Measuring Semantic Similarity Between Digital Forensics Terminologies Using Web Search Engines

Nickson M. Karie
Department of Computer Science,
University of Pretoria,
Private Bag X20, Hatfield 0028,
Pretoria, South Africa.
Email: menza06@hotmail.com

H.S. Venter
Department of Computer Science,
University of Pretoria,
Private Bag X20, Hatfield 0028,
Pretoria, South Africa.
Email: hventer@cs.up.ac.za

***Abstract*—Semantic similarity between different terminologies is becoming a generic problem that extends across numerous domains, touching applications developed for computational linguistics, artificial intelligence, cognitive science and, in the case of this paper, digital forensics. Despite the usefulness of semantic similarity measures in different domains, accurately measuring semantic similarity between any two terms remains a challenging task. The main difficulty lies in developing a computational method with the ability to generate satisfactory results close to how human beings perceive these terminologies, especially when used in their domain of expertise.**

This paper presents a novel approach of using the Web to measure semantic similarity between two terms x and y in the digital forensics domain. The proposed approach is based on the Euclidean distance, a mathematical concept used to calculate the distance between two points. This paper also shows how computing the absolute value of the difference of the logarithms of the hit count percentages of any given terms *x* and *y* relates to the computed Euclidean distance of *x* and *y*. Percentages are computed from the total number of hit counts reported by any Web search engine for the search terms *x*, *y* and the logical *x* AND *y* together. Finally, these concepts are used to deduce a formula to automatically calculate a semantic similarity measure coined as the Digital Forensic Absolute Semantic Similarity Value of the terms *x* and *y*, denoted as DFASSV(*x*, *y*).

Experiments conducted using the proposed DFASSV method focuses on the digital forensics domain. However, a comparison of the DFASSV approach with previously proposed Web-based semantic similarity measures shows that this approach is well suited for digital forensics domain terminologies. In the authors' opinion however, the DFASSV approach can be applied in other domains as well because it does not require any human-annotated knowledge. DFASSV is a novel approach to semantic similarity measure and constitutes the main contribution of this paper.

Keywords-Semantic similarity; digital forensics; digital forensic domain terminologies; Euclidean distance; absolute value; Web; Web search engines

I. INTRODUCTION

An accurate measurement of semantic similarity between terms is a matter of concern in many different domains. For example, due to the problem of ever-changing technological trends in digital forensics, new terms are constantly introduced into the domain and new meanings assigned to existing terms. Depending on the traditional knowledge-based approach, capturing the meaning of these new terms can be very hard, if not next to impossible. Such knowledge could be useful in, for example, the definition of new digital forensic terms, especially when attempting to standardise terms in the field of digital forensics. The authors are currently involved in the creation of an international standard for the digital forensic investigation process where the need arises to carefully define and reason about specific digital forensic terms.

This paper, therefore, proposes a method to compute the semantic similarity measure between two terms in the digital forensic domain using Web search engines. This method is referred as the Digital Forensic Absolute Semantic Similarity Value (DFASSV) in this paper. For the purpose of this paper and scalability of the semantic similarity measure, the terms that are paired using the proposed DFASSV method are rated on a scale of 0 to ∞ where 0 denotes identical semantic similarity between the two terms and ∞ denotes no semantic similarity. Experiments conducted using the proposed method have delivered impressive results.

The Web is a vast entity where an astronomical amount of information is amassed. It is also the largest semantic electronic database in the world [1]. This "database" is available to all and can be queried using any Web search engine that can return aggregate hit count estimates for a large range of search queries [1]. New information is also added to the Web on a daily basis. To tap into this rich bank of information, Web search engines are the most frequently used tools to query for information related to a particular term. To the authors' knowledge, there is so far no better or easier way to search for information on the World Wide Web than simply using Web search engines like Google. However, we do not dispute the existence of other techniques that can be used to search and extract information from the Web. For the purpose of this study, however, the Google search engine was used.

As for the remaining part of this paper, section II discusses related research work. In section III some technical background is explained, followed by a discussion of the proposed DFASSV method in section IV. Experimental results are considered in Section V, while conclusions are drawn in section VI and mention is made of future research work.

.

II. Related Work

There are several methods for measuring the semantic similarity between terms that have been proposed by other researchers. Some of these methods are based on taxonomy while others are Web-based. Taxonomy-based methods use information theory and hierarchical taxonomy such as the WordNet [4] to measure semantic similarity. Web-based methods, on the other hand, prefer the Web as a live and active corpus to a hierarchical taxonomy [5].

The concept of calculating similarity between two words based on the length of the shortest path connecting the two words in taxonomies is discussed in a paper by Roy Rada et al. [6]. If a word is polysemous (i.e. having more than one sense), then multiple paths may exist between the two words. In such cases only the shortest path between any two senses of the words is considered for calculating similarity. The problem of using this approach is that it assumes that, 'theoretically' all the paths in the taxonomy represent equal distances [7] (i.e. the path distance remains the same in all cases and at all times). In practice however, this assumption might not be true; hence the results of the computed semantic similarity measure may well be incorrect.

In another paper, Ming Li et al. [9], discuss the concept of using Web search engine hits for extracting social network information on the Web. Their approach measures the association between two personal names using the Simpson coefficient [9], [17] and [18] and is calculated based on the number of Web hits for each individual name and its conjunction. This approach however, focuses more on the strength of the relation, while the current paper focuses more on the automatic identification of the underlying semantic similarities.

