

Information Retrieval Architecture for Text Mining of Biomedical Literature

Report: Saurav Sahay
CS6400 - Spring 2005

1. Abstract

Text Mining for Biomedical domain is an active area of research that is seeing a plethora of new tools and techniques to accomplish its purpose. All these systems mostly try to integrate the Medline database, other databases that describe about the entities in the Medline abstracts and some form of natural language processing (NLP) to aid in interpretation of this information.

In this study, I am exploring many available biomedical information retrieval systems. I am also working on creating my own prototype system that accomplishes some of these challenges meted out in this domain. To begin with, I am looking at indexing capabilities and gene ontology database to combine and integrate with a sample local Medline database that I created for this purpose. I describe the general characteristics of the biomedical information retrieval system in the following sections followed by a study of existing systems and the proposed system to accomplish these challenges.

2. Text Mining

Text Mining is the process of discovering knowledge from unstructured text. All scientific text primarily remains in this form even today. Due to the exponential growth in size of this corpus, this field has emerged recently to develop schemes and methods for automated evaluation of text documents efficiently. The steps of text mining can be classified into the following components:

1. Gathering of text documents (automated/manual extraction from web sources)
2. Text preprocessing (semi-structuring the text by using databases, XML, etc.)
3. Natural Language Processing (Entity tagging or labeling, term identification)
4. Text Categorization (Classification or Clustering)
5. Visualization (Interface, graphical representation)
6. Analysis (evaluation of extracted information)

As we can see, each of the above steps are broad research domains in itself, the process of text mining needs an efficient integration of these steps for knowledge discovery. In the biomedical domain, the process of text mining is eased because of consolidated resources for text gathering and classification of the biomedical concepts using ontologies by active research community.

Nevertheless, we have problems pertaining to the usage of the terms and their synonyms, the unavailability of full text sources and probable loss of information, and the huge amount of literature that needs to be processed online for information extraction. There is a need to scale up the algorithms and make use of distributed processing and move beyond single term searches to phrase identification, efficient indexing and ranking results.

3. Natural Language Processing

Natural Language Processing (NLP) applied in the domain of Biomedical literature mostly consists of rules to resolve ambiguity in naming genes and proteins, terminology variance and finding semantic associations between entities. (An estimated 36% of acronyms in Medline have more than one definition) Most systems do not worry about misspellings, accented words or multi-lingual resolution because of the formal nature of standard English scientific publications.

A number of rule-based, linguistic, statistical, machine-learning, and hybrid approaches have been developed to mark up gene/protein terms (synonyms and full names) automatically in biological text. Some other approaches include referring to knowledge sources such as GenBank and Swissprot. Simple rules like morphological cues (upper case and special characters), use of suffixes such as -ase (as in kinase), rules such as "connect non-adjacent annotations if every word between them is either noun, adjective, or a numeral" has been applied to identify multi-word protein terms.

For finding semantic associations, knowledge sources such as UMLS (Unified Medical Language System), GO (Gene Ontology) and MeSH (Medical Subject Headings) have been used extensively. UMLS Semantic Network defines binary relations between semantic types. The concepts are related in several ways and are sometimes too general to convey meaning to the documents. Therefore, UMLS based annotations are needed to be improved and several NLP techniques are again employed for the purpose.

4. Ranking

Ranking attempts to measure how relevant documents are to particular queries by inspecting the number of times each search word appears in the document. There are several flavors of ranking, ranging from Google's complicated ranking scheme based on PageRank algorithm to a simple yet effective ranking based on tf-idf scheme. Keyword based tf-idf ranking can be performed at the database management system level itself using any of the opensource or commercial DBMS systems.

Ranking in the context of biomedical information retrieval may be questionable when the task is to find new relationships between entities. There may be a

need to look at all the documents/abstracts retrieved by the system. Still, the fact that ranking is based on keyword statistics in the corpus, we may hope to find relevant information from the highly ranked documents. Ranking using feedback mechanism (user feedback/automated feedback) is even more effective in terms of getting to the desired results. There are several relevance feedback algorithms defines for this purpose. Basic idea is to do an initial query, get feedback from the user as to what documents are relevant or select the top ranked results and then refine the search by adding words from known relevant document to the query.

