Received: December 4, 2017
Revision received: June 21, 2018
Accepted: June 25, 2018

www.estp.com.tr
DOI 10.12738/estp.2018.6.195 ♦ December 2018 ♦ 18(6) ♦ 2958-2966

Research Article

Application of Semantic Similarity Calculation Based on Knowledge Graph for Personalized Study Recommendation Service

Baoxian Jia[1] Xin Huang[2] Shuang Jiao[3]
Liaocheng University *Xi'an Eurasia University* *Liaocheng University*

Abstract

With the coming of the era of big data and the introduction of personalized education concept, how to provide students to value their valuable resources quickly has become hotspot. The efficiency of personalized recommendation service of large educational data is mainly reflected in the accuracy of the recommended algorithm. Semantic similarity computation is essential to be improved for the accuracy of calculation. The research will provide guidance for big data in education area. The development of the Semantic Web has led to new breakthroughs in many fields, such as semantic search, knowledge engineering, knowledge maps, and data connections. The core of the Semantic Web lies in the representation and representation of knowledge in the ontology layer. At the same time, it involves relevant rules and reasoning. Many research fields are based on the ontology lay-er and carry out related research. The semantic similarity technology is a major is-sue in these research fields. Due to the large-scale, heterogeneous, and loosely organized nature of Internet content, it poses a challenge for people to obtain in-formation and knowledge effectively. The Knowledge Graph has powerful open organization capabilities and semantic processing capabilities, laying the foundation for the knowledge-based organization and intelligent applications in the Internet era. At present, the work of the main semantic similarity methods focuses on the structure of the semantic network between concepts (eg, path length and depth), or only on the conceptual information content (IC), and at the same time uses the ontology-related properties for calculations. However, there are some flaws. Therefore, this paper proposes a semantic similarity method, wpath, which com-bines these two methods and uses IC to weight the shortest path length between concepts. In the experiment, it is verified that the proposed method has a certain degree of feasibility and credibility in computing semantic similarity in knowledge graphs. Compared with other methods, the results are superior to other methods.

Keywords

Ontology • Semantic Similarity • Knowledge Graph • Education Big Data • Personalized Recommendation • Collaborative Filtering

[1] **Correspondence to:** Liaocheng University, Liaocheng, 252059, China. Email: jiabaoxian1221@163.com
[2] Xi'an Eurasia University, Xi'an,710000, China. Email: huangxin@eurasia.edu
[3] Liaocheng University, Liaocheng, 252059, China. Email: 36593423@qq.com

Citation: Jia, B. X., Huang, X., & Jiao, S, Application of Semantic Similarity Calculation Based on Knowledge Graph for Personalized Study Recommendation Service. *Educational Sciences: Theory & Practice, 18(6),* 2958-2966. http://dx.doi.org/10.12738/estp.2018.6.195

The low precision problem in real-time recommendation and recommendation quality exists when applying traditional collaborative filtering algorithm to the education data resource. So we gave the optimization of education resources storage scheme based on an improved calculation method of similarity collaborative filtering recommendation system, improving the performance and quality of recommendation collaborative filtering algorithm. The research has important theoretical significance and application value for the personalized learning. In an HVAC system, the chiller consumes a large amount of energy to provide a cooling load. In this system, optimal loading of the cooling load is important for saving energy, and a better performance coefficient is good for the system. For a system cooling load, all the chillers provide the load. To increase energy conservation, the best combination of chillers should be determined. For the cooling load conditions, the optimal chiller loading (OCL) problem is setting the partial load ratio (PLR) of the chillers to reduce system power consumption. Heuristic optimization methods can be used to solve the problem, including the branch and bound method (Miller, 2015), Lagrangian method (Saruladha Aghila and Bhuvaneswary, 2011), and general algebraic modeling system (Slimani, 2013). The OCL must meet the system cooling load, and the constrained condition is transferred to a third objective function. Multi-objective optimization algorithms have been researched and applied for solving many types of industrial problems (Solé-Ribalta, Sánchez & Batet, 2014). Further, many meta-heuristics have been increasingly researched.

The remainder of this paper is organized as follows. Section 2 describes literature on MOEA/D. Section 3 describes the OCL problem. Section 4 presents modified algorithms to solve OCL. Section 5 is the experimental results which show that the performance of the proposed algorithm is better than algorithms from the literature. The final section is the conclusion.

Research Background and Significance

The rapid development of the Internet, online data content has grown in an explosive manner. Due to the large-scale, heterogeneous, and loosely organized nature of Internet content, it poses a challenge for people to obtain information and knowledge effectively. The Knowledge Graph has powerful open organization capabilities and semantic processing capabilities, laying the foundation for the knowledge-based organization and intelligent applications in the Internet era.

