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

Semantic Similarity Measurement Using Historical Google Search Patterns

Jorge Martinez-Gil and Jose F. Aldana-Montes

Received: date / Accepted: date

Abstract Computing the similarity between terms (or short text expressions) that have the same meaning but which are not lexicographically similar is a key challenge in the information integration field. The problem is that techniques for textual semantic similarity measurement often fail to deal with words not covered by synonym dictionaries. In this paper, we try to solve this problem by determining the semantic similarity for terms using the knowledge inherent in the search history logs from the Google search engine. To do that, we have designed and evaluated four algorithmic methods for measuring the semantic similarity between terms using their associated history search patterns. These algorithmic methods are: a) frequent co-occurrence of terms in search patterns, b) computation of the relationship between search patterns, c) outlier coincidence on search patterns, and d) forecasting comparisons. We have shown experimentally that some of these methods correlate well with respect to human judgment when evaluating general purpose benchmark datasets, and significantly outperform existing methods when evaluating datasets containing terms that do not usually appear in dictionaries.

Keywords Information Integration · Web Intelligence · Semantic Similarity

1 Introduction

Semantic similarity measurement relates to computing the similarity between terms or short text expressions, having the same meaning or related information, but which are not lexicographically similar [23]. This is an important problem in a lot of computer related fields, for instance, in data warehouse integration when creating mappings that link mutually components of data

University of Malaga
Department of Computer Science
Boulevard Louis Pasteur 35, Malaga (Spain)
{jorgemar, jfam}@lcc.uma.es

warehouse schemas (semi)automatically [4] or in the entity resolution field where two given text objects have to be compared [20]. But the problem is that semantic similarity changes over time and across domains [7]. The traditional approach for solving this problem has consisted of using manually compiled taxonomies such as WordNet [9]. The question is that a lot of (sets of) terms (proper nouns, brands, acronyms, new words, and so on) are not covered by these kinds of taxonomies; therefore, similarity measures that are based on this kind of resources cannot be used directly in these tasks. However, we think that the great advances in the web research field have provided new opportunities for developing accurate solutions.

On the other hand, Collective Intelligence (CI) is an active field of research that explores the potential of collaborative work in order to solve complex problems [36]. Scientists from the fields of sociology, mass behavior, and computer science have made important contributions to this field. It is supposed that when a group of individuals collaborate or compete with each other, intelligence or behavior 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 through their interactions with the web search engines, users provide a rich set of information that can be converted into knowledge reusable for solving problems related with semantic similarity measurement.

To do that, we are going to use Google Trends [10] which is a web application owned by Google Inc. based on Google Search [8]. This web application shows how often a particular search-term is entered relative to the total search-volume across various specific regions, categories, time frames and properties. We are working under the assumption that users are expressing themselves. This expression is in the form of searching for the same concepts from the real world at the same time but represented with different lexicographies. Therefore, the main contributions of this work can be summarized as follows:

- We propose for the first time (to the best of our knowledge) to use historical search patterns from web search engine users to determine the degree of semantic similarity between (sets of) terms. We are especially interested in measuring the similarity between emerging terms or expressions.
- We propose and evaluate four algorithmic methods for measuring the semantic similarity between terms using their historical search patterns. These algorithmic methods are: a) frequent co-occurrence of terms in search patterns, b) computation of the relationship between search patterns, c) outlier coincidence on search patterns, and d) forecasting comparisons.

The rest of this paper is organized as follows: Section 2 describes the related works that are proposed in the literature currently available. Section 3 describes the key aspects of our contribution, including the different ways of computing the semantic similarity. Section 4 presents a statistical evaluation of our approaches in relation to existing ones. Section 5 presents a discussion based on our results, and finally, Section 6 describes the conclusions and future lines of research.

2 Related Work

We have not found proposals addressing the problem of semantic similarity measurements using search logs. Only Nandi & Bernstein have proposed a technique which was based on logs from virtual shops for computing similarity between products [26]. However, a number of works have addressed the semantic similarity measurement [16], [28], [30], [34], [35], and the use of WI techniques for solving computational problems [19], [36], [37] separately.

With regards to the first topic, identifying semantic similarities between terms is not only an indicator of mastery of a language, but a key aspect in a lot of computer-related fields too. It should be taken into account that semantic similarity measures can help computers to distinguish one object from another, group them based on the similarity, classify a new object inside the group, predict the behavior of the new object or simplify all the data into reasonable relationships. There are a lot of disciplines where we can benefit from these capabilities [18]. Within the most relevant areas is the data warehouse field where applications are characterized by heterogeneous models that have to be analyzed and matched either manually or semi-automatically at design time [14]. The main advantage of matching these models consists of enabling a broader knowledge base for decision-support systems, knowledge discovery and data mining than each of the independent warehouses could offer separately [3]. There is also possible to avoid model matching by manually copying all data in a centralized warehouse, but this task requires a great cost in terms of resource consumption, and the results are not reusable in other situations. Designing good semantic similarity measures allows us to build a mechanism for automatically query translation (which is a prerequisite for a successful decouple integration) in an efficient, cheap and highly reusable manner.

