Semantic Numerical Operations on Strings

Taeho Jo
School of Computer and Information Engineering
Inha University
Incheon, South Korea
tjo018@inha.ac.kr

***Abstract*—This research is concerned with the definition, the analysis, and the simulation of the numeric semantic operations on strings. The motivation of this research is the dominance of textual data over numerical data in our reality and the necessity of defining semantic operations for analyzing meanings of words. In this research, we define and simulate the three operations: 'semantic similarity', 'semantic similarity average', and 'semantic similarity variance'. This research is expected to become the basis from which semantic analysis tools or systems of words, texts or corpus, are developed as its benefit. We present the simulations of carrying out operations on strings in the real corpus: NewsPage.com.**

***Keywords*-Semantic Operation; Similarity Semantic Average; Similarity Semantic Variance**

I. Introduction

The semantic operations are defined as the operations based on quantities indicating semantic relations among entities. The words in textual data are given as operands of the operations which were proposed in this research. As the basis of performing the operations, we use the similarity matrix which consists of semantic similarities indicating how much corresponding words are similar as each other. In this research, we define the following three operations: 'semantic similarity', 'semantic similarity average', and 'semantic similarity variance'. Each operation generates a normalized value between zero and one as its output.

Previously, we attempted to replace numerical vectors by string vectors in representing texts. The reason of the replacement is the three problems: huge dimensionality, sparse distribution, and poor transparency; they are described in detail in the literatures [1][2][3][4][5]. The replacement leads to the successful performance in text categorization and clustering. However, in order to use the string vectors more naturally and freely, we need more systematic mathematical analysis and definitions on strings. The previous research concerned with encoding of texts into string vectors will be mentioned in Section 2.

In this research, we define the three semantic operations on strings. The semantic similarity between two words indicating how much two words are similar as each other, is included as the basic operation. The SSA (Semantic Similarity Average) is proposed as the average over similarities of all possible pairs of words[5]. From the SSA, we derive SSV (Semantic Similarity Variance), as the variance over the similarities. In this research, we call the defined operations numerical semantic operations, since numerical values are generated from the operations as their outputs.

We expect the three benefits from this research. For first, the semantic operations are potentially used for developing string vector based approaches to tasks of text mining and information retrieval. For second, this research may provide the basis for developing automatic semantic analysis tool for words and texts. For third, the possibility of developing even digital computers only for text processing is available potentially. In order to take the benefits, we need to define more semantic operations and characterize them mathematically.

This article is composed of the five sections. In Section II, we explore the previous research relevant to this research. In Section III, we describe the proposed semantic operations formally and characterize them mathematically. In Section IV, the operations are simulated on the real corpus. In Section V, as the conclusion, we mention the significances and the remaining tasks of this research.

II. Previous Works

This section is concerned with the exploration for the previous works relevant to this research. In 2000, Jo invented a new neural network, proposing encoding documents into string vectors; it provides the motivation for doing this research [1]. The semantic relations between words are considered for doing information retrieval tasks such as ranking and term weighting. Even for doing other tasks, the semantic relations are also considered. Therefore, in this section, we will explore previous works in terms of string vector encoding and tasks involving the semantic relations between words.

This research is initiated from encoding documents into string vectors, instead of numerical vectors, for doing text mining tasks. Encoding documents so was initiated by Jo in 2000, inventing the new neural network, called NTC (Neural Text Categorizer), as a practical approach to text categorization [1]. Subsequently, in 2005, Jo and Japcowicz invented the unsupervised string vector based neural network which was called NTSO (Neural Text Self Organizer) [6]. In 2009, Jo modified the KNN and SVM into its string vector based versions where the similarity measure between string

ISBN: 978-1-61208-217-2

vectors was based on the semantic relations between words [7]. However, in order to use string vectors more freely, we need to define more semantic operations on strings and characterize them mathematically.

The semantic relations between words are considered especially in information retrieval tasks. In 2005, Shenkel et al. implemented the search engine which was called XXL, for searching for XML documents, using the semantic similarity between words, based on ontology and an index structure [8]. In 2005, Possas et al. proposed a term weighting scheme which was called ‘set based model’, considering the semantic relation between term [9]. In 2008, Vechtomova and Karamuftuoglu used the semantic relation between a query and terms for ranking retrieved documents [10]. The previous works show usefulness of the semantic relation between words in the domain of information retrieval.

The semantic relation between words may be considered in other tasks as well as the information retrieval tasks. In 1994, Kiyoki et al. defines metadata of image as words for representing their semantic relations for the image retrieval [11]. In 2004, Makkonen et al. defines semantic relations among words using ontology for doing the topic tracking and detection [12]. In 2007, Na et al. used the semantic relation between a query and terms for adjusting clustering results [13]. The previous works show that the semantic relation may be considered in various tasks.

