OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

COMPUTING SEMANTIC SIMILARITY OF CONCEPTS IN KNOWLEDGE GRAPHS

Ankita Thorat[1], Dr. A. C. Lomte[2]

PG Students, Department of Computer Engineering, JSPM's BSIOTR, Wagholi, Pune[1]

Associate Professor, Department of Computer Engineering, JSPM's BSIOTR, Wagholi, Pune[2]

ankitthorat1194@gmail.com[1], archanalomte@gmail.com[2]

ABSTRACT: ***The main idea of the wpath semantic similarity method is to encode both the structure of the concept taxonomy and the statistical information of concepts. Furthermore, in order to adapt corpus-based IC methods to structured KGs, graph based IC is proposed to compute IC based on the distribution of concepts over instances in KGs. The proposed a semantic similarity method, namely wpath, to combine these two approaches, using IC to weight the shortest path length between concepts. Conventional corpus-based IC is computed from the distributions of concepts over textual corpus, which is required to prepare a domain corpus containing annotated concepts and has high computational cost. As instances are already extracted from textual corpus and annotated by concepts in KGs, graph-based IC is proposed to compute IC based on the distributions of concepts over instances. Through experiments performed on well known word similarity datasets, we show that the wpath semantic similarity method has produced statistically significant improvement over other semantic similarity methods. Moreover, in a real category classification evaluation, the wpath method has shown the best performance in terms of accuracy and F score.***

Keywords- *Semantic similarity, semantic relatedness, information content, knowledge graph, WordNet, DBpedia*

I INTRODUCTION

Semantic similarity is a metric defined over a set of documents or terms, where the idea of distance between them is based on the likeness of their meaning or semantic content as opposed to similarity which can be estimated regarding their syntactical representation (e.g. their string format). These are mathematical tools used to estimate the strength of the semantic relationship between units of language, concepts or instances, through a numerical description obtained according to the comparison of information supporting their meaning or describing their nature. The term semantic similarity is often confused with semantic relatedness. Semantic relatedness includes any relation between two terms, while semantic similarity only includes is a relations. For example, "car" is similar to "bus", but is also related to "road" and "driving". The Conventional corpus-based IC requires to prepare a domain corpus for the concept taxonomy and then to compute IC from the domain corpus in offline. The inconvenience lies in the high computational cost and difficulty of preparing a domain corpus. More specifically, in order to compute corpus- based IC, the concepts in the taxonomy need to be mapped to the words in the domain corpus. Then the appearance of concepts is counted and the IC values for concepts are generated. In this way, the additional domain corpus preparation and offline computation may prevent the application of those semantic similarity methods relying on the IC values (e.g., res, lin, jcn, and wpath) to KGs, especially when the domain corpus is insufficient or the KG is frequently updated. Since KGs already mined structural knowledge from textual corpus, we

present a convenient graph-based IC computation method for computing the IC of concepts in a KG based on the instance distributions over the concept taxonomy. The graph-based IC is proposed to directly take advantage of KGs while retaining the idea of corpus-based IC representing the specificity of concepts. In consequence, the IC-based semantic similarity methods such as res, lin, jcn and the proposed wpath can compute the similarity score between concepts directly relying on the KG.

II LITERATURE SURVEY

Harshal Wanjari, A Hybrid Approach for Computing Semantic Similarity of Concepts in Knowledge Graphs Measuring semantic similarity of concepts is a crucial component in many applications which has been presented in the introduction. In this paper, we propose wpath semantic similarity method combining path length with IC. The basic idea is to use the path length between concepts to represent their difference, while to use IC to consider the commonality between concepts. The experimental results show that the wpath method has produced statistically significant improvement over other semantic similarity methods. Furthermore, graph-based IC is proposed to compute IC based on the distributions of concepts over instances. It has been shown in experimental results that the graph-based IC is effective for the res, lin and wpath methods and has similar performance as the conventional corpus-based IC. Moreover, graph-based IC has a number of benefits, since it does not requires a corpus and enables online computing based on available KGs. Based on the evaluation of a simple aspect category classification task, the proposed wpath method has also shown the best performance in terms of accuracy and F score. [1]

