A Survey on Vertical Search in the Biomedical Domain

Parikshit Sondhi and Chandra Ramachandran

Department of Computer Science
University of Illinois at Urbana-Champaign
1. Introduction

According to some estimates, nearly 4.5% of the searches on the web pertain to health related information [1]. This means that close to 10 million users, around 6.75 million on Google alone, search the web for biomedical information everyday. With an explosion of biomedical literature (nearly 17 million citations in PubMed [2]), and a plethora of biomedical websites, the biomedical domain is in an urgent need of tools for efficiently retrieving relevant information. In this report we present a systematic and thorough survey of the information retrieval tools and search engines available for the biomedical domain.

The report is organized as follows. In the second section we present a discussion on the kinds of information that are available in the domain and the various user expectations and requirements. Section three presents the domain specific issues that render the general search engines ineffective. In section four we discuss the different domain specific resources available, and in section five, we describe the current vertical search engines for the domain and discuss their features & deficiencies. Finally in section six, we present our conclusions and recommendations for the future.

2. Users and Information Preferences

Unlike most other vertical domains like travel or restaurant search, where all users usually look for similar kind of information, users in the biomedical domain can be divided into three distinct classes: the doctors, the researchers and the laypersons. Each category of users has a different level of expertise and information requirements. For example while doctors often need to search for clinical studies, experiments and a comparatively technical information on treatments, symptoms & drugs etc. at the point of care, laypersons usually want simple language documents about diseases, treatments, clinics, doctors etc. The kind of sources users search from, therefore depends on these information needs. A discussion of the various categories of information sources available in the domain is provided below:

2.1 Published Literature

This category comprises of all the biomedical literature that has been published in various journals and conferences over the years. PubMed [2] maintained by the National Library of Medicine maintains a database of all these publications and makes them freely available to the public. Over the past few years, there has been an explosion of published information. The database currently contains close to 17 million citations and has been growing exponentially over the past few years. This class of information is searched for almost exclusively by doctors and researchers.

2.2 Biomedical Websites

While we could not find any fixed statistics on the number of biomedical related websites that are out there, it can be reasonably estimated that their number is huge. The open directory project which maintains one of the largest human edited directories on the web (but is in no way comprehensive), contains around 61,500 websites marked under category “Health” [4]. The types of information available on these websites include articles, forums, blogs, advertisements, portals for clinics or companies selling biomedical related products etc. These websites are searched for by all three categories of users. However
while doctors and researchers need only focus on a very small set of reference websites, it is the
laypersons who are the major users in this area and most of the websites are targeted to them.

2.3 Other Sources

Apart from the textual information available in the form of published literature and websites, there are
also various databases pertaining to biomedical images, gene information or other research datasets.
These are meant mostly for researchers. For the purpose of this report we will ignore this category and
focus only on retrieval of textual information from the first two.

3. Domain Specific Characteristics

The biomedical domain possesses a unique set of characteristics that render most of the standard search
techniques ineffective. For example in a study conducted with a test set of 100 medical queries collected
from medical students in a specialized domain, a thorough search in Google was unable to obtain relevant
documents within top five hits for 40% of them [5]. A discussion of these characteristics is provided
below:

3.1 The Biomedical Language

The biomedical domain contains a well established, very specialized, highly developed and sophisticated
jargon. It comprises of synonyms, abbreviations, acronyms, common terms as well as scientific names
which are widely and very frequently used. Search engines therefore need to be able to understand such
complex language. For example:

“...there any recorded cases of endometriosis presenting with oral symptoms (gums bleeding) and
intracranial symptoms (migraine)...”

Moreover, the bio-medical literature also contains complex technical terms, long multi word expressions,
term order variations, abbreviations etc and uses them inconsistently. For example:

“Melanogaster 5'-phosphoribosylaminoimidazole carboxylase-5'-phosphoribosyl-4-(N-
succinocarboxamide)-5-aminoimidazole synthetase (Adel) mRNA, complete cds”

is a single multiword expression.

A biological term “egfr-1”, has a number of the variations such as “egfr1”, “egfr.1”, “(egfr-1)” or “egfr 1”

*Pulmonary Tuberculosis* is a multiword expression which is both a *Respiratory Disorder* and an
*Infectious Disorder*.

