Journal of Biomedical Informatics xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Context-driven automatic subgraph creation for literature-based discovery

Delroy Cameron [a,*], Ramakanth Kavuluru [b], Thomas C. Rindflesch [c], Amit P. Sheth [a], Krishnaprasad Thirunarayan [a], Olivier Bodenreider [c]

[a] Ohio Center of Excellence in Knowledge-Enabled Computing (Kno.e.sis), Wright State University, Dayton, OH 45435, USA
[b] Division of Biomedical Informatics, University of Kentucky, Lexington, KY 40506, USA
[c] National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894, USA

ARTICLE INFO

Article history:
Received 16 September 2014
Accepted 25 January 2015
Available online xxxx

Keywords:
Literature-based discovery (LBD)
Graph mining
Path clustering
Hierarchical agglomerative clustering
Semantic Similarity
Semantic relatedness
Medical Subject Headings (MeSH)

ABSTRACT

Background: Literature-based discovery (LBD) is characterized by uncovering hidden associations in non-interacting scientific literature. Prior approaches to LBD include use of: (1) domain expertise and structured background knowledge to manually filter and explore the literature, (2) distributional statistics and graph-theoretic measures to rank interesting connections, and (3) heuristics to help eliminate spurious connections. However, manual approaches to LBD are not scalable and purely distributional approaches may not be sufficient to obtain insights into the meaning of poorly understood associations. While several graph-based approaches have the potential to elucidate associations, their effectiveness has not been fully demonstrated. A considerable degree of *a priori* knowledge, heuristics, and manual filtering is still required.

Objectives: In this paper we implement and evaluate a context-driven, automatic subgraph creation method that captures multifaceted complex associations between biomedical concepts to facilitate LBD. Given a pair of concepts, our method automatically generates a ranked list of subgraphs, which provide informative and potentially unknown associations between such concepts.

Methods: To generate subgraphs, the set of all MEDLINE articles that contain either of the two specified concepts (A, C) are first collected. Then binary relationships or assertions, which are automatically extracted from the MEDLINE articles, called *semantic predications*, are used to create a labeled directed *predications graph*. In this predications graph, a *path* is represented as a sequence of semantic predications. The hierarchical agglomerative clustering (HAC) algorithm is then applied to cluster paths that are bounded by the two concepts (A, C). HAC relies on implicit semantics captured through Medical Subject Heading (MeSH) descriptors, and explicit semantics from the MeSH hierarchy, for clustering. Paths that exceed a threshold of semantic relatedness are clustered into subgraphs based on their *shared context*. Finally, the automatically generated clusters are provided as a ranked list of subgraphs.

Results: The subgraphs generated using this approach facilitated the rediscovery of 8 out of 9 existing scientific discoveries. In particular, they directly (or indirectly) led to the recovery of several *intermediates* (or B-concepts) between A- and C-terms, while also providing insights into the meaning of the associations. Such meaning is derived from predicates between the concepts, as well as the provenance of the semantic predications in MEDLINE. Additionally, by generating subgraphs on different thematic dimensions (such as *Cellular Activity*, *Pharmaceutical Treatment* and *Tissue Function*), the approach may enable a broader understanding of the nature of complex associations between concepts. Finally, in a statistical evaluation to determine the *interestingness* of the subgraphs, it was observed that an arbitrary association is mentioned in only approximately 4 articles in MEDLINE on average.

Conclusion: These results suggest that leveraging the implicit and explicit semantics provided by manually assigned MeSH descriptors is an effective representation for capturing the underlying *context* of complex associations, along multiple thematic dimensions in LBD situations.

* Corresponding author. Tel.: +1 937 775 5213; fax: +1 937 775 5133.
E-mail address: delroy@knoesis.org (D. Cameron).

http://dx.doi.org/10.1016/j.jbi.2015.01.014
1532-0464/

1. Introduction

Literature-based discovery (LBD) refers to the process of uncovering hidden connections that are implicit in scientific literature. Numerous hypotheses have been generated from scientific literature, using the LBD paradigm, which influenced innovations in diagnosis, treatment, preventions, and overall public health. The notion of LBD was proposed by *Don R. Swanson* (1924–2012) in 1986, through the well-known *Raynaud Syndrome–Dietary Fish Oils* Hypothesis *(RS-DFO)* [1]. By reading the titles of more than 4000 MEDLINE articles, Swanson serendipitously discovered that *Dietary Fish Oils (DFO)* lower *Blood Viscosity*, reduce *Platelet Aggregation* and inhibit *Vascular Reactivity* (specifically *Vasoconstriction*). Concomitantly, he observed that a reduction in both *Blood Viscosity* and *Platelet Aggregation*, as well as the inhibition of *Vasoconstriction*, appeared to prevent *Raynaud Disease*; a circulatory disorder that causes periods of severely restricted blood flow to the fingers and toes [2]. Swanson therefore postulated that *"dietary fish oil might ameliorate or prevent Raynaud's syndrome."* This hypothesis was clinically confirmed by DiGiacomo et al. [3] in 1989.

Swanson's discovery is interesting because explicit associations between *DFO* and these intermediate concepts (i.e., *Blood Viscosity, Platelet Aggregation* and *Vasoconstriction*) had long existed in the literature [4–8]. Likewise, explicit associations between the intermediates and *RS* had been well documented [9,2]. The serendipity in Swanson's Hypothesis lies in the fact that no explicit associations linking *DFO* and *RS* directly had been previously articulated in a single document.

