Noname manuscript No.
(will be inserted by the editor)

An Overview of Textual Semantic Similarity Measures Based on Web Intelligence

Jorge Martinez-Gil

Received: date / Accepted: date

Abstract Computing the textual similarity between terms (or short text expressions) that have the same meaning but which are not lexicographically similar is a key challenge in many computer related fields. The problem is that traditional approaches to semantic similarity measurement are not suitable for all situations, for example, many of them often fail to deal with terms not covered by synonym dictionaries or are not able to cope with acronyms, abbreviations, buzzwords, brand names, proper nouns, and so on. In this paper, we present and evaluate a collection of emerging techniques developed to avoid this problem. These techniques use some kinds of web intelligence to determine the degree of similarity between text expressions. These techniques implement a variety of paradigms including the study of co-occurrence, text snippet comparison, frequent pattern finding, or search log analysis. The goal is to substitute the traditional techniques where necessary.

Keywords Similarity measures · Web Intelligence · Web Search Engines · Information Integration

1 Introduction

Textual semantic similarity measurement consists of computing the similarity between terms, statements or texts, which have the same meaning, but which are not lexicographically similar [10]. This is an important problem in a lot of computer related fields, for instance, in data mining, information retrieval, or even, natural language processing. The traditional approach for solving this problem has been to use manually compiled dictionaries such as WordNet [4].

Jorge Martinez-Gil
University of Extremadura, Dpt. of Computer Science,
Av. de la Universidad s/n 10003, Caceres, Spain Tel.: +34 927257000-51642
E-mail: jorgemar@unex.es

The question is that a lot of (sets of) terms (acronyms, abbreviations, buzzwords, brand names, and so on) are not covered by these kinds of dictionaries; therefore, similarity measures that are based on this kind of resource cannot be used directly in these cases.

On the other hand, Collective Intelligence (CI) is a field of research that explores the potential that collaborative work has to solve a number of problems. It assumes that when a group of individuals collaborate with each other, intelligence that otherwise did not exist suddenly emerges. We use the name Web Intelligence (WI) when these users use the Web as a means of collaboration. We want to profit from the fact that web users provide rich sets of information that can be converted into knowledge reusable for solving problems related to semantic similarity measurement. To perform our experiments, we are going to use Google [3] which is a web search engine owned by Google Inc. and is currently the most popular search engine on the Web according to Alexa Ranking. However, we see no problem in using any other similar search engine.

So, in this paper we review and evaluate the most promising methods to determine the degree of semantic similarity between (sets of) terms using some kind of web intelligence. We are especially interested in those methods that are able to measure the similarity between emerging terms or expressions which are not frequently covered in dictionaries, including a new branch of methods designed by us which consists of using the historical search patterns from web search engines.

The rest of this paper is organized as follows: Section 2 describes related approaches that are proposed in the literature currently available. Section 3 describes the methods for semantic similarity measurement including the study of co-occurrence, text snippet comparison, frequent pattern finding, or trend analysis. Section 4 presents a statistical evaluation of the presented methods, and finally, we draw conclusions and put forward future lines of research.

2 Related Work

Much work has been developed over the last few years proposing different ways to measure semantic similarity. According to the specific knowledge sources exploited and the way in which they are used, different families of methods can be identified. These families are:

- **Edge Counting Measures**: taking into account the position of the terms in a given dictionary or taxonomy.
- **Information Content Measures**: measuring the difference of the information content of the two terms as a function of their probability of occurrence in a corpus.
- **Feature based Measures**: measuring the similarity between terms as a function of their properties or based on their relationships to other similar terms.
- **Hybrid Measures**: combining all of the above.

Now we propose creating a new category, called WI measures, for trying to determine the semantic similarity between terms using content generated by web users. The rest of this paper explains, evaluates, and discusses the semantic similarity measurement of terms using the Google search engine, but it is applicable to the rest of existing web search engines.

3 Google-Based Techniques

The problem which we are addressing consists of trying to measure the semantic similarity between two given (sets of) terms a and b. Semantic similarity is a concept that extends beyond synonymy and is often called semantic relatedness in the literature. According to Bollegala et al.; a certain degree of semantic similarity can be observed not only between synonyms (lift and elevator), but also between meronyms (car and wheel) or hyponyms (leopard and cat) [2].

