İSTANBUL TECHNICAL UNIVERSITY ★ INSTITUTE OF SCIENCE AND TECHNOLOGY

A TAXONOMY BASED SEMANTIC SIMILARITY OF DOCUMENTS USING THE COSINE MEASURE

M.Sc. Thesis by
Ainura Madylova

Department : **COMPUTER ENGINEERING**

Programme : **COMPUTER ENGINEERING**

JUNE 2009

İSTANBUL TECHNICAL UNIVERSITY ★ INSTITUTE OF SCIENCE AND TECHNOLOGY

A TAXONOMY BASED SEMANTIC SIMILARITY OF DOCUMENTS USING THE COSINE MEASURE

M.Sc. Thesis by
Ainura Madylova
(504071543)

Date of Submission : **04 May 2009**
Date of defence examination : **04 June 2009**

Supervisor : **Asst. Prof. Dr. Şule GÜNDÜZ ÖĞÜDÜCÜ**

Members of the Examining Committee : **Asst. Prof. Dr. Gülşen CEBİROĞLU ERYİĞİT (İ.T.Ü.)**

Asst. Prof. Dr. Banu DİRİ (Y.T.Ü.)

JUNE 2009

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLER ENSTİTÜSÜ

KOSİNÜS BENZERLİĞİNİ KULLANARAK BELGELER ARASI ANLAMSAL BENZERLİĞİ KAVRAMSAL SÖZLÜĞE DAYALI HESAPLAMA YÖNTEMİ

YÜKSEK LİSANS TEZİ

Ainura Madylova
(504071543)

Tezin Enstitüye Verildiği Tarih : 04 Mayıs 2009
Tezin Savunulduğu Tarih : 04 Haziran 2009

Tez Danışmanı : Asst. Prof. Dr. Şule GÜNDÜZ ÖĞÜDÜCÜ

Diğer Jüri Üyeleri : Asst. Prof. Dr. Gülşen CEBİROĞLU ERYİĞİT (İ.T.Ü.)

Asst. Prof. Dr. Banu DİRİ (Y.T.Ü.)

HAZİRAN 2009

FOREWORD

First of all, I would like to thank my advisor, Asst. Prof. Dr. Şule GÜNDÜZ ÖĞÜDÜCÜ, for her help and assistance throughout my education in ITU.

I would also like to thank Büşra AVCI and Sevilay ALKILIÇ for helping me on writing the Turkish part of this Thesis.

I would not be who I am today without the love and support of my family. Thank you for always being there.

June 2009

Ainura Madylova

Computer Engineer

TABLE OF CONTENTS

	Page
ABBREVIATIONS	**ix**
LIST OF FIGURES	**xi**
LIST OF TABLES	**xiii**
SUMMARY	**xvi**
ÖZET	**xviii**
1. INTRODUCTION	**1**
1.1. Problem Definition	1
1.2. Contribution of the Thesis	3
1.3. Organization of the Thesis	4
2. WORD THESAURUS	**5**
2.1. WordNets	5
2.2. Corpus	7
3. SEMANTIC SIMILARITIES BETWEEN WORDS	**9**
3.1. Corpus-based Semantic Similarity Metrics	9
3.1.1. LSA Semantic Similarity	10
3.2. Taxonomy-based Semantic Similarity Metrics	10
3.2.1. Wu-Palmer	11
3.2.2. Modified Wu-Palmer	11
3.2.3. Wan & Angryk	12
3.3. Hybrid Methods	12
3.3.1. Resnik	13
3.3.2. Jiang & Conrath	13
4. DOCUMENT SIMILARITIES	**15**
4.1. Document Semantic Similarity Metrics	16
4.1.1. SEMSIM document semantic similarity	16
4.1.2. THESUS	16
4.2. Single Term Similarity Metrics	17
4.2.1. Cosine similarity measure	18
4.2.2. Jaccard coefficient	18
5. PROPOSED DOCUMENT SEMANTIC SIMILARITY METHOD	**21**
6. EXPERIMENTS AND DISCUSSIONS	**27**
6.1. Document Sets and Text Document preprocessing	28
6.2. Cluster Validity Indices	29
6.2.1. Unsupervised cluster validity indices	30
6.2.1.1. Davies-Bouldin index	30
6.2.1.2. Silhouette index	31
6.2.2. Supervised cluster validity indices	31

6.2.2.1. Entropy index 31
6.2.2.2. Purity index 32
6.2.2.3. Jaccard index 32
6.3. Experiments 33
6.3.1. First experiment set 33
6.3.2. Second experiment set 40
6.4. A Walk-Through Example 42
7. CONCLUSION AND FUTURE WORK 45
REFERENCES 47
CURRICULUM VITA 51

ABBREVIATIONS

LSA	:	Latent Semantic Analysis
SVD	:	Single Vector Decomposition
TF	:	Term Frequency
IDF	:	Inverse Document Frequency
WSD	:	Word Sense Disambiguation
MFW	:	Most Frequent Words
DB	:	Davies-Bouldin
E	:	Entropy
P	:	Purity
J	:	Jaccard

LIST OF TABLES

Page

Table 2.1 : Distribution Parts of Speech of Turkish WordNet 6
Table 2.2 : Semantic Relations in Turkish BalkaNet 7
Table 6.1 : Properties of the document sets 28
Table 6.2 : The results of manually clustered documents of Dataset1 and Dataset2 34
Table 6.3 : Silhouette and Davies-Bouldin indices of Dataset1 for different number of clusters 35
Table 6.4 : Silhouette and Davies-Bouldin indices of Dataset2 for different number of clusters 35
Table 6.5 : Silhouette and Davies-Bouldin indices of Dataset3 for different number of clusters 36
Table 6.6 : Silhouette and Davies-Bouldin index of Dataset2 using all terms on the documents 36
Table 6.7 : The results of manually clustered documents of Dataset2 with Hybrid Corpus-Based Semantic Similarity metric 37
Table 6.8 : Silhouette and Davies-Bouldin indices of Dataset2 for different number of clusters with Hybrid Corpus-Based Semantic Similarity metric 37
Table 6.9 : The content of the clusters obtained using cosine similarity measure 38
Table 6.10: The content of the clusters obtained using the SEMSIM based on Wu-Palmer semantic similarity measure 38
Table 6.11: The results of manually clustered documents of Dataset1 and Dataset2 using LSA 39
Table 6.12: Silhouette and Davies-Bouldin indices of Dataset1 for different number of clusters using LSA 39
Table 6.13: Silhouette and Davies-Bouldin indices of Dataset2 for different number of clusters using LSA 39
Table 6.14: Silhouette and Davies-Bouldin indices of Dataset3 for different number of clusters using LSA 40
Table 6.15: The results of manually clustered documents of Dataset1 and Dataset2 41
Table 6.16: Silhouette and Davies-Bouldin indices of Dataset1 for different number of clusters 41
Table 6.17: Silhouette and Davies-Bouldin indices of Dataset2 for different number of clusters 42
Table 6.18: Silhouette and Davies-Bouldin indices of Dataset3 for different number of clusters 42
Table 6.19: Example: Small Document Set 43

Table 6.20: Clustering results using cosine and proposed semantic similarity measures 43

LIST OF FIGURES

Page

Figure 2.1: A Fragment of BalkaNet. Solid arrows represent IS-A links . . 6
Figure 5.1: Parent Vector of the term *apple* in a BalkaNet 22
Figure 6.1: Similarity matrix constructed for clustering 27

A TAXONOMY BASED SEMANTIC SIMILARITY OF DOCUMENTS USING THE COSINE MEASURE

SUMMARY

In this thesis, different document similarity measures are compared and a new method is proposed for calculating document semantic similarity of Turkish documents. This method is based on concept vectors, extracted from the large taxonomy, where different semantic relations between words are defined. The effects of the semantic and single term similarity metrics on the clustering of Turkish documents are studied and compared in terms of clustering validity indices.

Semantic similarities between two words differs in the thesaurus type used for calculation. According to the thesaurus, they can be grouped into three categories: Corpus-based, Taxonomy-based and Hybrid semantic similarity measures. Corpus-based semantic similarity measures are calculated using some pre-defined corpus, which contains a large number of words. This type of similarity measures are generally computed according to co-occurrence of given words in that corpus. Taxonomy-based semantic similarity metrics are calculated using a taxonomy, where the different semantic relations are defined between words. Hybrid semantic similarity measures combines the previous two types and make use of both a corpus and a taxonomy thesaurus. In this study, we focus on two Taxonomy-based semantic similarity measures which are Wu-Palmer and Modified Wu-Palmer semantic similarity metrics. These similarity metrics are based on edge-counting method over the IS-A relations of the given taxonomy. In addition to these, a corpus-based document semantic similarity measure based on the Latent Semantic Analysis (LSA) is examined, which uses document sets as a predefined corpus. One hybrid method proposed by Jaing and Conrath is also examined and compared with the Wu-Palmer and Modified Wu-Palmer similarity metrics.

Semantic similarities between documents are generally calculated using the combinations of semantic similarities between the words present in them. Different linear combination of word semantic similarities produces different document semantic similarity values. In this study we examined two document semantic similarity measures which are THESUS and SEMSIM. As underlying word semantic similarity metrics Wu-Palmer, Modified Wu-Palmer and Jaing-Conrath similarity measures are used. Moreover, single term similarity measure like cosine and Jaccard similarities are described and discussed. The LSA method is also applied to documents and similarity is then calculated using cosine similarity measure.

The document semantic similarity proposed in this study differs from existing semantic similarity measures. The word pairwise semantic similarities are not

used in calculation. Instead, concept vectors, produced using the taxonomy tree for each compared document, are used. This vectors then passed to cosine similarity calculations. Thus the proposed method is a combination of document semantics with cosine similarity measure. Besides, the proposed semantic similarity method outperforms existing document semantic similarity methods based on word pairwise similarity calculations in terms of low time complexity.

To compare the similarity metrics listed above we cluster Turkish documents using one of the described methods and record the clustering results. Comparison of clustering results is done in terms of five clustering validity indices: Entropy, Purity, Jaccard, Silhouette and Davies-Bouldin indices. Entropy, Purity and Jaccard indices use actual cluster labels of the document sets and measures the general "purity" of the clusters. Silhouette and Davies-Bouldin do not require class information of the documents and measures the quality of the produced clusters. Experiments are concentrated on document semantic similarity measures which include pairwise word semantic similarity calculations. Experimental results show that the proposed method outperforms other semantic similarity measures and single term similarities in terms of clustering quality while having the low competition time.

KOSİNÜS BENZERLİĞİNİ KULLANARAK BELGELER ARASI ANLAMSAL BENZERLİĞİ KAVRAMSAL SÖZLÜĞE DAYALI HESAPLAMA YÖNTEMİ

ÖZET

Bu tezde, Türkçe belgeler arasında anlamsal benzerlik hesaplamak için yeni bir yöntem önerilmektedir. Bu yöntem, kelimeler arası farklı anlamsal bağlantıların tanımlandığı kavramsal sözlükten çıkarılan kavram vektörlerine dayanmaktadır. Çalışma kapsamında anlamsal benzerlik ve sözcük benzerliği hesaplama yöntemlerinin Türkçe belgelerin demetlenmesi üzerindeki etkileri incelenmekte ve bu yöntemler demetleme göstergeleri aracılığıyla karşılaştırılmaktadır.

