Operationalizing Semantic Medline for meeting the information needs at point of care

Majid Rastegar-Mojarad, MS, Dingcheng Li, PhD, Hongfang Liu, PhD
Biomedical Statistics & Informatics, Mayo Clinic, Rochester, MN

Abstract

Scientific literature is one of the popular resources for providing decision support at point of care. It is highly desirable to bring the most relevant literature to support the evidence-based clinical decision making process. Motivated by the recent advance in semantically enhanced information retrieval, we have developed a system, which aims to bring semantically enriched literature, Semantic Medline, to meet the information needs at point of care. This study reports our work towards operationalizing the system for real time use. We demonstrate that the migration of a relational database implementation to a NoSQL (Not only SQL) implementation significantly improves the performance and makes the use of Semantic Medline at point of care decision support possible.

Introduction

Clinical Decision Support (CDS) systems aim to provide useful information to clinicians to enhance health and health care[1, 2]. This information should be available and accessible at point of care [3]. One of the common CDS resources is literature such as Medline that has potential to use in evidence-based medicine at point of care. As the volume of literature continuously expands, it is not possible for clinicians to update their medical knowledge [4]. During a physician’s lifetime, medical knowledge is increased by four-fold [5]. Here is the place that CDS systems could assist clinicians to make a better evidence-based judgment at point of care via retrieving relevant publications to their questions [6]. The systems should be able to filter unrelated publications and retrieve highly related studies. The current state of the art in natural language processing (NLP) incorporates semantics into information retrieval [3] that could be used in this problem. For example, Semantic Medline [7] utilizes NLP to summarize Medline citations and analyze salient content in titles and abstracts. There are multiple systems [8]–[10] quite successful in retrieving relevant publications to clinicians’ queries but failed to be operationalized at point of care due to the long response time.

Ask Mayo Expert (AME) [11] is a web-based system operationalized at Mayo clinic to provide vetted evidence-based clinical decision support at point of care. However, the content covered by AME is limited and about 30% of the queries entered by clinicians are null search, i.e., zero hits returned. Previously, our group developed a system to retrieve sentences from Semantic Medline to support clinicians’ information needs [8]. After processing a clinician’s query, the system retrieved related Medline abstracts via NCBI E-utils [12]. A MySQL implementation of Semantic Medline Database (SemMedDB) [7] used to extract relevant predications from the retrieved documents. The system was promising but the response time was not acceptable at point of care. To address this issue, we propose the use of document-based data store and search strategy for operationalizing the system. Specifically, we integrate all required resources into a schema-less and document-based store. Each document contains a sentence from Semantic Medline, predicates, and metadata about the publication. ElasticSearch (ES), a text search engine with the capability of running on a distributed environment, is used in our system to index and retrieve data. In the following, we provide necessary background information. The proposed system is described next. We then present our experiment demonstrating the applicability of the system as a point of care tool.

Background

In the big data era, there is a need for a database system that is able to store and retrieve huge amount of data efficiently and able to handle variations in data format. Built upon the advance in storing, exploring, and analyzing unstructured data, the movement in “Not only SQL” (NoSQL) has convinced data scientists that relational database (RDB) is not the best fit in the big data era. Like any other domains, there is a huge amount of data generated in healthcare such as clinical notes, literature, and etc. that have been used in a variety of biomedical applications [13], [14]. To make the data efficiently accessible, there is a need to adopt the big data technology. For example, in evidence-based medicine, one of the valuable resources could be Semantic Medline, semantically enhanced with predicates extracted by SemRep [15] from Medline titles and abstracts, that contains approximately 70 million semantic predicates. Currently, the predicates are stored in a relational database called Semantic Medline Database (SemMedDB). It has been used by many researchers to facilitate knowledge discovery. For example, Cairelli et al. implemented a system[16] that transferred Semantic Medline predications into an interactive graph of semantic
predications. Zhang et al. [17] applied Semantic Medline predications to find drug-drug interactions in clinical text. Workman and Stoddart [18] proposed to use Semantic Medline as a potential decision support system for point of care. Their system provided a graphic interface of semantic predications relevant to user’s information need. Previously, we also have built a system to handle null search for a point of care tool, Ask Mayo Expert (AME), but failed to be operationalized due to the long response time. In this project, we present our work on utilizing the big data technology to operationalize the use of Semantic Medline at point of care.

