A Taxonomy based Semantic Similarity of Documents using the Cosine Measure

Ainura Madylova
Department of Computer Engineering
Istanbul Technical University
Maslak, Istanbul TR34469 Turkey
Email: madylova@itu.edu.tr

Şule Gündüz Öğüdücü
Department of Computer Engineering
Istanbul Technical University
Maslak, Istanbul TR34469 Turkey
Email: sgunduz@itu.edu.tr

***Abstract*—In this paper, we present a new method for calculating semantic similarities between documents. This method is based on cosine similarity calculation between concept vectors of documents obtained from a taxonomy of words that captures IS-A relations. The calculation of semantic similarities between documents is a very time consuming task, since it is necessary first to calculate semantic similarities between each pair of words that appear on different documents. In this paper, we present a new method to calculate semantic similarities between documents which results in faster computational time. Both a taxonomy based semantic similarity and cosine similarity are employed. First, the concept vectors of documents are obtained by extending the terms in the document vectors with their corresponding IS-A concepts. Cosine similarity is then calculated between those concept vectors of documents. Thus, the overall similarity between documents is a combination of cosine similarity and semantic similarity. The proposed semantic similarity is tested in document clustering problem. The experimental results show that our method achieves a good performance.**

I. Introduction

With the recent growth and diversity of electronic data on the World Wide Web (www), it becomes more difficult for Internet users to find the useful information from these huge amounts of data. Search engines and recommender systems help people to reduce the information overload by finding relevant information on their search topic. Clustering of documents is one of the techniques used in search engines and in recommender systems for efficiently finding documents that have similar topics [1], for improving the performance of information retrieval systems [2], for assisting users on a web site [3] and for personalization of search engine results [4]. Formally, document clustering is an optimization problem where the input of the problem is a set of documents and a (dis)similarity measure between these documents. Thus, similarity plays an important role in document clustering.

Many traditional document clustering approaches rely on single term analysis of text, such as the vector space model (also called bag of word model or word space) [5]. In this model, the documents are represented as a feature vector of words (terms) that appear in the entire document set. Each feature vector contains term weights (usually Term Frequencies (TF) or Term Frequency/Inverse Document Frequency (TF-IDF)) of the terms appearing in that document. Clustering techniques based on this model make use of single term analysis only and do not consider semantically related words/terms appearing on different documents. This may cause a problem while clustering documents that have less terms in common, but are semantically related.

Most of the existing work on the semantic similarity is for calculating similarities between words [6][7][8]. In contrast to word semantic similarity, calculation of text or document semantic similarity is less investigated. At first glance, it seems to make sense to extend the semantic word similarity to the semantic document similarity. In this case, to calculate semantic similarities for a set of documents, each document can be represented as a feature vector of words (terms) that appear in the entire document set. For each pair of those terms, semantic similarities can be calculated using a semantic similarity measure. The semantic similarities between documents can then be calculated based on these pairwise similarity values of terms. However, we found that although semantic similarity measures improve the clustering results in terms of cluster validity indices, the main weakness of these methods are their high complexity [9].

In this paper, we propose a new method for effectively calculating semantic similarities between documents. Our proposed method has low computational complexity and gives comparable performance in document clustering. It combines a taxonomy structure with cosine similarity measure so that the semantic (dis)similarity between documents can be better quantified with a low computational complexity. The method inherits the edge-counting scheme in a taxonomy to calculate the semantic similarity. The taxonomy used in this paper is constructed by maintaining only IS-A relations between nodes in a taxonomy such as the WordNet [10]. Each node in this taxonomy is associated with the terms. Based on the experimental results in [11], the assessments of similarity in semantic networks can be thought of as involving just taxonomic (IS-A) links as Rada et al. [12] suggest.