In recent years, the research and application of large-scale knowledge map libraries have attracted extensive attention in academic and industrial circles. The main role of a knowledge map is to describe the entities that exist in the real world and the relationships between entities. In 2012, Google formally put forward the concept of knowledge maps. Its purpose is to improve the search engine's capabilities, improve the user's search quality and search experience, and will be popularized in the academic community and the industry after 2013. It will be a traditional keyword-base. The search model is upgraded to semantic-based search. With the development and application of artificial intelligence technology, knowledge maps as key technologies are widely used in intelligent search, intelligence analysis, intelligent question and answer, content distribution, personalized recommendation, and anti-fraud.

Research Content

The dissertation mainly proposes two different semantic similarity calculation methods. The first proposes a method for measuring semantic similarities between concepts in knowledge graphs (KGs); previous work on semantic similarity methods focused on the structure of semantic networks between concepts (eg, path length and depth). Or just focus on the conceptual information content (IC). This paper proposes a second semantic similarity method, wpath, which combines these two methods and uses IC to weight the shortest path length between concepts. Traditional corpus-based ICs are calculated from the distribution of concepts on a text corpus, which is a corpus that is prepared for a domain corpus containing the concept of annotations and that has a high computational cost. Since the instance has been extracted from the text corpora and annotated by the concept of the knowledge graph, a graph-based IC is proposed to calculate the IC based on the distribution of concepts. Through experiments on well-known word similarity data sets, this paper finds that the wpath semantic similarity method produces a statistically significant improvement over other semantic similarity methods. Moreover, in the actual class classification assessment, the wpath method shows the best performance in terms of accuracy and F-score.

Related Principles and Methods

Knowledge Graph

DBpedia is a large-scale multilingual encyclopedia that can be viewed as a structured version of Wikipedia. DBpedia uses a fixed pattern to extract information about Wikipedia's entities, including abstract, category, page link, and infobox. DBpedia currently has more than 28 million entities and hundreds of millions of RDF triples in 127 languages, and as the core of the linked data, there is an entity mapping relationship with many other data sets. DBpedia supports full download of datasets.

Semantic similarity measure

Corpus-based methods. The corpus-based approach measures the semantic similarity between concepts based on information obtained from large corpus (eg, Wikipedia). Following this idea, some work takes advantage of conceptual associations such as Point wise Mutual Information (Landauer and Dumais, 1997) or Normal Google Distance (Gabrilovich and Markovitch, 2007), while others use distributed semantic techniques to represent conceptual meanings in high-dimensional vectors such as latent semantics Analysis and explicit semantic analysis Gloov (Pennington Socher and Manning, 2014). Recent research based on distributed semantics considers advanced computational models, such as Word2Vec (Fellbaum and Miller, 1998), which use low-dimensional vectors to represent words or concepts.

The co-occurrence information of words having the same surrounding context will make a variety of words considered relevant. Because corpus-based methods rely mainly on the contextual information of words, they usually measure the general semantic relevance between words rather than relying on specific semantic

similarities of hierarchical relationships (Singh, 2004). In addition, corpus-based semantic similarity methods represent concepts as words without clarifying their different meanings. Compared with knowledge-based methods that rely on KGs, corpus-based methods generally have better coverage of vocabularies because their computational models can be effectively applied to a variety of newer corpora. Since they are based on word and text corpus instead of concept taxonomy, this article briefly introduces a corpus-based approach and details the main knowledge-based approach in the next section.

Knowledge-based approach. Let Paths(ci,cj)={P1;P2;...Pn} be a set of paths connecting basic concepts ci and cj by cardinality or size N. Let $|Pi|$ denote the path length Pi∈Paths(ci,cj), then length(ci,cj)=min(|Pi|) represents the shortest path length between the two concepts. The path method uses the shortest path length between concepts to represent their semantic distance, and the distance can be converted to similarity.

$$sim_{path}(c_i, c_j) = \frac{1}{1+length(c_i,c_j)} \tag{1}$$

The lch[25] method uses a non-linear function based on the shortest path length to represent the semantic similarity between concepts, as shown in the figure:

$$sim_{lch}(c_i, c_j) = -log(\frac{length(c_i,c_j)}{2*D}) \tag{2}$$

Where D is the maximum depth of concept classification in KG. The concept of the root and the path between given concepts through hierarchical relationships are called depths, because KGs can contain concepts that can be organized into hierarchical categories such as WordNet taxonomies, DBpedia ontology classes, and so on.

