

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

WEB ORTAMINDA MEKANSAL VERİ KAZANIMI

DOKTORA TEZİ
Lütfiye KUŞAK

Anabilim Dalı : Jeodezi ve Fotogrametri Mühendisliği

Programı : Geomatik Mühendisliği

ARALIK 2010

WEB ORTAMINDA MEKANSAL VERİ KAZANIMI

DOKTORA TEZİ
Lütfiye KUŞAK
(501022305)

Tezin Enstitüye Verildiği Tarih : 14 Eylül 2010

Tezin Savunulduğu Tarih : 15 Aralık 2010

Tez Danışmanı : Prof. Dr. Doğan UÇAR (İTÜ)
Diğer Jüri Üyeleri : Prof. Dr. Ali GÜNEŞ (İAÜ)
Prof. Dr. Necla ULUĞTEKİN (İTÜ)
Prof. Dr. Nebiye MUSAOĞLU (İTÜ)
Yrd. Doç. Dr. Feza BUZLUCA (İTÜ)

ARALIK 2010

Aileme,

ÖNSÖZ

Bu çalışmanın ortaya çıkması esnasında yaşadığım uzun süreçte bana maddi manevi her türlü desteği veren her zaman hayallerimi destekleyen, üzerimde emekleri çok fazla olan babam Mehmet Kuşak'a annem Hilal Kuşak'a, benim için bu sıcak yaz ayında tercih merkezi görevini üstlenen ve daha yaptığı pek çok fedakârlıklar ve destekler için canım kardeşim, akıl arkadaşım Bahriye Kuşak'a, sonsuz sabrı ve yönlendirmeleri, destekleri için hocam Prof. Dr. Doğan Uçar'a, beni her konuda yönlendiren değerli hocalarım Prof. Dr. Necla Uluğtekin, Prof. Dr. Ali Güneş, Prof. Dr. Nebiye Musaoğlu ve Yrd. Doç. Dr. Feza Buzluca hocalarıma, hiçbir zaman desteğini esirgemeyen, her zora düştüğümde hep yanımda olan canım arkadaşım gelecekte çok iyi yerlerde kendisini göreceğimden emin olduğum Gökhan Erdemir'e, panik halinde artık bu son diyerek adım adım yaklaştığımız hayallerimin asıl kaynağı Z. Damla Uça Avcı'ya (doktorada iyi ki seninle tanışmışım), her zaman destekleri ile yanımda olan Aytül Filiz, Müge Aksu ve Şafak Yüksel'e ve Prof. Dr. Zafer Aslan'a sonsuz teşekkürlerimi sunmayı bir borç bilirim. Ayrıca tezim esnasında veri desteği sağlayan İstanbul Büyükşehir Belediyesi Harita Müdürlüğü çalışanlarına ve İşlem Şirketler grubu yetkililerine verdikleri önemli bilgilerden ötürü teşekkür ederim.

ARALIK 2010

Lütfiye KUŞAK
Yük. Jeodezi ve Fotogrametri Müh.

İÇİNDEKİLER

Sayfa

ÖNSÖZ.....	v
İÇİNDEKİLER.....	vii
KISALTMALAR.....	ix
ÇİZELGE LİSTESİ.....	xi
ŞEKİL LİSTESİ	xiii
ÖZET	xv
SUMMARY.....	xvii
1. GİRİŞ	1
2. CBS'LER İÇİN VERİ KAZANIMI VE GÜNCELLEME SORUNU	3
2.1 CBS lerde Kullanılan Veriler.....	3
2.2 İnternet ve Değişen Bilgi Paylaşım Yaklaşımı	3
2.2.1 İnternetin kullanım alanları.....	4
2.2.2 İnternetteki verilerden faydalanma yöntemleri	4
2.2.3 Web sayfalarından veri elde edilmesinde kısıtlamalar	5
2.3 Mekânsal Veri Kavramı.....	5
2.3.1 Mekânsal veriler ve internetteki yeri.....	6
2.3.2 Web sayfalarından elde edilen mekânsal verilerin sistemlere katkıları	10
3. WEB ORTAMINDA MEKÂNSAL VERİ KAZANIMI.....	13
3.1 Veri Madenciliği	13
3.1.1 Veri madenciliğinin uygulama alanları	15
3.1.2. Veri madenciliğinde kullanılan yöntemler	15
3.1.1.1 Sınıflandırma	15
3.1.1.2 Kümeleme	17
3.1.1.3 Birliktelik analizi	17
3.1.2 Mekânsal veri madenciliği.....	18
3.2 Web Madenciliği	18
3.2.1 Web madenciliğinin içeriği.....	20
3.2.1.1 Web kullanım madenciliği	22
3.2.1.2 Web yapı madenciliği	22
3.2.1.3 Web içerik madenciliği	22
3.2.2 Uygulama alanları	25
3.2.3 İşlem adımları	26
3.2.3.1 Veri kaynaklarının bulunması, ...	27
3.2.3.2 Verilerin önışlemden geçirilmesi, ...	32
3.2.3.3 Genelleştirme	35
3.2.3.4 Çözümleme	38
4. VERİLERİN KAZANIMI	41
4.1 İBB Mevcut Adres Bilgi Sistemi Yapısı	44
4.1.1 Mahalle	46
4.1.2 Obje Grupları	49

4.2 İşyeri Adres Verilerinin Kazanımı.....	51
4.3 İşyeri Kullanım Verilerinin Belirlenmesi.....	59
4.3.1 Eğitici web sayfaları.....	62
4.3.2 Eğitilen web sayfaları.....	63
4.4 Web sayfalarında önişleme.....	66
4.4.1 Ontolojik yaklaşım.....	71
4.4.2 Belediye sözlüğü.....	71
4.4.3 Genişletilmiş sözlük.....	74
4.5 Sınıflandırma İşlemine Hazırlık	77
4.5.1 Belediye veri seti kullanımı.....	77
4.5.1.1 .Öklid yöntemine göre çalışmanın irdelenmesi	78
4.5.1.2 Kosinüs yöntemine göre çalışmanın irdelenmesi	79
4.5.2 .Ontolojik olarak desteklenen belediye veri seti kullanımı	81
4.5.2.1 Öklid yöntemine göre çalışmanın irdelenmesi	81
4.5.2.2 Kosinüs yöntemine göre çalışmanın irdelenmesi	82
4.6 Belediye ve ontolojik olarak desteklenmiş verilerin kullanım türleri sonuçlarının karşılaştırılması.....	88
5. VERİLERİN KARŞILAŞTIRILMASI.....	91
5.1 Örnek uygulama.....	91
6. SONUÇLAR	103
6.1 Çalışmanın Katkıları	106
6.2 Olası Çalışmalar.....	107
KAYNAKLAR.....	109
EKLER (CD)	115
ÖZGEÇMİŞ.....	117

KISALTMALAR

CBS	: Coğrafi Bilgi Sistemleri
GML	: Geographic Markup Language
KBS	: Kent Bilgi Sistemi
NLP	: Natural Language Processing
XML	: Extended Markup Language
SPIRIT	: Spatially-Aware Information Retrieval on the Internet
GPS	: Global Positioning System
PDA	: Personal Digital Assistant
HTML	: HyperText Markup Language
K-NN	: K-Nearest Neighbours
CART	: Classification and Regression Trees
WWW	: World Wide Web
DB	: Database
İBB	: İstanbul Büyükşehir Belediyesi
MEB	: Milli Eğitim Bakanlığı
AFFWeb	: Address Find From Web
LSI	: Latent Semantic Indexing

ÇİZELGE LİSTESİ

	<u>Sayfa</u>
Çizelge 3.1: Web madenciliğindeki verilerin sınıflandırılması (Berendt ve diğ. 2003), (Ayaz, 2007).....	22
Çizelge 3.2: Karışıklık matrisi gösterimi.....	38
Çizelge 4.1: Şişli Adres Bilgi Sistemi kapı numarası sayısı.....	45
Çizelge 4.2: Şişli Adres Bilgi Sistemi mahalle ve adres sayıları.....	48
Çizelge 4.3: Şişli Adres Bilgi Sistemi kullanım türü ve adres sayısı.....	50
Çizelge 4.4: Web sayfalarında adres sunumu örnekleri.....	52
Çizelge 4.5: 8 kelime için toplam kelime frekanslarını gösteren tablo.....	69
Çizelge 4.6: Eğitici dokümanlarında erişilen kelime frekansı.....	69
Çizelge 4.7: Ontolojik olarak desteklemiş belediye sözlüğü yardımı ile erişilen kelimeler.....	70
Çizelge 4.8: Belediye obje gruplarındaki kelimelerinin dağılımı.....	72
Çizelge 4.9: Şişli çalışma bölgesinde seçilen 6 farklı grup ve bunlara ait kelime dağılımları.....	73
Çizelge 4.10: Obje grupları ve kelime sayıları karşılaştırılması.....	77
Çizelge 4.11: Ontolojik olarak desteklenmiş olan web sayfalarının kullanım türlerinin tespitinde farklı K değerleri için duyarlılık sonuçları çizelgesi.....	85
Çizelge 4.12: K değerleri ve belediye ve ontolojik destekli sözlüklere göre doğru olarak erişilen sayfa sayıları.....	88
Çizelge 4.13: Duyarlılık değerleri.....	89
Çizelge 4.14: K=4 için hata matrisi.....	90
Çizelge 4.15: K=4 için duyarlılık ve anma matrisi.....	90
Çizelge 5.1: Mavi Dans Bale ve Dans Merkezi frekans bilgileri.....	93
Çizelge 5.2: Mavi Dans Bale ve Eğitim Merkezi adres bilgileri ve Adres Bilgi Sistemi veritabanı interaktif karşılaştırma sonucu.....	95
Çizelge 5.3: Kelime frekansları tablosu.....	97
Çizelge 5.4: Elite Optik adres bilgileri ve Adres Bilgi Sistemi veritabanı interaktif karşılaştırma sonucu.....	97
Çizelge 5.5: Web sayfası adres verileri (Url 9).....	99

ŞEKİL LİSTESİ

	<u>Sayfa</u>
Şekil 2.1: Bing Maps örneğinde harita ve bağlantı mimarisi gösterimi.....	10
Şekil 3.1: Veri madenciliğinde işlem adımları (Özkan, 2008).....	14
Şekil 3.2: Web sayfalarında ürün tanıtımı ve kullanıcı potansiyelinin belirlenmesi.....	20
Şekil 3.3: Web madenciliği sınıflandırması ve kullanım alanları (Özakar ve Püskülcü, 2002, Daş ve diğ, 2006, Gürcan ve Köse 2008).....	21
Şekil 3.4: Web içerik madenciliği.....	24
Şekil 3.5: Web madenciliğinde kullanılan very kaynakları (Daş ve diğ. 2006, Baykal.....	28
Şekil 3.6: Yapılandırılmamış web sayfası.....	29
Şekil 3.7: Yarı yapılandırılmış web sayfası.....	29
Şekil 3.8: Yapılandırılmış web sitesi.....	30
Şekil 3.9: Vektör uzay modeli gösterimi (Url2).....	33
Şekil 3.10: KNN algoritmasında sınıflandırma işlemi gösterimi (Url3).....	37
Şekil 4.1: Şişli İlçesinin İstanbul haritası üzerinde gösterimi (Url4).....	41
Şekil 4.2: Şişli ilçesi 1955-2009 yılları nüfus dağılım grafiği.....	42
Şekil 4.3: Adres Yönetim Şefliğinin, ilçe bazında sayısal ortamda Adres Bilgi Sistemi genel veritabanı şeması (2009).....	44
Şekil 4.4: Adres bilgi yönetimi örneği.....	45
Şekil 4.5: Şişli ilçesi, 28 mahalleden oluşmaktadır.....	46
Şekil 4.6: Adres Yönetim Şefliği, İlçe Bazında Sayısal Ortamda Adres Bilgi Sistemi Mahalle ID ve MahNoID gösterim örneği (2009).....	47
Şekil 4.7: İBB standart obje grupları örneği kavramsal gösterimi.....	49
Şekil 4.8: Adres Yönetim Şefliği, ilçe bazında sayısal ortamda Adres Bilgi Sistemi obje grupları örneği (2009).....	50
Şekil 4.9: AFFWeb akış şeması.....	53
Şekil 4.10: AFFWeb Adres karşılaştırma yazılımı arayüzü.....	54
Şekil 4.11: Dosya yükleme örneği.....	55
Şekil 4.12: Web sayfalarından adres yükleme.....	55
Şekil 4.13: Adres bileşenleri.....	56
Şekil 4.14: Belediye verileri ve bilinmeyen adres karşılaştırma veritabanı.....	57
Şekil 4.15: İBB-Şişli İlçesi adres verileri.....	57
Şekil 4.16: db1 veritabanı görünümü.....	58
Şekil 4.17: Eşleştirme sonrası görünüm.....	58
Şekil 4.18: Veritabanında eşleştirme sonrası görünüm.....	59
Şekil 4.19: AFFWeb (Address Find From Web) ve web madenciliği kavramsal akış şeması.....	60
Şekil 4.20: Genel doküman sınıflandırma şeması.....	61
Şekil 4.21: Vikipedi eğitim sitesi örnek dokümanı.....	62
Şekil 4.22: MEB sitesi örnek dokümanı.....	63
Şekil 4.23: Sınıfları belli olmayan web sayfalarına örnekler.....	64

Şekil 4.24:	Haber takip sitesi (Url5).....	65
Şekil 4.25:	Kelime frekanslarının tespit edilmesi (Yıldız ve diğ. 2007).....	65
Şekil 4.26:	Eğitici web sayfalarının gösterimi.....	66
Şekil 4.27:	Belediye ve ontolojik olarak desteklenmiş belediye sözlüğü.....	67
Şekil 4.28:	Eğitim ile ilgili 2 numaralı web sayfasına ait frekans bilgileri.....	67
Şekil 4.29:	Bilinen web sayfalarında kelime frekansları (web sayfasında en fazla geçen kelimeden en az geçen kelimeye doğru sıralanmış).....	68
Şekil 4.30:	Belediyeye ait kategorilerde kelimelerin dağılımı.....	72
Şekil 4.31:	Belediyeye ait test verisi olarak seçilen obje gruplarında kelime dağılım grafiği.....	73
Şekil 4.32:	Ontolojik olarak desteklenmiş ve belediye obje gruplarından oluşturulmuş sözlüğün karşılaştırılması.....	77
Şekil 4.33:	24 adet bilinen web sayfası ve 1 adet bilinmeyen web sayfası Öklid hesaplaması.....	78
Şekil 4.34:	K-NN yöntemine göre mesafelerin tespit edilmesi işlemi.....	79
Şekil 4.35:	Belediye kelimeleri yardımı ile Kosinüs değerlerinin hesaplanması.....	80
Şekil 4.36:	Belediye kelimeleri yardımı ile oluşturulan Kosinüs sonuçları.....	81
Şekil 4.37:	Ontolojik olarak desteklenmiş kelimeler ile Öklid hesabı.....	81
Şekil 4.38:	Öklid yönteminde aradaki mesafeler.....	82
Şekil 4.39:	Ontolojik destekli kelimeler yardımı ile Kosinüs yöntemi uygulaması.....	83
Şekil 4.40:	Benzerlik değerleri sonuçları.....	83
Şekil 4.41:	K değerlerinin ve erişilen sayfaların karşılaştırılması.....	88
Şekil 4.42:	K değerlerinin ve duyarlılık yüzdelerinin karşılaştırılması.....	89
Şekil 5.1:	Mavi Dans Bale ve eğitim merkezi resmi web sayfası.....	91
Şekil 5.2:	Adres veritabanı web sayfası (Url6).....	92
Şekil 5.3:	Ontoloji destekli belediye sözlüğü Kosinüs benzerlik sonuçları.....	93
Şekil 5.4:	Mavi Dans Bale ve Eğitim Merkezi AFFWeb adres eşleşmesi.....	94
Şekil 5.5:	Verilen adresin bileşenlerine ayrılması.....	94
Şekil 5.6:	İBB verileri karşılaştırma sonuçları.....	95
Şekil 5.7:	Tubiba web sayfası (Url 7).....	96
Şekil 5.8:	Firma rehberi web sayfası (Url8).....	96
Şekil 5.9:	K=4 için Elit1 ve Elit2 web sayfalarının kullanım türlerinin tespiti..	97
Şekil 5.10:	Web sayfası örneği.....	98
Şekil 5.11:	K=4 için Kosinüs yöntemi değerlendirmesi.....	98

WEB ORTAMINDA MEKANSAL VERİ KAZANIMI

ÖZET

İnternetin günlük hayatın neredeyse ayrılmaz bir parçası olması ile birlikte, sunulan bilginin kullanımı ve paylaşımına yönelik çalışmalar da hızla değişim göstermiştir. Bu değişim paralelinde var olan bilgi açığı ve kaynaklardan yararlanma rekabeti de yeniden şekillenmeye başlamıştır.

Maliyeti oldukça pahalı olan veri setlerini tekrar tekrar oluşturmak yerine, mevcut bulunan kaynakları çok daha iyi yönetmeye odaklanmış çalışmalar ön plana çıkmaya başlamıştır. Yerel yönetimler, özel sektör, eğitim sektörü ve bireysel kullanıcı bilginin paylaşımı ve kullanımına yönelik değişimler ve gelişmelerden direkt etkilenmektedir.

İnternet kullanımı ile gelişen ticaret ortamı, küçük ve büyük ölçekteki birçok şirkete müşterilerine ürünlerini tanıtabilmek ve hatta sipariş almak üzere iletişim bilgileri çerçevesinde şirket adreslerini web sayfalarında yayınlamaya başlamışlardır. Gerçekleştirilen bu araçlar sayesinde hem arama motoru firmaları destek görmekte hem de piyasa analizi mümkün hale gelmektedir. Özellikle arama motorları teknolojilerini daha etkin sonuçlar vermesi amacıyla mekânsal bilgilerin edinimi çalışmaları giderek hız kazanmaya başlamıştır. Mekânsal verilerin elde edilerek diğer başka çalışmalara destekleyici altlık olarak kullanılması düşüncesi diğer bir araştırma konusu olmuştur. Günümüzde sağlıklı ve sürdürülebilir kentsel gelişim için, teknolojinin sunduğu imkânlardan yararlanmanın gerekliliği gerek merkezi, gerekse yerel yönetimler tarafından kaçınılmaz hale gelmiştir.

Bu doğrultudaki Genel Enformasyon Teknolojisi araçları, mekânsal kullanım ihtiyaçları doğrultusunda da geliştirilmeye çalışılmaktadır.

Yapılan çalışmada özellikle gerek belediyelerde kullanılan CBS'ler, KBS'lerde kullanılan Adres Bilgi Sistemlerinin bilgilerin zenginleştirmesinin yanı sıra, mekânsal arama motorları için gerekli olan mekânsal veri setlerinin desteklenmesi hedeflenmiştir. Bu amaçla web sayfalarından elde edilen mekânsal veri setleri, yapılan tez çalışmasında çok amaçlı incelemeye alınmıştır.

Web sayfalarındaki mekânsal içerikli, veriler mekânsal referans bilgisi (örneğin koordinat) içermeyen veri özelliğindedir. Bu verilerin başında adreslerin, telefon numaralarının, posta alan kodlarının ve web sayfası içerik bilgilerinin geldiği açıktır. Bütün bu veriler yukarıda adı geçen sistemler için geliştirici veri setleri olarak kullanılabilirler. Bu tip verilere erişiminin sağlanabilmesi için, öncelikle web sayfaları üzerinden derlenmesi ve mekânsal verilerin gazetterler yardımı ile tespit edilmesi gerekmektedir. Elde edilen bu bilgilerin doğruluğu, İstanbul Büyükşehir Belediyesinin adres bilgi sistemi verileri ile karşılaştırılacaktır. Ayrıca web sayfalarından kazanılan adreslerin fonksiyonlarının da web içerik bilgisi sayesinde tespiti yapılacaktır. Tez çerçevesinde bu amaçla web madenciliği tekniklerinden faydalanılacaktır.

Web sayfalarından elde edilen mekânsal verinin gerek coğrafi arama motorları gerekse CBS'ler için kullanılabilecek olması güncellik, maliyet ve farklı sistemlere destek problemlerin çözümüne ciddi katkı sağlayacağı değerlendirilmektedir.

SPATIAL DATA GAIN FROM WEB

SUMMARY

Since internet is becoming vital for daily life, the modification of utilization and sharing tools of provided data has increased. In parallel with this development, the existent data/information/knowledge need and the competition on getting benefit from the resources has been reshaping.

Instead of acquiring and composing the datasets repeatedly, which is not cost effective; the methods which are based on administering the existing resources have drawn ahead. The local authorities, the private companies, sector of education and personal users are directly influenced by the alternations and improvement in data sharing and usage capabilities.

The emerging world of business has profits from the wide usage of internet technology. Most large scale companies and small establishments publish their addresses on websites to get orders of the products which they also introduce online. Activating these tools support the search engine companies and enable the market analyses. The uptrend in performance of search engines motivates the progress in extracting spatial data tools. Based on this, the idea of organizing and directing the extracted spatial data as fundamentals for other studies have become a new research field. Presently, the necessity of profiting from the opportunities of new technology became fatal for local and central authorities to maintain sustainable urban development.

General information technology tools are being implemented on purpose to meet the requirements of spatial related needs.

In this study, the aim is to enrich the collectable information in the geographical information systems and urban information systems that are used in municipalities. Besides, supporting the spatial datasets which are requisites for spatial search engines is intended. For this purpose, the spatial data which are extracted from web pages are analysed for multi-dimensional investigation.

The data on web pages have spatial content; however it may not include spatial reference as coordinates. It is obvious that this data sentences have addresses, phone numbers, and postal codes with web page contents. All these data can be used as developer datasets for the systems mentioned previously. To supply access to these data, first it have to be gathered and compiled over websites and spatial data have to be recognised by ontologic based dictionaries. The correction of attained data will be compared by the address information system of Istanbul Metropolitan Municipality. In addition to this, the building functions will be determined by web contents. In this context, the advantages of web mining technics will be taken.

The uses of extracting/mining of spatial data from web pages on both geographic search engines and geographical information systems is being evaluated as a cost effective up-to-date critical supplement for the solution of compatibility to various systems.

1. GİRİŞ

Dünya yüzeyinde mekânsal referanslı objelere ait konum bilgileri (genellikle koordinat bilgisi), öznitelik bilgileri ve topolojik ilişkiler vektör, raster ve tablolar halinde ifade edilmektedir.

Genel olarak mekânsal verilerin elde edilmelerinde, geçmişten günümüze üretilmiş olan analog ve sayısal haritalar, sayısallaştırma, GPS ölçmeleri, fotogrametrik veri üretim, uzaktan algılama ve üç boyutlu lazer tarama teknolojileri yoğun olarak kullanılmaktadır (Çelik ve diğ., 2005). Literatürlerde konum bilgilerinin kazanılması üzerinde sıklıkla durulmaktadır.

Mekânsal bir varlığı tanımlayan öznitelik verileri ise hızla değişim gösterebilmektedir. Bilgi sistemlerinin güncel ve doğru olarak sunulması için öznitelik verileri önem taşımaktadır. Fakat anılan verilerin klasik yöntemlerle toplanmasını maliyetinin oldukça yüksek ve zaman alıcı olduğu bir gerçektir.

İnternette yaşanan gelişmelere paralel olarak mekânsal verilerin web ortamından elde edilebilmesine yönelik yapılan çalışmalar da hız kazanmıştır. Veri kazanım kaynaklarından birisi olarak web sayfalarındaki yoğun mekânsal verilerden faydalanılması seçenek olarak araştırılmaya başlanmıştır. Mekânsal veri setlerinin web ortamından kazanılması ancak 1900'li yılların son dönemlerinde ilgi odağı haline gelmeye başlamıştır. Mekânsal arama motorlarını ve Coğrafi Bilgi Sistemlerini desteklemek amacı ile güncel mekânsal veri setlerine duyulan ihtiyaç giderek artmıştır. Fakat özellikle hangi tip verilerin mekânsal veri kapsamına gireceği, bu tip verilerin web ortamında nasıl elde edileceği ve elde edilen verilerin nasıl saklanılacağına yönelik çalışmalar hala devam etmektedir.

Bu tez çerçevesinde yapılan çalışmada özellikle gerek belediyelerde kullanılan CBS için oluşturulan bilgilerin zenginleştirmesinin yanı sıra, coğrafi arama motorları için gerekli olan mevcut mekânsal veri setlerinin desteklenmesi hedeflenmiştir. Bu amaçla web sayfalarından elde edilen mekânsal veri setleri yapılan tez çalışmasında çok amaçlı analiz edilmişlerdir.

Web sayfalarından işyeri adreslerinin kazanımı ve anılan işyerlerine ait kullanım fonksiyonlarının tespit edilmesi olmak üzere iki temel kısımdan oluşan çalışma kapsamında kent yönetiminin etkinleştirilmesinde önemli rol oynayan Kent Bilgi Sistemleri ile beraber kullanılan numarataj ve Adres Bilgi Sistemlerinde yer alan adres bilgilerinin web üzerinden kazanımının olanakları araştırılmaktadır. Ayrıca çalışmada web üzerinden kazanılan verilerin mevcut sistemlerdeki kayıtlı verileri güncelleştirilmesinde ve iyileştirilmesinde sağlayabileceği yararlar tartışılmaktadır. Bu çerçevede elde edilen bilgilerin doğruluğu, İstanbul Büyükşehir Belediyesinin bire bir arazide önemli maddi bir masraf yaparak oluşturduğu ve bu bakımdan daha doğru olduğu varsayılan Adres Bilgi Sistemi verileri ile karşılaştırılmaktadır. Mekânsal referanslı objeler ile alakalı özellikle öznitelik verilerinin kazanımında, alışlagelmiş yöntemlere ek olarak, web sayfaları da veri kaynağı olarak rol almaktadırlar. Çalışmanın diğer bir parçası ise web sayfalarında bulunan işyerlerinin kullanım fonksiyonlarının tespitinin yapılmasıdır. Bu amaçla web madenciliği tekniklerinden ne ölçüde yararlanılabileceği, çalışmanın temel araştırma noktalarından birini oluşturduğu söylenebilir.

Çalışmada 2. bölümde genel olarak CBS'ler için veri kazanımı ve güncelleme sorunu üzerinde durulmaktadır. Mekânsal veri kavramı ve internetteki yeri tartışılmaktadır. Bu amaçla web ortamında bulunan verilerden ne ölçüde faydalanılabileceği araştırılmaktadır. 3. bölümde web ortamında mekânsal verinin kazanımı ve çalışma için kullanılan web madenciliği gibi temel yöntemler incelenmektedir 4. bölümde benzerlik ve uzaklık hesaplaması yöntemleri karşılaştırılarak, elde edilen sonuçların, sınıflandırılması amacıyla, K-NN yöntemi ile çalışma için uygunluğu test edilmektedir. Seçilen iki farklı adres ve adreslerin bulunduğu 4 farklı web sayfası ontolojik olarak desteklenen belediye sözlüğü, kosinüs yöntemi ve K-NN sınıflandırması için seçilen K=4 parametresi ile test edildiği 5. bölümde kazanılan sonuçlar değerlendirilmektedir. Genel değerlendirmeler, bu tezden yola çıkılarak çalışmanın katkıları ve olası çalışmalara yönelik öneriler tezin 6. ve son bölümünü oluşturmaktadır.

2. CBS'LER İÇİN VERİ KAZANIMI VE GÜNCELLEME SORUNU

Günümüzde CBS lerin sürdürülebilirliğinin sağlanabilmesi için güncel veri setlerine duyulan ihtiyaç artmaktadır. CBS oluşturulurken elde edilen özellikle öznitelik verilerinin sistemin başarısı için devamlı olarak güncelliğinin kontrol edilmesi gerekmektedir. Problemin çözümünde verilerin güncelliğinin sağlanabilmesi amacıyla uygun olarak bir model önerilmektedir. Bu model işyeri adres verilerinin kazanılması ve işyeri kullanım fonksiyonları elde edilmesinden oluşmaktadır. Amaçların yerine getirilmesi doğrultusunda model modifiye edilerek ilgili işyerlerine ait diğer kitlesel veri olarak adlandırabileceğimiz öznitelik verilerinin kazanımı mümkün olacaktır.

2.1 CBS lerde Kullanılan Veriler

CSS'ler için günümüzde öznitelik verileri çalışmanın amacı kapsamında sıcaklık, maaş, yaş, toprak türü vb. pekçok veri olabilir. Verilerin güncel olması daha sonra yapılacak olan analiz ve sorgulamalar için daha iyi sonuçlar elde edilmesini sağlayacaktır. Bu bölümde İnternetin bu amaçlar için kullanılması genelde nasıl olmaktadır, mekansal verinin internetteki yeri ve önemi, internetten veri kazanımı esnasında getirilen zorunlu kısıtlamalar ve yapılan çalışmalar üzerinde durulmaktadır.

2.2 İnternet ve Değişen Bilgi Paylaşım Yaklaşımı

İnternetin günlük hayatın nerdeyse ayrılmaz bir parçası olması ile birlikte, sunulan bilginin kullanımı ve paylaşımına yönelik çalışmalar da hızla değişim göstermiştir. Bu değişim paralelinde var olan acil güncel bilgi ihtiyacını mevcut açık kaynaklardan yararlanarak kapatma isteği de yeniden şekillenmeye başlamıştır.

Klasik yöntemlerle oluşturulması maliyeti oldukça pahalı olan verileri güncelliğini yitirdikçe tekrar tekrar toplamak yerine, mevcut bulunan kaynakları çok daha iyi yönetmeye odaklanmış çalışmalar ön plana çıkmaya başlamıştır. Yerel yönetimler,

özel sektör, eğitim sektörü ve bireysel kullanıcı bilginin paylaşımı ve kullanımına yönelik değişimler ve gelişmelerden etkilenmektedir.

2.2.1 İnternetin kullanım alanları

İnternet kullanımı ile gelişen ticaret ortamında, küçük ve büyük ölçekteki şirketler, müşterilerine ürünlerini tanıtabilmek ve hatta sipariş almak üzere, iletişim bilgileri çerçevesinde, şirket adresleri, telefon numaraları gibi mekânsal özellik taşıyabilecek bilgileri web sayfalarında yayınlamaya başlamışlardır. Piyasa analizi yapmak üzere elde edilebilecek bilgilerden müşteri potansiyelinin belirlenmesi ve bütün bunların kazanılacak mekânsal çerçevede sunulması düşüncesi araştırmacıların ilgi odağı haline gelmiştir (Buyukokten ve diğ., 1999). Bu düşünce çerçevesinde gerçekleştirilen sistemler sayesinde hem arama motoru firmaları destek görmekte hem de piyasa analizi mümkün hale gelmektedir.

2.2.2 İnternetteki verilerden faydalanma yöntemleri

Mevcut bulunan açık kaynaklardan en iyi derecede yararlanılabilmesi için erişilmek istenilen bilgilerin kayıt altına alınarak, Elizio ve diğ. (2008) indekslenmesi çalışmaları yapılmaktadır. İndeksleme işlemi sayesinde anılan bilgilere erişim aşamasında kolaylık sağlanmaktadır. İnternetin büyümesi ile sağlık, ticaret, spor, alışveriş ve mekânsal bilgilerin de içinde bulunduğu geniş bir yelpazede farklı konularda zengin veri kaynaklarına erişim sağlanmış Borges ve diğ.(2003), Borges ve diğ. (2007), tarafından bilginin kullanımına yönelik çalışma alanları genişletilmiştir.

İnternet ve Değişen Bilgi Paylaşım Yaklaşımı

- İnternetin gelişmesi,
- Bilgi açığı problemi,
- Maliyet,
- Veri setlerinin güncel olarak yeniden oluşturulmasında zaman kaybı,
- İnternette var olan verilerin atıl olarak durması, fakat herkesin yararlanmasına açık ve büyük bir potansiyel olarak bulunması nedeniyle iştah kabartması beklentileri ve rekabet alanlarını da değiştirmiştir.

Değişen beklentiler sonucunda internette ticaret ortamı da gelişmiştir.

- Müşteri potansiyelinin belirlenmesi
- Piyasa analizi
- Ürün tanıtımı
- Siparişlerin ulaştırılması vb. konularda ne kadar değişim içerisinde olduğumuzun sadece küçük bir örneği olarak gözükmektedir.

Amazon.com da bir fotoğraf makinesi için araştırma yapıldığında aranılan fotoğraf makinesi ile ilgili sadece genel bilgilere ulaşmakla kalmayıp, ürüne ait ekstra eklenti bilgileri, benzer model örnekleri, fiyat karşılaştırmaları ve daha önceden bu ürünü alan kişilerin almış oldukları diğer ürünler hakkındaki bilgilere de ulaşmak şu andaki teknoloji sayesinde mümkün olmaktadır.

2.2.3 Web sayfalarından veri elde edilmesinde kısıtlamalar

Web sayfalarında verilerin kazanımı işlemlerinde web sayfalarını hazırlayan kurum ya da kişiler bilgilerin, resimlerin, dosyaların tamamının alınmasını engellemek amacı ile bir takım sınırlamalar getirmektedirler. Arama motorlarının indeksleme robotları bir web sitesine ulaştıklarında ilk olarak kök dizinde bulunan robots.txt dosyasını ararlar. Bu dosya robotlara hangi sayfaların indekslenip hangilerinin indekslenmeyeceğini anlatmak için yerleştirilir.

Kısıtlamalar dikkate alınarak yapılan çalışmalar yardımı ile günümüzde web sayfalarından kazanılan veriler yoğun olarak e-ticaret, bankacılık, güvenlik gibi çeşitli sektörlerde kullanılmaktadır.

2.3 Mekânsal Veri Kavramı

Dünya yüzeyinde mekansal referanslı objelere ait konum bilgileri (genellikle koordinat bilgisi), öznitelik bilgileri ve topolojik ilişkiler vektör, raster ve tablolar halinde ifade edilerek genel olarak “ mekânsal veriler” kavramı altında sunulmaktadır.

Genel olarak mekânsal verilerin elde edilmelerinde, geçmişten günümüze üretilmiş olan analog ve sayısal haritalar, sayısallaştırma, GPS ölçmeleri, fotogrametrik veri

üretim, uzaktan algılama ve üç boyutlu lazer tarama teknolojileri yoğun olarak kullanılmaktadır (Çelik ve diğ., 2005).

Diğer taraftan günümüzde hızla gelişme gösteren internette, web sayfalarından mekânsal verilerin kazanılarak kullanılması da klasik teknolojileri destekleyici yeni bir seçenek olarak düşünülmelidir.

2.3.1 Mekânsal veriler ve internetteki yeri

Dünyada özellikle e-ticaret uygulamalarında kullanılan çalışmaların mekânsal özellikteki çalışmaları da destekleyebileceği fikri bu çalışmanın da çıkış noktalarından biridir.

Yukarıdaki açıklamalar ışığında incelendiği zaman

- Yerel yönetimler,
- Özel sektör,
- Eğitim sektörü
- Bireysel kullanıcı bilginin paylaşımı ve kullanımına yönelik değişimler ve gelişmelerden direkt etkilenmektedir. Web sayfalarındaki mekânsal bilgilerin önemi Markowetz'in yapmış olduğu çalışma ile çok daha fazla ortaya çıkmaktadır.

Markowetz (2004), internetin tarihsel gelişimini genel olarak 3 farklı dönemde inceleyerek, Amazon gibi birkaç ticari kuruluşun hizmetlerini çevrimiçi olarak sunmasından günümüzde pizzacılar gibi mahallelerde yerleşmiş olan en küçük ticari şirketlerin dahi satışlarını internet hizmetlerinden faydalanarak sunduklarını belirtmektedir.

Web sayfalarındaki mekânsal içerikli veriler, mekânsal referans bilgisi (örneğin koordinat) içermeyen veri özelliğindedir. Bu verilerin başında adreslerin, telefon numaralarının, posta alan kodlarının ve web sayfası içerik bilgilerinin geldiği açıktır (Leidner ve diğ.,2003). Bütün bu veriler adı geçen sistemler için geliştirici, zenginleştirici veri setleri olarak kullanılabilirler.

Web sayfalarındaki mekânsal veriler, her zaman koordinat bilgileri ile birlikte (mekânsal referanslı objelerin geometrileri ile birlikte) sunulmayabilirler. Örneğin

adres bilgileri, telefon numaraları, posta ve alan kodları web sayfalarının içerdği mekânsal bilgiler arasında yer almaktadır (Leidner ve diğ., 2003).

Mekânsal veri toplama yöntemlerine destekleyici veya çeşitli alanlarda kullanılmak üzere aşağıda çalışmalarda sunulan web sayfalarından, mekânsal veri elde etme yöntemleri ve kullanım alanları 1999 yılından itibaren araştırılmaya başlayan konular arasında önemli yer tutmaktadır.

Mekânsal verilere yönelik yapılan çalışmaları, genel olarak, farklı kullanım alanlarında incelemek mümkündür.

- Arama motoru teknolojilerinde destekleyici olarak,
- Çeşitli alanlarda kullanılmak üzere destekleyici olarak,
- Sadece adres verilerini elde etmek,
- Analiz işlemleri için farklı kullanım alanları bulunmaktadır.

Arama motorları teknolojilerinde, sistemin daha etkin sonuçlar vermesi amacıyla 1990'lı yılların sonundan itibaren mekânsal bilgilerin web sayfalarından kazanılması çalışmaları giderek hız kazanmaya başlamıştır.

Google, Yahoo vb. gibi geleneksel arama motorlarında verilerin depolanması, indekslenmesi ve sorgulanması aşamaları mekânsal arama motorlarında da aynı aşamalarla ifade edilmektedir. Mekânsal arama motorlarında web sayfalarından kazanılan verileri indeksleme çalışmaları Ding ve diğ, (2000), Hill (2000), McCurley (2001), Waldinger ve diğ.(2003), Markowetz (2004), Heuer ve Dupke (2007), tarafından ele alınmıştır. 2005 yılında Markowetz ve diğ, (2005) yaptığı sistemde yukarıdaki özelliklere ek olarak ihtiyaç duyulması halinde kazanılan ve indekslenen veriler, koordinat verileri ile birleştirilerek zenginleştirilmiştir.

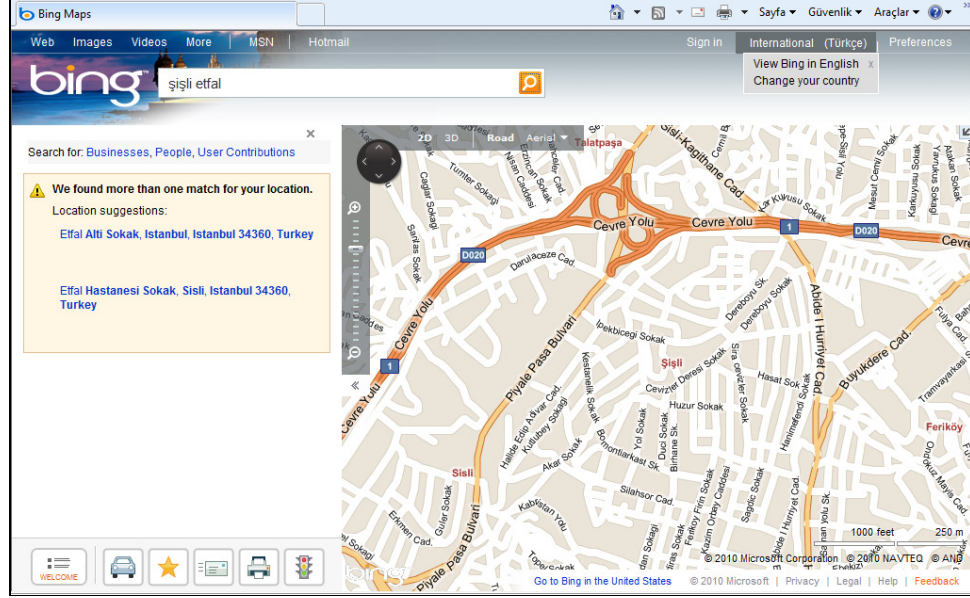
Mekânsal arama motorlarının içerik bakımından zenginleştirilmesi amacı ile Markowetz (2004) "xxxxx.de" uzantılı olan Almanya'ya ait web sayfalarını incelemiştir. Alan kodları, şehir isimleri ve posta kodları Markowetz tarafından detaylı biçimde incelenmiştir. Markowetz yaptığı bu çalışma ile internet kaynaklarında mevcut bulunan mekânsal özelliklerden faydalanma yolunu seçmiştir.

2002 yılından itibaren geliştirilmeye başlayan SPIRIT projesi kapsamında ise, web sayfalarından mekânsal veriler elde edilmekte ve elde edilen bu veriler gerekli düzenlemeler yapılarak daha sonra kullanılmak üzere SPIRIT coğrafi arama

motorunda saklanmaktadır (Jones ve diğ.,2004), (Clough, 2005), (Gelernter ve Lesk, 2008).

Tezuka ve diğ. (2006) da yapmış oldukları çalışmada özellikle Web ve CBS arasındaki sıkı ilişkiye dikkat çekmektedirler. Buna göre Web sayfalarının içeriğinde bulunan yerleşim bilgileri CBS ile ilişkilendirildiği zaman kullanıcı bu bilgiden faydalanabilecektir. Bu çalışmaya ek olarak, web sayfalarının içeriğinde bulunan mekânsal bilgilerin web madenciliği gibi işlemlerden geçirilerek çeşitli uygulamalarda kullanılmak üzere elde edilebilmesi üzerinde durulmaktadır. Web sayfalarından mekânsal veriyi elde etme çalışmalarında basit metin eşleştirme, morfolojik ve yapısal analiz gibi metnin doğal dil işleme yöntemleri kullanılarak işlenmesi ve ontoloji tabanlı edinim (ontology-based extraction) örneğin sözlük verisi kullanılması yöntemlerini sunmuşlar ve özellikle doğal dil işleme yöntemlerinden morfolojik analiz yöntemini seçmişlerdir. Blog verilerinden yola çıkılarak yapılan işlemler sonucunda doğruluk ve duyarlılık sonuçlarının uygunluğu web sayfalarından mekânsal bilginin çıkarılabileceği sonucunu ortaya koymaktadır. Çalışmada daha önceden bloglardan elde edilmiş önemli mekânların sürücünün yol güzergâhları üzerinde sunulmasını sağlayabilmek amacı ile de bir web araba radyosu geliştirmişlerdir.

Yukarıda adı geçen çalışmalar akademik çevrede veya araştırma projeleri olarak hayata geçirilen mekânsal arama motorları kapsamında yer almaktadır. Hızla büyüyen internetten ticari anlamda da mekânsal arama motorlarının desteklenmesi amacı ile Google, Yahoo, Ask Jeeves, MapQuest, Excite.com vb. ticari arama motorları da mekânsal platformda yer almaktadırlar. Bu tip arama motorları harita ve web sayfalarını birleştirerek (Şekil 2.1) kendi altyapılarında bulunan veritabanlarından faydalanmaktadırlar.



Şekil 2. 1: Bing Maps örneğinde harita ve bağlantı mimarisi gösterimi.

Mekânsal verilerin elde edilerek başka çalışmalara destekleyici altlık olarak kullanılması düşüncesi Dalli (2006) için de diğer bir araştırma konusu olmuştur. Günümüzde sağlıklı ve sürdürülebilir kentsel gelişim için, teknolojinin sunduğu imkânlardan yararlanmanın gerekliliği gerek merkezi, gerekse yerel yönetimler tarafından kaçınılmaz hale gelmiştir (Boll ve diğ., 2008). Bu doğrultudaki Genel Enformasyon Teknolojisi araçları, mekânsal kullanım ihtiyaçları doğrultusunda da geliştirilmeye çalışılmaktadır. Yapılan çalışmalarda web sayfalarının içerdiği mekânsal referanslı olabilecek veriler üzerinde durulmaktadır. Özellikle son yıllarda cep telefonları, GPS, PDA ve PALM gibi taşınabilir kişisel iletişim ve hesaplama araçlarında mekânsal referanslı veriler kullanılmaktadır. Zhang ve Gruwad yaptıkları bu çalışmada GML (Geographic Markup Language) tabanlı bir arama motoru geliştirmişlerdir.

