Computing Semantic Similarity of Documents Based on Semantic Tensors

Navid Bahrami
Department of Electrical, Computer and IT Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran
bahrami.navid@gmail.com
Amir Hossein Jadidinejad*
Department of Electrical, Computer and IT Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran
amir.jadidi@qiau.ac.ir
Mozhdeh Nazari
Department of Engineering, Guilan Science and Research Branch, Islamic Azad University, Rasht, Iran
mozhdeh_nazary@yahoo.com

Received: 14/Dec/2014 Revised: 17/Mar/2015 Accepted: 06/Apr/2015

Abstract

Exploiting semantic content of texts due to its wide range of applications such as finding related documents to a query, document classification and computing semantic similarity of documents has always been an important and challenging issue in Natural Language Processing. In this paper, using Wikipedia corpus and organizing it by three-dimensional tensor structure, a novel corpus-based approach for computing semantic similarity of texts is proposed. For this purpose, first the semantic vector of available words in documents are obtained from the vector space derived from available words in Wikipedia articles, then the semantic vector of documents is formed according to their words vector. Consequently, semantic similarity of a pair of documents is computed by comparing their corresponding semantic vectors. Moreover, due to existence of high dimensional vectors, the vector space of Wikipedia corpus will cause curse of dimensionality. On the other hand, vectors in high-dimension space are Usually very similar to each other. In this way, it would be meaningless and vain to identify the most appropriate semantic vector for the words. Therefore, the proposed approach tries to improve the effect of the curse of dimensionality by reducing the vector space dimensions through random indexing. Moreover, the random indexing makes significant improvement in memory consumption of the proposed approach by reducing the vector space dimensions. Additionally, the capability of addressing synonymous and polysemous words will be feasible in the proposed approach by means of the structured co-occurrence through random indexing.

Keywords: Information Retrieval; Natural Language Processing; Random Indexing; Semantic Similarity; Semantic Tensor.

1. Introduction

How similar are “Cat flu” and “Feline influenza”? Humans have initiate ability to compute semantic similarity due to their background knowledge about words and their interpretation ability. However, computing semantic similarity of words with multiple meanings is still remained as an obstacle. It must be noted that the meaning of each word is expressed according to the context that it appears and humans can interpret the meaning of a word according to its context. The main challenge refers to machines and how they deal with natural language and interpret concepts. In order to behave same as human, machines require human knowledge. Majority of natural language processing approaches leverage encyclopedias to transform knowledge and train machines. Moreover, there are many drawbacks in using encyclopedias. One of the obstacles is deep recognition of destination language for considering its syntax structure in processing. Another issue refers to extracting the meaning of words from encyclopedia. This problem is addressed here by considering the meaning of a word according to a given context.

The proposed approach is capable of extracting concepts from encyclopedia directly without any manual control. Whereas, encyclopedias contain wide range of documents, the meaning of each word can be expressed in high dimensional vector space using texts of documents. The most important achievement of the proposed approach is considering synonymy and polysemy. Indeed, it is able to disambiguate ambiguous and polysomic words.

The main characteristic of the proposed approach refers to employing simple texts of encyclopedias. In addition, it can limit the deep understanding of destination language to particular language structures such as punctuations, separators and etc. The main object of this approach is to compute semantic similarity of documents by extracting concepts from hierarchical structure of Wikipedia [1] and creating a semantic vector for each document and finally compare them. Due to Wikipedia's structure, the meaning of words can be expressed in different categories. Therefore, a three-dimensional vector space is created as a vector of words in various topics, which is organized by three-dimensional tensor structure. As an example, consider the meaning of the word “apple”. The fruit “apple” will be the first concept that is inspired

* Corresponding Author

in readers' mind. Nevertheless, if such a word is used with words such as “Ipad”, “computer” and “corporation”, then the meaning of fruit “apple” will not be visualized and imagined. A method like bag of words does not pay attention to the relations among words in the texts and considers the texts merely as a set of words without order and relation. Thus, in this approach the word “apple” has only one meaning and that is deducted from the repetition of this word in the text. However, as it was previously mentioned, the main idea of the proposed approach of this paper is based on considering the meaning of the words in different texts. The possibility of extracting the best concept from a large corpus (the Wikipedia corpus in this paper) and forming the semantic vector of the word “apple” can be done by determining the meaning of “apple” with the help of its neighboring words in the text. Thus, the meaning convergence to the word “apple” will be provided by Wikipedia corpus, if the word “apple” has the same meaning in the two different documents but its neighboring words are different. In other word, the semantic vector of the document will be obtained, if all the semantic vectors of available words in the document are gained. Therefore, the possibility to compare the documents will be created due to their semantic vectors.

On the other hand, finding the words' meaning in a high dimensional space is the neglected issue in this approach, which is doomed to failure due to curse of dimensionality [2]. Whereas, spaces of vectors in high dimensional space of Wikipedia corpus are very similar to each other, achieving the best semantic vector for each word would be a meaningless and vain task [3]. Although there is no final solution for this problem, reducing the dimensions of vector space is the most appropriate and acceptable method. The random indexing [4,5] is used in the proposed approach to extract the semantic vector of words from Wikipedia corpus and reduce the dimensions of vector space. In other word, this method is capable of computing the meaning of words and reducing the dimensions of vector space simultaneously, which can reduce the processing load and memory consumption.

This approach has widespread applications in natural language processing. Finding the most relevant documents to a query, classifying documents based on their semantic content and computing semantic similarity of documents in order to compare them are the most notable applications. Moreover, according to reduced dimensions of vector space, the proposed approach can efficiently be used in large-scale systems.

