Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2007)

SmartPortal™ for Biomedical Data Mining

Anna L. Buczak, *Member IEEE*, Charles Wan, Glenn Petry

Abstract— **Efficient data retrieval from large databases and the World Wide Web is an important task that has to be performed routinely in a wide range of applications. To facilitate the drug discovery process, the biomedical community needs tools that enable fast searching of databases and the web. SmartPortal™ assists users in their searches of biomedical information by quickly finding results of particular interest to the user in the deluge of data and the moving them to the top of the results list. SmartPortal™ achieves its goal through 1) constructing a user model for each particular user that captures the type of information of interest to that user; 2) using machine learning technologies to adapt the model to changing user needs, and to learn from user feedback what type of information is of interest to the user at any given moment; 3) automatic query expansion (using ontologies) to help the user construct useful queries faster and retrieve pertinent information.**

I. Introduction

Efficient data retrieval from large databases and the World Wide Web is an important task that has to be performed routinely in applications [1, 2, 3, 4, 5] ranging from intelligence analysis to drug discovery. In the past decade with the explosive development of electronic data accessible through the World Wide Web, the amount of data available to users became prohibitively large. It is highly desirable to develop tools that allow filtering the data by removing the irrelevant pieces of information for a given user and transmitting those which are relevant in a given context of the human decision-making. Such tools need to help users finding relevant information quickly in a rapidly increasing volume of available data.

The term *user lens*, directly related to the above filtering needs, has been introduced in the data retrieval literature by Vogt [6]. Vogt et al. use that term to emphasize that each user could have their own lens that is employed whenever they utilize the system and is trained with the user's relevance feedback. This lens is a rough model of the cognitive processes of the user when he or she is creating the query or interpreting a document. Our aim is creating such user lenses, to aid human users when interacting with the database in the form of search and data retrieval sequence.

This work is supported by the Defense Threat Reduction Agency and the U.S. Army Medical Research and Materiel Command under Contract No. W81XWH-06-C-0001, awarded by the U.S. Army Medical Research Acquisition Activity. The views, opinions, and/or findings contained in this report are those of the authors and should not be construed as an official Department of Army position, policy or decision unless so designated by other documentation.

Anna L. Buczak (corresponding author), Charles Wan and Glenn Petry are with Sarnoff Corporation, 201 Washington Road, Princeton, NJ 08543 (tel. 609-734-2667, fax: 609-734-2662, e-mail: abuczak@sarnoff.com)

The paper starts with a description of our biomedical application, and then continues with the system operation and architecture. Next, we elaborate on the user model representation in a form of a concept map, and the user model adaptation algorithms. Then, we describe the results of quantitative evaluation of SmartPortal™ and finish by conclusions.

II. Biomedical Application

A user lens can be useful in various applications dealing with data search and retrieval. The application that we are interested in, biomedical data search, is the search in which the user is looking for information related to a certain disease or pathogen, methods to respond to some biological threat, vaccines and other countermeasures, molecular pathways, drugs leads, etc. In this type of application the main data source of interest is the Entrez database [7] that provides users with integrated access to sequence, mapping, taxonomy, and structural data. The journal literature is available through Entrez PubMed [7], a web search interface that provides access to over 11 million journal citations in MEDLINE and contains links to full-text articles at participating publishers' Web sites.

SmartPortal™ will lead to a reduced drug discovery and development cycle by providing mechanisms for the user to quickly access relevant information. It helps users while performing searches of biomedical literature, such as the publications available through PubMed and information on the World Wide Web. The search of the web is performed through yahoo API.

III. System Operation

SmartPortal™ allows the user to quickly find data of interest in all the databases linked to it. The key advantage provided by SmartPortal™ is that searches are **personalized** (i.e., tailored to each individual) and **adaptive** (i.e., search is automatically modified as most appropriate for the user's previous search activity). SmartPortal™ achieves these goals by filtering and/or reordering the information retrieved from data sources and presenting it in the order of importance for a given user. SmartPortal™ uses user modeling and machine learning technologies to achieve its goals.

SmartPortal™ (Fig.1) comprises: 1) the user model that contains a set of features describing items of interest to the user; 2) a recommender engine that based on user model makes suggestions on new items that are of high interest; 3) machine learning mechanism that adapts the user model and/or the recommender engine to reflect user's current interests and to make more accurate recommendations; 4) link to data sources.

