Literature-Based Discovery by an Enhanced Information Retrieval Model

Kazuhiro Seki[1] and Javed Mostafa[2]

[1] Kobe University, Hyogo 657-8501, Japan
seki@cs.kobe-u.ac.jp
[2] Indiana University, Bloomington, IN 47405, USA

Abstract. The massive, ever-growing literature in life science makes it increasingly difficult for individuals to grasp all the information relevant to their interests. Since even experts' knowledge is likely to be incomplete, important findings or associations among key concepts may remain unnoticed in the flood of information. This paper brings and extends a formal model from information retrieval in order to discover those implicit, hidden knowledge. Focusing on the biomedical domain, specifically, gene-disease associations, this paper demonstrates that our proposed model can identify not-yet-reported genetic associations and that the model can be enhanced by existing domain ontology.

Keywords: Hypothesis discovery, Text data mining, Inference network, Implicit association, Gene Ontology.

1 Introduction

With the advance of computer technologies, the amount of scientific knowledge is rapidly growing beyond the pace we could digest. For example, Medline[1]—the most comprehensive bibliographic database in life science—currently indexes over 17 million articles and the number keeps increasing by 1,500–3,000 per day. Given the substantial volume of the publications, it is virtually impossible to deal with the information without the aid of intelligent information processing techniques, such as information retrieval (IR), information extraction (IE), and text data mining (TDM).

In contrast to IR and IE, which find information explicitly stated in documents, TDM aims to discover heretofore unknown knowledge through an automatic analysis on textual data [1]. A pioneering work in TDM, also known as literature-based discovery, was conducted by Swanson in the 1980's. He argued that there were two premises logically connected but the connection had been unnoticed due to overwhelming publications and/or over-specialization. To demonstrate the validity of the basic idea, he manually analyzed numbers of articles and identified logical connections implying a hypothesis that fish oil was effective for clinical treatment of Raynaud's disease [2]. The hypothesis was later supported by experimental evidence.

[1] http://www.ncbi.nlm.nih.gov/sites/entrez?db=pubmed

V. Corruble, M. Takeda, and E. Suzuki (Eds.): DS 2007, LNAI 4755, pp. 185–196, 2007.

This study is motivated by Swanson's work and attempts to advance the research in literature-based discovery. Specifically, we target implicit associations between genes and hereditary diseases as a test bed. Gene-disease associations are the links between genetic variants and diseases to which the genetic variants influence the susceptibility. For example, BRCA1 is a human gene encoding a protein that suppresses tumor formation. A mutation of this gene increases a risk of breast cancer. Identification of these genetic associations has tremendous importance for prevention, prediction, and treatment of diseases. To this end, we develop a discovery framework by extending the models and techniques developed for IR. Furthermore, we propose the use of domain ontologies for more robust predictions. To demonstrate the effectiveness of the proposed framework, we conduct various evaluative experiments on realistic benchmark data.

2 Related Work

Over two decades, Swanson has argued the potential use of a literature to discover new knowledge that has *implicitly* existed for years but has not been noticed by anybody. His discovery framework is based on a syllogism; i.e., two premises, "A causes B" and "B causes C," suggest a potential association, "A causes C," where A and C do not have a known, explicit relation. Such an association can be seen as a hypothesis testable for verification to produce new knowledge, such as the above-mentioned association between Raynaud's disease and fish oil. For this particular example, Swanson manually inspected two sets of articles concerning Raynaud's disease and fish oil and identified premises that "Raynaud's disease is characterized by high platelet affregability, high blood viscosity, and vasoconstriction" and that "dietary fish oil reduces blood lipids, platelet affregability, blood viscosity, and vascular reactivity," which together suggest a potential benefit of fish oil for Raynaud's patients.

Based on the groundwork, Swanson himself and other researchers developed computer programs to aid hypothesis discovery. The following briefly introduces some of the representative studies.

Weeber et al. [3] implemented a system, called DAD-system, taking advantage of a natural language processing tool. The key feature of their system is that the Unified Medical Language System (UMLS) Metathesaurus[2] was incorporated for knowledge representation and pruning. While the previous work focused on words or phrases appearing in Medline records for reasoning, DAD-system maps them to a set of concepts defined in the UMLS Metathesaurus using MetaMap [4]. An advantage of using MetaMap is that it can automatically collapse different wordforms (e.g., inflections) and synonyms to a single concept. In addition, using *semantic types* (e.g., "Body location or region") under which each Metathesaurus concept is categorized, irrelevant concepts can be excluded from further exploration if particular semantic types of interest are given. This

[2] UMLS is an NLM's project to develop and distribute multi-purpose, electronic knowledge sources and its associated lexical programs.