Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

Do Users Appreciate Novel Interface Features for Literature Search?

A User Study in the Life Sciences Domain

Anne Schneider[§], Rico Landefeld, Joachim Wermter, Udo Hahn
Jena University Language and Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena
Jena, Germany
http://www.julielab.de

***Abstract*— Faced with the challenges to design an easy-to-use, immediately comprehensible and powerful expert-user interface to search very large document collections in the life sciences, we developed several system prototypes. Their main features were faceting of the domain vocabulary for browsing and searching, flexible search-state-dependent drilling of the terminological hierarchy, dynamic query term auto-completions, and highlighting of matched terms (including synonyms and spelling variants). Under lab conditions we then evaluated these features in several task-based scenarios using camera recordings, thinking-aloud protocols and questionnaires. The results reveal that faceting and highlighting were very well received, while auto-completions seemed less important or were misconceptualized as spelling aids.**

Keywords – Information Search and Retrieval: Search process, User Interfaces: User-centered design, Life and Medical Sciences

I. INTRODUCTION

User interfaces of information retrieval systems have to mediate between the search problem implicit in the user's mind and the explicit content representations of a (usually large) document collection. The user's literature search problem is often only weakly structured - users can rarely exhaustively specify their search topic by enumerating all relevant search terms and, in addition, are often unaware of the terminological complexity and intricacies holding in their area of interest [1].

To cope with these problems, for many scientific fields, different forms of terminologies (thesauri, classifications, etc.) are available, which structure relevant domain terms through semantic relations (e.g., is-a or part-of), provide sets of semantically equivalent terms (i.e., synonyms and spelling variants), and thus end up normatively as so-called "controlled" vocabularies potential users have to adapt their search topic to.

The life sciences, a terminologically large and complex domain, come with an impressive amount of pre-structured terminological knowledge, as witnessed by resources such as the Medical Subject Headings (MeSH)[1] or the UniProt[2] protein repository, each of which amount to many thousands of (hierarchically structured) terms. Incorporating such large-scale and complex terminologies into user interfaces to support the query formulation process by the users results in enormous problems in terms of conceptual orientation and navigation strategies, in particular, when users run searches on collections with millions of documents such as Medline.[3] Additional challenges arise from the lacking link between controlled terminologies describing content at a meta level and the lexical intricacies and complexities of literal natural language use in documents (free-text search).

One example for such lexical complexity is the heavy use of abbreviations in the biomedical research literature. Since biomedical entities have often very long names, authors introduce abbreviations to save time and space. But abbreviations cause a high degree of ambiguity [2] - a hard challenge for literature search engines and their users. For example the abbreviation "APC" may stand for the protein "protein APC", the cell "antigen-presenting cell" (a blood cell), the chemicals "anaphase-promoting complex", "aphidicolin" and "allophycocyanin", or the diseases "adenomatous polyposis coli" and "atrial premature complexes".

Even if the ideal user were able to specify the search query in a terminologically perfect way, the form textual information is phrased or made searchable by an information retrieval system (e.g., through natural language free-text terms or terms taken from controlled vocabularies) is hard to anticipate and thus difficult to target. Therefore, interfaces carry a heavy burden to bridge both the user's and the system's perspective in order to contribute to the user's satisfaction with the outcome of a literature search.

The standard life science user, a clinician or a biomedical researcher, neither has complete knowledge about the methodological backbones of search engines, nor are such users fully aware of the terminological complexity of the domain. Hence, user interfaces for this audience have to be self-contained, self-explanatory, have to support established

[§] Anne Schneider is now affiliated with the Department of Computer Science at Trinity College, Dublin, Ireland (schneia@tcd.ie).

[1] http://www.nlm.nih.gov/mesh

[2] http://www.uniprot.org

[3] http://www.ncbi.nlm.nih.gov/sites/entrez

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

search metaphors and styles and must yield, most importantly, value-adding results for their daily work.

These observations have heavily influenced our system design. First, to cope with the terminological richness of this domain, facets [3] are supplied at a dynamically modified level of depth to orient the user what area of terminology is currently being explored by the search. Second, the consequences of term refinement are immediately displayed at the facet level by showing conceptually more specialized terms at the finer, drilled-down level of conceptual granularity. Third, suggestions for terms are provided by an auto-completion facility which supports the user's query formulation to avoid spelling mistakes and direct the user to suitable terms from a controlled vocabulary. Fourth, within the textual context of retrieved documents, matched terms (including synonyms and spelling variants) are highlighted.