Several works have been developed over the last few years proposing different ways to measure semantic similarity. Petrakis et al. stated that according to the specific knowledge sources exploited and the way in which they are used, different families of methods can be identified [30]. These families are:

- Edge Counting Measures: path linking the terms in the taxonomy and of the position of the terms in the taxonomy.
- Information Content Measures: measure the difference of information content of the two terms as a function of their probability of occurrence in a corpus.
- Feature based Measures: measure the similarity between terms as a function of their properties or based on their relationships to other similar terms.
- Hybrid Measures: combine all of the above.

Our proposal does not fit in well enough in any of these families of methods, so that it proposes a new one: Based on WI Measures. However, regarding the use of WI techniques for solving computational problems, we have found many approaches.

- Aggregate information that consists of creating lists of items generated in the aggregate by your users [12]. Some examples are a Top List of items bought, or a Top Search Items or a List of Recent Items.
- Ratings, reviews, and recommendations that consists of understanding how collective information from users can influence others [17].
- User-generated content like blogs, wikis or message boards that consist of extracting some kind of intelligence from contributions by users [24].

Now we propose using a kind of WI technique for trying to determine the semantic similarity between terms that consists of comparing the historical web search logs from the users. The rest of this paper consists of explaining, evaluating, and discussing the semantic similarity measurement of terms using historical search patterns from the Google search engine.

Finally, in order to compare our approaches with the existing ones; we are considering techniques which are based on dictionaries. We have chosen the Path Length algorithm [29] which is a simple edge counting technique. The score is inversely proportional to the number of nodes along the shortest path between the definitions. The shortest possible path occurs when the two definitions are the same, in which case the length is 1. Thus, the maximum score is 1. Another approach proposed by Lesk [22] which consists of finding overlaps in the definitions of the two terms. The score is the sum of the squares of the overlap lengths. The Leacock and Chodorow algorithm [21] which takes into account the depth of the taxonomy in which the definitions are found. An Information Content (IC) measure proposed by Resnik [32] and which computes common information between concepts a and b is represented by the IC of their most specific common ancestor subsuming both concepts found in the taxonomy to which they belong. Finally, the Vector Pairs technique [5] which is a Feature based measure which works by comparing the co-occurrence vectors from the WordNet definitions of concepts.

3 Contribution

Web searching is the process of typing freeform text, either words or small phrases, in order to look for websites, photos, articles, bookmarks, blog entries, videos, and more. People may search things on the Web in order to find information of interest related to a given topic. In a globalized world, our assumption is that large sets of people will search for the same things at the same time but probably from different parts of the world and using different lexicographies. We want to take advantage of this in order to detect similarities between terms and short text expressions. Although our proposal also works with longer text statements, we are going to focus on short expressions only.

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 (e.g. lift and elevator), but also between meronyms (e.g. car and wheel), hyponyms (leopard and cat), related words (e.g. blood and hospital) as well as between antonyms (e.g. day and night) [6]. To do this, we are going to work with time series. The reason is that Google stores the user queries in the form of time series in order to offer or exploit this information in an efficient manner in the future.

According to the Australian Bureau of Statistics[1], a time series is a collection of observations of well-defined data items obtained through repeated measurements over time. For example, measuring the value of retail sales each month of the year would comprise a time series. This is because sales revenue is well defined, and consistently measured at equally spaced intervals. In this way, data which is collected irregularly or only once are not time series.

The similarity problem in time series consists of being two sequences of real numbers representing the measurements of a real variable at equal time intervals defining and computing its similarity. However, this is not a trivial task, because even between different people, the notion of similarity varies. However, it is possible to offer a minimal notion of what is a similarity measure from a mathematical point of view:

Definition 1 (Similarity measure). *A similarity measure sm is a function* $sm : \mu_1 \times \mu_2 \mapsto R$ *that associates the similarity of two input terms* μ_1 *and* μ_2 *to a similarity score* $sc \in \Re$ *in the range [0, 1].*

A similarity score of 0 stands for complete inequality and 1 for equality of the input terms μ_1 and μ_2.

In this paper, we refer to 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 *card* and *car* are quite similar from a lexicographical point of view but do not share the same meaning at all. We are just interested in the real world concept that they represent.

Before beginning to discuss our proposal it is necessary to take into account that in this work we have worked under the assumption that Google has not suffered any transient malfunction when taking measurements of the user searches, so that the morphology of the search patterns is only due to user searches on the Web. Once the problem is clear, the first, and perhaps most intuitive solution, could consist of viewing each sequence as a point in n-dimensional Euclidean space, and define similarity between the two sequences, this solution would be easy to compute but there is an important problem because there are no actual scales used in the graphics due to the normalized results and, therefore it is not clear what the exact or absolute numbers are.

In order to avoid this kind of problem, we propose using four different ways to define and 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

[1] http://www.abs.gov.au/

great advantage of our proposal is that any of proposed methods take into account the scale of the results, but other kinds of characteristics like frequent co-occurrences, correlations, anomalies, or future trends respectively. Moreover, it should be taken into account that for the rest of this work, we are going to evaluate our four approaches using two benchmark datasets:

- Miller & Charles benchmark dataset which is a dataset of term pairs rated by a group of 38 human beings [25]. Term pairs are rated on a scale from 0 (no similarity) to 4 (complete similarity). Miller & Charles ratings has been considered as the traditional benchmark dataset to evaluate solutions that involve semantic similarity measures [6].
- Another new dataset that we will name Martinez & Aldana which is a dataset rated by a group of 20 people belonging to several countries, indicating a value of 0 for not 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 algorithm for comparing the semantic similarity of emerging words. This could be useful in very dynamic domains like medicine, finance, technology, and so on.