5. Clustering

In the biomedical domain, we are dealing with information retrieval from millions of documents. Clustering is an unsupervised learning problem where we need an automated way to organize this collection into documents relating to biomedical concepts. Documents in a cluster have 'similar content' defined by salient terms that are common to those documents. Documents can be organized as feature vectors consisting of words in the document. The vector is based on the bag of words model approach consisting of all the words represented in the document. Several clustering techniques have been applied to cluster documents, genes based on functional keyword association, and protein names. The clustering process has some notion of distance measure between feature vectors and documents close to each other are clustered together. We need to define the initial number of clusters and data points are chosen randomly and the closest data points are clustered together. Some statistical approaches have been applied to find the natural number of clusters in the collection.

Some of the current popular clustering techniques include hierarchical clustering, k-means, clustering based on SOM and graph based clustering. Clustering is also similar to the database segmentation problem where we have to effectively group similar data together for efficient querying. Ideas from this domain have also been applied for clustering of genes based on functional keyword similarity (Bond energy Algorithm)

6. Classification

Text classification is a supervised learning problem where we know the labels of the documents (specified by domain experts) and train the corpus to effectively predict unknown future data in the right classes automatically. Repositories like Medline are constantly increasing in size exponentially and they need to assign new documents in the right categories in an automated and reliable way.

Text classification is carried out by transforming documents, which typically are strings of characters, into representations suitable for the learning algorithm and the classification task. This corresponds to stop word elimination, stemming and preparation of feature vector. Each distinct word then corresponds to a feature. This representation scheme leads to very high-dimensional feature spaces. The dimensions of the feature vector can be reduced by using term frequency- inverse document frequency (TFIDF) thresholds and other dimensionality reduction techniques to improve performance. To abstract from different document lengths, each document feature vector can be normalized to unit length.

Current levels of precision of manual search and scan is about 12%. With machine learning approaches like Support Vector Machines, we get precision values in the range of 60-70%, a significant improvement. Furthermore, we get very high recall values meaning that we don't get much false negative values and have very few misses of relevant data.

7. Ontologies

Ontologies are an explicit formal specification representing objects, concepts, and the relationships among them within a defined area of interest. They are usually hierarchical and interconnected. Ontologies provide a standardized vocabulary for representing and communicating knowledge about objects and their relationships to one another. Ontologies exist in several different fields, but perhaps the best known effort is the gene ontology (GO) project. Ontologies help remove inconsistencies among gene names and their multiple function across many species. This inconsistency impairs the ability of users and computers in identifying genes with a relevant function. Because the ontological terms and the relationships between them are carefully defined by domain experts, the use of ontologies helps standardize annotations, improve information retrieval, and supports the construction of inference statements.

7.1 UMLS

The purpose of NLM's Unified Medical Language System[®] (UMLS[®]) is to facilitate the development of computer systems that behave as if they "understand" the meaning of the language of biomedicine and health. To that end, NLM produces and distributes the UMLS Knowledge Sources (databases) and associated software tools (programs) for use by system developers in building or enhancing electronic information systems that create, process, retrieve, integrate, and/or aggregate biomedical and health data and information, as well as in informatics research.

