Turk. J. Math. Comput. Sci.
15(2)(2023) 334–354

DOI : 10.47000/tjmcs.1240729

Revisiting Probabilistic Relation Analysis: Using Probabilistic Relation Graphs for Relational Similarity Analysis of Words in Short Texts

Dima Alnahas[1,*], Abdullah Ates[2], Ahmet Arif Aydin[2], Baris Baykant Alagoz[2]

[1]*R&D Department, Infina Software Inc., Istanbul, Turkey.*
[2]*Department of Computer Engineering, Inonu University, Malatya, Turkey.*

Received: 23-01-2023 • Accepted: 08-08-2023

Abstract. Relation graphs provide useful tools for structural and relational analyses of highly complex multi-component systems. Probabilistic relation graph models can represent relations between system components by their probabilistic links. These graph types have been widely used for the graphical representation of Markov models and bigram probabilities. This study presents an implication of relational similarities within probabilistic graph models of textual entries. The article discusses several utilization examples of two fundamental similarity measures in the probabilistic analysis of short texts. To this end, the construction of probabilistic graph models by using bigram probability matrices of textual entries is illustrated, and vector spaces of input word-vectors and output word-vectors are formed. In this vector space, the utilization of cosine similarity and mean squared error measures are demonstrated to evaluate the probabilistic relational similarity between lexeme pairs in short texts. By using probabilistic relation graphs of the short texts, **relational interchangeability** analyses of lexeme pairs are conducted, and confidence index parameters are defined to express the reliability of these analyses. Potential applications of these graphs in language processing and linguistics are discussed on the basis of the analysis results of example texts. The performance of the applied similarity measures is evaluated in comparison to the similarity index of the word2vec language model. Results of the comparative study in one of the illustrative examples reveal that synonyms with 0.18157 word2vec similarity value scored 1.0 cosine similarity value according to the proposed method.

2020 AMS Classification: 68T50, 65C20

Keywords: Bigram probability relations, probabilistic graph similarity, text similarity, relational interchangeability.

1. Introduction

Graph similarity has been widely utilized in the analysis of highly complex systems in several fields of science and engineering, such as chemistry [14, 38], biology [16, 25], computer vision [4], and social networks analysis [40, 50]. A common characteristic of these systems is that models of the systems can be expressed via a large network of many elements. These elements may involve very conditional interactions in complex networks. Classification of these types of systems and identification of their implicit or distinctive properties are achievable by exploring relational similarity in their graph models' structure. In linguistics, language structures or lexemes can establish highly sophisticated mutual relations because semantic relations, grammatical relations, correlated word sequences, etc., produce sophisticated

*Corresponding Author
Email addresses: dalnahas@infina.com.tr (D. Alnahas), abdullah.ates@inonu.edu.tr (A. Ates), arif.aydin@inonu.edu.tr (A. Aydin), baykant.alagoz@inonu.edu.tr (B. Alagoz)

relational graph models. Therefore, simplified analytical analysis techniques are needed to reach quantitative results on the properties of the language systems.

Short text in a language can be considered as representative of a complex lexical system of messaging and knowledge transcription, which can be encoded in textual patterns. For this reason, graph-based language models and graph similarity analysis may provide useful results related to the distinctive properties of a language, semantic relations among word sequences, and grammar formations that emerged around lexical items. Moreover, several studies have addressed the graph-based analysis of language properties in text analysis [6, 15, 32, 33, 39, 48]. A comprehensive list of applications of graph representations in text mining, Natural Language Processing (NLP), and information retrieval has been provided in [48]. Graph representations have been effectively utilized for event detection by using Twitter Streaming [27,28], similarity analysis in documents [34], keyword extraction in single document [44], and text classification [47]. All these works demonstrate practical use of graph modeling of language items.

For analysis of various diverse relations among language items, one of the effective ways is utilization of probabilistic graph models. These models allow computationally reduced representation of massive and repeated relations between lexemes of texts. It is evident that probabilistic graph models have advantages in reduction of model complexity when compared to the complexity of deterministic graph models. Probabilistic graph-modeling of short texts is widely performed by calculating bigram word transition probabilities [1,20,26]. Bigram probabilities of vocabulary set elements are represented by probabilistic directed graph structures, which can convey useful knowledge of word co-occurrences and the corresponding semantic relations between lexemes to carry out a lower complexity language processing application such as text similarity analysis [5, 12, 17, 19, 21–23, 31, 36, 46], grammar analysis [11, 35], exploration of grammar structures of a language [49] etc.

Graph representation of bigram probability matrix forms a probabilistic relation graph model, and this facilitates exploration of semantic and grammatical features in a finite-length text. Since word sequences of texts are directed, probabilistic relation graph should be a directed graph in order to preserve word transition information in the model. Nodes of the graph represent lexemes of text vocabulary, and edges of directed graph are weighted by bigram probabilities of words. Several statistical methods such as probabilistic latent semantic analysis [18], probabilistic latent classes [41] were employed to address NLP problems. Bigram probabilities of words have been generally estimated by calculating co-occurrence frequencies of bigram relations in a given text [1, 20]. In the literature, word-vector representations of co-occurrence probabilities and analysis in word-vector space have been successfully utilized in language processing applications [3, 7, 29, 30, 34, 37, 43–45, 47]. This has established our main motivation, and we consider word-vector couple that consists of input edge probabilities of a node to form input word-vector and the output edge probabilities of a node to form output word-vector. These word-vector couples can express probabilistic relations between lexemes with neighbor lexemes in the graph model of a short text and they establish a combined word-vector space to work on. Similarity measures are employed on this new word-vector space to express bigram relational similarities among lexemes of a text.

The majority of works make use of computational intelligence tools or machine learning methods to model semantic properties from the word-vector representation of text segments [3, 7, 29, 30, 37, 45]. Therefore, the learning skill of these tools explores and exploits the properties related to lexical items of texts. Those properties can vary between similarities, contrast, and irregularities [3, 29, 30]. This validates the case that semantic relations can be accessible by word-vector representations that are based upon the utilization of word-to-word co-occurrence probabilities or a wider window (n-gram) of relation probabilities. However, these studies do not attempt to conjecture the relational dynamics acting on the semantic properties of a language. Nevertheless, results of these successful works, which mostly rely on neural language processing [3,29,30,37] and distributional semantic models [7], demonstrated that semantic properties of a text can be absorbed by word vectors and these properties can be utilized for semantic analysis purposes. In these works, distance metrics (e.g., Euclidean distance, Manhattan distance, Minkowski p-distance, Kullback-Leibler (KL) divergence) were considered in a constructed reduced-dimension word vector space, which were based on frequencies of word co-occurrence [8]. These metrics were implemented effectively for geometrical exploration of word similarities and relatedness [7, 29]. However, geometrical analyses in metric spaces cannot entirely convey graph relations and word connectivity information. Another motivation of our study is to close this gap by combining word connectivity relations of graph models with word-vector space construction. Thus, the proposed hybrid methodology defines a word-vector space over probabilistic relation graphs of finite-length text entries. This allows the utilization of graph theory and probabilistic graph similarity techniques on word-vector analysis in order to gain more insight into relational

dynamics characterizing semantic features of a given short text. This type of effort can expand the distance-based word-vector analysis domain to graph-based probabilistic analysis domains and provide an opportunity to advance language analyses and language modeling tools. Furthermore, this can contribute to linguistic research and gain more insight into the mathematics of languages.

This study also provides a discussion of the utilization of probabilistic relation graphs in text analysis. An arithmetical foundation for relational interchangeability (RI) of lexemes and corresponding relational similarity analysis are introduced. Each word of vocabulary can be represented by a node of a directed probabilistic graph model of texts. In these graphs, relations among lexemes are expressed by two co-occurrence vector couples: input word vectors are formed by input edge bigram probabilities, and output word vectors are formed by output edge bigram probabilities. These word-vector couples construct an embedded space of lexemes, where cosine similarity (CS) and mean squared error (MSE) measures can be defined to perform relational similarity analysis. In our analysis, we observed that the relational interchangeability of lexemes could be beneficial for exploring word classes based on bigram relations in texts. However, the validity of relational similarity analysis strongly depends on the amount of knowledge that is gleaned from a given text content. Thus, we evaluated the relational information content of a given text and proposed a confidence index that is determined regarding the connectivity of lexemes in the probabilistic relation graphs. More connections infer more relational knowledge absorption of lexemes in text and increase the validity of relational similarity analysis for these lexemes. We hypothesize that when enough relation information is assimilated from the text, relational similarity of lexemes can, more confidently, express semantic relations among lexemes according to the context of the analyzed text. Moreover, illustrative examples are presented to demonstrate the utilization of CS and MSE measures for relational similarity analysis of lexemes in a given short text. Word classification and contextual connotation analysis according to word relations are studied in these examples.

Findings of this study can also be promising for natural language processing (NLP), for example, one can draw graph pictures of relational similarities that are extracted from words of a given text without any pre-knowledge and any training activity. These types of efforts can be an attempt to step towards establishing mathematical pictures of language use, and linguistic properties can be perceptible and theorized through graph pictures of languages.

2. Preliminaries and Problem Formulation

2.1. **Mathematical Foundation for Bigram Probabilities and Probabilistic Relation Graph Modeling of Finite Messages.** Probabilistic relational models have been used for learning or expressing probabilistic relations between objects in databases, and it has been considered as a learning tool to explore relations of noisy data in database [9, 13]. Probabilistic relation graphs have very frequent utilization in stochastic NLP studies, particularly for illustration of probabilistic relations among words of vocabulary [1, 8, 20, 26]. In general, Markov models and n-gram probabilities are used in stochastic NLP, and visualization of these models is carried out by probabilistic relation graphs.

Probability-weighted edges of relation graph express bigram probability distribution of word co-occurrences according to a given text. This section is devoted to fundamentals of co-occurrence matrix construction and bigram probability relation modeling of a text entry.

