ISSN: 2564-6761
Geomatik Dergisi
Journal of Geomatics
Araştırma Makelesi
DOI: 10.29128/geomatik.429344
2019; 4(1);14-22

Web Madenciliği ve Mekânsal İçerik Tespiti

Lütfiye KUŞAK[1*]

[1]İstanbul Aydın Üniversitesi, Mimarlık ve Tasarım Fakültesi, İç Mimarlık Bölümü, İstanbul
(lutfiyekusak@aydin.edu.tr); ORCID ID 0000-0002-7265-245X

Öz
Günümüzde gerek yazılım gerekse donanım alanlarındaki gelişmeler çok daha fazla verinin kayıt altına alınmasını olanaklı hale getirmektedir. Kayıt altına alınan verilerin anlamlı bir şekilde işlenmesi ve analizi güncel konular arasında yer almaktadır. Verilerin elde edilmesi çok maliyetli olduğundan mevcut kaynakların akıllı bir şekilde yönetilmesi giderek önem kazanmaktadır. Bu nedenle sadece bilişim alanında değil, yerel yönetimlerden ticaret sektörüne kadar geniş bir yelpazede teknolojinin sunduğu imkânların çok iyi değerlendirmesi gerekmektedir. Günümüzde sağlıklı ve sürdürülebilir kentsel gelişim için, teknolojinin sunduğu imkânlardan yararlanmanın gerekliliği çok açıktır. Yapılan çalışmada web sayfalarındaki adres bilgilerinin ve içeriklerinin mekânsal bilgilerin desteklenmesinde daha sonra da kontrol amaçlı kullanılıp kullanılamayacağı incelenmiştir. Bu amaçla web sayfalarındaki içeriklerden elde edilen mekânsal veri setleri, yapılan çalışmada çok amaçlı incelemeye alınmıştır. Çalışma bölgesi olarak İstanbul iline bağlı Şişli ilçesi seçilmiştir. Web sayfalarında sunulan adres ve içerik bilgileri web madenciliği yöntemleri ile analiz edilmiştir ve duyarlılık sonucu % 85 olarak bulunmuştur.

Anahtar Kelimeler: *Adres, web madenciliği, K-NN sınıflandırması, mekânsal içerik*

Web Mining and Spatial Content Detection

Abstract
Today, both software and hardware developments make it possible to record much more data. Significant processing and analysis of recorded data are among the actual topics. Since the acquisition of the data is costly, it is becoming increasingly important to manage existing resources in a smart way. For this reason, not only in the field of information, but also in the wide range of local government to the trade sector, it is necessary to evaluate the possibilities of technology. Today, it is clear that for healthy and sustainable urban development, it is necessary to utilize the opportunities provided by technology. In the study, it has been examined whether the address information and contents of the web pages can be used for the purpose of checking and then supporting the spatial information. For this purpose, the spatial data sets obtained from the contents of the web pages are taken for multi-purpose examination in the study. Şişli district of Istanbul province was chosen as the study area. The address and content information presented on the web pages were analyzed by web mining methods and the sensitivity result was found to be 85%.

Keywords: *Address, web mining, KNN classification, spatial content*

* Sorumlu Yazar

Geliş Tarihi: 01.06.2018
Kabul Tarihi: 05.07.2018

Geomatik Dergisi
Journal of Geomatics
2019; 4(1);14-22

1. GİRİŞ

Hayatımızın her alanında önemli bir yere sahip olan internet ortamında, sunulan bilginin en etkin şekilde kullanılması ve paylaşılması kaçınılmazdır. Gelişen teknoloji ile birlikte internette sunulan bilginin önemi daha da artmış; zamanla güncel bilgi ihtiyacının internette sunulan bilgiler kullanılarak ihtiyacın karşılanabileceği olgusu giderek yerleşmiştir.