In 2007, Cilibrasi and Vitànyi [1] introduced the concept of the Normalized Google Distance (NGD), which was based on a 2004 research paper on normalized information distance between two strings (discussed in [9]), and which calculates a distance metric between words using page counts indexed by a Web search engine. The NGD is evaluated in a word classification task (i.e. words are grouped based on their similarities according to the model referred to in [10]). This also means that the words in question usually display the same formal properties, especially their inflections and distribution. The problem with this method is that it uses a value (**M**) that can be defined as the total number of pages on the Web that Google will search when given a query. The value **M** is quoted as 8058044651Web pages [1]. According to an Official Google Blog [11], this number has increased significantly since 1998 when it was only 26 million. By 2000 the Google index had reached the one billion mark. Over the last decade, this number has been changing and, recently, even the Google search engineers stopped calculating it due to the sheer vastness of the Web these days [11]. The Google systems that process links on the Web recorded that 1 trillion (1,000,000,000,000) unique URLs exist on the Web at once. Therefore, it is the authors' opinion that, because of the ever changing nature of the Web, depending on this value to calculate **M** might produce unreliable similarity scores over time.

In their paper, Chen et al. [12] propose a Web-based double checking method to find similar words. They collect snippets for two words *x* and *y* from the Web search engine and use these to count the number of occurrences of *x* in the snippets of *y*, and *y* in the snippets of *x*. The two values are then combined nonlinearly to compute the similarity between *x* and *y*. The problem with this method is that it relies heavily on the search engine's ranking algorithm. Although two words may be similar, it is not a guarantee that one will find *y* in the snippets of *x* or vice versa [10]. This may also have some effect on the final computed similarity measure.

There are many other proposed methods for finding word similarity using the Web, but none of the cited references in this paper uses the reported Web hit count in the way that is introduced in this paper. Our approach uses the Web and the Web search engines to automatically calculate semantic similarity between two terms, based on the number of hit counts reported for each terms (rather than for each hit).

The hit count of a query is usually an estimated number of Web pages containing the queried term as reported by a Web search engine. The hit count, however, may not necessarily be equal to the term frequency, because the queried term may appear many times on a single Web page. Therefore, an additional hit count is computed where a search term *x* and another search term *y* appear both on the same Web page, indicated as a logical *x* AND *y* search query. The search results of this query therefore, can be considered as the global estimated value of the co-occurrence of the terms *x* and *y* together on the Web [2]. Logical *x* AND *y* is also used in this study to capture the context where both *x* and *y* are used together on the same Web page.

The presentation in this paper is a new approach to using the hit counts to calculate semantic similarity. This observation is also confirmed by the experimental results based on a benchmark data set of words from Miller and Charles (1998) [13]. This data set is a subset of Rubenstein and Goodenoughs' [15] original data set of 65 word pairs. Although the Miller and Charles experiment was carried out 25 years later than that of Rubenstein and Goodenough, the two sets of ratings are highly correlated with a correlation coefficient of 0.97 (on a scale of 0 to 1, where 0 indicates no correlation and 1 indicates complete correlation) [7]. Therefore, the Miller and Charles ratings can be considered as a reliable benchmark data set for evaluating semantic similarity measures.

III. Technical Background

Much of the theory explained in this paper is based on computing the Euclidean distance between any two points in the Euclidean space, and its relationship with the computed absolute value of the difference of any two real numbers in the number line. Different distance functions result in different distance measures. However, the Euclidean distance used in this paper is considered the most useful because it corresponds to the way objects are measured in the real world [25]. For

more information on Euclidean distance and absolute value, please refer to [22] and [23] respectively.

A. *The Euclidean distance*

The Euclidean distance is defined as the distance between any two points in a plane that one would measure using a ruler and it is given by the Pythagorean Theorem [19], [20], [21] and [22]. If, for example, $x = (x_1, x_2)$ and $y = (y_1, y_2)$ are two given points on the plane, then their Euclidean distance (d) can be defined as [19],

$$\sqrt{(x_1\text{-}x_2)^2 + (y_1\text{-}y_2)^2} \qquad (1)$$

Using this formula as distance, the Euclidean space becomes a metric space also called the distance space.

For any given two points x and y the Euclidean distance between them is the length of the line segment connecting them. In a Cartesian coordinate, for example, if $x = (x_1, x_2,\ldots,x_n)$ and $y = (y_1,y_2,\ldots, y_n)$ are two points in the Euclidean space, then the distance (d) from x to y, or from y to x can be defined by equation 2,which can be seen as a generalization of equation 1 [19] and [20].

$$d(x,y) = d(y,x) = \sqrt{(x_1\text{-}y_1)^2 + (x_2\text{-}y_2)^2 + \cdots + (x_n\text{-}y_n)^2} \quad (2)$$

where n represents any number denoting a point x_n and y_n in the Cartesian coordinate.

The position of any point in a Euclidean n-space is usually called a Euclidean vector. Therefore, the points x and y can be referred to as Euclidean vectors. Starting from the origin of the space, their tips indicate the distance between the two points (also called the magnitude or the norm). The Euclidean norm or the length of a vector x is the real number denoted as $||x||$ [24] and measures the distance of x as defined by equation 3 [26]:

$$\left||x|\right| = (x.x)^{1/2} = \sqrt{x.x} \qquad (3)$$

The distance, therefore, between x and y can be computed as [26]:

$$d(x,y) = \left||x - y|\right| \qquad (4)$$

The Euclidean norm and distance may as well be expressed in terms of components as shown in equation 5 [24] and [26]:

$$\left||x|\right| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} = \sqrt{x.x} \qquad (5)$$

If the length of a vector is considered as the distance from its tail to its tip, then it becomes clear that the Euclidean length of a vector is a special case of the Euclidean distance. Therefore, the distance between x and y is the Euclidean length of the distance vector defined as [24]:

$$\left||x\text{-}y|\right| = \sqrt{(x\text{-}y).(x\text{-}y)} \qquad (6)$$

Equation 6 is homogeneous to equations 3, 4 and 5 and can be used to compute the magnitude or the norm of the numerical difference between any two real numbers x and y in the number line, denoted as $\left||x\text{-}y|\right|$.

