Int. j. inf. tecnol. (February 2022) 14(1):311–322
https://doi.org/10.1007/s41870-019-00314-w

ORIGINAL RESEARCH

Concept based document similarity using graph model

Sheetal S. Sonawane[1,2] · **Parag Kulkarni**[2,3]

Received: 9 November 2017 / Accepted: 23 April 2019 / Published online: 7 May 2019

Abstract To address the process of document similarity, ontology based knowledge base such as WordNet and Wikipedia is used widely. However, there are still available different challenges, such as polysemy, synonym and high dimensionality. In this paper, a novel method for calculating the similarity of text documents is proposed. The proposed system exploits ontological framework to give correct assessment of the similarity between terms. A modified method for concepts extraction using WordNet and Wikipedia is proposed in this paper. Text document is represented as a conceptual coexistence graph. Index is constructed to handle scalability and easy computation based on large concepts and terms association. Graph similarity is calculated using vertex similarity. The integrated approach can find theme of documents based on disambiguated and extracted concepts. The experimental has been evaluated on 20 newsgroup dataset and self-generated datasets. Results show that our approach significantly improved compared to bag of words approach.

Keywords WordNet · Wikipedia · Document similarity · Graph model · Vertex similarity

✉ Sheetal S. Sonawane
sssonawane@pict.edu

[1] Pune Institute of Computer Engineering, Pune, India

[2] College of Engineering, Pune, Maharashtra, India

[3] iKnowlation Research Labs Pvt. Ltd., Pune, India

1 Introduction

The volume of text documents are growing day by day due to the extensive progress of the internet. It motivates the need of efficient machine learning and information retrieval algorithms for text mining. An important algorithm in document classification and clustering is finding the similarity or closeness between documents. Hence, document similarity is useful method in the field of information retrieval. Existing approaches represent documents using feature vector and similarity is calculated using TF-IDF (Term frequency-inverse document frequency) and vector space model. These approaches do not work well to find semantic and meaningful similarity between documents. Also such similarity measure has the problem of sparsity. There are major issues in existing approaches like documents which are dissimilar may have the same content and meaning of term is not considered.

As the text data is increasing day by day, it is highly dimensional and carries semantic information. Therefore it is essential to identify core semantic. Core semantic generally represents theme of document.

The core of this paper is to use concepts instead of terms to capture the topics of documents by creating a concept based document representation model. Concepts are units of knowledge [1], each with a unique meaning. There are three major benefits of using concepts in document similarity. Concepts are less redundant, ambiguity is considered in concepts and semantic relation between concepts can be used to compute document similarity.

There are number of semantic approaches proposed to handle the problem of document similarity. In ontology based approach WordNet and Wikipedia are commonly used ontologies to find the semantic relation of word. Redundancy and ambiguity problems are addressed using

ontologies. There are many challenges that still exist for semantic similarity computation. (i) Synonym and polysemy problems. (ii) Extraction of core semantics from text document. (iii) High dimensional terms in text. These challenges have been addressed by various researchers but have some weaknesses like the theme of document is not represented and concept representing term increases word space without increasing the performance of similarity.

This paper focuses on providing the similarity between documents using concept and their relations.

The contributions of the presented work are

(i) In this paper, the modified similarity measure based on WordNetsynset, WordNet gloss and Wikipedia context is proposed. Previous works have showed that the structural information on WordNet can improve the similarity measure, but the impact of adding textual information using WordNetsynset and Wikipedia are not still extensively researched. In this paper we explore combination of WordNetsynset, WordNet gloss and Wikipedia context can give a more accurate assessment of similarity computation.

(ii) The weighted conceptual graph of coexistence term is proposed for representing a text document. The existing document representation model neglects the association of terms with concepts related to the document. The proposed graph-based document model captures the valuable knowledge based on the relationship between terms and concepts. It also captures the semantic relationship between words and more accurately represents theme of the document.

(iii) We introduce inverted index mechanism for indexing concept association with terms. Although inverted index have been extensively used in the area of information retrieval, their potential impact on associating concepts to terms in document similarity has not been fully investigated. We have observed that it is the most successful data structure for this application. Concepts are disambiguated using Wu-palmer semantic similarity measure and Later, concept score is calculated based on its association with other concepts.

(iv) The proposed method can estimate better similarity as compared to traditional bag of words approach by observing experimental results.

The paper is organized as follows. Related work is explained in Sect. 2. Proposed work is described in Sect. 3. The detail phases such as graph construction, concept extraction, concept disambiguation and graph similarity is described in Sect. 3. Performance analysis is presented in Sect. 4 and conclusion is stated in Sect. 5.

2 Related work

Document collection is represented as

$$D = \{d_1, d_2, \ldots d_m\} \tag{1}$$

where m is number of documents in the collection.

Each document vector is represented as set of keywords

$$d_i = \{k_{i1}, k_{i2}, \ldots k_{in}\} \tag{2}$$

where n is number of index terms.

Hence document set is defined as,

$$D = \sum_{(i=0)}^{m} \sum_{(j=0)}^{n} d_{ij} \tag{3}$$

The basis approaches for document similarity works on features which captures important characteristics of the data.

1. Text based similarity: Text based document similarity [2] is measure by comparing the words in two documents as features. Vector space model uses cosine similarity to compute document similarity.
2. Co-occurrence based similarity: Document similarity is proposed where documents are represented by the set of candidate keywords [3]. Initial similarities are computer between documents and keywords and the weights of keywords to documents using cosine similarity. This system has a problem of Semantic ambiguity.