Son zamanlarda Foursquare, Google Map, Weibo vb. mekânsal tabanlı uygulamaların da artması ile birlikte mekânsal referanslı objelere ait öznitelik verilerinin elde edilmesinde anket vb. yöntemlere ek olarak, web sayfalarındaki verilerin de aktif bir şekilde kullanılabileceği çalışmalar göze çarpmaktadır (Borges vd., 2003; Borges vd., 2007; Tri vd., 2015; Xu vd., 2016). Mevcut verilere ek olarak mekânsal referanslı objelerin özniteliklerinin zenginleştirilmesinde, güncellenmesinde ve iyileştirilmesinde web madenciliği tekniklerinin faydalı olup olmayacağı çalışmanın temel araştırma noktasını oluşturmaktadır.

2000'li yıllara geçişte mekânsal veri setlerinin web sayfalarından elde edilme çalışmaları hız kazanmaya başlamıştır. Bu amaçla yapılan çalışmalar günümüzde de devam etmektedir (Li vd., 2014; Li, 2018).

İnternet tabanlı ticaret ortamının yaygınlaşması ile birlikte, bu işi yapmakta olan büyük ve küçük ölçekli şirketler müşterilerine erişebilmek ve sipariş alabilmek için şirket adres ve telefon bilgilerini web sayfalarında paylaşmaktadırlar. Piyasa analizi yapmak için web sayfalarından elde edilen bilgilerden müşteri potansiyelinin belirlenmesi ve mekânsal bilgiler ile ilişkilendirilmesi çalışmalarına rastlanmaktadır (Buyukokten vd., 1999).

İlk olarak Amazon tarafından sunulan çevrimiçi hizmetler bugün küçük ölçekli kuruluşlar tarafından da benimsenmektedir (Markowetz, 2004).

Mekânsal arama motorları, Google, Yahoo gibi geleneksel arama motorlarının kullandığı benzer işlem sıralarını takip etmektedir. Bu işlemler verilerin depolanması, indekslenmesi ve sorgulanması olarak kısaca sınıflandırılabilir.

Mekânsal arama motorlarında web sayfalarından elde edilen verilerin nasıl indekslenebileceği konusunda çalışmalar buulnmaktadır (Ding vd., 1999; Hill, 2000; McCurley, 2001; Waldinger vd., 2003;; Heuer ve Dupke, 2007). Bununla birlikte geliştirilen sistemlere ek olarak 2005 yılında Markowetz ve arkadaşlarının yaptığı sistemde kazanılan ve indekslenen veriler, koordinat verileri ile de birleştirilmiştir (Markowetz vd., 2005).

Web sayfalarından mekânsal veriler elde edilmesine ve elde edilen bu verilerin gerekli düzenlemeler yapıldıktan sonra bir coğrafi arama motorunda saklanması prensibine dayanan SPIRIT projesi 2002 yılında geliştirilmiştir. Mekânsal veriler sonradan kullanılmak üzere SPIRIT coğrafi arama motorunda kayıt altına alınmaktadır (Jones vd., 2004;Clough, 2005; Gelernter ve Lesk, 2008).

Web sayfalarındaki konumsal bilgilerin Coğrafi Bilgi Sistemleri (CBS) ile ilişkilendirilip kullanılabileceğine yönelik çalışmalar bulunmaktadır (Tezuka vd., 2006). Tezuka ve arkadaşları tarafından hazırlanan çalışmada bu tip verilerin işlenerek daha sonraki çalışmalarda kullanılabileceğine özellikle vurgu yapılmaktadır. Günümüzde blog vb. kişisel web sayfalarından çeşitli mekânsal bilgilere erişmek mümkündür. İnternette ticaret hizmetlerinin gelişmesine paralel olarak mekânsal verinin çok önemli olduğu göze çarpmaktadır. Google, Yandex, MapQuest, Excite.com gibi ticari arama motorları da mekânsal arama motorlarının hizmetlerinin iyileştirilebilmesini sağlamak amacıyla çalışmalarını sürdürmektedirler. Bu tip arama motorları geleneksel arama uygulamalarına ek olarak harita ve web sayfalarını birleştirerek hizmet sunmaktadırlar.

Teknolojinin vazgeçilmez parçamız olduğu günümüzde kentsel gelişim ve yenilemelerde de bu alandan faydalanmak kaçınılmazdır (Boll vd., 2008).

