Finding Hidden Relationships Between Medical Concepts by Leveraging Metamap and Text Mining Techniques

Weikang Yang[1], S. M. Mazharul Hoque Chowdhury[2], and Wei Jin[3](✉)

[1] Department of Computer Science, North Dakota State University, Fargo, USA
yangvigoo@gmail.com

[2] Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh
mazharul2213@diu.edu.bd

[3] Department of Computer Science and Engineering, University of North Texas, Denton, USA
wei.jin@unt.edu

Abstract. Text is one of the most common ways to store data in this computerized world. At a glance, it may seem that those data are not interconnected. But in reality, data can have hidden connections. Therefore, in this research, a new model has been presented that can find hidden relationships between two medical concepts by using MetaMap and appropriate text-mining techniques. Specifically, the model creates a new comprehensive index structure and can find cross-document hidden links connecting topics of interest that most existing approaches have ignored. Experiments show the effectiveness of the proposed model in discovering new connections between topics.

Keywords: Biomedical text mining · Hidden relationship discovery · Cross-document knowledge discovery

1 Introduction

If we consider the modern world, it is not possible to imagine it without the use of text. Therefore, text data plays a very important role in every sector of our lives, and most of the time a piece of text type data may contain some connection with other texts. In some cases, those connections are clearly visible and, in some cases, relations remain hidden. On the other hand, when people are interacting with text while reading books or newspapers or communicating with others via text messages, they try to make some connections between different things and try to get hidden information based on the text [1]. However, it is difficult for people to find all the connections based on a little information only. In some cases, a sentence indicating "A implies B" in one article and another sentence "B implies C" in another article, and people generally could not link them together to establish the relationship that "A may imply C", especially when they are facing a large volume of data. This idea was explained by Swanson in 1999 and he proposed a model named 'ABC model' to solve it [2].

W. Chen et al. (Eds.): ADMA 2022, LNAI 13725, pp. 41–52, 2022.
https://doi.org/10.1007/978-3-031-22064-7_4

The primary focus of this research is motivated by Swason's ABC model, and we extend it to a cross-document discovery setting and preform a finer granularity of relationship discovery. Additionally, to cope with a large-scale document collection, we introduce a new and efficient index structure that can support multi-level knowledge discovery and we believe it will also facilitate many other downstream natural language processing applications. We further develop a new approach that will be able to predict hidden relations between two input topics effectively. This model was tested against other existing models and showed its effectiveness. Moreover, instead of a single-thread implementation, we also explored a multi-thread implementation setting in order to reduce analysis time and increase system efficiency at the same time. Response time and accuracy can both be further improved by adding threads if computer hardware resource is concerned. In this research Medline database was used as the data source because of its large volume of articles on Bio-medicine and MetaMap was used to analyze the data. Output produced from this model will be combined with Swanson's model to find hidden connections.

The reminder of this paper is structured as follows. Section 2 discusses related work. In Sect. 3, we present an overview of our methods. In Sect. 4, we present details of our proposed techniques. Sections 5 and 6 present experiments and evaluation results. And finally, Sect. 7 brings conclusion and gives directions for future work.

2 Literature Review

In this modern world, data is everything and most of them are text-type data. There are a lot of tools available for text data analysis and researchers around the world have developed tools like MicroMeSH, SAPHIRE, and MetaMap to do this job. MetaMap is a public tool developed by the National Library of Medicine (NLM), which can map text to biomedical concepts using its enormous thesaurus [3]. This has been a popular tool applied to many Information Retrieval and text-mining applications [4]. When Metamap receives a sentence, it breaks a sentence into phrases and those phrases are used to find mapping concepts to calculate a mapping score for each concept. There are already many applications of MetaMap in the field of biomedical studies. For example, Wendy Marcelo et al. developed a system to detect patients with respiratory illness by processing patients' clinical reports with MetaMap [5]. Zuccon Holloway et al. tried to automatically identify disorders mentioned in health records such as discharge summaries by using MetaMap [6]. In order to test the performance of MetaMap, Pratt and Yetisgen conduct research, which aims to compare MetaMap's capability with that of people who are familiar with the biomedical field [7]. The result shows that MetaMap is capable to identify the biomedical concepts with a 93.3% recall compared with those tagged by biomedical experts.