The reminder of the paper is organized as follows: In section 2 the state of the art in computing semantic similarity are described. Way of constructing semantic space using word co-occurrences is presented in section 3. Key notions of the proposed approach such as extracting the meaning of words from Wikipedia and way of using them are indicated in section 4. Empirical experiments of the proposed approach for determining the effectiveness, Analyzing memory consumption and processing time are presented in section 5.

2. Related Work

Computing semantic similarity is one of the well-known agents in many fundamental tasks of computational linguistics such as word sense disambiguation, information retrieval and error correction [6]. Previous studies in this field can be classified into three main categories:

According to first category, texts are compared based on their common words using binary [7] and bag of words methods [8]. These methods are simple but whereas texts may contain many common words and express a concept with synonym words, they do not indicate any remarkable results.

On the other hand, knowledge-based methods leverage semantic relations of concepts defined in lexical resources such as WordNet [9] or Roget thesaurus [10] or network of concepts of Wikipedia [11] for computing semantic similarity and other applications. Then, the characteristics of graph structure of a lexicon are used for computing semantic similarity [6], such as method proposed by Resnik [12], Jiang and Conrath [13] and Lin [14]. These methods are confronted with some drawbacks. Noteworthy, they can only cover a limited range of vocabularies of a language and they do not include information of a particular filed either. Furthermore, knowledge-based methods are inherently limited to words and complex metrics are required for comparing texts. In contrast, these approaches are able to consider the contexts of the words. However, due to limitations of words in knowledge sources, considering the context of words and word sense disambiguation are also limited.

Other existing approaches employ statistical occurrences of words in a large corpus of unlabeled data. Latent Semantic Analysis (LSA) [15] is one of these approaches trained by word-document co-occurrence matrix. Vector space dimensions of this matrix are reduced using Singular Value Decomposition (SVD). This approach attempts to identify the most effective data known as implicit data in co-occurrence matrix in order to reduce dimensions. Therefore, interpreting its concepts is difficult and most of them are not commonly used by humans.

Another existing method which employs large corpus for computing semantic relatedness is Explicit Semantic Analysis (ESA) [16]. This approach leverages Wikipedia corpus as training set. Consequently, it uses Wikipedia concepts for considering concepts of words and documents. These concepts are directly defined by humans and are also consistent with natural concepts. Another advantage of determining concepts and their relations in Wikipedia corpus is presenting the related keywords to each concept, which has particular application in online advertisements of search engines [17].

The prominent idea of Explicit Semantic Analysis and Latent Semantic Analysis approaches is based on semantic kernel concept. Moreover, the aim of kernel methods is mapping data objects (o) of $D_{n\times m}$ matrix from a semantic space $(\emptyset(o))$ to a more comparable vector

space ($\tilde{\emptyset}(o)$) [18]. It means that data are transformed to a new computational semantic space and accordingly the result of calculation will have more accuracy and less complexity [19].

The rows of $D_{n\times m}$ present a set of data objects $o = \{o_1, o_2, \dots, o_n\}$ and its columns present semantic features $f = \{f_1, f_2, \dots, f_n\}$. Based on classic models (bag of words), data objects of matrix are corresponded to documents and the words are its features. Noteworthy, this approach does not take into consideration semantic relations among data objects. Consequently, to fill this lacuna, data objects must be mapped into a similar vector space which considers semantic relations among data objects. Therefore, for considering semantic content in vector space, transformation is employed as $\tilde{\emptyset}(o) = \emptyset(o)\, K_m$, where K_m is a semantic matrix. Different choices of the matrix K_m lead to different variants of vector space semantic kernels [18]. Such as creating K_m matrix explicitly (Wikipedia Semantic Kernel [20]) or implicitly (Latent Semantic Kernel [21]).

Despite the precision of these methods in computing semantic relatedness and similarity, they are confronted with some disadvantages. One of the fundamental obstacles of both ESA and LSA methods refers to their high processing time in encountering a new training document. Whereas the weighting method used in these approaches requires the computation of the probability of each word in all documents of a corpus, by adding a new document all weighting process must be recalculated practically.

Temporal Latent Semantic Analysis [22] is an extended form of LSA which contains time elements. It is capable of organizing weight of words in different time intervals using tensor structure. Moreover, this method has decreased the computational complexity of adding new training document to a time interval. Following the similar line of research, Temporal Semantic Analysis [23] is an extended form of ESA at words level where time is added as new component. Based on this method, meanings of words are expressed according to Wikipedia concepts at different states. This method has efficiently reduced processing time of ESA in encountering a new training document.

3. Semantic Space

Semantic space models are based on distributional hypothesis [24,25[24],[25]. This hypothesis indicates that semantic similarity of a pair of words is computing the similarity of co-occurrence distribution among them. Therefore, distributional hypothesis and vector space model are related to each other because the distributional hypothesis emphasizes on co-occurrences of words corresponded to word frequency in vector space model. Consequently, creating semantics for word co-occurrences depends on how an algorithm presents semantic alternations [5].

3.1 Semantics Using Word Co-occurrence

Due to semantic space, the meanings of words are mapped to a multidimensional space. Space dimensions represent the differences between the meanings of words. Therefore, semantically similar words specify close vector representation. To understand the meaning of semantic co-occurrence obviously, consider the word "apple". Suppose "Calories in red delicious apple" is a document describing the word apple that inspires the meaning of apple as a fruit in readers' mind. This meaning is obtained by co-occurrences of apple by words such as red, color and delicious. On the other hand, consider apple in "the Ipad is an apple product" document. In this document apple is used to represent the company that manufactures computer products. Consequently, the meaning of word "apple" is changed according to its co-occurrence with words such as Ipad and product. Simple two-dimensional representation of word "apple" in co-occurrence with other words is illustrated in Fig.1.