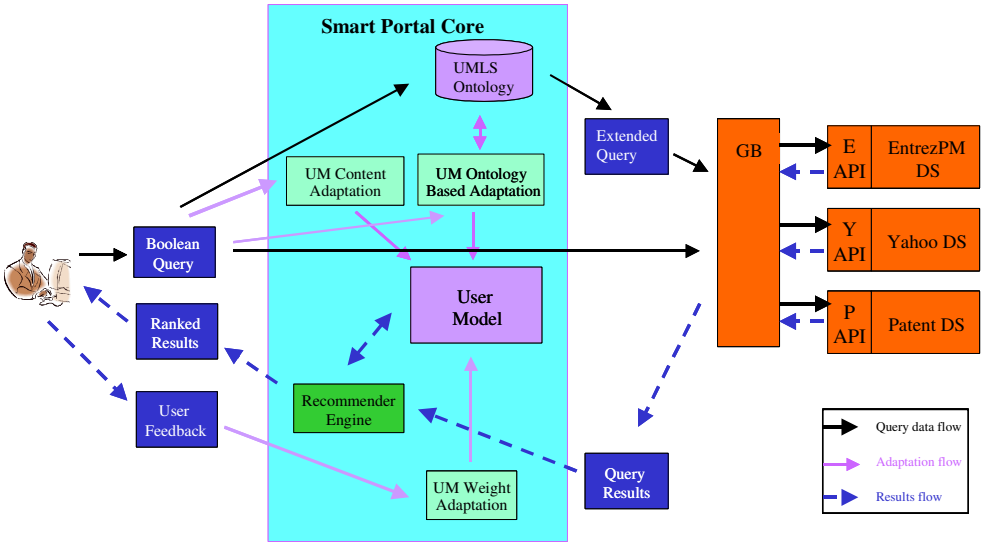

Figure 1 SmartPortal™ Architecture.

A. *Concept Map as a Representation of User Interests*

The user model is represented as a Concept Map (CM) [8, 9]. A Concept Map is a diagram consisting of concepts and relations. A concept denotes an object or an idea, and can usually be mapped to a node in an ontology. A relation denotes the relationship between two or more concepts. A binary relation is a triple {C1, C2, R}, where R is the relationship, C1 and C2 are the subject concept and object concept of the relationship, respectively. The Concept Map was extended to weighted concept map by Alonso & Li [10] and defined as a diagram consisting of concepts, relations, and their associated weights. Relations in a Concept Map can be named or unnamed. For example a named relation between concepts *streptomycin* and *plague* can be *cure*; meaning that streptomycin cures plague. When a Concept Map with unnamed relations is used, a relation between two concepts means that they are related is some fashion. This type of relationship could be hyponym, hypernym, antonym, etc. In some cases it is only important that the two concepts are related and the exact relationship does not matter.

The weight associated with a concept denotes the interest level the user has for this concept. Higher weight will be assigned to the concepts of most relevance to the user. The weight is a real value that can be positive or negative with larger positive value indicating higher relevance, and negative values representing items the user is not interested in.

An example Concept Map describing user's interests is shown on Figure 2. This Concept Map has 23 nodes and 20 unnamed relations. From the composition of this map, it seems that the user must be interested in botulinum since most of the concepts (16) are related to botulinum. The user might be interested in use of botulinum in bioterrorism since there are several concepts related to bioterrorism.

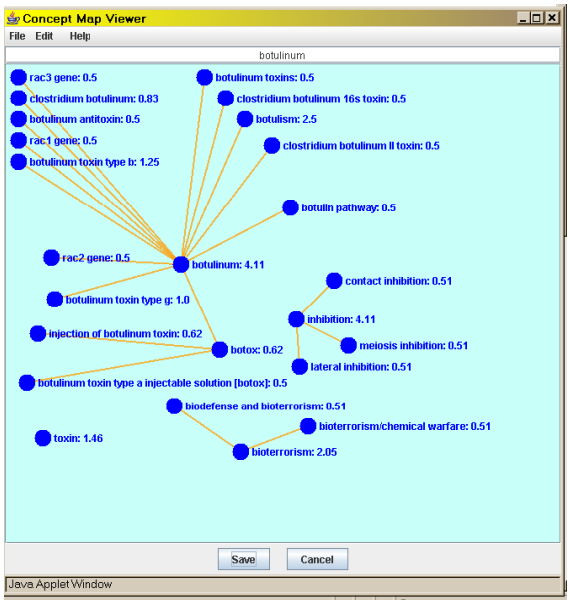

Figure 2. User Concept Map.

