Journal of Biomedical Informatics 57 (2015) 436–445

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Automatically finding relevant citations for clinical guideline development

Duy Duc An Bui [a,b,*], Siddhartha Jonnalagadda [b], Guilherme Del Fiol [a]

[a] Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA
[b] Department of Preventive Medicine-Health and Biomedical Informatics, Northwestern University, Chicago, IL, USA

ARTICLE INFO

Article history:
Received 2 June 2015
Revised 28 August 2015
Accepted 2 September 2015
Available online 10 September 2015

Keywords:
Information retrieval
PubMed
Practice guideline
Medical subject headings
Natural language processing

ABSTRACT

Objective: Literature database search is a crucial step in the development of clinical practice guidelines and systematic reviews. In the age of information technology, the process of literature search is still conducted manually, therefore it is costly, slow and subject to human errors. In this research, we sought to improve the traditional search approach using innovative query expansion and citation ranking approaches.

Methods: We developed a citation retrieval system composed of query expansion and citation ranking methods. The methods are unsupervised and easily integrated over the PubMed search engine. To validate the system, we developed a gold standard consisting of citations that were systematically searched and screened to support the development of cardiovascular clinical practice guidelines. The expansion and ranking methods were evaluated separately and compared with baseline approaches.

Results: Compared with the baseline PubMed expansion, the query expansion algorithm improved recall (80.2% vs. 51.5%) with small loss on precision (0.4% vs. 0.6%). The algorithm could find all citations used to support a larger number of guideline recommendations than the baseline approach (64.5% vs. 37.2%, $p < 0.001$). In addition, the citation ranking approach performed better than PubMed's "most recent" ranking (average precision +6.5%, recall@k +21.1%, $p < 0.001$), PubMed's rank by "relevance" (average precision +6.1%, recall@k +14.8%, $p < 0.001$), and the machine learning classifier that identifies scientifically sound studies from MEDLINE citations (average precision +4.9%, recall@k +4.2%, $p < 0.001$).

Conclusions: Our unsupervised query expansion and ranking techniques are more flexible and effective than PubMed's default search engine behavior and the machine learning classifier. Automated citation finding is promising to augment the traditional literature search.

1. Introduction

The practice of evidence based medicine requires integrating individual clinical expertise and the best available evidence in making decisions about patient care. However, health care practitioners have little time to keep up with the rapid growth in the biomedical literature. In 2009, there were about 25,400 peer-reviewed journals and the number increases 3.5% a year [1]. Citations indexed in PubMed have grown from 4 million (pre 1975) to 22 million today [2]. Each year, about 3000 clinical trial studies have posted results in ClinicalTrial.gov [3]. Fraser and Dunstan showed that it's almost impossible to keep up with the medical literature even within a narrow specialty [4]. In a review of information-seeking behavior, Davies showed that clinicians' lack of time, issues with information technology, limited search skills are top barriers for information searching [5]. As a result, most clinical questions raised by clinicians at the point of care remain unanswered. In a recent systematic review, Del Fiol et al. showed that clinicians raised roughly one question out of every two patients seen and over 60% of these questions were not answered [6]. To cope with information overload, clinicians rely on existing expert-compiled resources such as clinical practice guidelines (CPG) to fulfill their information needs [7]. However, the development and update of CPGs is costly, slow and unable to keep up with the rate of new evidence in the medical literature. In a 2003 survey of guideline developers, the average cost for CPGs development was $200,000 per guideline in the United States [8]. High quality guidelines that meet strict quality criteria [9,10] require more time and resources. Time required for finishing peer-review for a

* Corresponding author at: University of Utah, Biomedical Informatics, 421 Wakara Way, Ste 140, Salt Lake City, UT 84108-3514, USA. Tel.: +1 (801) 581 4080; fax: +1 (801) 581 4297.
E-mail addresses: duy.bui@utah.edu, bdaduy@gmail.com (D.D.A. Bui).

http://dx.doi.org/10.1016/j.jbi.2015.09.003
1532-0464/

cardiology guideline published by The American College of Cardiology (ACC) and American Heart Association (AHA) was from 12 to 18 months [11]. In summary, the rapid pace of new published literature can quickly make the CPGs outdated and suboptimal for clinical decision-making.