Let us express n-element vocabulary set $W = \{w_1, w_2, w_3, .., w_n\}$. Elements of a vocabulary set are lexemes that are commonly referred to as words. A finite-length word sequence, namely a message, is assumed to convey information by sequencing elements of a vocabulary set. A message can be composed of an ordered sequence of words and written by

$$M = \{"w_i w_j w_k ... w_p" : w_i, w_j, w_k, ..., w_p \in W\}.$$

A common vocabulary for $M_1, M_2, .., M_k$ message series can be defined by union of message elements as

$$W = \{w_i : w_i \perp M_1 \vee w_i \perp M_2 \vee ... \vee w_i \perp M_k\},$$

where the $\perp$ operator is occurrence operator and $w_i \perp M_j$ operation denotes that w_i item appears in M_j message at least one time. Bigram co-occurrence relation of message was expressed by relation frequency matrix [20] and the bigram co-occurrence frequency matrix R_{f2} of a message M is constructed by co-occurrence counts of vocabulary elements in the message. The elements of R_{f2} matrix can be written by [1, 20]

$$R_{f2} = \begin{cases} f_{i,j} = f_{i,j} + 1 \ , "w_i w_j" \perp M \wedge \{w_i, w_j\} \in W \\ f_{i,j} = f_{i,j} \qquad \text{else.} \end{cases} \tag{2.1}$$

Equation (2.1) indicates that count of "$w_i w_j$" occurrences in the message M is set to the co-occurrence frequency $f_{i,j}$. Here, $i, j \in Z^+$ is vocabulary index of words. These index associate matrix elements with vocabulary elements. Matrix form of R_{f2} includes all $f_{i,j}$ frequencies for all vocabulary elements as follows [7]

$$R_{f2} = \begin{bmatrix} f_{1,1} & f_{1,2} & f_{1,3} & \cdots & f_{1,n} \\ f_{2,1} & f_{2,2} & f_{2,3} & \cdots & f_{2,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{n,1} & f_{n,2} & f_{n,3} & \cdots & f_{n,n} \end{bmatrix}.$$

A probabilistic bigram relation model of a text can be identified by normalizing values of elements of R_{f2} into a range of [0, 1] for an infinite length of symbol sequence. The bigram probability matrix can be expressed as [1]

$$\begin{aligned} R_{p2} &= \lim_{\sum R_{f2} \to \infty} \frac{1}{\sum R_{f2}} R_{f2} \\ &= \lim_{\sum R_{f2} \to \infty} \frac{1}{\sum_{i=1,j=1}^{n,n} f_{i,j}} \begin{bmatrix} f_{1,1} & f_{1,2} & \cdots & f_{1,n} \\ f_{2,1} & f_{2,2} & \cdots & f_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ f_{n,1} & f_{n,2} & \cdots & f_{n,n} \end{bmatrix}, \end{aligned}$$

where $\sum R_{f2}$ is the sum of the element of matrix R_{f2} and calculated by $\sum_{i=1,j=1}^{n,n} f_{i,j}$. The sum term $\sum R_{f2} \to \infty$ refers to the case of infinite-length text that is required to reach exact probabilities. In practice, analyzed texts are finite in length, namely short texts. For a finite length symbol sequence, $\sum R_{f2}$ is finite, and probability theorem suggests that finite samples can provide an approximation to probability measurement, therefore bigram relation probability matrix can be approximately identified from a finite length text, which can be expressed as

$$R_{p2} \cong \frac{1}{\sum R_{f2}} R_{f2} \begin{bmatrix} p_{1,1} & p_{1,2} & p_{1,3} & \cdots & p_{1,n} \\ p_{2,1} & p_{2,2} & p_{2,3} & \cdots & p_{2,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{n,1} & p_{n,2} & p_{n,3} & \cdots & p_{n,n} \end{bmatrix}, \tag{2.2}$$

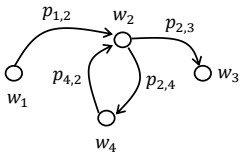
where $p_{i,j}$ is an estimation of bigram probability of i^{th} and j^{th} elements of a vocabulary set. It denotes the conditional probability of w_j occurrence under condition of being a follower of w_i occurrence in text and it can be calculated as [1, 8, 20]

$$P(w_j \| w_i) \cong p_{i,j} = \frac{f_{i,j}}{\sum R_{f2}}.$$

Simply, $p_{i,j}$ approximates to probability of "$w_i w_j$" co-occurrence in the message M, and it is employed to identify bigram relations within the message. Longer texts increase accuracy of bigram probabilities and accordingly, relevancy of bigram relation modeling from text. Calculation of R_{p2} matrix for a message implies unsupervised learning of relation probabilities among words of the message. The matrix R_{p2} represents a type of Markov chain, where the sum of all edge probabilities is 1. In conventional Markov chain graphs, which also express a finite state diagram [26], total probability of state transitions (edges) from a state (node) is 1, therefore sum of all edge probabilities depends on number of nodes in Markov chain.

The probabilistic bigram relation models are represented by a directed graph model $G_{p2} = (W, R_{p2})$, where vocabulary set stands for the node set of the graph, and bigram probability matrix R_{p2} defines probability-weighted edges among nodes of the graph. Here, both row and column elements of R_{p2} represent elements of the vocabulary set.

Element $p_{i,j}$ expresses the probability weight of the edge between node w_i and node w_j. Since G_{p2} is a directed graph, each node has input and output edges. Therefore, each edge in directed graphs contributes to an input relation of a node and an output relation of another node. In other words, the edge of $p_{i,j}$ is an output edge for the node w_i and an input edge for the node w_j. Figure 1 illustrates an example G_{p2} graph and its bigram probability matrix R_{p2}. In this graph representation, if $p_{i,j} \neq 0$, a probabilistic relational edge between node i and node j exists.



$$R_{p2} \cong \begin{bmatrix} 0 & p_{1,2} & 0 & 0 \\ 0 & 0 & p_{2,3} & p_{2,4} \\ 0 & 0 & 0 & 0 \\ 0 & p_{4,2} & 0 & 0 \end{bmatrix}$$

FIGURE 1. An example G_{p2} graph and its probabilistic bigram relation model R_{p2}

The output word vector for the node w_i is composed of the i^{th} row elements and it is expressed as

$$u_i = [p_{i,1} \quad p_{i,2} \quad p_{i,3} \quad \cdots \quad p_{i,n}].$$

The input word vector for the node w_i is composed of the i^{th} column elements and it is expressed as

$$v_i = [p_{1,i} \quad p_{2,i} \quad p_{3,i} \quad \cdots \quad p_{n,i}]^T.$$

Probabilistic relation of lexeme w_i is expressed by vector couple (u_i, v_i), where $u_i \in R^n$ and $v_i \in R^n$. For instance, the first row of R_{p2} matrix in Figure 1 is output word-vector for the lexeme w_1 that is obtained as $u_1 = [0 \quad p_{1,2} \quad 0 \quad 0]$, and the first column of R_{p2} matrix in Figure 1 is input word-vector for the lexeme w_1 that is obtained as w_1 that is obtained as $v_1 = [0 \quad 0 \quad 0 \quad 0]$. Probabilistic relations of the lexeme w_1 with other lexemes are expressed by word-vector couple (u_1, v_1).

These vector couples can be utilized in similarity analysis by measuring the probabilistic relation between lexemes using similarity metrics.

3. Similarity Analyses on Probabilistic Relation Graphs of Bigram Probability Matrices

In this section, commonly utilized similarity measures are applied to probabilistic relation graph models. Figure 2 depicts some example motifs for relationally interchangeable lexemes in probabilistic relation graphs in order to develop understanding of node connectivity-based relational similarity. In this context, relational interchangeability (RI) term infers similar input and output-edge connection patterns of neighbor nodes. Hence, relational similarity of nodes is defined according to similarity of their input word vectors and their output word vectors.

3.1. Relational Interchangeability in Probabilistic Relation Graphs of Short Texts.

When probabilistic relation vectors (u_i, v_i) and (u_j, v_j) of lexemes w_i and w_j are identical, they are called **relational interchangeable lexemes**. The fully relational interchangeable lexemes can be arithmetically defined by the conditions of $u_i = u_j$ and $v_i = v_j$. The relational interchangeability property of nodes can be evaluated by measuring similarity between output word vectors u_i and u_j and input word vectors v_i and v_j. In graph theory, degree of node is the number of edges that are connected to the node. The sum of degrees of w_i and w_j, which is written by $deg\{w_i\} + deg\{w_j\}$, has significance in RI analysis. In the following subsection, total degree of nodes is utilized for confidence analysis.

The higher degree of nodes increases validity of interchangeability analysis. Node degree is the number of relations that are established with other nodes: when the relational interchangeable lexemes w_i and w_j have fewer numbers of edges, this case can be accounted as a weaker analysis. However, relational interchangeable lexemes with a larger number of edges result in stronger analysis because it is based on richer correlation in terms of probabilistic relations with their neighbor nodes. Therefore, the confidence of RI analysis of lexemes depends on density of graph connectivity. The RI of lexemes can be used as an indicator that can detect relational word classes, and we have a prospect that possessing strong RI can be also an indication of semantic similarity of lexemes according to the content of analyzed text. Figure 2 shows two interchangeable lexemes, where the left-hand side graph allows a weaker RI analysis than that of the right-hand side graph because it is based on two identical edge groups ($p_{1,2} = p_{1,4}, p_{2,3} = p_{4,3}$) whereas the right-hand side graph has three identical edge groups ($p_{1,2} = p_{1,4}, p_{2,3} = p_{4,3}, p_{2,5} = p_{4,5}$). Accordingly, the sum of degree of w_i and w_j, which was given by $deg\{w_i\} + deg\{w_j\}$, can be considered an indication of strength of RI in the graph analysis.