Biomedical literature also has a characteristic feature of using numeric figures in naming concepts. For
example

NIH 3T3 fibroblasts

CYP1A1 promoter
3.2 Trustworthiness of Documents

Since the information returned by a biomedical question search engine is often of critical importance, for eg. it might influence the course of a patient’s treatment, the trustworthiness of the documents being used is essential. Due to this constraint most of the current systems, have restricted themselves to only the published biomedical literature. However web sources are essential for catering to the needs of laypersons. In fact currently there are no automated methods for ensuring a document’s trustworthiness. The users instead need to follow a set of guidelines such as checking the credentials of the authors, the credentials of the owners, last update date etc. It is clearly very cumbersome for a user to not only look for relevant results, but also ensure their trustworthiness. It would be much more convenient if the search engine could automatically assign a trust rating to the results.

3.3 Time Constraints

Another important characteristic of the domain is in context of doctors. Doctors work under serious time constraints and do not have much time to search for answers at the point of care. According to a study [6], physicians spend less than 2 minutes on average seeking an answer to a question. However, the average time required to obtain an answer ranges from 2.4 to 6.5 minutes [7]. Thus leaving most clinical queries unanswered.

3.4 Level of Difficulty

Finally due to the vast gap between the levels of expertise of different categories of users, the systems should have some mechanism for automatically identifying the difficulty level of a biomedical document and presenting it to the correct category of users. Some research in this area has been done by [8].
4. Resources

While resources for developing open domain search engines are well known, in this section we present a short survey of domain specific resources available for the biomedical domain.

- **MEDLINE**: MEDLINE [9] provides free access to over 4,800 journals published in the United States and more than 70 other countries primarily from 1966 to the present.

- **MedlinePlus**: MedlinePlus [3] includes over 4,000 articles about diseases, tests, symptoms, injuries, and surgeries. It also contains an extensive library of medical photographs and illustrations.

- **PubMed**: PubMed [2] is a free search engine offering access to the MEDLINE database of citations and abstracts of biomedical research articles. It is offered by the United States National Library of Medicine as part of the Entrez information retrieval system. PubMed also offers nearly 17 million citations which contain links to freely available full text articles, often on the PubMed Central digital library.

- **MeSH Terms**: MeSH [10] is the U.S. National Library of Medicine's controlled vocabulary used for indexing articles for MEDLINE/PubMed. MeSH terminology provides a consistent way to retrieve information that may use different terminology for the same concepts. MeSH database can be used to find Medical Subject Heading Terms and build a search strategy.

- **Gene Ontology**: GO [11] consists of 19,000 terms organized in three sub-ontologies for cellular location, molecular function and biological process.

- **Unified Medical Language System**: NLM’s Unified Medical Language System (UMLS) [12] project develops and distributes multi-purpose, electronic "Knowledge Sources" and associated lexical programs for system developers and researchers. They are useful in investigating knowledge representation and retrieval questions. UMLS contain three knowledge sources: Metathesaurus, Semantic Network and Specialist Lexicon. The main component is the Metathesaurus, which compiles and cross-references one hundred biomedical terminologies (in version 2003AA: more than 800,000 concepts and 2,000,000 strings), with their hierarchical and transversal relations. Its Semantic Network adds a common structure above these imported terminologies. Additionally, its Specialist Lexicon provides a large English lexicon with an emphasis on biomedical words, including derivational knowledge. Tools have been built around the UMLS to address terminological variation.

- **MMTX**: MMTX [13] is a tool to perform the task of mapping biomedical concepts to free text. The biomedical concepts used for mapping are taken from the Unified Medical Language System. The system is also capable of identifying multi-word expressions, synonyms, abbreviations, term variants and stop words.

- **SPECIALIST Tools**: The SPECIALIST NLP group [14] provides a large list of useful NLP tools for tasks like word sense disambiguation, spelling correction etc. meant specifically for the biomedical domain.

- **Health on Net Foundation**: The Health on Net Foundation (HON) is a consortium that certifies websites as trustworthy or not. In the process of certification, first the webmasters apply for
certification, the HON staff then rigorously analyze the website based on certain criteria [15, 16] and finally if satisfied, provide a digital HON certificate and also make the website searchable through their own search engine. While this is perhaps the only major effort for identifying trustworthy websites on the internet, it has its own drawbacks. Since the process of certification is manual, there’s only a limited number of websites that can be reviewed within a given period of time. Moreover reviews take place only once a year during which the content of a website may change considerably. Also at times certain rogue websites continue to display the HON certified seal inspite of their requests being rejected.