In this paper, we use the expression semantic similarity in order to express that we are comparing the meaning of terms instead of comparing their associated lexicography. For example, the terms *house* and *mouse* are quite similar from a lexicographical point of view but do not share the same meaning at all. We are only interested in the real world concept that they represent, considering that a similarity score of 0 stands for complete inequality and 1 for equality of the concepts being compared.

From our point of view, the methods for measuring semantic similarity using web search engines can be categorized as follows:

- **Co-occurrence methods**, which consist of measuring the probability of co-occurrence of the terms on the Web.
- **Frequent patterns finding methods**, which consist of finding similarity patterns in the content indexed by the web search engine.
- **Text snippet comparison methods**, which consist of determining the similarity of the text snippets from the search engines for each term pair.
- **Trend analysis methods**, which consist of comparing the time series representing the historical searches for the terms.

3.1 Co-ocurrence methods

On the Web, probabilities of term co-occurrence can be expressed by hits. In fact, these formulas are measures for the probability of co-occurrence of the terms a and b [5]. The probability of a specific term is given by the number of hits returned when a given search engine is presented with this search term divided by the overall number of web pages possible returned. The joint probability p(a, b) is the number of hits returned by a web search engine, containing both search term a and search term b divided by the overall number of web pages returned.

One of the most outstanding works in this field is the definition of the Normalized Google Distance (NGD) [5]. This distance is a measure of semantic similarity derived from the number of hits returned by the Google search engine for a given (set of) keyword(s).

$$NGD(a,b) = \frac{mx\{\log hit(a), \log hit(b)\} - \log hit(a,b)}{\log M - mn\{\log hit(a), \log hit(b)\}} \quad (1)$$

Other measures of this kind are: Pointwise Mutual Information (PMI), Dice, Overlap Coefficient, or Jaccard, all of which are explained by Bollegala et al.[2]. When these measures are used on the Web, it is necessary to add the prefix Web-; WebPMI, WeDice, and so on. All of them are considered probabilistic because given a web page containing one of the terms, these measures try to compute the probability of that web page also containing the other term. These are their corresponding formulas:

$$WebPMI(a,b) = log\frac{p(a,b)}{p(a) \cdot p(b)} \quad (2)$$

$$WebDice(a,b) = \frac{2 \cdot p(a,b)}{p(a) + (b)} \quad (3)$$

$$WebOverlap(a,b) = \frac{p(a,b)}{min(p(a), p(b))} \quad (4)$$

$$WebJaccard(a,b) = \frac{p(a,b)}{p(a) + p(b) - p(a,b)} \quad (5)$$

Despite its simplicity, the idea behind these measures is that terms with similar meanings tend to be close to each other because it seems to be empirically supported that synonyms often appear together in web pages [5], while terms with dissimilar meanings tend to be farther apart, and therefore, present low similarity values.

3.2 Frequent patterns finding

This group of techniques belongs to the field of machine learning, and consists of looking for similarity patterns in the websites that are indexed by the web search engines. One of the most popular techniques was proposed by Bollegala et al. [2] and proposes looking for such regular expressions as “a also known as b”, “a is a b”, “a is an example of b”, and so on. This is because this kind of expression indicates semantic similarity between the two (set of) terms.

A high number of occurrences of these kinds of patterns provides us with evidence for the similarity between the two terms, but it is necessary to perform some preliminary studies about what is ’a high number’ according to the problem that we wish to address. This can be done, for example, by studying the number of results offered by the web search engines for perfect synonyms. Moreover, it is necessary to take into account that these expressions should

be tested in two ways, because the similarity between a and b is by definition equal to the similarity between b and a.

In our study, we are going to use a method for measuring the occurrences of such expressions as “a is a b” OR “b is an a”. The maximum will be obtained after training the algorithm with some perfect synonyms. For example, try to imagine these perfect synonyms appear, on average, 1 million times together in the same regular expression. Then, 1 million occurrences will be the maximum and 0 occurrences the minimum. A pattern which appears 210,000 times on the web search engine results will present a similarity score of 0.21.

3.3 Text snippet comparison

This kind of technique comprises of capturing the text snippets which are generated by the web search engines when offering the results, just after searching for these terms. These text snippets can be processed in order to be compared using well-known algorithms for determining the similarity between short texts. In this way, we can determine the similarity between two terms based on their associated text snippets.

Moreover, one of the best algorithms for comparing the text snippets is Latent Semantic Analysis (LSA) which is a kind of statistical technique for representing the similarity of terms by analyzing a large text corpus. This technique uses a singular value decomposition approach, namely, a general form of factor analysis for condensing a very large matrix of text content into a smaller matrix[6].