Remark 3.1. Fully relational interchangeable nodes provide equal transition probability paths in a probabilistic graph and such paths are identical in terms of probabilistic relations.

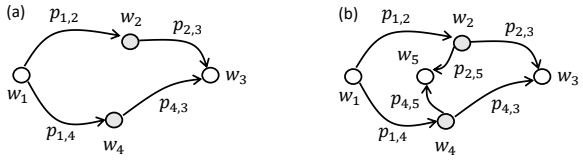


FIGURE 2. Relational interchangeable nodes are indicated by grey nodes in two graph configurations. Two input and two output edges are identical in (a), and three input and three output edges are identical in (b)

Proof. One can verify this remark by applying chain rule of probability in graphs. In Figure 2, for a transition from w_1 to w_3, transition probability of the path including w_2 node can be written by probability of $p_{1,2}p_{2,3}$, and transition probability of the path including w_4 node can be written by probability of $p_{1,4}p_{4,3}$. When w_2 and w_4 are fully relational interchangeable, one can easily state that $p_{1,2} = p_{1,4}$ (due to equality of input edge patterns) and $p_{2,3} = p_{4,3}$ (due to equality of output edge patterns). Accordingly, one can write, $p_{1,2}p_{2,3} = p_{1,4}p_{4,3}$ and this result indicates that path probabilities involving relational interchangeable nodes are equal and these paths are identical in term of probabilistic relations. □

In complex relational graphs, the relational similarity of lexemes w_i and w_j can be analyzed by similarity of probabilistic relation vectors (u_i, v_i). The following section addresses application of fundamental similarity measures such as cosine similarity and mean squared error of word-vector pair (u_i, v_i) to evaluate RI of words of short texts.

3.2. **Two Similarity Measures for Detection of Relationally Interchangeable Lexemes.** To measure RI of lexemes w_i and w_j, cosine similarity (CS) of output word vectors can be expressed as

$$c^u_{i,j} = \frac{u_i.u_j}{\|u_i\|\|u_j\|} = \frac{\sum_{h=1}^{n} p_{i,h}p_{j,h}}{\sqrt{\sum_{h=1}^{n} p^2_{i,h}}\sqrt{\sum_{h=1}^{n} p^2_{j,h}}}. \tag{3.1}$$

Also, CS of input word vectors can be expressed as

$$c^v_{i,j} = \frac{v_i.v_j}{\|v_i\|\|v_j\|} = \frac{\sum_{h=1}^{n} p_{h,i}p_{h,j}}{\sqrt{\sum_{h=1}^{n} p^2_{h,i}}\sqrt{\sum_{h=1}^{n} p^2_{h,j}}}. \tag{3.2}$$

Relational similarity based on CS can be expressed and normalized to the range of $[0, 1]$ as follows

$$c_{i,j} = \frac{1}{2}(c^u_{i,j} + c^v_{i,j}). \tag{3.3}$$

For $c_{i,j} = 1$, it infers that lexemes w_i and w_j are relational interchangeable according to analyzed graph. This similarity measure expresses only the pattern similarity by ignoring magnitude difference. One can calculate output CS matrix (C_{u2}) that is obtained for all lexemes in a vocabulary by

$$C_{u2} = R_{p2} \otimes R^T_{p2},$$

where the CS operator $\otimes$ performs calculation given by Equation (3.1) between vectors of R_{p2} and R^T_{p2}. Input CS matrix (C_{v2}) for all lexemes in a vocabulary can be written by using a CS operator as

$$C_{v2} = R^T_{p2} \otimes R_{p2}$$

By considering Equation (3.3), CS matrix C_2 can be expressed as

$$C_2 = \frac{1}{2}(C_{u2} + C_{v2}). \tag{3.4}$$

Another measure to assess RI of lexemes w_i and w_j is the mean squared error (MSE), for output-word vector, it can be expressed as

$$e^u_{i,j} = \frac{1}{n}\|u_i - u_j\|_2^2 = \frac{1}{n}\sum_{h=1}^{n}(p_{i,h} - p_{j,h})^2. \tag{3.5}$$

Also, MSE for input-word vector can be expressed as

$$e^v_{i,j} = \frac{1}{n}\|v_i - v_j\|_2^2 = \frac{1}{n}\sum_{h=1}^{n}(p_{h,i} - p_{h,j})^2. \tag{3.6}$$

MSE of lexeme pair can be expressed as follows

$$e_{i,j} = e^u_{i,j} + e^v_{i,j}. \tag{3.7}$$

For $e_{i,j} = 0$, it infers that lexemes w_i and w_j are fully interchangeable. One can calculate output MSE error matrix (E_{u2}), which is obtained for all lexemes in a vocabulary by MSE operator $\ominus$ in matrix form as

$$E_{u2} = R_{p2} \ominus R^T_{p2},$$

where the MSE operator $\ominus$ performs calculation defined by Equation (3.5) between vectors of R_{p2} and R^T_{p2}. Input MSE matrix (E_{v2}) of all lexemes in a vocabulary can be written by using MSE operator $\ominus$ for matrix as

$$E_{v2} = R^T_{p2} \ominus R_{p2},$$

where the elements of matrix E_{v2} can be calculated by Equation (3.6). By considering Equation (3.7), MSE matrix E_2 for single node RI can be expressed as

$$E_2 = E_{u2} + E_{v2}. \tag{3.8}$$

Figures 3 and 4 show a geometric interpretation of similarity operators in probabilistic relation space. Let us assume probabilistic relation vectors (u_1, v_1) and (u_2, v_2) of lexemes w_1 and w_2. Euclidean distance between input and output word vectors is the square root of the total squared error $n.e_{1,2}$ and cosine of angles between input and output word vectors is $c_{1,2}$. Zero value of MSE infers that input and output edges of these lexemes are exactly the same. However, CS does not suggest having exactly the same edges. It indicates that vectors are in the same direction, but the length of vectors (vector magnitudes) can differ. Therefore, CS is not sensitive to scaling of word vectors. CS can give value of one for the same input and output edge connections with different scaling factoring of probabilities. For example, for $u_2 = ku_1, k \in R^+$, CS of these input word vectors yields one because of scaling invariant property of CS. This property is useful for statistical insensitivity of the similarity measure to unbalanced repetition of some words in texts. This enables analysis of normalized relational patterns between words.

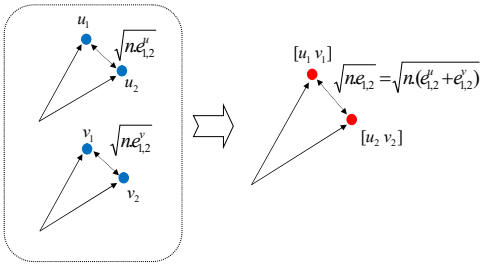


FIGURE 3. A geometric interpretation of MSE of relational similarity

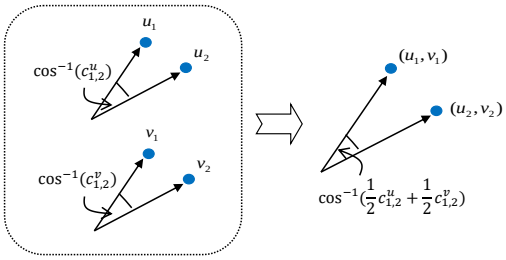


FIGURE 4. A geometric interpretation of CS of relational similarity

3.3. **Confidence Indices and Relational Interchangeability Analysis for Semantic Similarity.** Our tests indicated that the following two factors could play a role in the validity of RI index in order to express a semantic similarity relation of lexemes on the basis of analyzed text:

(i) *Connectivity of relational interchangeable lexemes with their neighbor lexemes:* The total degree of node pairs of lexemes w_i and w_j is expressed as sum of all edge counts for both pairs as $d_{i,j} = deg\{w_i\} + deg\{w_j\}$. Increase of node degrees with similar relational patterns increases possibility of RI to express semantic similarity of these lexemes.

(ii) *Information content absorbed by bigram probability matrix from short text:* When text content is not sufficient to absorb semantic relations, results of RI analysis cannot express semantic relations reliably. In a relation graph, edge density of nodes can be an indicator of richness of information absorption from the text. Therefore, average node degree of the probabilistic relation graph can be considered for assessing richness of relational semantic content of the graphs. In graph theory, node degree is a well-known property, and average node degree can be written for a probabilistic relation graph with a vocabulary size of n as

$$\gamma = \frac{\sum_{i,j}(deg\{w_i\} + deg\{w_j\})}{n}.$$

We proposed two confidence functions to estimate strength of a RI analysis to express a semantic relation. Let us assume that we have two interchangeable lexemes, w_i and w_j. A **capacitive confidence index** Co_C can be written

$$Co_C = (1 - e^{-d_{i,j}}) \tag{3.9}$$

and a **relative confidence index** can be calculated depending on the average node degree γ,

$$Co_R = \frac{d_{i,j}}{\gamma + d_{i,j}}. \tag{3.10}$$

Both confidence indices take values in the range of $[0, 1]$. When it is 1, the value expresses the highest confidence and when it is 0, it expresses the lowest confidence. To increase confidence, text length and, accordingly amount of text content should be increased to obtain denser relational interconnections between lexemes of text vocabulary.

Illustrative examples in the next section reveal that RI can be an effective tool to detect similar relational interaction motifs between words within a text in any language. We observed that these similar relational interaction motifs can be used for identification of connotations, synonyms, antonyms, grammar structure, etc.

4. Illustrative Examples for Short Text Relational Similarity Analysis

Let us process the following text:

Text 1: "Istanbul is a stunning cosmopolitan city full of museums, shopping, and world-class historical sites, and Mother Nature blessed it with many spots of natural beauty. Paris is a beautiful major city. It has many places of natural beauty as well. The city is also rich with museums and historical sites."