   Traditional algorithms do not consider the semantic relationships among words [4] so they cannot accurately represent the meaning of documents [5].
3. Semantic similarity: In case of documents where different terms are used to describe the same concept in different documents. Text based retrieval may give inaccurate and incomplete result. Semantic similarity states that two keywords are related because they share some aspects of their meaning.

Semantic similarity is useful term and measurement of semantic similarity between two keywords is a challenging task. It is computed using following approaches:

(i) Ontology based similarity

Ontologies have been of great interest for semantic similarity research group as they offer a structured and unambiguous representation of knowledge in the form of concepts connected by means of semantic meaning. These structures help to assess the degree of semantic proximity between terms.

To overcome the problem of redundancy and ambiguity, semantic information from ontology such as WordNet [6–10] and Wikipedia [11–13] has been widely used to improve the quality of similarity.

WordNet is a lexical knowledge base, containing information about synonyms, hypernyms and meronyms. WordNet is richer in content than other machine processing dictionaries and thesauri. It is easier to use for external applications and is freely available.

Wikipedia is a multilingual, web-based, freely available encyclopedia, constructed in a collaborative effort of voluntary contributors. Articles in Wikipedia form a heavily interlinked knowledge base enriched with a category system emerging from collaborative tagging, which constitutes a thesaurus. Wikipedia thus contains a rich body of lexical semantic information.

Ontology based similarity is carried out by using following approaches

1. *Edge-counting approach* Ontologies can be seen as a directed graph in which concepts are related by means of taxonomic (is-a) links. Limitation of such approach is only minimum path between taxonomies are considered.
2. *Feature-based measure* Similarity between concepts is assessed as function of their properties. This approach exploits more on semantic knowledge evaluating common and difference of compared concepts. Limitation of this approach is weighting parameter of features need to be tuned with ontology.

(ii) Corpus based similarity

To overcome the weakness of text similarity is to leverage the information derived from a large corpus, which is often the Web. Semantic similarity between web documents is calculated on the basis of page count and snippets retrieved using web search engine. Lexical patterns are extracted from the text snippets and clustered to identify the different patterns that describe same semantic relation. It is an automatic method based on lexical pattern generation works on named entity which is not possible using manual ontologies but it depends on search engine performance.

4. Conceptual similarity: Concept based document similarity is essential to find conceptual similarity between document. It is [14] used to handle similarity based on concepts it represents. Concept based similarity is find out using manually built thesauri, term concurrences data, by extracting latent word relationships and extracting concept from the corpus. Concept based document similarity is characterized using three parameters

1. Concept representation
2. Mapping method
3. Use in Information retrieval

The concept extraction is also used in generation of n-gram for matching two documents [15] using Wikipedia information.

Concept relations should be used to quantify the relevance of concept with respect to a term. Tingting et al. [5] proposed method to find conceptual representation using WordNet.

However, some important words which are not included in WordNet lexicon [1] are not considered as concepts for similarity evaluation. Also there still exist several challenges, such as synonym and polysemy, high dimensionality, extracting core semantics from texts. Hence there is need to extract concepts using world knowledge and rank them. The extracted concepts are large, so there is necessity to reduce the feature space. In this paper we quantify the closeness between the concepts and integrate into document similarity measure. Hence, documents do not have same words or concepts to judge their similarity. Also there is a need of structured representation of texts that combine fine grained semantic relations. Conceptual graph model is more suitable model for describing the relationship between terms and concepts. Edges in the semantic graphs should be considered weighted so as to capture the degree of associability between concepts. Once graph is constructed, researchers have proposed different graph similarity methods for computing document similarity and performance is good in comparison with state of the art methods.

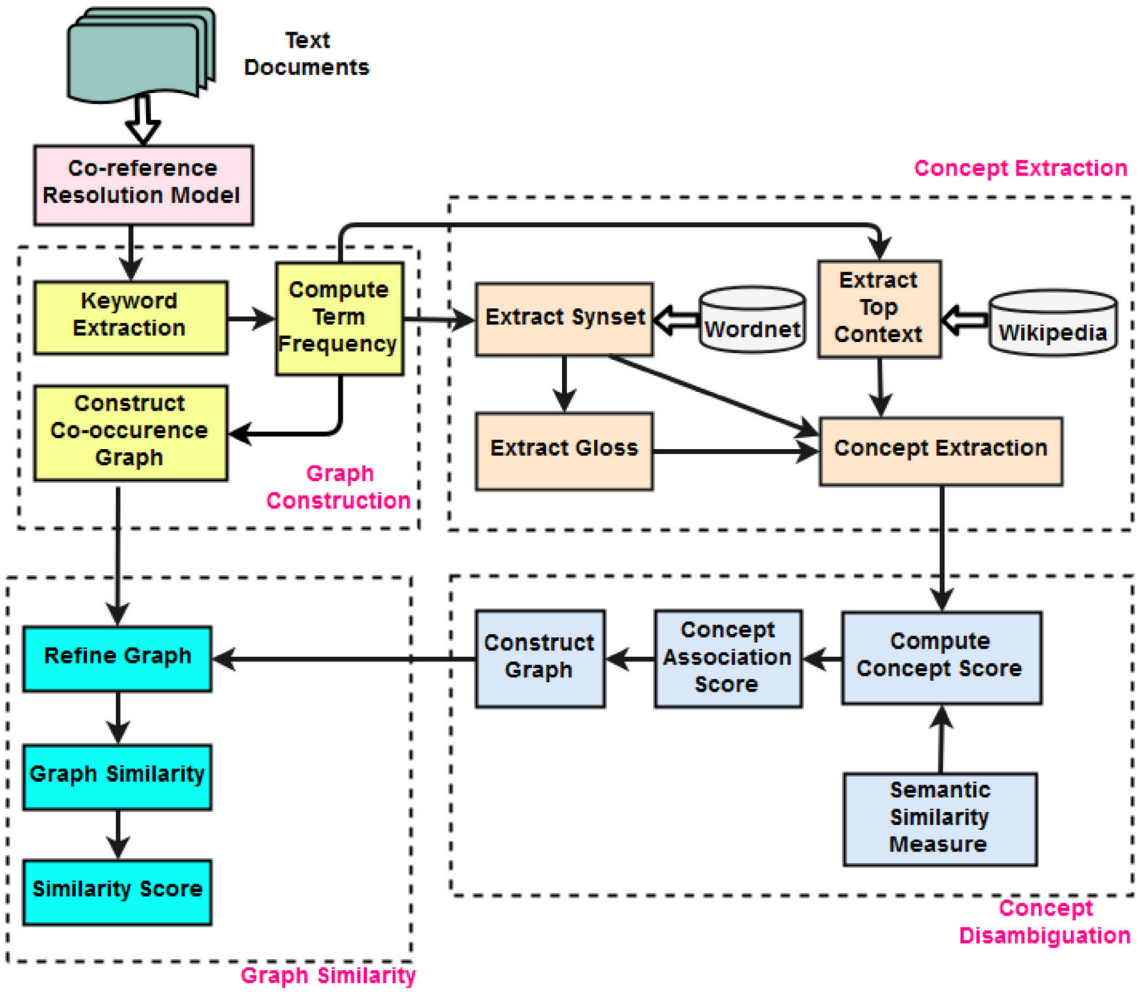
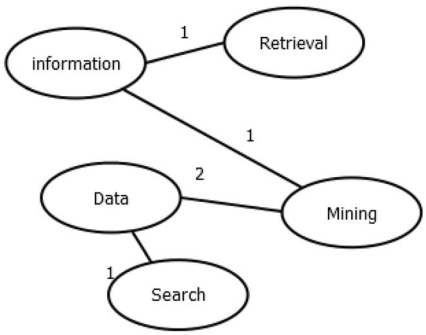


