Full Report

ibc Interdisciplinary Bio Central

Open Access, Open Review Journal

PubMine: An Ontology-Based Text Mining System for Deducing Relationships among Biological Entities

Tae-Kyung Kim[1], Jeong-Su Oh[1], Gun Hwan Ko[1], Wan-Sup Cho[2], Bo Kyeng Hou[1,*] and Sanghyuk Lee[1,3,*]

[1]Korean Bioinformation Center (KOBIC), Korea Research Institute of Bioscience (KRIBB) & Biotechnology, 111 Gwahangno, Yuseong-gu, Daejeon 305-806, Republic of Korea
[2]Department of Management Information System, Chungbuk National University, Cheongju, Republic of Korea
[3]Division of Life and Pharmaceutical Sciences, Ewha Womans University, 11-1 Daehyun-dong, Seodaemun-gu, Seoul 120-750, Republic of Korea

Subject areas:
Bioinformatics/Computational biology/Molecular modeling

Author contribution: T-K.K. designed the algorithm and wrote the manuscript; J-S.O. and G.H.K. carried out the programming; W-S.C. analyzed the result; B.K.H. and S.L. directed the entire study; all authors read and approved the final manuscript.

***Correspondence** and requests for materials should be addressed to B.K.H. (bkher71@kribb.re.kr) and S.L. (sanghyuk@kribb.re.kr).

Editor: Sun Shim Choi, Kangwon National University, Republic of Korea

Received April 08, 2011;
Accepted April 12, 2011;
Published April 25, 2011

Citation: Kim, T-K., et al. An Ontology-Based Text Mining System for Deducing Relationships among Biological Entities. IBC 2011, 3:7, 1-6. doi: 10.4051/ibc.2011.3.2.0007

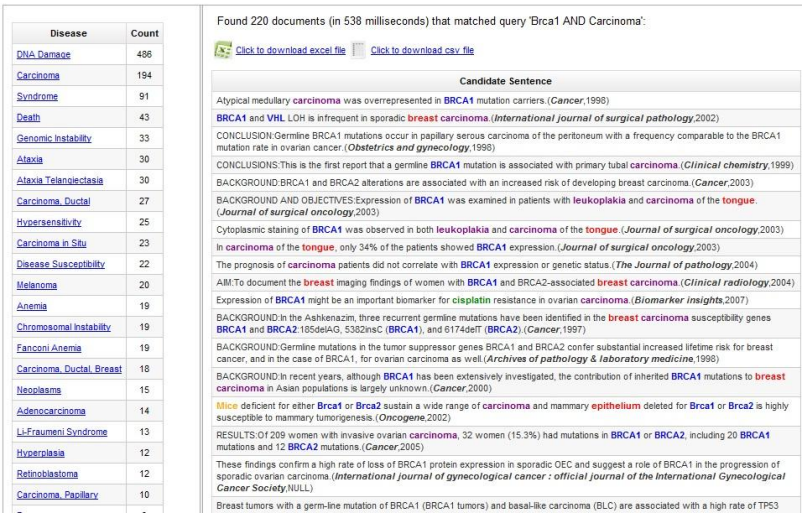
Funding: This work was supported by grants from "KRIBB Research Initiative Program" and "GIST Systems Biology Infrastructure Establishment Grant (2010)" through Ewha Research Center for Systems Biology (ERCSB).

Competing interest: All authors declare no financial or personal conflict that could inappropriately bias their experiments or writing.

This article is part of the special issue : the 9th Asia Pacific Bioinformatics Conference (APBC2011).

SYNOPSIS

Background: Published manuscripts are the main source of biological knowledge. Since the manual examination is almost impossible due to the huge volume of literature data (approximately 19 million abstracts in PubMed), intelligent text mining systems are of great utility for knowledge discovery. However, most of current text mining tools have limited applicability because of i) providing abstract-based search rather than sentence-based search, ii) improper use or lack of ontology terms, iii) the design to be used for specific subjects, or iv) slow response time that hampers web services and real time applications.
Results: We introduce an advanced text mining system called PubMine that supports intelligent knowledge discovery based on diverse bio-ontologies. PubMine improves query accuracy and flexibility with advanced search capabilities of fuzzy search, wildcard search, proximity search, range search, and the Boolean combinations. Furthermore, PubMine allows users to extract multi-dimensional relationships between genes, diseases, and chemical compounds by using OLAP (On-Line Analytical Processing) techniques. The HUGO gene symbols and the MeSH ontology for diseases, chemical compounds, and anatomy have been included in the current version of PubMine, which is freely available at http://pubmine.kobic.re.kr.
Conclusions: PubMine is a unique bio-text mining system that provides flexible searches and analysis of biological entity relationships. We believe that PubMine would serve as a key bioinformatics utility due to its rapid response to enable web services for community and to the flexibility to accommodate general ontology.