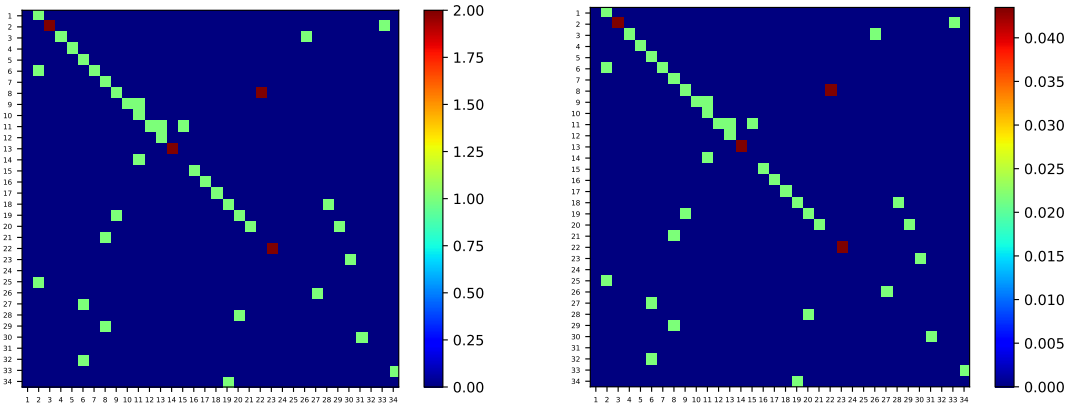
After Text 1 was cleaned from all punctuation marks (except full stops) and written entirely in lower case, a vocabulary set (W) of the text was obtained from a unique list of words as listed in Table 1.

Full stops have significance in co-occurrence relations of words. Therefore, a full stop is accounted as a lexeme in the vocabulary set for its role, indicating the end of a sentence and the beginning of another one. However, in this study, even though assigned a vocabulary index, the full stop frequency is assumed to be 0 to exclude it from lexical

Index	Word	Index	Word
1	Istanbul	18	it
2	is	19	with
3	a	20	many
4	stunning	21	spots
5	cosmopolitan	22	natural
6	city	23	beauty
7	full	24	.
8	of	25	Paris
9	museums	26	beautiful
10	shopping	27	major
11	and	28	has
12	world-class	29	places
13	historical	30	as
14	sites	31	well
15	mother	32	the
16	nature	33	also
17	blessed	34	rich

TABLE 1. Vocabulary of Text 1 in Example 1

relation calculation.
To represent words in matrices, vocabulary indices of lexemes are used for columns and rows index of the relation frequency matrix (R_{f2}) and probabilistic relation matrix (R_{p2}). The row index expresses output lexeme and the column index expresses input lexeme in relation. After normalizing values of R_{f2} elements by total frequency of R_{f2}, one obtains the probabilistic bigram relation model (R_{p2}) according to Equation (2.2). Figure 5a shows R_{f2} values and Figure 5b shows R_{p2} values. As Figure 5b demonstrates, row and column indices of the highest probability values indicate



(A) R_{f2} values for Example 1

(B) R_{p2} values for Example 1

FIGURE 5. R_{f2} and R_{p2} values for Example 1

lexeme couples that mostly co-occurred in the text, which are the word pairs (is, a), (historical, sites), (of, natural), (natural, beauty).
Figure 6 shows the probabilistic relation graph $G_{p2} = (W, R_{p2})$ that renders results of R_{p2} in Figure 5b. Figure 6 indicates one-node relational interchangeability of two lexemes (places, spots), as illustrated in Figure 2(a). As previously mentioned in Section 3, to evaluate RI of lexemes, two fundamental similarity measures are considered. These are

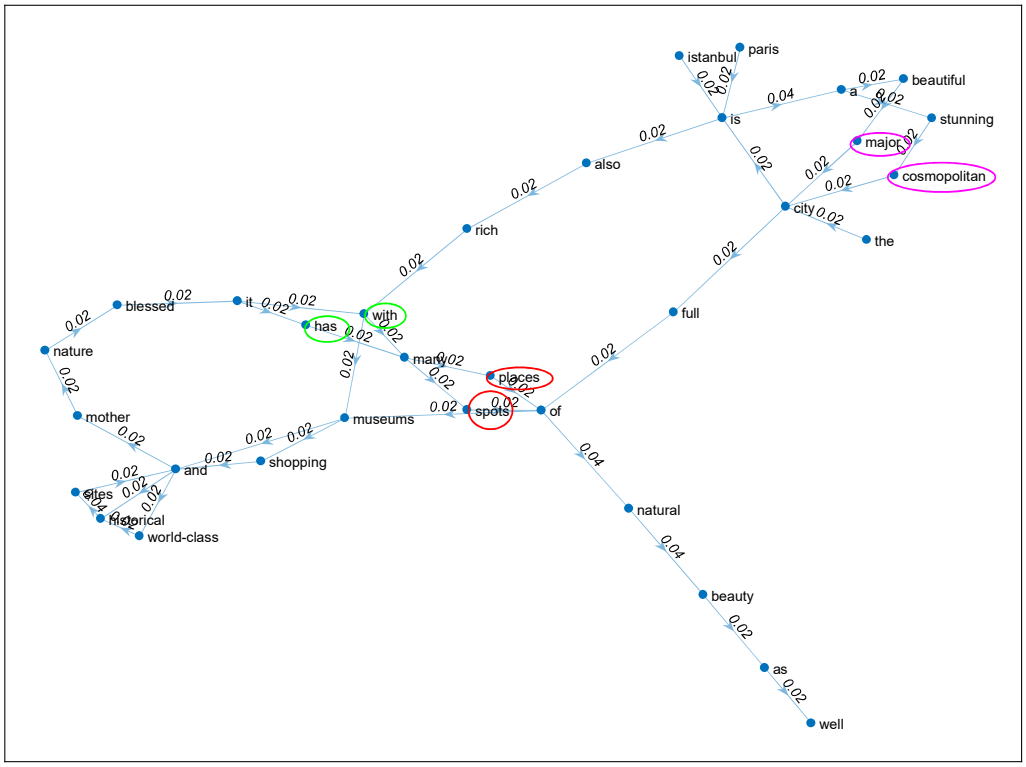


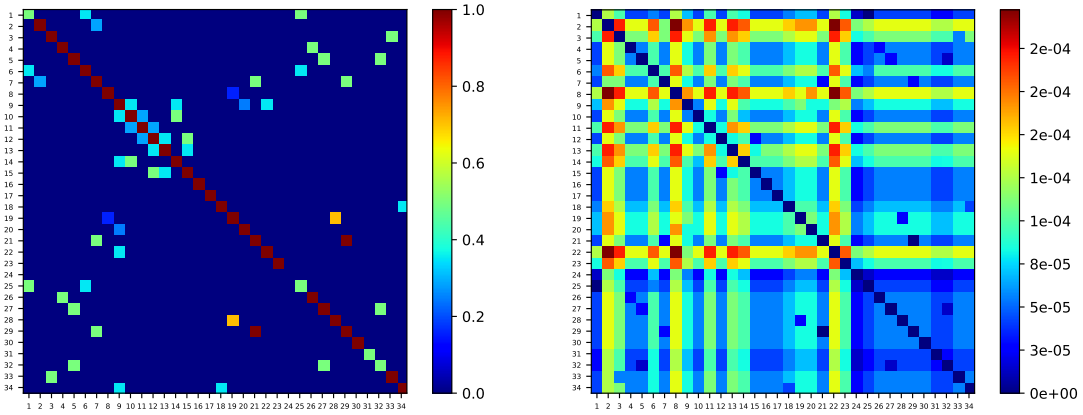
FIGURE 6. Probabilistic relation graph $G_{p2} = (W, R_{p2})$

cosine similarity and mean squared error. Besides, confidence indices are used to express validity of similarity analysis based on graph model of the text. The cosine similarity matrix (C_2) was calculated for R_{p2} according to Equation (3.4), and results are demonstrated in Figure 7a.

While calculating cosine similarities by using Equations (3.1) and (3.2), values of 0 in the denominator were changed to the value of 1 in order to avoid undefined operations such as $\frac{0}{0}$ and ∞ in calculation. Thus, we can avoid meaningless values in cosine similarity matrix during numerical calculations.

Elements of C_2 can take values between 0 and 1. The value of 1 infers corresponding lexemes presenting the same input and output relations according to probabilistic relation graph. In case of no relations with other lexemes, the corresponding elements of C_2 yields a zero value. One value on diagonal elements of C_2 matrix expresses similarity of lexemes by itself. When a diagonal element takes a value of 0.5, it infers that the lexeme has either a zero-input word-vector or a zero-output word vector. In other words, 0.5 values in diagonal elements indicate the lexemes that appear at the beginning or the end of a sentence. If it is mid-word of sentences, the diagonal element of lexeme becomes 1.

Table 2 shows that the lexeme pair (places, spots) has the highest level of cosine similarity with the value of 1 at the out-of-diagonal elements. This result suggests that these words are relationally interchangeable according to the graph. The lexeme pair (with, has) exhibits the second highest level of cosine similarity with a value of 0.70. These lexeme pairs may not be perfect synonyms of each other; however, these lexeme pairs can be categorized in similar word classes (connotation, synonyms, antonym, grammar structure etc.) in terms of co-occurrence relations between words of the text. When word counts and interconnection density of graph are increased by processing larger texts, the relational similarity analysis results can be more consistent and distinctive in terms of semantic similarity and grammatical relations. At this point, to evaluate information absorption level of probabilistic relation graph from Text 1, we used confidence indices that can be considered to evaluate validity of analysis results.



