İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

TÜRKÇE DOKÜMANLAR İÇİN ANLAMSAL BENZERLİK HESAPLAMA YÖNTEMİ

YÜKSEK LİSANS TEZİ

Müh. Bülent YÜCESOY

Anabilim Dalı : BİLGİSAYAR MÜHENDİSLİĞİ

Programı : BİLGİSAYAR MÜHENDİSLİĞİ

HAZiRAN 2007

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

TÜRKÇE DOKÜMANLAR İÇİN ANLAMSAL BENZERLİK HESAPLAMA YÖNTEMİ

YÜKSEK LİSANS TEZİ

Müh. Bülent YÜCESOY
(504041506)

Tezin Enstitüye Verildiği Tarih : 7 Mayıs 2007
Tezin Savunulduğu Tarih : 12 Haziran 2007

Tez Danışmanı : **Yrd.Doç.Dr. Şule Gündüz ÖĞÜDÜCÜ**

Diğer Jüri Üyeleri **Yrd.Doç.Dr. A. Şima Etaner UYAR (İ.T.Ü.)**

Yrd.Doç.Dr. Songül ALBAYRAK (Y.T.Ü.)

HAZiRAN 2007

ÖNSÖZ

Lisans ve yüksek lisans eğitimlerimi tamamladığım İTÜ Bilgisayar Mühendisliğinde benim üzerimde emeği geçen bütün hocalarıma teşekkürü bir borç biliyorum.

Bu çalışmanın başından sonuna beni destekleyen, bilgi ve tecrübeleriyle bana yol gösteren, değerli hocam Yrd. Doç. Dr. Şule Gündüz Öğüdücü' ye teşekkürü bir borç biliyorum.
Geliştirdikleri çizge tabanlı demetleme yönteminin kullanılmasına izin verdikleri için Yrd. Doç. Dr. Şule Gündüz Öğüdücü ve Yrd. Doç. Dr. A. Şima Etaner Uyar'a teşekkür ediyorum.

Balkanet projesi kapsamında Türkçe kavramsal sözlük çalışmaları için ve bu çalışmalarını bu araştırmada kullanılmasına izin verdikleri için Prof. Dr. Kemal Oflazer başkanlığında Sabancı Üniversitesi' ne teşekkür ediyorum.

Cümledeki anlama göre uygun sözcük kökünü seçen postagger yazılımını geliştiren ve bu araştırmada kullanımına izin veren Yrd. Doç. Dr. Deniz Yuret başkanlığında Koç Üniversitesi' ne teşekkür ediyorum.

Çalışma boyunca sundukları olanaklar ve verdikleri destek nedeni ile başta Sinan Çayır olmak üzere bütün İTÜ Bilgi İşlem Daire Başkanlığı çalışanlarına teşekkür ederim. İnternet sayfalarının benzerliklerine göre sınıflandırılması aşamasında bana yardım eden asistan öğrencilere de çok teşekkür ediyorum.

Benim için en önemlisi, yıllardır bana sevgi ve destek verip bana olan güvenlerini hiçbir zaman eksik etmeyen aileme minnettarım.

Haziran 2007 Bülent YÜCESOY

İÇİNDEKİLER

KISALTMALAR ..... **iv**
TABLO LİSTESİ ..... **v**
ŞEKİL LİSTESİ ..... **vi**
SEMBOL LİSTESİ ..... **vii**
ÖZET ..... **viii**
SUMMARY ..... **x**

1. GİRİŞ **1**

2. MEVCUT YÖNTEMLERİN İNCELENMESİ **5**
2.1 Kavramsal Sözlük Kullanan Yöntemler 5
2.2 Sözlük Kullanmayan Yöntemler 8

3. ÖNERİLEN YÖNTEMİN AÇIKLANMASI **12**
3.1 Balkanet ve Türkçe Kavramsal Sözlük 13
3.2 Ön İşleme 16
3.2.1 Html Sayfalarının Ayrıştırılması 17
3.2.2 Gürültünün Atılması 17
3.2.3 Sözcük Kökü Bulunması 18
3.3 Benzerlik Hesaplanması 20
3.3.1 Sözcüklerin Seçimi 20
3.3.2 Sözcükler Arası Benzerlik 22
3.3.3 Dokümanlar Arası Benzerlik 24
3.4 Çizge Tabanlı Demetleme 24

4. YAPILAN DENEYLER **26**
4.1 Deneylerde Kullanılan Demetleme Göstergeleri 27
4.2 İlk Grup Deneylerin Sonuçları 30
4.3 İkinci Grup Deneylerin Sonuçları 34

5. SONUÇLAR VE TARTIŞMA **38**

KAYNAKLAR **41**

ÖZGEÇMİŞ **43**

KISALTMALAR

ID	: IDentification
POS	: Part Of Speech
ILI	: Internal Language Index
ILR	: Internal Language Relations
DEF	: DEFinition
BCS	: Balkanet Common Set
SNOTE	: Special NOTE
HTML	: Hyper Text Markup Language
XML	: EXtended Markup Language
UTF	: Universal Transformation Format
DB	: Davies-Bouldin
İTÜ	: İstanbul Teknik Üniversitesi
BİDB	: Bilgi İşlem Daire Başkanlığı

TABLO LİSTESİ

		Sayfa No
Tablo 3.1	Türkçe Kavramsal Sözlük Kayıt Örnekleri ……	13
Tablo 3.2	Balkanet Bünyesindeki Bütün Sözcük İlişkileri……..…	15
Tablo 3.3	Balkanet Türkçe Kavramsal Sözlük İstatistiği…….….	16
Tablo 3.4	Dokümanlar Arası Benzerlik Hesabı Örneği …..……..	24
Tablo 4.1	Demetleme Jaccard Göstergesi Parametreleri ……	27
Tablo 4.2	İnternet Sitesinden Seçilen Alt Kümenin Özellikleri …..……..…	31
Tablo 4.3	İnternet Sitesi Alt Kümesi Deney Sonuçları ……..……	32
Tablo 4.4	İnternet Sitesi Bütün Sayfalar Demetleme Göstergeleri ……..…	34
Tablo 4.5	Deney 5 Demetleme Sonuçlarının İçeriği ……...	35
Tablo 4.6	Deney 6 Demetleme Sonuçlarının İçeriği ……...	35
Tablo 4.7	Deney 7 ve Deney 8 Demetleme Sonuçlarının İçeriği ……...	37

ŞEKİL LİSTESİ

Sayfa No

Şekil 2.1 **:** Örnek Kavramsal Sözlük..... 5
Şekil 2.2 **:** Kosinüs Benzerliği Hesabı 10
Şekil 3.1 **:** Sözcük Etiketleyici Örnek Girdisi ve Çıktısı..... 20
Şekil 3.2 **:** Önerilen Yöntem ile Wu-Palmer Formülü Kıyaslaması..... 23
Şekil 3.3 **:** Örnek Wu-Palmer ve Önerilen Yöntem Sözcük Benzerlikleri... 23

SEMBOL LİSTESİ

TF_{ij} **:** i. doküman j. sözcüğün terim frekansı
DF_j **:** j. sözcüğü içeren doküman frekansı
IDF_j **:** j. sözcüğü içeren ters doküman frekansı
W_{ij} **:** i. doküman j. sözcük terim ağırlığı
$D_1, \ldots, D_n$ **:** Kosinüs benzerliği uygulanan doküman uzayındaki dokümanlar
$T_1, \ldots, T_t$ **:** Kosinüs benzerliği uygulanan doküman uzayındaki terimler
W11, …,W14 : İkili doküman benzerlik hesabı örnek 1. doküman 4 anahtar sözcüğü
W21, …,W24 : İkili doküman benzerlik hesabı örnek 2. doküman 4 anahtar sözcüğü
S11, …, S44 **:** İkili doküman benzerlik hesabı örnek sözcük benzerlik değerleri
D_{nc} **:** Bütün demetler hesaplanarak bulunan Dunn göstergesi
$d(c_p, c_r)$ **:** Belirtilen demetler arasındaki farklılık değeri
diam(c) **:** Belirtilen demetlemenin çapı
R_g , R_{gh} **:** Davies-Bouldin göstergesinde belirtilen dokümanların R değeri
DB_{nc} **:** Bütün demetler hesaplanarak bulunan Davies-Bouldin göstergesi

TÜRKÇE DOKÜMANLAR İÇİN ANLAMSAL BENZERLİK HESAPLAMA YÖNTEMİ

ÖZET

Bu tez çalışmasında, Türkçe dokümanlar için yeni bir anlamsal benzerlik hesaplama yöntemi önerilmektedir. Anlamsal benzerlik hesabı, yakın anlama gelen dokümanların tespiti açısından çok önemlidir. Çalışmada incelenecek Türkçe dokümanlar internet ortamı üzerindeki sayfalar olacaktır. İnternet ortamının seçilme nedeni, internet üzerinde yaygın olarak doküman bulunabilmesidir. Fakat bu çalışma internet ortamı haricindeki dokümanlara da uygulanabilmektedir. Önerilen yeni yöntem, Türkçe kavramsal sözlük kullanmaktadır. Önerilen yöntem, kavramsal sözlük kullanan mevcut anlamsal benzerlik yöntemleri ile kıyaslanmaktadır. Bu sırada, kavramsal sözlük bünyesindeki bütün ilişkilerin kullanılıp kullanılmamasının cevabı da aranmaktadır. Önerilen yöntem, kavramsal sözlük kullanmayan mevcut anlamsal benzerlik yöntemleri ile de kıyaslanmaktadır.

Anlamsal benzerlik hesaplamak için, kavramsal sözlük kullanan veya kullanmayan iki türde yöntem mevcuttur. Kavramsal sözlük kullananlar da kendi içinde ayrıt temelli ve düğüm temelli olarak ikiye ayrılmaktadır. Düğüm temelli yöntemler, ayrıt temelli yöntemlerin ayrıtlara puan verilmiş halleri olarak yorumlanabilir ve birbirlerinden çok farklı yöntemler değildir. Bu nedenle, bu çalışma kapsamında ayrıt temelli yöntemler üzerinde çalışılmaktadır. Ayrıt temelli yöntemler arasında, temel ayrıt sayma yöntemi ve Wu-Palmer yöntemi gösterilebilir. Wu-Palmer yöntemi daha gelişmiş bir yöntem olduğu için, bu çalışmada geliştirilen yöntem Wu-Palmer yöntemi ile kıyaslanacaktır. Yapılan deneylerde önerilen yöntem, Wu-Palmer yöntemi ile elde edilen sonuçlardan daha başarılı sonuçlar üretmiştir.

Kavramsal sözlük kullanımı sırasında, bütün sözcük ilişkilerinin kullanılıp kullanılmaması incelenmiştir. Bu sorunun oluşma nedeni, mevcut çalışmalarda genellikle hep sadece üst kavram ilişkisinin kullanılmasıdır. Üst kavram ilişkisi dışında da birçok ilişki bulunmaktadır ve bütün ilişkilerin incelenmesinin fayda sağlayıp sağlamadığı araştırılmıştır. Kavramsal sözlükteki bütün ilişkileri inceleyince işlem zamanı oldukça uzamaktadır. Bununla birlikte, bütün ilişkileri inceleyince elde edilen kazanç aynı oranda büyük olmamaktadır. İlişkiler arasındaki farkları, ağırlık vererek yansıtmak suretiyle çalışılsa belki daha başarılı olabilirdi ama bu tamamen ayrı bir çalışma konusudur. Bütün ilişkilerin incelenmesinden büyük fayda sağlanamaması ardından kavramsal sözlükteki bütün ilişkileri incelemenin gerekli olmadığı sonucuna varılmıştır ve ilerideki deneyler sadece üst kavram ilişkisine bakılarak gerçeklenmiştir.

Bu çalışma kapsamında anlamsal benzerlik hesabı için kavramsal sözlük kullanmak veya kullanmamak arasındaki fark da incelenmek istenmiştir. Kavramsal sözlük kullanmadan hesap yapan iki temel yöntem söz konusudur. Bu yöntemler, dokümanlar arası benzerlik hesaplayan Jaccard ve kosinüs yöntemleridir. Jaccard yöntemi çok temel bir yöntemdir ve uygulaması da çok hızlı olduğu için yaygın olarak kullanılmaktadır. Ancak Jaccard yöntemini kullanabilmek için geniş bir metin

kütüphanesi olması gerekmektedir. Türkçe için hazırlanmış böyle bir kütüphane bulunmadığı için Jaccard yöntemini çalışmada kullanmak mümkün değildir. Kosinüs yöntemi ise Jaccard yönteminden daha gelişmiş bir yöntemdir. Ayrıca uygulaması hızlı olduğu için kosinüs yöntemi de oldukça yaygın olarak kullanılmaktadır. Bütün bu nedenlere dayanarak, kavramsal sözlük kullanmayan yöntemleri incelerken kosinüs yöntemi dikkate alınmıştır. Kavramsal sözlükler dil bilimciler tarafından ortaya konan net ve güvenilir kaynaklardır. Sözlüğün yetersiz kalmaması durumunda kullanılan sözlüğe dayalı yöntemin performansına da bağlı olarak sözlüksüz yöntemden iyi sonuç vermesi beklenmektedir. Deneyler sonucunda ise sözlüğün yeterli olduğu durumdaki deneyde de diğer deneylerde de kosinüs benzerliği daha iyi sonuçlar vermiştir. Bu durumun nedeni, önerilen yöntem ve Wu-Palmer yönteminin başarımının yeterli olmaması olabilir. Önerilen yöntem ve Wu-Palmer yöntemi kosinüs benzerliği kadar iyi sonuç vermediyse bile, ayrıt ve düğüm temelli yöntemleri birlikte inceleyen sözlüğe dayalı daha gelişkin yöntemlerin bu konudaki gelecek çalışmalarda tekrar incelenmesi makul olacaktır.

Benzerlikler hesaplandıktan sonra çok önemli bir adım, sonuçların yorumlanmasıdır. Yorumlama işlemini kolaylaştırmak için benzerliklere dayanarak dokümanlar arasında daha benzer dokümanlar aynı grupta olacak şekilde demetleme yapılmıştır. Çalışmada ikili doküman benzerlikleri elde edildiği için çizge tabanlı bir demetleme yöntemi kullanılmıştır. Çalışma kapsamında iki grup deney yapılmıştır. İlk grup deneyde, bütün dokümanlar arasından ufak bir alt küme seçilmiştir. Bu dokümanlar sağlıklı demetler oluşturmak için içeriklerine göre elle demetlenmiştir. Elle demetleme işlemi bu konuda uzman 3 kişi tarafından yapılmıştır. Deneylerde elde edilen demetleme sonuçları elle yapılan demetleme ile kıyaslanmıştır. İkinci grup deneyler ise bütün dokümanlar için incelenmiştir. Fakat bu durumda deneylerde kullanılan doküman sayısı fazla olduğundan insanlar tarafından bu dokümanların demetlenmesi zordur. Bu nedenle demetleme sonuçları, bu alanda kullanılan mevcut demetleme göstergeleri ile değerlendirilmiştir. Bu sırada 4 tane demet değerlendirme göstergesi kullanılmıştır. Bunlar; Jaccard göstergesi, Dunn göstergesi, Davies-Bouldin göstergesi ve Silhouette göstergesidir. Deneylerde önerilen yöntem Wu-Palmer yönteminden başarılı sonuçlar vermiştir. Kosinüs benzerliğinde ise önerilen yöntemden de Wu-Palmer yönteminden de başarılı sonuçlar elde edilmiştir. Sonuç olarak şu an için tek sözcüğe dayalı yöntemler, Türkçe dokümanlar arasında anlamsal benzerlik hesaplamak için daha elverişli yaklaşımlar olarak gözükmektedir.

SEMANTIC SIMILARITY METHOD FOR TURKISH DOCUMENTS

SUMMARY

In this thesis, a new semantic similarity calculation method is examined for Turkish documents. Semantic similarity calculation is very important to determine semantically similar documents. Turkish documents used in this research are documents on internet. The reason of internet choice is that there are widespread documents existing on the internet. However, this research can be implemented for any other kind of document, either on internet or not. The proposed new method is based on Turkish lexical dictionary. This method is compared with existing lexical methods. While examining lexical methods, the lack of considering all relations is also discussed. In addition, the new method is compared with existing non-lexical methods to find which one is better.

There are two kinds of algorithms to calculate semantic similarity, lexical methods and non-lexical methods. Lexical methods are categorized into edge-based and node-based algorithms. Node-based algorithms are not so different from edge-based algorithms and they can be considered like edge-based ones having different edge weights. Therefore, we focus on edge-based lexical methods in this research. Important edge-based lexical methods are basic edge counting method and Wu-Palmer method. It is stated that Wu-Palmer method is better than the basic edge counting method, so we compare Wu-Palmer method with the new lexical method. In our experiments, our new method outperforms Wu-Palmer method.