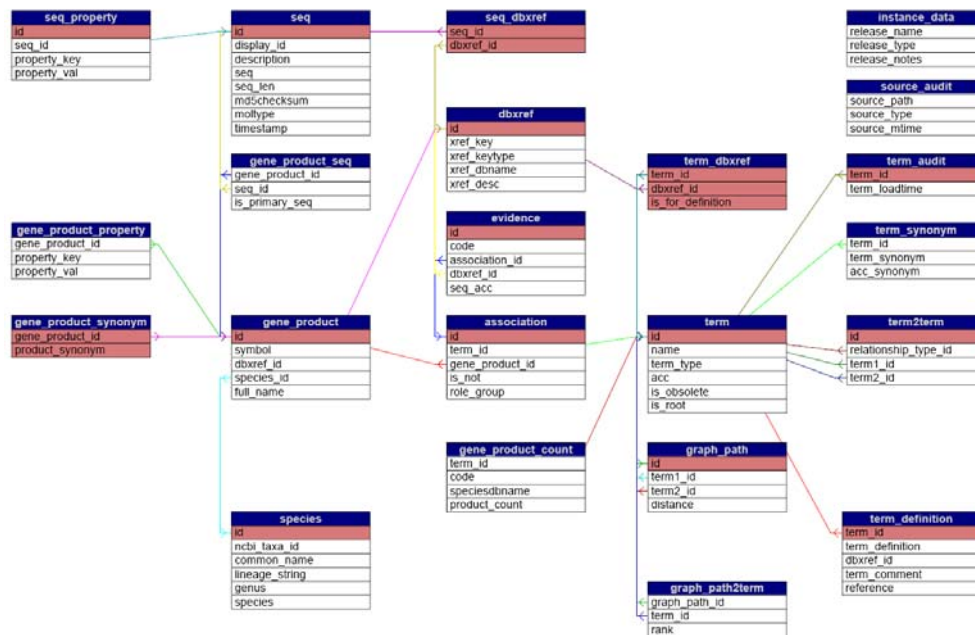
There are three UMLS Knowledge Sources: the Metathesaurus[®], the Semantic Network, and the SPECIALIST lexicon. The Metathesaurus is a very large, multi-purpose, and multi-lingual vocabulary database that contains information about

biomedical and health related concepts, their various names, and the relationships among them. The purpose of the Semantic Network is to provide a consistent categorization of all concepts represented in the UMLS Metathesaurus and to provide a set of useful relationships between these concepts. The SPECIALIST lexicon has been developed to provide the lexical information needed for the SPECIALIST Natural Language Processing System. [http://www.nlm.nih.gov/research/umls/about_umls.html]

7.2 GO

The Gene Ontology (GO) project is a collaborative effort to address the need for consistent descriptions of gene products in different databases. The three organizing principles of GO are molecular function, biological process and cellular component. A gene product has one or more molecular functions and is used in one or more biological processes; it might be associated with one or more cellular components.

In the UML diagram below, we see relationships between different terms, their definitions and their associations with gene products and their synonyms across species along with several other bookkeeping elements.



8. Existing IR Systems

Many IR systems have been developed recently for extracting knowledge from biomedical literature. Most of these systems work on a local copy of MEDLINE data. MEDLINE is the National Library of Medicine's premier bibliographic database covering medicine, nursing, dentistry, veterinary medicine, the

health care system, and the preclinical sciences. MEDLINE contains over 11 million bibliographic citations and author's abstracts from more than 4,000 biomedical journals dating back to the mid-1960's. Some of these systems have been described here to understand their capabilities.

8.1 MyMed

MyMed is a database system that mirrors Medline data locally on IBM DB2 infrastructure and uses DB2 XML Extender and Text Information Extender. The system supports advanced queries, wildcard searches, proximity searches and scoring functions. It uses huge hardware infrastructure of many distributed systems and 32 GB RAM to efficiently manage the search and retrieval process by multiple users. It does not have any additional capabilities of integrating ontologies, clustering and visualizing the data efficiently. Moreover, it is not available to the general public for research and testing.

8.2 Textpresso

Textpresso is an information extracting and processing package for *C. elegans* literature. Textpresso is a web-based application that runs off a 2800MHz Intel Xeon Pentium dual processor Linux server with about 3.0 Gb memory. Textpresso is a search tool to help retrieve and process information from a corpus of *C. elegans* abstracts and papers. Searches can be performed by entering keywords into the Textpresso search field, much like popular search engines such as Google and PubMed. Additionally, searches can also be performed by selecting classes from the Textpresso Ontology.

