Categorization of services for seeking information in biomedical literature: a typology for improvement of practice

Jung-jae Kim and Dietrich Rebholz-Schuhmann

Submitted: 16th January 2008; Received (in revised form): 3rd July 2008

Abstract

Biomedical researchers have to efficiently explore the scientific literature, keeping the focus on their research. This goal can only be achieved if the available means for accessing the literature meet the researchers’ retrieval needs and if they understand how the tools filter the perpetually increasing number of documents. We have examined existing web-based services for information retrieval in order to give users guidance to improve their everyday practice of literature analysis. We propose two dimensions along which the services may be categorized: categories of input and output formats; and categories of behavioural usage. The categorization would be helpful for biologists to understand the differences in the input and output formats and the tasks they fulfil in information-retrieval activities. Also, they may inspire future bioinformaticians to further innovative development in this field.

Keywords: information retrieval; biomedical literature; literature analysis; literature mining; information behaviour; text mining; natural language processing

INTRODUCTION

Web-based tools for exploring the biomedical literature in pursuit of information of interest are modelled on information-retrieval approaches for accessing the World Wide Web. The approaches for exploring the web have been studied and analysed in depth for meeting requirements in information retrieval [1, 2, 3]. However, little effort has been made to understand how literature mining services fulfil similar demands for biomedical research.

Innovative scientific work relies on a comprehensive knowledge of the current research status to generate novel and compelling hypotheses. Assessment of the scientific state of the art is primarily based on literature analysis. Unfortunately, there is no single web service that satisfies all demands on literature analysis by the biomedical research community. We assume that researchers will become more efficient in their literature analysis once they gain a better understanding of the ways the existing services meet their needs; they then can integrate these services into their workflow. We support solving the paradoxical situation that molecular biologists are used to questioning, standardizing and innovating in their experimental approaches, but they usually do not invest a significant portion of their research time into either systematically optimizing their approaches to literature analysis or exploring novel solutions to speed up access to the most relevant publications.

Some characteristics of biomedical literature are its huge volume, its high diversity and its high quality due to the peer-review system. These properties make it a unique source of information in comparison to other biomedical resources such as manually curated databases (e.g. UniProtKB [4], OMIM [5]) and widely used bioinformatics tools for large-scale experiments. The volume of available documents...
makes it almost impractical for biologists to locate information of interest without the use of automated tools such as search engines (e.g. PubMed, HubMed). Likewise, the constant increase of usage of PubMed (see the PubMed Usage Data, http://www.ncbi.nlm.nih.gov/About/tools/restable_stat_pubmed.html) is possible evidence of the perceived necessity of such tools.

Many automated tools have been developed over the past decade so that biomedical researchers now enjoy an abundance of them (see the list in Table 1). Among them, biologists would be most familiar with search engines, whose primary goal is to retrieve and rank relevant documents, so-called document retrieval. Information retrieval is often regarded as identical to document retrieval, but this description is too restrictive. We must consider additional functionalities as being part of information retrieval, such as document clustering, recognition of frequent co-occurrences between biomedical concepts and identification of the most relevant text fragments in retrieved documents. They are useful in seeking information from the literature and are supplementary to document retrieval in the sense that they are imposed on top of document retrieval.

The supplementary functionalities manipulate the results of document retrieval so that users may locate information of interest from the manipulated results more straightforwardly than from the original results. For example, a user may focus on a specific topic by choosing a particular cluster after document clustering. Furthermore, users will find it much faster to read through a selected set of the most relevant text fragments than to read through the whole set of retrieved documents. Since the functionalities have been developed recently and thus are not well-known by biomedical researchers, biologists may expect an explanation of the full range of available functionalities for biomedical information retrieval. They may also want more guidance on how to use the tools under different conditions.

To address these issues, we here present two dimensions of categorization for publicly available information-retrieval tools. The first dimension differentiates the input and output formats of the tools—the most evident features of the services. We also discuss the usability and the reliability of the tools as derived from their reported performance and estimated response time. We then generate assumptions as to which tools serve adequately on one or many behavioural modes of information-seeking activities (viz., starting, browsing, chaining, monitoring, differentiating and extracting) that have been proposed in a theoretical model describing behaviour patterns of information seeking from the web [6, 7]. These analyses may also give guidance to bioinformaticians suggesting promising research fields in biomedical information retrieval.

