TELKOMNIKA Indonesian Journal of Electrical Engineering
Vol. 12, No. 8, August 2014, pp. 6361 ~ 6368
DOI: 10.11591/telkomnika.v12i8.6034

Valuing Semantic Similarity

Abdoulahi Boubacar*[1], Zhendong Niu[2]
Beijing Institute of Technology, School of Computer Science
*Corresponding author, e-mail: abd_boubacar@yahoo.com[1], zniu@bit.edu.cn[2]

Abstract

Similarity is a tool widely used in various domains such as DNA sequence analysis, knowledge representation, natural language processing, data mining, information retrieval, information flow etc. Computing semantic similarity between two entities is a non-trivial task. There are many ways to define semantic similarity. Some measures have been proposed combining both statistical information and lexical similarity. It is difficult for a measure that performs well in a given domain to be applied with accuracy in another domain. A similarity measure may perform better with one language than another. Word is supposed to be not only similar to itself but also to some of its synonyms in a given context and some words with common roots. Our approach is designed to perform query matching and compute semantic relatedness using word occurrences. It performs better than classical measures like TF-IDF, Cosine etc. Although it is not a metric, the proposed similarity measure can be used for a wide range of content analysis tasks based on semantic distance and its efficacy has been demonstrated. The measure is not corpus dependent so it can establish directly the semantic relatedness of two entities.

Keywords: *semantic similarity, semantic relatedness, information retrieval*

1. Introduction

Measuring semantic similarity is the objective of many works. Many measures perform well in evaluation framework for a specific task like synonymy extraction [1], short text comparison [2], sentence similarity [3], and natural language processing [4] etc. The challenge is to find a single measure that performs accurately as long as semantic extraction is needed. The smallest unit for human communication is “word”. Semantic matching certainly cannot work without using words. Can word represent a unit that can express a thought? The answer is obviously no. Only the combination of words can really express an object or idea and by the same way give to a word a meaning. The association of words can be studied at a statistical level [5] using frequency estimation to define similarity measures. Corpus-based word similarity measures [6, 7] extract semantics using word frequency in the entire corpus. Words like stop-words which appear at the same frequency in almost all the documents do not make difference and are not related to any document in particular. A concept-based representation of documents [8] presents an alternative way in indexing. Concepts present more descriptions than words and represent a unit of knowledge from which semantic extraction is relatively easy. More descriptions need more words. For this reason it is very important to use a measure that can compare segments of texts [9] instead of a simple metric for words [10]. In order to measure accurately both query matching and document’s semantic relatedness directly, we had to implement a semantic similarity measure. Our measure is not based on corpus and does not estimate the distance between words and entities [11]. It can perform any task from text similarity [12, 13] to concept similarity [14, 15]. It performs better than the Jaccard similarity measure which is unable to rank accurately documents according to user query. It cannot be compared to corpus based similarity measures which are very poor measuring semantic relatedness. It is entirely based on occurrences therefore it uses extremely simple operators. Computing query matching or semantic relatedness becomes a very easy task using this measure.

Received February 28, 2014; Revised April 3, 2014; Accepted April 15, 2014

2. Related Work

Semantic similarity is an important and very popular tool for organizing and extracting information. Many approaches to estimate similarity have been developed in various domains. Information retrieval is one of the area that uses similarity measures. The most popular measures in this area are based on the vector model. The simplest vector approach to measure semantic similarity is the Cosine distance. The Cosine distance can be defined as:

$$\cos(\vec{a},\vec{b}) = \frac{\vec{a}.\vec{b}}{|\vec{a}||\vec{b}|} \quad (1)$$

In order to compute the similarity between two documents it is sufficient to consider the cosine value of their term vectors. The Jaccard model is a similar measure based on common words. It can be expressed as:

$$Jaccard\ (A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (2)$$

In this equality $A \cap B$ represents the number of words common to A and B while $A \cup B$ represents the total number of words in both$A$ and B. These measures are acceptable for large documents but not very appropriate when one have to deal with short segments of text. Comparing two short segments, the Dice similarity measure seems to be more suitable. The Dice similarity measure is different from the Jaccard measure by the fact that the common words are counted two times and the total length is the sum of the two lengths. The sum of the two lengths is indeed equal or greater than the length of the two texts because of the triangle inequality. The Dice measure can be expressed for two texts$A$ and B as follows:

$$Dice(A,B) = \frac{2|A \cap B|}{|A| \cup |B|} \quad (3)$$

The most popular measure for query matching is the TF-IDF. There are many ways to define the term frequency inverse document frequency. The simplest representation is:

$$tf - idf(t) = \frac{f(t,d)}{f(t,c)} \quad (4)$$

In this equality $f(t,d)$ and $f(t,c)$ represent respectively the frequency of term t in document d and a function of its frequency in an entire corpus c. The idea is to make a difference between common terms which occur in almost every document and related terms which are particular to a given document. The TF-IDF is not appropriate for computing documents similarity. Another measure combining and extending some vector approaches with Latent Semantic Indexing has been presented by J. M. Huerta for machine translation [16]. The measure is based on the Cosine distance and uses Singular Value Decomposition to express sentence similarity. For texts shorter than sentences, like query logs, surface matching is extremely difficult. We do need such similarity for example for query suggestion. Yih and Meek [17] have extended the TF-IDF measure by defining a weighting function according to web relevance scores. The measure is certainly applicable for query suggestion but works only if the words are frequently used. Web relevance can be applied to define co-occurrences in page-count based similarity measures. Bollegala et al [18] have reused popular distances like the Jaccard coefficient to define a page-count based similarity measure. The measure is introduced to determine the semantic similarity distance between two words with a probability function based on the likelihood principle. The semantic similarity between two words is likely to be extracted by corpus-based measures more accurately. Corpus based measures have been tested to determine the semantic relatedness of text segments, in order to detect paraphrases, by Mihalcea et al [19]. Probability is a way that has been explored by Lin [20] in order to measure commonality between two words. His work determines ways and rules for a probabilistic theoretical definition of similarity. An implementation has been made with a modified Dice measure in a taxonomy. Biomedical domain is an area that uses similarity measures frequently. Pedersen et al [21], Lord et al [22] have investigated in medical ontology. DNA sequence analysis [23] is an area of active research that continuously experiments

similarity measures. Gene ontology is the groundwork necessary for most of the DNA based technology. Ontology based similarity measures have been implemented for complex concept expressions over DL-lite knowledge based by Stuckenschmidt [24] and Hajian et al [25]. Concept's distance formalization is usually implemented with lattice theory. Zhang et al [26] have explored a topological approach with formal concept analysis as basis. The defined distance is used to characterize the concepts. A function of neighborhood acts to determine a separation process. Tracking information flow is similar to detecting paraphrases mentioned earlier. The most efficient topic to exercise is the semantic resemblance. Metzler et al [27] has investigated similarity measures analyzing the flow of events through a text corpus. Representing words as vertexes and the relationships between them as edges, Minkov and Cohen [28] have applied graph walk to define a semantic similarity measure. An inter-word similarity measure within a corpora have been tested based on the graph walk method. Latent semantic indexing has served as a basis to implement similarity measures for compliance analysis [29]. This large spectrum shows how diverse are the domains in which semantic similarity measures play a key role.

3. Semantic Similarity Measure

Let consider two texts A and B which semantic relatedness need to be measured. For each word common to A and B we count its occurrences in both A and B. Let $\cap_{A+B}$ denote the sum of the number of occurrences for all the common words. Let $\cup_{A+B}$ denote the sum of the number of words in A and the number of words in B including eventually their occurrences. We denote by δ the similarity measure such that:

$$\delta(A.B) = \frac{\cap_{A+B}}{\cup_{A+B}} \tag{5}$$

The occurrences are counted in both texts$A$ and B, and the text length is the sum of the lengths for the two texts therefore δ is symmetric. All the occurrences are taken for both texts.