Recently shared nodes (LCS) are the most specific concepts in the common ancestor of the two concepts. For example, the LCS of concept scientists and concept actors is the concept person. Let clcs be the LCS of the concepts c i and c j. The method uses the following formula to measure the semantic similarity of a given concept:

$$sim_{wup}(c_i, c_j) = \frac{2depth(c_{lcs})}{depth(c_i)+depth(c_j)} \tag{3}$$

The li method combines the shortest path length with the depth of the LCS. It uses a non-linear function to measure semantic similarity.

$$sim_{li}(c_i, c_j) = e^{-\alpha length(c_{lcs})} \cdot \frac{e^{\beta depth(c_{lcs})} - e^{-\beta depth(c_{lcs})}}{e^{\beta depth(c_{lcs})} + e^{-\beta depth(c_{lcs})}} \tag{4}$$

Where e is the Euler number and α, β are the parameters that contribute to the path length and depth, respectively. According to the experiment, the empirically optimal parameters are α=0.2 and β=0.6.

Research on Measure Method of Semantic Similarity of Knowledge Graph

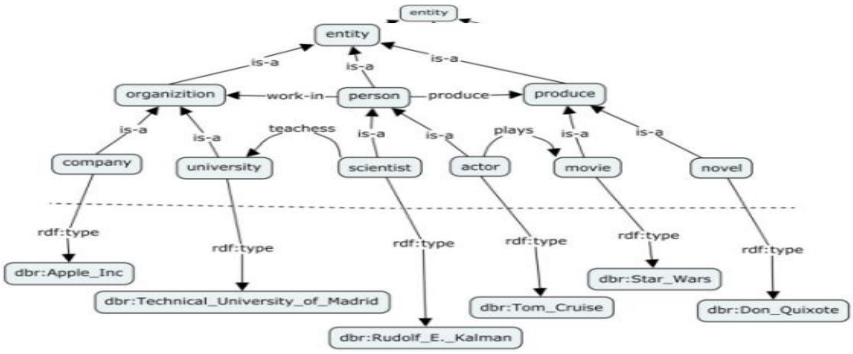

Figure 1. Knowledge graph example

Some traditional semantic similarity measures rely on the use of hierarchical relationships to measure semantic distances between concepts. The semantic similarity between two concepts is proportional to the length of the path connecting the two concepts. Path-based similarity measures require the structure of the semantic network to generate similarity scores that quantify the degree of similarity between the two concepts. Concepts that are physically close to each other in the classification are considered more similar than those that are located far away. Other semantic similarity measures take into account statistical information content (IC) of the concepts computed from corpus to improve the performance of similarity measures based only on conceptual classification structures. IC is a measure of the particularity of a concept. Higher values of the IC are associated with more specific concepts (eg, actor), while lower values are more common (eg, person). The IC is calculated based on the frequency count of the concepts that appear in the text corpora. The emergence of more specific concepts each time also means the emergence of more general ancestral concepts. In order to alleviate the disadvantages of path-based metrics and IC-based metrics, this paper proposes a new semantic similarity method.

Table 1
Entity and Entity Type Mapping Example

Entity	Type	Concept
dbr:Star_Wars	yago: Movie106613686, dbo: Film	Movie
dbr:Don_Quixote	yago: Novel106367879, dbo: Book	Novel
dbr:Tom_Cruise	yago: Actor109765278, dbo: Actor	Actor
dbr:Apple_Inc	yago: Company108058098, dbo: Company	Company

Semantic Similarity Measurement of Weighted Paths

Knowledge-based semantic similarity metrics are mainly used to quantify the semantic similarity of two concepts using information extracted from concept classification or IC. A metric takes a pair of concepts as input and returns a numeric value that represents its semantic similarity. Many applications rely on this similarity score to rank similarities between different pairs of concepts. Taking a segment of the concept taxonomy in Figure 2 as an example, given the conceptual pairs (beef, lamb) and (beef, octopus), the application

gives sim(beef, lamb) a higher similarity value than sim (Beef, octopus), because the concept beef and concept lamb are the kind of meat, and the concept octopus is a kind of seafood. From Table 2, it can be seen that the semantic similarity scores of some concept pairs calculated from the semantic similarity method. As can be seen from the table, the concept pair (beef, lamb) has a higher similarity score than the concept (beef, octopus).