5. Current Search Engines

In this section we describe the various search engines available for the domain. After a thorough survey, we found around 15 relevant domain specific search engines. Based on their characteristics (the target users, data sources etc.), we classified the prominent ones into five major categories as shown in table 1. Tables 2 and 3 provide the data sources and search features that these systems provide. In the following subsections, we give a brief overview of each category while focusing on the following:

- The main sources including medical journals, web, health blogs and forums from where information is gathered.
- The process by which new biomedical sources are added to each.
- Target users, in terms of doctors, researchers or laypersons.
- Technologies adopted in the information gathering, generation and display process.
- Search functions provided and additional features available such as the ability to ask questions online from doctors.

Table 1. Major categories of biomedical search engines

<table>
<thead>
<tr>
<th>PubMed Based</th>
<th>HON Sites Based</th>
<th>Site Specific</th>
<th>Google/Yahoo Custom Search</th>
<th>General Biomedical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rele-Med</td>
<td>MedHunt</td>
<td>WebMD</td>
<td>Vadlo</td>
<td>iMedix</td>
</tr>
<tr>
<td>GoPubMed</td>
<td></td>
<td>Medic8</td>
<td>OmniMedical Search</td>
<td>Healia</td>
</tr>
<tr>
<td>Entrez</td>
<td></td>
<td>MedNets</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medscape</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2. Data sources and target users for each category

<table>
<thead>
<tr>
<th>Category</th>
<th>Data Source</th>
<th>Target Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>PubMed Based</td>
<td>PubMed and MEDLINE</td>
<td>Doctors/Researchers</td>
</tr>
<tr>
<td>HON Site Based</td>
<td>HON Certified Websites</td>
<td>All Users</td>
</tr>
<tr>
<td>Site-Specific</td>
<td>Individual Websites</td>
<td>Mostly Laypersons</td>
</tr>
<tr>
<td>Custom Search</td>
<td>A selected set of Websites</td>
<td>All Users</td>
</tr>
<tr>
<td>General Biomedical</td>
<td>All Biomedical Websites</td>
<td>Mostly Laypersons</td>
</tr>
</tbody>
</table>

Table 3. Interesting features in each category of search engines

<table>
<thead>
<tr>
<th>Category</th>
<th>Interesting Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>PubMed Based</td>
<td>Categorization of Results, Relevance Ranking, MeSH term based search</td>
</tr>
<tr>
<td>HON Site Based</td>
<td>None</td>
</tr>
<tr>
<td>Site-Specific</td>
<td>Search in Blogs, forums, On-Call Doctors</td>
</tr>
<tr>
<td>Custom Search</td>
<td>Searching Powerpoint Presentations, Softwares etc.</td>
</tr>
<tr>
<td>General Biomedical</td>
<td>Social Networks, Personal Search</td>
</tr>
</tbody>
</table>

5.1 PubMed Based Search Engines

The first category is search engines based on the MEDLINE or PubMed databases. These search engines are predominantly meant for doctors and researchers. They index abstracts, titles, author-names and the content of each publication that they obtain from the PubMed databases. When a user performs a search, they do a direct keyword search across these indexed attributes and return results. Some of them also utilize domain specific sources such as MeSH or GO to provide a rich set of search features. The advantages of using articles indexed by PubMed are that their content is always reliable and freely available. However no statistics or studies are available on the relevance of their search results barring PubMed, where it is estimated that doctors require between 2.5 – 6 minutes to find relevant information [5].

ReleMed: The first search engine that we have surveyed in this category is ReleMed [17, 18]. This searches over 17 million articles from the MEDLINE database. Additional features available are options to do a Google search, and searching for specific entities like newsletters, editorials and clinical trials. No further data-processing is done in this search engine. The search functionality provided is limited, and hence it performs poorly as compared to other search engines in the same category. ReleMed was specifically created to handle multiword queries. Among the features provided, it uses sentence-level concurrences as a means of establishing relationships among query terms. The results provided are suitable for researchers and medical professionals. This is the only search engine in this category which estimates a relevance score for search results, and the results are ordered in decreasing relevance. The bulk of data gathering is in the form of XML data obtained with the help of a contract with NLM. Using standard XML parsers, it extracts title, abstract and citation information from each XML record. Table 4 shows a typical database schema used. Finally ReleMed is based on the LAMP architecture (Linux Apache MySQL Perl) [19] for its data gathering, processing with a back-end database and for displaying results.
Entrez: Entrez [20] is a life-sciences search engine hosted by the National Center for Biotechnology Information (NCBI) [21]. It permits search across multiple databases, primarily over PubMed. Entrez creates links these databases together for performing search. The databases not only include PubMed journals but also contain protein structure databases, population study datasets, gene expression data and so on. The entire list is available at [22]. The database also searches across NCBI websites and certain FTP sites. Similarity searches are performed using the BLAST algorithm [23] to compare the queried amino acid or DNA sequences to the remaining information available in the database. This is then linked to the corresponding PubMed citation where that sequence was published. An interesting feature here is that the search can be customized according to the type of database being searched. For example, to search the protein database we can search by 'molecular weight'. The target users here are medical professionals and researchers. Entrez uses a set of programming tools called as NCBI Entrez Programming Utilities (E-utilities) to automate search and retrieval operations across the entire set of Entrez databases. Even though the internal architecture of Entrez is not available, users can customize their own implementation of E-utilities using Perl and SOAP in Visual C#. Figure 1 shows the results of a query for 'DNA' in the Entrez protein sequence database, and this highlights the versatility of this search engine.

