Linked Data based Semantic Similarity and Data Mining

Hao Sheng, Huajun Chen, Tong Yu, Yelei Feng
College of Computer Science, Zhejiang University, China
{zjuhsh, huajunsir, ytcs, fengyelei}@zju.edu.cn

Abstract

As a part of the Semantic Web, Linked data is used to connect and share related data on the Web. Compared with traditional Web documents, it has following advantages: more structural; easily understood by humans; describing the things rather than documents or pages; stronger associations. For these reasons, it is more suitable for information search and data mining. In this paper, we proposed a novel approach for semantic similarity between linked data based on lexical taxonomy and corpus statistics. Our approach has been empirically tested by the linked data of Traditional Chinese Medicine (TCM). The experimental results show a good performance in finding and recommending similar herbs in TCM.

Keywords: Linked Data, Semantic Web, Semantic Similarity, Data Mining

1. Introduction

Linking Open Drug Data [1] (LODD) is an open data cloud (Figure 1) for biomedical life science. It integrates a variety of data source, such as drugs, traditional Chinese medicine, diseases, clinical trials, genes. TCMGeneDIT [2], one of the data resources, contains information about 848 different herbs and their associated diseases, effects, genes and ingredients. All these information are discovered from existing literature.

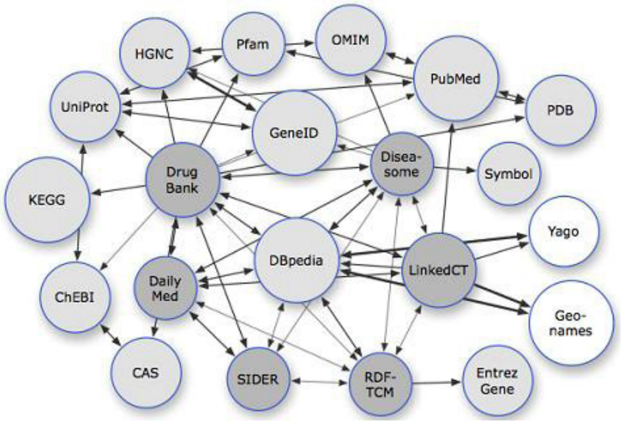

Figure 1. **Different datasets published by LODD into the linked data cloud**

In this paper, we aim to discover the potential associations between linked data entities and provide a method to calculate the degree of semantic similarity between two objects. The semantic similarity has been widely studied and discussed in many fields, such as Natural Language Processing (NLP), Information Retrieval (IR) and Data Mining. Several methods were proposed in the literature. These methods can be classified into three main categories: some using the lexical database, some using corpus statistics, and some hybrid approaches. Each of them had its own advantages and shortages. In this paper, we presented a novel approach based on the object attributes and combine the lexical taxonomy and corpus-based methods. First, we extracted the attributes of object from the linked data and created the index for retrieval. Second, we calculated the semantic distance between two attributes using WordNet [3]. Third, we measured the semantic similarity based on their co-occurrence frequency. At last, we combined the results above and got the final degree of semantic similarity.

The rest of this paper is organized as following: Section 2 briefly introduces the related work. Our method is described in Section 3. Section 4 gives a walk-through example of the method. Experimental results are shown in Section 5 and we conclude in the final section.

2. Related work

Many approaches for determining semantic similarity between two words, concepts or entities have been proposed in literature. In the line of lexical knowledge based approaches, R. Richardson et al. [4] proposed a method using WordNet as knowledge base for measuring semantic similarity between words. Similarly, C. Leacock and M. Chodorow [5] worked on Word Sense Identification in WordNet by combining local context and WordNet similarity. P. Resnik's method [6] was based on edge counting between two concepts in the taxonomy.

Using corpus statistics and co-occurrence frequency, E. Terra and C. Clarke [7] defined a method for similarity

978-1-4244-8099-9/10/$26.00 ©2010 IEEE

measure and frequency estimation. A. Islam and D. Inkpen [8] presented a new corpus-based method using PMI which is applied in our paper. J. Jiang and D. Conrath [9] combined a lexical taxonomy structure with corpus statistical information so that the semantic distance can be better quantified. P. Lord et al. [10] compared the semantic similarity between bioinformatics data using ontological annotations.