To develop this hypothesis, Swanson performed a Dialog® Sciserach using Raynaud and Fish Oil terms, on titles and abstracts of MEDLINE and Embase (Excepta Medica) citations, in November 1985. There were approximately 1000 articles in the Raynaud set and 3000 in the Fish Oil set. He found that only four articles among a reduced set of 489 articles (after filtering), contained cross-references spanning both sets. Among these four articles, only two articles [10,11] discussed relevant aspects of *RS* with *DFO*; although not in the context of Swanson's discovery. Swanson speculated that this phenomenon of logically related but noninteracting literatures alludes to the existence of *undiscovered public knowledge* [1]. Logically related information fragments may exist in the literature, but may have never been connected, or fully elucidated. He subsequently exploited his awareness of the existence of such undiscovered associations and investigated several other scenarios (three with Smalheiser [12–14]) that later led to new scientific discoveries [15,16]. Swanson grounded his observations in a paradigm now commonly known as the *ABC model* [1] for LBD, which is an integral part of LBD research, facilitating the generation of several hypotheses [1,15,16,12–14,17–25].

In current biomedical research, while finding unknown intermediates is an important task, domain scientists are often interested in developing a deeper understanding of causal relationships and mechanisms of interaction among concepts. For example, consider the complex scenario depicted in Fig. 1, in which *Dietary Fish Oils* produce several *Prostaglandins*, including *Prostaglandin I3* (PGI_3) and *Epoprostenol* (PGI_2, *the synthetic form of Prostacyclin*). The latter of these *Prostaglandins (Epoprostenol)* was known to treat *Raynaud Syndrome*. It was also known to disrupt *Platelet Aggregation*. Since *Platelet Aggregation* is deemed a cause of *Raynaud Syndrome*, one can reasonably conclude that a plausible mechanism by which *Dietary Fish Oils* treat *Raynaud Syndrome* is through the production of *Prostaglandins*, which actively disrupt *Platelet Aggregation*.

Aside from detecting such causal associations, it is known that complex associations may exist between concepts, in many different ways. For example, Fig. 2 shows that *Dietary Fish Oils* and *Raynaud Syndrome* are associated in at least the following three ways: (1) in terms of *Cellular Activity* involving *Blood platelets/Prostaglandins*, as shown in Fig. 2a, (2) through *Pharmaceuticals* that contain calcium channel blockers, such as *Nifedipine* and *Verapamil*, as shown in Fig. 2b, and (3) through *Lipids/Fatty Acids* from *Efamol* and *Evening primrose oil*, as shown in Fig. 2c.

In this paper, we build on our previous approach [26], in which we rediscovered and decomposed the *Raynaud Syndrome – Dietary Fish Oils* discovery. In our previous work, we manually created the multi-faceted subgraphs, by grouping together paths of *semantic predications*. Recall that a semantic predication is a binary relation between two concepts, expressed in the form (subject, predicate, object). Here, we present a method that uses rich representations to automatically create such subgraphs, by leveraging implicit and explicit semantics provided by MeSH descriptors.[1] To create the subgraphs, we first specify the context of a semantic predication and then use it to infer the context of a path. Paths are then clustered into coherent subgraphs on multiple thematic dimensions, based on their shared context.

The approach requires only three items from the user as input: (1) a list of concept labels for source (A) and target (C), (2) the maximum path length k of paths to be generated (default $k = 2$, for *ABC* associations), and (3) a cut-off date dt for articles to be included from the scientific literature. If no cut-off date is provided then all MEDLINE articles are used. The output of the approach is a ranked list of subgraphs S – i.e., create a function $\mathcal{F}$: $q \rightarrow S$, where $q = \{A, C, dt, k\}$.

To facilitate understanding the meaning of associations present in the subgraphs, the predicates of the semantic predications and their provenance in MEDLINE are provided (see Section 4). Relationships that are not explicit in the subgraphs, but are inferred, can be explored by composing MEDLINE queries (as we will show). The collective use of predicates, provenance and MEDLINE queries for knowledge exploration constitute the notion of *discovery browsing*, introduced by Wilkowski et al. [27] and extended by Cairelli et al. [28]. Discovery browsing is enabled when a system guides the user through their exploration of the literature in a process of cooperative reciprocity. The *"user iteratively focuses system output, thus controlling the large number of relationships often generated in literature-based discovery systems."*

To assess the efficacy of our approach, two forms of evaluation were conducted: (1) an evidence-based evaluation and (2) a statistical evaluation. The evidence-based evaluation showed that the generated subgraphs could facilitate the rediscovery of 8 out of 9 existing discoveries [1,15,16,12–14,29,30] (not recovered [28]). The statistical evaluation showed that an arbitrary association occurs only in approximately 4 articles in MEDLINE on average. This evaluation determines the *interestingness* of the subgraphs in general, as a way to assess whether a domain scientist might be interested in an arbitrary subgraph in the first place (see in Section 4.2). These results suggest that the subgraphs created using our approach provide an effective way of finding and elucidating poorly understood associations and may be of interest to domain scientists. In this paper we make the following specific contributions:

1. We develop a novel context-driven subgraph creation method for closed LBD (both A and C are known), capable of finding complex associations. Our approach is distinct from previous approaches, which are mainly based on statistical frequency, graph metrics, and specificity.

[1] MeSH is a controlled vocabulary (or thesaurus) of biomedical terms, organized in a hierarchical structure – https://www.nlm.nih.gov/mesh/.

Please cite this article in press as: Cameron D et al. Context-driven automatic subgraph creation for literature-based discovery. J Biomed Inform (2015), http://dx.doi.org/10.1016/j.jbi.2015.01.014