Fig. 1 System architecture

Fig. 2 Text representation using coexistence graph



	information	Retrieval	Data	Mining	Search
information	0	1	0	1	0
Retrieval	1	0	0	0	0
Data	0	0	0	2	1
Mining	1	0	2	0	0
Search	0	0	1	0	0

3 Document similarity

Document similarity measures are important components of many text analysis tasks like information retrieval, document classification and document clustering. Traditional measure the structural words overlap between documents. These methods ignore the deeper conceptual connections.

A novel method is proposed which uses ontology based approach and integrate Word-net and Wikipedia to measure semantic similarity between documents. Figure 1 illustrates the proposed system architecture.

Table 1 An example of concept description

Term	Synset	Gloss	Wikipedia
Country		The territory occupied by a nation; "he returned to the land of his birth"; "he visited several European countries"	A country is the territory controlled by or associated with a national government. A country may be an independent sovereign state or one that is occupied by another state, as a non-sovereign or formerly sovereign political division, or a geographic region associated with sets of previously independent or differently associated people with distinct political characteristics
	State	The group of people comprising the government of a sovereign state; "the state has lowered its income tax"	
	Land	The land on which real estate is located; "he built the house on land leased from the city"	
	The description of concept[a] is obtained by		
	$Synset(country_1) \cup Gloss(country_1) \cup Gloss(synset(country_1) = state) \cup Gloss(synset(country_1) = land) \cup \{Wikipedia - context(country)\}$		

[a]Here 1 subscript defines first synset of WordNet

First, document is preprocessed using NLP pre-processing steps: stop word removal, stemming, POS tagging using OPENNLP model. We applied co-reference resolution proposed in [16]. Co-reference resolution is a method of finding association of feature terms or mentions in the discourse that refers to the same entity. The noun terms (NN, NNS, NNPS, and NNP) are extracted using pre-processing model. Term frequency of each noun term is calculated and score is assigned to each term. The number of times terms occur together is used for defining edge weight. The weighted terms are used and graph is constructed based on its occurrence in the document. The graph construction is done using coexistence based representation and is described in Sect. 3.1.

Once the quality terms are extracted, we query ontologies like Wikipedia and WordNet using the terms and related concepts are extracted. Inverted index mechanism is used for indexing concepts with term. The detail concept extraction is explained in Sect. 3.2.

All extracted concepts are not useful hence to choose the most relevant concepts; the relatedness of the concepts is calculated. The semantic similarity is used for disambiguate the concepts. The extracted concepts are disambiguated using Wu palmer's semantic similarity measure. 90% similar concepts are chosen for given terms. The weights are assigned to the concepts according to its associated terms. The concept disambiguation module is described in Sect. 3.3. The constructed concept graphs are used and graph similarity is applied to compute the similarity score. The vertex similarity score computation is described in Sect. 3.4.

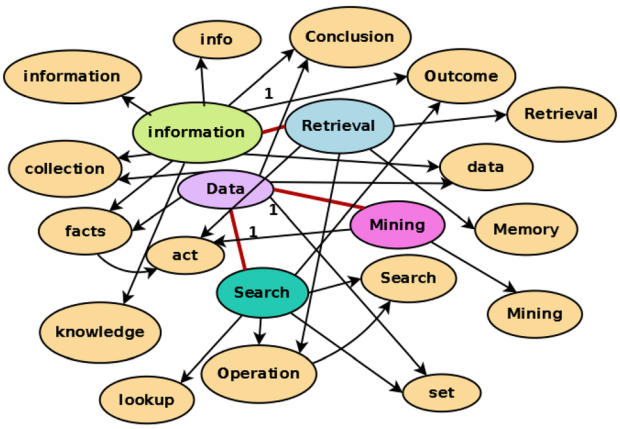


Fig. 3 Sample graph representation with concept

3.1 Graph construction

Text document can be represented as a graph in many ways. Nodes denote features and edge represent relationship between nodes [17–21].