Knowledge discovery in biomedical texts originates from Swanson's ABC model, based on which Jin and Srihari proposed a new type of query, namely, concept chain queries, attempting to detect hidden links between concepts [2]. Researchers Wei Jin and Rohini Srihari tried to generate concept chains connecting topics of interest from counterterrorism documents. The approach used a variant of the Term Frequency (TF) - Inverse Document Frequency (IDF) weighting scheme for evaluating each potential connecting term, achieving an approximate 82.5% recall [8]. This work is different from Jin's study by using a completely different context – the biomedical domain instead of the counterterrorism corpus. Another related work was proposed by Gopalakrishnan, Kishlay et al., which introduced a new approach that created a graph-based knowledge base and then used a bi-directional search for finding paths in the graph to answer concept chain queries in the biomedical field [9].

Researcher Vishrawas Gopalakrishnan et al. worked on Hypothesis generation and there he used medical data in order to find hidden relations [16]. In their research, they also focused on the hidden relation discovery based on given keywords to develop a hypothesis on a topic. Researcher Xiaohua Hu et al. presented a model that is capable to find hidden links from Medline data and generate novel hypothesis between these concepts [17]. Padmini Srinivasan and Bisharah Libbus worked on research where they used Medline citations to find connections between dietary substances and diseases [18]. In their research, they were trying to discover the therapeutic potential of curcumin/turmeric which is commonly used in Asia. Another research was done by Kishlay Jha and Wei Jin on the extraction of novel knowledge from biomedical literature and domain knowledge [19]. In their research, they presented a meticulous analysis of how manifold statistical information measures and semantic knowledge affect the knowledge discovery procedure.

Moreover, the use of hidden relation extraction covers a large area that includes relations between flood and hydrological variables, immunological markers in multiple sclerosis, and so on. In the year 2007, M.J. Sanscartier and E. Neufeld worked on a model that will be able to find hidden relationships between different variables in a dataset in order to correct models by finding hidden contextual variables [11]. Prakash & Surendran (2013) worked on the detection of social media hidden activity based on the connection between different types of social media profiles and the network traffic of a user [12]. Milton Pividori et al. worked on the discovery of hidden relations between different qualitative and quantitative data without standardization based on their influence on the cluster using a biological dataset [13]. In 2021, Mohamad Basel Al Sawaf et al. worked on a recurrent neural network model that will be able to analyze relationships between different hydrological variables related to flood in order to predict flooding [14]. Amir Hossein Rasekh et al. presented a model that can create descriptions of programming codes that have a relation to associated text documents [15]. Those descriptions can help a user to understand the program logically with little effort.

3 Methodology

This research applied the principle of pipe and filter software architectural design [10]. As like in this architecture the original data flow through different modules or filters and in each filter data gets modified and passed through the next one. Finally, the system produces the desired output at the end of the flow.

At first two topics will be provided to the system along with a number of preprocessed documents in order to discover hidden relations. The system will look for relations using the algorithm the can express both topics. Finally, all the findings will be presented as output. As an example, consider the task stated in [20], the first task that pioneered LBD in the biological domain - determining the relationship between fish oil and Raynaud's disease and how they are related.

The system contains three main phases – The MetaMap processing phase (will process and extract documents), the preparation phase (will generate S2C – Source to Concept and fetch the title and abstract), and the output phase (will find relations using Swanson's ABC model). Findings from the output module will be presented using a Graphical User Interface (GUI). This will show the generated concept chains and relevant evidence. Figure 1 presents the data flow of the system.

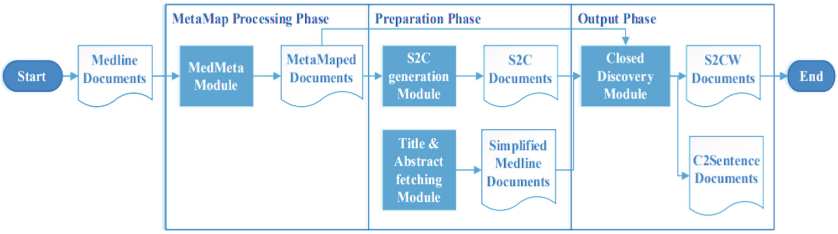


Fig. 1. Data flow diagram

4 Data Collection and Analysis

The Medline database contains a very large number of biomedical data-related citations and those were released by NLM on their official website. In this research, altogether nearly 22380000 citations were used to find hidden relations between two different topics. Here, each citation is an individual document. The Medline database includes a number of attributes such as ID, Publishing date, title, and abstract. For the Hidden Relationship Discovery title and abstract was used to find common links among topics.

4.1 Data Extraction from the Source

Data were extracted from the following link - ftp://ftp.ncbi.nlm.nih.gov/pubmed/baseline and XML tags as < PMID >, < PubDate >, < ArticleTitle >, and < AbstractText > were used to identify useful information.