In this research, it is questioned whether to care all semantic relations in taxonomy or not. This is examined because lexical methods generally just care about IS-A relations. However, there are many relations other than IS-A relation and utilization of considering all relations is investigated. Calculations with all relations make execution time higher. Additionally, utilization is not as big as execution time. Relation differences due to variant weights can improve the results but it is exactly another kind of research topic. Therefore, It is decided that considering all relations is not necessary. Afterwards, it is just worked with IS-A relations in further experiments.

Another question in this research is whether using lexical algorithms or non-lexical algorithms for semantic similarity. There are two widespread non-lexical algorithms. These are; Jaccard similarity method and cosine similarity method. Jaccard method is too basic and widespread due to its easy implementation. A huge text library of Turkish language is needed to implement Jaccard similarity. As no such Turkish text library exists, Jaccard similarity could not be implemented in this research. Cosine similarity is a better similarity method than Jaccard method. Cosine similarity is also widespread because it is easy to implement too. According to these explanations, it is decided to use cosine similarity for non-lexical methods. Lexical dictionaries are obvious, determined and reliable objects formed by language experts. If the dictionary is not small and lexical method performs well, it is expected to have better results with lexical methods than non-lexical ones. Our experimental results show

that non-lexical methods perform beter than the lexical ones whether the dictionary is enough or not. The reason of this result may be due to low performance of new proposed method and Wu-Palmer method. Although examined lexical methods may not perform better than the cosine similarity, working with hybrid lexical methods considering both edge-based and node-based algorithms can be examined in future researchs.

After similarity calculations, the important step is the analysis of results. To make this analysis easier, documents are clustered such that more similar ones will be in the same group. A graph based clustering method is used since pairwise similarities are examined. We have demonstrated our results on two sets of experiments. In our first set of experiments, a small subset of documents is selected. To introduce the accurate clusters, the documents are manually classified by contents. This manual clustering had been performed by three experts in the field. The clustering results produced by using our method as a similarity measure are compared to the manually produced clustering result. In our second set of experiments, we considered all the documents. As document number increased, manual clustering became more difficult. Therefore existing cluster validity indices in the literature are used to analyze the clustering results. We rely on four clustering measures – the Jaccard index, the Dunn index, the Davies-Bouldin index and the Silhouette index. Our proposed similarity measure outperforms the Wu-Palmer similarity measure. However, cosine similarity measure performs better than the Wu-Palmer and our proposed similarity measure. For the present, single-word models seem a better approach for calculating similarities between Turkish documents.

1. GİRİŞ

Dokümanlar arası benzerlik hesaplama işlemi, yaygın uygulama alanına sahiptir ve önemli bir gereksinimdir. Örneğin; internet ortamında kullanıcılar tarafından girilen anahtar sözcüklere en çok benzeyen dokümanların kullanıcıya listelenmesi beklenmektedir. Bilim toplumlarında arama ve araştırmanın hayati bir önemi vardır. Ayrıca her zaman için, verilerdeki benzerliklere dayanarak veriler sınıflandırılabilir. Dolaylı şekilde farklılıkların belirginleşip ayırt edici hale gelmesi sayesinde de tanılama yapılabilir. Bütün bu uygulama alanlarında yapılabileceklerin hızı ve performansı, benzerlik tespitinin başarımından etkilenecektir. Bilgi çağındaki dünyamızda bilgiye erişimin önemli bir unsuru olan dokümanlar arasındaki benzerlik, bu açıdan büyük önem taşımaktadır.

Dokümanların benzerliği denildiğinde akla ilk gelen dokümanların anlamsal benzerliğidir. Çünkü dokümanlar belirli bir konuyu anlatıp belirli bir anlam ifade ederler. Önemli olan da dokümanın içeriğine ve taşıdığı anlama göre benzerini bulmaktır. Anlamsal benzerliği hesaplamak için öncelikle anlam taşıyan temel parça olan sözcüklerin benzerliği incelenmektedir. Her doküman da sözcüklerden oluştuğu için, sözcükler arası benzerlikler birleştirilerek dokümanlar arası benzerlik hesaplanmaktadır. Bu çalışma kapsamında bahsedilen sözcüklere dayalı doküman benzerliği yöntemi, her çeşit dokümanın benzerliğini hesaplamak için uygulanabilecektir.

Doküman benzerliğini hesaplamak için sözcük benzerliğini kullanmak çok temel olması nedeniyle avantajlıdır. Fakat bu sistemin dezavantajları da vardır. Örneğin; dokümanda bazı sözcüklerin gereksiz kullanımı söz konusu olabilir. Gereksiz sayıda sözcük kullanımı hesaplamalarda oldukça yanıltıcı olabilir. Dokümanda bütün sözcüklerle işlem yapmak yerine dokümanı daha iyi nitelediği düşünülen anahtar sözcüklerle işlem yapılması sık uygulanan bir yaklaşımdır. Bu durumda anahtar sözcük seçimi önem kazanacaktır. Anahtar sözcük seçimi için dokümanda en çok kullanılan sözcükler dikkate alınırsa, gereksiz kullanımlar sebebiyle yöntemin başarımının düşmesi olağandır. Sözcük benzerliğini kullanmanın bir diğer

dezavantajı sözcüklerin çok çeşitli ön işlemlere ihtiyaç duymasıdır. Sözcüklerin Türkçe olup olmadığı, yanlış yazılıp yazılmadığı gibi pek çok kontrole ihtiyaç vardır. Bu kontrollerin ardından düzeltmeler yapılarak sözcükler kullanılacaktır fakat böylelikle düzeltmelerin başarımı da çalışmanın başarımını en başından etkilemiş olacaktır. Belirtilen ön işlemeler, araştırma kapsamında kullanılan araçlarla gerçeklenmektedir.

Dokümanların anlamsal benzerliğini hesaplamak için, çok kullanılmakta olan iki temel yaklaşım vardır. Bunlardan ilki, kavramsal sözlüklerin kullanımıdır [1–3]. Kavramsal sözlükler, her sözlük gibi bir dile ait sözcükleri ve anlamlarını açıklar. Sözcüklerin anlamını açıklarken sözcüklerin ait olduğu kavramlar arasındaki anlamsal ilişkileri kullanır. Sözlükteki anlamsal ilişkilere örnek olarak; eş anlam, zıt anlam, üst kavram, alt kavram gibi ilişkiler belirtilebilir. Aralarında herhangi bir anlamsal ilişki tanımlanmış olan sözcükler birbirine benzer olacaktır. Bu düşünceden hareketle, kavramsal sözlükler benzerlik hesaplamak için kullanılmaktadır.

Dokümanların anlamsal benzerliği, sözlük kullanmayan yöntemler [4–7] sayesinde de hesaplanabilir. Sözlük kullanmayan yöntemler arasında yaygın olarak incelemeye alınan parametre, sözcüklerin niceliğidir. Bu yöntemlerde bir dile ait pek çok dokümandan oluşan geniş metin kütüphaneleri incelenerek sözcüklerin o dildeki kullanım sıklığı belirlenmektedir. Daha sonra bu kulanım sıklıkları kullanılarak benzerlik yöntemine göre benzerlik değeri hesaplanabilir. Örneğin; sözcüklerin birlikte kullanım sıklığını, ayrı kullanım sıklığına bölerek bir benzerlik değeri önermek mümkündür. Metin kütüphanesinin büyüklüğü oldukça önemlidir çünkü ancak büyük bir kütüphanede hesaplanan nicelikler gerçeğe daha fazla yakınsayacaktır. Böylelikle yöntemlerin yanılma ihtimalini azaltacak ve daha sağlıklı sonuçlar alınabilecektir [7]. Çalışmanın yapıldığı dönemde Türkçe için bu şekilde hazır bir kütüphane bulunmamaktadır. İnternet ortamındaki Türkçe dokümanların taranması ile bu kütüphanenin oluşturulması düşünülebilir. Ancak bu işlem, pratik uygulama süresi olarak çok elverişli olmadığı gibi tamamen ayrı bir çalışma konusudur. Bu nedenle, geniş metin kütüphanesi kullanılamayacak ve sadece araştırmada incelenecek olan internet sitesine ait bütün sayfalar ile deneyler yapılacaktır.

Kavramsal sözlükler, dil bilimciler tarafından sözcükler ve kavramlar üzerinde yapılan incelemeler sonucunda ilişkileri tanımlanmış net ve güvenilir kaynaklardır.

Kavramsal sözlükler, benzerlik hesaplamasında da yaygın olarak kullanılmaktadır. Bu nedenle bu çalışmada kavramsal sözlük kullanan yöntemler üzerine araştırma yapılmaktadır. Kavramsal sözlük kullanan yöntemler araştırılırken, ayrıt temelli bakış açısında çalışmak hedeflenmiş ve yaygın olarak kullanılan Wu-Palmer yöntemi ile kıyaslanacak yeni bir yöntem önerilmiştir [3]. Benzerlik hesaplaması sırasında, kavramsal sözlükteki bütün ilişkilere bakılması veya bakılmaması da çalışma kapsamında incelenecektir. Bütün avantajlarına rağmen kavramsal sözlükler, içerdikleri sözcük sayısının az olması ile yetersiz kalabilirler. Sözlüğün yetersiz kalmaması durumunda, sözlüğe dayalı yöntemin performansına da bağlı olarak sözlüksüz yöntemlerden daha sağlıklı sonuç vermesi beklenebilir. Bu beklentinin sonucunu görebilmek için sözlüğün yetersiz kalmadığı durum da gerçeklenerek deney yapılacaktır. Bunun için dokümanları temsil etmek için seçilen anahtar sözcükler, sözlükte bulunan sözcükler arasından seçilmiştir.

Bulunan benzerlik değerlerinin yorumlanması çok önemlidir. Sistem başarımını yorumlamak için, sonuçları hazır olarak bilinen veriler üzerinde deneyler yapılıp deney sonuçlarının gerçek sonuçlarla kıyaslanması düşünülebilir. Verilerin benzerliklerini önceden tespit etmek için birkaç kişiden oluşan insan topluluğu dokümanlar arasındaki benzerlikleri inceleyebilir. Burada bir grup insanın çalışması önemlidir çünkü göreceden uzak herkesin üzerinde anlaştığı mantıklı bir sonuç kabul görmelidir. Bu şekilde sonuçlar üzerinde yorumlama yapmak mümkündür fakat çok büyük veri kümelerinde insan değerlendirmesi zorlaşacaktır. Bu durumlarda da demetleme yaparak yorumda bulunmak mümkündür. Bu çalışmada çizge tabanlı geliştirilmiş olan bir demetleme yöntemi [9] kullanılmıştır. Veriler daha benzer olan dokümanlar aynı grupta olacak şekilde demetlenirse; her demet kendi içinde ne kadar sağlıklı diye insan değerlendirmesine tabi tutulabilir. Farklı benzerlik yöntemleri kullanılarak oluşturulan demetleme sonuçlarının kalitesi, insan değerlendirmesi dışında demet değerlendirme göstergeleri kullanılarak da ölçülebilir. Böylece benzerlik yöntemlerinin demetleme performansı üzerindeki etkileri dolaylı olarak ölçülebilir.

Bu tez defterinin ikinci bölümünde veri madenciliğinde dokümanların anlamsal benzerliği için kullanılmakta olan mevcut yöntemler detaylı olarak açıklanacaktır. Üçüncü bölümde Balkanet [12] ve Türkçe kavramsal sözlük [13] yapıları detaylı olarak incelenecektir. Çalışma kapsamında yapılan ön işleme çalışmaları (ayrıştırma,

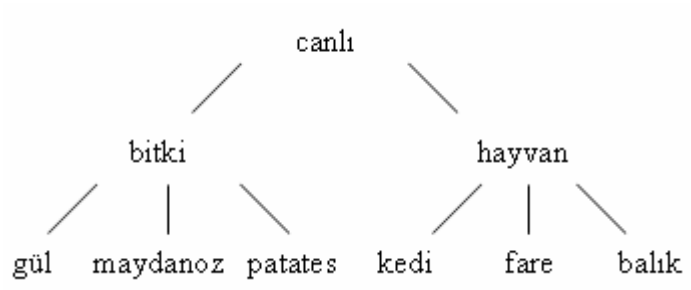
gürültü atılması ve sözcük kökü bulunması), doküman benzerliği hesabında kullanılacak sözcüklerin seçimi ve doküman benzerliğinin hesaplanması da üçüncü bölümde ele alınacaktır. Üçüncü bölüm dâhilinde son olarak doküman benzerlik değerlerinin çizge tabanlı bir yöntem sayesinde demetlenmesi konusundan bahsedilecektir. Dördüncü bölümde önerilen yöntem ve mevcut yöntemleri dikkate alarak gerçeklenen bütün deney sonuçları gösterilecektir. Beşinci ve son bölümde varılan sonuçlar yorumlanacak ve aynı çalışma konusu üzerinde ileride yapılabilecek araştırmalar belirtilecektir.

2. MEVCUT YÖNTEMLERİN İNCELENMESİ

Öncelikle veri madenciliğinde kullanılmakta olan yöntemleri incelemek yerinde olacaktır. Dokümanların anlamsal benzerliği açısından yöntemler, kavramsal sözlük kullanan yöntemler ve kavramsal sözlük kullanmayan yöntemler olarak iki kategoride incelenebilir. Bu bölümde, yöntemler arasında kıyaslamalar yapıp yöntemlerin avantajlı ve dezavantajlı yanları incelenerek açıklanacaktır.

2.1 Kavramsal Sözlük Kullanan Yöntemler

Kavramsal sözlükler, bir dile ait sözcükleri barındıran ve sözcüklerin anlamlarını kavramsal ilişkilerle açıklayan sözlüklerdir. Bu çalışmada Balkanet [12] bünyesindeki Türkçe kavramsal sözlük [13] kullanılmaktadır. Türkçe kavramsal sözlük 3.1' de detaylı açıklanmaktadır fakat şu anda somut olabilmesi için Şekil 2.1 ile gösterilen ufak bir kavramsal sözlük örneği verilmiştir.



Şekil 2.1 : Örnek Kavramsal Sözlük

Şekil 2.1' de gösterilen örnek kavramsal sözlük bir ağaç yapısındadır. Bu sözlük ağacı aşağıdan yukarıya doğru incelenirse; özel anlama sahip sözcüklerden genel anlama sahip sözcüklere veya kavramlara doğru ilerlendiği gözlenecektir. Özel anlamdan genel anlama doğru gidildikçe ağacın köküne yaklaşılarak anlam kaybı yaşanmakta ve sözcükler kavramlara dönüşmeye başlamaktadır. Bu ağaç yapısında, ayrıtlar anlamsal ilişkileri tanımlamaktadır. Ağaçtaki baba-çocuk ilişkisi, baba açısından düşünülerek üst kavram ilişkisi olarak tanımlanmıştır. Sözlük yazılırken bilgi tekrarı olmaması için, sadece üst kavram yönünde ilişki yazılabilir. Bu durumda baba açısından bakılarak üst kavram ilişkisi, çocuk açısından bakılarak alt kavram

ilişkisi şeklinde tekrarlama yapılmamış olur. Kavramsal sözlüklerde, sözcük-kavram ikilisi her zaman olacağı için üst kavram ilişkisi de her zaman bulunmak durumundadır. Fakat büyük kavramsal sözlüklerde üst kavram ilişkisi dışında ilişkiler bulunabilir. Örneğin Balkanet bünyesindeki Türkçe kavramsal sözlükte; parçası olma, bütünü olma, sebep olma, zıt anlam gibi pek çok farklı ilişki de mevcuttur. Ayrıca büyük kavramsal sözlüklerde düğümlerde, tek sözcük değil eş anlamlı birkaç sözcükten oluşan eş küme yapıları bulunmaktadır. Kavramsal sözlüklerde sadece üst kavram ilişkilerini inceleyen çalışmalar mevcuttur [1,2]. Bu çalışma kapsamında sadece üst kavram ilişkisine bakılmasının yanı sıra bütün ilişkilere bakılmasının sonuçları da yorumlanmaktadır.

Kavramsal sözlükler üzerinde çalışan yöntemler, ayrıt temelli [1] ve düğüm temelli [2] yöntemler olarak iki bölümde incelenebilir. Ayrıt temelli ilk yöntem, Eşitlik (2.1) ile gösterilen temel ayrıt sayma yöntemidir. Bu eşitlikte, Uzunluk(EnKisaYol(w1,w2)) değeri w1 ve w2 sözcükleri arasındaki en kısa yolun ayrıt uzunluğudur. MaxDerinlik değişkeni ise ağacın en derin seviyesini belirtir. Aralarında benzerlik hesaplanmak istenen iki sözcük de ağacın en derin seviyesinde bulunabilirdi ve bu durumda da aralarındaki en kısa yolun ayrıt uzunluğu 2 * MaxDerinlik olabilirdi. Bu nedenle formülde 2 * MaxDerinlik dikkate alınmaktadır. Yani formülde kesişim kümesi birleşim kümesinden çıkarılarak benzerlik hesaplanmaktadır.