Advanced Retrieval

In the advanced retrieval, searches of categories or keywords at any level of sophistication can be undertaken. Specify boolean operators, attributes, number of occurrences, etc.. In a keyword row, specify only one keyword per row. **This interface requires javascript. It does not work well with Netscape 4.79 or earlier versions.**

QUERY					
boolean operation	category or keyword	category or match attributes	specification	numerical comparison of matches	number of matches
required field	none	Please select a category or keyword	yes	greater than	0
and	none	Please select a category or keyword	yes	greater than	0
and	none	Please select a category or keyword	yes	greater than	0
and		Exact match		greater than	0
and		Exact match		greater than	0
and		Exact match		greater than	0

These criteria should be met in a sentence publication and searched in Abstract Full Text Title

Search

Reproduce query table with category row(s) and keyword row(s).

Again, this tool only provides boolean queries from different concepts and integrates the Gene Ontology with the abstracts and papers related to *C. elegans*. It does not provide additional Natural Language Processing or Text Mining capabilities.

8.3 BioRAT

BioRAT (Biological Research Assistant) is a stand alone java application that queries PubMed to download research articles and Google to download pdf documents and does text processing on these text files to highlight concepts using GO and NLP methods. It uses GATE (General Architecture for Text Engineering) which is a general purpose text engineering system based on NLP developed at Sheffield University. It takes 3-5 minutes to analyse each full length paper on a standard desktop pc with ~500 MB RAM and 1.7 GHz processor.

8.4 MedScan

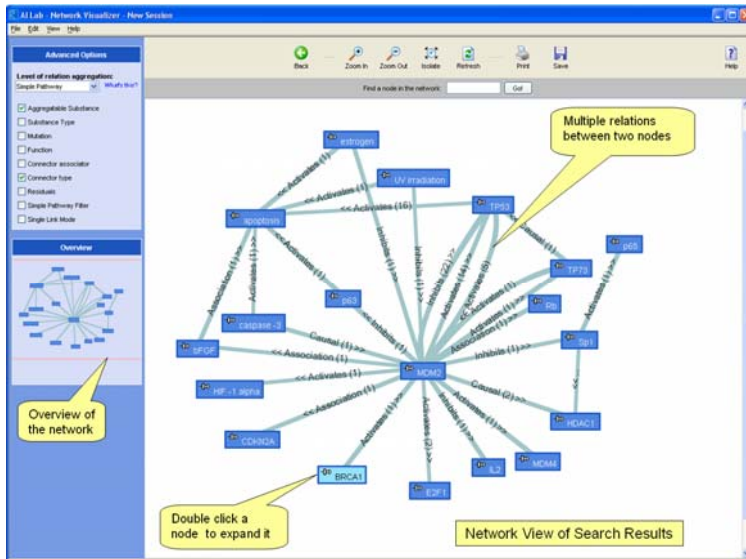
MedScan is a commercial tool that boasts of using specialized NLP engine to process the extracted text. It parses the full sentences and does not rely on just shallow parsers that are used in most current systems. These shallow parsers identify a subset of phrasal components (Nouns and verbs, for example) and are not able to reconstruct the structure of the entire sentence. They argue that these shallow parsers are only able to capture simple binary relationships and yield erroneous results for sentences containing complex relationships. This system uses the UMLS components to aid in extracting information.

8.5 BioMap

BioMap system, developed at Purdue university (not available on the internet) is again a text processing and NLP system that is built on Oracle 9i DBMS platform that is being built towards the goal of knowledge base creation rather than information extraction. It uses off the shelf components like Brill Tagger (for part of speech tagging) and machine learning tools like Hidden Markov Models for multi object tagging. It integrates Medline, UMLS and LocusLink for creating this knowledgebase.

8.6 GeneScene

Genescene is an impressive visualization tool built on Java technology that combines relations between entities in text extracted by a rule based parser and a corpus based co-occurrence analysis technique. Three ontologies, GO, UMLS and Human Genome Nomenclature (HUGO) are used to better integrate the relations.



8.7 Pubsearch

PubSearch is a literature curation management system designed to store and manage the available literature for an organism or system of interest. It provides database curators with a powerful literature search capability, stores relevant biological information, creates automatic associations between the biological information and the literature, and provides a user-friendly web interface for manual validation and curation. PubSearch is based on a simple MySQL relational database for the back-end, and Java Servlet and Java Server Pages for the API and front-end applications.