Disease	Count
DNA Damage	486
Carcinoma	194
Syndrome	91
Death	43
Genomic Instability	33
Ataxia	30
Ataxia Telangiectasia	30
Carcinoma, Ductal	27
Hypersensitivity	25
Carcinoma in Situ	23
Disease Susceptibility	22
Melanoma	20
Anemia	19
Chromosomal Instability	19
Fanconi Anemia	19
Carcinoma, Ductal, Breast	18
Neoplasms	15
Adenocarcinoma	14
Li-Fraumeni Syndrome	13
Hyperplasia	12
Retinoblastoma	12
Carcinoma, Papillary	10

Found 220 documents (in 538 milliseconds) that matched query 'Brca1 AND Carcinoma':

Click to download excel file Click to download csv file

Candidate Sentence
Atypical medullary **carcinoma** was overrepresented in **BRCA1** mutation carriers.(***Cancer***,1998)
BRCA1 and **VHL** LOH is infrequent in sporadic **breast carcinoma**.(***International journal of surgical pathology***,2002)
CONCLUSION:Germline BRCA1 mutations occur in papillary serous carcinoma of the peritoneum with a frequency comparable to the BRCA1 mutation rate in ovarian cancer.(***Obstetrics and gynecology***,1998)
CONCLUSIONS:This is the first report that a germline **BRCA1** mutation is associated with primary tubal **carcinoma**.(***Clinical chemistry***,1999)
BACKGROUND:BRCA1 and BRCA2 alterations are associated with an increased risk of developing breast carcinoma.(***Cancer***,2003)
BACKGROUND AND OBJECTIVES:Expression of **BRCA1** was examined in patients with **leukoplakia** and **carcinoma** of the **tongue**.(***Journal of surgical oncology***,2003)
Cytoplasmic staining of **BRCA1** was observed in both **leukoplakia** and **carcinoma** of the **tongue**.(***Journal of surgical oncology***,2003)
In **carcinoma** of the **tongue**, only 34% of the patients showed **BRCA1** expression.(***Journal of surgical oncology***,2003)
The prognosis of **carcinoma** patients did not correlate with **BRCA1** expression or genetic status.(***The Journal of pathology***,2004)
AIM:To document the **breast** imaging findings of women with **BRCA1** and BRCA2-associated **breast carcinoma**.(***Clinical radiology***,2004)
Expression of **BRCA1** might be an important biomarker for **cisplatin** resistance in ovarian **carcinoma**.(***Biomarker insights***,2007)
BACKGROUND:In the Ashkenazim, three recurrent germline mutations have been identified in the **breast carcinoma** susceptibility genes **BRCA1** and **BRCA2**:185delAG, 5382insC (**BRCA1**), and 6174delT (**BRCA2**).(***Cancer***,1997)
BACKGROUND:Germline mutations in the tumor suppressor genes BRCA1 and BRCA2 confer substantial increased lifetime risk for breast cancer, and in the case of BRCA1, for ovarian carcinoma as well.(***Archives of pathology & laboratory medicine***,1998)
BACKGROUND:In recent years, although **BRCA1** has been extensively investigated, the contribution of inherited **BRCA1** mutations to **breast carcinoma** in Asian populations is largely unknown.(***Cancer***,2000)
Mice deficient for either **Brca1** or **Brca2** sustain a wide range of **carcinoma** and mammary **epithelium** deleted for **Brca1** or **Brca2** is highly susceptible to mammary tumorigenesis.(***Oncogene***,2002)
RESULTS:Of 209 women with invasive ovarian **carcinoma**, 32 women (15.3%) had mutations in **BRCA1** or **BRCA2**, including 20 **BRCA1** mutations and 12 **BRCA2** mutations.(***Cancer***,2005)
These findings confirm a high rate of loss of BRCA1 protein expression in sporadic OEC and suggest a role of BRCA1 in the progression of sporadic ovarian carcinoma.(***International journal of gynecological cancer : official journal of the International Gynecological Cancer Society***,NULL)
Breast tumors with a germ-line mutation of BRCA1 (BRCA1 tumors) and basal-like carcinoma (BLC) are associated with a high rate of TP53

Keywords: bio-text mining, web service, ontology, systems biology, bio-resource search

Introduction

With the deluge of literature data in bioinformatics, biologists have difficulty in finding knowledge or relevant information from the literature. The volume of literature data is increasing almost exponentially recently. There are over 3,000 biological journals, in which about a million papers are published every year. For example, 130 journals are found to publish manuscripts related to breast cancer at the rate of 27 papers a week. It is estimated that 70% of biologists' office hours is devoted to examining literatures[1]. Extensive survey of preliminary literatures is directly connected to their research productivity in terms of time and cost savings by avoiding research overlaps and by getting significant scientific hints.

