Journal of Biomedical Informatics 46 (2013) 1060–1067

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

eTACTS: A method for dynamically filtering clinical trial search results ☆

Riccardo Miotto [a,1], Silis Jiang [a,1], Chunhua Weng [a,b,*]

[a] Department of Biomedical Informatics, Columbia University, New York, NY 10032, United States
[b] The Irving Institute for Clinical and Translational Research, Columbia University, New York, NY 10032, United States

ARTICLE INFO

Article history:
Received 21 May 2013
Accepted 27 July 2013
Available online 3 August 2013

Keywords:
Information storage and retrieval
Clinical trials
Dynamic information filtering
Interactive information retrieval
Tag cloud
Association rules

ABSTRACT

Objective: Information overload is a significant problem facing online clinical trial searchers. We present eTACTS, a novel interactive retrieval framework using common eligibility tags to dynamically filter clinical trial search results.

Materials and methods: eTACTS mines frequent eligibility tags from free-text clinical trial eligibility criteria and uses these tags for trial indexing. After an initial search, eTACTS presents to the user a tag cloud representing the current results. When the user selects a tag, eTACTS retains only those trials containing that tag in their eligibility criteria and generates a new cloud based on tag frequency and co-occurrences in the remaining trials. The user can then select a new tag or unselect a previous tag. The process iterates until a manageable number of trials is returned. We evaluated eTACTS in terms of filtering efficiency, diversity of the search results, and user eligibility to the filtered trials using both qualitative and quantitative methods.

Results: eTACTS (1) rapidly reduced search results from over a thousand trials to ten; (2) highlighted trials that are generally not top-ranked by conventional search engines; and (3) retrieved a greater number of suitable trials than existing search engines.

Discussion: eTACTS enables intuitive clinical trial searches by indexing eligibility criteria with effective tags. User evaluation was limited to one case study and a small group of evaluators due to the long duration of the experiment. Although a larger-scale evaluation could be conducted, this feasibility study demonstrated significant advantages of eTACTS over existing clinical trial search engines.

Conclusion: A dynamic eligibility tag cloud can potentially enhance state-of-the-art clinical trial search engines by allowing intuitive and efficient filtering of the search result space.

1. Introduction

Randomized controlled trials generate high-quality medical evidence for disease treatment and therapeutic development but still face longstanding recruitment problems. In fact, more than 90% of trials are delayed because of difficulties recruiting eligible patients [1–3]. Using Web applications, health consumers are becoming increasingly comfortable searching online for clinical research opportunities [4]. However, information overload is a common and significant problem with most existing clinical trial search engines (e.g., ClinicalTrials.gov [5], UK Clinical Trials Gateway [6]). For example, searching "diabetes mellitus, type II" on ClinicalTrials.gov returns a list of more than 5,000 trials (as of April 2013), which are sorted just by their probabilistic relevance to the search terms, with those containing the query in the title ranked highest [7]. Supplying additional parameters, such as location or study type, can only modestly improve search specificity, especially for searches of eligibility criteria. Moreover, identifying terms that are effective at retrieving relevant trials can be difficult for the average user [8].

One major limitation of existing clinical trial search engines is the underutilization of free-text eligibility criteria. This is mostly due to varied and complicated semantic structures (e.g., inclusion vs. exclusion and negation) that make it difficult to define standardized parsers as well as user-friendly representations to exploit in search applications [9–11]. Yet, we hypothesize that filtering clinical trials by eligibility criteria can greatly increase the specificity of the search engines.

1.1. Objective

This article presents eTACTS (eligibility TAg cloud-based Clinical Trial Search), a faceted search method to filter the list of clinical trials returned by any type of initial search (e.g., simple free-text query terms, advanced form-based). In particular, the resulting

* Corresponding author. Address: Department of Biomedical Informatics, Columbia University, 622 W 168th Street, VC-5, New York, NY 10032, United States. Fax: +1 (212) 305-3302.
E-mail address: cw2384@columbia.edu (C. Weng).

[1] These authors contributed equally to this work.

1532-0464/$ - see front matter © 2013 The Authors. Published by Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.07.014

trials are indexed through a small number of facets, each defining a distinct property of the text, and users can select facets to filter the search results [12–15]. eTACTS defines eligibility tags as facets for the clinical trial search results. An eligibility tag is a meaningful multi-word pattern, e.g., "breast carcinoma", "active malignancy", that frequently appears within the free-text eligibility criteria of clinical trials [16]. Eligibility tags are presented to users as a dynamic tag cloud to assist with iterative filtering of the resulting trials. A tag cloud is a visual representation of key concepts associated with textual documents. In this domain, individual tags are displayed as hyperlinks to a set of clinical trials that contain the tags in their eligibility criteria, with each tag's "importance" or relative frequency indicated by a mix of font size and color. When the user selects a tag, the cloud is updated according to the tag distribution in the remaining trials, which contain all of the selected tags in their eligibility criteria. By using common tags, we allow the users to quickly identify common and intuitive facets that lead to efficient and effective result filtering [16].