Mekânsal referanslı verilerin herhangi bir sistemi desteklemek üzere kazanıldığı durumların yanı sıra, bazı durumlarda sadece adres verilerinin elde edildiği çalışmalar da bulunmaktadır. Özellikle adreslerin hiyerarşisine dikkat ederek web sayfalarından adres bilgilerinin kazanıldığı Ahlers ve Boll (2008a, b), gerekli düzenlemelerin yapıldığı çalışmalar dikkat çekmektedir. Web sayfalarından mekânsal verilerin elde edilmesi üzerinde yoğunlaşan Morimoto ve diğ. (2003) uzay vektör modeli ile verileri sunmuşlar ve kümeleme yöntemleri ile verileri gruplama yoluna gitmişlerdir.

Borges ve diğ. (2007) tarihinde yapmış oldukları çalışmada web sayfalarında bulunan adres verilerini elde etmeyi hedefleyerek 4 milyon web sayfasını incelemişlerdir. Daha önceki yapılmış olan çalışmaları inceleyen ekip, web sayfalarında mekânsal verilerin bulunması üzerinde çalışmalar yapıldığını görmüşler ve daha önceki yaptıkları çalışmada kullanıcı destekli bir örnek ortaya koyarak web sayfalarından elde etmiş oldukları verileri XML formatında saklayarak gerektiği zaman kent bilgi sistemlerinde kullanılmaya hazır hale getirmişlerdir.

Kimler (2004) yılında yapılandırılmamış olan metinlerden mekânsal referansların tanınması ve görselleştirilmesi üzerinde çalışmıştır. İnternetin son 10 yıldaki baş döndürücü büyümesi ile birlikte kullanıcıların doğrudan erişemediği, ama internette bulunan çok miktardaki veriye erişim araştırma konusu olmaktadır. Bu çalışma ile yer adları sözlüğü (gazetteer) ve genişletilmiş veri setleri ile yapılandırılmamış metinlerden mekânsal verilerin tanınması çalışmaları desteklenmektedir.

Göbel ve diğ. (2006) tarafından yürütülen diğer bir çalışmada ise, mekânsal referanslı veriler aktif olarak kullanılmaktadır. Bu çalışmada temel alınan elektronik posta bilgileri sayesinde güvenlik ile ilgili bir tehdit durumunun ortaya çıkıp çıkmayacağı tespit edilmektedir.

Analiz amacı ile de web sayfalarından mekânsal verilerin elde edilmesi çalışmalarına başvurulmaktadır. Bölgesel ve ülkesel yapılan çalışmaların aksine Dalli (2006) tarihli çalışmasında blog ve haber sitelerine ait değişken mekânsal verilerin analizini bütün dünya çapında değerlendirmektedir. Elde edilen sonuçlar yardımı ile dünyada medyaya olan ilginin analizi hedeflenmektedir.

2.3.2 Web sayfalarından elde edilen mekânsal verilerin sistemlere katkıları

Yukarıda sunulan çalışmaların ışığında web sayfalarından elde edilen mekânsal verinin gerek coğrafi arama motorları gerekse CBS ve KBS'ler için kullanılabilecek olması, CBS'lerde veri güncelliği, veri toplama maliyeti ve farklı sitemlere destek olarak 3 farklı alana katkıda bulunabileceği düşünülmektedir.

Güncellik: Günümüzde belediyelerin yapmış oldukları adres numarataj sistemlerini büyük maliyeti olan hizmet alımı biçiminde çözümledikleri görülmektedir. Ayrıca bu tip bilgilerin elde edilme aşamasının mevsimsel kısıtlılıkların olumsuz etkisinde kaldığı ve dolayısıyla imar planlarının uygulama zamanı içerisinde güncellikten bir hayli uzak kalabildikleri görülmektedir. Özellikle kent planlarının onaylanması ve

işlemlerin başlaması için geçen süreç 2-5 sene arasında olabilmektedir. Bina kullanım türlerinin daha hızlı değişiklik gösterdiği bölgelerde bu durumun kent planlamacılar ve diğer karar vericiler için ciddi bir sorun yarattığı açıktır.

Maliyet: CBS lisanslı yazılımlarının çoğunun dış kaynaklı olması maliyetleri artırmaktadır. CBS uygulamalarında kullanılan sayısal verilerin hazırlanması ve ister yersel ister başka yöntemlerle olsun elde edilmesi oldukça maliyetli hizmetlerdir. Teknolojinin hızla değişmesi ve gerekli olan donanım ve yazılımın güncellenme maliyeti de belediyelere devamlı olarak bu alanda kaynak bulma zorunluluğu getirmektedir.

Farklı Sistemlere Destek: Elde edilecek veri setlerinin birden fazla sistemde (CBS, Coğrafi Arama Motorları, KBS) kullanılabilecek şekilde genişletilebilecek olması önemli katkılar sağlayacaktır. Örneğin; Herhangi bir kuruluş için (ihtiyaç duyduğu veriyi ilk elden), verileri her defasında en baştan toplamak yerine, mümkün olduğunca, hâlihazırdaki verilerinden ve o veriye sahip olan başka kaynaklardan elde edebilecektir,

Verileri belirli standartlara dönüştürmek, çok daha ekonomik, hızlı ve verimli yöntemleri destekleyecektir.

Bütün bu problemlerin çözümünde ise, mevcut veri setlerini destekleyici anlamda, web sayfalarından elde edilecek olan mekânsal referanslı objelerle ilişkilendirilebilecek verilerin çok düşük bir maliyet içinde kalınarak kullanılabileceği açıktır.

3. WEB ORTAMINDA MEKÂNSAL VERİ KAZANIMI

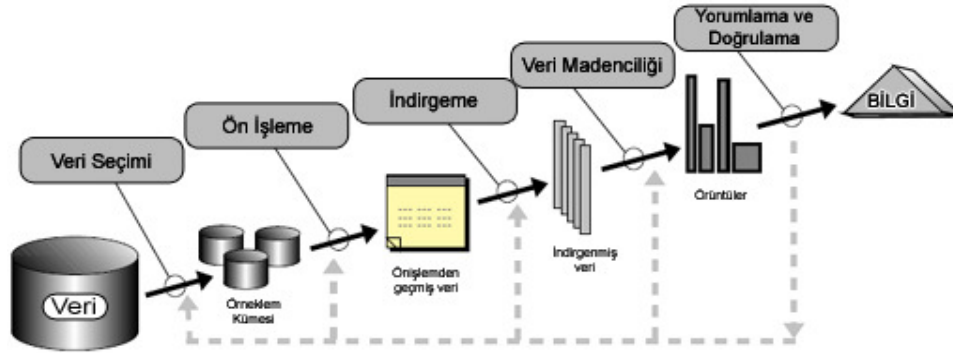
Haber vb. bilgilerin sunulmasında faydalanan portallar, kullanıcının bilgiye erişimini kolaylaştırmak amacı ile diğer web sitelerinden edinilen web içerikleri konularına göre sınıflandırılmaktadır. Spor, ekonomi, sağlık, kültür gibi temel konu başlıkları sayesinde bilgiye erişim kolaylaşmaktadır. Bu tür çalışmalarda sınıflandırma işlemleri için web madenciliği yöntemleri yaygın olarak kullanılmaktadır. Tez çalışması kapsamında web ortamında haber portalları ve e-ticaret uygulamaları temel alınarak web sayfalarının içerdiği mekânsal verilerin kullanım türlerini belirlemek için web madenciliği yöntemi kullanılmıştır. Web sayfalarından işyeri adres bilgilerine ait kullanım türleri veya diğer bir değişle kullanım fonksiyonları tespit edilmiştir. Genel çalışmalardan farklı olarak elde edilen bilgi sayesinde, yerel yönetimlerin ve mekânsal arama motorlarının ihtiyacına destek verecek olan çalışma kapsamında kullanılan yöntem ve formülleri açıklamak amacı ile öncelikle web madenciliğinin çıkış noktası olan veri madenciliği hakkında bilgi verilmektedir. 3. bölümde ise web madenciliğine daha ayrıntılı olarak değinilmektedir.

3.1 Veri Madenciliği

Çok büyük miktarda anlamsız veri yığınlarından veya işlenmemiş verilerden üstü kapalı, çok net olmayan, önceden bilinmeyen, fakat potansiyel olarak işe yarar bilginin elde edilmesi işlemlerinin tümüne “veri madenciliği” adı verilmektedir (Popov ve diğ., 2003), (Baykal ve Coşkun, 2009). Veri madenciliği deyimi ifadesi bazı literatürlerde, veritabanlarında bilgi madenciliği (knowledge mining from databases), bilgi çıkarımı (knowledge extraction), veri ve örüntü analizi (data/pattern analysis), veri arkeolojisi gibi farklı isimler ile de anılmaktadır.

Kullanılan yöntem bakımından klasik istatistiksel uygulamalar Berenndt ve diğ (2005) ve veri madenciliği benzer özellikler göstermesine karşın kullanılan veri yoğunluğu ve özelliği bakımından farklılıklar bulunmaktadır. İstatistiksel uygulamalarda düzenlenmiş ve özet veriler kullanılırken; veri madenciliğinde kullanılan veri setleri milyonlarca hatta milyarlarca olabilir. Veri sayısındaki yoğunluk sebebi ile veri madenciliği uygulamalarında özel analiz algoritmaları geliştirilmiştir (Gürcan ve Köse, 2008), (Özkan, 2008).

Kümeleme, veri özetleme, kümeler arasındaki ilişkilerin ve benzerliklerin bulunması, değişikliklerin analizi, aykırılıkların tespiti için yazılım ve istatistiksel yöntemler veri madenciliği uygulamalarında kullanılmaktadır (Ayaz, 2007). Uygulamada tek başlarına fazla anlam ifade etmeyen verilerin işlenip, anlamlı bilgi oluşturabilmesi sürecinde temel işlem adımlarından faydalanılmaktadır (Şekil 3.1).



Şekil 3. 1: Veri madenciliğinde işlem adımları (Özkan, 2008).

Birkaç veri kümesinin birleştirilmesi sonucu oluşturulan örnekleme kümesi oluşturulması aşaması “veri seçimi” olarak adlandırılmaktadır. Seçilen örnekleme kümelerinde yer alan hatalı kayıtların çıkarıldığı ve eksik nitelik değerlerinin değiştirildiği aşama olan “önişleme” adımı sayesinde keşfedilen bilginin kalitesinin artırılması amaçlanmaktadır. Veri madenciliğinde sorunun çalışma zamanını iyileştirmek amacı ile ilgisiz niteliklerin atıldığı ve tekrarlı kayıtların atıldığı “indirgeme” aşamasından sonra “veri madenciliği” uygulamalarından çalışma için uygun olan sınıflandırma, kümeleme vb. herhangi bir yöntem seçilmektedir. En son aşama olan “yorumlama ve doğrulama” bölümünde kazanılan bilginin geçerlilik, yenilik, yararlılık ve basitlik gibi kıstaslara göre değerlendirmesi yapılmaktadır.

3.1.1 Veri madenciliğinin uygulama alanları

Müşteri ilişkileri yönetimi, sepet analizi, çapraz satış, müşteri hatırlama, bankalarda kredi işlemleri, sahtekârlık saptama, tıbbi tanılar, hava tahminleri ve daha birçok alanda etkin olarak veri madenciliği uygulamalarına rastlanmaktadır. Kurumlar sahip oldukları veritabanlarındaki veri miktarı arttıkça, veriler arasındaki ilişkilerin ortaya çıkarılması çabası ile veri madenciliği yöntemlerinden faydalanma yoluna başvurmaktadırlar. Veri madenciliği yöntemlerinden birisi olan, ilişki analizi yöntemi sayesinde elde edilen sonuçlar; satış pazarlamadan, ürün katalog tasarımına kadar pek çok alanda kullanılmaktadır.

3.1.2. Veri madenciliğinde kullanılan yöntemler

Veri madenciliği uygulamalarında genellikle sınıflandırma, kümeleme, birliktelik kuralları yöntemleri kullanılmaktadır.

3.1.1.1 Sınıflandırma

Sınıflandırma yönteminin temeli; var olan veritabanının bir kısmının (3.2), eğitim amacı ile kullanılarak sınıflandırma kurallarının oluşturulması prensibine dayanır. Yeni bir durum ortaya çıktığı zaman (3.1), nasıl karar verileceği oluşturulan kurallar (3.3) yardımı ile belirlenmektedir (3.4). Yukarıdaki açıklamalar dikkate alındığında, sınıflandırma yönteminin matematiksel ifadesi aşağıdaki gibidir:

D veritabanı ve t her bir kaydı temsil etmek üzere

$$D=\{t_1, t_2, \dots, t_n\} \quad (3.1)$$

C_m adetten oluşan sınıflar kümesini temsil etmektedir;

$$C=\{C_1, C_2, \dots, C_m\} \quad (3.2)$$

$$F:D \rightarrow C \quad (3.3)$$

ve her bir t_i bir sınıfa dâhil olmalıdır.

Her bir C_j ayrı bir sınıftır ve her bir sınıf kendine ait kayıtları içerir.

$$C_j = \{t_i \mid f(t_i) = C_j, 1 \leq i \leq n, \text{ ve } t_i \in D\} \quad (3.4)$$

Sınıflandırma yöntemlerini uygulayabilmek için genelde istatistiksel yöntemler kullanılmaktadır. Geleneksel istatistiksel yöntemlerin dışında, veri madenciliği uygulamalarında Bayes sınıflandırma algoritması, karar ağaçlarına dayalı algoritmalar, yapay sinir ağları, K-NN (K en yakın komşu) algoritmaları sınıflandırma işlemleri için kullanılmaktadır. Sınıflandırma uygulamaları için geleneksel istatistiksel yöntemlerin yanı sıra veri madenciliğinde uygulanmak üzere geliştirilen algoritmalar ve özellikleri aşağıda sunulmaktadır.

“Karar ağaçları” entropiye dayalı olarak geliştirilen ID3 ve C4.5 algoritmaları, sınıflandırma, regresyon ağaçları, bellek tabanlı sınıflandırma modellerinden oluşmaktadır.

CART adı ile bilinen **“sınıflandırma ve regresyon”** yöntemlerinde Twoing algoritması ve Gini Algoritması kullanılan algoritmalar arasındadır. Bu algoritmaların, belirli ölçütlerin uygulanarak bölünme işlemlerinin gerçekleştirildiği bir yapısı vardır.

“K-En yakın komşu”(K-NN, K Nearest Neighbor) yönteminde, sınıfları belli olan bir örnek kümesindeki gözlem değerlerinden yararlanılarak, örneğe katılacak yeni bir gözlemin hangi sınıfa ait olduğunu belirleme amacı yatmaktadır.

Veri madenciliği uygulamalarında sınıflandırma işlemlerinde istatistiksel teknikler de kullanılmaktadır. **“Bayes sınıflandırıcıları”** da istatistiksel yöntemlerden birisidir. Değişkenlere ait alt kümeler arasındaki koşullu bağımsızlıkları tanımlamak üzere Bayes ağları da kullanılabilir.

“Optimizasyona dayalı sınıflandırma” modelleri arasında bulunan destek vektör makinesinde ise, sınıflandırma işlemi doğrusal veya doğrusal olmayan bir fonksiyon yardımıyla yapılmaktadır.

Sınıflandırma işlemi, veritabanlarındaki gizli örüntüleri ortaya çıkarmakta kullanılmaktadır. Gizli örüntülerin sık olarak rastlandığı resim, hastalık tanıları, dolandırıcılık tespiti, kalite kontrol çalışmaları ve pazarlama konuları, sınıflandırma yöntemlerinin kullanıldığı alanlardır. Sınıflandırma tahmin çalışmalarında kullanılan bir modeldir. Bir sonraki gün havanın nasıl olacağı vb. gibi, bir işaret veya sürecin geçmişteki gözlemlerine dayanarak gelecekteki muhtemel değerleri kestirme (öngörü) işlemlerinde sınıflandırma yöntemleri kullanılmaktadır (Silahtaroglu 2008), (Özkan 2008) .

Çalışmada web sayfalarında bulunan işyeri adres türlerinin belirlenmesi amacı sınıflandırma yöntemlerinden birisi olan K-NN algoritması kullanılmıştır. K-NN algoritması ve çalışma sistemi 4. bölümde detaylı olarak verilmektedir.

3.1.1.2 Kümeleme

Verilerin kendi aralarındaki benzerlikleri göz önüne alınarak, gruplandırılması (3.6) işlemi kümelemenin konusuna girmektedir. Sınıflandırma yönteminde, sınıflar önceden belirli iken, kümeleme yönteminde sınıflar önceden belirli değildir (3.5).

Kümeleme işleminin matematiksel tanımı aşağıdaki şekilde ifade edilir.

$$D=\{X_1, X_2, \dots, X_n\} \quad n=1,2, \dots, M \quad (3.5)$$

bir veritabanı olsun ve her bir X_n bir kaydı temsil etsin.

$$X=\{x_1, x_2, \dots, x_i\}, \quad i=1,2, \dots, m \quad (3.6)$$

her bir x_i , adı, soyadı, yaş ve gelir gibi özellikler olmak üzere;

Kümelemede D veritabanı j adet K kümesine bölünür (Silahtaroglu, 2008), (Özkan, 2008) .

Verilerin hangi gruplara/kümelere hatta kaç değişik gruba ayrılacağı eldeki verilerin birbirine olan benzerliğine göre belirlenir (Kaya ve Köymen, 2008). Kümeleme analizi biyoloji, tıp, antropoloji, pazarlama, ekonomi ve telekomünikasyon, anormal durumların tespiti gibi birçok ve birbirinden farklı alanlarda kullanılmaktadır.

3.1.1.3 Birliktelik analizi

Veritabanı içinde yer alan kayıtların birbirleriyle olan ilişkileri birliktelik kuralları yöntemiyle belirlenmektedir.

Birliktelik kurallarının matematiksel ifadesi aşağıdaki gibidir.

$$I=\{I_1, I_2, \dots, I_m\} \quad (3.7)$$

I bir dizi- nesne kümesi olmak üzere,

$$T=\{t_1, t_2, \dots, t_n\} \quad (3.8)$$

T veritabanındaki işlemleri gösterdiği düşünüldüğünde.

Her bir t_n nin alacağı değer 0 veya 1 dir. Eğer $t_n=0$ ise I_m satın alınmamış, eğer $t_n=1$ ise I_m satın alınmış demektir. Her bir işlem için veritabanında ayrı bir kayıt vardır (Silahtaroglu, 2008), (Özkan, 2008) .

3.1.2 Mekânsal veri madenciliği

Mekânsal veritabanlarında; mekânsal ve mekânsal olmayan veriler arasındaki ilişkilerin analizinde mekânsal veri madenciliği tekniklerinden faydalanılmaktadır. Mekânsal veri madenciliği kullanılarak veriler arasında var olabilecek, açıkça görülmeyen ilişkiler kolayca tespit edilebilmektedir. Trafik algılayıcılarının yerleştirildiği yerlerin incelenmesi, tehlikede olan yaşam türleri için yerleşim alanlarının modellenmesi, kanserli hastaların kümелendiği bölgelerin saptanması, flüorür ve diş sağlığı arasındaki ilişkilerin bölgesel olarak tespiti mekânsal veri madenciliği uygulama örnekleri arasında yer almaktadır (Shekhar ve diğ. 2003).

3.2 Web Madenciliği

Dünya Çapında Ağ (World Wide Web) kısaca WWW veya web olarak ta ifade edilen WWW; internet üzerinde çalışan ve "www" ile başlayan adreslerdeki sayfaların görüntülenmesini, yazı, resim, ses, film, animasyon gibi pek çok farklı yapıdaki verilere ulaşma olanağını sağlamaktadır.

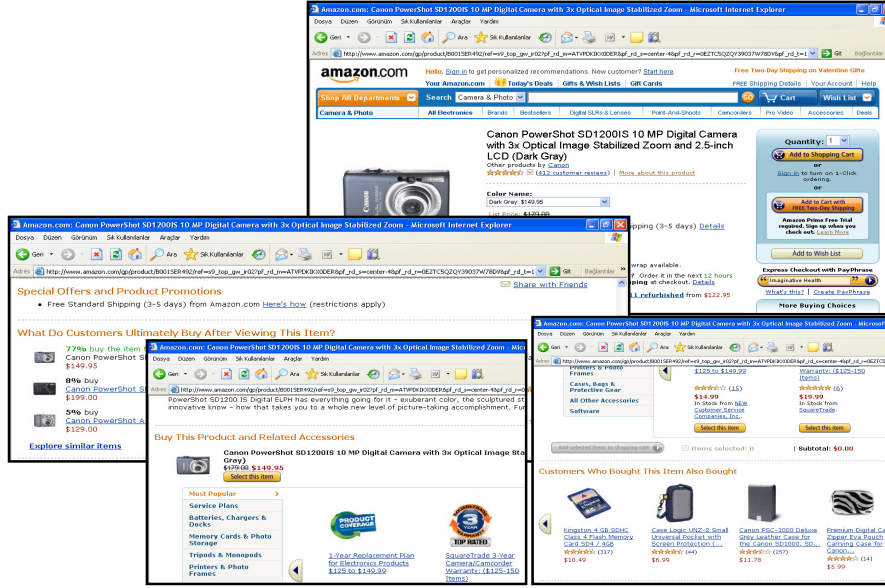
Birçok şirket, üniversite ve araştırma organizasyonu her geçen gün mevcut yapıya bir yenisi daha eklenerek webde yer almaktadırlar (Ayaz, 2007). 2009 Eylül verilerine göre 1,73 milyar kişinin dünya genelinde internet kullandığı, kullanıcı sayısının, 2008 yılına göre %18 arttığı saptanmıştır (Url1). Kurum ve kuruluşlar veri ve bilgi paylaşımlarını internet aracılığı ile sağlamaktadırlar. İnternetin ve içerdiği verinin sürekli ve düzensiz, yani yapılandırılmamış bir şekilde hızla büyümesine neden olmaktadır (Bhatia ve Kumar 2007), (Kunc, 2007), (Shakarmani ve diğ. 2010). Ayrıca artan kullanım ve paylaşım sonucunda internet sitelerinde yapılan işler değişmekte, ortak bir platformdan çok, ticari bir platform haline gelmektedir (Kosala ve Blockeel, 2000), (Daş ve diğ. 2006), (Baykal ve Coşkun 2009).

İnternette sunulan ve paylaşılan veriler ve bilgiler büyük olduğu kadar düzensiz ve birbirinden bağımsız bir yapıya sahiptirler. Veri ve bilgi bulutları internette son

derece dağınık ve düzensiz bir yapıda bulunmaktadır. Bu yönüyle web deki bilgiler tamamen anlamlı ve işe yarayan özellikler içermeyebilmektedir. Fakat insanoğlunun asırlar boyu edindiği bilgilerin ve günlük verilerin, 10-15 yıl gibi çok kısa bir sürede, internetten erişilebilir hale gelmesi pek çok kişiyi umutlandırmaktadır.

Tezuka ve Tanaka (2005) internet'in gelişmesi, modern grafik ara yüzlerinin ilerlemesi ve yüksek kapasiteli bilgi depolama cihazlarının geliştirilmesi ve ucuzlaması gibi gelişmeleri dikkate alarak, interneti büyük bir veritabanı olarak tanımlamışlardır. İnternetin sahip olduğu dağınık ve düzensiz bilgiyi yapılandırılmış (belirli bir yapıya uygun kayıtlı ve düzenli) olarak elde etmeyi amaçlayan çalışmalar gün geçtikçe artmaktadır (Bhatia ve Kumar, 2008).

Birçok şirket, üniversite ve araştırma organizasyonu internet siteleri ile ilgilenen kullanıcıları tanımak istemekte, haklarında bilgi toplamakta ve ihtiyaçları doğrultusunda yönlendirmektedir (Şekil 3.2). İnternet kullanıcılarının davranış bilgileri, internet sunucularında ham veriler şeklinde tutulması sayesinde Ayaz (2007) internet üzerinden satın alınan bir ürüne ait özelliklerin yanı sıra, ürün ile gerekli olabilecek yan ürünlerin kullanıcıya sunulması, ürün ile benzer özelliklere sahip farklı firmaların ürettiği diğer ürünler ve diğer kullanıcıların bu ürün hakkındaki yorum ve değerlendirmelerinin yanı sıra aldıkları ürünlerde kullanıcıya sunulan seçenekler arasında yer almaktadır.



Şekil 3. 2: Web sayfalarında ürün tanıtımı ve kullanıcı potansiyelinin belirlenmesi.

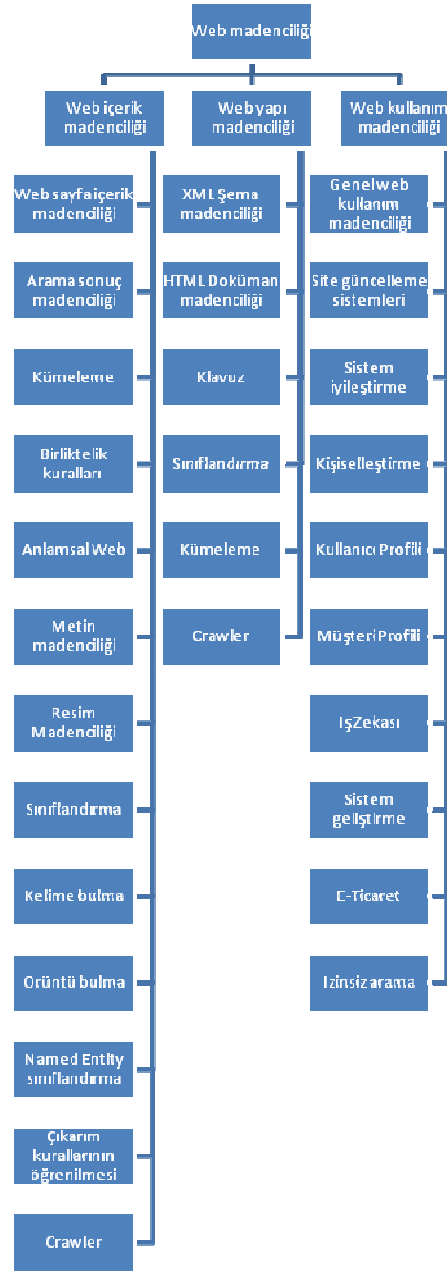
Yukarıda açıklanan oluşumların irdelenebilmesi ihtiyacı doğrultusunda ortaya çıkan, Web belgelerinden ve servislerinden, veri madenciliği tekniklerinin kullanılarak bilgilerin ayıklanması, ortaya çıkarılması ve analiz edilmesi işlemleri web madenciliği olarak tanımlanmaktadır (Daş ve diğ., 2006), (Kunc, 2007), (Bhatia ve Kumar, 2008).

3.2.1 Web madenciliğinin içeriği

Literatürlerde; web madenciliği, web sayfalarından elde edilecek analizlerin amaçlarına, incelenecek veri tiplerine (Çizelge 3.1) ve kullanılacağı alanlara göre, gerekli görülen durumlarda yöntemler tek başlarına veya birlikte kullanılmak üzere

- web içerik madenciliği,
- web yapı madenciliği ve
- web kullanım madenciliği yöntemleri olarak (Şekil 3.3) üç farklı bölüme

ayrılmaktadır (Özakar ve Püskülcü, 2002), (Kunc, 2007), (Gürcan ve Köse, 2008)



Şekil 3. 3: Web madenciliği sınıflandırması ve kullanım alanları (Özakar ve Püskülcü, 2002, Daş ve diğ., 2006, Gürçan ve Köse 2008).

Web içerik madenciliği genellikle kelime bulma, örüntü bulma, crawler ve anlamsal web vb. uygulamalarda kullanılırken, E-Ticaret çalışmaları, izinsiz aramaların tespiti, müşteri profili tespitinde web kullanım madenciliği teknikleri uygulanmaktadır.

Çizelge 3. 1: Web madenciliğindeki verilerin sınıflandırılması
(Berendt ve diğ. 2003), (Ayaz, 2007).

Veri Verinin yapısı Gösterimi	Web Madenciliği		
	İçerik Madenciliği	Yapı Madenciliği	Kullanım Madenciliği
	Metin belgeleri ve HTML Yapısız ve karışık veya DB İlişkisel ve sınıflandırılmalı	HTML linkleri Link yapısında Graph	server logs, browser logs... Kullanıcı etkileşimi İlişkisel tablolar ve graph

3.2.1.1 Web kullanım madenciliği

Cooley ve diğ. (1997) web kullanım madenciliğini, bir veya birden fazla web sunucusunu inceleyerek kullanıcı erişim desenlerinin otomatik keşfinin ve analizinin yapıldığı bir çeşit veri madenciliği etkinliği olarak tanımlamaktadır. Web kullanım madenciliği, kullanıcının internet sitelerinde dolaşırken geride bıraktığı erişim verilerinden bilgi üretmeyi amaçlar. Veriler, istemcilerde, sunucularda ve Proxy (Vekil sunucu) sunucularında depolanmakta, agent (üstlenici) kayıtları, istemci tarafında bulunan çerezler (cookies), kullanıcı profilleri (kayıt bilgileri), meta veri (metadata: sayfa özellikleri, içerik özellikleri, kullanılan veri), sunucu erişim kayıtları analizlerde veri kaynakları olarak kullanılmaktadır. Hedef müşteri kitlesinin belirlenmesinde, pazarlama uygulamalarında, ileriye yönelik planlamaların yapılmasında ve internet sitesinin hangi amaçla kullanıldığının gözlenmesi amacı ile internet kullanıcılarının erişim izlerinin analizleri tercih edilen uygulamadır (Kunc, 2007) .

3.2.1.2 Web yapı madenciliği

Web yapı madenciliği; dokümanlar arası belirli seviyelerde bağlantı (link) yapısının analiz sonuçlarını kullanılarak, web sitesinin ve web sayfalarının yapısal özetinin çıkarıldığı veri madenciliği çalışmasıdır (Cooley ve diğ., 1997)

3.2.1.3 Web içerik madenciliği

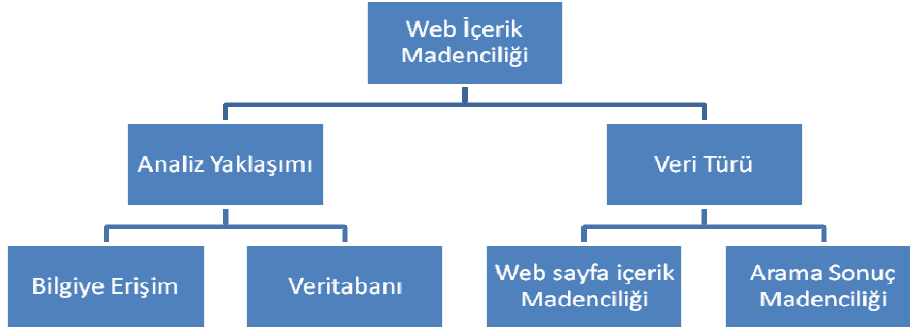
İnternette hızlı artış gösteren verilerden kullanım amaçlarına göre bilgilerin elde edilebilmesi için değişik yöntemler üzerinde araştırmalar yapılmaktadır (Egenhofer, 2002). Bu araştırmaların sonucu olarak ortaya çıkan web içerik madenciliği; WWW’deki dokümanların incelenmesi sonucu, metin, resim, ses, görüntü vb. içerik bilgilerine dayanılarak; erişim, benzerlik ve farklılıkların ortaya çıkarılmasını sağlayan bir yöntemdir. Yöntem kullanılarak sayfaların içeriği analiz edilmekte ve sayfanın temeline inilerek ilgili olduğu konu tespit edilmekte ve analizler sonucunda birbiriyle ilişkili ve aynı konuda olan sayfalar, kendi içinde sınıflandırılmaktadır.

Elektronik ticaret, online alışveriş gibi sistemlerin yaygınlaşması ile birlikte veri madenciliği kavramı ön plana çıkmaya başlamıştır.

Veritabanlarının veriyi sadece saklamak için tasarlandığı dikkate alındığında hiçbir işlem yapılmadan veri setleri incelendiği zaman verilerin bir anlam ifade etmediği görülmektedir. Veriler belirli bir amaç doğrultusunda işlendiği zaman anlamlı hale gelmektedir.

Web içerik madenciliği, web dokümanlarının belli istatistiksel metotlarla işlenip dokümanlardan istenen bilgilerin elde edilmesini sağlar. Web üzerindeki verilerin farklı türlerde olması, metin, ses, görüntü, resim, tablo, yardımcı veri, bağ verisi gibi farklı tiplerdeki yapılandırılmamış veri, Web içerik madenciliğinde verinin analizini zorlaştırmasından dolayı, verilerin daha kolay analiz edilebilmesi için farklı yaklaşımlar geliştirilmiştir.

Web içerik madenciliği çalışmaları literatürlerde, yapılan analiz çalışmalarına göre ve verinin türüne göre genel olarak ikiye ayrılmaktadır (Şekil 3.4). Yapılan analizlere göre bilgiye erişim (Information Retrieval-IR) yaklaşımında bilgiye erişim, bilginin analizi ve ayrıştırılması, analiz edilen bilgilerin amaca uygun olarak sınıflandırılması konularında yoğunlaşırken, veritabanı (dabase-DB) yaklaşımları web üzerinde bulunan verilerin bir veri tabanı üzerine kaydedilip, bu veri tabanı üzerinden modellenmesi, sorgulanması ve filtrelenmesi üzerinde çalışır. Web içerik madenciliğinde kaynak veri, web sayfalarının içeriklerinden oluştuğu için kullanılan veri tamamen sınıfsız ve karmaşık bir veridir. Bu yüzden web sayfalarının bir veri tabanı üzerinde modellenmesi işleri daha kolay hale getirir. Sorgulama, filtreleme ve sınıflandırma işlemleri veritabanı üzerinde daha kolay yapılmaktadır. Veritabanı yaklaşımları, web sayfalarındaki yarı yapılandırılmış veri üzerinde standart sorgulama ve veri madenciliği tekniklerini kullanma alanlarına odaklanmıştır (Cooley ve diğ., 1997), (Kunc, 2007), (Gürcan ve Köse, 2008), (Pate ve diğ., 2008).



Şekil 3. 4: Web içerik madenciliği.

Web içerik madenciliği işlenen verinin türüne göre iki gruba ayrılır:

Web sayfa içerik madenciliği, web sayfalarının içeriklerinden farklı bilgiler elde etmek için kullanılır. Bu bilgi metin olabileceği gibi sayfalardan resim, video veya ses gibi değişik formatlardaki bilgilerinde çıkarımı sağlanabilir.

Arama sonuç madenciliği, arama motorlarından elde edilen arama sonuçlarının sınıflandırılmasını sağlar. Bu sayede arama sonuçlarından, aranan hedefe daha yakın olanlarının saptanması mümkün olabilmektedir. Sadece kelimelerin istatistiksel değerlendirilmelerine dayanan içerik madenciliği bazı durumlarda başarısız olabilmektedir. Örneğin, seçilen kelimelerin aranan sayfada çok az sayıda bulunması durumunda başarısız sonuçlar üretilebilmektedir. Burada, anlamsal içerik analizi, konu madenciliğinde başarının artırılmasına önemli katkılar sağlayabilmektedir. Bu durumda, seçilen kelime ve kelime gruplarına dayanan istatistiksel sonuçlarla, anlamsal analiz sonuçları ile birlikte değerlendirilerek daha doğru sonuçlara ulaşılabilir.

Web içerik madenciliği analizlerini tek başına tanımlamak yetersiz kalmaktadır. Analizi gerçekleştiren veri türleri veri madenciliği konusuna girmemesine rağmen veri madenciliği teknikleri kullanılması dolayısı ile ilgili veri madenciliği ile bazı bakımlardan benzer özellikler taşımaktadır. Veri madenciliğinde, tam olarak yapılandırılmış veriler kullanılırken; web verileri yapılandırılmış ve yapılandırılmamış verilerden oluşur (Pol ve diğ., 2008). Web içerik madenciliğinin metin madenciliği ile benzer özellikler göstermesinin nedeni ise verilerin büyük bir çoğunluğunun yukarıda da açıklandığı üzere yarı yapılandırılmış veya yapılandırılmamış metin özelliği taşımasından kaynaklanmaktadır. Web içerik madenciliği ile metin madenciliği arasındaki fark ise metin madenciliğinin tamamen yapılandırılmamış veriler üzerine odaklanmış olmasıdır.

3.2.2 Uygulama alanları

Kömeşli (2007) de yapmış olduğu çalışmada ilişkili veya anlamsal olarak birbirine yakın bilgilerin, araştırılması ve çıkarılması sırasında insanın fiziksel ve düşünsel araçlarını kullanarak işlemesinin zaman alıcı ve zor olmasından dolayı yıllar geçtikçe otomatik bilgi çıkarımına olan ihtiyacının arttığını belirtmektedir. Ayrıca büyük bir bilgi kaynağı olarak görülen Internet'teki bilgilere ulaşmanın önündeki en büyük engelin içerdiği bilgilerin %90'ının doğal dillerle ifade edilmiş olmasından kaynaklandığını vurgulamaktadır. Bilgisayarların doğal (konuşma ve yazılı) dili henüz yeterli derecede anlayamıyor olmasından dolayı sayfalardaki ilişkili veya anlamsal olarak yakın bilgilerin bulunması büyük ölçüde insanlara düşmektedir.

Bilgi miktarının hızlı artışı sonucunda istenilen bilgiyi bulmak, işlemek, kullanmak için yapılan analizlerde değişmektedir. Geçmişten günümüze kadar büyük düzensiz verilerden anlamlı bilgilerin elde edilebilmesi için günümüze kadar değişik yöntemler denenmektedir. Hedeflenen çalışmalar arasında sayılan insanlığın yazılı kaynaklarının internet ortamına aktarılmasının tamamlanmasından sonra bu ihtiyaç daha da belirginleşecektir. İnternet ortamındaki verilerin büyük olması kadar, bu bilgilerin düzensiz oluşu da bu noktada web madenciliğine ayrı bir önem kazandırmaktadır.

Aranılan bir bilgiye ulaşmak işleminde, bilginin bulunabileceği dokümanlar, veri tabanları, tablolar, e-postalar, web sayfaları gibi ortamlar taranarak aranılan bilgi, bu ortamlar dâhilinde bulunan verilerden bilgi çıkarımıyla süzülmemektedir. Yıllar boyu kataloglama ve veri yönetimi gibi günlük işler bilgisayarlar tarafından gerçekleştirilmiştir. Ancak incelenen belgelerin çok fazla olması durumunda bu işlemin çok zaman alacağı da bir gerçektir. İnsanın ihtiyacı olan bilgiyi büyük bilgi yığınlarının içinden bulup çıkarmak işlemini bilgisayarlara yaptırabilmek her geçen gün daha fazla istenmektedir. Bilgi çıkarma işlemi bilgisayarlar tarafından gerçekleştirileceği için “otomatik bilgi çıkarma” terimi kullanılmaktadır. Otomatik bilgi çıkarmaya yönelik çalışmaların bir sonucu olarak ortaya çıkan Web madenciliği, veri madenciliği tekniklerinin kullanılarak web belgelerinden ve servislerinden bilginin ayıklanması, ortaya çıkarılması ve analiz edilmesi işlemidir.

Fakat etkili veri çıkarma sorunu hiçbir zaman çözülmemiştir. Bilgi çıkarma sistemleri insan aklı yerine bu incelemeyi yapmakta ve sadece istenilen amaç için gerekli

bilgiyi vermektedir. Otomatik bilgi çıkarma işleminin stratejisi bütün ilgili dokümanları önceden belirlenmiş bir formata uygun bilgilerin büyük metin dosyaları arasından bulunup çıkartılmasıdır (Kömeşli, 2007). Bu noktada “Web Madenciliği” için kabaca “WWW’de bulunan bilgilerin keşfedilmesi ve yorumlanması” da denilmektedir. Bir dokümanın karakteristiği üzerinde çalışıldığında, eğer doküman sorgu ile alakalı ise, doküman sorgulama yapıldığında cevap verebilecek şekilde olmalıdır.

Satış işlemlerinin desteklenmesi, çevrimiçi (online) veritabanlarında veri tekrarının (dublikasyon) önlenmesi, web sayfalarından finansal ve emlak verilerinin elde edilmesi, özel iş başvurularının incelenmesi, pazar araştırmaları ve analizleri, çevrimiçi gazete aramak ve haberlerini sunabilmek için web madenciliği tekniklerinden aktif olarak faydalanılmaktadır. Günümüzde “iş ilanları içeren bir gazete sayfasında mühendis arayanların telefon numaralarını listelemek” gibi bir görevi insanların yerine makinelere yaptırmak çok daha mantıklı bir hale gelmiştir (Daş ve diğ., 2006), (Ayaz, 2007). (Gürcan ve Köse, 2008), (Baykal ve Coşkun, 2009).

3.2.3 İşlem adımları

Etzioni (1996) Web madenciliği terimini “The World Wide Web: quagmire or gold mine?” isimli makalesinde ilk defa tartışmaya açarak, veri madenciliği tekniklerinin kullanılması ile world wide web de bulunan dosya ve servislerden otomatik örüntülerin bulunabileceğini ifade etmiştir. Daha sonraki çalışmalar için başlangıç noktası olarak kabul edilen makalede Etzioni, işlem adımlarını 3 temel kısma ayırmıştır. Günümüzde yapılan çalışmalarda işlem adımları aşağıda sırasıyla belirtildiği gibi 4 farklı kısımda ele alınmaktadır.

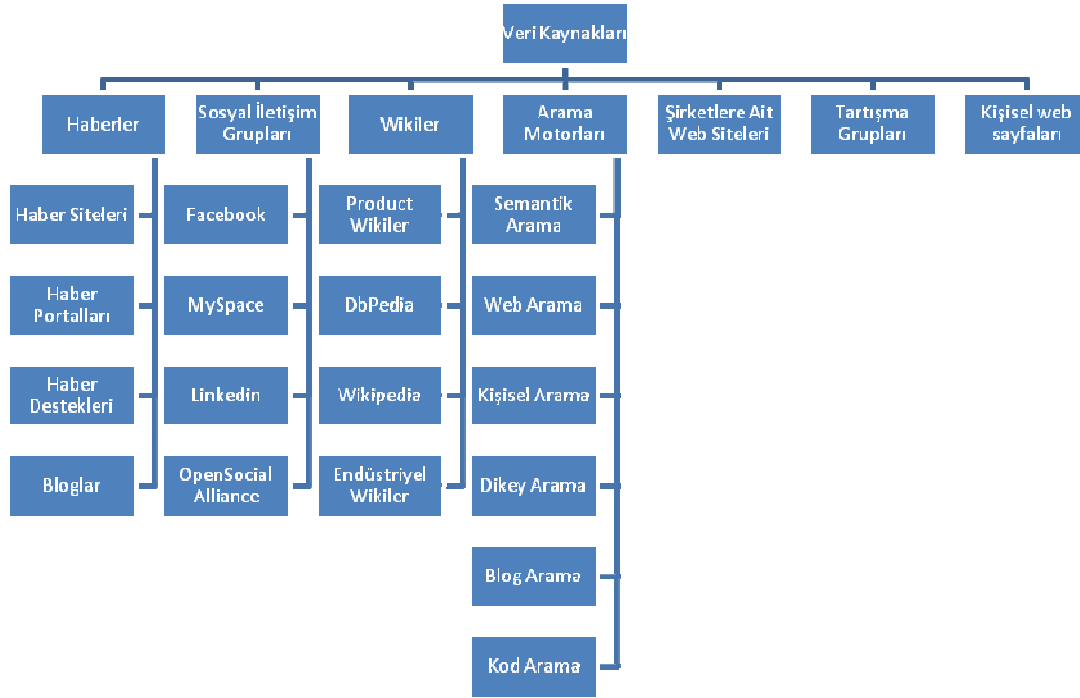
- Kaynak bulma (Resource Finding)
- Bilgi Çıkarımı ve Ön İşleme (Information Extraction and Pre-Processing)
- Genelleştirme (Generalization)
- Çözümleme (Analysis).

3.2.3.1 Veri kaynaklarının bulunması,

Kaynak bulma, kısaca işlenmemiş bilginin kazanımı, elde edilmesi kısmı olarak açıklanmaktadır. Web 'de mevcut bulunan bloglar, siteler, haber grupları Emel ve Taşkın (2002) vb. yerlerden verilerin toplanıp, HTML etiketlerinden ayrılarak (ham veri) çevrimiçi ya da çevrimdışı olmasına bakılmadan aramaları desteklemek amacı ile kayıt altına alınma aşaması veri kaynaklarının bulunması adımıdır.