This research is intended to define various semantic relations between words, assuming that each string has its own meaning. In the previous works, the semantic relations have been considered not mathematically but informally or implicitly. In other words, the mathematical foundations are not founded, yet; the computation of the semantic similarity has depended on very heuristic computations. Even if the modification and creation of string vector based approaches in favor of text categorization and clustering was successful, it was limited to process string vectors because of no more systematic mathematical foundations. Therefore, the goal of this research is to define more semantic operations on strings and characterize them algebraically, in order to overcome the limitation.

III. Numerical Semantic Operations

This section describes the semantic operations in detail and consists of the four sections. In Section III-A, we describe the similarity matrix as the basis of carrying out the semantic operations. In Section III-B, we mention the two opposite operations: semantic similarity and semantic distance. In Section III-C, we define the SSA formally and characterize it mathematically. Section III-D covers the SSV like the SSA.

A. Similarity Matrix

Before entering the semantic operations on strings, we will describe the similarity matrix in this section. The similarity matrix is used as the basis for performing the semantic operations on strings. In the similarity matrix, each of its rows and columns corresponds to a string. The matrix has the two properties: its elements are symmetry and its diagonal elements are 1s. The similarity matrix defines the semantic similarity of each of all possible pairs of strings, and it assumes that the matrix is always given before doing the operations on strings.

The similarity matrix refers to the square matrix which defines the semantic similarity of each of all possible pairs of strings, and it is denoted as follows:

$$\begin{pmatrix} s_{11} & s_{12} & \dots & s_{1N} \\ s_{21} & s_{22} & \dots & s_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ s_{N1} & s_{N2} & \dots & s_{NN} \end{pmatrix}$$

The similarity matrix is given as the N by N matrix, and N indicates the total number of strings. Each of the columns and the rows corresponds to its unique string; both the ith row and the ith column correspond to the identical string. The element of the similarity matrix, s_{ij} indicates the semantic similarity between the string corresponding to the ith column and that corresponding to the jth row. Following the two properties, the similarity matrix may be built manually or automatically.

The first property of the similarity matrix is that its elements are symmetrical to each other. In other words, the rule $s_{ij} = s_{ji}$ applies to all elements in the similarity matrix. We already mentioned that the string corresponding the ith column is identical to that corresponding to the ith row. The two strings which correspond to the ith column and the jth column is same to those which correspond to the vice versa. The commutative raw is applicable to the semantic similarity between two strings.

The second property of the similarity matrix is that its diagonal elements are always given 1.0. In other words, s_{ii} is given as 1.0 as the maximum similarity. Every element in the similarity matrix is given as a normalized value between 0 and 1. The value, 1.0, signifies the maximum similarity between two strings. In the context of this research, it is assumed that the two identical strings have their maximal similarity.

The similarity matrix may be constructed, manually or automatically. A finite set of strings and the size of the similarity matrix are decided in advance. The 1.0 values are absolutely assigned to the diagonal elements of the similarity matrix. Keeping its symmetry property, normalized values between 0 and 1 are assigned to the off-diagonal elements. In other literatures, the process of building the similarity matrix from a corpus is mentioned; refer to the literatures for the detail description of the automatic construction.

B. Semantic Similarity and Distance

This subsection is concerned with the two base semantic operations on strings. One operation covered in this section is for evaluating how much two strings are similar based on their meanings. The other is for doing how much they are different from each other with respect to their meanings. The commutative law is applicable to both operations; the result is identical to the different order of the input strings. Therefore, in this subsection, we will describe the both operations with respect to their definition and properties.

The first base operation is for evaluating a semantic similarity between two strings. We already described the similarity matrix in Section 1, as the basis of these operations. It is possible to construct automatically the similarity matrix from a corpus, but the detail process is not covered in this article. As shown in Figure 1, the semantic similarity is carried out by retrieving directly the corresponding element from the similarity matrix as follows:

$$sim(str_i, str_j) = s_{ij}$$

This operation becomes the fundamental one for deriving more advanced operations, later.