While these features, at the first sight, might seem intuitively plausible, it remains to be shown whether they are really appreciated by the targeted user group. Therefore, we conducted two user studies - the first one was rather exploratory using paper mock-ups and informal conversations about design decisions, the second one was much more structured through questionnaires and carefully recorded and analyzed exploiting thinking-aloud protocols taken under lab conditions. In this paper, we report on the outcome of these studies and the implications they had on (re)designing the search interface for our semantic search engine SEMEDICO.

II. SYSTEM DESCRIPTION

Our search interface complements the now classical ranked document list interface with the faceted search approach. It currently contains about 20 categories (facets) with over 900,000 hierarchically organized concepts. This term set is based on semantic metadata which were automatically acquired via text mining, but we also grabbed Medline-supplied bibliographic metadata (such as author names, publication dates, etc.). Crucially, both the search interface and the text mining engine use a common vocabulary which has been compiled from different sources (see below). From a design perspective, our system is built upon the facet metaphor from the FLAMENCO system [3], while it uses a standard layout inspired by major web search engines like GOOGLE or YAHOO!. Still, our interface has been considerably augmented by new features to handle the vast size of domain knowledge and life science terminologies.

A. User Interface

A typical literature search starts with the query formulation where the user's information problem is translated into a set of query terms. Here, the user faces the difficulty to guess which terms might literally occur in relevant articles. Our system facilitates this task by showing – dependent on already typed input – possible term auto-completions, separated in categories, along with other synonyms (see Figure 1) for the already typed input. The underlying assumption is that this may help avoid parallel searches in other resources, if the user is uncertain about the correct spelling or naming. If a user selects a term from the auto-completion list, the system maps it to the underlying terminology, immediately triggering a "controlled" term search and thus easing query term disambiguation. If no term is selected, a free text search is run instead, using the query terms "as is".

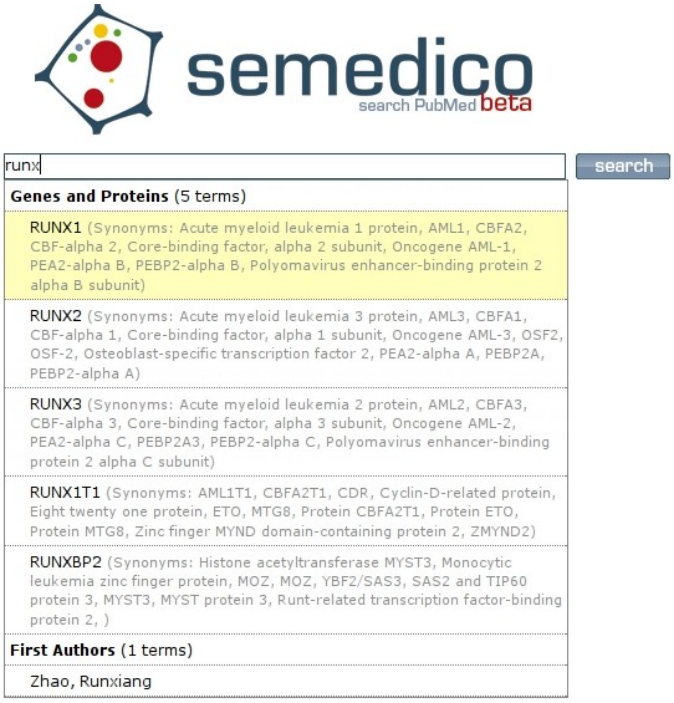

Figure 1. System-supplied auto-completions for user input "**runx**" on the start screen.

To facilitate browsing, in particular, when the user's information problem is only vaguely described, our search interface displays the terminological metadata and the textual context of the user query to suggest possible search directions and relationships between terms. More concrete, in addition to a ranked document list, the system displays a term list split into category facets and further divided into facet tabs (see Figure 2). These terms occur in the document hit list and are ordered by their frequency (i.e., in how many documents a term occurred). Only the three most frequent terms are displayed in each facet (note that this list must be updated whenever the query is modified). A mouse hover shows a description of the term and, if available, also synonymous terms. In this way, facet terms are related to the user query and create a semantic context based on the result set. The textual context of the query terms is shown in the document list with text snippets containing the highlighted query terms or their synonyms in a keywords-in-context (KWIC) style. We use a list-based KWIC visualization as favoured by Aula [4].

