Effective Navigation of Question Results Primarily based on Conception Hierarchies

Ajitha S[1], Nayana V[2]

[1]Asst.Professor Dept. Of Computer Science and Engineering, MountZion Institute of Science and Technology, Chengannur, Kerala
[2]Asst.Professor Dept. Of Computer Science and Engineering, MountZion Institute of Science and Technology, Chengannur, Kerala

---***---

Abstract - The proposed system is aimed at solving the problem of information overload faced by researchers in the biomedical domain. When the user gives a query on the interface of the system, the corresponding abstracts are retrieved from the MEDLINE database using the Utils tool. The usage of a proxy that is run in locally. It intercepts all HTTP traffic, extracts queries, question chains, i.e., finally posed queries, end result sets, clicked end result pages, in addition to the complete click on movement of finally visited internet pages, and shops this statistics to a neighborhood database document which we discuss with because the neighborhood index withinside the following. Accordingly, searches with Google (the equal technique may be without problems implemented to another seek engine as well) are intercepted and seek consequences are re-ranked in keeping with private choices. offers a top level view of our seek personalization architecture.

Key Words: Categorization, feature extraction, cosine similarity,

1. INTRODUCTION

Database structures are being an increasing number of used for interactive and exploratory data retrieval .In such retrieval; queries frequently bring about too many solutions. Not all of the retrieved objects are applicable to the person; typically, best a tiny fraction of the end result set is applicable to person. Unfortunately, person frequently desires to look at all or maximum of the retrieved objects to locate the ones thrilling ones. This too-many-solutions phenomenon is normally stated as “statistics overload”. In this paintings we endorse a way to the trouble of statistics overload in MEDLINE database that's the storehouse of scientific journals. So on this thesis the trouble of statistics overload in MEDLINE is treated via way of means of categorizing the MEDLINE question consequences primarily based totally on MeSH idea hierarchy. A dynamic navigation scheme is employed withinside the MeSH idea hierarchy in order that the customers can locate question consequences applicable to their area of interest.

2. LITERATURE REVIEW

The proposed system [2] plays computerized undertaking of MeSH key phrases for a given clinical article reference primarily based totally on associative class, that's a information mining method derived from affiliation rule mining. It includes following 3 steps: • Data preparation Associative Classification calls for record withinside the shape of a transaction. A transaction has parts: a hard and fast of MeSH key phrases & a hard and fast of phrases extracted from title & summary. After the tree fields are extracted, the title & summary are processed one by one from the MeSH key phrases one by one. The key phrases are normalized to shape the MeSH tree identifiers. Each identifier is a sequence of wide variety separated via way of means of dots being the identifiers of department of MeSH tree from root node to suitable node indicated via way of means of the enter keyword. Title & summary are normalized via way of means of forestall word pruning & stemming. • Rule era This step generates common object units among article’s phrases & corresponding MeSH key phrases ensuing in a hard and fast of regulations. Minimum aid is first used throughout the rule era step to first of all lessen the wide variety of regulations that's in addition decreased throughout class via way of means of self belief threshold. • Tuning of parameters in this step a no of algorithm’s parameters are examined to find out an most beneficial set of values used in addition in class. This set of parameters are accountable for prunning , ranking & choice of regulations.

GoPubMed [3] makes use of each MeSH & Gene Ontology to categorize the question effects of PubMed. In this paintings a unique time period extraction set of rules is used to retrieve the MeSH & Gene Ontology phrases that are then used to assemble the sub-ontology primarily based totally at the question. If a phrase withinside the summary fits multiple time period the set of rules will go back the fit that has the very best degree with inside the ontology hierarchy and is the shortest some of the identical degree. Once the phrases are extracted from the summary the ontology must be offered for surfing abstracts. GoPubMed presentations the sub-ontology tree as a result derived & for every idea node with inside the tree it exhibits pinnacle 10 standards most effective. In our scheme every idea node exhibits a selective & dynamic listing of descendant nodes ranked most effective relevance to person question now no longer always children In Automatic textual content categorization the use of Neural networks, [6] the MEDLINE articles are classified primarily based totally on MeSH terms .The neural community is educated to assign MeSH terms primarily based totally at the time period frequency of phrases from the title & summary. In this paintings the trouble of spotting MeSH phrases for a particular file given a hard and fast of phrases with inside the file is solved through the use of back-propagation or counter propagation. Right here the enter to the neural community is the file vectors similar to every file to be classified & output is the class to which the file belongs. XPLORMED [7] is a device this is used to question on MEDLINE citations. It is a question refinement device .Method utilized in XPLORMED is primarily based totally on locating a subset of abstracts through the use of fuzzy binary relations. Following steps are hired on this paintings: • A MEDLINE seek produces abstracts • This is given as enter to the machine. 6 • The machine selects the phrase from abstracts primarily based totally on their electricity of affiliation to different phrases. • Selected phrases are joined to shape the elegance of related phrases. • One or extra phrase training are used to choose the subset of abstracts. The affiliation among phrases in summary are described through fuzzy binary relations: • S w-Degree of relatedness among phrases • I w - Degree of inclusion of 1 phrase into another So end of this literature survey is that works centered on categorization of MEDLINE files are there. But none of them offer an green navigation scheme the use of which the person can undergo the question effects .The closest to the proposed machine is GoPubMed machine which implements a static navigation technique at the effects of PubMed. GoPubMed lists a predefined listing of high-degree MeSH standards, such as "Chemicals and Drugs," "Biological Sciences," and so on, and for every one in every of them presentations the pinnacle-10 standards.