Kurt Bollacker, Colin Evans, Freebase: A Collaboratively Created Graph Database For Structuring Human Knowledge Freebase is a practical, scalable tuple database used to structure general human knowledge. The data in Freebase is collaboratively created, structured, and maintained. Freebase currently contains more than 125,000,000 tuples, more than 4000 types, and more than 7000 roperties. Public read/write access to Freebase is allowed through an HTTP based graph-query API using the Metaweb Query Language (MQL) as a data query and manipulation language. MQL provides an easy-to-use object-oriented interface to the tuple data in Freebase and is designed to facilitate the creation of collaborative, Web-based data-oriented applications. [2]

Jorn Hees, An Evolutionary Algorithm to Learn SPARQL Queries for Source-Target-Pairs Finding Patterns for Human Associations in DBpedia In this paper we presented an evolutionary graph pattern learner. The algorithm can successfully learn a set of patterns for a given list of source-target-pairs from a SPARQL endpoint. The learned patterns can be used to predict targets for a given source. We use our algorithm to identify patterns in DBpedia for a dataset of human associations. The prediction quality of the learned patterns after fusion reaches a Recall10 of 63.9% and MAP of 39.9 %, and significantly outperforms PageRank, HITS and degree based baselines. [3]

Sebastian Hellmann Knowledge Base Creation, Enrichment and Repair In this chapter we have presented tools for conversion and extraction of data into RDF that were developed in the context of the LOD2 project. Specifically, the DBpedia Extraction Framework supports the extraction of knowledge from Wikis such as Wikipedia, the RDFa, Microdata and Micro formats Extraction Framework crawls and collects data from the Web and Rozeta enables users to create and refine terminological data such as dictionaries and thesauri from natural language text. Once this data has been extracted and lifted to RDF, tools such as ORE and RDFUnit can analyse data quality and repair errors via a GUI. The presented tools are open source and make part of the Linked Data. [4]

III PROBLEM STATEMENT

The problem of measuring semantic similarity between concepts in KGs. we focus on the problem of computing the semantic similarity between concepts in KGs. We propose a method for measuring the semantic similarity between concepts in KGs. We propose a method to compute IC based on the specificity of concepts in KGs.

IV SYSTEM ARCHITECTURE

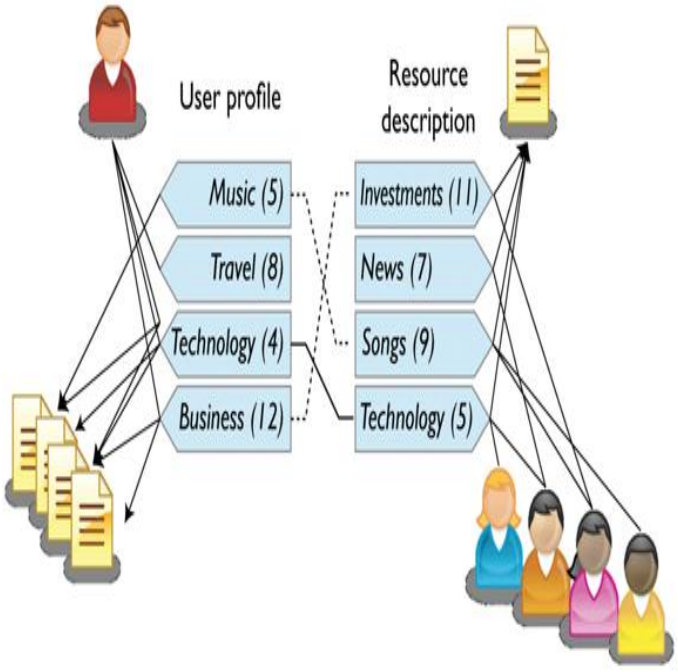

Figure 1: System Architecture

The proposed method aims to give different weights to the shortest path length between concepts based on their shared information, where the path length is viewed as difference and the common information is viewed as commonality. For identical concepts, their path length is 0 so their semantic similarity reaches the maximum similarity 1. As the path length between concepts in the concept taxonomy becomes bigger (bigger value of path length), the semantic similarity between concepts becomes smaller.

