Integrated Term Weighting, Visualization, and User Interface Development for Bioinformation Retrieval

Min Hong*, Anis Karimpour-Fard*, Steve Russell*, and Lawrence Hunter

Bioinformatics, University of Colorado Health Sciences Center,
4200 E. 9th Avenue Campus Box C-245, Denver, CO 80262, USA
{Min.Hong,Anis.Karimpour-Fard,Steve_Russell,Larry.Hunter}
@UCHSC.edu

Abstract. This project implements an integrated biological information website that classifies technical documents, learns about users' interests, and offers intuitive interactive visualization to navigate vast information spaces. The effective use of modern software engineering principles, system environments, and development approaches is demonstrated. Straightforward yet powerful document characterization strategies are illustrated, helpful visualization for effective knowledge transfer is shown, and current user interface methodologies are applied. A specific success of note is the collaboration of disparately skilled specialists to deliver a flexible integrated prototype in a rapid manner that meets user acceptance and performance goals. The domain chosen for the demonstration is breast cancer, using a corpus of abstracts from publications obtained online from Medline. The terms in the abstracts are extracted by word stemming and a stop list, and are encoded in vectors. A TF-IDF technique is implemented to calculate similarity scores between a set of documents and a query. Polysemy and synonyms are explicitly addressed. Groups of related and useful documents are identified using interactive visual displays such as a spiral graph that represents of the overall similarity of documents. K-means clustering of the similarities among a document set is used to display a 3-D relationship map. User identities are established and updated by observing the patterns of terms used in their queries, and from login site locations. Explicit considerations of changing user category profiles, site stakeholders, information modeling, and networked technologies are pointed out.

1 Introduction

The volume of biological and medical information has been growing rapidly, leading to increased interest and research in information retrieval and presentation, knowledge extraction, and the enhanced discovery of new ideas. The text indexing effort here examines terms used in a database of Medline papers to determine internal consistencies and patterns. Text ranking applies the frequencies of the constituent text terms in order to characterize documents. Different user types have varied interests in documents and in the distinctive terms in these documents. Similarities in users and documents are used here as indicators of the technical arenas, concepts, timeframes, and research directions with the highest value to the reader.

* Contributed equally.

T.G. Kim (Ed.): AIS 2004, LNAI 3397, pp. 673–682, 2005.

The frequencies of words in each document and in the user's query are used to derive a measure of information pertinence. Documents closest to a query which is also similar to the user's overall word preferences, are ranked and returned in order. The ratings based on key terms are also visualized in a spiral display. The care that is expended in biomedical information site implementation (as in commercial websites) is directly related to the ultimate success and utility of the site.

2 Methods

The objective here is to illustrate effective document characterization, helpful information visualization, and adaptive awareness to user categories and individual interests.

2.1 Text Indexing and Analysis

A vector space model is implemented to process document terms. Similarity of publications is represented using the term frequency vector for each document. The test domain here is Medline abstracts for articles on breast cancer. The word characterization method is based on stemming text terms, using a stop list to exclude common terms, and on creating a term-frequency vector for each document. Such vector space models are widely used based on TF-IDF (term frequency – inverse document frequency) rules for calculating similarity between documents. There has been growing research in biological text processing focusing on different areas such as clustering [11], categorization using predefined categories [22], detection of keywords [8, 16], and the detection and extraction of relations [4, 26, 27]. These advances are applied for the tasks here, using text terms and user profiling, and in the next phases it is anticipated that categorization of terms and direct query modifications will also be implemented.

There are many biological terms that have multiple synonyms, and there is a lack of uniformity in the set of terms used by researchers. One consequence is that identical queries from dissimilar users can result in their getting the identical set of returned documents, which is rarely the best set of publications for their separate needs. Documents and queries are represented as bags of terms and statistical values. Each document vector shows the presence or absence of a term, or the weight of each term within the document. The construction of the document-term vector space can be divided into three different stages: document indexing, weighting of index terms, and document ranking. In the document indexing stage, non-significant terms and words that do not describe context are removed. Word stemming reduces terms to a root form. Such feature reduction can improve efficiency as well as accuracy in document clustering and classification [28, 29].