Before describing the application, we like to discuss some of RDB’s features and consider their usage in SemMedDB. One of RDB’s advantages is supporting transaction and providing ACID (Atomicity, Consistency, Isolation, Durability) compliance. Nevertheless, is it required for SemMedDB? As SemMedDB is semi-static, meaning most of SQL queries on SemMedDB are Select queries, so ACID compliance isn’t required. Of course, time to time there are bunch of Inserts (for new publications) but they do not jeopardize the database consistency. Does increasing data size affect RDB performance? The size of SemMedDB keeps growing over the years and the performance of RDB will decrease. Is using traditional relational database efficient to develop a text search engine on the top of this repository? There are several tools that implemented for this purpose and they could be used as database and search engines such as: Lucene, Indri, Solr, and Elasticsearch (that all are based on Lucene[19]).

**Methods**

Our goal is to retrieve relevant sentences from scientific articles for answering clinicians’ questions at point of care. The query could be as simple as a single word, phrase, or sentence. In our previous systems [8]–[10], we proposed several approaches to retrieve relevant sentences and rank them based on different measures such as tf-idf, Journal information, publication type, and etc.

There are multiple steps in the previous system[8] as shown in the top part of Figure 1. The first step, query processing, involved: tokenization, lexical normalization, UMLS Metathesaurus look-up, and concept screening and Medical Subject Heading (MeSH) conversion. After processing and expanding user’s query, the system used NCBI E-utils to retrieve relevant Medline abstracts. E-utils returned PubMed identifier (PMID) and meta-data of relevant abstracts. In the next step, the system queried SemMedDB to retrieve sentences that appeared in these abstracts containing at least one semantic predication. With respect to the SemMedDB design, the system needed to join two tables that one of them has more than 17 million rows and the other one, more than 143 million rows. Then the system ranked the retrieved sentences. In the new system, we integrated all needed resources into one place and instead of using a relational database, one of the common search engines, ElasticSearch (ES), is utilized to store and search the data. ES is a search engine built upon Apache Lucene and supports distributed implementation. To implement the new system, we first downloaded needed resources including Metadata for Medline abstracts (retrieved from PubMed), predication information from SemMedDB, and SCImago journal and country ranking information [20]. After downloading the resources, we indexed and stored them in ES. As ES is a document-based search engine, we first formed documents and then indexed them. Unlike RDB, which requires a carefully defined schema, a document in ES contains a record or tuple without a predefined schema. A document in ES is equivalent to a row in relational database. Each document in our index contains: sentence, abstract’s metadata, and journal’s rank information. Figure 2 illustrates the original source of each field in ES’s document. After building the index, we were able to query the index and retrieve relevant sentences to user query. As mentioned earlier, the query could be one word, multiple words, or sentence. The system followed the same method, we used in the previous system[8], to process and expand the query. ES retrieved and ranked relevant sentences. As we integrated Medline abstracts...
metadata and SemMedDB data, the system did not need to query two resources. Like the previous system, we uses publication type and journal score to rank the retrieved sentences (more detail about the ranking method in [9], [10]). Integrating the ranking and searching is one of the advantages of using ES. The bottom part of Figure 1 illustrates the architecture of our new implementation.

Results

We implemented our system in Java and compared response time for 2,750 queries for two approaches. These queries were asked by clinicians with no hits in AME. To get the top five relevant sentences for each of these queries, the new approach took 5 minutes and 19 seconds total. The average response time for each query was about 116 milliseconds with the median response time less than 100 milliseconds. Table 1 shows ten of these queries, response time, and three top relevant sentences retrieved by the new system. In addition, the response time for the previous system is mentioned. Figure 3 illustrates the distribution of response time for the queries.

Discussion

Previously, our group developed several systems to retrieve relevant publications automatically for expert-written content [8], [10]. In the previous systems, we focused on finding and ranking the most relevant sentences to user’s query. The main bottleneck for operationalizing the previous system [8] was the response time. Our motivation for
this work was to take advantage of the big data technology and make our system applicable at point of care by shortening the response time. We migrated all needed resources from RDB to NoSQL environment. The new design decreased the system response time significantly and made the system applicable at point of care. Of course, we sacrificed some of RDB features such as minimum redundancy. We should add that for updating the data on the regular basis, the migration to the NoSQL environment does not affect the process, as Semantic Medline is updated every three months.

There are several limitations of this study. 1) We did not evaluate quality of responses in this study because our focus was the response time with no change of basic retrieval and ranking algorithms. 2) Another approach to integrate all the resources is, ignoring normalization in RDB and populating one table in RDB with all the resources. We did not compare our system with this design, because RDB is not as powerful as ES in searching text. 3) We did not evaluate different type of metrics to rank retrieved sentences, as we reported them previously.