Our overall approach can be summarized as follows. Each document in a document set is represented as a feature vector of terms that appear in the entire document set. For every word in the document set, the corresponding node is found in the taxonomy to construct the vector of concepts of the word by following IS-A relations where each concept in this vector is a node on the path to the root of the taxonomy. Each

978-1-4244-5023-7/09/$25.00 ©2009 IEEE

ISCIS September 14-16, 2009
METU
Northern Cyprus Campus

document is then represented with the terms appearing on this document and concept vectors corresponding to these terms. A rank-order weighting scheme is employed for the concept vector in order to perform edge-counting method. Finally, the pairwise similarities between documents are calculated using cosine similarity measure. The documents are then clustered based on the calculated similarities. The performance of our similarity measure is evaluated by the quality of clusters produced using that metric. Experimental results show that our similarity measure outperforms cosine similarity in terms of cluster validity indices while time complexity is the same for both measures.

The rest of the paper is organized as follows. In Section II we introduce the related work and in Section III we give the background of the problem to be solved. Section IV presents our proposed semantic similarity measure. Section V provides detailed experimental results. Finally, in Section VI we conclude and discuss future work.

II. Related Work

Semantic similarity is studied in the context of word similarity. There are mainly three approaches to determine the semantic similarity between terms. The first approach relies on a large corpus of text and the term similarities are often derived from their co-occurrence distribution in this corpus [13][14]. The second approach calculates the semantic similarity based on the conceptual distance in a taxonomy such as the WordNet [10]. The third approach is a hybrid method which combines the first two approaches. A detailed review on term similarity can be found in [13][15]. As stated in [9][16], semantic similarity between documents has been less studied compared to semantic similarity between terms.

There have been several proposals to extend the semantic word similarity to the semantic document similarity by calculating first pairwise term similarities appearing on different documents [9][17][18]. In [11], a new semantic similarity measure was introduced. The experiments in that study showed that using all the relations in semantic network during similarity calculation in a taxonomy does not increase the clustering quality in terms of cluster validity indices. An experimental study was also conducted to examine the effects of the proposed measure and single term similarity measures to the clustering results of Turkish documents. It is found in that study that the major drawback of semantic similarity approaches is that they are computationally expensive.

In several studies for information retrieval, a variant of the vector space model is used to calculate similarities between documents and queries. Latent Semantic Indexing (LSI) is a variant of the vector space model which converts a high dimensional space into a low dimensional space [19]. The aim of LSI is to construct a semantic space using singular value decomposition (SVD) wherein documents that share many associated words with a keyword are placed near one another. However, for large-scale data sets, the computing and storage costs of SVD may be prohibitive.

III. Background

In this section, we first overview several similarity measures for text documents. Given a set $\mathcal{D}$ of documents the aim is to calculate pairwise similarities between each pair of documents in $\mathcal{D}$. Based on the vector space model, each document $d_i \in \mathcal{D}$ can be represented by a vector of frequencies of terms/words:

$$\vec{d_i} = \{(c_1, w_{i1}), (c_2, w_{i2}), \ldots, (c_n, w_{in})\} \quad (1)$$

where $c_1, c_2 \ldots, c_3$ are words that appear in $\mathcal{D}$ and w_{ij} is the weight of word c_j in a document d_i. The weight w_{ij} can be represented either by Term Frequency (TF) or Term Frequency-Inverse Document Frequency (TF-IDF). TF is calculated simply by dividing the number of times that the word c_j appears in document d_i by the total number of words in d_i. TF-IDF is equal to TF multiplied by IDF$= \log \frac{n_j}{n}$, where n_j is the number of times that the word c_j appears in the whole document set $\mathcal{D}$, and n is total number of words present in that set.