4.2 MetaMap Module – Processing Phase

The functional structure of the MedMeta module is divided into four major steps. They are:

a. Extracts useful information from the data source – Medline documents.
b. Inserts MetaMap API in the code and sends titles and abstracts to the MetaMap Server.

c. Reads from MetaMap output and builds indexes of Medline documents based on concept occurrence relationship.
d. Write index files in XML format.

The MedMeta Module consists of seven modules – MedMeta, XMLParser, MedlineCite, MetaMapProc, PostProc, OutputXML, and Thread. The sequence is presented in Fig. 2.

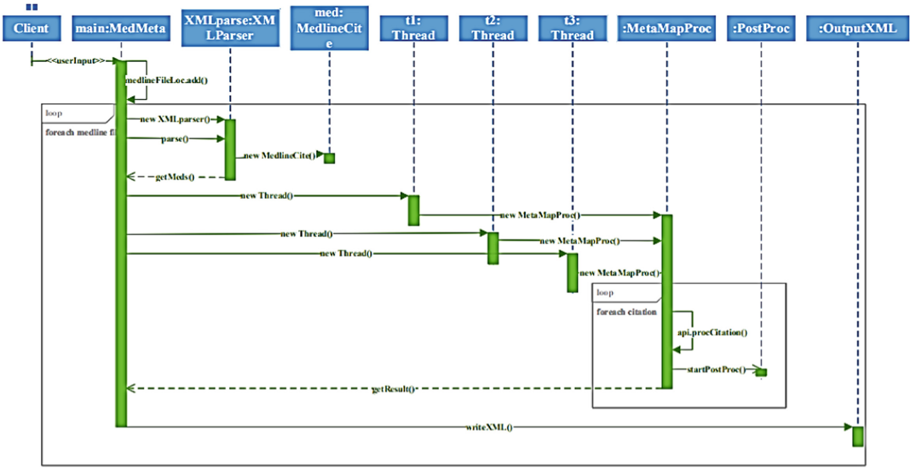


Fig. 2. Sequence diagram for MedMeta module

MedMeta's primary object is to regulate the creation of other modules and XMLParser can process oversized XML files. In this research total data used was 93.7 GB in total, where a single file was around 200 MB. After the parsing, XMLParser stored data into MedlineCite (a placeholder module to store data) objects. MetaMapProc will communicate with server and build index to initiate post-processing in PostProc. The PostProc is divided into 3 levels of data structure. The first level is Semantic type (holds one or many Concepts objects), the second one is Concept type (holds one or many Occurrence objects) and the third one is Occurrence type (holds the entity information). OutputXML uses an XML file that will hold the data. The thread module will apply parallel processing to improve performance.

4.3 MetaMap Module – Preparation Phase

In the preparation phase, a simplified S2C document will be produced using semantic and concept type data structure to reduce access time complexity. During the data collection phase, OutputXML generated a big amount of data and those were converted to 2 level S2C to reduce size. According to Fig. 3, S2C generation module consists of four modules – S2CGenerator (forwards the document through XMLParser), XMLParser (storing data in the server), SemanticType (uniquely leveling each hash to hold the S2C relationship), and WriteXMLFile (writes relationships in the XML file). Here, S2CGenerator is responsible for forwarding the document through XMLParser. There are 133 semantic types in total and each of them holds a hash set of concepts.

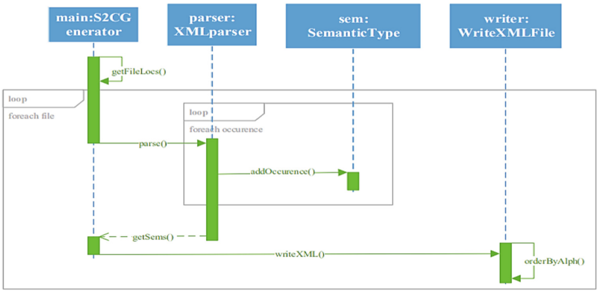


Fig. 3. Sequence diagram for S2C generation module

4.4 Title and Abstract Fetching Module

Aim of this module is to fetch the title and abstract from the original document in order to reduce the access time complexity. Figure 4 shows the title and abstract fetching module.

