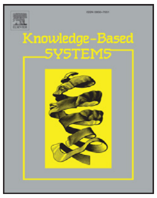
Knowledge-Based Systems 000 (2016) 1–18

Contents lists available at ScienceDirect

Knowledge-Based Systems

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



A semantic similarity measure for linked data: An information content-based approach

Rouzbeh Meymandpour*, Joseph G. Davis

School of Information Technologies, The University of Sydney, Sydney, Australia

ARTICLE INFO

Article history:
Received 11 November 2015
Revised 5 July 2016
Accepted 6 July 2016
Available online xxx

Keywords:
Semantic Web
Linked Data
Linked Open Data
Similarity measures
Semantic similarity
Information content
Ranking
Recommender systems
Collaborative filtering
Content-based filtering

ABSTRACT

Linked Data allows structured data to be published in a standard manner so that datasets from diverse domains can be interlinked. By leveraging Semantic Web standards and technologies, a growing amount of semantic content has been published on the Web as Linked Open Data (LOD). The LOD cloud has made available a large volume of structured data in a range of domains via liberal licenses. The semantic content of LOD in conjunction with the advanced searching and querying mechanisms provided by SPARQL has opened up unprecedented opportunities not only for enhancing existing applications, but also for developing new and innovative semantic applications. However, SPARQL is inadequate to deal with functionalities such as comparing, prioritizing, and ranking search results which are fundamental to applications such as recommendation provision, matchmaking, social network analysis, visualization, and data clustering. This paper addresses this problem by developing a systematic measurement model of semantic similarity between resources in Linked Data. By drawing extensively on a feature-based definition of Linked Data, it proposes a generalized information content-based approach that improves on previous methods which are typically restricted to specific knowledge representation models and less relevant in the context of Linked Data. It is validated and evaluated for measuring item similarity in recommender systems. The experimental evaluation of the proposed measure shows that our approach can outperform comparable recommender systems that use conventional similarity measures.

1. Introduction

The rapid development of Semantic Web technologies such as *resource description framework (RDF)* [1] have facilitated the publishing of structured data in a standard way that can be readily consumed and reused by machines and shared across diverse applications. This has transformed the conventional *Web of Documents* associated with Web 1.0, into the *Web of Data* (also referred to as *Linked Data*) – publishing and interlinking of structured data on the Web.

Linked Data can be private or public. It can be used inside organizations and enterprises, and shared among business partners to provide easier integration and to facilitate interoperability. Linked Data can also be open. *Linked Open Data (LOD)* is a recent community-driven effort that provides access to a large and increasing amount of diverse structured data using open Semantic Web standards and through liberal licenses [2]. The LOD cloud currently provides free access to 570 datasets[1] in areas such as media, geography, government, publications and life sciences. Using Semantic Web standards and LOD protocols (see Berners-Lee [3]), these datasets have been made publicly available for machine and human consumption. This data offers unprecedented opportunities for developing novel and innovative applications. It also makes semantic application development more efficient and cost-effective.

In order for Semantic Web-based applications to be able to systematically search, retrieve and analyze Linked Data, specific tools and technologies are required. Semantic Web crawlers and search engines are useful tools for browsing and searching semantic data (e.g. Swoogle search engine [4] and Semantic Web Search Engine [SWSE] [5]). However, the explicit graph matching-based approach of SPARQL, the main query language for the Semantic Web and Linked Date [6], is not able to address complex requirements such as prioritizing and ranking search results.

Answering questions such as which of the retrieved results better match the reference query is fundamental to some of interesting Linked Data applications like recommendation provision, matchmaking, social network analysis, visualization, semantic navigation, and data clustering. These require specific measures to analyze and compare the entities in Linked Data. This paper addresses

* Corresponding author.
E-mail address: rouzbeh.meymandpour@sydney.edu.au (R. Meymandpour).

[1] As of November 2015 according to http://lod-cloud.net/.

http://dx.doi.org/10.1016/j.knosys.2016.07.012
0950-7051/© 2016 Published by Elsevier B.V.

this problem based on a systematic assessment of semantic similarity between entities.

Similarity measures evaluate the degree of overlap between entities based on a set of pre-defined factors such as taxonomic relationships, particular characteristics of the entities, or statistical information derived from the underlying knowledge base. They have been proposed and used in diverse areas such as cognitive psychology, computational linguistics, artificial intelligence (AI), and natural language processing (NLP) to assess the similarity (or dissimilarity) between domain concepts or entities. However, besides the fact that each similarity measure is dependent on the implicit or explicit assumptions in its design and formulation, they are largely limited to the specifications and knowledge representation models of particular application domains. These limitations make them less applicable in the Linked Data context.