Kullanılan kaynaklar

Internet'te var olan verilerin sürekli olarak değişmesi, güncellenmesi ve yeni bilgilerin eklenmesi web den bilgi çıkarımı (kazanımı) işleminde karşılaşılan en büyük zorluklardan birisi olarak belirtilmektedir. Web sayfalarının dinamik yapısından dolayı web den bilgi çıkarımı, normal metin tabanlı dokümanlara göre daha zordur, (Şekil 3.5). Web barındırdığı veriler ile değişken ve büyük bir veritabanı olarak kabul edilmektedir. Dokümanlar, veri tabanları, tablolar, e-postalar, kişisel bloglar, facebook, wikipedia vb. web sayfaları gibi yarı yapılandırılmış veya yapılandırılmamış özellikleri ile makinelerin anlayıp doğrudan bilgi çıkarabileceği şekilde değildir. Yapılandırılmış veriler ile işlem yapan veri madenciliği yöntemleri ancak web dokümanları ön işlem safhasından geçirildikten sonra uygulanabilmektedir.



Şekil 3. 5: Web madenciliğinde kullanılan veri kaynakları (Daş ve diğ. 2006, Baykal ve Coşkun 2009).

Web madenciliğinde kullanılan kaynaklar genel olarak web sayfalarındaki verilerin sunum özelliklerine göre yapılandırılmamış web sayfaları, yarı yapılandırılmış web sayfaları ve yapılandırılmış web sayfaları olmak üzere 3 farklı bölümde incelenmektedir.

a) Yapılandırılmamış web sayfaları

Yapılandırılmamış web sayfalarında (Şekil 3.6.) içerik düz metin şeklinde hazırlanmıştır. Web sayfasının düzenlenmesinde belli bir kural bulunmamaktadır. Bu tür web içeriklerinde ilgili veriye erişim için metin madenciliği teknikleri kullanılmaktadır. Yapısal olmayan veri (unstructured data) bilgisayarla işlenen bir veri yapısına sahip olmayan ya da makine tarafından kolay okunamayan bilgi yığındır.



Şekil 3. 6: Yapılandırılmamış web sayfası.

Ses (örnek telefon kayıtları) ve video verileri, e-posta içerikleri, kelime işlemci dokümanları, web sayfalarında yer alan forum verileri anket cevapları, müşteri, kamu kurumları vs. bildirimleri, öneri ve şikâyet bilgileri, wiki, çevrimiçi chat vb. yapısal olmayan veri örnekleri arasında yer almaktadır.

b) Yarı yapılandırılmış web sayfaları

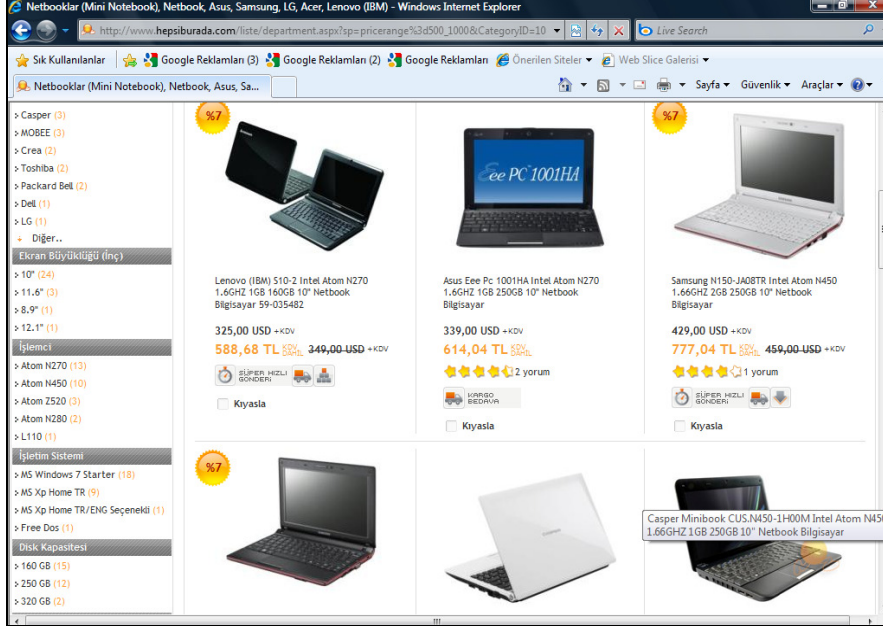
Web sayfalarındaki bilgilerin bir kısmı veritabanlarında hazır bir şekilde bir kısmı ise metin olarak tutulduğu web sayfaları yarı yapılandırılmış olarak adlandırılmaktadır (Şekil 3.7).



Şekil 3. 7: Yarı yapılandırılmış web sayfası.

c) Yapılandırılmış web sayfaları

Bütün bilgileri bir veritabanında tutulan web sayfalarına yapılandırılmış web sayfaları olarak adlandırılmaktadır. Bu tip web sayfalarına örnek olarak www.amazon.com ve www.hepsiburada.com vb. gibi alışveriş siteleri verilebilir (Şekil 3.8).



Şekil 3. 8: Yapılandırılmış web sitesi.

Web içerik madenciliğinde kaynak veri, web sayfalarının içeriklerinden oluştuğu için kullanılan veri tamamen sınıflandırılmamıştır ve karmaşık bir düzende sunulmaktadır (Daş ve diğ., 2006). Video, ses, görüntü, bağlantılı ve bağlantısız metinler içeren ve çoğu belli bir düzene sahip olmayan çoklu web dokümanlarından otomatik bilgi çıkarımı, web içerik madenciliği ilgi alanına girmektedir. Web üzerindeki verilerin farklı türlerde olması, metin, ses, görüntü, resim, tablo, yardımcı veri, bağ verisi gibi farklı tiplerdeki yapılandırılmamış veri, web içerik madenciliğinde verinin analizini zorlaştırır. Verinin yapısı ve verinin sunumundaki karmaşıklıklardan dolayı web içerik madenciliğinde hem veri madenciliği, hem de metin madenciliğinde kullanılan teknikler birleştirilip, web sayfalarının yarı yapılandırılmış özellikleri de dikkate alınarak amaca uygun bilgiler kazanılmaktadır (Özakar ve Püskülcü, 2002), (Ayaz, 2007), (Baykal ve Coşkun, 2009).

Web içerik madenciliği; e-ticaret, arama motorları gibi alanlarda yaygın olarak kullanımına başvurulmuş bir yöntemdir (Pater ve diğ., 2008). Google, Lycos, Altavista

gibi arama motorlarında web de bulunan dokümanların içeriğine bakılarak sınıflandırılma ve konularına göre düzenleme işlemleri web içerik madenciliği yöntemlerinin kullanılması sayesinde arama sonuçlarının aranılan konular ile ilgili olanları desteklemek için kullanılmaktadır (Daş ve diğ., 2006), (Ayaz, 2007), (Kunc, 2007), (Gürcan ve Köse, 2008).

Web içerik madenciliği; world wide web deki bütün dokümanları (metin, resim, ses, görüntü vb.) incelenmesi, dokümanların içerikleri arasındaki anlamsal (semantik) ilişkisel benzerliklerin ve farklılıkların ortaya çıkarılabilmesine yardımcı olan yöntemdir. Dolayısıyla web içerik madenciliğin amacı, bu kaynaklar arasından bilginin bulunması veya filtrelenmesidir. Web sayfa içerik madenciliği, web sayfalarının içeriklerinden farklı bilgiler elde etmek için kullanılır. Bu bilgiler metin biçiminde olabileceği gibi sayfalardan resim, video veya ses gibi değişik formatlardaki bilgilerin çıkarımı da olabilir.

Web içerik madenciliğinde yapılan analizler sonucunda dokümanlar arasındaki ilişkiler ve benzerlikler ortaya çıkartılmaktadır. Verinin analizi aşamasında sınıflandırma, kümeleme, birliktelik kuralları gibi yaygın olarak veri madenciliğinde kullanılan yöntemlere başvurulmaktadır.

Bu süreçte bilginin toplanıp, yorumlanmasında kullanıcının karşılaştığı birtakım problemler bulunmaktadır.

- Çok veri bulunması (çok fazla sayıda tekrar edilmiş bilgi olması):
Bunun sonucunda birçok arama sorgusuna karşılık olarak alakasız metinlerin de gelmesi ya da benzer sayfaların gelmesi,
- Arama kriterlerinin iyi belirlenememesi,
- Verinin kişiselleştirilememesi (Ayaz, 2007),
- Kelime tabanlı sınırlı sorgulama ara yüzü problemleri örnek olarak verilebilir (Bhatia ve Kumar, 2007), (Bhatia ve Kumar, 2008).

Web madenciliği teknikleri yardımı ile kullanıcıların karşılaştığı yukarıda anılan söz konusu problemlerin çözülmesi çalışmaları devam etmektedir (Kunc, 2007).

3.2.3.2 Verilerin önışlemden geirilmesi,

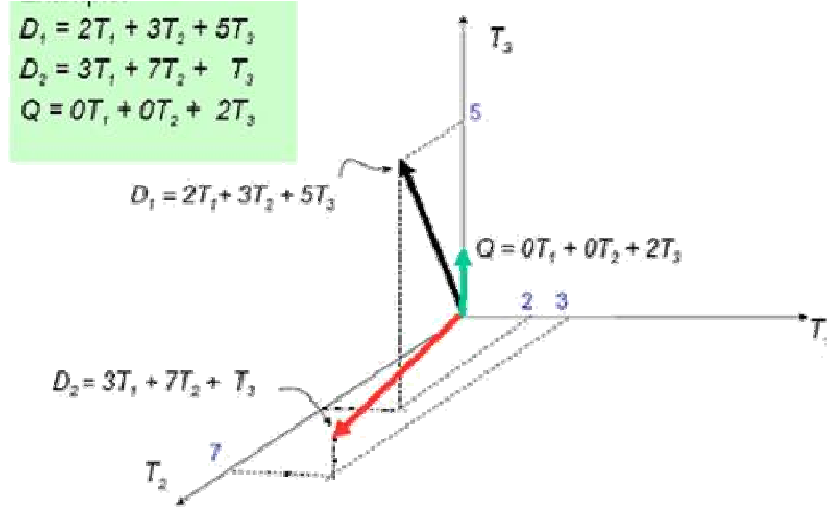
Genelleřtirme ařaması, iin gerekli verinin, asıl nemli bilgilerin, iřlenmemiř temel veriden elde edilmesi iin kullanılabilecek her trl dnřtrme ve temizlenme iřlemlerinin uygulandıėı ařama olarak tanımlanmaktadır. Verilerin nışlemden geirilmesi kaynakların elde edilmesi ařamasında ulařılan ham dokmanların ieriėine ve yapısına baėlı olarak deėiřiklik gstermektedir (Ayaz, 2007), (Pater ve diė., 2008). Verilerin nışlemden geirilmesi sırasında, yapılan iřlemler genelde ařaėıdaki gibi sıralanmaktadır.

- Html ifadelerinin temizlenmesi ařamasında (a href, br, b, p, font, table, div vb.) ifadeler metin dosyalarından ayıklanmaktadır.
- “ karakter grubunun temsil ettiėi noktalama iřaretlerinden (ift tırnaklar (“) gibi) karakter kodları ve noktalama iřaretlerinin metin dosyalarından temizlenmesi saėlanır.
- Nokta (.) ve tek tırnak (') dıřındaki noktalama iřaretlerinin temizlenmesi. Noktanın, ilgili metni cmlelere ayırmada ayırıcı (separator) olarak kullanmasından, tek tırnaėın ise, kendisinden sonraki karakterlerin iřleme alınmamasını saėlamasından dolayı temizlenmezler. Stop- Functional kelimeler, frekansı yksek ama bir anlam ifade etmeyen sadece yer kaplayan bir, ben, sen, o, ve vb. kelimelerdir. Yapılan iřlemler sayesinde web madenciliėi analiz srecinde kelime olarak anlam ifade etmeyen veriler temizlenerek daha az veri iřlenmektedir.

Vektr Uzak Modeli (Vector Space Model)

Metinlerin sınıflandırılması iřlemini gerekleřtirebilmek iin eėitim dokmanları ve eėitilmek istenilen dokmanlar vektrsel olarak uzakda ifade edilebilir duruma getirilmelidir. Salton (1975) yılında yaptıėı alıřmada, metinlerdeki terimleri tanımlayabilmek amacı ile vektr uzak modeli kavramını ifade etmiřtir. Vektr uzak modeli; dokman ierisindeki terimlerin aėırlıklarının saptanması ve aėırlıkların vektrn oluřturulmasında kullanılması prensibine dayanmaktadır. Bugn yapılan alıřmaların oėu vektr uzak modeli ile ifade edilmektedir (řekil 3.9). Vektr uzak modeli kategorilerin ierdiėi kelimeler ve bu kelimelerin bulunabileceėi eėitim dokmanları ve eėitilmek istenilen dokmanların eksenler halinde sunulması iřlemlerinin btndr. Kelimeler szlk adı verilen kelimeler tablosunda

tutulmaktadır. Genelde her bir terimin ağırlığının hesaplanması işlemi; terimlerin belgede geçme sıklığı ve bütün belgede geçme sıklığının oranlanması ile yapılmaktadır. Fakat terimlere ait ağırlılıklar hesaplanırken uzun belgeler kısa belgelere göre avantajlı bir durum sağlayabileceğinden dikkat edilmesi gereken bir durum oluşturmaktadır. Bilgi çıkarımı, bilgi filtreleme, indeksleme gibi alanlarda kullanılan model cebirsel özellik taşımaktadır.



Şekil 3. 9: Vektör uzay modeli gösterimi (Url2).

Dokümanlar kelimelerin vektörleri olarak ifade edilmektedir (Şekil 3.9). Şekilde T harfi kelimeleri ifade ederken D eğitim dokümanlarını (belge) Q bilinmeyen dokümanı (sorgu) temsil etmektedir. Uzak vektör modeli sayesinde belgeler ve sorgular matematiksel olarak ifade edilebilir hale gelmektedir. Her bir belge, anlamlı kelimelerden oluşan ve kelimelerin her birinin ağırlığı olan birer terim vektörü olarak sunulmaktadır. Oluşturulan belge ve sorgu vektörleri bir araya getirilerek bütün belgeleri içeren terim matrisi haline dönüştürülmektedir (3.9).

$$D = \begin{bmatrix} & T_1 & T_2 & \dots & T_i \\ D_1 & f_{11} & f_{21} & \dots & f_{i1} \\ D_2 & f_{12} & f_{22} & \dots & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} \quad (3.9)$$

D belge terim matrislerini ifade etmektedir.

Her bir D_1 , D_2 , D_n satırı eğitim amacı ile kullanılan dokümanları temsil etmektedir.

$T_1, T_2, \dots, T_i$ terim vektörleri olarak ifade edilmektedir.

Terim Ağırlığı Bulma

Dokümanların vektör olarak temsil edebilmesi amacı ile metin içerisinde geçen kelimelerin bir takım işlemlere sokulması gereklidir. Metinlerin genelleştirme işlem adımıyla kullanılmak üzere vektör haline getirilebilmesi amacı ile yararlanılan 3 farklı yöntem hakkında geniş açıklamalar aşağıda bulunmaktadır. Farklı yöntemler kullanılarak tespit edilebilen terim ağırlığı her bir terimin o belge için önemini belirtir.

- Terim Sayısı (Bit)

Anahtar sözcük sözlüğünde yer alan sözcüklerin metinde yer alıp almadığı gösteren vektörel bir gösterge oluşturulmaktadır. Metinde sözlükteki kelimelerin bulunup bulunmamasına göre 1 veya 0 rakamları ile tanımlanmaktadır.

$D=(0,1,0,1,0,0)$

- Terim Sıklığı- Devrik Belge Sıklığı (Frekans)

Sözcüklerin metinlerde kaç defa kullanıldığına dayanarak vektörel göstergenin oluşturulduğu bir yöntemdir.

$D=(0,7,0,1,6,0)$

- Terimin ağırlığı (Terim Ağırlıklandırma) (Tf-Idf)

Belge içerisinde bir terimin geçme sayısı ile doğru orantılıyken; bütün belge havuzu içerisinde o terimin geçme sıklığıyla ters orantılı olmasına dayanan bir yaklaşımdır (3.10).

$$w_i = tf_i * \log \frac{D}{df_i} \quad (3.10)$$

Tf-Idf ağırlıklandırmasında her bir dokümandaki sözcüklerin frekansı analiz işlemlerinde esas alınan noktadır. Dokümanda daha fazla geçen (Tf değeri büyük sözcükler) o doküman için daha değerli olmaktadır. Ayrıca Idf tüm dokümanlarda seyrekleşen sözcükler ile ilgili bir ölçü vermektedir. Bu değer tüm eğitim dokümanları ele alınarak hesaplanmaktadır. Bu yüzden eğer bir sözcük dokümanlarda sık geçiyorsa, o doküman için belirleyici olmadığı düşünülebilir. Eğer sözcük dokümanlarda çok sık geçmiyorsa, o sözcüğün o doküman için belirleyici özelliği olduğu kabul edilebilir.

3.2.3.3 Genelleştirme

Kullanılan yaklaşımlar

Genelleştirme aşamasında modelleme, sınıflandırma, filtreleme ve kümeleme gibi analizlerin yapıldığı veri madenciliği ve metin madenciliği yöntemleri kullanılmaktadır (Ayaz 2007), (Pater ve diğ., 2008). Sınıflandırma işleminin uygulanabilmesi için terim sıklıkları veya diğer yöntemler ile ön işlemden geçirilerek düzenlenen verilerin uzaklıklarının veya benzerliklerinin hesaplanması gerekmektedir. Hesaplamalar için en çok kullanılan yöntemler aşağıda listelenmektedir.

Uzaklık Hesabı

- Öklid
- Mahalanobis
- Minkowski

Benzerlik Hesabı

- Dice,
- Kosinüs
- Jaccard
- Hamming
- Basit eşleştirme yöntemi.

Bu yöntemlerden biri seçilerek ağırlıklandırılmış veriler işlemi yapılır ve sınıflandırma işlemleri için hazır hale getirilir. Uygulamanın amacına göre yukarıda belirtilen yöntemlerden herhangi birisi seçilerek veya birden fazlası karşılaştırılarak sınıflandırma işleminde en iyi doğruluk, duyarlılık elde edilmeye çalışılmaktadır. Yukarıda adı geçen yöntemler arasında yeralan Öklid ve Kosinüs yöntemleri çalışmada kullanılması nedeni ile ayrıntılı olarak incelenmektedir.

Öklid Yöntemi

p eğitim sınıfını ve q sınıfı bilinmeyen veri setlerini ifade etmek üzere;

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + \dots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^n (p_i - q_i)^2} \quad (3.11)$$

Aradaki uzaklık farkının karelerinin toplamının karekökünün hesaplanması yöntemine dayanan (3.11) bu yaklaşımda bulunan sonuçlar arasında en küçük değere sahip olanlar alınmaktadır. Böylece istenilen örneğe yakın olduğu için bu sınıfa girdiği önerisi sunulmaktadır.

Kosinüs Yöntemi

$\text{Sim}(D1, Q) = \cos\theta$

$$\text{sim}(D1, Q) = \frac{D_i \cdot Q}{|D_i| \cdot |Q|} = \frac{\sum W_{ij} * W_{qj}}{\sqrt{\sum_j W_{i,j}^2} * \sqrt{\sum_j W_{q,j}^2}} \quad (3.12)$$

Benzerliklerin hesaplanması yöntemine dayanan Kosinüs yönteminde ise benzerlik değerlerinin 1 e en yakın olduğu değer seçilmektedir (3.12). Bütün yukarıdaki çalışmaların nedeni web sayfalarındaki metinlerin sayısal ifadelere dönüştürülerek veri madenciliği yöntemlerinin kullanılabilmesini sağlayabilmektir. Verilerin düzenlenmesi işlemini sonrasında çalışmanın amacına uygun olarak sınıflandırma, kümeleme vb. analizler uygulanmaktadır. Aşağıda kullanılan analizlere örnekler sıralanmaktadır.

- Navie Bayes,
- Destek vektör makinesi (Support Vector Machine),
- K- en yakın komşuluk algoritması,
- C 4.5,
- Rastgele orman sınıflandırma,
- Karar ağacı algoritmaları teknikleri ağırlıklandırma aşamasından sonra sınıflandırma amacı ile kullanılmaktadır.

Sınıflandırma ve kümeleme kavramları birbirine bir hayli yakın kavramlar olmalarına karşın; sınıflandırma işlemlerinde referans sınıfların bulunmasına karşılık kümeleme işlemlerinde referans sınıfların bulunmaması en önemli farktır. Sınıflandırma işleminde referans sınıflardan faydalanılarak, yeni kayıtların hangi sınıfa uygun olacağı tahmin edilmeye çalışılmaktadır. Kümelendirme işleminde ise

referans sınıfların olmaması sadece verilerin içindeki anlamlı kümeler bulunmaya çalışılmaktadır. Tez çalışmasının amacına uygun olarak yukarıda açıklanan nedenle çalışmalarda kümeleme işlemleri yerine sınıflandırma işlemi kullanılmıştır. Yukarıda sayılan sınıflandırma analizlerinden çalışmada kullanılmak üzere KNN analizi seçilmektedir. Seçilen analizin kolay uygulanabilir olması ve daha önceki web madenciliği çalışmalarındaki uygun sonuçlar getirmesi, KNN'in seçilmesinde etkin olmuştur.

Ağırlıklandırma işlemi yapıldıktan sonra üzerinde Kosinüs ve Öklid yöntemleri denenilen veriler K-NN sınıflandırması için hazır hale getirilmektedir.

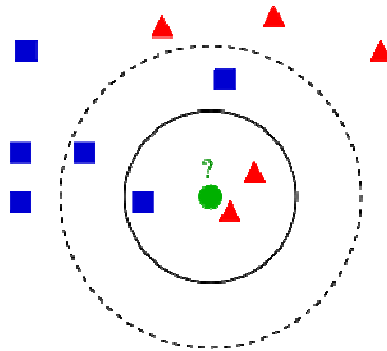
K-NN (K En Yakın Komşuluk Analizi) Sınıflama Yöntemi

Avantajları

- Uygulanabilirliği basit bir algoritmadır.
- Gürültülü eğitim dokümanlarına karşı dirençlidir.
- Eğitim dokümanları sayısı fazla ise etkilidir.

Dezavantajları

- K parametreye ihtiyaç duyar (şekil 3.10).
- Uzaklık bazlı öğrenme algoritması, en iyi sonuçları elde etmek için, hangi uzaklık tipinin ve hangi niteliğin kullanılacağı konusunda açık değildir.
- Hesaplama maliyeti gerçekten çok yüksektir. Bunun nedeni her bir sorgu, örneğinin tüm eğitim örneklerine olan uzaklığını hesaplamak gerekmektedir.



Şekil 3. 10: KNN algoritmasında sınıflandırma işlemi gösterimi (Url3).

Değerlendirme adımı genellikle karışıklık matrisi (Confusion matrix) kullanılmaktadır. Matristen faydalanılarak doğruluk, hata, anma, duyarlılık ve f

ölçütü değerleri bulunur. Karışıklık matrisi ilk kez Kohavi ve Provast tarafından 1998 yılında kullanılmaya başlanmıştır (Çizelge 3.2).

Çizelge 3. 2: Karışıklık matrisi gösterimi

		Öngörülen Sınıf	
		Negatif	Pozitif
Doğru Sınıf	Negatif	A (TP)	B (FN)
	Pozitif	C (FP)	D (TN)

$$\text{Doğruluk}=(A+D)/(A+B+C+D) \quad (3.13)$$

$$\text{Hata}=(C+B)/ (A+B+C+D) \quad (3.14)$$

$$\text{Anma}=D/(B+D) \quad (3.15)$$

Anma (Recall): Elde edilen dokümanın ilgili olma durumu

$$\text{Duyarlılık}=A/(D+C) \quad (3.16)$$

Duyarlılık (Prezisyon): Bütün dokümanlardaki ilgili dokümanın bulunması işlemidir.

$$\text{F ölçütü (Kümeleme ölçütü)} =(2*\text{duyarlılık}*anma)/(\text{duyarlılık}+anma) \quad (3.17)$$

Karşılaştırma yapılırken A modeline ait anma ve duyarlılık değerlerine bakıldığında B modelinden yüksek değer elde ediliyorsa “A modeli B modelinden daha iyi bir modeldir” denilmektedir.

Ayrıca K değeri seçilirken yüksek doğruluk değeri veren parametre de seçilmelidir. Ama her zaman mutlaka K değeri bir işlemle belirlenmez. Önceden bir değer varsayılarak buna göre işlemler yapılabilir (Bhatia ve Kumar, 2008).

3.2.3.4 Çözümleme

Web madenciliğinin son bölümü olup genelleştirme işlemlerinin analiz sonuçlarının tartışıldığı ve insan faktörünün en büyük rolü üstlendiği safhadır (Ayaz 2007), (Pater 2008).

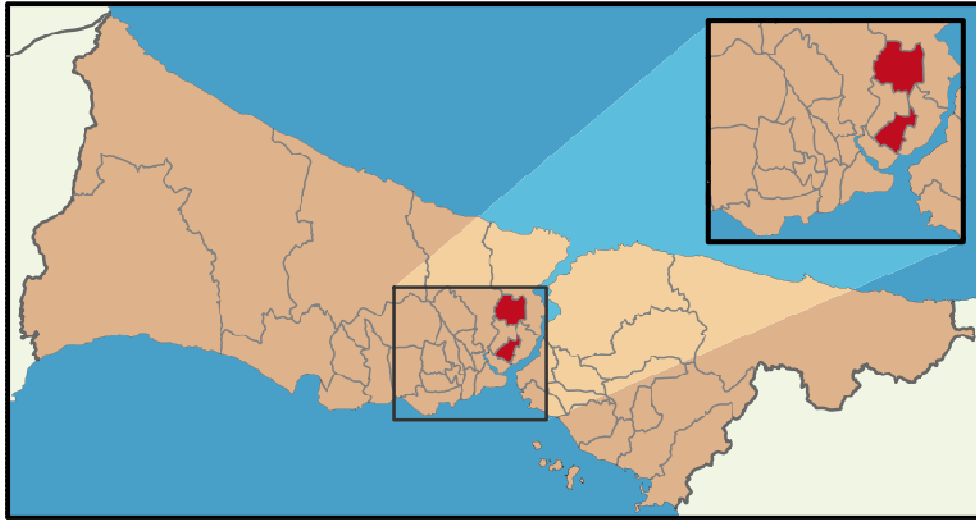
Bhatia ve Kumar (2008) anlamsal web madenciliği üzerinde durmaktadırlar. Web dokümanları ile işlem yapılması, web dokümanlarının sahip olduğu yarı yapısal veya yapısal olmayan özelliklerinden dolayı verilerin işlenmesi ve analizi bakımından veri madenciliğinden web madenciliğini ayırmaktadırlar.

Miah (2009), yapılan daha önceki çalışmaları incelendiğinde çeşitli metin sınıflandırma amacı ile kullanılan algoritmalara rastlamıştır. Bazıları var olan diğer algoritmalar ile performans bakımından karşılaştırılan yeni algoritmalar olduğu kadar, var olan diğer algoritmaların karşılaştırılması şeklinde de olmaktadır. Çeşitli sınıflandırma yöntemleri kullanılmaktadır. Bunlar Naive Bayes, Karar ağaçları (decision trees), karar kuralları (decision rules), regresyon yöntemleri, sinir ağları (Neural Networks), K-NN, destek vektör makinesi (Support Vector Machine- SVM), Rocchio sınıflandırıcıları örnek olarak verilebilir.

Miah (2009) da yapmış olduğu araştırmada benzerlik ölçümünün yapılabilmesi için terim frekansı yöntemini kullanmaktadır.

4. VERİLERİN KAZANIMI

Kültür, ekonomi, sağlık, turizm gibi sektörlerde hızlı değişim gösteren Şişli İlçesi, 30 km lik yüzölçümü, 28 mahallesi ile İstanbul'da Avrupa kıtasında yer almaktadır. Şişli İlçesinin doğuda Beşiktaş, Kuzeyde Sarıyer, Batıda Eyüp ve Kağıthane, Güneyde Beyoğlu ilçeleri ile komşuluğu vardır.



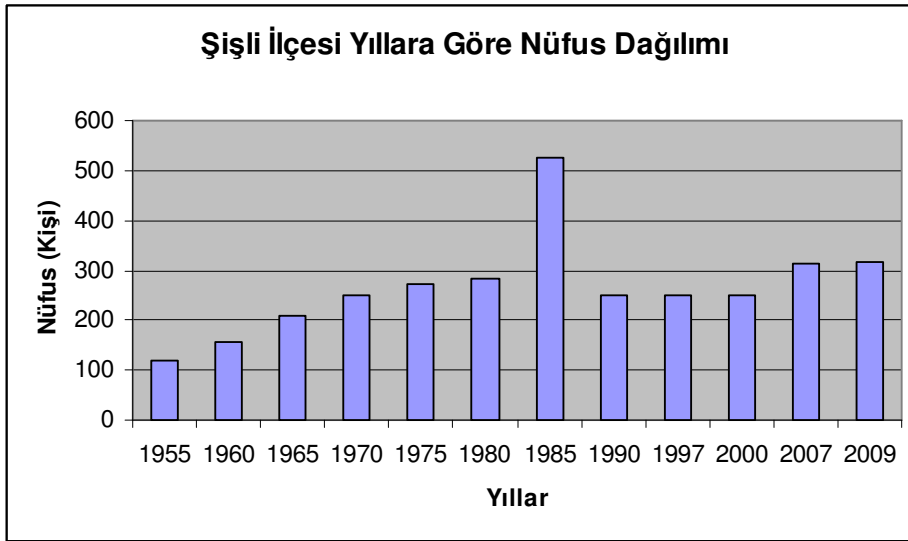
Şekil 4. 1: Şişli İlçesinin İstanbul haritası üzerinde gösterimi (Url4).

Tezin doğrudan konusu olmasa da, araştırma konusu olarak seçilen eğitim, sağlık, ticaret, turizm, kültür ve spordan oluşan 6 farklı obje grubunun ilçe açısından öneminin anlaşılabilmesi bakımından, Şişli ilçesinin yapısal özellikleri hakkında kısaca bilgi vermenin yararlı olacağı değerlendirilmektedir.

Denize sahili bulunmayan Şişli ilçesinde, çok sayıda tarihi eser, işyeri, modern ticaret, kültür ve sanat merkezleri bulunmaktadır. 1987 yılında, Kağıthane'nin ilçe olması ile ikiye bölünen Şişli (Şekil 4.1) kurulduğu 16 yy.'dan günümüze kadar ciddi değişimlerden geçmiştir. 16 yy. da çok az yerleşim bulunan Şişli; 19 yy. da özellikle Harbiye, Pangaltı ve Maçka mahallelerine doğru yayılma göstermiştir. Şişli Etfal Hastanesi, Bomonti Bira Fabrikası 19. yy. ın başında kurulan önemli ticaret ve sağlık kurumları arasındadır. Şişli, 19 yy.lın son çeyreğinde Taksim ve Şişli Meydanı

arasında atlı tramvay seferleri ile ulaşım bakımından da ilklerin yaşandığı bir bölgedir.

Özellikle 1920'lerden sonra bahçe içerisindeki evler ve konaklar yerini yavaş yavaş apartmanlara bırakmaya başlamıştır. 1950 yılından sonra Çağlayan, Gültepe semtleri yoğun göç bölgeleri olarak dikkat çekmektedir. Kâğıthane bölgesinin 1960'lı yıllarda sanayi merkezi olması, Beyoğlu'nda bulunan mağaza ve alışveriş merkezlerinin Harbiye, Nişantaşı, Osmanbey ve Şişli semtlerine kayması, Dolapdere'den 1970'lerde oto tamirhanelerinin kaldırılması amacı ile Çeliktepe'nin kuzeyinde sanayi alanının oluşturulması 50 senelik değişimi gözler önüne sermektedir.



Şekil 4. 2: Şişli ilçesi 1955-2009 yılları nüfus dağılım grafiği.

İstanbul'un diğer bütün ilçeleri gibi, sürekli bir nüfus artışına sahne olan Şişli ilçesinin 1980'lerde ikiye katlanan nüfusu (Şekil 4.2) Kâğıthane'nin ilçeden ayrılmasıyla 250 000'e kadar gerilemiştir. Yatay mekânsal genişleme imkânı olmadığından, nüfus sayısında yıllara göre önemli değişiklik görülmemektedir.

Nüfusun yüzde 85'i ilçenin güneyinde yaşar. Ayazağa, Maslak ve Huzur mahallelerinden oluşan kuzey kesiminde nüfus yoğunluğu düşüktür. Maslak'ın gelişen bir finans merkezi olmasından dolayı nüfusunda artış olmasına rağmen, güney kesime göre daha az yoğunluk göstermektedir. Çünkü güneydeki Mecidiyeköy ve diğer mahalleler daha büyük sanayi, finans ve eğlence merkezleridir.

Şişli İlçesi eğitim kurumları açısından oldukça zengindir. Nişantaşı Anadolu Lisesi (eski English High School), Notre Dame de Sion Fransız Lisesi, Nişantaşı Nuri Akın

Anadolu Lisesi (eski Nişantaşı Kız Lisesi), Saint Michel Lisesi, Şişli Anadolu Lisesi ilçedeki köklü ortaöğretim kurumlarıdır.

İlçe sınırları içindeki yükseköğretim kurumları arasında İstanbul Teknik Üniversitesi (İTÜ) önemli bir yer tutar. Marmara, Yıldız Teknik, İstanbul Bilgi, Beykent, Bahçeşehir, Haliç, İstanbul Bilim üniversitelerinin de ilçe sınırları içinde yerleşkeleri bulunmaktadır.

Sağlık kurumları bakımından zengin bir yapıya sahip olan Şişli’de en çok bilinen hastane Şişli Etfal hastanesidir.

Şişli ilçesi hızla değişim gösterirken 21. yüzyıl İstanbul'unun hatta Türkiye ekonomisinin iş ve finans merkezi olma durumundadır. Ticaret ve hizmet sektöründe çalışanlar ağırlıklı bir paya sahiptir. İmalat sanayi daha çok Ayazağa'daki Cendere Deresi kıyılarında yoğunlaşmıştır. Şişli-Mecidiyeköy-Esentepe-Levent-Maslak aksı üzerindeki ve çevresindeki bölge, Türk finans sektörünün kalbi durumuna gelmiştir. Diğer ticaret merkezleri arasında, Bomonti Bira Fabrikası, Cevahir Alışveriş Merkezi ve Perpa yer almaktadır.

İstanbul Hilton Oteli ve ilçede bulunan seyahat acenteleri ve oteller turizm sektörünün önemli yapılarıdır.

Mecidiyeköy'de bulunan Ali Sami Yen Stadı, Ayazağa Yusuf Tunaoğlu Stadı, Feriköy Stadı, İTÜ Ayazağa Yerleşkesi içindeki spor kompleksi ile Ayazağa'daki Eczacıbaşı Spor Salonu önemli Şişli spor tesisleri arasında yer almaktadır. Galatasaray Spor Kulübü için inşa edilen Aslantepe Türk Telekom Arena, Şişli İlçesinin kuzeyinde, Ayazağa - Seyrantepe arasında yapılmaktadır.

Birçok sinema ve tiyatro salonu bulunan Şişli, İstanbul'un başlıca kültür merkezlerinden biridir. İlçedeki önemli kültür merkezleri arasında Harbiye Açıkhava Tiyatrosu, Cemal Reşit Rey Konser Salonu, Lütüfî Kırdar Uluslararası Sergi ve Kongre Sarayı sayılabilir.

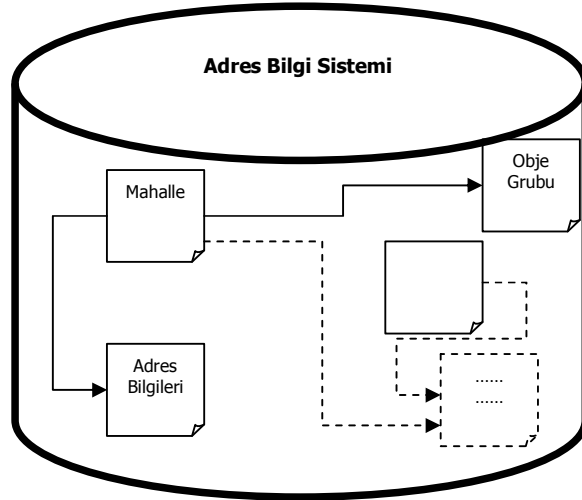
Bunların dışında Harbiye Muhsin Ertuğrul Sahnesi, Gönül Ülkü-Gazenfer Özcan Tiyatrosu, Kenter Tiyatrosu, Harbiye Cep Tiyatrosu, Profilo Kültür Merkezi, Stüdyo Tiyatrosu, Tiyatrokare ve Efe Sanat Evi diğer tiyatro ve performans mekânlarıdır. Askeri Müze ve Atatürk Müzesi bu sektörün diğer iki önemli tesisidir.

Türkiye’de bütün illere ve ilçelere uygulanabilir yapısı bulunan web madenciliği ve AFFWeb uygulaması öncelikle Şişli bölgesi verileri üzerinde denenmiştir. Bölgenin hızlı değişim göstermesi ve İBB’deki adres bilgi sistemi verilerinin eksiksiz olarak sunulması, Şişli ilçesinin öncelikli seçilme nedenleri arasında yer almaktadır.

Yapılan çalışmada temel olarak 2 farklı veri seti kullanılmaktadır. 1. veri seti İstanbul Büyükşehir Belediyesi Adres Yönetim Şefliği tarafından hazırlanan Adres Bilgi Sistemi verileridir. 2. veri seti ise, kayıt altına alınmış web sayfalarıdır. Web sayfalarından daha sonra da geniş olarak ifade edileceği gibi hem eğitici hem de eğitilen veriler olmak üzere iki farklı biçimde faydalanılmaktadır. 4. bölüm içerisinde adı geçen her iki veri setine de detaylı olarak değinilmektedir.

4.1 İBB Mevcut Adres Bilgi Sistemi Yapısı

Çalışmada belediye verilerinden, web madenciliği ve kontrol amaçlı olmak üzere iki farklı alanda faydalanılmaktadır. Web madenciliği alanında, web sayfalarının önışlem adımlarından birisi olan kelime sıklıklarının belirlenmesi aşamasında kullanılan sözlüğün oluşturulmasında “belediye obje grupları” aynen alınmıştır. Kontrol kısmında ise, web sayfalarından kazanılan adres bilgilerinin doğruluğunu ölçebilmek amacı ile belediye mahalle ve obje kodu verileri ile birlikte adres bilgileri kullanılmaktadır.



Şekil 4. 3: Adres Yönetim Şefliğinin, ilçe bazında sayısal ortamda Adres Bilgi Sistemi genel veritabanı şeması (2009).

Çalışma içerisinde kullanılan alanlar dikkate alındığında, faydalanılmış olan belediye veri setinin özelliklerine ve istatistiksel bilgilere kısaca değinmek uygun olacaktır.

Adres bilgi sisteminde (Şekil 4.3) yer alan veritabanı içerisinde yaklaşık 10 adet tablo bulunmaktadır. Çalışma için seçilen mahalle, obje kodu, adres bilgileri tabloları Adres Bilgi Sisteminin küçük bir parçasını oluşturmaktadır.

İBB ye ait Adres Bilgi Sistemi veritabanı ARCGIS Coğrafi Bilgi Sistemi yazılımı kullanılarak hazırlanmıştır.

OBJECTID	MahNoID	MahAdID	KAPI_NO	ESKI_KAPI_NO	TAHSIS_KAPI_NO	ESKI_YOL_ADI	KAPI_ADI	YOL_ADI	OBJE_ANA_GRUP
1	214	KÜŞTEPE	14/3			Alev Sk.	Konut	Alev Sokağı	
2	214	KÜŞTEPE	131/1			Yavrukus Sk.		Yavrukus Sokağı	
3	214	KÜŞTEPE	131/2			Yavrukus Sk.		Yavrukus Sokağı	
4	214	KÜŞTEPE	133/4			Yavrukus Sk.		Yavrukus Sokağı	
5	214	KÜŞTEPE	133/2			Yavrukus Sk.		Yavrukus Sokağı	
6	214	KÜŞTEPE	12			Doğanay Sk.		Kartopu Çıkmağı	
7	214	KÜŞTEPE	14			Doğanay Sk.		Kartopu Çıkmağı	
8	305	FERİKÖY	7	7		Al Karanfil Sk.		Bayır Sokağı	
9	305	FERİKÖY	9	9		Al Karanfil Sk.		Bayır Sokağı	
10	305	FERİKÖY	11	11		Al Karanfil Sk.		Bayır Sokağı	
11	305	FERİKÖY	13	13		Al Karanfil Sk.	Yüksel Apartmanı	Bayır Sokağı	
12	305	FERİKÖY	15	15		Al Karanfil Sk.	Balkancı Apartmanı	Bayır Sokağı	
13	305	FERİKÖY	4	8		Al Karanfil Sk.	Şahinler Apartmanı	Bayır Sokağı	
14	305	FERİKÖY	6	10		Al Karanfil Sk.		Bayır Sokağı	
15	305	FERİKÖY	8	12		Al Karanfil Sk.	Mertaş Apartmanı	Bayır Sokağı	
16	305	FERİKÖY	10	14		Al Karanfil Sk.	Emek Apartmanı	Bayır Sokağı	
17	305	FERİKÖY	12	16		Al Karanfil Sk.	Ayaz Apartmanı	Bayır Sokağı	
18	305	FERİKÖY	4	4		Açık Yol Sk.	Hidroglu Apartmanı	Açıkyol Sokağı	
19	305	FERİKÖY	6	6		Açık Yol Sk.	Yasemin Apartmanı	Açıkyol Sokağı	
20	305	FERİKÖY	10	10		Açık Yol Sk.	Erdoğan Apartmanı	Açıkyol Sokağı	
21	305	FERİKÖY	12	12		Açık Yol Sk.	Sözen Apartmanı	Açıkyol Sokağı	
22	305	FERİKÖY	14	14		Açık Yol Sk.		Açıkyol Sokağı	
23	305	FERİKÖY	16-18	16-18		Açık Yol Sk.	Çağlar Cafe	Açıkyol Sokağı	
24	305	FERİKÖY	22	22		Açık Yol Sk.		Açıkyol Sokağı	
25	305	FERİKÖY	24	24		Açık Yol Sk.		Açıkyol Sokağı	
26	305	FERİKÖY	30A	30		Açık Yol Sk.	Arka Kullanılmayan Kapı	Açıkyol Sokağı	
27	305	FERİKÖY	3	3		Açık Yol Sk.	Oto Tamir	Açıkyol Sokağı	
28	305	FERİKÖY	15	17		Açık Yol Sk.		Açıkyol Sokağı	
29	305	FERİKÖY	17	19		Açık Yol Sk.		Açıkyol Sokağı	
30	305	FERİKÖY	152	222		Savaş Sk.	Günaydın Apartmanı	Savaş Sokağı	
31	305	FERİKÖY	154	224		Savaş Sk.		Savaş Sokağı	
32	305	FERİKÖY	156	226		Savaş Sk.	Soylu Apartmanı	Savaş Sokağı	
33	305	FERİKÖY	158	228		Savaş Sk.	Kibar Apartmanı	Savaş Sokağı	
34	305	FERİKÖY	160	230		Savaş Sk.	Suna Apartmanı	Savaş Sokağı	

Şekil 4. 4: Adres bilgi yönetimi örneği.

Veritabanı tabloları arasında mahalle bilgileri, adres bilgileri (Şekil 4.4), obje grupları bilgileri ve ARCGIS uzantısı olan shape dosyaları bulunmaktadır. 2009 İstanbul Büyükşehir Belediyesi verilerine göre adresler tablosunda gösterildiği biçimiyle Şişli İlçesinde 53951 kayıtlı adres bilgisi bulunmaktadır. Bu adreslere ait tablo bilgileri aşağıdaki gibidir.

Çizelge 4. 1: Şişli Adres Bilgi Sistemi kapı numarası sayısı.

ADRES SAYISI			
Eski Kapı Numarası Bulunan	Değişmeyen Kapı Numarası Bulunan	Tahsis Edilmiş Kapı Numarası Bulunan	Hiç Kapı Numarası Bulunmayan
22936	11065	2826	3073

İBB tarafından yaptırılan ve yenilenen Adres Bilgi Sistemi verilerine göre kapı numarasına sahip adreslerden 11065 adetinin eski numaralarında bir değişim olmamıştır. 2826 adrese ise yeni kapı numarası tahsis edilmiştir. Herhangi bir kapı numarası bilgisine rastlanmayan adres sayısı ise 3073 dür (Çizelge 4.1).