Therefore, it is essential to take advantage of text mining tools to handle huge biological literature data strategically. Although various text mining tools such as Entrez[2], iHOP[3], MedMiner [4], PolySearch[1], FACTA[5], and GoPubMed[6] are being widely used, they were developed for specific aims, thus having limited ability in terms of advanced search and analysis.

NCBI's Entrez[2], the most widely used tool for retrieving biological literatures, supports simple and advanced searches. In spite of its powerful search for specific fields of literatures (e.g. title, author, journal, affiliation, date, etc.), it does not support elaborate sentence-based search that is helpful in reducing false-positives efficiently. iHOP (Information Hyperlinked over Proteins)[3] generates an interaction network built from co-citation of genes or proteins within a sentence. However, it is a major disadvantage to support gene or protein names only as a query type. PolySearch[1] and FACTA[5] retrieve relationships among biological entities of proteins, diseases, drugs, and organisms, based on sentence pattern recognition. PolySearch is based on heuristic weighting of differential co-occurrence to infer significant relationships between terms. Disadvantages of these programs may be the limited use of ontology terms, slow response time, late update, and a priori defined query types. GoPubMed[5] provides the state of the art web interfaces with the ontology support. But the search type is rather fixed with no web service capability.

In this paper, we propose an elaborate sentence-based text-mining system called PubMine, which supports flexible and diverse types of query, full structured ontologies, fast and easy-to-use interfaces. It also supports multi-dimensional analysis of relationships between genes, diseases and chemical compounds by using the OLAP (On-Line Analytic Processing) technique[7]. OLAP is an approach to swiftly answer multidimensional analytical queries, mainly used in business intelligence, which encompasses relational reporting and data mining. The multidimensional data model, an integral part of OLAP, is designed to solve complex queries in real time.

In the following sections, we describe the algorithmic details and implemented features of PubMine.

Results

Key features of PubMine

PubMine is a sentence-based text mining system that provides a fast and easy-to-use web application with flexible and diverse types of query. It also supports multi-dimensional analysis of relationships between genes, diseases and chemical compounds with full structured ontologies.

PubMine supports diverse types of powerful searches such as the fuzzy search, proximity search, wildcard search, range search, and the Boolean combinations as shown in Table 1. The fuzzy search finds similar terms based on Levenshtein distance algorithm and edit distance algorithm[8]. For example, search for BRCA1~ identifies all BRCA1-related terms such as BRCA-1, BRCA-I, BRCA1/a, BRCA1/b, etc. This expansion improves search coverage significantly by including ambiguous terms. The proximity search limits the distance (the number of words) between two terms. This is useful in removing false positives. Our experiment shows that 98% of the hits with distance of over 10 words are identified as false. The wildcard search finds derivatives or ambiguous words with an asterisk (*) for a multiple-character wildcard search and a question mark (?) for a single-character wildcard search. The range search can be applied on the alphanumeric value of date or *pmid* to limit the range of query results. Finally, biologists can make elaborate queries by combining these search functions in the Boolean logical operators such as 'AND', 'OR', 'NOT', '(', and ')'.

Table 1. Sentence-based query examples

Query Type	Example Query	Query Description
Fuzzy Search	*BRCA1~*	• Retrieve sentences containing words similar to BRCA1 *i.e.* BRCA1a, BRCA-1, BRCA Ⅰ ,BRCA- Ⅰ
Proximity Search	*"BRCA1 p53"~10*	• Retrieve sentences that word distance between BRCA1 and p53 in each sentence is lower than 10.
Range Search	*mod_date (20020101 to 20030101)*	• Retrieve sentences within specific ranges of publication date.
Wildcard Search	*interact**	• Retrieve sentences including *"interact"* derivatives. *i.e.* interact, interacts, interacted, interaction,
Logical Expression	*"BRCA1 p53"~10 AND (bind* OR interact* OR associate*)*	• Retrieve sentences, which contain two protein names, BRCA1 and p53, 10 of distances between the two, and derivatives of *bind*, *interact*, and *associate*.

Users can extract meaningful information in combination of fuzzy search, proximity search, range search, wildcard search, and logical expression operators.