A. Cosine Similarity

Cosine similarity measures the angle between two vectors and is calculated by dividing the inner product of two vectors by multiplication of vectors' length. The cosine similarity between documents d_i and d_j is formulated as follows:

$$sim_{cos}(d_i, d_j) = \frac{\overrightarrow{d}_i \bullet \overrightarrow{d}_j}{\| \overrightarrow{d}_i \| \cdot \| \overrightarrow{d}_j \|} \quad (2)$$

$$= \frac{\sum_{k=1}^{n} w_{ik} \times w_{jk}}{\sqrt{\sum_{k=1}^{n} w_{ik}^2} \times \sqrt{\sum_{k=1}^{n} w_{jk}^2}} \quad (3)$$

where $\bullet$ denotes the vector dot product and $\| \ \|$ is the length of a vector. The cosine similarity values range between $[0, 1]$, where a cosine similarity value of 0 means that the documents are unrelated and a cosine similarity value close to 1 means that the documents are closely related. It is obvious, that in order to have cosine similarity greater than 0, documents should have some words in common. When all of the words and their associated weights are same in two different documents, then the cosine similarity between these two documents is equal to 1.

The time complexity of cosine similarity calculation is $O(N) = N^2 \cdot c$ where N is the number of documents in a document set and c is the number of words in the document set. For $c << N$, the complexity of the cosine similarity metric can be generalized to $O(N) = N^2$.

B. Semantic Similarity

Most of the existing semantic similarity calculations use WordNet [10] ontology and inter-word relations defined in it. Hypernym/hyponym (IS-A) type relation accounts for approximately 80% of those relations [20]. Therefore, most of the semantic similarity metrics are based on path traversal in a semantic network involving just IS-A relations between words. In this study, a taxonomy of IS-A relationships of words is constructed from BalkaNet [21] which is a multilingual lexical

database comprising of individual WordNets for the Balkan languages, including Turkish. Fig. 1 illustrates a fragment of this taxonomy which is mapped to the Princeton WordNet, where lines represent hierarchical IS-A links between words (concepts).

Wu-Palmer similarity is one of the word semantic similarity measures using IS-A type relations of a taxonomy [6]. The principle of this similarity computation is based on the edge counting method. In [13], it is stated that the Wu-Palmer similarity measure is fast to compute, and is arguably as good as the others (see [13] for details). The Wu-Palmer similarity between a pair of concepts c_1 and c_2 is measured as:

$$sim_{W\&P}(c_1, c_2) = \frac{2 \times N_3}{2 \times N_3 + N_1 + N_2} \tag{4}$$

where N_1 and N_2 are the number of IS-A links from c_1 and c_2 to their most specific common superclass C ($lso(c_1, c_2)$), N_3 is the number of IS-A links from C to the root of taxonomy. The Wu-Palmer similarity of $apple$ and $pear$, $sim_{W\&P}(apple, pear)$, is equal to 0.83, where $N_3 = 5$, $N_1 = 1$ and $N_2 = 1$ (Fig. 1). In the same way $sim_{W\&P}(apple, vegetables)$ is equal to 0.73.

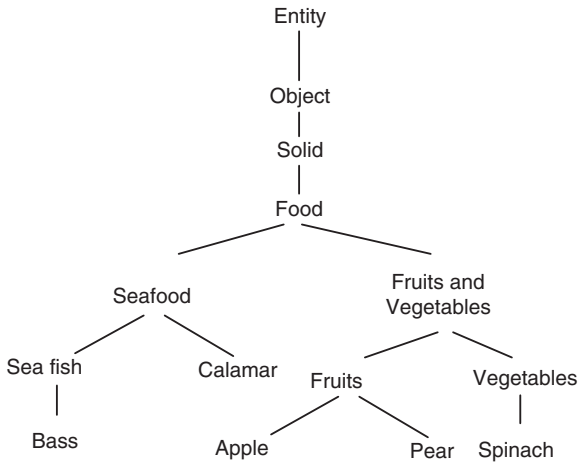

Fig. 1. A Fragment of BalkaNet. Solid lines represent IS-A links

There exist several methods to calculate semantic similarities between documents. In [9], the methods proposed in [11] and [17] were examined and compared via the clustering results they produced. According to that study, document similarity measure formulated in Eq. 5 performed better than others when used as a similarity measure for clustering documents. It is called ***SEMSIM*** in this paper. The semantic similarity between documents d_i and d_j is calculated as:

$$sim_{SEMSIM}(d_i, d_j) = \\ = \frac{\sum_{r=1}^{n} \sum_{u=1}^{n} w_{ir} \times w_{ju} \times sim_{W\&P}(c_r, c_u)}{\sqrt{\sum_{k=1}^{n} w_{ik}^2 \times \sum_{k=1}^{n} w_{jk}^2}} \tag{5}$$

where $sim_{W\&P}(c_r, c_u)$ refers to the Wu-Palmer similarity given in Eq. 4. When using ***SEMSIM***, the complexity of the pairwise similarity calculation in a document set is $O(N) = N^2 \cdot c^2 \cdot d$, where N and c are the number of documents and words in the document set respectively and d is the depth of the taxonomy (i.e. the number of taxonomic levels).

IV. Proposed Method

As have been shown in Section III, one of the main advantage of the cosine similarity metric over the document semantic similarity measures is its low time complexity. In order to decrease the computation time of document semantic similarity calculations, we propose a new method, that extends document vectors using parent nodes from the IS-A taxonomy and then computes cosine similarity between newly formed vectors.

Definition 1: (Parent Vector) A vector $\vec{P_{c_i}} = (c_i, p_{i1}, p_{i2}, \ldots, p_{in})$ is called parent vector of term c_i if p_{i1} is a parent node of c_i and there exist a chain of terms $p_{i1}, p_{i2}, \ldots, p_{in}$ in a taxonomy, connected by direct IS-A relation links, where term $p_{i(j+1)}$ is a parent node of the term p_{ij}.

Parent vectors corresponding to terms in a document are constructed for each document. Each term of each parent vector in a document are weighted using an appropriate weighting scheme. The calculation of these weights consists of the following steps:

1) Extract the parent vector $\vec{Pc_j}$ of the word c_j that appear on a document d_i.
2) Take only first k closest parents where $k \leq 10$.
3) The weight of the m.th parent of c_i is calculated by $pw_{jm} = w_{ij} \times (10 - m) \times 0.1$, where w_{ij} is the weight of c_j in document d_i as in Eq. 1.

Definition 2: (Concept Vector) Let document d_i be represented as shown in Eq. 1. Then the *concept vector* of d_i is formed by merging all parent vectors corresponding to terms appeared in d_i. That is, $\vec{concept\_d_i} = \vec{P_{c1}}, \vec{P_{c2}}, \ldots, \vec{P_{cn}}$

A concept vector, $\vec{concept\_d_i}$, is formed for each document d_i in the document set to represent the documents. If the parent vectors of a document share some terms in common, then each of these terms takes part only once in the $\vec{concept\_d_i}$ and its weight in that vector is updated by summing up its weight values in the corresponding parent vectors. The semantic similarity between documents can be then calculated as follows:

$$sim_{CT}(d_i, d_j) = \frac{\vec{concept\_d_i} \bullet \vec{concept\_d_j}}{\|\vec{concept\_d_i}\| \cdot \|\vec{concept\_d_j}\|} \tag{6}$$

The aim of selecting only first k nodes is based on the fact that as we go up in the taxonomy, the generality of concepts increases. In order to obtain only meaningful relations representing human cognitive limitations, the length of the parent vector is limited.

Our proposed semantic similarity formulation using concept vectors is inspired by Wu-Palmer similarity measure. The Wu-Palmer semantic similarity between two concepts c_i and c_j depends on the position of $lso(c_i, c_j)$ with respect to concepts c_i, c_j and the taxonomy root. It is maximized when $lso(c_i, c_j)$ is close to the concepts, namely where N_3 is high and $N1$,

$N2$ are low; and minimized otherwise. Our proposed semantic similarity measure behaves in the same way. Let d_1 and d_2 be two documents consisting of only one term, c_1 and c_2 respectively. If the minimal path of IS-A links between these two terms in the taxonomy is long, then the N_3 value is low, that means $lso(c_1, c_2)$ is high and the number of terms between $lso(c_1, c_2)$ and the root of the taxonomy is small. This results in a small value of Wu-Plamer similarity between these two terms. In this case our proposed semantic similarity calculation leads to a small value either, since the concept vectors of two documents share only the words that are between $lso(c_1, c_2)$ and the root of the taxonomy. It is obvious that the cosine similarity between concept vectors of documents that share less words in common is small.

