IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

Survey for Mining Biomedical data from HTTP Documents

Shally HR[*1], Rejimol Robinson R R[2]

[*1,2]Dept.Computer Sceince & Engineering ,Sree Chitra Thirunal College of Engineering Trivandrum, India

shalu.it88@gmail.com

Abstract

Medical informatics has been rapidly growing in the last decade. Despite of major advantages in the science and technology of health care it seems that medical informatics discipline has the potential to improve and facilitate the ever- changing and ever-broadening mass of information concerning the etiology and prevention , treatment of diseases as well as the maintenance of health. With the rapid development of computer technologies and medical industries, the sharing and exchanging of medical information has become increasingly important. Data mining is a process of discovering useful information from a database and analysis of extracted information. Goal of this paper is to know the new techniques which is used to mine the biomedical documents from hypertext documents.

Keywords: Ontology UMLS,Pagerank,Hypertext Induced Topic Search

Introduction

Web is a formidable task which contains millions of documents about a specific term. Finding a useful document, which really contains the information that are looking for, from this huge web database is not an easy task. Web also contains documents on medical science. Medical documents are different from normal documents on the basis of data. Biomedical words are different from natural English language word. Google is one of a popular search engine used today. Google uses various techniques for ranking and retrieving documents from the web.Here we are going to see how the biomedical words can mine from the hypertext documents.

Commonly two type of algorithm is used for retrieving the web documents. They are pagerank[2] and Hypertext induced topic search[3]. Keyword Query is used to find some related document. If two documents have same keyword then a comparison is done between the hyperlinks of two documents and it will rank the document based on the hyperlink. This method is not fair at all because websites can add so many fake links so this will lead to increase the ranking of a particular document.

Biomedical named entity recognition(NER) refers to the task of automatically identifying occurrences of biological or medical terms in unstructured text. Common entities of interest include gene and protein names ,medical problems and treatments, drug names etc.

Data mining is a young and interdisciplinary field, drawing from fields such as database systems, data warehousing, machine learning, statistics, signal analysis, data visualization, information retrieval, and high performance computing. It has been successfully applied in diverse areas such as marketing, finance, engineering, security and games. Another fact is that some users may visit first twenty to thirty documents returned by google for a query. Rest of the documents that are discarded by users may contain more information Biomedical documents are searched by medical personals and medical institutes for collecting the newly discovered information. Medical ontologies are developed to solve problems such as the demand for reusing and sharing patient data or the transmission of these data. The unambiguous communication of complex and detailed medical concepts is a crucial feature in current medical information systems. In these systems, several agents must interact in order to share their results and, thus, they must use a medical terminology with a clear and non-confusing meaning. But the development of ontology is a difficult task. The more the biomedical term is present the more the document is important and relevant. Medical ontology is used for counting the total number of biomedical terms which is present in the document. Then re-rank the documents based on the count of

medical terms. Ontology is similar to a dictionary or glossary, but with greater detail and structure that enables computers to process its content. Ontology consists of a set of concepts, axioms, and relationships that describe a domain of interest. amount of domain knowledge. Much of this information is contained in biomedical ontology. consists of a set of concepts, axioms, and relationships that describe a domain of interest. amount of domain knowledge. Much of this information is contained in biomedical ontology.

Mining Techniques

Web Mining

Web mining is the integration of information gathered by traditional data mining methodologies and techniques with information gathered over the World Wide Web. (Mining means extracting something useful or valuable from a baser substance, such as mining gold from the earth.) Web mining allows you to look for patterns in data through content mining, structure mining, and usage mining. Content mining is used to examine data collected by search engines and Web spiders. Structure mining is used to examine data related to the structure of a particular Web site and usage mining is used to examine data related to a particular user's browser as well as data gathered by forms the user may have submitted during Web transactions. Web usage mining mines secondary information extracted from user interactions with the web while surfing. Web structure mining is the discovery of interesting patterns from the hyperlink structure of the web. It involves mining the web document's structure and links aimed at increasing coverage and giving accurate response to web search queries . Web content mining deals with classifying web pages into categories based on their contents so that similar pages can be grouped together to enhance performance. Web structure and web content mining aim at summarizing formation on web pages to facilitate efficient and effective information retrieval.