Because users under time pressure typically inspect only few top-ranked documents on the hit list [5], the faceted terms can be used to expand or refine a query in an informed way in order to reduce the result set size in a convenient way. In contrast to manual query reformulation, this never leads to

empty result sets. The facet terms are arranged in hierarchies for drill-down during query refinement (bibliographic facets are inherently flat). Each click on a facet term updates the query and starts a new controlled term search. Each step is also recoverable and hierarchies can be drilled up. Thus, facets serve as a flexible means to navigate the document space according to the user's information state.

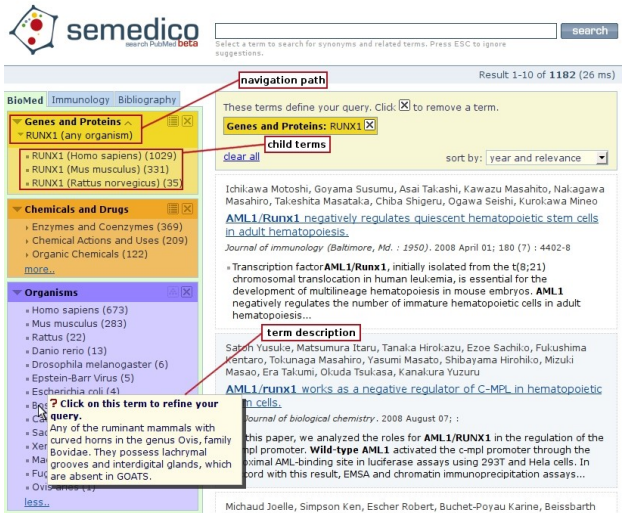

Figure 2. The result list for the query `**RUNX1'** with facet boxes on the left side. The `Genes and Proteins' facet box is drilled down and shows child terms of "**RUNX1**".

On the one hand, facets may improve the effectiveness of a search (which still has to be shown empirically), on the other hand, they also add cognitive complexity load to the interface. Furthermore, biomedical researchers work in different sub-domains and may thus only need a limited, specially tailored selection from all available facets. Accordingly, our facets are adaptable to deal with this problem in that they can be hidden and collapsed on demand. Their term lists can be filtered and navigated, if they become too large.

Since some term hierarchies, especially from the MESH, are quite deep and, hence, are likely to trigger long drill-downs, facets can be displayed in a flat modus which only features the most specific terms, the leaves of the term hierarchy (Figure 3).

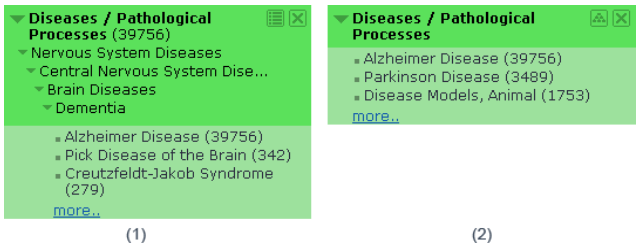

Figure 3. The **Diseases** facet in (1) hierarchic mode and in (2) flat mode

B. *Text Mining Engine and Controlled Vocabularies*

All semantic metadata accessible in SEMEDICO are automatically generated by the JCORE text mining engine [6]. Technically speaking, this system is a JAVA Web server application based on a LUCENE-derived[4] search index. The engine processes MEDLINE texts by recognizing and indexing (via the LUCENE search engine) terms from crucial biomedical terminologies, i.e., MESH and UNIPROT. MESH is a comprehensive and well-curated terminology with approximately 25,000 biomedical terms and 140,000 (non-curated) chemical substances covering various sub-domains – they range from molecular biology and chemistry over translational medicine to applied health care. UNIPROT is the most comprehensive terminology for proteins across several species (350,000 entries). The content-related facets (see Table 1) were compiled from the more general biomedical categories and the immunology-specific categories of MESH as well as from the protein entries of UNIPROT. The selection process was supported by immunologists and other biomedical practitioners in order to ensure proper domain coverage. These terminologies also contain hierarchies of varying degrees (ranging from depth 2 to depth 10), thus making them apt to browsing. Technically, this is accomplished by adding all parent terms to the search index.

