

# Research Statement

Gheorghe Muresan, Ph.D.

## ***Introduction***

I have a strong interest in investigating interaction models, and in designing, building and evaluating systems that support information-seeking activities. I am particularly interested in personalization of information retrieval, i.e. in building systems that display adaptive, intelligent behavior; such systems should learn a user's short and long term interests, particular preferences and shifts in preferences, and should be aware of the user's context at a given time when a search for information is initiated. Consequently, the result of the search depends not just on the query specified by the searcher, but also on the user profile and the search context. Moreover, I am interested in supporting social filtering, or recommender functionality in information systems, so that a person's interests or preferences can be inferred based on membership to a certain community, or from actions or decisions taken by people with similar profiles.

The consequence is that my research work spans a variety of very different areas of investigation:

1. Interaction models – investigating information retrieval functionality and search strategies that support a variety of information-related tasks. This also covers different forms of relevant feedback, from explicit ticking of checkboxes to indicate relevance, to implicit feedback, which means interpreting behavioral cues or actions such as printing, bookmarking, or saving, as well as spending time reading, scrolling, etc.
2. User Interfaces and Information Visualization – investigating tools, metaphors, or display organizations that support a variety of information retrieval tasks, from exploratory browsing in search for serendipitous information nuggets, to fast identification of known items, to comprehensive search for documents covering a certain topic.
3. Mathematical/statistical models for representing, summarizing or retrieving documents, for representing topics and identifying the topics covered by a document, and for estimating semantic similarity between documents.
4. Methods and algorithms for organizing information, from unsupervised clustering based on inter-document similarity, to supervised, training-based classification.

The next sections briefly describe the objectives of the relevant research projects that I have been involved in recently, and the achievements of those completed.

## ***Recent or current projects***

### **Mediated Information Retrieval**

My involvement in Information Retrieval started during my doctoral work, when I became interested in providing system support for exploratory searches, i.e. for information-seeking situations when the searcher is uncertain about what kind of information may be useful for solving a certain task, or is unfamiliar with the problem domain and its terminology, and therefore unable to specify a query. I designed and implemented WebCluster, a reification of the mediation model proposed and investigated with my supervisor, Prof. David Harper. The idea is to elicit the searcher's information need, to gradually learn her interests, to build a

user profile, and to subsequently assist the user in conducting searches on the web or other target collection, or to monitor a document collection for new information relevant to her.

During my Ph.D. I only evaluated the usability of a mediation system, not its effectiveness. I focused instead on running extensive search simulations, comparing different strategies and investigating the effect on performance of various parameters such as the term weighting model, the clustering algorithm, or the cluster labeling formula. The conclusion was that, with some user training, a mediation system has the potential to significantly improve retrieval effectiveness.

I recently co-supervised (with Prof. Nick Belkin) a Rutgers Ph.D. doctoral student who conducted a user experiment to investigate the effectiveness, efficiency and usability of a mediation system, and the effect of the search results organization on the retrieval performance. I designed and built the four user interfaces needed for the factorial experimental design, with the within-subjects test comparing two mediated with two non-mediated interfaces, and the between-subjects test comparing two interfaces with standard ranked list output against two interfaces that combined the ranked list with a hierarchy generated by a clustering algorithm. The statistical analysis of the results is ongoing. More future work is needed in investigating how well cluster structures generated by different algorithms, as well as document and cluster labels generated by different methods, can support the exploration of the information space.

### **Integrated approach to interaction design, log analysis and user modeling**

Most often, the specification of an interactive system is in the designer's natural language, such as English, accompanied by a set of the sketches of the interface at different stages of the interaction. Unfortunately, natural-language specifications tend to be lengthy, vague and ambiguous, and therefore are often difficult to prove complete, consistent and correct. Formal and semiformal languages have proven their value in fields such as mathematics, physics, circuit design and command language systems.

In graphical user interface (GUI) design, several approaches have been proposed for modeling the interaction, such as menu-tree structure, statecharts, transition diagrams and user action notation. They specify an interaction grammar, i.e. the states of the system, possible actions available, and state transitions when various actions take place. However, none of these approaches has gained widespread adoption as each of them has advantages and disadvantages, plus there is no widespread understanding of their advantage.

The long-term vision of this project is to define an *interaction modeling language* (IML) based on UML, XML and XMI that formally describes the user interactions that an interactive system can support. Modeling tools that use statecharts, transition diagrams or any other visual tools for modeling will be based on the common IML, and will therefore be able to inter-operate. Apart from offering inter-operability between modeling and design tools, IML will support computer-assisted software engineering (CASE) tools that generate the code based on the model, after checking the model's correctness and consistency.

Apart from these practical outcomes, IML will also support research in human-computer interaction (HCI): IML will be used to log the states of the systems, the user actions and the state transitions. The log analysis will support models of user behavior based on Hidden Markov Models (HMM) which will also be represented graphically so that behavioral patterns can more easily be observed. Also, usability problems can be detected (such as functionality never used), or combinations of actions can be observed, and optimizations provided.

