres that both are useful for supporting information access tasks and that can be reliably classified and to develop a set of algorithms for doing the classification. We will discuss in turn the methods to be applied in the user studies and in the natural language processing (NLP) work, and how we expect the two studies to inform each other.

User studies

The main user study activities in year 1 will be experiments to assess the utility of different sets of genres. We anticipate that each run of an experiment will take one to two months to complete, meaning that we will able to carry out four to six experiments per year. The initial experiments will be used to test the utility of an interface that clusters documents into a small set of genres. We will be using the corpus developed in previous work. The genres will be selected based on a review of several factors:

1. the distribution of genres in the current corpus;
2. consideration of which genres were most useful for the information seeking tasks in the experiments in our prior work;
3. the literature on information access to identify which genres are likely to be useful; and
4. the literature on automatic classification to identify reliably classifiable text types.

The experimental tasks will be taken from the set of experiments being used in the prior work reviewed above. An example experimental task is to find a relevant Web page for a given problem scenario. Each subject will perform a set of search tasks using the subset of genres of interest, and then complete an on-line questionnaire. Since the experimental interfaces will use a predetermined database of documents, we can identify which documents are relevant, and thus assess the success of a subject’s search (relevance and recall). In addition to success, for each task we will measure the time taken, number of pages explored, time spent using various aspects of the interface, such as cluster descriptions. Finally, the quantitative results will be augmented with qualitative “debriefing” sessions [88]. Baseline data on performance can be obtained from the current set of experiments, so these experiments can be run with just one treatment, though it may be desirable to use a more powerful within-subjects design. We anticipate needing several runs of the experiment to arrive at a final set of genres that seem useful in supporting information access and that perform well in the NLP experiments. However, if we settle on a useful set of genres more quickly, we will move up some experiments planned for year 2 of the study.

We anticipate running experiments with 30 to 60 subjects. A sample size of 60 per group will provide an 80% chance of detecting an effect size of about 0.5 standard deviation (i.e., 80% power) within subjects groups at a 5% confidence level (assuming equal variance). The experiments can be designed so the number of subjects is determined dynamically as the experiment progresses, stopping an experiment with fewer subjects if the effect size is large, and using more only if the effect size is small. We estimate that a total of approximately 360 subjects per year will be needed across all experiments. In order to provide an incentive to take part, we plan to offer a small payment for participation. As well, we plan to run some experiments over the Web, thus enabling us to draw on a wider population of potential subjects.

Natural language processing studies

The goal in year 1 of the NLP studies is to develop algorithms that can reliably classify documents into a small set of text types that correspond to genres found useful in the user studies. The approach is as follows:

1. Implement a wide range of features as suggested by previous work in this area (shown in Table 1). Prior to extracting features, the web pages will be processed by natural language processing techniques to identify words, POS tags, phrases, syntactic structures of text, and HTML structures.

2. Use hierarchical and SOM clustering to identify naturally occurring text types and compare with genre labels using pages in the current corpus. Use k-Means clustering iteratively to identify a number k for which the number of text types can be reliably recognized automatically, in the next step.

3. Determining a useful subset of text types:
   a) Identify a subset of the top-level genres that seem to correspond to text types suitable for automatic learning, train classifiers and determine level of accuracy.
   b) Compare the accurately identifiable text types from this track to the useful genre from the user study track, and revise the genre set for learning.
   c) Examine web pages of documents that are not accurately classified and identify, if possible, additional features
   d) Iterate over this process.
At the end of Year 1, a subset of top-level text types should be identified and classifiers trained to automatically recognize them as accurately as possible.

Year 2

The goal of year 2 of the project is to develop and test an integrated prototype, genre-enhanced information access system. However, as time permits we will extend the results of the earlier studies, as described in this section.

NLP studies

In the first part of the year, the focus of work will be on integrating the NLP text type classification algorithms developed in year 1 into an information access prototype. We anticipate implementing this prototype as a meta-search engine that processes queries through some commercial search engine then processes and presents the results to the user.