4.1.1 Mahalle

28 mahalleden oluşan Şişli İlçesi iki farklı kısımda incelenebilir (Şekil 4.5). Ayazağa, Maslak ve Huzur mahallerinin bulunduğu kısım Kâğıthane bölgesi olarak adlandırılmaktadır.



Şekil 4. 5: Şişli ilçesi, 28 mahalleden oluşmaktadır.

Adres Yönetim Şefliği tarafından hazırlanan ilçe bazında sayısal ortamda Adres Bilgi Sistemi mahalle tablosu incelendiğinde Şişli ilçesinde bulunan bütün mahalleler,

standart olarak belirlenen ID bilgileri, sahip oldukları veri tipleri, kapladıkları alan bilgileri görülmektedir (Şekil 4.6).

OBJECTID	MahAdID	MahNoID	SHAPE	SHAPE_Length	SHAPE_Area
1	MAHMUT ŞEVKET PAŞA	307	Uzun ikili veri	2558,86340160488	340005,371472326
2	İNÖNÜ	301	Uzun ikili veri	1883,56881953552	183820,051145723
3	ESKİŞEHİR	303	Uzun ikili veri	1738,22379535606	184106,773562543
4	FERİKÖY	305	Uzun ikili veri	2708,43999050673	242525,260800267
5	HALİL RIFAT PAŞA	294	Uzun ikili veri	2974,43532682847	435118,637628361
6	ESENTEPE	211	Uzun ikili veri	6804,49056596371	1507922,26148818
7	GÜLBAHAR	212	Uzun ikili veri	3426,42735175112	416275,244472179
8	KUŞTEPE	214	Uzun ikili veri	3173,64791853462	404504,876294677
9	KAPTANPAŞA	293	Uzun ikili veri	2526,04905975898	278310,529382238
10	İZZET PAŞA	215	Uzun ikili veri	1824,16615092018	176689,327187017
11	PAŞA	306	Uzun ikili veri	4141,55973820239	744178,784064805
12	DUATEPE	308	Uzun ikili veri	1804,46776503366	138072,95764942
13	ERGENEKON	300	Uzun ikili veri	1590,48632171516	102521,154367159
14	BOZKURT	302	Uzun ikili veri	1853,40033125213	181253,031081085
15	MEŞRUTİYET	296	Uzun ikili veri	2311,638115611	301241,046274092
16	AYAZAĞA	207	Uzun ikili veri	20205,9355280879	15857229,0871461
17	TEŞVİKİYE	297	Uzun ikili veri	3038,24932607425	450164,683788403
18	MECİDİYEKÖY	213	Uzun ikili veri	3425,51536753286	454929,442876473
19	YAYLA	304	Uzun ikili veri	2669,41759573144	241269,115585011
20	MERKEZ	218	Uzun ikili veri	6199,18406046898	1127973,22831837
21	HALİDE EDİP ADIVAR	292	Uzun ikili veri	2758,60959175485	350672,665562178
22	CUMHURİYET	295	Uzun ikili veri	3145,99823360698	362214,854417711
23	19 MAYIS	217	Uzun ikili veri	2997,59660213463	435052,681318125
24	FULYA	216	Uzun ikili veri	3568,47492014352	679581,020277796
25	MASLAK	210	Uzun ikili veri	12490,8103258918	6089050,68750242
26	HUZUR	208	Uzun ikili veri	8169,59284154244	2408599,89840592
27	HARBIYE	299	Uzun ikili veri	4126,44868223217	806585,451254458
28	HALASKARGAZI	298	Uzun ikili veri	1629,2914930249	99741,0570088858
* omatikSayı)					

Şekil 4. 6: Adres Yönetim Şefliği, İlçe Bazında Sayısal Ortamda Adres Bilgi Sistemi Mahalle ID ve MahNoID gösterim örneği (2009).

Şişli İlçesi Adres Bilgi Sistemi mahalle bazında adres verilerine göre Ayazağa, Kuştepe, Mecidiyeköy ve Maslak mahallelerine ait adres bilgilerinin diğer mahallere oranla daha fazla olduğu görülmektedir (Çizelge 4.2).

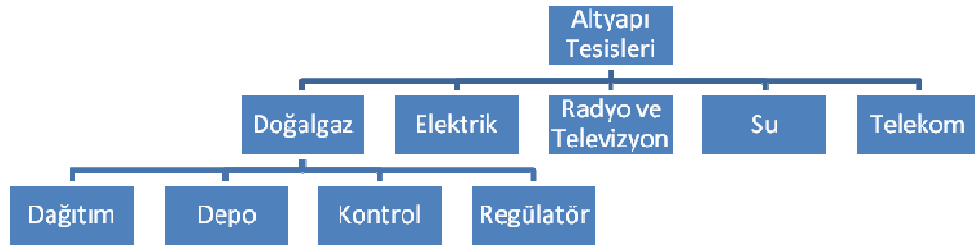
Çizelge 4. 2: Şişli Adres Bilgi Sistemi mahalle ve adres sayıları.

SIRA NO	MAHALLE ID	MAHALLE ADI	ADRES SAYISI
1	207	AYAZAĞA	5552
2	208	HUZUR	1371
3	210	MASLAK	3132
4	211	ESENTEPE	1664
5	212	GÜLBAHAR	2387
6	213	MECİDİYEKÖY	3505
7	214	KUŞTEPE	3423
8	215	İZZETPAŞA	1513
9	216	FULYA	2137
10	217	19 MAYIS	1715
11	218	MERKEZ	2663
12	292	HALİDE EDİP ADIVAR	1753
13	293	KAPTANPAŞA	249
14	294	HALİL RIFAT PAŞA	685
15	295	CUMHURİYET	1233
16	296	MEŞRUTİYET	2283
17	297	TEŞVİKİYE	1726
18	298	HALASKARGAZİ	782
19	299	HARBİYE	676
20	300	ERGENEKON	828
21	301	İNÖNÜ	1430
22	302	BOZKURT	1493
23	303	ESKİŞEHİR	1905
24	304	YAYLA	1578
25	305	FERİKÖY	1535
26	306	PAŞA	2655
27	307	MAHMUT ŞEVKET PAŞA	2982
28	308	DUATEPE	1095
TOPLAM			53950

4. bölümde de vurgulandığı gibi, ticaret alanında gelişmekte olan Maslak Mahallesi ve yerleşim bakımından Şişli'nin en eski mahallelerinden olan Mecidiyeköy için elde edilen sonuçlar normal olarak değerlendirilebilir.

4.1.2 Obje Grupları

Adresler tablosu hazırlanırken, adreslere ait kullanım türleri ve alt grupları bir hiyerarşi içerisinde ifade edilmektedir (Şekil 4.7). Grupların düzgün bir biçimde sunulabilmesi amacı ile Büyükşehir Belediyesi tarafından standart bir OBJE_KOD tablosu hazırlanmıştır.



Şekil 4. 7: İBB standart obje grupları örneği kavramsal gösterimi.

İstanbul Büyükşehir Belediyesi Adres Bilgi Sistemi kapsamında hazırlanan OBJE_KOD tablosunda tespit edilen 14 farklı temel obje grubu ve adres bilgileri bu obje grupları kullanılarak değerlendirilmektedir (Şekil 4.8). Tabloda alt grupları ile birlikte 505 adet kayıt bulunmaktadır.

OBJE_ANA_GF	OBJE_ANA_GF	OBJE_GRUP_H	OBJE_GRUP_A	OBJE_KOD	OBJE_ADI
1	Altyapı Tesisi	1	Doğalgaz	1	Doğalgaz Dağıt
1	Altyapı Tesisi	1	Doğalgaz	2	Doğalgaz Depo
1	Altyapı Tesisi	1	Doğalgaz	3	Doğalgaz Kontr
1	Altyapı Tesisi	1	Doğalgaz	4	Doğalgaz Regü
1	Altyapı Tesisi	2	Elektrik	5	Hidroelektrik S
1	Altyapı Tesisi	2	Elektrik	6	Termik Santral
1	Altyapı Tesisi	2	Elektrik	7	Trafo
1	Altyapı Tesisi	3	Radyo-televizy	8	Radyo-TVAnter
1	Altyapı Tesisi	4	Su	9	Su Deposu
1	Altyapı Tesisi	4	Su	10	Terfi Merkezi
1	Altyapı Tesisi	5	Telekom	11	Telefon Anten
1	Altyapı Tesisi	5	Telekom	12	GSM Anten
1	Altyapı Tesisi	5	Telekom	13	Telefon Santral
1	Altyapı Tesisi	5	Telekom	14	Telefon Kulübes
2	Dini Tesis	6	İbadethane	15	Cami
2	Dini Tesis	6	İbadethane	16	Havralar
2	Dini Tesis	6	İbadethane	17	Kilise
2	Dini Tesis	6	İbadethane	18	Manastır
2	Dini Tesis	6	İbadethane	19	Mescid
2	Dini Tesis	6	İbadethane	20	Patrikhane
2	Dini Tesis	6	İbadethane	21	Sinagog
2	Dini Tesis	7	Mezarlık	22	Hristiyan Meza
2	Dini Tesis	7	Mezarlık	23	Musevi Mezarlı
2	Dini Tesis	7	Mezarlık	24	Müslüman Mez
2	Dini Tesis	7	Mezarlık	25	Şehitlik
2	Dini Tesis	7	Mezarlık	26	Türbe
3	Eğitim Kurumu	8	Dershane	27	Dershane
3	Eğitim Kurumu	9	Halk Eğitim Me	28	Halk Eğitim Me
3	Eğitim Kurumu	10	İlköğretim	29	İlköğretim Devle
3	Eğitim Kurumu	10	İlköğretim	30	İlköğretim Özel
3	Eğitim Kurumu	10	İlköğretim	31	İlköğretim Vakıf
3	Eğitim Kurumu	11	Kuran Kursu	32	Kuran Kursu
3	Eğitim Kurumu	12	Lise	33	Lise Devlet
3	Eğitim Kurumu	12	Lise	34	Lise Özel

Şekil 4. 8: Adres Yönetim Şefliği, ilçe bazında sayısal ortamda Adres Bilgi Sistemi obje grupları örneği (2009).

Temel obje gruplarının sahip olduğu adreslerin sayısı (Çizelge 4.3) dikkatli bir biçimde incelendiğinde aşağıdaki sonuçlar ortaya çıkmaktadır.

Çizelge 4. 3: Şişli Adres Bilgi Sistemi kullanım türü ve adres sayısı.

KULLANIM TÜRÜ ID	KULLANIM TÜRÜ	ADRES SAYISI
1	ALTYAPI TESİSİ	187
2	DİNİ TESİS	124
3	EĞİTİM KURUMU	304
4	GENEL KULLANIM	32515
5	KÜLTÜREL TESİS	69
6	REHBER	1
7	RESMİ KURUM	156
8	SAĞLIK KURUMU	431
9	SANAYİ VE ÜRETİM ALANI	607
10	SOSYAL AMAÇLI KURUM	235
11	TİCARET ALANLARI	16143
12	TURİZM	97
13	ULAŞIM	903
14	YEŞİL ALAN, SPOR VE DİNLENME YERLERİ	87
15	TANIMSIZ	2092
TOPLAM		53951

Adres kullanım türlerine göre adres bilgileri değerlendirildiğinde genel kullanım amaçlı adreslerin çok daha baskın olduğu görülmektedir. Genel kullanım bilgileri içerisinde arsa, boş alan, tarla – bahçe, konut + küçük sanayi, konut + ticaret, çatı katı, harap bina, inşa halinde bina girmektedir. Kapı numaraları olmasına rağmen, İBB tarafından kullanım alanı belirlenmemiş 2092 adet adres bilgisi, yani obje bulunmaktadır. Çalışmada web sayfalarından adres kazanımlarının yorumlanmasında obje grupları ve bağlı bulundukları adres bilgileri önemli yer tutmaktadır.

Web sayfalarındaki mekânsal verilerin kullanım türlerinin belirlenebilmesi amacı ile uygulanan web madenciliği kapsamında öncelikle, terim sıklıklarının tespitinin yapılması gerekmektedir (Bkz 3.bölüm). Terim sıklıklarının saptanması amacı ile temel bir sözlük kullanılmalıdır. Genellikle web madenciliği çalışmalarında Miller (1995) yöneticiliği altında geliştirilen, kelimelerin türlerinin, kısa tanımlarının ve eş anlamlılarının sunulduğu WordNet veritabanı kullanılmaktadır. Çalışmada ise Adres Yönetim Şefliği tarafından oluşturulan standart obje gruplarından temel sözlük oluşturulmasında faydalanılmaktadır. Çalışma kapsamında daha önceden hazırlanmış herhangi bir sözlüğün kullanılmamasının nedeni, web sayfalarının İBB obje gruplarına göre kullanım türlerinin veya diğer bir deyişle fonksiyonlarının tespit edilmesinin amaçlanmasıdır. Bu nedenle obje grupları ayrı bir başlık altında incelenmiştir.

4.2 İşyeri Adres Verilerinin Kazanımı

Borges ve diğ., 2007 yılında yaptıkları çalışmada adresin sahip olduğu yapıyı ontolojik olarak sokak ismi gibi temel adres, bina numarası örneğinde olduğu gibi tamamlayıcı ve adresin posta kodu ve şehir ismine sahip olmasına göre tanımlayıcı olmak üzere 3 farklı kısımda incelemektedirler. AFFWeb’de adres yapılarının incelenmesi yerine web sayfalarının ait olabileceği obje gruplarının kavramsal olarak desteklenmesi yöntemi kullanılmıştır. Adreslere ait kavramsal bir yöntem uygulanamamasına neden olarak Türkiye’de adres bilgi sisteminin çok sağlıklı sonuçlar vermemesi ve yeni yeni bir sistem oluşturma çalışmalarının gündemde olması gösterilebilir. Belirli standartlar oluşturulmaya çalışılmasına rağmen özellikle web ortamında verilen adresler incelendiğinde standart bir yapının olmadığı kesin olarak görülmektedir.

Çizelge 4. 4: Web sayfalarında adres sunumu örnekleri.

Örnek No	Örnek Adres
1	Merkez Mah. Kocamansur Sok. No:62/ Şişli
2	Şişli Sağlık Grup Başkanlığı 2.Taşocağı cad. Bahçelievler Sok. No:2 Mecidiyeköy Şişli
3	Şişli Sağlık Merkezi Posta Adresi : 19 Mayıs Caddesi No:8 34360 Şişli-İSTANBUL

Web sayfalarında sunulan adreslerde belirli standardın kullanılmaması 3 basit örnek ile açıklanabilir (Çizelge 4.4) .

- 1. örnekte / işareti kullanılırken diğerlerinde böyle bir işarete yer verilmemiştir.
- 3. örnekte “Caddesi” diye açık olarak adres bilgisi verilmesine rağmen 2. adres örneğinde “cad.” kısaltması kullanılmıştır.
- 3. örnekte il bilgisi bulunmasına rağmen diğerlerinde yoktur. Verilen örneklerde de açıkça görüldüğü üzere adres bilgilerinin kavramsal olarak belirli bir hiyerarşi içerisinde sunulması şu an içerisinde bulunduğumuz şartlarda nerede ise olanaksızdır.

31 Temmuz 2006 tarihli 26245 nolu Resmi Gazetede yayınlanan ADRES VE NUMARALAMAYA İLİŞKİN YÖNETMELİK kapsamında;

(1) Adres standardı; il, ilçe, bucak, köy, mezra, belediye adı; mahalle adı ve sabit tanıtım numarası, meydan, bulvar, cadde, sokak ve küme evlerin adı ve sabit tanıtım numarası ile site, blok, mevki adı, dış kapı numarası, iç kapı numarası, posta kodu bileşenlerinden oluşur.

(2) Bu bileşenlerden il, ilçe adı, dış kapı numarası ve posta kodunun her adreste bulunması zorunludur.

Diğer taraftan Yeni Türk Ticaret Kanununun 1424. maddesine göre bütün anonim, limitet ve paylı komandit şirketlere bir internet sitesi açma zorunluluğu getirilmiştir. Bu maddeye göre web sayfalarına koymaları gereken bilgiler arasında örneğin,

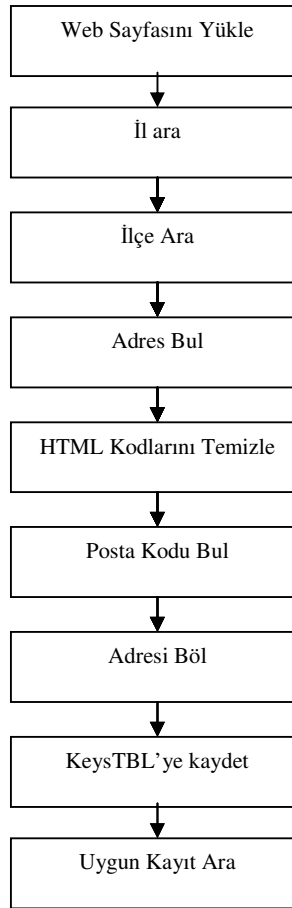
- a) Şirketçe kanunen yapılması gereken ilanlar,
- b) Pay sahipleri ve ortaklar açısından önem taşıyan açıklamalar,
- c) Yönetim ve müdürler kurulu ile genel kurul toplantıları hazırlıkları,
- d) Anılan kurulların yapılmasına ilişkin bilgiler,

e) Pay sahiplerine veya ortaklara sunulması gereken belgeler,

f) Genel kurullara ait olanlar dâhil her türlü çağrılar, bulunmasına karşın adres bilgilerinin deklare edilmesi ile ilgili herhangi bir uyarı ve sorumluluk bulunmamaktadır. Genellikle küçük veya büyük özel şirketler müşterilerine ulaşabilmek adına web sayfalarında adres bilgilerine yer vermektedirler. Bu da web sayfalarından adres kazanımı kapsamında karşılaşılan bir diğer probleme örnek olarak verilebilir.

Çalışma için geliştirilen bir sistem olan AFFWeb web sayfalarından adres verilerine ulaşarak bir veritabanına kayıt eden ve İBB veritabanı ile karşılaştıran özelliğe sahiptir. Bu özellikler daha sonraki bölümlerde detaylı olarak incelenmektedir.

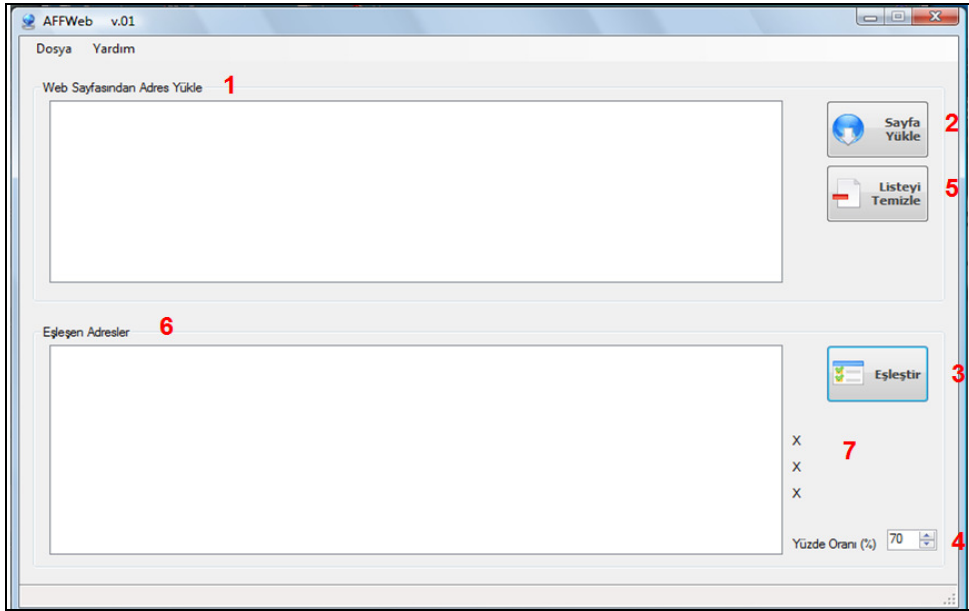
AFFWeb Web sayfalarından adres bilgilerine ulaşabilmek için geliştirilmiş bir yazılımdır (Şekil 4.9).



Şekil 4. 9: AFFWeb akış şeması.

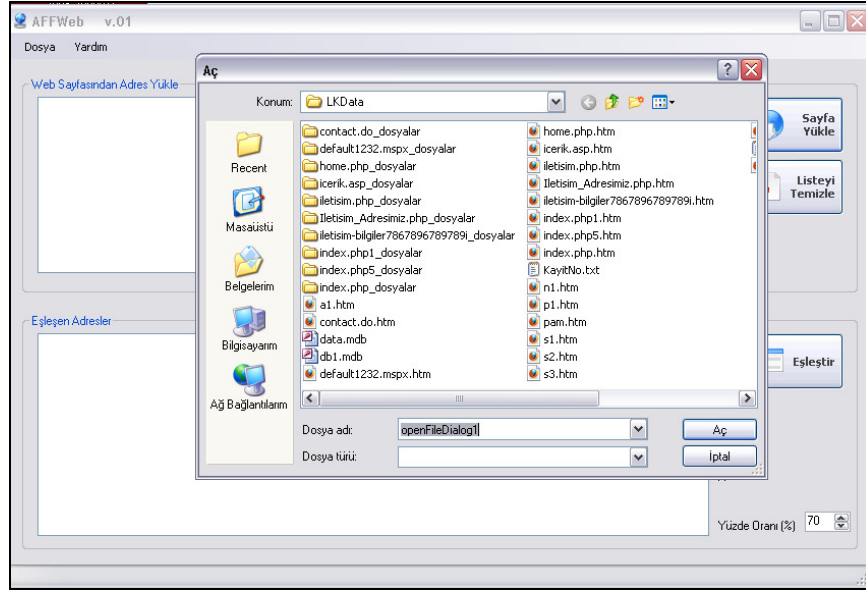
AFFWeb akış şemasına göre adres bul kısmında sistem il, eğer yoksa ilçe ya ait bölgenin sağında 50 karakter, solunda 250 karakteri web sayfasından alarak bir sonraki adım olan HTML kodlarından temizleme işlemi için adres verileri hazırlanmaktadır. Sağ tarafından 50 karakter alınmasının nedeni, il bilgisinden sonra gelebilecek olan ülke veya posta kodu bilgilerine de erişimin sağlanabilmesidir.

Adresi böl kısmında ise adres içerisinde bulunan mahalle, cadde sokak bilgilerine göre veritabanında kısaltmalar adı altında tutulan basit bir gazetteer olarak sunulabilecek tablo yardımı ile gerçekleştirilmektedir. Bu sayede adres bilgileri kelimeler halinde KeysTBL'ye atılarak veritabanında kayıt altına alınabilmektedirler. Uygun kayıt ara kısmında ise KeysTBL ve İBB adres tablosu kullanıcının belirleyebileceği “benzerlik oranı” ile karşılaştırılmaktadır.



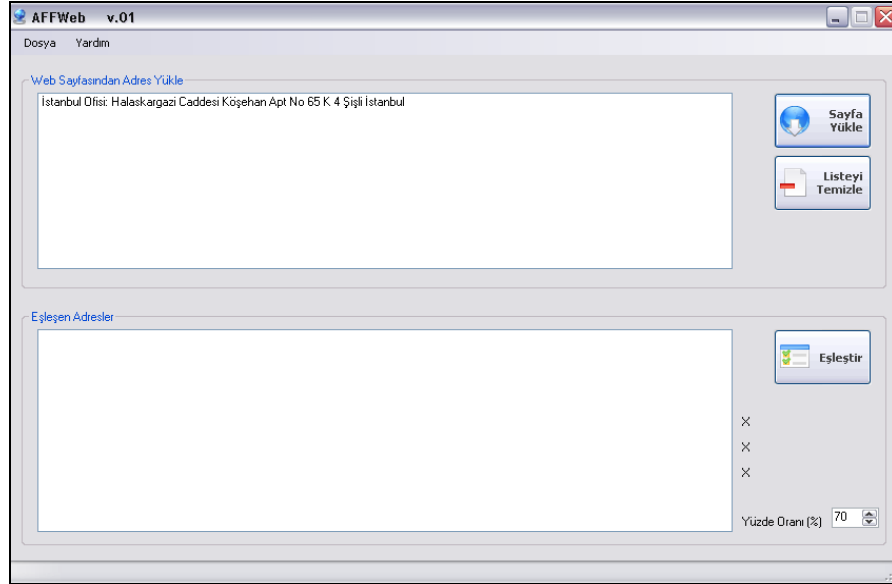
Şekil 4. 10: AFFWeb Adres karşılaştırma yazılımı arayüzü.

Adres karşılaştırma arayüzü (Şekil 4.10) kullanıcının verilere ulaşmadaki başlangıç ve bitişlerini gösterir zaman ve erişim sürecini gösterir bir sayaca sahiptir (7 numara). Kullanıcıya verilerin eşleşmesinde beklediği doğruluğu sağlayabilmek amacı ile yüzde oranı (4 numara) kısmı ilave edilmiştir. Bilgisayarda kayıt altına alınan web sayfalarında bulunan adres bilgileri 2 numara ile gösterilen “Adres Yükle” butonu yardımı ile 1 numaralı kısma HTML kodlarından ve diğer içeriklerden temizlenerek yüklenmektedir.



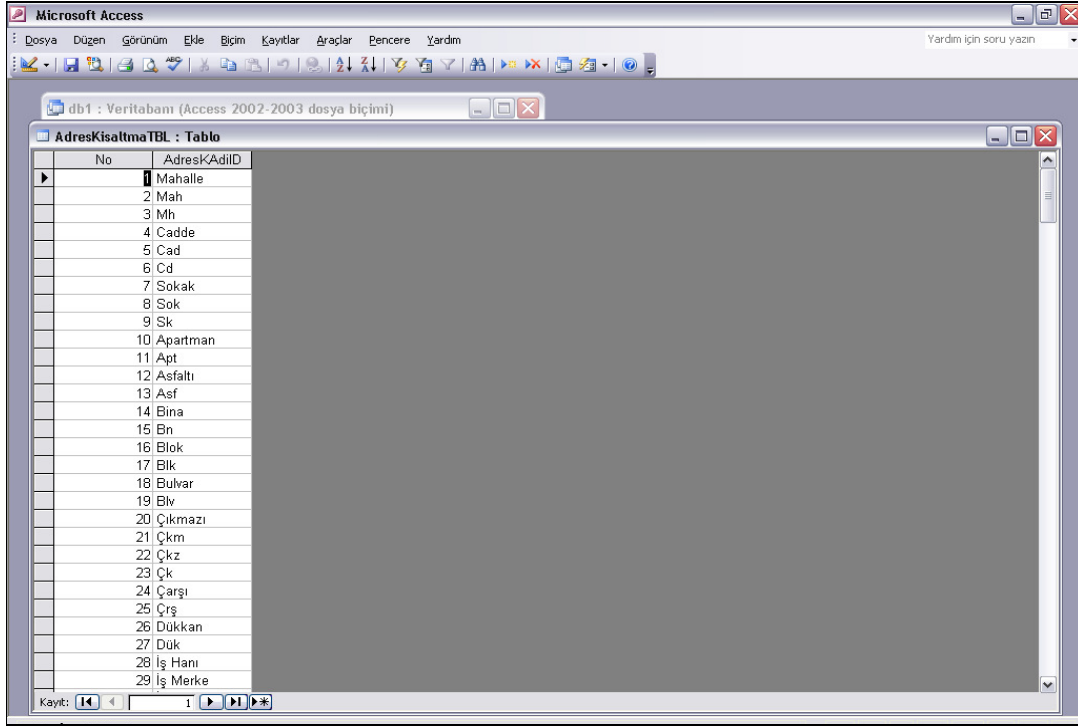
Şekil 4. 11: Dosya yükleme örneği.

Obje türlerine göre sınıflandırılmış olan web sayfalarının kaydedildiği kısımda kayıtlı web sayfalarına erişim imkânı bulunan yazılımda içerdiği adres bilgisine erişilmek istenilen web sayfası dosya yükleme kısmından sisteme alınmaktadır (Şekil 4.11).



Şekil 4. 12: Web sayfalarından adres yükleme.

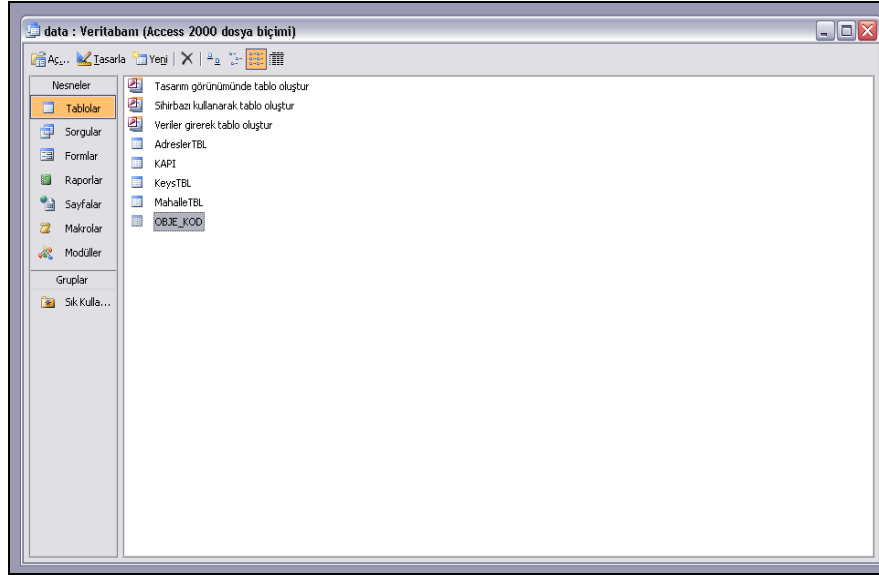
Seçilen web sayfasının içermiş olduğu adres bilgisi HTML kodlarından temizlenerek ekrana yüklenir. Adres temizleme işlemi sırasında akış şemasında sunulan işlem adımları takip edilir (Şekil 4.12).



Şekil 4. 13: Adres bileşenleri.

Adreslerin yüklenmesi esnasında daha önceden belirlenmiş küçük bir gazetteer yapısına sahip adreslerin içerebileceği cadde, cad, mah. vb. kısaltma veya tam bilgilerin olduğu standart bir adres yapısında bulunması gereken adres bileşenleri kullanılır. Resmi gazetede yayınlanan ilgili yönetmeliğin öngördüğü adres bileşenlerinin baz alınarak oluşturulduğu küçük gazetteer ilerleyen çalışmalarda ihtiyaçlar doğrultusunda çok daha fazla genişletilebilir. Adres bileşenleri (Şekil 4.13) sayesinde web sayfalarında bulunan adreslerin bulunduğu kısmın yazılım tarafından algılanabilmesi sağlanmaktadır.

Sisteme ait veritabanı görünümü (Şekil 4.14) te görülmektedir. Veritabanı içerisinde İBB ye ait obje kodları, mahalle bilgileri, adres bilgileri (Şekil 4.15) ve programın çalıştırılmasından sonra adres eşleşmelerinin kayıt altında tutulduğu KeysTBL tablosu bulunmaktadır.

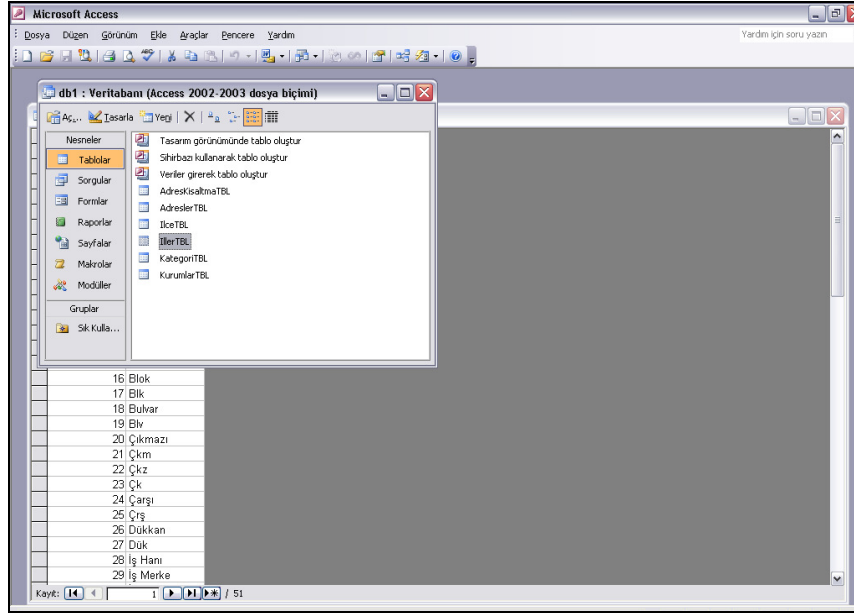


Şekil 4. 14: Belediye verileri ve bilinmeyen adres karşılaştırma veritabanı.

AFFWeb yazılımının veritabanı kısmında Adresler TBL olarak adlandırılan Şişli Adres Bilgi Sistemi verileri, obje kodlarının bulunduğu kısım, mahalle bilgileri tabloları bulunmaktadır. Bu bilgiler tamamen İBB Adres Bilgi Sistemi verilerinden oluşmaktadır. KeysTBL tablosu ise eşleştirme sonrası yüzde oranını ve adres numarasını tutan kısımdır.

OBJECTID	MahNoID	MahAdID	KAPI_NO	ESKİ_KAPI_NO	TAHSİS_KAPI_NO	ESKİ_YOL_AD	KAPI ADI	
1	214 KÜŞTEPE	14/3				Alev Sk.	Konut	Aİ
2	214 KÜŞTEPE	131/1				Yavrukus Sk.		Y
3	214 KÜŞTEPE	131/2				Yavrukus Sk.		Y
4	214 KÜŞTEPE	133/4				Yavrukus Sk.		Y
5	214 KÜŞTEPE	133/2				Yavrukus Sk.		Y
6	214 KÜŞTEPE	12				Doğanay Sk.		K
7	214 KÜŞTEPE	14				Doğanay Sk.		K
8	305 FERİKÖY	7	7			Aİ Karanfil Sk.		B
9	305 FERİKÖY	9	9			Aİ Karanfil Sk.		B
10	305 FERİKÖY	11	11			Aİ Karanfil Sk.		B
11	305 FERİKÖY	13	13			Aİ Karanfil Sk.	Yüksel Apartmanı	B
12	305 FERİKÖY	15	15			Aİ Karanfil Sk.	Balkancı Apartmanı	B
13	305 FERİKÖY	4	8			Aİ Karanfil Sk.	Şahinler Apartmanı	B
14	305 FERİKÖY	6	10			Aİ Karanfil Sk.		B
15	305 FERİKÖY	8	12			Aİ Karanfil Sk.	Mertaş Apartmanı	B
16	305 FERİKÖY	10	14			Aİ Karanfil Sk.	Emek Apartmanı	B
17	305 FERİKÖY	12	16			Aİ Karanfil Sk.	Ayaz Apartmanı	B
18	305 FERİKÖY	4	4			Açık Yol Sk.	Hidroğlu Apartmanı	A
19	305 FERİKÖY	6	6			Açık Yol Sk.	Yasemin Apartmanı	A
20	305 FERİKÖY	10	10			Açık Yol Sk.	Erdoğan Apartmanı	A
21	305 FERİKÖY	12	12			Açık Yol Sk.	Sözen Apartmanı	A
22	305 FERİKÖY	14	14			Açık Yol Sk.		A
23	305 FERİKÖY	16-18	16-18			Açık Yol Sk.	Çağlar Cafe	A
24	305 FERİKÖY	22	22			Açık Yol Sk.		A
25	305 FERİKÖY	24	24			Açık Yol Sk.		A
26	305 FERİKÖY	30A	30			Açık Yol Sk.	Arka Kullanılmayan Kapı	A
27	305 FERİKÖY	3	3			Açık Yol Sk.	Oto Tamir	A
28	305 FERİKÖY	15	17			Açık Yol Sk.		A
29	305 FERİKÖY	17	19			Açık Yol Sk.		A

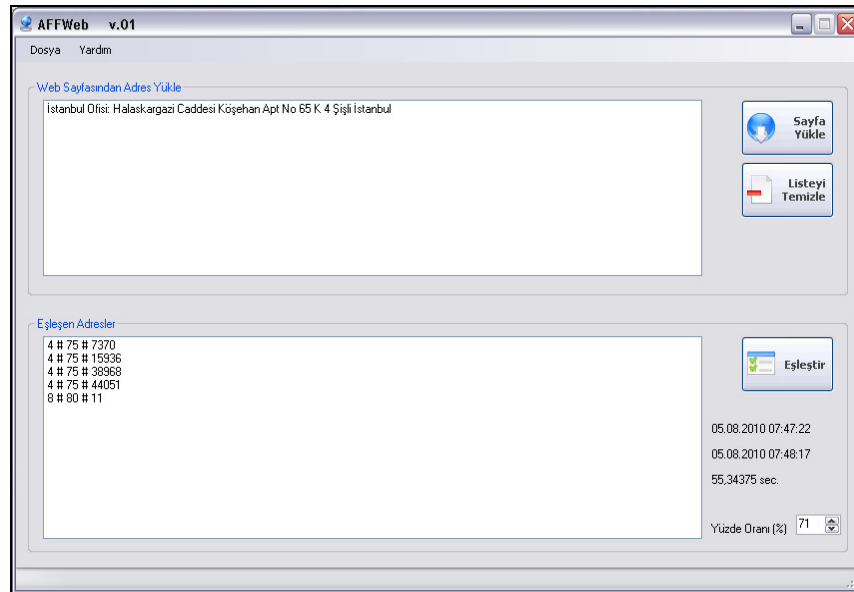
Şekil 4. 15: İBB-Şişli İlçesi adres verileri.



Şekil 4. 16: db1 veritabanı görünümü.

Diğer bir veritabanı ise adres bileşenlerinin, ileriye dönük olarak yazılımın genişletilebilmesini sağlayacak olan iller, ilçeler ve kurumlar tablolarının bulunduğu db1 (Şekil 4.16) dir.

4 numaralı kısım yardımı ile AffWeb ve İBB adreslerinin ne kadar uyumla eşleşeceğini belirlemek amacı ile gerekli yüzde ayarları yapıldıktan sonra 3 numaralı kısımda bulunan “Eşleştir” butonu yardımı ile kayıtlı adresler ve İBB adresleri karşılaştırılır.



Şekil 4. 17: Eşleştirme sonrası görünüm.

Örneğe göre web sayfasından bulunan 4 numaralı adres verisi %75 doğrulukla 44051 numaralı İBB verisi ile eşleşmiştir (Şekil 4.17).

KeyID	ObjekKodID	YuzdeOranID	k1	k2	k3	k4	k5	k6	k7
1			Axa	Oyak	Sanat	Galerisi	Şişli	#	#
2			Eylül	Sanat	Galerisi	Şişli	#	#	#
3			Milli	Reasürans	Sanat	Galerisi	Tesvikiye	#	#
4	44051	75	Abide	Hürriyet	Istasyon	183	#	#	#
5			Ayazağa	Esen	Spor	Klübü	Adres	Atatürk	Kurt
6			Feniköy	Spor	Klübü	Adres	Feniköy	Barutçular	Sokak
7			Mecidiyeköy	Esentepe	Caddesi	Özlem	Pasajı	Mecidiyeköy	9
8	11	80	FERİKÖY	Al Karanfil	Yüksel	Bayır	Konut	#	#
23	0	0	İZZETPAŞA	AYAZMA	ŞİŞLİ	İSTANBUL	#	#	#
24	0	0		HALAARGAZIC	ŞİŞLİ	İSTANBUL	#	#	#

Şekil 4. 18: Veritabanında eşleştirme sonrası görünüm.

Eşleştirme sonrası ayrı bir kayıt dosyasında tutulan sonuçlar yukarıda görüldüğü gibidir (Şekil 4.18).

4.3 İşyeri Kullanım Verilerinin Belirlenmesi

Belediyelerin kent bilgi sistemleri başta olmak üzere hâlihazırda kullandıkları veri setlerini zenginleştirmek ve güncellemek amacı ile web sayfalarından,

- işyerleri,
- kamu kurumları,
- kültür ve spor merkezleri,
- dinlenme alanları, lokantalar,
- eğlence mekanları vb.

yerlere ait adres bilgileri ve bu mekanlara ilişkin kullanım türleri tespit edilebilir. Bu amaçla web madenciliği yöntemi ile web sayfalarının kullanım türleri belirlenebilir, düzenli ifadeler (regular expression) yardımı ile Wang ve diğ.(2005) elde edilen adres bilgileri ile var olan adres bilgi sistemlerinde güncelleme ve eksikliklerin giderilmesi çalışmaları yapılabilir. Çalışma bağlamında adres bilgilerinin kazanılması amacıyla AFFWeb (Address Find From Web) uygulaması gerçekleştirilmiştir. Kayıt altına alınan web sayfaları kullanım türlerinin tespiti amacı ile sınıflandırma amaçlı genelleştirme işleminin ardından, adres bilgilerinin tespiti ve kayıt altına alınması işlemlerinden faydalanılmaktadır. Elde edilen adres bilgileri

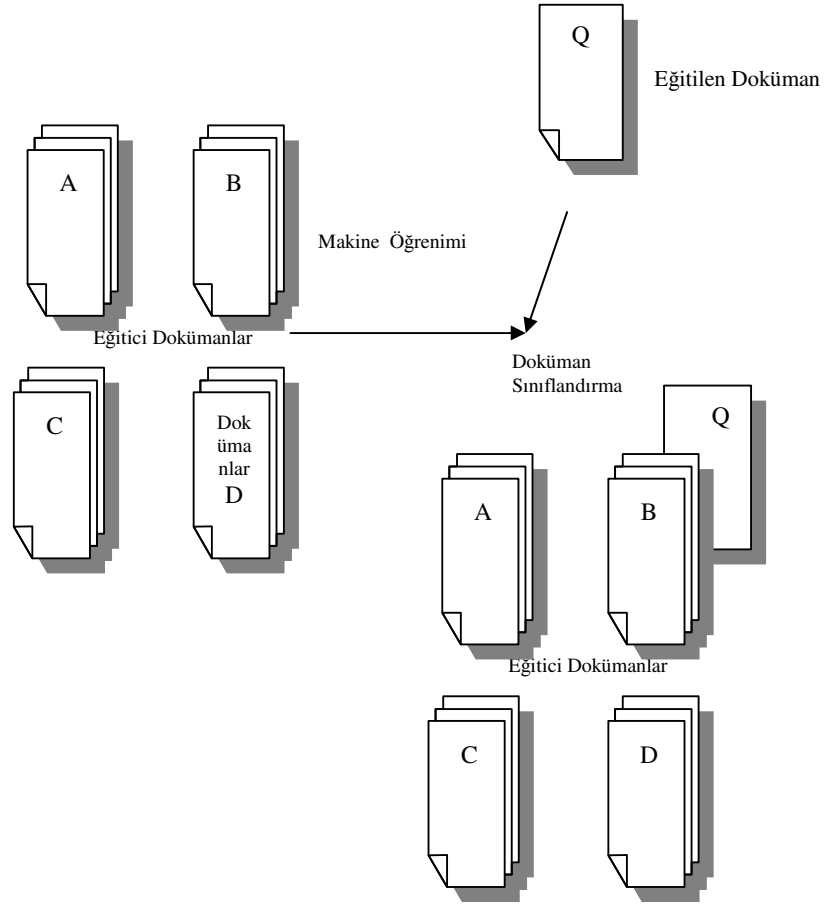
- Kayıt altına alınan web sayfalarının kullanım türlerinin tespit edilmesi
- Kullanım türleri tespit edilen web sayfalarından adreslerin kazanılması
- Kazanılan adreslerin Adres Bilgi Sistemi ile karşılaştırılması



60

yapılmıştır. Şişli İlçesi adres verilerinin hizmet alımı suretiyle üretildiği ve tamamen doğru olduklar varsayılmıştır.