It is also clear from equation 6 that the one-dimensional Euclidean distance between x and y can be realized.

1) *One-dimensional Euclidean distance*

In the case of one dimension, the distance between two points x and y on the real number line is equivalent to the absolute value of their numerical difference. Thus, if x and y represents two real numbers, then the distance between them can be computed as [27]:

$$\sqrt{(x\text{-}y)^2} = |x\text{-}y| \qquad (7)$$

In addition, in one-dimension there is usually a single homogeneous, translation-invariant distance function, which is the Euclidean distance and defines the distance between elements of a set. Translation-invariant implies that starting from the origin, at least in one direction, the object is infinite. In higher dimensions, up to n-dimensions, there are other possible distance functions but these are beyond the scope of this paper. We therefore consider only up to the two-dimensional Euclidean distance in this paper.

2) *Two-dimensional Euclidean distance*

In the Euclidean plane, if $x = (x_1, x_2)$ and $y = (y_1, y_2)$, then the distance (d) between the two points x and y is given by equation 8 [28], which is homogeneous to equations 1 and 2.

$$d(x,y) = \sqrt{(x_1\text{-}y_1)^2 + (x_1\text{-}y_2)^2} \qquad (8)$$

At this point the discussion of the Euclidean distance presents us with the foundation of establishing its relationship with the absolute value.

B. *The relationship between Euclidean distance and absolute value*

Having gained an understanding of the Euclidean distance, we now establish its relationship with the absolute value of any real number x denoted as $|x|$. The absolute value $|x|$ is the numerical value of x without regard to its sign. For example, the absolute value of $+x$ is x, and the absolute value of $-x$ is also x. This simply means, the absolute value of any real number x may be thought of as its distance from zero (i.e. how far x is from zero on the number line) [29], [30] and [31].

In practice, the absolute value of all real numbers is always positive. See equation 9. The concept of absolute value is closely related to the notion of distance in various mathematical and physical contexts. In this paper, therefore, the relationship between the Euclidean distance and the absolute value is established first. This relationship is then used to generate a formula that automatically calculates a semantic similarity measure (the distance) between any two terms x and y in the digital forensics domain. For any real number x its absolute value denoted by $|x|$ can be defined as shown in equation 9 [32]:

$$|x| = \begin{cases} x, & \text{if } x \geq 0 \\ -x, & \text{if } x < 0 \end{cases} \qquad (9)$$

Based on this definition, the absolute value of x is always either positive or zero, but never negative. In addition, the absolute value of the difference of any two real numbers x and

y defines the distance between x and y denoted as $|x - y|$ which is equivalent to the Euclidean distance of x and y. Since in mathematics the square-root of a number x without regard to its sign represents a positive square root, and the absolute value of x is always either positive or zero, but never negative, it follows that [32]:

$$|x| = \sqrt{x^2} \qquad (10)$$

Equation 10 is homogeneous to equation 7 and is sometimes used as a definition of the absolute value of any real number [32]. For any real numbers x and y, the absolute value will always have the following four fundamental properties [32]: See Figure 1.

$\|x\| \geq 0$	Non-negativity
$\|x\| = 0 \Leftrightarrow x = 0$	Positive-definiteness
$\|xy\| = \|x\|\|y\|$	Multiplicativeness
$\|x + y\| \leq \|a\| + \|b\|\|$	Sub-additivity

Figure 1: Fundamental properties of absolute value

C. *Deriving the similarity distance*

From the discussions above, it should now be clear that the absolute value of any real number is closely related to the idea of distance. The absolute value of any real number, therefore, is the distance from that number to the origin, along the real number line. For any given two real numbers x and y, the absolute value of the difference of x and y is the distance between them. The standard Euclidean distance between two points, for example x and y, defined in equation 2 affirms this, where $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$. In the Euclidean n-space, the distance is defined as [27]:

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (11)$$

Note that equation 11 is homogenous to equation 2. This can be viewed as a generalisation of $|x - y|$, since if x and y are two real numbers, then (from equation 10) we can define:

$$|x-y| = \sqrt{(x-y)^2} \qquad (12)$$

Equation 12 is homogeneous to equation 7 and equation 10. Equation 7 is used when computing the one-dimension Euclidean distance while equation 10 is used as a definition of the absolute value. Thus, equations 7, 10 and 12 prove that the 'absolute value' distance for any real numbers is equal to the Euclidean distance, defined in equation 7, when you consider them as either one and/or two-dimensional Euclidean spaces both defined in equation 7 and 8 respectively. Hence, the properties of the absolute value of the difference of any two real numbers (non-negativity, identity of indiscernibles, symmetry and the triangle inequality See Figure 2) agree with the concept of the distance function used to define the distance between the elements of a set. For any real value function f on a set X × X is called a distance function (or a metric) on X, if it satisfies the following four axioms [35] and [36]. See Figure 2.