In the second part of the year (during the final user tests) additional NLP studies can be carried out to further enhance the algorithms. In particular, we will investigate the faceted hierarchy of genres, will focus on which features are useful in detecting genre, and will continue development of genre accuracy through getting feedback from the user study.

1. Install a trained classifier on the text type subset identified in year 1 to be used in information searches in the user study. Obtain feedback about usefulness and accuracy of genres and continue development.

2. Increase the number of text types that can be accurately classified.

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Content Cues</strong></td>
<td>Bag-of-Words, using words in the text visible to the user</td>
</tr>
<tr>
<td></td>
<td>Phrases</td>
</tr>
<tr>
<td><strong>Form Cues</strong></td>
<td>Number of prepositions, adverbs, adjectives, nouns, nominalizations,</td>
</tr>
<tr>
<td></td>
<td>verbs with tense</td>
</tr>
<tr>
<td></td>
<td>Number of first or second person pronouns</td>
</tr>
<tr>
<td></td>
<td>Syntactic complexity of sentences: average number of words, character,</td>
</tr>
<tr>
<td></td>
<td>syntactic levels</td>
</tr>
<tr>
<td><strong>Lexical Cues (closed classes)</strong></td>
<td>Structural words: <em>therefore, that, …</em></td>
</tr>
<tr>
<td></td>
<td>Terms of address</td>
</tr>
<tr>
<td></td>
<td>Words used in expressing dates</td>
</tr>
<tr>
<td></td>
<td>Words specific to individual genre</td>
</tr>
<tr>
<td><strong>Character Level Cues</strong></td>
<td>Punctuation, delimiters, smilies</td>
</tr>
<tr>
<td></td>
<td>Acronyms</td>
</tr>
<tr>
<td></td>
<td>Spelling</td>
</tr>
<tr>
<td><strong>Web Document Metadata</strong></td>
<td>Web Page URL, header, size, title, &lt;meta&gt; tags</td>
</tr>
<tr>
<td><strong>HTML structure</strong></td>
<td>Hyperlinks: number of links, hypertext structure, anchors</td>
</tr>
<tr>
<td></td>
<td>Inline graphics: dimensions, names, alternative text, format, number of</td>
</tr>
<tr>
<td></td>
<td>images</td>
</tr>
<tr>
<td></td>
<td>Position of document in hypertext structure</td>
</tr>
<tr>
<td></td>
<td>Attribute values in HTML tags</td>
</tr>
</tbody>
</table>
a) In this year, we will use the faceted hierarchy with a hierarchical classification technique to train local classifiers on sub-text type.

b) From the user study, we will identify genres that are useful but that cannot be accurately identified due to lack of training data examples. The two techniques involving semi-supervised training of unlabelled data and co-training will be used to compensate for sparse training data. If necessary, additional documents will be collected from the web and labelled to increase the number of examples in an important genre type.

3. Use feature selection techniques in a set of experiments to determine which features are useful in accurately identifying text type. This step is important to establish future practical use of the automatic text type identification.

4. Perform a set of evaluation experiments on the accuracy of the classification of the final set of text types identified during this year of the project.

User studies

The main activity during year 2 will be to carry out a final test of the finished prototype. However, while the prototype is being integrated, we will carry out several follow-on studies.

1. Test alternative user interfaces that use genre/text type. The current prototype uses genre metadata to cluster documents but we hope to experiment as well with text type as an aspect of relevance feedback (e.g., “show me more like this”) or in prioritization of search results (e.g., it has been claimed that most searchers are not interested in getting personal home pages [9] so these could be moved down in the results set).

2. Experiment with a corpus where the genre identification has been intentionally degraded to assess the required degree of accuracy for a system to be useful.

3. Run experiments with different user communities to assess the generality of usefulness of the set of genres.

As noted above, these additional experiments will be started during year 1 if we arrive at a suitable set of genres/text types more quickly than anticipated.