3. METHODOLOGY

The proposed system is aimed at solving the problem of information overload faced by researchers in the biomedical domain. When the user gives a query on the interface of the system, the corresponding abstracts are retrieved from the MEDLINE database using the eUtils tool. The usage of a proxy that is run in locally. It intercepts all HTTP traffic, extracts queries, question chains, i.e., finally posed queries, end result sets, clicked end result pages, in addition to the complete click on movement of finally visited internet pages, and shops this statistics to a neighborhood database document which we discuss with because the neighborhood index with inside the following. Accordingly, searches with Google (the equal technique may be without problems implemented to another seek engine as well) are intercepted and seek consequences are re-ranked in keeping with private choices. Offers a top level view of our seek personalization architecture.

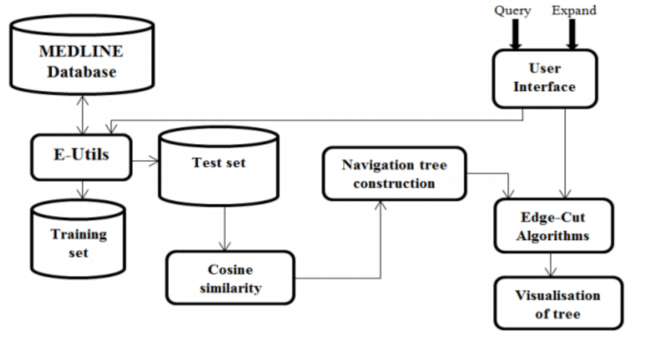


Fig 1: Methodology

The retrieved abstracts are categorized into the MeSH concept hierarchy based on cosine similarity of TFIDF vector of the result document set and the training set documents .The resulting tree is given as input to the edge-cut algorithms of the dynamic navigation scheme to reduce the tree size. Thus the navigation cost of the user is minimized. The main two modules of this system are described in detail below:

3.1 Categorization of Documents

A concept hierarchy is created by using the MESH concept hierarchy. For each MESH concept in the hierarchy corresponding training documents are assigned. When a query is given in the MEDLINE browser the corresponding abstracts are retrieved and categorized according to the MESH concepts. The categorization is done by computing cosine similarity of the training documents and the retrieved documents.