The comparison between these two benchmark datasets 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 (totally different values) and 1 represents the best case (totally equivalent values). Note that all tables, except those for the Miller & Charles ratings, are normalized into values in [0, 1] range for ease of comparison. Pearson's correlation coefficient is invariant against a linear transformation [6]. As a general rule, for all the table below the two first columns represent each of the term of the pair to be studied, the third column presents the results from the benchmark dataset, and finally the fourth column represents the value returned by our algorithm.

3.1 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 [6], [11], [34]. We propose adapting this paradigm for our purposes. To do that, we are going to compute the joint probability $p(a, b)$ so that a user query may contain both the search term a and the search term b over the time. Figure 1 shows a example for the co-occurrence of the terms *car* and *automobile* along the time. As can be seen, the terms car and automobile appear together 6 years and the search log is 6 years old, so the resulting score is 6 divided by 6, thus 1. Therefore, we have evidence of their semantic similarity.

The method that we propose to measure the similarity using the notion of co-occurrence consists of using the following formula:

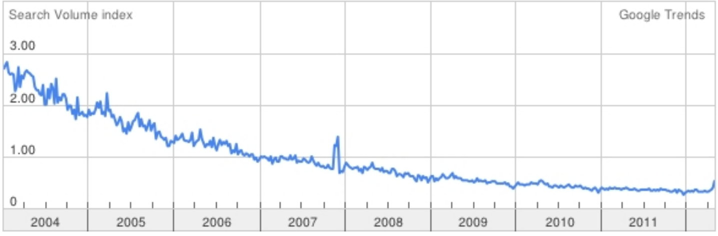

Fig. 1 Search pattern containing both terms car and automobile. User queries have included both terms at the same time frequently so that there is evidence that the both terms represent the same object

		Miller-Charles	**Co-occurrence**
rooster	voyage	0.080	0.000
noon	string	0.080	0.000
glass	magician	0.110	0.000
cord	smile	0.130	0.000
coast	forest	0.420	0.625
lad	wizard	0.420	0.000
monk	slave	0.550	0.000
forest	graveyard	0.840	0.000
coast	hill	0.870	0.750
food	rooster	0.890	0.000
monk	oracle	1.100	0.000
car	journey	1.160	0.750
brother	lad	1.660	0.000
crane	implement	1.680	0.000
brother	monk	2.820	0.000
implement	tool	2.950	0.000
bird	crane	2.970	0.625
bird	cock	3.050	0.000
food	fruit	3.080	1.000
furnace	stove	3.110	0.875
midday	noon	3.420	0.000
magician	wizard	3.500	0.125
asylum	madhouse	3.610	0.000
coast	shore	3.700	0.750
boy	lad	3.760	0.250
journey	voyage	3.840	0.375
gem	jewel	3.840	0.500
automobile	car	3.920	1.000
Score		1.000	0.364

Table 1 Results for the study of the co-occurrence using the Miller & Charles dataset

$$\frac{n.\ years\ terms\ co - occur}{n.\ years\ registered\ in\ the\ log} \tag{1}$$

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.

		Martinez-Aldana	**Co-occurrence**
peak oil	apocalypse	0.056	0.000
bobo	bohemian	0.185	0.000
windmills	offshore	0.278	0.000
copyleft	copyright	0.283	0.000
whalewatching	birdwatching	0.310	0.000
tweet	snippet	0.314	0.000
subprime	risky business	0.336	0.000
imo	in my opinion	0.376	0.000
buzzword	neologism	0.383	0.000
quantitave easing	money flood	0.410	0.000
glamping	luxury camping	0.463	0.000
slumdog	underprivileged	0.482	0.500
i18n	internationalization	0.518	0.000
vuvuzela	soccer horn	0.523	0.125
pda	computer	0.526	1.000
sustainable	renewable	0.536	0.625
sudoku	number place	0.538	0.000
terabyte	gigabyte	0.573	0.625
ceo	chief executive officer	0.603	0.375
tanorexia	tanning addiction	0.608	0.000
the big apple	New York	0.641	0.500
asap	as soon as possible	0.661	0.000
qwerty	keyboard	0.676	1.000
thx	thanks	0.784	0.375
vlog	video blog	0.788	0.000
wifi	wireless network	0.900	1.000
hi-tech	high technology	0.903	0.000
app	application	0.915	1.000
Score		1.000	0.523

Table 2 Results for the study of the co-occurrence using the Martinez & Aldana dataset

Table 1 shows us the results obtained using this method. The problem is that there are terms that are not semantically similar but are searched together frequently, for instance: *coast* and *forest*, or *coast* and *hill* in this dataset. However, our technique provides good results most cases, therefore, the correlation of this technique with respect to human judgment is moderate and could be useful in such cases where a dictionary or thesaurus do not exist.