İki sözcük arasında anlamsal benzerlik hesaplama yöntemleri, kullandıkları kavramsal sözlüğe göre üçe ayırırlar: Derlem tabanlı, Kavramsal sözlük tabanlı ve Karma anlamsal benzerlik hesaplama yöntemleri. Derlem tabanlı anlamsal benzerlikler önceden tanımlanmış, çok sayıda sözcük içeren bir derlem kullanarak hesaplanmaktadır. Bu yöntemde benzerlik ölçütü genellikle verilen sözcüklerin derlemde yer alma sayısına göre hesaplanır. Kavramsal sözlük tabanlı yöntemlerde sözcükler arasında anlamsal benzerliklerin tanımlandığı bir kavramsal sözlük kullanılır. Karma anlamsal benzerlik hesaplama yöntemleri derlem ve kavramsal sözlükleri birlikte kullanarak önceki iki yöntemi birleştirir. Bu çalısmada iki farklı kavramsal sözlük tabanlı yöntem incelenmektedir: Wu-Palmer ve Modified Wu-Palmer anlamsal benzerlik hesaplama yöntemleri. Bu yöntemler kavramsal sözlükte bulunan IS-A bağlantılarının ayrıt hesabına dayanmaktadır. Bunlara ek olarak, belge kümelerini derlem olarak kullanan, LSA derlem tabanlı anlamsal benzerlik hesaplama yöntemi de incelenmektedir. Ayrıca, bu çalışma kapsamında Jiang ve Conrath tarafından önerilen bir karma anlamsal benzerlik hesaplama yöntemi de incelenmekte ve adı geçen iki yöntemle karşılaştırılmaktadır.

Belgeler arasındaki anlamsal benzerlikler genel olarak içerdikleri kelimeler arasındaki anlamsal benzerlik bileşimleri kullanılarak hesaplanmaktadır. Kelimeler arasındaki benzerliklerin her bir doğrusal bileşimi farklı bir belge benzerliği değeri ortaya koyar. Bu çalışmada iki farklı belge anlamsal benzerlik hesaplama yöntemi kullanılmaktadır: THESUS ve SEMSIM. Bu hesaplamalarda, kelimeler arasındaki anlamsal benzerlikleri tespit etmek için daha önce adı geçen Wu-Palmer, Modified Wu-Palmer ya da Jiang-Conrath yöntemlerinden biri kullanılmaktadır. Ayrıca, sözcük benzerliği hesaplama yöntemlerinden kosinüs ve Jaccard benzerlikleri de tanımlanmakta ve tartışılmaktadır. LSA yöntemi belgeler üzerine de uygulanarak belgeler arasındaki benzerlik kosinüs benzerliği kullanılarak hesaplanmıştır.

Bu çalışmada önerilen belge anlamsal benzerlik yöntemi, mevcut yöntemlerden farklıdır. Hesaplamada iki kelime arasındaki anlamsal benzerlikler kullanılmamaktadır. Onun yerine, incelenen her döküman için kavramsal

sözlük ağacı yoluyla elde edilen kavram vektörleri kullanılmaktadır. Daha sonra bu vektörler üzerinden kosinüs benzerliği hesaplanmaktadır. Böylece, önerilen yöntem, belgelerin anlamsallığının kösinüs benzerlik hesaplama yöntemi ile bütünleştirilmesinden oluşmaktadır. Bununla birlikte önerilen yöntemin mevcut belge anlamsal benzerlik yöntemlerine asıl üstünlüğü zaman karmaşıklığının düşük olmasıdır.

Yukarıda adı geçen benzerlik yöntemlerinin karşılaştırılması için Türkçe belgeler, tanımlanan yöntemlerden biri kullanılarak demetlenmekte ve demetleme sonuçları kaydedilmektedir. Demetleme sonuçlarının karşılaştırılması beş farklı demetleme göstergesiyle yapılmaktadır. Bunlar: Entropy, Jaccard, Purity, Sihouette ve Davies-Boulding göstergeleridir. Entropy, Purity ve Jaccard göstergeleri belge kümelerinin gerçek etiketlerini kullanır ve demetlerin genel "arılığını" ölçer. Sihouette ve Davis-Boulding göstergeleri ise belgelerin sınıf bilgisini gerektirmez ve oluşan demetlerin kalitesini ölçer. Deneysel çalışmalar önerilen yöntemin demetleme kalitesi bakımından diğer anlamsal benzerlik ve sözcük benzerliği hesaplama yöntemlerinden daha iyi olduğunu ve aynı zamanda da zaman karmaşıklığının düşük olduğunu göstermektedir.

1. INTRODUCTION

1.1 Problem Definition

With the rapid growth and high diversity of the online data the concepts like data organization, access and retrieval have become a crucial part of many software applications. Most of the times, the data organization precedes the latter ones, simplifying and speeding up them. The data organization generally implies its classification or clustering. Classification and clustering of the raw text subsume many different techniques from different areas of computer science like machine learning, natural language processing and data mining. Several such techniques and combinations of them exist in order to simplify and speed up the process of data management. Web page classification algorithms (supervised learners), classify unprocessed data according to some predefined models. Those models are constructed using the previously classified data, which contains web pages with the classes these pages belong to. After the model is constructed, the classifiers can use both web page text data and inter-page link connection information for class prediction of unseen data[1],[2]. The web document clustering, which is a from of an unsupervised learning, does not require class label information of web pages for classification. Instead, it organizes the data in clusters or groups so that pages in the same clusters are more similar to one another than the pages from the different clusters. Clustering is widely used in several applications like information retrieval [3], topic or keyphrase extraction [4],[5], personalization of web search engines results [6] and assistance of users on the web sites [7]. Web pages or documents are generally clustered according to (dis)similarities among them. The document similarity, in turn, depends on how similar the words composing them.

Most of the text(document) clustering algorithms use single term analysis of the text, such as vector space model (bag of words model)[8]. In this model documents

are represented as feature vectors. Each dimension of these vectors corresponds to a distinct word in a document set. Term Frequency or Term Frequency - Inverse Document Frequency metrics are used to represent the numeric value of a word in a vector. This model is then passes to a single term document similarity calculation, which corresponds to cosine or Jaccard similarity measure between vectors. Saying in other words, similarity between two documents directly depends on the number of words they both have in common.

In contrast to single term similarity measures, most of document semantic similarity metrics rely on semantic relatedness between words they contain. As an example, according to the cosine similarity measure, similarity between terms "nurse" and "doctor" will be 0 as these two words represent different dimensions in a vector space. However it is obvious that these two terms are highly semantically related. Calculation of semantic similarity between two terms is a widely studied area and several methods are proposed and described in details [9],[10],[11],[12],[13]. Despite the differences in these methods, they can be generalized in three main categories : corpus-based, taxonomy-based and hybrid semantic similarities. Corpus-based semantic similarity metrics [14],[15] uses a large corpora to calculate similarity between words, while taxonomy-based approaches [9],[10],[13],[16] make use of a spacial large word thesaurus, in which certain semantic relations between words are defined. Hybrid methods [17],[18] combine both corpus and ontology based metrics to calculate semantic relatedness between words. Many studies were conducted on definition and comparison of different semantic similarity metrics [19],[20],[21]. Those studies have shown that even different linear combination of same algebraic terms in equations produces different similarity results.

The semantic similarity between documents is generally calculated using semantic similarities between words they contain. All existing document semantic similarity metrics differs in linear combination of word semantic similarity measures used in calculations. Those combinations generally infer the selection of certain words to represent documents in a similarity estimation. After inter-document similarities are computed, results are gathered into the similarity matrix. Each entry in the similarity matrix represents the similarity between two

documents. The document clustering is performed over the similarity matrix. Thus, both underlaying word semantic similarity measure and the methods used to combine them directly effect the clustering results. In this study, different similarity measures are used in the clustering of Turkish documents, and compared in terms of clustering validity indices.

1.2 Contribution of the Thesis

Despite the fact that the semantic similarity metrics better identify relations between distinct documents, single term similarity measures like cosine similarity metric remain as most widely reported measure of text similarity. The reason of preference of the latter is not only simplicity, but also lower time complexity of calculation. Therefore, the aim of the current study is the development of a semantic similarity measure between documents, that possess the properties of semantic relatedness and has the low time complexity. In this thesis, a new method for calculating semantic similarity between is proposed. It makes use of concept vectors of words, extracted from taxonomy tree, which are then used in calculation of cosine similarity between documents. As concept vector of any document is calculated offline (before the similarity calculations), the time complexity of proposed method is the same with cosine similarity measure. As a result, precise and fast similarity measure is developed to be used in clustering of Turkish documents.

This study contains two main parts and each of them described in Section 6. The first part consist of comparison of existing document similarity measures in terms of effects they have on Turkish document clustering. The second part describes the proposed method and compares it with cosine similarity measure in terms of cluster validity indices.

Overall approach in all parts of the study can be generalized as follows. First of all, documents, collected from the web, pass through preprocessing stage. Then each document is transformed into the vector space model, where it is represented as a vector of words(terms). For each pair of document degree of similarity is calculated using specified similarity measure. The similarities then

are gathered into one similarity matrix of a document set. Clustering is done over that similarity matrix and results are evaluated using the cluster validity indices.

1.3 Organization of the Thesis

This Thesis consists of seven sections. In the Section 2, word thesaurus used in this study are described. In Section 3, word semantic similarity metrics based on those thesaurus and the ones used in this study are described. Next section, Section 4, presents the document similarity metrics which are compared and studied in this thesis. In the Section 5, a new method for calculating document semantic similarity is proposed and a simple example illustrates the steps of calculations. In the Section 6, the results of conducted experiments are shown and discussed. Finally, the last section, Section 7, concludes this thesis and includes the proposition of a future work.

2. WORD THESAURUS

The semantic relation or similarity of words is defined as a relation between *meaning* of those words. To be able to calculate the semantic similarity, a large set of words with semantic relations between them should be provided. The Word Thesaurus is a large collections of terms or concepts, constructed by information specialists. Different kinds of relations between terms can be defined on the thesaurus. In the following sections, two most known thesaurus types are described. Both of them are used in the semantic similarity calculations later on.

2.1 WordNets

English WordNet [22] is a large lexical database for English words. It was created and is being maintained at the Cognitive Science Laboratory of Princeton University under the direction of psychology professor George A. Miller. In the WordNet, words are grouped in synonym sets (*synsets*), having short definitions (*glossary*) and several semantic relations between each synset. Nouns and verbs are organized into hierarchies, defined by hypernym/hyponym relation (*IS-A type relation*). The subsumption hierarchy, i.e., IS-A relation network, accounts for approximate to 80% of all relations defined in the English WordNet. That's why most of taxonomy-based semantic similarity measures based on WordNet use only IS-A type relation to calculate the similarity[9],[10],[13],[16].

In this study, the taxonomy of words from Turkish WordNet [1] is used. It is a multilingual lexical database comprising of individual WordNets for the Balkan languages, including Turkish. Fig. 2.1 illustrates a fragment of this taxonomy, where arrows represent hierarchical hypernym/hyponym (IS-A) links between words (concepts).

The statistics of Turkish WordNet [24] is given in Tables 2.1 and 2.2.