V RESULT

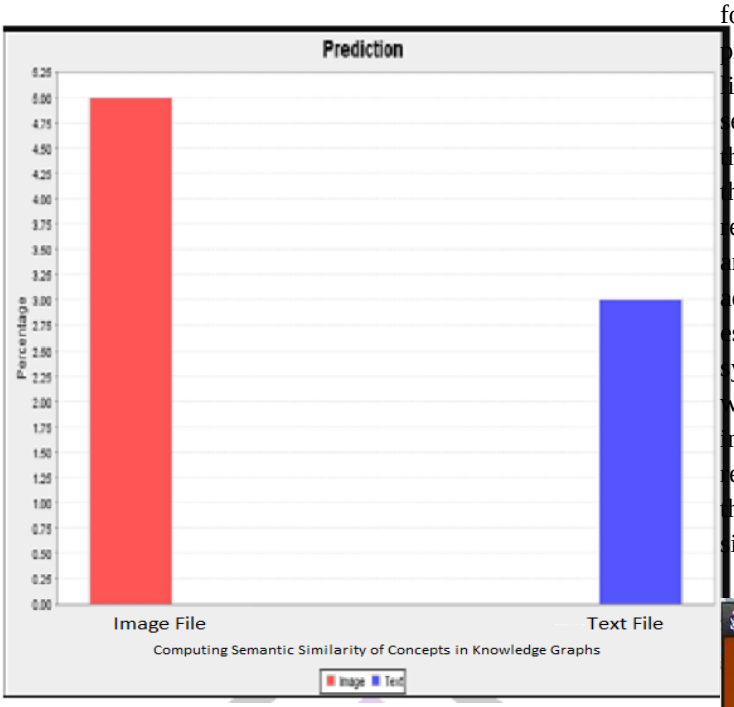

Figure 2: Experimental Analysis

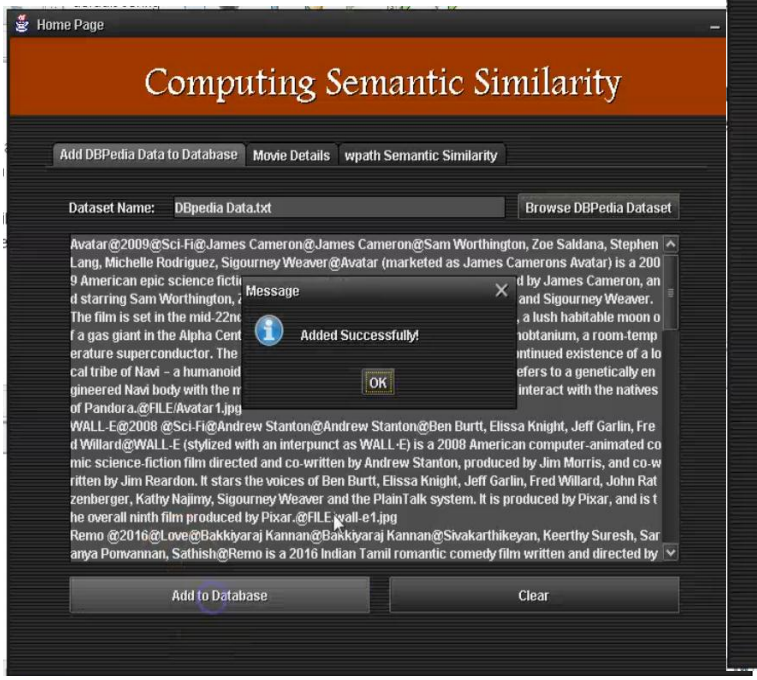

Figure 3: Add DBPedia Data to Database

WPath Semantic Similarity Metric: The information based semantic closeness estimations referenced in the past fragment are generally made to assess how much two thoughts are semantically near using information drawn from thought logical grouping or IC. Estimations take as data two or three thoughts, and reestablish a numerical regard appearing semantic likeness. Various applications rely upon this likeness score to rank the equivalence between different arrangements of ideas. Chart Based Information Content: Customary corpus-based IC requires setting up a space corpus for the thought logical characterization and after that to procedure IC from the region corpus in detached. The trouble lies in the high computational expense and inconvenience of setting up a space corpus. Even more explicitly, remembering the true objective to figure corpus-based IC, the thoughts in the logical order ought to be mapped to the words in the region corpus. By then the nearness of thoughts is checked and the IC regards for thoughts are delivered. In this way, the additional room corpus preparation and disconnected estimation may keep the use of those semantic similarity systems relying upon the IC regards (e.g., res, lin, jcn, and wpath) to KGs, especially when the territory corpus is insufficient or the KG is as regularly as often as possible refreshed. The results are generated in java language. Finally the proposed methodology shows computing semantic similarity of concept in knowledge graphs.