Çalışmada web sayfalarının kullanım türlerinin belirlenebilmesi amacıyla kullanılan web madenciliğinde yöntem olarak sınıflandırma seçilmiştir. Yapılandırılmış verilerde genel sınıflandırma işleminde eğitici ve eğitilen olmak üzere iki farklı veri seti kullanılmaktadır. Doküman sınıflandırma işleminde de bu süreç benzer şekilde işlemektedir. Sınıf isimleri A, B, C, D olan dokümanlar arasına hangi sınıfa dâhil olacağı bilinmeyen Q dokümanı koyulduğu zaman makine öğrenim yöntemleri kullanılarak ilgili sınıfının tespiti yapılabilmektedir (Şekil 4.19).



Şekil 4. 20: Genel doküman sınıflandırma şeması.

Web sayfalarının kullanım türlerinin tespit edilebilmesi amacı ile genel sınıflandırma işlemleri temel alınmaktadır. Eğitici veri seti adı altında obje grubu belirli örnek web sayfaları kullanılacaktır. Eğitilen veri seti kapsamında ise, dâhil olduğu obje grubu belirli olmayan web sayfaları kullanılmaktadır. Hem eğitici hem de eğitilen veriler

özellikleri ve seçilirken dikkat edilmesi gereken durumlar geniş olarak incelenmektedir.

Önceden hangi obje grubuna (sınıf) ait olduğu bilinmeyen web sayfaları, eğitici web sayfaları ile karşılaştırarak ilgili olduğu grup tespit edilebilir. Bu yöntem sayesinde web sayfalarının içeriğinde bulunan adres bilgileri de ilgili obje grubuna dâhil olmaktadır.

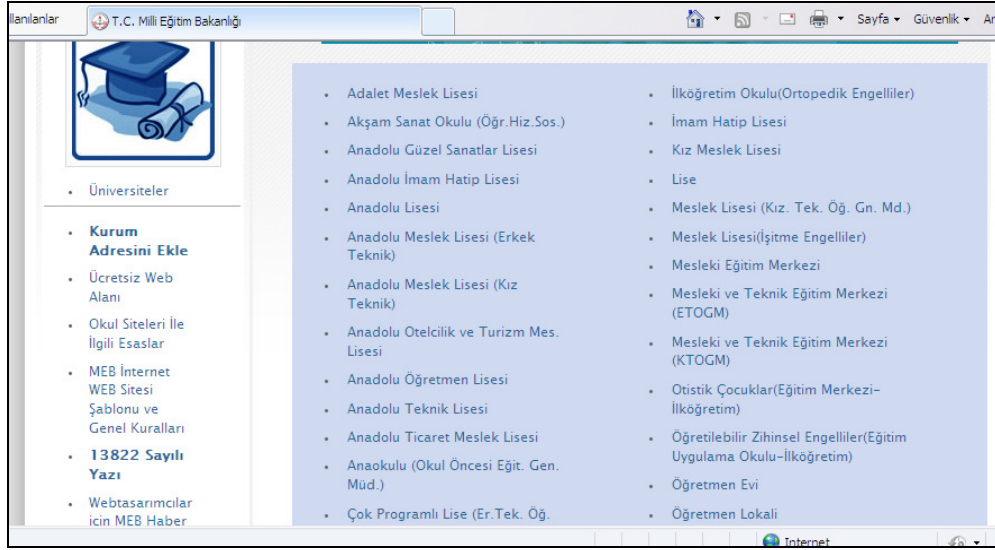
4.3.1 Eğitici web sayfaları

Çalışmada eğitim, sağlık, ticaret, turizm, spor ve kültür olmak üzere 6 farklı obje grubu incelenmektedir. Her bir obje grubu için, eğitici veri grubu olarak 4 farklı web sayfası kullanılmıştır. Bu web sayfaları örnek veri seti olarak adlandırılmaktadır. Sınıflandırma işleminin başarısı, önemli derecede eğitici amacı ile seçilen örnek veri setlerine bağlıdır. Her bir grupta incelemeye alınan web sayfaları aşağıda gösterilmiştir. Homojenliği sağlamak amacıyla her bir grupta sayfaları oluşturan kurumların aynı yapıda olmasına özen gösterilmiştir.

Eğitim: Eğitime ait örnek dokümanlar seçilirken içerisinde olabildiğince çok eğitim ile ilgili kelimelerin geçmesine dikkat edilmelidir. Bu zorunluluk dikkate alındığı zaman seçilen web sayfaları arasında, Vikipedi (Şekil 4.21), MEB web sitesi, özel eğitim kurumlarına ait olan web sayfaları bulunmaktadır.



Şekil 4. 21: Vikipedi eğitim sitesi örnek dokümanı.



Şekil 4. 22: MEB sitesi örnek dokümanı.

Vikipedide yer alan eğitim dokümanı eğitim ile ilgili olabilecek birçok kelimeyi içerisinde barındırmaktadır (Şekil 4.22). Web sayfası seçiminde daha önceden belirlenmiş olan sözlük ile olabildiğince uyuşmasına dikkat edilmelidir. Eğitim sayfalarının seçimi daha sonra yapılacak olan sınıflandırma işleminin başarısını ciddi oranda etkilemektedir.

Sağlık: Vikipedi, Sağlık Bakanlığı, sağlık ve hastanelere ait web siteleri.

Ticaret: İstanbul Ticaret Odası, Mefruşat, A.Ş. web sayfaları ve vikipedi.

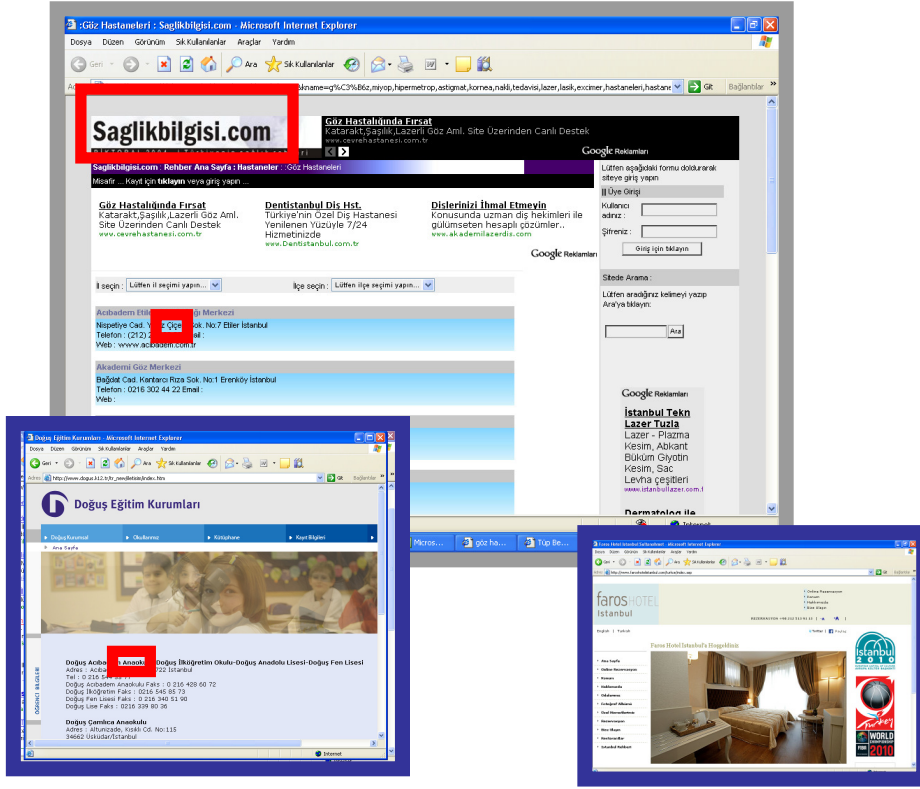
Spor: Spor Bakanlığı, spor ile ilgili diğer siteler ve vikipedi.

Kültür: Kültür Bakanlığı sitesi, vikipedi.

Gönüllü olarak dünyanın her yerinden insanların katılımı ile geliştirilen büyük bir ansiklopedi olan Wikipedia sayfalarından günümüzde web madenciliği ve arama motorları çalışmalarında faydalanılmaktadır. Bu nedenle Wikipedia sitesinin Türkiye temsilcisi olan Vikipedi çalışmada eğitici web sayfası kapsamında yoğun bir biçimde kullanılmaktadır.

4.3.2 Eğitilen web sayfaları

Her sayfada 1 veya 1 den fazla adres bilgisinin olabileceği düşüncesi ile bu kayıtların dışında olmak üzere 33 adet web sayfasına ait sonuçlar da incelemeye alınmıştır.



Şekil 4. 23: Sınıfları belli olmayan web sayfalarına örnekler.

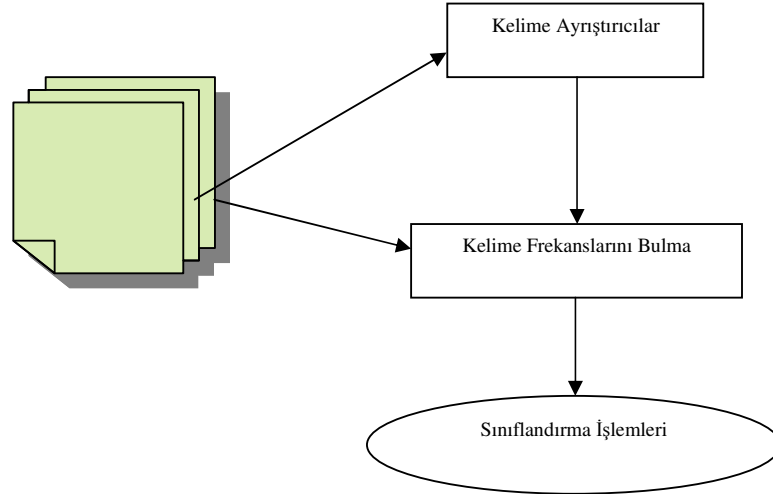
Örnek olarak seçilen Doğu Eğitim Kurumları ve Faros Hotel'e ait web sayfalarında (Şekil 4.23) en az bir veya birden fazla adres bilgisi olabileceği düşüncesinden hareketle kullanım türü tespit edilebilen bir web sayfasında yer alan adres bilgilerinin de kullanım türlerinin tespit edilebileceği yaklaşımı ortaya çıkmaktadır. Sistemin işlemesi durumunda Adres Bilgi Sistemlerinin desteklenmesi çalışmalarında sağlayacağı güncellik ve adres bilgilerinin kontrolü gibi yararlar göz ardı edilemez.

Manuel olarak web sayfalarının dâhil oldukları obje gruplarının belirlenmesi yerine otomatik olarak dâhil olacakları obje gruplarının tespit edilebilmesi yapılacak çalışmaları hızlandıracaktır. Özellikle web sayfalarının sayıları 10-20 gibi küçük rakamlarla kalmayıp haber arama motorlarında (Şekil 4.24) olduğu gibi günlük hatta dakikalık olarak 100 den hatta 1000 den fazla sayfaların sunulması çalışmalarında olduğu gibi işlemleri kolaylaştıracaktır.



Şekil 4. 24:Haber takip sitesi (Url5).

Web kullanım türlerinin tespit edilebilmesi web madencisinde sınıflandırma işlem ile desteklenebilir. Fakat web sayfalarının yapılandırılmamış özelliğinden ve sayısal bilgi değil de kelime bilgileri içermeleri sınıflandırma işleminin doğrudan yapılmasının önüne geçmektedir. Web sayfalarını sınıflandırmaya hazırlayabilmek amacı ile önışlemden geçirmek gerekmektedir.



Şekil 4. 25: Kelime frekanslarının tespit edilmesi (Yıldız ve diğ. 2007).

Web sayfalarındaki kelime frekanslarının tespit edilmesi önışlem adımlarından birisidir. Frekans işlemlerinin ardından web sayfaları sınıflandırma işlemi için hazır hale gelmektedir (Şekil 4.25). Kelime sıklıklarının diğer bir değişle frekanslarının tespiti web sayfalarının sayısal bir ortamda sunulabilmesine olanak sağlamaktadır.

Önişlem adımlarından diğerleri ise web sayfalarından HTML kodlarının atılması, sık sık tekrarlanan ve, ile vb. gibi bağlaçların atılmasını sağlamaktır.

Microsoft Excel - belediyetiblenler															
Dosya Düzen Görünüm Ekle Bçim Araçlar Yeri Pencere Yardım															
B006 =TOPLAM(B\$1:B800)															
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	
1	Eğitim	Aydın	at	(KKTG)	Eğitimi	A.O.L	2005-2010	HAKKINDA	bilgiye	adresiniz		2009-2010	EksenEğit	RAM	Buyc
2	2-A	Aydın	Avrupa		A.L	Ahlak	aradığınız	internete	Diyarbakır	Bekleyen	Lisesi	Kayseri	Elitshare	Öğretiyoru	C
3	A	Aydınsarı	B		Adresler	ANASAYF	B	kolay	Doğu	güncel	M Koyuncu	Kayseri	e-mail ben	Öğretmeni	Cans
4	Aal	Ayrancıpl	Bahçesi	85 Yıl	Anadolu	AOL	daha	Onay	Doğu	için	Manas	Kirgizistan	EmineÖm	Öğretmeni	CeBi
5	Abant	Ayşe	Balcı	adına	Anadolu	Aydinus	Eğitimde	SITE	Dokuz	kaybetmec	Matematik	Kırık	Emirgazim	Öğretmeni	Celal
6	Acıpayam	Ayvakti	Bankası	Ahlak	Anaokulu	Bankası	I.O.	yayın	Dosyabak	Site	Matematik	Kırıkkale	Engelli	Öğretmeni	Coğr
7	Açık	Babaeskie	Başan	Ahmet	Anlatımları	egitimsiter	I.O.	Yeri	Dumlupınar	Matematik	Kırıkkale	Engelliler	Öğretmenlersitesi		Coğr
8	Adamsiti	Babaro	Başan	Akdeniz	AOL	Eğitim	I.O.	Yöcel	Dumlu	Mays	Kırıkkale	e-okul	Öğretmenet		Coğr
9	Adana	Bahçeşehi	Başan	ALES	Argivi	Eğitimi	I.O.	Düz	Dursunbey	Mekani	Kimyaders	E-okul	Öğretmensayfası		Cum
10	Adıyaman	Bakırköy	Bayar	Ali	Bayşal	Eğit	I.O.O	ECA	Mekanları	Kişisel	ErdekMYC	Öğretmeni	C	Net	Foru
11	Adilcevaz	Balkesir	Bg	Amerikan	Beden	Forum	içerikli	Edebi	Menderes	Kişisel	Erdoğan	OSYM	Çağ	net	Gaffa
12	Adnan	Basketbol	Bg	Anadolu	Bilgi	I.O	İHL	Edebiyat	Mevsimi	Kocaeli	Etkineğitir	Öykü	Çan	Net	Gen
13	Afacançoc	Basketbol	Bilgi	Anadolu	Cumhuriyet	I.O	İlan	Edebiyat	İ Milli	Koç	Etkinlik	Özel	Çankaya	Net	Gen
14	Afyon	Başanm	Bilgi	Anadolu	Cumhuriyet	I.O	İlçe	Edebiyat	İMM	Kolayfizik	Etkintasarı	Özel	Çardak	Net	Gen
15	Afyon	BaşanYolu	Bilgi	Anadolu	Cumhuriyet	I.O	İlgili	Edebiyat	İ Müzik	KolayTES	Excel-Wor	Özel	Çardak	Net	Gen
16	Ahi	Başkent	Bilgi	Anadolu	Çok	I.O	İlkoğretim	Edebiyat	Necati	Kolej	Eyuboglu	Özel	Çardak	Net	Gezri
17	Ahiatıpl	Bayfenci	Bilgi	Anadolu	CPL	I.O	İmam	E-egitim	Nolu	Kolejpark	Eyüp	Özel	Cay	Net	Gov
18	Ahmet	Beden	Bir-Sen	Anaokulu	Dairesi	I.O	İMKB	Ege	İnolun	Kozlukmei	Fatih	Özel	Çiftlik	Net	Gölb
19	Ahmet	Beden	buraya	Anaokulu	Danışmanlık	I.O	İngilizce	EgeLisesi	Notu	Köksal	Fatih	Özel	Çanlı	Net	Günç
20	AhmetAy	Bendeokul	Cd	Anaokulu	Demirel	I.O	İngilizce	Eğitim-Site	O.	Körfez	Fatsalman	Özel	Çizgi	Net	Güve
21	Akademik	BenimYuv	Cd	and	Dergileri	I.O	İst.	Eğitim	Ofisi	Kpss	Fen	Özel	ÇocukDün	net	Habe
22	Akdeniz	Beyfen	Cevahir	and	Ders	I.O.O	İstanbul	Eğitim	Okul	KPSS	Fen	Özel	Çocuklara	Net	Habe
23	Akdenizli	Beykent	Cumhuriye	Ansikiöpec	Ders	I.O.O	İstiklal	Eğitim	Okul	KPSS	Fen	Özel	Çok	Net	Halk
24	Aksaray	Biga	Cumhuriye	Ansikiöpec	Ders	I.O	İzle	Eğitim	Okul	Kpss	Fenbilimle	ÖzelDersE	Çukurova	Net	Halk
25	Al	Biga	De	AOL	ders	İHL	İzzet	Eğitim	Okulu	Kurs	Fenokulu6	Özgün	Dargeçit	Net	Halk
26	Ali	Bilecik	Dr	Araçları	Ders	ile	K Paşa	Eğitim	Okulu	kurulumus	FenveTekn	Özgür	Dagerli	Net	Has
27	Ali	BilgiEğitim	Edu	Ağartır	Ders	İLETİŞİM	Kağıthane	Eğitim	Onsekiz	Kültür	Ferit	Özlem	Demethoc	Org	Hazır
28	Alibeyköy	Bilgimce	Edu	Ağartır	Ders	İlgili	kaliteli	Eğitim	Osmangaz	Kültür	Firat	Phanokul	Özmen	Org	HEM
29	Alisko	BilgininEfe	Eğitim	Dershanesi	İmam	Karalmas	Eğitim		Osmangaz	Kültür-San	Fizik	Phanokul	Özmen	Org	Hizir
30	Almanyadı	Bilgisayar	eğitim	Eğitimi	Dershaneleri	İO	Karne	Eğitim	Öğren	Kürşat	Fizikciyiz	Pekiye	Ders	Org	Hizir
31	Alperakipe	Bilgisayar	Eğitim	engelli	Dershaneleri	İzleme	Kavram	Eğitim	Öğretim	Kütahya	FloryaKole	Pinokyo	Ders	Org	HOC
32	Anadolu	Bilgisayar	Eğitim	Enstitüsü	Dershaneleri	Kazak	Kaynak	Eğitim	Öğretmen	Kütüphane	forumbaku	PratikSözl	Ders	Org	Hoca
33	Anadolu	Bilgiyurt	Eğitim	Ergil	Dershanesi	köy	Kemal	Eğitim	Öğretmeni	Leenglish	Galatasara	Prof	Ders	Org	İman
34	Anadolu	Bilimge	Eğitim	Ergen	Dershanesi	kursları	Kübn	Eğitim	Öğretmeni	Leenglish	Galatasara	Prof	Ders	Org	İman
35	Anadolu	Bilimge	Eğitim	Fakültesi	Dershanesi	Kursu	Kübn	Eğitim	Öğretmeni	Leenglish	Galatasara	Prof	Ders	Org	İman

Şekil 4. 26: Eğitici web sayfalarının gösterimi.

Eğitici veri seti olarak seçilen 24 adet web sayfası içerdikleri HTML kodlarından ayrılarak düzenlenmektedir. Düzenlenen web sayfaları MS Excel’de tutulmaktadır. Çalışmanın boyutu için MS Excel, tezde yeterli desteği sağlamaktadır fakat elde edilen sonuçlar incelendiğinde Visual Studio. Net programlama dilinde yazılacak olan bir ara yüzle elde edilecek kelime frekansları sonuçlarının, Öklid ve Kosinüs hesaplama sonuçlarının kolaylıkla veritabanlarına entegre edilebilecek özelliğe sahip olduğu görülmektedir. Excel’de e1, e2, e3 ve e4 sayfaları eğitici olarak seçilmiş MEB, Wikipedi vb. sayfaları temsil etmektedir (Şekil 4.26).

4.4 Web sayfalarında önişleme

Excel sayfalarında kelimeler halinde sunulan web sayfaları içerdikleri kelimelere ait frekanslarının bulunabilmesi için sözlük ile karşılaştırılmaktadır. Çalışmanın amacına uygun olarak 2 farklı sözlük kullanılmaktadır (Şekil 4.27).

Belediye Obje Grupları Sözlüğü			
Acenta	Kütüphane	Eşya	Prefabrik
Acil	Laboratuvar	Ev	Pvc
Aile	Lamba	Evcil	Radio
Ajansları	Lastik	Factoring	Rehabilitasyo
Akdeniz	Leasing	Faks	Reklam
Aktar	Lise	Fast	Rekreasyon
Alanları	Lokanta	Fayansçı	Resmi
Aletler	Lostra	Film	Restoran
Aletleri	Lunapark	Finans	Saatçı
Alışveriş	Maddeler	Firmaları	Sabun
Alüminyum	Maddeleri	Food	Sağlayıcıları
Ambar	Mağaza	Fotoğraf	Sağlığı
Ambulans	Makineleri	Fotoğrafçı	Sağlık
Ana	Mal	Fotokopi	Saha
Anaokulu	Mali	Fotokopici	Sahası
Ant	Malzemeleri	Fuar	Salonu
Antikacı	Mamül	Galeri	Sanat
Antik kent	Manav	Galerileri	Sanatoryum
Antrepolar	Manifaturacı	Gazeteler	Sanayi
Aracı	Marangoz	Gazino	Sandığı
Araştırma	Marina	Gece	Santral
Armatörler	Market	Genel	Saray
Armatörücü	Market	Giyim	Satıcıları
ATM	Marketi	Görüntüleme	Satis
Av	Matbaa	Gözlükçü	Satış
Avize	Medrese	Grafikerler	Satış
Avukatlar	Medya	Gümrükçüler	Sauna
AYYM	Mekanları	Günöbrlik	Savunma

ONTOLOJİK OLARAK DESTEKLİ BELEDİYE SÖZLÜĞÜ				
Vazo	Lezzet	Etüt	Değerli	Bandaj
Adıvar	Lider	Farsça	Demografik	Baskül
Adli	Liquidamber	Fem	Demokratik	Basur
Adsl	Logos	Fen	Divan	Bere
Afet	Lokum	Fizik	Döşeme	Besin
Agop	Lonca	Fmv	Ebru	Beslenme
Ağa	Loryma	Fonksiyon	Efsane	Bevliye
Alev	Lostra	Forma	Egemen	Biyokimya
Amir	Loto	Fotokopi	Eminönü	Biyolojik
Amos	Lotus	Fotokopya	Ender	Boğmaca
Andak	Madalyon	Genç	Ermeni	Yutak
Antifriz	Mahkeme	Geocentre	Eser	Zatürree
Antre	Makam	Giriş	Etkileyici	Zayıflama
Apollon	Makedon	Gönye	Etnik	Zona
An	Maksillofasiyal	Grafik	Evrensel	Zührevi
Arma	Makyaj	Grafiker	Evrim	Aşı
Bağkur	Mankenlik	Gramer	Fakirhanesi	Bakıcı
Asker	Köşe	Harita	Fashion	Atak
Zemin	Kristal	Hedef	Flamenkçe'de	Ateş
Zincirlikuyu	Kriz	Hamlı	Gelenek	Baş
Cenaze	Krmkd	Hukuk	Gladatör	Bebek
Atık	Kroki	İngilizce	Göktürk	Biçak
Bagaj	Kuaför	Işık	Gravür	Bisturi
Bey	Kulube	İçmimar	Grek	Boğur
Blot	Kum	İlkoğretim	Hamam	Bronş
Boşaltma	Kundura	İmlem	Hamamlar	Bulantı
Böcek	Kuşam	İöo	Han	Bulaşıcı
Börülce	Kutlaması	İstatistik	Harika	Buluğ

Şekil 4. 27: Belediye ve ontolojik olarak desteklenmiş belediye sözlüğü.

Belediye sözlüğü ve ontolojik olarak desteklenen belediye sözlüğü ile elde edilen sonuçlar ayrı ayrı değerlendirilmektedir. Kelime sıklıklarının tespiti işlemi (Şekil 4.28) 24 adet eğitici web sayfası ve 33 adet eğitilen web sayfası için ayrı ayrı yapılmaktadır.

Kelimeler	Frekans
Abone	0
acenta	0
Acentelerl	0
Acil	0
Açık	1
Adli	0
Adliye	0
Ağaç	0
Ahır	0
ahşap	0
Aile	0
Ait	0
ajanları	0
Akarsu	0
Akbil	0
Akdeniz	3
aksesuar	0
Aktar	0
Alan	0
Alanı	0
Alanı	0
Alanları	0
aletler	0
Aletleri	0
Alışveriş	0
Altyapı	0
Alüminyum	0
Amaçlı	0
Ambar	0
Ambulans	0
Amirliği	0
Ana	20
Anaokulu	5

Şekil 4. 28: Eğitim ile ilgili 2 numaralı web sayfasına ait frekans bilgileri.

Sözlüğün boyutunu ifade eden ve içerdği kelime sayısı 446 olan belediye sözlüğü yardımı ile eğitici ve eğitilen matrisler oluşturulmaktadır. Bu durumda eğitici setlerde matris 446X24 ile eğitilen sette 446X1 olarak ifade edilmektedir.

Elde edilen sonuçlara göre sınıflandırma ile doğrudan ilgisi olmayan fakat yapılan frekans işleminin sonuçlarının daha kolay anlaşılmasını sağlayabilecek bazı bulgular aşağıda sunulmaktadır.

Ontolojik olarak desteklenmiş belediye sözlüğü yardımı ile oluşturulan diğer bir sette ise matris 1747X24 ile ifade edilmektedir.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1		eğ1		eğ2		eğ3		eğ4		sa1		sa2		sa3		sa4		ti1		ti2	
2	eğitim	150	eğitim	59	eğitim	103	eğitim	54	sağlı	28	sağlı	29	hasta	21	çocuk	15	ana	46	tic	89	tic
3	ç	19	ders	33	top	37	yaz	37	hasta	20	ç	17	kan	20	sağlı	14	bilgi	34	mal	29	elektronik
4	geliş	16	ise	26	baş	30	dil	32	tip	9	diş	12	sağlı	19	bakan	9	ç	28	baş	18	internet
5	top	16	site	26	gen	25	yaşam	28	kan	8	bitki	9	diş	13	hasta	9	veri	27	iş	18	iş
6	yapı	14	ana	20	sistem	24	üniversite	18	cerrahi	7	kan	8	göz	13	ç	9	elektronik	21	ç	15	bilgi
7	birey	12	okul	14	insan	23	kur	13	yardım	6	hasta	7	kalp	12	soğuk	8	iş	20	kur	14	dünya
8	değerle	12	özel	14	yer	23	doktor	12	çocuk	5	kalp	7	kar	11	tür	8	kur	20	pazar	13	yapı
9	kavram	12	bilgi	13	ç	19	sertifika	12	ç	5	uzman	7	gen	9	türk	8	tic	20	yapı	13	pazar
10	bilim	11	fen	13	iş	19	okul	10	kalp	5	beslenme	6	kadın	9	kan	6	eleman	14	geliş	11	aras
11	kur	11	öğretmen	13	kur	19	tür	10	bilim	4	iş	6	bitki	7	kar	6	kişi	14	av	10	ç
12	tür	11	ingiliz	11	aras	16	türk	10	damar	4	anne	5	geliş	7	kur	6	yaz	14	yol	10	kur
13	türk	11	ödev	10	genel	16	alman	9	değerli	4	baş	5	bey	6	site	6	iletişim	12	aras	9	tör
14	faaliyet	10	gen	9	okul	16	ülke	9	diş	4	çocuk	5	bilgi	6	uzman	6	sistem	10	üret	9	mal
15	gen	10	kur	9	yaşam	16	aile	7	enfeksiyon	4	çözüm	5	ç	6	bey	5	üret	10	alış	8	şirket
16	madde	9	matematik	9	kar	15	av	7	göz	4	haber	5	kişi	6	depresyon	5	yapı	10	anlam	8	üret
17	okul	9	rehber	8	bilgi	14	gen	7	üye	4	kar	5	lazer	6	gebe	5	güven	9	ekonomi	8	ürün
18	baş	8	sınıf	8	geliş	14	genel	7	ana	3	saç	5	tedavi	6	haber	5	internet	8	tarih	8	kar
19	genel	8	tür	7	öğretmen	13	sınav	7	baş	3	tedavi	5	vitamin	6	hastane	5	kar	8	tüccar	8	model
20	öğrenci	8	türk	7	yapı	13	yurt	7	biyokimya	3	bebek	4	anne	5	tedavi	5	aras	7	uygar	8	müşteri
21	anlam	7	baş	6	tür	12	han	6	burun	3	doğu	4	baş	5	teşhis	5	firma	7	yer	8	özel
22	bilgi	7	cumhuriye	6	ülke	12	kar	5	gen	3	erkek	4	çap	5	trafik	5	sayısal	7	bölge	7	tüketici
23	tarih	7	din	6	yaz	12	üye	5	giriş	3	güzel	4	özel	5	yapı	5	sektör	7	deniz	7	veri
24	teknoloji	7	yay	6	kültür	11	engel	4	göğüs	3	kriz	4	yol	5	ansikloped	4	şifre	7	kar	7	alış
25	dil	6	dil	5	özel	11	etkinlik	4	hekim	3	risk	4	alerji	4	bebek	4	baş	6	top	7	birli
26	din	6	edebiyat	5	büyük	10	ç	4	hemsire	3	soğuk	4	çay	4	bilim	4	mal	6	değişim	6	ekonomi
27	kar	6	sınav	5	ekonomi	10	lise	4	kadın	3	vücu	4	dönem	4	dr	4	ülke	6	dönem	6	hızlı
28	öğrenim	6	atatürk	4	kişi	10	öğreti	4	kar	3	ab	3	iş	4	gen	4	yorum	6	kavram	6	küresel
29	öğreti	6	halk	4	yüksek	10	tartı	4	kulak	3	cilt	3	kara	4	hafta	4	ab	5	para	6	rekabet
30	ülke	6	kar	4	din	9	vize	4	roma	3	hamile	3	kilo	4	hekim	4	değişim	5	sektör	6	sistem
31	insan	5	kps	4	merkez	9	amerika	3	saç	3	kilo	3	uzman	4	ilaç	4	fiyat	5	taş	6	taraf
32	müdür	5	kültür	4	meslek	9	baş	3	soğuk	3	nokta	3	ürün	4	kadın	4	içeri	5	ürün	6	türk
33	doğa	4	meslek	4	bilim	8	çalışma	3	su	3	psikolojik	3	yaşam	4	kış	4	kara	5	birli	5	yara

Şekil 4. 29: Bilinen web sayfalarında kelime frekansları (web sayfasında en fazla geçen kelimeden en az geçen kelimeye doğru sıralanmış).

Web sayfalarında kelimelere ait frekans değerlerine göre eğitim ile ilgili sayfalarda ilk 10 satırın, eğitim ile ilgili kelimelerden oluştuğu görülmektedir. Diğer obje grupları arasında sağlık ile ilgili bir web sitesi olan sağlık3 web sitesi de ilk 10 satırda sağlık ile kelime frekanslarının yoğunlaştığı görülmektedir. Elde edilen sonuçlara göre 6 obje grubunda rastgele seçilen 8 kelimede frekans değerlerinin sonuçlarına göre toplam 4 eğitim sayfasında geçen anaokulu kelimesine ait frekans değeri 5 iken aynı kelime için diğer sınıflarda 0 olduğu dikkat çekmektedir. Sözlük oluşturulurken kelimelerin seçiminin önemi (Şekil 4.29) ve Çizelge 4.5’ de açık bir biçimde görülmektedir.

Çizelge 4. 5: 8 kelime için toplam kelime frekanslarını gösteren tablo.

	ΣEğitim	ΣSağlık	ΣTicaret	ΣTurizm	ΣSpor	ΣKültür
Aktar	5	0	2	0	0	2
Ambulans	0	0	0	0	0	0
Anaokulu	5	0	0	0	0	0
ATM	0	0	0	0	1	0
Bakkal	0	0	0	0	0	0
Banka	2	0	2	0	0	0
Çocuk	11	27	1	0	1	0
Dershane	10	0	0	0	0	1

Seçilen 6 farklı obje grubu ve belediye sözlüğü kullanılarak yapılan frekans değerleri aşağıda sunulmaktadır (Çizelge 4.6).

Çizelge 4. 6: Eğitici dokümanlarında erişilen kelime frekansı.

Eğitici Dokümanlar	Kelime sayısı	Belediye kelimeleri ile erişilen sayı	Belediye kelimeleri ile kelime (%)	Ortalama (%)
e1	1487	300	20,17	19,82
e2	1219	260	21,33	
e3	2450	398	16,24	
e4	1036	223	21,53	
s1	1172	114	9,73	11,27
s2	738	101	13,69	
s3	1136	115	10,12	
s4	1067	123	11,53	
ti1	1840	306	16,63	15,49
ti2	1362	284	20,85	
ti3	1041	192	18,44	
ti4	1272	77	6,05	
tu1	543	104	19,15	20,38
tu2	1950	356	18,26	
tu3	429	116	27,04	
tu4	1161	198	17,05	
sp1	1231	167	13,57	10,40
sp2	1092	149	13,64	
sp3	544	42	7,72	
sp4	1978	132	6,67	
kü1	493	119	24,14	20,36
kü2	542	85	15,68	
kü3	776	117	15,08	
kü4	689	183	26,56	
Toplam	27248	4261	Ortalama	16,29

Çizelgede ifade edilen kelime sayısı kısmı önişlemden geçirilen web sayfalarındaki mevcut kelime sayısına işaret eder. Örneğin eğitim ile ilgili 1. web sayfasında 1487 adet kelime bulunmaktadır. Belediye destekli erişilen kelime sayısı kullanılan sözlük ile elde edilen toplam kelime sayısını vermektedir. 1. web sayfasından belediye destekli sözlük kullanılarak erişilen toplam kelime sayısı 300 adettir. Erişilen yüzde hesaplandığında 1487 adet kelimedenden 300 adetine ulaşılması sonucunda %20,17 lik bir oran yakalanmaktadır.

Çizelge 4. 7: Ontolojik olarak desteklemiş belediye sözlüğü yardımı ile erişilen kelimeler.

Eğitici Dokümanlar	Kelime sayısı	Ontolojik olarak desteklenen belediye kelimeleri ile erişilen sayı	Ontolojik olarak desteklenen belediye kelimeleri ile kelime (%)	Ortalama (%)
e1	1487	672	45,19	46,58
e2	1219	614	50,37	
e3	2450	1074	43,84	
e4	1036	486	46,91	
s1	1172	379	32,34	37,79
s2	738	329	44,58	
s3	1136	439	38,64	
s4	1067	380	35,61	
ti1	1840	710	38,59	37,06
ti2	1362	662	48,60	
ti3	1041	390	37,46	
ti4	1272	300	23,58	
tu1	543	221	40,70	47,80
tu2	1950	919	47,13	
tu3	429	245	57,11	
tu4	1161	537	46,25	
sp1	1231	454	36,88	36,92
sp2	1092	434	39,74	
sp3	544	218	40,07	
sp4	1978	613	30,99	
kü1	493	239	48,48	46,74
kü2	542	228	42,07	
kü3	776	301	38,79	
kü4	689	397	57,62	
Toplam	27248	11241	Ortalama	42,15

Ontolojik olarak desteklenmiş belediye sözlüğünde bu oran değişmektedir. Erişilen kelime sayıları e1 sayfasında 672 e2 sayfasında 614 olarak tespit edilmektedir (Çizelge 4.7).

Sonuç olarak ontolojik yaklaşımla çok daha fazla kelimeye ulaşılabilirdiği görülmektedir. Fakat daha fazla kelimeye erişimin sınıflandırmanın duyarlılığını arttıracakı söylenemez. Sözlük oluşturulurken dikkat edilmesi gereken nokta kelimelerin sınıflarını tam anlamıyla temsil edecek özellikte oluşturulmasıdır. Belediye sözlüğü ve ontolojik olarak desteklenmiş belediye sözlüğü kullanılarak elde edilen frekanslar ayrı ayrı değerlendirilmiş, sınıflandırmadaki başarı sonuçları ölçülmüştür.

4.4.1 Ontolojik yaklaşım

Çalışmada web sayfalarındaki adreslerin kullanım türlerinin belirlenebilmesi amacı ile K-NN sınıflandırma yöntemi seçilmiştir. Web madenciliğindeki sınıflandırma işlemi, yapılandırılmış veritabanlarındaki sınıflandırma işleminden verilerin ön işlemlerden geçirilmesi bakımından farklılıklar göstermektedir. Web sayfalarındaki veriler yapılandırılmamış özelliğe sahip olduklarından, öncelikle içermiş oldukları kelimelerin metin madenciliğinde olduğu gibi sayısal formata dönüştürülmesi gerekmektedir. Sayısal formata dönüştürme işlemi özel bir sözlük yardımı ile yapılmaktadır. Kullanılan sözlük yardımı ile her bir web sayfasında yer alan kelimeler ve kelimelerin web sayfasındaki geçme sıklıkları tespit edilir. Çalışma süresince 2 farklı sözlükten bahsedilmektedir. Belediye sözlüğü olarak adlandırılan sözlükte bulunan kelimeler, sağlık, ticaret vb. temel obje grupları ve bu obje gruplarına ait alt obje gruplarında geçen kelimeler yardımı ile oluşturulmaktadır. Ontolojik olarak desteklenmiş belediye sözlüğünde ise temel obje grupları ve bu obje gruplarında temel kavramı temsil edebilecek bütün kelimelerin sözlük içerisinde yer almasına çalışılmaktadır. İki farklı sözlüğün kullanılmasının nedenleri ve neden buna ihtiyaç duyulduğu sonraki kısımlarda geniş olarak açıklanmaktadır.

4.4.2 Belediye sözlüğü

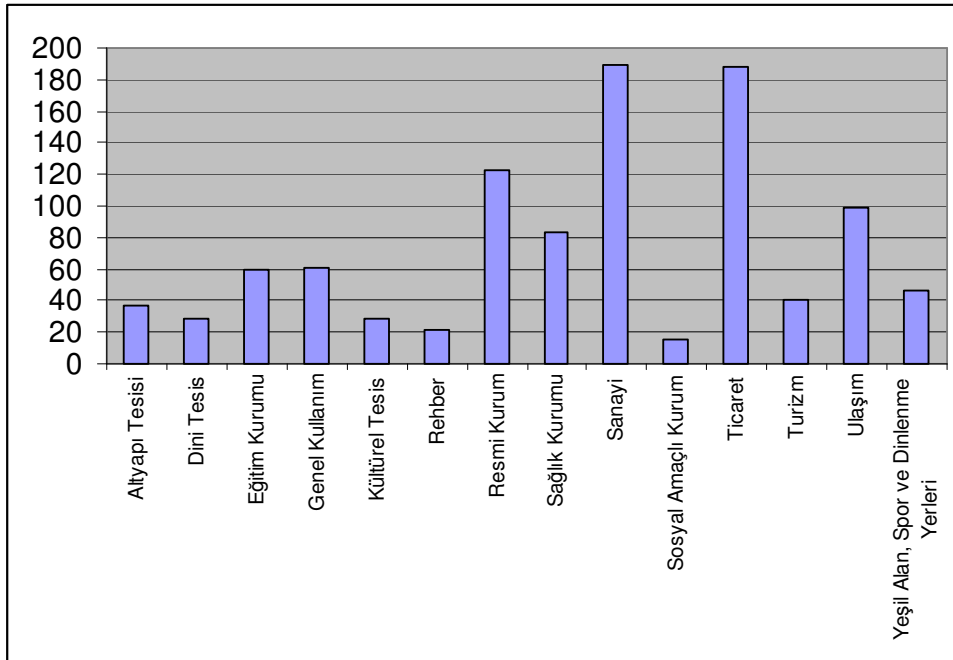
Belediye sözlüğünde kullanılmak üzere temel obje grupları ve bu obje gruplarına ait kelimeler sınıflandırma çalışmasının ilk aşamasında belirlenmektedir. Sağlık ile ilgili temel obje grubu içerisinde bulunan kelimeler arasında eczane, hastane, klinik vb. merkezler yer almaktadır. Değerlendirme sonucunda ortaya çıkan çizelgede 14 farklı

temel obje grubu ve kelime sayıları arasında ciddi farklılıklar dikkat çekmektedir (Çizelge 4.8).

Çizelge 4. 8: Belediye obje gruplarındaki kelimelerinin dağılımı.

OBJE GRUBU İSMİ	KELİME SAYISI
Altyapı Tesisi	37
Dini Tesis	29
Eğitim Kurumu	60
Genel Kullanım	61
Kültürel Tesis	28
Rehber	22
Resmi Kurum	123
Sağlık Kurumu	83
Sanayi	189
Sosyal Amaçlı Kurum	15
Ticaret	188
Turizm	41
Ulaşım	99
Yeşil Alan, Spor ve Dinlenme Yerleri	46
Toplam	1021

Toplamda 1021 adet kelimedenden oluşan sözlükte sağlık kurumları ile ilgili geçen kelime sayısı 83, ticaret ile ilgili kelime sayısı 188 olarak görülmektedir. Aradaki farklılığı daha ciddi biçimde algılayabilmek amacı ile oluşturulan grafikte farklılıklar net olarak görülmektedir (Şekil 4.30).

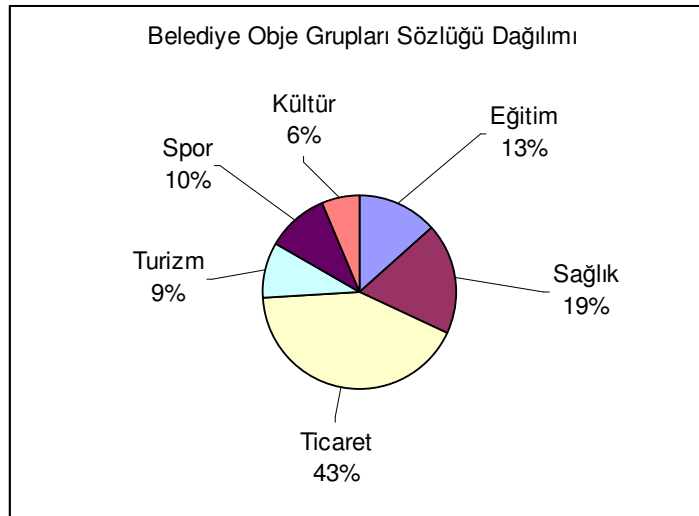


Şekil 4. 30: Belediyeye ait kategorilerde kelimelerin dağılımı.

Belediye obje grupları yardımı ile oluşturulan sözlüğün kullanılması ile yapılan işlemler ve analiz sonuçları ilerleyen bölümlerde daha geniş kapsamda sunulmaktadır. 14 farklı obje grubuna sahip olan belediye verileri arasından, çalışma kapsamında değerlendirilmek üzere web sayfalarında sunulabilecek ve diğer 8 gruba göre çok daha hızlı değişim gösteren 6 obje grubu seçilmiştir (Çizelge 4.9).

Çizelge 4. 9: Şişli çalışma bölgesinde seçilen 6 farklı grup ve bunlara ait kelime dağılımları.

OBJE GRUBU İSMİ	KELİME SAYISI
Eğitim Kurumu	60
Kültürel Tesis	28
Sağlık Kurumu	83
Ticaret	188
Turizm	41
Yeşil Alan, Spor ve Dinlenme Yerleri	46



Şekil 4. 31: Belediyeye ait test verisi olarak seçilen obje gruplarında kelime dağılım grafiği.

Yapılan inceleme sonucunda seçilen 6 obje grubu arasında ticaret ile ilgili kelimelerin sayısının 6 obje grubu arasında % 64 gibi yüksek bir orana sahip olduğu ortaya çıkmaktadır. Belediye tarafından standart olarak saptanan obje gruplarındaki (Şekil 4.31) mevcut kelimelerden oluşturulan sözlük (EK CD)'de sunulmaktadır. Kelimelerin düzensiz dağılımının engellenmesi ve dolayısı ile web sayfalarının kullanım türlerinin belirlenmesinde daha başarılı sonuçlar elde edebilmek amacı ile obje gruplarına ait kelimeler anlamsal olarak yeni kavramlarla desteklenmesi gereği ortaya çıkmaktadır.