[1] BalkaNet[23]

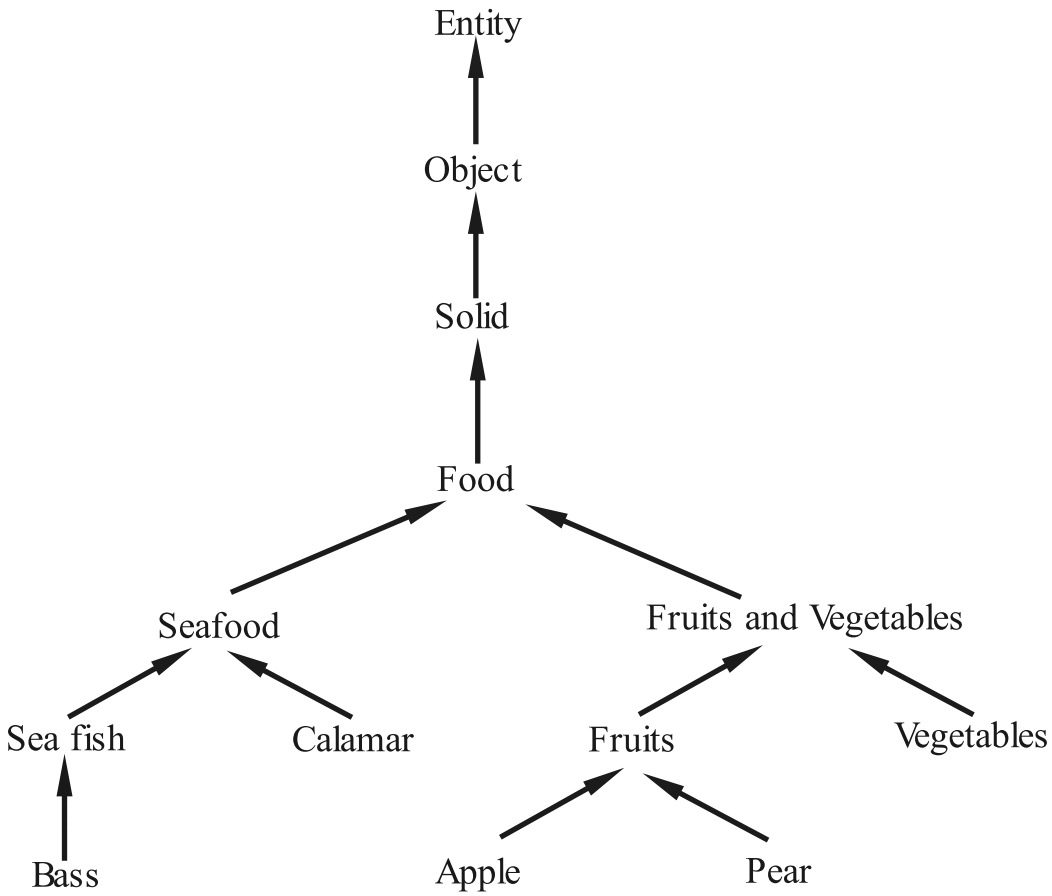

Figure 2.1 : A Fragment of BalkaNet. Solid arrows represent IS-A links

Table 2.1 : Distribution Parts of Speech of Turkish WordNet

Synset Type	Number	Percentage
Nouns	8 691	74.7%
Verbs	2 556	22.0%
Adjectives	381	3.3%

Table 2.1 presents the content of Turkish WordNet in terms of Parts of Speech. As can be seen from this table, 3/4 of the Turkish WordNet are nouns. Table 2.2 presents the relation types between all synsets in Turkish WordNet. Semantic relations defined for nouns are Hypernym, Holo_part, Holo_member, Near_antonym, Category_Domain and Usage_Domain. In total, Hypernym/hyponym (IS-A) relation subsumes approximately 70% of semantic relations existing between noun words in Turkish WordNet. Referring on this

Table 2.2 : Semantic Relations in Turkish BalkaNet

Relation Type	Number
Hypernym	12 907
Holo_part	1 815
Holo_member	1 245
Also_see	1 018
Similar_to	2 487
Near_antonym	1613
Category_domain	403
Be_in_state	617
Sub_event	131
Causes	100
Usage_domain	32
Total	22 368

statistics, semantic similarity metrics used in this study use only IS-A relation network of Turkish WordNet.

2.2 Corpus

The corpus (plural *corpora*) is a large and structured set of texts. According to Tognini [25] corpus is a thesauri of words, which is defined as a collection of texts assumed to be representative of a given language, put together so that it can be used for linguistic analysis. There exist several semantic similarity metrics that uses corpora to calculate the similarity between two words[14],[15]. Those similarities measures are based on corpus statistics.

For calculation of corpus-based semantic similarity based on the LSA of documents, experimental document sets are used as a corpora. For calculation of hybrid semantic similarity measure, we use a METU-Turkish Corpus, that is a collection of 2 million words of post-1990 written Turkish samples [26].

Other most known thesaurus like Roget's Thesaurus [27], Macqurie Thesaurus [28], group words in a structure based on categories within which there are several levels of finer clustering [20]. As these thesaurus are defined only for English words, it was not possible to use them in this study, which is concentrated on the similarities of Turkish documents.

3. SEMANTIC SIMILARITIES BETWEEN WORDS

As have been mentioned before, semantic similarity between two terms has been studied widely. There exist several studies where these metrics are described and compared [10],[19],[20],[21]. The efficiency of proposed methods are generally evaluated with respect to the human interpretation of the relatedness degree between processed words. There exist several applications, which use semantic similarities between words. Synonym extraction of words from a large dictionary [29] relays on the idea that the words with semantically close definitions are likely to be synonyms. Budanitsky and Hirst [20] provide the large overview and comparison of the different semantic similarity metrics on the detection and correction of real-word spelling errors in open-class words, i.e., malapropism. Solving TOEFL-style Synonym Questions and Detecting Speech Recognition errors by selecting words that do not fit into their context is another application where word semantic similarity is used [30].

Word semantic similarity measures are categorized according to the thesaurus that is used in similarity calculation. Three main categories of semantic similarity measures are *Corpus-based*, *Taxonomy-based* and *Hybrid* similarity metrics. In following sections some methods in these categories are explained. In this study one corpus-based, two taxonomy-based and one hybrid semantic similarity measuring methods are further examined.

3.1 Corpus-based Semantic Similarity Metrics

Corpus-based semantic similarity measures between words are identified using information derived from large corpora. In general, the degree of semantic similarity is measured according to co-occurrence of two words in a given corpus. The most widely used corpus-based semantic similarity measure, which is the

Latent Semantic Analysis (also called Latent Semantic Indexing), is explained below.

3.1.1 LSA Semantic Similarity

The Latent Semantic Analysis (LSA) method is proposed by Landauer [15]. LSA is a varient of the vector space model, where term co-occurrences in a corpus are captured by means of a dimensionality reduction using the Single Vector Decomposition (SVD) method. SVD is a linear algebra operation, which can be applied to any rectangular matrix in order to find the correlations between rows and columns[19]. LSA improves the standard vector space model by reducing the sparseness and high dimensionality. The cosine similarity, which is later described in Section 4, can be used to calculate the similarity between documents in a newly formed vector space.

LSA modifies the word vector space model, thus the connection between documents and words contained in them is lost. That is why LSA can not be used in applications like topic identification or text summarization.

3.2 Taxonomy-based Semantic Similarity Metrics

Taxonomy-based semantic similarity measures make use of the Semantic Networks like WordNet and generally calculated by edge-counting methods. Different properties of a semantic network or a taxonomy can be included in similarity calculation. All metrics discussed below use WordNet taxonomy for calculation. As was mentioned before, IS-A type relation subsumes for approximate of 70% of all relation defined for noun words in Turkish WordNet. Thus, first two semantic similarity metrics discussed in this section use only IS-A type relation to calculate word semantic similarity. The third semantic similarity measure described here introduces another approach to calculate semantic similarity degree between two words in a taxonomy. This method is slightly similar with the document semantic similarity measure proposed in this thesis in the way that taxonomy and cosine similarity measures are used together to obtain the semantic similarity between any two entity.

3.2.1 Wu-Palmer

One of the well-known semantic similarity metric that uses IS-A taxonomy is a similarity proposed by Wu and Palmer [9] in a paper on translating English verbs into Mandarin Chinese. The Wu-Palmer similarity between pair of concepts (words) c_1 and c_2 uses the depth information of concepts(words) in a taxonomy relatively to their lowest super-ordinate($lso(c_1,c_2)$) and the depth of $lso(c_1,c_2)$ itself (path from $lso(c_1,c_2)$ to the root of taxonomy). $lso(c_1,c_2)$ can be simply interpreted as the lowest "common" parent of two nodes in heirarchy.

$$sim_{W\&P}(c_1,c_2) = \frac{2 \times N_3}{2 \times N_3 + N_1 + N_2} \tag{3.1}$$

where N_3 is the depth of $lso(c_1,c_2)$, N_1 and N_2 are the path length from c_1 and c_2 to $lso(c_1,c_2)$ repectevly.

3.2.2 Modified Wu-Palmer

In [10] a slightly modified version of Wu-Palmer similarity metric was poroposed by Gunduz and Yucesoy. It is defined as follows:

$$sim_{G\&Y}(c_1,c_2) = \frac{N_3}{N_3 + max(N_1,N_2)} \tag{3.2}$$

As can be seen from Equation 3.2 it produces smaller than (or equal to) results of Equation 3.1. The idea behind this is to decrease semantic relatedness between dissimilar concepts.

To illustrate the calculation method and the difference between Wu-Palmer and Modified Wu-Palmer similarity metrics, consider the example of calculating the semantic similarity between two words *apple* and *seafish*. The Fig. 2.1 shows the part of Turkish WordNet needed for calculation. In this example $lso(apple,seafish)$ is a term "Food", thus $N_3 = 3$, $N_1 = 3$ and $N_2 = 1$. Replacing these numbers in the Equation 3.1 and Equation 3.2, the semantic similarities are equal to $sim_{W\&P}(apple,seafish) = 0.60$ and $sim_{modified}(apple,seafood) = 0.50$. A smaller word semantic similarity will leads to the smaller document semantic

similarity. Thus, the aim is to decrease semantic similarity between documents having unrelated terms.

3.2.3 Wan & Angryk

The semantic similarity measure proposed by Wan and Angryk [31] uses extended context vectors retrieved from the WordNet to calculate semantic similarity between two words. For each compared word c the vector of concepts $\vec{v}$ is formed by the help of all relations (hypernyms,meronyms, attributes etc.) and glossary information existing in WordNet. When such vectors for compared words are formed, cosine of the angle between two context vectors is used to measure the degree of similarity. The formulation is shown below:

$$sim_{W\&A}(c_1, c_2) = cos(\vec{v_{c_1}}, \vec{v_{c_2}}) = \frac{\vec{v_{c_1}} \cdot \vec{v_{c_2}}}{|\vec{v_{c_1}}||\vec{v_{c_2}}|} \tag{3.3}$$

The proposed document similarity measure, described later in Section 5 works similar to the word semantic similarity metric proposed by Wan and Angryk. They both make use of term vectors extracted from WordNet and cosine similarity measure. The Wan-Angryk semantic similarity metric extracts all semantic information of the word from the WordNet which is a very time consuming task. That's why it is not used in the experimental part of this thesis.

3.3 Hybrid Methods

Hybrid methods combines corpus and taxonomy based semantic similarities. Some of the hybrid similarity metrics use different taxonomy properties like density and node depth in semantic similarity calculations. All of hybrid similarity methods use the word frequency information that is obtained from the large corpora. This frequency is simply calculated by dividing the number of occurrence of certain concept c by the total number of words present in corpora. In [32] different combinations of taxonomy depth and density, path length between words and information content of them are studied in order to identify the best combination. The details of two hybrid semantic similarity methods are discussed below.

3.3.1 Resnik

Resnik [18] defines the similarity between two concepts as the extent to which they share information in common. As both of those concepts are found in the same ontology hierarchy (taxonomy), this extent or common information "carrier" of concepts c_1 and c_2 is their lowest super-ordinate($lso(c_1,c_2)$). The value of information content of the $lso(c_1,c_2)$ is obtained from corpora. The formulation of this semantic similarity metric is shown below:

$$sim_{Res}(c_1,c_2) = IC(lso(c_1,c_2)) = log^{-1}P(lso(c_1,c_2)) \tag{3.4}$$

where $P(c)$ is a probability of encountering an instance of concept c. It is calculated by dividing the number of occurrence of c in a corpus by the total number of concepts present in that corpus. As can be seen from the equation 3.4 as probability of concept c increase the information content of that concept decrease. Because of hierarchical structure of ontology (taxonomy), concepts lower in hierarchy are subsumed by the upper ones, thus, $P(c)$ monotonically increase, and information content $IC(c)$ monotonicaly decrease from bottom of ontology to it's top.