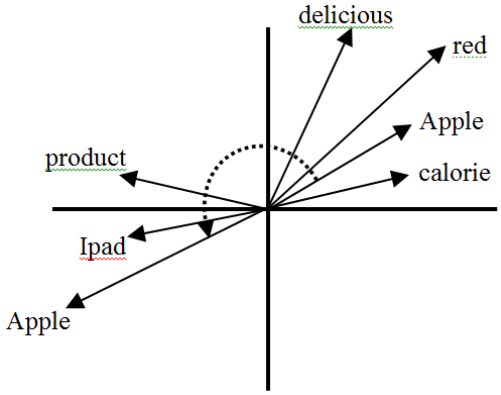

Fig. 1. Representation of word "apple"

3.2 Random Indexing

Simple co-occurrence can be efficiently used in a large corpus. Based on this model, each word attempts to assign its dimensions to words which are co-occurred. The results can be hundreds, thousands or perhaps millions of dimensions. Fundamentally, considering the number of dimensions according to the number of unique words (considering one repetition of each word) can be a complex issue in a large corpus and considerable efforts have been made to reduce space dimensions. Random indexing is one of these approaches which employs random projection of co-occurrence matrix to a space with less dimensions. Based on this technique, an index vector is assigned to a unique word. This vector is a random vector of numbers of 1, 0 and -1 in a space with constant dimensions (e.g. 500). The size of random index vector indicates the number of dimensions in semantic space.

Index vectors are constructed in such a way that any two arbitrary index vectors are orthogonal to each other with high probability. This feature is essential for accurate approximation from a word co-occurrence matrix to a low dimensional matrix. The meaning of each word is computed by summation of random index vector of co-occurred words in a small window of text. Random

indexing method can work efficiently in reducing the dimensions of a corpus that has already been processed [4]. Formally, consider word w, then w_i represents the co-occurrences of current word to a word with distance of i and $index\ (w_i)$ is the index vector of co-occurred word. For a word w, a window size n is defined which is considered as the number of co-occurred words. Therefore, the meaning of word w is presented as:

$$Semantics(w) = \sum_{o \in D} \sum_{-n \le i \le n} index(w_i) \quad (1)$$

4. Where o is All Occurrence of Word w in the Corpus D $(\forall o \in D)$.The Proposed Approach

The prominent goal of the proposed approach is using Wikipedia corpus in order to map documents into a high dimensional vector space. Constructed vector space comprises semantic discrimination and consequently the meanings of documents and words can be expressed based on topics.

The proposed approach is divided into two main phases. Training phase contains extracting documents from Wikipedia corpus based on specific categories and creating discriminative semantic space for each word of documents in each category. The created semantic space is then used in second phase. This stage is called test phase where semantic vectors of existing words in input documents are specified based on specific category and semantic vectors of input documents are computed based on them. The architecture of the proposed approach is presented in Fig. 2.

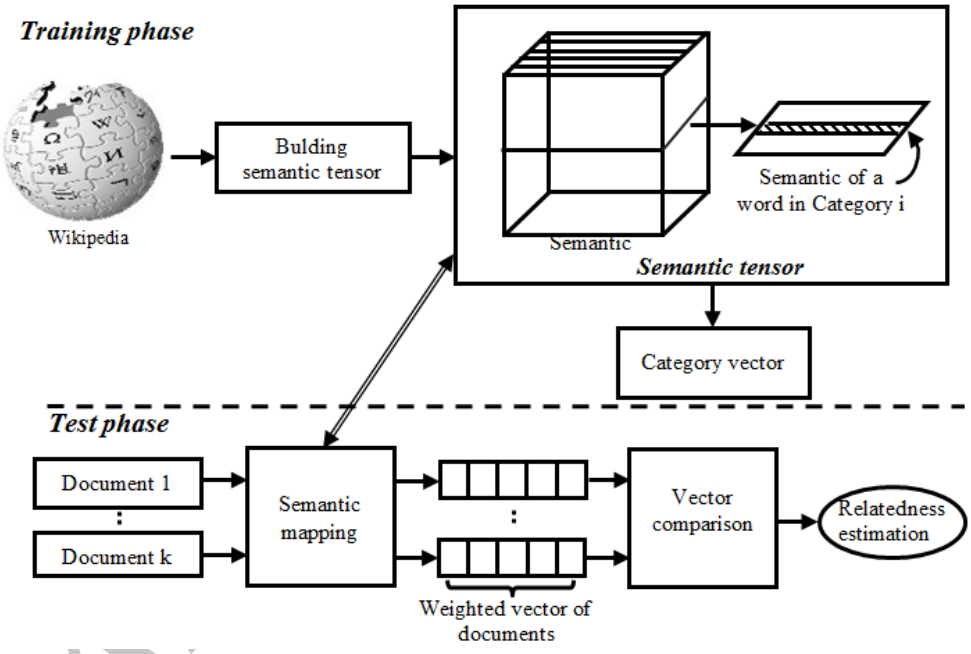

Fig. 2. The proposed approach

4.1 Category of Semantic Space

Adding categories to a semantic space causes word semantic discrimination according to different thematic areas. Therefore, the meaning of each word is recognized with respect to the topic of a document. The main part of training phases is focused on creating this separated space. According to test phase on Fig. 2, in the beginning documents are extracted from Wikipedia corpus based on categories. Categories must be selected in such a way that the meaning of each category is reasonably discriminative towards another. Considering directed acyclic graph of Wikipedia, it is possible to obtain categories with high degree of semantic discrimination and high level of access.