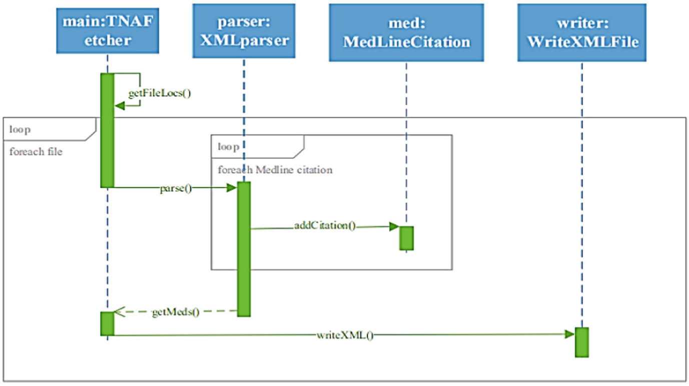
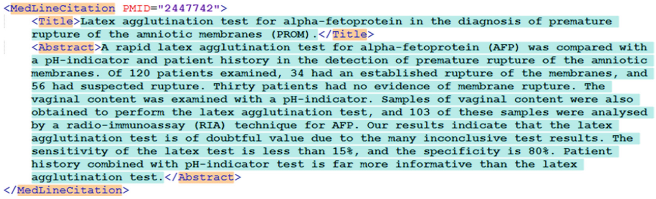


Fig. 4. Sequence diagram for title & abstract fetching module

The Title and Abstract fetching module consist of four modules. Three of them are TNAFetcher, XMLParser, and WriteXMLFile and their design are the same as their counterpart in the S2C generation module. Besides, the fourth one is MedLineCitation and it has two data fields – Title and Abstract, where data fetched from XML documents gets stored. Figure 5 shows a simplified Medline Document.



```
<MedLineCitation PMID="2447742">
    <Title>Latex agglutination test for alpha-fetoprotein in the diagnosis of premature
    rupture of the amniotic membranes (PROM).</Title>
    <Abstract>A rapid latex agglutination test for alpha-fetoprotein (AFP) was compared with
    a pH-indicator and patient history in the detection of premature rupture of the amniotic
    membranes. Of 120 patients examined, 34 had an established rupture of the membranes, and
    56 had suspected rupture. Thirty patients had no evidence of membrane rupture. The
    vaginal content was examined with a pH-indicator. Samples of vaginal content were also
    obtained to perform the latex agglutination test, and 103 of these samples were analysed
    by a radio-immunoassay (RIA) technique for AFP. Our results indicate that the latex
    agglutination test is of doubtful value due to the many inconclusive test results. The
    sensitivity of the latex test is less than 15%, and the specificity is 80%. Patient
    history combined with pH-indicator test is far more informative than the latex
    agglutination test.</Abstract>
</MedLineCitation>
```

Fig. 5. A simplified medline document.

4.5 Closed Discovery Module

This module uses user inputs, MetaMaped documents, S2C documents, and simplified Medline documents as inputs and after processing generates S2CW files and displays outputs through a Graphical User Interface. Figure 8 shows the Sequence diagram of the Closed Discovery module (Fig. 6).

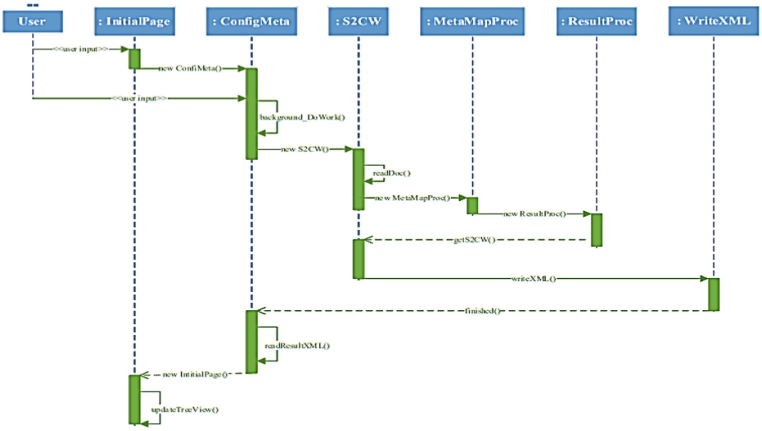


Fig. 6. Sequence diagram of closed discovery module.