B. *Recommender Engine*

The Concept Map is used by the Recommender Engine to rank documents in an order reflecting their relevance to the user. The Recommender score S for a given document D is computed as:

$$S(D) = \sum_i \alpha_i f_i \qquad (1)$$

where α_i is the weight for a user model parameter i of the user model (its current value in the Concept Map) and f_i is the parameter relevance for i in doc D. f_i is the widely used in Information Retrieval measure: Term Frequency – Inverse

Document Frequency (TFIDF) [11] that measures how important a given word is to a document in a corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. Given a document corpus D, a word w, and an individual document $d \in D$, TFIDF is defined as:

$$TFIDF_d = f_{w,d} \cdot \log \frac{|D|}{f_{w,D}} \quad (2)$$

where $f_{w,d}$ equals the number of times w appears in d, $|D|$ is the size of the corpus, and $f_{w,D}$ equals the number of documents in which w appears in D.

We use Lucene [12] to compute the TFIDF of the document snippets returned by the search for a given query. The recommender engine orders all the documents returned by the query from the various data sources from the highest ranked to the lowest ranked. One of the pages with SmartPortal™ results is shown on Figure 3.

C. *User Model Adaptation Algorithms*

In order for it to better reflect the current interests of the user, SmartPortal™ performs three types of adaptation of the User Model (UM). The first type of adaptation, *UM Content Adaptation*, adds new concepts to the existing concept map based on the new queries that the user makes. Every time the user issues a query that has some new terms or expressions they are added to the Concept Map, if they are not already there. The weight of a newly added concept is set to an initial value, such as the average of weights of the terms already present in the Concept Map.

The second adaptation mechanism, the *UM Ontology-Based Adaptation*, is responsible for augmenting the User Model based on information contained in a bio-medical ontology. The ontology that we are using is the Unified Medical Language System® (UMLS) [13] that is composed of a Metathesaurus, a Semantic Network, and a Specialist Lexicon. SmartPortal™ is relying on the UMLS Metathesaurus and on the Semantic Network. The Metathesaurus is a multi-purpose, and multi-lingual vocabulary database that contains information about biomedical and health related concepts, their various names, and the relationships among them. It is built from the electronic versions of many different thesauri, classifications, code sets, and lists of controlled terms used in patient care, health services billing, public health statistics, indexing and cataloging biomedical literature, and /or basic, clinical, and