TABLE 1. MESH AND UNIPROT BASED FACETS WITH NUMBER OF TERMS

Biomedical Facets	Terms	Immunology Facets	Terms
Genes and Proteins	113,898	Immunoglobulins / Antibodies	695
Signs and Symptoms	73	Transplantation	32
Cellular Processes	93	Hematopoietic Progenitor Cells	36
Gene Expression	38	Immune Processes	67
Organisms	26	Blood Cells	42
Chemicals and Drugs	141,469		
Investigative Techniques	717		
Therapies and Treatments	431		
Diseases / Pathological Processes	4,186		

Bibliographic metadata already provided in MEDLINE such as author and journal names, publication dates, etc. is also added to the pool of metadata. In order to keep computation of faceted result lists fast, we had to limit the amount of metadata which is dominated by the bibliographic metadata, especially the authors. For this reason, we consider only the first and the last author for facet computation and only if they have at least three publications. This makes sense inasmuch both first and last author are considered to be the most prominent positions in the authors' list of biomedical publications.

III. EVALUATION

Our first study was conducted with paper mock-ups of a preliminary interface draft complemented by thinking-aloud

[4] http://lucene.apache.org

protocols [7]. The purpose of this pre-test was to help assess the understandability of two key interface features, *viz*. auto-completion of query terms and the document hit list complemented by facets. Furthermore, we tested three different designs of facet drill-downs. All 30 subjects (Bachelor students, PhD students and post-docs) had a biology or medicine background and were required to have expertise in the use of PUBMED.

For the auto-completion feature, three printed screenshots were presented and subjects were asked what kind of interaction was triggered in the previously seen screenshot. All understood what auto-completion was about. Then, a screen shot with a document hit list along with facets was presented. Two thirds grasped the role of facets immediately and one additional quarter after they had been shown the mouse hover help text for a facet term. Finally, detailed screenshots were shown with a facet in initial state and in drilled-down state. The latter ones showed the navigation path either as a horizontal list (the original FLAMENCO design), a vertical list or an indented tree-like list (see Figure 4). The indented tree-like navigation path was preferred over the other variants. The findings and feedback from this study were the basis of our original interface design.

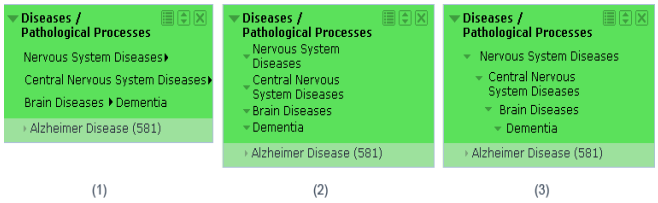

Figure 4. The facet designs presented to the participants. All designs show the **Diseases** facet (1) original horizontal list-like FLAMENCO design (2) the hierarchy as list (3) the hierarchy as indented tree-like list.

The goal of the second user study was to examine the usage of our interface features in real-life search scenarios and to test under much more rigorous conditions the system's usability as a whole in terms of user satisfaction, efficiency and error rates. This study was conducted with 14 participants, all being biomedical professionals (PhD students and post-docs) from a major German university hospital with an associated medical graduate school (Medical School Hanover). To get an impression of their Internet and domain search experience, [5] the participants were asked about the time they usually spent on private surfing and domain specific searches. 8 of the participants stated to surf less than 10 hours, 5 between 10 and 15 hours and 1 participant more than 15 hours per week. 4 stated that they spent less than 2 hours with domain specific searches, 6 between 2 and 6 hours and 4 more than 6 hours. Two typical search scenarios were randomly chosen from the TREC Genomics [8] data set and were presented to the participants in a lab setting. Before the users had to accomplish the tasks they were given ten minutes to become acquainted with the interface's "look and feel", without any further guidance.

The first search task was representative of a typical biomedical information problem (how to find information about the relationship between a certain gene and a disease), while the second one was designed to test more specific design decisions, i.e., whether subjects would recognize the different facet tabs and use them properly. The participants were instructed to stop, once they had found a relevant document and to justify why this document answered the question. Thinking-aloud protocols were taken and recorded during both tasks to gain further insights into the users' understanding of the system. All sessions were recorded with a screen camera. After task completion, the users were introduced to the features of the interface, in particular to the facets and the auto-completion function but also to term highlighting and controlled term mapping (which automatically handles synonyms). They were then asked to state whether these functions had been useful or not. Answers could be positive, neutral or negative; the results are displayed in Table 2.