The tools we analyse in this article are listed in Table 1 together with their web addresses and references if available. We do not explain the details of any underlying text-mining techniques, since they

---

**Table 1: Information retrieval tools examined**

<table>
<thead>
<tr>
<th>Name</th>
<th>Homepage</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ali Baba</td>
<td><a href="http://alibaba.informatik.hu-berlin.de/">http://alibaba.informatik.hu-berlin.de/</a></td>
<td>[8]</td>
</tr>
<tr>
<td>BioText</td>
<td><a href="http://biosearch.berkeley.edu">http://biosearch.berkeley.edu</a></td>
<td>[10]</td>
</tr>
<tr>
<td>CiteXplore</td>
<td><a href="http://www.ebi.ac.uk/citexplore/">http://www.ebi.ac.uk/citexplore/</a></td>
<td></td>
</tr>
<tr>
<td>eTBLAST</td>
<td><a href="http://invention.swmed.edu">http://invention.swmed.edu</a></td>
<td>[12]</td>
</tr>
<tr>
<td>iHOP</td>
<td><a href="http://www.ihop-net.org/UniPub/iHOP/">http://www.ihop-net.org/UniPub/iHOP/</a></td>
<td>[15, 16]</td>
</tr>
<tr>
<td>Info-PubMed</td>
<td><a href="http://www-tsujii.is.s.u-tokyo.ac.jp/info-pubmed/">http://www-tsujii.is.s.u-tokyo.ac.jp/info-pubmed/</a></td>
<td>[17]</td>
</tr>
<tr>
<td>McSyBi</td>
<td><a href="http://textlens.hgc.jp/McSyBi">http://textlens.hgc.jp/McSyBi</a></td>
<td>[18]</td>
</tr>
<tr>
<td>MedEvi</td>
<td><a href="http://www.ebi.ac.uk/tc-test/textmining/medevi">http://www.ebi.ac.uk/tc-test/textmining/medevi</a></td>
<td>[19]</td>
</tr>
<tr>
<td>MEDIE</td>
<td><a href="http://www-tsujii.is.s.u-tokyo.ac.jp/medie/">http://www-tsujii.is.s.u-tokyo.ac.jp/medie/</a></td>
<td>[17]</td>
</tr>
<tr>
<td>Protein Corral</td>
<td><a href="http://www.ebi.ac.uk/Rebholz-srv/pcorral/index.jsp">http://www.ebi.ac.uk/Rebholz-srv/pcorral/index.jsp</a></td>
<td></td>
</tr>
<tr>
<td>PubMed Central</td>
<td><a href="http://www.pubmedcentral.nih.gov/">http://www.pubmedcentral.nih.gov/</a></td>
<td></td>
</tr>
<tr>
<td>Textpresso</td>
<td><a href="http://www.textpresso.org/">http://www.textpresso.org/</a></td>
<td>[20]</td>
</tr>
<tr>
<td>Twease</td>
<td><a href="http://twease.org/medline/app">http://twease.org/medline/app</a></td>
<td></td>
</tr>
</tbody>
</table>

Tools are listed in an alphabetical order.
can be found in the references for the tools and in review publications about biomedical text mining [21–26]. Table 1 does not include web-based services in the general domain, such as Google Scholar (http://scholar.google.com/), Science Direct (http://www.sciencedirect.com/) and IngentaConnect Complete (http://www.ingentaconnect.com/). We focus on tools that have been introduced specifically for the biomedical domain. Furthermore, we exclude information-extraction systems, which are primarily aimed at populating databases with selected types of information, and knowledge discovery systems, which generate unpublished but potentially valid relations between biomedical terms [27], since the systems themselves do not provide interactive interfaces for a continued search, even though their results can be used for seeking information from the literature.

CLASSIFICATION BY INPUT AND OUTPUT FORMATS

Input format

Users of an information-retrieval tool have to specify their query in textual form. The text may be short (e.g. a single word, a database entry identifier) or very long (e.g. a complete document) and may require a rather complex format unsuitable to persons with little computational experience (e.g. Boolean queries). The inputs to the tools convey the demands of the users and steer the outcome. The designer of a tool has the choice of making it very restrictive by imposing strong constraints on the input format or very tolerant by using only basic assumptions during the processing of the input.

Inputs that are delivered to the tools listed in Table 1 can be classified into three categories:

(i) Single-keyword queries—usually named entities (e.g. protein name, disease name, protein database identifier),
(ii) Boolean queries, where keywords or query terms are combined with Boolean operators (e.g. AND, OR, NOT) and
(iii) text queries composed of any type of text (e.g. a whole MEDLINE abstract, a set of documents).

The tools that only accept single-keyword queries as inputs (e.g. iHOP, Info-PubMed, BioContrasts) generally interpret the input as a named entity and then query their own databases containing information about the named entities. The information in the database has been extracted or annotated from the literature in an off-line manner. The tools accept the single-keyword queries and display the extracted and gathered information about the entities with links to the sources of the information. While the tools of this category in Table 1 only deal with genes and proteins, LitMiner [28] deals in addition with chemical compounds, diseases and tissues. We have, however, not included this tool for the categorization work, since it does not provide links to the sources of the information, which we believe are critical for verifying the information found in the text. The genes and proteins dealt with by the tools are matched with entries from well-known gene/protein databases such as Entrez Gene and UniProtKB/Swiss-Prot.