Pubsearch Test environment

[Search for Articles / Genes / Hits / Terms / All] [Add Article / Articles in Bulk / Gene / Term]
 [Find Help / Curation Guide] [Browse Gene Ontology / Tax Ontology]
 [User demo Logout] [Display all Annotation Issues] [Submit SourceForge Bug Report]

Search for Hits

This search allows you search hits between articles and terms

Filter based on validation status

- Retrieve hits marked as "valid"
- Retrieve hits that haven't been looked at
- Retrieve hits marked as "maybe valid"
- Retrieve hits marked as "invalid"
- Retrieve with any validation status

Output format

- List Hits Individually
- List Hits Grouped by Article

Terms	Articles
Term name <input type="text"/> Exactly <input type="text"/> External ID <input type="text"/> Exactly <input type="text"/> Filter based on term type <input type="text" value="only allow 'gene'"/> Filter based on obsolescence <input checked="" type="radio"/> Retrieve only non-obsolete terms <input type="radio"/> Retrieve only obsolete terms <input type="radio"/> Retrieve both obsolete and non-obsolete terms	Title <input type="text"/> Contains <input type="text"/> Journal <input type="text"/> Contains <input type="text"/> Let the Year span From <input type="text" value="Any"/> To <input type="text" value="Present"/> Volume <input type="text"/> Issue <input type="text"/> Page Start <input type="text"/> Restrict publication types to <input type="text" value="Any"/> Restrict article types to <input type="text" value="Any"/> Filter based on obsolescence <input checked="" type="radio"/> Retrieve only non-obsolete articles <input type="radio"/> Retrieve only obsolete articles <input type="radio"/> Retrieve both obsolete and non-obsolete articles Filter based on PDF availability <input checked="" type="radio"/> Retrieve articles with and without PDF links <input type="radio"/> Retrieve only articles with PDF links <input type="radio"/> Retrieve only articles without PDF links

