Provenance and Knowledge Abstraction for Reachability: A Framework for Knowledge Discovery

Final Report

Student: Delroy Cameron
Mentor: Olivier Bodenreider

An intuitive preliminary step when developing techniques for Knowledge Discovery is to first evaluate the feasibility of your approach by observing its effectiveness when applied to the task of rediscovering already discovered knowledge. The fundamental premise is that techniques for rediscovery should demonstrate practicability when applied to the more challenging task of open knowledge discovery. As it relates to Don R. Swanson's scientific discoveries using biomedical literature, we adopted a systematic approach to rediscovering the 'Magnesium-Migraine' and 'Raynaud's Disease-Fish Oil' connections, using background knowledge as a key enabler. Our attempts at this task, led to striking yet perhaps obvious realization. Since the time of Swanson's discoveries in 1986, the number of direct links between 'Magnesium-Migraine' appearing in Biomedical Literature expressing the previously unknown connections is in fact rather overwhelming. Hence the task of rediscovering Swanson's discovery is now trivial. We contend that any attempt at such rediscovery using background knowledge, must necessarily freeze the knowledge realm to a period before and up to Swanson's discovery in 1986. In our case, to practically achieve this using the resources at NLM, would require restricting predications extracted from the literature to those pre-1986, and use a 1986-versioned UMLS Metathesaurus. Obtaining the relevant predications for this task is not straightforward, since SemRep has not processed predications outside the range (2000-2009).

Another scenario for observing the efficacy of techniques for knowledge discovery exists within the context of the Question-Answering challenges posed by the TREC Genomics Track. We exploit these Question-Answering scenarios to understand how well our semantics-based approaches can rediscover the documents pertaining to each question in the gold standard document set. Hence, by exploiting *Provenance* and the notion of *Knowledge Abstraction* (which we define formally) to establish *Reachability* among the documents pertaining to a question, we speculate that a generic framework for knowledge discovery can be derived. In particular, by extending the notion of vertex reachability to documents whose predications express the connections between concepts using labeled edges in a graph, we were able to establish reachability with a high reachability ratio (which we also define formally) on 10/26 questions in the TREC dataset. While this is by no means a representative sample of the entire dataset, our preliminary results are sufficiently encouraging to continue the collaboration between Kno.e.sis and NLM along this line of using provenance and knowledge abstraction for reachability with implications on knowledge discovery. In particular, the inclusion of Dr. Thomas Rindflesch will facilitate improved cumulative precision and recall through improvements in predication extraction using SemRep.

For the moment a direct outcome of this work is establishing a mirror of the BKR at Kno.e.sis. In the long term, we expect that at maturation, our analytics for reachability and knowledge discovery could be implemented in a live production system, used by biomedical researchers across the research spectrum. In the short term, we target the World Wide Web (WWW) 2011 Conference (deadline October 29, 2010) as a venue for submission, with the minimal expectation of processing all of the remaining questions and contrasting our approach to systems void of knowledge abstraction and comparable semantics. The full conference paper will be appended to this report as soon as it becomes available.