For a given document, its graph representation is defined as a vertex corresponds to unique terms of the document and edges represent coexistence [22] between the terms

Document 1: Information retrieval has changed recently. Data Search is important.
Terms: {information, retrieval, data, search}

(a)

Task	Outcome
Concept of (information)	{I1:info, I2:facts}
Concepts of (retrieval)	{R1:operation,R2: retrieval}
Concept of (data)	{D1:facts,D2:info}
Concept of (search)	{S1:lookup,S2:operation, S3:search }

(b)

Concept	Terms
I1,D1:info(0)	Information(0.8), Data(0.6)
I2,D1:facts(0)	Data(0.67)
R1:S2:Operation(0)	Retrieval(0.7),Search(0.5)
S1:Lookup(0)	Search(0.8)
S3:Search(0)	Search(0.9)

(c)

Fig. 4 **a** Input document. **b** Extracted concepts for the terms (output of **a**). **c** Inverted index constructed for concepts (output of **b**)

within a document. Edge weight is number of occurrences of the terms in the document.

Let $T = \{(t_1,f_1),(t_2,f_1),\ldots(t_p,f_p)\}$ where $t_i \in \{NN, NNP, NNS, NNPS\}$ and $f_i = TF$ $G = (V,E)$where $V = \{(t_1,f_1),(t_2,f_1),\ldots(t_p,f_p)\}$ $E = \{(e_1,f_1),\ldots(e_p,f_p)\}$ t_1 is neighbor of t_2, if t_1 occurs together with t_2 and number of times it occurs in given document is represented using the edge weight.

$e_1 = \{(t_1,t_2),w_1\}$

Example Doc 1: information retrieval has changed recently. Data mining is an essential for data search. Information mining is part of data mining.

$$V = \{(\text{information},2),(\text{mining},3),(\text{data},3),(\text{retrieval},1),(\text{search},1)\}$$

The graph constructed with adjacency matrix for this example is shown in Fig. 2.

3.2 Concept extraction

Goal of our system is to provide meaningful representation of document. Hence concept plays an important role in representing the document. Terms are taken as input to this module and WordNet and Wikipedia are used for querying related concepts.

WordNet is the largest lexical English database which is used widely across the world. It provides sense of words and relation between words. However the WordNet is insufficient as it does not provide all the words. Wikipedia is currently the largest knowledge repository on the web.

Most of the researchers are used Wikipedia and WordNet individually as knowledge base for extracting concepts. Advantages of both the ontologies are considered in this work. We attempted to use both the system to extract the appropriate concepts.

Definition 1 (*WordNetSynset + WordNet Gloss + Wikipedia Context*) Let WordNet $Synset = Si = \{S_{i1}, S_{i2}, \ldots S_{il}\}$ where l is number of senses for t_i such that $t_i \in T$.Each Synset in WordNet has a gloss [23] associated with it that contains one or more definitions and some examples. Let Synset-Gloss $Synset - GlossS_{ij}$ be the definition and examples of S_{ij}. Let Wikipedia-context (t_i) be the union of concepts occurs in the top paragraph. Then the extracted concepts C_i is defined as,

$$C_i = S_i \cup Gloss(t_i) \cup \left\{ \sum_{i=1}^{p} \sum_{j=1}^{l} Synset - GlossS_{ij} \right\} \cup \{Wikipedia - context(t_i)\} \quad (4)$$

The formula (4) is a representation of concept extraction. Table 1 shows example of Definition 1.

The extracted concepts are added in the graph. The example of a concept graph is shown in Fig. 3. For efficiency and easier access, inverted indexing is used.

Inverted indexes are the most basic and commonly used data structure in the field of information retrieval. To avoid a lengthy sequential scan through each term in a document and to improve run-time performance of retrieval, inverted index is used.

An entry for each of the n concepts according to Eq. (1) is stored in a structure called an index. For each concept, a pointer references is called a posting list. The posting list contain the terms of a document which has relation with concept. In the Fig. 4, the posting list contains both the terms and the term frequency. The concepts are indexed and terms are appended in the file. The inverted index file provides faster access to the associated terms.

The Fig. 4a–c represents the steps in creation of inverted index. The concepts are indexed and initially weight given for concepts is 0. The steps are given in Algorithm 1.

Algorithm 1: Inverted index Construction for document d.

Input: Set of documents $D = \{d_1, \ldots, d_n\}$ where n is number of documents

Terms $T = \{(t_i, f_i), \ldots (t_p, f_p)\}$ where t_i is set of terms and f_i is term frequency of document d. Concepts $\{(c_i, w_i), \ldots, (c_l, w_l)\}$ is set of extracted concepts based on T in document d. Initially $w_i = 0$

SC = stored concept list into index table

Output: Inverted index table IT. Each table entry consist of terms and concepts list

1: **Procedure** INVERTED_INDEX (Document set D)
2: $SC = 0,\ i = 0,\ j = 0,\ index = 0;$
3: **For** each document
4: **For** each Term t_i do
5: **For** each Concept c_i of t_i do
6: **If** $c_i \in SC$ then
7: $index = getindex(c);$
8: $IT_{index}.T = \cup d_i.t_i$
9: $IT_{index}.T = \cup d_i.f_i$
10: **else**
11: $w_i = 0;$
12: $IT_j.C = (c_i, w_i)$
13: $IT_j.T = \cup d_i.t_i;$
14: $IT_j.T = \cup d_i.f_i$
15: $SC_j = (c_i, w_i);$
16: $j{+}{+};$
17: **end if**
18: **end for**
19: **end for**
20: **end for**
21: return IT
22: **end procedure**

In order to extract the relevant concepts of the term, the relation between the concepts is calculated. If the concepts of a given document are highly related to each other, they are more related to the theme of the document. To find similarity between concepts, semantic similarity is required. Semantic similarity computation is used in many applications in the information retrieval research area.

