New Similarity

K M N Mahyuddin¹, O S Sitompul², S Nasution³, and H Ambarita⁴

¹,²Informatin Technology, Fasilkom-TI, Universitas Sumatera Utara, Padang Bulan USU Medan 20155, Indonesia.
³Program Studi Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Sumatera Utara, Padang Bulan USU Medan 20155, Indonesia.
⁴Sustainable Energy Research Centre, Faculty of Engineering University of Sumatera Utara, Jl. Almamater Kampus USU Medan 20155, Indonesia.

*mahyuddin@usu.ac.id

Abstract. Either space or object consists of parts, where a part of it is interchangeable with part from other objects based on similarities in the nature of the parts. However, the similarity of the two objects in advance must quantified to determine the difference and the closeness. There are many ways to measure similarity and dissimilarity, but there is no measurement of the similarity of two objects by the parts semantically. In this article, we will reveal new similarity formally based on the concept of multilevel sections and triangular equation. The new similarity applied to the URL address to measure the similarity of identity of web pages.

1. Introduction
Each space consists of subspaces (sections), as well as each object contains the components be builder itself [1]. In each object, either the largest or the smallest components have possible sameness with the components of the other objects. Based on that principle, it is possible that two objects to be similar or not, as well as the spaces [2]. Similarity of two objects is measured based on the concept of proximity, which is used to deliver the limited value in the range [0,1]. The use of this range expressed in fuzzy theory, which functions in the information management semantically [3]. There are many similarity measurements, but it involves a sequence of components of the objects needs a special approach [4]. This paper aims to express formally something in common whereby the possibility of the components may sorted.

2. Basic Concept and Motivation
Generally similarity measurement is based on the concept of distance [5,6]. Each distance has opposition to the similarity. Thus, if the concept of distance grouped into different principles and uses, the similarities are also in the same group [7]. The distance based on the mathematical theory: Although the distance is always familiar by the principles of geometry, but in classical mathematic the distance involves concepts such as algebra; strings; numbers; polynomials; matrices; analysis functions; and probability, while based on the application the distance is categorized according to graph theory or coding theory [4]. The distance is also related to the world of computers that function in the Internet, for example [4,8]. Likewise, there are distances for cosmological, geography, physics and biology. However, the distances are also related to the distance scale and figurative sense [4,9].
In the world of information technology, extracting the social network involves the basic measuring the distance between two social actors [8,10], which can be used to trace the origin of the information [11], or it can be used to predict the social behavior [12]. While in the scientific world, the distance measurement is refer to the similarity between the documents and improperly disclose the possible behavior of scientific writing is plagiarism. The semantic technology or knowledge technology in data mining, it includes the problems about big data, require techniques involving resemblance for clustering data or the documents in certain groups [13]. Therefore, the development of literacy in general and the semantic technology needs the similarities based measurements.

No matter small or big, either object or space is composed of parts called components or subspaces section [14]. Formally, for example, object A is composed of components a or may be disclosed that a set A contains a, i = 1,…,n. Indirectly a set A can also be broken down into subspaces based on the components i.e. empty set, subsets contains one component of A, and so on, or {{a1},{a,a2},…} [15]. Each component in A has advantages or disadvantages compared to other components if an operation implemented, or between one with another have a relationship so that for all relationships we need a consideration for determining a way to find the components are closer to the others [16]. Therefore, on a sequence of components that likely to be existed in two or more objects need to be measured similarity between the rows of components of two or more objects.

3. An Approach: Toward Similarity
The concept of similarity relates to the symmetry form like line, equilateral triangle, squares or rectangles. One of symmetrical forms is the right triangle where there are three sides a, b, and c, where c as the hypotenuse, and then

\[ c^2 = a^2 + b^2 \] (1)

Therefore, each parameter of similarity measurement directed to the square form (the parameters rank two). First, we define the representation of any objects or any space.

Definition 1. A set A is a representation of object with an operation \( \rightarrow \) which assigns to each pair of components a,b in A exactly a relation \( h = (a \rightarrow b) \).

As an objects a set A consists of \( a_i, i = 1,\ldots,n \), or a space a set A can be divided into subspaces of A. The size of A is \( |A| = n \).

Lemma 1. If a set \( A = \{a_i|i=1,\ldots,n\} \) is the representation of an object where \( a_i \) are the components of A, then each component has the relations with others components.

