Journal of Theoretical and Applied Information Technology
31st October 2017. Vol.95. No 20

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

COMPUTING THE DISTANCE AMONG BUSINESS DOMAINS USING SEMANTIC SIMILARITY MEASURES

[1]ELHAM ABD AL LATIF AL BADAWI , [2]AHMAD KAYED,[3]SIDDIQ AHMED BABIKIR

[1]Sudan University of Science and Technology, Computer Science Department, Sudan
[2]Middle East University, Computer Science Department, Jordan
[3]University Technology Malaysia(UTM), Information System Department, Malaysia

[1]elhambadawi@gmail.com,[2]drkayed@ymail.com,[3]Siddiqb@gmail.com

ABSTRACT

Cloud service providers maintain multiple-tenant databases in the cloud computing environment. They reduce the cost by sharing computing and storage resources. Mixing tenants' databases violates security protocols. Databases should be segregated to provide security, confidentiality, reliability, and availability. The competitors of different business domains will sit aside, according to the logical distance of their business domain. The farthest distance will be given to closest business domains. To compute the logical distance among business domains a "good" measure should be defined. There are several types of similarity measures that could be used to measure the similarity among concepts. This paper evaluates several measures to determine the best one that is suitable to be used in our segregation technique. We evaluated the applicability of using five different semantic similarity measures. These measures are path-length measure, Wu & Palmer's measure, Leacock & Chodorow measure, Resnik measure and Lin measure. Taxonomy for business domains has been built to use these measures. This taxonomy mimics the WordNet taxonomy. Several experiments have been conducted to define the best measure among these measures. This paper finds that the shortest path is the best measure to compute the logical distance among businesses.

Keywords: *Cloud Computing, Semantic Similarity Measuare, Path Length Measure, Information Content Measure, Taxonomy.*

1. INTRODUCTION AND MOTIVATION

The cloud computing has become a major IT trend. National Institute of Standards and Technology (NIST) defines cloud computing as follows "Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction" [1].

Storing data is one of the most critical issues in the cloud computing. Cloud providers in order to store the customer's data, they have created a large-scale data centers distributed across the world consisting of thousands of computer nodes (servers) and a large shared storage system.

According to S. Subashini et al [2] there is an obvious challenge in keeping user's data protected within a shared environment. Risks and security threats of cloud storage such as the challenges in handling privileged user access, ensuring privacy and ensuring data segregation should be considered. In the shared environment the user data may be exposed to disclosure, if the information of competing businesses within the same domain stored in the same storage area, then the risk will increase and it has a major effect on competing companies. Businesses within similar domain are on the higher level of risk than the businesses with no similar domains, because data disclosure and data leakage will not have any major effect on businesses that are not in similar domains.

1.1 Data Segregation

Multi tenancy is an important characteristic of cloud computing. In multi-tenant, multiples users and organizations reside at the same location. So it becomes possible for the malicious users to gain access to the other users' data. The segregation means classification of business domain according to their similarities. For

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

example, if a hospital's database is stored with the database of a university within the same storage area, there is a small risk of data disclosure or leakage. But if a bank database is stored with competing bank's database, then there will be a huge risk of data disclosure. So keeping data separate and maintain isolation among the users is an important issue. The isolation means less sharing & less scalability. These issues can be solved by creating a robust segregation Technique [3]. The main goal of the segregation is to enhance the cloud data security and prevent similar business domain to be allocated in the same storage area.

In this paper, we want to choose the best similarity measure. We deployed five different semantic similarity measures to measure the similarity between fifty pairs of concepts. These concepts represent different business domain taken from the taxonomy described below in section 2. Then the same fifty concepts are evaluated by human expert, results have been compared in order to determine the best measure to be used in our segregation technique to enhance the data security for the cloud computing.

1.2 Semantic Similarity

The term semantic similarity indicates the computation of the similarity amount among the concepts. The semantic similarity measure is a method used to compute the similarities between concepts/terms included in the knowledge sources, in order to perform estimation and to identify the relatedness percentage between these concepts. The similarities help in identifying concepts that are having the common characteristics [4]. Semantic similarity between the concepts is a fundamental issue and plays an important role in many applications of artificial intelligence, knowledge sharing and web mining [2]. It is also a fundamental concept used in many Natural Language Processing (NLP) tasks and in the Information Retrieval (IR) to find documents and extract information that is most relevant and most similar to the user query [4].

Lin (1998) [5] states, three intuitions which should be considered when defining a similarity measure between two concepts these are:

- The more commonality they share, the more similar they are.
- The more differences they have, the less similar they are.
- Maximum similarity is reached when the two items are identical.

Many of the existing measures are focusing on the first and the third principal, ignoring the second one. But the differences between the concepts will appear in a way, when the commonality is measured. The similarity between concepts or entities can be identified if they share common attributes or if they are linked to other semantically related entities in the ontology [6].

The structure of this paper is as follows: Section2 will present the business domain taxonomy, section 3 shows the related work in the semantic similarity measure field, section 4 discusses our experiments which describe the implementation of the similarity measures equations to calculate the similarity. Section 5 shows the result of our experiment and section 6 concludes the paper.

2.THE TAXONOMY

Taxonomy for business domains has been built to use semantic similarity measures. This taxonomy mimics the WordNet taxonomy.WordNet is an electronic lexical database which is considered to be the most important resource available to the researchers in computational linguistics [4].