$w1, w2$: Benzerliği hesaplanacak olan sözcükler

$$Benzerlik(w1, w2) = 2 * MaxDerinlik - Uzunluk(EnKisaYol(w1, w2)) \quad \textbf{(2.1)}$$

Şekil 2.1' den hareketle örnek vermek gerekirse "gül" ile "patates" arasındaki benzerlik 2 * 2 – 2 = 2 olacaktır. Aynı şekilde "gül" ile "balık" arasındaki benzerlik 2 * 2 – 4 = 0 olacaktır.

Ayrıt temelli yöntemler için daha başarılı bir örnek olarak Eşitlik (2.2) ile belirtilen Wu-Palmer yöntemi incelenebilir [3]. Buradaki KesisimDugumuDerinliği(w1,w2) ifadesi iki sözcüğün sözlükte bulunduğu noktalardan yukarıya doğru çıkılırken en yakında buluştukları düğümün ağaçtaki derinliğini göstermektedir. Derinlik(w1) ve Derinlik(w2) ifadeleri ise benzerlikleri hesaplanacak iki sözcüğün ağaçtaki derinliklerini göstermektedir. Formülde 2 * KesisimDugumuDerinligi(w1,w2) hesaplanma nedeni, toplam olarak kesişim hesaplanırken iki sözcük için de bu

değerin dikkate alınmasıdır. Burada da kesişim değeri birleşim değerine bölünerek benzerlik hesaplanmaktadır.

$w1, w2$: Benzerliği hesaplanacak olan sözcükler

$$Benzerlik(w1, w2) = \frac{2 * KesisimDugumuDerinligi(w1, w2)}{Derinlik(w1) + Derinlik(w2)} \quad \textbf{(2.2)}$$

Wu-Palmer yöntemi, kavramsal sözlük temelli benzerlik hesaplamalarında yaygın olarak kullanılan bir yöntemdir. Eşitlik (2.2) ifadesinde pay kısmı benzerlikleri, payda kısmı farklılıkları içermektedir. Benzerlikler ile farklılıkları birlikte incelediği için ve benzerlikleri farklılıklara göre normalize ederek değerlendirdiği için tercih edilen bir yöntem olmaktadır. Wu-Palmer yöntemine Şekil 2.1 üzerinden örnek vermek gerekirse; "gül" ile "patates" arasındaki benzerlik 2 * 1 / (2 + 2) = 0,5 olmaktadır. "gül" ile "patates" sözcüklerinin kesişim düğümü "bitki" olup derinliği 1'dir. Benzeri şekilde "gül" ile "balık" arasındaki benzerlik 2 * 0 / (2+2) = 0 olmaktadır. "gül" ile "balık" sözcüklerinin kesişim düğümü ağacın kökü olan "canlı" düğümü olmaktadır. Ağacın kökünde buluştukları için kesişim düğümü derinliği 0 olup benzerlik de 0 çıkmaktadır.

Düğüm temelli yöntemlerde, düğümlerdeki sözcüklere o dilin geniş metin kütüphaneleri incelenerek frekans değerleri atanır. Düğüm frekanslarının hesaplanması Eşitlik (2.3a) ile gösterilmiştir. Kavramsal sözlüğü bir ağaç olarak düşünüp T ile gösterebiliriz. Öncelikle benzerliği hesaplanacak *w1* ve *w2* sözcüklerinin kesişim düğümü bulunmalıdır. Kesişim düğümünün kök olduğu alt ağaca da T_{alt} diyebiliriz. T' nin her elemanı kavramsal sözlük ağacındaki düğümlerdir ve y ile gösterilmiştir. T_{alt} ile gösterilen alt ağacın her düğümü de x ile gösterilmektedir. Geniş metin kütüphaneleri incelenerek o dilde o sözcüğün kaç kez kullanıldığı tespit edilir. Tespit edilen sonuçlar Eşitlik (2.3a) da KullanimSayisi(x) veya KullanimSayisi(y) ile gösterilmektedir. Kesişim düğüm frekansını bulmak için ise; tespit edilen kullanım sayıları Eşitlik (2.3a)' da görülen oran ile hesaplanır. Son olarak da Eşitlik (2.3b) ile kesişim düğümü frekans değerinden benzerlik elde edilir.

$w1, w2$: Benzerliği hesaplanacak olan sözcükler

$$KesisimDugumuFrekansi(w1, w2) = \frac{\sum_{x \in T_{alt}} KullanimSayisi(x)}{\sum_{y \in T} KullanimSayisi(y)} \quad \textbf{(2.3a)}$$

$$Benzerlik(w1, w2) = -\log KesisimDugumuFrekansi(w1, w2) \quad \textbf{(2.3b)}$$

Eşitlik (2.3a), (2.3b) ile belirtilen düğüm temelli yöntem, ayrıt temelli yöntemden çok farklı değildir. Düğüm temelli yöntemde düğümlere frekans verilmektedir. Bir ayrıtın iki düğümü arasındaki frekans değerlerinden hareketle o ayrıta bir ağırlık verilerek işlem yapılmış gibi bir dönüşümü düşünerek yöntemlerin birbirine yakınlığını görebilmek mümkündür. Bu çalışma kapsamında ise ayrıt temelli yöntemler ile çalışılmak hedef seçilmiştir. Düğüm temelli yöntemleri kullanabilmek için de geniş bir metin kütüphanesinin varlığına ihtiyaç duyulduğu daha önce belirtilmişti. Ancak çalışmanın yapıldığı dönemde Türkçe için böyle bir kütüphane bulunmamaktadır. Kavramsal sözlük kullanarak anlamsal benzerlik hesaplayan ve yaygın olarak kullanılarak başarılı sonuçlar veren ayrıt temelli Wu-Palmer yöntemi ile kıyaslanmak üzere yeni bir yöntem önerilmektedir. Önerilen yöntem ile yapılan benzerlik hesabı 3.3.2' de detaylı olarak açıklanmaktadır.

2.2 Sözlük Kullanmayan Yöntemler

Dokümanların anlamsal benzerliğini hesaplamak için kavramsal sözlük kullanımı yanı sıra sözlük kullanmadan sözcüklerin nicelik miktarlarına göre çalışan yöntemler de mevcuttur [4-7]. Sözlüksüz yöntemlerin başında Jaccard benzerliği incelenebilir [4]. A ve B iki sözcük kümesi ise; Jaccard benzerliği Eşitlik (2.4) ile görüldüğü üzere hesaplanmaktadır. Eşitlik (2.4) yöntemin çok temel olduğunu, uygulamasının hızlı ve basit olduğunu hissettirmektedir. Yöntem temel benzerlik ifadesi olarak kesişimin birleşime bölünmesini söylemektedir. Benzerliği aranan iki sözcük açısından düşünülürse; kesişim iki sözcüğün birlikte kullanıldığı doküman sayısı, birleşim iki sözcüğün de geçtiği toplam dokümanların sayısı olacaktır. Benzerlik değeri de kesişimin birleşime bölünmesi ile kolayca bulunabilecektir. Uygulaması hızlı ve basit olduğu için, Jaccard benzerliği en yaygın olarak kullanılan sözlüksüz yöntemdir [5].

A, B : Benzerliği hesaplanacak iki sözcük

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A - B| + |B - A| + |A \cap B|} \quad \textbf{(2.4)}$$

Jaccard benzerliğini örneklemek gerekirse; “gül” sözcüğü o dilde 1000 dokümanda geçiyor diye varsayalım. “patates” sözcüğü de 2000 dokümanda geçiyor diye varsayalım. “gül” ve “patates” sözcüklerinin birlikte geçtiği doküman sayısı da 300 kabul edelim. Bu durumda J(gül,patates) = 300 / (1000+2000-300) = 0,11 olarak bulunabilecektir. Bu örnekte sözcük benzerliği bulunmuştur. Bundan sonra da 3.3.3’te anlatılacağı üzere sözcük benzerliğinden doküman benzerliğine geçilmesi mümkündür.

Sözlüksüz yöntemlerden yaygın olarak kullanılan diğer bir benzerlik ise kosinüs benzerliğidir [4, 6, 7]. Kosinüs benzerliği, dokümanlar arası benzerlik hesaplamadan önce terim frekansı ve ters doküman frekansını kullanarak terim ağırlığı hesaplar. Daha sonra da birçok dokümandan oluşan kümede, her doküman için terim ağırlıklarını matris halinde yazarak oluşan vektörler arasındaki açıdan hareketle benzerlik tespit eder. Terim frekansı, o dokümandaki her sözcük için sözcüğün dokümandaki sayısının dokümandaki toplam sözcük sayısına oranıdır. Terim frekansı Eşitlik (2.5) ile gösterilmiştir. Ters doküman frekansı ise bütün dokümanlar arasından o sözcüğü içeren dokümanların sayısını belirtmeye yarayan bir değerdir. Ters doküman frekansı değeri Eşitlik (2.6a) ve (2.6b) ile gösterilmiştir. Benzerlik değeri hesaplanması da Eşitlik (2.8) ile gösterilmektedir.

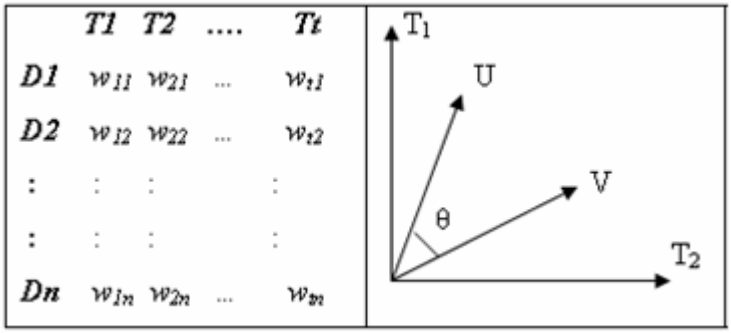
i. doküman j. sözcük için;

$$\text{Terim Frekansı : } TF_{ij} = \frac{F_{ij}}{\sum_j F_j} \qquad \text{(F: Frekans)} \tag{2.5}$$

$$\text{Doküman Frekansı : } DF_j \text{ : j sözcüğünü içeren doküman sayısı} \tag{2.6a}$$

$$\text{Ters Doküman Frekansı : } IDF_j = \log\frac{N}{DF_j} \quad \text{(N: Toplam doküman sayısı)} \tag{2.6b}$$

$$\text{Terim Ağırlığı : } W_{ij} = TF_{ij} * IDF_j \tag{2.7}$$



Şekil 2.2 : Kosinüs Benzerliği Hesabı

U, V : İki vektör

$|U|, |V|$: Vektör boyları

Q : İki vektör arasındaki açı

$$0 \le benzerlik = \cos Q = \frac{U.V}{|U|.|V|} \le 1 \qquad \textbf{(2.8)}$$

Kosinüs benzerliğinde geçen terim frekansı, dokümanda yoğun olarak geçen anahtar sözcükleri bulabilmek içindir ve yüksek değerleri tercih edilir. Terim frekansı, dokümandaki toplam sözcük sayısına bölüm şeklinde normalize edilerek hesaplanmaktadır. Bu sayede hata toleransı arttırılmak istenmiştir. Örneğin; doküman için anahtar sözcük niteliği taşımamasına rağmen gereksiz yere çok kez kullanılmış ve bu sayede frekansı yüksek çıkarak benzerlik yönteminin başarımını düşürecek durumlar söz konusudur [7]. Bunlar kaçınılmaz durumlardır ama en azından bu şekilde etkisi normalize edilmiştir. Ters doküman frekansı ise, terimin birçok doküman arasından az sayıda dokümanda geçmesi itibariyle ayırt ediciliğini belirlemeye yöneliktir ve düşük olması tercih edilir. Terim ağırlığı ise terim frekansı ile ters doküman frekansı çarpımıdır. Bu sayede bütün dokümanlar arasından az sayıda dokümanda bulunan fakat bulunduğu doküman içinde de yoğun olarak bulunan sözcükler bulunmak istenmiştir. Ters doküman frekansı hesabındaki logaritmanın nedeni ise; terim ağırlığı belirlenirken terim frekansına ile çarpıldığı için çok yüksek değerler çıkmaması ve terim frekansının etkisinin biraz daha azaltılarak daha dengeli sonuçlar elde edilmesini sağlamak içindir.

Bu çalışmada; kavramsal sözlük kullanan yöntemler esas alınmıştır. Çünkü kavramsal sözlük dil bilimcilerce ortaya konmuş daha güvenilir bir kaynaktır. Kavramsal sözlük incelenirken 2.1' de belirtildiği nedenlerle ayrıt temelli yöntemler

seçilmiştir. Ayrıt temelli yöntemlerde başarımı iyi olan ve yaygın olarak kullanılan Wu-Palmer yöntemi ile kıyaslanacak bir yöntem önerilmiştir. Bu araştırmada kavramsal sözlük kullanmayan yöntemler de önerilen yöntem ve Wu-Palmer yöntemi ile kıyaslanmaktadır. Bu kıyaslama, önerilen yöntemi sözlüksüz yöntemlerle karşılaştırırken kavramsal sözlük kullanıp kullanmama açısından da bir sonuç verecektir. Bu çalışmada sözlüksüz yöntemler adına kosinüs benzerliği kullanılacaktır. Jaccard benzerliğinin kullanılamama nedeni olarak, geniş Türkçe bir metin kütüphanesi olmaması söylenmişti. Buna ilaveten kosinüs benzerliği Jaccard benzerliğinden başarılıdır ve uygulaması kolay olup oldukça yaygındır [4,6]. Kosinüs benzerliğini hesaplamak için geniş metin kütüphanelerine gerek duymadan, üzerinde çalıştığımız siteye ait bütün dokümanları kullanacağız. Ama kosinüs yöntemindeki terim ağırlıklarının daha sağlıklı çıkması için kümede ne kadar daha fazla doküman varsa o kadar sağlıklı sonuç alınacağı unutulmamalıdır. Yani burada da geniş metin kütüphaneleri varlığı gereksinimi ortaya çıkmaktadır. Buna karşın kosinüs benzerliği bütün dokümanlardan global bir vektör uzayı oluşturarak değerlendirme yaptığı için avantajlıdır ve kullanılabilir [4,6]. Kosinüs benzerliğinin hesaplanması için Java dilinde geliştirilen “SciLib” kütüphanesi kullanılmıştır [17].

3. ÖNERİLEN YÖNTEMİN AÇIKLANMASI

Bu çalışmada geliştirilen yöntem bu bölümde açıklanmaktadır. Öncelikle Balkanet ve Türkçe kavramsal sözlük açıklanacaktır. Bu çalışmada internet ortamındaki sayfalarla çalışıldığı için html sayfalarının bulunup içerisinde tarama yapıldıktan sonra sözcüklerin işlenebilmesi gereklidir. İşlenen sözcüklerin hepsi Türkçe olmayabilir ya da Türkçe olan kelimelerde de karakter bozukluğu veya yazım yanlışı yapılmış olabilir ve bu hatalar düzeltilmelidir. Sözcüklerin ayıklanması ve düzeltilmesi işlemleri, çalışmanın başlangıcındaki ön işleme adımlarından gürültünün atılması kapsamı içinde düşünülebilir. İşlenecek sözcüklerin gürültülerden arındırılmasının ardından, uygun sözcük kökü bulunarak benzerlik hesaplamasına geçilmelidir. Sözcük kökü bulunmasının nedenleri, Türkçe kavramsal sözlük gibi veri tabanlarında kök halinde veriler bulunması ve farklı çekim ekleri almış aynı köke sahip sözcüklerin (örneğin; "gelindi" ve "gelinmiştir") tamamen benzer olduğu düşüncesi gösterilebilir.