f

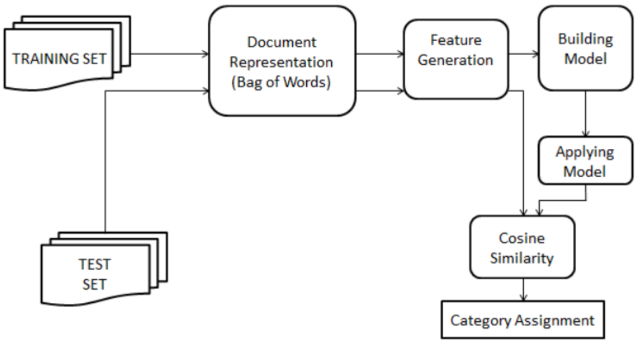


Fig 2 : steps in categorization

Representing features of each node to convert set of training documents representing a MESH concept into a standard form, first the documents are subjected to a set of stemming and stopping procedures to obtain a Bag of Words representation. Stopping procedure removes the common words from the set of documents. Stemming procedure converts the words of a document into its root word. A vocabulary is formed consisting of the MESH concepts. Then the following frequencies of a document are computed: • Term Frequency (TF): The term frequency of the ith word Wi in the vocabulary is the no of times the word appears in the document • Document Frequency(DF): Document frequency of the ith word Wi in the vocabulary is the no of documents in which the word appears. 15 • Inverse Document Frequency(IDF): Inverse document frequency of a word w is calculated by : IDF (w)=log(|d|)/DF(w) Here d is the total documents of in the node • Term Frequency × Inverse Document Frequency(TFIDF): TFIDF(w , d)=TF(w ,d) × IDF(w) The feature vector of each node is computed by F=«TFIDF(w1,d)...

3.2 Dynamic navigation scheme: Edge-Cut Algorithms

The Ideal Edge-Cut Algorithm computes the navigation cost of each node of the input tree traversing it in post-order .For each node of the input tree excluding the leaf the algorithm stores all possible edge-cuts of the tree rooted at the node and the set of all possible sub-trees rooted at the node. Compute the cost of all valid edge cuts for the set of sub-trees and take the minimum cost edge cut as the ideal one.

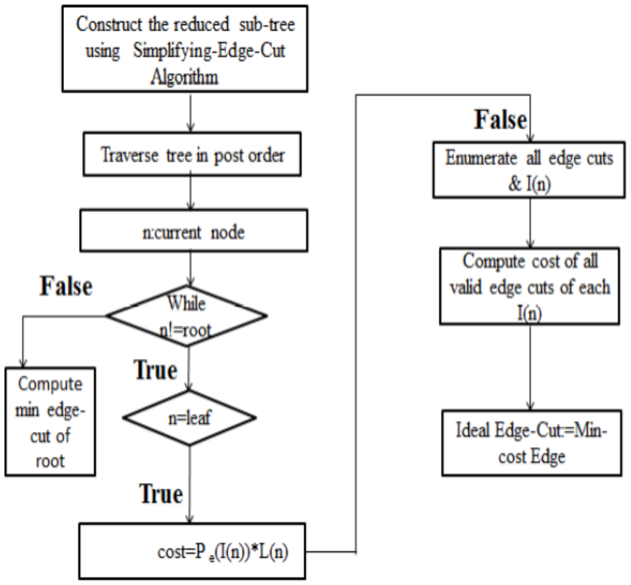

Construct the reduced sub-tree using Simplifying-Edge-Cut Algorithm
Traverse tree in post order
n:current node
False
While n!=root
True
n=leaf
True
Compute min edge-cut of root
cost=P e(I(n))*L(n)
False
Enumerate all edge cuts & I(n)
Compute cost of all valid edge cuts of each I(n)
Ideal Edge-Cut=Min-cost Edge

Fig 3 : Ideal edge cut algorithm

The algorithm to compute the minimal navigation cost, Ideal-Edge-Cut is exponential and hence infeasible for the navigation trees of most queries. So Ideal Edge-Cut is to be run on a reduced navigation tree to be run on real time. This reduced tree is obtained by Simplifying-Edge-Cut Algorithm which partitions the tree .For each tree node n, the algorithm removes the ”heaviest” children of n one-by-one until the weight of n falls below k.

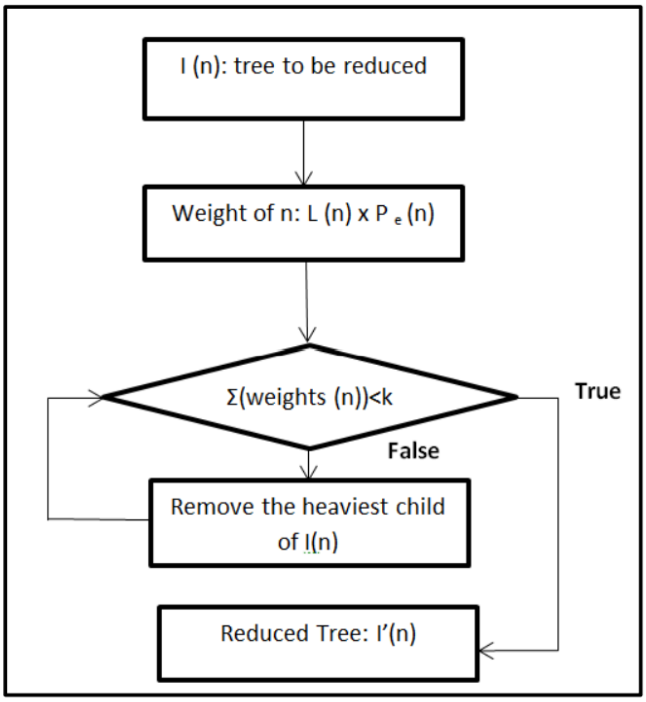

I (n): tree to be reduced
Weight of n: L (n) x P e (n)
Σ(weights (n))<k
True
False
Remove the heaviest child of I(n)
Reduced Tree: I'(n)