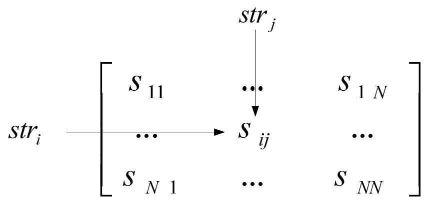

Figure 1. The Process of Retrieving the Semantic Similarity from the Similarity Matrix

The second operation is the semantic distance which is opposed to the previous operation. Like the semantic similarity, this operation generates a normalized value between 0 and 1 as the output. The semantic distance between two strings is computed by subtracting the semantic similarity from 1.0 as follows:

$$dis(str_i, str_j) = 1.0 - sim(str_i, str_j) = 1.0 - s_{ij}$$

The value generated from the semantic distance is the 1.0's complement of the semantic similarity. We may build the semantic distance matrix by subtracting each element from 1.0 as follows:

$$\begin{pmatrix} 1.0 - s_{11} & 1.0 - s_{12} & \dots & 1.0 - s_{1N} \\ 1.0 - s_{21} & 1.0 - s_{22} & \dots & 1.0 - s_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ 1.0 - s_{N1} & 1.0 - s_{N2} & \dots & 1.0 - s_{NN} \end{pmatrix}$$

Both operations are characterized as the fact that the commutative law is applicable. In the case of the semantic similarity, the commutative law applies because the similarity matrix is symmetry, as follows:

$$sim(str_i, str_j) = s_{ij} = s_{ji} = sim(str_j, str_i)$$

The commutative law also applies because the same value is subtracted from 1.0 as follows:

$$dis(str_i, str_j) = 1.0 - s_{ij} = 1.0 - s_{ji} = dis(str_j, str_i)$$

The similarity distance matrix becomes symmetry, but its diagonal elements are 0 values instead of 1.0 values. If the similarity distance matrix is given, the semantic distance is carried out by retrieving the corresponding element from the matrix.

C. Semantic Similarity Mean

This subsection is concerned with the first n-ary semantic operation on strings. The n-ary semantic operation refers to the class of semantic operations which takes an arbitrary number of strings as the input. In this operation, all possible pairs of strings are generated and the semantic similarity to each pair is computed. The semantic similarity mean of the strings is computed by averaging the similarities of the all possible pairs. In this subsection, we will describe the operation with respect to the definition, the properties, the procedure, and the utility.

This operation is denoted as follows:

$$avgsim(str_1, str_2, \dots, str_n) = \frac{2}{n(n-1)} \sum_{i<j} sim(str_i, str_j)$$

When n strings are given as the input, we generate $n(n-1)/2$ pairs of strings as all possible ones. For each pair, we may compute the similarity by retrieving it from the similarity matrix as shown in Figure 1. We obtain the average semantic similarity by summing the similarities of all pairs and dividing the sum by the number of all possible pairs, $n(n-1)/2$. The average semantic similarity signifies the semantic cohesion of the group of strings.

The properties of this operation are as follows:

- If all strings are identical, the average semantic similarity is given as 1.0 values, since the diagonal elements of the similarity matrix are given 1.0.
- $\frac{2}{n(n-1)} \sum_{i<j} sim(str_i, str_j) = \frac{2}{n(n-1)} \sum_{i>j} sim(str_i, str_j)$, since the similarity matrix is symmetry one.
- If all pairs of the strings are complementary (lowest similarity), the average semantic similarity becomes the minimum.
- The average semantic similarity is always given as a normalized value, since the similarities of all possible pairs are given as normalized values.

This operation takes an arbitrary number of strings as the input. Among the strings, all possible pairs are generated; if the number of strings is n, $n(n-1)/2$ pairs are generated.

For each pair, its similarity is retrieved from the given similarity matrix. The average semantic similarity is obtained by summing the similarities of the all possible pairs and dividing the sum by the number of pairs. Therefore, the averaged semantic similarity which is given as a normalized value is the output of the operation.

Figure 2 illustrates the two different groups of strings. The left group in Figure 2 contains the strings within the domain of computer science. The right group in Figure 2 contains the strings spanning over various domains. Intuitionally, the left group of strings has the higher average semantic similarity than the right group. Through the example illustrated in Figure 2, this operation may be used for estimating the cohesion of groups of strings.

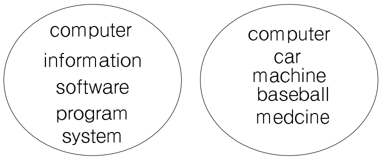

Figure 2. Two Groups of Words: One in a specific domain and the other in various domains

D. Semantic Similarity Variance

This subsection is concerned with the second n-ary semantic operation on strings. Under an identical average semantic similarity, there exist different distributions of similarities of pairs of strings. The pairs of strings may concentrate on the average semantic similarity, or they may disperse from it. We need the measure how much the similarities of the pairs concentrate on the average semantic similarity. In this subsection, we describe the operation in detail with respect to its definition, properties, and procedure.