4.4.3 Genişletilmiş sözlük

Çalışmanın amacına uygun işyeri fonksiyonlarının belirlenmesinde kullanılan Web madenciliği uygulaması kapsamında belediye verilerinin yapılandırılmasında kullanılan kavramların yetemeyeceği açıktır. Bu yüzden bu kavramların ek sözcüklerle desteklenmesi gerekmektedir (EK CD). Aksi takdirde web madenciliği uygulamasında hazır bir veritabanı kullanmayıp, insanlar tarafından oluşturulan web sayfalarını kullandığı düşünüldüğünde verilerin çok başarılı bir biçimde eşleşmeyeceği açıktır. Günlük konuşma diline bakıldığında, örneğin eğitim kurumları obje grubu için bir metnin içerisinde geçen sadece okul değil, kalem, öğretmen, öğretim, öğrenim gibi kelimeler de o metnin eğitim ile ilgili olduğunu gösterir. Bütün bu eksikliklerin giderilmesi ve bilgilerin sağlıklı sonuçlar vermesi amacıyla çalışmada kullanılacak olan sözlük ontolojik olarak desteklenmiştir.

Temel olarak ontoloji kavramı nedir? Hangi alanlarda ne gibi amaçlar için kullanılır? gibi sorulara yanıtlar bulabilmek amacı ile yapılan araştırmada aşağıdaki bilgilere ulaşılmaktadır. Ontoloji, bir bütün olarak varlığı ele alan ve varlığın en temel niteliklerini inceleyen bir felsefi disiplindir (Hochmair 2005, Lim ve Sun 2005, Golemanti ve diğ, 2007, Henriksson ve diğ, 2008).

- Varoluş nedir?
- Fiziksel nesneler nelerdir?
- Bir fiziksel nesnenin var olduğunu söylemek ne demektir? Sorusunun cevabını vermek mümkün müdür?
- Bir nesnenin özellikleri veya ilişkileri nedir ve bunlar nesneyle nasıl ilişkilidir?
- Var oluş bir özellik midir?
- Bir nesne ne zaman yok olur, ne zaman değişir?

gibi sorulara yanıt bulmak ta ontolojinin konusudur.

Bilgisayar biliminde “kavramsallaştırmanın açıkça belirtilmesidir” şeklinde Gruber (1993) tarafından açıklanan ontoloji, soyut model oluşturma anlamına gelmektedir. Terimlerin anlamları, terimler arasındaki ilişkileri, terimler arasındaki eşanlamlılıkları ve sıradüzen kavramı ontoloji olarak sunulmaktadır.

Golemati ve diğ (2007), ontoloji oluşturulurken seçilen genel kavramın genişletilebilir ve özelleştirilebilir olmasına dikkat edilmesi gerektiği konusunda, ontoloji ile ilgilenenleri açık bir biçimde uyarmaktadırlar.

Ontolojiler yapay zeka, anlamsal web, ve yazılım mühendisliğinde dünyanın tamamının veya bir kısmının temsil edilmesinde kullanılır ve tanımladığı kavramlar

- Bireyler: temel nesneler
- Sınıflar: kümeler, nesne tipleri
- Özellikler: nesnenin sahip olduğu nitelikler, karakteristikler ve parametreler
- İlişkiler: nesnelerin birbirleri arasındaki etkileşim yollarının tanımları olarak sıralanabilir.

Günümüzde ontolojiler, ifade edildiği tanımlama diline bağlı olmaksızın birçok benzer özelliğe sahiptir. Yukarıda da belirtildiği gibi, birçok ontoloji birey, sınıf, özellik, ilişki, kural ve kısıtlamalardan oluşur (Stumme ve diğ 2006, Cheng ve Du 2008).

Mekânsal özellikteki çalışma alanlarında da ontoloji kullanılmaktadır. Lemmens ve Vries, (2004) yerleşim çalışmalarında, Henriksson ve diğ (2008) web sayfalarından doğal ve yapay yerleşimlerin kazanımı çalışmalarında, Lemmens ve diğ (2006) ise yaptıkları çalışmalarda ontoloji ile ilgilenmişlerdir. Barb ve Shyu, (2008) uydu görüntülerinin kayıtlı tutulduğu veritabanları için oluşturmuş oldukları kavramsal (ontolojik) şema ile anlamsal (semantik) sorgulamaları desteklemeyi hedeflemiştir.

Ontoloji ile ilgili bilgiler ışığında belediye obje gruplarının yardımı ile oluşturulan sözlük incelendiği zaman, kelime olarak yetersiz ve seçilen obje gruplarında kelime dağılımı bakımından da dengesizlikler göstermektedir. Bu dengesizlik kısaca bir işlemle ortaya koyulabilir (4.1).

k_s -Sözlükteki bütün kelimelerin sayısı
 k_t -Sözlükte bulunan turizm ile ilgili kelime sayısı
 t_t -turizm ile ilgili kelimelerin dağılımı
 k_{ti} -Sözlükte bulunan ticaret ile ilgili kelime sayısı
 t_{ti} -ticaret ile ilgili kelimelerin dağılımı

$$t_t = \frac{k_t}{k_s} * 100 \quad (4.1)$$

$$k_s=446 \quad k_t=41 \quad t_t=\%9,1.$$

$$k_s=446 \quad k_{ti}=188 \quad t_{ti}=\%43.$$

Bu dengesizliği giderebilmek için belediye obje gruplarından oluşturulan sözlüğün ontolojik olarak desteklenmesi yoluna gidilmiştir. Ontolojik olarak desteklenen belediye kelimeleri için yeniden aynı işlemlerin (4.2) yapılması durumunda karşılaşılan sonuçlar aşağıdaki gibidir.

k_s -Sözlükteki bütün kelimelerin sayısı
 k_t -Sözlükte bulunan turizm ile ilgili kelime sayısı
 Ot_t -turizm ile ilgili kelimelerin dağılımı
 k_{ti} -Sözlükte bulunan ticaret ile ilgili kelime sayısı
 Ot_{ti} -ticaret ile ilgili kelimelerin dağılımı

$$Ot_t = \frac{k_t}{k_s} * 100 \quad (4.2)$$

$$k_s=1747 \quad k_t=200 \quad Ot_t=\%11,4.$$

$$k_s=1747 \quad k_{ti}=293 \quad Ot_{ti}=\%16,8.$$

Bu sayede 2 farklı soruna çözüm getirilmiş olmaktadır.

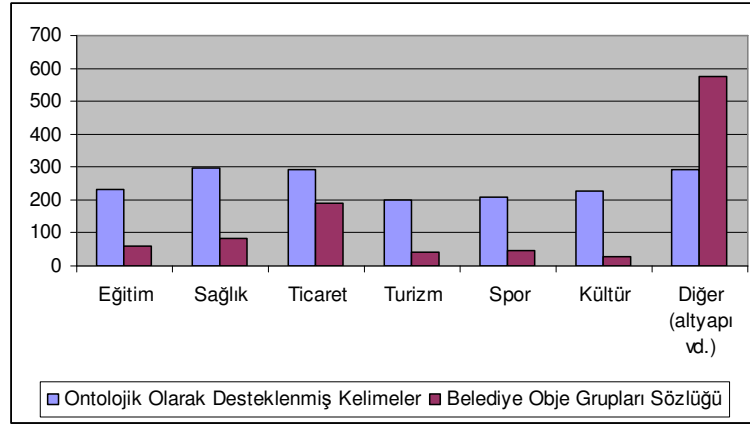
- Kelimelerin dengesiz dağılımının optimum düzeye indirgenmesinin sağlanması,
- Web sayfalarının kullanım türlerinin saptanmasında daha doğru sonuçlar elde edilmesi.

Genelde bilginin sunulması üzerine ontoloji oluşturulurken farklı seviyeler seçilmektedir. Çalışmada üst, orta ve alt seviyelerde olmak üzere 3 seviye belirlenmiştir. Bu seviye dağılımına göre; sağlık, eğitim, kültür, turizm, spor ve ticaret obje grupları, çalışmada üst seviyede yer almaktadır. Hastane, poliklinik, sağlık ocağı gibi kelimeler orta seviyeyi ifade etmektedir. Alt seviyede ise doğrudan bina kullanım türü kapsamına girmeyen fakat kelime olarak sağlık sektörünü çağrıştırabilecek olan hemşire, doktor, diş, eklem vb. gibi kelimeler kullanılmaktadır. Belediye ve ontolojik olarak desteklenen belediye sözlüğünde yer alan kelimeler EK-CD kısmında verilmektedir. Ontolojik olarak desteklenen kelimelerin obje gruplarındaki dağılımı artmış ve dengelenmiştir (Çizelge 4.10).

Çizelge 4. 10: Obje grupları ve kelime sayıları karşılaştırılması.

OBJE GRUPLARI	ONTOLOJİK OLARAK DESTEKLENMİŞ KELİMELER	BELEDİYE OBJE GRUPLARI SÖZLÜĞÜ
Eğitim	230	60
Sağlık	295	83
Ticaret	293	188
Turizm	200	41
Spor	210	46
Kültür	226	28
Diğer (altyapı vd.)	293	575
Toplam	1747	1021

Belediye sözlüğündeki sağlık ile ilgili kelimelerin sayısı 83 adet ve sadece bina fonksiyonları ile sınırlandırılmış olmasına karşın, ontolojik olarak desteklenen sözlükte bu sayı 295 e kadar yükselmektedir (Şekil 4.32). Sadece bina fonksiyon bilgileri ile sınırlandırılmadan sözlükte sağlık ile ilgili olabilecek bütün destekleyici kavramların yer alması sağlanmaya çalışılmıştır.



Şekil 4. 32: Ontolojik olarak desteklenmiş ve belediye obje gruplarından oluşturulmuş sözlüğün karşılaştırılması.

Yapılan bütün bu işlemler sonucunda ontolojik olarak desteklenen kelimeler ile birlikte obje gruplarına ait kelimeler arasındaki düzensiz dağılım en aza indirgenmektedir (Şekil 4.32).

4.5 Sınıflandırma İşlemine Hazırlık

4.5.1 Belediye veri seti kullanımı

Çalışmada daha önce de belirtildiği gibi sözlük olarak belediye ve ontoloji destekli belediye sözlükleri kullanılmaktadır. Çalışmanın bu bölümünde belediye sözlüğü

Yönteme göre eğitim setleri ve eğitilecek olan setler arasındaki mesafeler dikkate alınarak veriler K-NN uygulaması için hazır hale getirilmektedir. Bu amaçla Öklid yöntemi kullanılmaktadır. Web sayfaları arasındaki mesafelerin ölçümü prensibine dayana Öklid yöntemi 3. Bölümde ayrıntılı olarak verilmiş bulunmaktadır.

	eğ1	eğ2	eğ3	eğ4	eğ5	eğ6	sa1	sa2	sa3	sa4	sa5	sa6
1	eğ1	eğ2	eğ3	eğ4	eğ5	eğ6	sa1	sa2	sa3	sa4	sa5	sa6
2	216,081	160,120	173,078	153,753	161,149	147,149	155	165,781	174,105	157,510	164,402	157,0
3	160,120	101,587	99,214	72,526	81,288	71,414	77,916	77,756	109,731	83,624	99,307	80,89
4	173,078	101,587	126,340	103,286	112,867	95,061	104,407	104,355	127,314	108,700	117,422	109,8
5	153,753	99,214	104,818	75,863	87,544	74,833	73,824	73,300	108,555	79,974	97,933	76,40
6	161,149	82,994	46,051	67,282	66,803	28,757	39,974	87,275	50,822	69,785	44,8	
7	147,149	32,143	93,289	46,615	67,082	66,832	39,736	87,395	52,611	73,591	45,38	
8	155	51	93,225	47,191	63,906	58,489	36,386	39,225	87,194	52,744	76,517	44,36
9	165,781	20,757	91,707	39,785	54,761	48,238	31,749	36,290	106,106	81,167	95,164	79,88
10	174,105	55,932	109,393	61,571	69,022	68,406	74,696	72,952	106,106	81,167	95,164	79,88
11	157,510	91,380	101,725	60,547	75,980	70,456	87,541	86,595	87,342	99,577	63,72	
12	164,402	27,477	91,613	45,771	65,482	62,890	38,548	35,872	50,557	69,785	44,8	
13	157,0	29,982	92,749	47,979	65,667	57,008	39,484	38,935	89,593	51,971	81,104	43,48
14	160,120	108,600	139,554	114,332	122,674	117,209	109,796	111,202	136,040	116,081	129,027	113,3
15	173,078	34,525	94,138	50,675	68,154	60,386	42,201	41,761	89,088	54,378	79,724	47,42
16	153,753	82,783	107,847	73,034	84,816	79,435	67,059	67,720	104,847	75,818	95,467	70,88
17	161,149	59,101	105,380	69,332	83,390	78,085	64,946	64,303	102,430	72,856	95,420	67,30
18	147,149	31,448	93,150	48,528	67,334	67,262	40,199	40,877	89,241	53,253	79,214	45,36
19	155	15,620	89,474	41,207	61,018	54,258	32,480	30,789	88,434	46,378	77,562	36,04
20	165,781	34,058	93,872	47,707	67,637	68,639	49,439	42,731	90,559	54,525	81,867	47,12
21	174,105	55,932	109,393	61,571	69,022	68,406	74,696	72,952	106,106	81,167	95,164	79,88
22	157,510	91,380	101,725	60,547	75,980	70,456	87,541	86,595	87,342	99,577	63,72	
23	164,402	27,477	91,613	45,771	65,482	62,890	38,548	35,872	50,557	69,785	44,8	
24	157,0	29,982	92,749	47,979	65,667	57,008	39,484	38,935	89,593	51,971	81,104	43,48
25	160,120	108,600	139,554	114,332	122,674	117,209	109,796	111,202	136,040	116,081	129,027	113,3
26	173,078	34,525	94,138	50,675	68,154	60,386	42,201	41,761	89,088	54,378	79,724	47,42
27	153,753	82,783	107,847	73,034	84,816	79,435	67,059	67,720	104,847	75,818	95,467	70,88
28	161,149	59,101	105,380	69,332	83,390	78,085	64,946	64,303	102,430	72,856	95,420	67,30
29	147,149	31,448	93,150	48,528	67,334	67,262	40,199	40,877	89,241	53,253	79,214	45,36
30	155	15,620	89,474	41,207	61,018	54,258	32,480	30,789	88,434	46,378	77,562	36,04
31	165,781	34,058	93,872	47,707	67,637	68,639	49,439	42,731	90,559	54,525	81,867	47,12
32	174,105	55,932	109,393	61,571	69,022	68,406	74,696	72,952	106,106	81,167	95,164	79,88
33	157,510	91,380	101,725	60,547	75,980	70,456	87,541	86,595	87,342	99,577	63,72	
34	164,402	27,477	91,613	45,771	65,482	62,890	38,548	35,872	50,557	69,785	44,8	
35	157,0	29,982	92,749	47,979	65,667	57,008	39,484	38,935	89,593	51,971	81,104	43,48
36	160,120	108,600	139,554	114,332	122,674	117,209	109,796	111,202	136,040	116,081	129,027	113,3
37	173,078	34,525	94,138	50,675	68,154	60,386	42,201	41,761	89,088	54,378	79,724	47,42
38	153,753	82,783	107,847	73,034	84,816	79,435	67,059	67,720	104,847	75,818	95,467	70,88
39	161,149	59,101	105,380	69,332	83,390	78,085	64,946	64,303	102,430	72,856	95,420	67,30
40	147,149	31,448	93,150	48,528	67,334	67,262	40,199	40,877	89,241	53,253	79,214	45,36
41	155	15,620	89,474	41,207	61,018	54,258	32,480	30,789	88,434	46,378	77,562	36,04
42	165,781	34,058	93,872	47,707	67,637	68,639	49,439	42,731	90,559	54,525	81,867	47,12
43	174,105	55,932	109,393	61,571	69,022	68,406	74,696	72,952	106,106	81,167	95,164	79,88
44	157,510	91,380	101,725	60,547	75,980	70,456	87,541	86,595	87,342	99,577	63,72	
45	164,402	27,477	91,613	45,771	65,482	62,890	38,548	35,872	50,557	69,785	44,8	
46	157,0	29,982	92,749	47,979	65,667	57,008	39,484	38,935	89,593	51,971	81,104	43,48
47	160,120	108,600	139,554	114,332	122,674	117,209	109,796	111,202	136,040	116,081	129,027	113,3
48	173,078	34,525	94,138	50,675	68,154	60,386	42,201	41,761	89,088	54,378	79,724	47,42
49	153,753	82,783	107,847	73,034	84,816	79,435	67,059	67,720	104,847	75,818	95,467	70,88
50	161,149	59,101	105,380	69,332	83,390	78,085	64,946	64,303	102,430	72,856	95,420	67,30
51	147,149	31,448	93,150	48,528	67,334	67,262	40,199	40,877	89,241	53,253	79,214	45,36
52	155	15,620	89,474	41,207	61,018	54,258	32,480	30,789	88,434	46,378	77,562	36,04
53	165,781	34,058	93,872	47,707	67,637	68,639	49,439	42,731	90,559	54,525	81,867	47,12
54	174,105	55,932	109,393	61,571	69,022	68,406	74,696	72,952	106,106	81,167	95,164	79,88
55	157,510	91,380	101,725	60,547	75,980	70,456	87,541	86,595	87,342	99,577	63,72	
56	164,402	27,477	91,613	45,771	65,482	62,890	38,548	35,872	50,557	69,785	44,8	
57	157,0	29,982	92,749	47,979	65,667	57,008	39,484	38,935	89,593	51,971	81,104	43,48
58	160,120	108,600	139,554	114,332	122,674	117,209	109,796	111,202	136,040	116,081	129,027	113,3
59	173,078	34,525	94,138	50,675	68,154	60,386	42,201	41,761	89,088	54,378	79,724	47,42
60	153,753	82,783	107,847	73,034	84,816	79,435	67,059	67,720	104,847	75,818	95,467	70,88
61	161,149	59,101	105,380	69,332	83,390	78,085	64,946	64,303	102,430	72,856	95,420	67,30
62	147,149	31,448	93,150	48,528	67,334	67,262	40,199	40,877	89,241	53,253	79,214	45,36
63	155	15,620	89,474	41,207	61,018	54,258	32,480	30,789	88,434	46,378	77,562	36,04
64	165,781	34,058	93,872	47,707	67,637	68,639	49,439	42,731	90,559	54,525	81,867	47,12
65	174,105	55,932	109,393	61,571	69,022	68,406	74,696	72,952	106,106	81,167	95,164	79,88
66	157,510	91,380	101,725	60,547	75,980	70,456	87,541	86,595	87,342	99,577	63,72	
67	164,402	27,477	91,613	45,771	65,482	62,890	38,548	35,872	50,557	69,785	44,8	
68	157,0	29,982	92,749	47,979	65,667	57,008	39,484	38,935	89,593	51,971	81,104	43,48
69	160,120	108,600	139,554	114,332	122,674	117,209	109,796	111,202	136,040	116,081	129,027	113,3
70	173,078	34,525	94,138	50,675	68,154	60,386	42,201	41,761	89,088	54,378	79,724	47,42
71	153,753	82,783	107,847	73,034	84,816	79,435	67,059	67,720	104,847	75,818	95,467	70,88
72	161,149	59,101	105,380	69,332	83,390	78,085	64,946	64,303	102,430	72,856	95,420	67,30
73	147,149	31,448	93,150	48,528	67,334	67,262	40,199	40,877	89,241	53,253	79,214	45,36
74	155	15,620	89,474	41,207	61,018	54,258	32,480	30,789	88,434	46,378	77,562	36,04
75	165,781	34,058	93,872	47,707	67,637	68,639	49,439	42,731	90,559	54,525	81,867	47,12
76	174,105	55,932	109,393	61,571	69,022	68,406	74,696	72,952	106,106	81,167	95,164	79,88
77	157,510	91,380	101,725	60,547	75,980	70,456	87,541	86,595	87,342	99,577	63,72	
78	164,402	27,477	91,613	45,771	65,482	62,890	38,548	35,872	50,557	69,785	44,8	
79	157,0	29,982	92,749	47,979	65,667	57,008	39,484	38,935	89,593	51,971	81,104	43,48
80	160,120	108,600	139,554	114,332	122,674	117,209	109,796	111,202	136,040	116,081	129,027	113,3
81	173,078	34,525	94,138	50,675	68,154	60,386	42,201	41,761	89,088	54,378	79,724	47,42
82	153,753	82,783	107,847	73,034	84,816	79,435	67,059	67,720	104,847	75,818	95,467	70,88
83	161,149	59,101	105,380	69,332	83,390	78,085	64,946	64,303	102,430	72,856	95,420	67,30
84	147,149	31,448	93,150	48,528	67,334	67,262	40,199	40,877	89,241	53,253	79,214	45,36
85	155	15,620	89,474	41,								

Microsoft Excel - Belediyecisün

<

Şekil 4. 35: Belediye kelimeleri yardımı ile Kosinüs değerlerinin hesaplanması.

Kosinüs değerlerinin hesaplanmasında aradaki benzerlik oranı hesaplanmaktadır. Genel Kosinüs yöntemi yapılan çalışmaya uyarlandığında bağıntılar aşağıdaki gibi ifade edilebilir (Şekil 4.35).

$k_1, k_2, \dots, k_{1276}$ dahil $e_1, e_2, \dots, sa_1, sa_2, \dots, ti_1, ti_2, \dots, \dots$ olmak üzere
 fk_i =Eğitici sayfalardaki kelime frekansı $i=1, 2, \dots, 1276$
 fq_i =Eğitilen sayfalardaki kelime frekansı
 fk_j =Eğitici sayfalardaki kelime frekansı $j=1, 2, \dots, 24$
 dQ_j =Öklid yöntemine göre hesaplanan değer $j=1, 2, \dots, 24$

$$dQ_j = fk_1 * fq_1 + fk_2 * fq_2 + fk_3 * fq_3 + \dots + fk_{1276} * fq_{1276} \quad (5.2)$$

$$fk_j^2 = fk_1^2 + fk_2^2 + fk_3^2 + fk_4^2 + \dots + fk_{1276}^2 \quad (5.3)$$

$$fQ^2 = fq_1^2 + fq_2^2 + fq_3^2 + fq_4^2 + \dots + fq_{1276}^2 \quad (5.4)$$

$$\text{Benzerlik}(W, Q) = dQ_j / (fk_j * fQ) \quad (5.5)$$

Kullanım türü bilinen web sayfalarındaki kelime sıklıkları ve kullanım türü belli olmayan web sayfasındaki kelime sıklıkları arasında bir benzerlik oranı hesaplanır (Şekil 4.36).

1	eğ1	eğ2	eğ3	eğ4	eğ5	eğ6	sa1	sa2	sa3	sa4										
2	eğ1	0.126708832	eğ1	0.174845282	eğ1	0.055210985	eğ1	0.125591426	eğ1	0.055843009	eğ1	0.238866717	eğ1	0.045835519	eğ1	0.003930258	eğ1	0.014068206	eğ1	0.053106222
3	eğ2	0.380739322	eğ2	0.520742934	eğ2	0.282646917	eğ2	0.304323477	eğ2	0.279886265	eğ2	0.410870964	eğ2	0.058873095	eğ2	0.042188991	eğ2	0.059268513	eğ2	0.065193821
4	eğ3	0.255279119	eğ3	0.193381159	eğ3	0.145926586	eğ3	0.204430078	eğ3	0.130865417	eğ3	0.403277363	eğ3	0.116206503	eğ3	0.106782624	eğ3	0.106243666	eğ3	0.098153849
5	eğ4	0.209926533	eğ4	0.196904474	eğ4	0.134815564	eğ4	0.121726908	eğ4	0.083773591	eğ4	0.280497296	eğ4	0.041224492	eğ4	0.034881444	eğ4	0.081283627	eğ4	0.050991054
6	sa1	0.036171717	sa1	0.070780619	sa1	0.035014334	sa1	0.205355036	sa1	0.009589243	sa1	0.211541312	sa1	0.548845707	sa1	0.087683522	sa1	0.131233749	sa1	0.124775272
7	sa2	0.071248543	sa2	0.020598043	sa2	0.023694997	sa2	0.192685273	sa2	0.025484847	sa2	0.260131723	sa2	0.095130205	sa2	0.095130205	sa2	0.139812339	sa2	0.065360069
8	sa3	0.17328322	sa3	0.0	sa3	0.0183356	sa3	0.144289779	sa3	0.122656565	sa3	0.141402309	sa3	0.263473089	sa3	0.070085736	sa3	0.127217552	sa3	0.028972949
9	sa4	0.145277773	sa4	0.085545648	sa4	0.057339946	sa4	0.43966172	sa4	0.072401836	sa4	0.494255696	sa4	0.410743707	sa4	0.168825777	sa4	0.243847392	sa4	0.153821962
10	ti1	0.129173279	ti1	0.438125998	ti1	0.085285477	ti1	0.497019742	ti1	0.088220397	ti1	0.461009682	ti1	0.087424289	ti1	0.123760017	ti1	0.093600462	ti1	0.058965397
11	ti2	0.076575989	ti2	0.309371147	ti2	0.014113357	ti2	0.100987433	ti2	0.033643793	ti2	0.131187885	ti2	0.064420795	ti2	0.064391568	ti2	0.042366425	ti2	0.027737695
12	ti3	0.140590617	ti3	0.027427809	ti3	0.015656632	ti3	0.148363005	ti3	0.073116304	ti3	0.11040191	ti3	0.055009635	ti3	0.031201883	ti3	0.022258141	ti3	0.007911979
13	ti4	0.044153844	ti4	0.016656914	ti4	0.042281483	ti4	0.135283962	ti4	0.008737414	ti4	0.317787194	ti4	0.05758871	ti4	0.066510496	ti4	0.085272189	ti4	0.042450207
14	tu1	0.049153133	tu1	0.037626983	tu1	0.016143348	tu1	0.077376435	tu1	0.034253389	tu1	0.176939761	tu1	0.070293271	tu1	0.029633851	tu1	0.021159578	tu1	0.025047699
15	tu2	0.047444548	tu2	0.019756498	tu2	0.009831378	tu2	0.042331443	tu2	0.026113483	tu2	0.08486953	tu2	0.069520787	tu2	0.023286217	tu2	0.032085599	tu2	0.027477631
16	tu3	0.090729323	tu3	0.063713093	tu3	0.032220998	tu3	0.100908139	tu3	0.034170495	tu3	0.136730094	tu3	0.130058438	tu3	0.098709023	tu3	0.099142627	tu3	0.059263499
17	tu4	0.072752325	tu4	0.015097001	tu4	0.017083759	tu4	0.045803935	tu4	0.048490828	tu4	0.08130812	tu4	0.0711550128	tu4	0.020394248	tu4	0.025435157	tu4	0.027850583
18	sp1	0.040379992	sp1	0.015820455	sp1	0.021672534	sp1	0.065161198	sp1	0.015034826	sp1	0.09803222	sp1	0.029809946	sp1	0.022879399	sp1	0.02828498	sp1	0.019775845
19	sp2	0.075377774	sp2	0.077242662	sp2	0.034767844	sp2	0.120855703	sp2	0.016066408	sp2	0.204345499	sp2	0.130452099	sp2	0.042407441	sp2	0.060993264	sp2	0.040850056
20	sp3	0.138877642	sp3	0.119629721	sp3	0.04389682	sp3	0.216150517	sp3	0.077222731	sp3	0.205358953	sp3	0.049961297	sp3	0.039939692	sp3	0.03526792	sp3	0.02156816
21	sp4	0.09125255	sp4	0.075227717	sp4	0.038845585	sp4	0.203154724	sp4	0.048560059	sp4	0.254935274	sp4	0.071858887	sp4	0.049184711	sp4	0.050078924	sp4	0.049504508
22	ku1	0.058811747	ku1	0.078052408	ku1	0.044094628	ku1	0.114711464	ku1	0.029576652	ku1	0.19370595	ku1	0.083202621	ku1	0.191110461	ku1	0.111220675	ku1	0.12419189
23	ku2	0.05698119	ku2	0.039403906	ku2	0.017295652	ku2	0.0767899	ku2	0.033652935	ku2	0.193272325	ku2	0.032160659	ku2	0.022763359	ku2	0.020846616	ku2	0.067206231
24	ku3	0.095297255	ku3	0.07031137	ku3	0.032944002	ku3	0.108441557	ku3	0.044223246	ku3	0.117176315	ku3	0.091720613	ku3	0.214277502	ku3	0.058841582	ku3	0.31827781
25	ku4	0.074519115	ku4	0.087847945	ku4	0.018128441	ku4	0.117350799	ku4	0.04000352	ku4	0.170488845	ku4	0.101983485	ku4	0.074344302	ku4	0.071736541	ku4	0.076278821

Şekil 4. 36: Belediye kelimeleri yardımı ile oluşturulan Kosinüs sonuçları.

Şekil 5.10'te gösterildiği gibi aradaki benzerlik oranları değerlendirilmektedir. Buna göre 1. bilinmeyen sayfa ile bilinen eğitim sayfası arasındaki benzerlik (e1-Q1) 0,126 olarak hesaplanmıştır. e2-Q1 arasındaki benzerlik ise 0,960 olarak hesaplanmıştır. 1 sayısına en yakın olan değere çok daha benzerdir ifadesinden yola çıkarak $e2Q1 < e1Q1$ olduğundan Q1 e2 sayfasına çok daha benzerdir denilebilir. Bütün benzerlikler Kosinüs yöntemine göre hesaplanarak aralarındaki mesafeler 1 sayısına yakınlıklarına göre büyükten küçüğe doğru sıralanır.

4.5.2 .Ontolojik olarak desteklenen belediye veri seti kullanımı

4.5.2.1 Öklid yöntemine göre çalışmanın irdelenmesi

Çalışmanın ikinci analiz kısmında, web sayfalarında doğru olarak erişilen kelime sayısının artırılması ve böylece hesaplamalarda daha yüksek doğruluklar hedeflenmesi amacı ile ontolojik olarak desteklenmiş kelimeler sözlük belediye sözlüğü kullanılmaktadır (Şekil 4.37).

1	eğ1	eğ2	eğ3	eğ4	sa1	sa2	sa3	sa4	ti1	ti2	ti3	ti4	tu1	tu2	tu3	tu4	sp1	sp2	sp3	sp4	ku1	ku2	ku3	ku4	qe1	eğ1	qe1-eğ2	qe1-eğ3	qe1-eğ4	qe	
2	a.ş.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	ab	0	1	2	1	0	3	1	0	5	1	0	1	0	2	0	4	0	1	0	3	0	0	0	1	25	16	9	16	0	
4	access	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
5	acent	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
6	acil	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
7	acta	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
8	acsap	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
9	ada	0	2	0	0	1	0	1	3	0	0	0	0	0	0	1	2	1	4	0	0	0	0	0	1	1	1	1	1	1	
10	advar	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
11	adli	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
12	adsl	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
13	afet	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
14	afiş	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
15	afrika	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
16	afrodit	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
17	agop	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
18	agora	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
19	ağa	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	2	0	0	0	0	0	0	1	0	0	0	0	0	0	
20	agaoglu	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
21	agir	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0
22	aguz	0	0	0	0	2	2	1	2	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	
23	ahlak	1	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
24	ands	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
25	ajie	1	0	4	7	1	1	0	2	0	2	0	0	0	2	0	0	0	0	0	0	0	0	0	0	1	0	16	0	46	
26	ajan	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
27	akademi	1	1	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	1	
28	akar	0	0	0	0	0	0	0	0	0	0	0	0	2	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	
29	akaryakt	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	
30	akbank	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
31	akdeniz	0	3	0	0	2	0	0	0	0	2	0	0	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
32	akgöl	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
33	akgün	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
34	ak	eğ1	eğ2	eğ3	eğ4	sa1	sa2	sa3	sa4	ti1	ti2	ti3	ti4	tu1	tu2	tu3	tu4	sp1	sp2	sp3	sp4	ku1	ku2	ku3	ku4	qe1	eğ1	qe1-eğ2	qe1-eğ3	qe1-eğ4	qe

Ontolojik olarak desteklenen belediye kelimelerinin oluşturduğu sözlük kullanılarak elde edilen eğitici ve eğitilen web sayfalarına ait matrisler sınıflandırma işlemi için öncelikle ön değerlendirmelerden geçirilmektedir. Öklid yöntemi kullanılarak 33 eğitilen web sayfası için eğitici web sayfaları ile olan mesafeler hesaplanmaktadır. Hesaplama sonuçlarına ait sonuçlar bir tabloda sunulmaktadır (Şekil 4.38).

	eş1	eş2	eş3	eş4	eş5	eş6	sa1	sa2	sa3	sa4
1	eş1	eş2	eş3	eş4	eş5	eş6	sa1	sa2	sa3	sa4
2	eş1	eş2	eş3	eş4	eş5	eş6	sa1	sa2	sa3	sa4
3	eş1	eş2	eş3	eş4	eş5	eş6	sa1	sa2	sa3	sa4
4	eş1	eş2	eş3	eş4	eş5	eş6	sa1	sa2	sa3	sa4
5	eş1	eş2	eş3	eş4	eş5	eş6	sa1	sa2	sa3	sa4
6	eş1	eş2	eş3	eş4	eş5	eş6	sa1	sa2	sa3	sa4
7	eş1	eş2	eş3	eş4	eş5	eş6	sa1	sa2	sa3	sa4
8	eş1	eş2	eş3	eş4	eş5	eş6	sa1	sa2	sa3	sa4
9	eş1	eş2	eş3	eş4	eş5	eş6	sa1	sa2	sa3	sa4
10	eş1	eş2	eş3	eş4	eş5	eş6	sa1	sa2	sa3	sa4
11	eş1	eş2	eş3	eş4	eş5	eş6	sa1	sa2	sa3	sa4
12	eş1	eş2	eş3	eş4	eş5	eş6	sa1	sa2	sa3	sa4
13	eş1	eş2	eş3	eş4	eş5	eş6	sa1	sa2	sa3	sa4
14	eş1	eş2	eş3	eş4	eş5	eş6	sa1	sa2	sa3	sa4
15	eş1	eş2	eş3	eş4	eş5	eş6	sa1	sa2	sa3	sa4
16	eş1	eş2	eş3	eş4	eş5	eş6	sa1	sa2	sa3	sa4
17	eş1	eş2	eş3	eş4	eş5	eş6	sa1	sa2	sa3	sa4
18	eş1	eş2	eş3	eş4	eş5	eş6	sa1	sa2	sa3	sa4
19	eş1	eş2	eş3	eş4	eş5	eş6	sa1	sa2	sa3	sa4
20	eş1	eş2	eş3	eş4	eş5	eş6	sa1	sa2	sa3	sa4
21	eş1	eş2	eş3	eş4	eş5	eş6	sa1	sa2	sa3	sa4
22	eş1	eş2	eş3	eş4	eş5	eş6	sa1	sa2	sa3	sa4
23	eş1	eş2	eş3	eş4	eş5	eş6	sa1	sa2	sa3	sa4
24	eş1	eş2	eş3	eş4	eş5	eş6	sa1	sa2	sa3	sa4
25	eş1	eş2	eş3	eş4	eş5	eş6	sa1	sa2	sa3	sa4
26	eş1	eş2	eş3	eş4	eş5	eş6	sa1	sa2	sa3	sa4

Şekil 4. 38: Öklid yönteminde aradaki mesafeler.

Öklid yöntemi sonuçlarına göre 1. bilinmeyen sayfa ile bilinen eğitim sayfası arasındaki mesafe (e1-q1) 223,349 olarak hesaplanmıştır. e2-q1 arasındaki uzaklık ise 171,593 olarak hesaplanmıştır. $E2q1 < e1q1$ olduğundan q1 e2 sayfasına çok daha yakındır. Bütün mesafeler Öklid yöntemine göre hesaplanarak aralarındaki mesafeler küçükten büyüğe doğru sıralanır. K değerleri ise en yakın komşuluğu ifade etmektedir.

4.5.2.2 Kosinüs yöntemine göre çalışmanın irdelenmesi

Kosinüs yöntemi kullanılarak, ontolojik olarak desteklenen belediye sözlüğünün yardımı ile elde edilen matrisler arasındaki benzerlik oranları değerlendirilmiştir (Şekil 4.39).

Sayfa 64

	e _{g1}	e _{g2}	e _{g3}	e _{g4}	e _{g5}	e _{g6}	e _{g7}	e _{g8}	e _{g9}	e _{g10}	e _{g11}	e _{g12}	e _{g13}	e _{g14}	e _{g15}	e _{g16}	e _{g17}	e _{g18}	e _{g19}	e _{g20}	e _{g21}	e _{g22}	e _{g23}	e _{g24}	e _{g25}	e _{g26}	e _{g27}	e _{g28}	e _{g29}	e _{g30}	e _{g31}	e _{g32}	e _{g33}	e _{g34}	e _{g35}	e _{g36}	e _{g37}	e _{g38}	e _{g39}	e _{g40}	e _{g41}	e _{g42}	e _{g43}	e _{g44}	e _{g45}	e _{g46}	e _{g47}	e _{g48}	e _{g49}	e _{g50}	e _{g51}	e _{g52}	e _{g53}	e _{g54}	e _{g55}	e _{g56}	e _{g57}	e _{g58}	e _{g59}	e _{g60}	e _{g61}	e _{g62}	e _{g63}	e _{g64}	e _{g65}	e _{g66}	e _{g67}	e _{g68}	e _{g69}	e _{g70}	e _{g71}	e _{g72}	e _{g73}	e _{g74}	e _{g75}	e _{g76}	e _{g77}	e _{g78}	e _{g79}	e _{g80}	e _{g81}	e _{g82}	e _{g83}	e _{g84}	e _{g85}	e _{g86}	e _{g87}	e _{g88}	e _{g89}	e _{g90}	e _{g91}	e _{g92}	e _{g93}	e _{g94}	e _{g95}	e _{g96}	e _{g97}	e _{g98}	e _{g99}	e _{g100}	e _{g101}	e _{g102}	e _{g103}	e _{g104}	e _{g105}	e _{g106}	e _{g107}	e _{g108}	e _{g109}	e _{g110}	e _{g111}	e _{g112}	e _{g113}	e _{g114}	e _{g115}	e _{g116}	e _{g117}	e _{g118}	e _{g119}	e _{g120}	e _{g121}	e _{g122}	e _{g123}	e _{g124}	e _{g125}	e _{g126}	e _{g127}	e _{g128}	e _{g129}	e _{g130}	e _{g131}	e _{g132}	e _{g133}	e _{g134}	e _{g135}	e _{g136}	e _{g137}	e _{g138}	e _{g139}	e _{g140}	e _{g141}	e _{g142}	e _{g143}	e _{g144}	e _{g145}	e _{g146}	e _{g147}	e _{g148}	e _{g149}	e _{g150}	e _{g151}	e _{g152}	e _{g153}	e _{g154}	e _{g155}	e _{g156}	e _{g157}	e _{g158}	e _{g159}	e _{g160}	e _{g161}	e _{g162}	e _{g163}	e _{g164}	e _{g165}	e _{g166}	e _{g167}	e _{g168}	e _{g169}	e _{g170}	e _{g171}	e _{g172}	e _{g173}	e _{g174}	e _{g175}	e _{g176}	e _{g177}	e _{g178}	e _{g179}	e _{g180}	e _{g181}	e _{g182}	e _{g183}	e _{g184}	e _{g185}	e _{g186}	e _{g187}	e _{g188}	e _{g189}	e _{g190}	e _{g191}	e _{g192}	e _{g193}	e _{g194}	e _{g195}	e _{g196}	e _{g197}	e _{g198}	e _{g199}	e _{g200}	e _{g201}	e _{g202}	e _{g203}	e _{g204}	e _{g205}	e _{g206}	e _{g207}	e _{g208}	e _{g209}	e _{g210}	e _{g211}	e _{g212}	e _{g213}	e _{g214}	e _{g215}	e _{g216}	e _{g217}	e _{g218}	e _{g219}	e _{g220}	e _{g221}	e _{g222}	e _{g223}	e _{g224}	e _{g225}	e _{g226}	e _{g227}	e _{g228}	e _{g229}	e _{g230}	e _{g231}	e _{g232}	e _{g233}	e _{g234}	e _{g235}	e _{g236}	e _{g237}	e _{g238}	e _{g239}	e _{g240}	e _{g241}	e _{g242}	e _{g243}	e _{g244}	e _{g245}	e _{g246}	e _{g247}	e _{g248}	e _{g249}	e _{g250}	e _{g251}	e _{g252}	e _{g253}	e _{g254}	e _{g255}	e _{g256}	e _{g257}	e _{g258}	e _{g259}	e _{g260}	e _{g261}	e _{g262}	e _{g263}	e _{g264}	e _{g265}	e _{g266}	e _{g267}	e _{g268}	e _{g269}	e _{g270}	e _{g271}	e _{g272}	e _{g273}	e _{g274}	e _{g275}	e _{g276}	e _{g277}	e _{g278}	e _{g279}	e _{g280}	e _{g281}	e _{g282}	e _{g283}	e _{g284}	e _{g285}	e _{g286}	e _{g287}	e _{g288}	e _{g289}	e _{g290}	e _{g291}	e _{g292}	e _{g293}	e _{g294}	e _{g295}	e _{g296}	e _{g297}	e _{g298}	e _{g299}	e _{g300}	e _{g301}	e _{g302}	e _{g303}	e _{g304}	e _{g305}	e _{g306}	e _{g307}	e _{g308}	e _{g309}	e _{g310}	e _{g311}	e _{g312}	e _{g313}	e _{g314}	e _{g315}	e _{g316}	e _{g317}	e _{g318}	e _{g319}	e _{g320}	e _{g321}	e _{g322}	e _{g323}	e _{g324}	e _{g325}	e _{g326}	e _{g327}	e _{g328}	e _{g329}	e _{g330}	e _{g331}	e _{g332}	e _{g333}	e _{g334}	e _{g335}	e _{g336}	e _{g337}	e _{g338}	e _{g339}	e _{g340}	e _{g341}	e _{g342}	e _{g343}	e _{g344}	e _{g345}	e _{g346}	e _{g347}	e _{g348}	e _{g349}	e _{g350}	e _{g351}	e _{g352}	e _{g353}	e _{g354}	e _{g355}	e _{g356}	e _{g357}	e _{g358}	e _{g359}	e _{g360}	e _{g361}	e _{g362}	e _{g363}	e _{g364}	e _{g365}	e _{g366}	e _{g367}	e _{g368}	e _{g369}	e _{g370}	e _{g371}	e _{g372}	e _{g373}	e _{g374}	e _{g375}	e _{g376}	e _{g377}	e _{g378}	e _{g379}	e _{g380}	e _{g381}	e _{g382}	e _{g383}	e _{g384}	e _{g385}	e _{g386}	e _{g387}	e _{g388}	e _{g389}	e _{g390}	e _{g391}	e _{g392}	e _{g393}	e _{g394}	e _{g395}	e _{g396}	e _{g397}	e _{g398}	e _{g399}	e _{g400}	e _{g401}	e _{g402}	e _{g403}	e _{g404}	e _{g405}	e _{g406}	e _{g407}	e _{g408}	e _{g409}	e _{g410}	e _{g411}	e _{g412}	e _{g413}	e _{g414}	e _{g415}	e _{g416}	e _{g417}	e _{g418}	e _{g419}	e _{g420}	e _{g421}	e _{g422}	e _{g423}	e _{g424}	e _{g425}	e _{g426}	e _{g427}	e _{g428}	e _{g429}	e _{g430}	e _{g431}	e _{g432}	e _{g433}	e _{g434}	e _{g435}	e _{g436}	e _{g437}	e _{g438}	e _{g439}	e _{g440}	e _{g441}	e _{g442}	e _{g443}	e _{g444}	e _{g445}	e _{g446}	e _{g447}	e _{g448}	e _{g449}	e _{g450}	e _{g451}	e _{g452}	e _{g453}	e _{g454}	e _{g455}	e _{g456}	e _{g457}	e _{g458}	e _{g459}	e _{g460}	e _{g461}	e _{g462}	e _{g463}	e _{g464}	e _{g465}	e _{g466}	e _{g467}	e _{g468}	e _{g469}	e _{g470}	e _{g471}	e _{g472}	e _{g473}	e _{g474}	e _{g475}	e _{g476}	e _{g477}	e _{g478}	e _{g479}	e _{g480}	e _{g481}	e _{g482}	e _{g483}	e _{g484}	e _{g485}	e _{g486}	e _{g487}	e _{g488}	e _{g489}	e _{g490}	e _{g491}	e _{g492}	e _{g493}	e _{g494}	e _{g495}	e _{g496}	e _{g497}	e _{g498}	e _{g499}	e _{g500}	e _{g501}	e _{g502}	e _{g503}	e _{g504}	e _{g505}	e _{g506}	e _{g507}	e _{g508}	e _{g509}	e _{g510}	e _{g511}	e _{g512}	e _{g513}	e _{g514}	e _{g515}	e _{g516}	e _{g517}	e _{g518}	e _{g519}	e _{g520}	e _{g521}	e _{g522}	e _{g523}	e _{g524}	e _{g525}	e _{g526}	e _{g527}	e _{g528}	e _{g529}	e _{g530}	e _{g531}	e _{g532}	e _{g533}	e _{g534}	e _{g535}	e _{g536}	e _{g537}	e _{g538}	e _{g539}	e _{g540}	e _{g541}	e _{g542}	e _{g543}	e _{g544}	e _{g545}	e _{g546}	e _{g547}	e _{g548}	e _{g549}	e _{g550}	e _{g551}	e _{g552}	e _{g553}	e _{g554}	e _{g555}	e _{g556}	e _{g557}	e _{g558}	e _{g559}	e _{g560}	e _{g561}	e _{g562}	e _{g563}	e _{g564}	e _{g565}	e _{g566}	e _{g567}	e _{g568}	e _{g569}	e _{g570}	e _{g571}	e _{g572}	e _{g573}	e _{g574}	e _{g575}	e _{g576}	e _{g577}	e _{g578}	e _{g579}	e _{g580}	e _{g581}	e _{g582}	e _{g583}	e _{g584}	e _{g585}	e _{g586}	e _{g587}	e _{g588}	e _{g589}	e _{g590}	e _{g591}	e _{g592}	e _{g593}	e _{g594}	e _{g595}	e _{g596}	e _{g597}	e _{g598}	e _{g599}	e _{g600}	e _{g601}	e _{g602}	e _{g603}	e _{g604}	e _{g605}	e _{g606}	e _{g607}	e _{g608}	e _{g609}	e _{g610}	e _{g611}	e _{g612}	e _{g613}	e _{g614}	e _{g615}	e _{g616}	e _{g617}	e _{g618}	e _{g619}	e _{g620}	e _{g621}	e _{g622}	e _{g623}	e _{g624}	e _{g625}	e _{g626}	e _{g627}	e _{g628}	e _{g629}	e _{g630}	e _{g631}	e _{g632}	e _{g633}	e _{g634}	e _{g635}	e _{g636}	e _{g637}	e _{g638}	e _{g639}	e _{g640}	e _{g641}	e _{g642}	e _{g643}	e _{g644}	e _{g645}	e _{g646}	e _{g647}	e _{g648}	e _{g649}	e _{g650}	e _{g651}	e _{g652}	e _{g653}	e _{g654}	e _{g655}	e _{g656}	e _{g657}	e _{g658}	e _{g659}	e _{g660}	e _{g661}	e _{g662}	e _{g663}	e _{g664}	e _{g665}	e _{g666}	e _{g667}	e _{g668}	e _{g669}	e _{g670}	e _{g671}	e _{g672}	e _{g673}	e _{g674}	e _{g675}	e _{g676}	e _{g677}	e _{g678}	e _{g679}	e _{g680}	e _{g681}	e _{g682}	e _{g683}	e _{g684}	e _{g685}	e _{g686}	e _{g687}	e _{g688}	e _{g689}	e _{g690}	e _{g691}	e _{g692}	e _{g693}	e _{g694}	e _{g695}	e _{g696}	e _{g697}	e _{g698}	e _{g699}	e _{g700}	e _{g701}	e _{g702}	e _{g703}	e _{g704}	e _{g705}	e _{g706}	e _{g707}	e _{g708}	e _{g709}	e _{g710}	e _{g711}	e _{g712}	e _{g713}	e _{g714}	e _{g715}	e _{g716}	e _{g717}	e _{g718}	e _{g719}	e _{g720}	e _{g721}	e _{g722}	e _{g723}	e _{g724}	e _{g725}	e _{g726}	e _{g727}	e _{g728}	e _{g729}	e _{g730}	e _{g731}	e _{g732}	e _{g733}	e _{g734}	e _{g735}	e _{g736}	e _{g737}	e _{g738}	e _{g739}	e _{g740}	e _{g741}	e _{g742}	e _{g743}	e _{g744}	e _{g745}	e _{g746}	e _{g747}	e _{g748}	e _{g749}	e _{g750}	e _{g751}	e _{g752}	e _{g753}	e _{g754}	e _{g755}	e _{g756}	e _{g757}	e _{g758}	e _{g759}	e _{g760}	e _{g761}	e _{g762}	e _{g763}	e _{g764}	e _{g765}	e _{g766}	e _{g767}	e _{g768}	e _{g769}	e _{g770}	e _{g771}	e _{g772}	e _{g773}	e _{g774}	e _{g775}	e _{g776}	e _{g777}	e _{g778}	e _{g779}	e _{g780}	e _{g781}	e _{g782}	e _{g783}	e _{g784}	e _{g785}	e _{g786}	e _{g787}	e _{g788}	e _{g789}	e _{g790}	e _{g791}	e _{g792}	e _{g793}	e _{g794}	e _{g795}	e _{g796}	e _{g797}	e _{g798}	e _{g799}	e _{g800}	e _{g801}	e _{g802}	e _{g803}	e _{g804}	e _{g805}	e _{g806}	e _{g807}	e _{g808}	e _{g809}	e _{g810}	e _{g811}	e _{g812}	e _{g813}	e _{g814}	e _{g815}	e _{g816}	e _{g817}	e _{g818}	e _{g819}	e _{g820}	e _{g821}	e _{g822}	e _{g823}	e _{g824}	e _{g825}	e _{g826}	e _{g827}	e _{g828}	e _{g829}	e _{g830}	e _{g831}	e _{g832}	e _{g833}	e _{g834}	e _{g835}	e _{g836}	e _{g837}	e _{g838}	e _{g839}	e _{g840}	e _{g841}	e _{g842}	e _{g843}	e _{g844}	e _{g845}	e _{g846}	e _{g847}	e _{g848}	e _{g849}	e _{g850}	e _{g851}	e _{g852}	e _{g853}	e _{g854}	e _{g855}	e _{g856}	e _{g857}	e _{g858}	e _{g859}	e _{g860}	e _{g861}	e _{g862}	e _{g863}	e _{g864}	e _{g865}	e _{g866}	e _{g867}	e _{g868}	e _{g869}	e _{g870}	e _{g871}	e _{g872}	e _{g873}	e _{g874}	e _{g875}	e _{g876}	e _{g877}	e _{g878}	e _{g879}	e _{g880}	e _{g881}	e _{g882}	e _{g883}	e _{g884}	e _{g885}	e _{g886}	e _{g887}	e _{g888}	e _{g889}	e _{g890}	e _{g891}	e _{g892</}
--	-----------------	-----------------	-----------------	-----------------	-----------------	-----------------	-----------------	-----------------	-----------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	------------------------

