Chapter 2.21
Information Retrieval by Semantic Similarity

Angelos Hliaoutakis
Technical University of Crete (TUC), Greece

Giannis Varelas
Technical University of Crete (TUC), Greece

Epimenidis Voutsakis
Technical University of Crete (TUC), Greece

Euripides G. M. Petrakis
Technical University of Crete (TUC), Greece

Evangelos Milios
Dalhousie University, Canada

ABSTRACT

Semantic Similarity relates to computing the similarity between conceptually similar but not necessarily lexically similar terms. Typically, semantic similarity is computed by mapping terms to an ontology and by examining their relationships in that ontology. We investigate approaches to computing the semantic similarity between natural language terms (using WordNet as the underlying reference ontology) and between medical terms (using the MeSH ontology of medical and biomedical terms). The most popular semantic similarity methods are implemented and evaluated using WordNet and MeSH. Building upon semantic similarity, we propose the Semantic Similarity based Retrieval Model (SSRM), a novel information retrieval method capable for discovering similarities between documents containing conceptually similar terms. The most effective semantic similarity method is implemented into SSRM. SSRM has been applied in retrieval on OHSUMED (a standard TREC collection available on the Web). The experimental results demonstrated promising performance improvements over classic information retrieval

methods utilizing plain lexical matching (e.g., Vector Space Model) and also over state-of-the-art semantic similarity retrieval methods utilizing ontologies.

INTRODUCTION

Semantic Similarity relates to computing the similarity between concepts, which are not necessarily lexically similar. Semantic similarity aims at providing robust tools for standardizing the content and delivery of information across communicating information sources. This has long been recognized as a central problem in Semantic Web where related sources need to be linked and communicate information to each other. Semantic Web will also enable users to retrieve information in a more natural and intuitive way (as in a "query-answering" interaction).

In the existing Web, information is acquired from several disparate sources in several formats (mostly text) using different language terminologies. Interpreting the meaning of this information is left to the users. This task can be highly subjective and time consuming. To relate concepts or entities between different sources (the same as for answering user queries involving such concepts or entities), the concepts extracted from each source must be compared in terms of their meaning (i.e., semantically). Semantic similarity offers the means by which this goal can be realized.

This article deals with a certain aspect of Semantic Web and semantics, that of semantic text association, and text semantics respectively. We demonstrate that it is possible to approximate algorithmically the human notion of similarity using semantic similarity and to develop methods capable of detecting similarities between conceptually similar documents even when they don't contain lexically similar terms. The lack of common terms in two documents does not necessarily mean that the documents are not related. Computing text similarity by classical information retrieval models (e.g., Vector Space, Probabilistic, Boolean (Yates & Neto, 1999)) is based on lexical term matching. However, two terms can be semantically similar (e.g., can be synonyms or have similar meaning) although they are lexically different. Therefore, classical retrieval methods will fail to associate documents with semantically similar but lexically different terms.

In the context of the multimedia semantic Web, this article permits informal textual descriptions of multimedia content to be effectively used in retrieval, and obviates the need for generating structured metadata. Informal descriptions require significantly less human labor than structured descriptions.

In the first part of this article, we present a critical evaluation of several semantic similarity approaches for computing the semantic similarity between terms using two well-known taxonomic hierarchies namely WordNet[1] and MeSH[2]. WordNet is a controlled vocabulary and thesaurus offering a taxonomic hierarchy of natural language terms developed at Princeton University. MeSH (Medical Subject Heading) is a controlled vocabulary and a thesaurus developed by the U.S. National Library of Medicine (NLM)[3] offering a hierarchical categorization of medical terms. Similar results for MeSH haven't been reported before in the literature. All methods are implemented and integrated into a semantic similarity system, which is accessible on the Web.

In the second part of this article, we propose the "Semantic Similarity Retrieval Model" (*SSRM*). *SSRM* suggests discovering semantically similar terms in documents (e.g., between documents and queries) using general or application specific term taxonomies (e.g., WordNet or MeSH) and by associating such terms using semantic similarity methods. Initially, *SSRM* computes *tf·idf* weights to term representations of documents. These representations are then augmented by semantically similar terms (which are discovered from WordNet or MeSH by applying a range query in the neighborhood of each term in the taxonomy) and by

Related Content

A Stroke Information System (SIS): Critical Issues and Solutions

Subana Shanmuganathan (2010). *Biomedical Knowledge Management: Infrastructures and Processes for E-Health Systems (pp. 177-191).*

www.irma-international.org/chapter/stroke-information-system-sis/42606

Computer-Based Health Information Systems: Projects for Computerization or Health Management? Empirical Experiences from India

Ranjini C.R.and Sundeep Sahay (2009). *Medical Informatics: Concepts, Methodologies, Tools, and Applications (pp. 1265-1288).*

www.irma-international.org/chapter/computer-based-health-information-systems/26296

Mental Task Classification Using Deep Transfer Learning with Random Forest Classifier

Sapna Singh Kshatri, Deepak Singh, Mukesh Kumar Chandrakarand G. R. Sinha (2022). *International Journal of Biomedical and Clinical Engineering (pp. 1-17).*

www.irma-international.org/article/mental-task-classification-using-deep-transfer-learning-with-random-forest-classifier/301215

Finding Impact of Precedence based Critical Attributes in Kidney Dialysis Data Set using Clustering Technique

B.V. Ravindra, N. Sriraamand Geetha Maiya (2015). *International Journal of Biomedical and Clinical Engineering (pp. 44-50).*

www.irma-international.org/article/finding-impact-of-precedence-based-critical-attributes-in-kidney-dialysis-data-set-using-clustering-technique/136235

Neural Network Based Automated System for Diagnosis of Cervical Cancer

Seema Singh, V. Tejaswini, Rishya P. Murthyand Amit Mutgi (2015). *International Journal of Biomedical and Clinical Engineering (pp. 26-39).*

www.irma-international.org/article/neural-network-based-automated-system-for-diagnosis-of-cervical-cancer/138225