Fig 4 : Simplifying the edge cut algorithm

The experiments have been performed on a Toshiba Satellite C640 device with 2. four GHz CPU and four GB of most important reminiscence going for walks Windows7.All the algorithms are applied the usage of java and SQL database is used. The MeSH idea hierarchy is applied as a tree in jsp. The tree nodes are saved in an SQL database. When the consumer offers a question at the browser, the corresponding abstracts are retrieved from the MEDLINE database the usage of the eUtils. Initially the usage of the eUtils the PubMedIDs similar to the question are retrieved. PubMedIDs for this reason retrieved are furnished as enter to the eUtils question to retrieve the corresponding abstracts. In the experimental set-up the no of abstracts which can be retrieved similar to a question is 100.It is performed with the aid of using putting the retmax of the eUtil question as 100 Initially a education record set is there for every MeSH idea. The TFIDF vector of every MeSH idea is constituted of the education set. Next, the TFIDF of the retrieved abstracts are computed .The retrieved abstracts are labeled into the MeSH idea hierarchy primarily based totally on cosine similarity of TFIDF of the end result record and the education set files that is computed with the aid of using the categorization set of rules. Initially simplest the foundation node is shown. When the consumer plays the EXPAND motion on the foundation node then the Simplifying Edge-Cut set of rules is administered to lessen the tree size. The ensuing tree is given to the Idea lEdge-Cut set of rules to locate the minimal value side reduce which minimizes the navigation value of the consumer

4. RESULTS

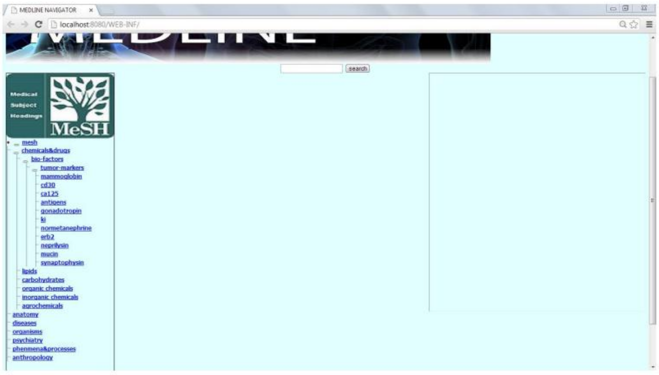


Fig 5: MeSH tree before Edge-cut operation

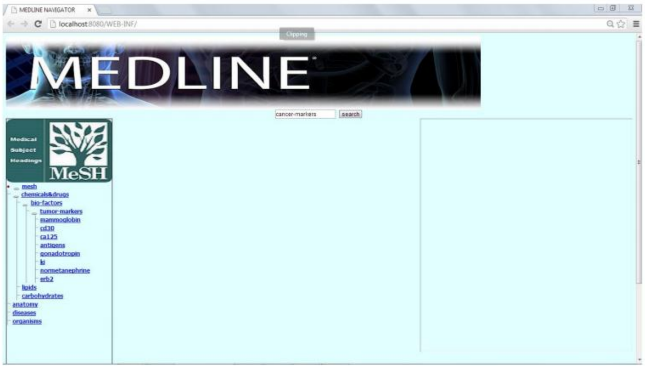
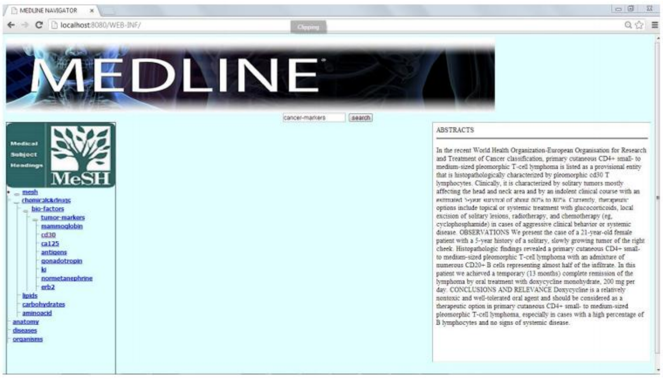