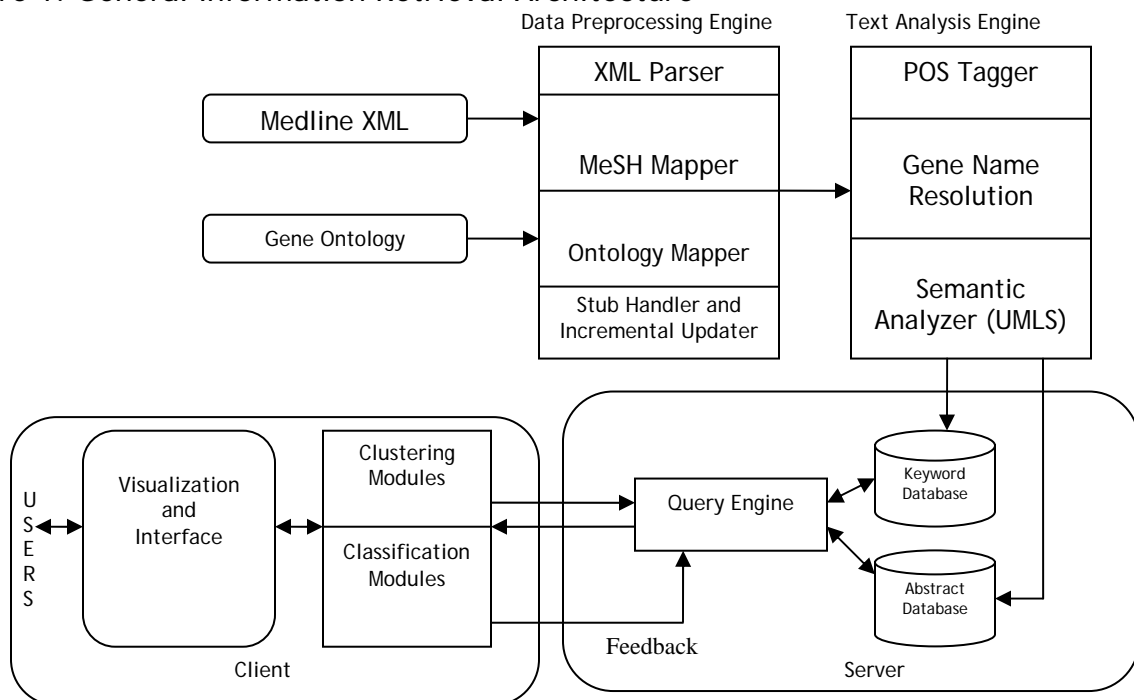
9. Proposed System

There is a need for a single unified system that integrates all the components of biomedical text mining from database development to clustering and classification to effective visualization of results to help the biologists infer knowledge from it. Such a system would give flexibility for the researcher to see the results in various dimensions.

9.1 Defining the constituent components and framework for the system.

The generic information retrieval framework for Text Mining involves management and processing of unstructured information to build databases that present the digested information in a structured form so that it can be mined, searched and visualized better and gives a richer understanding of the original unstructured content. Since we have most of the information about every published gene, their comparison and results and various discussions in literature, we can make use of this domain-specific information to build more efficient information retrieval system tailored for this purpose. Moreover, there is a huge collection of relationships and ontological information about these entities (for ex., UMLS, GO, COG, MeSH, GeneRIF, etc.) in separate databases that can be used to bridge the gap between discrete single database searches and the available information to build knowledge systems for biomedical information retrieval.

Figure 1: General Information Retrieval Architecture



The database and related components is organized as (shown in Figure 1) an inverted index on all the gene names and contain functional keywords along with their relevance weights in this proposed architecture. Another inverted index list on important keywords is proposed for efficient retrieval of gene names associated with this list. This inverted index needs to be updated regularly to provide most current and enhanced information on gene-keyword and keyword-gene relationships. This persistent storage of functional keyword association would help the researcher immediately search the entire Medline and retrieve top keywords from the database. Tf-idf or its variant weighting scheme called Okapi weighting should be incorporated as the preferred keyword weighting scheme. Okapi weighting technique is the state of the art scoring method used extensively in Information Retrieval community. This probabilistic model weighting method relies heavily on better estimation of various probabilities. It uses dampened frequency (eg., logarithmic tf function) and is dependent on document lengths and some other parameters to be a more effective tf-idf weighting scheme.

We would be able to build hierarchical gene and keyword networks once our indices are built to explore the cross linkages between keywords associated with various genes and other genes associated with these keywords along with the annotated abstracts information. This kind of visualization will aid the researcher to understand the complex network and infer useful information from it.

The proposed architecture for this system can be divided into two components - a one time static processing component to populate our functional keyword database and annotated abstract database and an interactive runtime component for querying this database to identify gene clusters along with other classes of attributes and relationships as well as add annotations to the meta-data field in the abstract database. The ontology mapper will give consistent description to the gene products identified in the abstracts and add this information in the database. In order to extend the database in future with richer sources of information, we will provide for incremental update facility and a stub handler to provide for handling future integration with the extension of our database. This would correspond to integrating new batch updates of Medline database in our system. The XML parser would parse around 45 GB of Medline data (as of today) to load into the relational database using a SAX (Simple API to XML) parser. (I have used a tool called BioTextEngine for parsing some Medline XML files into a local PostgreSQL database that uses SAX parser) These huge XML files do not render themselves well for processing and searching and need proper indexing and searching utilities that is provided by relational systems.

The text analysis engine has a lot of scope to add semantic associations using NLP and other entity extraction methods based on machine learning techniques. A simple Part-of-speech tagger would identify the noun, verb and other phrases and would help understand the linguistic information associated with abstracts. For example, we would like to screen out words like 'not a gene' instead of associating keyword information with such a gene. The Gene Name Resolution component would disambiguate the use of several alias names for one gene name in the abstracts. The Semantic Analyzer would use the Unified Medical Language System (UMLS) provided by NLM to infer various lexical and semantic relationships between the biomedical concepts extracted by our system. All this rich information would be loaded one-time into the database system containing the abstract information. In order to integrate this rich information resource with the Medline database, we need 100s of GB of storage space for UMLS and a more than 2 GB RAM for efficient processing.