In guideline development, experts perform systematic reviews of the available evidence, which involves a series of scientifically rigorous steps [11]. The two first and important steps are a systematic literature search followed by screening for relevant citations. Literature search involves identifying possibly relevant studies from electronic literature databases. Citation screening involves quickly scanning abstract and full-text manuscripts to assess the eligibility of studies. Informatics research has investigated automated and semi-automated methods to aid with citation screening [12–16]. Fiszman et al. were among the first research groups introducing informatics solutions to support clinical guideline development [15,16]. They developed a semantic filter to automatically classify relevant citations. Similarly, Cohen et al. investigated a machine learning approach to solve a classification problem in drug effectiveness reviews [12,17]. To meet the needs of citation screening, those methods aimed for a balance between recall and precision. However, recall is more important than precision in systematic literature search. The 2011 ACCF/AHA's manual for clinical guideline development described the need for literature search to be comprehensive, and key to the development of valid guidelines [11]. The Cochrane handbook for systematic reviews highlights that "searches should seek high sensitivity, which may result in relatively low precision." [18] In the present study, we investigated the literature search stage and aimed to maximize recall while controlling the impact on precision. We developed and assessed query expansion and ranking methods to enhance information retrieval performance in the context of clinical guideline development. The solution was based on an extension of PubMed's search engine, optimized to retrieve and rank relevant studies for cardiovascular guidelines.

There have been previous works that we leveraged to inform our system [15–17,19–21]. Fiszman's gold standard included citations that were used to support 30 clinical questions [16]. Our work sought for a larger gold standard, which includes citations to support more than 600 guideline recommendations. Research on query expansion showed that using MeSH concepts and MeSH hierarchy can improve performance of image retrieval and biological question retrieval [19,20]. Our query expansion method was also based on finding relevant MeSH concepts, but was optimized to retrieve guideline conditions.

Traditional information retrieval or question answering systems rank documents by relevance or similarity to the user query. Generic queries (e.g., "heart failure") can generate thousands of documents that share the search keywords. PubMed by default sorts the results by recently added date, without considering relevancy and scientific quality. Informatics research has investigated machine learning approaches to prioritize citation screening in systematic reviews [14,22,23]. Yet, machine-learning approaches are arguably not flexible since they require sufficient high-quality training data and often do not generalize well to new domains. Unsupervised ranking methods have been investigated in the citation retrieval studies by Jonnalagadda et al. [24,25]. Their method assigned weights based on journal impact measures; however, the method validation was limited to the "heart failure" topic. In the present research, we developed novel unsupervised query expansion and citation ranking methods with a larger gold standard that includes cardiovascular conditions. We then compared the performance of these methods with PubMed's query expansion and ranking, and a machine learning classifier.

2. Materials and methods

Our study design consisted of three main parts: (1) development of a gold standard composed of studies used in the development of cardiovascular guidelines; (2) iterative development of a citation finding system composed of two main components: query expansion and citation ranking; and (3) evaluation of each system component using standard information retrieval metrics and comparison with baseline approaches. Fig. 1 depicts the summarization of our system architecture and study design.

2.1. Gold standard

The gold standard consisted of citations that have been used to support guideline practice recommendations. We focused on the cardiovascular guidelines published by the American College of Cardiology (ACC) and the American Heart Association (AHA). The full revision cardiovascular guidelines developed by the ACC/AHA and published from 2010 to 2014 were retrieved using a PubMed search. Since the majority of guideline topics are about complete management of a condition, we focused on retrieving condition topics in this study. Topics about interventions or diagnostic procedures are reserved for future research. For those guidelines discussing the comprehensive management of cardiovascular conditions, we performed the following steps to build the gold standard: (1) Extracted all the citations listed in the "References" section of the guideline; (2) extracted the guideline recommendations whose evidence sources were provided in the guideline and the citations that were used as evidence sources to support each recommendation; and (3) automatically mapped those citations in free-text to PubMed IDs using the NCBI Batch Citation Matcher tool [26]. Manual mapping was performed to supplement the citation IDs that could not be matched by the NCBI tool. Table 1 shows examples of guideline recommendations, supporting citations, and their corresponding PMIDs.

2.2. System overview

The system is an extension of PubMed's search engine to enhance the ability to retrieve citations for clinical guideline development. The system has a preprocessing stage and two other main stages: query expansion and document ranking. The query expansion stage aims to improve recall while the document ranking aims to improve precision on top-ranked documents.

2.2.1. Preprocessing

This step takes the title of the guideline as input and extracts the conditions of interest. Since there is little variation among guideline titles, we used simple regular expression rules such as words following "Patients With", "diagnosis and treatment of", and "management of" to extract main conditions from guideline titles (e.g. "Guideline for the Management of Patients With Atrial Fibrillation", "Guideline for the diagnosis and treatment of hypertrophic cardiomyopathy", "Guidelines for the diagnosis and management of patients with Thoracic Aortic Disease"). This step also detects whether a particular guideline focuses on one or more conditions. For instance, the phrase "Extracranial Carotid and Vertebral Artery Disease" was broken into two conditions: "Extracranial Carotid Disease" and "Vertebral Artery Disease".

2.2.2. Query expansion

Based on the extracted condition terms, we conducted a search using PubMed's default search behavior. When entering a query on the PubMed search interface, PubMed automatically expands the query to maximize recall. For instance, PubMed expands the query