Çalışma kapsamında her zaman için ikili doküman benzerliklerinin tespit edilmesi söz konusudur. Dokümanların benzerliği hesaplanırken dokümanlarda geçen bütün sözcükler kullanılmayıp sadece dokümanı en iyi nitelediği düşünülen anahtar sözcükler kullanılmaktadır. Bu durumda sözcükler arasında seçim yapılmaktadır. Seçilen sözcükler, dokümanda en çok kullanılan sözcükler ile doküman içinde anahtar sözcük olarak tanımlanmış sözcüklerdir. Sözcükler arası benzerliği hesaplarken önerilen yöntem ve Wu-Palmer yöntemi kullanılmaktadır. Daha sonra da sözcükler arası benzerlikler birleştirilerek doküman benzerlikleri hesaplanmaktadır. Kavramsal sözlük kullanan yöntemler için doküman benzerliği bu şekilde önce sözcük benzerliği sonra doküman benzerliği şeklinde iki adımda hesaplanmaktadır. Buna karşın, bu çalışmada sözlüksüz yöntemlerin örneği olarak incelenecek olan kosinüs benzerliğinde ise doğrudan doküman benzerliği hesaplanmaktadır. Kosinüs benzerliğinde ikili doküman benzerlikleri bulunmamakta bütün dokümanlar tek seferde çok boyutlu uzayda incelenerek sonuçlar elde edilmektedir. Doküman benzerlikleri de hesaplandıktan sonra çizge tabanlı bir yöntem kullanılarak

demetleme yapılmaktadır. Demetleme ile daha benzer dokümanların aynı gruplara yerleştirilmesi amaçlanmaktadır. Demetlemenin sonuçlarından hareketle yöntemlerin başarımları çıkartılıp araştırmanın sonuçları değerlendirilebilecektir.

3.1 Balkanet ve Türkçe Kavramsal Sözlük

Balkanet [12] projesi, Bulgaristan, Çek Cumhuriyeti, Yunanistan, Romanya, Türkiye ve Sırbistan'dan oluşan Balkan ülkelerinin dillerine ait, bilgisayar ortamında işlenebilecek şeklinde sözcükler arası anlam ilişkilerini barındıran bir sözlük oluşturmayı amaçlamaktadır. Sözlükteki anlam ilişkileri; kavramlar aracılığıyla tanımlanmaktadır. Bu sayede daha kolay işlenebilir, sağlıklı ve net bir yapı oluşmaktadır. Kavramlar da sözlük yapısı içine katılınca doğal dil işleme çalışmalarında kullanılabilecek kavramsal sözlükler ortaya çıkmaktadır.

Türkçe kavramsal sözlük [13] kayıt örnekleri Tablo 3.1' de gösterilmiştir.

Tablo 3.1 : Türkçe Kavramsal Sözlük Kayıt Örnekleri

<SYNSET><ID>BILI-00000022</ID> <SYNONYM><LITERAL>sipahi<SENSE>1</SENSE></LITERAL> <LITERAL>tımarlı sipahi<SENSE>1</SENSE></LITERAL></SYNONYM> <DEF>Tımar sahibi, atlı Osmanlı askeri.</DEF><SNOTE>Ottoman soldier who joins the army against the revenue of land that brings 3,000 to 20,000 akchas annually. (SEE PICTURE)</SNOTE><POS>n</POS><ILR>ENG20-09929597-n<TYPE>hypernym</TYPE></ILR><STAMP>orhanb 2004/04/22</STAMP></SYNSET>
<SYNSET><ID>ENG20-07338012-n</ID><POS>n</POS> <ILR>ENG20-07336473-n<TYPE>hypernym</TYPE></ILR> <BCS>1</BCS><STAMP>orhanb 2004/03/19</STAMP></SYNSET>
<SYNSET><ID>ENG20-09071304-n</ID><POS>n</POS><SYNONYM> <LITERAL>Yahudi<SENSE>1;Musevi</SENSE></LITERAL></SYNONYM> <ILR>ENG20-00006026-n<TYPE>hypernym</TYPE></ILR> <ILR>ENG20-07964905-n<TYPE>holo_member</TYPE></ILR> <DEF>Musa dinine bağlı olan kimse</DEF><BCS>2</BCS></SYNSET>

Şekil 2.1 ile basit bir kavramsal sözlük örneği incelenmişti. Tablo 3.1' de xml biçimde yazılmış Türkçe kavramsal sözlük verisi ağaç yapısında düşünülüp Şekil 2.1 ile kıyaslanırsa yapısının benzer olduğu görülebilir. Aralarındaki fark, Türkçe kavramsal sözlük düğümlerinde eşkümelerin bulunması ve kavramsal sözlükte üst kavram ilişkisi dışında ilişkilerin varlığıdır.

Tablo 3.1' deki ilk örnek üzerinden sözlüğün temel yapısı incelenebilir. "LITERAL" alanında sözcük yazmaktadır. "SENSE" alanı, o sözcüğün dildeki kaçıncı anlamı olduğu tanımlanmaktadır. Örnekte belirtilen iki sözcük eş anlamlıdır, bu nedenle "SYNONYM" etiketi içinde gösterilmektedirler. Bu eş anlamlı sözcükler, en dışta "SYNSET" etiketi ile gösterilmektedirler çünkü bir eşküme oluşturmak suretiyle sözlükte bir düğüm olarak bulunmaktadırlar. "ID" alanı bu eşkümenin ayırt edici tanımlama numarasıdır. "ID" değerinde yazan değer, kavramsal sözlükler için "ILI" numarası olarak da belirtilmektedir. "POS" etiketi isim, sıfat gibi sözcüğün durumunu belirtir. Örneğimizde sözcükler isimdir. "DEF" alanı ile eşkümedeki sözcüklere ait tanımlama girilmiştir. "SNOTE" alanı ile sözcükler için özel not eklenmiştir. "STAMP" değeri, sözlük kaydını giren kullanıcı adı ve zamanı belirten mühür bilgisidir. "ILR" alanı eşkümenin, sözlükteki diğer eşkümelerle olan ilişkisini belirtmektedir. İlişkisi olan eşkümenin "ILI" numarası ve "TYPE" alanı ilişkinin tip bilgisini içermektedir. Bu örnekte belirtilen ilişki tipi, üst kavram ilişkisidir.

Tablo 3.1' deki ikinci ve üçüncü örnekler daha istisna örneklerdir. İkinci örnekte, "SYNONYM" etiketi olmadığı görülmektedir çünkü "ID" değeri ile İngilizce kavramsal sözlüğünde belirtilen sözcüğün Türkçede karşılığı yazılamadığı için Türkçe kavramsal sözlükte bulunmamaktadır. Ayrıca "BCS" alanı karşımıza çıkmaktadır. "BCS" alanı da Balkanet çalışması içinde sözcüklerin gruplandığı ortak küme numarasıdır. "BCS" değeri 1, 2, 3 aralığında sayılar olabilir. Üçüncü örnekte ise "SENSE" alanında sözcüğün o dildeki kaçıncı anlamı olduğunun tam sayı değeri yazılması gerekirken tam sayı ve sözcük karışık bir değer yazıldığı görülmektedir.

Kavramsal sözlüklerde üst kavram ilişkisi dışında da ilişkiler vardır. Balkanet bünyesinde toplam 20 çeşit ilişki vardır. Bu ilişkiler için Tablo 3.2 incelenebilir.

Tablo 3.2: Balkanet Bünyesindeki Bütün Sözcük İlişkileri

İlişki	**Açıklama**	**Örnek**
Synonmy	Eş anlam	(canlılık, hayatiyet)
Near_antonym	Yakın zıt anlam	(almak) - (satmak)
Category_domain	Aynı kategoride olma	(hücre) - (biyoloji)
Also_see	Ayrıca ilişkisi	(giyinmek) - (şık giyinmek)
Be_in_state	Durumunda olma	(uzun:1) - (boy pos:2, yükseklik:3)
Subevent	Alt olay	(bulmak) - (mantık yürütmek)
Causes	Sebep olma	(zayıflatmak) - (zayıflamak)
Hypernym (superordinat)	Üst kavram	(reçel) – (gül reçeli)
Hyponym (subordinat)	Alt kavram	(gül reçeli) – (reçel)
Verb_group	Fiil grubu	(act:5 play:8 act-as:2) - (dissemble:3 pretend:2 act:9)
Similar_to	Yakın anlam	(müteakip) - (sıradaki) - (bir sonraki)
Usage_domain	Aynı kullanım alanı	
Holo_member	Üye - bütünü	(Enterobacteriales) - (bakteri)
Mero_member	Bütün – üyesi	(bakteri) - (Enterobacteriales)
Holo_part	Kısmi Parça – bütünü	(zeytin) - (zeytin ağacı)
Mero_part	Bütün - kısmi parçası	(zeytin ağacı) - (zeytin)
Derived_From	Türetilmiş	(hızlıca) – (hızlı)
Region_domain		
Holo_madeof	Hammadde - bütünü	(kâğıt) – (kitap)
Mero_madeof	Bütün – hammaddesi	(masa) – (tahta)

Balkanet çalışmasında toplam 20 ilişki olmasına rağmen Türkçe kavramsal sözlükte bunlardan 13 tanesi bulunmaktadır. Balkanet bünyesindeki Türkçe kavramsal sözlüğe ait güncel istatistikler Tablo 3.3' de gösterilmiştir.

Tablo 3.3: Balkanet Türkçe Kavramsal Sözlük İstatistiği

	Türkçe
Eşkümeler	14626
İsimler	11059
Filler	2725
Sıfatlar	802
Zarflar	40
Sözcükler	20310
Sözcük/Eşküme	1.39
BC1	1220
BC2	3479
BC3	3794
Etki alanına özgü Eşkümeler	300
Balkanlara özgü Eşkümeler	103
Dile özgü eşkümeler	204
Dil içi ilişkiler	19834

Kavramsal sözlüğün işlenmesi sırasında xml dosyadan arama yapmak yavaş olduğu için veritabanına aktarılması tercih edilmiştir. Veritabanına aktarım sırasında veya sonrasında sorun oluşmaması için, Tablo 3.1' de bahsedilen bütün durumların kontrol edilmesi gerektiği unutulmamalıdır. Ayrıca veritabanına aktarım sırasında da dosyadaki sözcüklerde bulunan olası Türkçe karakter sorunları düzeltilerek aktarım yapılmıştır. Sözlüğün bütün alanları birebir veritabanına aktarılmamış, sadece gerekli görülen alanlar ("ID", "POS", "LITERAL", "SENSE", "BCS", "ILR", "TYPE") aktarılmıştır.

3.2 Ön İşleme

Dokümanların çalışmada gerekli şekilde işlenebilmesi için, ön şart olan bazı düzenlemelerin yapılarak biçim verilmesi gerekmektedir. Bu amaç ön işlemeler

sayesinde gerçekleştirilmektedir. Burada gerçeklenecek olan ön işleme adımları; html sayfaların ayrıştırılması, gürültünün atılması ve sözcük kökü bulunması şeklinde üç alt başlıkta detaylı olarak incelenecektir.

3.2.1 Html Sayfalarının Ayrıştırılması

Bu çalışmada bir web sitesinden elde edilen dokümanlar için benzerlikler hesaplanmıştır. Öncelikle web sitesinin bütün sayfaları, Teleport Pro Web Dizinleme programı aracılığıyla yerel diske kaydedilmiştir. Daha sonra, yerel diskteki dizin içerisinden sadece htm ve html sayfalar bir filtre ile filtrelenerek incelemeye alınmıştır. İncelenen sayfaların içinin taranması ve sözcüklerin html kodundan ayrılarak düzenlice alınabilmesi için Java html parser [15] kullanılmıştır. Bu aracın seçilmesinin nedeni HTML kodundaki bütün etiketlerin düzgün yazılmadığı durumda bile çalışabilmesidir. Ayrıca Java html parser [15] programı, bu çalışma kapsamında kullanılmamış olsa bile dosyadaki linkleri de çıkarabilme özelliğine sahiptir ve olası ileriki çalışmalarda bu özelliği de kullanılabilir. Bu çalışma şu anda sadece htm ve html sayfalar için çalışmaktadır. Fakat başka sayfalar için de çalışması istenirse, ilgili tarayıcı programlar aracılığıyla düzenlemeler yapılarak kapsam genişletilebilir.

3.2.2 Gürültünün Atılması

Tarayıcı programlar ile dosyaların içeriğine ulaşabildikten sonra sözcükler arasındaki gürültü olarak adlandırılan istenmeyen kısımları ayıklamaya yönelik adımlar gerçeklenmelidir. Öncelikle Türkçe olmayan her türlü içerik atılmaktadır. Burada atılanlar, html etiketleri arasındaki veriler, @ # ~ gibi daha pek çok normal sözcük içerisinde geçmeyecek özel karakterlerin ve rakamların içinde bulunduğu sözcükler ve bazı temel Türkçe sözcük kurallarına aykırı durumdaki sözcüklerdir. Türkçe kuralları hakkında bilgi için; ses bilgisi (fonetik), şekil bilgisi (morfoloji), cümle bilgisi (sentaks), köken bilgisi (Etimoloji), anlam bilgisi (Semantik) gibi pek çok açılardan internet üzerinde Türk dilini anlatan eğitim sitelerinden yararlanılmıştır [16]. Temel Türkçe kuralları olarak, sadece sözcük sonunda “b,c,d,g” harfleri olamayacağı ve iki sesli harfin yan yana gelemeyeceği dikkate alınmıştır. İki sesli harfin yan yana gelemeyeceği konusunda “istisnai, mesai” gibi ender örnekler bulunsa da bunların sayısı fazla olmayıp genel olarak bu kıstasların doğru olması sebebiyle kullanılmasında sakınca görülmemiştir. Bunun dışında da İngilizce

sözcüklerin özellikleri olarak Türk alfabesinde olmayan “w, q, x” gibi harfleri içeren veya sonu “ing” eki ile biten sözcükler de incelenmeden atılmıştır. Son olarak da kalan sözcükler arasından en az üç harften oluşan sözcükler incelenmektedir. Bu şekilde de bir veya iki harften oluşan bağlaçlar gibi yapılar süzülmek istenmiştir. Bu sırada “ak, al” gibi sözcükler de kaybedilmiştir fakat yine bunların sayısı fazla olmadığından bu işlemi yapmakta bir sakınca bulunmamıştır. Burada belirtilen kurallara dayanarak filtreleme işleminin yapılması, düzenli ifadelerin yazılımı sayesinde hızlı bir şekilde yapılabilmektedir.

Belirtilen gürültülerden arındırıldıktan sonra kalan sözcüklerdeki Türkçe karakter bozuklukları giderilmiştir. Bunun için utf-8 karakter kümesi ile yazılmış eclipse kodunda bozuk karakterler düzgün Türkçe karakterlerle değiştirilmiştir. Türkçe kelime olduğu halde “ıstanbul” örneğinde olduğu gibi yanlış yazılmış kelimeler olabilir. Bu şekilde yanlış yazımlar zemberek programının öner fonksiyonu sayesinde “istanbul” halinde düzeltilmiştir [14]. Zemberek programı OpenOffice yazılımındaki Türkçe işlemlerini yapan bileşendir ve java ortamında kodu sunulmaktadır. Bu nedenle araştırmada Zemberek kullanılmıştır. İlk adımda kurallar tanımlayarak Türkçe olmayan sözcüklerin ayıklanmış olmasına rağmen hala gözden kaçan sözcükler çıkarsa bunların da ayıklanması gerekecektir. Bunun için de zemberek programının TürkçeMi fonksiyonu kullanılarak Türkçe olmayan sözcükler ayıklanmıştır [14]. Zemberek programının TürkçeMi fonksiyonunu öner fonksiyonundan sonra kullanmak çok önemlidir çünkü öner fonksiyonu ile “ıstanbul” değeri “istanbul“ yapılmadan önce TürkçeMi fonksiyonu çalıştırılsaydı; “ıstanbul” Türkçe değil diyecekti ve sözcük kaybedilmiş olunacaktı. Bütün bunlardan sonra sitede incelenecek sayfalar ve sayfalardaki sözcükler sayıları ile birlikte veri tabanına atılmıştır. Burada yapılanlara kısaca tekrar toplu bir örnek vermek gerekirse; herhangi bir dokümanda geçebilecek olan “C# hashing fonksiyonu kullanilarak ıstanbul sözcugü için 33 değeri elde edilmiştir” cümlesinden “fonksiyonu, kullanılarak, istanbul, sözcüğü, için, değeri, elde, edilmiştir” sözcükleri çıkarılarak çalışmaya devam edilecektir.

3.2.3 Sözcük Kökü Bulunması

Ön işleme aşamasının son işlemi sözcüklerin köklerinin bulunmasıdır. Sözcük kökü bulunmasının nedeni, Türkçe kavramsal sözlük gibi veri tabanlarında kök halinde

veriler olması ve farklı çekim ekleri almış aynı köke sahip sözcüklerin benzerliğinin köklerden hareketle daha sağlıklı şekilde hesaplanacak olmasıdır.