3.3.2 Jiang & Conrath

In [12] Jiang and Conrath proposes another hybrid method that combines WordNet ontology with a large corpus. Semantic similarity between words is calculated using edge-based notation by adding the information content factor as a decision factor. The link strength between child concept c_i and parent node p is defined as:

$$LS(c_i,p) = -log(P(c_i|p)) = IC(c_i) - IC(p) \tag{3.5}$$

Considering other factors, such as local density, node depth, and link type, the overall edge weight $wt(c,p)$ for a child node c and its parent p can be determined as follows:

$$wt(c,p) = \left(\beta + (1-\beta)\frac{\bar{E}}{E(p)}\right)\left(\frac{d(p)+1}{d(p)}\right)^{\alpha}[IC(c) - IC(p)]T(c,p), \tag{3.6}$$

where $d(p)$ denotes the depth of the node p in the hierarchy, $E(p)$ the number of edges in the child links (i.e local density), $\bar{E}$ the average density in the whole hierarchy, and $T(c,p)$ the link realtion/type factor. The parameters $\alpha(\alpha \geq 0)$ and $\beta(0 \leq \beta \leq 1)$ control degree of how much the node depth and density factor contribute to the edge weighting. Optimal parameters for α and β are $\alpha = 0.5$ and $\beta = 0.3$ [12]. The overall distance between two nodes would thus be the summation of edge weights along the shortest path linking two nodes:

$$Dist(c_1, c_2) = \sum_{c \in (path(c_1,c_2) - lso(c_1,c_2))} wt(c,p) \tag{3.7}$$

where $path(c_1, c_2)$ is the set that contains all the nodes in the shortest path from c_1 to c_2. One of the elements of the set is $lso(c_1, c_2)$, which denotes the lowest super-ordinate of c_1 and c_2. Considering the special case where only link strength is considered, by setting $\alpha = 0, \beta = 1$ and $T(c,p) = 1$, the distant function can be simlified as follows:

$$Dist_{J\&C}(c_1, c_2) = IC(c_1) + IC(c_2) - 2 \times lso(c_1, c_2) \tag{3.8}$$

As most of the studies are generally concentrated not on distance but on similarity between two concepts, any linear inverse of Eq.3.7 or Eq.3.8 can be used to calculate semantic similarity between two concepts. As an example, similarity can be calculated as follows [21]:

$$Sim_{J\&C}(c_1, c_2) = 1 - Dist(c_1, c_2) \tag{3.9}$$

Another transformations of Jiand&Conrath distance into similarity can be found in [19] and [33]. In this study, the Jiand&Conrath similarity is calculated using the distance measure described in Eq.3.7 in combination with Eq.3.9.

4. DOCUMENT SIMILARITIES

Before applying any similarity metric to the documents, those documents must be transformed to a vector space model in order to construct the base for calculations. In the vector space each document is represented by an individual vector and all distinct words in a whole document set represent distinct dimensions. The value of word or dimension in a document vectors is assigned to some numeric weight, calculation of which is described below.

Let $d_i \in D$ be a document in a document set D. Then vector representation of the document d_i is as follows:

$$\vec{d_i} = \{(c_1, w_{i1}), (c_2, w_{i2}), \ldots, (c_n, w_{in})\} \tag{4.1}$$

where $c_1, c_2 \ldots, c_n$ are words or concepts appear in D and w_{ij} is the weight of word c_j in a document d_i (if c_j does not exist in document d_i, the corresponding weight $w_{ij} = 0$).

There exist several weighting schemas for representing the weight of term in a document vector. Two of them are described here. The Term Frequency (TF) of any term c_j in a document d_i is calculated as follows:

$$TF_{ij} = \frac{f_{ij}}{\sum_{k=1}^{n} f_{ik}} \tag{4.2}$$

where f_{ik} is the number of times that the term c_k appears in document d_i.

Inverse Document Frequency, which measures the general 'importance' of the term c_j in a document set D is calculated as:

$$IDF_j = \log \frac{n_j}{n} \tag{4.3}$$

where n_j is the number of documents where term c_j appears , and n is total number of documents in the document set D. Then TF-IDF weight of term c_j in

document d_i is computed as :

$$tfidf_{ij} = TF_{ij} \cdot IDF_j \tag{4.4}$$

4.1 Document Semantic Similarity Metrics

In contrast to word semantic similarity measures, semantic similarities between documents are less investigated. There exist only a few studies that compare those metrics among each other [21].

After documents in a document set are represented in a vector space model they are passed to the similarity calculation algorithms. All of existing document semantic similarity metrics includes word pairwise similarity calculations within themselves. They generally differs in the way these calculations are combined. The combination includes the selection and weight assignments to word that would represent a document in similarity calculations. In this section, two document semantic similarity measures are described.

4.1.1 SEMSIM document semantic similarity

One method to calculate the semantic similarity between documents is proposed in [10]. It is formulated as below:

$$sim_{SEMSIM}(d_i, d_j) = \frac{\sum_{r=1}^{n} \sum_{u=1}^{n} w_{ir} \times w_{ju} \times Sim(c_r, c_u)}{\sqrt{\sum_{k=1}^{n} w_{ik}^2 \times \sum_{k=1}^{n} w_{jk}^2}} \tag{4.5}$$

where $Sim(c_r, c_u)$ refers to a semantic similarity between two concepts(words) c_u and c_r. Each word on two different documents has an impact on the pairwise document similarity value proportional to their weights in these documents. The similarity measure is normalized by the multiplication of document vector lengths.

4.1.2 THESUS

In [34], another method for calculating the document semantic similarity is proposed. It is described as follows:

Let Ω represents the ontology (taxonomy) and Eq. 4.1 represents a document in a document set D such as $c_u \in \Omega$. According to [34] the similarity between two documents can be calculated as follows:

$$sim_{THESIM}(d_i,d_j) = \frac{1}{2}\left[\left(\frac{1}{K}\sum_{r=1}^{n}\max_{u\in[1,n]}(\lambda_{r,u}Sim(c_r,c_u))\right) + \left(\frac{1}{H}\sum_{r=1}^{n}\max_{u\in[1,n]}(\mu_{r,u}Sim(c_r,c_u))\right)\right] \quad \textbf{(4.6)}$$

where

$$\lambda_{r,u} = \begin{cases} \frac{w_{ir}+w_{ju}}{2\times max(w_{ir},w_{ju})} & \text{if} \quad w_{ir}>0 \wedge w_{ju}>0 \\ 0 & otherwise \end{cases}$$

and K is a normalizing factor that is the sum of all the $\lambda_{u,r}$ that were used. $Sim(c_u,c_r)$ refers to the semantic similarity between two concepts(words) c_u and c_r. In a similar way, $\mu_{r,u}$ and H are defined. The weight factors, $\lambda_{u,r}$ and $\mu_{r,u}$, give less importance to terms that do not describe the document with a high weight.

As can be seen from document similarity calculation discussed above, all of them includes word pairwise semantic similarity calculation within themselves. Those calculations generally include WordNet tree traversals, that's why the main drawback of document semantic similarity calculation is their time complexity. If the number of documents to be compared is N, and each document is represented by c terms, then time the complexity for document semantic similarity is $O(N) = N^2 \cdot c^2 \cdot d$, where d is overall depth of the WordNet tree. Even if c is small number, the time complexity generally equals to $O(N) = N^3$, because of quadratic form of c and multiplication by d.

4.2 Single Term Similarity Metrics

Single term similarity metrics do not use any semantic relation between words and are calculated directly from vector space model representations of the documents. Thus, they are generally faster than document semantic similarity measures. Time complexity of single term similarity metrics is $O(N) = N^2 \cdot c$, where N is the number of documents used in calculations, and c is the number of term in each document. As c is generally small number ($c << N$) time complexity of single term similarity measure calculation can be generalized to $O(N) = N^2$. Being simple and fast is the reason of the preference of single term similarity

metrics over the semantic ones. Below, two most known and widely used single term similarity metrics are described.

4.2.1 Cosine similarity measure

Cosine similarity between two documents is calculated using vector representation of documents as shown in equation 4.1. It measures the angle between two vectors and calculated by dividing inner product of those vectors by multiplication of their length. The cosine similarity between documents d_i and d_j is formulated as follows:

$$sim_{cos}(d_i, d_j) = \frac{\overrightarrow{d}_i \bullet \overrightarrow{d}_j}{\| \overrightarrow{d}_i \| \cdot \| \overrightarrow{d}_j \|} = \frac{\sum_{k=1}^{n} w_{ik} \times w_{jk}}{\sqrt{\sum_{k=1}^{n} w_{ik}^2 \times \sum_{k=1}^{n} w_{jk}^2}} \tag{4.7}$$

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 common words, which play the role of dimensions. When all of the words are same and have the same weight assignment, i.e. documents are identical, cosine similarity is equal to 1.

4.2.2 Jaccard coefficient

Jaccard Coefficient is another metric used to calculate single term similarity between two documents. Like cosine similarity metric, it is calculated directly on document vector representation as shown in equation 4.1. Jaccard Coefficient between two documents d_i and d_j is calculated as follows:

$$sim_{JC}(d_i, d_j) = \frac{\sum_{k=1}^{n} w_{ik} \times w_{jk}}{\sum_{k=1}^{n} w_{ik}^2 + \sum_{k=1}^{n} w_{jk}^2 - \sum_{k=1}^{n} w_{ik} \times w_{jk}} \tag{4.8}$$

As can be seen from the above equation the Jaccard Coefficient measure differs from the cosine measure in the way it normalizes the inner product of two document vectors. The cosine similarity is sensitive to the relative importance of each word [35]. The Jaccard Coefficient, in contrast, measures similarity as the proportion of (weighted) words two texts have in common versus the words they do not have in common [36].Similar to the cosine similarity measure the Jaccard

Coefficient takes value between $0 \leq sim_{JC} \leq 1$. Like the cosine similarity, in order to have the Jaccard Coefficient greater than 0, documents must have some words in common.

Although semantic similarity measure are expected to be more efficient then single term document similarity metrics, second ones are more preferable because of the simplicity and low computation time. Time complexity is a very important aspect of many software applications, and it becomes more important when these applications includes online data organization.

5. PROPOSED DOCUMENT SEMANTIC SIMILARITY METHOD

As was mentioned before, the main drawback of all document semantic similarity measures is their high time complexity. Clustering of the documents includes a high number of pairwise document similarity calculations. The clustering time directly proportional to the time spent for similarity computations. Taking the clustering of large document set as a core objective, the main point of this study is the development of a semantic similarity measure that will have a low time complexity; at least the one that the cosine similarity metric has. In this thesis a new method for calculating the semantic similarity between documents is proposed.

The idea behind the proposed method is to merge semantics of documents with cosine similarity measure for document similarity calculations. It is done in order to keep the semantic relations between documents while using the single term similarity measures, which, in terms, have lower computation time. The proposed method make use of the concept vectors, extracted from the taxonomy tree. Those vectors then represent the documents in pairwise cosine similarity calculations. The concept vectors of a document is constructed from the parent vectors of the words present in that document. The definitions for parent vectors and concept vectors are given bellow.

Definition 5.1(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}.

Fig.5.1 illustrates a parent vector of the term *apple* extracted from the BalkaNet taxonomy tree. Calculation of parent vector 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.