The next step in training phase is organizing the meanings of words according to the documents existing in each category, which is done by a particular structure called semantic tensor. Semantic tensor adds categories to vector space. Therefore, instead of using two-dimensional matrix of word×semantic, a three-dimensional matrix of word×semantic×category can be used. Two-dimensional vector of word w and three-dimensional representation of category vector are illustrated in Fig. 3.

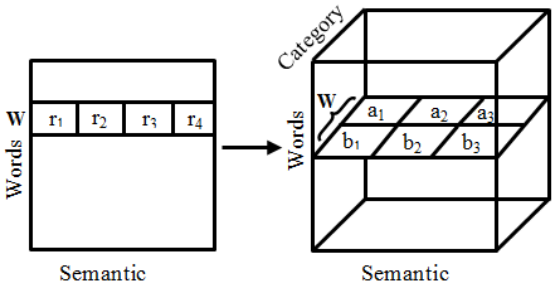

Fig. 3. Two and three dimensional vector represenation of word w

Semantic tensor based on this model has three main advantages:

1. It is possible to add new documents to each category.
2. Semantic tensor representation enables semantic meaning of a word to be compared in different categories.
3. Random indexing method used for weighting words in a document and reducing dimensions of vector space can perform processing on each document separately. Consequently, time and memory can be saved efficiently.

In order to add categories to semantic tensor, addition operation for weighted vector of each word cannot be added to Eq. (1) immediately. In other word, addition must be done at different categories separately. The summation of categories defines a semantic part for each word. The semantic part of word "apple" is presented in Fig. 4.

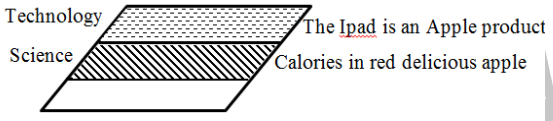

Fig. 4. Semantic slice of word "apple"

Therefore, the meaning of each word in a category is equivalent to the results of random indexing where all its values in that category are summed together. Finally, a single vector is obtained for each word in each category.

Fig. 4 presents how a word such as apple can have two different meanings. First semantic vector is constructed from summation of word apple vectors in category of technology that introduces apple as a manufacturer of computer components and accessories and the second semantic vector in categoy of science expresses the claories in apple.

Therefore, the semantic tensors can be defined more formally. The input is a set of documents as follows:

$$D=(CT_0,d_0),(CT_1,d_1),(CT_2,d_2),\dots,(CT_i,d_i) \quad (2)$$

Where d_i is the set of documents occurring at category CT_i. If W_D is considered as a set of unique words in the collection $(w \in W_D)$, a unique index $(index(w))$ is assigned to each word in the set. Consequently, the meaning of each word in each category is defined as follows:

$$Semantics(w,CT)=\sum_{o_{CT}\in d_i}\sum_{-n\le i\le n} index(w_i) \quad (3)$$

Where o_{CT} is the context for an occurrence of word w at category CT. Moreover, $index(w_i)$ is an index vector of co-occurred words with distance of i to the main word. The semantic slice can be defined as follows:

$$Slice(w)=\{(CT_i, Semantics\ (w,CT_i)) \mid w \ni d_i, i=1,k\} \quad (4)$$

The final step of training phase is creating weighted vectors for categories of semantic tensor. This operation is done by summing the existing words vector in each category of semantic tensor. Therefore, category vector CT is equal to:

$$Semantics(CT)=\sum_{1\le i\le k} index(w_i) \quad (5)$$

Where i is index identifier of all words in the category CT.

According to this, input documents with semantic tensor category are allowed to be compared in test phase based on their weights.

4.2 Computing Semantic Similarity

Mapping word semantic vector of semantic tensor to the corresponding word of input document and creating a semantic vector for each document is the basis of test phase. Consequently, vectors can be compared for computing semantic similarity.

As it is indicated in Fig. 2, the initial step in testing phase is receiving the input documents, which will be compared for computing semantic similarity. These documents contain simple and explicit text, which can be easily interpreted by humans. Nevertheless, the most important step in test phase is semantic mapping. According to Fig. 5, the semantic mapping for each input document performs the three following functions:

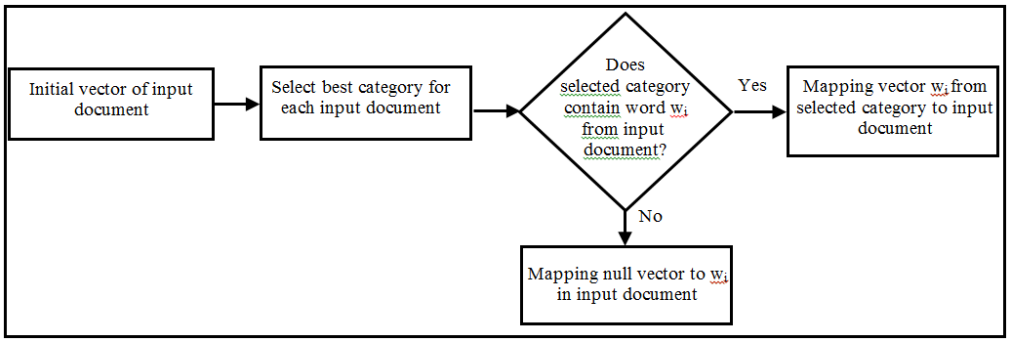

Fig. 5. Flowchart of semantic mapping

1. Creating initial index vector for input documents. Therefore, the preliminary weighting of input documents is accomplished using random indexing method by considering existing words of each document. As a result, the primary index vector for each document is computed by summation of words' weights. This method is simple but logical because words are components of sentences and documents. Index vector created for each document is leveraged in order to be compared to index vector of categories.
2. Detecting target category from semantic tensor for each input document is done by cosine comparison of weighted vector of documents and weighted vector of semantic tensor categories. Consequently, the most appropriate category is determined for each document.

$$Cosine\ (index(d_{input})\ , index(CT_i)) \quad (6)$$

This action provides a background for weighting of input documents using semantic tensors.