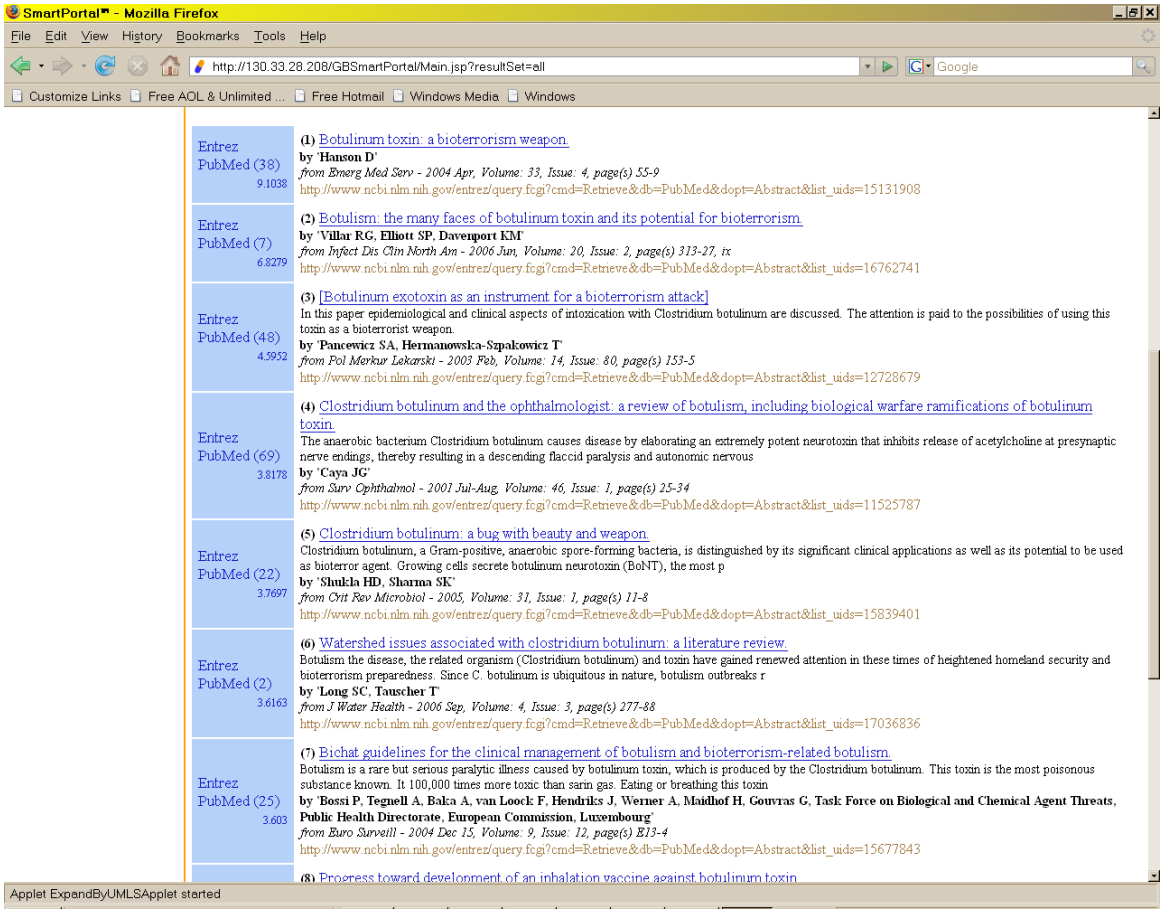

Figure 3. One of the 20 pages of SmartPortal results for the query '(botulinum OR "botulinum toxin") AND bioterrorism'.

health services research. The Metathesaurus contains concepts, concept names, and other attributes from more than 100 terminologies, classifications, and thesauri. Each concept in the Metathesaurus has a unique and permanent concept identifier (CUI).

The purpose of the Semantic Network is to provide a consistent categorization of all concepts represented in the UMLS Metathesaurus and to provide a set of useful relationships between these concepts. The Semantic Network is an upper-level ontology that provides information about the set of basic semantic types, or categories, which may be assigned to these concepts, and it defines the set of relationships that may hold between the semantic types. Currently the Semantic Network contains 135 semantic types and 54 relationships. Among the different types of relationships, the ones of special interest to us are child (CHD) and parent (PAR). The first of them means that a given term has child relationship to some other term in a Metathesaurus source vocabulary, and the second that it has a parent relationship in a Metathesaurus source vocabulary.

The UM Ontology-Based Content Adaptation algorithm operates on the query that the user entered. When a user enters a query each of the terms in the query is checked in the Metathesaurus and its CUIs are retrieved. The algorithm for building the tree checks the semantic type of each of the CUIs in the Semantic Network. The CUIs are shown to the user as a tree with the semantic type as a higher level node of the tree, and all CUIs that are of that type (see Figure 3). The user chooses which CUIs are relevant to include them into the expanded query. The original query is expanded with the CUIs that the user has chosen. Those CUIs are added with OR between them, and the original relation (AND, OR) between the terms of the query is kept. The same CUIs are added to the Concept Map with the weight equal to the average weight of concepts already present in the Concept Map. The links between concepts in the Concept Map are added if the concepts are related in the Semantic Network. The original Concept Map and the map after expansion by UMLS are shown in Figure 4.

The third adaptation algorithm, the *UM Weight Adaptation*, learns the weights for different concepts in the User Model based on user's feedback. Search results are presented to the user in the order from the highest to the lowest recommender score. An example of the results was shown in Figure 3. When the user clicks on the result that seems of interest, the full result is shown to the user and subsequently the user has the option to give feedback on what he/she has just read (Figure 5). There are four types of feedback: the first two "Relevant – Save the link" and "Relevant" mean that the user is interested in that document; in the first case also the link to the document is saved for future reference. The third one, "Neutral", means that the user has no opinion on that document; the last one "Irrelevant" means that the document is of no interest to the user in the context of the search. Based on the type of feedback given to the document, the UM Weight Adaptation algorithm adapts the concept weights in the user model, differently.

For the weight adaptation, SmartPortal™ is using a modified version of Tailored Winnow 2 (TW2) algorithm. We have chosen TW2 because of its tolerance to errors in user feedback and small updating complexity. TW2 [2] performs weight promotion for documents judged by the user as relevant, and weight demotion – for documents judged as irrelevant. TW2 maintains non-negative weights $(w_1, \ldots, w_n)$ for binary features $att_1, \ldots, att_n$, respectively. Initially all weights have a value zero. TW2 classifies documents whose vectors $x = (x_1, \ldots, x_n)$ satisfy $\sum_{i=1}^{n} w_i x_i > \theta$ as relevant, and all others as irrelevant. Let $w_{i,b}$ and $w_{i,a}$ denote the weight w_i before the current update and after, respectively.