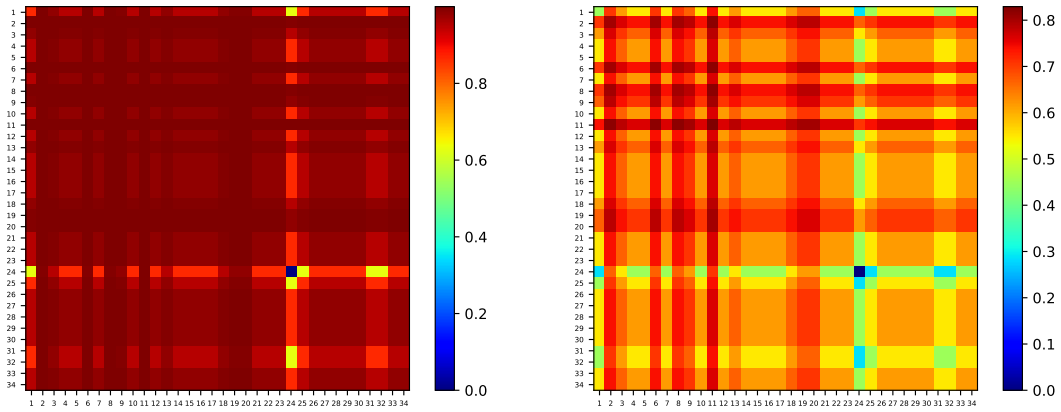
(A) C_2 values for Example 1 (B) E_2 values for Example 1

FIGURE 7. *C_2 and E_2 values for Example 1*

The MSE is another index to evaluate relational similarity of lexemes in probability relation graphs. MSE matrix E_2, which is calculated for each lexeme couple by using Equation (3.8), is shown in Figure 7b.

For MSE analysis, zero values of MSE matrix indicate RI of lexeme pairs according to co-occurrence relations of Text 1. Similar to cosine similarity results given in Figure 7a, MSE matrix indicated that the same word pairs (places, spots) and (Istanbul, Paris) were found relationally interchangeable so that they have zero MSE values. Corresponding to previous analysis, the word pairs (cosmopolitan, major) and (with, has) have slightly higher MSE values that are the value of 2.7710^{-5}.

Figure 8a shows capacitive confidence index values (Co_C) that were calculated by Equation (3.9). These confidence index values can be used to evaluate validity of the given similarity analysis. As observed in Figure 8a, values of



(A) Co_C values for Example 1 (B) Co_R values for Example 1

FIGURE 8. *Co_C and Co_R values for Example 1*

capacitive confidence index are not quite discriminative, to obtain more discriminative confidence values, Relative confidence index values are calculated (Co_R) as described in Equation (3.10). Figure 8b and Table 2 show this effect. Their values are calculated relatively with respect to average node degree of the graph. Table 2 shows a comparison between capacitive confidence index and relative confidence index for selected lexeme pairs, which have various values

of cosine similarity index. According to results in Table 2, the relational similarity analysis, given for (with, has), has the highest confidence with ACI of 0.99 and RCI of 0.70, because analysis result is based on the highest connectivity (with 6 relations) in the graph. The pairs (places, spots) and (cosmopolitan, major) have lower confidence. Due to less connectivity (relation) in the graph, confidence of analysis given for (Istanbul, Paris) lexeme pair is less.
Word2vec language modeling provides word embedding to metric spaces, which became a popular tool for distance-based semantic similarity analyses of very large corpora. The main advantage of word2vec scheme for text processing is that it allows a reduced-dimension language modeling by performing optimal word embedding to low-dimensional metric spaces and allows low-complexity geometric interpretations of semantic correspondences [10, 29, 30]. We performed word2vec similarity analyses for example, text and results of these analyses are added to Tables 2, 3, and 4 for evaluation of correspondence and contrasts between word2vec similarity analysis and the proposed probabilistic relation graph-based analysis. (To obtain a comparable result for similar setting, word2vec data was calculated for a window size of 2, dense vector size is set to 2, and minimum frequency word counts is set to 1 [10]).

In Table 2, the semantic correspondence between (places, spots) lexeme pair is suggested by C_2 with a high confidence index. Word2vec similarity index detects higher similarity for (Istanbul, Paris) lexemes. Word2vec similarity index can not detect strong semantic relation for (places, spots) pair. A possible reason for this effect is that word2vec language model relies on a distance-based optimal word embedding and it cannot preserve graph relations of words. Due to optimal mapping of low-dimension data to a low-dimension vector space, word2vec language models tend to generalize data in a reduced-dimension, continuous metric spaces, and therefore it has a tendency for preserving more common relational knowledge in language modeling, that is, when reducing dimension of word2vec model, it inherently discards scarce or sparse relational knowledge among lexemes of the text and considers them as noisy data. This property is commonly known as generalization of data when fitting to a reduced complexity model, and it makes word2vec models an effective and computationally efficient tool for very large and noisy corpora in practical NLP applications such as word classification, and online machine translation applications.

Probabilistic relation graph modeling enables to preserve all probabilistic relations in a corpus. Hence, its computational complexity grows depending on corpus size and therefore, it is more convenient for a detailed analysis of short texts. This property is achieved by proposing a hybrid approach that combines word-vector analysis with probabilistic relation graphs in order to analyze more detailed and sparse relational knowledge in short texts. This is a major contribution of this study.

It is noteworthy to state that probabilistic relation graph modeling promises numerous recourses for word relation analyses by involving graph properties, for instance, transitive semantic relation [1,8]. Some properties associated with transitive paths and cycles of lexemes within a graph structure can be investigated by using connectivity matrices [2], which are calculated by only taking power of relation matrices. The k-transition (edge) connectivity matrix R^k_{p2} is calculated in a recursive form as follows [2,42]

$$R^k_{p2} = R^{k-1}_{p2} R_{p2}; R^1_{p2} = R_{p2}$$

Lexeme pairs	Word Class	Degree sum of lexemes	Capacitive confidence index (Co_C)	Relative confidence index (Co_R)	Cosine similarity index (C_2)	Word2vec similarity index
places, spots	synonym	4	0.98168	0.61818	1	0.18157
with, has	connotation	6	0.99752	0.70833	0.70711	0.52245
cosmopolitan, major	connotation	4	0.98168	0.61818	0.5	0.39953
Istanbul, Paris	Similar object	2	0.8647	0.44737	0.5	0.94855

TABLE 2. A list of (Co_C), (Co_R) and (C_2) values of selected lexeme pairs for Example 1

The matrix R^k_{p2} expresses probability of k-transition paths through graphs and it reveals deeper transitive semantic relations that are stored in the matrix R_{p2} [1]. The diagonal elements of R^k_{p2} show k-transition cycle probabilities of lexemes, and the non-diagonal elements show k-transition path probabilities of lexemes. It is very useful to detect the most probable and the least probable k-transition cycles and paths within relation graph models of short texts [2]. Figure 9 shows values of R^3_{p2} that are calculated for Text 1 in this example. It presents three transition probabilities, which are cyclic triple probabilities on diagonal elements and acyclic word quadruples on non-diagonal elements. The high probable cyclic triples in the relation graph (Figure 6) is found (and, historical, sites) by considering high probability elements on diagonal of R^3_{p2}. This indeed indicates a most probably cycling emergence of the triple (and, historical, sites) comes from the parts "...of museums, shopping, and world-class historical sites, and Mother Nature blessed..." and "... is also rich with museums and historical sites. " in Text 1. By considering non-diagonal elements of R^3_{p2}, some high probable acyclic word quadruples are found as (places to beauty), (spots to beauty), and (full to beauty). The related parts in Text 1 are "... has many places of natural beauty.. ". "... Nature blessed it with many spots of natural beauty. ". The quadruples (full to beauty) are a form of connotation from "... city full of museums, shopping and.. " to "... of natural beauty" by bridging " of " words in the graph.

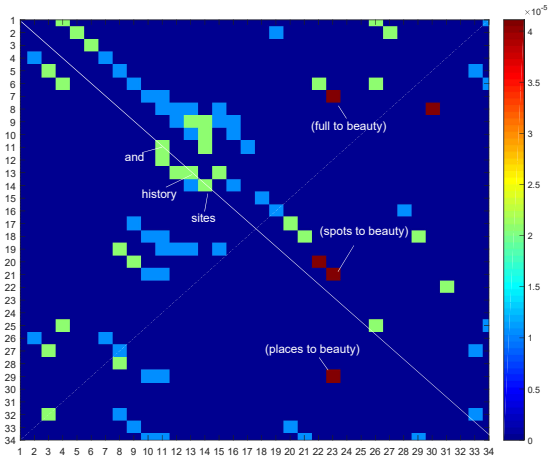


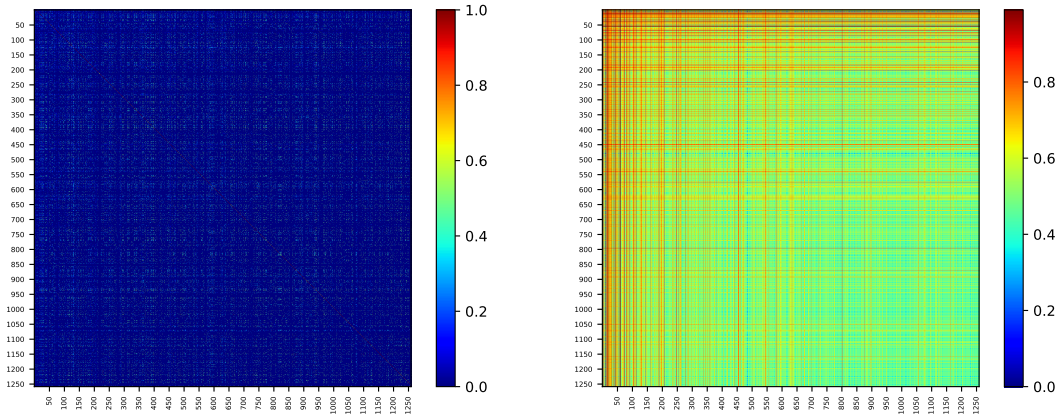
FIGURE 9. Some of 3-transition cycle and 3-transition path probabilities in R^3_{p2} for Text 1

In order to expand the amount of vocabulary set and allow the absorption of more information from the text entry, we analyze a larger text that consists of 3739 words and composes biographies of several famous scientists. By following the same stages in the first example, we calculated R_{f2} and R_{p2} matrices and obtained a vocabulary set of 1259 lexemes. Then, we calculated cosine similarity matrix C_2 of R_{p2} as depicted in Figure 10b. Figure 11 shows relative confidence indices for Example 2.