Management plan

Based on preliminary assessment of the effort required, we are requesting funding for two graduate students, some support for a professional research staff member and a small amount of summer support for 3 PIs (approximately 1 summer months per PI). All three PIs, Drs. Kevin Crowston and Barbara Kwasnik, will work during the summer on project management and research design, and devote 10% of effort during the academic year to project management and oversight (1/2 day per week) as part of their regular responsibilities. Dr. McCracken, a full-time research professor, will be funded for 3 months during the calendar year (approximately 25% effort). All three PIs will share in overall project design and report writing. Each PI will be responsible for designing specific aspects of the project and overseeing those aspects:

- Dr. Crowston will direct the project and be responsible for project oversight and reporting.
- Drs. Kwasnik and Crowston will lead the user experiments to determine useful genres.
- Dr. McCracken will lead and contribute to the computer/information science research team in
NLP tool development for text type classification and final system integration.

We will employ two main project management techniques. First, we will have regular meetings of the project members to share findings and to plan the work. Initially, these will be every other week, but the frequency of meetings will be adjusted depending on the pace of the work being carried out at the time. These formal meetings of all project participants will augment the regular interaction of the teams of PIs and students working on the data collection and analysis. Second, an initial project activity will be the development of a more detailed timeline (based on the initial one in the budget justification section) against which progress will be measured. The budget includes support during summer and academic year to support these activities.

Expected scientific, societal and educational impacts

Our project will have three kinds of societal impacts:

1. tackling problems facing users in their interactions with large digital collections, thus advancing knowledge and understanding within information science, and specifically digital libraries, and contributing to the intellectual development of this field;
2. developing useful reusable intellectual products, thus contributing to society in general; and
3. contributing to the advancement of the education of doctoral students.

Our project will address several important, persistent and difficult problems. First, we will address the bottleneck of information overload in information access due to undifferentiated documents and noisy collections by providing another dimension on which to differentiate documents, namely genre/text type. Our project will also create several reusable intellectual tools that will disseminate our research beyond the usual journal publication and conference presentations, as important as those are. Specifically, we will develop algorithms for automatic classification of text type and an integrated system. Finally, our project will contribute to education by providing a training ground for doctoral students at Syracuse University.

Additional results from prior NSF funding

In addition to the work reviewed above, two of the PIs for this grant, Crowston and McCracken, have been funded by an additional NSF grant within the past 48 months: HSD 05–27457 ($684,882, 2005–2008, with R. Heckman and E. Liddy), Investigating the Dynamics of Free/Libre Open Source Software Development Teams. Crowston also received funding for IIS 04–14468 ($327,026, 2004–2006) and SGER IIS 03–41475 ($12,052, 2003–2004), both entitled Effective work practices for Open Source Software development. These grants have supported a study of the evolution of effective work practices for distributed groups, specifically, for teams of free/libre open source software (FLOSS) developers. The funding enabled travel to conferences (e.g., ApacheCon and OSCon) to observe and interview FLOSS developers and to present preliminary results, and for the purchase of data analysis software and equipment. This work has resulted in nine journal papers [17–19, 21–23, 25, 27, 46] with others under review [1], multiple conference papers [e.g., 12, 15, 20, 24, 26, 28, 43, 48, 49, 61, 78] and workshop presentations [10, 13, 14, 16, 47]. These grants have supported a total of six PhD students; several others have been involved in specific aspects of the work. The HSD grant included a component applying NLP techniques to analyze large corpora of email and provided significant experience working in an interdisciplinary team.
References


Kevin Crowston  
Curriculum Vitae

Education


Appointments

1991–1996 Assistant Professor of Computer and Information Systems, School of Business, University of Michigan.

Publications (from a total of 53 peer reviewed journal and conference papers)


Other significant publications


**Synergistic activities**

1. **Mini-track chair** (with Carina Ihlström Eriksson and Tero Paivarinta). *Mini-track on Genres for Digital Documents, Hawai‘i International Conference on System Science*, January 2008 (service to scientific community)

2. **Editor, special issue** on Genre of Digital Documents, *Information, Technology & People*.