Figure 1. Figure showing the results of a query for 'DNA' in the Entrez protein sequence Database

<table>
<thead>
<tr>
<th>PubMedID</th>
<th>Sentence ID</th>
<th>Text</th>
<th>Citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12</td>
<td>...</td>
<td>Author A, Author B...</td>
</tr>
</tbody>
</table>
**Medscape:** Medscape [24] is a primarily a health resource rather than a search engine and it is aimed at medical professionals and physicians. The primary source of information is journals from MEDLINE. It also has daily medical news, conference coverage and also search from a drug database called as Medscape Drug Database or MDR. Medscape provides results from both its internal sources (that have been contributed on by medical professionals in its parent organization, WebMD [25]) as well as results from MEDLINE for a particular search query. Medscape provides categorization of results into 'Book Reviews', 'Case Reports', 'Clinical Cases' and so on. Information regarding the internal architecture of Medscape is currently not available. Figure 2 shows the results of a query for 'cancer' in Medscape, and this serves to highlight the limited functionality provided by Medscape. This search engine also does not provide any relevance ranking of results, and as such is not very powerful as compared to the other search engines.

![Medscape Search](https://search.medscape.com/jk-search?queryText=cancer)

Figure 2. Figure showing the results of a query for 'cancer' in Medscape

**GoPubMed:** GoPubMed [26] is one of the first knowledge based search engines for biomedical data, and is based on the PubMed database. It is aimed at medical professionals primarily and allows for easy navigation to the appropriate articles by using a Gene Ontology (GO) and MeSH categorization of abstracts. For example, a query for 'AIDS' provides categories such as 'organisms', 'diseases', 'geographicals' etc. GoPubMed features an extensive use of AJAX to provide suggestions to users as they type their queries. A term extraction algorithm has been developed which uses local sequence alignment of words in both the abstract and the GO terms, and helps in ranking the importance of a particular word.
Figure 3 shows the results of a sample query for 'cancer' in GoPubMed with the top categories for the query highlighted in red. Figure 4 shows the other categories for the query highlighted, and obtained using the GO and MeSH database. From a video demonstration of GoPubMed [27], it is clear that its relevance ranking is significantly higher as compared to a simple keyword-based search of the PubMed database.
5.2 Search Engines Based on HON-Certified Websites

In order to ensure the reliability of search results, certain web based search engines either search only within the Health on Net Foundation certified websites or at least specifically mark whether a displayed result is HON certified or not. However as we have already described above, reliability of the HON certification itself is sometimes questionable and it does not cover the entire set of biomedical websites.

**MedHunt:** One of the major websites searching through only HON-certified websites is MedHunt [28], the search engine developed at HON itself. This is aimed at both the general public as well as medical professionals. MedHunt uses the terminology provided by MeSH. Users are also offered a list of further medical or health-related options that appear related to a given query and a spell checker to help refine their search.

The data gathering for MedHunt is performed using MARVIN. MARVIN (Multi-Agent Retrieval Vagabond on Information Networks) searches the Web and selects only documents that are relevant to a specific and chosen domain. The relevance of a document is computed based on a formula that considers the number of MeSH terms that MARVIN finds in the document, as well as their place in the document. It finally stores the selected documents in a database that users can then query via MedHunt.

Whenever a query is entered, the following relevant information is displayed.