The taxonomy was built using Thomson Reuters Business Classification (TRBC) which is an industry classification of global companies [7], used primarily in the financial investment and advisory space for the description of business domain. It is owned and operated by Thomson Reuters. TRBC covers over 70,000 public companies from 130 countries and provides over 10 years of classification history.

The five-level hierarchical structure starts with the topmost container Business which contains economic sectors, each of which has a number of business sectors which are classified into industry groups containing different companies.

TRBC consists of 10 economic sectors, 25 business sectors, 52 industry groups, 124 industries and 837 Activities.

The Thomson Reuters Business Classification describes a large number of business domain types. For our experiment we have built the taxonomy describing only three basic business sectors. These are: Basic Materials, Consumer Goods & Services and Financial. We choose these

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

branches randomly and it's sufficient for our experiment.

For example, as being described in figure 1 below: Basic material economic sector is being classified into three basic business sectors which are Applied Resources, Chemicals and Mineral Resources. Applied Resources is being further classified into Paper & Forest Products and Containers & Packaging.

Paper & Forest Products are further classified into two industries Non-Paper Containers & Packaging and Paper Packaging. Containers & Packaging Industry are then further classified into two industries Forest & Wood Products and Paper Products.

The taxonomy is a type of ontology which is used to represent data. The following Figure 1 describes business domain type.

Journal of Theoretical and Applied Information Technology
31st October 2017. Vol.95. No 20

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

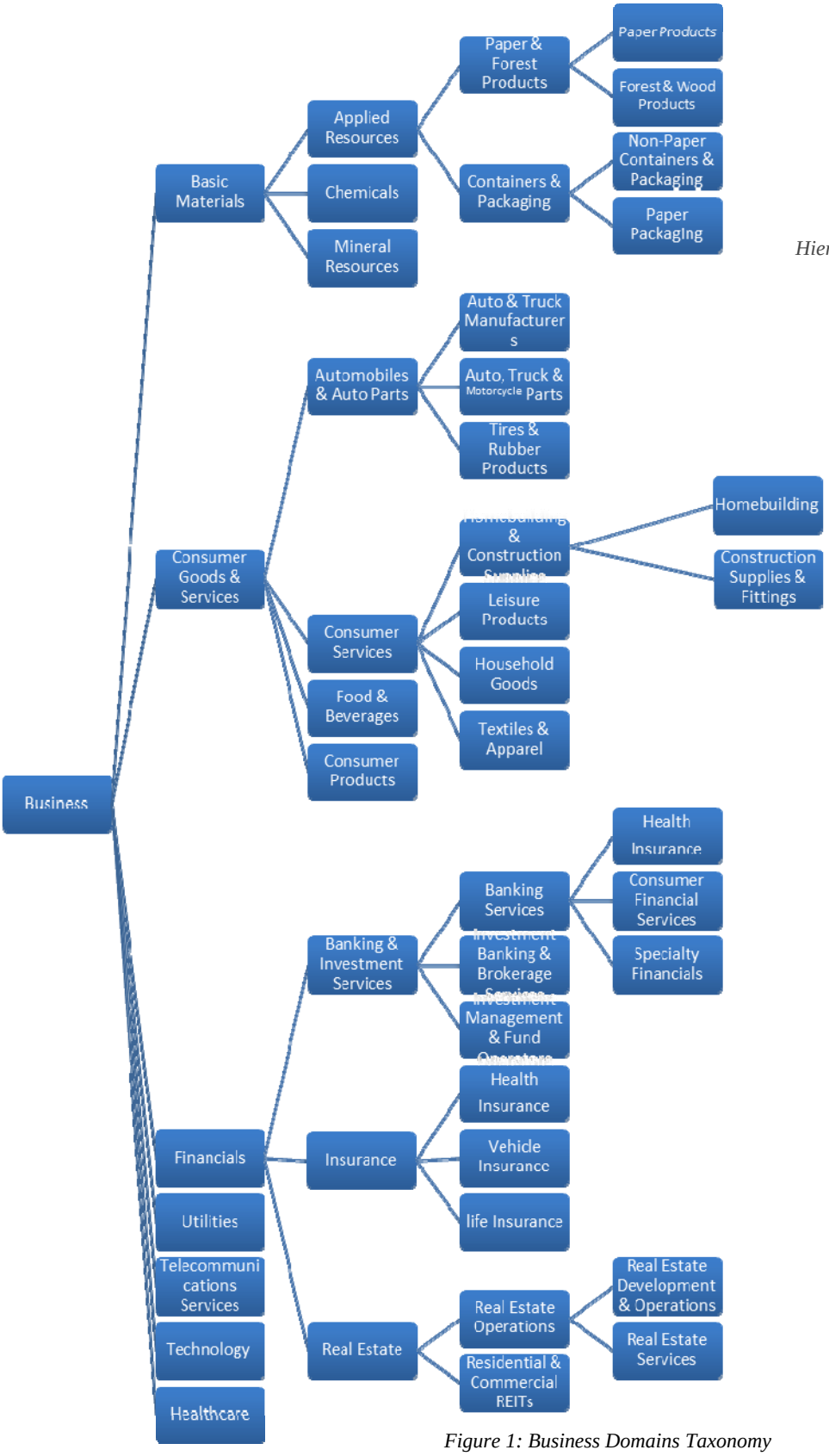

Figure 1: Business Domains Taxonomy

Hierarchy Of Business Type According To Thomson Reuters Business Classification (TRBC)

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3. RELATED WORK

Several methods and research work for determining semantic measures have been proposed in the last few decades, similarity between nodes is also called as relatedness. To measure the relatedness between two concepts C1 and C2, researchers have used several types of measures.