The most popular type of inputs to information-retrieval systems is keyword-based Boolean queries. It is obvious that the set of Boolean queries is a superset of the set of single-keyword queries. A multi-word query (e.g. ‘breast cancer’, ‘MAPK pathway’) is generally interpreted as if the multiple words are concatenated by AND operators, thus having to be contained in the same document. The OR operator can be used to specify keyword alternatives, for example synonyms and term variants, in order to retrieve all documents that make reference to at least one of the alternatives. Accordingly, combinations of keywords and Boolean operators can be used to retrieve an improved set of documents, compared to single-keyword queries.

For increased coverage of document retrieval, some tools from the second category provide the functionality of an automatic expansion of queries with synonyms (e.g. CiteXplore, PubMed), with frequently co-occurring terms (e.g. HubMed), or with full names of abbreviations (e.g. Twease). Other tools such as EBI Med and MedEvi allow users to input concept identifiers such as database identifiers (e.g. UniProt accession numbers) instead of listing all synonyms of a gene name.

Further restrictions can be imposed on Boolean queries to improve the results of document retrieval. For instance, the queries can be constrained to selected search fields in the meta-data of scientific articles (e.g. title, author, publication date, abstract). For this purpose, a keyword in the query can be labelled with the name of a field so that it is only applied to the selected field of the articles. Furthermore, tools making use of Lucene indexing
(e.g. CiteXplore, EBIMed) usually provide fuzzy search, proximity search, range search and boosting as features to the Lucene queries (http://lucene.apache.org).

In addition, several tools allow semantic restrictions on the queries. For instance, MedEvi offers the use of semantic variables (e.g. ‘[gene]’, ‘[disease]’, ‘[cell]’, ‘[drug]’) as keywords in queries. With the example query ‘‘breast cancer’ AND [gene]’, MedEvi returns text fragments that include not only ‘breast cancer’ but also any mention of a gene name in the local context of the disease name mention. This supports finding genes associated with a disease. The Semantic Search interface of MEDIE has an entry page where users can fill in semantic slots (i.e. subject, verb, object) with terms that should comply with the specified semantic roles and the implied semantic relations in sentences. For instance, if a gene name and a verb are provided for the subject slot and the verb slot, respectively, the pair is matched to all subject–verb pairs that have been identified in sentences from MEDLINE and gathered in the MEDIE database. This approach supports finding sentences in which the gene plays certain roles indicated by the verb. Textpresso allows users to specify semantic types of query terms with predefined semantic classes, many of which are adopted from Gene Ontology [29].

Any type of text or set of documents can serve as inputs to the tools belonging to the third category. For instance, eBLAST receives a text and returns a list of documents that are relevant to the topic of the given text. McSyBi allows users to input a set of documents and returns hierarchical or non-hierarchical clusters of the documents. These tools make use of the fact that a text can be transformed into a vector of terms extracted from the text, and that similarity between texts can be calculated from the mathematical similarity of their vectors.

**Output format**

Outputs of the information-retrieval tools should comply with the user’s requests as specified by their inputs. As the underlying information is contained in stretches of text, for example in the retrieved documents, the outputs are usually in the same format as the inputs. In the other cases, the tools produce outputs in a format that is likely to be expected by users, for example no other information than a list of gene names that are found in the retrieved documents and then associated with a given disease. The outputs vary in length and structure, and we may assume that a short and well-structured representation of the output allows the user to better locate the targeted information than a long and coarse representation of the output [30].

We classify the output formats of the tools into the following three categories:

(i) A set of documents that match the input queries (document retrieval),
(ii) A set of text fragments that are identified from the retrieved documents as being relevant to the given inputs (passage retrieval) and
(iii) A summary of relations between biomedical concepts recognized in the retrieved documents (relation retrieval).

Tools belonging to the first category are mainly search engines functioning like the Google search engine. They receive a query, retrieve documents whose content matches the query, and rank the documents based on their relevance to the query. Some of the tools from this category (e.g. HubMed, GoPubMed) cluster and select the documents based on keywords that have been identified and selected as being relevant since they are statistically over-represented in the retrieved documents, and those that match terms from well-known ontologies (e.g. MeSH, Gene Ontology). For instance, GoPubMed identifies Gene Ontology terms in the retrieved documents and clusters the documents based on the identified terms. The resultant clusters can be then displayed in the same hierarchy as that of the ontology. By selecting a Gene Ontology term in the hierarchy, users can locate documents that contain the term.

The retrieved documents, however, are often too many for users to browse and search for detailed information. Therefore, some available tools reduce the documents into text fragments, mostly sentences that have high relevance to the given queries. Once a user has spotted interesting text fragments, they can return to the original documents to explore for the details of the information found in the fragments.