In the area of Information Retrieval (IR), Y. Li et al. [11] presented a new approach for semantic similarity using multiple information sources. Variety of strategies for determining the availability of different information sources were implemented in the paper. An ontology based method for information retrieval is discussed in A. M. Rinaldi et al [12].

3. Semantic similarity model

3.1 WordNet based semantic similarity

WordNet is a large English lexical knowledge database developed at Princeton University. Terms in WordNet are grouped into different sets called Synsets. Each synset includes a list of synonyms. There are four relationships in WordNet:

Hypernym: X is a hypernym of Y, if Y is a (kind of) X.
Hyponym: X is a hyponym of Y, if X is a (kind of) Y.
Holonym: X is a holonym of Y, if Y is a part of X.
Meronym: X is a meronym of Y, if X is a part of Y.

In this paper, we use the *hypernym* relationship to construct *is-a* hierarchical taxonomy graph. Figure 2 shows an example: car is a kind of vehicle, elevator is a kind of device, and entity is the root of the graph. In discussing the semantic similarity algorithm, we use the following notations and definitions:

- ***lch*($\boldsymbol{w_1, w_2}$)**: the lowest common hypernym of word w_1 and w_2, For example, in Figure 1, *lch(car, train)= vehicle, lch(car, elevator) = artifact.*
- ***dis*($\boldsymbol{w_1, w_2}$)**: the shortest path length between two nodes in the hierarchical graph. i.e., *dis(car, artifact)=2, dis(elevator, entity) = 3.*
- ***dep(w)***: the depth of the node in the hierarchical graph. We define that *dep(root) = 1, dep(w) = dis(w, root) + 1.*

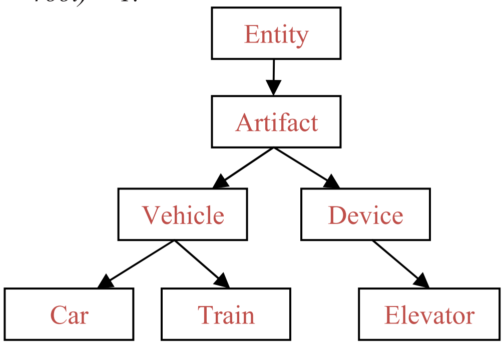

Figure 2. **The *is-a* hierarchical taxonomy graph**

Then, we can define the WordNet based semantic similarity:

$$Sim_{wordnet}(w_1, w_2) = \frac{2 \times dep\big(lch(w_1, w_2)\big)}{dis\big(w_1, lch(w_1, w_2)\big) + dis\big(w_2, lch(w_1, w_2)\big) + 2 \times dep\big(lch(w_1, w_2)\big)}$$

As we can see that the deeper common hypernym the pairs have, and the shorter distance from them to the common hypernym, the higher similarity they get.

e.g.,

$$Sim_{wordnet}(car, train) = \frac{2 \times dep(vehicle)}{dis(car, vehicle) + dis(train, vehicle) + 2 \times dep(vehicle)} = \frac{2 \times 3}{1 + 1 + 2 \times 3} = 0.75$$

$$Sim_{wordnet}(car, elevator) = \frac{2 \times dep(artifact)}{dis(car, artifact) + dis(elevator, artifact) + 2 \times dep(artifact)} = \frac{2 \times 2}{2 + 2 + 2 \times 2} = 0.5$$

3.2 Corpus based semantic similarity

Many different corpus based methods for determining the semantic similarity between two words have been proposed. PMI-IR [13] uses Pointwise Mutual Information. It simply calculated the probability that two words co-occur in the same document. The problem with this method is that if the words are less used than their synonyms, the similarity between them will be underestimated probably. So we use the Second Order Co-occurrence PMI which not only measures the co-occurrence frequency of the two words but also considers their synonyms. First we define some notations as following:

- $\boldsymbol{f(t)}$, which means how many times the word *t* appears in the all corpus.
- $\boldsymbol{f(t, w)}$, which means how many times the word *t* and *w* both appear in the same document.
- PMI is represented by $\boldsymbol{f^{pmi}(t_i, w)}$, $f^{pmi}(t_i, w) = log_2 \frac{f(t_i, w) \cdot m}{f(t_i) \cdot f(w)}$,

where, m is the total number of the words. Then we assumed that W_1 and W_2 are the two words need to be calculated the semantic similarity. For W_1, we define a set of words called *X*, which contains the synonyms of the W_1 and is sorted in descending order by their PMI values with W_1. We take the top-n synonyms having $f^{pmi}(t_i, w) > 0$. $X = \{X_i\}$, *Where* $i = 1,2,3 \ldots n_1$, $f^{pmi}(X_1, W_1) \geq f^{pmi}(X_2, W_1) \geq \ldots f^{pmi}(X_{n_1}, W_1) \geq 0$.
Then we define the n-PMI summation function:
$f(W_1, W_2, n_1) = \sum_{i=1}^{n_1} (f^{pmi}(X_i, W_2))^{\gamma}$,

where, $f^{pmi}(X_i, W_2) > 0$.
It sums all the positive PMI values of words in *X* with W_2. γ is a positive integer greater than 1. The higher value of γ is, the greater influence is on the words having high PMI values. Here we choose $\gamma = 3$.

Similarly, for W_2 we define a set of words *Y*. $Y = \{Y_i\}$, *where i=1,2,...,* n_2,
$f^{pmi}(Y_1, W_2) \ge f^{pmi}(Y_2, W_2) \ge \dots f^{pmi}(Y_{n_2}, W_2) \ge 0$, and the summation function for W_2,

$$f(W_2, W_1, n_2) = \sum_{i=1}^{n_2} (f^{pmi}(Y_i, W_1))^{\gamma}.$$

Now, we can get our corpus based semantic similarity:

$$Sim_{pmi}(W_1, W_2) = \frac{f(W_1, W_2, n_1)}{n_1} + \frac{f(W_2, W_1, n_2)}{n_2}$$

We normalize the corpus based semantic similarity, so the range of result score can be [0, 1] as the WordNet based semantic similarity.

3.3 Semantic similarity model in TCMGeneDIT

As mentioned in Section 1, TCMGeneDIT is one of the databases in LODD about Traditional Chinese Medicine. Each herb has four association relationships with diseases, genes, effects and its ingredients. In order to measure the similarity between two herbs, we take these relationships as the attributes of the herb and calculate the semantic similarity between the corresponding attributes of two herbs. We take the *herb-disease* relationship for example.

Step 1. We consider two herbs, *P* and *R*. The numbers of diseases they treat are *n* and *m* respectively. That is, $P = \{p_1, p_2, \dots p_n\}$, $R = \{r_1, r_2, \dots r_m\}, n \le m$, if not, switch them.

Step 2. Calculate the WordNet based similarity between two diseases lists. We construct a $n \times m$ similarity matrix said $A = \{a_{ij}\}$, *where* $i = 1,2,..n,\ j = 1,2, \dots m$, $a_{ij} = Sim_{wordnet}(p_i, r_j)$.

Step 3. Calculate the corpus based similarity described in Sec.3.2. Similarly, we construct another $n \times m$ similarity matrix said $B = \{b_{ij}\}$, *where* $i = 1,2,..n,\ j = 1,2, \dots m$, $b_{ij} = Sim_{pmi}(p_i, r_j)$.

Step 4. We get a joint matrix said $C = \{c_{ij}\}$, *where i = 1,2,...,n, j = 1,2,...,m,* $c_{ij} = \varphi_1 a_{ij} + \varphi_2 b_{ij}$, $\varphi_1 + \varphi_2 = 1$, where φ_1 is the WordNet based semantic similarity matrix weight factor, φ_2 is the corpus based semantic similarity matrix weight factor.

Step 5. Find out the maximum value c_{ij} in matrix *C* and add it to a list said *S*, $S \to S \cup c_{ij}$. Then remove the *i*th row and the *j*th column from *C*. Repeat above operations until $c_{ij} = 0$ or $n - |S| = 0$.

Step 6. Sum up all the elements in *S* and get the average score, $Sim(X,Y) = \frac{\sum_{i=1}^{|S|} s_i}{|S|}$.

4. A walk through example

We choose *Angelica sinensis* and *Ginkgo biloba* for herb *P* and *R* respectively. The diseases they treat are listed in Table 1.