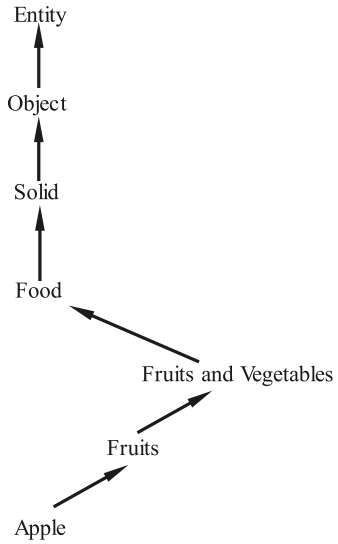

Figure 5.1 : Parent Vector of the term *apple* in a BalkaNet

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. 4.1.

Definition 5.2(Concept Vector) Let document d_i be represented as shown in Equation 4.1. Then the *concept vector* $\vec{concept\_d_i}$ of the document d_i is formed by merging all parent vectors of terms, present in d_i. That is, $\vec{concept\_d_i} = \vec{P_{c1}}, \vec{P_{c2}}, \ldots, \vec{P_{cn}}$

As staited in Def. 5.1, to form a concept vector $\vec{concept\_d_i}$ for document d_i, we merge all parent vectors $\vec{P_c}$ constructed for words found in d_i. While merging, if for some $\vec{Pc_i}$ and $\vec{Pc_j}$ there exist same words, they are recorded only once and their weights are summed up. After $\vec{concept\_d_i}$ is constructed, document d_i is represented by newly formed vector and this vector used in cosine similarity calculations and is formulated as follows:

$$sim_{CT}(d_i, d_j) = \frac{\vec{concept\_d_i} \bullet \vec{concept\_d_j}}{\|\vec{concept\_d_i}\| \cdot \|\vec{concept\_d_j}\|} \qquad \textbf{(5.1)}$$

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 increase. Thus the "importance" of relatedness between subsumed nodes decrease. Therefore, in order to obtain only

meaningful relations representing human cognitive limitations the length of the parent vector is limited.

Semantic similarity formulation using concept vectors is inspired by Wu-Palmer similarity measure. 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 concepts, where N_3 is high and N_1, N_2 are low; minimized otherwise. Proposed semantic similarity measure behaves in the same way. The lower position of $lso(c_i, c_j)$ leads to the greater overlapping between concept vectors of c_i and c_j, which will turn out in a higher semantic similarity. The proposed similarity is minimized when $lso(c_i, c_j)$ is very high in a taxonomy and concepts vectors of terms share only a few number of common words.

Concepts vectors of documents in a document sets are constructed offline, before the similarity calculations. The complexity of this procedure is $O(N) = N \cdot c \cdot d$, where N is number of documents in the document set, c is the number of terms used to represent each document in calculation and d is overall depth of taxonomy tree, which is used in word semantic similarity calculation. As was mentioned in Section 4 the complexity of cosine simiarity is $O(N) = N^2$. Then, the time complexity of proposed method is equal to $O(N) = N^2 + N \cdot c \cdot d$, or more precise $O(N) = N^2$, which is the same with the time complexity of cosine similarity measure.

A Walk-Through Example In order to illustrate the calculation steps of concept vectors and similarity calculation between documents a simple walk-through example is constructed.

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 $\vec{d}_l = (bass, 0.30), (squid, 0.20)$ respectively. By setting $k = 5$ the parent vectors of d_i are calculated as

$Pc_{apple,0.30}$={(apple, 0.30),
(fruits, 0.27),
(fruits and vegetables, 0.24),

(food, 0.21),
(solid, 0.18),
(object, 0.15)}

and

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

So at the end 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 it is clear that these documents are semantically related. The semantic similarities calculated with our proposed method are $sim_{CT}(d_i, d_j) = 0.62$ and $sim_{CT}(d_i, d_l) = 0.46$. As can be seen from calculation results, the proposed similarity metric decreases when $lso(c_i, c_j)$ goes up the taxonomy.

As can be seen from the discussions above, the proposed document semantic similarity measure successfully identifies the semantic similarity between distinct documents and has the time complexity as low as the cosine similarity's one.

6. EXPERIMENTS AND DISCUSSIONS

Two different experiment sets are conducted in this study. The first experiment set consists of comparison of document semantic similarity measures described in Eq. 4.5 and 4.6 with single term similarity measures that are cosine similarity and Jaccard Coefficient, described in Section 4. Moreover, the LSA method is also examined and compared with other similarity measures. Comparison is done over the clustering results produced using one of the similarity measures mentioned above. The second experiment set is conducted to compare the proposed method described in Section 5 with cosine similarity metric and document semantic similarity measure described in Equation 4.5. In all experiments same document sets are used. Similarity matrix of documents is constructed for each document set using one of the similarity measure listed above. This matrix is then passes to the clustering operation. The structure of similarity matrix is shown below.

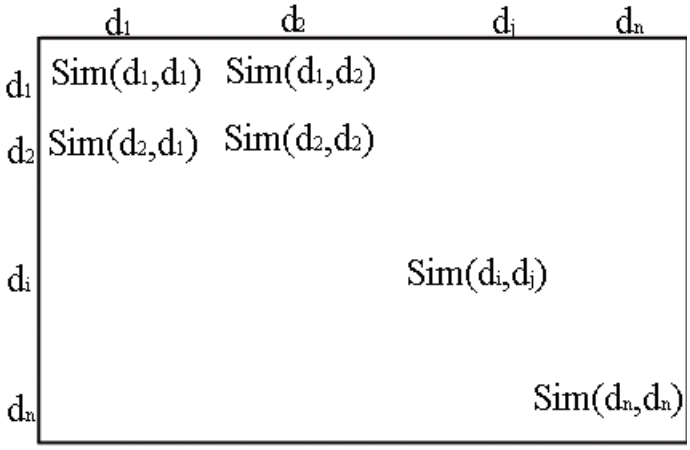

Figure 6.1 : Similarity matrix constructed for clustering

The clustering of similarity matrices is done using Cluto software package[1]. Embedded clustering method of Cluto tool that operates on document's similarity

[1]http://www-users.cs.umn.edu/ karypis/cluto/

space is chosen. Clustering results are evaluated and compared using different cluster validity indices.

6.1 Document Sets and Text Document preprocessing

In all experiments three different document sets are used. All of them are retrieved from the web. First document set (*Dataset1*) contains 2382 documents and categorized manually into seven clusters by en expert[2]. Second document set (*Dataset2*) contains 481 documents. It is retrieved manual from the web by setting predefined 5 topics and using the search engines. These topics are chosen in such a way that the grouped documents would have the minimum informational intersection and have clear differences in textual content. Web search engines are queried using words like "cars", "illnesses", "culture", "fashion" and "animals". Results of the queries are processed to obtain homogeneous groups of related documents that differ in size and are suitable for clustering experiments. Third dataset (*Dataset3*) is collected automatically from the web site of a Turkish Internet Service Company[3]. It consists of 1987 documents and do not have predefined cluster number. The properties of document sets are given in the Table 6.1.

Table 6.1 : Properties of the document sets

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

Retrieved web pages are not suitable for direct similarity estimation. That is why the preprocessing stage is applied to them to make calculation possible. First of all web pages are parsed to remove HTML tags and tokenized into individual terms using HTML parser. A morphological analyzer [37] and postagger [38] that are developed for Turkish are used to transform all terms into the most probable stem terms. The output of these operations is a document sets, consisting only

[2]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)

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

of text files, corresponding to the web pages. All stop words, defined for Turkish, and other noise are removed from the text files. Document feature vectors are then extracted from these text files.

The calculation of semantic similarity between documents is time and space consuming task due to the high dimension of the vector space model. To simplify measuring process, only m most frequent words (MFW) of each document are selected to represent those document in the calculations. The weights of MFW are assigned using normalized TF weighting schema.

The next step of preprocessing stage is a word sense disambiguation (WSD) phase. It is done not only to identify correct sense of the words but also to fix document term vectors that would be used in calculations. WSD is a large topic of natural language processing and it lays beyond the scope of our project. To simplify the WSD stage, we select a method, that is correlated with our study. If some word w from MFW of document d_i has more than one sence, the one, which has maximum semantic similarity with all other MFW is selected. As an example consider the document d which has MFW as $[burun(nose), kulak(ear))]$. In Turkish language, the word *burun* can stand for either a "nose" or a "foreland". Using Wu-Palmer semantic similarity metric, the similarities between $burun(nose), kulak(ear)$ and $burun(foreland), kulak(ear)$ are calculated, which are equal to 0.57 and 0.0 respectively. As first semantic similarity is greater than last, $burun(nose)$ is selected to represent the document d.

WSD is the last step of the preprocessing stage. After this step, all documents are ready for the similarity calculation and, in the sequel, for clustering.

6.2 Cluster Validity Indices

Clustering results are evaluated using five different cluster validity indices. Cluster validity indices are traditionally classified into three types: Unsupervised, Supervised and Relative [39]. Unsupervised validity indices measures the goodness of clustering structure without the respect to the external information. In this study two unsupervised cluster evaluation measures are used: *Silhouette*

and *Davies-Bouldin*. Supervised validity indices measures the extent to which the clustering structure discovered by clustering method matches some external structure. *Entropy*, *Purity* and *Jaccard* supervised evaluation measures are used in this study. The details of calculations of every validity indices is given below.

6.2.1 Unsupervised cluster validity indices

6.2.1.1 Davies-Bouldin index

DB index is a function of the ratio of the sum of intra-cluster dispersion to inter-cluster separation [40]. The DB index aims at identifying sets of clusters that are compact and well separated. The original DB index has been modified in [41] to be used in case of having pairwise similarities between data points in the data set. Let $\mathscr{D} = \{d_1, \ldots, d_N\}$ be the document set and let $\mathscr{C} = \{C_1, \ldots, C_K\}$ its clustering into K clusters. The DB index is calculated using Eq. 6.1, Eq. 6.2, Eq. 6.3, Eq. 6.4.

$$\Delta(C_i) = \frac{1}{|C_i| * (|C_i| - 1)} \sum_{d_i, d_j \varepsilon C_i, d_i \neq d_j} dis(d_i, d_j) \tag{6.1}$$

where $\Delta(C_i)$ is the average diameter of cluster C_i, $|C_i|$ denotes the number of documents in cluster C_i and $dis(d_i, d_j)$ is the dissimilarity between d_i and d_j. In this study the dissimilarity between two documents d_i and d_j is computed as $(1 - \sigma(d_i, d_j))$ where $\sigma(d_i, d_j)$ is the similarity between d_k and d_l.

$$\delta(C_i, C_j) = \frac{1}{|C_i| * |C_j|} \sum_{d_i \varepsilon C_i, d_j \varepsilon C_j} dis(d_i, d_j) \tag{6.2}$$

where $\delta(C_i, C_j)$ is the average linkage between the two clusters.

$$DB_j(C_j) = \max_{i \neq j} \left\{ \frac{\Delta(C_i) + \Delta(C_j)}{\delta(C_i, C_j)} \right\} \tag{6.3}$$

where DB_j is the average similarity between cluster C_j and its most similar one.

$$DB(\mathscr{C}) = \frac{1}{K} \sum_{j=1}^{K} DB_j(C_j) \tag{6.4}$$

where $DB(\mathscr{C})$ gives the DB index value of the clustering solution $\mathscr{C}$. A lower value of DB index indicates a good clustering solution.

6.2.1.2 Silhouette index

Another cluster validity index that can be used to judge the quality of any clustering solution is the Silhouette index [42]. As the DB index, it takes into account the compactness of the resulting clusters and the separation between them. The Silhouette index can be calculated as in Eq. 6.5, Eq. 6.6 and Eq. 6.7:

$$s(d_i) = \frac{b_i - a_i}{max(b_i - a_i)} \tag{6.5}$$

where a_i is the average dissimilarity between $d_i \in C_j$ and other documents in C_j, b_i is the minimum average dissimilarity between d_i and other clusters. A silhouette index S_j is assigned to each cluster C_j as in Eq. 6.6.

$$S_j = \frac{\sum_{d_i \in C_j} s(d_i)}{|C_j|} \tag{6.6}$$

The Silhouette index of the clustering (GS) solution can be calculated as in Eq. 6.7.

$$GS(\mathscr{C}) = \frac{\sum_{j=1}^{K} S_j}{K} \tag{6.7}$$

The GS takes values between -1 and 1 where greater values of it means a better clustering solution.

