Search interface feature evaluation in biosciences

Anna Divoli
University of Chicago
annadivoli@gmail.com

Alyona Medelyan
Pingar
alyona.medelyan@pingar.com

ABSTRACT
This paper reports findings on desirable interface features for different search tasks in the biomedical domain. We conducted a user study where we asked bioscientists to evaluate the usefulness of autocomplete, query expansions, facetted refinement, related searches and results preview implementations in new pilot interfaces and publicly available systems while using baseline and their own queries. Our evaluation reveals that there is a preference for certain features depending on the search task. In addition, we touch on the current pain point of facetted search: the acquisition of faceted subject metadata for unstructured documents. We found a strong preference for prototypes displaying just a few facets generated based on either the query or the matching documents.

Categories and Subject Descriptors
D.2.2 [**Software Engineering**]: Design Tools and Techniques – *user interfaces.*

General Terms
Measurement, UI Design, Human Factors

Keywords
Search user interfaces, qualitative user-study

1. INTRODUCTION
Interface features are elements of search user interfaces, which facilitate the search process. Examples of such features are auto-complete and query expansion suggestions, facetted navigation, and document surrogates in search results previews. Vast research exists on the usefulness of interface features on the web [6], although less so in the biomedical domain.

We identified and addressed two open questions in studies of search user interfaces: Queries and search tasks can be classified into categories [7], but how should interface differ depending on the task? Facetted navigation has been demonstrated useful for search in structured data [11], but which approach to generating facet categories for unstructured documents works best?

We conducted a qualitative user study to systematically evaluate techniques to compute and present individual search features. In extended interviews, bioscientists rated the usefulness of features in common interfaces on baseline and their own queries, which we classified as browsing, gathering information, and search for facts. Side-by-side comparison allowed us to identify clear preferences in interface features depending on these tasks.

After an overview of state-of-the art in computing the features and user-study outcomes, we discuss search tasks in the biomedical domain. We then turn to the experiment and participating systems. Finally, we discuss how the study was conducted and its findings.

2. RELATED WORK
Related work can be grouped based on the studied features.

Autocomplete (autosuggest) provides dynamic search suggestions as the user types the query. Commonly suggestions originate from existing query logs, but could also be computed using biomedical terminology resources [9]. Users intuitively interact with autocomplete, increasingly so with time [1]. It is recommended to display results before suggesting new queries and to compute suggestions starting from all existing terms [6].

Query expansion suggests alternative query terms when users' guesses result in only a few or incorrect results. Such terms can be computed from query logs or thesauri. Users react positively to search expansions as long as their number is limited [6].

Facetted refinement helps to narrow down results based on a dimension (facet) of the searched item. Computing facets for products, accommodation or any structured data search is straight-forward; however search in unstructured documents limits facets to pre-existing metadata such as author, subject headings or social bookmarking tags, and such may not exist. Various approaches address this shortcoming. Named entity extraction can generate facets such as people and organizations names [12]. Hierarchical clustering of search results allows using clusters' labels as facet categories (e.g. clusty.com). In biomedicine, existing controlled vocabularies and ontologies are used to derive facets [5, 9] (see section 4.1.1). Facetted refinement is welcomed by users in all studies, but good execution is the key [11].

Related searches are query suggestions that lead to new searches by either changing the query focus or refining it. Such suggestions can be derived dynamically from top search results [2]. A quarter of search sessions made use of these suggestions, but their effectiveness is questionable. A study of related searches in the biomedical domain reports a strong desire for gene and organism names that can be driven from bioscience ontologies [4].

Results preview help users judge the relevance of their searches by listing surrogates for each document containing its title, URL, preview and sometimes keywords. Document previews can be most relevant sentences derived via query-based summarization [10] or snippets that combine multiple sentences while replacing their irrelevant parts with ellipses [3]. Studies indicate that users prefer non-truncated sentences and the preview including document summary should put all query terms in context [6].

The above studies provide insights into usefulness and effectiveness of interface features, but do not tell if there is a preference in certain features depending on the search task. Another gap is a comparison of different techniques that implement facetted navigation in unstructured documents.

3. SEARCH TASKS IN BIOSCIENCE

Kellar et al. classify information seeking tasks into four major categories and analyze how often people conduct and repeat these tasks [7]. Nearly 50% of search queries relate to Transactions (email, banking, shopping), all of which are frequently repeated. Other queries are somewhat equally split between *Browsing* (blogs, news), *Information Gathering* (e.g. graduate schools to apply) and *Fact Finding* (e.g. weather forecasts). The latter three are conducted by bioscientists in their daily work when they browse for new publications, gather information on a particular genes, proteins or diseases, or search for facts, for instance in biomedical databases. Our study takes this differentiation into account when analyzing scientists' rankings of interface features.