In our study, we are going to use the first text snippet for each term and the LSA algorithm, this LSA algorithm has been borrowed from [17]. More elaborated techniques can be applied. For example, it is possible to capture the n first snippets and try to look for similarities one by one, in this way the uncertainty of dealing with an appropriate/relevant text snippet can be avoided.

3.4 Trend analysis

People may search things on the Web in order to find information related to a given topic. We want to take advantage of this in order to detect similarities between terms and short text expressions. To do this, we are going to work with time series, i.e. collections of observations of well-defined data items obtained through repeated measurements, because the web search engines often store the user queries in this way in order to offer or exploit this information in an efficient manner in the future.

The similarity problem in time series implies that by using two sequences of real numbers representing the measurements of a variable at equal time intervals; the similarity can be defined and computed. Maybe the most intuitive solution could consist of viewing each sequence as a point in n-dimensional

Euclidean space, and defining similarity between sequences as $Lp(X, Y)$, this solution would be easy to compute but there is a problem because there are no actual scales used in data from the web search engines due to the results being normalized and, therefore it is not clear what the exact numbers are.

In order to avoid this kind of problem, we propose using four different ways to compute the semantic similarity: Co-occurrence of Terms in Search Patterns, Computing the Relationships between Search Patterns, Outlier Coincidence on Search Patterns, and Forecasting comparisons. The great advantage of our proposal is that any of the proposed methods take into account the scale of the results, but other kinds of characteristics.

3.4.1 Co-occurrence of Terms in Search Patterns

The first algorithmic method that we propose consists of measuring how often two terms appear in the same query. Co-occurrence of terms in a given corpus is usually used as an indicator of semantic similarity in the literature. We propose adapting this paradigm to our purposes. To do this, we are going to compute the joint probability $p(a, b)$ so that a user query may contain the search terms over time. For example, if we look for the co-occurrence of the terms *lift* and *elevator* over time, we can see that these two terms appear frequently, so we have evidence of their semantic similarity.

The method that we propose for measuring the similarity using the notion of co-occurrence means using the following formula:

$$\frac{n.\ years\ terms\ co-occur}{n.\ years\ registered\ in\ the\ log} \tag{6}$$

We think that the proposed formula is appropriate because it computes a score according to the fact that the terms never appear together or appear together every year. In this way a similarity score of 0 stands for complete inequality and 1 for equality of the input terms.

3.4.2 Correlation between Search Patterns

The correlation between two variables is the degree to which there is a relationship between them. Correlation is usually expressed as a coefficient which measures the strength of a relationship between the variables. We propose using two measures of correlation: Pearson and Spearman.

The first measure of correlation that we propose, i.e. Pearson's correlation coefficient, is closely related to the Euclidean distance over a normalized vector space. Using this measure means that we are interested in the shape of the time series instead of their quantitative values. The philosophy behind this technique is that similar concepts may present almost exactly the same shape in their associated time series and, therefore, semantic similarity between them is presumed to be very high. This similarity can be computed as follows (where the terms a and b are substituted by their corresponding time series):

$$Sim(a,b) = \frac{\text{cov}(a,b)}{\sigma_a \sigma_b} = \frac{E[(a-\mu_a)(b-\mu_b)]}{\sigma_a \sigma_b} \tag{7}$$

The second measure of correlation that we propose using is the Spearman coefficient which assesses how well the relationship between two variables can be described using a monotonic function. If there are no repeated values, a perfect Spearman correlation occurs when each of the variables is a perfect monotone function of the other. This is the formula used to compute it:

$$Sim(a,b) = 1 - \frac{6 \sum d_i^2}{n(n^2-1)} \tag{8}$$

3.4.3 Outlier Coincidence on Search Patterns

There is no rigid mathematical definition of what constitutes an outlier. Grubbs said that "An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs" [7].

So our proposal suggests looking for elements of a time series that distinctly stand out from the rest of the series. Outliers can have many causes. Once we have discarded a Google malfunction, we have to assume that outliers in search patterns occur due to historical events, and that users search for information related to this historical event at the same time but maybe using different lexicographies.

Various indicators are used to identify outliers. We are going to use the proposal of Rousseeuw and Leroy who affirm that an outlier is an observation which has a value that is more than 2.5 standard deviations from the mean [16].