6.2.2 Supervised cluster validity indices

6.2.2.1 Entropy index

The entropy and purity measures are also frequently used external validation measures. Both of them measure the "purity" of the clusters with respect to the given class labels. Given a particular cluster C_k, the entropy of this cluster is defined to be [43]:

$$E(C_k) = -\frac{1}{log\ M} \sum_{i=1}^{M} \frac{n_k^i}{|C_k|}\ log \frac{n_k^i}{|C_k|} \tag{6.8}$$

where n_k^i is the number of data items of the ith class in $\mathscr{P}$ that were assigned to the kth cluster in $\mathscr{C}$. The entropy of the entire clustering solution is then defined

as the sum of the individual entropies of each cluster weighted by the number of data items assigned to the cluster:

$$Entropy(\mathscr{C}) = \sum_{i=1}^{K} \frac{|C_i|}{N} E(C_i) \tag{6.9}$$

where N is the number of data items in the data set. Lower entropy value indicates a better clustering performance.

6.2.2.2 Purity index

The purity of a cluster C_k is defined as the fraction of the number of data items of the cluster to the largest number of data items assigned to that cluster [43]:

$$Pr(C_k) = \frac{1}{|C_k|} max_i(n_k^i) \tag{6.10}$$

The overall purity of the clustering solution defined to be:

$$Purity(\mathscr{C}) = \sum_{i=1}^{K} \frac{|C_i|}{N} Pr(C_i) \tag{6.11}$$

In general higher purity value indicates a better clustering performance.

6.2.2.3 Jaccard index

The Jaccard index is one of the cluster validity indices, which uses document class label in calculation. The equation for calculating Jaccard index is shown below:

$$J(\mathscr{C}) = \frac{a}{a+b+c} \tag{6.12}$$

where $\mathscr{P} = \{P_1, \ldots, P_M\}$ be manually determined clusters in the data set $\mathscr{D}$, a denotes the number of pairs of documents with the same label in $\mathscr{P}$ and assigned to the same cluster in $\mathscr{C}$, b denotes the number of pairs of documents with the same label in $\mathscr{P}$, but in different clusters in $\mathscr{C}$ and c denotes the number of pairs of documents in the same cluster in $\mathscr{C}$, but with different class labels in $\mathscr{P}$. Clustering solution that results in a high Jaccard index is desirable.

6.3 Experiments

6.3.1 First experiment set

The aim of this experiment set is to compare semantic similarity metrics with single term similarity measures in terms of cluster validity indices. For every data set Silhouette and Davies-Bouldin indices are calculated. In addition to these, for data sets, whose actual class labels are known (*Dataset1,Dataset2*), Entropy, Purity and Jaccard indices are also computed.

The documents sets were preprocessed as mentioned in Section 6.1. The number m of MFW were empirically set to 20, since for large numbers the computing time grows unacceptably.

For each data set, pairwise similarities of documents are calculated using two different single term similarity measures, namely cosine and Jaccard similarities. In the experimental tables, cosine similarity measure is denoted as σ_{cos} and Jaccard coefficient is denoted as σ_{JC}.

Four different semantic similarities between documents based on taxonomy-based word semantic similarities are calculated: (1) THESIM based on Wu-Palmer similarity ($\sigma_{THESIM_{W\&P}}$ in the experimental tables); (2) THESIM based on Modified Wu-palmer similarity proposed by Gunduz and Yucesoy($\sigma_{THESIM_{modified}}$ in the experimental tables); (3) SEMSIM based on Wu-Palmer similarity ($\sigma_{SEMSIM_{W\&P}}$ in the experimental tables) and (4) SEMSIM based on Modified Wu-Palmer similarity ($\sigma_{SEMSIM_{modified}}$ in the experimental tables). THESIM based on Wu-Palmer similarity and Modified Wu-Palmer similarity are calculated using Equation 4.6 by substituting Wu-Palmer and Modified Wu-Palmer similarity, respectively, into the formula instead of semantic similarity values of terms ($Sim(c_u, c_r)$). In a similar way, SEMSIM based on Wu-Palmer similarity and Modified Wu-Palmer similarity is calculated.

The first part of this experiment set is aimed to identify the similarity metric that produces the clustering solution, matching the actual class labels of the document sets the best. As *Dataset3* do not has pre-defined class labels, this part

of experiments is not performed on it. Experiments are conducted on *Dataset1* and *Dataset2* using the pre-defined class numbers, that are 7 and 5 repectively. Supervised cluster validity indices are used to evaluate the clustering results, which are Entropy(E), Purity(P) and Jaccard(J) (notice that best values are large for Jaccard and Purity, small for Entropy).

These results are illustrated in Table 6.2. As can be seen from results, all semantic similarity measures lead to the clustering results which are very close to each other. The reason for this may be that, as Jain et al. [44] have pointed out, clustering is a subjective process and the same set of data items often needs to be partitioned differently for different applications. Eventhough, it can be said that single term similarities are performed better than document semantic similarity metrics. This may imply, that single term similarity measure matches the human judgements better than semantic similarities, which in term, can find semantic relations between documents that exceed human comprehension. Among document semantic similarities SEMSIM based on Wu-Palmer similarity is generally better than other semantic similarity metrics.

Table 6.2 : The results of manually clustered documents of Dataset1 and Dataset2

Similarity Measure	Dataset1			Dataset2		
	E	*P*	*J*	*E*	*P*	*J*
σ_{cos}	**0.77**	**0.40**	**0.14**	**0.62**	**0.51**	0.23
σ_{JC}	0.79	0.40	0.13	0.64	0.51	**0.24**
$\sigma_{THESIM_{W\&P}}$	0.83	0.33	0.10	0.75	0.40	0.18
$\sigma_{THESIM_{modified}}$	0.90	0.32	0.11	0.87	0.40	0.16
$\sigma_{SEMSIM_{W\&P}}$	0.78	0.34	0.11	0.72	0.40	0.16
$\sigma_{SEMSIM_{modified}}$	0.83	0.33	0.11	0.73	0.40	0.16

In the second part of these experiment set, document set are evaluated without the usage of pre-defined class labels, thus all data sets are considered. Unsupervised evaluation measures, which are Silhouette and Davies-Bouldin indices, are used to evaluate the clustering results (notice that the best values are high for Silhouette and small for Davies-Bouldin). Clustering is done using different number of clusters. Tables 6.3, 6.4, 6.5 show these results. As can be seen from the table, number of clusters that have the best Silhouette and Davies-Bouldin values do not match the actual class numbers of *Dataset1* and *Dataset2*. Only for *Dataset2*

both Silhouette and Davies-Bouldin indices are agree on the same number of clusters (10) for SEMSIM based on Wu-Palmer similarity measure. Only for *Dataset3* cosine similarity measure leads to a better clustering solution. The reason for these can be the fact that most of words in document sets could not be found in a Turkish WordNet tree, that leads to the worse results for semantic similarity measures.

Table 6.3 : Silhouette and Davies-Bouldin indices of Dataset1 for different number of clusters

Similarity Measure	Silhouette index					Davies-Bouldin index				
	number of clusters					number of clusters				
	5	7	10	15	20	5	7	10	15	20
σ_{cos}	0.02	0.04	0.03	0.03	0.04	1.94	1.91	1.92	1.92	1.91
σ_{JC}	0.01	0.02	0.02	0.02	0.03	1.97	1.96	1.96	1.95	1.93
$\sigma_{THESIM_{W\&P}}$	0.01	0.01	0.01	0.01	0.01	1.93	1.93	1.92	1.92	1.92
$\sigma_{THESIM_{modified}}$	0.01	0.01	0.01	0.01	0.01	1.93	1.93	1.92	1.92	1.92
$\sigma_{SEMSIM_{W\&P}}$	**0.07**	0.05	0.04	0.05	0.03	1.81	1.83	1.83	1.81	1.82
$\sigma_{SEMSIM_{modified}}$	**0.07**	0.04	0.03	0.05	0.03	1.81	1.83	1.84	**1.80**	1.81

Table 6.4 : Silhouette and Davies-Bouldin indices of Dataset2 for different number of clusters

Similarity Measure	Silhouette index				Davies-Bouldin index			
	number of clusters				number of clusters			
	4	5	10	15	4	5	10	15
σ_{cos}	0.04	0.15	0.14	0.11	1.89	1.80	1.80	1.82
σ_{JC}	0.13	0.11	0.07	0.06	1.84	1.87	1.89	1.91
$\sigma_{THESIM_{W\&P}}$	0.09	0.11	0.14	0.09	1.83	1.81	1.78	1.83
$\sigma_{THESIM_{modified}}$	0.07	0.12	0.08	0.09	1.85	1.79	1.84	1.83
$\sigma_{SEMSIM_{W\&P}}$	0.10	0.12	**0.16**	0.13	1.80	1.80	**1.74**	1.77
$\sigma_{SEMSIM_{modified}}$	0.10	0.12	0.10	0.10	1.82	1.80	1.81	1.80

Considering these result, it can be said that SEMSIM based on Wu-Palmer similarity measure slightly outperforms all other document semantic similarities, investigated in this study. Taking this fact into account, we decide to farther examine this similarity measure and conduct second part of the experiments, where we compare SEMSIM based on Wu-Palmer similarity measure and cosine similarity metric using all terms that appear in documents using only *Dataset2*. The reason for choosing this document set is that it is the smallest among all.

Table 6.5 : Silhouette and Davies-Bouldin indices of Dataset3 for different number of clusters

Similarity Measure	Silhouette index					Davies-Bouldin index				
	number of clusters					number of clusters				
	5	7	10	15	20	5	7	10	15	20
σ_{cos}	**0.10**	0.07	0.07	0.05	0.06	**1.81**	1.87	1.87	1.87	1.86
σ_{JC}	0.05	0.05	0.04	0.03	0.03	1.91	1.90	1.92	1.93	1.92
$\sigma_{THESIM_{W\&P}}$	0.04	0.03	0.03	0.03	0.03	1.90	1.90	1.90	1.90	1.91
$\sigma_{THESIM_{modified}}$	0.04	0.03	0.02	0.03	0.03	1.90	1.90	1.90	1.90	1.91
$\sigma_{SEMSIM_{W\&P}}$	0.06	0.03	0.04	0.03	0.03	1.83	1.84	1.85	1.85	1.84
$\sigma_{SEMSIM_{modified}}$	0.05	0.03	0.04	0.04	0.06	1.86	1.85	1.86	1.86	**1.81**

The results of the clustering are evaluated using Silhouette and Davies-Bouldin indices. Table 6.6 illustrates the results. As can be seen from this table SEMSIM based on Wu-Palmer semantic similarity metric outperforms cosine similarity in terms of both cluster validity indices for the correct number of classes, that is 5.

Table 6.6 : Silhouette and Davies-Bouldin index of Dataset2 using all terms on the documents

Similarity Measure	Silhouette index				Davies-Bouldin index			
	number of clusters				number of clusters			
	4	5	10	15	4	5	10	15
σ_{cos}	0.16	0.13	0.12	0.11	1.78	1.81	1.82	1.83
$\sigma_{SEMSIM_{W\&P}}$	0.10	**0.23**	0.12	0.03	1.73	**1.54**	1.65	1.76

To verify the clustering results in terms of clusters' content we further examined them to compare with pre-defined cluster assignments. Tables 6.9 and 6.10 illustrates the content of clusters produced by cosine and SEMSIM based on Wu-Palmer semantic similarity measure, respectively. The representation of clusters is produced using following steps: five documents from each cluster that have the highest similarity between the rest of the documents in the same cluster are selected. The content of each cluster is then represented by the most important terms (terms with the highest normalized TF weights) appear on that documents. As can be seen from results, cosine similarity measure produces the clustering results that are more homogeneous according to the given cluster topic.