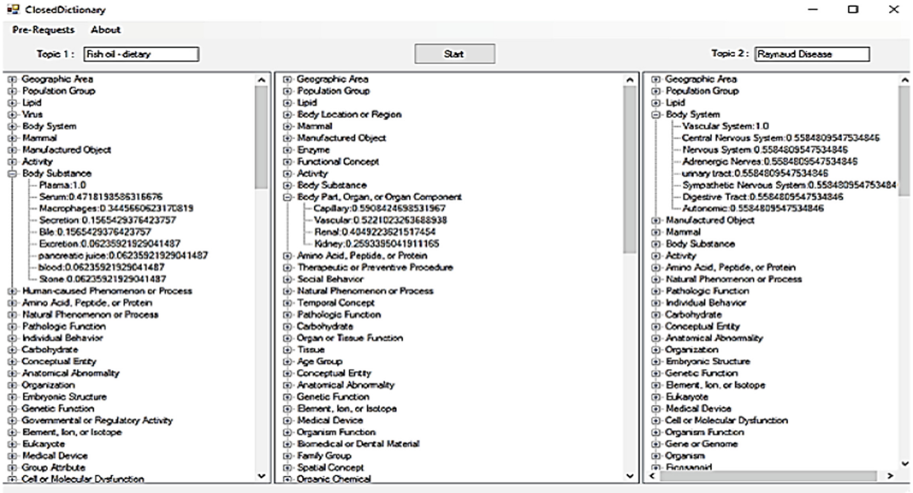
The ClosedDiscovery Module consists of eight modules – InitialPage, ConfigMeta, S2CW, MetaMapProc, ResultProc, WriteXML, SentPage, and MergedSentPage. SentPage and MergedSentPage are not included in Fig. 8, but they are used to display additional graphical user interfaces. Here, the InitialPage of GUI takes user inputs "Topic 1" and "Topic 2" and through ConfigMeta user can select target data. ConfigMeta is also responsible for building XML documents to store sentences related to Topic 1 and Topic 2. S2CW passes the XML document to MetaMapProc, which works as a gateway to the server that return findings. ResultProc uses those findings to build concept chains by calculating Term Frequency (TF) and Inverse Document Frequency (IDF). The IDF measures the importance of the concept and is measured for concept c by –

$$IDF(c) = log_e(Total\ number\ of\ sentences\,/\,Number\ of\ sentences\ with\ concept\ c\ in\ it) \tag{1}$$

The assumption is that the rarer a concept appears in the sentences, the higher the IDF value is, meaning the more important the concept is to the context. Then, the ResultProc multiplies TF and IDF to get concept weights (i.e. $\text{Weight}_c = \text{TF}_c * \text{IDF}_c$). Each MetaMap may contain several concepts and ResultProc analyze them to build C2W (Concept to Weight) and C2Sents (Concept to Sentence) relationship. Together they create S2CW (Semantic to Concept Weight) relationship, based on which the weight of each concept is further normalized by:

$$Normalized\ Weight(c) = \frac{the\ weight\ of\ the\ concept\ c}{maximum\ weight\ in\ this\ Semantic\ Type} \tag{2}$$

After the normalization data will be arranged in the descending order and by merging S2CW for topic 1 (S2CWA) and S2CW for topic 2 (S2CWC) intermediate level linking (common things between Topic 1 and 2) S2CWB will be generated. At this point, WriteXML will build XML files for S2CWA, S2CWB, S2CWC, C2SentsA, and C2SentsC using DocumentBuilderFactory Java API. SentPage is responsible for viewing all sentences related to a certain concept and MergedSentPage will list out all sentences related to concept chains A-B and B-C. Figure 7 represents the GUI which takes the input from the user and visualize the output and Fig. 8 represents linking sentences between topic 1 and 2.



Topic 1 : Fish oil - dietary

Topic 2 : Raynaud Disease

```
⊞ Geographic Area
⊞ Population Group
⊞ Lipid
⊞ Virus
⊞ Body System
⊞ Mammal
⊞ Manufactured Object
⊞ Activity
⊟ Body Substance
    Plasma:1.0
    Serum:0.4718193586316676
    Macrophages:0.3445660623170819
    Secretion 0.1565429376423757
    Bile 0.1565429376423757
    Excretion 0.06235921929041487
    pancreatic juice:0.06235921929041487
    blood:0.06235921929041487
    Stone 0.06235921929041487
⊞ Human-caused Phenomenon or Process
⊞ Amino Acid, Peptide, or Protein
⊞ Natural Phenomenon or Process
⊞ Pathologic Function
⊞ Individual Behavior
⊞ Carbohydrate
⊞ Conceptual Entity
⊞ Anatomical Abnormality
⊞ Organization
⊞ Embryonic Structure
⊞ Genetic Function
⊞ Governmental or Regulatory Activity
⊞ Element, Ion, or Isotope
⊞ Eukaryote
⊞ Medical Device
⊞ Group Attribute
⊞ Cell or Molecular Dysfunction
```