Different weighting schemes can be used to discriminate one document from the other. Experimentally it has been shown that TF-IDF discrimination factors lead to more effective retrieval [25]. Terms that occur in only few documents are more valuable for characterization than the terms that repeatedly appear in many documents (IDF = Inverse document frequency). On the other hand, the more often that a term occurs within a certain document, the more likely it is that that term is important to that document (TF = term frequency). The formula that is used here is [18]:

$$IDF(t) = \log (N / N_t)$$

N = total number of documents in the set

N_t = number of documents in which the term t occurs

The weighting of the ith term in a document is given from its frequency f_i by:

$$W_i = \text{weight of term } i = f_i * IDF$$

One of the methods of adding concepts to augment the vector space approach is latent semantic indexing (LSI). This approach creates a matrix of terms that is analyzed by singular value decomposition (SVD) for the most different and predictive items. This is similar to clustering in that it solves the synonym problem but not polysemy [3, 7]. Other methods such as clustering or categorization of documents can be applied after documents vectors are created. Manual categorization is becoming impractical [14, 15] due to the volume of documents. So, clustering is most often accomplished using an unsupervised learning algorithm. Categorization generally achieves better results, using human supervision.

A simple approach to add conceptual information is the method used by the Semantic And Probabilistic Heuristic Information Retrieval Environment (SAPHIRE). SAPHIRE predefines a set of allowable terms [10], and maps the words and phrases to a set of canonical terms. Another approach is Northern Light, which classifies the documents into pre-defined subjects [34]. Another example using added concepts is the DYNOCAT system, which uses a knowledge server [14, 15]. It maps words and phrases to canonical terms, but it does not weight the term vector with frequency information [2]. This dynamic categorization uses a semantic base representation of terms, from the semantic type of each term. The most traditional way for adding conceptual information is natural language processing. Using this approach after preprocessing, tagging is applied. Many systems have been implemented using this approach [8, 17, 32]. Vector space modeling such as the SMART project does not place restrictions on the subject matter. SMART uses relevance feedback to tighten its search, but its user interface is quite unfriendly [19, 30]. Another vector system is Fox, where concepts other than normal content terms are used. An extended vector is assembled with classes of information about documents. Subvectors represent different concept classes, and similarity between extended vectors is calculated as a linear combination of corresponding subvectors [6].

Here, vectors are created from Medline abstracts to see how eliminating common words and infrequent terms affect document classification. The steps are:

1. Eliminate common English words:
   a. A “stop- list” containing 582 words is identified, ignoring words with little information content. Different lists are gathered from various sources and are merged together and sorted.
   b. The vector length is reduced by selecting a subset of key words. Then, a refinement is performed to find biomedical terms.
   c. The abstracts are parsed, stemming is applied, and a frequency count is tallied for each term. The number of documents in which each term appears is determined and the terms that occur most are identified. The importance of each term is assumed to be inversely proportional to the number of documents that contains that term. This IDF weighing scheme specifies that terms with low document membership are less discriminating [28].

2. Abstracts are chosen if they contain “breast”, “cancer”, “brca”, or “gene”.
3. Terms are normalized to lower case; non-alphanumeric terms are removed.
4. Common biomedical terms are removed.
5. For terms with the length six or more [11] forms (like cancer, cancerous) are replaced with their stem (cancer), using the Porter algorithm [12, 13].
6. Less frequent terms are deleted using a cut-off threshold, reducing the dimensionality of the document vectors, since infrequent terms have little resolving power (Zipf’s Law) [1, 25, 28].

If the document collection contains n unique terms, then each document is represented as a vector of length n: (t_1, t_2, t_3, t_4, … t_n) where t_i is proportional to the frequency of term i. The vector space is also represented in a simplified form by using 0s and 1s if the term is absent or present [1, 21- 25, 28]. The jth document d_j is thus written as an n-tuple $d_j = <w_{1j}, w_{2j}, w_{3j}, w_{4j}, \ldots, w_{nj}>$, and the ith keyword ti is written as an m-tuple $t_i = <w_{i1}, w_{i2}, w_{i3}, w_{i4}, \ldots, w_{im}>$.