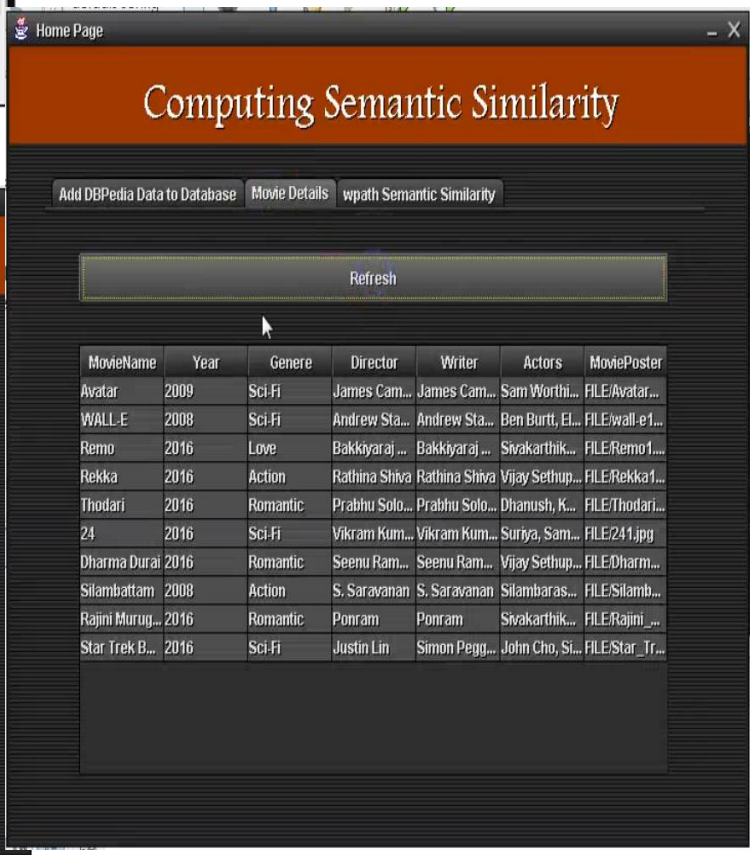

Figure 4: Movie Details

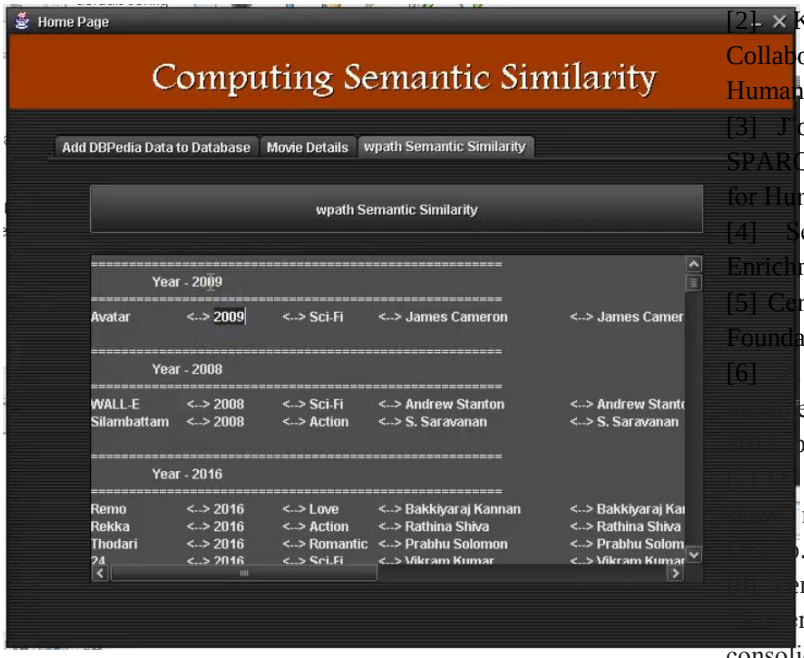

Figure 5: wpath Semantic Similarity

VI CONCLUSION AND FUTURE WORK

Measuring semantic similarity of concepts is a crucial component in many applications which has been presented in the introduction. In this paper, we propose wpath semantic similarity method combining path length with IC. The basic idea is to use the path length between concepts to represent their difference, while to use IC to consider the commonality between concepts. The experimental results show that the wpath method has produced statistically significant improvement over other semantic similarity methods. Furthermore, graph-based IC is proposed to compute IC based on the distributions of concepts over instances. In this system, we evaluated the proposed method in the word similarity dataset and simple classification using the most established evaluation method. More evaluation of semantic similarity methods in other applications considering the taxonomical relation could be useful and can be one of our future works. Furthermore, this paper mainly discussed semantic similarity rather than general semantic relatedness. Therefore, another future work could be in studying the combination of knowledge-based methods with the corpus-based methods for semantic relatedness. Finally, since we combined Word-Net and DBpedia together in this paper, we would further explore using the proposed approaches for measuring the entity similarity and relatedness in KGs.