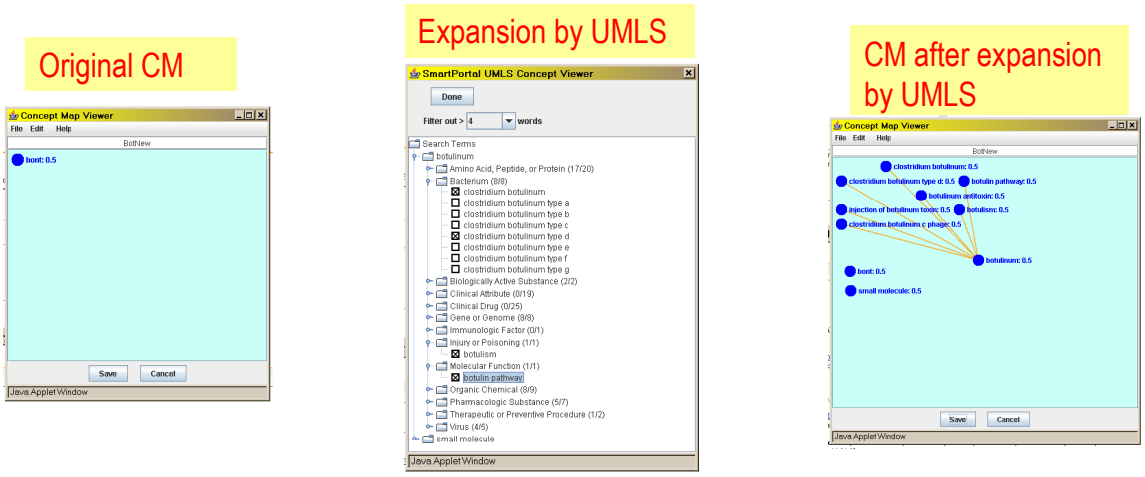

Figure 4. User Concept Map Before and after Ontology-Based Content Adaptation.

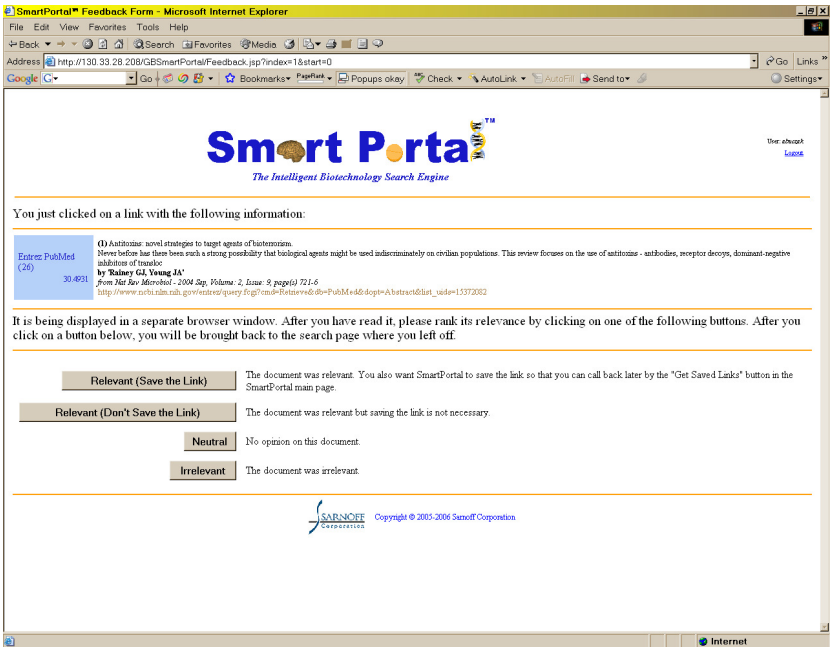

Figure 5. Types of feedback in the SmartPortal™.