3.1. Measuring Semantic Relatedness

Let consider the following data where $D_i$and $W_j$represent respectively documents and words. The frequency for each word and document is indicated at the intersection. For example document D_i contains one time the word $W_1$and four times the word W_2.

Table1. Eight Documents are Presented

Data	W_1	W_2	W_3	W_4	W_5	W_6	W_7	W_8
D_1	1	4	0	0	0	0	0	0
D_2	1	2	7	0	0	0	0	0
D_3	0	4	2	11	0	0	0	0
D_4	0	0	3	2	9	0	0	0
D_5	0	0	0	2	1	13	0	0
D_6	0	0	0	0	3	2	8	0
D_7	0	0	0	0	0	4	2	10
D_8	0	0	0	0	0	0	2	1

For the documents presented by table1 we can calculate the semantic relatedness with the δ similarity measure. $\delta(D_i, D_j) = \delta(D_j, D_i) \forall i.j \in \mathbb{N}$. It is consequently possible to complete the table2 by symmetry.

Table 2. The Semantic Relatedness of Documents $D_1, \dots, D_8$ using the δ Similarity Measure

δ	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8
D_1	1							
D_2	0.53	1						
D_3	0.36	0.56	1					
D_4	0	0.42	0.58	1				
D_5	0	0	0.45	0.47	1			
D_6	0	0	0	0.48	0.66	1		
D_7	0	0	0	0	0.57	0.70	1	
D_8	0	0	0	0	0	0.63	0.79	1

The δ measure is not corpus dependent therefore can directly measure the semantic relatedness. It can be applied for any language as long as words are separated by blank space.

3.2. Comparison between δ and the Jaccard Similarity Measure

The δ measure can be used as long as surface matching is needed. If we observe the two tables, we can see clearly the lack of accuracy in Table 3: very different similarities present the same value. In Table 2 the similarities are all different and proportionality is respected.

Table 3. The Semantic Relatedness of Documents using the Jaccard Similarity Measure

$Jaccard$	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8
D_1	1							
D_2	0.67	1						
D_3	0.25	0.50	1					
D_4	0	0.20	0.50	1				
D_5	0	0	0.20	0.50	1			
D_6	0	0	0	0.20	0.50	1		
D_7	0	0	0	0	0.20	0.50	1	
D_8	0	0	0	0	0	0.25	0.67	1

The similarity between D_1 and D_2 presents the highest score with the Jaccard measure. That is not the case with the δ measure. If we look at word's frequencies in both D_1 and D_2 we note that common words represent almost the half of the total of words. Our measure is therefore more accurate than the Jaccard measure. Unfortunately, we can remark from table2 that the highest score of relatedness recorded is $\delta(D_8, D_7)$ even though W_8 is present only one time in D_8. As a result, there is a need to balance the relatedness between the two entities. For this reason we derive from δ another similarity measure denoted as Δ such that for all documents D_i, D_j we have:

$$\Delta(D_i, D_j) = \frac{\delta(D_i,D_j)+Jaccard(D_i,D_j)}{2} \quad (6)$$

3.3. Comparison between Δ and the Dice Similarity Measure

Table 4 represents the semantic relatedness for the same data with the Δ similarity measure. We can remark that $\Delta(D_8, D_7)$ has been balanced compare to $\delta(D_8, D_7)$. Now we can compare the Δ measure to the Dice similarity wich is designed to better estimate the relatedness for short segments of text.