Relevance feedback technique is based on the principle that it is easy for users to judge documents as being relevant or non-relevant for their search query. A system can then automatically generate a better query using such relevance judgments for further searching. Relevance feedback has been shown to work quite effectively across test collections. We will be using this technique in our Enhanced Medline search interface to be able to refine the results given by the system based on user preference.

Researchers are concerned with a specific set of questions about entities that are extracted from experimental observations and are predictive in nature. These questions conform to predictable language patterns and most questions do not involve much open-ended reasoning. We can model our databases specifically to answer these types of questions by using simple predictive annotation tags like GENE, PROTEIN, DRUG, DISEASE etc in our information representation. We will provide authorized users to add meta-data tags in the abstract database that will help identify special properties of these abstracts and associate them with other relevant information. Automating this process of annotation is again a huge challenge that involves application of many machine learning and information retrieval principles.

The purpose of this annotated abstract database will be to create an enriched search engine. This system would extract abstracts, functional keywords along with other relevant features of the abstracts based on input information which could be the disease name, function name or the gene name, for example. This system would learn continually with user feedback and weight adjustment to be a rich source of information.

Biologists currently waste a lot of time and effort in searching for all of the available information about each small area of research. This is hampered further by the wide variations in terminology that may be common usage at any given time, and that inhibits effective searching by computers as well as

people. Some of the Medline abstracts have information regarding their Medical Subject headings and the qualifier name. We can also use this information to our advantage in describing the abstract information and the associated keywords.

10. Exploring DBMS indexing techniques to speed-up abstracts search.

Medline contains over 15 Million citations and to retrieve documents in an ad-hoc manner from databases that we build requires efficient indexing to speed-up the process of document retrieval. Searching over Medline abstracts that are generally stored as character large objects takes considerable time (as tested in my initial experiments), hence we need to explore this area deeper to do efficient search and retrieval. For example, we can organize the abstracts into a set of well-defined concepts by using feature extraction methods (stop word elimination, statistical keyword relevance) from these abstracts.

In order to achieve performance speed-ups over inherent slow full text abstract searches (using the LIKE operator to match keywords in abstracts), we need to perform full text indexing of our Medline abstracts. Full text indexing is not a inbuilt feature in PostgreSQL (as opposed to DB2 or Oracle). I am exploring tsearch2 package (an extension in PostgreSQL for full text indexing) for incorporating this feature in my local Medline and gene ontology database. These modules can be integrated with PostgreSQL. These modules have options to specify dictionaries for spell check, synonym dictionary, stop word recognition and even weighting and ranking of keywords. If we can integrate all these features straight into the DBMS, it would become very efficient to carry out our information retrieval task getting all the advantages of the DBMS system.

11. Discussion

The application of information science and technology to computational biology has mainly focused on three primary areas: database development for managing diverse information related to biology, algorithms development for biological data analysis and software applications for accessing data over the internet. There is a lot of scope for further development in all these active areas of research for biomedical information retrieval that broadly incorporates indexing, querying, comparison and feedback. With the information deluge we are facing in biomedical science, it is a challenge for computer scientists to scale up to meet the requirements of this field.

There are several open issues in this domain that come to my mind:

- (i) Scalability of algorithms and databases
- (ii) mapping and interface of different data sources

- (iii) visual text mining to enhance productivity
- (iv) integration of different biomedical databases
- (v) bag of words approach to Natural Language Processing
- (vi) new measures of user effectiveness of system
- (vii) Full-text search and efficient indexing
- (viii) automatic updating of different ontologies
- (ix) borrowing ideas from bibliometric methods and digital libraries
- (x) Summarizing and Question Answering

Some of the above points have not been discussed in this report. Research in the above domains will further lead to leverage the knowledge discovery process effectively.