3. Mapping weights of words existing in target category of semantic tensor to their corresponding words in input document. Therefore, the weight of word *w* for input document of category *k* is defined as follows:

$$Semantics(d_{input}(w)) = Semantics(w, CT_k) \quad (7)$$

It should be noted that if a word of input k document does not exist in selected category, it will be considered null and it actually will have no effect on weighting of input document.

According to Fig. 2, after semantic mapping step, words of each input document are weighted based on target category of semantic tensor and only a single vector is remained for each document. Consequently, weighted vector of each document can be obtained by summing the weight of its words. Indeed, it can be mentioned that the main purpose of test phase is using semantic tensor in order to determine the semantic meaning of input documents.

The complementary step of testing phase is comparing documents' semantic vector to compute the degree of their similarity done by comparing documents in vector space using cosine similarity measure. Accordingly, the proposed measure is a corpus based approach which is capable of comparing documents using input documents and creating quantity values in a high dimensional vector space.

5. Empirical Experiments

In order to evaluate the proposed method, various experiments are carried out to reveal the efficiency of the proposed measure in comparison to other existing semantic similarity measures. These experiments contain two fundamental approaches. Experiments are performed to determine the potentiality of the proposed approach in computing semantic similarity of documents. The evaluations of these experiments are done by computing the Pearson correlation coefficient between empirical results and human judgments on Lee benchmark dataset.

Other experiments include the analysis of memory consumption of the proposed method in comparison to other commonly used semantic similarity measures. The effect of increasing the number of training documents and unique words on memory consumption of the proposed method and other existing measures are also highlighted in these experiments.

The required processing time for executing the proposed method is presented in the following of this section using two various experiments.

5.1 Corpus

The proposed method is capable of using a corpus where hierarchical structure of categories and related documents to each category are specified. The reason of this issue refers to the possibility of extracting required documents based on semantic tensor's categories.

The proposed method is implemented using 2011 Wikipedia version containing 3573789 articles which are organized in 739980 categories. The English version of Wikipedia has been employed in our experiments but other languages can be also used.

Before using Wikipedia corpus for constructing semantic tensors, preprocessing is accomplished on documents of different categories as follows:

1. Removing bookmarks
2. Removing stop words
3. Stemming using Porter stemmer

These processes efficiently eliminate documents' disorders. Therefore, high frequently words which do not express a particular meaning are removed and in order to have uniform text, other words are transformed to their basic forms.

5.2 Benchmark Dataset

For comparing the precision of the proposed method in computing semantic similarity of texts, lee dataset [26] has been used, which contains a collection of 50 documents from Australian Broadcasting Corporation's new mail service. The length of these documents is between 51 to 128 words and they include large number of topics. Judgment had been done by 83 students of Adelaide University of Australia. These documents were paired in all possible ways and each of the 1225 pairs has 8-12 human judgment. Finally, the average of obtained values was considered as degree of semantic similarity of each pair of documents.

5.3 Empirical Result

Two following tools were employed for empirical experiments of this paper:

1. Wikipedia miner [27] based on Java for extracting documents from Wikipedia corpus
2. S-Space [28] library based on Java for leveraging implementations of random Indexing, LSA and

ESA existing in this library and applying the required process of the proposed approach.

Experiments are divided into two main categories. The main reason for choosing categories in each set of documents is the ability of document semantic discrimination of each category to the others. The first set of experiments are accomplished using extracted documents of four categories of matter, life, concept and society. These four categories exist in depth one of Wikipedia's hierarchical structure after fundamental category and all of documents are organized in subcategories of these four main categories. The second set of experiments is performed on seven categories of arts, biography, geography, history, mathematics, science and society. These categories are chosen from Wikipedia's documents classification in English Wikipedia website (at the time of experiments). The results of experiments present acceptable semantic discrimination among these seven decided categories. It is due to that if semantic discrimination was not correctly obtained, result of experiments would confront with significance decreasement.

In both sets of experiments the same number of documents are extracted from each category (500 documents) and the length of random index vector of semantic tensor has been considered 120 with a window size of ± 3. Then unequal number of documents has been extracted from each category (more than 3000 documents for each category) and the length of random index vector in semantic tensor has been set 150 with word window size of ± 3. Empirical experiments conducted by Karlgren and Sahlgren [29] indicated that short length word window often provides better functionality. This issue also seems reasonable because sentences with length less than eight words provide high readability and are able to explain the meaning of a word clearly.

The results of experiments are presented in table 1. For the aim of comparison, the results of experiments conducted by the Bag of Word approach, Basic random indexing, LSA and ESA on Lee dataset are also shown in this table.

Table 1. Pearson correlation coefficient between various semantic similarity measures and human judgments on Lee dataset

Algorithm	Pearson correlation with human judgments
Bag of words	0.50
Random indexing	0.52
Latent Semantic Analysis (LSA)	0.60
Explicit Semantic Analysis (ESA)	0.72
Our approach (4 category, 500 documents)	0.64
Our approach (4 category, different number of documents)	0.61
Our approach (7 category, 500 documents)	0.60
Our approach (7 category, different number of documents)	0.62

Table 1. Pearson correlation coefficient between various semantic similarity measures and human judgments on Lee dataset

Comparison results indicate - the effectiveness of the proposed method in comparison to other semantic similarity measures. According to the results obtained by the proposed approach, it can perform better than Bag of Words, random indexing and LSA methods and it only presents lower performance than ESA method.