Bölgesel ve ülkesel ölçekte yapılan çalışmaların tersine, dünya çapında yapılmış çalışmalara rastlamak mümkündür. Dalli (2006) çalışmasında blog ve haber sitelerine ait değişken mekânsal verilerin analizini bütün dünya ölçeğinde değerlendirmek suretiyle dünyada medyaya olan ilginin analizini ortaya koymuştur.

Geomatik Dergisi
Journal of Geomatics
2019; 4(1);14-22

Yukarıda sunulan araştırmaların ışığında hazırlanan bu çalışmanın, verilerin güncellenmesinde, kazanım maliyetlerinin azaltılmasında, standartlaştırılmaları durumunda farklı sistemlere destek olabileceği, internet sayfalarında paylaşılan verilerin kontrolü de dahil olmak üzere farklı alanlara katkıda bulunabileceği düşünülmektedir. Bu çalışmada özellikle web sayfalarında paylaşılan verilerin kontrolü üzerinde durulmuştur.

Ülkemizde de diğer ülkelerde olduğu gibi web sayfalarında gerek büyük gerekse küçük şirketler adres bilgilerini ve içeriklerini müşterileri ile paylaşmaktadırlar. 6102 sayılı Türk Ticaret Kanununa (TTK) göre bağımsız denetime tabi olan sermaye şirketleri ticaret siciline tescil edildikten sonra üç ay içerisinde internet sitesi açmak ve bilgilerini paylaşma zorunlulukları bulunmaktadır (TTK, 2018). Bunun yanı sıra Türkiye'de adreslerin kayıt altına alınarak belirli bir standart oluşturulması çalışmaları sürdürülmektedir. İlgili çalışmalar İçişleri Bakanlığı Nüfus ve Vatandaşlık İşleri Genel (NVİGM) Müdürlüğü tarafından yürütülen Mekansal Adres Kayıt Sistemi (MAKS) ile devam etmektedir (MAKS, 2018).

MAKS (Mekânsal Adres Kayıt Sistemi) projesi sayesinde bağımsız denetime tabi olan sermaye şirketlerinin hepsi bir standart koda sahip olacağı için, şirketlerin internet sitelerindeki adres bilgilerinin doğruluğunu kontrol etmek çok daha kolay olacaktır. Bağımsız denetime tabii olan şirketlerin yanı sıra internet sayfası açma zorunluluğu olmayan küçük işletmeler ve kurumlar da adres ve içerik bilgilerini web sayfalarından yayımlamaktadırlar. MAKS çalışmasının bitmesi ve bütün adres bilgilerinin bütünleşmesi sayesinde web sayfalarının içeriklerinde sunulan adres bilgilerinin doğruluğunun en yüksek seviyeye ulaşması beklenilmektedir. Fakat bu kontrol mekanizmalarına rağmen oldukça geniş alana yayılmış olan web sitelerinin içeriklerini kontrol eden sabit bir sistem bulunmamaktadır. Günümüzde Fake Address Generator gibi FAG, 2018) web sayfaları ile gerek oyun amaçlı gerekse başka nedenler ile gerçek adres bilgilerinin kombinasyonundan oluşan gerçek olmayan adresler üretilebilmektedir. Bu tip adreslerin, web sitelerinin içeriklerinde paylaşıma açılması mümkündür. Yapılmış olan çalışma sabit, doğru bir altyapıya sahip olan MAKS projesine web tabanlı yardımcı bir araç gibi yazılımı geliştirilerek gerçekte var olmayan, güvenirliği tartışmalı veya hatalı adres bilgilerinin ve ilgili içeriklerin ortaya çıkarılmasına da katkıda bulunabilir. MAKS gibi bir büyük bir sistemine benzerliği olduğu düşünülen ve bölgesel olarak geliştirilen İstanbul Büyükşehir Belediyesi (İBB) Adres Bilgi Sistemi verileri çalışmada kullanılmıştır.

2. YÖNTEM

Web sayfalarının içeriğinde yazı, resim, ses, film, animasyon gibi pek çok farklı yapıda veriler yer almaktadır.