REFERENCES

[1] Harshal Wanjari, "A Hybrid Approach for Computing Semantic Similarity of Concepts in Knowledge Graphs".

[2] Kurt Bollacker, Colin Evans, "Freebase: A Collaboratively Created Graph Database For Structuring Human Knowledge".

[3] J rn Hees, "An Evolutionary Algorithm to Learn SPARQL Queries for Source-Target-Pairs Finding Patterns for Human Associations in DBpedia".

[4] Sebastian Hellmann "Knowledge Base Creation, Enrichment and Repair".

[5] Cengiz Karakoyunlu, "Toward a Unified Object Storage Foundation for Scalable Storage Systems".

[6] Cheng, et al., "ERMS: An elastic replication ement system for HDFS," in Proc. Cluster Workshops, pp. 32–40.

[7] Chen, J. Yao, and Z. Xiao, "Libra: Lightweight data mitigation in MapReduce," IEEE Trans. Parallel . Syst., vol. 26, no. 9, pp. 2520–2533, Sep. 2014.

[8] erma, G. Dasgupta, T. Nayak, P. De, and R. Kothari, r workload analysis for power minimization using consolidation," in Proc. Conf. USENIX Annu. Tech. Conf., 2009, pp. 28–28.

[9] J. Luo, L. Rao, and X. Liu, "Temporal load balancing with service delay guarantees for data center energy cost optimization," IEEE Trans. Parallel Distrib. Syst., vol. 25, no. 3, pp. 775–784, Mar. 2014.

[10] K. Shvachko, K. Hairong, S. Radia, and R. Chansler, "The hadoop distributed file system," in Proc. IEEE 26th Symp. Mass Storage