TABLE 2. RESULTS OF THE FEATURE ASSESSMENT

Feature	Positive	Neutral	Negative
auto-completion	65%	21%	14%
facets	100%	0%	0%
term highlighting	100%	0%	0%
synonym handling	60%	0%	40%

Whereas facets and term highlighting were unanimously judged positively, the picture is less clear with respect to auto-completion and synonym handling (in terms of term mapping), although a clear majority still deems them positive. Table 3 shows how many of these features were actually used both during the initial "look and feel" phase and during the task completion phase. These results were obtained by analyzing the screen camera-recorded sessions.

TABLE 3. USAGE OF FEATURES DURING THE ``LOOK AND FEEL'' AND THE TASK COMPLETION PART OF THE STUDY. USAGE FREQUENCY IS GIVEN AS THE PERCENTAGE OF USERS WHICH USED A CERTAIN FEATURE.

Feature	"Look and Feel"	Task Completion
auto-completion	60%	30%
facet tabs	9%	80%
facets	10%	40%

Interestingly, in contrast to a 100% positive judgment in what concerns its functionality, auto-completion was actually used only 60% of the time during the "look and feel" phase and, even more surprising, only 30% during the completion of the two tasks. The usage statistics for the other 100%-positively judged 'facet' feature (as well as their arrangement in tabs) indicates that they were employed much more often during task completion.

Finally, participants were asked to fill in a Software Usability Measurement Inventory (SUMI)-based [9] questionnaire to evaluate the overall satisfaction, the

[5] We assume that participants with high personal Internet exposure are more knowledgeable in the use of auto-completion features and some form of facets.

effectiveness and efficiency as well as the control and error rate experienced by the user. Each of these usability features were elicited by five questions which could be answered positively, negatively or neutrally (see Table 4).

TABLE 4. RESULTS OF THE SUMI-BASED QUESTIONNAIRE

Feature	Positive	Neutral	Negative
overall satisfaction	69%	30%	1%
efficiency / effectiveness	59%	37%	4%
control / error rate	53%	34%	12%

While the majority of answers on all three SUMI dimensions were positive, the amount of negative assessments was fairly low, with the strongest negative reactions relating to control and error rates.

IV. RELATED WORK

The DYNACAT [10] system organizes search results in dynamically generated categories corresponding to predefined types of user queries. Its focus is to support lay users (e.g., patients) to get answers to typical non-expert questions. The category generation is based on MESH terms of MEDLINE abstracts. The evaluation showed that users find answers quicker than with a classical ranked document list interface.

Similarly, Hearst [11] describes a MESH-based search interface for biomedical texts as an example for a metadata-driven browsing interface. Both systems, however, use only a small subset of the available MESH terminology.

GOPUBMED [12] is an interface which categorizes the search results of the PUBMED search engine with the GENE ONTOLOGY [6] to enable browsing and query refinement. Unfortunately, studies considering usability aspects of that system have not been conducted up until now.

Divoli *et al.* [13] tested several different mock-up screen designs. They found that biologists like to see additional suggestions for gene/protein names when searching for biomedical literature. The lack of intuitively plausible and comprehensible search-improving functionalities appears to be a problem for almost all current biomedical document retrieval systems.

Our system responds to some of these desiderata. Unlike DYNACAT we focus on expert users and thus refrains from any expert-layman vocabulary mapping problem. It excels on very large and diverse vocabulary systems covering large portions of the life sciences domains. Its design is also dedicated to support users by easily comprehensible search facilities and thus improve search results by intelligible interface design.

V. DISCUSSION AND CONCLUSION

We considered various design aspects for an expert-level user interface in the life sciences domain. Its terminological complexity and the kinds of information access problems to be solved diverge considerably from the mainstream of "everyday" GOOGLE-style searches. We focused on various selected search features – faceted browsing and navigation, query term auto-completions, textual term high-lighting (including their synonyms and spelling variants) –, integrated them in interface prototypes and assessed them empirically in two user studies.

The results of a first (paper mock-up) study backed up our hypotheses that auto-completion and facets were well understood by expert users. The second study under much more controlled lab conditions showed that facets and term highlighting where considered very useful. Furthermore, the separation of facets into tabs was understood to a large extent and seems to be a suitable approach to handle even large numbers of facets. Also the auto-completion feature and the automatic term mapping for synonyms were considered as useful by two-thirds of the participants. These results, however, were not reflected by the usage of these features in real search tasks – possibly a limitation of our lab setting where users only spent a few minutes getting used to the system before task completion was required.