Kökü bulunmak istenen sözcük, Sabancı Üniversitesi tarafından geliştirilen biçimbirimsel çözümleyiciye [13] girdi olarak verilmektedir. Biçimbirimsel çözümleyici, sözcüğün biçimine göre olası kök ve ek durumlarını çözümleyerek alternatiflerin hepsini sonuç olarak sunmaktadır. Biçimbirimsel çözümleyicinin sonuç sunmaması mümkün değildir çünkü sözcüğü çözümleyemezse sözcük tamamen kök halinde olabilir diyerek sözcüğü aynen geri vermektedir. Bundan sonra incelenen doküman, her sözcüğü yerine biçimbirimsel çözümleyicinin verdiği bütün sonuçları koyarak yeniden yazılmaktadır. Alternatif çözümlemelerin hepsi birer satırda yazmaktadır. Sözcükler arasındaki boşluk karakterler bir boş satır ile gösterildiği gibi cümle sonlarındaki noktalar da kesinlikle unutulmamalıdır. Bu şekilde yeniden yazılmış doküman, biçimbirimsel çözümleyicinin çözümleme sonuçlarını girdi olarak işleyebilen bir sözcük etiketleyici yazılımına verilmelidir. Sözcük etiketleyici yazılımın amacı, dokümanın her cümlesindeki sözcüklerin kök alternatifleri arasından o cümleye en uygun olan kökünü bulmaktır. Bu çalışmada Koç Üniversitesi tarafından geliştirilen, morfolojik analiz ile uygun sözcük kökünü bulabilen bir sözcük etiketleyici yazılımı kullanılmıştır [10]. Belirtilen sözcük etiketleyici yazılım, uygun sözcük kökünü sağlıklı olarak bulabilmektedir fakat bu bütün sözcükler için doğrudan bu yazılım çalıştırılmamaktadır. Sözcüğü bu yazılıma vermeden önce sözcüğün kavramsal sözlükte bulunup bulunmadığı kontrolü yapılmaktadır. Eğer sözcük sözlükte varsa sözlükte zaten kök halinde veriler bulunduğu için sözcük bütünüyle kök halinde demektir ve sözcük etiketleyiciye vermeye gerek yoktur. Eğer sözcük bütün haliyle sözlükte bulunmuyorsa sözcük etiketleyici yazılıma verilerek buradan gelen sonuç kullanılmaktadır. Sözcük etiketleyici yazılımın sonuç vermemesi de mümkün değildir, ancak istisnai programcılık hataları nedeniyle çalıştırılamaması durumunda sonuç alınamayabilir. Bu duruma örnek olarak programa girdi olarak verilen dosyada # karakteri geçmesi durumunda programın perl kodu olması ve bunu yorum satırı olarak algılaması nedeniyle hata vererek çalışmasını yarım bırakıp sonlanması gösterilebilir. Bu tarz sorunların yaşanmaması için yapılan ön işlemenin önemi bu örnek ile daha net anlaşılmaktadır. Bu tarz istisnai programcılık hataları sebebiyle sonuç alamama durumunda ise biçimbirimsel çözümleyicinin kök alternatiflerinden en uzun kök

seçilerek işleme devam edilmektedir. Bu şekilde bir seçim yapılmasında hata bulunmamaktadır. Çünkü en uzun kök, sözcüğün bütününe en yakın köktür ve en özel anlamdaki olasılıktır. En kısa kökün seçilmesi durumunda daha genel bir kök seçilerek daha fazla anlam kaybı yaşanacaktır. Örneğin "gözlükçüler" sözcüğü için "göz" , "gözlük" , "gözlükçü" kök alternatifleri varken istisnai hata durumları ile ilişkin kök seçilemezse sözcüğe en yakın olarak en özel anlama sahip ve onu daha iyi niteleyen en uzun kök olan "gözlükçü" seçilmektedir. Burada anlatılanlar Şekil 3.1 ile örneklendirilmiştir. Şekil 3.1'de sol tarafta sözcük etiketleyiciye verilen girdi dosyasının bir kısmı, sağ tarafta da alınan sonuç dosyasının ilişkin kısmı görülmektedir.

```
bugün   bugün   +Noun+A3sg+Pnon+Nom

hava    hav     +Noun+A3sg+Pnon+Dat
hava    hava    +Noun+A3sg+Pnon+Nom

güzel   güzel   +Adj

<\S>    <\S>

<S>     <S>

buna    bu      +Pron+Demons+A3sg+Pnon+Dat
buna    buna    +Verb+Pos+Imp+A2sg

göre    gör     +Verb+Pos+Opt+A3sg
göre    göre    +Postp+PCDat

giyin   giy     +Verb+Pos+Imp+A2pl
giyin   giy     +Verb+Reflex+Pos+Imp+A2sg

<\S>    <\S>

<S>     <S>
```

```
bugün bugün+Noun+A3sg+Pnon+Nom
hava hava+Noun+A3sg+Pnon+Nom
güzel   güzel   +Adj
. .+?
<\S> <\S>
<S> <S>
buna bu+Pron+Demons+A3sg+Pnon+Dat
göre göre+Postp+PCDat
giyin giy+Verb+Pos+Imp+A2pl
. .+?
<\S> <\S>
<S> <S>
```

Şekil 3.1 : Sözcük Etiketleyici Örnek Girdisi ve Çıktısı

3.3 Benzerlik Hesaplanması

Benzerlik hesaplanabilmesi için öncelikle benzerlik hesabında kullanılan sözcükler seçilecek, sonra sözcükler arası ikili benzerlikler hesaplanıp en sonunda da dokümanlar arası ikili benzerlik tespit edilecektir. Bu durum kavramsal sözlük kullanan yöntemler için geçerlidir. Sözlüksüz yöntem örneği olan kosinüs benzerliğinde arada sözcükler arası benzerlik hesaplanmadan doğrudan doküman benzerliği hesaplanacaktır.

3.3.1 Sözcüklerin Seçimi

Benzerlik hesabında kullanılacak sözcükler, dokümanlardaki bütün sözcükler değildir. Burada dokümanı en iyi şekilde tanımlayabilen sözcüklerin incelenmesi

yeterli olacaktır. Bu durumda da bu sözcüklerin seçimi önem kazanmaktadır. Seçim sırasında sözcüklerin dokümanda kaç kez kullanıldığı sayıları incelenecektir. Sözcük kökleri üzerinden işlem yapıldığı için buradaki sayıların kök halleri için belirlenmiş değerler olduğu unutulmamalıdır. Dokümandaki ortalama sözcük sayısından fazla kullanılan bütün sözcüklerin seçilmesi şeklinde deneyler yapılmıştır. Bu şekilde işlem zamanının uzaması nedeniyle daha sonraki deneylerde ise en fazla kullanılan ilk 10 sözcük işleme alınmıştır.

En sık kullanılan sözcüklere ek olarak, internet sayfalarının başında "META KEYWORD" html etiketi ile belirtilen ön tanımlı anahtar sözcükler varsa bunlar da dikkate alınmaktadır. Tanımlanmış anahtar sözcüklerden de tek sözcükten oluşan kök halindeki örnekler işleme alınmaktadır. Anahtar sözcükler, en başta belirtilen "meta keyword" tanımından seçildiği için dokümanda kullanım sayıları gibi bir değerleri yoktur. Anahtar sözcüklere hesaplamalarda kullanılabilecek değer atanması gereklidir. Bunun için Eşitlik (3.1)' de gösterilen formül kullanılmaktadır. Eşitlik (3.1) ile anlatılmak istenen, olabildiğince az sayıda doküman için anahtar sözcük olarak belirtilmiş sözcüklerin ayırt ediciliğinin daha sağlıklı olması sonucunda daha yüksek puan alacağıdır. Deneylerin farklı şekillerde yapılabilmesi için Eşitlik (3.1) dışında değerler vermek de mümkündür. Eğer ön tanımlı anahtar sözcüklerin etkisi ayrıca görülmek istendiği için onlar olmadan sonuç hesaplanmak istenirse ön tanımlı anahtar sözcüklerin hepsine 0 puanı verilebilir. Bunun yanı sıra, Eşitlik (3.1) ile puan hesaplanmış olsa bile gözlemci herhangi bir sözcüğün o doküman için ayırt ediciliğinden emin ise o sözcüğe 1 puanını vererek de sonuç alabilir.

w : Herhangi bir ön tanımlı anahtar sözcük

$$AnahtarSozcukDegeri(w) = 1 - \frac{AnahtarSozcukOlduguSayfaSayisi(w)}{SitedekiToplamSayfaSayisi} \qquad \textbf{(3.1)}$$

Bu çalışma için seçilen web sitesinde bütün dokümanlarda aynı ön tanımlı anahtar sözcükler bulunmaktadır. Bu nedenle Eşitlik (3.1) ile hesaplama yapınca da anahtar sözcüklere 0 puanını vermiştir ve doğrudan 0 değerini vererek anahtar sözcüklerin incelenmediği duruma denk gelmiştir. Çalışmada da Eşitlik (3.1)' e göre çalışılmıştır. Gözlemler sonucunda sözcüklerin ayırt ediciliğinin doğrudan 1 değeri ile denenmesine ise gerek görülmemiştir.

Dokümanı iyi niteleyen sözcüklere ek olarak başlıkta yer alan sözcüklerin incelenmesi düşünülmüştür. Çünkü başlıklar dokümanları özetleyen kısa ifadeler olup önemlidir. Fakat başlıkları ayırt etmek için html h1-h6 değerlerinin varlığı tercih edilmemiştir çünkü günümüz internet sayfaları saf html kodu olarak yazılmayıp bu değerler çoğunlukla bulunmamaktadır. Başlığı ayırt edebilmek için, metin kısmından ayrı olarak koyu biçimde yazılmış olabilecek sözcükleri bulmak istenmiştir. Ayrıca görselliğin oldukça önemli olduğu günümüz internet sayfalarında görünümden bağımsız olarak başlığın sağlıklı şekilde bulunabilmesi amaçlanmıştır. Fakat görünümden bağımsız olarak başlık tespiti tamamen ayrı bir çalışma konusu olduğu için başlıkların incelenmesinden vazgeçilmiştir [8].

Burada anlatılan seçim işlemleri teknik olarak ele alınan elemelerdir. Teknik seçim şartları bu kadarken pratikte de bir seçim olduğu söylenebilir. Örneğin; kavramsal sözlük kullanan yöntemlerde seçilen sözcük sözlükte bulunmuyorsa benzerlik 0 dönecektir ve doküman benzerliği hesabında etkisi olmayarak sözcük hiç seçilmemiş gibi bir etki yapabilecektir. Bunun dışında 0 puanını alan ön tanımlı anahtar sözcükler de hiç seçilmemiş etkisini yapacaktır. Ayrıca dokümandaki ortalama sözcük kullanım sayısının üzerindeki bütün sözcüklerle çalışma yapılmıştı fakat programın pratik çalışma zamanı kısıtı nedeniyle sadece belirli bir sayıdaki en yaygın kullanılan sözcük kümesi ile işlem yapılmıştır. Bu örnekler teorik sözcük seçimi etmenlerinin haricinde pratik nedenlerle sözcük seçimi yapıldığını da göstermektedir.

3.3.2 Sözcükler Arası Benzerlik

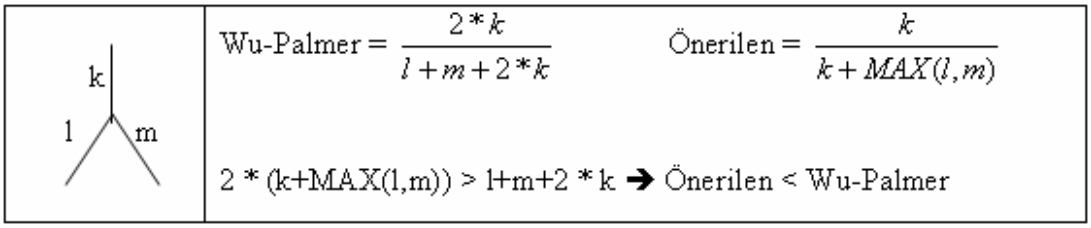
Sözcükler arası ikili benzerlikler hesaplarken Eşitlik (2.2)' de gösterildiği üzere Wu-Palmer yöntemi kullanılacaktır. Bunun yanı sıra da bu çalışmada önerilen yöntem ile hesaplama yapılacaktır. Önerilen yöntem Eşitlik (3.2) ile gösterilmektedir. Önerilen yöntemin ve Wu-Palmer yönteminin farkları, önerilen yöntemin pay kısmında 2 çarpanı olmaması ve paydasındaki ifadedir. Pay kısmında 2 çarpanı, iki sözcük için hesap yapılmasından gelmekteydi. Burada aynı uzaklığı iki kez saymak gerekli görülmediği için 2 çarpanı kullanılmadı. Paydadaki ifade ise, kesişim düğümüne gelene kadar sözcüklerin ne kadar anlam kaybettiğine ağırlık verilmesinin sonucudur. Çünkü iki sözcük için de, kesişim düğümüne olan uzaklıklardan daha büyük olan değer seçilmektedir. Burada incelenen sözcüklerden oluşan grubun sağlamlığı (benzerlik değeri), grubun en zayıf halkası kadardır yaklaşımı

sergilenmektedir. Wu-Palmer yönteminde kesişim ayrıt uzunluğu, birleşim ayrıt uzunluğuna bölünmekteydi. Önerilen yöntemde bahsedilen farklılıklar nedeniyle daha performanslı sonuçlar beklendiği için bu çalışma yapılmıştır.

$n1, n2$: İkili sözcük benzerliği hesaplanacak olan sözcükler

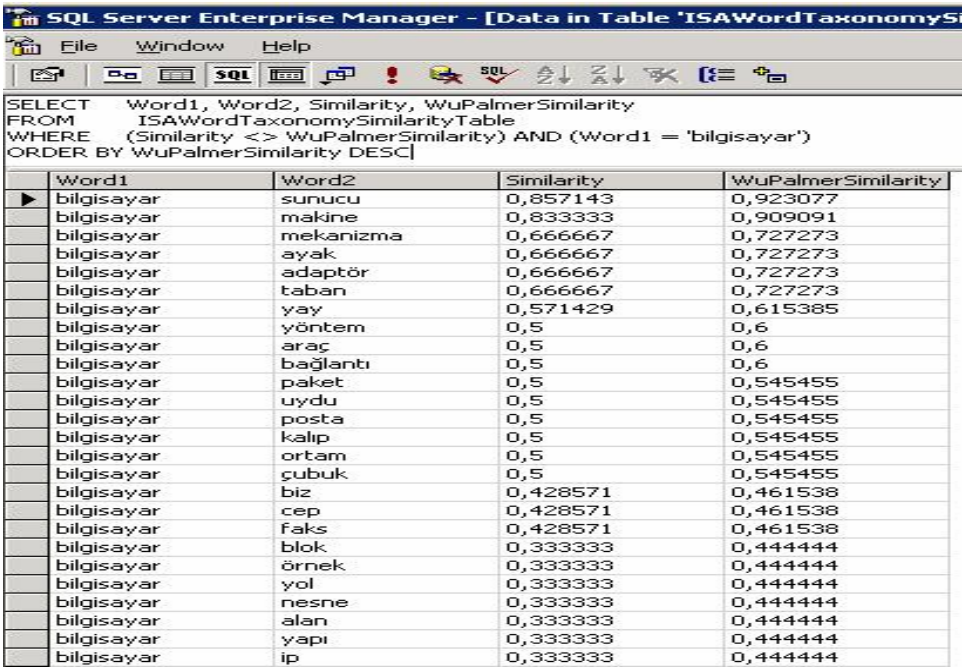
$$Benzerlik(n1,n2) = \frac{Derinlik(KesisimDugumu(n1,n2))}{MAX(KesisimDugumuneEnKisaYol(n1,n2))} \quad \textbf{(3.2)}$$

Şekil 3.2 ile Wu-Palmer yöntemi ve önerilen yöntem matematiksel açıdan incelenmektedir. Yöntemlerin matematiksel ifadeleri karşılaştırılırsa, önerilen yöntem sonucunda her zaman Wu-Palmer dan daha düşük sayısal değerde bir benzerlik sonucu hesaplandığı görülmektedir.



Şekil 3.2 : Önerilen Yöntem ile Wu-Palmer Formülü Kıyaslaması

Şekil 3.3 ile çalışmada hesaplanan gerçek değerlerden bazıları görülmektedir.