By filtering lexeme pairs that have a cosine similarity value of 1 (diagonal values were excluded from the filtered results), 309 lexeme pairs are obtained, which have a confidence index of 0.46, and most of them could not be evaluated as truly relational interchangeable such as; (1925, Switzerland), (respect, 1643), (addition, helping), (come, nothing).

To deepen our analysis, we filtered lexeme pairs that have confidence index values in the range of [0.8-0.9] and cosine similarity equal or greater than 0.5. Then, we obtained 77 lexeme pairs. When these pairs are reviewed, we observe that some of them present semantic relations and they can be interchangeable according to the relational knowledge absorbed from the context of this text. Table 3 shows some of the obtained results from Example 2.

Table 3 also includes word2vec similarity index of lexeme pairs for comparison purpose. One can observe that although the proposed method can detect connotation between (evolution, theory) pair with the highest similarity index C_2 and confidence index, the (evolution, theory) connotation is not so highly suggested in word2vec similarity index analysis. As mentioned in previous example, due to natural result of the word embedding in reduced-dimensional metric spaces, word2vec analysis is more effective for detection of generalizable or common semantic relations (e.g. word classes) in case of enough large text entry to explore common trends. For example, (heat, light) pair is detected



(A) C_2 values for Example 2

(B) Co_R values for Example 2

FIGURE 10. C_2 and Co_R values for Example 2

as the most similar words according to word2vec similarity index because the text is large enough to expose closeness between heat and light lexemes. Connotations between (light, waves), (achievements, discoveries), (become, be) are not strongly indicated by word2vec similarity index for this short text example.

The word2vec analysis needs processing on larger text to relevantly establish optimal distances between lexemes when embedding to a reduced dimension metric space. The short text can reduce consistency of the established word2vec spaces and accordingly semantic similarity analyses. The detection of connotations in short text is more straightforward by forming word vectors on the probabilistic relational graph model.

Lexeme pairs	Word Class	Degree sum of lexemes	Relative confidence index $0.8 \leqslant Co_R \leqslant 0.9$	Cosine similarity $C_2 \geqslant 0.5$	Mean Squared Error (1.0×10^{-9})	Word2vec similarity index
evolution, theory	connotation	22	0.82938	0.66822	7.2542	0.63266
light, waves	connotation	22	0.82938	0.53724	2.1635	0.53950
achievements, discoveries	connotation	19	0.80762	0.64589	1.0818	0.65553
heat, light	connotation	19	0.8076	0.56498	2.4181	0.98062
become, be	connotation	19	0.80762	0.60864	2.5453	0.68074

TABLE 3. List of lexeme pairs that have confidence index values in the range of [0.8-0.9] and cosine similarity equal or greater than 0.5

We present analysis results for a larger text that is composed of several short essays with various topics such as science, engineering, tourism, etc. It consists of 6916 words. By following the same stages in the previous examples, we calculated R_{f2} and R_{p2} matrices and obtained a vocabulary set of 2198 lexemes, and then, we calculated cosine similarity matrix C_2 and relative confidence indices. Figure 11 shows relative confidence index values for this text.

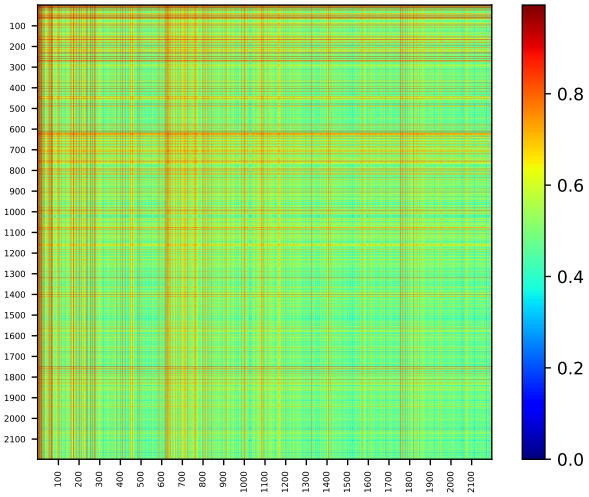


FIGURE 11. Co_R values for Example 3

By following the same stages in the second example, we filtered lexeme pairs that have a cosine similarity value of 1 (diagonal values were excluded from the filtered results), and 479 lexeme pairs were obtained, which have a confidence index of 0.45. To better evaluate results, we filtered lexeme pairs that have confidence index values in the range of [0.8-0.9] and cosine similarity equal or greater than 0.6. As a result, 43 lexeme pairs, which present relational similarity with higher confidence, were obtained. Table 4 shows some of the obtained results from Example 3. Upon reviewing these pairs, we observed that majority of them present semantic relations, and these pairs are detected as interchangeable according to relational context of the text.

Table 4 also presents word2vec similarity indices that are calculated for Example 3. In contrast to the previous example, the connotation between (evolution, theory) pair is highly suggested by word2vec similarity index because of increase in the length of text entry and diversity of topics in this example. This enables more optimal establishment of distances between word clusters in reduced-dimension metric space of word2vec model. Therefore, results of the proposed index C_2 and results of the word2vec similarity index are more consistent in this example. Word2vec similarity index highly suggests (cannot, can) pair because they are in the same grammar word classes. However, weak connotations (science, light) and (mechanics, physics) are not revealed by word2vec similarity index. We observed that the probabilistic relation graph-based approaches can be more effective to detect weaker connotations in short text analyses. On the other hand, Word2vec-based approaches are rather effective for analysis of word classes in case of an enough large text entry that allows better generalizations of lexeme relations. Cosine similarity is mainly related to similarity of probabilistic connectivity patterns of lexemes in the relational graph model and confidence index implies denser connections of lexemes within the graph. Similarity of connectivity patterns of densely connected lexemes is more reliable to detect interchangeable words such as synonyms, antonym, negations, or connotations in the context of the given text. Therefore, the region with higher values of both cosine similarity and confidence index is preferable for RI of words in the text and this region is shown to be filtering region in Figure 12. Table 4 demonstrates some results from this region and indicates importance of higher levels of confidence index for semantic interchangeability.

Figure 12 shows a distribution of lexeme pairs in two classes. The first one is non-grammar class which does not include any grammatical words and the second one is grammar class which includes at least one grammar word in the pair. In this figure, grammar class is represented by blue dots and the yellow dots indicate pairs of non-grammar class. Figure 12 demonstrates that grammar class and non-grammar class are clustered in different regions of cosine similarity and confidence index space and this allows separation of these regions by using Co_R and C_2 indices. This property can be utilized to detect grammar words for languages, where grammatical words are solely written in English and frequently repeated. The main reason of this effect can be the use of Co_R because a frequent repetition of grammatical words causes high average degree of lexeme pairs and leads to decreasing Co_R values of non-grammatical class pairs.

Lexeme Pairs	Word Class	Degree sum of lexemes	Relative confidence index $0.8 \leqslant Co_R \leqslant 0.9$	Cosine similarity $C_2 \geqslant 0.6$	Mean Squared Error (1.0×10^{-9})	Word2vec similarity index
cannot, can	similar grammar (negation)	38	0.88793	0.60672	1.7422	0.90688
science, light	connotation	30	0.86216	0.6454	3.3585	0.34704
time, universe	connotation	28	0.85376	0.64758	0.4513	0.47086
mechanics, physics	connotation	20	0.80658	0.66194	0.25189	0.26411
evolution, theory	connotation	24	0.83344	0.66482	1.4169	0.72257

TABLE 4. List of lexeme pairs that have confidence index values in the range of [0.8-0.9] and cosine similarity equal or greater than 0.6

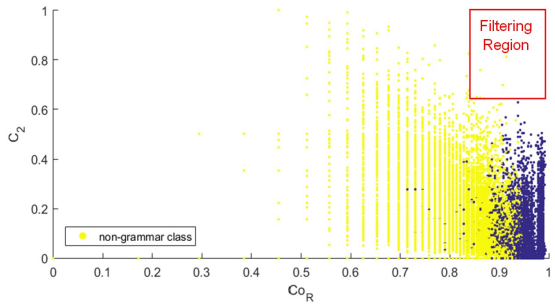


FIGURE 12. Distribution of grammar and non-grammar classes of lexeme pairs from Example 3

Furthermore, word probability is another useful index for detection of grammar words because these words appear with a high probability in texts.

In this example, we presented probabilistic relation analysis for a multi-language short text. For this purpose, a short English text is translated into several languages such as Turkish, Spanish, French, Arabic, and German. Figure 13 shows R_{p2} for concatenated corpus. This figure clearly shows that each language forms its own lexeme domain in R_{p2}. This indicates that extraction of probabilistic relation in an unsupervised manner naturally allocates different parts of R_{p2} matrix for different languages. This behavior interestingly corresponds to the suggestions on multilingualism which has been reported that while languages do share some areas of the brain, they also retain some separate neural areas in the brain [24]. R_{p2} matrix can model similar regional discrimination of language allocations based on word relations. Probabilistic relation graph analysis can be used as numerical tool for multilingualism research.