Types of Semantic Similarity Measures

According to Joe Raad et al [8] several measures of semantic similarities have been proposed. These semantic similarity measures can be generally partitioned into five categories:

- First category is based on how close the two concepts in the taxonomy are? By measuring the length of the path, which has linked the two concepts and it's called Path Length based measures.
- Second category is based on how much information these two concepts share and it is called Information Content based measures.
- Third category is based on the properties of these two concepts and it is called Feature based measures.
- Fourth category is based on the assumption that the semantically close terms tend to appear in similar context which called Distributional based measures.
- Fifth category is based on combinations of the previous options and it is called Hybrid measures.

Rada [9] defined the conceptual distance between two words in the "**IS-A**" hierarchy relationships as the length of the shortest path connecting the two words. Then Sussna [10] extended the Edge Counting Measure with weighted edges, which means that the edge between adjacent nodes A and B has a weight that must be considered in calculating the similarity.

There is also new approach called Hybrid method which combines the shortest path approach and information content to improve the similarity measurement performance. The second approach that employs Information Contents is firstly applied by Resnik [11]. He stated that the more information content two words share, the more similar they are. There are several other measures proposed following Resnik's method. Lin [5] extends Resnik's measure by considering the information content of individual concepts together with the Information content of Nearest Common Ancestor (NCA).

Developing a semantic similarity measure is a complex task, particularly the one which totally resembles human assessment is very difficult to be designed.

First the paper shows the used definitions to clarify their meaning.

1) C1 and C2: concept1 and concept2.
2) Len (ci, cj): the length of the shortest path connecting the nodes ci and cj, represented by the number of edges in the path.
3) depth(ci): the length of the path from the global root entity to concept ci, depth(root)=1.
4) deep_max or max: the maximum depth of the taxonomy. In our taxonomy it is 5.
5) NCA (C_1, C_2): Information content of the Nearest Common Ancestor of c1 and c2.
6) LCS (C_1, C_2) determines that the Least Common Super means the information content of the nearest shared parent.
7) LSO(ci,cj): the Lowest Common Subsumer of ci and cj.
   NCA(C_1, C_2) =LSO(C_1,C_2) = LCS (C_1, C_2).
8) Sim=similarity

The main semantic measures could be classified into two categories; Structure-based measures and Information content measures. These types of measures will be explained in details in the following section:

3.1 Structure-based measures (Path length based measures)

In this method, the quantification of similarity measurement among concepts is determined according to the path distance, which separates the concepts on the taxonomy or ontology structure and it includes the following types:-

3.1.1 Shortest Path

It represents the measures that are used in the hierarchy structure to compute the minimum sum of edge weight in hierarchical structure linking the terms C1 and C2. Rada et al. estimate the distance between two concepts (u, v) as the shortest-path linking them **(sp (u, v))** and they used biomedicine domain to evaluate their work in the information retrieval tasks using shortest path measure [9].

The similarities between two concepts **C1** and **C2** can be formulated as follows [10]:

Sim (c1, c2)= 2 * Max – len (c1,c2) (1)

Where Max is the maximum depth of the taxonomy. Len (c1, c2) is the shortest path length between C1 and C2.

This equation explained in details in section 4.

3.1.2 Weighted Links

This measure is an extension of the above measure. The distance between two concepts is obtained by summing up the weights of links [12].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The weight of a link may be affected by the following:

(a) The density of the taxonomy at that point.

(b) The depth of the hierarchy.

(c) The strength of connotation between parent and child nodes. Then, computing the distance between these two concepts is translated into summing up the weights of the traversed links instead of counting them [13].

3.1.3 Hirst and St-Onge Measure (HSO)

Hirst and St-Onge (1998) determines the similarity among concepts based on the path distance between the two concepts, considering the number of changes in direction of the path connecting these concepts and the allowableness of the path [14]. Hirst and St-Onge (1998) classifies the semantic relations in the WordNet lexical ontology into three main relations as follows: Extra Strong Relations, Strong Relations and Medium Strong Relations.

3.1.4 Wu and Palmer

This measure calculate the similarity of two concepts **C1** and **C2** considering the position of the most specific common concept C which is the nearest shared parent or ancestor of both concepts [15].

$$\mathbf{Sim(c1,c2)_{wup}} = \frac{2*N}{N1+N2+2*N} \qquad (2)$$

Where **N1** and **N2** are the distance from the Specific Common Concept (nearest shared parents) to concept C1 and C2 respectively. N is the distance which separates the closest common ancestor of C1 and C2 from the root node. The equation [11] and its application explained later using numeric values from our taxonomy in section 4.

3.1.5 Slimani

Improve Wu and Palmer measure by proposing a penalization factor of two concepts **C1** and **C2** placed in the neighborhood to be multiplied by Wu and Palmer measure. This function aims to penalize or to reduce the value of similarity measure where two concepts are not in the same hierarchy [16].