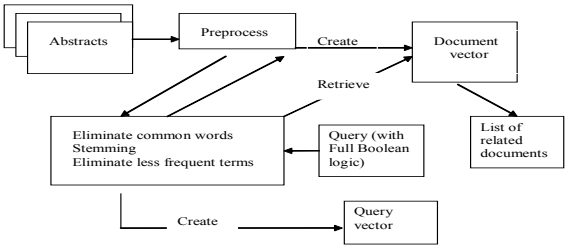


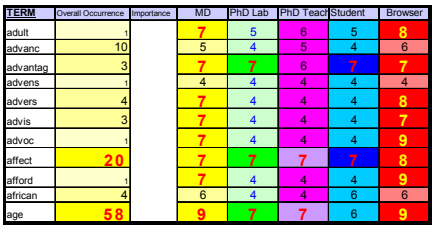
Fig. 1. Creating query vectors from abstracts

2.2 Query Process

The words in a user query are analyzed in a manner similar to that performed on the document abstracts. In addition, there are methods for improving information retrieval that involve adjusting the user query. For instance, the query can be expanded by adding all the terms that occur in the same category or cluster as each of the query terms [9, 31]. The list of returned documents is determined by Boolean operands “and” or “or”. Synonyms are entered manually, but in further versions of the system this will done automatically. The steps in processing a query are 1) eliminate common English words, 2) apply stemming, 3) create a query vector, 4) compare to the space of document term vectors, and 5) return the most related documents.

2.3 Document Ranking Using User Personalization

The relative importance-weighting levels of terms for a particular user (or group) are obtained and applied at query time. If the user query contains a given term, the slot for that term is set to one – otherwise the slot is set to zero. Profile information relates to terms that are important to a user. Medline users could be, say, medical doctors (MDs), college students, college teachers, PhD researchers, or even information-browsing patients. A list of terms is assembled that reflects those words commonly found in queries from such users. User evaluations of the relative importance of terms give a scale from zero to ten.



TERM	Overall Occurrence	Importance	MD	PhD Lab	PhD Teach	Student	Browser
adult	1		7	5	6	5	8
advanc	10		5	4	5	4	6
advantag	3		7	7	6	7	7
advens	1		4	4	4	4	4
advers	4		7	4	4	4	8
advis	3		7	4	4	4	7
advoc	1		7	4	4	4	9
affect	20		7	7	7	7	8
afford	1		7	4	4	4	9
african	4		6	4	4	6	6
age	58		9	7	7	6	9

Fig. 2. Term importance for user type

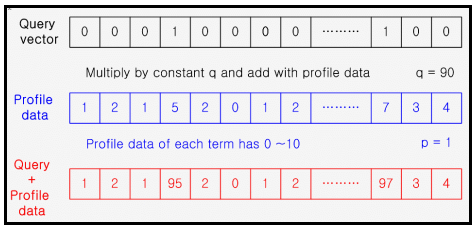


Fig. 3. Query vector and profile data

The vector for the query is combined with the preference-profile vector for that type of user. The influence of the profile is initially set so that 90% of the importance in the final query+profile vector is derived from the actual query, and the other 10% is derived from the profile vector. The system offers five variations of term-weighting: 1) standard tf (default), 2) best fully weighted, 3) classical tf-idf, 4) best weighted probabilistic, and 5) coordination level binary vector [23], as shown in the following table.