SQL Server Enterprise Manager - [Data in Table 'ISAWordTaxonomyS

File Window Help

```
SELECT    Word1, Word2, Similarity, WuPalmerSimilarity
FROM      ISAWordTaxonomySimilarityTable
WHERE     (Similarity <> WuPalmerSimilarity) AND (Word1 = 'bilgisayar')
ORDER BY WuPalmerSimilarity DESC
```

Word1	Word2	Similarity	WuPalmerSimilarity
bilgisayar	sunucu	0,857143	0,923077
bilgisayar	makine	0,833333	0,909091
bilgisayar	mekanizma	0,666667	0,727273
bilgisayar	ayak	0,666667	0,727273
bilgisayar	adaptör	0,666667	0,727273
bilgisayar	taban	0,666667	0,727273
bilgisayar	yay	0,571429	0,615385
bilgisayar	yöntem	0,5	0,6
bilgisayar	araç	0,5	0,6
bilgisayar	bağlantı	0,5	0,6
bilgisayar	paket	0,5	0,545455
bilgisayar	uydu	0,5	0,545455
bilgisayar	posta	0,5	0,545455
bilgisayar	kalıp	0,5	0,545455
bilgisayar	ortam	0,5	0,545455
bilgisayar	çubuk	0,5	0,545455
bilgisayar	biz	0,428571	0,461538
bilgisayar	cep	0,428571	0,461538
bilgisayar	faks	0,428571	0,461538
bilgisayar	blok	0,333333	0,444444
bilgisayar	örnek	0,333333	0,444444
bilgisayar	yol	0,333333	0,444444
bilgisayar	nesne	0,333333	0,444444
bilgisayar	alan	0,333333	0,444444
bilgisayar	yapı	0,333333	0,444444
bilgisayar	ip	0,333333	0,444444

Şekil 3.3 : Örnek Wu-Palmer ve Önerilen Yöntem Sözcük Benzerlikleri

3.3.3 Dokümanlar Arası Benzerlik

Sözcük benzerliğinin doküman benzerliğine dönüştürülmesi için Tablo 3.4' de ufak bir örnek ile gösterilen ağırlıklı ortalama yöntemi seçilmiştir.

Tablo 3.4 : Dokümanlar Arası Benzerlik Hesabı Örneği

ÖRNEK - 4 sözcükle tanımlanan 2 dokümanın sözlük kullanılarak benzerliği					
		W21	**W22**	**W23**	**W24**
	W11	S11	S12	S13	S14
	W12	S21	S22	S23	S24
	W13	S31	S32	S33	S34
	W14	S41	S42	S43	S44
Benzerlik	=	W11 * W21 * S11 + W11 * W22 * S12 + ... + W12 * W21*S21 + W12 * W22 * S22 + ...			

Tablo 3.4' te satırlarda birinci dokümanı niteleyen sözcükler, sütunlarda ikinci dokümanı niteleyen sözcükler yazılmıştır. İki doküman için de sözcük seçimi sonucunda 4 tane anahtar sözcük üzerinden hesaplama yapılacağı varsayılmıştır. Sözcüklerin ağırlıkları, sözcüğün dokümanda kullanım sayısının dokümandaki toplam sözcük sayısına bölünmesi ile bulunmaktadır. Matrisin hücrelerinde yazan değerler ise sözcükler arası ikili benzerlik değerleridir ve 3.3.2'de anlatıldığı üzere hesaplanan değerlerdir. Kavramsal sözlük kullanılarak bulunan benzerlik değerleri bu şekilde ağırlıklı ortalama şeklinde ağırlıklarla çarpılıp toplanmak suretiyle doküman arası ikili benzerlik değerleri bulunmaktadır.

3.4 Çizge Tabanlı Demetleme

Demetleme, benzerlik sonuçlarının yorumlanmasını kolaylaştırmak için yapılmaktadır. Bütün internet sitesine ait dokümanları insan gözü ile incelemek mümkün olmayacakken demetleme sonrasında demetlerin içini inceleyerek ufak parçalar halinde tüm veri analiz edilebilmiş olur. Buna ek olarak demetleme göstergelerini kullanarak demetleme performanslarından hareketle dolaylı olarak benzerlik sonuçları yorumlanabilir [11]. Bu çalışma kapsamında ikili doküman benzerlikleri söz konusu olduğu için, sonuçlar çizge tabanlı bir demetleme yöntemi ile analiz edilmiştir [9]. Burada kullanılan çizge tabanlı yöntem sezgisel çalışmaktadır. Sezgisel yöntemlerde her çalıştırmada farklı sonuçlar alınabilir. Bu

nedenle güvenilir sonuçlar ile çalışmak için, sezgisel yöntemler arka arkaya birkaç kez çalıştırılıp bütün sonuçlar arasından daha çok tekrarlanan ve çeşitli başarım ölçütlerine göre daha başarılı sonuçlar veren iki alternatif çözüm üzerinde işlemler yapılmıştır.

4. YAPILAN DENEYLER

Bu çalışmada dokümanlar www.bidb.itu.edu.tr adresinden alınmıştır. Bu web sitesinde toplam 480 doküman bulunmaktadır. Dokümanlar İTÜ BİDB hakkında bilgi vermektedir ve bilgisayar sistemlerine ait çeşitli teknik konuları anlatmaktadır. Deneyler, bu araştırma kapsamında cevap bulunmak istenen sorulara göre gerçeklenmiştir. Öncelikle kavramsal sözlükteki bütün ilişkilerin kullanılıp kullanılmayacağı araştırılmıştır. Bu sıradaki deneylerde önerilen yöntem ile Wu-Palmer yönteminin ikisi birden gerçeklendiği için iki yöntemi kıyaslamak da mümkün olmuştur. Deney 1 ile önerilen yöntem sözlükteki bütün ilişkileri kullanarak yapılmıştır. Deney 2 ile Wu-Palmer yöntemi sözlükteki bütün ilişkileri inceleyerek işlem yapmıştır. Aynı şekilde Deney 3 ile önerilen yöntem sadece üst kavram ilişkileri ile denenirken; Deney 4' te ise Wu-Palmer yöntemi sadece üst kavram ilişkileri ile uygulanmıştır. Birinci grup olarak adlandırılan bu dört deneyde de sözcükler dokümanlardaki ortalama üzerindeki sözcüklerden seçilmiştir. Sözcük seçimi ve incelenen ilişkiler fazla olduğu için işlem zamanı yüksek çıkmaktadır. Bu nedenle bütün dokümanlar incelenmeyip en çok sayıda sözcük içeren belirli bir doküman alt kümesi ile deneyler yapılmıştır. İlk grup deneylerde elde edilen sonuçlar yorumlanarak ileriki deneyler için fikir vermiştir. Örneğin ilk grup deneylerde bütün ilişkileri incelemenin büyük fayda sağlamaması sebebiyle göz ardı edileceği sonucu çıkmış ve ileriki deneylerde sadece üst kavram ilişkileri incelenmiştir. Bunun dışında ilk grup deneylerde önerilen yöntem ve Wu-Palmer yöntemi kıyaslanmıştır ve önerilen yöntemin üstün geldiği durumlar gözlenmiştir.

İkinci grup deneylerde, dokümanlardaki en çok kullanılan 10 sözcük dikkate alınarak kavramsal sözlük kullanan yöntemler ile sözlüksüz yöntemler kıyaslanmıştır. Bu araştırmalar sırasında önerilen yöntem ile Wu-Palmer yönteminin ikisi birden gerçeklenip kıyaslama yapılmıştır. Deney 5 ile önerilen yöntem sadece üst kavram ilişkileri ile çalıştırılmıştır. Deney 6 ile Wu-Palmer yöntemi sadece üst kavramlarla çalıştırılmaktadır. Deney 7 ve Deney 8 de kosinüs benzerliği hesaplanmaktadır. Deney 8 normal olarak kosinüs benzerliği hesaplaması iken Deney 7 de kavramsal

sözlükte bulunan sözcükler arasından seçilerek test yapıldığı için kavramsal sözlüğün yetersiz kalmadığı durumu örneklemektedir. Bu deneylerde de önerilen yöntemin Wu-Palmer yönteminden başarılı olduğu sonuçlar alınmasının yanı sıra sözlüksüz yöntemler ile hep daha iyi sonuçlar alınmıştır.

4.1 Deneylerde Kullanılan Demetleme Göstergeleri

Bu çalışma sırasında iki grup deney yapılmıştır. İlk grup deneylerde belirli bir doküman alt kümesi üzerinden kavramsal sözlükteki bütün ilişkilere bakılıp bakılmaması değerlendirilmiştir. İkinci grup deneylerde ise doküman benzerliği hesaplanması için kavramsal sözlük kullanan yöntemler ile sözlüksüz yöntemler arasında inceleme yapılmıştır. İki grup deneyde de önerilen yöntem Wu-Palmer yöntemi ile karşılaştırılmıştır. Deneyler sırasında kullanılan demetleme göstergeleri; Jaccard göstergesi, Dunn göstergesi, Davies-Bouldin göstergesi ve Silhouette göstergesidir [11, 18].

İlk grup deneylerde doküman alt kümesi kullanılma nedeni hem uygulamanın işlem zamanını azaltmak hem de dokümanların insanlar tarafından elle demetlenmesini sağlamak içindir. İlk grup deneylerde kullanılan 28 dokümanlık alt küme 3 kişilik bir ekip tarafından daha benzer dokümanlar aynı grupta olacak şekilde demetlenmiştir. Deneylerde hesaplanan benzerlikler çizge tabanlı demetleme yöntemi ile demetlendikten sonra, insanlar tarafından yapılan gruplama ile Jaccard göstergesine göre kıyaslanmaktadır [11]. Jaccard göstergesi, Jaccard benzerliği ile tamamen aynıdır. Bu gösterge, bulunan deney sonuçlarının insanların yaptığı gruplamaya olan benzerliğini ölçerek deney sonuçlarının başarımını belirlemektir.

Jaccard göstergesinin hesaplanması için Tablo 4.1 ve Eşitlik (4.1), (4.2), (4.3) incelenebilir. Eşitlik (4.3)' deki ifadenin 2.4' te anlatılan Jaccard benzerliği ifadesiyle aynı olduğuna dikkat ediniz.

Tablo 4.1 : Demetleme Jaccard Göstergesi Parametreleri

Demetleme Sonucu	**Grup Bilgisi**	**Sembol**
Aynı	Aynı	a
Aynı	Farklı	b
Farklı	Aynı	c
Farklı	Farklı	d

N : Toplam doküman sayısı

M : İkili doküman benzerliği sayısı

$$M = \frac{N * (N-1)}{2} \quad \textbf{(4.1)}$$

$$a+b+c+d = M \quad \textbf{(4.2)}$$

$$\text{Jaccard göstergesi} = \frac{a}{a+b+c} \quad \textbf{(4.3)}$$

İkinci grup deneylerde Dunn, Davies-Bouldin ve Silhouette göstergeleri kullanılmıştır. Dunn göstergesi Eşitlik (4.4), (4.5) ve (4.6) ile gösterilmektedir. Bu eşitliklerde nc toplam demet sayısını, c_p bir demeti, p ise demet içindeki bir veriyi göstermektedir. $s(q,w)$ ifadesi de q ve w dokümanları arasındaki benzerliktir.

$$D_{nc} : \min_{p=1,\ldots,nc} \{ \min_{r=1,\ldots,nc} (\frac{d(c_p, c_r)}{\max_{t=1,\ldots,nc} diam(c_t)}) \} \quad \textbf{(4.4)}$$

$$\text{Demetler arası uzaklık} : d(c_p, c_r) = \min_{q \in c_p, w \in c_r} s(q,w) \quad \textbf{(4.5)}$$

$$\text{Demetin çapı} : diam(c) = \max_{q,w \in c} s(q,w) \quad \textbf{(4.6)}$$

Demetleme başarımı olarak demet içindeki dokümanlar birbirine en fazla benzerken, demetler arasındaki dokümanların da birbirinden en fazla farklılık göstermesi istenmektedir. Demet içi doküman benzerliğinin en fazla olması, Dunn göstergesindeki Eşitlik (4.6) ile demetin çapı olarak gösterilmiştir. Demetler arasındaki uzaklığın (yani farklılığın) en fazla olması da; Dunn göstergesindeki Eşitlik (4.5) ile en az benzerlik değeri olarak gösterilmiştir. Zaten uzaklık ile benzerlik arasında böyle ters bir ilişki vardır. Aralarındaki uzaklık değeri en yüksek olan dokümanlar birbirine en az benzerdir. Dunn göstergesi tanımında da bu durum açısından çelişki yaratacak bir durum bulunmamaktadır. Dunn göstergesi sonuç hesabında da demetler arası en büyük uzaklık, demetler arasındaki en büyük çapa bölünmektedir. Bu hesaptan hareketle Dunn gösterge değeri büyük olduğu sürece performansın daha yüksek olacağı sonucu çıkarılabilir.

Davies-Bouldin (DB) göstergesi de Eşitlik (4.7), (4.8), (4.9) ile gösterilmektedir. Bu eşitliklerde nc toplam demet sayısı, c_g ise herhangi bir demettir. R_g, g demeti için tanımlanmış olan DB göstergesi R değeridir. R_{gh} ise g ve h demetleri için tanımlanmış olan ikili Davies-Bouldin göstergesi R değeridir. R değeri ise DB göstergesinde demet içi ve demetler arası uzaklıkları dikkate alarak tanımlanmış olan bir benzerlik ifadesini göstermektedir.

$$R_{gh} = \frac{diam(c_g) + diam(c_h)}{d(c_g, c_h)} \tag{4.7}$$

$$R_g = \max_{g=1,\ldots,nc, g \neq h} R_{gh} \tag{4.8}$$

$$DB_{nc} = \frac{1}{nc} \sum_{g=1}^{nc} R_g \tag{4.9}$$

DB göstergesi ise Dunn göstergesinin daha geliştirilmiş halidir. Buradaki demet çapı ve demetler arası uzaklık Dunn göstergesinde olduğu gibi Eşitlik (4.5) ve Eşitlik (4.6) ile hesaplanmaktadır. Burada yine demetin çapı ve demetler arası farklılığa göre hesap yapılmaktadır. Fakat her demet için olası tüm durumlar arasından en yüksek R değeri seçilerek işlem yapılmaktadır. Bu açıdan daha ayrıntılı bir işlem söz konusudur. En sonda bütün demetler için hesaplanan değerlerin ortalaması alınarak gösterge sonucu elde edilmiştir. Dunn göstergesindeki orantının tersi şeklinde bir orantı hesabı olduğu için DB göstergesi açısından daha düşük değerlere sahip olmanın daha performanslı olacağı sonucu çıkarılabilir.

Silhouette göstergesi yöntemi de Eşitlik (4.10), (4.11), (4.12), (4.13), (4.14) ile açıklanmıştır. Bu eşitliklerde N toplam doküman sayısı, c_i bir demet, i demet içindeki bir veri, N_{c_i} c_i demeti içindeki toplam doküman sayısını göstermektedir. $s(i,j)$, i ve j dokümanları arasındaki benzerlik değeridir. $d(i,j)$ ise i ve j dokümanları arasındaki uzaklıktır. $a(i)$ i verisinin bulunduğu demet içindeki diğer verilere olan ortalama uzaklık, $b(i)$ ise kendisine en yakın demete olan uzaklığıdır.

$s(i)$: i. doküman için silhouette katsayısı

$$d(i,j) = 1 - s(i,j) \tag{4.10}$$

$$a(i) = \frac{1}{N_{c_k}} \sum_{i,j \in c_k} d(i,j) \quad \textbf{(4.11)}$$

$$b(i) = \min_{l \neq k} \{ \frac{1}{N_{c_l}} \sum_{i \in c_k, j \in c_l} d(i,j) \} \quad \textbf{(4.12)}$$

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad \textbf{(4.13)}$$

$$\text{Silhouette Göstergesi} = \frac{1}{N} \sum s(i) \quad \textbf{(4.14)}$$

Silhouette göstergesinin incelenmesinin nedeni, Dunn ve DB den farklı bir bakış açısı ile de değerlendirme yapılmak istenmesidir. Dunn ve DB göstergeleri, demetler arası benzerlik az demet içi benzerlik çok olsun düşüncesiyle değerlendirir. Silhouette göstergesinde ise bütün dokümanlar için tek tek Eşitlik (4.13) ile gösterilen silhouette katsayısı hesaplanıp bütün dokümanların katsayılarının ortalaması alınarak sonuç gösterge değeri bulunur. Buradan da anlaşılabileceği üzere silhouette göstergesi yerel bir bakış açısına sahiptir. Dokümanların belirlenen demetler arasından hangisinde olmasının daha uygun olmasına karşın hangisine yerleştirildiği arasındaki uyumun ölçüsü olarak sayısal bir değer verir. Silhouette göstergesinin çalışabilmesi için birden fazla demet olmalıdır ve her demette de en az iki eleman olmalıdır. Eşitlik (4.13)' ten de anlaşılabileceği üzere $|s(i)| \leq 1$ olacaktır. Burada değer 1 değerine ne kadar yakın çıkarsa o kadar olması gereken ile uyumlu, sağlıklı ve başarılı bir sonuç var demektir. Silhouette göstergesinin yerel uyum açısından bakış açısı, ancak demetler belli iken her dokümanın en uygun demetini tespitte başarım sağlar [18]. Fakat genel demetleme performansı açısından yorumu DB göstergesi kadar başarılı değildir [19].