4. EXPERIMENT DESCRIPTION

The aim of this study is to identify which search interface features are useful for searching the biomedical literature. Additionally, we strived to understand which approaches to facetted navigation for this domain work best. Based on the current knowledge of user preferences we hypothesized that users prefer different interface features depending on the search task.

4.1 Interface Features & Evaluated Systems

To test our hypothesis we implemented two prototype systems. To test the usefulness of the features more broadly, we included two systems primarily used in the biomedical domain (PubMed) and general search (Google), as well as additional publicly available systems that handle such features in different ways in bioscience (GoPubMed, Semedico, NextBio) and general search (Bing).

4.1.1 Overview of the studied systems

PubMed (ncbi.nlm.nih.gov/pubmed) is the primary search engine used by bioscientists for their research, as it comprises over 20 million citations for biomedical literature from MEDLINE, life science journals, and books. Articles are indexed with Medical Subject Headings (MeSH) [8]. **GoPubMed** (gopubmed.com) is a semantic search engine for the biomedical domain. It provides refinement of PubMed search results using the original hierarchy of structured vocabularies: Gene Ontology (GO) and MeSH [5]. **Semedico** (semedico.org) [9] is a faceted biomedical search system with a ranked list interface. The facets are populated from a semantic index generated by disambiguating words in articles to corresponding concepts in MeSH and UniProt. Its hierarchy of top 20 categories was defined by biologists. **NextBio** (nextbio.com) is a commercial ontology-based semantic framework based on gene, tissue, disease and compound ontologies. It combines literature with data such as clinical trials. **Google** (google.com) is the second most common search engine used by scientists for their work. According to our findings it is often preferred to PubMed for searching methodology and techniques (laboratory protocols). **Bing** (bing.com) handles queries, ranking and some of the features we study in this paper somewhat different to Google.

Our **two prototypes** were built to test additional ways of implementing and representing features of a search user interface. We used the **Pingar API** (pingar.com) for semantic analysis of queries and documents and **Apache Solr** (lucene.apache.org/solr) for full-text indexing and searching. We indexed the 85,000 articles in the Open Access PubMed dataset for this purpose (ncbi.nlm.nih.gov/pmc/tools/openftlist). The prototypes allowed us to test Pingar's tools for generating query expansions, related searches, keywords, summaries and taxonomy mapping, as well as Solr's built-in facetted search and snippet extraction features.

4.1.2 Implementation of tested interface features

Only certain systems were selected for testing each feature. Here, we list how each feature is supported by the systems we tested. Information provided on systems' websites and publications.

Autocomplete: **PubMed** employs Automatic Term Mapping that compares and maps user's search terms to lists of pre-indexed terms. **GoPubMed** matches typed terms to MeSH and GO terms. **Semedico** places the suggestions in a taxonomy tree allowing users to select a broader term as their query. Synonyms are listed in brackets. **NextBio** lists matching genes, compounds, SNPs, diseases, tissues, biogroups and authors. **Google** predicts suggestions based on other users' search activities – for certain queries it analyzes just the last two words. **Bing** also computes suggestions using user's queries and boosts trending queries.

Query expansion: **PubMed** displays the "search details" that combine (sub)headings, fields and Boolean. Users can edit them and re-submit. **Semedico** displays terms identified in the query and users may remove one from the search. **Pingar** suggests misspellings, grammatically similar terms and synonyms as checkboxes to add to the query using *OR*.

Facetted search: PubMed allows filtering by "free full text" or "reviews" and shows the number of matching results in brackets. Suggestions are displayed as links. **GoPubMed** categorizes filtering suggestions into "Top Terms" (more specific) and "Knowledge Base" (more generic), ordered by relevance. Suggestions are displayed as checkboxes allowing multiple selections. **Semedico** displays 9 top level MeSH terms as facet categories each in a differently colored box. Per category, top 3 most frequent terms are shown (numbers in brackets). Expanding leads to more terms, or their child terms. **Solr** was chosen to evaluate facetted refinement based on indexed metadata: journal year and title, and keywords, generated by Pingar. Single and multiple selections were tested (links vs. checkboxes). **Pingar** dynamically generates facet categories by first mapping top 10 search results to terms in multiple biomedical taxonomies and then walking up the taxonomy tree to find common broader terms. A different variation of top 3 most relevant facets is displayed for each query. Each facet lists top 5 most frequent terms and can be expanded to see more. Some screenshots showed terms computed by analyzing only text surrounding the query terms (QB), others the entire content of the document (DB). The intention was to evaluate whether search query should aid as a context when computing facet values. We also tested preference over choosing one or multiple terms per facet category (links vs. checkboxes).