Semantic similarity is calculated on the basis of information content and structure based. The score of sample concepts by using different measures is given in Table 2. The information content based approach is based on probability of occurrence of word in instance of Synset it occurs in. This measure has sparse data problem. The other type is using hierarchical structure of WordNet with path length property. This method combines the length of the path with the depth of the concepts in a weighted and non-linear manner. This method is not having sparse data problem.Hence Wu-palmers [24] method is used for disambiguating concepts in this paper. The description of Wu palmer's measure is given in Definition 2. In order to get the quality concepts, 90% similarity between concepts representing vertices is used. The integrated graph of terms and concepts example is described in Fig. 5.

Definition 2 Wu-palmers Semantic similarity computation: Let N_1 and N_2 is the number of IS-A links from C_1 and C_2 respectively to the most specific common concept C, and H is the number of IS-A links from C to the root of the taxonomy.

$$Sim_{wu\_palmer}(C_1, C_2) = \left(\frac{2H}{N1 + N2 + 2H}\right) \tag{5}$$

3.3 Concept weight calculation

The disambiguated nouns may increase the dimensionality of the feature space. Hence we need to find a way to reduce the dimensionality and to get the core concepts. The one concept is associated with many terms in a document hence we use this property to assign weight to concept. The importance of the word senses within a given document should be considered. Hence, we focused on assigning weight to concepts based on the association with the terms.

Table 2 Semantic relatedness using different measure (WP: Wu palmers, LCH: Leacock and Chodorow, Path: Shortest Path, Lin: Lin et al., Res: Resnik, JCN: Jiang & Conrath)

Concepts	Structure based			Information content based			
	WP	LCH	Path	Lin	Res	JCN	Lesk
Data-information	1.0	3.68	1.00	1.0	7.43	1.0	670
Data-concept	0.33	1.49	0.11	0.13	0.77	0.09	34
Data-conclusion	0.461	1.60	0.12	0.11	0.77	0.07	28
Data-outcome	0.20	1.49	0.11	0.10	0.77	0.07	28
Data-knowledge	0.439	1.89	0.16	0.15	0.77	0.11	26
Data-lookup	0.170	1.29	0.09	0.00	0.00	0.05	16
Data-fact	0.360	1.60	0.12	0.11	0.77	0.08	33
Data-collection	0.889	2.99	0.50	0.82	5.25	0.45	233

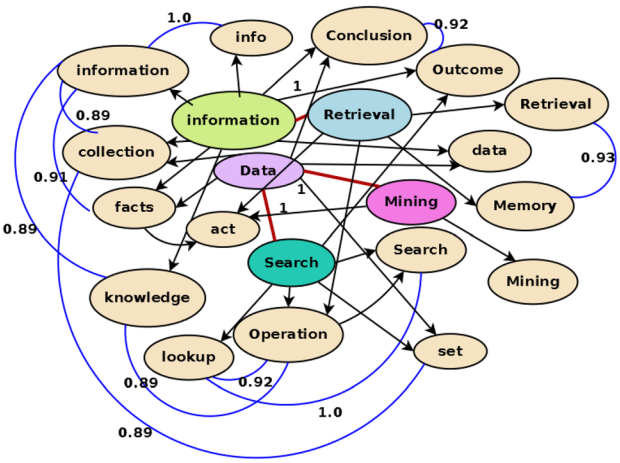


Fig. 5 Concept association score based on Wu-palmers method of example given in Fig. 2

Description of weight calculation of concept is given in Definition 3.

We have Inverted index file which has concepts and associated terms. The weight associated with term is useful for describing the weight of concepts. For example the terms information, data and retrieval has concept "information". Hence information (concept) = 08 + 0.3 + 0.54 = 1.64. The example with inverted index is described in Fig. 6.

Definition 3 (*Concept weight*) given document d, let $\{(t_1,f_1)\ldots(t_p,f_p)\}$ be the terms with frequencies and $\{(c_1,w_1)\ldots(c_q,w_q)\}$ be the disambiguated concepts in C. Hence the mapping $f : T \rightarrow C$, for t_k and t_m $(t_k, t_m \in T)$, weight of c_i is computed by

$$w_i = \{f_k + f_m\} \tag{6}$$

3.4 Concept score calculation

In the previous step we calculated the concept weight based on the terms associated with it. The concepts are semantically related to each other. To get quality concepts, there is need to rank these concepts based on its association with other concepts. The concept is important if it is linked to other concepts of higher weight. Hence, Concept score is calculated on the basis of weight assigned to terms and association score. The weighted average of concept node is

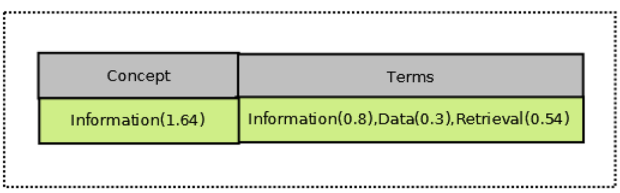


Fig. 6 Inverted index example where information (concept) = 0.8 + 0.3 + 0.54 = 1.64