Another major feature of PubMine is to analyze the multidimensional relationships among biological entities of genes, diseases and chemical compounds. To effectively retrieve relationships among biological entities, we use a collection of OLAP queries based on a star schema[7]. It is noteworthy that multidimensional analysis enables users to perform sophisticated analysis for the biological entities. For example, a typical search of disease-related genes is one dimensional analysis (i.e. X→Y style, where X is a disease and Y is a gene). Virtually all currently available tools including PolySearch, FACTA, and GoPubMed belong to this category. PubMine is unique to extend conventional search into multidimensional one. As an example of two dimensional search, we can find chemical compounds that are related to a specific disease and a gene simultaneously (i.e. X, Y → Z style, where X is a disease, Y is a gene, and Z is a chemical compound).

User interface

PubMine's web interface was developed in standard HTML and JSP. A screenshot of the query interface for a simple search is shown in Figure 1. In the simple query, users can choose the *'in sentence'* option for the sentence-based search, or the *'in abstract'* option for the abstract-based search as the query target. Many false positive results can be reduced by the sentence-based search because most of the meaningful relationships appear within a sentence (e.g. protein-protein, protein-drug, and gene-disease relationships).

The output for the simple query is shown in the lower part of

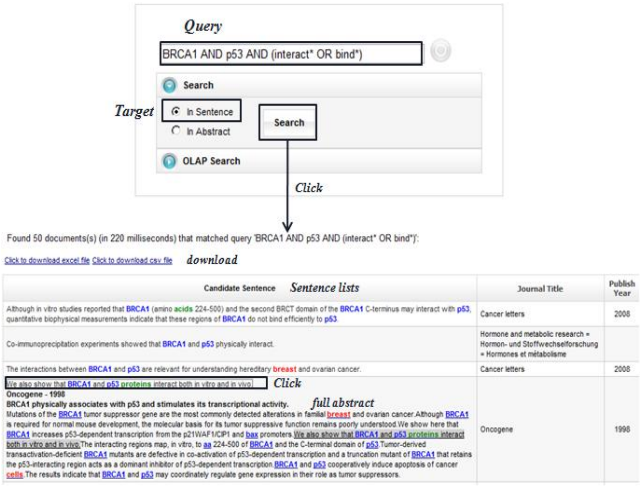


Figure 1. A screenshot of PubMine's query interface and the result view of a simple query. The simple search interface allows users to create their queries for both sentence and abstract targets. For example, a query, '*BRCA1 AND p53 AND (interact* OR bind*)*', can find interaction information between *BRCA1* and *p53* in sentence unit.

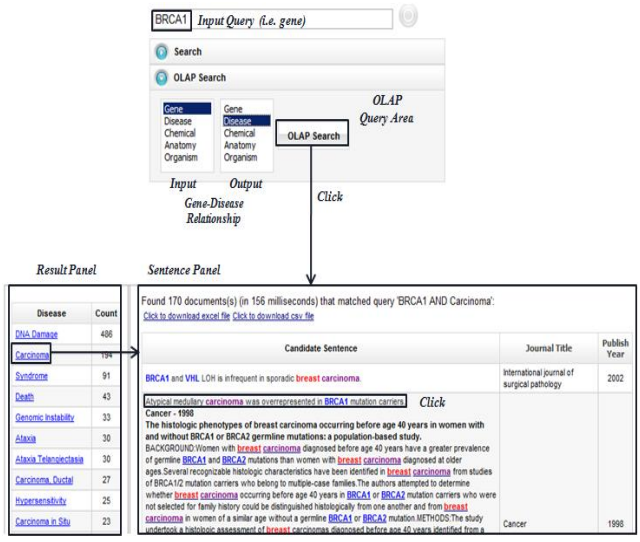


Figure 2. A screenshot of OLAP query and results. After entering a query (i.e. BRCA1 as a gene name) and clicking the 'OLAP Search' button, users can examine the summary of result in the *'Result panel'* (i.e. Disease List) and detailed information in the *'Sentence Panel'* (i.e. BRCA1-Carcinomas related sentences).

Figure 1. Initial output shows the list of target sentences that are clickable to show the full abstract. Additional information such as journal name, publication year, and impact factor is included as well. Note that terms belonging to the ontologies (anatomy, organisms, genes, diseases, and chemical compounds) appear in colored text to enhance the visibility and hyperlinked for user's convenience. Users can download the search results in text or Excel formats.

In OLAP query, users can find corresponding sentences to identify relationships between biological entities. Figure 2 shows the search for diseases related to the BRCA1 gene. Input query type is the gene name and users select the associated term category (disease in this example). The result shows the summary table of term occurrences and the retrieved sentences in a similar format of the previous result. We count the number of sentences where the given keyword appears together with any ontology terms in the selected category (e.g. MeSH C terms for disease). Clicking on the identified term leads users to examine the candidate sentences.