3.1.6 Li et al.

This similarity measure combines the **shortest path length (SP)** between two concepts **C1** and **C2**, and the **depth (N)** of the most specific common concept **C** in the taxonomy, in a non-linear function [17].

$$\mathbf{Sim(c1,c2)} = e^{-\alpha*SP} * \frac{e^{\beta*N} - e^{-\beta*N}}{e^{\beta*N} + e^{-\beta*N}} \qquad (3)$$

Where $\alpha \geq 0$ and $\beta \geq 0$ are parameters scaling the contribution of shortest path length and depth respectively [10]. The optimal parameters are $\alpha = 0.2$ and $\beta = 06$ it is therefore obvious that this measure scores between 1 (for similar concepts) and 0 (for not similar concepts).

3.1.7 Leacock and Chodorow similarity measure:

According to Leacock and Chodorow (1998), firstly the similarity between two concepts is determined by discovering the shortest path length, which connects these two concepts in the taxonomy. The similarity is calculated as the negative algorithm of this value (Batet et al., 2011). The equation of Leacock and Chodorow (1998) can be written as follows [18]:

$$\mathbf{Sim_{lc}(C1,C2)} = \mathbf{- \log (((len\ (C1,C2))/(2*deep\_max))} \qquad (4)$$

This equation explained in details later in section 4.

3.2 Information Content (IC) measures

Following is the standard argumentation of information theory [Ross, 1976], the information content of a concept **c** can be quantified as the negative log likelihood **[6]**,

$$\mathbf{IC(c) = -\log p(c)} \qquad (5)$$

3.2.1 Resink Measure

Resnik (1999) proposed a similarity measure using Information content, this measure signifies that the more information two terms share in common, the more similar they are, and the information shared by two terms is indicated by the information content of the term that subsumes them both in the taxonomy[13].

In this measure, the similarity of two concepts **(c1, c2)** is defined as the **Information Content** (IC) of their LCS(Least Common Super) which determines the information content of the nearest shared parents , as shown in the following equation[18]:

Equation:

$$\mathbf{Sim_{Resnik}(C_1, C_2) = -\log p(LCS(C_1, C_2)) = IC(LCS(C_1, C_2))} \qquad (6)$$

Where the

$$IC(C) = ((\log (depth(C))/ (\log (\mathbf{deep\_max})) \qquad (7)$$

This equation explained in details later in section 4.

From formula (6) noted that:

(1) The values only rely on the lowest subsumer of the pair of concepts from the taxonomy.

(2) LCS(Banking & Investment Services, Insurance) = LCS (Insurance, Real Estate)= Financials, therefore SimResnik (Banking & Investment

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Services, Insurance)= simResnik (Insurance, Real Estate)=IC(Financials).

3.2.2 Lin [18]

The authors proposed a measure of similarity which takes into account the information shared by two concepts like Resnik, but presents a better ranking of similarity than the Resnik measure.

The authors defined the similarity between two concepts **(c1, c2)** as: dividing twice the IC of the concepts LCS by the sum of the individual IC of each concept

Equation:

Sim lin (C1, C2) =

2*IC (LCS (C1, C2))/ ((IC (C1) +IC (C2)) (8)

This equation explained in details later in section 4.

The aim of this paper is to compute the logical distance between two concepts using the five different semantic similarity measures. The best measure must be chosen to measure the similarity for our segregation technique to enhance the data security for the cloud computing

The contribution of this work is the new deployment of semantic measure to compute how much businesses are closed to each other.

4. EXPERIMENT DESIGN AND RESULT

The experiment was done by applying some of semantic similarity measures equations to compute the similarity between the business domains represented by our taxonomy that shown in figure 1 section 2. We have evaluated the result of these measures against human expert evaluation, and then we compute the error rate of all the measures to test the effectiveness and to find the best semantic similarity measure which gives a less error rate, to be used in our segregation technique.

Determining similarity value is necessary as we are going to utilize it in our segregation method that will be discussed in our next research paper. The next research paper will be about how to distribute the stored data between cloud servers according to their similarity value such that the more they are unrelated business domain the chance to place them in the same server will increase and vice versa. Since the cloud providers usually store the information of multiple users in the same server in order to reduce the storage cost, this will lead to a higher level of security risk. Especially if they place the same business domain type's data close to their competing companies, the information will be exposed to disclosure. Therefore, since the need for a scientific approach arises, the research chooses data segregation in order to prevent the placement of the same business domain type at the same server and provide more security.

All the semantic similarity measures take as inputs, a pair of concepts representing business domain , and return a value indicating the semantic relatedness between them.

The results have been shown in six tables. Table 1 presents results from path length similarity measures and the similarity values between 50 different pairs of business domain. Table 2 presents results obtained by Wu & Palmer's measure. Table 3 presents the results we obtained from Leacock & Chodorow. Table 4 presents results from Resnik similarity. Table 5 presents results from Lin similarity measure. Table 6 presents results from all previous similarity methods and their correlation with human expert judgment.

We conclude our evaluation using the following word similarity metrics: path length, Wu & Palmer's Measure, Leacock & Chodorow, Resnik, and Lin.

These five measures were selected based on their observed performance in other language processing applications, and also for their relatively high computational efficiency.