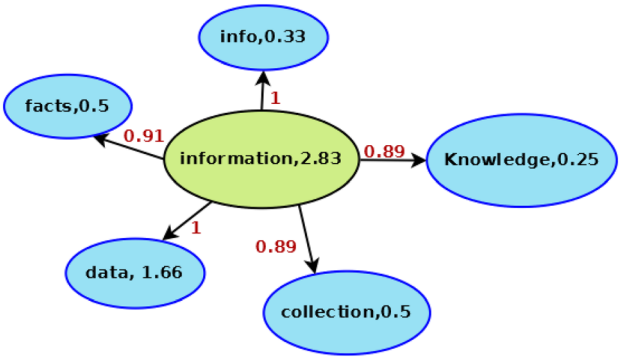


Fig. 7 Concept score calculations

calculated to calculate score of concept. The formula is described in Definition 4 and the example is given in Fig. 7.

Definition 4 For given document d, let $\{(t_1,f_1)\ldots(t_p,f_p)\}$ be the terms with frequencies and $\{(c_1,w_1)\ldots(c_q,w_q)\}$ be the disambiguated concepts of document d, Hence

$$Score\ (c_i) = w_i \times \frac{\sum_{j=1}^{l} w_j \times c_j}{\sum_{j=1}^{l} c_j} \tag{7}$$

A large value of $Score(c_i)$ indicates that c_i is a conceptually important concept in a document.

4 Graph similarity

Many researchers have worked to compute similarity between graphs which is representation of document. Euclidean distance is commonly used method to find similarity based on vertex label. Schumacher et al. [14] proposed graph similarity using Graph edit distance and graph isomorphism. Weighted graph similarity for the application of document similarity is unexplored area. As weighted graph is used for modeling document, we have explored use of vertex cosine similarity [25] for finding similarity between two documents. The formula is given in Eqs. (8) and (9).

$$Label - distance(v_i, v_j) = \begin{cases} 1, if\ v_i = v_j \\ 0, if\ v_i \neq v_j \end{cases} \tag{8}$$

$$Vertex - Similarity(v_i, v_j) = \frac{\left|common - neighbours\ (v_i, v_j)\right|}{\sqrt{deg\ (v_i) \times deg\ (v_j)}} \tag{9}$$

Algorithm: Graph Similarity

Input: Graph $G = \{G_1, G_2, \ldots G_n\}$ where $G_1 = (V_1, E_1)$

Output: Semantic similarity score

Set $sim_{score} = 0.0$

For each graph $G_i \grave{o}\, G$ *and* $G_j \grave{o}\, G$ where $(i \neq j)$ do

For each vertex $v_{i1}\ of\ G_i$ *and* $v_{j1}\ of\ G_j$ do

$dist = label - distance(v_{i1}, v_{j1})$ (Equation 4)

$if\ (dist == 1)$

$sim - score = vertex - similarity(v_{i1}, v_{j1})$ (Equation5)

End if

End for

Combine score of all vertices for graph G_i

return sim_{score}

End for

5 Performance analysis

In this section, we evaluate our system on two different dataset, compare the results with traditional vector space model and discuss results.

Table 3 Description of self-constructed dataset

Dataset	No of categories	Name of the category	Description
Self-generated dataset 1	4	Information retrieval, education, sports and data structures	Mixture of various domain categories
Self-generated dataset 2	9	Physics, chemistry, mathematics, biology, physics, social science, database, geography, history	All technical categories are considered

5.1 Dataset

Performance comparison is done with two set self-generated dataset (One is of four categories and other is of 8 categories where 100 documents are in each category) and 20 news group dataset. 20 news group dataset is widely used in existing work of document similarity. For the analysis the data from five different categories is considered. Self-generated dataset description is given in Table 3.

5.2 Evaluation metrics

In experiment analysis, the similarity is computed using precision, recall and F measure. F-measure combines the information of precision and recall. The confusion matrix is constructed for successful and unsuccessful retrieval and performance is analyzed.

The cluster quality is evaluated using purity. Purity assumes that all the document of a cluster is the member of actual class of cluster. It is calculated using following formula

$$purity(W, C) = \frac{1}{N}\sum_{k} \max j|W_k \cap C_j| \tag{10}$$

where $W = \{W_1, W_2, \ldots W_k\}$ the set of clusters, C is $= \{C_1, C_2, \ldots C_j\}$ is set of classes and N is number of documents.

Correlation between two documents is calculated using Matthew's correlation coefficient (MCC). It is the metric used for accessing the quality of the predicted value to the observed value. As a metric for comparing documents, it takes into account the correctness AND the incorrectness (true/false positives/negatives) of the comparison, allowing for a more balanced score among different comparisons.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{11}$$

where TP = true positives, TN = true negatives, FP = false positives and FN = false negatives.

Table 4 Similarity performance using self-generated dataset 1

	Precision	Recall	F-measure
IR			
VSM	0.42	0.33	0.423
CS	**1**	**1**	**1**
Education			
VSM	0.6	0.712	0.651
CS	**0.84**	**0.85**	**1**
Sports			
VSM	0.8	0.8	0.8
CS	**1**	**1**	**1**
DS			
VSM	0.4	0.4	0.4
CS	**0.8**	**0.8**	**1**

Bold values indicate improvement in the performance

Table 5 Similarity Performance using self-generated dataset 2

Method	P	R	F	Accuracy	Purity
2 categories					
VSM	0.85	0.85	0.85	0.85	0.85
CS	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
3 categories					
VSM	0.57	0.54	0.51	0.68	0.81
CS	**0.80**	**0.63**	**0.68**	**0.78**	**0.85**
9 categories					
VSM	0.80	0.37	0.46	0.52	0.55
CS	0.80	**0.56**	**0.58**	**0.61**	**0.61**

Bold values indicate improvement in the performance

5.3 Results and analysis

The self-constructed dataset 1 is mostly the semantically related documents. Hence the terms in the documents are not similar. Table 4 shows detail analysis comparison with vector space model. In vector space model the terms similarity is considered hence our system is improved as concept are taken into consideration. Table 5 describes the performance on self-generated dataset 2. The performance is compared on 2 categories, 3 categories and 9 categories.