3. **Webmaster**, International Federation for Information Processing (IFIP) Working Group 8.2 on *Information Systems and Organizations* (service to scientific community)

**Collaborators in the past 48 months**

- Eileen Allen (Syracuse)
- Hala Annabi (Ohio)
- Kathy Chudoba (Florida State)
- You-Lee Chun (Syracuse)
- Megan Conklin (Elon)
- John D’Ignazio (Syracuse)
- U. Yeliz Eseryel (Syracuse)
- Claudio Garavelli (Polytechnic of Bari)
- Robert Heckman (Syracuse)
- James Howison (Syracuse)
- Carina Ihlström (Halmstad)
- Bernhard Katzy (UniBW Munich)
- Barbara Kwasnik (Syracuse)
- Chei Sian Lee (National University of Singapore)
- Qing Li (Syracuse)
- Elizabeth D. Liddy (Syracuse)
- Chengetai Masango (Syracuse)
- Nelson Massad (Florida Atlantic)
- Nora Misiolek (Marist College)
- Kangning Wei (Syracuse)
- Michael Myers (Auckland)
- Dmitri Roussinov (Arizona State)
- Joseph Rubleske (Syracuse)
- Steve Sawyer (Penn State)
- Barbara Scozzi
- Sandra Sieber (IESE)
- Mary-Beth Watson-Manheim (Illinois Chicago)
- Kangning Wei (Syracuse)
- Rolf Wigand (Arkansas)
- Eleanor Wynn (Intel)

**Thesis advisors**

Professor Thomas W. Malone (Chair), Deborah Ancona and John Carroll (all of the Sloan School of Management, Massachusetts Institute of Technology).

**Thesis advisees (5 current advisees and 2 graduates)**

Marcel Allbritton (consultant), Naybell Hernandez, Chengetai Masango, Kangning Wei, James Howison, Qing Li (all of the School of Information Studies, Syracuse University); Hala Annabi (Ohio)
BARBARA HANNA KWAŚNIK

EDUCATION

Queens College. CUNY  B.A., English Literature  June, 1969
SUNY Binghamton,  M.A. English Literature  May, 1972
Queens College. CUNY  M.L.S.  June, 1981
Rutgers. The State University of New Jersey. SCILS  Ph.D. Info and Library Studies  May, 1989

APPOINTMENTS

School of Information Studies. Syracuse University.
   Professor. 2003-present
   Assistant and Associate Professor. 1987-2003.
   Director, Ph.D. Program. 1994-98 and 1999-2001
   Director MSLIS Program, 1991-94
School of Communication, Information, and Library Studies. Rutgers. The State University of N.J.
   Teaching and Research Assistant. 1985-87. Department of Library and Information Studies and Department of Journalism and Mass Media.

PUBLICATIONS RELATED TO PROJECT


OTHER SELECTED PUBLICATIONS


**SYNERGISTIC ACTIVITIES**

**Panelist:** “Research Issues in Genres.” Organized by Mark Rosso and Stephanie Haas. ASIST Annual Meeting, Oct. 21-25, 2007, Milwaukee, WI

**Mini-track co-chair** Mini-track on Genres for Digital Documents, Hawai‘i International Conference on System Science, January 2004, 2005, & 2006 (service to scientific community)


**Collaborators in past 48 months**
Clare Beghtol (Univ. Toronto)
Joseph Bush (Taxonomist, Inc.)
Kevin Crowston, Elizabeth Liddy, John D'Ignazio, Min-Chun Ku, Victoria Rubin, Jian Qin, Keisuke Inoue, David Pimentel, Joseph Rubleske (i-School, Syracuse University)
Jonathan Furner (UCLA)
Stephanie Haas, (UNC Chapel Hill)
Natalia Levina (NYU)
Ia McIlwaine (Emerita, City University, London, UK)
Mark Rosso (Meredith College, NC)
Dmitri Roussinov (Univ. Arizona)
Michael Shepherd (Dalhousie University, NS Canada))

**Graduate Advisors**
James D. Anderson. (Chair). Nicholas Belkin, Daniel O'Connor (all at School of Communication, Information and Library Studies. Rutgers. The State University of New Jersey).
Michael Nilan. School of Information Studies. Syracuse University.