Discussion and Conclusion

Various programs are being used for text mining for biological literatures. In Table 2, we provide a broad qualitative comparison of PubMine with other text mining systems such as Entrez, MedMiner,

Table 2. Feature comparison of various bio-textmining systems[1]

	Entrez	MedMiner	iHOP	PolySearch	PubMine
Type of Search Supported	Literature, Disease, Gene, Structure, Taxonomy, SNP, Compound, Etc.	Gene, Drug, Text Word	Gene	Gene, Disease, Drug, Metabolite, Tissues/Organs, Subcellular, Localization, Text Word	Gene, Disease, Anotomy, Chemical Compounds, and elaborate user-defined search
Extensive hyperlinking	Most Extensive	Less Extensive	More Extensive	More Extensive	More Extensive
User-Defined Query Design	Yes, but limited on Abstract	Limited	Limited	Limited	Flexible
Query Coverage	Wide. But limited on Abstract	Narrow	Narrow	Narrow	Wide
Performance	Good	Medium	Good	Low	Good
Text and sentence highlighting	No	Yes	Yes	Yes	Yes
Co-occurrence scoring scheme	None	No	Sentence Level	Sentence Level	Sentence Level
Use of keywords for association words	None	Predefined keywords	Predefined keywords	Predefined & custom association words	Fully custom association words
Sentence pattern recognition	No	No	Yes	Yes	Yes
Thesaurus query synonym expansion	Yes, limited	Yes, limited	Yes, for genes only	Yes	Yes

iHOP, and PolySearch. Diverse aspects of programs are covered including 'type of search supported', 'extensive hyper-linking', 'response time', 'user-defined query design', 'ontology support', and so on. Each program has its own merits and disadvantages according to the design spirit, but the performance of PubMine is pretty decent in overall aspects. Particularly, it has an excellent grade for 'user-defined query design' to support flexible queries, 'query coverage' to cover general ontology terms of biological interest, 'response time' due to implementation of the star schema, and 'type of search supported' to support multidimensional analysis. Even though PubMine has similar characteristics with PolySearch, our program supports more flexible types of user-defined query and full structure of ontology terms in MeSH. Furthermore, PubMine's response time is usually within seconds, whereas PolySearch takes much longer time to retrieve relevant sentences that contain relationships between two biological entities except pre-calculated cases.

Even though PubMine is competitive with other text mining tools, it still has many aspects to improve both in terms of performance and user convenience. The website will be enhanced significantly to support the full ontology structures as in the GoPubMed and to implement several user-convenient features such as personal account to save queries and results. We plan to develop a curation-aid tool that allows users to examine and select meaningful hits rapidly. Furthermore, tools for statistical analysis will be included as well to increase intuitive understanding of query results.

We also plan to provide the web services based on SOAP (Simple Object Access Protocol) so that other programs can utilize PubMine using the web API. This would enhance the utility of our program significantly since no text mining program is available for web services. Rapid response of PubMine makes the web services possible.

PubMine is a unique and powerful text mining system for biological literatures with many merits such as flexible searches, ontology-based relationship analysis among biological entities, and fast response time with OLAP implementation. We believe that PubMine, available at http://pubmine.kobic.re.kr, would be a useful web application for both biologists and bioinformaticians.

Methods

Text mining usually consists of several steps: Text preparation and pre-processing, mining query evaluation and result verification.

Text preparation and pre-processing

This step is the most time consuming step in text mining and essential for efficient text retrieval and analysis since enormous amount of biomedical literatures are stored in unstructured format.

PubMine has four steps for the text preparation and pre-processing tasks: *abstract file creation, sentence tokenization, inverted index creation, and dimensional index creation.* Figure 3 shows these steps. We use the PubMed literatures as the source data, which includes over 19 million abstracts from MEDLINE and other life science journals[2].

In the abstract file creation step, we downloaded the abstract files from NCBI's PubMed to our local database periodically. After download, we removed any duplicate data based on the key field ***pmid*** and filtered out empty abstracts that do not include any meaningful contents. The amount of empty abstract was an approximately 44% (8,487,565) of the entire abstracts (19,347,443).

In the sentence tokenizing step, PubMine tokenizes each abstract into sentence units by utilizing the SPECIALIST NLP[12].