$f(x, y) \geq 0$	Non-negativity (i)
$f(x, y) = 0 \Leftrightarrow x = y$	Identity of indiscernibles (ii)
$f(x, y) = f(y, x)$	Symmetry (iii)
$f(x, y) \leq f(x, z) + f(z, y)$	Triangle inequality (iv)

Figure 2: Distance function axioms

Note that condition (i) and (ii) together produce positive definiteness and the first condition is implied by the others. The technical background discussed in this section, especially the relationship between Euclidean distance and absolute value therefore simplifies the understanding of the proposed DFASSV method.

IV. The Proposed DFASSV Method

Hit counts reported by Web search engines are useful information sources for this study and, as such, are used as input to this study. This is first explained in the next section, where after the calculation of the DFASSV method is explained.

A. *Understanding the concept of 'hit Counts'*

The hit count of a query as discussed earlier, is an estimated number of Web pages containing the queried term as reported by a Web search engine.

In addition, the Web constitutes the largest semantic electronic "database" available on earth. Information can be accessed and extracted via any Web search engine that can return aggregate hit count estimates for a large range of search queries [1].The Web also provides semantic information for almost every known word or term. In some cases, semantics associated with each term or word is also described. In our approach, however, as explained earlier, we do not consider just the hit count for the logical x AND y search query as the only parameter for assessing the semantic similarity, but we also include the hit counts for the individual terms x and y before computing the semantic similarity value. We will, therefore, adopt the following notations in this paper:

$f(x)$ denotes the hit count for the queried term x,

$f(y)$ denotes the hit count for the queried term y and

$f(x, y)$ denotes the hit count for the logically x AND y search query where both x and y appears together on the same Web page.

To calculate the Digital Forensic Absolute Semantic Similarity Value of x and y, denoted as DFASSV(x, y), we do not need to know the number of Web pages indexed by the Web search engine quoted as 8058044651 in [1]. This is so because according to [14] the process of estimating the number of pages indexed by a search engine can be a very difficult task. This paper, however, does not discuss the process of estimating the number of pages indexed by a search engine in any further detail. (For more information in this regard, please refer to [11]).

Our approach however, replaces the number of pages indexed by a search engine with a simple computed value (**T**) defined as the sum of the hit counts reported by the Web

search engine for the search terms *x*, *y* and logical *x* AND *y* together.

Thus, $\mathbf{T} = f(x) + f(y) + f(x,y)$ (13)

where $f(x)$, $f(y)$ and $f(x, y)$ are as defined earlier.

Recalling the concept of the Euclidean distance and the absolute value at this point, we now establish their relationship with the proposed DFASSV method.

B. *Digital Forensic Absolute Semantic Similarity Value (DFASSV)*

In order to enhance communication among domain experts and also enable faster computation of meaning between computers in a computer digestible form, many long-term projects have been initiated to try and establish semantic relations between common objects and/or names of these objects. Good examples of these projects include the CYC project [3] and the WordNet [4]. The idea is to create a semantic Web of such vast proportions that rudimentary intelligence and knowledge about the real world objects emerge spontaneously. However, to achieve this, structures have to be properly designed with the ability to manipulate knowledge, and high quality contents have to be entered in these structures by knowledgeable human experts. While these efforts are good and take a long-term view, the overall information entered is very small when compared to what is available on the Web today [1]. We, therefore, take advantage in this study of the freely-available information on the Web and use it to calculate a semantic similarity measure between terms used in the digital forensics domain.

The proposed method in this paper, computes the semantic similarity value between two terms *x* and *y* in digital forensics, based on finding the one-dimensional Euclidean distance defined in equation 7, which is equal to finding the absolute value of the difference of any two real numbers. See equations 7 and12.

To begin with, the hit counts $f(x)$, $f(y)$, $f(x, y)$ and the value of **T** for any two digital forensic terms *x* and *y* is obtained using the Google search engine. These parameters are then used as input to the proposed DFASSV method. **T** is however, computed using equation 13. There are four input parameters defined as $f(x)$, $f(y)$, $f(x, y)$ and **T**. Using equation 12, which is similar to one-dimensional Euclidean distance (see equation 7); only two real numbers are needed as input. In order to establish a 1:1 mapping of the values of *x* and *y* in equation 12, DFASSV replaces the values of *x* and *y* with the percentage values of $f(x)$ and $f(y)$ computed as:

$\left(\frac{f(x)}{\mathbf{T}} * 100\right)$= percentage of the hit counts for the search term *x* and

$\left(\frac{f(y)}{\mathbf{T}} * 100\right)$= percentage of the hit counts for the search term *y*.

Substituting these values in equation 12 gives equation 14

$$|x - y| = \sqrt{\left(\left(\frac{f(x)}{\mathbf{T}} * 100\right) - \left(\frac{f(y)}{\mathbf{T}} * 100\right)\right)^2} \qquad (14)$$

The value obtained from equation 14 is in the fixed range of 0 per cent to 100 per cent. Treating the points *x* and *y* as Euclidean vectors, starting from the origin (0%) of the space, their tips (100%) indicate the distance between the two points.