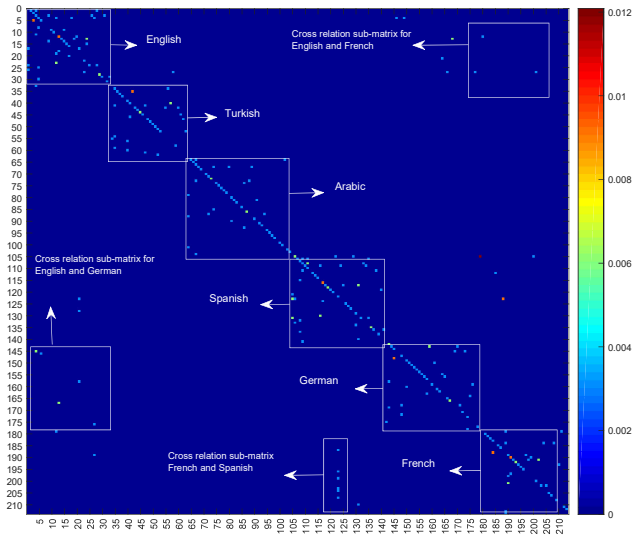


FIGURE 13. R_{p2} values for Example 4

Languages	1: English	2: Turkish	3: Arabic	4: Spanish	5: German	6: French
1: English	1.5258	.030517	0	0	.17320	.052314
2: Turkish	.030517	1.3092	0	0	0	0
3: Arabic	0	0	.88064	0	0	0
4: Spanish	.049486	0	0	1.3903	0	.20996
5: German	.17320	0	0	0	.99307	0
6: French	.052314	0	0	.16331	0	1.0605

TABLE 5. Probabilistic cross-relation density indices for probabilistic relation matrix in Figure 13. (All values in the table are normalized by 10^{-4})

Figure 14 shows probabilistic relation graph $G_{p2} = (W, R_{p2})$. The allocation of different domains can be observed in this graph as denser connection regions of the text of each language. Connections between different language domains are established by common words of languages. Such cross-relations between languages can be interpreted as a natural result of coming from the same Proto-language ancestor such as Spanish (blue) and French (black), and German (cyan) and English (red).

To illustrate application potentials of inter-language analyses on probabilistic relation graphs, we define a cross-relational probability density index that measures common word probability densities within cross-relation sub-matrices. The cross-relation sub-matrices appear in non-diagonal inter-language domains of multi-language probabilistic relation matrix. Some examples of cross-relation sub-matrices $Rs_{i,j}$ are illustrated in Figure 13. The probabilistic cross-relation density index for (i, j) language couple in a multi-language probabilistic relation matrix can be expressed as

$$LC(i, j) = \frac{1}{S(Rs_i, j)} \sum_{i,j \in Rs_{i,j}} p_{i,j},$$

where the operator $S(Rs_{i,j})$ represents size of $Rs_{i,j}$ sub-matrix (number of elements in $Rs_{i,j}$). Table 5 shows probabilistic cross-relation density index $LC(i, j)$ for assessment of probabilistic cross-relations among the six languages given in Figure 13. The value of $LC(4, 6)$ indicates high cross-relation from Spanish to French languages according to the analyzed text. Figure 15 shows graphical picture of data in Table 5. Edges are probabilistic cross-relation density between languages, which is calculated according to $(LC(i, j) + LC(j, i))/2$. It shows relational closeness between six languages in this example.

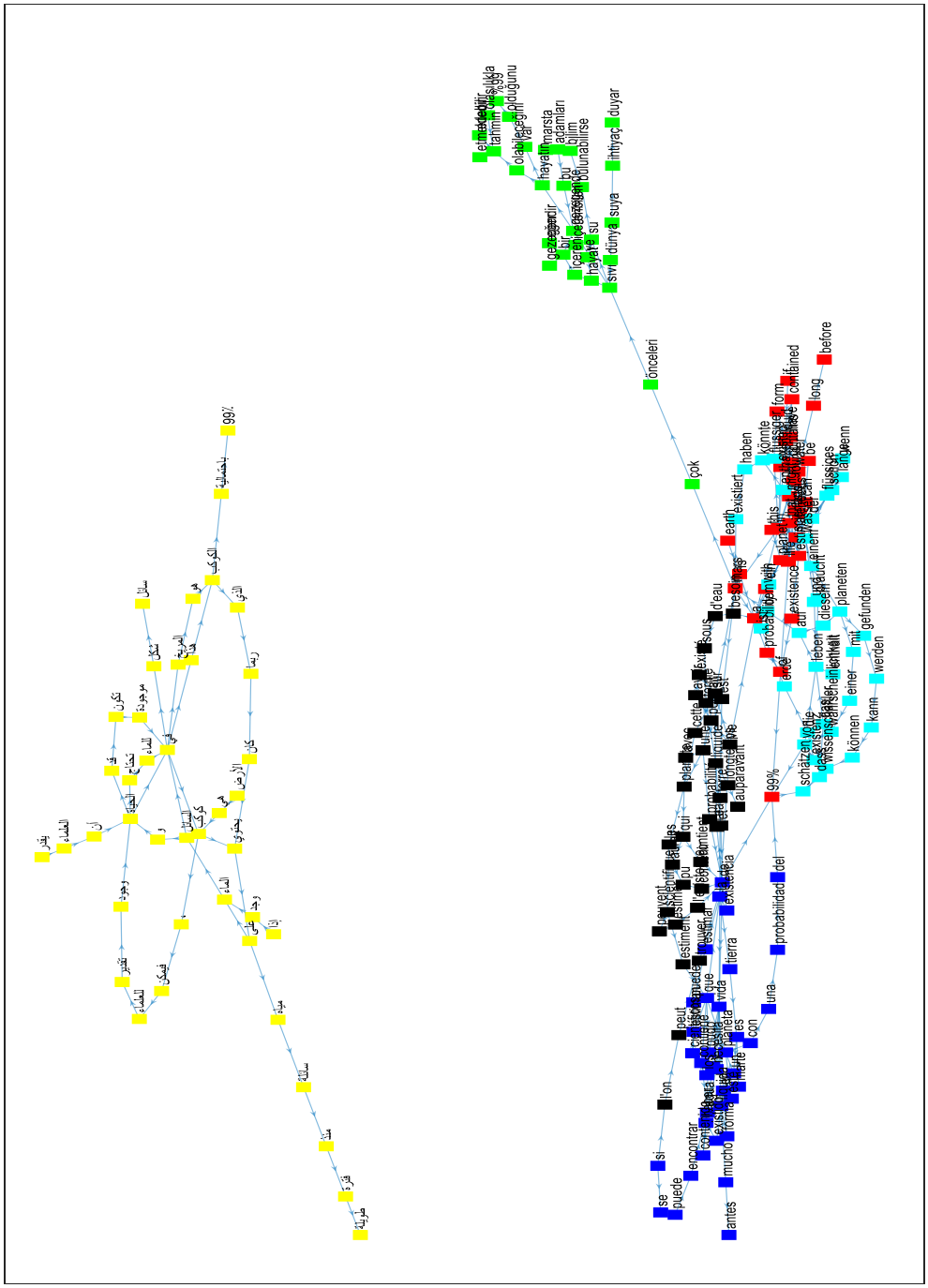


FIGURE 14. Probabilistic relation graph $G_{p2} = (W, R_{p2})$ for Example 4

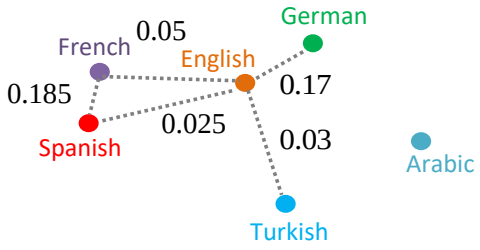


FIGURE 15. A probabilistic cross-relation density mapping of six languages according to multi-language text in Example 4

5. CONCLUSIONS

In this study, bigram probability analysis is revisited, and some opportunities related to graph-based analysis of texts are presented. The main focus of this article is relational similarity analysis of lexemes regarding node connectivity. CS and MSE similarity measures were derived according to input edge and output edge patterns for nodes of the directed graph. These patterns are defined in a space of word-vector couples that are input word vectors and output word-vectors. To evaluate validity of performed analysis, RI of lexemes is investigated, and a confidence index is proposed.

Some remarks of this study can be summarized as follows:

(i) Relational knowledge, which is contained by words of a short text, can be absorbed by probabilistic relation graph model without any pre-knowledge of a language or supervised training efforts. Probabilistic relation graph model of a text is considered as a mathematical depiction of word relations within a message. In fact, construction of bigram relation probability matrix can be considered as unsupervised learning of lexeme relations from a given text.

(ii) Graph properties can be used for formation of word vectors. This study demonstrates a word vector space construction based on input and output edge patterns of nodes. Thus, connectivity relations among lexemes can be conveyed in word vector spaces. We observed that graph similarity analysis based on input and output edge patterns similarity presents useful properties such as RI. The interchangeable lexemes have equal transition probability paths in the graph, and it can be effective for finding semantic similarities, relational similarities, and weak connotations in a short text. CS and MSE similarity indices are derived in relational word-vector space of bigram relation probability matrix.

(iii) Increasing connection density of a graph by providing and processing longer texts allows better representation of lexeme relations of a language and it increases confidence of analysis based on the relation graphs.

(iv) We observed that the proposed approach can be applicable for inter-language analysis that can allow investigation of relations or closeness of different language domains by using data in probabilistic relation matrix of multi-language messages. This analysis can also promise development of useful mathematical tools for linguistics researches and provide useful implications in linguistics and etymology.

(v) This probabilistic language model enables to create a language-independent mathematical picture of received messages or short texts. This feature makes this method a language-independent analysis tool for language exploration and inter-language analyses. A future work can address analysis of a message that is generated by mixture of sequential data from multiple unknown sources.

CONFLICTS OF INTEREST

The authors declare that there are no conflicts of interest regarding the publication of this article.

Authors Contribution Statement

All authors contributed to the design and implementation of the research, to the analysis of the results, and to the writing of the manuscript. The authors are responsible for correctness of the statements provided in the manuscript. All authors reviewed the results and approved the final version of the manuscript.