Other tools go beyond pure text retrieval and extract pre-defined types of information from the search results. The most popular strategy is the extraction of co-occurrences between concepts (e.g. gene, disease, cell). Co-occurrences have been extensively exploited based on the assumption that they may represent biologically meaningful
relations [31]. Ali Baba and iHOP display graphs of the co-occurrences, while EBIMed and Protein Corral summarize the co-occurrences in the form of tables. BioContrasts focuses on a specific type of co-occurrences—contrastive relations between proteins. Contrastive relations reveal explicit differences and implicit similarities of the contrasted proteins and thus are particularly useful for knowledge discovery [9].

In summary, documents retrieved by search engines provide all the information that users wish to find in the literature. However, as the literature has already grown so much, it is unrealistic to read all retrieved documents from beginning to end to find the targeted information. To address this issue, many novel tools focus on more specific types of results. These tools, however, tend to encounter the opposite problem: their results are too specific to be useful to users who are interested in other types of information. The categorization of the tools by their input and output formats is summarized in Table 2.

Figure 1 depicts a graphical summary of the categorization. It shows that queries with a more restrictive structure are less expressive, and that with a large number of pieces of data output from an unrestricted query each is on average less informative than those from less exhaustive outputs. Also, it suggests that narrower specifications for the output findings induce more informative outputs, but require more processing for the identification of the findings to meet the specifications.

It is clear that there is a high degree of variability in the ways the tools pick up the query and deliver the results. However, the answer to the question ‘which is the best tool of all’ depends on the information needs of the user, since no single tool can efficiently meet all demands. One might be tempted to assume that the right solution would be the integration of all tools into one web portal and the generation of combined results [e.g. Vivisimo BioMetaCluster (http://vivisimo.com/products/biometacluster)]. Certainly it would be a great achievement, but it could also be a rather inefficient solution to meet the variety of researcher’s needs. There is still no guide available that explains how to locate the target information in the vast number of results generated by the different tools. This open issue leads to the following questions: what types of information-seeking behaviour patterns do users have and which type of solutions can deliver the best results for these types of behaviour patterns? Answers to these questions may be found in the recent research work that assesses the information-seeking patterns for information retrieval on the web.

CLASSIFICATION BY INFORMATION-SEEKING BEHAVIOUR PATTERNS

Behaviour patterns of users in seeking information on the web have been extensively studied to achieve improvements on information-retrieval systems. We adopt a model of information-seeking behaviour [6, 7] to characterize the tools in question with respect to their applicability to different behavioural activities.

Ellis and colleagues have developed a model describing information-seeking behaviour based on records obtained from monitoring researchers’ use of information-retrieval systems [6, 7]. They suggest that information-seeking behaviour can be well characterized with only a small number of distinguishable types of activities: starting, chaining, browsing, differentiating, monitoring and extracting. The behavioural model describes relations between these activities, but does not define a set of stages that any or all researchers follow when seeking information [6]. The relations between different activities are not examined by this article. We here report on our attempt to associate features of the information-retrieval tools with the types of activities in the behavioural model. Such associations may give biologists guidance in choosing the most appropriate tools for particular tasks of information seeking in their research. Figure 2 depicts the six types of behavioural activities proposed in the model.

For better understanding, we present example queries of MedEvi for the first three types of behavioural activities. We chose MedEvi as an example tool following our experience that MedEvi positions the information of interest, if available, in the first page of its outputs (see the section ‘Browsing’ for its comparison with search engines). A series of examples for the single tool, rather than mixed examples for different tools, may help us present a coherent narrative.