4.2 İlk Grup Deneylerin Sonuçları

İlk grup deneylerde kullanılan dokümanlara ait bilgiler Tablo 4.2' de gösterilmektedir. Tablo 4.2' deki gruplama 3 kişi tarafından ortaklaşa yapılan analizlerin sonucunda belirlenmiştir. Deney sonuçlarının çizge tabanlı yöntemle demetlenmesi ardından bulunan gruplar Tablo 4.2' deki insan gruplaması sonuçları

ile Jaccard göstergesine göre karşılaştırılmaktadır. Deney sonuçları ve bulunan Jaccard indeksi değerleri Tablo 4.3' te gösterilmektedir.

Tablo 4.2: İnternet Sitesinden Seçilen Alt Kümenin Özellikleri

Doküman No	**Toplam Sözcük**	**Seçilen Sözcük**	**Doküman Adı**	**Grup No**
1	3655	95	Donanım Bilgileri	0
2	1942	81	İnternet tarihçesi	0
3	1555	46	Windows sistem hesapları	0
4	1427	51	Dağıtımlı hesaplama sistemi	0
5	1239	42	MSSQL Query Analyzer	0
6	1201	60	Kablosuz iletişim teknolojileri	0
7	1149	40	xp uzak masaüstü	0
8	786	41	İnternet explorer açıkları	1
9	672	30	Truva atları	1
10	610	29	İTÜ E-posta liste kullanım politikaları	1
11	588	31	İTÜ ögrenci lab kullanım politikası	1
12	487	24	Spyware	1
13	481	24	İçerik Yönetim Sistemi	1
14	421	47	IIS 6.0	1
15	177	10	YGG projeleri–1	2
16	177	10	YGG projeleri–2	2
17	101	13	İTÜ/NET projesi	2
18	100	17	İTÜ BİDB Proje Yönetim Sistemi Projesi	2
19	91	11	İTÜ E-posta Kimlik Doğrulama Projesi	2
20	89	9	İTÜ DS Otomasyon Projesi	2
21	88	13	ITU MBS Projesi	2
22	248	9	Spybot casusu	3
23	224	21	Blackmal solucanı	3
24	220	21	Mytob.ig virüsü	3
25	202	22	Beagle virus	3
26	199	17	Mytob.hl virüsü	3
27	193	14	kiman virüsü	3
28	192	19	sasser solucanı	3

Tablo 4.3: İnternet Sitesi Alt Kümesi Deney Sonuçları

Deney1 (Önerilen yöntemin bütün ilişkilerle incelenmesi)		**Jaccard İndeksi**
Sonuç 1	0 0 0 0 0 0 1 0 0 0 0 0 1 1 0 0 1 0 0 0 0 0 0 0 0 1 0	0,189
Sonuç 2	0 1 1 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0,278
Deney2 (Wu-Palmer yönteminin bütün ilişkilerle incelenmesi)		
Sonuç 1	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 1 0	0,229
Sonuç 2	0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 1 0 1 1 1 1 0 0 0 0 0 0 0	0,255
Deney3 (Önerilen yöntemin sadece üst kavramlarla incelenmesi)		
Sonuç 1	0 0 0 0 1 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1	0,204
Sonuç 2	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 1 1 0 0 0 1 2 0 2	0,229
Deney4 (Wu-Palmer yönteminin sadece üst kavramlarla incelenmesi)		
Sonuç 1	0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 1 0 0	0,209
Sonuç 2	0 0 0 0 0 0 0 0 0 1 0 0 1 0 1 1 0 1 1 1 1 0 0 0 1 0 0 0	0,271

İlk grup deneylerde Tablo 4.2' de gösterildiği üzere internet sitesinin en çok sayıda sözcük içeren bir alt kümesi olan 28 doküman için çalışma yapılmıştır. Dokümanlardaki sözcük seçimi, dokümandaki ortalama sözcük sayısından fazla kullanılan sözcüklerin tamamı şeklinde yapılmıştır. İki doküman arasında benzerlik hesaplanırken sözcükler arasında kartezyen çarpım yapılıp bütün ilişkiler incelendiği için sözcük sayısındaki az bir artış bile yöntemin çalışma zamanını ciddi derecede uzatmaktadır. Tablo 4.2' ye bakılırsa ilk doküman için 95 tane sözcük olduğu görülmektedir. Bu da kartezyen çarpım açısından oldukça büyük değerlere götüren yüksek bir sayıdır ve çalışma zamanını ciddi derecede yavaşlatmaktadır.

Tablo 4.3' te Deney 1 ile gösterilen çalışma sırasında önerilen yöntem, Türkçe kavramsal sözlükte bulunan 13 ilişkinin hepsini inceleyerek gerçeklenmiştir. Deney 2 ile gösterilen çalışma sırasında Wu-Palmer yöntemi, Türkçe kavramsal sözlükte bulunan 13 ilişkinin hepsini inceleyerek gerçeklenmiştir. Deney 1 ile 140 saatte sonuç alınabilmiştir. Bütün ilişkiler incelendiği için Türkçe kavramsal sözlük 37220 veriden oluşmaktadır. Bu çalışma sırasında, 28 tane dokümana ait 28 * (28-1) / 2 = 378 tane ikili doküman benzerliğini hesaplayabilmek için 59188 tane ikili sözcük benzerliği hesaplanmıştır. Ayrıca Deney 1 sırasında, önerilen yöntem bütün ilişkilere göre benzerlik hesaplarken Wu-Palmer sözcük benzerliğini de hesaplayıp kaydetmektedir çünkü yapılan iki işlem birbirine oldukça benzerdir. Bu durum sebebiyle de zaten çalışma zamanı uzun olan uygulama biraz daha geç sonuç üretmiştir. Bu durumda Wu-Palmer yöntemi kullanılarak yapılan doküman benzerliklerini hesaplama süresi de 3,5 saat olmuştur.

Deney 1 sonucunda çizgesel demetleme yönteminin önerdiği en olası 2 çözüm Tablo 4.3' te gösterilmiştir. Jaccard benzerliği daha yüksek olması sebebiyle 2. olasılık daha muhtemeldir. Deney 2, bu deney ile kıyaslanabilir. Deney 2 sonucunda çizgesel demetleme yönteminin önerdiği en olası 2 çözüm Tablo 4.3' te gösterilmiştir. Jaccard benzerliği daha yüksek olması sebebiyle 2. olasılık daha muhtemeldir. Deney 1' e ait 2. sonuç ile Deney 2' ye ait 2. sonuç incelenirse; Deney 1 sonucunun daha iyi olduğu görülmektedir. Yani sözlükteki bütün ilişkiler kullanılarak çalışıldığında önerilen yöntem, Wu-Palmer yönteminden iyi sonuç vermiştir.

Tablo 4.3' te Deney 3 ile gösterilen çalışma sırasında önerilen yöntem, Türkçe kavramsal sözlükteki üst kavram ilişkisine bakarak çalışmaktadır. Deney 4 ile gösterilen çalışma sırasında Wu-Palmer yöntemi, Türkçe kavramsal sözlükteki üst kavram ilişkisine bakarak çalışmaktadır. Deney 3' ün toplam çalışma zamanı 12,25 saattir. Deney 1'in 140 saatlik çalışma zamanı kıyaslanarak Deney 1'de yapılan işlem yükü daha iyi anlaşılabilir. Bütün ilişkiler incelenmeyip sadece üst kavram ilişkileri incelendiği için Türkçe kavramsal sözlük [13] 21042 veriden oluşmaktadır. Bu çalışma sırasında, 28 tane dokümana ait 28 * (28–1) / 2 = 378 tane ikili doküman benzerliğini hesaplayabilmek için 41731 tane ikili sözcük benzerliği hesaplanmıştır. Deney 3 ile Wu-Palmer yönteminin sözcükler arası benzerliği hesaplanmıştı. Deney 4 kapsamında sadece ağırlıklı ortalamalar ile sözcük benzerlikleri birleştirilerek Wu-Palmer doküman benzerliğine geçilmiştir. Bu işlem toplam 6 saat sürmüştür.

Deney 3 sonucunda çizgesel demetleme yönteminin önerdiği en olası 2 çözüm Tablo 4.3' te gösterilmiştir. Jaccard benzerliği daha yüksek olması sebebiyle 2. olasılık daha muhtemeldir. Deney 4, bu deney ile kıyaslanabilir. Deney 4 sonucunda çizgesel demetleme yönteminin önerdiği en olası 2 çözüm Tablo 4.3' te gösterilmiştir. Jaccard benzerliği daha yüksek olması sebebiyle 2. olasılık daha muhtemeldir. Deney 3' e ait 2. sonuç ile Deney 4' e ait 2. sonuç seçilirse; Deney 4 sonucu daha iyi olduğu görülmektedir. Yani sadece sözlükteki üst kavram ilişkileri incelendiği zaman Wu-Palmer yöntemi daha iyi sonuç vermiştir.

Tablo 4.3' e genel olarak bakılırsa bütün Jaccard değerlerinin 0,2 civarında olduğu görülecektir. Bütün ilişkileri inceleyip incelememek arasında büyük zaman farkı olduğu da görülmektedir. Bu nedenle, bütün ilişkileri incelemek ile sadece üst kavram ilişkilerini incelemek arasında büyük bir fark olmadığı ve ileriki deneylerde sadece üst kavram ilişkileri incelenerek çalışılabileceği kararına varılmıştır.

4.3 İkinci Grup Deneylerin Sonuçları

İkinci grup deneylerde alınan bütün sonuçlar için demetleme gösterge değerleri Tablo 4.4' te gösterilmektedir.

Tablo 4.4 : İnternet Sitesi Bütün Sayfalar Demetleme Göstergeleri

Deney5 (Önerilen yöntem sadece üst kavram ve sözlükten ilk 10 anahtar sözcük)						
Sonuç 1	**Dunn**	0,00027	**DB**	7118,85	**Silhouette**	0,6375
Sonuç 2	**Dunn**	0,00027	**DB**	1179,71	**Silhouette**	0,1973
Deney6 (Wu-Palmer yöntem sadece üst kavram ve sözlükten ilk 10 anahtar sözcük)						
Sonuç 1	**Dunn**	0,00031	**DB**	4431,12	**Silhouette**	0,4381
Sonuç 2	**Dunn**	0,00031	**DB**	3176,11	**Silhouette**	0,2651
Deney7 (Kosinüs benzerliği ve sözlükten ilk 10 anahtar sözcük)						
Sonuç 1	**Dunn**	0,0052	**DB**	127,53	**Silhouette**	0,6423
Sonuç 2	**Dunn**	0,0053	**DB**	192,557	**Silhouette**	0,5417
Deney8 (Kosinüs benzerliği ve dokümanlardaki ilk 10 anahtar sözcük)						
Sonuç 1	**Dunn**	0,0025	**DB**	802,054	**Silhouette**	0,8018
Sonuç 2	**Dunn**	0,0025	**DB**	802,054	**Silhouette**	0,833

Tablo 4.4' te Deney 5 ile gösterilen çalışmada önerilen yöntem, Türkçe kavramsal sözlükte üst kavram ilişkileri ile çalışmıştır ve dokümanın en çok kullanılan sözcükleri arasından kavramsal sözlükte de bulunan ilk 10 tanesi incelenmiştir. Deney 5, 150 saatte çalışmıştır. Toplam hesaplanan sözcük benzerliği sayısı 94627 olmuştur. Bu deney sonucunda 480 doküman için 480 * (480–1) / 2 = 114960 tane ikili doküman benzerliği hesaplanmıştır. Tablo 4.4' te Deney 6 ile gösterilen çalışmada Wu-Palmer yöntemi, Deney 5'e benzer şekilde sadece Türkçe kavramsal sözlükteki üst kavram ilişkileri ile ve sözlükte bulunan ilk 10 sözcük ile çalışmıştır. Deney 6, 160 saatte çalışmıştır. Çalışma süresi, büyük veri arasından sorgulama nedeniyle uzun çıkmıştır.

Deney 5 ile Deney 6 önerilen yöntem ile Wu-Palmer yöntemini kıyaslamak için incelenebilir. İki deneyde de en iyi sonuç veren değerleri karşılaştırabiliriz. Dunn göstergesinde değerin yüksek olması istenir. DB göstergesinde düşük, Silhouette göstergesinde de yüksek olması istenir. Dunn göstergesine göre Wu-Palmer yöntemi daha iyi sonuç vermektedir fakat değerler birbirine yakındır. Bunun dışında DB ve Silhouette göstergelerinde ise önerilen yöntem daha iyi sonuç vermektedir. DB göstergesinin, incelenen göstergeler arasındaki en iyi gösterge olduğu unutulmamalıdır [19].

Deney 5 için demetleme yönteminin sonuçları Tablo 4.5' te incelenebilir. Deney 6 için demetleme sonuçları içeriği de Tablo 4.6 ile incelenebilir. Bu tablolar demetlerin içinden rasgele seçilen bazı dokümanların başlıklarını göstermektedir.

Tablo 4.5 : Deney 5 Demetleme Sonuçlarının İçeriği

DENEY 5 **(Önerilen yöntem sadece üst kavram ve sözlükten ilk 10 anahtar sözcük)**				
Sonuç 1				
Demet1(362 syf.)		**Demet2 (118 syf.)**		
VLAN Banleed Virüsü Bilgisayar Güvenliği Telnet-ftp-ssh Kullanıcı Şifre İşlemleri		İTÜ'de Akademik Hesaplama Enterprise Manager İTÜ BİDB Projeler İTÜ Ayniyat Projesi Yazıcı Kota İşlemleri		
Sonuç 2				
Demet1 **(334 syf)**	**Demet2** **(98 syf)**	**Demet3** **(4 syf)**	**Demet4** **(42 syf)**	**Demet5** **(2 syf)**
Mydoom.fs Vir. Mydoom.cf Vir. Pine Ayarları SQL Platformu Kull. VT Proj.	OE Kimlik Yap. Opera IMAP Fedora IP Ayarı Query Designer Win98'de MAC	PST yapma Kişisel web syf. İYS Projesi İTÜ Web Hizm.	Akıllı Kart Proj. Pers. Perf. Proj. Öğr. Lab Kuralı Kull. Politikalar Şifre İşlemleri	Yapımda – 1 Yapımda – 2

Tablo 4.6 : Deney 6 Demetleme Sonuçlarının İçeriği

DENEY6(Wu-Palmer sadece üst kavram ve sözlükten ilk 10 anahtar sözcük)		
Sonuç 1		
Demet1(362 syf.)	**Demet2(118 syf.)**	
Yaha.f Virüsü Suclove.a Virüsü Linux Kök Sert. Kurulumu Thunderbird E-posta Ayarları Bağlı Tablolardan Veri Silme	SQL Security XP'de MAC adresi Windows Kök Sert. Kurulumu İTÜ DS Otomasyonu Projesi Finans Çerçevesi Projesi	
Sonuç 2		
Demet1(302 syf.)	**Demet2(105 syf.)**	**Demet3(73 syf.)**
Trojan.Lisentkey Symbos.Doomboot.C Virüsü Win98 Çoklu Dil Desteği W32.Looked.F virüsü Bootable Linux CD	.NET Framework XP'den Yazılıma Erişim Putty Kurulum Kullanım SQL Tabloları Yazılım patentleri ve AB	İTÜ OSH Projesi İYS Projesi SQL Sorgulama Teknikleri SQL Operatörleri Birkaç Tablo Birleştirme

Tablo 4.4' te Deney 7 ve Deney 8 ile gösterilen çalışmalar kosinüs benzerlik yöntemine ilişkin test sonuçlarıdır. Kosinüs benzerliği kavramsal sözlük

kullanmadan çalışan bir yöntemdir. Bu çalışma kapsamında kavramsal sözlük kullanmak veya kullanmamak da araştırılmaktadır. Bu nedenle sözlüksüz yöntemlerle de testler yapılmıştır. Tablo 4.4' e bakarak her deneyde en yüksek başarımı veren sonuçlar arasında kıyaslama yapılabilir. Dunn göstergesinde sonuçların yüksek olması istenir. DB göstergesi Dunn göstergesinin gelişmiş halidir ve sonuçların düşük olması istenir. Silhouette göstergesi ise her dokümanın tanımlı demetler arasından hangisinde olması en uygundur mantığına göre toplam bir uyum değeri verir ve değerin yüksek olması istenir. Bu göstergeler arasından en başarılı olan DB göstergesidir [19]. Bu bilgiler ışığında Tablo 4.4' ten seçim yapılırsa ilginç bir sonuç göz önüne çıkmaktadır. Bütün göstergelerde kosinüs benzerliğini kullanan Deney 7 ve Deney 8 daha iyi sonuç vermiştir. Kavramsal sözlük yetersiz kalmıyorsa ve yöntemin performansı yeterli ise sözlük kullanmayan yöntemlerden iyi sonuç vermesi beklenmekteydi. Buradaki kavramsal sözlük yetersiz değilken durumunu gerçeklemek için kavramsal sözlükte bulunan sözcükler seçilerek Deney 7 uygulanmıştır. Buna rağmen hem Deney 7 hem Deney 8 de sonuçlar daha başarılı çıkmıştır. Ayrıca Dunn ve DB göstergeleri, demetler içi benzerlik fazla demetler arası benzerlik az olmasına göre çalışmaktadır. Kosinüs benzerliğinde de terim ağırlığı hesabı ile sitedeki bütün dokümanlar arasında az geçen ama bulunduğu dokümanda çok kullanılan sözcükler irdelenmektedir. Bu nedenle Dunn ve DB göstergeleri yöntemi ile kosinüs benzerliği yönteminde paralellik söz konusu mudur ve bu göstergelerde kosinüs benzerliği için artı yönde çalışma olabilir mi düşüncesi doğmuştur. Bu düşüncenin ardından farklı bir bakış açısına sahip olan Silhouette göstergesi ile de deneme yapılmış fakat burada da kosinüs benzerliğinde daha iyi sonuç alınmıştır. Bütün bu denemelere dayanarak kosinüs benzerliğinin çalışmada kullanılan Wu-Palmer yönteminden de önerilen yöntemden de daha iyi sonuç verdiği söylenebilir. Ayrıca kosinüs benzerliği yöntemi bütün sayfalar için yirmi dakikada çalışmıştır. Çalışma zamanının da diğer yöntemlere göre bu şekilde düşük olması sonucunda, performans / zaman ölçütü açısından bu yöntemin kullanımı tercih edilebilir. Bu çalışma kapsamında kullanılan yöntemler arasında böyle bir sonuç alınmış olsa bile kavramsal sözlük kullanan ve kullanmayan pek çok yöntem açısından bu şekilde bir inceleme yapılabilirse bu konuda daha sağlıklı bir yorum getirmek mümkün olacaktır.