Example 1. Let us consider three documents d_i , d_j and d_l with vector representations $\vec{d_i} = (apple, 0.30), (vegetables, 0.20)$, $\vec{d_j} = (pear, 0.30), (spinach, 0.20)$ and $\vec{d_l} = (bass, 0.30), (calamar, 0.20)$ respectively. By setting $k = 5$ the parent vectors of d_i can be calculated as

$P_{apple,0.30}$={(apple, 0.30),
(fruits, 0.27),
(fruits and vegetables, 0.24),
(food, 0.21),
(solid, 0.18),
(object, 0.15)}

and

$P_{vegetables,0.20}$={(vegetables, 0.20),
fruits and vegetables, 0.18),
(food, 0.16),
(solid, 0.14),
(object, 0.12),
(entity, 0.10)}.

So the concept vector $\vec{concept}\_d_i$ of document d_i is equal to

$\vec{concept}\_d_i$ ={(apple, 0.30),
(vegetables, 0.20),
(fruits, 0.27),
(fruits and vegetables, 0.42),
(food, 0.37),
(solid, 0.32),
(object, 0.27),
(entity, 0.10)}

In the same manner concept vectors $\vec{concept}\_d_j$ and $\vec{concept}\_d_l$ can be constructed. It is obvious, that $sim_{cos}(d_i, d_j) = 0.0$ and $sim_{cos}(d_i, d_l) = 0.0$ as there are no common words between these documents. However these documents are highly semantically related. The semantic similarities calculated with our proposed method are $sim_{CT}(d_i, d_j) = 0.62$ and $sim_{CT}(d_i, d_l) = 0.46$. The semantic similarity between documents using Eq. 5 is equal to $sim_{SEMSIM}(d_i, d_j) = 1.50$ and $sim_{CT}(d_i, d_l) = 1.04$. As can be seen from calculation results, the proposed similarity metric decreases when $lso(c_i, c_j)$ goes up the taxonomy.

V. Experimental Results

The experiments are performed on three different document sets, containing the data collected from the web. The first data set (*Dataset1*) contains 2382 documents and categorized manually into seven clusters by an expert[1]. The second document set (*Dataset2*) contains 481 documents. It is retrieved manually from the web by setting predefined 5 topics and querying a major search engine. The third data set (*Dataset3*) is collected automatically from the web site of a Turkish Internet Service Company[2]. It consists of 1987 documents in which the groups of documents are unknown. The properties of these document sets are given in Table I.

TABLE I
Properties of the document sets

	Number of	
	documents	classes
Dataset1	2382	7
Dataset2	481	5
Dataset3	1987	-

All document sets pass through the same preprocessing stage. First of all, each web page is parsed to remove HTML tagging and is tokenized into individual terms. A morphological analyzer [22] and postagger [23] that are developed for Turkish are used to transform all terms into the most probable stem terms. The stop words defined for the Turkish language are removed from the documents. In our previous work [9], a big amount of words present in documents could not be found in the BalkaNet taxonomy. Therefore the semantic similarity between those words can not be calculated. Since the aim of this study is to develop a semantic similarity for documents, document sets were modified in a way, that every document are represented by words that exist in BalkaNet; all words that could not be found in the BalkaNet tree are removed. In addition to these, only nouns are selected in order to obtain meaningful relations between documents.