Table 2 shows us the results obtained using the study of co-occurrence over the specific benchmark. The problem is that there are terms that are not semantically similar but are searched together frequently, for instance the terms *sustainable* and *renewable* or *slumdog* and *underprivileged*. However, the global score is fine what confirm us that it could be used for identifying similarities when dictionaries or other kinds of external resources do not exist.

3.2 Correlation between Search Patterns

The correlation between two variables is the degree to which there is a relationship between them [1]. Correlation is usually expressed as a coefficient which

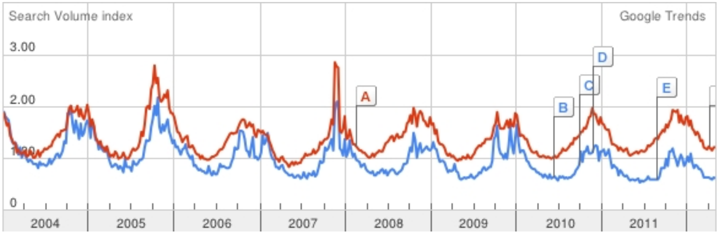

Fig. 2 Historical search log for the terms Furnace and Stove. According to Pearson coefficient, similarity between these temporal series is high which shows us that maybe the two words represent a quite similar object

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 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 can be appreciated in Figure 2, where the terms *furnace* and *stove* present almost exactly the same shape and, therefore, semantic similarity between them is supposed to be very high. The Pearson correlation coefficient can be computed as follows:

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \tag{2}$$

Table 3 shows us the results for the general purpose benchmark dataset. As can be seen, some term pairs present negative correlation, i.e. one of them presents an ascendant pattern while the other presents a descendant one, so the final quality of the method is going to be decreased. Therefore, negative correlations worsen the final score.

Table 4 shows us the results for the specific benchmark dataset. As in the Miller & Charles benchmark dataset, some term pairs present negative correlation, i.e. one of them presents an ascendant pattern whist the other presents a descendant one, so the final quality of the method is not good.

The second measure that we propose using is the Spearman correlation coefficient which assesses how well the relationship between two variables can be described using a monotonic function. If there are no repeated data values, a perfect Spearman correlation occurs when each of the variables is a perfect monotone function of the other [1]. This is the formula to compute it:

$$\rho_{X,Y} = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{3}$$

After using this correlation coefficient for our experiments, we have determined that is not useful for our purposes, because no correlation was detected (a value near to zero). We have discovered that an increment in the web searches for a term does not suppose an increment on the web searches for a

		Miller-Charles	**Pearson**
rooster	voyage	0.080	-0.060
noon	string	0.080	0.338
glass	magician	0.110	0.405
cord	smile	0.130	-0.007
coast	forest	0.420	0.863
lad	wizard	0.420	0.449
monk	slave	0.550	0.423
forest	graveyard	0.840	0.057
coast	hill	0.870	0.539
food	rooster	0.890	0.128
monk	oracle	1.100	0.234
car	journey	1.160	-0.417
brother	lad	1.660	0.101
crane	implement	1.680	0.785
brother	monk	2.820	0.121
implement	tool	2.950	0.771
bird	crane	2.970	0.610
bird	cock	3.05	0.507
food	fruit	3.080	0.286
furnace	stove	3.110	0.728
midday	noon	3.420	0.026
magician	wizard	3.500	0.622
asylum	madhouse	3.610	0.149
coast	shore	3.700	0.183
boy	lad	3.760	0.090
journey	voyage	3.840	0.438
gem	jewel	3.840	0.155
automobile	car	3.920	0.840
Score		1.000	0.163

Table 3 Results for the Pearson's correlation using the Miller & Charles dataset

synonym, so this kind of correlation is not good for trying to determine the semantic similarity between terms using historical search logs and therefore is not going to be considered further in the paper.

3.3 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" [15].

So our proposal consists of looking for elements of a time series that distinctly stands 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.

Figure 3 shows us a screenshot form Google Trends where the time series representing the terms *gem* and *jewel* can be seen. There is a common outlying

		Martinez-Aldana	**Pearson**
peak oil	apocalypse	0.056	-0.100
bobo	bohemian	0.185	-0.147
windmills	offshore	0.278	0.779
copyleft	copyright	0.283	-0.127
whalewatching	birdwatching	0.310	-0.090
tweet	snippet	0.314	0.159
subprime	risky business	0.336	0.000
imo	in my opinion	0.376	0.831
buzzword	neologism	0.383	0.459
quantitave easing	money flood	0.410	0.165
glamping	luxury camping	0.463	0.000
slumdog	underprivileged	0.482	-0.010
i18n	internationalization	0.518	0.966
vuvuzela	soccer horn	0.523	0.828
pda	computer	0.526	0.900
sustainable	renewable	0.536	0.640
sudoku	number place	0.538	-0.220
terabyte	gigabyte	0.573	-0.060
ceo	chief executive officer	0.603	0.163
tanorexia	tanning addiction	0.608	0.000
the big apple	New York	0.641	0.200
asap	as soon as possible	0.661	0.455
qwerty	keyboard	0.676	0.124
thx	thanks	0.784	-0.272
vlog	video blog	0.788	0.838
wifi	wireless network	0.900	-0.659
hi-tech	high technology	0.903	0.867
app	application	0.915	0.473
Score		1.000	0.106

Table 4 Results for the Pearson's correlation using the Martinez & Aldana dataset

observation in the year 2007. We do not know the reason, but this information is not necessary for our purpose. We only look for overlapping outliers in order to determine the similarity between search patterns, and therefore, their associated terms.