The third part of this experiment set is conducted to evaluate the performance of the Jiang-Conrath hybrid semantic similarity measure as described in Equation 3.8. The time complexity of this document semantic similarity metric is $O(N) = N^2 \cdot c^2 \cdot d^2 \cdot e^2$ where N is the number of documents in a document set, c is the number of words used to represent documents in calculations, d is an overall depth of the taxonomy tree and e is the maximum number of child nodes in the taxonomy tree. Due to this high time complexity, the experiments are conducted only for the smallest document set, that is *Dataset2*. Tables 6.8 and 6.8 show the results. As can be seen from the results, SEMSIM using Jiang-Conrath similarity metric outperforms all others in terms of Entropy index, but still performs worse the cosine similarity metric in terms of Purity and Jaccard indices. Evaluation of clustering solution using Silhouette and Davies-Bouldin cluster validity indices shows that SEMSIM based on Jiang-Conrath hybrid similarity metric performs better than all other metrics. As stated in [12], the reason for this can be the fact that using density and depth factors of the ontology tree in combination with the information content factor as a decision factor outperforms both information content approach proposed in [18] and traditional edge counting methods. However, high time complexity of this method makes the pairwise document similarity calculation very expensive.

Table 6.7 : The results of manually clustered documents of Dataset2 with Hybrid Corpus-Based Semantic Similarity metric

Similarity Measure	Dataset2		
	E	*P*	*J*
$\sigma_{SEMSIM_{corpus}}$	0.57	0.41	0.18
$\sigma_{THESIM_{corpus}}$	0.52	0.46	0.22

Table 6.8 : Silhouette and Davies-Bouldin indices of Dataset2 for different number of clusters with Hybrid Corpus-Based Semantic Similarity metric

Similarity Measure	Silhouette index number of clusters				Davies-Bouldin index number of clusters			
	4	5	10	15	4	5	10	15
$\sigma_{SEMSIM_{corpus}}$	0.23	0.14	0.05	-0.01	1.61	1.72	1.82	1.83
$\sigma_{THESIM_{corpus}}$	0.27	0.21	0.10	0.07	1.67	1.73	1.82	1.85

Table 6.9 : The content of the clusters obtained using cosine similarity measure

Cluster 1 31 pages	Cluster 2 124 pages	Cluster 3 121 pages	Cluster 4 103 page	Cluster 5 108 pages
Sezen Aksu's Concert	Squint	fish in aquarium	Aşık Veysel	Honda's Prize
Orhan Pamuk's book	mouth cancer	chemistry in aquarium	theater festival	Fiat
book fair	asthma	bees	symphony orchestra	new Seat Ibiza
photography exhibition	tonsillitis	ants	exhibition	Ferrari
science-fiction competition	oversleeping	goose	museum formation	rent-a-car drive

Table 6.10 : The content of the clusters obtained using the SEMSIM based on Wu-Palmer semantic similarity measure

Cluster 1 8 pages	Cluster 2 160 pages	Cluster 3 42 pages	Cluster 4 139 pages	Cluster 5 138 pages
new printed books	alzheimer	theater festival	dog feeding	Opel
photography exhibition	hand eczema	symphony orchestra	bees	Mercedes cars
sculptural prize	leather dresses	Kemeraltı Bazaar	fish	Japanese museum
musical	winter clothing	Rock-and-Coke	what is BARF	second hand cars
concerts in Istanbul	fashion fair	theater stages	owls	undersea photographs

The fourth and last part of this experiment set is conducted to evaluate the LSA of the document sets. Each document set is treated as a distinct corpus, and LSA is done over the vector space model constructed from that corpus. The construction of the vector space model (which form a document matrix) is done in the following way: assume that documents in a document set D are represented as shown in Eq. 4.1. We set n (the number of words; vector space dimension) to 5000 and select 5000 most frequent words of the document set D. Thus the document matrix with dimensions $m \times 5000$ is formed, where m is the number of documents in a document set D. This document matrix is then passed to SVD operation and the dimensions of the matrix are reduced to $m \times 20$. After

having costructed the reduced document matrix, cosine similarity measure is used to calculate the paiwise similarities between documents using the formula descirebed in Eq. 4.7. Produced similarity matrices are clustered as described previously, and the clustering results are evaluated using different cluster validity indeces. Tables 6.11, 6.12, 6.13, 6.14 show the experimental results. It can be seen, that LSA method outperforms all the others in terms of every cluster validity indices. The LSA is generally used by search engines for matching the query to text datasets. It overcomes the problem of synonymy and polysomy, which generally cause mismatches in normal semantic similarity calculations. Even though LSA produces good clustering results, the connections between words and documents are lost after dimension reduction, thus the further analysis of the clusters becomes impossible.

Table 6.11 : The results of manually clustered documents of Dataset1 and Dataset2 using LSA

Similarity Measure	Dataset1			Dataset2		
	E	*P*	*J*	*E*	*P*	*J*
σ_{LSA}	0.53	0.54	0.23	0.28	0.79	0.53

Table 6.12 : Silhouette and Davies-Bouldin indices of Dataset1 for different number of clusters using LSA

Similarity Measure	Silhouette index number of clusters					Davies-Bouldin index number of clusters				
	5	7	10	15	20	5	7	10	15	20
σ_{LSA}	0.20	0.24	0.23	0.23	0.19	1.51	1.40	1.39	1.39	1.42

Table 6.13 : Silhouette and Davies-Bouldin indices of Dataset2 for different number of clusters using LSA

Similarity Measure	Silhouette index number of clusters				Davies-Bouldin index number of clusters			
	4	5	10	15	4	5	10	15
σ_{LSA}	0.51	0.41	0.43	0.42	1.16	1.23	1.11	1.07

Table 6.14 : Silhouette and Davies-Bouldin indices of Dataset3 for different number of clusters using LSA

Similarity Measure	Silhouette index number of clusters					Davies-Bouldin index number of clusters				
	5	7	10	15	20	5	7	10	15	20
σ_{LSA}	0.27	0.26	0.23	0.20	0.19	1.32	1.30	1.37	1.37	1.32

6.3.2 Second experiment set

The aim of this experiment set is to evaluate the proposed method and to compare it with cosine similarity metric in terms of clustering validity indices. Furthermore, the comparison is also done with document similarity measure described in Equation 4.5. The SEMSIM document semantic similarity measure is chosen because it produced better results in the previous experiment set.

The preprocessing stage is slightly modified for this set of experiments. As calculating taxonomy-based semantic similarity completely based on taxonomy usage, the overlapping between document words and used thesaurus was aimed to be maximized. That's why all words which could not be found in BalkaNet ontology are removed from documents. Further more, only noun words are selected to obtain meaningful relations between document and increase coincidence between document representation and human comprehension. This process highly reduce document set dimension, that's why we empirically set the number m of MFW to 10. For greater m some documents simply would not have enough terms to be represented with.

In the experimental tables cosine similarity is represented as σ_{cos}, proposed modified cosine similarity as σ_{CT} and document semantic similarity described in Equation 4.5 with Wu-Palmer similarity metric is represented as $\sigma_{SEMSIM_{W\&P}}$. We omit setting k value, taking the whole parent vector to provide full matching with Wu-Palmer similarity metric.

For the datasets with predefined class numbers Entropy, Purity and Jaccard indeces are used to evaluate the clustering results (notices that the best values are small for Entropy, high for Purity and Jaccard). Silhouette and Davies-Bouldin

indices are used to evaluate the clustering results of all datasets (notice that the best values are high for Silhouette and small for Davies-Bouldin). Clustering is done using different number of clusters. Table 6.15 illustrates the results of manually clustered document sets which are *Dataset*1 and *Dataset*2. These results shows that cosine similarity produces clustering solutions that match manual clustering the most. The results of proposed methods are a little worse. Tables 6.16, 6.17 and 6.18 presents the results of clustering in terms of the unsupervised cluster validity indices. As can be seen from the experimental tables, for all datasets the proposed method produces the best result when compared to cosine and SEMSIM similarity measures. Even if for the third document set, *Dataset3*, cosine similarity produced better Silhouette index, the proposed method outperforms it in terms of better correlation between Silhouette and Davies-Bouldin indices.

Table 6.15 : The results of manually clustered documents of Dataset1 and Dataset2

Similarity Measure	Dataset1			Dataset2		
	E	*P*	*J*	*E*	*P*	*J*
σ_{cos}	**0.69**	**0.42**	**0.15**	**0.16**	**0.93**	**0.77**
σ_{SEMSIM}	0.90	0.33	0.11	0.37	0.81	0.56
σ_{CT}	0.81	0.33	0.11	0.19	0.91	0.73

Table 6.16 : Silhouette and Davies-Bouldin indices of Dataset1 for different number of clusters

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

Table 6.17 : 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
σ_{cos}	0.15	0.14	0.17	0.18	1.75	1.73	1.70	1.67
σ_{SEMSIM}	0.21	0.17	0.15	0.15	1.58	1.59	1.64	1.60
σ_{CT}	0.24	0.23	0.24	**0.30**	1.57	1.53	1.50	**1.41**

Table 6.18 : 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
σ_{cos}	**0.24**	0.16	0.17	0.15	1.65	1.74	1.71	1.72
σ_{SEMSIM}	0.11	0.06	0.08	0.05	1.73	1.76	1.74	1.71
σ_{CT}	0.10	0.15	**0.20**	0.08	1.68	1.63	**1.56**	1.68

6.4 A Walk-Through Example

In order to illustrate the operational difference between proposed method and cosine similarity measure a simple experiment is conducted. To compare clustering results between similarity metrics we use the document set, described in Table 6.19, which consist of only 9 documents, each represented by equally weighted five terms. This documents are constructed manually with high correlation between words within documents.

Documents are grouped under three topics, which are “fruits and vegetables”,”medicine” and “animals”. Results of clustering using cosine similarity measure and proposed semantic similarity metric are illustrated in Table 6.20. Clustering using proposed similarity measures ends with results, that are exactly the same with the human comprehension.

All experiments in this experiment sets justifies the efficiency of the proposed method. The aim of reducing the computation time is also archived, which simplifies and speeds up the clustering process of large document sets.

Table 6.19 : Example: Small Document Set

d1	jam:0.2, cherry:0.2, apple:0.2, tree:0.2, pear:0.2
d2	potato:0.2, tomato: 0.2, garden:0.2, home:0.2,soil:0.2
d3	vegetables:0.2, tomato:0.2, spinach:0.2, pomogranate:0.2, grass:0.2
d4	doctor:0.2, clinic:0.2, pills:0.2, pacient:0.2, ache:0.2
d5	pharmacy:0.2, clinic: 0.2, pills:0.2, nurse:0.2, room:0.2
d6	influenza:0.2, home:0.2, illness:0.2, ache:0.2, ambulance:0.2
d7	cat:0.2, dog:0.2, home:0.2, pet:0.2, offspring:0.2
d8	fish:0.2, aquarium: 0.2, dog:0.2, home:0.2, sheep:0.2
d9	bear:0.2, soil:0.2, forest:0.2, crocodile:0.2, river:0.2

Table 6.20 : Clustering results using cosine and proposed semantic similarity measures

	Cosine Similarity	Proposed Similarity
cluster 1	d1, d2, d3, d9	d1, d2, d3
cluster 2	d4, d5	d4, d5 , d6
cluster 3	d6, d7, d8	d7, d8, d9

The experiments conducted throughout this study show that LSA corpus-based semantic similarity measure performs better than described taxonomy-based and hybrid document semantic similarity measures. However, it modifies the document vector space, making further analysis of document clusters impossible. Hybrid method proposed by Jaing&Conrath performs better than taxonomy-based similarity measures. This fact identifies the advantage of using the taxonomy tree density and depth information among with corpus statistics. However, the computation time of this method becomes very time costly because of many tree traversals. Single term document similarity measures identifies the predefined cluster number better than taxonomy-based semantic similarity measures. Experimental results have also shown that the proposed

taxonomy-based semantic similarity measure outperforms single term similarity measures in terms of quality of produced clusters. Moreover, it has lower time complexity than other document semantic similarity measures, based on pairwise word semantic similarity calculations.