The operation, called semantic similarity variance, is denoted as follows:

$$var(str_1,\ldots,str_n)$$
$$=\frac{2}{n(n-1)}\sum_{i<j}(sim(str_i,str_j)-avgsim(str_1,\ldots,str_n))^2$$

If n strings is given as the input, the number of all possible pairs becomes $n(n-1)/2$. Before performing this operation, the average semantic similarity should be computed by the operation which was mentioned in Section III-C. This operation focuses on the individual square of difference between a similarity of each pair and the average semantic similarity. This operation corresponds to the variance in the context of statistics.

The properties of this operation are as follows:

- $\frac{2}{n(n-1)}\sum_{i<j}(sim(str_i,str_j) - avgsim(str_1,\ldots,str_n))^2 = \frac{2}{n(n-1)}\sum_{i>j}(sim(str_i,str_j) - avgsim(str_1,\ldots,str_n))^2$

  It means that swapping indexes of elements does not influence on computing the average semantic variance, since the similarity matrix is symmetric as follows:
- $sd(str_1,\ldots,str_n)=\sqrt{var(str_1,\ldots,str_n)}$

$sd(str_1,\ldots,str_n)$ is called the semantic similarity standard deviation.

In this operation, an arbitrary number of strings is given as the input. Using the operation which was mentioned in Section III-C, the semantic similarity average is computed. For each pair, the difference square between its similarity and the average semantic similarity is computed. The difference squares are averaged into the semantic similarity variance. The square root of the semantic similarity variance becomes the semantic similarity standard deviation. Whether it is the variance or standard deviation, the value is always given as a normalized value.

The operation may be used for judging whether words are distributed, randomly or not. Let us consider the two groups of words with their identical semantic similarity. One group whose semantic similarities are concentrated on the average semantic similarity has very small the semantic similarity variance. However, the other whose semantic similarities are dispersed very much has the larger semantic similarity variance. In this case, the latter group is judged as the random distribution of words.

IV. Simulations

This section is concerned with the set of simulations of carry out the semantic operations on strings. We used the collection of news articles called 'NewsPage.com' in this research as the source from which the similarity matrix is built. The similarities among words are computed automatically based on the number of texts where the words are collocated with each other. We selected words from the corpus at random and we applied the semantic operations to them. In this section, we present and discuss the simulation results from applying the semantic operations.

We illustrate the specification of the collection of news articles called 'NewsPage.com' in Table I. The collection was constructed by copying and pasting individual news articles provided by the web site, 'newspage.com' as plain texts files. The five categories were predefined and 1,000 articles are available in the collection. Previously, it had been used as the test data for evaluating the approaches to text categorization. However, in this research, we use it as the source from which the similarity matrix is built.

This set of simulations is carried out with three steps: indexing the corpus, constructing the similarity matrix, and carrying out the semantic operations on strings. The corpus, which is the collection of texts, is indexed into a list of words and their frequencies as shown in Figure 3. We selected 100 words randomly and built the 100 X 100 similarity matrix by computing semantic similarities among words based on the number of texts where the words collocates with each

Table I
CATEGORIES AND NUMBER OF ARTICLES IN CORPUS: NEWSPAGE.COM

Category Name	#Articles
Business	400
Health	200
Law	100
Internet	300
Total	1000

other. We made 16 lists each of which consists of five words by selecting words randomly among the selected 100 words and applied the semantic similarity average and variance to each list. We generated values of the semantic similarity averages and variances as results of this set of simulations.

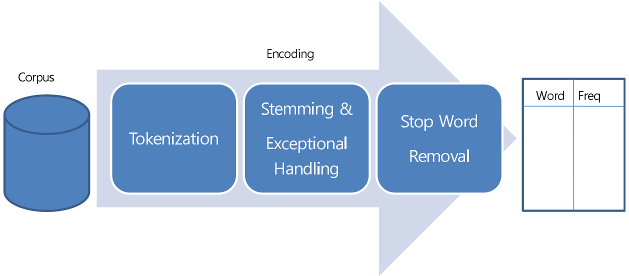

Figure 3. The Process of Indexing Corpus

In Figure 4, we illustrate the simulation results from carrying out the operations whose basis is constructed from NewsPage.com. In Figure 4, each position in the x-axis corresponds to a list of words. In the y-axis, each value indicates a normalized one between zero and one of the two operations. The gray bar and the white bar indicate values of SSA (Semantic Similarity Average) and SSV (Semantic Similarity Variance), respectively.