Starting

Starting comprises all activities that are generic to the initial retrieval of information to obtain an overview of a topic, or to locate key elements of
<table>
<thead>
<tr>
<th>Input format</th>
<th>Output format</th>
<th>Tool</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene (name, database concept identifier)</td>
<td>Gene—gene interactions and sentences about the</td>
<td>iHOP</td>
<td>• Browse sentences that contain a given gene name of a given species</td>
</tr>
<tr>
<td></td>
<td>interactions</td>
<td></td>
<td>• Highlight and provide links to co-occurring gene names</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Info-PubMed</td>
<td>• Display graphs of gene-gene interactions selected by users</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Provide a drag-and-drop user interface</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BioContrasts</td>
<td>• Focus on protein—protein contrastive relations</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Display graphs of contrastive relations of a given gene</td>
</tr>
<tr>
<td>Query (Boolean operator, search field restriction)</td>
<td>Documents</td>
<td>BioText</td>
<td>• Search even figure captions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CiteXplore</td>
<td>• Search MEDLINE, Patents, C.B.A., Citeseer</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Expand query terms with synonyms</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PubMed</td>
<td>• Official site for MEDLINE</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Allow user customization through My NCBI</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PubMed Central</td>
<td>• Provide programming utilities to access MEDLINE</td>
</tr>
<tr>
<td>Document clusters</td>
<td>An ordered list of documents</td>
<td>HubMed</td>
<td>• Provide Atom and RSS feeds for regular search updates on a given query</td>
</tr>
<tr>
<td>Hierarchical clusters of documents</td>
<td></td>
<td>GoPubMed</td>
<td>• Expand query with frequently co-occurring terms</td>
</tr>
<tr>
<td>Text fragments (usually sentences)</td>
<td></td>
<td>MedEvi</td>
<td>• Cluster search results by MeSH and Gene Ontology terms</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Allow semantic restriction using ontology hierarchy</td>
</tr>
<tr>
<td>Textpresso</td>
<td></td>
<td></td>
<td>• Provide various statistics on publications of a topic</td>
</tr>
<tr>
<td>Twease</td>
<td></td>
<td></td>
<td>• Focus on <em>C. elegans</em></td>
</tr>
<tr>
<td>Co-occurrences of frequent terms</td>
<td></td>
<td>Ali Baba</td>
<td>• Allow semantic restriction of query terms with pre-defined semantic classes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Expand abbreviations with full names</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EBIMed</td>
<td>• Display graph of terms that are frequent in retrieved documents</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Browse sentences that contain both given queries and the frequent terms</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Protein Corral</td>
<td>• Display a table of co-occurrences between terms of pre-defined types found in retrieved documents</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Browse sentences containing the co-occurrences</td>
</tr>
<tr>
<td></td>
<td></td>
<td>eBLAST</td>
<td>• Display a table of co-occurrences between frequent protein names</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Browse sentences containing the co-occurrences</td>
</tr>
<tr>
<td></td>
<td></td>
<td>McSyBi</td>
<td>• Return a set of texts that are similar to a given text</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Return a hierarchy of clusters of given documents</td>
</tr>
</tbody>
</table>

The features listed are selected preferably from the perspective of information retrieval. Tools in a category are ordered alphabetically.
the topic, from the literature. During this phase, the information-seeking behaviour typically profits from queries with a low degree of specificity that produce exhaustive information. More specific queries would restrict the results and thus exclude potentially important information.

Users may initiate the search with general keywords that are already known to be relevant to the topic. For instance, with the query ‘Huntington’s disease’ to MedEvi, we can find definitions of the disease and links to review papers of the disease amongst the top-ranked results. Search engines in general support these activities.

**Chaining**

Once key elements of a subject have been identified, users may follow referential connections between sources to identify new sources of information. In other words, information found in the retrieved documents is used to modify or improve the previous query to better meet information needs of users. Virtually all the tools in question can be used for chaining activities, in that the textual search results of the tools contain not only the key elements, but also other elements related to the key elements.

While search engines can be used for these activities, reading all retrieved documents is not feasible due to the time that would have to be spent on this process. Tools that output co-occurrences of terms would be more suitable for this task than search engines, on the assumption that frequently co-occuring terms might have semantic relations amongst themselves and with the query terms. Navigation through the co-occurrence relations would help users to reach the target information. However, these tools have the limitation that the co-occurrences found are only between concepts of pre-defined types (e.g. genes, diseases).

Some of the tools that output documents and text fragments (e.g. HubMed, GoPubMed, MEDIE, MedEvi) highlight terms that are semantically or statistically significant in the results. These highlighted terms can be used for the reformulation of queries. In MedEvi’s results for the query ‘Huntington’s disease’, for example, we find ‘huntingtin’ highlighted, which is the key protein in the pathway of the disease. We can further inquire about the relation between the disease and the protein by expanding the query to ‘“Huntington’s disease” AND huntingtin’. eBLAST provides a special method for chaining such that users do not need to formulate queries, but can

**Figure 1:** Overview on the categorization of information retrieval tools on the basis of their input and output formats.

**Figure 2:** Ellis’ model of information seeking behaviour.
submit text of any length which includes important keywords for a subject.

**Browsing**

Browsing is semi-directed searching in an area of potential interest. This requires that the tools supporting these activities should have the capability of processing the retrieved documents to filter potentially relevant information. In this sense, the tools that identify the most relevant text fragments from retrieved documents (e.g. MedEvi, MEDIE, Textpresso, Twease) are more helpful for browsing than those delivering a list of documents. For instance, receiving the query “‘Huntington’s disease’ AND huntingtin”, MedEvi outputs results in which we find information about the roles of the protein in the disease pathway in the first page. In comparison to it, biomedical search engines output a list of documents from which users should select interesting articles on the basis of their titles and follow hyperlinks to read the content for detailed information.