Various indicators are used to identify outliers. For our purposes we are going to use the proposal of Rousseeuw and Leroy that affirm that an outlier is an observation which has a value that is more than 2.5 standard deviations from the mean [33].

Table 5 shows us the results obtained by this method using the Miller & Charles benchmark dataset. The obtained correlation for this benchmark dataset is low, because only terms which have suffered a search boom in their search histories can be identified as similar.

Table 6 shows us the results obtained by this method using the Martinez & Aldana benchmark dataset. The obtained correlation for the this benchmark datasets is low, because only terms which which present outliers can be compared, thus, it cannot be outlier coincidence if there is not outliers in the historical search pattern.

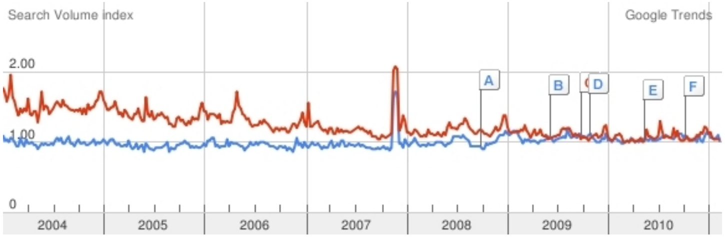

Fig. 3 Historical search log for the terms Gem and Jewel which are considered for the Miller & Charles benchmark dataset as synonyms. There is a perfect coincidence on their respective outliers which is represented in the interval from Nov 18 2007 to Dec 2 2007

		Miller-Charles	**Outlier**
rooster	voyage	0.080	0.000
noon	string	0.080	0.000
glass	magician	0.110	0.000
cord	smile	0.130	0.000
coast	forest	0.420	0.000
lad	wizard	0.420	0.000
monk	slave	0.550	0.000
forest	graveyard	0.840	0.000
coast	hill	0.870	0.000
food	rooster	0.890	0.000
monk	oracle	1.100	0.000
car	journey	1.160	0.000
brother	lad	1.660	0.000
crane	implement	1.680	0.307
brother	monk	2.820	0.000
implement	tool	2.950	0.037
bird	crane	2.970	0.000
bird	cock	3.050	0.000
food	fruit	3.080	0.000
furnace	stove	3.110	0.500
midday	noon	3.420	0.000
magician	wizard	3.500	0.000
asylum	madhouse	3.610	0.000
coast	shore	3.700	0.000
boy	lad	3.760	0.000
journey	voyage	3.840	0.889
gem	jewel	3.840	1.000
automobile	car	3.920	0.000
Score		1.000	0.372

Table 5 Results from outlier coincidence using the Miller & Charles dataset

So we have seen that the major problem for this technique is that not all terms present outliers. It cannot be outlier coincidence if outliers do not exist. Therefore, our method does not fit well enough to all situations. However, score shows us that this kind of technique could be very useful in situations where outlier exists, e.g. *sustainable* and *renewable*, *i18n* and *internationalization*, and so on.

		Martinez-Aldana	**Outlier**
peak oil	apocalypse	0.056	0.000
bobo	bohemian	0.185	0.000
windmills	offshore	0.278	0.400
copyleft	copyright	0.283	0.000
whalewatching	birdwatching	0.310	0.000
tweet	snippet	0.314	0.000
subprime	risky business	0.336	0.000
imo	in my opinion	0.376	0.000
buzzword	neologism	0.383	0.000
quantitave easing	money flood	0.410	0.000
glamping	luxury camping	0.463	0.454
slumdog	underprivileged	0.482	0.000
i18n	internationalization	0.518	0.375
vuvuzela	soccer horn	0.523	0.333
pda	computer	0.526	0.000
sustainable	renewable	0.536	0.800
sudoku	number place	0.538	0.000
terabyte	gigabyte	0.573	0.000
ceo	chief executive officer	0.603	0.000
tanorexia	tanning addiction	0.608	0.009
the big apple	New York	0.641	0.000
asap	as soon as possible	0.661	0.000
qwerty	keyboard	0.676	0.000
thx	thanks	0.784	0.000
vlog	video blog	0.788	0.000
wifi	wireless network	0.900	0.000
hi-tech	high technology	0.903	0.308
app	application	0.915	0.000
Score		1.000	0.007

Table 6 Results from outlier coincidence using the Martinez & Aldana dataset

3.4 Forecasting Comparison

Our forecasting comparison method consists of comparing the prediction of the (sets of) terms for the months following. 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 cars, because it is a kind of non-stationary good. Anyway, we wish to obtain a quantitative result for the quality of this method in order to compare it with the other ones since we can extract positive hints.