The following is the application of the semantic similarity measures and their results

4.1 Shortest Path Measure

It represents the measure that uses the path distance to compute the similarity among concepts.

Similarity (c1, c2) =

(2 * Max-Depth) – length (c1, c2) (1)

Where, **Max** is the Maximum depth of the taxonomy .

Length (c1, c2) is the shortest path length c1 and c2.

For example, to calculate the similarity between **Paper Products and Forest & Wood Products** the **shortest path length = 2** (from the taxonomy in section 2),

Max-Depth for our taxonomy **= 5.**

So,

Sim (**Paper Products, Forest & Wood Products**) =

(2*5) -2=8 = 80%,

some results of the calculation of path length equation are shown in the following Table 1.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1: Sample of the similarity values between the concepts using the Path length Measure

ID	Concept1	Concept2	Length (c1,c2)	Similarity
1	Paper Products	Paper Products	0	100%
2	Paper Products	Forest & Wood Products	2	80%
3	Paper Products	Health Insurance	8	20%
4	Health Insurance	Auto & Truck Manufacturers	7	30%
5	Real Estate Development & Operations	Health Insurance	6	40%
. . . 50	Real Estate Development & Operation	Life Insurance	5	50%

4.2 Wu & Palmer's Measure

This similarity measure considers the position of C1 and C2 to the position of the most specific common concept C. Several parents can be shared by C1 and C2 by multiple paths. The most specific common concept is the closest common ancestor.

$$\mathbf{Sim(c1,c2)_{wup}} = \frac{2*N}{N1+N2+2*N} \qquad (2)$$

Where **N1** and **N2** are the distance from the specific common concept to concept **C1** and **C2** respectively. **N** is the distance which separates the closest common ancestor of **C1 and C2** from the root node [10].

For example: to calculate the similarity between (**Paper Products** and **Forest & Wood Products**). Firstly, we determine that the specific common subsume of **Paper Products**, **Forest & Wood Products** is as shown in our taxonomy is **Paper & Forest Products.**

Since **N1** and **N2** are the distance from the specific common concept to concept **C1** and **C2**, **N1=1**, **N2=1** and **N=3** since N is the distance which separates the closest common ancestor of C1 and C2 from the root node (**Business**).

Similarity (Paper Products, Forest & Wood Products) = $\frac{2*3}{1+1+(2*3)}$**=0.75 =75%,**

Some results of the calculation of Wu & Palmer's Measure equation are shown in the following table2.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 2: Sample Of The Similarity Values Between The Concepts Using The Wu & Palmer's Measure

ID	Concept1	Concept2	Similarity
1	Paper Products	Paper Products	100%
2	Paper Products	Forest & Wood Products	75%
3	Paper Products	Health Insurance	0%
4	Health Insurance	Auto & Truck Manufacturers	0%
5	Real Estate Development & Operations	Health Insurance	25%
. . . 50	Real Estate Development & Operation	Life Insurance	29%

4.3 Leakcock& Chodorow's Measure

The similarity is calculated as the negative algorithm of the shortest path length.

$$Sim_{lc}(C1,C2) = -\log(((len(C1,C2))/(2*deep\_max)) \quad (4)$$

Where **len** is the length of the shortest path connecting the two **concepts (C1, C2)** and **deep_max** is the maximum depth of the taxonomy [18].

For example: to calculate the similarity between Paper Products and Forest & Wood Products.

Firstly, we determine the length between **Paper Products**, **Forest & Wood Products** as shown in our taxonomy,

The shortest path length is 2

The deep_max of our taxonomy is 5

The similarity is calculated as the negative algorithm of this value

Similarity (Paper Products, Forest & Wood Products) = -log(2/(2*5))= 0.698970004=70%, some results of the calculation of Leakcock & Chodorow's Measure requation are shown in the following table 3.

Table 3: Sample Of The Similarity Values Between The Concepts Using The Leakcock & Chodorow's Measure

ID	Concept1	Concept2	Length (c1,c2)	Similarity
1	Paper Products	Paper Products	0	100%
2	Paper Products	Forest & Wood Products	2	70%
3	Paper Products	Health Insurance	8	10%
4	Health Insurance	Auto & Truck Manufacturers	7	15%
5	Real Estate Development & Operations	Health Insurance	6	22%
. . . 50	Real Estate Development & Operation	Life Insurance	5	30%

4.4 Resnik's Measure

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

This measure uses the Information content of the shared parents. The principle of this measure is as follows: two concepts are more similar if they present more amount of shared information. The information shared by two concepts C1 and C2 is indicated by the information content of the concepts that subsume them in the taxonomy [18].
Equation:

$$Sim_{Resnik}(C_1, C_2) = -\log p(LCS(C_1, C_2)) = IC(LCS(C_1, C_2)) \quad (6)$$

Where, **LCS (C_1, C_2)** refer to **Least Common Super**, and **P (c)** is the probability of finding an instance of concept c in a large corpus. For calculating the similarity we need to apply the following equations that have been explained earlier in section 3.

$$IC(C) = IC(LCS(C_1, C_2)) = ((\log(depth(C))/(\log(deep\_max)) \quad (7)$$

Example: To calculate the similarity between **Paper Products** and **Forest & Wood Products** we need the following ,which has been taken from our taxonomy:
1) Deep_Max=5.
2) IC (LCS (C_1, C_2)) = IC(Paper & Forest Products).
3) Depth (C) = depth (Paper & Forest Products) = 4.