Table 4. The Semantic Relatedness using the Δ Measure

Δ	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8
D_1	1							
D_2	0.60	1						
D_3	0.31	0.53	1					
D_4	0	0.31`	0.54	1				
D_5	0	0	0.33	0.49	1			
D_6	0	0	0	0.34	0.58	1		
D_7	0	0	0	0	0.39	0.60	1	
D_8	0	0	0	0	0	0.44	0.73	1

Table 5. The Semantic Relatedness using the Dice Measure

$Dice$	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8
D_1	1							
D_2	0.80	1						
D_3	0.40	0.67	1					
D_4	0	0.33	067	1				
D_5	0	0	0.33	0.67	1			
D_6	0	0	0	0.33	0.67	1		
D_7	0	0	0	0	0.33	0.67	1	
D_8	0	0	0	0	0	0.40	0.80	1

We can remark again that the Dice similarity measure is not accurate. Very different similarities has the same score. Some scores like $Dice\ (D_2, D_1)$ are abnormally high. The two measures δ and Δ are more accurate than the Jaccard measure and the Dice measure. The two measures can be used for extracting semantic relatedness. We'll remark in the next section that the δ measure is more suitable than Δ in query processing while the opposite is observed in the case of semantic relatedness. The difference is fairly tiny in both semantic relatedness and query processing.

3.4. Query Processing using either δ or Δ

Both Jaccard and Dice similarity measures are unable to process user query. Our measures can be used to process user query.

For that goal we need to consider the query as a document and measure its relatedness to the documents. Let consider the previous data and a query Q such that $Q = 1W_2 + 2W_4 + 1W_6$. W_4 is repeated twice in the query Q. The relevance of the query to the documents are presented by Table 7 for the δ measure, table8 for the Cosine measure, and table9 for the Δ measure.

Table 6. The query and the documents

Data	W_1	W_2	W_3	W_4	W_5	W_6	W_7	W_8
D_1	1	4	0	0	0	0	0	0
D_2	1	2	7	0	0	0	0	0
D_3	0	4	2	11	0	0	0	0
D_4	0	0	3	2	9	0	0	0
D_5	0	0	0	2	1	13	0	0
D_6	0	0	0	0	3	2	8	0
D_7	0	0	0	0	0	4	2	10
D_8	0	0	0	0	0	0	2	1
Q	0	1	0	2	0	1	0	0

Table7 Relevance of Q using the δ similarity measure.

δ	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8
Q	0.56	0.21	0.86	0.20	0.90	0.18	0.25	0

Table8 Relevance of Q using the $Cosine$ similarity measure.

$Cosine$	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8
Q	0.40	0.11	0.89	0.17	0.53	0.09	0.15	0

We note here a proportionality between the Cosine similarity and the δ measure for almost all the cases. The relevance $Cosine(D_5, Q) = 0.53$ while the relevance $\delta(D_5, Q) = 0.90$. If we use the Cosine similarity measure the most relevant document to the query Q is the document D_3 while using the δ measure the most relevant document to the query Q is D_5. We can see using the Δ similarity measure that we'll have similar difference because Δ and δ are proportional.

Table9 Relevance of Q using the Δ measure.

Δ	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8
Q	0.41	0.21	0.68	0.20	0.70	0.19	0.23	0

4. Discussion

This study was motivated by the fact that we were looking for an appropriate measure for query and query expansion for a concept based information retrieval system. Concepts semantic relatedness is the key feature for query expansion in the model we are implementing. We have represented concepts as vertexes and their relatedness as edges. We have already developed a method to extract concepts from user queries and documents. It has been easy to measure the similarity between a query and each of the concepts with Apache Lucene which has the tf-idf as similarity measure. For that task, it was sufficient to consider the entire collection of concepts as a corpus. Unfortunately the tf-idf measure cannot measure accurately

the semantic relatedness for two texts. In order to measure the semantic relatedness between each pair of concepts we were obliged to choose the Dice similarity measure. The Dice measure can effectively compare two texts. The choice is justified by the fact that we need to measure directly the relatedness of two concepts. A corpus based measure is not suitable for this task. As we have shown in our examples, even though the Dice measure can solve the problem it remains inaccurate. We had to compute the relatedness between documents and concepts using the Cosine similarity measure for a sake of accuracy. In order to compute the relatedness of a query to the documents, we have to compute the path between them through concept nodes which link them. It appeared very uncomfortable to sum scores expressed with three different similarity measures (tf-idf, Dice, and Cosine). Beside all we were not satisfied because the tf-idf measure cannot express directly how a query is related to a concept and the Cosine measure cannot express exactly how a document is related to a concept. In addition, the Dice measure supposes that the words appear at least one tine in each concept. That is true, and the Dice measure is better than the Jaccard measure (3) but it does not indicate the exact degree of relatedness. Our goal was to find an appropriate measure which can compute the three semantic similarities, in order to compute the sum and express the path. We have achieved that goal using new semantic similarity measures.