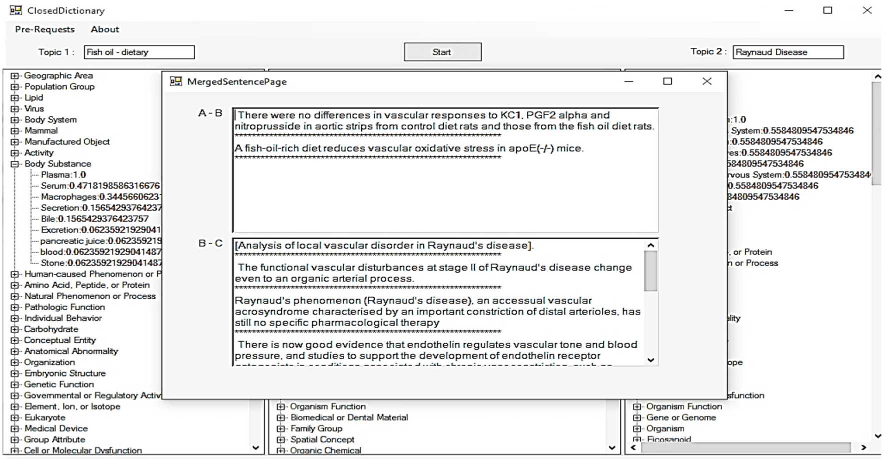
```
⊞ Geographic Area
⊞ Population Group
⊞ Lipid
⊞ Body Location or Region
⊞ Mammal
⊞ Manufactured Object
⊞ Enzyme
⊞ Functional Concept
⊞ Activity
⊞ Body Substance
⊟ Body Part, Organ, or Organ Component
    Capillary:0.5908424698531967
    Vascular:0.5221023263688938
    Renal:0.4049223621517454
    Kidney:0.2593395041911165
⊞ Amino Acid, Peptide, or Protein
⊞ Therapeutic or Preventive Procedure
⊞ Social Behavior
⊞ Natural Phenomenon or Process
⊞ Temporal Concept
⊞ Pathologic Function
⊞ Carbohydrate
⊞ Organ or Tissue Function
⊞ Tissue
⊞ Age Group
⊞ Conceptual Entity
⊞ Anatomical Abnormality
⊞ Genetic Function
⊞ Element, Ion, or Isotope
⊞ Medical Device
⊞ Organism Function
⊞ Biomedical or Dental Material
⊞ Family Group
⊞ Spatial Concept
⊞ Organic Chemical
```

```
⊞ Geographic Area
⊞ Population Group
⊞ Lipid
⊟ Body System
    Vascular System:1.0
    Central Nervous System:0.5584809547534846
    Nervous System 0.5584809547534846
    Adrenergic Nerves:0.5584809547534846
    urinary tract:0.5584809547534846
    Sympathetic Nervous System:0.558480954753484
    Digestive Tract:0.5584809547534846
    Autonomic:0.5584809547534846
⊞ Manufactured Object
⊞ Mammal
⊞ Body Substance
⊞ Activity
⊞ Amino Acid, Peptide, or Protein
⊞ Natural Phenomenon or Process
⊞ Pathologic Function
⊞ Individual Behavior
⊞ Carbohydrate
⊞ Conceptual Entity
⊞ Anatomical Abnormality
⊞ Organization
⊞ Embryonic Structure
⊞ Genetic Function
⊞ Element, Ion, or Isotope
⊞ Eukaryote
⊞ Medical Device
⊞ Cell or Molecular Dysfunction
⊞ Organism Function
⊞ Gene or Genome
⊞ Organism
⊞ Eicosanoid
```

Fig. 7. GUI for visualizing relationship between topics 1 and 2

5 System Evaluation

During the development of MedMeta module a multi-thread program (3 threads) was used to communicate with the MetaMap server instead of single thread to evaluate system performance. An experiment was conducted to find the impact of threads against



Topic 1 : Fish oil - dietary

Topic 2 : Raynaud Disease

A - B

There were no differences in vascular responses to KC1, PGF2 alpha and nitroprusside in aortic strips from control diet rats and those from the fish oil diet rats.
****

A fish-oil-rich diet reduces vascular oxidative stress in apoE(-/-) mice.
****

B - C

[Analysis of local vascular disorder in Raynaud's disease].
****

The functional vascular disturbances at stage II of Raynaud's disease change even to an organic arterial process.
****

Raynaud's phenomenon (Raynaud's disease), an accessual vascular acrosyndrome characterised by an important constriction of distal arterioles, has still no specific pharmacological therapy
****

There is now good evidence that endothelin regulates vascular tone and blood pressure, and studies to support the development of endothelin receptor

Fig. 8. GUI for sentences that are creating the relationship

Table 1. Performance comparison

Number of threads	Processing time	Average memory usage	Average CPU usage	Performance compare to single thread	Performance (%)
1	145 s	98 MB	25%	1	
2	87 s	110 MB	40%	1.67	85%
3	62 s	128 MB	73%	2.34	78%

performance with a system that has eight-core i7-4790k CPU, 16GB RAM and Windows Pro 10 x64 OS. The number of threads varies from one to three (Table 1).