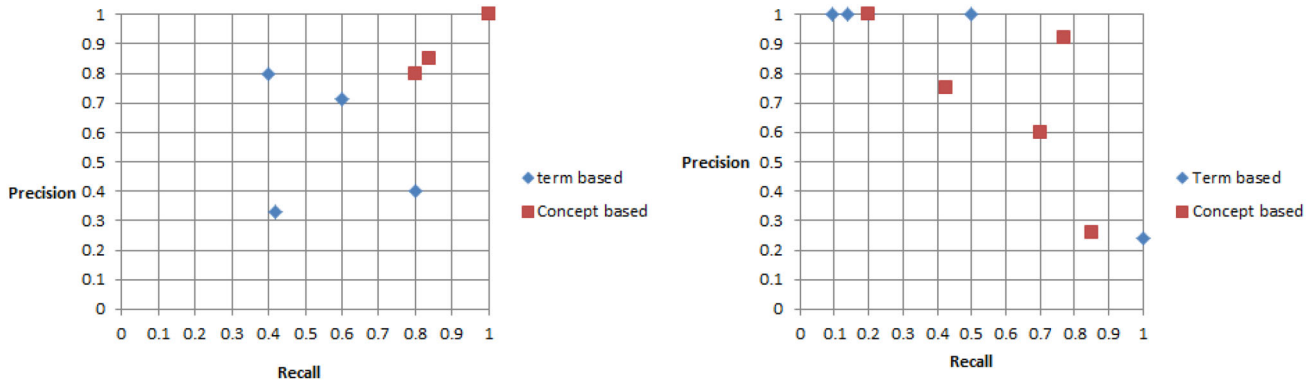


Fig. 8 Precision and Recall of concept based similarity method (CS) (on self-generated dataset 1 and 20 newspaper dataset respectively) in comparison with TF-IDF

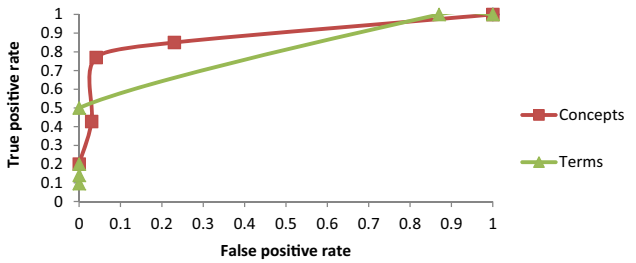


Fig. 9 ROC curves of the similarity using different methods on 20 newspaper dataset

We found the performance is consistent among all. Even though, this dataset has mostly related document, our system has improved performance in all metrics.

Figure 8 describes precision and recall of concept based similarity method (CS) in comparison with vector space model. It shows that the precision and recall range is from 0.8 to 1.0 for the proposed system while for traditional TF-IDF is from 0.4 to 0.8.

Figure 9 shows concept based ROC area under curve is 1. Also the performance of these algorithms at different false positive regions varies from the traditional TF-IDF scheme. At a low false-positive point i.e. FPR = 0.12, the concept based similarity performs better than considering terms. This phenomenon can be clearly observed from the ROC curves.

The similarity of proposed system is compared with two BOW and two concept based state of the art algorithms.

Table 6 Similarity performance on 20 newsgroup dataset

Method	F
Group-average agglomerative clustering (GAAC)	0.47
K-Means	0.54
Bi-secting K-means	0.50
N-gram and Wikipedia in GAAC	0.59
CS (proposed system)	**0.64**

Bold value indicates improvement in the performance

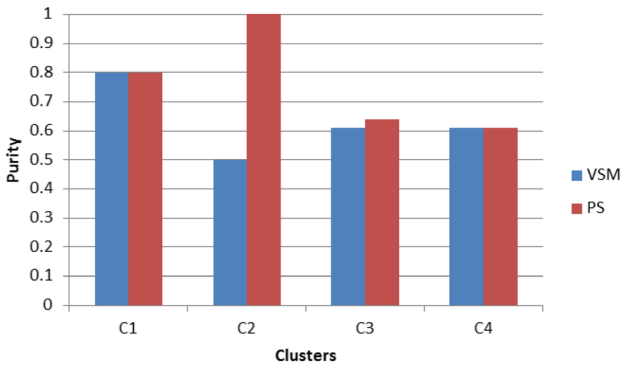


Fig. 10 Clustering performance on self-generated dataset. *PS* proposed system, *VSM* vector space model

The analysis is shown in Table 6. The analysis shows that the proposed work achieve high performance on 20 newsgroup dataset.

The application of document similarity is also evaluated using clustering method. The self-generated dataset 1 and 2 are used for performing document clustering. For self-generated dataset 1, the cluster 1 and cluster 2 performances is same but for cluster 3 and 4, the performance is improved. Figure 10 shows the clustering performance using purity measure on self-generated dataset 1.