To reduce the vector space and make calculations more efficient and fast, 10 most frequent words (MFW) are empirically selected to represent every document. This number is chosen proportionally to the number of words, remained in documents after the preprocessing stage. All terms are weighted by normalized TF. Parent vectors for MFW are constructed in a way described in Section IV and parent weights are assigned accordingly. Since the words could have more than one senses, a sense disambiguation phase is performed to select the real meaning of the MFW and to fix the document vectors. The selection of the correct sense s for a word w is done in a way that semantic similarity between s and all other MFW

[1]Dr. A. Cüneyd Tantuğ: a member of the Natural Language Processing Group of Department of Computer Engineering, Istanbul Technical University (http://ddi.ce.itu.edu.tr)

[2]http://www.mynet.com.tr

TABLE II
SILHOUETTE AND DAVIES-BOULDIN INDICES OF DATASET1 FOR DIFFERENT NUMBER OF CLUSTERS

Similarity Measure	Silhouette number of clusters				Davies-Bouldin number of clusters			
	5	7	10	15	5	7	10	15
sim_{cos}	0.03	0.05	0.05	0.07	1.93	1.91	1.90	1.86
sim_{SEMSIM}	0.11	0.09	0.07	0.06	**1.71**	1.74	1.82	1.81
sim_{CT}	0.08	**0.14**	**0.14**	**0.14**	1.83	1.75	**1.71**	1.72

TABLE III
SILHOUETTE AND DAVIES-BOULDIN INDICES OF DATASET2 FOR DIFFERENT NUMBER OF CLUSTERS

Similarity Measure	Silhouette number of clusters				Davies-Bouldin number of clusters			
	5	7	10	15	5	7	10	15
sim_{cos}	0.15	0.14	0.17	0.18	1.75	1.73	1.70	1.67
sim_{SEMSIM}	0.21	0.17	0.15	0.15	1.58	1.59	1.64	1.60
sim_{CT}	0.24	0.23	0.24	**0.30**	1.57	1.53	1.50	**1.41**

TABLE IV
SILHOUETTE AND DAVIES-BOULDIN INDICES OF DATASET3 FOR DIFFERENT NUMBER OF CLUSTERS

Similarity Measure	Silhouette number of clusters				Davies-Bouldin number of clusters			
	5	7	10	15	5	7	10	15
sim_{cos}	**0.24**	0.16	0.17	0.15	1.65	1.74	1.71	1.72
sim_{SEMSIM}	0.11	0.06	0.08	0.05	1.73	1.76	1.74	1.71
sim_{CT}	0.10	0.15	**0.20**	0.08	1.68	1.63	**1.56**	1.68

is maximum among other senses of w. Since a word sense disambiguation is not the aim of the study, we will not discuss the effectiveness of this method, saying only that it performed well comparing with the results obtained without using the word sense disambiguation method.

Cluto software package[3] is used for document clustering. To compare clustering results *Davies-Bouldin* and *Silhouette* cluster validity indices are used. Davies-Bouldin index is a function of the ratio of sum of intra-cluster dispersion to inter-cluster separation. A lower value of Davies-Bouldin index indicates a good clustering solution. Silhouette index takes into account the compactness of the resulting clusters and separation between them. It takes values between -1 and 1, where greater values of Silhouette index means a better clustering solution. We set the same number of clusters for all data sets. In experimental tables cosine similarity is represented as sim_{cos}, proposed semantic similarity as sim_{CT} and document semantic similarity described in Eq. 5 as sim_{SEMSIM}. We omit setting k value, taking the whole parent vector to provide fully overlapping with Wu-Palmer similarity metric.

As can be seen from the results in Tables II, III and IV, proposed document semantic similarity measure outperforms both cosine and previously proposed similarity metric in terms of clustering validity indices. Only for $Dataset3$ the cosine similarity ends with a slightly better Silhouette index whereas our proposed similarity metric yields a clustering result with a better Davies-Bouldin index value.