Also, document clustering (e.g. HubMed, GoPubMed, McSyBi) is useful for browsing since users can select the document clusters that they are most interested in. In addition, it is helpful if the tools highlight query terms, text-mining results and semantically or statistically significant terms in retrieved documents. Half of the tools in question provide highlighting in one way or another.

**Differentiating**

Differentiating is characterized by activities that use distinguishing features of source documents to filter the documents according to their nature and quality. Generally, such filtering can be performed with search options provided by most search engines, including filtering by publication dates and filtering by authors. However, there are only a few services in the biomedical domain that exploit the differences that can only be identified through processing the content of the source documents. For example, BioContrasts exploits one aspect of such differences—contrastive relations. It identifies contrastive relations between proteins from the literature by using language patterns such as ‘A but not B’.

There are indirect ways to use tools for differentiating. Document clustering can be used to compare documents from different semantic categories. The comparison may help users to understand the characteristics of each category. The feature of displaying the history of user queries with result numbers (e.g. PubMed) can be used, for example to know how frequently the genes in relation with a disease have been studied and then to focus on less studied genes for new discovery. However, there is clearly the need to develop more and better solutions for differentiating.

**Monitoring**

Monitoring is used for keeping up awareness of the latest developments in a field by automatic monitoring of particular sources. PubMed, through My NCBI, and HubMed support monitoring activities, in that they enable users to get updates of results for search queries provided either via e-mail or via RSS feed. There are certainly demanding challenges to using monitoring—for example, how to accurately set search queries for monitoring (e.g. restriction to semantic categories) and how to efficiently manage the citations received through monitoring.

**Extracting**

Extracting is characterized by working through sources to locate information of interest. There is no general pattern describing these activities since in principle, any type of information can be a target of information seeking. It is thus advisable for users to select those tools that generate exactly the type of outputs that they require.

In summary, we have proposed potential associations between the behavioural patterns of information seeking and the tools of biomedical information retrieval. For example, search engines can be used for starting, and also for monitoring if they have such functionalities. Tools that deliver document clusters can be used for differentiating and selective browsing. Those that output text fragments would be effective in browsing. Tools that identify gene–gene interactions and co-occurrences of terms would support chaining. These associations can be a general guide to help choose those tools that are most appropriate for a selected behavioural activity. The choice of the most suitable tool among those of a particular type would, in addition, require understanding of the details of the functionalities of the tools. We expect that in the future a more fine-grained classification will lead to tighter associations between the classification systems.
COMPARISON BY USABILITY AND RELIABILITY

In this section, we explore the question of how much the tools under consideration in this review would be useful and reliable in practice for the information-retrieval tasks. As the basis of this analysis we rely on the reported performance and the estimated response time of the tools.

Precision and recall

The performance evaluation of information-retrieval systems usually employs the two prevalent measures of precision and recall. Precision is the ratio of the number of truly relevant retrieved documents to the total number of retrieved documents for a query. Recall is the ratio of the number of truly relevant retrieved documents to the total number of relevant documents in a test set for evaluation. The trade-off between precision and recall in information-retrieval systems is well-known [1].

Many efforts have been made to evaluate the systems against those measures in common metrics [32]. Community-based efforts have brought forth evaluation competitions including KDD Challenge Cup [33], TREC Genomics Track [34, 35], BioCreAtIvE [36] and BioNLP [37]. The evaluation challenges define shared tasks, such as gene/protein name recognition and protein–protein interaction recognition, and provide training and test corpora for implementation and evaluation of the systems. The metrics particularly for evaluating gene/protein name recognition methods would be useful to evaluate tools that deal with genes and proteins, if we carefully consider for example species of genes and proteins. However, it is almost impossible to evaluate all the tools using a single metric, due to the high degree of their diversity. Furthermore, other evaluation corpora, especially for evaluating systems that identify protein–protein interactions, can be found inconsistent to each other in the sense that a system shows significantly varying performances when evaluated on different corpora [38]. Thus, we here introduce reported performances of the tools each by each, while it should be noted that the tools even belonging to a specific category cannot be directly compared.

The tools that output gene–gene interactions and co-occurrences of terms have been more thoroughly evaluated than the others. It has been reported that the gene name recognition of iHOP shows 87% recall, as assessed against the LocusLink database [39] and 94% precision, as manually evaluated with about 400 gene name occurrences per organism [16]. However, its gene–gene interaction recognition has not been evaluated. In the BioContrasts database 90% of the contrastive relations are estimated to be correct, as evaluated with 100 relations [9]. The recall of the contrastive relation recognition is reported to be lower than 61.5%. The tools displaying gene–gene interactions tend to give higher priority to precision than to recall, for higher coherence of the database content. The high coherence ensures usefulness and reliability of the content, meeting the expectations of the users. Ali Baba is reported to achieve a maximum recall of 52% at 75% precision in extracting protein–protein interactions, as evaluated on the SPIES corpus [40]. The protein–protein interactions recognized by EBIMed are compared to Wnt pathway described in the Kyoto Encyclopedia of Genes and Genomes (KEGG) [41], and it is reported that 37% of the identified interactions are meaningful [11].