According to the table it is clearly visible that with the increase of the number of threads performance increases greatly, where for 1 thread it took 145 s and for 3 threads it took 62 s. Moreover, 3 threads achieve 2.34 times improved performance over a single thread. It shows, in case of hardware limitations three-threads is a preferred option.

6 Result Evaluation

In order to justify the accuracy of the model, we used the widely adopted evaluation queries in the related literature, e.g., comparison was made between this research and Gopalakrishnan and Kishlay's study [9] using the gold standard query pair "Fish-oil" and "Raynaud's Disease". Gopalakrishnan and Kishlay had found connecting words such as platelet aggregation, vascular reactivity, blood viscosity and Prostaglandin. Our model not only finds all these impronat connecting terms, but also assigns each a high rank in its associated semantic types. Table 2 presents our partial result of evaluating the above query.

Table 2. Evaluating the query pair: fish oil and raynaud disease

Connecting concepts	Find?	Weight	Semantic type rank
Platelet aggregation	True	0.87	2
Vascular reactivity	True	1.27	1
Prostaglandin	True	0.309	6
Arthritis rheumatoid	True, it appears as “Rheumatoid Arthritis”	0.43	2

Moreover, our model also found other important linking terms such as “Hemodynamic” and “Atherosclerosis”. However, these links were not detected by them [9].

Another query was made for the pair “Schizophrenia” and “Phospholipase A2”. Gopalakrishnan and Kishlay found linking terms as chlorpromazine, receptors dopamine, prolactin, arachidonic acid, phenothiazines, and norepinephrine [9]. Table 3 shows how our framework behaves in this case. This model was able to find relevant data as them and again this system found interesting linking terms, such as “PGE2”, which [9] could not detect.

Table 3. Evaluating the query pair: Schizophrenia and Phospholipase A2

Connecting concepts	Find?	Weight	Rank in the semantic type
Chlorpromazine	True	0.22	11
Receptors dopamine	True	0.33	5
Prolactin	True	1.17	2
Arachidonic acid	True	1.18	1
Norepinephrine	True	0.3	11

For the standard query pair Migraine and Magnesium, Gopalakrishnan and Kishlay had found connecting words such as propranolol, adenosine triphosphate, calcium, ergotamine, serotonin, norepinephrine, adenine nucleotides, and epinephrine [9]. Table 4 shows how our framework behaves.

Table 4 indicates that the system has also found all those eight connecting concepts found in Gopalakrishnan and Kishlay’s model. Our system also found some other linking terms that were not found in the previous study, such as “Insulin”.

Table 4. Evaluating the query pair: Migraine and Magnesium

Connecting concept	Find?	Weight	Rank in the semantic type
Propranolol	True	0.41	8
Adenosine triphosphate	True	0.52	10
Calcium	True	1.50	1
Ergotamine	True, the system found "Ergot Alkaloids"	0.34	5
Serotonin	True	1.49	1
Norepinephrine	True	0.79	6
Adenine nucleotides	True	0.32	14
Epinephrine	True	0.49	13

7 Conclusion and Future Work

The goal of this research is to find the hidden connections between concepts of interest. This will assist researchers in the early discovery of hypotheses that may lead to important findings by initiating a deep understanding of the hidden information in the biomedical sector. Experiments show that the developed model can find meaningful logical connections between topics of interest and can visualize discovered hypotheses in a user-friendly way. In future work, we will expand the chains to multiple levels, which we call concept graph queries, and MetaMaped files and other additional domain-specific resources will be combined for easier access and more comprehensive knowledge discovery.

References

1. Belkin, N.J.: Interaction with texts: Information retrieval as information seeking behavior. In: Information Retrieval. p. 55–66 (1993). 10.1.1.50.6725
2. Swanson, D.R.: Complementary structures in disjoint science literatures. In: Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, Chicago, IL, pp. 280–289 (1991). https://doi.org/10.1145/122860.122889
3. Aronson, A.R.: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: Proceedings of AMIA Annual Symposium, pp. 17–21 (2001). https://pubmed.ncbi.nlm.nih.gov/11825149/
4. Kay Deeney. MetaMap - A Tool for Recognizing UMLS Concepts in Text. U.S. National Library of Medicine (2017). https://metamap.nlm.nih.gov/
5. Chapman, W.W., Fiszman, M., , Dowling, J.N., Chapman, B.E., Rindflesch, T.C.: Identifying respiratory findings in emergency department reports for biosurveillance using MetaMap. Studies in Health Technology and Informatics, **107**(Pt 1), pp. 487–91 (2004). https://pubmed.ncbi.nlm.nih.gov/15360860/
6. Zuccon, G., Holloway, A., Koopman , B., Nguyen, A.: Identify disorders in health records using conditional random fields and metamap. In: Proceedings of the CLEF 2013 Workshop on Cross-Language Evaluation of Methods, Applications, and Resources for eHealth Document Analysis, pp. 1–8 (2013). https://eprints.qut.edu.au/62875/