**Graduate Advisees (Graduates and Current)**
Swati Bhattacharyya, You-Lee Chun, Hannah Francis, Min-Chun Ku, Makiko Miwa, David Pimentel, Silvia Barcellos Southwick, Claire Urfels, Sarah Webb.
Nancy Jean McCracken  
School of Information Studies at Syracuse University  
209 Hinds Hall, Syracuse NY 13244-4100  
(315) 443 3955  Fax: (315) 443 5806  njm@ecs.syr.edu  

Education  
Ph.D. Computer and Information Science, Syracuse University, Syracuse, New York, 1979  
B.A. Mathematics, Earlham College, Richmond, Indiana, 1971  

Appointments  
2007 - Research Associate Professor, School of Information Studies (IST), Syracuse Univ. (SU)  
2003 - 2007 Research Associate Professor, Center for Natural Language Processing (CNLP), IST, SU  
1987 - Affiliated Faculty, Dept. of Engineering and Computer Science (EECS), SU  
2000 - 2003 Senior Research Associate, CNLP, IST, SU  
1992 - 2000 Senior Research Scientist, Northeast Parallel Architectures Center (NPAC), SU  
1987 - 1992 Manager of Research Consulting, NPAC, SU  
1987 - 1989 Research Consultant, NPAC, SU  
1984 - 1987 Senior Research Associate, School of Computer & Information Science (CIS), SU  
1982 - 1984 Associate Professor, CIS, SU  
1978 - 1982 Assistant Professor, CIS, SU  

Related Publications  

Other Publications  
Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA)


**Synergistic Activities**
Director and Training Coordinator for an NSF Research Experiences for Undergraduates (REU) site program in computational science (1990-96): Responsible for developing training plan, lecture materials and primary lecturer (90-96). Co-director (94-95). Director (96). Oversaw 2-3 student research projects per year, including writing a research paper.

Primary responsibility for the NSF Regional Alliance grant: Retooling for Parallel Computing, 1995-98, in collaboration with the Center for Research in Parallel Computing (CRPC), an NSF Research Center, in a "train the trainer" project for supercomputing center trainers. Directed the development of curriculum materials with the CRPC and the National Center for Supercomputing Applications (NCSA).


Educational activities directed at community or minority students: Developed a one-day math/computer science workshop for the Sonia Kovalesky festival to interest high school girls in mathematics, October 1999, and again in 2006. Directed and taught the Young Scholars Program (95-97), a one month summer program for Syracuse middle school students, and the Java Academy (98-99), a 10 week set of Saturday classes for high school students in Java.

Community Service: Member of the Board of Directors of the Technology Club of Syracuse, which supports science and technology education for K-12 students, 1996-2002.

**Collaborators & Other Affiliations:**

**a. Collaborators and Co-Editors:** (non-SU)
Chen, Jiangping
University of North Texas

Ittycheriah, Abraham
IBM, TJ Watson Research Center

Latto, Andrew
Gensym, Inc.

Light, Marc
University of Iowa and MITRE

**b. Graduate and Postdoctoral Advisors:**
Reynolds, John C.
Carnegie-Mellon University

Fox, Geoffrey C.
Indiana University

Lockwood Morris, Luis Sanchis, Otway Pardee (emeriti) Syracuse University

**c. Thesis Advisor and Post-graduate Scholar Sponsor**
O'Farrell, William
IBM, Toronto Research Centre

Sen, Mehmet
KOC University, Istanbul

Necati E. Ozgencil
Syncsort, New Jersey
Budget Justification

A. Salaries and Wages – Senior Personnel
   Funding is requested for 1 month summer support for all three PIs as well as 1 academic year month for Research Professor Dr. McCracken (10% effort during the AY). The PI, Dr. Crowston, will be responsible for overall project direction and coordination and for assuring successful project completion, including submission of NSF progress reports, as required. Drs. Kwasnik and Crowston will lead the user experiments to determine useful genres. Dr. McCracken will lead the computer/ information science research team in NLP tool development and integration and will be directly involved in implementation during the academic year. The PIs will jointly be responsible for the review of the data and preparation of manuscripts for publication.