ölçülmesi amacı ile çapraz doğrulama yöntemi uygulanmaktadır. Tez çalışmasında da çapraz doğrulama yöntemi ile elde edilen sonuçların başarısı ölçülmüştür. Çapraz doğrulama yöntemi elde edilen ve normalde elde edilmesi gereken sonuçların karşılaştırılması işlemine dayanmaktadır. Çapraz doğrulama ile ilgili geniş bilgi web madenciliği kısmında detaylı olarak işlenmiştir. Elde edilen sonuçlar için KNN yöntemine göre K değerinin belirlenmesi gerekmektedir. Çapraz doğrulama yöntemi K değerinin belirlenmesinde yararlanılabilecek yöntem arasında yer almaktadır. K değerinin belirlenmesi çalışmanın amacına, büyüklüğüne bağlı olarak kişi tarafından deneme yanılma yöntemi ile belirlenebildiği gibi, çapraz doğrulama ile elde edilen duyarlılık sonuçları dikkate alınarak ta saptanabilmektedir.

Ontolojik olarak desteklenmiş sözlük yardımı elde edilen frekans değerleri kullanılarak sayfalar arasındaki Kosinüs benzerlik yöntemi sonuçları aşağıdaki çizelgede toplu halde verilmektedir. KNN sınıflandırma işleminin yapısı gereği çalışma için uygun K değerinin seçilmesi amacı ile 1-12 arası K değerleri duyarlılık sonuçları incelenmiştir (Çizelge 4.11).

Çizelge 4. 11: Ontolojik olarak desteklenmiş olan web sayfalarının kullanım türlerinin tespitinde farklı K değerleri için duyarlılık sonuçları çizelgesi.

	EĞİTİM	SAĞLIK	TİCARET	TURİZM	SPOR	KÜLTÜR	TOPLAM	
K=1								DUYARLILIK
EĞİTİM	4			2			6	66,66666667
SAĞLIK		6				1	7	85,71428571
TİCARET			3	1	1		5	60
TURİZM				4		2	6	66,66666667
SPOR				1	2	1	4	50
KÜLTÜR			1			4	5	80

ORTALAMA 68,17460317

K=2								DUYARLILIK
EĞİTİM	4		2				6	66,66666667
SAĞLIK		6				1	7	85,71428571
TİCARET			3	1	1		5	60
TURİZM				4		2	6	66,66666667
SPOR				1	2	1	4	50
KÜLTÜR			1			4	5	80

ORTALAMA 68,17460317

K=3								DUYARLILIK
EĞİTİM	4		2				6	66,66666667
SAĞLIK		7					7	100
TİCARET			3	1	1		5	60
TURİZM				4		2	6	66,66666667
SPOR				1	3		4	75
KÜLTÜR						5	5	100

ORTALAMA 78,05555556

K=4								DUYARLILIK
EĞİTİM	6						6	100
SAĞLIK		7					7	100
TİCARET			3	1	1		5	60
TURİZM				4		2	6	66,66666667
SPOR				1	3		4	75
KÜLTÜR						5	5	100

ORTALAMA 83,61111111

K=5								DUYARLILIK
EĞİTİM	5		1				6	83,33333333
SAĞLIK		7					7	100
TİCARET			3	2			5	60
TURİZM				4		2	6	66,66666667
SPOR				1	2	1	4	50
KÜLTÜR						5	5	100

ORTALAMA 76,66666667

K=6								DUYARLILIK
EĞİTİM	5		1				6	83,33333333
SAĞLIK		7					7	100
TİCARET			4	1			5	80
TURİZM				4		2	6	66,66666667
SPOR				2	1	1	4	25
KÜLTÜR						5	5	100

ORTALAMA 75,83333333

Çizelge 4. 11 (Devam): Ontolojik olarak desteklenmiş olan web sayfalarının kullanım türlerinin tespitinde farklı K değerleri için duyarlılık sonuçları çizelgesi.

	EĞİTİM	SAĞLIK	TİCARET	TURİZM	SPOR	KÜLTÜR	TOPLAM	
K=7								DUYARLILIK
EĞİTİM	5		1				6	83,33333333
SAĞLIK		7					7	100
TİCARET			2	2	1		5	40
TURİZM				4		2	6	66,66666667
SPOR				1	2	1	4	50
KÜLTÜR						5	5	100

ORTALAMA 73,33333333

K=8								DUYARLILIK
EĞİTİM	5		1				6	83,33333333
SAĞLIK		7					7	100
TİCARET			2	2	1		5	40
TURİZM				4		2	6	66,66666667
SPOR					2	1	3	66,66666667
KÜLTÜR						5	5	100

ORTALAMA 76,11111111

K=9								DUYARLILIK
EĞİTİM	5	1					6	83,33333333
SAĞLIK		7					7	100
TİCARET			2	2	1		5	40
TURİZM				4		2	6	66,66666667
SPOR				2	2		4	50
KÜLTÜR			1			4	5	80

ORTALAMA 70

K=10								DUYARLILIK
EĞİTİM	5	1					6	83,33333333
SAĞLIK		7					7	100
TİCARET			2	1	2		5	40
TURİZM				4		2	6	66,66666667
SPOR				2	2		4	50
KÜLTÜR			1			4	5	80

ORTALAMA 70

K=11								DUYARLILIK
EĞİTİM	5	1					6	83,33333333
SAĞLIK		7					7	100
TİCARET			2	2	1		5	40
TURİZM				4		2	6	66,66666667
SPOR				2	2		4	50
KÜLTÜR			1			4	5	80

ORTALAMA 70

K=12								DUYARLILIK
EĞİTİM	5	1					6	83,33333333
SAĞLIK		7					7	100
TİCARET			2	3			5	40
TURİZM				4		2	6	66,66666667
SPOR				2	2		4	50
KÜLTÜR			1			4	5	80

ORTALAMA 70

Toplamda 33 adet kullanım türü bilinmeyen web sayfası değerlendirmeye alınmıştır. Ortaya çıkan sonuçlar web sayfalarının dolayısı ile içermiş oldukları adres verilerinin kullanım türlerinin saptanmasında kullanılmaktadır (Çizelge 4.11). $K=1$ değeri için benzerlik oranları dikkate alınarak K-NN sınıflandırması yapıldığında 4 sayfanın kesin eğitim sayfası olarak gözüktüğü görülmektedir. En yüksek duyarlılık değerinin % 80 ile 4 web sayfasının tam olarak doğru tespit edildiği ve bunların kültür sayfaları olduğu sonucu ortaya çıkmaktadır. Ortalama duyarlılık değeri incelendiği zaman $K=1$ yakınlık değeri için % 68’lik bir oran yakalanmaktadır.

$K=2$ değeri için benzerlik oranları dikkate alınarak K-NN sınıflandırması yapıldığında $K=1$ değerinde olduğu gibi 4 sayfa eğitim sayfası olarak sınıflandırılmaktadır. En yüksek duyarlılık değerinin % 85 ile 6 web sayfasının tam olarak doğru tespit edildiği ve bunların sağlık sayfaları olduğu sonucu ortaya çıkmaktadır. 33 sayfada $K=2$ değeri için ortalama duyarlılık değeri % 68 dir. $K=3$ değeri için benzerlik oranları dikkate alınarak K-NN sınıflandırması yapıldığında en yüksek duyarlılık değerinin % 100 ile 7 web sayfasının tam olarak doğru tespit edildiği ve bunların sağlık sayfaları olduğu sonucu ortaya çıkmaktadır. Ortalama duyarlılık değeri incelendiği zaman %78 lik bir oran yakalanmaktadır. $K=4$ değeri için benzerlik oranları dikkate alınarak K-NN sınıflandırması yapıldığında 6 sayfanın kesin eğitim sayfası olarak gözüktüğü görülmektedir. En yüksek duyarlılık değerinin % 100 ile 7 web sayfasının tam olarak doğru tespit edildiği sağlık, 6 web sayfası doğru olarak tespit edilen eğitim ve 5 web sayfası tam olarak tespit edilen kültür olduğu sonucu ortaya çıkmaktadır. Ortalama duyarlılık değeri incelendiği zaman % 83,6 lik bir oran yakalanmaktadır. $K=5$ değeri için ortalama duyarlılık değeri incelendiği zaman % 76,7 lik bir oran yakalanmaktadır. Sırası ile diğer ortalama duyarlılık değerlerinin % 70 ile % 76,11 arasında değiştiği görülmektedir.

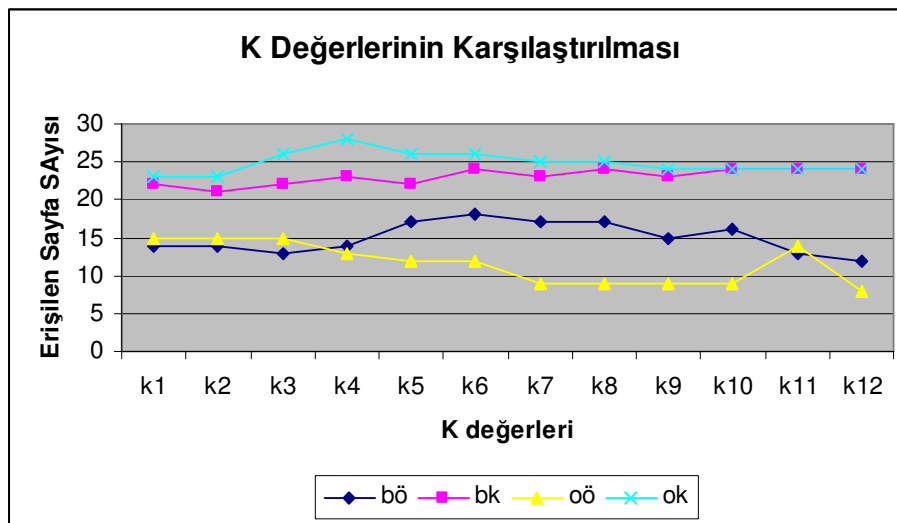
Farklı K değerleri dikkate alınarak duyarlılık ve ortalama duyarlılık sonucu değerlendirildiğinde en yüksek erişimin K değerinin 4 olarak kabul edilmesi durumunda sağlandığı görülmüştür. Bu durumda çalışmada $K=4$ olarak seçilebilir ve çok daha fazla web sayfası için $K=4$ değeri alınarak K-NN işlemi yapılabilir.

4.6 Belediye ve ontolojik olarak desteklenmiş verilerin kullanım türleri sonuçlarının karşılaştırılması

Çizelge 4. 12: K değerleri ve belediye ve ontolojik destekli sözlüklere göre doğru olarak erişilen sayfa sayıları.

	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10	K11	K12
bö	14	14	13	14	17	18	17	17	15	16	13	12
bk	22	21	22	23	22	24	23	24	23	24	24	24
oö	15	15	15	13	12	12	9	9	9	9	14	8
ok	23	23	26	28	26	26	25	25	24	24	24	24

bö= belediye Öklid, bk=belediye Kosinüs, oö=ontolojik Öklid, ok=ontolojik kosinüsü ifade etmektedir. En yüksek erişimin K=4 ile ontolojik Kosinüs işlemi sonucunda olduğu görülmektedir. En düşük erişim ise ontolojik Öklid işlemi ile K=12 seçilmesi durumunda olduğu sonucu ortaya çıkmaktadır (Çizelge 4.12).



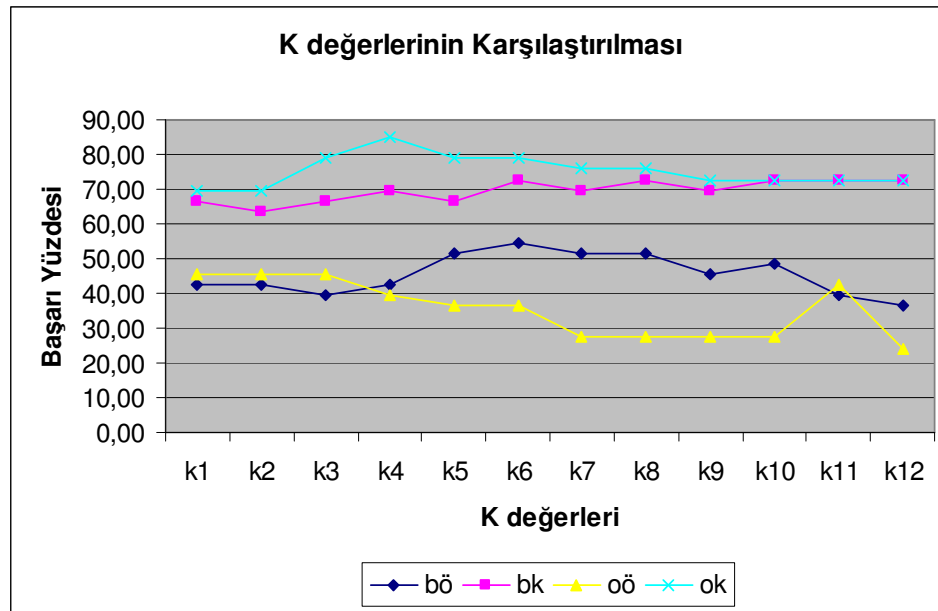
Şekil 4. 41: K değerlerinin ve erişilen sayfaların karşılaştırılması.

Yakınlık değeri 1 ile 12 arasında olan doğru sayfalara erişim grafiği incelendiği zaman K=4 te en yüksek erişimin olduğu görülmüştür. Kosinüs yöntemi kullanılarak yapılan çalışmada K değerinin 8 ve 12 arasındaki değerleri için belediye veya ontolojik olarak desteklenmiş sözlük kullanılarak yapılan sınıflandırma işlem sonuçları örtüşmektedir (Şekil 4.41).

Çizelge 4. 13: Duyarlılık değerleri.

	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10	K11	K12
bö	42,42	42,42	39,39	42,42	51,52	54,55	51,52	51,52	45,45	48,48	39,39	36,36
bk	66,67	63,64	66,67	69,70	66,67	72,73	69,70	72,73	69,70	72,73	72,73	72,73
oö	45,45	45,45	45,45	39,39	36,36	36,36	27,27	27,27	27,27	27,27	42,42	24,24
ok	69,70	69,70	78,79	84,85	78,79	78,79	75,76	75,76	72,73	72,73	72,73	72,73

Çizelge 4.13 de duyarlılık değerleri sonuçları toplu olarak sunulmaktadır. Buna göre Belediye kelimelerinin web sayfalarına erişimde yetersiz kaldığı görülmektedir. Ayrıca Öklid yönteminin yapılan çalışma ve benzerleri için uygun olmadığı da ortaya çıkmaktadır. Belediye kelimelerinin ana obje grupları değiştirilmeden ontolojik olarak desteklenmesi sonucunda başarının çok daha fazla arttığı görülmektedir.



Şekil 4. 42: K değerlerinin ve duyarlılık yüzdelerinin karşılaştırılması.

Şekil 4.42 de de görüldüğü gibi şekle göre başarı yüzdeleri adı altında verilen duyarlılık sonuçlarının K=4 ve ontolojik olarak desteklenen bir sözlükle Kosinüs yönteminin uygulanması ile arttırılabileceği açıktır.

K=4 seçilmesi durumunda 33 web sayfasını birbirlerine göre sınıflandırma doğrulukları Çizelge 4.14. te hata matrisi ile gösterilmektedir.

Çizelge 4. 14: K=4 için hata matrisi.

GERÇEK SINIF	TAHMİN EDİLEN						HATA
	EĞİTİM	SAĞLIK	TİCARET	TURİZM	SPOR	KÜLTÜR	
EĞİTİM	6	0	0	0	0	0	0,00
SAĞLIK	0	7	0	0	0	0	0,00
TİCARET	0	0	3	1	1	0	0,40
TURİZM	0	0	0	4	0	2	0,33
SPOR	0	0	0	1	3	0	0,25
KÜLTÜR	0	0	0	0	0	5	0,00
						ORTALAMA:	0,16

Yukarıdaki tabloda görüldüğü üzere en fazla hata oluşabilen sınıf “Ticaret” olarak gözükmemektedir. Daha sonra “Turizm” amaçlı işyerleri gelmektedir. Bu tip hatalara kelimelerin çok daha iyi belirlenmemesi ve genelde turizm şirketlerinin de ticaret sınıfından sayılması neden olarak gösterilebilir.

Çizelge 4. 15: K=4 için duyarlılık ve anma matrisi.

K=4	EĞİTİM	SAĞLIK	TİCARET	TURİZM	SPOR	KÜLTÜR	TOPLAM	DUYARLILIK (PRECISION)
EĞİTİM	6						6	100
SAĞLIK		7					7	100
TİCARET			3	1	1		5	60
TURİZM				4		2	6	66,66
SPOR				1	3		4	75
KÜLTÜR						5	5	100
ANMA (RECALL)	100	100	100	66,6667	75	71,4286		
ORTALAMA								83,611

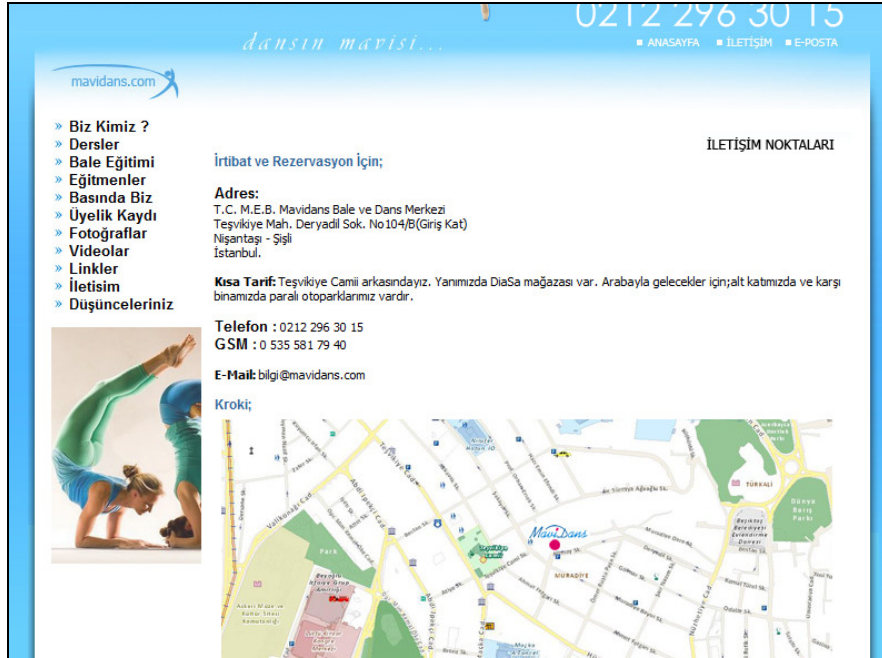
Yukarıdaki çizelgeler ve grafikler değerlendirildiği zaman web sayfalarının kullanım türlerinin belirlenmesinde metin madenciliğinde olduğu gibi benzerlik yaklaşımları çok daha etkin sonuçlar verecektir (Çizelge 4.15). Çalışmada kullanılan Kosinüs benzerlik yöntemi bunlardan birisidir ve sonuçları mesafe değerlerinin karşılaştırılması yöntemine dayanan Öklid yönteminden çok daha iyi sonuçlar ortaya çıkarmaktadır. Bu bağlamda çalışmanın büyültmesi için herhangi bir sakınca bulunmamaktadır. Çok daha fazla web sayfası kullanımı ile benzer çalışmada kapsam genişletilebilir. Ontolojik olarak desteklenen sözlükte uygulama sonuçlarında zayıf kalan turizm ve ticaret gibi obje grupları daha belirgin kavramlar ile desteklenebilir. Daha sonraki çalışmalar için diğer sınıflandırma yöntemleri arasında yer alan Bayes yöntemi ve LSI (Latent Semantic Indexing-Gizli Anlam İndeksi) kullanılabilir.

5. VERİLERİN KARŞILAŞTIRILMASI

Çalışmanın bu bölümünde elde edilen veriler ile belediyenin mevcut verilerinin içeriksel doğruluğu ve tutarlılığı yorumlanmaktadır. Bu amaçla Mavi Dans Bale ve Eğitim Merkezi ve Elite Optik adreslerini içeren rastgele 4 farklı web sayfası örnek olarak incelenmiştir. K-NN sınıflandırması sonucunda en yüksek duyarlılık değerlerinin elde edildiği $K=4$ yakınlık değeri dikkate alınarak ontolojik olarak desteklenmiş belediye sözlüğü kullanılmıştır. Hesaplama yöntemi olarak web sayfalarının kullanım türlerinin çok daha doğru tespit edilmesini sağlayan Kosinüs seçilmiştir.

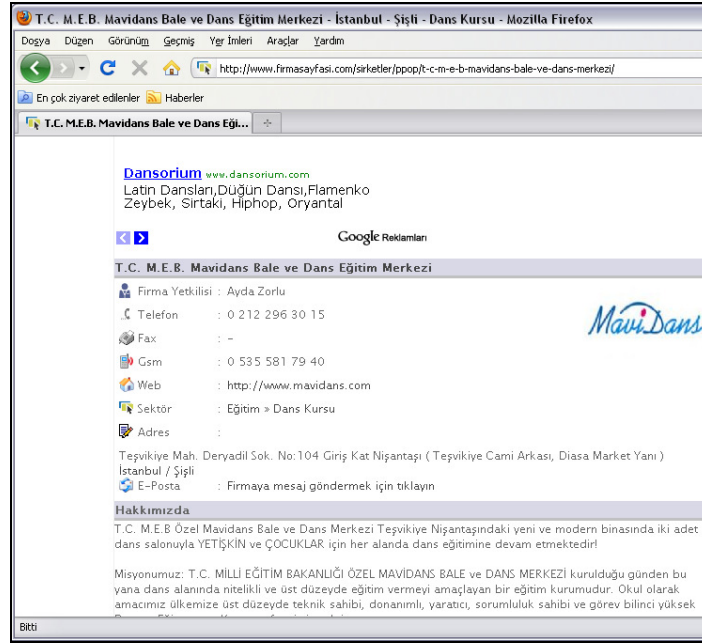
5.1 Örnek uygulama

“Mavi Dans Bale ve Eğitim Merkezi” Şişli ilçesinde Teşvikiye mahallesinde yer almaktadır. Merkeze ait web sitesi (www.mavidans.com) iletişimi sağlayabilmek ve merkez hakkında bilgiler verebilmek amacı ile hazırlanmıştır (Şekil 5.1).



Şekil 5. 1: Mavi Dans Bale ve eğitim merkezi resmi web sayfası.

Web sayfasındaki adres dikkate alındığında merkezin Teşvikiye mahallesinde, Deryadil sokağında ve 104 numaralı kapı numarasında bulunduğu anlaşılmaktadır. www.firmasayfasi.com (Url 6) web sitesinden aynı merkeze ait adres bilgilerine erişim sağlandığında merkezin resmi web sayfası ile aynı olduğu sonucu ortaya çıkmaktadır (Şekil 5.2).



Şekil 5. 2: Adres veritabanı web sayfası (Url6)

Her iki web sitesine ait verilerden faydalanılarak web sitesinin kullanım türünün tespit edilmiştir. Kelime frekanslarının tespitinde ontolojik olarak desteklenmiş belediye sözlüğü kullanılmış benzerliliklerin hesaplanmasında kosinüs yöntemi uygulanmıştır. KNN sınıflandırması için $K=4$ seçilmiştir. Kelime frekanslarının tespit edilmesi işlemi sonucunda merkeze ait web sitesinde “dans” kelimesinin 45 defa yinelendiği görülmektedir. Firma adreslerinin yer aldığı 2. web sayfasına ait kelime frekans bilgileri kapsamında “salon” kelimesinin 8 defa web sayfasında geçtiği görülmektedir (Çizelge 5.1).

Çizelge 5. 1: Mavi Dans Bale ve Dans Merkezi frekans bilgileri.

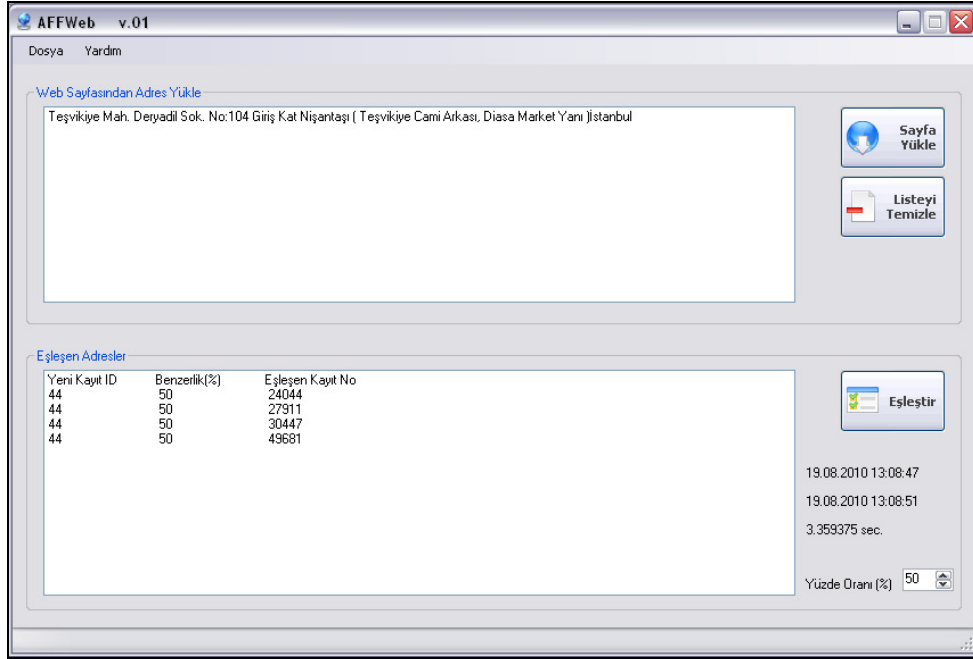
Kelime	Q (frekans)	Kelime	Q (frekans)
Dans	45	Kar	8
Ders	30	Salon	8
Bale	19	Baş	7
Eğitim	19	Bey	6
Firma	13	Çocuk	6
Kur	13	Dönem	6
Mavi	13	Kategori	6
Merkez	12	Kayıt	6
İç	11	Kum	6
Özel	11	Okul	6
İlan	9	Program	5
M.e.b	9	Şirket	5

Kelime sıklıkları yardımı ile oluşturulan matrisler kullanılarak yapılan Kosinüs yaklaşımı ile elde edilen benzerlik değerlerinin 1 sayısına yakınlık değerleri büyükten küçüğe doğru sıralanmıştır (Şekil 5.3). K=4 olarak seçildiği için en yakın 4 değer eğitim ile ilgili olduğu tespit edilmiştir.

	mavidans
eğ2	0,41822
eğ3	0,33985
eğ1	0,31578
eğ4	0,26498
ku1	0,22032
ti1	0,18941
sp2	0,18353
sa4	0,18342
ti4	0,17949
ku3	0,17074
ku4	0,15418
ti3	0,127
ku2	0,12005
tu1	0,1187
ti2	0,11795
sa2	0,11695
sa3	0,11021
sp4	0,10838
tu3	0,1041
tu4	0,10021
tu2	0,09303
sp1	0,08443
sp3	0,08348
sa1	0,07773

Şekil 5. 3: Ontoloji destekli belediye sözlüğü Kosinüs benzerlik sonuçları.

Elde edilen sonuçlar web sayfasının eğitim obje grubuna dâhil olduğunu göstermektedir. Adres bilgilerinin İBB adres bilgileri ile karşılaştırılması işlemi AFFWeb yazılımı kullanılarak gerçekleştirilmiştir.



Şekil 5. 4: Mavi Dans Bale ve Eğitim Merkezi AFFWeb adres eşleşmesi.

AFFWeb yazılımı tez çalışması için geliştirilmiştir. Mavi Dans Bale ve Eğitim Merkezi adres bilgileri İBB veritabanı ile eşleştirme amacı ile bileşenlerine ayrılmaktadır. Adresin % 50 benzerlikle örtüşmesi istenildiğinde 4 farklı adresle benzeştiği ortaya çıkmıştır (Şekil 5.4).

KeyID	ObjekodID	YuzdeOranID	k1	k2	k3	k4	k5	k6	k7	k8	k9
1			Axa	Oyak	Sanat	Galerisi	Şişli	#	#	#	#
2			Eylül	Sanat	Galerisi	Şişli	#	#	#	#	#
3			Milli	Reasurans	Sanat	Galerisi	Teşvikiye	#	#	#	#
4.44051		75	Abide	Hürriyet	İstasyon	183	#	#	#	#	#
5			Ayazağa	Esen	Spor	Klübü	Adres	Atatürk	Kurt	Yolu	1
6			Feniköy	Spor	Klübü	Adres	Feniköy	Barutçular	Sokak	Şişli	7
7			Mecidiyeköy	Esentepe	Caddesi	Ozlem	Pasajı	Mecidiyeköy	9	#	#
8.11		80	FERİKÖY	Al Karanfil	Yüksel	Bayır	Konut	#	#	#	#
34.0		0	MAA	ŞİŞLİ	İSTANBUL	#	#	#	#	#	#
35.46207		75	Maçka	Şişli	İstanbul	#	#	#	#	#	#
30447		50	Teşvikiye	Deryadil	Yanı	İstanbul	#	#	#	#	#

Şekil 5. 5: Verilen adresin bileşenlerine ayrılması.

Adres bilgileri her kelimeye göre bir alanda kayıt altına alınmaktadır. Vt1 adı ile sunulan veritabanı kısmında sadece sokak, cadde vb. bilgilerin önündeki kelimeler dikkate alınmaktadır. Bu nedenle Diasa Market, Teşvikiye Camii arkası gibi bilgiler kayıt altına alınmamaktadır (Şekil 5.5). Yazılım geliştirilebilir özellikte olduğundan bu gibi eksiklikler kolaylıkla daha sonraki çalışmalarda giderilebilecektir. Bu durumda adres bilgilerinin interaktif olarak sistemi sorgulayan kişi tarafından yorumlanması da gerekmektedir.

OBJECTID	MahNoID	MahAdID	KAPI_NO	ESKİ_KAPI_NO	TAHSİS_KAPI_NO	ESKİ_YOL_ADI	KAPI_ADI	YOL_ADI	OBJE_ANA_GRI
29686	297	TEŞVİKİYE	27C			İhlamur Teşvikiye Sk.	Levent Kuaför	Av. Süreyya Ağ	
30037	297	TEŞVİKİYE	15B			Akkavak Sk.		Akkavak Sokağ	
30038	297	TEŞVİKİYE	24A			Akkavak Sk.	Yalçın Kitapevi	Akkavak Sokağ	
30040	297	TEŞVİKİYE	14F			Teşvikiye Cami Sk.	Loxy	Osman F. Sede	
30102	297	TEŞVİKİYE	14E			Teşvikiye Cami Sk.		Osman F. Sede	
30259	297	TEŞVİKİYE				İhlamur Teşvikiye Sk.		Av. Süreyya Ağ	
30447	297	TEŞVİKİYE				Teşvikiye Cd.	İstanbul Moda Akademisi	Teşvikiye Cadd	
30772	297	TEŞVİKİYE				Arslan Süreyya Ağaçlı Sk.	Shop Concept	Av. Süreyya Ağ	
30779	297	TEŞVİKİYE	15A			İhlamur Teşvikiye Yolu		Av. Süreyya Ağ	
30781	297	TEŞVİKİYE	21			Öğretmen Hasım Çeken Cad.	Shell Shop	Hakkı Yatan Cd	
30782	297	TEŞVİKİYE	21C			Güzelbahçe Sk.	Eczane Feyza	Güzel Bahçe S	

Şekil 5. 6: İBB verileri karşılaştırma sonuçları.

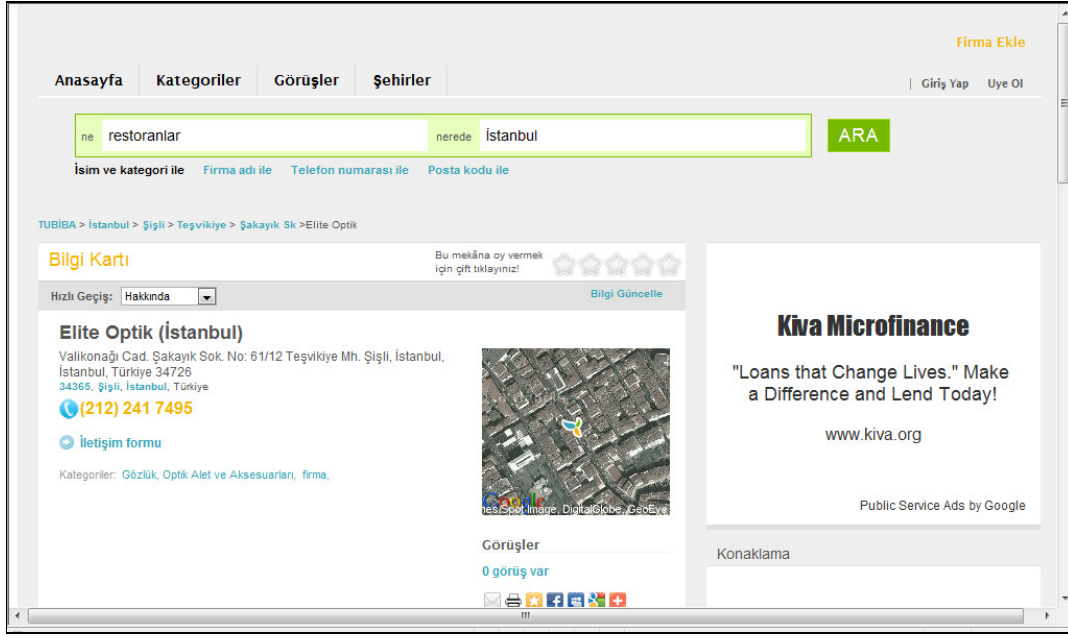
AFFWeb yazılım sonucuna göre 30447 ID numarasına sahip obje ile verilen adres %50 oranında örtüşmektedir (Şekil 5.6).

Çizelge 5. 2: Mavi Dans Bale ve Eğitim Merkezi adres bilgileri ve Adres Bilgi Sistemi veritabanı interaktif karşılaştırma sonucu.

OBJECTID	MahNoID	MahAdID	KAPI_NO	ESKİ_KAPI_NO	TAHSİS_KAPI_NO	ESKİ_YOL_ADI	KAPI_ADI	YOL_ADI	OBJE_ANA_GRP_KOD	OBJE_ANA_GRP_UP_ADI	OBJE_GRP_KOD	OBJE_GRP_P_ADI	OBJE_KOD	OBJE_ADI	X	Y
5239	297	TEŞVİKİYE	104	17		Hostes Rona Altınay Sk.	Şeren Apartmanı	Deryadil Sokağı	4	Genel Kullanım	19	Yapı	52	Konut	415653,41	4546725,309
27745	297	TEŞVİKİYE	104A			Hostes Rona Altınay Sk.	Topoğlu Otopark	Deryadil Sokağı	13	Ulaşım	93	Karayolu Ulaşımı	339	Otopark Kapalı	415657,9076	4546727,104
34503	297	TEŞVİKİYE	104B			Hostes Rona Altınay Sk.		Deryadil Sokağı	4	Genel Kullanım	19	Yapı	51	İşyeri	415655,8878	4546726,235
24651	297	TEŞVİKİYE	104C			Hostes Rona Altınay Sk.	Diasa Market	Deryadil Sokağı	11	Ticaret Alanları, Satış Noktaları, İşyerleri	70	Alışveriş / İş Merkezi	176	Süpermarket	415649,0696	4546724,145

Web sayfasında verilen adres manuel olarak İBB kayıtlarında aranmıştır. Arama sonucunda 104 kapı numarasına sahip olan adreslerin 5239, 27745, 34503 ve 24651 obje ID'lerine sahip oldukları ortaya çıkmıştır. Dans merkezine ait herhangi bir bilgiye rastlanamamıştır. Kayıtlarda görülen en uygun kayıt 104B kapı numarasına sahip 34503 numaralı adrestir (Çizelge 5.2). Genel kullanım adı altında verilen adres kaydında net bir bilgiye erişilememektedir. Ayrıca merkez MEB'e bağımlı olmasına rağmen Şişli eğitim kurumları arasında da yer almamaktadır. Yukarıda verilen sonuçlar ışığında merkezin web sayfası yardımı ile belediye verilerini hazırlayan şirket, kurum ya da kuruluşlar tarafından tekrar gözden geçirilmesi yararlı olacaktır.

Detaylı olarak incelenen diğer bir adres bilgisi olan Elite Optik Teşvikiye mahallesinde yer almaktadır. Yapılan incelemede 2 farklı web sayfası incelenmiştir (Şekil 5.7).



Şekil 5. 7:Tubiba web sayfası (Url 7).



Şekil 5. 8:Firma rehberi web sayfası (Url8).

Elite Optik adres bilgilerinin bulunduğu her iki web sayfası ontoloji destekli belediye sözlüğü yardımı ile Kosinüs yöntemi kullanılarak incelenmiştir (Şekil 5.8). Elde edilen sonuçlar aşağıdaki gibidir. Elit1 ismi ile sunulan Tubiba web sayfası (Url17) kelime frekanslarına göre en yoğun kelimeler firma, göz, optik kelimeleri olarak elde edilmiştir. Elit2 isimli Firma rehberi sayfasındaki (Url 8) firma, bilgi, optik kelimelerine ait frekanslar çok daha fazladır (Çizelge 5.3).

Çizelge 5. 3: Kelime frekansları tablosu.