As mentioned earlier, in one-dimensional Euclidean distance there is usually a single homogeneous, translation-invariant distance function (i.e. starting from the origin at least in one direction the object is infinite). For a similarity distance of 0 to ∞ instead of 0 per cent to 100 per cent, equation 14 is further modified as follows:

The values $\left(\frac{f(x)}{\mathbf{T}} * 100\right)$ and $\left(\frac{f(y)}{\mathbf{T}} * 100\right)$, denoted as percentage of the hit counts for the search term *x* and *y* respectively, are substituted by their computed logarithms as:

$\log\left(\frac{f(x)}{\mathbf{T}} * 100\right)$ and $\log\left(\frac{f(y)}{\mathbf{T}} * 100\right)$ respectively

Logarithm is a useful arithmetic concept used in all areas of science to help simplify the understanding of many scientific ideas. For example, logarithms may be defined and introduced in different ways as a means to simplify calculations. For the purposes of this study, we adopt a simple approach to simplify the computation of the Euclidean distance based on finding the absolute value of the difference of the logarithms of the hit count percentages of the terms *x* and *y*. There are no limits imposed on logarithms, thus their inputs and outputs can be in any range. Therefore, substituting these values in equation 14 gives rise to equation 15:

$$|x - y| = \sqrt{\left(\log\left(\frac{f(x)}{\mathbf{T}} * 100\right) - \log\left(\frac{f(y)}{\mathbf{T}} * 100\right)\right)^2} \qquad (15)$$

Equation 14 and 15 are both analogous to equation 7 and 12.

Equation 15 therefore, gives a value in the range of 0 to ∞ and can be re-written as equation 16, which is used to automatically compute the Absolute Semantic Similarity Value of the terms *x* and *y* in digital forensics denoted as DFASSV(*x*, *y*). Using the left hand side of equation 15, equivalent to the right hand side we can define DFASSV(*x*, *y*) as,

$$\text{DFASSV}(x, y) = \left|\log\left(\frac{f(x)}{\mathbf{T}} * 100\right) - \log\left(\frac{f(y)}{\mathbf{T}} * 100\right)\right| \qquad (16)$$

where

$f(x)$ = the hit counts for the search term *x*, $f(y)$ = the hit counts for the search term *y* and **T** = the sum of hit counts for the search terms *x* and *y* as defined in equation 13.

Equation 16, therefore, defines DFASSV(*x*, *y*), a new approach for calculating the semantic similarity between two terms *x* and *y* in digital forensics using Web search engines. In other words equation 16 denotes DFASSV as the computed absolute value of the difference of the logarithms of the hit count percentages of terms *x* and *y*. The experimental results obtained using the new proposed DFASSV approach was found to be remarkable and are discussed in the section that follows.

V. EXPERIMENTAL RESULTS

While the theory discussed in this paper is rather intricate, the resulting method is simple enough. Knowing that any given two digital forensics terms are perceived to be similar, the computed absolute semantic similarity value denoted by

equation 16 can be used as a quick guide (proof) to show that the two given terms are truly semantically similar or not.

For example, given any two digital forensic terms *x* and *y*, we find the number of hit counts for search term *x* denoted as *f*(*x*), the number of hit counts for search term y denoted as *f*(*y*), the number of hit counts for logical *x* AND *y* both appearing together on one page denoted as *f*(*x*, *y*), and finally the sum of hit counts denoted as (**T**). **T** is computed using equation 13.

As a concrete example, let search term *x* be 'Digital evidence' and search term *y* be 'Electronic evidence'. Using the Google search engine with hit counts as reported for the search terms *x* and *y* as on 14 April 2012, it follows that:
"Digital evidence" *f*(*x*) =659000,
"Electronic evidence" *f*(*y*) =575000,
"Digital evidence"AND "Electronic evidence" *f*(*x*, *y*) =53900.
Therefore **T**= $f(x) + f(y) + f(x,y)$ =1287900.

Substituting these values in equation 16 gives a semantic similarity measure of the terms 'Digital Evidence' and 'Electronic evidence' of **0.0592**. Since this value is relatively close to 0, it proves that the two terms are very closely related to the human-perceived meaning when used in digital forensics. It can also mean that, in case of a digital forensics investigation, the term 'Digital Evidence' can be used in the place of 'Electronic Evidence' without misleading the receivers of such information.

To further analyse the performance of the proposed method, we conducted two sets of experiments. First we compared the similarity scores produced by the proposed DFASSV method against the Miller and Charles benchmark data set [13] and [14]. Secondly, the proposed DFASSV approach was tested using digital forensics domain terms to measure its performance against the human-perceived meaning of the selected terms. These two experiments are discussed in the two sub-sections that follow respectively.

A. The Miller and Charles Benchmark Data Set

To assess the performance of the proposed DFASSV method, we evaluated it against the Miller and Charles data set [13]. The latter is a subset of Rubenstein and Goodenough's original data set of 65 word pairs [15]. As stated earlier, the Miller and Charles ratings are considered one of the most reliable benchmarks for evaluating semantic similarity measures.

The term pairs using the proposed DFASSV method are rated on a scale of 0 to ∞ (infinite), where 0 means identical semantic similarity and ∞ means no similarity. This is the opposite of the Miller and Charles dataset where word pairs are rated on a scale of 0 (dissimilarity) to 4 (identical semantic similarity). In summary, infer from the results that the smaller the value computed by the proposed DFASSV method, the more similar the terms (See Table I).This is also true from the correlation coefficient value of **-0.2777.** (Note that a negative correlation coefficient indicates that as one variable increases, the other decreases, and vice-versa.) This is further depicted by a graphical representation of the similarity measures in Table I, shown in Figure 3.

According to Cilibrasi and Vitànyi [1] Google events capture all background knowledge about the search terms concerned available on the Web. The Google event *x*, consists of a set of all Web pages containing one or more occurrences of the search term *x*. Thus, it embodies, in every possible sense, all direct context in which *x* occurs on the Web. This constitutes the Google semantics of the term *x* [1]. For this reason, in our experiments, the Google search engine was used.