İnternette sunulan ve paylaşılan veriler ve bilgiler büyük olduğu kadar karışık ve birbirinden bağımsız bir yapıya sahiptir. Bu yönüyle web deki bilgiler ilk bakışta tamamen anlamlı ve işe yarayan özellikler içermeyebilmektedir.

Veri madenciliğinde olduğu gibi web madenciliğinin de temel özeliklerinden bir tanesi büyük veri yığını içerisinden anlamlı bir sonuç ortaya çıkarmak istemektir. Web belgelerinden ve servislerinden, veri madenciliği tekniklerinin kullanılarak bilgilerin temizlenmesi, ortaya çıkarılması ve çözümlenmesi işlemleri web madenciliği olarak tanımlanmaktadır (Daş vd., 2006; Kunc, 2007; Bhatia ve Kumar, 2008).

Web madenciliği terimi "The World Wide Web: quagmire or gold mine?" isimli makalede ilk defa tartışmaya açılmış ve veri madenciliği yöntemlerinin kullanılması ile internet ortamında bulunan dosya ve servislerden otomatik örüntülerin bulunabileceğini ortaya koyulmuştur (Etzioni, 1996).

Çalışmada adres bilgilerini içerisinde barındıran web sayfalarının içerikleri tespit edilerek, adreslerin kullanım türlerinin de tespiti yapılması hedeflenmektedir. Bunun için öncelikle verilerin sayısal ortamda işlenebilmesi için vektör uzay modeli oluşturulmuştur. Vektör uzay modelinin oluşturulmasında daha önceden oluşturulmuş sözlüklerden faydalanılmıştır.

Daha sonraki aşamada Web sayfalarının içeriklerinin tespiti için Öklid ve Kosinüs yöntemlerinden faydalanılmıştır. Son aşamada

web madenciliğinde kullanılan sınıflandırma yöntemlerinden birisi olan k-nearest neighbors algorithm (KNN) den de faydalanılmıştır.

2.1.Vektör Uzay Modeli

Web madenciliğinin alt kollarından birisi olan içerik analizi ile web sayfalarındaki verilerin elde edilmesi işlemleri yapılmaktadır. Aynı metin madenciliğinde olduğu gibi web içerik madenciliğinde de metinlerin sınıflandırılması için öncelikle dokümanların vektörsel olarak uzayda ifade edilebilir duruma getirilmesi gerekmektedir.

Metinlerdeki terimleri işleyebilmek ve tanımlamak için yapılan ilk çalışma 1975 yılında Salton tarafından geliştirilen vektör uzay modelidir (Salton, 1975).

Bu modelde metin içerisindeki terimlerin bir matris halinde sunulması ilkesi hedeflenmektedir.

$$D=\begin{bmatrix} & T_1 & T_2 & \cdots & T_T \\ D_1 & f_{11} & f_{21} & \cdots & f_{t1} \\ D_2 & f_{12} & f_{22} & \cdots & f_{t2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ D_n & f_{1n} & f_{2n} & \cdots & f_{tn} \end{bmatrix} \quad (1)$$

D belge terim matrislerini (1) ifade etmektedir. Her bir D_1, D_2, … D_n satırı eğitim amacı ile kullanılan dokümanları, T_1, T_2, …, T_t terim vektörlerini ifade edilmektedir.

Dokümanlar vektör formatında sunulmadan önce kelimeler eklerinden, noktalamalardan temizlenir ve bu işlemler önişlem olarak adlandırılır. Çalışmada bu işlemlerin sonrasında metinlerin vektör halinde ifade edilebilmesi için "*Terim Sıklığı- Devrik Belge Sıklığı (Frekans)"* yöntemi kullanılır.

2.2.Öklid Yöntemi

Öklid yönteminde aşağıdaki formül (2) kullanılmaktadır. Burada p eğitim sınıfını ve q eğitilecek veri setlerini, d ise uzaklığı ifade etmektedir.

$$d(p,q)==\sqrt{\sum_{i=1}^{n}(p_i-q_i)^2} \quad (2)$$

Öklid yöntemi; en küçük sonuç değerinin istenilen örneğe en yakın sınıfa ait olduğu prensibine dayanmaktadır.