As mentioned in Section IV one of the main drawbacks of document semantic similarity calculation is the running time of algorithm. As parent vectors of each documents are calculated only once and before inter-document similarity calculations, the running time of our proposed method is $O(N) = N^2 \cdot c + N \cdot c \cdot d$ or more precise $O(N) = N^2 \cdot c$ where N is number of documents used in calculation, c number of words in each document and d is a overall WordNet tree depth. For $c << N$, the complexity of the proposed semantic similarity metric is $O(N) = N^2$. Thus, the time complexity is the same with cosine similarity calculation.

Example 2. To compare clustering results obtained using cosine similarity measure and our proposed semantic similarity metric, an example document set is manually constructed (Table V). This document set consist of only 9 documents. Each document is represented by equally weighted five terms. Documents are grouped under three topics, which are "fruits and vegetables" (d_1, d_2, d_3), "medicine" (d_4, d_5, d_6) and "animals" (d_7, d_8, d_9). The clustering results when using cosine similarity measure and our proposed semantic similarity metric are illustrated in Table VI. The clusters obtained with our proposed similarity match exactly with the predefined groups.

[3]http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview

TABLE V
EXAMPLE: SMALL DOCUMENT SET

d_1	jam:0.2, cherry:0.2, apple:0.2, tree:0.2, pear:0.2
d_2	potato:0.2, tomato: 0.2, garden:0.2, home:0.2,soil:0.2
d_3	vegetables:0.2, tomato:0.2, spinach:0.2, orange:0.2, grass:0.2
d_4	doctor:0.2, clinic:0.2, pills:0.2, pacient:0.2, ache:0.2
d_5	pharmacy:0.2, clinic: 0.2, pills:0.2, nurse:0.2, room:0.2
d_6	influenza:0.2, home:0.2, illness:0.2, ache:0.2, ambulance:0.2
d_7	cat:0.2, dog:0.2, home:0.2, pet:0.2, offspring:0.2
d_8	fish:0.2, aquarium: 0.2, dog:0.2, home:0.2, sheep:0.2
d_9	bear:0.2, soil:0.2, forest:0.2, crocodile:0.2, river:0.2

TABLE VI
CLUSTERING RESULTS USING COSINE AND PROPOSED SEMANTIC SIMILARITY MEASURES

	Cosine Similarity	Proposed Similarity
cluster 1	d_1, d_2, d_3, d_9	d_1, d_2, d_3
cluster 2	d_4, d_5	d_4, d_5, d_6
cluster 3	d_6, d_7, d_8	d_7, d_8, d_9

VI. CONCLUSION AND FUTURE WORK

Recent growth in both online and offline data increases the requirements in precise data organization. Moreover, time spent for these calculations becomes a crucial point due to the enlargement of processed data. Many different similarity metrics were proposed and compared. Most of them are based on word pairwise semantic similarity calculation, few on inter-document relatedness. The calculation of semantic similarity between words does not consumes much time, thus time complexity generally is not considered in those studies. But the time becomes really important when the amount of inter-word similarity calculation increases, as in calculating inter-document similarities.

In this paper we present a new document semantic similarity calculation method that combines cosine similarity measure with simple extended vectors, obtained from IS-A taxonomy. The effect of the proposed method on clustering of Turkish documents is studied and compared with the cosine similarity measure. The results of experiments are presented in terms of clustering validity indices. As those results show, our proposed method outperforms both cosine similarity measure. The time complexity of the proposed method is same with the cosine similarity measure.

Further research can be done on preprocessing stage for word sense disambiguation. Moreover, the effects of the proposed method on text summarization and identification of cluster topics can be investigated.

REFERENCES

[1] R. Saraçoğlu, K. Tütüncü, and N. Allahverdi, "A fuzzy clustering approach for finding similar documents using a novel similarity measure," *Expert Systems with Applications*, vol. 33, no. 3, pp. 600–605, 2007.

[2] H. X. W. Wu and S. Shekhar, Eds., *Clustering and Information Retrieval*. Kluwer, 2003.