B. Salaries and Wages – Other Personnel
   Funding is requested for two Ph.D. students, 50% academic year and 100% summer effort, for a total of 2200 hours/year (4400 hours in two years). The graduate students will support the principal investigators in experiment design and will have primary responsibility for data collection and analysis, under the oversight of the PIs.

C. Fringe Benefits
   Fringe Benefits are calculated as direct costs in accordance with Syracuse University’s indirect cost rate agreement (Department of Health and Human Services, 32.4% for faculty academic year, 17.0% for faculty during the summer, 17.2% for graduate students). Actual rates in place at the time of an award will be charged.

E2. Travel:
   Travel support is requested for students and PIs to disseminate results at academic conferences (one trip each, $1800/trip).

G. Other Direct Costs
   1. Materials and supplies:
      $500 per year is requested for supplies, plus an additional $3000 in year 1 to purchase a Linux server to be used to run the machine learning experiments. A dedicated machine is required because some of the experiments, e.g., feature selection, are extremely time consuming to run and so should be run on a separate machine.
   6. Other
      A total of $8400 is requested for payments for experiment participants. $80,028 is requested for partial support of tuition for two graduate students (12 credit hours per year per student at $1,079/credit for Year 1 and $1,144/credit for Year 2).

I. Indirect Costs
   Indirect Costs are calculated in accordance with Syracuse University’s federally negotiated indirect cost rate agreement (Department of Health and Human Services), which is currently 46% of modified total direct costs (MTDC).
The following proposed timeline for the project indicates how the requested resources will be applied.
Facilities, equipment and other resources

Syracuse University is one of the largest and most comprehensive independent universities in the United States. Founded in 1870, Syracuse offers excellent facilities, equipment and other resources for research and study in many academic and professional disciplines.

The School of Information Studies is a leading center for innovative programs in information policy, information behavior, information management, information systems, information technology and information services. Its approach stands out from other institutions that offer computer science, management, information science and related programs in that our focus is on users and user information needs as a starting point for integrating information and information technology into organizations. The faculty of the School cross disciplinary boundaries to integrate the common elements of information management in business, government, education, and nonprofit settings, including the relationship of information and knowledge, electronic and traditional libraries, information systems and technology, information resources management, information policy and services, and the study of information users.

The School of Information Studies space plan includes providing (1) a space for a community of learning, research, and education for students and faculty; (2) space that supports economic development and growth in Central New York; (3) space that supports research, development and economic growth through the School’s research centers; (4) common spaces that are inviting to students and visitors; (5) space that supports communication and connections between floors to preserve the strong feelings among students, faculty, and staff of being on the IST team; (6) a building that supports state of the art technology including broadband and wireless in offices, classrooms and centers; (7) space with the flexibility to change to meet the needs of a changing networked economy, changing technology, research, and faculty and student needs; (8) classroom space that supports student access to technology and/or classroom discussions in a room such as a case management classroom; (9) sufficient conference and meeting room space for a school enriched by its faculty and staff commitment to team meetings, service, and collaborative research; and (10) space that supports a collaborative learning environment for students.

SU’s Library serves the information and research needs of the academic community. The collections exceed 2.6 million volumes, 11,330 serials and periodicals, and 3.4 million microforms, located in several libraries on campus. Library services include information and reference, online database searching, access to bibliographic and other data on CD-ROM and interlibrary loan.

Computing Services helps researchers, faculty and students use computing by providing personal computers, mainframe computers, data communication networks, software, training and advice. Most equipment and services are available without a direct charge.