This paper begins with a review of the related work on LOD-based similarity measurement and recommendation provision (Section 2). It proceeds to provide an overview of the previous approaches for semantic similarity measurement (Section 3) and describes their limitations in the new context of Linked Data (Section 4). By drawing on a formal, mathematical definition of Linked Data, we present our LOD-based semantic similarity measure (Section 5). In order to validate the proposed measure and to demonstrate its applicability and value, it is applied for developing an LOD-based recommender system (Section 6). We compare the performance of our recommender system with that of conventional and state-of-the-art systems. We conclude by discussing the limitations of this study (Section 7). Finally, future research directions and conclusions are discussed in Section 8.

2. Literature review

2.1. Entity and relation ranking in linked data

As mentioned in the foregoing, a key limitation associated with SPARQL is that it does not provide capabilities such as comparing and prioritizing the retrieved search results. Approaches such as f-SPARQL [7] and iSPARQL [8] aimed to provide extensions to the language for ranking and similarity analysis using SPARQL. This research presents a similarity measure based on an information content-based measure which can be used to prioritize and compare the search results of SPARQL queries. In an earlier study [9], we provided examples on how our measure of information content can be used for ranking entities (i.e. resources) and entity properties (i.e. relations) and also demonstrated its applicability for faceted browsing. In another study [10], we have demonstrated how this measure can be applied for providing rankings of universities. The experiments showed comparable results to those provided by international ranking systems such as Quacquarelli Symonds (QS), Times Higher Education (THE), and Academic Ranking of World Universities (ARWU) that require extensive manual efforts to collect, weight, and analyze a large volume of diverse data.

Several methods have been developed by extending traditional Web-based ranking methods such as PageRank [11] and HITS [12] to enable entity ranking in Linked Data. For example, ObjectRank [13] and PopRank [14] have extended PageRank to enable ranking in directed labeled graphs. Ontology Rank [4] and Ontology Dictionary [15] are ranking methods developed for the Swoogle Semantic Web search engine [4]. They use a link analysis-based approach derived from PageRank to rank Semantic Web objects, namely, documents, terms and RDF graphs. ReConRank [16] consists of two ranking methods: ResourceRank that gives PageRank-based scores to resources in the RDF graph and ContextRank that incorporates the provenance of semantic content into the ranking computation. Bamba and Mukherjea [17] modified the hub and authority scores of HITS for ranking the results of Semantic Web queries. Another generalization of HITS in the context of Linked Data is TripleRank [18]. The algorithm was evaluated for faceted browsing and filtering semantic relations for providing a better Linked Data exploration experience.

As will be discussed in more detail in subsequent sections, many of the existing ranking methods are based on the network structure of Linked Data. Delbru et al. [19] argued that link-based approaches such as Swoogle's Ontology Dictionary [15] and ReConRank [16] are computationally expensive and more importantly, do not consider the "semantics of datasets" defined as relationships among datasets that create the Web of Data. The authors proposed DING (Dataset rankING) [19] – a two-layer ranking model for the Web of Data. Its first layer consists of high-level datasets and links between them and the second layer includes entities in the datasets. An extension of PageRank is first applied in the dataset layer to determine the importance of datasets. The importance of each dataset is then distributed to its resources and combined with local entity ranks computed using various methods depending on the dataset.

TRank [20] employed Linked Open Data for the ranking of entity types with applications in natural language processing (NLP). Entity types are properties such as being a person, location, movie, director, etc. that describe the nature of the entity. In LOD, they are related to the main resource using relations such as rdf:type.

Machine learning-based ranking techniques (known as learning to rank [LTR]) (see Liu [21]) have also been used in the context of structured data. In Dali et al. [22], centrality-based features extracted using PageRank and HITS algorithms are combined with features based on statistics related to the RDF graph. Graph-based features of nodes include the number of subjects and objects associated with the node and the diversity of incoming and outgoing relations reachable at different distances from the node. Their main aim was to rank entity search results of the RDF graph. Another approach employed LTR for ranking semantic associations among nodes in RDF graphs [23].

2.2. Semantic similarity and relatedness measurement in linked data

Section 3 below provides an overview of several approaches developed for measuring semantic similarity and relatedness using WordNet and Wikipedia. However, the application of similarity measures in Linked Data is a relatively new research trend. In this section, we review some of the methods developed for RDF graphs and Linked Data.

Before the emergence of Linked Open Data (LOD), RDF graphs were created based on the relationships between entities and used for recommendation provision. Fouss et al. [24] presented a movie recommender system based on a random walk model based on a weighted graph. This graph consists of three types of nodes, namely, people, movies and movie categories, connected by the 'has_watched' and 'belongs_to' relations. Several similarity measures such as short distance and vector-based cosine were evaluated using the proposed graph model. In another study [25], the authors aimed to construct an RDF graph using the information extracted by crawling online stores such as Amazon and iTunes and created. The graph combined with RDF-based user profiles was used for providing recommendations. However, no details of the similarity computation method, the recommendation algorithm and systematic evaluation of the system were provided.

A number of studies have proposed approaches for using LOD for providing recommendations in various domains. However, they are generally restricted to DBpedia and were not properly evaluated. In our earlier work [26], we described some of the key challenges such as heterogeneity and data quality issues involved in