In addition, we also observed that the different system responses to (and result sets for) free text search and term search were a real pitfall and may have influenced the usability assessments (see Table 4). This was caused by the fact that term search was only triggered by selecting auto-completion and by clicking on the facet terms. We observed that some users did not even recognize the auto-completion list because they were looking at the keyboard while typing their queries, whereas others noticed the list and used it as a query spelling aid. This seems to suggest that the auto-completion feature alone is not sufficient for query disambiguation and term search because, after all, typed keyword queries are the preferred interaction style and must be supported. Thus, as users lack a profound understanding of the difference between "controlled" term and free-text search, these search styles definitely need a much tighter integration and far more directed computational support.

Nevertheless, the adaption of facets for biomedical literature search along with text mining-driven term (synonym and variation) resolution appears to be a promising avenue to enhance the document search capabilities in the life sciences.

At the moment our interface undergoes a re-design to include the findings of the user studies in the next prototype. Automatic query disambiguation and free-text search with a complete coverage of all document attributes have already been deployed. We plan to rerun the last study with this refined version of our system to test our claim that these changes will improve the usability. If the system reaches a reasonable usability performance, we plan to conduct a longitudinal study to observe the usage of the extended features in realistic document retrieval scenarios. Particularly, the many facet customization-related features need further examination.

ACKNOWLEDGMENT

This work has been supported by the German Federal Ministry for Research and Education (BMBF) in the STEMNET

[6] http://www.geneontology.org/

project (www.stemnet.de) under contract no. 01DS001 ("Wissensmanagement für die Stammzellbiologie").

REFERENCES

1. G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The vocabulary problem in human-system communication. *Communications of the ACM*, pages 964-971, *30* (11), 1987.
2. H. Liu, Y. A. Lussier and C. Friedman. A study of abbreviations in the UMLS. In Proceedings of the AMIA Symposium 2001, pages 393-397, 2001.
3. K.-P. Yee, K. Swearingen, K. Li, and M. Hearst. Faceted metadata for image search and browsing. In CHI '03: Proceedings of the 2003 SIGCHI Conference on Human Factors in Computing Systems, pages 401-408, 2003.
4. A. Aula, Enhancing the readability of search result summaries, In Proceedings of the Conference HCI 2004: Design for Life, pages 1–4, 2004.
5. B. J. Jansen and U. W. Pooch. A review of Web searching studies and a framework for future research. *Journal of American Society for Information Science and Technology*, pages 235-246, *52* (3), 2000.
6. U. Hahn, E. Buyko, R. Landefeld, M. Mühlhausen, M. Poprat, K. Tomanek, and J. Wermter. An overview of JCoRe, the JULIE lab UIMA component repository. In Proceedings of the LREC'08 Workshop 'Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP', pages 1-7, 2008.
7. J. Nielsen, Usability Engineering. Morgan Kaufmann, pages 195-200, 1993
8 W. Hersh, A. Cohen, J. Yang, R. T. Bhupatiraju, P. Roberts, and M. Hearst. TREC Genomics track overview. In Proceedings of the 14th Text REtrieval Conference (TREC 2005), National Institute of Standards and Technology, 2005.
9. E. P. W. M. van Veenendaal. Questionaire based usability testing. In Conference Proceedings of the European Software Quality Week, 1998.7.
10. W. Pratt, M. A. Hearst, and L. M. Fagan. A knowledge-based approach to organizing retrieved documents. In Proceedings of the 16th National Conference on Artificial Intelligence, pages 80-85, 1999.
11. M. Hearst. Next generation web search: Setting our sites. *IEEE Data Engineering Bulletin*, pages 38-48, *23* (3), 2000.
12. A. Doms and M. Schroeder. GoPubMed: Exploring PubMed with the Gene Ontology. *Nucleic Acids Research*, pages 783-786, *33*, 2005.
13. A. Divoli, M. Hearst, and M. Wooldridge. Evidence for showing gene/protein name suggestions in bioscience literature search interfaces. In PSB 2008 – Proceedings of the Pacific Symposium on Biocomputing, pages 568-579, 2008.