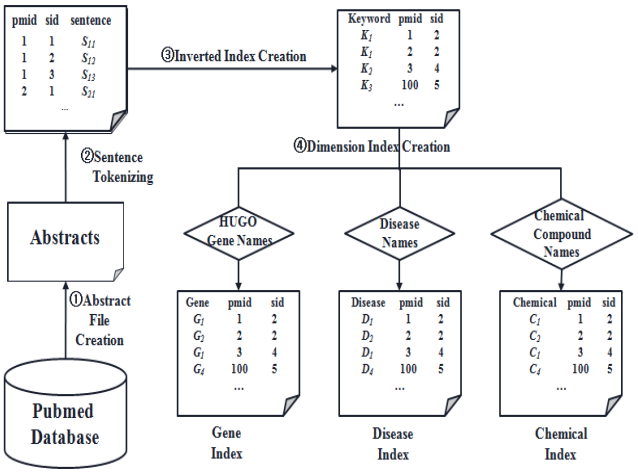


Figure 3. Text preparation and pre-processing procedure.

However, sentence detection is not quite reliable because the tool splits sentences by semicolon (";"), colon (":") and a period ("."). In fact, we found that 10% of all tokenized sentences were incomplete. As incomplete sentences cause a negative impact on a sentence-based search, we developed an incomplete sentence complement algorithm**.** After applying this algorithm, the percentage of incomplete sentence was under 1% and the number of the tokenized sentences was reduced from 92817421 to 82385179.

In the inverted index creation step, PubMine builds an inverted index for the tokenized sentences to allow fast full-text searches. The inverted index maintains mapping information for the document references of keywords (usually noun, verb, adjective and adverb) in each sentence, excluding stop words like prepositions and conjunctions[13]. PubMine utilizes the inverted index for elaborate sentence searches, which include wildcard search, fuzzy search, proximity search, range search, and logical expression search. Furthermore, we use the inverted index for fast dimensional index creation in the next step.

In the dimension index creation step, we build dimensional indexes for biological entities such as genes, diseases, and chemical compounds by using the MeSH ontology[14] and the HUGO gene symbols[15]. Table 3 shows the summary statistics of biological ontology terms. The OLS (Ontology Lookup Service) will be added to cover more biological concepts in the future. To create each dimensional index, it is critical to remove the garbage terms like ALL, UP, DO, OUT, etc. as shown in Table 4. These terms have a negative impact on detecting semantic relationships between biological entities. In addition, they cause a performance delay due to the many occurrences. These terms were removed by manual inspection after ordering all terms according to frequencies. As a result, we found that these terms (0.03% of all terms) occupy 30% of index size and that over 99% of these terms are not associated

Table 3. Summary statistics of ontology terms for biological entities

Category	Unique Entities	Synonyms	Source
Anatomy	1615	9925	MeSH (A)
Organisms	3570	19396	MeSH (B)
Diseases	4408	39610	MeSH (C)
Chemical compounds and Drugs	8815	72645	MeSH (D)
Genes	28733	106735	HUGO

PubMine uses the MeSH terms for anatomy, organisms, diseases, chemical compounds and drugs; and HGNC gene symbols and synonyms for gene names.

Table 4. Partial list of garbage terms

Division	Terms
Genes	CELL, SIMPLE, ALL, T, CAN, MICE, FAT, HAS, NE, TYPE, PH, TH, END, CT, II, FACT, FIND, LIGHT, HE, DAMAGE, LARGE, PER, LI, PR, Men, CD, UP, OUT, DO, IV, II, gamma, GAMMA, SEX, GO, STEP, ODD, DELTA, SHE, ER, IMPACT, CAT, MED, beta, BETA, AM, KILLER, MINOR, NUDE, SET, BEST, RED, FISH, GREAT, MASS, RANK, STOP, CAR, REST, STEP, CI, NM, SE, IP, lobe, CUT
Chemical compounds	DNA, RNA, Proteins, Water, Lead, Elements, Acids
Anatomy	CELLS, Brain, Blood, Serum, Heart, Liver, Plasma
Diseases	Disease, Infection, Syndrome, Pain, Death, Body Weight

These terms are not considered in building dimensional indexes. They are usually common words, not biological keywords.

to any biological meaning.

Mining query evaluation and result verification

Data model and query for multidimensional analysis among biological entities

PubMine uses star schema as shown in Figure 4 for multidimensional analysis among biological entities. A star schema consists of a fact table and multiple dimension tables that are connected by foreign keys. In this study, the star schema consists of a fact table (*Sentence*) and four dimension tables (*ChemicalIndex, GeneIdx, KeywordIdx* and *DiseaseIdx*). In addition to the star schema, a source table (Abstract) and four dictionary tables (*Keyword, Chemical, Disease* and *Gene*) are maintained in the database.