I led a group of undergraduate students that successfully applied this approach to designing a simple juke-box system in Java, and produced a graphical output (SVG) of the logged sequence of user actions and system state transitions. I also applied this approach in the Interactive track of the Text Retrieval Conference (TREC 2003) and, most recently, in the Mediated Information Retrieval (MIR) project described above.

However, improvements are needed in order to help this approach gain acceptance: the process of converting UML state and class diagrams (obtained with a commercial tool such as Rational Rose) into DTD and XML schemas needs to be automated in a robust and flexible way. Also, while the XML parser for processing interaction logs can be generated automatically, the outcome of the analysis is just numeric data, and we have been building the graphical representation manually. The generation of visual representations for the results also needs to be automated.

### **Interactive and HARD tracks of TREC**

Logging search interactions, facilitated by the project described above, facilitated a very different area of research: log analysis. In Interactive TREC, in which I participated in 2002 and 2003, and in which the Rutgers IR group had participated every year since 1992, a number of subjects are assigned a number of information tasks and are asked to search for relevant information. We thus have access to a large number of interaction logs which, unlike web search logs, have been created in controlled laboratory conditions, so we have clear search session boundaries, know the tasks assigned to the subjects, and have information about the searcher's familiarity with the search topics, their satisfaction with the search results and, in many cases, relevance judgments associated with the documents retrieved. The analysis of the queries submitted by different people for a certain topic, and of the documents retrieved and judged relevant or not relevant, allows us to build topic models and relevance models so that, given a user query, the intended topic can be predicted, as well as the documents or document paragraphs relevant for that topic. This project, for which I received a Rutgers University Research Council grant, is ongoing.

In 2004 and 2005 I participated in the Highly Accurate Retrieval of Document (HARD) track of TREC, which provided an excellent opportunity to investigate some aspects of personalization in IR. One aspect was to explore questions that can elicit extra information from a searcher, via *clarification forms*, in order to gauge aspects of the searcher's interests, preferences or context. Apart from comparing the various questions or combinations of questions among themselves, we also compared their effectiveness in improving search performance (via query expansion) against automatic methods, such as simple pseudo-relevance feedback, or query expansion methods based on using the web as a training corpus.

Another line of research was making use of user profiles, in the form of *metadata* which specified the subjects' level of familiarity with a number of topics, and their preference for a certain document genre or geographic coverage. We investigated various approaches (logistic regression, SVM, statistical language modeling) to predicting a document's class, as well as formulae for combining evidence from different sources.

Our overall conclusion was that algorithms are better than humans at estimating collection statistics and generating good queries. An obvious consequence is that, rather than spending time and requiring the searcher's effort to specify good queries, a better way is to use implicit relevance feedback and to let the machine generate the queries.

We are continuing that line of work with an investigation of the correlation between implicit and explicit source of relevance, in order to build confidence levels for the former, and of weighting schemes that allow relevance feedback and query reformulation methods to assign different levels importance to sources of evidence and to expansion terms.

### **Recommender systems**

A natural extension to implicit relevance feedback approaches that analyze behavioral cues in order to predict user interests and to build user profiles, is to take advantage of user membership to real or virtual communities, and to derive information from investigating patterns of behavior or preferences.

My colleague, Prof. Paul Kantor, is the author of AntWorld, an information system based on *swarm intelligence*, or on using *digital pheromonic schemes* to encode the searchers' exploration of the information space. When a searcher initiates a *quest* by introducing a query, the system reveals *trails* taken by previous searchers, together with results that seemed to produce user satisfaction.

In 2003 I supervised a group of Master students who conducted a user study to evaluate AntWorld as part of their IR term project. Our conclusion was that, while there were usability problems and issues related to trust, there was potential to significantly improve performance and satisfaction. Apart from improvements in usability, a longitudinal study was proposed to evaluate the value of the recommender system if and after it gains acceptance.

In 2004 I contributed to a Rutgers proposal to the National Science Foundation (NSF) to extend the desktop-based AntWorld into an AntWeb of profile- and context-aware webpages, where the content and format of the pages are personalized for the user of the web browser, and are informed by behavior and actions of users with similar profile. While the proposal was rejected, it contained interesting ideas that are worth pursuing in the future.

I am also supervising a PhD student's semester-long independent study; the plan is to design, implement and test the prototype of a recommender bibliographic system, envisaged to be used by (mainly graduate) students working on assignments or doing project work. Apart from evaluating the usability and effectiveness of the system, we are interesting in studying how the users' relevance judgments are affected, and whether they will tend to start trusting other people's judgments and go with the flow, instead of making their own judgments.

### **Conclusion**

While the projects that I have conducted are relatively small, due to the constraints of the academic environment, they have yielded results that are interesting for IR research. Also, participating in these projects and, moreover, proposing and leading many of them, have developed and are a proof of the knowledge and skills relevant for a research position in the industry.