Kelime	Q(Frekans)	Kelime	Q(Frekans)
Firma	15	Firma	14
Göz	5	Bilgi	6
Optik	5	Optik	6
Ana	3	Rehber	4
Bilgi	2	Tic	4
İletişim	2	Şirket	3
İnternet	2	Telefon	3
Lens	2	Alet	2
Nişan	2	Ana	2
Rehber	2	Foto	2

	O	P	Q	R	S
		elit1		elit2	
ti1	0,27587	ti1	0,32419		
sa3	0,14091	ti3	0,25084		
ti3	0,13409	ti2	0,20455		
ku2	0,12775	eg2	0,13207		
ti4	0,12044	ku2	0,13106		
sp2	0,11222	tu4	0,12998		
sa4	0,11117	ti4	0,10933		
eg2	0,10429	tu3	0,08945		
tu3	0,10334	sp4	0,08733		
tu4	0,09906	sa4	0,07507		
sp4	0,09776	sp2	0,07353		
sa1	0,08789	eg3	0,07126		
sa2	0,07768	ku1	0,06238		
ku4	0,07417	sa3	0,05543		
eg3	0,07215	sa2	0,04951		
ku1	0,06476	tu1	0,04936		
ku3	0,06083	ku4	0,04762		
sp3	0,06019	tu2	0,04321		
ti2	0,05589	sp1	0,04244		
tu2	0,05323	sa1	0,04004		
tu1	0,04644	eg1	0,03369		
sp1	0,04407	ku3	0,03084		
eg1	0,03966	sp3	0,02822		
eg4	0,03406	eg4	0,02129		

Şekil 5. 9: K=4 için Elit1 ve Elit2 web sayfalarının kullanım türlerinin tespiti.

K=4 için her iki web sayfası da ticaret obje grubunda yer almaktadır. Bu durumda web sayfaları ticaret ile ilgilidir denilebilir (Şekil 5.9).

Adres bilgileri incelendiği zaman İBB verileri arasında sadece 38 adet optik ile ilgili adrese rastlanmıştır bu adresler içerisinde yer alan Elite Optik ise farklı kapı numarasına sahiptir (Çizelge 5.4).

Çizelge 5. 4: Elite Optik adres bilgileri ve Adres Bilgi Sistemi veritabanı interaktif karşılaştırma sonucu.

OBJECTID	MahNoID	MahAdID	KAPL_NO	ESKİ_KAPL_NO	TAHSİS_KAPL_N	Q	ESKİ_YOL_ADI	KAPL_ADI	YOL_ADI	OBJE_ANA_GRUP_KOD	OBJE_ANA_GRUP_ADI	OBJE_GRUP_KOD	OBJE_GRUP_ADI	OBJE_KOD	OBJE_ADI	X	Y
37855	297	TEŞVİKİYE	79H				Sakayık Sk.	Elite Optik	Şakayık Sokağı	11	Tic Al. Sat. Nok. İş.	79	1/7/KM*	241	G/O*	415341,2906	4547083,356
33816	297	TEŞVİKİYE	28B				Teşvikiye Cd.	Odak Optik	Teşvikiye Caddesi	11	Tic Al. Sat. Nok. İş.	79	1/7/KM	241	G/O	415262,9412	4546981,633
28443	297	TEŞVİKİYE	36C				Valikonagı Cad.	Murat Optik	Vali Konağı Caddesi	11	Tic Al. Sat. Nok. İş.	79	1/7/KM	241	G/O	415260,7237	4547001,109

İ*:İlaç T*:Tıp KM*:Kimyasal Maddeler G*:Gözlükçü O*:Optisyen

Yukarıdaki eksikliklere rağmen bazı web sayfalarındaki adres bilgileri İBB verileri ile tam olarak örtüşmektedir. Farklı adresler ve İBB verileri ile örtüşüp örtüşmemeleri bilgileri bir tablo halinde sunulmaktadır. www.istanbulburda.com sitesinden (Url9) (Şekil 5.10) alınan özel eğitim kurumları ve dershanelerine ait bir web sayfası incelenmiş ve İBB verileri ile karşılaştırılmıştır (Çizelge 5.5).

The screenshot shows the Şişli website with a sidebar menu on the left containing categories like 'Alışveriş', 'Belediyeler', 'Cafe-Restoran', 'Foto Galeri', 'Hastaneler', 'Konaklama', 'Kültür Sanat', 'Lezzet', 'Lojistik', 'Muhtarlar', 'Müzeler', 'Müzik-Eğlence', 'Okullar', 'Oteller', 'Sinema', 'Sizden Gelenler', 'Tarihi Yerler', 'Tatil', 'Ulaşım', 'Vakıf Ve Dernekler', 'Video Galeri', 'İstanbul Firmaları', 'İstanbul Haritası', and 'Yurtlar'. The main content area is titled 'Şişli' and features a 'KURSLAR' section with a list of courses including 'ŞİŞLİ'DE SOKAK İFTARLARI', 'Şişli Sürücü Kursları', 'Şişli İlköğretim Okulları', 'Uğur Ekey İlköğretim Okulu', 'Şişli Muhtarlıklar', 'Şişli İETT Otobüs Hattı Tarifesi', 'MAĞKA TAŞIŞLA TELEFERİK', 'Şişli Sinema Salonları', 'Şişli WHIRLPOOL SERVİSLERİ', and 'Şişli AEG-Electrolux Yetkili Ser..'. There is also a 'Şişli'de bulunan özel kurslar' section with details about 'Özel Barış D. Özel Eğitim ve Rehabilitasyon Merk.' and contact information. The right sidebar contains 'İç Mimarî Dekorasyon Kursu', 'Kariyer Eğitim Kurumları', 'DOĞA Şişli-Feriköy'de', 'Çalışanlar için İngilizce', and 'Ön+Genel Muhasebe'.

Şekil 5. 10: Web sitesi örneği.

eğ2	0,30761
eğ3	0,29988
ti1	0,29581
eğ4	0,26564
ti3	0,23007
sa4	0,20513
ku4	0,19923
sp2	0,19844
eğ1	0,19525
ku2	0,19284
ku3	0,17331
ku1	0,16953
ti2	0,16551
sa3	0,15424
sp3	0,15298
tu4	0,14743
tu1	0,13802
sp4	0,11217
sa2	0,10719
ti4	0,10595
tu3	0,09296
tu2	0,08799
sp1	0,07319
sa1	0,05648

Şekil 5. 11: K=4 için Kosinüs yöntemi değerlendirilmesi.

K=4 için değerlendirme sonuçlarına göre web sayfası eğitim ile ilgilidir sonucu çıkarılmıştır (Şekil 5.11).

Çizelge 5. 5: Web sayfası adres verileri (Url 9).

Adres	İBB verileri içerisinde VAR/YOK
1 Özel Barış D. Özel Eğitim ve Rehabilitasyon Merk. Büyükdere Cad. Raşit Rıza Sok. No:1 Kat:8-9 Mecidiyeköy/ŞİŞLİ	Yapılan inceleme sonucunda adres bilgisine rastlanamamıştır.
2 Özel Bilginç Bilgisayar Kursu Eski Büyükdere C ad. No:73/4 Şişli	Yapılan inceleme sonucunda adres bilgisine rastlanamamıştır.
3 Özel Birlik Özel Eğitim ve Rehabilitasyon Merkezi Kar kuyusu No:49 Kuştepe /Şişli	Kapı No:41
4 Özel Lotus Yabancı Dil Bilgisayar Kursu Ergenekon Mah. Ölçek Sok. No:106 Harbiye /Şişli	Yapılan inceleme sonucunda adres bilgisine rastlanamamıştır.
5 Özel Mecidiyeköy Çınar Ö. Eğitim ve Rehabilitas M. Selahattin Pınar C. Naci Kasım S. Önügören İş Merkezi No:7 Kat:3 Şiş li	Kapı No:18/İş Merkezi olarak verilmekte.
6 Özel Mircan Akademi Kuaförlük Estetisyenlik Kursu Halaskar Gazi Cad. Yasan İş Merkezi No:243/5 Osmanbey/Şişli	İş merkezi bilgileri var fakat iş merkezi içerisindeki mevcut adresler hakkında İBB verileri arasında bilgi bulunmamakta.
7 Özel Sürekli Eğitim İng. Bilgisay. ve Muh. Kur. Dua Tepe Mah. Ergenekon Cad. Bozkurt S. No:1/3 Kat:4 Kurtuluş/ Şişli	Yapılan inceleme sonucunda adres bilgisine rastlanamamıştır.
8 Özel Turhan Armutçu Yabancı Dil Kursu Halaskar Gazi Cad. No:299-301 Kat:3 Şişli	ObjelD:15513218 /Yol_Adı/Halaskargazi Cad./Apartmanı/Halaskargazi Caddesi/GenelKullanım/Yapı/Konut
9 Özel Yeni Devir Lisan Eğitim Değişim Kursu 2. Taşocağı Cad. Urfalı İş Merkezi No:4/4 Mecidiyekö y	Yapılan inceleme sonucunda adres bilgisine rastlanamamıştır.
10 Özel Yeni Tarz Yabancı Dil Kursu Büyükdere Caddesi Altınova Apt. No:54 Kat:8 Şişli /İSTANBUL	ObjelD:16597216/FULYA/Sisli BüyükdereCad./AltınovaApartmanı/Büyükdere Caddesi/Ticaret Alanları, Satış Noktaları, İşyerleri/Alışveriş / İş Mekezi/İş Merkezi
11 Özel İstanbul Murat Açıköğretim Kursları Şişli Şub Mecidiyeköy Mah. Naci Kasım Sok. No:4	42826213/MECİDİYEKÖY/No:4/NaciKasım Sk./Murat/Açıköğretim/NaciKasımSokağı/EğitimK urumu/ Dershane/Dershane

Çizelge 5. 6 (Devam): Web sayfası adres verileri (Url 9).

Adres	İBB verileri içerisinde VAR/YOK
12 Özel Özgem Özel Eğitim ve Rehabilitasyon Merkezi Gülbahar Mah. Avni Dilligil Sok. No: 12/A Mecidiyeköy/İSTANBUL	Yapılan inceleme sonucunda adres bilgisine rastlanamamıştır.
13 Özel (YADÖM) Yabancı Dil Merkezi Kursu Şişli Şb. Yeniköy Cad. Çevreyolu Çıkışı Gökfiliz İşhanı Kat:3 D:7 Mecidiyeköy Şişli	14194213/MECİDİYEKÖY/811G-33. Sk./Gökfiliz İş Merkezi/Mecidiyeköy Yolu Caddesi/ Ticaret Alanları, Satış Noktaları, İşyerleri/Alışveriş / İş Mekezi/İş Merkezi
14 Özel Başkent İletişim Bilimleri Akademisi Kursu (İst.Şubesi) Büyükdere Cd. No:15/Ab Blok K:1 Hür Han Şiş li	Yapılan inceleme sonucunda adres bilgisine rastlanamamıştır.
15 Özel Deneyim İktisat ve Muhasebe Kursu Fulya Mah. Ortaklar Cad. No:1 Daire 6 Mecidiyeköy Şişli	Yapılan inceleme sonucunda adres bilgisine rastlanamamıştır.
16 Özel Tolga Han Dans Kursu Esentepe Mah. Salih Tozar Sok. No:2/11 Mecidiyeköy Şişli	41882211/ESENTEPE/10A/Salih Tozan Sk. Tolgahan Dans Okulu/Salih Tozan Sokağı/Genel Kullanım/Yapı/İşyeri
17 Özel Tolga Sağ Müzik Kursu 19 Mayıs Cad. Alkarafil Sok. No:6/2 Şişli	31003217/19 MAYIS/5B/Hilal Sk./Tolga Sağ Muzik Kursu/Hilal Sokağı/Ticaret Alanları, Satış Noktaları, İşyerleri/Kağıt / Kitap / Fotoğraf/ Müzik/ Basım/ Müzik Evi (Kaset, CD vb.)
18 Özel Verdil İtalyanca Yabancı Dil Kursu Halaskar Gazi Cad. No:281/2 Osmanbey Şişli	Yapılan inceleme sonucunda adres bilgisine rastlanamamıştır.
19 Özel İlkumut Bilgisayar ve İktisat Kursu Büyükdere Cad. Osman Apt. No:58 Kat.3 D:9 Mecidiyeköy Şişli	16599216/FULYA/58/Sisli Büyükdere Cad.Osman Apartmanı/Büyükdere Caddesi/Genel Kullanım/Yapı /Konut
20 Özel İngilizce Zamanı Dil Kursu Esentepe Mah. Matbuat Sok. Gazeteciler Sitesi No:2 Şiş li	Yapılan inceleme sonucunda adres bilgisine rastlanamamıştır.
21 Özel Şakir Tarkan Sırma Bale ve Dans Kursu Halaskar Gazi Cad. Osmanbey İşhanı No:182/2 Osmanbey Şişli	795298/HALASKARGAZİ/70/182/Halaskargazi Cad./Osmanbey İş Hanı/Halaskargazi CaddesiTicaret Alanları, Satış Noktaları, İşyerleriAlışveriş / İş Mekezi/İşhanı
22 Özel Şevval Özçelik Güzellik Uzmanlığı Yetiş Kursu Avni Dilligil Sok. No:12 D:5/6 Mecidiyeköy/ŞİŞLİ	Yapılan inceleme sonucunda adres bilgisine rastlanamamıştır.

Yapılan alıřmalar sonucunda web sayfalarından elde edilen adreslerin kullanım trlerinin belirlenmesi alıřmaları ortalama % 83,6 lık duyarlılık deęeri edilmiřtir. Adres kullanım fonksiyonu tespit edilirken kullanılan ontolojik yaklařım ve Kosins yntemi alıřmanın doęruluęunu arttırıcı yntemler olarak dikkat ekicidir. Ayrıca yapılan testler sonucunda belirlenen K=4 deęeri web sayfalarının kullanım trlerinin belirlenmesinde doęru sonulara eriřimi saęlamaktadır. alıřmanın dięer bir parası olan adreslerin kazanımı alıřmaları ise bazı problemleri beraberinde getirmektedir. Daha nce de belirtildięi gibi adreslerin bir standarda sahip olmaması, řirketler iin zorunlu adres beyanının bulunmaması alıřmada karřılařılan en byk zorluk olmuřtur. Adres bilgilerinin belediye verilerinde, web sayfalarında farklı kapı numaraları, sokak isimleri ile beyan edilmeleri alıřmayı hazırlayan kurum ve kuruluřlar tarafından dikkatle incelenmesi gereken problemler arasında yer almaktadır. Yapılan incelemeler sonucunda eęitim ile ilgili adreslerde devlet ve zel lise, niversite, ilkokul gibi kurumlarda adresler bakımından herhangi bir eřleřme problemi bulunmamasına raęmen zellikle dershane ve zel kurslarda bazı sorunlar ortaya ıkmaktadır. Karřılařılan sorunlar ařaęıdaki gibidir:

- Web sayfalarındaki kapı numaraları ve Adres Bilgi Sistemindeki kapı numaraları farklı olabilmektedir.
- zellikle iř merkezleri ierisinde yer alan eęitim kurumlarının Adres Bilgi Sistemi ierisinde yer almadıęı grlmřtr.
- Objeye ait kullanım tr kısmı dershane olmasına raęmen genel kullanım olarak geiřtirilmiřtir (izelge 5.5).
- Adres Bilgi Sisteminde dikkat eken dięer bir zellik ok fazla genel kullanım obje trnn yer almasıdır. Detaylandırılabilir adreslerin bulunmasına karřılık bazı adres bilgileri sadece genel kullanım adı altında kayıt altına alınmıřtır.
- alıřma ierisinde saptanan bu durum ok daha fazla benzer rnekler ile oęaltılabilir. Fakat ortaya ıkan durumun ciddiyetinin arařtırılması gerekmektedir.

6. SONUÇLAR

İnternetin günlük hayatın ayrılmaz bir parçası olması ile birlikte, her ortamda sunulan bilginin kullanımı ve paylaşımına yönelik çalışmalar da hızla değişim göstermiştir. Maliyeti oldukça pahalı olan veri setlerini en azından güncellemek ya da tekrar baştan oluşturmak yerine, mevcut bulunan kaynakları çok daha iyi yönetmeye odaklanmış çalışmalar ön plana çıkmaya başlamıştır. Diğer bir ifade ile var olan bilgi açığını kapatmak amacıyla yasal kaynaklardan yararlanma rekabeti de yeniden şekillenmektedir.

Mekânsal verilerin, özellikle yapılandırılmamış biçimde bulunduğu, web dokümanlarından elde edilerek diğer başka çalışmalara destekleyici altlık olarak kullanılması düşüncesi bu bağlamda ele alınan önemli bir araştırma konusudur. Günümüzde sağlıklı ve sürdürülebilir kentsel gelişim için, teknolojinin sunduğu imkânlardan yararlanmanın gerekliliği hem merkezi, hem de yerel yönetimler tarafından kaçınılmaz hale gelmiştir. Bu doğrultudaki Genel Enformasyon Teknolojisi araçları, mekânsal kullanım ihtiyaçları doğrultusunda da geliştirilmeye çalışılmaktadır.

Web sayfalarından hem veri kazanımı hem de bilgi kazanımı bakımından faydalanmak mümkündür. Web sayfalarındaki adres bilgilerine erişim veri kazanımını demektir. Bu verilere ulaşım, çalışmanın ikinci kısmında adres verilerinin kazanımı ve doğruluklarının İBB verileri ile testleri sayesinde yapılmaktadır. Örnek bölge olarak alınan ve 4. bölümün girişinde kısaca değinildiği üzere Şişli İlçesi gibi hızlı büyüyen ve değişen bir bölgede, diğer veriler yanı sıra adres bilgilerinin zamanında güncellenememesi planlayıcılar ve yöneticiler için ciddi bir problem olabilmektedir.

Bilgi kazanımı bakımından incelendiği zaman ise içerisinde adres verilerini barındıran web sayfaları eğer belediyenin obje grupları doğrultusunda sınıflandırılabilir ise, içerisinde barındırdıkları bilgilerin çok daha değer kazanacağı açıktır.

Yapılan çalışmada özellikle belediyelerdeki CBS ve KBS'lerde kullanılan Adres Bilgi Sistemlerinin bilgilerin zenginleştirmesinin yanı sıra, mekânsal arama motorları için gerekli olan mekânsal veri setlerinin desteklenmesi hedeflenmiştir. Bu amaçla web sayfalarından elde edilen mekânsal veri setleri, yapılan tez çalışmasında çok amaçlı incelemeye alınmıştır. Borges ve diğ. yapmış oldukları çalışmada web sayfalarından elde edilen mekânsal veriler, sadece adres kazanımı ve bu adres verilerinin daha sonraki CBS çalışmaların da değerlendirilmek üzere saklanması prensibine dayanmaktadır. Markowetz yapmış olduğu çalışmada Borges ve diğ. yaptığı çalışmaya benzer biçimde adres verilerinin kazanımı üzerinde durmuş ve karşılaşılan zorluklara geniş biçimde değinmişlerdir. Markowetz ve Borges ve diğ. yapmış oldukları çalışmalardaki özellikler incelenerek geliştirilen AFFWeb sayesinde web sayfalarındaki adres verilerine erişim olanaklı hale getirilmiştir. Yapılan analizlerden elde edilen sonuçlar ışığında web sayfalarından adres verilerinin kazanımının mümkün olduğu ve Türkiye için de uygulanabilir olduğu görülmüştür. Fakat adres verilerinin kazanımı esnasında karşılaşılan bir takım güçlükler de yok değildir. Adres kazanımının başarısını düşüren faktörler arasında

- Çizelge 4.4 te belirtildiği gibi adreslerin belirli bir standart yapıya sahip olmaması,
- Adreslerde yapılan değişimlerin web sayfalarına çabuk yansımaması verilebilir.

Adres bilgilerinin kontrolü aşamasında, taşeron firmalar tarafından yapılan ve doğruluğu kabul edilen adres verileri ile web sayfalarından kazanılan adreslerin karşılaştırılması esnasında özellikle değişen kapı numaralarından firmaların haberdar olmaması söz konusudur. Kuruluşların adres bilgilerini müşterilerine erişim amacı ile web sayfalarında eksiksiz olarak sundukları düşünüldüğünde ortaya ciddi bir problem çıkmaktadır. Bu durumun ortaya çıkmasındaki sebeplerden bir tanesi, her kuruma ait web sayfası bulunmasının yasalar tarafından zorunlu tutulmasına karşın, adres bilgilerini doğru olarak vermesinde yasal zorunluluktan ve bir standarttan bahsedilmemesi adres bilgilerinin ciddiyetini azaltmaktadır. Adres bilgilerini toplayan taşeron firmaların ve belediyenin toplu adres bilgilerini, adres arama motorları, şehir rehberleri gibi web sayfalarında sunulmak üzere pazarladıkları düşünüldüğünde depolanan ve sunulan adres bilgilerinin doğruluk ve güncellik bakımından ciddi biçimde mercek altına alınması gerekmektedir.

Genelde yapılan diğerk bir alıřma alanı ise, web sayfalarının ierik bilgileri yardımı ile gruplandırılması, bylece sayfalara eriřimde bir dzenin saėlanmasıdır. Gnmzde pek ok arama motoru ve haber portalları anılan sistemi desteklemektedir.

Yukarıdaki alıřmalar doėrultusunda her iki trde yapılan alıřmaların birleřtirilerek sunulması bu alıřmanın esas konusunu oluřturmaktadır. Web sayfalarından sadece adreslerin kazanılması ile sınırlı kalınmadan, adres bilgilerini ierisinde barındıran web sayfalarının ana konuları da tespit edilerek, adreslerin kullanım trlerinin de tespiti yapılmaktadır. Elde edilen verilerin doėruluėu, İBB ye ait doėruluėu kesin olarak kabul edilen Adres Bilgi Sistemi ile kontrol edilmektedir.

İki farklı ařamadan oluřan alıřmanın sonularına gre farklı kombinasyonlar ortaya ıkmaktadır. 1. Ařamada web sayfalarının fonksiyonları veya diğerk bir ifade ile kullanım trleri tespit edilmektedir. 2. Ařamada ise kullanım trleri tespit edilen web sayfalarından adres bilgileri kazanılmaktadır. Bu durumda ařaėıdaki gibi sonular ortaya ıkabilir,

- Web kullanım tr doėru tespit edilmiř, adres bilgisi doėru kazanılmıř meknsal bilgiler,
- Web kullanım tr doėru tespit edilmiř, adres bilgisi doėru kazanılmamıř meknsal bilgiler (adres eřleřmesi eksik veya hi olmayan),
- Web kullanım tr doėru tespit edilememiř, adres bilgisi doėru kazanılmıř meknsal bilgiler,
- Web kullanım tr ve adres bilgisi doėru kazanılamamıř meknsal bilgiler ((adres eřleřmesi eksik veya hi olmayan))

alıřmanın 4. blm 3. kısmında analizlerinin yapıldıėı zere web madenciliėi yntemleri kullanılarak, kullanım tr tespit edilmiř sayfaların deėerlendirildiėi alıřmada klid ynteminin Kosins yntemine gre duyarlılık sonuları bakımından ok daha kt sonular verdiėi tespit edilmiřtir. Bu sonulara gre %39 luk bir duyarlılık sz konusudur. Web madenciliėi sonuların lldė bir yntem olan apraz doėrulama yntemine gre duyarlılık sonularının daha fazla olduėu web sayfaları iin doėru iřlemler yapılmıřtır denilebilir. Ayrıca yapılan analizler, belediye obje gruplarından oluřturulan ve kelime sıklıklarının tespit edilmesinde kullanılan

sözlüğün geliştirilmesi ve ontolojik olarak desteklenmesi sonucunda çok daha iyi duyarlılık sonuçlarına erişildiğini göstermiştir. Web sayfalarının içeriğinin incelenmesine dayanan bir çalışma olan web kullanım türü tespiti çalışmasında sözlüğün olabildiğince geniş ve obje grupları bakımından vurgulayıcı kelimelerden oluşturulması duyarlılık sonuçlarını arttırmaktadır. Yapılan çalışmada EK-CD de sunulan ontolojik olarak desteklenmiş kelimelerle ve Kosinüs yöntemi uygulanmış sayfalarda ortalama %83,61 lik duyarlılıkla web sayfalarında kullanım türleri doğru olarak tespit edilmektedir. K-NN sınıflandırma yöntemi kullanılarak yapılan işlemler sonucunda elde edilen duyarlılık değeri eğitim, sağlık, kültür ile ilgili web sayfalarında % 100 doğru sonuçlar vermektedir. Bunun temelinde kullanılan sözlüğün içermiş olduğu ilgili obje gruplarına ait kelimelerin eğitim, sağlık ve kültür ile ilgili konularda yeterli doğrulukta seçilmesidir. Diğer obje grupları için kelimeler çok daha vurgulayıcı olacak şekilde geliştirilebilir.

Adres ile ilgili eşleşmeler dikkate alındığında gerek Adres Bilgi Sistemindeki bazı eksikliklerden gerekse adreslerde standart uygulama zorunluluğu olmasına rağmen uyulmaması gibi nedenlerden dolayı eşleşmelerde oranlar %70-%85 oranında değişmektedir. Eşleşme oranının yazılımda kullanıcıya bırakılması ayrıca eşleşmede istenilen doğruluğun sağlanabilmesini de sağlamaktadır.

6.1 Çalışmanın Katkıları

Web sayfalarından bilgi ve veri kazanımları sayesinde İBB verilerinin çok daha zenginleştirilebileceği, hataların azaltılabileceği elde edilen sonuçlar ile açık ve net olarak görülmektedir. İBB verilerinin pek çok mekânsal arama motorları, CBS ve KBS uygulamalarına destek verdiği düşünüldüğü zaman web madenciliğinin bu alanda bu göz ardı edilemeyecek yeni imkânlar sunabileceğine şüphe yoktur.

Benzer sorunların minimize edilmesinde web madenciliği teknikleri Coğrafi Bilgi Sistemi yöneticilerine yeni olanaklar sunmaktadır.

Web sayfalarından elde edilen mekânsal verinin gerek coğrafi arama motorları gerekse CBS'ler için kullanılabilmesi ile,

- Veri güncelleme,
- Veri kazanım maliyetlerinin düşürülmesi ve

- Başka sistemlere destek gibi alanlarda çözüme önemli katkılar sağlayacağı değerlendirilmektedir. Konuya zenginlik kazandıracak aşağıdaki kriterler ya da araçların da çalışmalara entegre edilmesi ile elde edilecek sonuçların daha da verimli olacağına şüphe yoktur.

Sonuç olarak; mekânsal objeleri kullanım türlerine göre sınıflandırmada başarı oranını arttırmak amacı ile ontolojik yaklaşım kullanılarak kelimelerdeki zenginlik arttırılmalıdır. Çalışmada çizelgeler ve grafikler değerlendirildiği zaman web sayfalarının kullanım türlerinin belirlenmesinde metin madenciliğinde olduğu gibi benzerlik yaklaşımları mesafeye dayalı yaklaşımlardan çok daha etkin sonuçlar verecektir.

6.2 Olası Çalışmalar

Çalışmada kullanılan Kosinüs benzerlik yöntemi bunlardan birisidir ve sonuçları mesafe değerlerinin karşılaştırılması yöntemine dayanan Öklid yönteminden çok daha iyi sonuçları ortaya çıkarmaktadır. Çalışmanın bu doğrultuda büyültmesi için herhangi bir problem gözükmemektedir. Çok daha fazla web sayfası kullanımı yeni çalışmalar için önerilebilir. Ontolojik olarak desteklenen sözlükte uygulama sonuçlarında zayıf kalan turizm ve ticaret gibi obje grupları daha belirgin kavramlar ile desteklenebilir. Daha sonraki çalışmalar için diğer sınıflandırma yöntemleri arasında yer alan Bayes yöntemi ve LSI (Latent Semantic Indexing-Gizli Anlam İndeksi) kullanılabilir. Çalışmada AFFWeb ve web madenciliği uygulaması tek bir yazılım çatısı altında toplanabilir.

Diğer taraftan;

- Web sayfalarında aynı adrese birden fazla rastlanması durumunda Büyükşehir Belediyesine ve örneğin İSKİ'ye ait veri setlerinden mutlaka faydalanılmalıdır.
- Yersel olarak adresler, özellikle elde edilen kapı numaraları kontrol edilmelidir.

Böylece pek çok konuya güvenlik, verimlilik kazandırılması ve dolayısıyla düzenin etkinliği artırılabilir. Bu doğrultudaki çalışmaların disiplinler arası bir ekip tarafından projelendirilmesi üzerinde ciddi olarak durulmalıdır.

KAYNAKLAR

- Ahlers D., and Boll S.,** 2008: “Urban Web Crawling”, *Proceedings Of The First International Workshop On Location And The Web*, Beijing, China, Pages 25-32, Year of Publication: 2008, ISBN:978-1-60558-160-6
- Ahlers, D., and Boll, S.,** 2008. “Retrieving Address-Based Locations From The Web”, *GIR’08*, October 29-30, 2008, Napa Valley, California, USA
- Ayaz R.,** 2007. “Web Madenciliğine Bir Bakış”, Yıldız Teknik Üniversitesi, Bilgisayar Mühendisliği, İstanbul.
- Barb A.S., C. -R. Shyu,** 2008. “Ontology Driven Content Mining and Semantic Queries for Satellite Image Databases”, in *Proc. of International Geoscience and Remote Sensing Symposium*.
- Baykal, A., and Coşkun, C.,** 2009. “Web Madenciliği Teknikleri” Akademik Bilişim 2009, s.59, Şanlıurfa
- Berendt, B., Hotho A., Mladenic D., Someren M., Spiliopoulou M., and Stumme G.,** 2003. “A Roadmap for Web Mining: From Web to Semantic Web”, *Proceedings of the 1st European Web Mining Forum (EMWF 2003)*, Cavtat-Dubrovnik, Croatia, 2003; 1-22.
- Bhatia MPS. and Kumar A.,** 2008. “Information Retrieval And Machine Learning: Supporting Technologies For Web Mining Research And Practice”, *Webology*, Volume 5, Number 2, June, 2008
- Bhatia MPS., and Kumar A.,**2007. “Mining Paradigms: Roadmap For Semantic Web Mining”, *The 31th Annual Conference of the German Classification Society on Data Analysis, Machine Learning, and Applications (GFKI’07)*, March 2007, Freiburg i. Br., Germany.
- Boll S., Jones C., Kansa E., Kishor P., Naaman M., Purves R., Scharl A., and Wilde E.,** 2008. “Location And The Web (Locweb 2008)”, *WWW 2008*, April 21-25, Beijing, China
- Borges A. V. , Leander A. H. F., Medeiros C. B., Silva A. S., and Davis C.,**2007. “Discovering Geographical Locations In Web Pages Using Urban Addresses”, *GIR’07*, Lisboa, Portugal
- Borges, A. V. , Leander, A. H. F., Medeiros, C. B., Silva, A. S., and Davis, C.,** 2003. “The Web As A Data Source For Spatial Databases” *V Simpósio Brasileiro de Geoinformática - GeoInfo 2003*, Campos do Jordão (SP)
- Buyukokten, O., Cho, J., and Garcia-Molina, H.,** 1999. “Exploiting Geographical Location Information Of Web Pages”, *ACM SIGMOD Workshop on The Web and Databases (WebDB’99)*, June 3-4, 1999 , Philadelphia, Pennsylvania.

- Çelik R, N, Özlüdemir, M,T, Doğru A, Ö, Güney C**, 2005. "Mekansal Veri Toplama Teknolojileri Ve Büyük Ölçekli Harita Ve Harita Bilgileri Üretim Yönetmeliği", *TMMOB Harita ve Kadastro Mühendisleri Odası 10. Türkiye Harita Bilimsel ve Teknik Kurultayı*, Ankara
- Cheng G., and Du Q.**, 2008. "Ontology for Geographical Names Management and Retrieval", *2008 International Symposium on Knowledge Acquisition and Modeling*
- Clough P.**, 2005. "Extracting Metadata For Spatially- Aware Information Retrieval On The Internet", *GIR'05*, Bremen, Germany
- Cooley R., Mobasher B., and Srivastava J.**, 1997. "Web Mining: Information And Pattern Discovery On The World Wide Web Tools with Artificial Intelligence", *IEEE International Conference on In Tools with Artificial Intelligence, Proceedings*, Ninth pp. 0558-567.
- Dalli A.**, 2006. "System For Spatio-Temporal Analysis Of Online News And Blogs", *WWW2006*, Edinburg, Scotland
- Daş R., Türkoğlu İ., and Poyraz M.**, 2006. "Genetik Algoritma Yöntemiyle İnternet Erişim Kayıtlarından Bilgi Çıkarılması", *SAÜ Fen Bilimleri Enstitüsü Dergisi* 10. Cilt, 2. Sayı, s. 67-72
- Ding J., Gravano L., and Shivakumar N.**, 2000. "Computing Geographical Scopes of Web Resources", *Proceedings of the 26th VLDB Conference*, Cairo, Egypt
- Egenhofer M. J.**, 2002. "Toward Semantic Geospatial Web", *GIS'02*, November 8-9, 2002, McLean, Virginia, USA
- Elizio C., Campelo C., and Baptista C.**, 2008. "Geographic Scope Modelling For Web Documents", *GIR'08*, Napa Valley, California, USA
- Emel G, and Taşkın Ç.**, 2002. "Genetik Algoritmalar Ve Uygulama Alanları", Uludağ Üniversitesi, *İktisadi İdari Bilimler Fakültesi Dergisi*, CiltXXI, sayı 1, s.129-152
- Etzioni, E.**, 1996. "The World Wide Web: Quagmire Or Gold Mine", *Communcations of the ACM*, 39(11):65-68
- Gelernter J., and Lesk M. E.**, 2008. "Traditional Resources Help Interpret Text", *BooksOnline'08*, Napa Valley, California, USA
- Göbel R., Colaiacomo L, and Cruz A.**, 2006. "A Spatial Search Engine Supporting Security Application", *Proceedings Fourth Conference Image Information Mining, ESA Workshop Proceedings*
- Golemati M., A. Katifori, C. Vassilakis, G. Lepouras, C. Halatsis**, 2007. "Creating An Ontology For The User Profile: Method And Applications". In: *Proceedings of the First IEEE International Conference on Research Challenges in Information Science (RCIS)*, Morocco

- Gruber T R.**, 1993. "Toward principles for the design of ontologies used for knowledge sharing", Originally in N. Guarino and R. Poli, (Eds.), *International Workshop on Formal Ontology, Padova, Italy. Revised August 1993. Published in International Journal of Human-Computer Studies*, Volume 43 , Issue 5-6 Nov./Dec. 1995, Pages: 907-928,
- Gürcan F., and Köse C.**, 2008. "Web Content Mining And Subject Classification (Web İçerik Madenciliği Ve Konu Sınıflandırılması)", *Akademik Bilişim Çanakkale Onsekiz Mart Üniversitesi*
- Henriksson R., T. Kauppinen and E. Hyvönen**, 2008. "Core Geographical Concepts: Case Finnish Geo-Ontology Location and the Web (LocWeb)" 2008 Workshop, *17th International World Wide Web Conference WWW 2008, ACM International Conference Proceeding Series*; Vol. 300, Pages 57-60, Beijing, China
- Heuer J.T., and Dupke S.**, 2007. "Towards a Spatial Search Engine Using Geotags" *GI-Days 2007 - Young Researchers Forum.IfGIprints* 30. ISBN: 978-3-936616-48-4
- Hill L.** 2000. "Core Elements Of Digital Gazetteers: Placenames, Categories, And Footprints", *ECDL 2000*, LNCS 1923, pp.280-290, Springer-Verlag Berlin, Heidelberg
- Hochmair H.**, 2005. "Ontology matching for spatial data retrieval from internet portals", *GeoS 2005*, LNCS 3799, pp. 166-182
- Jones, C., Abdelmoty, A., Finch, D., Fu G., and Vaid, S.**,2004. "The Spirit Spatial Search Engine: Architecture, Ontologies And Spatial Indexing" *Proc Third International Conference on Geographic Information Science GIScience 2004*, Maryland, USA, Lecture Notes in Computer Science 3234,125-139.
- Kaya H. and Köymen K.**, 2008. "Veri Madenciliği Kavramı Ve Uygulama Alanları", *Doğu Anadolu Bölgesi Araştırmaları*
- Kimler M.**, 2004. "Geo-Coding: Recognition Of Geographical References In Unstructured Text, And Their Visualization" *Diplomarbeit, Fachhochschule Hof*
- Kohavi R., Provost F.**, 1998. "Glossary Of Terms", *Machine Learning*, Vol. 30, No. 2/3, pp. 271-274.
- Kömesli M.**, 2007. "Coğrafi Bilgi Sistemlerinde Ontolojilerin Kullanımı" *TMMOB Harita ve Kadastro Mühendisleri Odası Ulusal Coğrafi Bilgi Sistemleri Kongresi, KTÜ, Trabzon*
- Kosala R., and Blockeel H.**, 2000. "Web Mining Research: A Survey", *ACM SIGKDD*
- Kunc M.**,2007. "Web Mining Overview", *Proceedings of the 13th Conference STUDENT EEICT 2007* Volume 4, Brno, CZ, VUT v Brne, p. 391-395, ISBN 80-214-3410-3
- Leidner J. L , G. Sinclair, B. Webber**, 2003. "Grounding Spatial Named Entities For Information Extraction And Question Answering", *Proceedings of the HLT-NAACL Workshop on Analysis of geographic references*, p.31-38

- Lemmens R., and de Vries M.**, 2004. "Semantic Description of Location Based Web Services Using an Extensible Location Ontology" *Proceedings of Münster GI-days In Geoinformation and mobility*, Vol. IFGI prints (2004), pp. 261-262.
- Lemmens R., Granell C., Wytzisk A., de By R., Gould M., and Oosterom P.**, 2006. "Semantic And Syntactic Service Descriptions At Work In Geo-Service Chaining", *9th AGILE Conference on Geographic Information Science*, Visegrád, Hungary
- Lim E. P., and Sun A.**, 2005. "Web Mining - The Ontology Approach" *The International Advanced Digital Library Conference (IADLC)*, Nagoya
- Markowetz A.**, 2004. "Geographic Properties Of Internet Resources", *Diplomarbeit*, Philipps-Universität Marburg Fachbereich Mathematik und Informatik Geographic Properties of Internet Resources
- Markowetz A., Chen Y., Suel T., Long X., and Seeger B.**, 2005. "Design And Implementation Of Geographic Search Engine", *Technical Report*, TR-CIS-2005-03
- McCurley M.** 2001. "Geospatial Mapping And Navigation Of The Web", *WWW10*, May 1-5, 2001, Hong Kong
- Miah M.**, 2009. "Improved K-NN Algorithm For Text Classification", *International Conference on Data Mining (DMIN) 2009*.
- Miller G. A.**, 1995. "Wordnet: A Lexical Database For English", *In Communications of the ACM*, volume 38, pages 39–41, 1995.
- Morimoto, Y., Aono, M., Houle, M., and McCurley, K.**, 2003. "Extracting Spatial Knowledge from the Web", *Proceedings of the 2003 Symposium on Applications and the Internet*, p.326
- Özakar B., and Püskülcü H.**, 2002. "Tikların Dili- Web İçerik Ve Web Kullanım Madenciliği Tekniklerinin Entegrasyonu İle Oluşmuş Bir Veritabanından Nasıl Yararlanılabilir?", *İnternet Konferansı*, ISTANBUL
- Özkan Y.**, 2008. "*Veri Madenciliği Yöntemleri*", 216 s., ISBN 978-975-6797-82-2, Papatya Yayıncılık, İstanbul
- Pater M., Popescu D., and Maştei D.**, 2008. "Pattern Discovery Techniques In Web Mining", *Proceedings of Renewable Sources And Environmental Electro-Technologies RSEE'08*, Romania
- Pol, K.; Patil, N.; Shreya Patankar; Chhaya Das; Datta Meghe Coll. of Eng., Airoli, Mumbai**, 2008. "A Survey On Web Content Mining And Extraction Of Structured And Semistructured Data", *Emerging Trends in Engineering and Technology, 2008. ICETET '08*. First International Conference on Nagpur, Maharashtra
- Popov B., Kiryakov A., Manov D., Kirilov A., Ognyanoff D., and Goranov M.**, 2003. "Towards Semantic Web Information Extraction", *In Proceedings of ISWC (Sundial Resort, Vol. 20 (2003)*
- Salton G., A. Wong, C. S. Yang**, 1975. "A Vector Space Model For Automatic Indexing", *Communications of the ACM*, v.18 n.11, p.613-620

- Shakarmani R., Kedar N. and Khandelwal N.,** 2010. Article: Performance Assessment using Text Mining. International Journal of Computer Applications 1(12):1–6, February 2010. Published By Foundation of Computer Science.
- Shekhar S., Pusheng Z., Yan H., and Ranga R. V.,** 2003. "Trends In Spatial Data Mining", *As A Book Chapter To Appear In "Data Mining: Next Generation Challenges And Future Directions"*, Hillol Kargupta and Anupam Joshi(editors), AAAI/MIT Press, 2003
- Silahtaroglu G.,** 2008. “*Kavram Ve Algoritmalarıyla Temel Veri Madenciliği*”, 174 s., ISBN 978-975-6797-81-5, Papatya Yayıncılık, İstanbul
- Stumme G., Hotho A., and Berendt B.,** 2006. “Semantic Web Mining State Of Art And Future Directions” *Science, Services and Agents on the World Wide Web*, Vol. 4, No. 2. pp. 124-143.
- Tezuka, T., and Tanaka, K.,** 2005. “Landmark Extraction: A Web Mining Approach”, COSIT 2005, LNCS 3693, pp. 379-396
- Tezuka, T., Kurashima, T., and Taka, K.,** 2006. “Toward Tighter Integration Of Web Search With A Geographic Information System”, *WWW2006*, May 23-26, 2006, Edinburg, Scotland
- Url-1** <<http://www.internetworldstats.com/>>, alındığı tarih, 17.07.2009
- Url-2** <http://en.wikipedia.org/wiki/Vector_space_model>, alındığı tarih, 17.07.2009
- Url-3** <http://en.wikipedia.org/wiki/K-nearest_neighbor_algorithm> alındığı tarih, 18.09.2009
- Url-4** <<http://upload.wikimedia.org>>, alındığı tarih 18.09.2009
- Url-5** <<http://www.habertakip.com/> arama motoru>, alındığı tarih 06.09.2010
- Url-6** <www.firmasayfasi.com>, alındığı tarih, 06.09.2010
- Url-7** <<http://tubiba.turkcebilgi.com/elite-optik-tesvikiye>>, alındığı tarih, 06.09.2010
- Url-8** <<http://www.firmarehberi.tv.tr/elite-optik-sanayi-ve-ticaret/adresi-telefon-numarasi.aspx>>, alındığı tarih, 06.09.2010
- Url-9** <http://www.istanbulburda.com/haber_detail.php?id=4226>, alındığı tarih, 06.09.2010
- Waldinger, R., Jarvis, P., and Dungan, J.,** 2003. “Pointing to Places in a Deductive Geospatial Theory”, *In Workshop on Analysis of Geographical References; Human Language Technology Conference*, Edmonton, Canada.
- Wang C., Xie X., Wang L., Lu Y., and Ma W.,** 2005. “Detecting Geographic Locations From Web Resources”, *GIR'05*, Bremen, Germany

EKLER (CD)

EK A. Bina Fonksiyonları kullanım bilgileri

EK B. Şişli İlçesi Mahalle ve Sokak isimleri

EK C. İstanbul Büyükşehir Belediyesi adres bilgi sistemi (Şişli Örneği)

EK D. Belediye obje grupları sözlüğü

EK E. Ontolojik olarak desteklenmiş belediye obje grupları sözlüğü

EK F. Web sayfalarından adres verisi elde etmek için kullanılan kod bilgisi

ÖZGEÇMİŞ

Ad Soyad: Lütfiye KUŞAK

Doğum Yeri ve Tarihi: Çayeli/RİZE 15.12.1976

Lisans Üniversite: Yıldız Teknik Üniversitesi/Jeodezi ve Fotogrametri Mühendisliği

Y. Lisans Üniversite: İstanbul Teknik Üniversitesi/Jeodezi Ve Fotogrametri Mühendisliği

Çalıştığı Kurum: İstanbul Aydın Üniversitesi/Emlak ve Emlak Yönetimi Programında öğretim görevlisi olarak çalışmaktadır.