Table 1. Five different term-weighting methods

Methods	Document term weighting	Query term weighting
txc/txx	$\frac{tf}{\sqrt{\sum (tf_i)^2}}$	tf
tfc/nfx	$\frac{tf \log \frac{N}{n}}{\sqrt{\sum \left(tf \log \frac{N}{n}\right)^2}}$	$\left(0.5 + \frac{0.5tf}{\max tf}\right) \log \frac{N}{n}$
bxx/bxx	1 (if term is present)	1 (if term is present)
tfx/tfx	$tf \log \frac{N}{n}$	$tf \log \frac{N}{n}$
nxx/bpx	$0.5 + \frac{0.5tf}{\max tf}$	$\log \frac{N-n}{n}$

Documents are ranked and presented in an order that matches the query and the user's inferred interests. There are two well-known methods for such purposes, cosine similarity and distance similarity. The cosine vector similarity (inner product or dot product) is used here [33]. Each document vector is compared with the query-profile vector to give a -1 to +1 value. The similarity is computed as:

$$sim(Q, D_i) = \frac{\sum_{j=1}^{t} w_{q_j} \cdot w_{d_{ij}}}{\sqrt{\sum_{j=1}^{t} (w_{q_j})^2 \cdot \sum_{j=1}^{t} (w_{d_{ij}})^2}}$$

Fig. 4. Similarity between documents

When the documents are well-aligned, the angle between them is small and the cosine is near positive one.

2.4 Interactive Display and Intuitive Refinement

A spiral ranking graph is adopted [35], where each document is shown by a small dot, starting with the highest ranked (similar) document in the center. The user can see if the scores are distributed evenly or whether just a few documents are good matches. Documents that cluster together have a similar interest value to the user.

Fig. 5. Spiral ranking graph (left), clustering menu (middle), 3D vector visualization (right)

The interface includes sliders that enable refinement of the query term-weighting and group-profiling. The graph then dynamically changes the corresponding arrangement of the dots. Visualization of high dimensional term factors is narrowed through a manual selection of three selected term-dimensions. A k-means clustering algorithm is applied to distinctively color groups of similar documents.

2.5 Web Site Design

Website development should be based on principles, engineered to be flexible, and grounded in measurability. Commercial sites generally involve teams with special skills, integrating their efforts during the site's design, construction, evaluation, and support. Among the key areas are usability, User Experience Design (UXD), and personalization. Specific personas are walked through mainstream and side-path visits, and storyboards are used to explicitly trace out the key page-clicks and navigation. Novice web site designers often suffer from "developer's fallacy" in assuming that all users perceive the site like they do. Unfortunately, foolproofing a site is quite frustrating. Unlike commercial sites, academic sites are often in transitional situations over time, without a sustained vision or a way of measuring value and choosing the most important upgrades.

Several popular and important biologically-related Internet sites were reviewed, with a starting point of usability features and aids. Only a subset of all review approaches and metrics could be considered, so there were several informal interviews with various users of the sites. A walkthrough session by a typical user is common in commercial website evaluations, where the person makes selections and follows typical paths with good and not so good services from the site. Generally, bio-sites were found to be indifferent to user needs, differences, and difficulties.

The development here was undertaken to expose ideas and clarify alternatives related to the above objectives. In achieving modest site-development and information-access goals, a more central system coding purpose was also achieved - that of following explicit examples of design choices, time tradeoffs, and user awareness - as a contribution to the partnership of informatics and the pursuit of biological understanding.

The site's method of retrieval in many cases gave better results than Medline according to user evaluations. Several models underlie the user interface for the site.

- User: profile, persona, group, affective concern, motivation, goal, value, expertise, country, language, pet peeve
- Task: workflow, component, difficulty, failure
- System: database, processor, program, network, policy
- Site owner: institution, finder, administrator, professor
- Developer: coder, consultant, architect, tester, graphic artist, student
- Document: page layout, image map, link, abstract, summary, title, rating, size
- Document writer: profession, motivation, group, institution, country
- User interface: session, page, visit, program, timing

2.6 User Acceptance and Term-Associated User Profile Vectors

Another issue addressed explicitly is that of the measurement of success. The site was pre-determined to have a standard of satisfactory capability if a subset of documents could be accessed in an individual manner by disparate users.