In this paper, we (1) describe the design of a novel interactive clinical trial search framework named eTACTS; (2) demonstrate that a dynamic tag cloud can efficiently reduce the trial search results based on interactive search parameters expressed by eligibility tags; (3) demonstrate that eTACTS helps users discover trials not highlighted by conventional search engines; and (4) demonstrate that searching by eTACTS effectively produce more relevant results than other available search engines.

1.2. Related work

Prior studies proposed automatic techniques to transform clinical trial specifications into a computable form that can be efficiently reused for classification, clustering, and retrieval [17–22]. A number of efforts also focused on formally representing free-text clinical trial eligibility criteria for computational processing [10,16,23–27]. Consequently, several projects are underway to improve clinical trial recruitment with Web-based information technologies [28–30]. These methods either help clinicians find relevant trials for their patients [31] or help patients identify trials themselves [5,6,32–38]. Some tools provide general search facilities that query public trial repositories (e.g., ClinicalTrials.gov [5], UK Clinical Trials Gateway [6], Search Clinical Trials [34], TrialReach [35], ASCOT [38]). Others employ user provided medical history to recommend suitable trials (e.g., PatientsLikeMe [33], Corengi for "type II diabetes" trials [37]) or match users with research coordinators (e.g., ResearchMatch [36]). Alternatively, TrialX employs a question/answer mechanism (i.e., AskDory!) to provide users with a list of actively-recruiting trials, whose recruiters users can then call to verify eligibility [32].

Most of these systems use only pre-structured information (e.g., condition, location, title) or limited manual annotations of the eligibility criteria for clinical trial searches. Only ASCOT [38] provides searches with discriminative power based on automatic processing of eligibility criteria. In particular, ASCOT annotates each clinical trial with the Unified Medical Language System (UMLS) [39] terms extracted from its eligibility criteria. The annotations related to the trials retrieved by an initial search are then displayed as a list beside the search results, allowing the user select those he/she considers effective at reducing the number of results. The most frequent annotations in the clinical trial repository are also provided as a static tag cloud (i.e., related neither to the initial search nor to the user interaction) to initially filter the results.

Interactive information retrieval has gained popularity lately [40] and presenting tag clouds has become a well-established data visualization technique [41–43]. While some criticisms have been raised by the Internet community about the use of tag clouds in general domain and social applications [44–46], they were effectively used as a data-driven aid for users searching and browsing pertinent information in more specific scenarios, e.g., to discern credible content in online health message forums [47], music [48] and image retrieval [49]. Our method differs from ASCOT in that the cloud of eligibility tags, which gets updated after each user tag selection, is the main filtering tool. Additionally, while ASCOT mines annotations from each trial independently, we use a controlled vocabulary composed only of frequent and common tags that are mined across multiple trials. This leads to a higher level and more intuitive representation designed to simplify searches and to help users interact with the search system.

2. Material and methods

The eTACTS framework consists of two components (see Fig. 1): (1) eligibility tag mining and clinical trial indexing, and (2) online tag cloud-based dynamic trial search. In the following sections, we present the main components of the proposed framework and the evaluation design. The detailed design and evaluation for unsupervised tag mining and eligibility criteria indexing were reported previously [16] and hence will only be briefly reviewed here.

2.1. Tag mining and eligibility criteria indexing

eTACTS automatically mines tags from the free-text eligibility criteria of a representative set of clinical trials. Text processing techniques are used to extract relevant *n*-grams from each criterion, where the *n*-gram relevance is defined by the grammatical role of the words, limited presence of stop words, and matching of at least one word with the UMLS lexicon. Terms that match the UMLS are also normalized into preferred UMLS terms. Only the most frequent *n*-grams of the collection are retained as potential tags. This set is then automatically polished—not-discriminative *n*-grams and irrelevant substrings are removed—to obtain the final controlled vocabulary of eligibility tags. At indexing time, each clinical trial available in the repository is annotated with only those tags extracted from their eligibility criteria.

Tags are mined and assigned to trials regardless of their role being inclusion or exclusion. In fact, with eTACTS, tags are meant to identify high-level concepts mentioned in the text rather than structured semantic patterns (e.g., "concept-X *greater than* N", "*not* concept-X") in eligibility criteria. While distinguishing between inclusion and exclusion roles can be useful with semantic patterns, it is not always useful for tags. For example, a tag appearing frequently in clinical trial eligibility criteria is "body mass index" (BMI), which is usually followed by a value (e.g., "inclusion: BMI > 40"). Without indexing the entire pattern, distinguishing BMI between inclusion/exclusion would be misleading in the filtering process. In fact, for example, a user could select "inclusion: BMI" aiming to find trials enrolling participants with a high BMI value; however, in this way, the user might miss those trials where the same concept is expressed as an exclusion criterion (e.g., "exclusion: BMI $\leqslant$ 40"). In contrast, tags related to medical conditions, e.g., "breast cancer", might benefit from identifying their role. Nevertheless, in this study we treated all tags identically because our objective was to assess the general feasibility of our approach, which is based on the hypothesis that natural language processing-based semantic pattern recognition and processing (which can be error-prone and lead to noisy representations) are not necessary for information filtering, which is our focus.

2.2. Tag cloud-based retrieval

The objective of tag-based retrieval is to refine the results of a simple search. The key feature of eTACTS is to present to users a