**Conclusion**

This study builds upon our previous work where we showed that using Semantic Medline could return better answers to clinicians who search for evidence. To make the system applicable at point of care, we migrated Semantic Medline to big data environment. Comparing our new system with the old one illustrated a significant decrease in response time.

Table 1: A comparison of response time for a randomly selected ten queries. The top three relevant sentences retrieved by the new system are also presented.

| Query                                      | Time (milliseconds) | Top three relevant sentences                                                                 |
|--------------------------------------------|---------------------|---------------------------------------------------------------------------------------------|
| is barrett esophagus a precursor to cancer | 69745               | 1) Esophageal adenocarcinoma (EAC) is the most rapidly increasing cancer in the Western world and Barrett’s esophagus (BE) is the only known precursor lesion for this lethal cancer.  
2) Patients experiencing gastroesophageal reflux may be predisposed to developing Barrett’s esophagus, which is thought to be a precursor for the development of esophageal cancer.  
3) Barrett’s esophagus (BE) is the only established precursor lesion in the development of esophageal adenocarcinoma (EAC) and it increases the risk of cancer by 11-fold. |
| Left atrial enlargement                     | 318201              | 1) left atrial enlargement is more pronounced in mitral insufficiency;  
2) Left atrial enlargement and right ventricular hypertrophy in essential hypertension.  
3) HYPOTHESIS: Airway collapse is independent of left atrial enlargement. |
| c. difficile                                | -                   | 1) We measured airborne and environmental C. difficile adjacent to patients with symptomatic C. difficile infection (CDI).  
2) C. difficile infection can cause serious complications and death.  
3) Genomes of individual strains of C. difficile are highly divergent. |
| mri and titanium                            | 131139              | 1) Effects of new titanium cerebral aneurysm clips on MRI and CT images.  
2) Comparative MRI compatibility of 316 L stainless steel alloy and nickel-titanium alloy stents.  
3) The purpose of this study is to assess the presence and extent of artifacts seen on postoperative MRI scans in patients with titanium spinal implants. |
| sexually transmitted infections (stis)     | 66714               | 1) Young adults have high rates of sexually transmitted infections (STIs).  
2) BACKGROUND: Incidence of sexually transmitted infections (STIs) among young people in the United Kingdom is increasing.  
3) BACKGROUND: Improved treatment of sexually transmitted infections (STIs) is associated with decreased HIV incidence. |
| urinary tract infections, uncomplicated cystitis | -                 | 1) Acute uncomplicated cystitis (AUC) is one of the most common bacterial urinary tract infections.  
2) Acute uncomplicated cystitis and acute uncomplicated pyelonephritis are two frequently encountered urinary tract infections (UTI) in premenopausal, healthy females.  
3) Acute uncomplicated cystitis (AUC) and acute uncomplicated pyelonephritis (AUP) are two common urinary tract infections (UTI) in otherwise healthy young women. |
| afib stroke risk                            | 67600               | 1) Nonrheumatic atrial fibrillation (AFib) is the most potent common risk factor for stroke, raising the risk of stroke 5-fold.  
2) Atrial fibrillation increases the risk of stroke.  
3) atrial fibrillation also increases the risk of stroke. |
| pancreatic function test                    | 64810               | 1) Limits of the evocative pancreatic function test in the diagnosis of low-grade pancreatitis.  
2) Endoscopic pancreatic function test using combined secretin and cholecystokinin stimulation for the evaluation of chronic pancreatitis.  
3) The pancreatic function test at the Gastro-intestinal Clinic, Groote Schuur Hospital—a historical perspective |
| reactive airway disease                     | 405413              | 1) Montelukast does not prevent reactive airway disease in young children hospitalized for RSV bronchiolitis.  
2) Reactive airway disease in patients with prolonged exposure to industrial solvents.32)  
3) An association of GER with "awake apnea," reactive airway disease, and recurrent pneumonia has
been demonstrated.

1) CER1 gene variations associated with bone mineral density, bone markers, and early menopause in postmenopausal women.
2) Effects of menopause on bone mineral density in women with endemic fluorosis.
3) Do lifestyle choices explain the effect of alcohol on bone mineral density in women around menopause?

Acknowledgments
The authors would like to thank Soheil Moosavinasab and Siddhartha Jonnalagadda for providing us the implementation of their systems.