Related searches: PubMed suggests variations of the query in the "also try" area, but such terms are not always available. **Google** offers two kinds of related searches in different parts of the interface: Searches for things of similar kind (e.g. "aquaporin" for "connexin") and more specific searches (e.g. "connexin 26" for "connexin"). **Bing**'s two areas designated to related searches show the same suggestions formatted as one or two columns. The suggestions are variations of the original query with added or modified parts. **Pingar** also computes related searches, but instead of query logs, top search results are analyzed for suggestions.

Results preview: We limited the evaluation of this interface feature to prototype systems only and tested the following features: **(1) Keywords** are usually defined by authors or extracted automatically to represent the key topics in an article. We used Pingar API to compare two cases: extracting keywords from the text surrounding the query terms and from the entire document. The intention was to evaluate whether search query should aid as a context when computing keywords. **(2) Document**

preview is commonly implemented using sentence snippets containing the query terms. One prototype used Solr to extract top 3 such snippets per document. Another one used Pingar's query-based summary extraction tool to display the top scoring sentence in the document, and the top scoring paragraph on mouse-over.

4.2 The Study

We run an exploratory short study with 6 bioscientists (2 faculty, 2 postdocs, 2 PhD students), where we explained them the 3 types of search and asked them for examples of such searches they use for their work: queries and resources/systems. The study was conducted in person with each participant (10-15 min sessions).

We recruited 10 bioscientists to participate in the main study. All of them are researchers in academia in various biological areas: Developmental, Molecular, Cell, Evolutionary, Transcriptional and Systems Biology, as well as Biochemistry, Immunology, Genetics, Population Genetics, and Neuroscience. 2 of them are faculty; 7 are postdoctoral researchers (having received PhDs from 2005 to 2010) and 1 final year PhD student.

We selected the interface features we wanted to study for the 3 different types of search and identified systems that handle these features in different manner. We asked the participants, via email prior the sessions, to supply us with 4 different queries they use. We also asked for use frequency, resources, to elaborate on what information they are looking for, and to fill out an informed consent form. We selected two baseline queries based on our exploratory study. We also selected one personal query from each participant. The selection was done aiming at having a range for the 3 search types. For each query we took screenshots in different systems available for that query and isolated the part of the system that shows the interface feature in question (logos were removed).

Table 1. Queries and search type classification

Query (Participant ID)	Information sought after
Browsing	
meiotic sex chromosome inactivation (4)	papers published in the genetics general area
cisternal maturation AND yeast (6)	new publications on the mechanism that underlies Golgi cisternal maturation in yeast
cerebral cortex (7)	new publications
"chromatin looping"(10)	papers
Fact finding	
animal models of huntington's disease (1, 3, 4, 6 7, 8, 9)	[baseline query]
Information gathering	
connexin (2, 5, 10)	[baseline query]
ecdysone receptor (1)	functional, genetic, disease and binding studies
NKT (2)	general information
adaptation diet human genetic (3)	studies of candidate genes that explain how humans with various lifestyles adapted to different diets
myelin (5)	information on mechanisms that promote myel repair in demyelinating disorders
connexin (8)	publications by others on connexins & how the relate to our studies
'systems biology' proteins (9)	new methods & discoveries in systems biology of proteins (experiments, format, conclusions)

During the in person sessions (each lasted 1-2 hours), we showed participants in PowerPoint presentation screenshots for one baseline query (of their choice) and for one of their own queries (Table 1). For each feature we asked them to rate overall usefulness and aesthetics using a 5 point Likert scale. Then for each system demonstrating possible handling of each feature, we asked them to rate usefulness and aesthetics using again the Likert scale. Finally for each feature, we asked them to rank the systems in order of preference. Throughout the sessions we applied the talk aloud protocol and encouraged comments and suggestions.