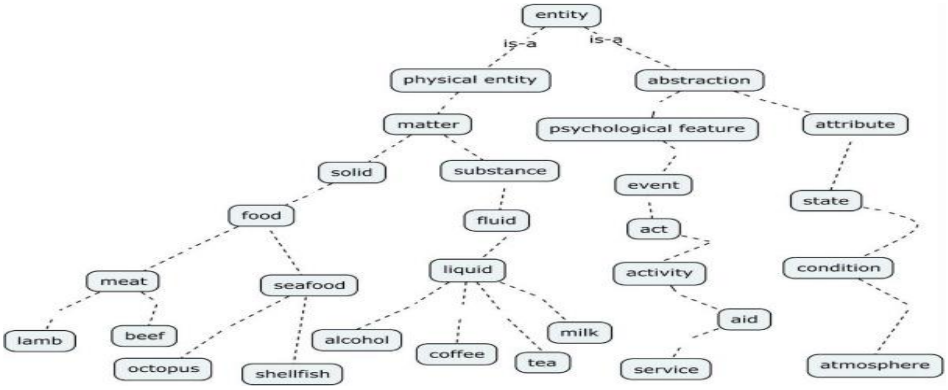

Figure 2. Basic Framework Example of Concept Classification

Table 2
Some Conceptual Pairs Examples in Different Semantic Similarity Methods

Concept Pairs	path	lch	wup	li	res	lin	jcn	wpath
beef-octopus	0.200	2.028	0.714	0.442	6.109	0.484	0.07	0.494
beef-lamb	0.333	2.539	0.857	0.667	6.725	0.591	0.09	0.692
meat-seafood	0.333	2.539	0.833	0.659	6.109	0.760	0.20	0.662
octopus-shellfish	0.333	2.539	0.857	0.667	9.360	0.729	0.12	0.801
beef-service	0.071	0.999	0.133	0.000	0.000	0.000	0.05	0.071
beef-atmosphere	0.083	1.153	0.154	0.000	0.000	0.000	0.05	0.083
beef-coffee	0.111	1.440	0.429	0.168	3.337	0.319	0.06	0.208
food-coffee	0.143	1.692	0.500	0.251	3.337	0.411	0.09	0.260

One of the drawbacks of traditional knowledge-based methods (eg, paths or lch) when dealing with such tasks is that the semantic similarity of any two concepts with the same path length is the same (uniform distance problem). As shown in Figure 2 and Table 2, based on the path and lch semantic similarity method, sim (meat, seafood) is the same as sim(beef, lamb) and sim(octopus, shellfish) because these pairs of concepts have the same shortest path length. Some knowledge-based methods (such as wup or li) try to solve the shortcomings by including depth information in the concept category. Given that higher-level concepts are more common than lower-level concepts in hierarchies, these methods use the depth of the concept to assign higher similarity values to those pairs of concepts that are deeper in the classification. For example, based on the semantic similarity methods of wup and li, the concepts of lamb and concept beaf lie deeper in the conceptual taxonomy (the similarity between lamb and beef is higher than that of meat seafood). Although the performance of using depth is improved compared to the pure path length approach, for the classification criteria given in Figure 2, many concepts share the same depth (level), resulting in the same similarity. For example, as shown in Table 2, sim (lamb, beef) is equal to sim (octotpus, shellfish) based on the semantic similarity method of wup and li, because they have the same depth.

To solve the same path length and depth issues, some knowledge-based methods (eg, res, lin, or jcn) were proposed to include ICs because different concepts usually have different IC values (eg, the IC for meat is 6.725, and food The IC is 6.109), so sim (lamb, beef) is different from sim(octopus, shellfish). Please note that the ICs in this section are based on corpus ICs, which are statistical methods for measuring the informativeness of concepts. The general concept has a lower amount of information, so the IC's value is lower, and more specific concepts have a higher IC value. For example, meat's IC is higher than food's IC because meat is a sub-concept of food. The idea of using ICs to calculate semantic similarity is that the more information two concepts share, the more similar they are. ICs using LCS alone in the res method can represent common information shared by two concepts, but the problem is that the similarity of any two concepts with the same LCS is the same. For example, based on semantic similarity, although the concept pairs (beef, lamb) and (octopus, shellfish) have different similarity scores, the concept pairs (meat, seafood) and (beef, octopus), (beef, coffe) and (food) The similarity scores for coffee and coffee are the same because the LCS of the concept pair is the concept food and matter. Other methods (eg, lin or jcn) attempt to solve this shortcoming by the IC including the concept being compared. However, the mere use of conceptual informativeness to indicate differences between concepts may lose valuable distance information provided by human experts who have created concept taxonomy. In the preliminary experiments of this paper, it has been shown that the path length between concepts in classification is a very effective feature to measure the semantic similarity of concepts. In addition, when the LCS of the concept pair is a root concept entity, the li, res, and lin methods fail, and the similarity value is 0, such as a concept pair (beef, service) and (beef, atmosphere). In addition, the lin and jcn methods still lack layer sub-information. For example, because the concept (meat, seafood) is more common than (octopus, shellfish) and (meat, seafood) is assumed to be less similar, the lin and jcn methods provide higher similarity scores.