Only a few of the tools that output either a set of documents or a set of text fragments have been evaluated with respect to precision and recall. Mueller and colleagues measured the performance of Textpresso in extracting genetic interaction data from journal articles, focusing on Caenorhabditis elegans [20]. They reported that the tool shows 35% precision and ~62% recall for a query composed of semantic categories and keywords and 19.5% precision for a large-scale retrieval from 3307 articles. Lewis and colleagues estimated the performance of eTBLAST with inputs from 10 novice users and reported that the average precision is 76.8%, while the highest precision is 85.0% [12]. They also compared various algorithms of text similarity search with the evaluation data of TREC 2003 [34], reporting that cosine similarity achieves the highest mean average precision (MAP) value 0.27.

A few more tools (e.g. GoPubMed) have been evaluated but only on specific parts of their systems [13, 42]. Other tools emphasize their unique features as web services instead of reporting estimated performance of the features, as summarized in Table 2. Interestingly enough, Twease, as a web service, allows users to adjust the outputs according to their requirements of precision and recall by moving the ‘Precision/Recall Slider’ to the desired setting.

Response time

We also attempted to measure the average response time of the tools. Since topics related to genes and
proteins are the only domains that are commonly tackled by the tools, we used the protein name ‘p53’ as the query for the measurement. This protein name represents the well-known product of a tumour suppressor gene, and is thus most suitable as a query for the tools to deliver a large body of information enough to make the measurement sensible. Table 3 summarizes the response time taken by the tools in question. Note that the reported response time may significantly exceed the expected average response time across all queries, since p53 is a much studied topic in the biomedical literature, thus generating a much larger set of retrieved documents than other queries. We have not attempted to estimate the average response time due to the variety in the tools’ input format which hinders all standardization efforts.

As shown in Table 3, the tools that output an ordered (or ranked) list of documents show short response times similar to search engines (e.g. Google), except eTBLAST. The search engines build up indices for keywords in an off-line manner, which link the keywords with documents that contain the keywords. They are fast since, given a query, they only look up the query terms in the indices and display the best-ranked documents associated with the query terms. eTBLAST is slow due to the process of computing similarity between word vectors that represent the input text and the retrieved documents.

The tools that output document clusters are generally slower than search engines, due to the additional processing time for the clustering. The response time of the tools that output text fragments varies, depending on the degree of complexity of the post-processing performed after the searching of the indices. For example, MedEvi takes most of the response time to rearrange search results according to the order of the query terms and to locate statistically and semantically significant terms in the results.

The tools that make use of their own databases of gene–gene interactions deliver their information just as fast as search engines. They first retrieve a list of genes that match an input gene name, and then allow users to follow links for the retrieved genes.
to reach the information of gene–gene interactions. The tools that output co-occurrences are generally much slower than the others, since they compute the co-occurrences on the fly. The advantage of the ‘on-the-fly’ generation of data lies in avoiding all the maintenance for the databases that contain the recognized co-occurrences.

The tools with slow response time usually provide status bars that show the remaining time for the retrieval. Some provide off-line services for users to get the results via e-mail (e.g. eTBLAST) or by later revisiting the service page with unique session IDs (e.g. McSyBi).

In summary, our analysis shows trade-offs between precision, recall, coverage and speed. For instance, the tools focusing on gene–gene interactions tend to show higher precision and speed, but lower coverage, than those for general co-occurrences. It is difficult to compare the tools that output document or text fragments due to the lack of reported performance. Nonetheless, we assume that search engines in comparison to tools that output text fragments achieve higher coverage and speed but lower precision as result of lower information density. Users should consider these trade-offs when choosing tools for their information-retrieval tasks.

DISCUSSION

The fact that the biomedical literature is too large to oversee and to navigate without the aid of search engines is a big challenge for bioinformaticians. In this section, we relate some of the research topics in natural language processing to under-explored aspects of biomedical information retrieval, placing stress on promising directions of future developments in the field.

Information extraction is the task of extracting from unstructured text the information that fits a given template for events or facts of interest [43]. It has primarily been used for the population of databases. For example, the databases of tools such as iHOP, Info-PubMed and BioContrasts have been populated by information-extraction systems. The databases store references to the text from which the facts and events have been extracted, and these references can be used to locate detailed information again in the text. As explained in ‘Input format’ section, however, those tools share the limitation that they only deal with genes and proteins. As many systems that extract information about other biological types such as diseases and drugs have recently emerged [21, 24, 26], there is the need to integrate their databases into web-based services for information-retrieval purposes.