Fig 6 : MeSH tree after Edge-cut operation



ABSTRACTS

In the recent World Health Organization-European Organisation for Research and Treatment of Cancer classification, primary cutaneous CD4+ small- to medium-sized pleomorphic T-cell lymphoma is listed as a provisional entity that is histopathologically characterized by pleomorphic cd30 T lymphocytes. Clinically, it is characterized by solitary tumors mostly affecting the head and neck area and by an indolent clinical course with an estimated 5-year survival of about 60% to 80%. Currently, therapeutic options include topical or systemic treatment with glucocorticoids, local excision of solitary lesions, radiotherapy, and chemotherapy (eg, cyclophosphamide) in cases of aggressive clinical behavior or systemic disease. OBSERVATIONS We present the case of a 21-year-old female patient with a 5-year history of a solitary, slowly growing tumor of the right cheek. Histopathologic findings revealed a primary cutaneous CD4+ small- to medium-sized pleomorphic T-cell lymphoma with an admixture of numerous CD20+ B cells representing almost half of the infiltrate. In this patient we achieved a temporary (13 months) complete remission of the lymphoma by oral treatment with doxycycline monohydrate, 200 mg per day. CONCLUSIONS AND RELEVANCE Doxycycline is a relatively nontoxic and well-tolerated oral agent and should be considered as a therapeutic option in primary cutaneous CD4+ small- to medium-sized pleomorphic T-cell lymphoma, especially in cases with a high percentage of B lymphocytes and no signs of systemic disease.

Figure 7: Categorized query results

4. CONCLUSION

Information explosion is a chief hassle confronted with the aid of using the researchers whilst looking biomedical databases even as searching for applicable works of their vicinity of hobby. In this paintings the hassle is addressed with the aid of using categorization and navigation of question outcomes the usage of MeSH idea hierarchy. Categorization is completed primarily based totally at the cosine similarity of the retrieved abstracts and the MeSH standards. So the question outcomes are organised right into a dynamic MeSH tree which has MeSH standards as its nodes .A dynamic navigation scheme is supplied withinside the MeSH tree such that every node growth exhibits most effective a small subset of the idea nodes primarily based totally at the consumer question as a way to decrease the navigation cost. The consumer can navigate the MeSH tree, discover the standards which might be of hobby to them and the articles classified below the idea.

REFERENCES

[1] A. Kashyap, V. Hristidis, M. Petropoulos, and S. Tavoulari, "Effective navigation of query results based on concept hierarchies," IEEE Transactions on Knowledge and Data Engineering, vol. 23, no.4,pp. 540–553, April 2011.

[2] R. Rak, L. A. Kurgan, and M. Reformat, "Multilabel associative classification categorization of medline articles into mesh keywords," IEEE Engineering in Medicine a nd Biology vol. 26, pp. 47–55, April 2007.

[3] A. Doms and M. Schroeder, "Gopubmed: exploring pubmed with the gene ontology," Oxford Journal on Life Sciences, vol. 33,no.2, pp. 783–786, April 2005.

[4] K. Chakrabarti, S. Chaudhuri, and S.W. Hwang, "Automatic Categorization of Query Results," Proc. ACM SIGMOD, pp. 755- 766, 2004

[5] J.S. Agrawal, S. Chaudhuri, G. Das, and A. Gionis, "Automated Ranking of Database Query Results", First Biennial Conf. Innovative Data Systems Research,2003.

[6] Miguel.E.Ruiz and P. Srinivasan, "Automatic text categorization using neural networks," Springer:Information Retrival, no.5,pp. 87–118, August 2002. 23

[7]Medical Subject Headings (MeSH), http: //www.nlm.nih.gov/ mesh/, 2010.

[8]Entrez Programming Utilities, http://www.ncbi.nlm.nih.gov/entrez/query/static /eutils_help.html, 2008.

[9] W. T. Chuang, A. Tiyyagura, J. Yang, and G. Giuffrida, "A fast algorithm for hierarchical text classification," Springer:Data Warehousing and Knowledge Discovery,pp.409-418,September 2000

[10] C. Perez-Iratxeta, P. Bork, and M.A. Andrade, "Exploring MEDLINE Abstracts with XplorMed," Drugs of Today, vol. 38, pp. 381-389,