Table 1. **Example data of *Angelica sinensis'* and *Ginkgo biloba's* diseases**

Medicine	Diseases
Angelica sinensis	*Thrombosis*
	Pneumonia
	Glioblastoma
Ginkgo biloba	*Glaucoma*
	Encephalitis
	Epilepsy
	Cataract

Step 1. P={Thrombosis, Pneumonia, Glioblastomas}, R={Glaucoma, Encephalitis, Epilepsy, Cataract}. So, *n=3, m=4.*

Step 2. Calculate the WordNet based semantic similarity between *Angelica sinensis'* and *ginkgo biloba's* diseases. *A=*

	Thrombosis	*Pneumonia*	*Glioblastoma*
Glaucoma	0.191	0.833	0.720
Encephalitis	0.381	0.273	0.261
Epilepsy	0.286	0.545	0.522
Cataract	0.261	0.833	0.720

Step 3. Calculate the corpus based semantic similarity between *Angelica sinensis'* and *ginkgo biloba's* diseases. *B=*

	Thrombosis	*Pneumonia*	*Glioblastoma*
Glaucoma	0.306	0.361	0.340
Encephalitis	0.366	0.400	0.404
Epilepsy	0.286	0.331	0.323
Cataract	0.289	0.342	0.372

Step 4. Get a joint matrix by setting both φ_1 and φ_2 to 0.5. *C=*

	Thrombosis	*Pneumonia*	*Glioblastoma*
Glaucoma	0.249	**0.597**	0.530
Encephalitis	0.374	0.337	0.333
Epilepsy	0.286	0.438	0.423
Cataract	0.275	0.588	0.546

Step 5. Find the maximum element in matrix *C*, c_{12}=0.597, and add it to *S*, that *S={*0.597*}*. Then remove *Row 1* and *Column 2*. After that, *C=*

	Thrombosis	*Glioblastoma*
Encephalitis	0.374	0.333
Epilepsy	0.286	0.423
Cataract	0.275	**0.546**

Now, 0.546 is the maximum value. So add it to *S*, that *S={*0.597, 0.546*}*. The new matrix after removing *Row 3* and *Column 2* is:

	Thrombosis
Encephalitis	**0.374**
Epilepsy	0.286

Here, the maximum value is 0.374. So, *S={*0.597, 0.546, 0.374*}*. We go through to the next step as $n - |S| = 0$. *(n* $= |S| = 3$*)*

Step 6. We get the final score, *Sim(Angelica sinensis, Ginkgo biloba)* = 1.517/3 = 0.506.

5. Experimental results

Our experimental data are extracted from TCMGeneDIT and the data structure is depicted in Figure 3.

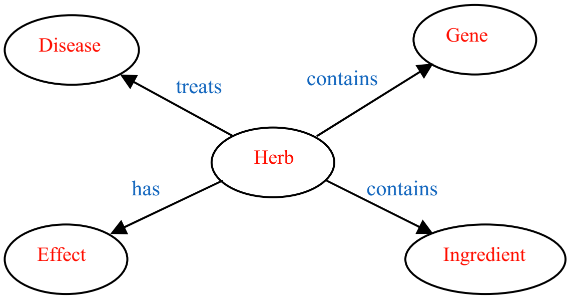

Figure 3. **Data structure**

We take the four association relationships as four attributes of herb. And the weight of each attribute is set by empirical experience on medical knowledge, shown in Table 2.

Table 2. **Weight of relationships**

Subject	Object	Relationship	Weight
herb	disease	treats	0.25
herb	effect	has	0.15
herb	gene	contains	0.25
herb	ingredient	contains	0.35

The experiments consist of two parts:

5.1 Comparison of semantic similarity between herbs

We choose two methods to compare with our approach. One is edge counting method; the other is corpus based method using PMI-IR. We calculated the semantic similarity between *Angelica*, one of the most frequently used herb, and other four herbs. Then we compared the results with the one given by domain experts.

As we can see from the results shown in Figure 4, our proposed method achieves a better performance which is closer to the result given by domain experts than others.