References
[1] E. H. Shortliffe, “Computer programs to support clinical decision making,” JAMA, vol. 258, no. 1, pp. 61–66, Jul. 1987.
[2] E. Berner, “Clinical Decision Support Systems: State of the Art,” Rockville, Maryland: Agency for Healthcare Research and Quality., AHRQ Publication No. 09-0069-EF., Jun. 2009.
[3] D. Demner-Fushman, W. W. Chapman, and C. J. McDonald, “What can natural language processing do for clinical decision support?,” J Biomed Inform, vol. 42, no. 5, pp. 760–772, Oct. 2009.
[4] J. W. Williamson, P. S. German, R. Weiss, E. A. Skinner, and F. Bowes, “Health science information management and continuing education of physicians. A survey of U.S. primary care practitioners and their opinion leaders,” Ann. Intern. Med., vol. 110, no. 2, pp. 151–160, Jan. 1989.
[5] R. Smith, “What clinical information do doctors need?,” BMJ, vol. 313, no. 7064, pp. 1062–1068, Oct. 1996.
[6] J. W. Ely, J. A. Osheroff, S. M. Maviglia, and M. E. Rosenbaum, “Patient-care questions that physicians are unable to answer,” J Am Med Inform Assoc, vol. 14, no. 4, pp. 407–414, Aug. 2007.
[7] S. R. Jonnalagadda, G. Del Fiol, R. Medlin, C. Weir, M. Fiszman, J. Mostafa, and H. Liu, “Automatically extracting sentences from Medline citations to support clinicians’ information needs,” J Am Med Inform Assoc, vol. 20, no. 5, pp. 995–1000, Oct. 2013.
[8] S. Moosavinasab, M. Rastegar-Mojarad, hongfang Liu, and S. Jonnalagadda, “Towards Transforming Expert-based Content to Evidence-based Content,” presented at the AMIA Clinical Research Informatics Summit, 2014.
[9] D.C. Li, H. Liu, C. G. Chute, and S. Jonnalagadda, “Towards assigning references using semantic, journal and citation relevance,” in Bioinformatics and Biomedicine (BIBM), 2013.
[10] D. A. Cook, F. Enders, J. A. Linderbaum, D. Zwart, and F. J. Lloyd, “Speed and accuracy of a point of care web-based knowledge resource for clinicians: a controlled crossover trial,” Interact J Med Res, vol. 3, no. 1, p. e7, 2014.
[11] Entrez Programming Utilities Help. National Center for Biotechnology Information (US), 2010.
[12] H. Kilicoglu, D. Shin, M. Fiszman, G. Rosemblat, and T. C. Rindflesch, “SemMedDB: a PubMed-scale repository of biomedical semantic predications,” Bioinformatics, vol. 28, no. 23, pp. 3158–3160, Dec. 2012.
[13] M. Rastegar-Mojarad, M. E. Bales, and H. Yu, “researchermap: a tool for visualizing author locations using Google maps,” Stud Health Technol Inform, vol. 192, p. 1187, 2013.
[14] S. Hebbring, M. Rastegar-Mojarad, Z. Ye, J. Mayer, C. Jacobson, and S. Lin, “Application of Clinical Text Data for Phenome-Wide Association Studies (PheWASs),” presented at The American Society of Human Genetics, 2014.
[15] T. C. Rindflesch and M. Fiszman, “The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text,” J Biomed Inform, vol. 36, no. 6, pp. 462–477, Dec. 2003.
[16] M. J. Cairelli, C. M. Miller, M. Fiszman, T. E. Workman, and T. C. Rindflesch, “Semantic MEDLINE for discovery browsing: using semantic predications and the literature-based discovery paradigm to elucidate a mechanism for the obesity paradox,” AMIA Annu Symp Proc, vol. 2013, pp. 164–173, 2013.
[17] R. Zhang, M. J. Cairelli, M. Fiszman, G. Rosemblat, H. Kilicoglu, T. C. Rindflesch, S. V. Pakhomov, and G. B. Melton, “Using semantic predications to uncover drug-drug interactions in clinical data,” J Biomed Inform, vol. 49, pp. 134–147, Jun. 2014.
[18] T. E. Workman and J. M. Stoddart, “Rethinking information delivery: using a natural language processing application for point-of-care data discovery,” J Med Libr Assoc, vol. 100, no. 2, pp. 113–120, Apr. 2012.
[19] “Lucene.” http://lucene.apache.org/.
[20] “SCImago Journal & Country Rank.” http://www.scimagojr.com/.