Proof. Let \( |A| = n \), then we get \( n(n-1) \) relations or there is a set of relations \( H = \{h_j|j=1,\ldots,m\} \) for \( a_i \) in A \( i=1,\ldots,n \). Each relation can be expressed in binary \{0,1\} for justifying membership of two components in A. Thus, each components has at least one relation with another.

Lemma 2. If a set \( A = \{a_i|i=1,\ldots,n\} \) is the representation of an object, then the components in A have different or same relation between one to another.

Proof. In binary, \( h=0 \) or \( h=1 \) for each pair of components in A. Let \( R \) is real values for weights \( w \) of \( (a_i \rightarrow a_j) \) where \( a_i,a_j \) in A or \( w(a_i \rightarrow a_j) \), we have three: (a) \((a_i,a_j) > (a_j,a_i)\), (b) \((a_i,a_j) = (a_j,a_i)\), or (c) \((a_i,a_j) < (a_j,a_i)\). Thus \( w(h_1) > w(h_2), w(h_1) = w(h_2), \) or \( w(h_1) < w(h_2) \), for all \( h_1, h_2, h_3 \) in \( H \).

Proposition 1. Let \( H = \{h_j|j=1,\ldots,m\} \) is a set of relations between the components in A, then the relations between components be a sequence.

Proof. Each relation between two components in A has a weight. The relation weights in range strongest to weakest, and the relations be one order (Lemma 2). Each relation has two points (Lemma 1): the beginning and the end. The strongest relation be first part of sequence, and then the stronger
relation be second part, and so forth. In case, there are the ends of two or more relations with common beginning, then we throw the weakest relation, or if there are two or more same relations we can choose one of them so that we get one sequence of $h_j, j=1,...,m$. Thus, all of relations exist in a sequence.

**Definition 2.** Let $Z = \{z_1, z_2, \ldots, z_2\}$ is a set of attributes, and a set of transaction $M_i$ are subsets of attributes, or $Mi$ be subset of $Z$. The implication $X \rightarrow Y$ with two possible values $T =$ TRUE or $F = $ FALSE as an association rule if $X$, $Y$ are subset of $Z$ and $X \cap Y$ is the empty set.

**Theorem 1.** If $A$ represents an object, then there is a sequence of components in $A$ based on the association rule.

**Proof.** The direct consequence of Proposition 1 to Definition 2.

### 3.1. Symmetry form

**Definition 3.** Let $H$ is a vector space. Each pair of $h_i, h_j$ in $H$ has a resultant $r_h$ as follow

$$r_h = h_i^2 + h_j^2 + 2h_i h_j \cos \alpha$$

(2)

where $\alpha$ is angle between $h_i$ and $h_j$.

**Corollary 1.** If $H$ is a set of relations for components in $A$, then the resultant of relations is

$$r_h = \Sigma_{j=1\ldots m} h_j^2.$$  

(3)

**Proof.** To address the symmetrical forms, we assume that the angle $\alpha$ between two vectors in $H$ is $90^\circ$. Therefore, based on Equation (2) in Definition 3 we have $r_{h12} = h_1^2 + h_2^2$ for $h_1, h_2$ in $H$. Recursively based on Theorem 1 we have $r_{h123} = r_{h12} + h_3^2 = h_1^2 + h_2^2 + h_3^2$ for $h_1, h_2, h_3$ in $H$, or for all $h_j$ in $H$, $j=1,\ldots, m$, $r_h = h_1^2 + h_2^2 + \cdots + h_m^2 = \Sigma_{j=1\ldots m} h_j^2$ or Equation (3).

In this case, for each vector $h$ in $H$ we assume that $h = w(h)$, or each weight of relation be a distance among two components of $A$.

**Corollary 2.** If $H$ is a set of relations for components in $A$ and $h_i = h_j$, then the resultant of two relations $h_i, h_j$ in $H$ is

$$r_h^2 = 2h_i h_j$$

(4)

**Proof.** To address the symmetrical forms, we use the concept of the right triangle where $a = b$. Thus, we use Equation (1) to adapt the position of two vectors $h_i, h_j$ in $H$ such that $r_h$ as the hypotenuse $c$. Let $h_i = h_j$, we have $r_h^2 = h_i^2 + h_j^2 = 2h_i^2 = 2h_i h_i$, or we reposition one of $h_i h_j$ be $h_j$, so $r_h^2 = 2h_i h_j$ or Equation (4).