Although four conducted experiments are not very different from each other, the amount of difference can express some points. Considering the results obtained by four categories, by increasing the number of documents the results are decreased. This is due to negative impact of added documents to each category. These documents had not only negative impact on weighting of documents' words, but also made some difficulties for selecting a target category for each document. This issue is probably due to discrimination reduction of the general meaning of each category by adding new documents to them. As it was noted, reducing efficiency by increasing the amount of documents in these four categories is negligible. However, it presents that increasing the number of documents of each category requires a lot of precision.

On the other hand, in experiments containing seven categories, by increasing the number of training documents, the results have been improved It seems that it refers to inadequate number of training documents towards covered topics in text documents. It cannot be definitely expressed that how many categories would present better results. However it must be noted that training documents in each category must be selected in such a manner to cover a wide range of topics in that field. In addition, the semantic discrimination of each category must be observed towards another one. On the other hand, it must also be mentioned that the large number of categories is one of the main factor of increasing error detection of target category because semantic discrimination process and appropriate document selection for each category is difficult and error probability in selecting related documents to each topic would be increased.

5.4 Memory Analysis

In this section the memory consumption of the proposed method is compared to ESA and LSA methods by considering a set of specific documents and their unique words. The proposed method contains seven categories in this experiment and the length of random index vector is 100. These two factors along with the number of documents and their unique words are the main factors affecting the memory size of the proposed method. Two sets of experiments have been carried out in order to examine the effectiveness of increasing the number of documents and words on memory consumption increment. These two sets of experiments have some specific features as follows:

1. The first set contains 26989 documents and 159947 unique words.
2. The second set contains 228312 documents and 501436 unique words.

3. Number of extracted documents from search in hierarchical structure of Wikipedia to specific depth considering determined topic categories. In order to increase the number of documents in second experiments, depth of search is also increased.
4. The rate of document increment from first series to second series is equal to 45.8 times and the rate of word increment is equal to 3.13 from first series to second series.

Memory consumption of the proposed method, ESA and LSA are illustrated in Fig. 6 according to mentioned features.

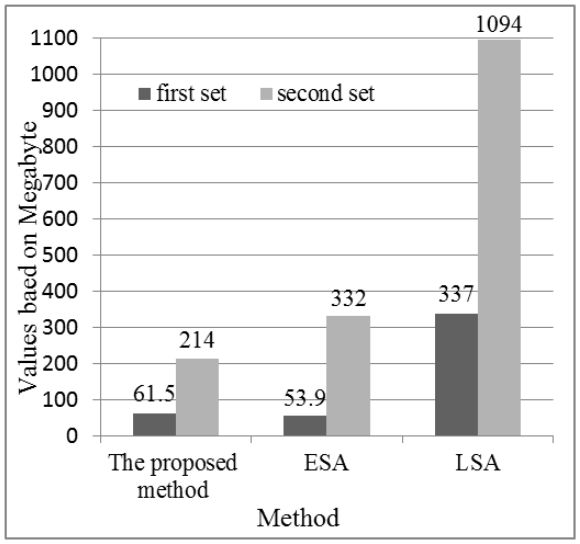

Fig. 6. Histogram of memory consumption of the proposed method, LSA and ESA

By analyzing the results, the rate of memory consumption increment of the proposed method from first series to second series of experiment is equal to 3.47. Moreover, the rate of memory consumption increment of ESA is equal to 6.15 and this rate is equal to 3.24 for LSA. The growth rate of the proposed method largely depends on the growth rate of words, whereas based on the proposed method a wide range of words can be initialized using a vector with the length of 100 (with respect to unique and orthogonal vectors) and the number of vectors is only increased by escalating the number of words. However, this rate highly depends on the number of considered categories.

On the other hand, ESA has a particular structure with various numbers of posting which are dependent on the number of documents and unique words. By increasing the number of documents and unique words, the probability of word occurrence in documents is increased and the length of postings in inverted index is also increased frequently. The rate of memory consumption of this method confirms this issue. Consequently, LSA depends on the number of words; whereas by decreasing dimensions, vectors with the same length are created for each word (vector with length of 100 in this particular example). The memory consumption of this method refers memory consumption of each cell in each word vector according to the required decimal precision. Accordingly, it can be stated that the proposed method is more optimal in comparison to LSA and ESA in memory consumption.

5.5 Time Analysis

The required time for creating the final vector for each input document based on the proposed method is presented in this section. Effective steps for estimating the required time are illustrated in Figure 7.

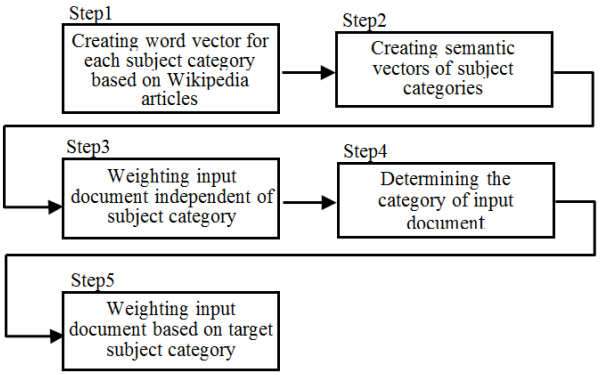

Fig. 7. Effective steps in time analysis of the proposed method

As an example, consider two datasets of section 5-4. Time analysis of the first set containing 26989 documents and the second set containing 228312 documents based on time steps of the proposed method (Figure7) are presented in Table 2. It must be noted that the characteristics of computer hardware that experiments were conducted on it is CPU Core2Duo E4600 and 3GB of RAM.