5. RESULTS

Below we discuss participants' reactions to content and overall usefulness of the interface features. Given the space limitations, we plan to present our aesthetics preference findings in a different venue. Overall, participants told us that aesthetics are important (and need to be "good enough" to use a system) but what really matters is the content. We would also like to emphasize that all scores presented in this paper are for specific features supported by each system and are not reflective of the systems as whole.

5.1 Interface Features vs. Search Tasks

Figure 1 summarizes how participants judged the usefulness of interface features for their queries. The length of each bar equals the number of ratings. All 4 browsing participants liked **autocomplete**, whereas 5 out of 6 participants with info gathering queries rated it as neutral and 1 out of 6 as not useful. This shows a clear difference in usefulness of this feature. For **query expansions** participants expressed positive or neutral opinions for browsing and mixed opinions for the other search types. **Facetted refinement** was rated mostly useful for all search types with equal neutral scores for information gathering. **Related searches** got predominantly neutral or negative reactions, except for browsing, for which they are equally spread. Not surprisingly, it's useful to see **document previews** for all search types. Comments analysis shows that **snippets** are better for browsing, whereas **summaries** with full sentences for finding specific or specialized information.

The results are intuitive and participant's comments confirm these. We encourage the interface designers to give priority to "green" elements of the search display and do not bother much about related searches for biologists. Most of them told us that their searches are usually specific and even correctly suggested related searches are not of interest. Access to query expansions is important to experts, but should not be a default feature.

5.2 System Rankings

For each interface feature participants ranked the systems in order

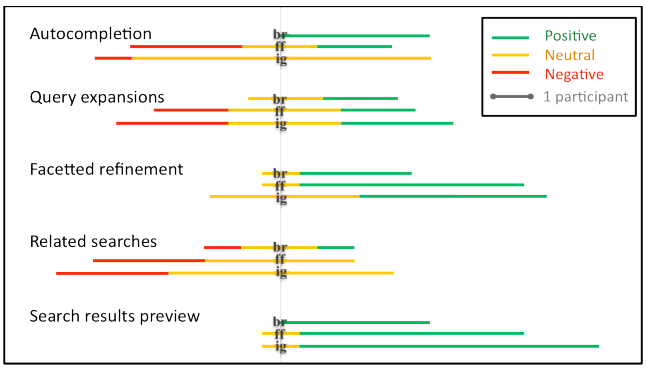

Figure 1. Usefulness ratings for interface features & search tasks: browsing (br), fact finding (ff) & information gathering (ig)

of preference. Table 2 shows the systems ranked at the top and at the bottom (top/bottom two were considered if 5 or more systems were compared, top/bottom one if 4 or less were compared).

5.3 Comparisons with Previous Findings

Our results agree with those reported by Scheider et al. that facets are well received by bioscientists, but autocomplete is less important [9]. Scheider et al. also argue that in biosciences, a large number of facets are needed per query, which they grouped into collapsible tabs in Semedico. While our participants were positive on collapsible tabs (judging by comments but not explicit testing), 9 out of 10 did not want to see a large number of facetted groups. The large number of choice and the inevitable redundancy overwhelmed them. They commented that they would not spend time inspecting the facets despite speculating that some might be useful. Divoli et al. found that bioscientists like to refine searches by organism names [4] and two our participants also commented they really liked Semedico's facet "Organisms" (again not explicitly tested). Our results also confirm findings in [4] that users prefer selecting multiple suggestions using checkboxes.

5.4 More Findings on Interface Features

During the course of the study, participants provided us with interesting suggestions that are not currently implemented by the systems. Below we categorize them by feature.

Autocomplete was preferred when the major part of the query is typed and users feel pigeon holed if suggestions come up with the first characters. Specific suggestions work best here. **Query expansions** do not need to include misspellings and close grammatical forms. These should be included in search automatically. Overall biologists mostly refine and focus searches, but not expand them. **Facetted refinement** is always desired, with checkboxes that enable multiple selections. Too much information and too many categories scare them away, as does redundancy of terms (across categories and within each category). Simpler designs are better (e.g., not too many colours) - this is why Pingar's top 3 ranked facets with a few values each scored highly. Users expect facet categories to reflect query types, e.g. if the query mentions a disease, conditions should be shown, but not other diseases. Many liked the ability to refine search by a specific keyword related to their query, offered by GoPubMed's "top terms" and Pingar's keywords. Some commented that year, publication and even the entire facetted refinement column should not be displayed by default. PubMed's option to refine by reviews was highly favored. Some explained that besides offering comprehensive information, reviews help to discover important papers by navigating through the references. Pingar DB's **document preview** was preferred for general searches (baseline queries) and Pingar QB for specific searches (their own queries). **Related searches** are not desired, except for browsing. Scientists dislike clicking on links leading them to new or broader searches.