In PubMine, various relationships among biological entities can be analyzed by means of OLAP queries, which support sophisticated multi-dimensional analysis. Table 5 shows examples of OLAP queries that retrieve gene-disease relationships and gene-chemical relationships.

There is obvious contrast between PubMine and conventional mining tools in the data analysis. PubMine can perform various

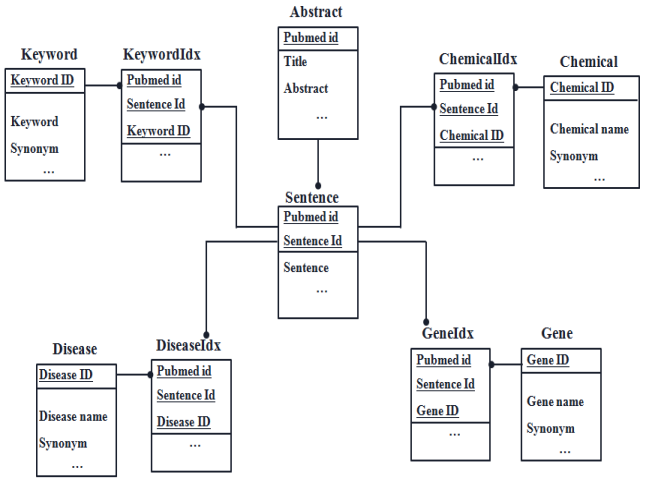


Figure 4. Data model of star schema for multidimensional analysis. The *Abstract* table (source table) contains the full abstracts of PubMed. The *Sentence* table (fact table) keeps sentence information from the sentence tokenizing step, and four dimension tables (*GeneIdx*, *DiseaseIdx, ChemicalIdx*, and *KeywordIdx*) maintain index information for gene, diseases, chemical compounds, and terms in sentences. Four dictionary tables (*Gene*, *Disease*, *Chemical,* and *Keyword*) include ontology information such as standard keywords, synonyms and annotation information.

Table 5. Relationship analysis between biological entities using star-join SQL queries

(A) gene-disease relationship query

```
SELECT term, COUNT(*)
FROM ( SELECT DI.pmid, DI.sid, DI._mid
        FROM DiseaseIndex DI, (SELECT pmid, sid
                               FROM GeneIndex
                               WHERE gname='p53'
                               LIMIT 50000) GI
        WHERE DI.pmid = GI.pmid AND
              DI.sid = GI.sid
      ) DD, mesh_term MT
WHERE DD._mid = MT.no
GROUP BY term
ORDER BY COUNT(*) DESC;
```

(B) gene-chemical relationship query

```
SELECT MM.term, COUNT(*)
FROM ( SELECT DISTINCT pmid,sid,gname
        FROM GeneIndex
        WHERE gname='p53'
        LIMIT 50000) AA,
        ChemicalIndex CC,
        mesh_term MM
WHERE AA.pmid=CC.pmid AND
        AA.sid = CC.sid AND
        MM.no=CC.cid
GROUP BY MM.term
ORDER BY COUNT(*) DESC;
```

(A) gene-disease relationship query retrieves diseases related to a gene '*p53*';
(B) gene-chemical relationship query shows chemical compounds related to a gene '*p53*'.

relationship analyses among biological entities just by writing SQL queries. However, conventional tools such as PolySearch and FACTA need program development or data management because their data models define queries in a static manner. A key problem in PubMine is the performance when large number of records is in the fact table or dimension tables. In fact, the numbers of records in the fact table *Sentence* and the index table *GeneIdx* are about 80 million and 40 million, respectively. To solve performance problem, we optimize the database parameters to reduce I/O cost, maximize cache hits, and build SQL queries carefully to utilize indexes, and process sub-range efficiently. Experiments show that response time for most queries has been reduced to less than 3~5 seconds by these optimizations.

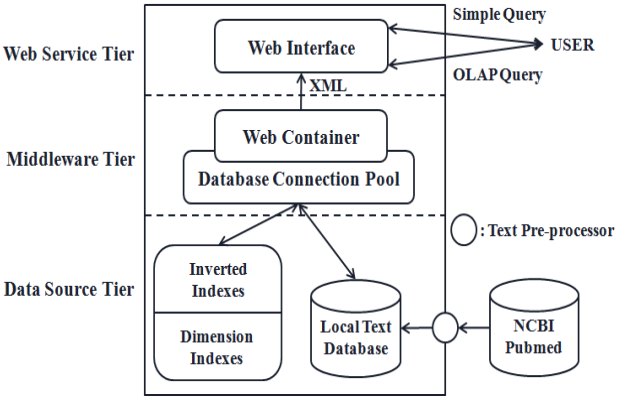


Figure 5. PubMine architecture. PubMine consists of three tiers: a data source tier, a middleware tier and a web service tier. The data source tier manages original text data and related indexes (dimension indexes and inverted index). The middleware tier provides an application server for java servlets and database pooling. Finally, the web service tier provides a user-friendly web interface including an easy query window for simple and OLAP searches.