Data Mining

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

Text Mining

Text mining, sometimes alternately referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output.

Outlier Mining

Web outlier mining targeting web datasets has not received similar attention in the mining community. There are three types of web outliers; web content outliers, web structure outliers, and web usage outliers . A web content outlier is described as a web document with different contents compared to similar documents taken from the same category.

Link Analysis Algorithms

There are three famous link analysis Algorithms are available

1.HITS Algorithm
2.PageRank Algorithm
3.WCOW-Mine Algorithm\

Hyperlink-Induced Topic Search (HITS)

Hyperlink-Induced Topic Search (HITS) is a link analysis algorithm which helps in rating Web pages also known as Hubs and authorities It conclude two main values for a page:

1. Page authority, which estimates the value of the content of the page.

2. Page hub value, which estimates the value of its links to other pages.

First it retrieve the set of results to the search query so that the computation is performed only on this result set and not across all Web pages. Authority and hub values are defined in terms of one another in a mutual recursion. An authority value is calculated as the sum of the scaled hub values that point to that page. A hub value is the sum of the scaled authority values of the pages it points to. Some implementations also consider the relevance of the linked pages.

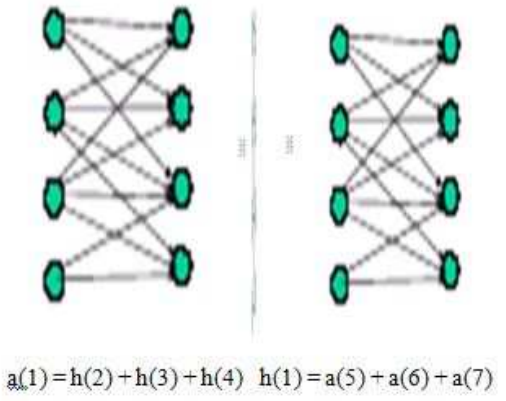
The algorithm performs a series of iterations, each consisting of two basic steps:

Authority Update: Update every node's Authority score to be equal to the sum of the Hub Score's of every node that points to it. That is, a node is given a high authority score by being linked to by pages that are recognized as Hubs for information.

Hub Update: Update every node's Hub Score to be equal to the sum of the Authority Score's of every node that it points to.
That is, a node is given a high hub score by linking to nodes that are considered to be authorities on the subject.
The Hub score and Authority score for a node are defined with the following algorithm:

1. Start with every node having a hub score and authority score of 1.
2. Run the Authority Update Rule
3. Run the Hub Update Rule
4. Normalize the values by dividing every Hub score by the sum of the squares of all Hub scores, and dividing each Authority score by the sum of the squares of all Authority scores.
5. Repeat from the second step as necessary.

HITS focus on both authoritative pages and good hub pages, but Pagerank only focus on the authoritative pages. HITS is query dependent. while querying, it will cost time to calculate the authoritative pages and hubs based on the query. So HITS may not be quite efficient. But the score of pagerank is static until new pages are added. In addition to the efficiency mentioned above, HITS only consider the first group of random relevant pages, but the content relevance of later expanded pages are ignored. Besides, it likes to return more general pages than specific answers.



STEPS

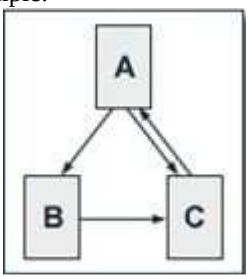
1. Determines a base set S
2. let set of documents returned by a standard search engine be called the root set R
3. Initialize S to R
4. Add to S all pages pointed to by any page in R.
5. Add to S all pages that point to any page in R
6. Maintain for each page p in S:
7. Authority score: ap (vector a)
8. Hub score: hp (vector h)
9. For each node initiliaze the ap and hp to 1/n
10. In each iteration calculate the authority weight for each node in S
11. In each iteration calculate the hub weight for each node in S
12. After new weights are computed for all nodes, the weights are normalized:

PageRank Algorithm

The original PageRank algorithm was described by Lawrence Page and Sergey Brin in several publications. It is given by
PR(A) = (1-d) + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))
where
PR() is A the PageRank of page A,
PR(Ti) is the PageRank of pages Ti which link to page A, C(Ti) is the number of outbound links on page Ti and d is a damping factor which can be set between 0 and 1.
So, first of all, that PageRank does not rank web sites as a whole, but is determined for each page individually. Webcrawler is used to fetch the page from web. Further, the PageRank of page A is recursively defined by the PageRanks of those pages which link to page A. The PageRank of pages Ti which link to page A does not influence the PageRank of page A uniformly. Within the PageRank algorithm, the PageRank of a page T is always weighted by the number of outbound links C(T) on page T. This means that the more outbound links a page T has, the less will page A benefit from a link to it on page T.The weighted PageRank of pages Ti is then added up. The outcome of this is that an additional inbound link for page A will always increase page A's PageRank.Finally, the sum of the weighted PageRanks of all pages Ti is multiplied with a damping factor d which can be set between 0 and 1. Thereby, the extend of PageRank benefit for a page by another page linking to it is reduced.

The Characteristics of PageRank

The characteristics of PageRank shall be illustrated by a small example.



Assume that a small web consisting of three pages A, B and C, whereby page A links to the pages B and C, page B links to page C and page C links to page A. According to Page and Brin, the damping factor d is usually set to 0.85, but to keep the calculation simple set it to 0.5. The exact value of the damping factor d admittedly has effects on PageRank, but it does not influence the fundamental principles of PageRank. So, we get the following equations for the **PageRankcalculation:**

PR(A)=0.5+0.5PR(C)

PR(B)=0.5+0.5(PR(A)/2) PR(C) = 0.5 + 0.5 (PR(A) / 2 + PR(B))

These equations can easily be solved. We get the following PageRank values for the single pages:

PR(A)=14/13=1.07692308

PR(B)=10/13=0.76923077 PR(C) = 15/13 = 1.15384615

It is obvious that the sum of all pages' PageRanks is 3 and thus equals the total number of web pages. For simple three-page example it is easy to solve the according equation system to determine PageRank values. In practice, the web consists of billions of documents and it is not possible to find a solution by inspection.

WCOND-Mine algorithm

WCOND-Mine algorithm is used to mine the webcontent.WCOND-Mine algorithm exploits the advantages of ngrams and the html structure of the web. The algorithm does not require a domain dictionary. The n-gram frequency profile for each document is generated separately. The weight of document is based on their ngram frequency distribution and the html tags that enclosed their root words.

Input:documents(Di),weights w(Nkit)

Outputs:top-n outliers

1. Extract/Preprocess documents
2. Read preprocessed documents(Di)
3. Generate n-gram frequency profile
4. For(int i=0;i<NoOfDoc;i++){
5. Compute document dissimilarities(DISi) using equations (1) to (5)
6. }//end for
7. Sort first- n documents in descending
8. Order of DLSi,label them as top-n
9. For (int k=0;k<NoOfDoc-n;k++){
10. If(DISk < maxDIS){
11. Delete document
12. Else
13. Delete document with maxDIS from top-n
14. Add document to top-n and re-sort
15. maxDIS=New maximum
16. }//End if
17. }//End for
18. Print the top-n outliers

Natural language processing (NLP) is a field of computer science, artificial intelligence (also called machine learning), and linguistics concerned with the interactions between computers and human (natural) languages. Specifically, it is the process of a computer extracting meaningful information from natural language input and/or producing natural language output. In theory, natural language processing is a very attractive method of human-computer interaction.

Supporting Tool:UMLS The Unified Medical Language System (UMLS) is a compendium of many controlled vocabularies in the biomedical sciences It provides a mapping structure among these vocabularies and thus allows one to translate among the various terminology systems; it may also be viewed as a comprehensive thesaurus and ontology of biomedical concepts. UMLS further provides facilities for natural language processing. It is intended to be used mainly by developers of systems in medical informatics. UMLS is used to remove stopwords which are present in the documents. UMLS consists of Knowledge Sources (databases) and a set of software tools.

