Semantic Search as Inference:
Applications in Health Informatics

Bevan Koopman[1,2]
[1]Australian e-Health Research Centre, CSIRO, Brisbane, Australia
[2]Queensland University of Technology, Brisbane, Australia
bevan.koopman@csiro.au

Abstract

In this thesis, we present models for semantic search: Information Retrieval (IR) models that elicit the meaning behind the words found in documents and queries rather than simply matching keywords. This is achieved by the integration of structured domain knowledge and data-driven information retrieval methods.

The research is set within health informatics to tackle the unique challenges within this domain; specifically, how to bridge the 'semantic gap'; that is, how to overcome the mismatch between raw medical data and the way human beings interpret it. Bridging the semantic gap involves addressing two issues: *semantics*; that is, aligning the meaning or concepts behind words found in documents and queries; and leveraging *inference*, which utilises semantics to infer relevant information.

Three semantic search models — all utilising concept-based rather than term-based representations — are developed; these include: the Bag-of-concepts model, which utilises concepts from the SNOMED CT medical ontology as its underlying representation; the Graph-based Concept Weighting model, which captures concept dependence and importance in a novel weighting function; and the core contribution of the thesis, the Graph INference model (GIN): a unified theoretical model of semantic search as inference, achieved by the integration of structured domain knowledge (ontologies) and statistical, information retrieval methods. It is the GIN that provides the necessary mechanism for *inference* to bridge the semantic gap. All three models are empirically evaluated using clinical queries and a real-world collection of clinical records taken from the TREC Medical Records Track (MedTrack).

Our evaluation shows that the use of concept-based representations in the Bag-of-concepts model leads to improved retrieval effectiveness. When concepts are combined within the Graph-based Concept Weighting model, further improvements are possible. The evaluation of GIN highlighted that its inference mechanism is suited to hard queries — those that perform poorly on a term-based system. In-depth analysis also revealed that the GIN returned many new documents not retrieved by term-based systems and therefore never evaluated for relevance as part of the TREC MedTrack. This highlights that using current IR test collections, where semantic search systems did not contribute to the pool, may underestimate the effectiveness of semantic search systems.

This work represents a significant step forward in the integration of structured domain knowledge and data-driven information retrieval methods. Furthermore, the thesis provides

an understanding of inference — when and how it should be applied for effective semantic search. It shows that queries with certain characteristics benefit from inference, while others do not. The detailed investigation into the evaluation of semantic search systems shows how current IR test collections may underestimate effectiveness of such systems and new techniques for evaluation are suggested. The Graph Inference model, although developed within the medical domain, is generally defined and has implications in other areas, including web search, where an emerging research trend is to utilise structured knowledge resources for more effective semantic search.

Supervisors: Peter Bruza, Laurianne Sitbon, Michael Lawley

Available: `http://koopman.id.au`