The input to the DFASSV method is therefore the reported Google hit counts for any paired terms *x* and *y* from the digital forensics domain. The DFASSV method works by calculating the Euclidean distance between the terms *x* and *y*, equated to the computed absolute value of the difference of the logarithms of the hit count percentages of *x* and *y* as shown earlier in equation 16. Given any two terms *x* and *y* as points in the Euclidean plane, the associated computed absolute value of the difference of the logarithms of the hit count percentages of *x* and *y*, determines the similarity between the terms *x* and *y*.

Word Pair	**M&C**	**Web Jaccard**	**Web Dice**	**Web Overlap**	**Web PMI**	**Proposed DFASSV**
cord-smile	0.13	0.102	0.108	0.036	0.207	0.756
rooster-voyage	0.08	0.011	0.012	0.021	0.228	0.828
noon-string	0.08	0.126	0.133	0.060	0.101	0.524
glass-magician	0.11	0.117	0.124	0.408	0.598	1.399
monk-slave	0.55	0.181	0.191	0.067	0.610	0.389
coast-forest	0.42	0.862	0.870	0.310	0.417	0.055
monk-oracle	1.1	0.016	0.017	0.023	0	0.457
lad-wizard	0.42	0.072	0.077	0.070	0.426	0.400
forest-graveyard	0.84	0.068	0.072	0.246	0.494	1.258
food-rooster	0.89	0.012	0.013	0.425	0.207	1.778
coast-hill	0.87	0.963	0.965	0.279	0.350	0.248
car-journey	1.16	0.444	0.460	0.378	0.204	0.865
crane-implement	1.68	0.071	0.076	0.119	0.193	0.418
brother-lad	1.66	0.189	0.199	0.369	0.644	0.970
bird-crane	2.97	0.235	0.247	0.226	0.515	0.051
bird-cock	3.05	0.153	0.162	0.162	0.428	0.024
food-fruit	3.08	0.753	0.765	1	0.448	0.223
brother-monk	2.82	0.261	0.274	0.340	0.622	0.966
asylum-madhouse	3.61	0.024	0.025	0.102	0.813	0.945
furnace-stove	3.11	0.401	0.417	0.118	1	0.180
magician-wizard	3.5	0.295	0.309	0.383	0.863	0.638
journey-voyage	3.84	0.415	0.431	0.182	0.467	0.238
coast-shore	3.7	0.786	0.796	0.521	0.561	0.411
implement-tool	2.95	1	1	0.517	0.296	0.838
boy-lad	3.76	0.186	0.196	0.601	0.631	0.271
automobile-car	3.92	0.654	0.668	0.834	0.427	0.975
midday-noon	3.42	0.106	0.112	0.135	0.586	0.855
gem-jewel	3.84	0.295	0.309	0.094	0.687	0.027
Correlation	**1**	**0.259**	**0.267**	**0.382**	**0.548**	**-0.27 77**

TABLE I.

COMPARISON OF SEMANTIC SIMILARITY OF HUMAN RATINGS AND BASELINES ON MILLER AND CHARLES' DATASET WITH DFASSV

The distance measure shown in Table II depicts the relatedness of the terms in question. Table I on the other hand,

was used mainly for the purpose of comparison in order to indicate different semantic similarity measures from previously-proposed methods compared to those of the proposed DFASSV method. This was done to provide a clear picture of the performance and accuracy of DFASSV.

From Table I, the first column shows the word pairs used and column two indicates the ratings from Miller and Charles. Columns 3 to 6 also used for comparison show the ratings from previously proposed sematic similarity methods and the last column depicts the equivalence similarity measure computed using the proposed DFASSV. For example, the word pair 'gem-jewel' (See Table I) with a similarity measure of **3.84** from Miller and Charles and **0.027** from the proposed DFASSV clearly depicts the accuracy of DFASSV. This is also depicted in the other columns and indicates a better performance than that of some of the previously proposed methods. See Figure 3 for a graphical representation of the Table I results.

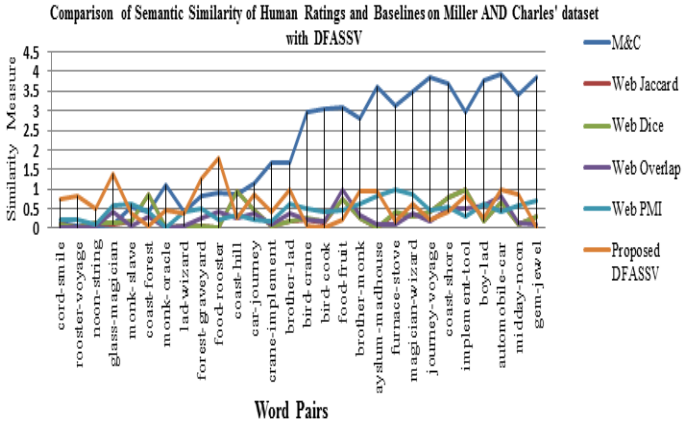

Figure 3: Comparison graph of the semantic similarity ratings and baselines on Miller and Charles' dataset with DFASSV (See Table I).