To do that, we propose training a neural network in order to predict the results for the searches. We can establish the similarity between two terms on 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 get older. The reason for our choice is that neural networks

		Miller-Charles	**Forecast**
rooster	voyage	0.080	0.661
noon	string	0.080	0.108
glass	magician	0.110	0.235
cord	smile	0.130	0.176
coast	forest	0.420	0.703
lad	wizard	0.420	0.647
monk	slave	0.550	0.971
forest	graveyard	0.840	0.355
coast	hill	0.870	0.218
food	rooster	0.890	0.770
monk	oracle	1.100	0.877
car	journey	1.160	0.478
brother	lad	1.660	0.707
crane	implement	1.680	0.083
brother	monk	2.820	0.154
implement	tool	2.950	0.797
bird	crane	2.970	0.315
bird	cock	3.050	0.893
food	fruit	3.080	0.229
furnace	stove	3.110	0.876
midday	noon	3.420	0.932
magician	wizard	3.500	0.595
asylum	madhouse	3.610	0.140
coast	shore	3.700	0.631
boy	lad	3.760	0.369
journey	voyage	3.840	0.690
gem	jewel	3.840	0.940
automobile	car	3.920	0.689
Score		1.000	0.193

Table 7 Results from search forecasting comparison using the Miller & Charles dataset

have been widely used successfully as time series forecasters for real situations [27]. In order to train the Neural Network we have chosen the following parameters:

- For neurons count: an input layer of 12, a hidden layer of 12 and an output layer of 1
- For the learning: a learning rate of 0.05, a Momentum of 0.5 and Max Iteration of 10000
- For the activation function: bipolar sigmoid
- The period of time chosen is 6 months

In order to compare the predictions we have chosen the Pearson correlation coefficient because in our previous experiment we have shown that is better than the Spearman coefficient. Table 7 shows us the results obtained for the the Miller & Charles benchmark dataset once again.

Table 8 shows us the results obtained for the Martinez-Aldana once again. The final score obtained is not particularly good due to the partial negative correlations for some term pairs.

		Martinez-Aldana	**Forecast**
peak oil	apocalypse	0.056	0.359
bobo	bohemian	0.185	0.671
windmills	offshore	0.278	-0.731
copyleft	copyright	0.283	-0.352
whalewatching	birdwatching	0.310	0.626
tweet	snippet	0.314	0.010
subprime	risky business	0.336	-0.011
imo	in my opinion	0.376	-0.136
buzzword	neologism	0.383	0.924
quantitave easing	money flood	0.410	0.548
glamping	luxury camping	0.463	0.166
slumdog	underprivileged	0.482	-0.701
i18n	internationalization	0.518	-0.401
vuvuzela	soccer horn	0.523	-0.374
pda	computer	0.526	0.964
sustainable	renewable	0.536	0.869
sudoku	number place	0.538	0.137
terabyte	gigabyte	0.573	0.896
ceo	chief executive officer	0.603	-0.396
tanorexia	tanning addiction	0.608	0.267
the big apple	New York	0.641	-0.830
asap	as soon as possible	0.661	0.711
qwerty	keyboard	0.676	0.879
thx	thanks	0.784	0.760
vlog	video blog	0.788	0.752
wifi	wireless network	0.900	0.204
hi-tech	high technology	0.903	-0.117
app	application	0.915	-0.322
Score		1.000	0.027

Table 8 Results from search forecasting comparison using the Martinez & Aldana dataset

4 Evaluation

In order to evaluate the considered approaches, we adopt the Pearson correlation coefficient [13] as a measure of the strength of the relation between human ratings of similarity and computational values. However, Pirro stated that to have a deeper interpretation of the results is also necessary to evaluate the significance of this relation [31]. To do this, we are going to use the p-value technique, which shows how unlikely a given correlation coefficient, r, will occur given no relation in the population [31]. Note that the smaller the p-level, the more significant the relation. Moreover, the larger the correlation value the stronger the relation. The p-value for Pearson's correlation coefficient is based on the test statistics defined as follows:

$$s = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-n^2}} \tag{4}$$

where r is the correlation coefficient and n is the number of pairs of data. When the p-value is less than 0.05, then we can say the obtained value is

Algorithm	**Score**
Resnik	0.814
Leacock	0.782
Path length	0.749
Vector pairs	0.597
Outlier	0.372
Co-ocurr.	0.364
Lesk	0.348
Prediction	0.193
Pearson	0.163

Table 9 Results for the statistical study concerning to the general purpose benchmark dataset

Algorithm	**Score**
Co-ocurr.	0.523
Vector pairs	0.207
Pearson	0.106
Lesk	0.079
Path length	0.061
Prediction	0.027
Outlier	0.007
Leacock	0.005
Resnik	-0.016

Table 10 Results for the statistical study concerning to the benchmark dataset which contains terms that do not appear in dictionaries very frequently

statistically significant. We have obtained that, for our benchmark datasets, all values above 0.25 are statistically significant.

Before to explain the obtained results it is necessary to state that all results have been obtained from data collected before 22nd may 2011. Results from third party approaches have been obtained by the tool offered by Pedersen [2].

Table 9 shows the results for the general purpose benchmark dataset, i.e. Miller & Charles. Existing techniques are better than most of our approaches. However, Outlier and Co-occurrence techniques present a moderate accuracy. The rest of approaches do not seem to be as good as most of the techniques based on synonym dictionaries when identifying the semantic similarity for well known terms. The reason is that knowledge represented in a dictionary is considered to be really good, and therefore, it is not possible for artificial techniques to surpass it.