- Query relevance
- The weblink and its description
- Place where the server is hosted (for e.g. DNA biosciences is hosted at chelmsford, UK)
- Keywords related to the query
- Whether the site is HON certified or not
- When was the link last monitored by HON
- Language of the website
- Date the site registered with HON
- Date it was last crawled by MARVIN

These details are highlighted in Figure 5. MARVIN searches two types of websites: annotated websites which are updated daily and created manually by the HON and auto-indexed websites which are automatically retrieved from the web by MARVIN. The internal architecture of MedHunt is currently not available.
5.3 Site-specific Search Engines

The next category of search engines search only from within their own websites. While search on a single website is too small in context of a vertical search engine, we list them here just to contrast the kind of features they provide. These engines gather and collect data from various sources such as in-house doctors, registered users and bloggers, through site-specific forums etc. However they are not always scalable as some rely more on manual rather than automatic methods of gathering information. One of the advantages of site-specific search engines is that there is a high-degree of control over the kind of results being displayed.

**Medic8**: Medic8 [29] is a UK medical information resource for healthcare professionals and consumers. All content has been reviewed by qualified UK doctors prior to listing. It provides online medical guides (such as guides to cosmetic surgery) which have been contributed by the registered users and medical professionals employed by Medic8. It organizes health topics by themes such as allergies and also provides search over its online medical dictionary. The data gathering and extraction process is manual.

**WebMD**: WebMD [25] is also a UK based resource which provides health news, articles (written by registered users, the community and medical professionals) and displays the top 12 health topics on the main page. It is targeted mainly at the general public more than medical professionals. It has its own logs and community boards, and displays the top 3 items (about which people are talking about). It also provides a directory of health topics. Users can ask questions to the community and to certified doctors of WebMD. The answers displayed are a mix of doctor's answers, news headlines and articles related to the question and community responses. It also categorizes its articles into pre-defined categories such as...
healthy living, drugs, first aid and so on. Additional search features include 'find a doctor', 'find a hospital' and 'user reviews on pills and drugs'. This is unique to WebMD and is not provided by any other search engine. It does not index published literature and the results returned are from its own website only.

**MedNets:** MedNets [30] is a search engine targeted at general users. Currently, the search engine is incomplete and many of the queries that we tried returned no results. It is based on a technology called MedExplore and has its own searchable database of 20000 articles. Right now it focuses on internal medicine. An interesting aspect of MedNets is that it provides links to Searchable databases, journals, associations and societies, news sources, jobs and books, all of which have been gathered manually. The current search functionality searches the MedNets website, NIH guidelines and National Library of Medicine [31].

### 5.4 Google or Yahoo! Powered Custom Search

This category of search engines are powered by Google Custom Search [32] or Yahoo! alpha [33]. These search engines feed a specific set of URLs to the corresponding search engine and the user queries are then search in the index based on these URLs. Some of these websites provide additional features such as social networks, ability to search through protocols, online tools and software. However a drawback of these sites is that while they are able to improve search results by searching only within a small set of websites, these lists of websites are often not comprehensive. Moreover such custom search engines suffer from the same set of drawbacks as the general search engines and are unable to handle the complicated biomedical language.

**Vadlo:** The first search engine in this category is Vadlo [34]. Vadlo indexes five categories of information: Protocols, Online Tools, PowerPoints, Databases, and Software and provides searches across these categories. Figure 6 shows a typical results page for a query 'DNA'. Among these categories, the ability to search through powerpoints is an interesting feature provided by Vadlo, and is not available in any other search engine. Vadlo indexes PowerPoints presentation which refer primarily to: biology research, academia, bioinformatics, biology education and medical/clinical information.
OmniMedicalSearch: OmniMedicalSearch [35] is a multipurpose biomedical search engine which provides the maximum number of options for search among the search engines that we have surveyed. It searches over web, medical journals, dictionaries, acronyms, groups and associations, forums, medical image databases. Users can suggest additional links to either of the above categories. The data sources are gathered using a Google custom search. For other categories such as acronyms, it uses popular toolkits like medilexicon [36], and for dictionaries it uses OneLook [37], yet another toolkit in this domain. This search engine caters to all categories of users.