Question answering solutions are information-retrieval solutions that make use of natural language queries to retrieve relevant text fragments at high precision [44]. The solutions follow the assumption that users can make use of information-retrieval solutions more efficiently and more effectively, if they are enabled to express their need in a natural language query. The question answering systems aim to produce text fragments that precisely answer the question of the user. The output can be either named entities, if the query asks for entities that meet certain properties (e.g. ‘Which genes are involved in the Alzheimer’s disease?’), often with supporting text fragments and documents, or the relevant text alone. This approach fits perfectly well to the information retrieval needs of users in that we may translate most of information requests from users precisely into natural language queries. Since question answering technologies have to overcome an overwhelming degree of complexity in question analysis, document retrieval and answer identification, however, there are only a few on-line services in the biomedical domain, such as EAGLi (http://eagl.unige.ch/EAGLi/) and BioQA (http://cbioc.eas.asu.edu/bioQA/v2/). Even though they may have great potential, it is still too early to judge their performance. Both systems are still under development.

One of the most interesting characteristics of the biomedical literature is that the biomedical domain has several well-recognized types of objects, including genes and diseases, which are of primary concern to domain experts. It is not surprising to see that much effort has been poured in to building up manually curated databases for these types (e.g. UniProtKB, OMIM) and to develop methods for recognizing relations between objects of these types [21–26]. Recently, the Genomics Track of TREC 2007 has employed a question-answering task that introduces natural language queries with variables for pre-defined types of entities (e.g. ‘What centrosomal [GENES] are implicated in diseases of brain development?’) [35]. We think that these variables can be
effectively used for the purpose of semantic type restriction of Boolean queries in biomedical information retrieval. As for the implementation of the restriction, the instances of the variables can be recognized by existing methods for named entity recognition \[45\], while the semantic relations between the variable instances and their context have been dealt with to a limited extent by the information-extraction techniques. Textpresso and MedEvi are early adopters of the semantic type restrictions in Boolean queries.

The existing tools often provide a means for users to navigate throughout the results of the tools, for example, with graphical interfaces (e.g. Ali Baba), with hypertext links to related documents or database entries (e.g. EBIMed, iHOP, MedEvi), and with semantic clustering of retrieved documents (e.g. HubMed, GoPubMed). Since the number of retrieved documents can easily exceed the capacity of users for individual reading even with the help of the navigation interfaces, it is desirable to generate a shortened text from the retrieved documents, which is as informative as possible \[26\]. Text summarization is the field of natural language processing that has the same goal \[46\]. In the biomedical domain, we find a few scientific publications about automatic generation of textual summaries for genes from MEDLINE abstracts \[47–49\]. This approach would be applicable to summarization of any textual search results.

One of the most wanted aspects in the biomedical information retrieval is the behavioural mode of ‘differentiating’, as explained in ‘Differentiating’ section. If we assume that research is mainly driven by the comparison of new discoveries with known ones, it is rather awkward to see that there are few tools for aiding the comparison task. This lack might be due to the assumption that the comparison requires deep semantic analysis of text, which is seen as computationally hard at the moment. If this is indeed the case, BioContrasts may be an example to show how to focus on specific types of semantic information for the differentiating tasks with simplified computational methods \[9\].

**CONCLUSION**

Biomedical literature is an invaluable repository of public knowledge of high quality in biomedicine with remarkable characteristics such as the vast number of records, the well-defined underlying typology for both physical and functional concepts and the increasing association with well-established databases. The tools for seeking information from the biomedical literature are major assistants for biomedical researchers to explore the literature for their purposes. What we have presented in this article are, at least, minimal guidelines for researchers to do so in a competitive way. In general, information retrieval is a creative activity; it cannot be directed by simple step-by-step guidelines. We have rather introduced the existing tools in a ‘user-friendly’ way without detailed explanation of incorporated techniques, and have revealed under-developed and still much wanted research fields in biomedical information retrieval.

**Key Points**

- No single service for information retrieval can meet all information demands of biologists for their research.
- We first cluster the existing services for seeking information from the biomedical literature, based on their input and output formats—the basic requirements for the usage of the services.
- We also compare the unique features of the services with different behavioural types of information-seeking activities, for better understanding for their potential use.
- We describe the usability and the reliability of the tools, along with their reported performance and estimated response time.
- The typological explanation of the services may give guidance to bioinformaticians suggesting promising research fields in biomedical information retrieval.

**Acknowledgements**

We express our gratitude to Adam Bernard, Vivian Lee, Piotr Pezik, Antonio Jimeno Yepes, and anonymous reviewers for their valuable comments on the paper. This work was sponsored by the EC STREP project “BOOTstrep” (FP6-028099, www.bootstrep.org).

**References**

34. Voorhees E. Overview of TREC. In: *Proceedings of the Text Retrieval Conference* 2003, pp. 1–14, Gathersburg, Maryland, USA.