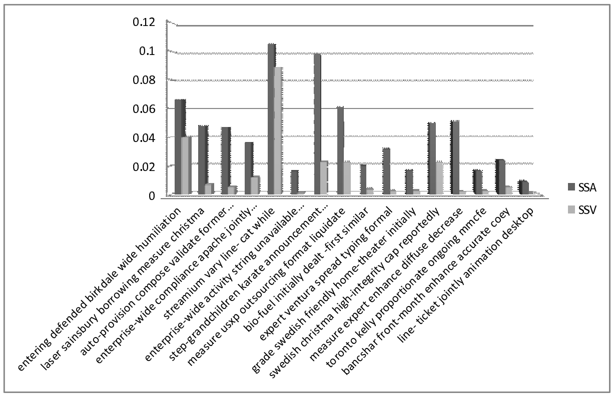

Figure 4. The Simulation Results from the SSM and SSV from the Corpus: NewsPage.com

Let us consider the results from simulating the two operations illustrated in Figure 4. The list which contains step-grandchildren, karte, announcement, and other two words, has the high SSA and the low SSV, as shown in Figure 4. The list is characterized by its values as dense semantic relations among the five words. The list with streaminum, vary, line-cat, and other two words, have large values of both SSA and SSV; it indicates that majority of words are related semantically densely, and minority are related loosely. The list with line-ticket, jointly, animation, desktop, and so on has low values of both SSA and SSV; it is characterized as loose semantic relation among them.

V. CONCLUSION AND FUTURE WORKS

Let us consider the significances of this research. From this research, we obtain the chance to measure semantic relations among words by the two simple operations: semantic similarity and semantic distance. We are able to observe the semantic cohesion of words through the operation called SSA. It gets possible to observe the distributions over semantic similarities of words, through the operation called SSV. This research provides potentially the way of developing semantic analyzer of textual data.

In spite of the above significances, let us consider remaining tasks for proceed further research. We need to make more simulations of carrying out the operations in other domains. More semantic operations will be defined and characterized mathematically. When the complexity of performing the semantic operation is high, it is necessary to reduce the complexity by developing their approximating algorithms. The operations will be applied to text processing tasks in information retrieval systems and text mining systems.

REFERENCES

[1] T. Jo, “NeuroTextCategorizer: A New Model of Neural Network for Text Categorization”, pp. 280-285, The Proceedings of ICONIP 2000, 2000.

[2] T. Jo, ”The Implementation of Dynamic Document Organization using Text Categorization and Text Clustering”, PhD Dissertation of University of Ottawa, 2006.

[3] T. Jo and D. Cho, “Index Based Approach for Text Categorization”, pp. 127-132, International Journal of Mathematics and Computers in Simulation, vol 2, no 1, 2008.

[4] T. Jo, “Topic Spotting to News Articles in 20NewsGroups with NTC”, pp. 50-56, Lecture Notes in Information Technology, vol 7, 2011.

[5] T. Jo, “Definition of String Vector based Operations for Training NTSO using Inverted Index”, pp. 57-63, Lecture Notes in Information Technology, vol 7, 2011.

[6] T. Jo and N. Japkowicz, “Text Clustering using NTSO”, pp. 558-563, The Proceedings of IJCNN, 2005.

[7] T. Jo, “Modification of Classification Algorithm in Favor of Text Categorization”, pp. 13-23, International Journal of Computer Science and Software Technology, vol 2, no 1, 2009.

[8] R. Schenkel, A. Theobald, and G.Weikum, “Semantic Similarity Search on Semistructured Data with the XXL Search Engine”, pp. 521-545, Information Retrieval, vol 8, no 4, 2005.

[9] B. Possas, N. Ziviani, W. Meira, Jr., and B. Ribeiro-Neto, “Set-based vector model: An efficient approach for correlation-based ranking”, pp. 397-429, ACM Transactions on Information Systems, vol. 23, no. 4, 2005.

[10] O. Vechtomova and M. Karamuftuoglu, “Lexical cohesion and term proximity in document ranking”, pp. 1485-1502, Information Processing & Management, vol 44, no 4, 2008.

[11] Y. Kiyoki, T. Kitagawa and T. Hayama, “A Metadatabase System for Semantic Image Search by a Mathematical Model of Meaning”, pp. 34- 41, ACM SIGMOD Record, vol 23, no 4, 1994.

[12] J. Makkonen, H. Ahonen-Myka, and M. Salmenkivi, “Simple Semantics in Topic Detection and Tracking”, pp. 347-368, Information Retrieval, vol 7, no 3-4, 2004.

[13] S. Na, I. Kang and J. Lee, “Adaptive document clustering based on query-based similarity”, pp. 887-901, Information Processing & Manage- ment, vol 43, no 4, 2007.