When a user gives a positive feedback to a document classified incorrectly, weight promotion occurs as follows: Where $\alpha > 1$ is a promotion and demotion factor.

$$w_{i,a} = \begin{cases} w_{i,b} & \text{If } x_i = 0 \\ \alpha & \text{If } x_i = 1 \text{ and } w_{i,b} = 0 \\ \alpha w_{i,b} & \text{If } x_i = 1 \text{ and } w_{i,b} \neq 0 \end{cases} \quad (3)$$

When a user gives a negative feedback to a document classified incorrectly, weight demotion occurs as follows:

$$w_{i,a} = \begin{cases} w_{i,b} & \text{If } x_i = 0 \\ \dfrac{w_{i,b}}{\alpha} & \text{If } x_i = 1 \end{cases} \quad (4)$$

In our modification of TW2 algorithm we are starting from weights set to 0.5 and the weights can be positive or negative. Also we are not requiring feedback on 10 documents judged as relevant and 10 documents judged as irrelevant as WebSail does [2]. Rather, we ask the user to give feedback, only when he/she opens a document and only on that one document. This way the user is less burdened with giving feedback.

An example of how the weights are adapted based on user's feedback is shown in Figure 6. After feedback the weights of several concepts went up: botulinum, botulism, neurotoxin, bioterrorism, clostridium botulinum. The change of those weights causes a change in the order of documents recommended.

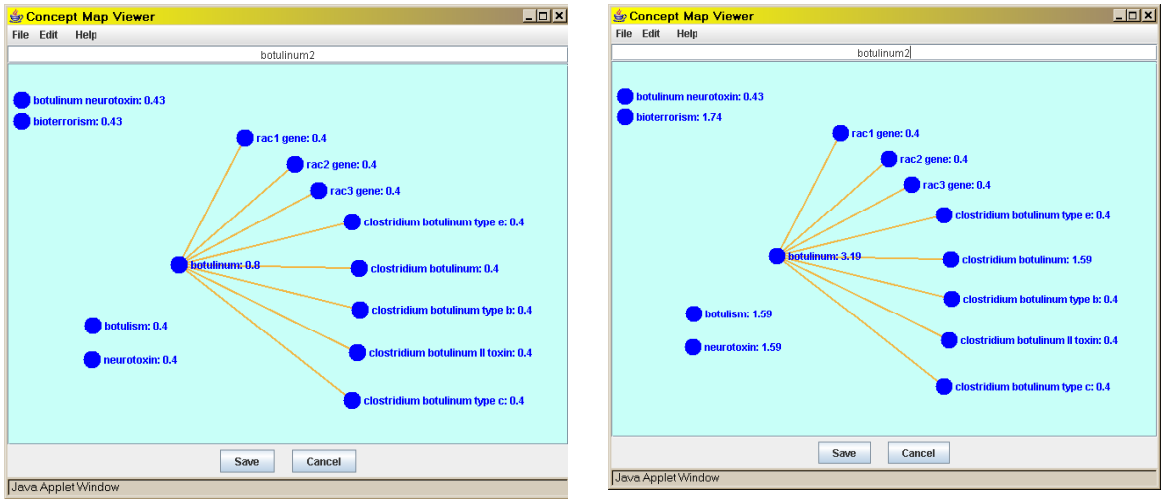

Figure 6. Original User Concept Map and the same map after a few feedbacks.

D. *Fostering Collaboration*

Collaboration in SmartPortal™ is achieved through sharing of Concept Maps between users. A specialist in a given discipline can develop a Concept Map, make it public and allow other users to import it and modify for their own purposes.

IV. EXPERIMENTAL RESULTS

In this section we describe our ongoing evaluation experiments with SmartPortal™. The goal of these experiments is to quantitatively assess the performance of SmartPortal™ in the context of biomedical search applications. The initial experiments described compare the performance of PubMed with and without SmartPortal™.

Similarly to [2], we are using Relative Recall as a measure of how well the system pinpoints the information of interest to the user. Relative Recall^n is defined as:

$$R^n_{\mathrm{Re}call} = \frac{R_n}{\min(n, R)} \qquad (5)$$

where R_n is the number of relevant documents ranked among top n search results; R is the total number of relevant documents among the list of m retrieved documents. In our experiments m was set to 100. The Relative Recall metric measures how well the system moves the documents of interest to the user (i.e. relevant documents) to the top of the list.

In order to determine which of the documents were relevant to the user, we developed click collection software that stores links to the documents that the user gave "Relevant - Save" or "Relevant" feedback to. These documents constitute the set R for each query.

In the experiments carried out, biomedical users (not involved in the project) were performing searches on botulinum, anthrax, Ebola and plague. They were instructed to perform searches in the usual fashion, i.e. to open and give feedback only as they would usually do (without rating all the documents). The users did not know that for SmartPortal™ system it is actually easier to provide relevant results when more feedback is given. When performing the searches the users judged an average of 6.2 documents in relevance feedback for each query.