$Sim_{Resnik}(C_1, C_2)$ = IC(Paper & Forest Products) = ((log (depth(Paper & Forest Products))/ (log (deep_max))
Sim(Paper Products, Forest & Wood Products) = ((log (4)/ (log (5)) = 0.861353 =86%, some results of the calculation are shown in the following table 4.

Table 4: Sample Of The Similarity Between Concepts Using Resnik's Measure.

ID	Concept1	Concept2	IC(LCS (C1,C2))	Sim
1	Paper Products	Paper Products	1.00	100%
2	Paper Products	Forest & Wood Products	0.86	86%
3	Paper Products	Health Insurance	0.00	0%
4	Health Insurance	Auto & Truck Manufacturers	0.00	0%
. . . 50	Real Estate Development & Operation	Life Insurance	0.43	43%

4.5 Lin's Measure
This measure of similarity takes into account the information shared by two concepts same as Resnik measure, but the similarity between the two concepts (c1, c2) is calculated by: dividing twice the IC of the concepts LCS by the sum of the individual IC of each concept
Equation:

$$Sim\ lin(C1,C2) = 2*IC(LCS(C1,C2))/((IC(C1)+IC(C2)) \quad (8)$$

Example: To compute the similarity between (Paper Products) and (Forest & Wood Products), we need the following from our taxonomy:
1) IC (C1) = IC (Paper Products)
2) IC (LCS (C1, C2) = IC (Paper & Forest Products).
3) IC (C2) = IC (Forest & Wood Products).
4) Depth (Paper & Forest Products) = 4.
5) deep_max=5.

$$IC(C) = ((\log(depth(C))/(\log(deep\_max)) \quad (7)$$

IC (Paper Products) = ((log (depth (Paper Products))/ (log (deep_max))= ((log (5))/ (log (5)) = 1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

IC (Forest & Wood Products) =
((log (5))/ (log (5)) = 1
IC (LCS (C1, C2)) =
IC (Paper & Forest Products) =

((log (depth (Paper & Forest Products))/ (log (deep_max)) =
((log (4)/ (log (5)) = 0.861353.

Similarity (Paper Products, Forest &Wood Products) =
2*IC(LCS(C1, C2))/ ((IC (C1) +IC (C2)) = 2*0.861353/(1+1) = 0.86 =86% ,
Sample of the results has been shown in Table 5.

Table 5: Sample Of The Similarity Between The Concepts Using The Lin's Measure

.

ID	Concept1	Concept2	IC(c1)	IC(C2)	LCS (C1,C2)	Similarity
1	Paper Products	Paper Products	1.00	1.00	1.00	100%
2	Paper Products	Forest & Wood Products	1.00	1.00	0.86	86%
3	Paper Products	Health Insurance	1.00	1.00	0.00	0%
4	Real Estate Development & Operations	Health Insurance	1.00	1.00	0.43	43%
. . . 50	Real Estate Development & Operation	Life Insurance	1.00	0.86	0.43	46%

5. RESULTS

In our experiment a sample of fifty pairs of business domains has been taken from our business taxonomy. Five measures have been computed among these pairs. The results of all these measures, which presented in the previous tables (table1..table5) have been aggregated in the table 6. The last column of the table 6 shows the average evaluation of similarity between each pairs by group of human experts. in business domain.
Table 6 shows the absolute error for the each one of five measures in column 3,5,7,9,and 11 respectively. The error computes by the absolute difference for each measure with the human evolution (column 12).
For example, the second column represents the similarity between ((**Paper Products**), (**Paper Products**)) using the path measure. The fourth column represents the similarity value for same pair using the Wu & Palmer measure, the sixth, eighth and tenth represent the result of (Leakcock & Chodorow's Measure, Resnik measure, and Lin measure) respectively.

The results of the similarity value of all five measures for the first pair of concepts are 100% as we compared the concept with the same concept ((**Paper Products**), (**Paper Products**)).
The second row indicates the similarity between the pair of concepts ((**Paper Products**), (**Forest & Wood Products**)) and the result of similarity for (Path, W & P, L& Ch, Resnik, and Lin) are 80%, 75%, 70%, 86%, and 86%) respectively . While the expert human evaluation for the same pair((**Paper Products**), (**Forest & Wood Products**)) is **80%**.
The third column (Error) in the second row represents the error for the Path measure when compared with human evaluation = |80%-80%| =0%.
The error for W&P measure is 5% which = |80%-75%| (The absolute value for the difference between human expert evaluation and W&P measure evaluation), by the same way all other errors have been calculated.
The last row of the table shows the average of errors for the fifty pairs calculated using (Average) function for each Error column of each similarity measure.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 6: Shows The Similarity Value For Fifty Pairs Of Business Domains And Their Error Rate.