Steps For Mining For Mining Biomedical

By using the above three algorithm we will collect all the HTML pages then we will go for biomedical data mining

1. Collect all the HTML documents ,then convert the HTML document as a text Documents.
2. HTML documents will contain several tags that are not necessary .we know that Google documents contains several text and un-necessary word those words are called stop words so we have to eliminate all the stop words in the document else it will make our search incorrect and it will add so many fake links .this is the preprocessing stage .Now we have simplified our documents.

3. Here we are going to identify the biomedical term, For identifying biomedical word we use UMLS thesaurus . The identification of biomedical entity from the metathesaurus will be done by simple keyword querying the database based on the word from text document. If any word is found as biomedical by the help of UMLS database then it is save in a document.
4. After getting the biomedical document The next step is to calculate the importance of the page according to the frequency (number of occurrence) of biomedical terms. The more the bio-medical terms identified the more the important or relevant the page is. Terms of the documents are counted and saved in a separate file .Documents containing highest number of files are ranked first.

Conclusion

Biomedical document retrieval method is showing nearly optimal results. However, the importance of a web document greatly depends on user's need i.e., how much relevant the web document is according to the user query. More efficient web text processing (NLP) may bring more optimal result for the experiment. Efficiency depends on time, the time of calculating the ranking could be an important issue. So far we are thinking about off line measurement, which is done by most of the existing crawler. As web contains millions of documents, there is no other better options to search for accurate result.

References

[1] Agyemang, Malik, "Web Content Outlier Mining:Motivation, Framework, and Algorithms",Proceedings of the 2004 ACM symposium on Applied computing, 2004.
[2] Lawrence and Brin, Sergey and Motwani, Rajeev and Winograd, Terry, "The Page Rank Citation Ranking: Bringing Order to the Web", Technical Report. Stanford InfoLab, 1999.
[3] Jon M. Kleinberg, "Authoritative sources in a hyperlinked environment", J. ACM, 46(5):604-632, 1999,
[4] UMLS® Knowledge Source Server , 07 July 20 I O.
[5] Musen, M.A., "Medical Informatics: Searching for Underlying Components, " Methods InfMed 41(1):12-19, 2002.
[6] Hahn, U., Romacker, M., and Schulz, S, "How Knowledge Drives Understanding-Matching Medical Ontologies with the Needs of Medical Language Processing, " Artif Intell Med 15(1):25-51, 1999.
[7] Degoulet, P, Sauquet, D., Jaulent, M.e., Zapletal,E., and Lavril, M., "Rationale and Design Considerations for a Semantic Mediator in HealthInformation Systems", Methods Inf Med 37(4-5):5 18-526, 1998.
[8] Pisanelli, D.M., Gangemi, A., Battaglia, M., and Catenacci, e., "Coping with Medical Polysemy in the Semantic Web: The Role of Ontologies," Medinfo 2004:416-419, 2004.
[9] Sougata Mukherjea, Saurav Sahay: "Discovering Biomedical Relations Utilizing the World-Wide Web", Pacific Symposium on Biocomputing , 164-175, 2006.
[10] H. Shatkay and R. Feldman. "Mining the Biomedical Literature in the Genomic Era: An Overview", Journal of Computational Biology (JCB), V.lO, Issue: 6, pp. 821-856, December,2003.
[11] He Tan and Patrick Lambrix, "Selecting an ontology for biomedical text mining", Workshop on BioNLP, 2009.
[12] Agyemang M., Ezeife, C.I. “LSC-Mine: Algorithm for Mining Local Outliers” Proc. of 15th IRMA International Conference, Idea Group, New Orleans,USA, May 2004, pp (2)5-8
[13] Breunig M.M., Kriegel H-P., Ng R.T., Sander J. “LOF: Identifying Outliers in Large dataset” Proc. ACM SIGMOD 2000 Int. Conf. on Management of Data,Dallas, TX, USA, 2000, 29(2): 93-104
[14] Barnett V., Lewis T. “Outliers in Statistical Data” John Willey, 1994
[15] Chakrabarti, S., Berg, M., Dom, B. “Focused Crawling: A New Approach to Topic-specific Web Resource Discovery” Computer Networks, Netherlands, 1999
[16] Chakrabarti S., Dom B., Gibson D., Kleinberg J.,Kumar S, Raghavan P., Rajagopalan S., and Tomkins A. “Mining the Link Structure of the World Wide