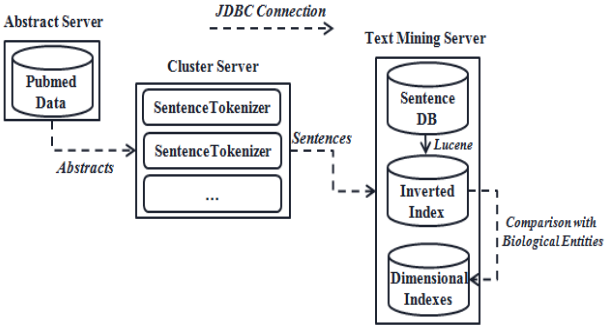


Figure 6. Data processing procedure for sentence tokenizing and dimensional indexing. The procedure consists of three steps: the first step is to get abstracts from the local PubMed database; the second step is to tokenize and deliver the sentences to the text mining server; the third step is to make dimensional indexes for biological entities. As the data delivery is performed only on network, additional works in each node of the cluster system are unnecessary.

System architecture

PubMine has three tiers architecture (*data source, middleware, and web service*) as shown in Figure 5. It provides web-based user-friendly interface. The data source tier manages the ontology terms, PubMed abstracts, and indexes. The local text database contains a sentence table and several dictionary tables. The sentence table contains the pre-processed sentences extracted from PubMed abstracts. PubMine uses a MySQL DBMS for the local text database and the dimension indexes, and the Lucene indexer[9] for the inverted indexes. The middleware tier is composed of an Apache web server, Tomcat as a web container, and a database connection pool. The database connection pool includes buffer caches to reduce the connection overhead. The web service tier provides users with an easy-to-use web interface for querying and reporting. The web interface uses Ajax[10] and ICEfaces[11] framework for asynchronous communications and easy maintenance.

Due to the large quantity of documents, the sentence tokenizing step and dimensional index creation step require severe time complexity. PubMine has a fast and automated data processing flow for these steps, based on a cluster system and remote data connections using JDBC (Java Database Connectivity) as shown in Figure 6. At least 40 cluster nodes are used due to the limited number of connections in JDBC. Our data processing flow enables PubMine to deal with the latest PubMed data effectively without any manual intervention.

Acknowledgements

This work was supported by grants from "KRIBB Research Initiative Program" and "GIST Systems Biology Infrastructure Establishment Grant (2010)" through Ewha Research Center for Systems Biology (ERCSB).

References

1. Tanabe, L., Scherf, U., Smith, L.H., Lee, J.K., Hunter, L., and Weinstein, J.N. (1999). MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. *Biotechniques* 27, 1210-1214, 1216-1217.
2. Cheng , D., Knox, C., Young, N., Stothard, P., and Damaraju, S. (2008). PolySearch: a webbased text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Res* 36, 399-405.
3. Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Helmberg, W., et al. (2005). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 33, D39-45.
4. Hoffmann, R., and Valencia, A. (2005). Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics* 21 Suppl 2, ii252-258.
5. Tsuruoka, Y., Tsujii, J., and Ananiadou, S. (2008). FACTA: a text search engine for finding associated biomedical concepts. *Bioinformatics* 24, 2559-2560.
6. Andreas, D.a.M., S. (2005). GoPubMed: expoloring PubMed with the Gene Ontology. *Nucleic Acids Res* 33, 1210-1217.
7. O'Brien, J. (2009). “Marakas G: Management Information Systems", 9th Edition (New York: McGraw-Hill Higher Education).
8. Navarro, G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys* 33, 31-88.
9. Lucene, http://lucene.apache.org/java/docs/index.html.
10. AJAX, http://www.w3schools.com/ajax/default.asp.
11. ICEfaces, http://www.icefaces.org/main/home/.
12. Specialist NLP Tool, http://lexsrv3.nlm.nih.gov/Specialist/Home/index.html.
13. Zobel: Inverted Files for Text Search Engines. ACM Computing Surveys 2006, 38 (2):6.
14. MeSH, http://www.nlm.nih.gov/mesh/meshhome.html.
15. HUGO Gene, http://www.genenames.org/.