3 Results

188 documents were used to evaluate the system, 29 of which did not contain any abstract. Only a small number of the most important words within a document provide the key information for classification, so a next phase of this project will address the elimination of terms that provide little semantic content or significance. A query ("breast" & "cancer" & "brca" & "gene") was entered as a test of the system's re-

Table 2. Identification and pruning of key terms

Step	# of terms
1.a Create stop-list	582 terms in stop-list
1.b Identify common terms within corpus (weight threshold = .15)	3 terms added to 1a
2 Terms within the corpus of documents	34832
3 After eliminating numeric and non-alphanumeric terms	30875
4 After eliminating common words	17884
5 After stemming and eliminating redundant terms	2334
6 After eliminating less frequent terms (threshold = .01)	1649

trieval. In this case, 106 out of the 159 documents were successfully retrieved. 34 out of the 53 documents that were not retrieved did not contain the term “gene”. Two of the documents did not contain any of the exact terms. The system here missed some documents that Medline retrieved, since the Medline search also uses concept enhancements rather than just the query or document terms.

It is possible for the user to adjust the weightings to assess the sensitivity of the with respect to key terms. The sensitivity related to the term changed can be calculated from the number of positions in the returned ranking that each document changes. The prototype site here is adaptable in that it encourages user self-identification for improved query handling. The user can also tell the site what terms were favored or to be excluded. The site was also set up to be automatically adaptive, with adjustments based on user activities, in the form of web visit logs. Different term-weighting methods give different ranking results.

4 Discussion

The speed of computer processors has grown at a rate of an 18 month doubling (Moore’s Law). The amount of data has also grown, but it is poorly understood that this rate exponentially exceeds processing power ("Russell's Law") and that the ability to transport data is also lagging behind. The increasing percentage of unused data is a concern that makes encyclopedic understanding ever less possible in areas like bio-tech and medical research. There is an increasing reliance on search engines, yet queries often return thousands of documents in less than optimal orderings. There is a need to provide different users with focused information to reduce the complexity of details.

In this paper a system is shown that incorporates adaptive term-frequency alignment with user profile data. The system also provides interactive visualization to give an immediate overview of the retrieval space. Query-term clustering indicates the relationships between documents, so the user can readily find similar documents.

Improvements to the system here could involve adding concepts and semantic types to the term processing by using UMLS [5]. In addition, categorizing results from queries using the MeSH hierarchy (Medical Subject Headings) could be of value.

References

1. V. Anh, O. Kretser, and A. Moffat, Vector-space ranking with effective early termination, Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, p.35-42, September 2001, New Orleans, Louisiana, United States
2. D. Berrios, R. Cucina, P. Sutphin, L. Fagan, Methods for Semi-Automated Indexing For High Precision Information Retrieval. 2002
3. Berry, M., Dumais, S., and Letsche, T., Computational methods for intelligent information access, Proc. of Supercomputing '95, San Diego, CA: USA, 1995.
4. C. Blaschke, M. Andrade, C. Ouzounis, and A. Valencia, Automatic extraction of biological information from scientific text: protein-protein interactions. Intelligent Systems for Molecular Biology, Heidelberg, p. 60, 1999.