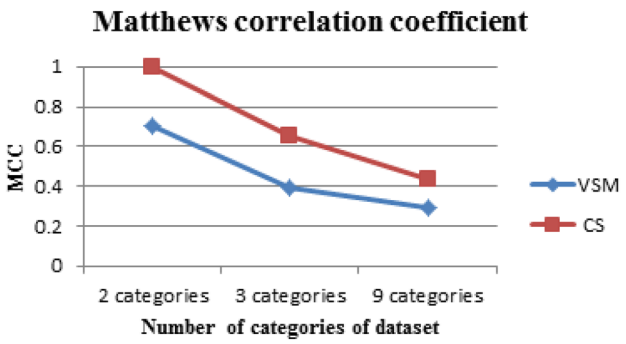


Fig. 11 MCC performance on self-generated dataset 2

Table 3 shows detail analysis of self-generated dataset 2. The performance of our system is better than the base across all categories of dataset.

Self-generated dataset 2 is also analyzed to find correlation between documents using MCC. The analysis indicates the improvement in the correlation using concept based similarity across all categories. Figure 11 shows the analysis.

6 Conclusions

In this study, an important problem of document similarity is addressed. Concept based similarity using graph model is proposed in this paper. This paper presents a methodology for Concept extraction using WordNetsynset, WordNet gloss and Wikipedia context and represented structural information using conceptual graph. An important data structure an inverted index is constructed to maintain the association of concepts and terms. The proposed approach uses WordNet based concept disambiguation method. Further concept score is calculated to reduce the dimensions of feature space. The coexistence graph is refined using integrating concepts. Graph model possesses different basic properties.

Simple graph property using degree and adjacency property is used for calculating vertex similarity to measure similarity between two graphs.

The four important problems are addressed in this paper. The problems are appropriate concept extraction, handling disambiguation between terms, representing conceptual description using graph and similarity computation. Experimental results on three real datasets show that the proposed approach outperforms the term based similarity.

Future work can be done in this topic for using fuzzy graph that represent weight of vertex and edge. Another interesting topic to do further is to use an efficient graph similarity measure.

References

1. Huang L et al (2012) Learning a concept-based document similarity measure. J Am Inf Res 63(8):1593–1608
2. Salton G, Wong A, Yang C (1975) A vector space model for automatic indexing. Commun ACM 18(11):613–620
3. Hammouda K, Kamel M (2004) Document similarity using a phrase indexing graph model. Knowl Inf Syst 6(6):710–727
4. Buttler D (2004) A short survey of document structure similarity algorithms. In: Proceedings of international conference on internet computing, pp 3–9
5. Wei T et al (2015) A semantic approach for text clustering using WordNet and lexical chains. Expert Syst Appl 42(4):2264–2275
6. Cimiano P, Hotho A, Staab S (2005) Learning concept hierarchies from text using formal concept analysis. J Artif Intell Res (JAIR) 24:305–339
7. Hensman S (2004) Construction of conceptual graph representation of texts. In: Proceedings of student research workshop at HLT-NAACL, Boston, pp 49–54
8. Sajgalk M et al (2013) From ambiguous words to key-concept extraction. In: IEEE proceedings of 10th international workshop on text-based information retrieval at DEXA 2013, pp 63–67
9. Warin M, Volk HM (2004) Using WordNet and semantic similarity to disambiguate an ontology
10. Dennai A et al (2015) A new measure of the calculation of semantic distance between ontology concepts. Int J Inf Technol Comput Sci IJITCS 7(7):48
11. Hadi A et al (2008) Key word suggestion using conceptual graph construction from Wikipedia rich documents. In: Proceedings of international conference on information and knowledge engineering (IKE'08), Las Vegas, USA, pp 14–17
12. Ratinov L et al (2011) Local and global algorithms for disambiguation to Wikipedia. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies-volume 1. Association for Computational Linguistics
13. Huang A et al (2009) Clustering documents using a Wikipedia-based concept representation. Advances in knowledge discovery and data mining. Springer, Berlin, pp 628–636
14. Schumacher M, Ponzetto SP (2014) Knowledge-based graph document modeling. In: Proceedings of the 7th ACM international conference on Web search and data mining. ACM
15. Kumar N, Vemula VVB, Srinathan K, Varma V (2010) Exploiting n-gram importance and additional knowledge based on Wikipedia for improvements in GAAC based document clustering. KDIR
16. Sonawane SS, Kulkarni PA (2016) Context-based co-reference resolution for text document using graph model (cont-graph). Int J Knowl Eng Data Min 4(1):1–17
17. Sonawane SS, Kulkarni PA (2014) Graph based representation and analysis of text document: a survey of techniques. Int J Comput Appl 96(19):1–8
18. Blanco R, Lioma C (2011) Graph-based term weighting for information retrieval. Inf Retr 15(1):54–92
19. Sonawane SS (2014) Graph based information retrieval. Int J Adv Comput Knowl Discov III(I):2278–5698
20. Mihalcea R, Radev D (2011) Graph-based natural language processing and information retrieval. Cambridge University Press, Cambridge
21. Rousseau F, Vazirgiannis M (2013) Graph-of-word and TW-IDF: new approach to ad hoc IR. In: Proceedings of the 22nd ACM international conference on information & knowledge management, pp 59–68, ACM
22. Schenker A et al (2003) Classification of web documents using a graph model. In: Proceedings of 7th international conference on document analysis and recognition (ICDAR2003). Computer Society Press, Scotland
23. Banerjee S et al (2003) Extended gloss overlaps as a measure of semantic relatedness. IJCAI 3:805–810
24. Wu Z et al (1994) Verbs semantics and lexical selection. In: Proceedings of the 32nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics
25. Weale T et al (2009) Using the Wiktionary graph structure for synonym detection. In: Proceedings of the 2009 Workshop on the people's web meets NLP: collaboratively constructed semantic resources. Association for Computational Linguistics