Table 10 shows the results for the specific purpose benchmark dataset, i.e. Martinez & Aldana. Our approaches present, in general, a better quality than those currently in existence. It is the case for the co-occurrence techniques which significantly beat all others. Moreover, we have experimentally confirmed our hypothesis related to the fact that using historical search patterns could be more beneficial when the terms to be analyzed are not covered by dictionaries.

[2] http://marimba.d.umn.edu/

5 Discussion

Search trends in users web search data have traditionally been shown to very useful when providing models of real world phenomena. Now, we have proposed another way to reuse these search patterns. We have propose comparing search patterns in order to determine the semantic similarity between their associated terms. Despite the results that we have obtained, there are several problems related to the use of historical search patterns for determining the semantic similarity between text expressions:

1. Terms typed by the users that can have multiple meanings based on their context
2. Users use multiple terms to look for both singular and plurals
3. Many of these results rely on the careful choice of queries that prior knowledge suggests should correspond with the phenomenon

On the other hand, our proposal has a number of additional advantages with respect to other existing techniques. It is not time consuming since it do not imply that large corpora should be parsed. We have shown that it correlates well with respect to human judgment (even better than some other preexisting measures). Moreover, our work could be considered as seminal for new research lines:

- The time series representing the historical search pattern for a given term could be used as a kind of semantic fingerprint, thus, some kind of data which identifies a term on the Web. If two semantic fingerprints are similar, it could be supposed that the terms could represent a similar real world entity.
- The results of this work are also applicable to study the stability of ontology mappings. This means that it is possible to establish semantic correspondences between any kind of ontologies according to time constraints.

6 Conclusions

In this paper, we have proposed a novel idea for determining the semantic similarity between (sets of) terms which consists of using the knowledge inherent in the historical search logs from the Google search engine.

To validate our hypothesis, we have designed and evaluated four algorithmic methods for measuring the semantic similarity between terms using their associated history search patterns. These algorithmic methods are: a) frequent co-occurrence of terms in search patterns, b) computation of the relationship between search patterns, c) outlier coincidence in search patterns, and d) forecasting comparisons.

We have shown experimentally that the method which studies the co-occurrence of terms in the search patterns correlates well with respect to human judgment when evaluating general purpose benchmark datasets, and significantly outperform existing methods when evaluating datasets containing terms that do not usually appear in dictionaries. Moreover, we have found than the other three additional methods seem to be better than most of the existing ones when dealing with this special kind of emerging terms.

As future work, we want to keep working towards applying new time series comparison algorithms so that we can determine which are the best approaches for solving this problem and implement them in real information systems where the automatic computation of semantic similarity between terms may be necessary. Moreover, we want to analyze the possibility to smartly combine our algorithmic methods in order to determine if two terms are or no semantically similar.

Acknowledgments

We would like to to thank in advance the reviewers for their time and consideration. We thank Lisa Huckfield for proofreading this manuscript. This work has been funded by Spanish Ministry of Innovation and Science through: ICARIA: From Semantic Web to Systems Biology, Project Code: TIN2008-04844 and by the Regional Government of Andalucia through: Pilot Project for Training and Developing Applied Systems Biology Technology, Project Code: P07-TIC-02978.

References

1. Aitken, A. Statistical mathematics. Oliver & Boyd. 2007.
2. Badea, B., Vlad, A. Revealing Statistical Independence of Two Experimental Data Sets: An Improvement on Spearman's Algorithm. ICCSA 2006: 1166-1176.
3. Banek, M., Vrdoljak, B., Min Tjoa, A., Skocir, Z. Automating the Schema Matching Process for Heterogeneous Data Warehouses. DaWaK 2007: 45-54.
4. Banek, M., Vrdoljak, B., Tjoa, A.M. Using Ontologies for Measuring Semantic Similarity in Data Warehouse Schema Matching Process. *CONTEL* 2007: 227-234.
5. Banerjee, S., Pedersen, T. Extended Gloss Overlaps as a Measure of Semantic Relatedness. *IJCAI* 2003: 805-810.
6. Bollegala, D., Matsuo, Y., Ishizuka, M. Measuring semantic similarity between words using web search engines. *WWW* 2007: 757-766.
7. Bollegala, D., Honma, T., Matsuo, Y., Ishizuka, M. Mining for personal name aliases on the web. *WWW* 2008: 1107-1108.
8. Brin, S., Page, L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks* 30(1-7): 107-117 (1998).
9. Budanitsky, A., Hirst, G. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics* 32(1): 13-47 (2006).
10. Choi, H., Varian, H. Predicting the present with Google Trends. Technical Report, Economics Research Group, Google (2009).
11. Cilibrasi, R., Vitnyi, PM. The Google Similarity Distance. *IEEE Trans. Knowl. Data Eng.* 19(3): 370-383 (2007).
12. Dhurandhar, A. Improving predictions using aggregate information. KDD 2011: 1118-1126.
13. Egghe, L., Leydesdorff, L. The relation between Pearson's correlation coefficient r and Salton's cosine measure CoRR abs/0911.1318: (2009).