Tablo 4.4 ile gösterilen Deney 7 ve Deney 8' e ait sonuçlar için çizgesel demetleme yönteminin oluşturduğu demetlerin içerikleri Tablo 4.7' de incelenebilir.

Tablo 4.7 : Deney 7 ve Deney 8 Demetleme Sonuçlarının İçeriği

<table>
<tr><th colspan="4">DENEY 7 (Kosinüs benzerliği ve sözlükten ilk 10 anahtar sözcük)</th></tr>
<tr><th colspan="4">Sonuç 1</th></tr>
<tr><th>Demet1(415 syf.)</th><th>Demet2(55 syf.)</th><th>Demet3(6 syf.)</th><th>Demet4(4 syf.)</th></tr>
<tr><td>Kiman.b Virüsü
Zotob.E Virüsü
Areses.F Virüsü
SQL Kullanıcıları
Virüslü E-posta
Yön.</td><td>Kümeli VT Proj.
İŞAL Proj.
Proje Yön. Sis.
Proj.
Pers. Otomasyonu
IP Yön. Sis. Proj.</td><td>Öğrenci Kullanıcı
İşl.
İTÜ Kullanıcı İşl.
Akademik/İdari İşl.
LAN
Yazıcı İşlemleri
Enterprise Manager</td><td>Destek Sayfaları
Atölye Hizmeti
Kişisel Web Syf.
Lisanslı
Yazılımlar</td></tr>
<tr><th colspan="4">Sonuç 2</th></tr>
<tr><th colspan="2">Demet1(383 syf.)</th><th>Demet2(73 syf.)</th><th>Demet3(24 syf.)</th></tr>
<tr><td colspan="2">Looked.F Virüsü
Zotob.H Virüsü
SSH Kullanımı
Autex.C Virüsü
Patran ve Nastran</td><td>Kümeli E-posta Proj.
Akıllı Kart Proj.
Ayniyat Otomasyon Proj.
SQL Kimlik Doğrulama
SQL 2005</td><td>MS Outlook Ayarları
Güvenli E-posta Ayarları
MS OE Ayarları
Opera'da SSL Desteği
Yapım Aşamasındadır</td></tr>
<tr><th colspan="4">DENEY 8 (Kosinüs benzerliği ve dokümanlardaki ilk 10 anahtar sözcük)</th></tr>
<tr><th colspan="4">Sonuç 1</th></tr>
<tr><th colspan="2">Demet1(404 syf.)</th><th colspan="2">Demet2(76 syf.)</th></tr>
<tr><td colspan="2">Beagle.CE Virüsü
Welchia Virüsü
SQL Roller
Fluent 5.5
SQL Replication</td><td colspan="2">İTÜ EKP Projesi
E-eğitim Yön. Sis. Projesi
İTÜ BİDB Perf. Monitor Projesi
Mathematica
Yazılım Derlemeleri</td></tr>
<tr><th colspan="4">Sonuç 2</th></tr>
<tr><th colspan="2">Demet1(418 syf.)</th><th colspan="2">Demet2(62 syf.)</th></tr>
<tr><td colspan="2">Nibu.N Virüsü
Alcra.D Virüsü
DHCP Ayarları
İTÜ Sertifika Servisleri
Mocbot.A Virüsü</td><td colspan="2">İTÜ BAP Projesi
IP Yön. Sis. Projesi
KYS 2.0 Projesi
Erişim Sorunları
Yurtlar Destek</td></tr>
</table>

5. SONUÇLAR VE TARTIŞMALAR

Bu çalışma kapsamında kavramsal sözlükteki bütün ilişkileri inceleyip incelememe sorusuna cevap bulmak için ilk grup deneylere bakılabilir. İlk grup deneylerin sonuçları açısından Tablo 4.2' de daha yüksek Jaccard göstergesi değeri veren sonuçlar seçilebilir. Tablo 4.2'de görüldüğü üzere bütün deneyler için 2. sonuçlar daha başarılıdır. Bu sonuçlar için incelemek gerekirse; kavramsal sözlükteki bütün ilişkilerin önerilen yöntem ve Wu-Palmer için incelendiği Deney 1 ve Deney 2, sadece üst kavram ilişkilerinin yine bu yöntemler için incelendiği Deney 3 ve Deney 4' ten daha iyi sonuç vermiştir. Bu bağlamda bütün ilişkileri incelemek faydalı olmuştur denebilir ama bütün değerler 0,2 civarındadır ve sonuçlar arasında büyük farklar yoktur. Ayrıca bütün ilişkileri incelemenin daha yüklü bir işlem olduğu da düşünülürse ilerideki deneylerde sadece üst kavram ilişkisine bakma kararı alınmıştır. Kavramsal sözlükteki bütün ilişkilerin ağırlık verilerek değerlendirmesi bütün ilişkilere göre araştırmada daha büyük fayda sağlayacaktır yalnız bu tamamen ayrı bir çalışma konusudur. Bu nedenle bu çalışmada bu şekilde inceleme yapılmamıştır ve yapılan inceleme çerçevesinde de ufak bir fayda sağlanmasına karşın ilerideki çalışmalarda sadece üst kavramlara bakarak devam etme kararı alınmıştır.

Bu araştırmada önerilen yöntem ile Wu-Palmer yönteminin başarımı kıyaslanmak istenmiştir. Bu sorunun sorulma nedeni, Eşitlik (2.2) ile gösterildiği üzere Wu-Palmer yönteminin benzerliği en temel şekilde kesişim / birleşim şeklinde tanımlaması ve bu tanımın geliştirilebileceği inancıdır. Bu çalışmada önerilen yöntem ise Eşitlik (3.2) ile gösterilmektedir. Önerilen yöntemin avantajı olarak kesişim düğümüne gidilen yollardan en uzununu dikkate alması söylenebilir. Bu sayede incelenen sözcüklerden oluşan grubun sağlamlığını, yani benzerliğini, en zayıf halkayı dikkate alarak belirlemesi ve daha gerçekçi olması amaçlanmıştır. Deney sonuçlarını yorumlayarak önerilen yöntemin başarımı kıyaslanmak istenirse, öncelikli olarak Jaccard göstergesi daha yüksek değerleri veren sonuçlarla birlikte ilk grup deneylere bakılabilir. Bu şekilde Tablo 4.2' de Deney 1 ve Deney 2 kavramsal

sözlükteki bütün ilişkiler kullanılarak bu iki yöntemi kıyaslamaktadır ve burada önerilen yöntem daha iyi sonuç vermiştir. Benzeri şekilde Tablo 4.2' de Deney 3 ve Deney 4 sözlükteki üst kavram ilişkilerini kullanarak bu iki yöntemi kıyaslamaktadır ve burada da Wu-Palmer yöntemi daha iyi sonuç vermiştir. İlk grup deneylerde yakın sonuçlar alınmıştır ve buradan hareketle iki yöntemi kıyaslamak adına kesin bir şey söylenememiştir. Bu durumda ikinci grup deneyler incelenebilir. İkinci grup deneylerin incelemesinde de göstergelerin daha iyi gösterdiği değerler seçilerek inceleme yapılabilir. Dunn göstergesi yüksek, DB göstergesi düşük, Silhouette yüksek değerler ister. Bu şekilde Tablo 4.3' e bakılarak Deney 5 ve Deney 6 kıyaslanabilir. Bu durumda Dunn göstergesi için Wu-Palmer yöntemi daha başarılı çıkmıştır ama değerler oldukça birbirine yakındır. DB göstergesinde de Silhouette göstergesinde de önerilen yöntem daha iyi sonuç vermiştir. DB göstergesinin de bu göstergeler arasındaki en iyi gösterge olduğu hatırlatılmalıdır [19]. Tablo 4.3' teki gösterge değerleri arasındaki sayısal oranlar, birebir olarak deneylerde uygulanan yöntemlerin performansları arasındaki oranlar olarak düşünülmemelidir. Fakat Tablo 4.3' teki değerlere göre sıralama yapılarak yorumlamalarda bulunulabilir. Bu yorumlamalar sonucunda ikinci grup deneylerde ağırlıklı olarak önerilen yöntemin Wu-Palmer yönteminden başarılı sonuçlar verdiği anlaşılmıştır. Sonuç olarak bütün deneyler açısından da önerilen yöntemin Wu-Palmer yönteminden başarılı çıktığı söylenebilir.

Bu çalışma kapsamında araştırılan diğer bir soru, kavramsal sözlük kullanan bir yöntem ile çalışmak veya sözlüksüz yöntemlerle çalışmak arasındaki performans farkını bulmaktı. Bu sorunun cevabı için ikinci grup deneylerin sonuçları incelenebilir. Tablo 4.3' teki göstergelerin verdiği en iyi sonuçlar için kıyaslama yapılırsa; bütün göstergelerde kosinüs benzerliğini kullanan Deney 7 ve Deney 8' in daha iyi sonuç verdiği gözlenecektir. Kavramsal sözlük yetersiz kalmıyorsa ve yöntemin performansı yeterli ise sözlük kullanmayan yöntemlerden iyi sonuç vermesi beklenmekteydi. Buradaki kavramsal sözlük yetersiz değilken durumunu gerçeklemek için Deney 7 uygulanmıştır. Buna rağmen hem Deney 7 hem Deney 8 de sonuçlar daha başarılı çıkmıştır. Ayrıca Dunn ve DB göstergeleri, demetler içi benzerlik fazla demetler arası benzerlik az olmasına göre çalışmaktadır. Kosinüs benzerliğinde de terim ağırlığı hesabı ile sitedeki bütün dokümanlar arasında az geçen ama bulunduğu dokümanda çok kullanılan sözcükler irdelenmektedir. Bu

nedenle Dunn ve DB göstergeleri yöntemi ile kosinüs benzerliği yönteminde paralellik olabileceği düşüncesi doğmuştur. Bu düşüncenin ardından farklı bir bakış açısına sahip olan Silhouette göstergesi ile de deneme yapılmış fakat burada da kosinüs benzerliğinde daha iyi sonuç alınmıştır. Bütün bu denemelere dayanarak kosinüs benzerliğinin çalışmada kullanılan Wu-Palmer yönteminden de önerilen yöntemden de daha iyi sonuç verdiği söylenebilir. Bu çalışma kapsamında kullanılan yöntemler arasında böyle bir sonuç alınmış olsa bile kavramsal sözlük kullanan ve kullanmayan pek çok yöntem açısından bu şekilde bir inceleme yapılması ardından bu konuda bir yorumda bulunmanın daha sağlıklı olacağı kararına varılmıştır.

Bu çalışmadan sonra bu alandaki gelecek çalışmalarında daha pek çok geliştirme yapılabilir. Ayrıt temelli ve düğüm temelli yöntemler birleştirilerek bir kavramsal sözlük yöntemi önerilebilir. Kavramsal sözlükteki bütün ilişkiler incelenerek, her ilişkiye farklı ağırlık verecek şekilde bir yöntem önerilebilir. Kavramsal sözlükler yetersiz kalabilme sorunu ile karşı karşıya olduğu için, yetersiz kalındığı durumda kavramsal sözlük kullanmayan yöntemlerle birleştirilerek melez bir yöntem önerilebilir. Dokümanlardan seçilen anahtar sözcükler arasına görünümden bağımsız olarak tespit edilecek başlıklar da eklenebilir.

KAYNAKLAR

[1] **Lee, J. H., Kim, M.H. and Lee, Y. J.,** 1993. Information Retrieval Based on Conceptual Distance in IS-A Hierarchies, *Journal of Documentation,* **49**, 188-207

[2] **Resnik, P.,** 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy, *Proceedings of the 14th International Joint Conference on Artificial Intelligence,* Montreal, Canada, August 20-25, 448–453

[3] **Wu, Z. and Palmer, M**., 1994. Verb Semantics and Lexical Selection, *Proc. 32nd Annual Meetings of the Association for Computational Linguistics,* Las Cruces, New Mexico, 27 June - 1 July, 133–138.

[4] **Tan, P., Steinbach, M. and Kumar, V.**, 2006. Introduction to Data Mining, *Similarities Between Data Objects*, pp. 73-83, Addison-Wesley

[5] **Chu, C. H.,** 1989. Cluster Analysis in Manufacturing Cellular Formation, *International Journal of Management Science,* **17**, 289-295

[6] **Friburger, N. and Maurel, D.,** 2002. Textual Similarity Based on Proper Names, *SIGIR 2002 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Tampere, Finland, August 11-15, 155–167

[7] Cosine Similarity Tutorial, http://www.miislita.com/information-retrieval-tutorial/cosine-similarity-tutorial.html, 10.2006

[8] **Tatsumi, Y. and Asahi, T.,** 2005. Analyzing Web Page Headings Considering Various Presentation, *The 14th International World Wide Web Conference 2005*, Chiba, Japan, May 10-14

[9] **Gündüz-Öğüdücü, Ş. and Uyar, A. Ş.,** 2004. A Graph Based Clustering Method using a Hybrid Evolutionary Algorithm, *6th WSEAS International Conference on Applied Mathematics,* Corfu, Greece, August 17-18

[10] **Yuret, 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*, New York, USA, June 4-9, 328-334

[11] **Haldiki, M., Batistakis Y. and Vazirgiannis M.,** 2001. On Clustering Validation Techniques, *Journal of Intelligent Information Systems,* **17**, 107-145

[12] Balkanet Project Home Page, http://www.ceid.upatras.gr/Balkanet/, 06.2006

[13] S.Ü. Türkçe Sözcük Veritabanı, http://www.hlst.sabanciuniv.edu/TL/, 01.2007

[14] Zemberek Projesi Ana Sayfası, https://zemberek.dev.java.net/, 01.2007

[15] Java HTML Parser Home Page, http://htmlparser.sourceforge.net/, 08.2006

[16] Türk Dil Bilgisi Eğitim Sayfası, http://www.karatekin.net/2turkdili/, 11.2006

[17] SciLib Sayfası, http://www.ia.hiof.no/~kyvh/SciLib/docs/scilib.html, 03.2007

[18] **Rousseeuw, P.J.,** 1987. Journal of Computational Applications inMath, *Silhouettes: a graphical aid to the interpretation andvalidation of cluster analysis.*, **20**, 53–65

[19] **Kim, M. and Ramakrishna, R.S.,** 2005. Pattern Recognition Letters, *New indices for cluster validity assessment*, **26**, 2353-2363

ÖZGEÇMİŞ

Ankara doğumlu olan Bülent Yücesoy, ilköğrenimi İzmir Aliağa Petkim İlkokulunda okuyarak 1993 yılında tamamlamıştır. 1997 yılında ortaokulu İzmir Bergama Akif Ersezgin Anadolu Lisesinde tamamlamıştır. Lise öğrenimini İzmir Fen Lisesinde 2000 yılında tamamlamıştır. Öğrenim hayatına İstanbul Teknik Üniversitesi Bilgisayar Mühendisliği bölümü ile devam edip 2004 senesinde mezun olmuştur. Halen İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği programında yüksek lisans öğrenimine devam etmektedir. Bülent Yücesoy şu anda İTÜ Bilgi İşlem Daire Başkanlığında çalışmaktadır.