3.4.4 Forecasting Comparison

Our forecasting comparison method compares the prediction of the (sets of) terms for the coming months. There are many methods for time series forecasting, but the problem is that people's behavior cannot be predicted, or at least, can be notably influenced by complex or random causes. For example, it is possible to predict searches related to holidays every summer, but it is not possible to predict searches related to balls. Anyway, we wish to obtain a quantitative result for the quality of this method in order to compare it with the others.

To do that, we propose training a neural network in order to predict the results of the searches. We can establish the similarity between two terms on the basis of the similarity between these predictions. We have chosen a forecasting based on neural networks and discarded such techniques as moving average or exponential smoothing. Moving average uses past observations weighted equally, while exponential smoothing assigns exponentially decreasing weights as the observation gets older. The reason for our choice is that neural networks have been widely used successfully as time series forecasters for real situations [12].

Term	Term	Score
peak oil	apocalypse	0.056
bobo	bohemian	0.185
windmills	offshore	0.278
copyleft	copyright	0.283
tweet	snippet	0.314
subprime	risky business	0.336
imo	in my opinion	0.376
buzzword	neologism	0.383
quantitave easing	money flood	0.410
glamping	luxury camping	0.463
slumdog	underprivileged	0.482
i18n	internationalization	0.518
vuvuzela	soccer horn	0.523
pda	computer	0.526
sustainable	renewable	0.536
sudoku	number place	0.538
terabyte	gigabyte	0.573
ceo	chief executive officer	0.603
tanorexia	tanning addiction	0.608
the big apple	New York	0.641
asap	as soon as possible	0.661
qwerty	keyboard	0.676
thx	thanks	0.784
vlog	video blog	0.788
wifi	wireless network	0.900
hi-tech	high technology	0.903
app	application	0.915

Table 1 Benchmark dataset containing the similarity scores for a set of terms and expressions which are not frequently covered by dictionaries

4 Evaluation

We have created a new dataset which has been rated by a group of 20 people who come from several countries, indicating a value of 0 for non similar terms and 1 for totally similar terms. This dataset is specially designed to evaluate terms that are not frequently included in dictionaries but which are used by people daily. In this way, we will be able to determine the most appropriate method for comparing the semantic similarity of emerging terms. This could be useful in very dynamic domains such as medicine, finance, sms language, social networks, technology, and so on. Table 1 shows the term pairs and the mean for the values obtained after asking the people to comment on their similarity.

The comparison between this dataset and our results is made using the Pearson's Correlation Coefficient, which is a statistical measure for the comparison of two matrices of numeric values. Therefore the results can be in the interval [-1, 1], where -1 represents the worst case and 1 represents the best case. This coefficient allows us to measure the strength of the relation between human ratings of similarity and computational values. However, Pirro stated that it is also necessary to evaluate the significance of this relation [15]. To do

Ranking	Algorithm	Score
1	**WebOverlap**	0.531
2	**Patterns**	0.525
3	**Co-ocurr.**	0.523
4	**WebDice**	0.403
5	**WebJaccard**	0.383
6	**WebPMI**	0.366
7	**NGD**	0.271
8	**Snippet comp.**	0.247
9	Vector pairs	0.207
10	**Pearson**	0.106
11	Lesk	0.079
12	Path length	0.061
13	**Prediction**	0.027
14	**Outlier**	0.007
15	Leacock	0.005
16	**Spearman**	≈ 0
17	Resnik	-0.016

Table 2 Ranking for the algorithms tested using the benchmark dataset which contains terms that do not appear in dictionaries very frequently

that, we have used the p-value technique, which shows how unlikely a given correlation coefficient will occur given no relation in the population. We have obtained that, for our sample, all values above 0.3 are statistically significant. A larger dataset would be necessary to confirm the significance of the rest of the tests.

On the other hand, in order to compare the emerging methods with the existing ones; we consider techniques which are based on dictionaries. We have chosen the Path Length algorithm which is a simple node counting approach. The similarity score is inversely proportional to the number of nodes along the shortest path between the definitions. The shortest path occurs when the two definitions are the same [14]. A dictionary-based approach proposed by Lesk which consists of finding overlaps in the definitions of the two terms. The relatedness score is the sum of the squares of the overlap lengths [9]. An ontology-based technique from Leacock and Chodorow which takes into account the depth of the taxonomy in which the definitions are found [8]. An information-based technique proposed by Resnik, which computes common information between concepts represented by their common ancestor subsuming both concepts found in the taxonomy to which they belong [11]. Finally, the Vector Pairs technique which works by comparing the co-occurrence vectors from the WordNet definitions of concepts [1].

