Journal of Informetrics 9 (2015) 686–703

Contents lists available at ScienceDirect

Journal of Informetrics

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

SemPathFinder: Semantic path analysis for discovering publicly unknown knowledge

Min Song[a,*], Go Eun Heo[a], Ying Ding[b]

[a] *Department of Library and Information Science, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul, Republic of Korea*
[b] *Department of Information and Library Science, Indiana University, Bloomington, IN, USA*

ARTICLE INFO

Article history:
Received 16 March 2015
Received in revised form 22 June 2015
Accepted 22 June 2015
Available online 1 September 2015

Keywords:
Literature based discovery
Named entity recognition
Relation extraction
Semantic path analysis
Semantic relatedness score

ABSTRACT

The enormous amount of biomedicine's natural-language texts creates a daunting challenge to discover novel and interesting patterns embedded in the text corpora that help biomedical professionals find new drugs and treatments. These patterns constitute entities such as genes, compounds, treatments, and side effects and their associations that spread across publications in different biomedical specialties. This paper proposes SemPathFinder to discover previously unknown relations in biomedical text. SemPathFinder overcomes the problems of Swanson's ABC model by using semantic path analysis to tell a story about plausible connections between biological terms. Storytelling-based semantic path analysis can be viewed as relation navigation for bio-entities that are semantically close to each other, and reveals insight into how a series of entity pairs is organized, and how it can be harnessed to explain seemingly unrelated connections. We apply SemPathFinder for two well-known use cases of Swanson's ABC model, and the experimental results show that SemPathFinder detects all intermediate terms except for one and also infers several interesting new hypotheses.

1. Introduction

In response to a surge in biomedical search, the demand of automated text processing methods has risen to deal with massive amount of text data. Text mining, discovering meaningful information from text data, facilitates efficient text processing and analysis through automated techniques. In particular, biomedical text mining, defined as the "Conceptual Biology" (Blagosklonny & Pardee, 2002), becomes a pivotal methodology to discover hidden biological meanings and treatments from the biomedical literature. Mining new knowledge from a meaningful relationship among concepts buried in published manuscripts helps prevent medical research from reinventing the wheel.

Literature-based discovery, such as Swanson's ABC model, which generates meaningful hypotheses by mining the biomedical literatures based on term co-occurrence and transitive relationship of documents, has proven very useful to discover hidden relations between two disparate fields by common intermediate terms that occur in both areas. In 1986, Swanson used word co-occurrence and citation analysis for the implicit linkage between Raynaud disease and fish oil (Swanson, 1986a). According to Swanson (Swanson, 1986b, 1989), publicly undiscovered knowledge exists because

[*] Corresponding author.
E-mail address: min.song@yonsei.ac.kr (M. Song).

http://dx.doi.org/10.1016/j.joi.2015.06.004

independently created fragments of information are not retrieved, interpreted, and studied together even if they are logically related. Swanson formalizes the procedure for identifying undiscovered public knowledge in the biomedical literature as follows: consider two separate literature sets, ALit and CLit, where the documents in CLit discuss concept C and the documents in ALit discuss concept A. Each of these two literature sets discusses the relationship of its main concept (A or C) with some intermediate concepts B, or bridge concepts. However, the possible connections between the bridge concepts are not discussed. Swanson's ABC model identifies relationships such as "A implies B" and "B implies C," then speculates that "A implies C." This derived knowledge that "A implies C" is not conclusive but hypothetical.

The usefulness of automatically generated hypotheses is verified through clinical research (DiGiacomo, Kremer, & Shah, 1989; Gallai et al., 1992). Since Swanson's ABC model, many researchers have conducted ABC model related studies. Previous studies of Swanson ABC model focused on extracting core entities by streamlining common intermediate terms. Those include statistical techniques, ontology-based concept mapping, and predicative extraction by dependency rules. Recently, graph-based approaches were applied to Swanson's ABC model to spot the relationship among entities.

However, existing approaches either require intensive manual processing or a semi-automatic procedure to find and select biomedical entities. In addition, they are limited to just one dimension—that is, the association relationship between two concepts—and network analysis is not feasible in most of existing approaches. Due to these issues, it is difficult to understand the whole structures, or to interpret the results systematically.

To overcome these challenges, we propose the new knowledge discovery system, SemPathFinder, to provide semantic path analysis that enables storytelling for plausible hypothesis generation. Storytelling-based semantic path analysis can be viewed as relation navigation for bio-entities that are semantically close to each other, and reveals insight into how a series of entity pairs is organized, and how it can be harnessed for explaining unexpected connections. With SemPathFinder, we demonstrate how to effectively discover various relationships among source and target concepts and their intermediate concepts by expanding intermediate concepts to multiple levels. In particular, this study makes use of two different types of path between two nodes: the shortest path and the cheapest path. The shortest path identifies the most cost-effective path, whereas the cheapest path is calculated by semantic relatedness among entities at different levels (the number of intermediate nodes is one, two, three, etc.).

In this paper we present an outline of related research; describe SemPathFinder; explain the results for two well-known cases of literature-based knowledge discovery and discuss the analysis of the experiments; and summarize our conclusions and suggest future work.

2. Related work

Several approaches have been proposed to overcome the limitations of Swanson's approach, which is based on co-occurrence. One key limitation of a simple co-occurrence-based approach is the identification of false positive connections. A second limitation is the coarse granularity of the identified relationships. Despite these limitations, co-occurrence is successfully used to generate networks including protein–protein interaction networks, gene–disease networks, and regulatory gene expression networks (Friedman, Kra, Yu, Krauthammer, & Rzhetsky, 2001; Šarić, Jensen, Ouzounova, Rojas, & Bork, 2006).

2.1. Ontology-based approach

Mukhopadhyay, Palakal, and Maddu (2010) made the text-based Apriori algorithm. They noticed that the method is to extract hyper graphs representing multiway associations among biological entities. Also, the hyper graphs with visualization provide information about gene–gene associations relevant to specific disease. Around the same time, Baker and Hemminger (2010) suggested a mining method for inferring the connections between chemical, protein, and disease from medical subject heading (MeSH) annotations using Swanson's ABC model. The researchers validated their method using time split published literature from PubMed. However, the method is limited: it only uses MeSH terms for the inference. Atkinson and Rivas (2008) manually curated sentence templates to infer cause–effect relations through a Bayesian network after processing texts with Natural Language Processing (NLP) techniques. The resulting relations were validated by medical experts. Coulet et al. (2011) applied the complex templates of sentence structure to extract biological relationships. They showed that the method can normalize and integrate heterogeneous relationships from the literature. To analyze adverse drug reactions (ADRs), Lin, Xiao, Huang, Chiu, and Soo (2010) computed the complexity of drug–drug target interactions by the number of common targets and the average clustering coefficient. Wang et al. (2011) proposed the Bio-LDA method to identify latent topics and uncover relationships among topics and biological terms.

Weeber, Klein, de Jong-van den Berg, and Vos (2001) proposed a two-step model of the discovery process in which hypotheses are generated and subsequently tested with Natural Language Processing techniques. They use the semantic information that is provided with these concepts as a powerful filter to successfully simulate Swanson's discoveries of connecting disjointed literature sets. Sun and Han (2013) proposed a meta-path based mining approach to detect significant paths based on predefined path patterns. Meta-paths systematically capture numerous semantic relationships in heterogeneous networks.