5. Conclusion and Future Work

δ and Δ are as accurate as the Cosine similarity for query matching and they express more accurately the degree of semantic relatedness. In addition they are not corpus dependent. The weakness of corpus dependent measures is that they cannot express an absolute value for relevance or semantic relatedness. All the results they provide are corpus dependent. We have proven that they are good tools for query matching as well as for semantic relatedness. The particularity of our measures is that they can be used as long as semantic similarity is needed. By using unique measure, comparison becomes very easy. Both δ and Δ can be used for the same tasks. Our future work is to use them to process user query, establish concepts semantic relatedness, and study a concept based information retrieval.

References

[1] Olivier Ferret. *Testing semantic similarity measures for extracting synonyms for a corpus*. Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10) Valletta, Malta. European Language Resources Association (ELRA). 2010; 3338-3343.
[2] Mehran Sahami, Timothy D Heilman. *A web-based Kernel Function for Measuring the Similarity of Short Text Snippets*. WWW '06 Proceedings of the 15th international conference on World Wide Web. ACM New York, NY, USA. 2006; 377-386.
[3] Xiaohua Hu, and Shen Xiajiong. *The Evaluation of Sentence Similarity Measures. Palakorn Achananuparp.* DaWaK '08 Proceedings of the 10th international conference on Data Warehousing and Knowledge Discovery. Pages Springer-Verlag Berlin, Heidelberg. 2008; 305-316.
[4] Angela Schwering. *Evaluation of a Semantic Similarity Measure for Natural Language Spatial Relations*. Spatial Information Theory. Lecture Notes in Computer Science. 2007; 4736: 116-132. Lecture Notes in Computer Science. 2007; 4736: 116-132.
[5] Egidio Terra, CLA Clarke. *Frequency Estimates for Statistical Word Similarity Measures*. *Proceeding* NAACL '03 Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. 2003; 1: 165-172.
[6] Aminul Islam, and Diana Inkpen. *Semantic Text Similarity Using Corpus-Based Word Similarity and String Similarity.* ACM Transactions on Knowledge Discovery from Data (TKDD) TKDD Homepage archive. 2008; 2(2); Article No. 10.
[7] Aminul Islam, Diana Ikpen, Iluju Kiringa. *Applications of corpus-based semantic similarity and word segmentation to database schema matching.* The VLDB Journal. 2008; 17(5): 1293-1320.
[8] Anna-Lan Huang, David Milne, Eibe Frank, Ian H Witten. Learning a Concept-based Document Similarity measure. *Journal of the American Society for Information Science and Technology.* 2012; 63(8): 1593-1608.
[9] Donald Metzler, Susan Dumais, Christopher Meek. *Similarity Measures for Short Segments of Text*. ECIR'07 Proceedings of the 29th European conference on IR research, Springer-Verlag Berlin, Heidelberg. 2007; 16-27.
[10] Ming Li, Xin Chen, Xin Li, Bin Ma, and Paul M.B. Vitanyi. *The Similarity Metric*. IEEE Transactions on Information Theory. Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms. Baltimore, Maryland, USA. 2003.