7. CONCLUSION AND FUTURE WORK

Text (document) clustering has become an important part of web data organization with the rapid growth of the World Wide Web. Several online software applications, like search engines and web recommendation systems use document clustering to simplify their work. The document clustering is generally done over the similarity matrix of the given document set. Thus, the clustering results directly depend on the similarity measure used for estimation of the document similarity matrix. Besides the precise clustering results, the time complexity of the text clustering also gains the high importance when the amount of processed data is very high.

Document semantic similarity is the metric that can be used for text clustering. Even though semantic relations among documents seem to be carrying more information, single term similarity measures remain the main techniques for similarity calculations. Single term similarity measures are calculated directly from vector representation of the documents, thus they are simple and fast. Throughout this study, semantic similarities are investigated in terms of effects they produce over the clustering results of Turkish documents. Moreover, a new method for calculating document semantic similarity is proposed, which has a low time complexity.

The effects of the semantic and single term similarity measures on clustering of Turkish documents are compared and evaluated in the first experiment set of these study. This set of experiments has shown that single term similarity measures produce clustering results that are closer to pre-defined cluster labeling. However, document semantic similarity metric proposed in [10] and described in 4.5 produces more compact and separated clusters.

Besides the difference in clustering solutions, another core difference between cosine and semantic similarities is their time complexity. As cosine similarity

measure calculates the inter-document similarity directly from the document term vectors, its computation time is much less than the computation time of semantic similarity measures. The drawback in time complexity of the latter one is caused by the WordNet tree traversals, used in calculations of pairwise word semantic similarities.

To decrease the computation time of document semantic similarity a new method is proposed, which make use of content vectors of documents and cosine similarity measure. By offline calculation of the concept vectors, which includes WordNet tree traversals, the time complexity of proposed method is reduced to the time complexity of the cosine similarity measure. Experimental results, illustrated in the second experiment set of this study, have shown that proposed method outperforms both the cosine similarity and the semantic similarity measure proposed in [10] and described in 4.5.

Further research can be done on improvement of preprocessing stage, especially on WSD. The WSD specify the words which will be used in semantic similarity calculations that is why directly affect the clustering results. In our study we used simple disambiguation technique and improving it can cause different and better results.

The small number of semantic relations in Turkish BalkaNet prevents the usage of relations other than hypernym/hyponym. With the enlargement of Turkish BalkaNet, additional study can be conducted to examine semantic similarity metrics that combine more than one semantic relations (for example, meronym/holonym).

In addition to this, the text summarization and the definition of cluster topic can be investigated with the usage of the proposed document semantic similarity metric.

REFERENCES

[1] **Sen, P.**, **Namata, G.M.**, **Bilgic, M.**, **Getoor, L.**, **Gallagher, B. and Eliassi-Rad, T.**, 2008. Collective Classification in Network Data, Technical Report CS-TR-4905, University of Maryland, College Park.

[2] **Mccallum, A.K. and Mitchell, T.**, 2000. Text classification from labeled and unlabeled documents using EM, Machine Learning, pp. 103–134.

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

[4] **Hammouda, K.M.**, **Matute, D.N. and Kamel, M.S.**, 2005. CorePhrase: Keyphrase Extraction for Document Clustering, In IAPR 4th International Conference on Machine Learning and Data Mining, pp. 265–274.

[5] **Saraçoğlu, R.**, **Tütüncü, K. and Allahverdi, N.**, 2007. A fuzzy clustering approach for finding similar documents using a novel similarity measure, *Expert Systems with Applications*, **33(3)**, 600–605.

[6] **Jiang, Z.**, **Joshi, A.**, **Krishnapuram, R. and Yi, L.**, 2000. Retriever: Improving Web Search Engine Results Using Clustering, Technical report, University of Maryland Baltimore County.

[7] **Bade, K. and Nurnberger, A.**, 2006. Personalized Hierarchical Clustering, Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, IEEE Computer Society, Washington, DC, USA, pp. 181–187.

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

[9] **Wu, Z. and Palmer, M.**, 1994. Verb semantics and lexical selection, 32nd. Annual Meeting of the Association for Computational Linguistics, New Mexico State University, Las Cruces, New Mexico, pp. 133 –138.

[10] **Yucesoy, B. and Gunduz Oguducu, S.**, 2007. Comparison of Semantic and Single Term Similarity Measures for Clustering Turkish Documents, ICMLA '07: Proceedings of the Sixth International Conference on Machine Learning and Applications, IEEE Computer Society, Washington, DC, USA, pp. 393–398.

[11] **Lin, D.**, 1998. An Information-Theoretic Definition of Similarity, In Proceedings of the 15th International Conference on Machine Learning, Morgan Kaufmann, pp. 296–304.

[12] **Jiang, J.J. and Conrath, D.W.**, 1997. Semantic similarity based on corpus statistics and lexical taxonomy, International Conference Research on Computational Linguistics.

[13] **Hirst, G. and St-Onge, D.**, 1997, Lexical Chains as representation of context for the detection and correction malapropisms, `citeseer.comp.nus.edu.sg/109361.html`.

[14] **Turney, P.D.**, 2001. Mining the Web for Synonyms: PMI-IR Versus LSA on TOEFL* , Proc. European Conference on Machine Learning.

[15] **Landauer, T.K.**, **Foltz, P.W. and Laham, D.**, 1998. Introduction to Latent Semantic Analysis, *Discourse Processes*, **25**, 259–284.

[16] **Leacock, C. and Chodorow, M.**, 1998. Combining Local Context and WordNet Similarity for Word Sense Identification, *An Electronic Lexical Database*, 265–283.

[17] **Jiang, J.J. and Conrath, D.W.**, 1997, Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy, `cmp-lg/9709008`.

[18] **Resnik, P.**, 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy, In Proceedings of the 14th International Joint Conference on Artificial Intelligence, pp. 448–453.

[19] **Mihalcea, R. and Corley, C.**, 2006. Corpus-based and knowledge-based measures of text semantic similarity, In AAAI'06, pp. 775–780.

[20] **Budanitsky, A. and Hirst, G.**, 2006. Evaluating WordNet-based Measures of Lexical Semantic Relatedness, *Computational Linguistics*, **32(1)**, 13–47.

[21] **Madylova, A. and Gunduz Oguducu, S.**, 2009. Comparison of Similarity Measures for Clustering Turkish Documents, *Intelligent Data Analysis*, **13(5)**, in press.

[22] **Miller, A.G.**, 1990. WordNet: An on-line Lexical Database, *International Journal of Lexicography*, **3(4)**, 235–244.

[23] BalkaNet Project, http://www.ceid.upatras.gr/Balkanet/.

[24] **Orhan Bilgin, O.C. and , K.O.**, 2004. Building a Wordnet for Turkish, *Romanian Journal of Information Science and Technology Volume 7, Numbers 1-2, 2004*, 163–172.

[25] **Tognini-Bonelli, E.**, 2001. Corpus Linguistics at Work, volume 6 of *Studies in Corpus Linguistics*, Benjamins, Amsterdam.

[26] **Bilge Say, Deniz Zeyrek, K.O.U.z.**, 2002. Development of a Corpus and Treebank for Presentday Written TUrkish, Eleventh International Conference of Turkish Linguistics.

[27] **Roget, P.M.**, 1977. Roget's Thesaurus of English Words and Phrases, P Shalom Pubns; Indexed edition.

[28] **Bernard, J.N.L.**, editor, 2007. The Macquarie Thesaurus, Macquarie Dictionary Publishers, 2nd edition.

[29] **Muller, P.**, **Hathout, N. and Gaume, B.**, 2006. Synonym Extraction Using a Semantic Distance on a Dictionary, Proceedings of TextGraphs: the Second Workshop on Graph Based Methods for Natural Language Processing, Association for Computational Linguistics, New York City, pp. 65–72.

[30] **Inkpen, D.**, 2007. Semantic Similarity Knowledge and Its Applications, *Studia Universitatis Babes-Bolyai, Informatica*, **LII(1)**.

[31] **Wan, S. and Angryk, R.A.**, 2007. Measuring semantic similarity using wordnet-based context vectors., SMC, IEEE, pp. 908–913.

[32] **Li, Y.**, **Bandar, Z.A. and Mclean, D.**, 2003. An approach for measuring semantic similarity between words using multiple information sources, *Knowledge and Data Engineering, IEEE Transactions on*, **15(4)**, 871–882.

[33] **Seco, N.**, **Veale, T. and Hayes, J.**, 2004, An Intrinsic Information Content Metric for Semantic Similarity in WordNet.

[34] **Halkidi, M.**, **Nguyen, B. and Varlamis, I.**, 2003. THESUS: Organizing Web Document Collections Based on Link Semantics, *VLDB J*, **12**, 320–332.

[35] **Hersh, W. and Bhupatiraju, R.T.**, 2003b. TREC Genomics track overview., The twelfth Text Retrieval Conference, TREC 2003b, National Institute of Standards and Technology, pp. 1–14.

[36] **van Rijsbergen, C.J.**, 1979. Information retrieval, Butterworths, London, 2 edition.

[37] **Oflazer, K.**, 1994. Two-level description of Turkish Morphology, *Literary and Linguistic Computing*, **9(2)**, 137–148.

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

[39] **Tan, P.N.**, **Steinbach, M. and Kumar, V.**, 2005. Introduction to Data Mining, Addison Wesley.

[40] **Davies, D. and Bouldin, D.**, 1979. A Cluster Separation Measure, *IEEE Transactions on Pattern Analysis and machine Intelligence*, **1(2)**, 224–227.

[41] **Speer, N.**, **Spieth, C. and Zell, A.**, 2005. Biological Cluster Validity Indices Based on the Gene Ontology, **A.F.F. et al.**, editor, Advances in Intelligent Data Anylsis VI: 6th International Symposium on Intelligent Data Analysis (IDA 2005), volume 3646 of *Lecture Notes in Computer Science (LNCS)*, Springer, pp. 429–439.

[42] **Rousseeuw, P.**, 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *J. Comput. Appl. Math.*, **20(1)**, 53–65.

[43] **Zhao, Y. and Karypis, G.**, 2004. Empirical and Theoretical Comparisons of Selected Criterion Functions for Document Clustering, *Mach. Learn.*, **55(3)**, 311–331.

[44] **Jain, A.K.**, **Murty, M.N. and Flynn, P.J.**, 1999. Data clustering: a review, *ACM Comput. Surv.*, **31(3)**, 264–323.

CURRICULUM VITA

Candidate's full name:	Ainura Madylova
Place and date of birth:	Jalalabad / KYRGYZSTAN, 10 Mart 1986
Universities and Colleges attended :	Middle East Technical University (Bs.)

Publications:

Madylova, A.; Gunduz Oguducu S., Comparison of Similarity Measures For Clustering Turkish Documents, Intelligent Data Analysis, vol. 13, no. 5, in press, 2009

Madylova, A.; Gunduz Oguducu S., A Taxonomy based Semantic Similarity of Documents using the Cosine Measure, submitted to ISCIS 2009, METU, Ankara, Turkey