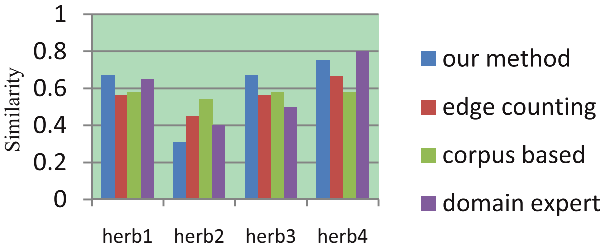

Figure 4. **Results of semantic similarity between herbs**

5.2 Recommendation of similar herbs

For a given herb, we get two recommendation lists which include the most related herbs. One is recommended by our approach, the other by the domain experts. Then we calculate the percentage of the common herbs both appear in two lists.

Table 3. **Recommendation results**

Method	Percentage of common herbs
Our method	43.4%
Edge counting	32.5%
Corpus based	38.3%

In this experiment, we selected five usual herbs. For each herb, we recommended top n most related herbs. As the result in Table 3, our method got 43.4% of common recommendation herbs with domain experts while the edge counting and corpus based methods were 32.5% and 38.3% respectively.

6. Conclusion and future work

In this paper, we have proposed a novel approach for measuring the semantic similarity between two objects which are described in Linked data. It combined the WordNet based method and corpus based method so that it had complementary advantages of both. Furthermore, other form of semantic similarity formula can be derived from our method by setting the factors. The experimental results show that our methods had a better performance and a high similarity with the results given by domain experts.

This work can be extended to multiple databases of linked data. We intend to mine the association relationships between concepts or entities from different data source, like drugs in DailyMed [14] and DrugBank [15]. Second, a visualization system will be developed to display the recommendation results.

References

[1] Linking Open Drug Data: http://esw.w3.org/HCLSIG/LODD.

[2] Y.-C. Fang, H.-C. Huang, H.-H Chen, H.-F Juan, "TCMGeneDIT: a Database for Associated Traditional Chinese Medicine, Gene and Disease Information Using Text Mining", *BMC Complementary and Alternative Medicine*, 8(1):58, 2008.

[3] WordNet: http://wordnet.princeton.edu/

[4] R. Richardson, A.F. Smeaton and J. Murphy, "Using WordNet as a Knowledge Base for Measuring Semantic Similarity between Words", *Report Working paper CA-1294*, School of Computer Applications, Dublin City University, Dublin, Ireland, 1994.

[5] C. Leacock and M. Chodorow, "Combining Local Context and WordNet Similarity for Word Sense Identification in WordNet", *An Electronic Lexical Database*, pages 265–283, MIT Press, 1998.

[6] P. Resnik, "Using Information Content to Evaluate Semantic Similarity", *In Proceedings of the IJCAI05*, pages 488-453, 1995.

[7] E. Terra and C. L. A. Clarke, "Frequency Estimates for Statistical Word Similarity Measures", *Proceedings of HLT-NAACL*, Main Papers, pp. 165-172, May-June 2003.

[8] A. Islam and D. Inkpen, "Second Order Co-occurrence PMI for Determining the Semantic Similarity of Words", *In Proceedings of the International Conference on Language Resources and Evaluation*, Genoa, Italy, 1033—1038, 2006.

[9] J. Jiang and D. Conrath, "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy", *In International Conference on Research in Computational Linguistics*, Taiwan, 1998.

[10] P. Lord, R. Stevens, A. Brass, and C. Goble, "Investigating Semantic Similarity Measures across the Gene Ontology: the Relationship between Sequence and Annotation", *Bioinformatics*, 19(10):1275–83, 2003.

[11] Y. Li, Z.A. Bandar and D. Mclean, "An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources", *IEEE Transactions on Knowledge and Data Engineering.* Vol. 15, no. 4, pp. 871-882. July-Aug. 2003.

[12] Antonio M. Rinaldi, "An Ontology-driven Approach for Semantic Information Retrieval on the Web", *ACM*, Vol. 9, No. 10, July 2009.

[13] P. Turney, "Mining the Web for Synonyms:PMI-IR versus LSA on TOEFL", *In Proceedings of the 12th European Conference on Machine Learning*, 2001.

[14] DailyMed: http://www4.wiwiss.fu-berlin.de/dailymed/

[15] DrugBank: http://www4.wiwiss.fu-berlin.de/drugbank/