7. Pratt, W., Yetisgen-Yildiz, M.: A study of biomedical concept identification: MetaMap vs. people. In: AMIA Annual Symposium Proceedings, pp. 529–33 (2003). https://pubmed.ncbi.nlm.nih.gov/14728229/
8. Jin, W., Srihari, R.K.: Knowledge discovery across documents through concept chain queries. In: Proceedings of the Sixth IEEE International Conference on Data Mining – Workshops (ICDMW'06), pp. 448–452 (2006). https://doi.org/10.1109/ICDMW.2006.105
9. Gopalakrishnan, V., Jha, K., Jin, W., Zhang, A.: A survey on literature based discovery approaches in biomedical domain. In: Journal of Biomedical Informatics, **93**, 103141 (2019). doi: https://doi.org/10.1016/j.jbi.2019.103141
10. Philipps, J., Rumpe, B.: Refinement of pipe-and-filter architectures. In: Wing, J.M., Woodcock, J., Davies, J. (eds.) FM 1999. LNCS, vol. 1708, pp. 96–115. Springer, Heidelberg (1999). https://doi.org/10.1007/3-540-48119-2_8
11. Sanscartier, M.J., Neufeld, E.: Identifying hidden variables from contextspecific independencies. In: Proceedings of the Twentieth International Florida Artificial Intelligence Research Society Conference, pp. 472–477 (2007). 10.1.1.329.7687, Florida, USA
12. Prakash, D., Surendran, S.: Detection and analysis of hidden activities in social networks. International Journal of Computer Applications (0975–8887), **77**(16), 34–38 (2013). https://doi.org/10.5120/13570-1404
13. Pividori, M., Cernadas, A., de Haro, L.A., Carrari, F., Stegmayer, G., Milone, D.H.: Clustermatch: discovering hidden relations in highly diverse kinds of qualitative and quantitative data without standardization. Bioinformatics **35**(11), 1931–1939 (2019). https://doi.org/10.1093/bioinformatics/bty899
14. Sawaf, M.B.A., Kawanisi, K., Jlilati, M.N., Xiao, C., Bahreinimotlagh, M.: Extent of detection of hidden relationships among different hydrological variables during floods using data-driven models. Environ. Monit. Assess. **193**(11), 1–14 (2021). https://doi.org/10.1007/s10661-021-09499-9
15. Rasekh, A.H., Arshia, A.H., Fakhrahmad, S.M., Sadreddini, M.H.: Mining and discovery of hidden relationships between software source codes and related textual documents. Digital Scholarship in the Humanities. **33**(3), 651–669 (2018). https://doi.org/10.1093/llc/fqx052
16. Gopalakrishnan, V., Jha, K., Zhang, A., Jin, W.: Generating hypothesis: Using global and local features in graph to discover new knowledge from medical literature. In: Proceedings of the 8th International Conference on Bioinformatics and Computational Biology, Las Vegas, Nevada, USA. pp. 23–30 (2016). 978–1–943436–03–3
17. Hu, X., Zhang, X., Yoo, I., Zhang, Y.: A semantic approach for mining hidden links from complementary and non-interactive biomedical literature. In: Proceedings of the Sixth SIAM International Conference on Data Mining, Bethesda, MD, USA, pp. 200–209 (2006). https://doi.org/10.1137/1.9781611972764.18
18. Srinivasan, P., Libbus, B.: Mining MEDLINE for implicit links between dietary substances and diseases. In Bioinformatics. **20**, i290–i296 (2004). https://doi.org/10.1093/bioinformatics/bth914
19. Jha, K., Jin, W.: Mining novel knowledge from biomedical literature using statistical measures and domain knowledge. In: Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (BCB '16). Association for Computing Machinery, New York, NY, USA, pp. 317–326 (2016). https://doi.org/10.1145/2975167.2975200
20. Swanson, D.R.: Fish oil, raynaud's syndrome, and undiscovered public knowledge. Perspect. Biol. Med. **30**(1), 7–18 (1986). https://doi.org/10.1353/pbm.1986.0087