14. Fong J., Shiu H., Cheung D. A relational-XML data warehouse for data aggregation with SQL and XQuery. *Softw., Pract. Exper.*38(11): 1183-1213 (2009).

15. Grubbs, F. Procedures for Detecting Outlying Observations in Samples. *Technometrics* 11(1): 1-21 (1969).

16. Hliaoutakis, A., Varelas, G., Petrakis, EGM., Milios, E. MedSearch: A Retrieval System for Medical Information Based on Semantic Similarity. ECDL 2006: 512-515.

17. Hu, N., Bose, I., Koh, NS., Liu, L. Manipulation of online reviews: An analysis of ratings, readability, and sentiments. Decision Support Systems (DSS) 52(3):674-684 (2012).

18. Hjorland, H. Semantics and knowledge organization. ARIST 41(1): 367-405 (2007).

19. Jung, JJ., Thanh Nguyen, N. Collective Intelligence for Semantic and Knowledge Grid. J. UCS (JUCS) 14(7):1016-1019 (2008).

20. Kopcke, H., Thor, A., Rahm, E. Evaluation of entity resolution approaches on real-world match problems. *PVLDB* 3(1): 484-493 (2010).

21. Leacock, C., Chodorow, M., Miller, GA. Using Corpus Statistics and WordNet Relations for Sense Identification. *Computational Linguistics* 24(1): 147-165 (1998).

22. Lesk, M. Information in Data: Using the Oxford English Dictionary on a Computer. *SIGIR Forum* 20(1-4): 18-21 (1986).

23. 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).

24. Liu, B., Zhang, L. A Survey of Opinion Mining and Sentiment Analysis. Mining Text Data 2012:415-463.

25. Miller, G., Charles, W. Contextual Correlates of Semantic Similarity. *Language and Cognitive Processes* 6(1): 1-28 (1991).

26. Nandi, A., Bernstein, PA. HAMSTER: Using Search Clicklogs for Schema and Taxonomy Matching. *PVLDB* 2(1): 181-192 (2009).

27. Patuwo, BE., Hu, M. Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting* 14(1): 35-62 (1998).

28. Patwardhan, S., Banerjee, S., Pedersen, T. Using Measures of Semantic Relatedness for Word Sense Disambiguation. *CICLing* 2003: 241-257.

29. Pedersen, T., Patwardhan, S., Michelizzi, J. WordNet::Similarity - Measuring the Relatedness of Concepts. *AAAI* 2004: 1024-1025.

30. Petrakis, EGM., Varelas, G., Hliaoutakis, A., Raftopoulou, P. X-Similarity: Computing Semantic Similarity between Concepts from Different Ontologies. JDIM 4(4): 233-237 (2006).

31. Pirro, G. A semantic similarity metric combining features and intrinsic information content. *Data Knowl. Eng.* 68(11): 1289-1308 (2009).

32. Resnik, P. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *IJCAI* 1995: 448-453.

33. Rousseeuw, PJ., Leroy, AM. *Robust Regression and Outlier Detection*, John Wiley & Sons, Inc. (2005).

34. Sanchez, D., Batet, M., Valls, A. Web-Based Semantic Similarity: An Evaluation in the Biomedical Domain. Int. J. Software and Informatics 4(1): 39-52 (2010).

35. Sanchez, D., Batet, M., Valls, A., Gibert, K. Ontology-driven web-based semantic similarity. J. Intell. Inf. Syst. 35(3): 383-413 (2010).

36. Scarlat, E., Maries, I. Towards an Increase of Collective Intelligence within Organizations Using Trust and Reputation Models. *ICCCI* 2009: 140-151.

37. Sparck Jones, K. Collective Intelligence: It's All in the Numbers. IEEE Intelligent Systems (EXPERT) 21(3):64-65 (2006)

38. Tuan Duc, N., Bollegala, D., Ishizuka, M. Using Relational Similarity between Word Pairs for Latent Relational Search on the Web. Web Intelligence 2010: 196-199.

Bio

Dr. Jorge Martinez-Gil was a senior researcher in the Department of Computer Languages and Computing Sciences at the University of Malaga (Spain) at the time of this work was done. His main research interests are related with the interoperability in the World Wide Web. In fact, his PhD thesis has addressed the ontology meta-matching and reverse ontology matching problems. Dr. Martinez-Gil has published several papers in prestigious journals like SIGMOD Record, Knowledge and Information Systems, Knowledge Engineering Review, Online Information Review, Journal of Computer Science & Technology, Journal of Universal Computer Science, and so on. Moreover, he is a reviewer for conferences and journals related to the Data and Knowledge Engineering field.

Prof. Dr. José F. Aldana-Montes is currently a professor in the Department of Languages of Computing Sciences at the Higher Computing School from the University of Malaga (Spain) and Head of Khaos Research, a group for researching about semantic aspects of databases. Dr. Aldana-Montes has more than 20 years of experience in research about several aspects of databases, semistructured data and semantic technologies and its application to such fields as bioinformatics or tourism. He is author of several relevant papers in top bioinformatic journals and conferences. Related to teaching, he has been teaching theoretical and practical aspects of databases at all possible university levels: from undergraduate courses to PhD.