Table 2. Top and bottom ranked systems

		Ranked top	Ranked bottom
Autocomplete	All (baseline)	Google (7/10), Bing (5/10)	GoPubMed (7/10), Semendico (7/10)
	All (own)	Google (6/10), PubMed (4/8)	NextBio (6/8), Semendico (4/6)
	Browsing	Google (3/4)	NextBio (3/3)
	Fact finding	Google (5/7)	GoPubMed (7/7)
	Info gathering	Google (5/9), GoPubMed, PubMed (4/8)	Semendico (5/7), NextBio (4/7)
Query expansions	All (baseline)	Pingar (5/10), Semendico (3/7)	
	All (own)	Pingar (5/10), PubMed (4/8)	
	Browsing	PubMed (2/4)	
	Fact finding	Semendico (4/7)	
	Info gathering	Pingar (7/9)	
Facetted refinement	All (baseline)	Pingar DB (8/10), Pingar QB, Solr (4/10)	PubMed (7/10), Semendico (6/10)
	All (own)	GoPubMed (6/10), Pingar QB (5/11)	Semendico (6/9), PubMed (4/6)
	Browsing	GoPubMed (3/5)	Semendico (3/4)
	Fact finding	Pingar DB(6/7)	PubMed (6/7)
	Info gathering	Pingar QB (6/9)	Semendico (5/8), Solr (5/9)
Related searches	All (baseline)	Pingar (7/10)	Bing (5/10)
	All (own)	Google (6/10)	Pingar (4/10), Bing (4/9)
	Browsing	Google (4/5)	Pingar (4/4)
	Fact finding	Pingar (5/7)	Bing (4/7)
	Info gathering	Google (4/5)	Google (3/5), Bing (3/7), PubMed (3/7)
Doc preview	All (baseline)	Solr (6/10)	
	All (own)	Pingar (7/10)	
	Browsing	Pingar (3/4)	
	Fact finding	Solr (5/7)	
	Info gathering	Pingar (6/9)	

6. CONCLUSIONS & FUTURE WORK

This paper demonstrates user preferences for different search features depending on search types in the biomedical domain. Although the search tasks bioscientists perform are not clearly distinct from each other, in the future we would like to study which tasks to prioritize and how to integrate features on the interface to allow optimization for switching search types.

7. ACKNOWLEDGMENTS

We are extremely grateful to all participants for their contribution.

8. REFERENCES

[1] Anick, P. and Kantamneni, R. 2008. A longitudinal study of real-time search assistance adoption. In *SIGIR'08*: 701-702.

[2] Anick, P. 2003. Using terminological feedback for web search refinement. In *SIGIR'03*: 88-95.

[3] Cutrell, E. and Guan, Z. 2007. An eye-tracking study of information usage in web search. In *CHI'07*: 407-416.

[4] Divoli, A., Hearst, M.A. and Wooldridge, M.A. 2008. Evidence for Showing Gene/Protein Name Suggestions in Bioscience Literature Search, PSB 2008.

[5] Doms, A. and Schroeder, M. 2005. GoPubMed: Exploring PubMed with the Gene Ontology.*Nucl Acids Res*,33:783-786

[6] Hearst, M.A. *Search User Interfaces*. Cambridge UP, 2009.

[7] Kellar, M., Watters, C. and Shepherd, M. 2007. A field study characterizing web-based information seeking tasks. *JASIST*, 58(7), 999-1018.

[8] Lu, Z., Wilbur, J.W., *et al.* 2009. Finding query suggestions for PubMed, *AMIA Annu Symp Proc.* 2009: 396–400.

[9] Schneider, A., Landefeld, R., Wermter, J. and Hahn, U. (2009) Do users appreciate novel interface features for literature search? Systems, Man and Cybernetics, SMC 2009.

[10] Tombros, A. and Sanderson, M. 1998. Advantages of query biased summaries in IR. In *SIGIR'98*: 2-10.

[11] Tunkelang, D. *Faceted Search*. Morgan and Claypool, 2009.

[12] Tunkelang, D. 2006. Dynamic category sets: An approach for facetted search. In *SIGIR'06 Faceted Search Workshop*.

Columns on Last Page Should Be Made As Close As Possible to Equal Length