Table 2 shows the results for the benchmark dataset. As can be seen, the emerging methods are much better than those based on dictionaries. The reason is that by using Google, it is possible to have access to fresher and up to date content. On the other hand, we can see that the best methods are those based on co-occurrence, pattern finding and in part for trend analysis. Text snippet comparison seems to less effective, but these results may be influenced by the fact that our implemented method is simple. More complex methods

based on this paradigm could be better, at least when solving specific scenarios. Finally, we have found that the classic methods (Vector pairs, Lesk, Path Length and Resnik), thus, those based on dictionaries are much worse than the majority of the emerging ones, thus, our initial hypothesis is confirmed. Moreover, it is necessary to take into account that most of the methods explained here are apt for optimization, although this step is beyond the scope of this work.

5 Conclusions

In this paper, we have presented and evaluated a set of novel techniques for determining the semantic similarity between (sets of) terms which consists of using knowledge from web search engines. All methods reviewed have been evaluated using a benchmark dataset for terms which are not often included in dictionaries, taxonomies or thesaurus. As a result, we have shown experimentally that some of the methods based on Web Intelligence significantly outperform existing methods when evaluating this kind of dataset.

For future work, we want to avoid the cognitive bias associated with the fact that people rate our term pairs in many different ways according to their cultural background. There are terms that are perfect synonyms to us, but people from other cultures do not agree (and vice versa), so in the future it will be necessary to reach a common agreement on the data used to evaluate the different approaches. Moreover, we are going to keep working towards applying novel time series comparison algorithms, because we think that is an area little explored and can lead to success if the appropriate time series algorithms are used. The final goal is to determine which the best approaches for solving this problem are, and implement them in real information systems where the automatic computation of semantic similarity between terms may be necessary.

References

1. Banerjee, S., Pedersen, T. Extended Gloss Overlaps as a Measure of Semantic Relatedness. *IJCAI* 2003: 805-810.
2. Bollegala, D., Matsuo, Y., Ishizuka, M. Measuring semantic similarity between words using web search engines. *WWW*: 757-766 (2007).
3. Brin, S., Page, L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Networks 30(1-7): 107-117 (1998).
4. Budanitsky, A., Hirst, G. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics* 32(1): 13-47 (2006).
5. Cilibrasi, R., Vitnyi, PM. The Google Similarity Distance. *IEEE Trans. Knowl. Data Eng.* 19(3): 370-383 (2007).
6. Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, Richard A. Harshman: Indexing by Latent Semantic Analysis. *JASIS* 41(6): 391-407 (1990).
7. Grubbs, F. Procedures for Detecting Outlying Observations in Samples. *Technometrics* 11(1): 1-21 (1969).
8. Leacock, C., Chodorow, M., Miller, GA. Using Corpus Statistics and WordNet Relations for Sense Identification. *Computational Linguistics* 24(1): 147-165 (1998).

9. Lesk, M. Information in Data: Using the Oxford English Dictionary on a Computer. *SIGIR Forum* 20(1-4): 18-21 (1986).
10. Li, Y., Bandar, A., McLean, D. An approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. *IEEE Trans. Knowl. Data Eng.* 15(4): 871-882 (2003).
11. Resnik, P. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *IJCAI*: 448-453 (1995).
12. Patuwo, BE., Hu, M. Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting* 14(1): 35-62 (1998).
13. Patwardhan, S., Banerjee, S., Pedersen, T. Using Measures of Semantic Relatedness for Word Sense Disambiguation. *CICLing*: 241-257 (2003).
14. Pedersen, T., Patwardhan, S., Michelizzi, J. WordNet::Similarity - Measuring the Relatedness of Concepts. *AAAI*: 1024-1025 (2004).
15. Pirro, G. A semantic similarity metric combining features and intrinsic information content. *Data Knowl. Eng.* 68(11): 1289-1308 (2009).
16. Rousseeuw, PJ., Leroy, AM. *Robust Regression and Outlier Detection*, John Wiley & Sons, Inc. (2005).
17. Wolfe, MB., Goldman SR. Use of Latent Semantic Analysis for Predicting Psychological Phenomena: Two Issues and Proposed Solutions. *Behavior Research Methods* 35: 22-31 (2003).