B. *Digital Forensics Terminologies*

In Table II, a part of the experimental findings is presented using the digital forensics domain terminologies. Each term enclosed in double quotes " " is used as a single Google search term denoted in Table II as $f(x)$ and $f(y)$ respectively. The computed DFASSV(x, y) using equation 16 therefore shows the semantic similarity measures obtained to ascertain the performance of the DFASSV method with the human-perceived meaning of the terms. The authors have no knowledge of other experiments of this kind in the digital forensics domain that can be used as a baseline to judge the performance of DFASSV. This is, therefore, a novel approach of using a Web search engine to determine the semantic similarity of terms in digital forensics.

The selected terms used are: 'digital evidence', 'digital forensics', 'electronic evidence', 'digital and multimedia evidence' [33] and [34] among other terms. The authors found that these terms are mostly used in discussions that involve the digital forensic investigation process and also in the accreditation of digital forensics laboratories, hence the motivation for the experiment indicated in Table II. In all the experiments conducted, DFASSV showed remarkable results.

To determine the semantic similarity measure of the terms as shown in Table II, the proposed DFASSV was used in all our experiments. The first two columns of Table II shows the digital forensics terms used for the experiments and their equivalent similarity measure indicated in the last column. Using the results in Table II, a random interview was conducted to a few digital forensics researchers and their understanding of these terms seemed to agree with the results of the proposed DFASSV method.

Digital Forensics Terms		**Computed *DFASSV(x, y)***
f(x)	***f(y)***	
Digital evidence	Electronic evidence	0.059217
Digital forensics	Digital evidence	0.431534
Digital forensics	Electronic evidence	0.490752
Electronic evidence	Digital and multimedia evidence	1.833840
Digital evidence	Digital and multimedia evidence	1.893057
Digital forensics	Digital and multimedia evidence	2.324592
Attacker	Adversary	0.357051
Cracker	Attacker	0.361608

TABLE II.

SEMANTIC SIMILARITY RATINGS OF DIGITAL FORENSIC TERMS BASED ON DFASSV

In the case of a digital forensic investigation for example, DFASSV can be used to determine the usage of terms where a similarity measure closer to 0 means that the two terms are closely related in meaning. The terms 'Digital evidence' and 'Electronic evidence', for example, with a similarity measure of **0.059217** indicates that they can be used interchangeable without causing confusion to the stakeholders. On the other hand a semantic similarity measure far from 0 would mean that the two terms are not closely related in meaning and therefore, one cannot replace the other. For example the terms 'Digital forensics' and 'Digital and multimedia evidence' with a similarity value of **2.324592** means they cannot be used interchangeable.

C. *Application of The Proposed DFASSV Method in the Digital Forensics Domain*

The proposed DFASSV method as demonstrated in this paper can be used in the digital forensics domain for example, to determine the semantic relatedness of terms and also as a way towards resolving the semantic disparities that exist in the domain. In addition, DFASSV can be used to help determine the most relevant and appropriate terminologies to use or included for example when building a specific ontology in the digital forensics domain. In addition, other future relevant undertakings in the digital forensics domain, in the authors' opinion, might as well benefit from applying such a method as DFASSV.

VI. CONCLUSION

The problem that this paper addressed was that of the ever-changing technological trends in digital forensics where new terms are constantly introduced into the domain and new meanings assigned to existing terms.

In this paper a method was presented to automatically calculate a semantic similarity value between any two given digital forensics terms, using a new approach. Unlike previous methods, the Digital Forensic Absolute Semantic Similarity Value (DFASSV) approach proposed in this paper is unsupervised. No special background information is needed to understand and use this method because it utilises the existing bank of information from the Web by simply incorporating the hit counts between two digital forensics terms reported by any Web search engine. In addition, the authors also found that DFASSV is well suited for terminologies that originate from within the same domain.

Though the initial experiments were carried out on the digital forensics domain terms, the authors believe that the DFASSV method can be extended to other domains as well. This is due to the fact that the results of the experiments conducted to evaluate this method using the digital forensics domain terminologies are remarkable. The results show that this approach of measuring semantic similarity between two terms significantly outperforms some of the previous proposed measures.

As part of future research work, the authors are now planning to conduct an investigation in order to find out whether there are existing parameters other than hit counts reported by search engines that can be used with DFASSV to enhance the accuracy delivered by this method even more as a way towards resolving semantic disparities in the digital forensics domain

REFERENCES

[1] R.L. Cilibrasi and P.M.B. Vitànyi. 2007. The Google Similarity Distance. IEEE Transactions on Knowledge and Data Engineering, Vol. 19, No 3, March 2007, pp. 370–383.

[2] D.Thiyagarajan, N. Shanthi and S. Navaneethakrishnan, A Web Search Engine-Based Approach To Measure Semantic Similarity Between Words. International Journal of Advanced Engineering Research and Studies. E-ISSN2249–8974.

[3] D.B. Lenat, CYC: A large-scale investment in knowledge infrastructure, Communications of the ACM, November 1995/Vol. 38, No. 11.

[4] G.A. Miller, WordNet, A Lexical Database for the English Language, Cognitive Science Lab, PrincetonUniversity.

[5] Zheng Xu et al. 2011. Measuring semantic similarity between words by removing noise and redundancy in web snippets. Concurrency and Computation: Practice and Experience, 23:2496–2510. Published online 22 September 2011 in Wiley Online Library (wileyonlinelibrary.com) DOI: 10.1002/cpe.1816.

[6] R. Rada, H. Mili, E. Bichnell and M. Blettner. 1989. Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man and Cybernetics, 9(1):17–30.

[7] S. Vijay, 2012. A Combined Method to Measure the Semantic Similarity between Words, International Journal of Soft Computing and Engineering (IJSCE), ISSN: 2231–2307, Volume 1, Issue ETIC2011, January 2012.