5.5 General Biomedical Search Engines

This category includes those search engines which build their own index over a large collection of biomedical websites, not restricted to HON and are not based on Google or Yahoo custom search. We found only two search engines in this category iMedix [38] and Healia [39, 40]. They seemed to be the best suited for laypersons. While both of them allow searching over a broad base of biomedical websites, they do specify which of the results are HON certified.

iMedix: iMedix [38] is a biomedical search engine with a patient-to-patient social network. The site is based on the open-source Lucene search engine, which then uses relevance feedback to improve results. The site uses AJAX to suggest and guide the user when a query term is typed. Along the side column are profiles of iMedix members who may be interested in health topics related to the search. Relevant health articles contributed by registered users are also displayed. iMedix is mainly targeted to the general public rather than medical professionals. For ensuring trustworthiness of results, all HON certified links contain a “Accredited by: HON” tag. Figure 7 shows an example results page for a query, ‘cancer’ in iMedix. There however appeared to be a problem with the display of HON certification tags. Every result seems to get a HON accreditation. When we tried a query ‘Barack Obama’, even results from Wikipedia were shown as HON certified! Figure 8 shows results of this query.
Figure 7. Figure showing the results of a query for 'cancer' in imedix.

Figure 8: Results for the query “Barak Obama”
Healia: Healia [39, 40] is a multipurpose search engine which searches the web, medical journals (from PubMed) and clinical trials. In addition it provides health guides and community blogs contributed by users and medical professionals. Users can also ask questions from doctors who are available and registered with Healia. The major functionality in Healia comes from the ability to search through medical journals indexed from PubMed. HON certified results are tagged with “HONcode accredited”. Community blogs contain discussions on the most commonly occurring ailments such as allergies. The most popular user queries to the community are also displayed along with their responses. Users participate actively in the community blogs and question-answer sessions. The main users in Healia are patients and consumers. One interesting feature in Healia is the ability to do personal search. For example, one can filter search based on: Professionals (those with a higher level of biomedical knowledge), Gender, Age groups, Ethnic groups and so on. Alternate results are also suggested in Healia. Details regarding the internal architecture of Healia are proprietary (pending patents) and hence are unavailable for analysis.

6. Conclusion and Future Recommendations

In this report we presented a thorough survey of vertical search in the biomedical domain. We identified the different categories of users and information sources and described the various domain specific issues that render general search engines ineffective. We also discussed the available tools & resources and analyzed the currently available search engines. Our analysis makes it clear that while quite a few promising vertical search engines have surfaced, they still need to go a long way to satisfy the needs of users. Some of the prominent issues and our recommendations in this regard are discussed below.

6.1 Lack of Trustworthiness

The first major issue pertains to trustworthiness of web based content. While HON certification is useful to some extent, its not entirely reliable. Moreover, since most of the general public is not even aware of it, the users simply don’t care about checking the HON certifications. Automated systems capable of identifying a websites trustworthiness based on the NLM guidelines are therefore required for ensuring delivery of reliable results.

6.2 Lack of Comprehensiveness

Another major issue with the current breed of biomedical search engines pertains to their lack of comprehensiveness. Out of all search engines we surveyed, only one (Healia) appeared to index both a very large collection of biomedical websites and the PubMed collection, all others were either based on PubMed or a small subset of biomedical websites. However, even in case of Healia, a search for a slightly technical term “intercranial bleeding” retrieved just 15 results, compared to Google which produced around 1 million. This lack of comprehensiveness drives the users to simply use Google, even though its results are not always relevant and lack trust ratings.

6.3 Fragmentation of User Base

From the survey it is fairly clear, that different search engines cater to different sets of users. Even search engines that claim to use both PubMed and web based sources, are not well optimized for both cases. This is not only bad for the users, but also for the search engine companies themselves, as they only target a subset of the market share. For example, while search engines such as Entrez or GoPubMed are fairly good for searching published literature, even doctors frequently need to search the web and they must therefore use multiple search engines to cover all sources. This is both inconvenient and inefficient, especially when answers are required within two minutes at the point of care.
6.4 Text Analysis
While automatic analysis the technical level of documents might be a fairly useful is delivering results based on a user’s expertise, we did not find any search engines implementing it. From the literature it appears that fairly good solutions for the problem are available [8] and therefore we feel that it should be implemented.

6.5 Use of Domain Specific Resources
Unlike most other domains, a rich set of tools and resources is available for the biomedical domain. These can be used to perform a fairly deep level of text analysis on the available documents and their use can help take biomedical information retrieval to an entirely new level. It is for example possible through the use of MMTX and SPECIALIST tools to identify different categories of biomedical entities like disease names, drug names etc. within texts. Moreover the UMLS lexicon also defines relationships between all these categories. This information could be used for example to better organize the indexes, display a summary of information pertaining to any given entity or at the bare minimum to improve search results. However we rarely saw search engines using them. Only GoPubMed and Entrez for example used ontologies to search and organize the display of results.
7. References