In recursive, the general form of Equation (4) for each $h_i = h_j, i=1,\ldots,n, j=1,\ldots,m$, is

$$r_h^2 = 2\Sigma_{i=1\ldots n, j=1\ldots m} h_i h_j.$$  

(5)

### 3.2. Proof of new similarity
As expressed by Equation (1) we obtain a normalized compression distance between the hypotenuse and two sides of the right triangle, i.e.

\[ c^2 - (a^2 + b^2) = 0, \]

while the comparison between the hypotenuse and two sides of the right triangle we get normalized information distance as follows

\[ c^2/(a^2+b^2) = 1. \]

Let us can transform a distance (dissimilarity) \( d \) from Equation (6) and Equation (7), then we can define the distance as follow [4].

**Definition 4.** A distance is

\[ d = 1 - c^2/(a^2 + b^2) \]

where \( a > 0 \) and \( b > 0 \).

Based on Equation (1) \( d = 0 \). Suppose \( c^2 \) and \( a^2 + b^2 \) are not same or \( c^2 < a^2 + b^2 \), then \( 0 \leq d \leq 1 \). With that, we define a proximity or similarity as follow [4].

**Definition 5.** Let \( A \) be a set as representation of an object. A function \( s : A \times A \rightarrow R \) is called similarity on \( A \) if \( s \) is non-negative, symmetric, and if \( s(a,b) \leq s(a,a) \) holds for all \( a,b \) in \( A \), with equality if and only if \( a = b \). Or,

\[ s = 1 - d. \]

**Theorem 2.** [8,1,17] For all objects \( A \) in \( O \), the similarity of \( A_i \) and \( A_j \) in \( O \) is

\[ s = 2|ab|/(|a|+|b|), \text{ } s \text{ in } [0,1] \]

where \( |ab| \) is the cardinality of \( |a\cap b| \), also \( |a| \) and \( |b| \) are the cardinalities of vectors of \( A_i \) and \( A_j \), respectively.

**Proof.** By substituting Equation (8) into Equation (9) we have the formula as follows

\[ s = c^2/(a^2 + b^2) \]

Based on Equation (8) in Corollary 1, Equation (9) in Corollary 2, and Equation (1) we can substitute Equation (3) and Equation (4) into Equation (11) where \( c^2 = 2hh_j \). Or based on Equation (5) \( c^2 = 2\Sigma_{i=1,\ldots,n}h_ih_j \) for \( h_i = h_j \), and based on Equation (4) \( a^2 = \Sigma_{i=1,\ldots,n}h_i \) and \( b^2 = \Sigma_{j=1,\ldots,m}h_j \), we have

\[ s = (2\Sigma_{i=1,\ldots,n}h_ih_j)((\Sigma_{i=1,\ldots,n}h_i + \Sigma_{j=1,\ldots,m}h_j), \text{ } s \text{ in } [0,1] \]

or \( s = (2|ab|)/(|a|+|b|) \) where \( |ab| = \Sigma_{i=1,\ldots,n}h_ih_j \), \( |a| = \Sigma_{i=1,\ldots,n}h_i \) and \( |b| = \Sigma_{j=1,\ldots,m}h_j \).

### 4. Discussion

Tim Berners-Lee [18] has set that each web page has a URL address in canonical form as follows [17]:

\[ u = \{s, d, p, q\} = \{\text{scheme, authority, path, query}\} \text{ in a line as follows} \]

\[ \text{http(s)://d}_{\text{m}}\text{.d}_{\text{2}}\text{.d}_{\text{1}}\text{/p}_{\text{1}}\text{/p}_{\text{2}}/\ldots/\text{p}_{n-2}/x \]
we have two groups of URL, i.e.

(i) ieeexplore.ieee.org (m=1), ieeexplore.ieee.org/iel5 (m=2), ieeexplore.ieee.org/iel5/5976947 (m=3), ieeexplore.ieee.org/iel5/5976947/5995747 (m=4), and ieeexplore.ieee.org/iel5/5976947/5995747/05995766.pdf (m=5).

(ii) ieeexplore.ieee.org/iel7/7562830 (m=3), ieeexplore.ieee.org/iel7/7562830/7577456 (m=4), and ieeexplore.ieee.org/iel7/7562830/7577456/07577513.pdf (m=5).