No.	Path	Error	W & P	Error	L& Ch	Error	Resnik	Error	Lin	Error	Human expert
1	100%	0%	100%	0	100%	0%	100%	0%	100%	0%	100%
2	80%	0%	75%	5%	70%	10%	86%	6%	86%	6%	80%
3	20%	5%	0%	15%	10%	5%	0%	15%	0%	15%	15%
4	30%	10%	0%	20%	15%	5%	0%	20%	0%	20%	20%
5	40%	0%	25%	15%	22%	18%	43%	3%	43%	3%	40%
6	70%	10%	40%	40%	52%	28%	43%	37%	56%	24%	80%
7	80%	0%	75%	5%	70%	10%	86%	6%	86%	6%	80%
8	30%	10%	0%	20%	15%	5%	0%	20%	0%	20%	20%
9	90%	5%	86%	1%	100%	15%	43%	42%	50%	35%	85%
10	90%	5%	86%	9%	100%	5%	86%	9%	93%	2%	95%
11	40%	10%	0%	50%	22%	28%	0%	50%	0%	50%	50%
12	40%	10%	0%	30%	22%	8%	0%	30%	0%	30%	30%
13	40%	10%	0%	30%	22%	8%	0%	30%	0%	30%	30%
14	70%	0%	33%	37%	52%	18%	68%	2%	73%	3%	70%
15	60%	0%	33%	27%	40%	20%	43%	17%	50%	10%	60%
16	50%	0%	0%	50%	30%	20%	0%	50%	0%	50%	50%
17	50%	0%	0%	50%	30%	20%	0%	50%	0%	50%	50%
18	40%	10%	0%	30%	22%	8%	0%	30%	0%	30%	30%
19	60%	10%	50%	0%	40%	10%	68%	18%	68%	18%	50%
20	20%	10%	0%	30%	10%	20%	0%	30%	0%	30%	30%
21	80%	10%	33%	57%	70%	20%	43%	47%	51%	39%	90%
22	40%	10%	0%	30%	22%	8%	0%	30%	0%	30%	30%
23	30%	10%	0%	20%	15%	5%	0%	20%	0%	20%	20%
24	20%	10%	0%	10%	10%	0%	0%	10%	0%	10%	10%
25	60%	0%	0%	60%	40%	20%	0%	60%	0%	60%	60%
26	90%	0%	80%	10%	100%	10%	68%	22%	88%	2%	90%
27	70%	5%	40%	35%	52%	23%	43%	32%	56%	19%	75%
28	50%	10%	0%	40%	30%	10%	0%	40%	0%	40%	40%
29	80%	0%	67%	13%	70%	10%	68%	12%	81%	1%	80%
30	60%	10%	0%	50%	40%	10%	0%	50%	0%	50%	50%
31	90%	5%	80%	15%	100%	5%	68%	27%	88%	7%	95%
32	20%	10%	0%	10%	10%	0%	0%	10%	0%	10%	10%

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

No.	Path	Error	W & P	Error	L& Ch	Error	Resnik	Error	Lin	Error	human expert
33	20%	10%	0%	10%	10%	0%	0%	10%	0%	10%	10%
34	80%	0%	57%	23%	70%	10%	68%	12%	73%	7%	80%
35	30%	10%	0%	20%	15%	5%	0%	20%	0%	20%	20%
36	30%	10%	0%	20%	15%	5%	0%	20%	0%	20%	20%
37	30%	10%	0%	20%	15%	5%	0%	20%	0%	20%	20%
38	50%	10%	0%	40%	30%	10%	0%	40%	0%	40%	40%
39	60%	10%	33%	37%	40%	30%	43%	27%	46%	24%	70%
40	90%	5%	86%	9%	100%	5%	86%	9%	93%	2%	95%
41	70%	10%	40%	40%	52%	28%	43%	37%	56%	24%	80%
42	40%	10%	0%	30%	22%	8%	0%	30%	0%	30%	30%
43	60%	0%	0%	60%	40%	20%	0%	60%	0%	60%	60%
44	90%	0%	80%	10%	100%	10%	68%	22%	88%	2%	90%
45	70%	10%	40%	20%	52%	8%	43%	17%	56%	4%	60%
46	70%	10%	40%	20%	52%	8%	43%	17%	56%	4%	60%
47	20%	10%	0%	10%	10%	0%	0%	10%	0%	10%	10%
48	30%	10%	0%	40%	15%	25%	0%	40%	0%	40%	40%
49	40%	10%	0%	30%	22%	8%	0%	30%	0%	30%	30%
50	50%	20%	29%	1%	30%	0%	43%	13%	46%	16%	30%
Error Rate	7%		25%		11%		25%		22%		

6. DISCUSSION AND LIMITATION OF THE STUDY

The shortest path has been achieved 7% error rate, which is the lowest rate among all other measures. For that reason it has been chosen as the best measure to be utilized in our proposed segregation technique.

The other reason for choosing the shortest path measure that many researchers advice to use it, according to (Rada et al) a natural way to evaluate semantic similarity in a taxonomy is to evaluate the distance between the nodes corresponding to the items being compared — the shorter the path from one node to another, the more similar they are. Given multiple paths, one takes the length of the shorter one [19]. Caviedes et al stated that the length of the shortest path (PL) between two terms in a given ontology has been proved to be a good indicator of the semantic distance between them [20].

There were some limitations in calculating these results. First, human intervention that used to evaluate the similarity and to compute the error. We try to minimize this by consulting three experts and then take the average of their evaluation.