[3] K. Bade and A. Nurnberger, "Personalized hierarchical clustering," in *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*. Washington, DC, USA: IEEE Computer Society, 2006, pp. 181–187.

[4] Z. Jiang, A. Joshi, R. Krishnapuram, and L. Yi, "Retriever: Improving web search engine results using clustering," University of Maryland Baltimore County, Tech. Rep., 2000.

[5] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., 1986.

[6] Z. Wu and M. Palmer, "Verb semantics and lexical selection," in *32nd. Annual Meeting of the Association for Computational Linguistics*, 1994, pp. 133 –138. [Online]. Available: citeseer.ist.psu.edu/wu94verb.html

[7] C. Leacock and M. Chodorow, "Combining local context and wordnet similarity for word sense identification," *An Electronic Lexical Database*, pp. 265–283, 1998.

[8] A. Pandya and P. Bhattacharyya, "Text similarity measurement using concept representation of texts." in *Pattern Recognition and Machine Intelligence*, ser. Lecture Notes in Computer Science, S. K. Pal, S. Bandyopadhyay, and S. Biswas, Eds., vol. 3776. Springer, 2005, pp. 678–683.

[9] A. Madylova and S. G. Oguducu, "Comparison of similarity measures for clustering turkish documents," *Intelligent Data Analysis*, vol. 13, no. 5, p. in press, 2009.

[10] C. Fellbaum, Ed., *Wordnet: An Electronic Lexical Database*. MIT Press, 1998.

[11] B. Yucesoy and S. G. Oguducu, "Comparison of semantic and single term similarity measures for clustering turkish documents," in *ICMLA '07: Proceedings of the Sixth International Conference on Machine Learning and Applications*. Washington, DC, USA: IEEE Computer Society, 2007, pp. 393–398.

[12] R. Rada, H. Mili, E. Bicknell, and M. Blettner, "Development and application of a metric on semantic nets," in *IEEE Transactions on Systems, Man and Cybernetics*, 1989, pp. 17–30.

[13] D. Lin, "An information-theoretic definition of similarity," in *Proc. 15th International Conf. on Machine Learning*. Morgan Kaufmann, San Francisco, CA, 1998, pp. 296–304. [Online]. Available: citeseer.ist.psu.edu/95071.html

[14] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *In Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1995, pp. 448–453.

[15] Y. Li, Z. A. Bandar, and D. McLean, "An approach for measuring semantic similarity between words using multiple information sources," *IEEE Trans. on Knowl. and Data Eng.*, vol. 15, no. 4, pp. 871–882, 2003.

[16] C. Corley and R. Mihalcea, "Measuring the semantic similarity of texts," in *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, Ann Arbor, Michigan, June 2005, pp. 13–18.

[17] M. Halkidi, B. Nguyen, and I. Varlamis, "Thesus: Organizing web document collections based on link semantics," *VLDB J*, vol. 12, pp. 320–332, 2003.

[18] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," in *In AAAI06*, 2006, pp. 775–780. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.65.3690

[19] C. A. Kumar and S. Srinivas, "Latent semantic indexing using eigenvalue analysis for efficient information retrieval," *International Journal of Applied Mathematics and Computer Science*, vol. 16, no. 4, pp. 551–558, 2006.

[20] A. Budanitsky and G. Hirst, "Evaluating wordnet-based measures of lexical semantic relatedness," *Computational Linguistics*, vol. 32, pp. 13–47, 2006.

[21] B. Project, http://www.ceid.upatras.gr/Balkanet/.

[22] K. Oflazer, "Two-level description of Turkish morphology," *Literary and Linguistic Computing*, vol. 9, no. 2, pp. 137–148, 1994.

[23] D. Yüret and F. Türe, "Learning morphological disambiguation rules for Turkish," in *Proceedings of the Human Language Technology conference and North American chapter of the Association for Computational Linguistics annual meeting (HLT-NAACL)*, New York, NY, 2006, pp. 328–334.