For number of duplication \(c_{ua} = c_{ub} = 4\) and \(n_{ua} = n_{ub} = 5\) we have \(|a| = 0.64 + 2.56 + 5.76 + 10.24 + 16\) and \(|b| = 0.64 + 2.56 + 5.76 + 10.24 + 16\), while for \(|ab|\) where \(m=1\) we obtain \(|ab| = 0.64\), and by using Equation (10) the similarity of two URLs above is 0.036.

Experiments involving many URL addresses have not done to demonstrate in real terms and appraise new formula of similarity. In addition, each query does not only generate the URL addresses from web pages but also is to presents a piece of information that is recognizable as a snippet. The similarity does not happen for two or more web page URLs but also the content of the snippets. Therefore, if the similarity based on the concept of two-dimensional, then the other information also accompany the query can be an additional dimension and may be measured based on extra dimensions. This assumption may expressed as follows.

**Conjecture.** Similarity between two two-dimensional objects be information of objects if and only if the resultant of similarities be the similarity of the multidimensional objects.

5. Conclusions

We have produced a formula of similarity between two objects based on the components of the object. The similarity based on the concept of two-dimensional and the symmetry forms. By applying to the URL addresses, this work will be generating the validation that are appropriate for this similarity. Future work will be assessing what to be conjecture.
References

[1] Mahyuddin K M N 2011 Kolmogorov complexity: clustering objects and similarity *Bulletin of Mathematics* 3 (1) 1-16
[2] Farhadinia B 2014 Distance and similarity measures for higher order hesitant fuzzy sets *Knowledge-Based Systems* 55 43-48
[3] Mahyuddin K M N, Opim S S, Emerson P S and Shahrul A N 2016 An Extracted Social Network Mining *IEEE Proceeding of SAI Computing Conference* 1168-1172
[4] Michel-Marie Deza and Elena Deza 2006 *Dictionary of Distance* Elsevier
[5] W Bas d H, Frans W and Remco C V 2013 A geometrical distance measure for determining the similarity of musical harmony *International Journal of Multimedia Information Retrieval* 2 189-202
[6] Victor B, Anestis G H and Baisil K P 2007 Distance and similarity measures for fuzzy operators *Information Sciences* 177 2336-2348
[7] Grzegorz Kondrak 2005 *N*-Gram Similarity and Distance *String Processing and Information Retrieval* 3772 115-126
[8] Mahyuddin K M N and Syahrul A M N 2010 Superficial method for extracting social network for academic using web snippets *Rough Set and Knowledge Technology* LNAI 6401 483-490
[9] Mahyuddin K M N 2016 Social network mining: A definition of relation between the resources and SNA *M-CAIT* Universiti Kebangsaan Malaysia: Malaca
[10] Mahyuddin K M N and Shahrul A N 2011 Extraction of academic sosial network from online database *IEEE Proceeding of 2011 International Conference on Semantic Technology and Information Retrieval* 64
[11] Mahyuddin K M N, Melvani H and Runtung S 2016 Using Social Networks to Assess Forensic of Negative Issues *IEEE International Conference on Cyber and IT Service Management*
[12] Mahyuddin K M N, Marischa E, Rahmad S and Syahrul A M N 2015 Behavior of the resources in the growth of social network *IEEE Proceeding of the 5th International Conference on Electrical Engineering and Informatics*
[13] Aaron, Opim S S and Romi F R 2015 Distributed autonomous Neuro-Gen Learning Engine for content-based document file type identification *IEEE International Conference on Cyber and IT Service Management*
[14] Mahyuddin K M N, Rahmad S and Marischa E *to appear* Studies on Behavior of Information to Extract the Meaning behind the Behavior
[15] Mahyuddin K M N 2014 New method for extracting keyword for the social actor, *Intelligent Information and Database Systems* LNAI 8397 83-92
[16] Mahyuddin K M N, Melvani H, Rahmad S *to appear* Mining of the social network extraction
[17] Mahyuddin K M N 2013 Superficial method for extracting academic sosial network from the Web (Ph.D Dissertation: Universiti Kebangsaan Malaysia)
[18] Lee S H, Kim S J and Hong S H 2005 On URL normalization *Computational Science and Its Application - ICCSA* LNCS 3481 1076-1085