The Relative Recall without and with SmartPortal™ (Figure 7) shows a very promising performance of our method. SmartPortal™ achieves 0.658, 0.786, and 0.942 Relative Recall of the type 5, 10, and 20, respectively. The same Relative Recall numbers for native PubMed (i.e. without SmartPortal) are only a pale 0.45, 0.538, and 0.673, respectively.

The percentage improvement in Relative Recall of searches with SmartPortal™ over searches without it (i.e. PubMed only) is 40-46% (see Figure 8). This is a very prominent improvement, especially given the fact that the user gave an average of 6.2 feedbacks only.

It is important to note that the method of computing results that we used is actually unfavorable to SmartPortal™. The reason is that we are computing the value of Relative Recalls for all the queries starting when no feedback was given to the system yet. As such at the beginning SmartPortal™ did not have the time to learn the user preferences yet. In our next set of experiments, we will allow the SmartPortal™ to stabilize the user model first by getting a pre-specified number of feedbacks (e.g. 5) and compute the Relative Recall metrics staring from that point onward only.

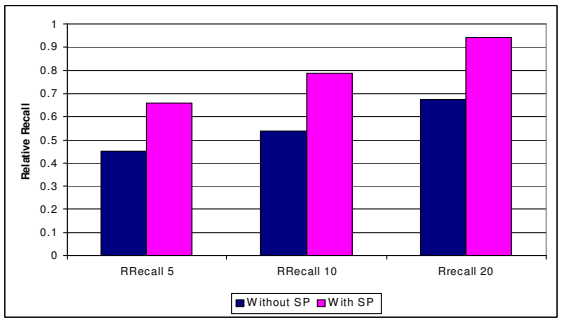

Figure 7. Relative Recall without and with SmartPortal™

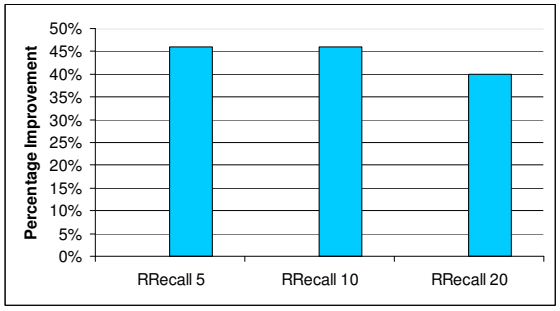

Figure 8. Percentage Improvement in Relative Recall with SmartPortal™ over PubMed only.

We have also performed an initial investigation of how much SmartPortal™ improves the results for biomedical searches performed on the web through yahoo (our second data source). The initial results are very similar to the ones reported here for PubMed. However, more experiments with yahoo need to be performed before those results can be published.

V. RELATED WORK

Recommender systems (user lenses) are a form of artificial intelligence technology that provides the user with personalized suggestions about the items of interest to the individual, based on previous examples of the user's likes and dislikes. Recommender systems can suggest information of any type: web pages, news articles, books, movies, TV shows, images, news articles, etc. [3, 4, 5, 14].

User models in recommender systems range from very simple to sophisticated. An example of a simple user model is a list of items that the user found interesting (set of ids for documents in case of a collaborative recommender), possibly with the addition of how many times a given document was opened, saved, etc. It could also contain a list of items that were of no interest to the user. A more sophisticated user

model can contain some content information such as stems of the words used in user queries, or stems of the words employed in titles and abstracts of documents accessed by the user; again these could be augmented by relative frequency of the stems. Some other information that could be used in the user profile is the metadata describing the items of interest (such as genre in case of TV-shows or movies) [5]. User models represented as Concept Maps belong to the more sophisticated methods of representing users interests. They were introduced by Alonso and Li [10] for the use in recommender systems for intelligence analysts.

System adaptivity can be achieved by adjusting user models themselves or by adapting the way the recommender engine works. Sometimes both methods are used in a recommender system. Most of the methods adjust the model weights. These methods range from least-squares [15], pseudo-inverse methods, gradient descent, conjugate gradient [16], bubble-up [10], Tailored Winnow 2 [2], Genetic Algorithms [17] to Simulated Annealing [18]. The adaptation mechanism needs to be fast, reliable, and preferably work in an incremental fashion.

In SmartPortal™ the adaptation mechanisms not only adapt the User Model weights but also perform the adaptation of concepts in the profile based on the information in user queries and in the biomedical ontology. This tri-fold user adaptation leads to very good performance results.