5. K. Campbell, D. Oliver, and E. Shortliffe, The Unified Medical Language System: Toward a Collaborative Approach For Solving Terminologic Problems Submitted to a Special Issue of the Journal of the American Medical Informatics Association
6. C. Crouch, S. Apte, and H. A. Bapat. An IR approach to XML retrieval based on the extended vector model, http://www.dumn.edu/~ccrouch/pubs/XML_Paper_final.pdf
7. Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41, 391-407, 1990
8. K. Fukuda, A. Tamura, T. Tsunoda, and T. Takagi, Toward information extraction: identifying protein names from biological papers, Pac. Symp. Biocomput, p. 707, 1998
9. A. H. Griffiths, Claire Luckhurst, and Peter Willett, P Using inter-document similarity information in document retrieval systems, Journal of the American Society for Information Science, vol. 37, 3-11, 1986.
10. W. Hersh, and R. Greenes, SAPHIRE – An information retrieval system featuring concept matching, automatic indexing, probabilistic retrieval, and hierarchical relationships. Computers and Biomedical Research 23: 410-425, 1990
11. Iliopoulos, A. J. Enright, C. A. Ouzounis; TEXTQUEST: Document Clustering of Medical Abstracts for Discovery in Molecular Biology; Proceedings of the Sixth Annual Pacific Symposium on Biocomputing (PSB 01), 384-395, 2001
12. B. Lovins, Development of a stemming algorithm. Mechanical Translation and Computational Linguistics, 11, 22-31, 1968
13. M.E. Porter, An algorithm for suffix stripping program, 14(3), 130-137, 1990
14. Pratt, L. Fagan, The Usefulness of Dynamically Categorizing Search Results, Journal of the American Medical Informatics Association (JAMIA), 7(6), 605-617, 2000
15. W. Pratt, H. Wasserman, QueryCat: Automatic Categorization of MEDLINE Queries. Proceedings of the American Medical Informatics Association (AMIA) Fall Symposium 2000
16. D. Proux, F. Rechenmann, L. Julliard, V. Pillet, and B. Jacq, Detecting gene symbols and names in biological texts: a first step toward pertinent information extraction, Genome Informatics Workshop, Tokyo, p. 72., 1998
17. TC Rindflesch L. Tanabe, JN Weinstein, L. Hunter, EDGAR: extraction of drugs, genes and relations from the biomedical literature, Pac Symp Biocomput, 517-28.
18. Robertson, S., Walker, S., Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval, Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, 232-241, July 3-6, 1994
19. J. Rocchio, The SMART retrieval system - experiments in automated document processing, Relevance feedback information retrieval, ed. G. Salton, p. 313, Prentice-Hall, Englewood Cliffs NJ, 1971.
20. S. Russell, Knowledge Liquidity, Proceedings of Knowledge World, July, 1999.
21. G. Salton, Automatic content analysis in information retrieval, University of Pennsylvania, PA, 1968.
22. G. Salton, Developments in automatic text retrieval, Science 253, 974, 1991.
23. G. Salton, Term-weighting approaches in automatic text retrieval, Information Processing & Management, Vol 24, 5, 513-523, 1988
24. G. Salton., Automatic Text Processing: the transformation, analysis, and retrieval of information by computer, (Addison-Wesley, Reading MA, 1989)
25. G. Salton., A. Wang, and C. Yang, A vector space model for information retrieval. In Journal of the American Society for Information Science, vol 18, 613-620, 1975
26. T. Sekimizu, H. Park, and J. Tsujii, Identifying the interaction between genes and gene products based on frequently seen verbs in Medline abstracts, Genome Informatics Workshop, Tokyo, p. 62, 1998

27. J. Thomas, D. Milward, C. Ouzounis, S. Pulman, and M. Carroll, Automatic extraction of protein interactions from scientific abstracts, Pac. Symp. Biocomput, p. 538, 2000
28. T. Tokunaga, and M. Iwayama, Text categorization based on weighted inverse document frequency. Technical Report 94 TR0001, Department of Computer Science, Tokyo Institute of Technology, Tokyo, Japan, 1994.
29. S. Uger, and S. Gaucho, Feature reduction for document clustering and classification. Technical report, Computing Department, Imperial College, London, UK, 2000.
30. E.M. Voorhees, in WordNet: an electronic lexical database, Using WordNet for text retrieval, ed. C. Fellbaum, p. 285, MIT Press, Cambridge MA, 1998.
31. P. Willett, An algorithm for the calculation of exact term discrimination values, Information Processing and Management: an International Journal, v.21 n.3, 225-232, 1985
32. Yoshida, K. Fukuda, T. Takagi, PNAD-CSS: a workbench for constructing a protein name abbreviation dictionary, Bioinformatics, 16:169-75, Feb 2000.
33. Information Retrieval lecture note (University of Massachusetts Amherst) http://ciir.cs.umass.edu/cmpsci646
34. Northernlight, http://www.northernlight.com/
35. Interactive 3D Visualization for Document Retrieval, http://zing.ncsl.nist.gov/~cugini/uicd/viz.html