[8] Y. Matsuo, T. Sakaki, K. Uchiyama and M. Ishizuka. 2006. Graph-based word clustering using web search engine. In Proceedings of EMNLP 2006.

[9] M. Li, X. Chen, X. Li, B. Ma and P.M.B. Vitànyi. 2004. The similarity metric. IEEE Transactions on Information Theory, 50(12):3250–3264.

[10] D. Bollegala, Y. Matsuo and M. Ishizuka. 2009. A Relational Model of Semantic Similarity between Words using Automatically Extracted Lexical Pattern Clusters from the Web.

[11] Anon., We knew the web was big... | Official Google Blog. Available at: http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html [Accessed April 17, 2012].

[12] H. Chen, M. Lin and Y. Wei. 2006. Novel association measures using web search with double checking. In Proceedings of the COLING/ACL 2006, pp. 1009–1016.

[13] G.A. Miller and W.G. Charles. 1998. Contextual correlates of semantic similarity. Language and Cognitive Processes. Volume 6, Issue 1.

[14] Z. Bar-Yossef and M. Gurevich. 2006. Random sampling from a search engine's index. In Proceedings of the 15th International World Wide Web Conference.

[15] H. Rubenstein and J.B. Goodenough. 1965. Contextual Correlates of Synonymy. Computational Linguistics. Decision Sciences Laboratory, L.G. Hanscom Field, Bedford, Massachusetts.

[16] W.G. Charles, Contextual Correlates of Meanings. Applied Psycholinguistics 21 (2000) 505–524. Cambridge Journals. Available at: http://journals.cambridge.org/action/displayFulltext?type=1&fid=66932&jid=APS&volumeId=21&issueId=04&aid=66931 [Accessed April 20, 2012]. p. 514.

[17] Anon., Simpson's Rule. Available at: http://pages.pacificcoast.net/~cazelais/187/simpson.pdf [Accessed May 7, 2012].

[18] Anon., Distance and Similarity Coefficients. Available at: http://paleo.cortland.edu/class/stats/documents/11_Similarity.pdf [Accessed May 7, 2012].

[19] P. Sanchez, R. Milson and M. Slone. Euclidean distance (version 11). Available at: http://planetmath.org/EuclideanDistance.html [Accessed May 7, 2012].

[20] A. Bogomolny, Pythagorean Theorem and its many proofs.Interactive Mathematics Miscellany and Puzzles. Available at: http://www.cut-the-knot.org/pythagoras/index.shtml [Accessed May 7, 2012].

[21] A. Bogomolny, The Distance Formula.Interactive Mathematics Miscellany and Puzzles. Available at: http://www.cut-the-knot.org/pythagoras/DistanceFormula.shtml [Accessed May 7, 2012].

[22] D.T. Larose. 2005. Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley & Sons, Hoboken, NJ.

[23] G.B. Dantzig and M.N. Thapa. 1997. Linear Programming: Introduction. p. 147. Section 6.3. Hamilton Printing, Rensselaer, NY.

[24] Anon, Vectors in Euclidean Spaces. Available at: http://scottmccracken.weebly.com/uploads/9/0/6/6/9066859/vectors-print_version.pdf [Accessed May 10, 2012].

[25] D.G Bailey, An Efficient Euclidean Distance Transform**,** IWCIA 2004, LNCS 3322, pp. 394–408, 2004.

[26] Anon, Complex Vector Spaces and Inner Products. Available at: http://college.cengage.com/mathematics/larson/elementary_linear/4e/shared/downloads/c08s4.pdf [Accessed May 11, 2012].

[27] R. Balu and T. Devi, Identification Of Acute Appendicitis Using Euclidean Distance On Sonographic Image. International Journal of Innovative Technology & Creative Engineering (ISSN: 2045-8711**).** VOL.1 NO.7 JULY 2011

[28] Per-Erik Danielsson, Euclidean Distance Mapping, Computer Graphics and Image Processing 14, 227-248 (1980)

[29] Anon, Absolute Value. Available at: http://www.purplemath.com/modules/absolute.htm [Accessed May 11, 2012].

[30] Anon, Absolute Value Functions. Available at: http://hotmath.com/hotmath_help/topics/absolute-value-functions.html [Accessed May 11, 2012].

[31] Anon, Absolute Value Functions. Department of Mathematics, College of the Redwoods Available at: http://msenux.redwoods.edu/IntAlgText/chapter4/chapter4.pdf [Accessed May 11, 2012].

[32] Anon,Absolute Value. Available at: http://math.ucalgary.ca/sites/math.ucalgary.ca/files/courses/F07/MATH251/lec5/MATH251-F07-LEC5-Appendix_E.pdf [Accessed May 11, 2012].

[33] G. Palmer, A Road Map for Digital Forensic Research. DFRWS TECHNICAL REPORT. DTR - T001-01 FINAL. Report from the First Digital Forensic Research Workshop (DFRWS). November 6th, 2001 - Final.

[34] NATA, Proficiency Testing Policy in the Field of Forensic Science. Available at: http://www.nata.asn.au/phocadownload/publications/Technical_publications/Policy_Tech_circulars/technical-circular-15.pdf [Accessed March 12, 2012].

[35] J. Fennell, Axiomatics Through The Metric Space Axioms**.** Metric Space Axiomatics. University College Cork

[36] T. Margush, Distances Between Trees**.** Discrete Applied Mathematics 4 (1982) 281-290North-Holland Publishing Company.