Table 2. The required time for each step of the proposed method

	First set (Second)	Second set (Second)
Step1	25.11	154.561
Step2	9.183	36.651
Step3	0.097	0.097
Step4	0.19	0.146
Step5	1.238	3.709
Total time	35.818	195.164

Time analysis with mentioned hardware for ESA approach on the first set is equal to 110.447 seconds and on the second set is equal to 1401.952 seconds. Moreover, time analysis for LSA approach on the first set is equal to 126.295 seconds and on the second set is equal to 1329.941 seconds. According to empirical experiments, ESA approach on first set requires 3.08 more processing time in comparison to the proposed method and this value is equal to 7.183 on second set. Furthermore, the processing time of LSA approach is 3.52 more in comparison to the proposed method on first set and this value is equal to 6.81 on the second set. The results represent the significant improvements in time consumption of the proposed method in comparison to LSA and ESA approaches.

Implementing the ESA and LSA approaches on mentioned hardware was done using S-Space package. English Porter Stemmer was employed for stemming step of both approaches. Additionally, vector space dimensions of LSA approach are considered 300. This

approach reduces the dimensions of word-document matrix using SVD method. Wikipedia version 2011 has been employed as primary dataset of these approaches.

Long processing time of LSA and ESA approaches in comparison to the proposed method is due to their complexity in constructing training vector using Wikipedia corpus. ESA approach creates m×n table of words and concepts. Each elements of this table presents the weight tf.idf of a word in a particular concept. Noteworthy, computing tf.idf for each word is a time consuming task. Moreover, in order to reduce the dimensions of vector space, ESA approach requires inverse vector algorithm, which subsequently increases the processing time. LSA approach is confronted with the same drawbacks. The created word-document matrix is weighted using entropy measure and the dimensions are reduced using SVD algorithm which requires long processing time. The reason of this issue is due to the complexity of this algorithm, whereas the time complexity of the fastest implemented SVD algorithm is equal to $O(m.n^2)$ [5].

On other hand, whereas the proposed method employs random indexing method it has not only significant reduction in memory consumption, but also considerable reduction in processing time. Random indexing method considers random vectors with constant length for words existing in documents. These random vectors contain normal numbers and only require meeting unique vectors condition. Moreover, vectors with constant length cause reduction in vector space dimensions. Therefore, weighting and reducing dimensions of vector space are done simultaneously and simply. As result, the proposed method prospers significant time reduction in creating semantic vectors of words.

6. Discussion

The most important achievement of this paper refers to determining the effectiveness of the proposed method and improving its results in comparison to basic random indexing methods. This is due to the nature of random indexing method where indexing employs neighbor words to express the meaning of a key word. It requires large sets of documents to identify key concepts and balance their weights. This issue is clearly marked by comparing the results of experiments because by adding a large corpus of documents to basic random indexing method in the proposed method, the results have been significantly improved.

One of the major advantages of the proposed method in comparison to ESA and LSA is that it does no impose significant processing load during adding or changing the training set. In the view of fact it leverages random indexing method. Moreover, by employing this method, words in new documents are weighted independently and these new weighted values are added to previous weights. Since the used weighting method in LSA and ESA (usually TF.IDF) requires to consider the weight of a word in all training set documents and by adding even a new training document the weight of all words must be recalculated and therefore high processing load is imposed based on them.

Although the proposed method employs three-dimensional tensor, it uses less memory than approaches with two-dimensional structure. This indicates the high potentiality of random indexing method in reducing the dimensions of vector space towards other existing methods. Moreover, it can significantly reduce the memory consumption of the proposed method.

Besides the advantages mentioned for the proposed method, it confronts with some limitations. Initially, it requires a rich corpus which contains many words. This limitation is due to random indexing method. In other word, according to random indexing method if a word does not exist in the corpus, no weight will be considered for it. Therefore, word would not have any effect in computing semantic similarity. The second limitation refers to determining the number of documents in each category. If the documents of each category are semantically close to each other and their semantic discrimination is low, error probability in choosing the best category will increase. Furthermore, having large number of categories with various topics causes complexity in identifying the appropriate category for extracting the meaning of words. Accordingly, if a target category is not selected properly, the overall performance of the proposed method will decrease.

7. Conclusions

In this paper a novel method based on semantic tensor is proposed for computing semantic similarity of texts. This is a semantic technology for natural language processing. The proposed method is based on Wikipedia corpus where articles are categorized in different topics and documents are extracted from these categories. The most important aspect of the proposed method is its ability for identifying synonymy and polysemy, which are one of the most important issues in natural language processing. Therefore, this method does not merely rely on common word frequency in texts and it can identify the value of association between two texts that express a topic with various texts.

The evaluation results revealed acceptable performance of the proposed method in computing sematic similarity according to optimal memory consumption. Consequently, the Pearson Correlation coefficient of the proposed method and human judgments is between 0.54 and .064. Although ESA has better performance than the proposed method, the other existing methods show lower performance.

According to experiments, memory consumption of the proposed method is 80% less than the memory required by LSA and 30% less than the amount of memory required by ESA. By increasing the amount of documents of a corpus, this value would improve. Consequently, the efficiency is simultaneously improved by decreasing memory consumption.