References

Alexander S. Yeh, Lynette Hirschman and Alexander A. Morgan. Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup

Allen C. Browne Guy Divita Chris Lu Lynn McCreedy Destinee Nace Lexical Systems; A report to the Board of Scientific Counselors September 2003, NLM

Amit Singhal, Modern Information Retrieval: A brief overview, Google, Inc.

Barry R. Zeeberg, Weimin Feng, Geoffrey Wang, May D. Wang, Anthony T. Fojo, Margot Sunshine, Sudarshan Narasimhan, David W. Kane, William C. Reinhold, Samir Lababidi, Kimberly J. Bussey, Joseph Riss, J. Carl Barrett, and John N. Weinstein. GoMiner: A Resource for Biological Interpretation of Genomic and Proteomic Data. *Genome Biology*, 2003 4(4):R28

Berry de Bruijn and Joel Martin. Finding Gene Function using LitMiner, TREC 2003 Report

Berry de Bruijn and Joel Martin. Getting to the (c)ore of knowledge: mining biomedical literature

Brigitte Mathiak and Silke Eckstein. Five Steps to text mining in Biomedical Literature

David P. A. Corney, Bernard F. Buxton, William B. Langdon and David T. Jones. BioRAT: extracting biological information from full-length papers

Eric W. Brown, Andrew Dolbey, Lawrence Hunter IBM Research and the University of Colorado TREC 2003 Genomics Track

G Bhalotia, PI Nakov, AS Schwartz, MA Hearst BioText Team Report for the TREC 2003 Genomics Track, TREC 2003 Report

Gondy Leroy et al. GeneScene: Biomedical Text and Data Mining

Hans-Michael Müller, Eimear E. Kenny, Paul W. Sternberg Textpresso: An Ontology-Based Information Retrieval and Extraction System for Biological Literature, *PLOS Biology* Volume 2, Issue 11, NOVEMBER 2004

Jonathan D. Wren. The emerging in-silico scientist: how text-based bioinformatics is bridging biology and artificial intelligence

Kamal Kumar, Mathew J Palakal, Snehasis Mukhopadhyay, Mathew J Stephens, Huian Li, BioMap: Toward the Development of a Knowledge Base of Biomedical Literature, 2004 ACM symposium on applied computing.

K. N. Lewis, M. D. Robinson, T. R. Hughes, C. W. V. Hogue MyMED: A database system for biomedical research on MEDLINE data, IBM SYSTEMS JOURNAL, VOL 43, NO 4, 2004

L. Tanabe, U. Scherf, L. H. Smith, J. K. Lee, L. Hunter and J. N. Weinstein, MedMiner: an Internet Text-Mining Tool for Biomedical Information, with Application to Gene Expression Profiling, *BioTechniques* 27:1210-1217

L. Venkata Subramaniam, Sougata Mukherjea, Pankaj Kankar, Biplav Srivastava, Vishal S. Batra, Pasumarti V. Kamesam, Ravi Kothari. Information Extraction from Biomedical Literature: Methodology, Evaluation and an Application

M. J. Schuemie et al. Distribution of information in biomedical abstracts and full text publications

Mathew Palakal, Snehasis Mukhopadhyay, Javed Mostafa, Rajeev Raje, Mathias N'Cho and Santosh Mishra. An intelligent biological information management system

Robert Baud, Patrick Ruch. The future of Natural Language processing for biomedical applications

Robert Mack and Michael Hehenberger. Text-based knowledge discovery: search and mining of life-sciences documents

Salton, G. (1971) The SMART retrieval system-experiments in automatic document processing, Prentice-Hall, NJ.

Salton, G., Wong, A., and Yang, C. S. (1975) A vector space model for automatic indexing. Communications of the ACM, 18: 613-620.

Sudeshna Adak, Vishal S Batra, Deo N Bhardwaj, P V Kamesam, Pankaj Kankar, Manish P Kurhekar, Biplav Srivastava A System for Knowledge Management in Bioinformatics, IBM India Tech. Report

Svetlana Novichkova, Sergei Egorov and Nikolai Daraselia. Medscan, a natural language processing engine for Medline abstracts