VI. Conclusions

Recommender systems are an important artificial intelligence technology for helping users deal with information overload. SmartPortal™ helps biomedical users finding quickly the information of interest in huge data bases. It employs user modeling and machine learning technologies to provide the user with personalized suggestions about the items of interest to the individual, based on previous examples of the user's likes and dislikes. The user model created by the system is a Concept Map that captures the concepts the user is interested in and relations among them.

Machine learning methods for adapting user profiles in a fast, reliable, and preferably incremental fashion are important research areas. SmartPortal™ achieves the personalization of user results by three adaptation mechanisms 1) UM Content Adaptation adds new concepts to the existing concept map based on the new queries that the user makes; 2) UM Ontology-Based Adaptation augments the User Model based on information contained in a bio-medical ontology; 3) UM Weight Adaptation learns the weights for different concepts in the User Model based on user's feedback.

The adaptation mechanisms developed for SmartPortal™ are very powerful and result in high Relative Recall metrics in comparison to the same searches performed on PubMed. The methodology and algorithms developed are general in nature and are applicable to other domains than biomedical. The only element that needs change in case of a different domain is the ontology.

Acknowledgment

The authors would like to thank Dr. Lynne Gilfillan for suggestions concerning the UMLS® ontology.

References

[1] Perkowitz, M., Etzioni, O. (2000) "Towards adaptive Web sites: Conceptual framework and case study," *Artificial Intelligence* 118, 245–275.

[2] Chen, Z., X. Meng, B. Zhu, R. Fowler, (2002) "WebSail: From On-line Learning to Web Search", *Knowledge and Information Systems*, 4(2):219—227.

[3] Breese, J.S., D. Herlocker, C.Kadie, (1998), "Empirical Analysis of Predictive Algorithms for Collaborative Filtering", *Fourteenth Conference on Uncertainty in Artificial Intelligence*, Madison, WI, Morgan Kaufman.

[4] Mooney, R. J., Roy, L., "Content-Based Book Recommending Using Learning for Text Categorization", *Fifth ACM Conference on Digital Libraries*, pp. 195-240, San Antonio, TX, June 2000.

[5] Buczak, A.L., J. Zimmerman, K. Kurapati, (2002), "Personalization: Improving Ease-of-Use, Trust and Accuracy of a TV Show Recommender", *International Conference on Adaptive Hypermedia and Adaptive Web Based Systems, 2nd Workshop on Personalization in Future TV*.

[6] Vogt, C.C., Garrison W. Cottrell, Richard K. Belew, and Brian T. Bartell (1999) "User Lenses – Achieving 100% Precision on Frequently Asked Questions," *User Modeling UM'99*.

[7] Entrez PubMed http://www.ncbi.nlm.nih.gov/entrez/query.fcgi

[8] Ausubel, D. P. *The Psychology of Meaningful Verbal Learning.* New York: Grune and Stratton, 1963.

[9] Novak, J. D., and Gowin, D.B. (1984) "*Learning How to Learn*." New York and Cambridge, UK: Cambridge University Press.

[10] Alonso, R., and Li, H. (2005) "Model-Guided Information Discovery for Intelligence Analysis", *14th Int. Conf. on Information & Knowledge Management, CIKM'05*, Bremen, Germany.

[11] Salton, G. and McGill, M. J. (1983) *Introduction to modern information retrieval*. McGraw-Hill.

[12] Lucene http://lucene.apache.org/java/docs/

[13] UMLS UMLS® KNOWLEDGE SOURCES, November Release 2005AC DOCUMENTATION, http://www.nlm.nih.gov/research/umls/umlsdoc.html

[14] Billsus, D., M.J. Pazzani, (1998), "Learning Collaborative Information Filters", *Fifteenth International Conference on Machine Learning*, Wisconsin, USA, pp.46-54.

[15] Zhang, T., Iyengar, V. (2002). "Recommender Systems Using Linear Classifiers", J. of Machine Learning Research, 2.

[16] B.T. Bartell, G.W. Cottrell, R.K. Belew "Optimizing Parameters in a Ranked Retrieval System Using Multi-Query Relevance Feedback," J. Am. Soc. Inf. Sci. 46, 254-271, 1995.

[17] Sheth, B., Maes, P. (1993) "Evolving Agents for Personalized Information Filtering." 9th Conf. Artificial Intelligence Applications.

[18] Jansen, B.J. "Using Simulated Annealing to Prioritize Query Results," ACM Conf. Computer Science Education, 1997.