Considering the advantages and disadvantages of traditional knowledge-based semantic similarity methods, this paper proposes a weighted path length (wpath) method to measure the semantic similarity between concepts by combining path length and IC. ICs using two concept LCSs weight their shortest path lengths so that concept pairs with the same path length can have different semantic similarity scores if they have different LCSs. The wpath semantic similarity method is as follows

$$sim_{wpath}(c_i, c_j) = \frac{1}{1+length(c_i,c_j)*k^{IC(c_{lcs})}} \tag{3}$$

Where k∈(0,1) and k=1 means that the IC does not contribute to the shortest path length. The parameter k represents the contribution of the IC of the LCS and represents the common information shared by the two concepts.

The proposed method aims to assign different weights to the shortest path length between concepts based on their shared information, where the path length is considered as a difference and the common information is regarded as versatile. For the same concept, their path length is 0, so their semantic similarity reaches the maximum similarity 1. With the concept of the concept of path length between the concepts become larger and larger (the path length value is greater), the concept The smaller the semantic similarity between. The w path similarity score range is (0, 1), which improves the similarity score range of the lch method and the res method.

Conclusion

Since IC-based metrics (for example, res, lin, and jcn) do not involve the hierarchy of concepts, their calculated similarity scores lack information about hierarchy levels and concept distances. Since structural knowledge is retained in the wpath method, it can provide higher similarity scores for more specific concepts, but it can also provide higher similarity scores for concepts that share the same IC and is located closer to the classification. s position. In the (beef, octopus), (meat, seafood) example, because they share the same IC and (beef, octopus), they are closer in taxonomy, and the wpath method gives a higher similarity score. This shows the improvement of the wpath method over the res method. The (octopus, shellfish) and (meat, seafood) examples show that the wpath method solves the grading problem of the lin and jcn methods by giving a more specific concept for higher similarity scores when the two concept pairs have the same path length. .

In summary, the wpath semantic similarity approach uses structure-based methods (eg, path, lch, wup, and li) to represent the distance between concepts in a classification, overcome the same path and depth issues, and lead to the same similarity score for many concepts. Correct. By weighting their path lengths using shared information (IC) between concepts, wpath can not only retain the ability to display distances between concepts based on taxonomies, but also obtain statistical information to represent the commonality between concepts between concepts. The structure in the sexual classification is the same.

References

Fellbaum, C., & Miller, G. (1998). Combining local context and wordnet similarity for word sense identification. *An Electronic Lexical Database*. 265-283. https://books.google.com/books

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. *International Joint Conference on Artificial Intelligence*. 1606-1611. http://www.aaai.org/Papers/IJCAI/2007/IJCAI07-259.pdf

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211-240. http://doi.apa.org/journals/rev/104/2/211.html

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39-41. https://dl.acm.org/citation.cfm?id=219748

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation, in. *Empiricial Methods Natural Language Process*, *12*, 1532-1543. http://www.aclweb.org/anthology/D14-1162

Rada, R., Mili, H., & Bicknell, E. (1989). Development and application of a metric on semantic nets. *IEEE Transaction on System Man & Cybernetics*, *19*(1), 17-30.

Saruladha, K., Aghila, G., & Bhuvaneswary, A. (2011). Information content based semantic similarity approaches for multiple biomedical ontologies. *Advances in Computing and Communications*, 327-336. https://link.springer.com/chapter/10.1007/978-3-642-22714-1_34

Singh, P. (2004). Web ontology to facilitate semantic web. *Inflibnet Centre Ahmedabad*, 11-13. https://www.inflibnet.ac.in/publication/newsletter/v21n4.pdf

Slimani, T. (2013). Description and evaluation of semantic similarity measures approaches. *International Journal of Computer Applications*, *80*(10), 25-33. https://arxiv.org/abs/1310.8059

Solé-Ribalta, A., Sánchez, D., & Batet, M. (2014). Towards the estimation of feature-based semantic similarity using multiple ontologies. *Knowledge-Based Systems*, *55*, 101-113. https://www.sciencedirect.com/science/article/abs/pii/S0950705113003262