[11] Jay J Jiang, David W. Conrath Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. *Proceedings of the International Conference Research on Computational Linguistics (ROCLING X).* Taiwan. 1997.
[12] Alberto Barron-Cedeno, Andreas Eiselt, Paolo Rosso. Monolingual Text Similarity Measures: A Comparison of Models over Wikipedia Articles Revisions. *Proceedings of the 7th international Conference on Natural Language ICON.* 2009.
[13] Wen-tau Yih, Kristina Toutanova, John C Platt, Christopher Meek. Learning Discriminative Projections for Text Similarity Measures. *Proceeding CoNLL '11 Proceedings of the Fifteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics Stroudsburg,* PA, USA. 2011; 247-256.
[14] *Wenjie Li, Qiuxiang Xia.* A Method of Concept Similarity Computation Based on Semantic Distance. *In Procedia Engineering.* 2011; 15.
[15] Dolf Trieschnigg, Edgar Meij, Maarten de Rijke and Wessel Kraaij. Measuring Concept Relatedness Using Language Models. *SIGIR, ACM.* 2008; 823-824.
[16] Juan M Huerta. Vector based Approaches to Semantic Similarity Measures. *Advances in Natural Language Processing and Applications, Citeseer.* 2008; 163.
[17] Wen-tau Yih, Christopher Meek. Improving Similarity Measures for Short Segments of Text. *AAAI'07 Proceedings of the 22nd national conference on Artificial intelligence - AAAI Press.* 2007; 2: 1489-1494.
[18] Danushka Bollegala, Yutaka Matsuo, Mitsuru Ishizuka Measuring Semantic Similarity between Words Using Web Search Engines. *Proceedings of the 16th international conference on World Wide Web.* New York, NY, USA, ACM. 2007; 757-766.
[19] Rada Mihalcea, Courtney Corley, Carlo Strapparava. Corpus-based and Knowledge-based Measures of Text Semantic Similarity. *AAAI'06 Proceedings of the 21st national conference on Artificial intelligence - AAAI Press.* 2006; 1: 775-780.
[20] Dekang Lin. An Information Theoretic Definition of Similarity. *Proceedings of the 15th International Conference on Machine Learning.* Morgan Kaufmann, San Francisco, CA. 1998; 296-304.
[21] Ted Pedersen, Serguei Pakhomov, Siddharth Patwardhan, Christopher G Chute. Measures of Semantic Similarity and Relatedness in the Medical Domain. *Journal of Biomedical Informatics archive.* 2007; 40(3): 288-299.
[22] PW Lord, RD Stevens C. A. Goble Semantic Similarity Measures as tools for Exploring the Gene Ontology. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing.* 2003: 601-612.
[23] Computation of Similarity Measures for Sequential Data using Generalized Suffix Trees. Konrad Rieck, Pavel Laskov, Sören Sonnenburg. *Advances in Neural Information Processing Systems.* 2007; 19 (NIPS).
[24] Heiner Stuckenschmidt. A Semantic Similarity Measure for Ontology Based Information. *FQAS '09 Proceedings of the 8th International Conference on Flexible Query Answering Systems.* Springer-Verlag Berlin, Heidelberg. 2009; 406-417.
[25] Behnam Hajian, Tony White. Measuring Semantic Similarity using a Multi-Tree Model. *IJCAI 22nd International Joint Conference on Artificial Intelligence. Barcelona.* 2011.
[26] Lishi Zhang, Shengzhe Gao, Liyan Qi. Topological Distance Function in Formal Concept Lattice. *FSKD '08 Proceedings of the 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery IEEE Computer Society Washington.* DC, USA. 2008; 05: 570-574.
[27] Donald Metzler, Yaniv Bernstein, W Bruce Croft, Alistair Moffat, Justin Zobel. Similarity Measures for Tracking Information Flow. *CIKM '05 Proceedings of the 14th ACM international conference on Information and knowledge management.* ACM New York, NY, USA. 2005; 517-524.
[28] Learning graph walk based similarity measures for parsed text. *EMNLP '08 Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics Stroudsburg,* PA, USA. 2008; 907-916.
[29] Asad Sayeed, Soumitra Sarkar, Yu Deng, Rafah Hosn, Ruchi Mahindru, Nithya Rajamani. Characteristics of document similarity measures for compliance analysis. *CIKM '09 Proceedings of the 18th ACM Conference on Information and Knowledge Management.* ACM New York, NY, USA. 2009; 1207-1216.