References

[1] Medelyan O., D. Milne, C. Legg, and I.H. Witten, "Mining meaning from Wikipedia", International Journal of Human-Computer Studies, Vol. 67, No. 9, 2009, pp. 716-754.
[2] Kriegel H.P., P. Kröger, and A. Zimek, "Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering", ACM Transactions on Knowledge Discovery from Data, Vol. 3, No.1, 2009, pp. 1-58.
[3] Assent I., "Clustering high dimensional data", WIREs Data Mining Knowl Discov, Vol. 2, 2012, pp. 340-350.
[4] Chatterjee N., and S. Mohan. "Extraction-based single-document summarization using random indexing", 19th IEEE International Conference on Tools with Artificial Intelligence, Vol. 2, IEEE, 2007, pp.448-455.
[5] Jurgens D., and K. Stevens, "Event detection in blogs using temporal random indexing", Proceedings of the Workshop on Events in Emerging Text Types(eETTs), Association for Computational Linguistics, 2009, pp. 9-16.
[6] Budanitsky A., and G. Hirst, "Evaluating WordNet-based Measures of Lexical Semantic Relatedness", Computational Linguistics, Vol. 32, No. 1, 2006, pp. 13-47.
[7] Pincombe B., "Comparison of human and latent semantic analysis (LSA) judgements of pairwise document similarities for a news corpus", DTIC Document, 2004.
[8] Baeza-Yates R., and B. Ribeiro-Neto, "Modern information retrieval", ACM press New York, Vol. 463. 1999.
[9] Fellbaum C., "WordNet: An Electronic Lexical Database", MIT Press, Cambridge,1998.
[10] Roget P., "Roget's Thesaurus of English Words and Phrases", Longman Group Ltd, 1852.
[11] Jadidinejad A.H., and F. Mahmoudi, "Unsupervised Short Answer Grading Using Spreading Activation over an Associative Network of Concepts", Canadian Journal of Information and Library Science, Vol. 38, No. 4, 2014, pp. 287-303.
[12] 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, 1999. Vol. 11, No. 1, 1999, pp. 95-130.
[13] Jiang J.J., and D.W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy", Proceedings of International Conference Research on Computational Linguistics (ROCLING X), 1997, pp. 19-33.
[14] Lin D., "An Information-Theoretic Definition of Similarity", Proceedings of the Fifteenth International Conference on Machine Learning, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998, pp. 296-304.
[15] Landauer T.K., P.W. Foltz, and D. Laham, "An introduction to latent semantic analysis", Discourse Processes, Vol. 25, 1998, pp. 259-284.
[16] Gabrilovich E., and S. Markovitch, "Wikipedia-based Semantic Interpretation for Natural Language Processing", Journal of Artificial Intelligence Research, 2009.Vol. 34, 2009, pp. 443-498.
[17] Jadidinejad A.H., and F. Mahmoudi, "Advertising Keyword Suggestion Using Relevance-Based Language Models from Wikipedia Rich Articles", Journal of Computer & Robotics, Vol. 5, No.1, 2014, pp.29-35.
[18] Jadidinejad A.H., F. Mahmoudi, and M.R. Meybodi, "Clique-based semantic kernel with application to semantic relatedness", Natural Language Engineering, Cambridge University Press, Vol. 1, No. 1, (To appear), pp.1-18;
[19] Jadidinejad A.H., and Marza V., "Building a Semantic Kernel for Persian Text Classification with a Small Amount of Training Data", Journal of Advances in Computer Research, Vol. 6, No. 1, 2014, pp.125-136.
[20] Wang P., and C. Domeniconi, "Building semantic kernels for text classication using wikipedia", Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, 2008, pp. 713-721.
[21] Cristianini N., and J. Shawe-Taylor, and H. Lodhi, "Latent Semantic Kernels", Journal of Intelligent Information Systems, Vol. 18, No. 2-3, 2002, pp. 127-152.
[22] Wang Y., and E. Agichtein, "Temporal latent semantic analysis for collaboratively generated content: preliminary results", Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, ACM, 2011, pp. 1145-1146.
[23] Radinsky K., E. Agichtein, E. Gabrilovich, and S. Markovitch, "A word at a time: computing word relatedness using temporal semantic analysis", Proceedings of the 20th international conference on World wide web, ACM, 2011, pp. 337-346.
[24] Baroni M., and A. Lenci, "Distributional memory: A general framework for corpus-based semantics", Computational Linguistics, Vol. 36, No. 4, 2010, pp. 673-721.
[25] Turney P.D., and P. Pantel, "From frequency to meaning: vector space models of semantics", Journal of Artificial Intelligence Research, Vol. 37, 2010, pp. 141-188.
[26] Lee M.D., B. Pincombe, and M. Welsh, "An Empirical Evaluation of Models of Text Document Similarity", Proceedings of the 27th annual meeting of the Cognitive Science Society (CogSci'05), Erlbaum: Mahwah, NJ, 2005, pp. 1254-1259.
[27] Milne D., and I.H,Witten, "An open source toolkit for mining wikipedia", Artificial Intelligence,Vol. 194, 2012, pp. 222-239.
[28] Jurgens D., and K Stevens, "The S-Space package: An open source package forword space models", Proceedings of the ACL 2010 System Demonstrations, Uppsala, Sweden, 2010, pp. 30-35.
[29] Karlgren J., and M. Sahlgren, "From Words to Understanding", Foundations of real-world intelligence, CSLI Publications, 2001.

Navid Bahrami received the B.Sc degree in Software Engineering from Ghiasodin Jamshid Kashani University of Abyek, Qazvin, Iran in 2010 and the M.Sc degree in Software Engineering from Islamic Azad University of Qazvin, Qazvin, Iran in 2014. His research interests are Machine learning, Information retrieval and Data mining.

Amir Hossein Jadidinejad is faculty member of Islamic Azad University of Qazvin (QIAU). His research interest include Information Retrieval, Machine Learning and Statistical Data Analysis. He is also interested in applying state-of-the-art models of Information Retrieval and Machine Learning to very large collections, such as WWW.

Mojdeh Nazari Soleimandarabi received the B.Sc degree in Computer Software Engineering from Guilan University and M.Sc degree in the same field of study from Science and Research University, Guilan, Iran in 2012 and 2015, respectively. Her research interests include information retrieval, semantic relatedness, semantic web and data mining.