Second, the data used for building the taxonomy were in very specific domain, since it is very hard and it is a time consuming to do these experiments for many domains. We try to minimize this by using a bigger sample (fifty pairs) that represents a large span of domains. The data have been collected from the internet to reduce the time of building our taxonomy. Future research may consider the development of the taxonomy that specifies more business domains as ontology concepts.

Conclusion

This paper aimed to find a good method to compute the logical distance among businesses. The aim is to allocate the businesses which are closed to each other (competitors) away from each other. The novelty of this paper included in

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

the idea of deploying the existing SM to reach to that goal. Five semantic similarity measures from different families have been chosen. These measures are path length measure, Wu & Palmer's measure, Leacock & Chodorow measure, Resnik measure and Lin measure.

Several experiments have been designed and run to find which SM will be fit our goals. A business domain taxonomy has been built to use these SMs. The similarity of fifty pairs of business domains that has been taken from the business domain taxonomy has been computed using the different five measures. Human expert in business domain has computed the logical distances among these fifty pairs. Then the SMs results have been compared with the human experts. This paper finds out that the shortest path measure is the most suitable measure which could be used to find the logical distances among business domains. This enabled us to build upon this measure several segregation techniques to allocate similar businesses away from each other. This is a vital issue in allocating several tenants DB away from each in the cloud severs.

The taxonomy was built using an industry classification called Thomson Reuters Business Classification (TRBC) which is used primarily in the financial investment for the description of business domain. We used this taxonomy in order to measure the distance and the similarity between the business domains which is represented as nodes of the taxonomy.

ACKNOWLEDGEMENT

We would like to thank Dr. Ahmed Arab from the department of business administration at Applied Science University, for his effort in evaluating the similarity between the concepts according to his knowledge and expertise in the business domain.

REFERENCES

[1] P.Mell, and Grance,. "The NIST Definition of Cloud computing, National Institute of Standards and Technology", T. 2009.

[2] S. Subashini, V.Kavitha. "A survey on security issues in service delivery models of cloud computing", Journal of Network and Computer Applications 34(2011)1-11

[3]Aggarwal, Navdeep, et al. "Cloud computing: data storage security analysis and its challenges." *International Journal of Computer Applications* 70.24 (2013).

[4] Elavarasi, S. Anitha, J. Akilandeswari, and K. Menaga. "A Survey on Semantic Similarity Measure." International Journal of Research in Advent Technology, Vol.2, No.3, March 2014 E-ISSN: 2321-9637

[5] D. Lin, "An Information-Theoretic Definition of Similarity," in Proceeding of the 15th International Conference on Machine Learning, Madison, 1998, p. 296-304.

[6] Mihalcea, Rada, Courtney Corley, and Carlo Strapparava. "Corpus-based and knowledge-based measures of text semantic similarity." AAAI. Vol6. 2006.

[7]http://thomsonreuters.com/en/products-services/financial/market-indices/business-classification.html.

[8] Raad, J., Bertaux, A., & Cruz, C. (2015, July). A survey on how to cross-reference web information sources. In Science and Information Conference (SAI), 2015 (pp. 609-618). IEEE.

[9]R. Rada, H. Milli, E. Bicknell, M. Blettner, "Development and Application of a metric on Semantic Nets", IEEE Trans. on Systems, Man, and Cybernetics, 19(1): 17-30 (1989)

[10] Thabet Slimani Slimani, Thabet. "Description and evaluation of semantic similarity measures approaches." arXiv preprint arXiv:1310.8059 (2013).

[11] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy", in: Proc. of 14th International Joint Conference on Aritificial Intelligence, 1995,PP 448-453.

[12] Richardson, R, Smeaton, A. & Murphy, J. 1994. Using WordNet as a Knowledge Base for Measuring Semantic Similarity Between Words. Technical Report Working paper CA-1294, School of Computer Applications, Dublin City University, Dublin, Ireland.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[13] Varelas, Giannis, et al. "Semantic similarity methods in wordNet and their application to information retrieval on the web." Proceedings of the 7th annual ACM international workshop on Web information and data management. ACM, 2005.

[14] Hirst, G. and St-Onge, D. 1998. Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms. In Proceedings of Fellbaum, pages 305-332.

[15] Wu, Zhibiao, and Martha Palmer. "Verbs semantics and lexical selection."Proceedings of the 32nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics,), pages 133{138, Las Cruces, New Mexico, 1994.

[16] Slimani, T. Ben Yaghlane, B. and Mellouli, K. 2006. "A New Similarity Measure based on Edge Counting" World Academy of Science, Engineering and Technology, PP 34-38.

[17] Y. Li, Z. Bandar, D. McLean, "An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources", IEEE Transactions on Knowledge and Data Engineering, 15(4):871-882, July/August. (2003)

[18] Meng, Lingling, Runqing Huang, and Junzhong Gu. "A review of semantic similarity measures in wordnet." International Journal of Hybrid Information Technology 6.1 (2013): 1-12.

[19]Nessah, Djamel, Okba Kazar, and Aïcha-Nabila Benharkat. "Towards a hybrid semantic similarity measure to set the conceptual relatedness in a hierarchy." International Journal of Metadata, Semantics and Ontologies 11.3 (2016): 155-164.

[20]CAVIEDES, Jorge E.; CIMINO, James J. Towards the development of a conceptual distance metric for the UMLS. Journal of biomedical informatics, 2004, 37.2: 77-85.