References

[1] Alnahas, D., Alagoz, B.B., *Probabilistic relational connectivity analysis of bigram models*, In 2019 International Artificial Intelligence and Data Processing Symposium (IDAP) (Malatya, Turkey, 2019), 379–384.

[2] Alnahas, D., Alagoz, B.B., *A theoretical study on event spreading prediction by probabilistic connectivity analysis in dispersive networks*, In 2019 International Artificial Intelligence and Data Processing Symposium (IDAP) (Malatya, Turkey, 2019), 590–595.

[3] Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C., *A neural probabilistic language model*, Journal of Machine Learning Research, **3**(2003), 1137–1155.

[4] Conte, D., Foggia, P., Sansone, C., Vento, M., *Thirty years of graph matching in pattern recognition*, International Journal of Pattern Recognition and Artificial Intelligence, **18**(2004), 265–298.

[5] Dogus, B., Guzel, G., *Development of matlab tool for text analysis*, Capstone Project presented at Inonu University, Computer Engineering Department, (2018).

[6] Erkan, G., Radev, D. R., *Lexrank: Graph-based lexical centrality as salience in text summarization*, Journal of Artificial Intelligence Research, **22**(2004), 457–479.

[7] Evert, S., Baroni, M., Lenci, A., *Distributional semantic models*, In Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT): Tutorial Abstracts (Los Angeles, CA, USA, June 2010), Association for Computational Linguistics, 15–18.

[8] Fallucchi, F., Zanzotto, F.M., *Transitivity in semantic relation learning*, In Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering (NLPKE-2010), (2010), IEEE, 1–8.

[9] Friedman, N., Getoor, L., Koller, D., Pfeffer, A., *Learning probabilistic relational models*, In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (1999), IEEE, 1300–1309.

[10] Ganesan, K., Gensim word2vec tutorial - full working example, 2018.

[11] Ganesh, B.R., Gupta, D., Sasikala, T., *Grammar error detection tool for medical transcription using stop words parts-of-speech tags ngram based model*, In Proceedings of the Second International Conference on Computational Intelligence and Informatics (Singapore, 2018), Springer, 37–49.

[12] Gardner, M., Mitchell, T., *Efficient and expressive knowledge base completion using subgraph feature extraction*, In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, (2015), 1488–1498.

[13] Getoor, L., Friedman, N., Koller, D., Taskar, B., *Learning probabilistic relational of relational structure*, In Proceedings of the Eighteenth International Conference on Machine Learning, (2001), 170–177.

[14] Hall, R.J., Murray, C.W., Verdonk, M.L., *The fragment network: A chemistry recommendation engine built using a graph database*, Journal of Medicinal Chemistry, **60**(2017), 6440–6450.

[15] Herrmannova, D., Knoth, P., Stahl, C., Patton, R., Wells, J., *Text and graph based approach for analyzing patterns of research collaboration: An analysis of the trueimpactdataset*, In 1st Workshop on Computational Impact Detection from Text Data (CIDTD) (Miyazaki, Japan, 2018).

[16] Heymans, M., Singh, A.K., *Deriving phylogenetic trees from the similarity analysis of metabolic pathways*, Bioinformatics, **19**(2003), i138–i146.

[17] Higgins, D., *Which statistics reflect semantics? rethinking synonymy and word similarity*, Linguistic Evidence: Empirical, Theoretical and Computational Perspectives, (2005), 265–284.

[18] Hofmann, T., *Probabilistic latent semantic analysis*, In Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, (1999), Morgan Kaufmann Publishers Inc, 289–296.

[19] Huang, A., *Similarity measures for text document clustering*, In Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008), **4**(Christchurch, New Zealand, 2008), 9–56.

[20] Jurafsky, D., Martin, J.H., *Speech and language processing: An introduction to natural language processing*, Computational Linguistics, and Speech Recognition, (2008).

[21] Konopik, M., Pražák, O., Steinberger, D., Brychcín, T., *Uwb at semeval-2016 task 2: Interpretable semantic textual similarity with distributional semantics for chunks*, In Proceedings of the 10th International Workshop on Semantic Evaluation, (2016), 803–808.

[22] Lin, Y.-S., Jiang, J.-Y., Lee, S.-J., *A similarity measure for text classification and clustering*, IEEE Transactions on Knowledge and Data Engineering, **26**(2013), 1575–1590.

[23] Lopez-Gazpio, I., Maritxalar, M., Gonzalez-Agirre, A., Rigau, G., Uria, L. et al. *Interpretable semantic textual similarity: Finding and explaining differences between sentences*, Knowledge-Based Systems, **119**(2017), 186–199.

[24] Lorenzen, B., Murray, L., *Bilingual aphasia: A theoretical and clinical review*, American Journal of Speech-language Pathology, (2008).

[25] Mall, R., Cerulo, L., Bensmail, H., Iavarone, A., Ceccarelli, M., *Detection of statistically significant network changes in complex biological networks*, BMC Systems Biology, **11**(2017), 32.

[26] Manning, C.D., Schütze, H., Foundations of Statistical Natural Language Processing, MIT press, 1999.

[27] Meladianos, P., Nikolentzos, G., Rousseau, F., Stavrakas, Y., Vazirgiannis, M., *Degeneracy-based real-time sub-event detection in twitter stream*, In Ninth International AAAI Conference on Web and Social Media, (2015), 248–257.

[28] Meladianos, P., Xypolopoulos, C., Nikolentzos, G., Vazirgiannis, M., *An optimization approach for sub-event detection and summarization in twitter*, In European Conference on Information Retrieval, (Cham, 2018), Springer, 481–493.
[29] Mikolov, T., Chen, K., Corrado, G., Dean, J., *Efficient estimation of word representations in vector space*, arXiv preprint arXiv:1301.3781, (2013).
[30] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., *Distributed representations of words and phrases and their compositionality*, In Advances in Neural Information Processing Systems, (2013), 3111–3119.
[31] Mikolov, T., tau Yih, W., Zweig, G., *Linguistic regularities in continuous space word representations*, In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (2013), 746–751.
[32] Mnih, A., Hinton, G.E., *Three new graphical models for statistical language modelling*, In Proceedings of the 24th International Conference on Machine Learning, (2007), 641–648.
[33] Nabhan, A.R., Shaalan, K., *A graph-based approach to text genre analysis*, Computación y Sistemas, **20**(2016), 527–539.
[34] Nikolentzos, G., Meladianos, P., Rousseau, F., Stavrakas, Y., Vazirgiannis, M., *Shortest path graph kernels for document similarity*, In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, (2017), 1890–1900.
[35] Niwattanakul, S., Singthongchai, J., Naenudorn, E., Wanapu, S., *Using of jaccard coefficient for keywords similarity*, In Proceedings of the International Multiconference of Engineers and Computer Scientists, **1**(2013), 380–384.
[36] Ozdikis, O., Senkul, P., Oguztuzun, H., *Semantic expansion of tweet contents for enhanced event detection in twitter*, In 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, (2012), IEEE, 20–24.
[37] Pennington, J., Socher, R., Manning, C.D., *Glove: Global vectors for word representation*, In Proceedings of the 2014 Conference on Empirical Methods in Batural Language Processing, (EMNLP), (2014), 1532–1543.
[38] Raymond, J.W., Willett, P., *Maximum common subgraph isomorphism algorithms for the matching of chemical structures*, Journal of Computer-aided Molecular Design, **16**(2002), 521–533.
[39] Resnik, P., *Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language*, Journal of Artificial Intelligence Research, **11**(1999), 95–130.
[40] Rong, H., Ma, T., Tang, M., Cao, J., *A novel subgraph k+ -isomorphism method in social network based on graph similarity detection*, Soft Computing, **22**(2018), 2583–2601.
[41] Rooth, M., Riezler, S., Prescher, D., Carroll, G., Beil, F., *Inducing a semantically annotated lexicon via em-based clustering*, In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, (1999), Association for Computational Linguistics, 104–111.
[42] Rosen, K.H., Discrete Mathematics and Its Applications, McGraw-Hill, 2007.
[43] Rosenthal, G., Váša, F., Griffa, A., Hagmann, P., Amico, E. et al., *Mapping higher-order relations between brain structure and function with embedded vector representations of connectomes*, Nature Communications, **9**(2018), 2178.
[44] Rousseau, F., Vazirgiannis, M., *Main core retention on graph-of-words for single-document keyword extraction*, In European Conference on Information Retrieval (Cham, 2015), Springer, 382–393.
[45] Sahlgren, M., *Vector-based semantic analysis: Representing word meanings based on random labels*, In ESSLI Workshop on Semantic Knowledge Acquisition and Categorization, (2001).
[46] Shibuya, Y., Jensen, K.E., *Mining for constructions in texts using n-gram and network analysis*, Globe: A Journal of Language, Culture and Communication, (2015).
[47] Skianis, K., Malliaros, F., Vazirgiannis, M., *Fusing document, collection and label graph-based representations with word embeddings for text classification*, In NAACL-HLT Workshop on Graph-Based Natural Language Processing (TextGraphs) (New Orleans, Louisiana, United States, 2018), 382–393.
[48] Vazirgiannis, M., Malliaros, F.D., Nikolentzos, G., *Graphrep: Boosting text mining, nlp and information retrieval with graphs*, In Proceedings of the 27th ACM International Conference on Information and Knowledge Management, (2018), 2295–2296.
[49] Wang, Y.-Y., Mahajan, M., Huang, X., *A unified context-free grammar and n-gram model for spoken language processing*, In 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings Cat. No. 00CH37100, **3**(New Orleans, Louisiana, United States, 2000), IEEE, 1639–1642.
[50] Watts, D.J., Small Worlds: The Dynamics of Networks Between Order and Randomness, vol. 9. Princeton University Press, 2004.