2.3.Kosinüs Yöntemi

Sim(D1, Q)=cosθ

$$sim(D1,Q)=\frac{D_i.Q}{|D_i|.|Q|}=\frac{\sum W_{ij}*W_{qj}}{\sqrt{\sum_j W_{i,j}{}^2}*\sqrt{\sum_j W_{q,j}^2}} \quad (3)$$

Kosinüs yönteminin (3) temel prensibi benzerlikler üzerine kuruludur. Bu yöntemde benzerlik değeri 1'e yakınlaştıkça yüksek benzerlikten bahsetmek mümkündür.

Elde edilen sonuçların değerlendirilmesi aşamasında genellikle karışıklık matrisi (Confusion matrix) kullanılmaktadır (Tablo 1). İlk defa Kohavi tarafından 1998 yılında kullanılmaya başlanılan karışıklık matrisi yardımı ile, doğruluk, hata, anma (4), duyarlılık (5) ve f ölçüt değerleri hesaplanır (Kohavi, 1998).

Tablo 1. Karışıklık Matrisi Gösterimi

		Öngörülen sınıf	
		Negatif	Pozitif
Doğru sınıf	Negatif	A (dp)	B (ön)
	Pozitif	C (öp)	D (dn)

Anma=D/(B+D) (4)

Anma (Recall): Elde edilen dokümanın ilgili olma durumu

Duyarlılık=A/(D+C) (5)

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

Web sayfasının içeriğinin tespitinde KNN sınıflandırma yöntemi kullanılmıştır. KNN yönteminde K ayırt edicilik parametresi seçimi yüksek doğruluk değeri veren sonuçlara göre seçilmelidir.

2.4.Veri Seti

Çalışmada İstanbul iline bağlı Şişli ilçesi test bölgesi olarak seçilmiştir. Bölgenin ticari açıdan zengin ve dinamik olması, İstanbul Büyükşehir Belediyesi (İBB) adres bilgi sistemi verilerinin çalışmanın yapıldığı tarihlerde Şişli ilçesi için eksiksiz olarak sunulması ilçenin test bölgesi olarak

Geomatik Dergisi
Journal of Geomatics
2019; 4(1);14-22

seçilmesindeki en önemli faktörler arasında yer almaktadır.

İBB Adres Yönetim Şefliği tarafından hazırlanan Adres Bilgi Sistemi verileri ve web sayfalarının kayıt altına alınmasından oluşan iki farklı veri seti çalışmanın temelini oluşturmaktadır. Çalışmada web sayfası içeriklerinden elde edilen veriler hem eğitici hem de eğitilen kısım olmak üzere iki farklı biçimde kullanılmaktadır.

Adres Bilgi Sistemi verilerinden iki aşamada faydalanılmaktadır. İlk aşama web sayfalarının içeriklerinin tespit edilmesi diğeri elde edilen içeriklerin kontrol edilmesi aşamasıdır. Web madenciliğinin önişlem adımlarından birisi olan terim sıklıklarının belirlenmesinde öncelikle bir sözlük oluşturulması gerekmektedir. Adres Bilgi Sistemi verileri içerisinde bulunan obje grupları bölümünün yardımı ile sözlük oluşturulmuştur.

Yapılan çalışmanın ilk adımında Adres Bilgi Sistemi içerisinde bulunan OBJE_KOD tablosundan faydalanılarak 14 farklı temel obje grubu oluşturulmuştur. (Şekil 1). OBJE_KOD tablosunda temelde 14 alt grup olarak 505 adet kayıt bulunmaktadır.

	OBJE_ANA_GF	OBJE_ANA_GF	OBJE_GRUP_K	OBJE_GRUP_
▶	1	Altyapı Tesisi	1	Doğalgaz
	1	Altyapı Tesisi	1	Doğalgaz
	1	Altyapı Tesisi	1	Doğalgaz
	1	Altyapı Tesisi	1	Doğalgaz
	1	Altyapı Tesisi	2	Elektrik
	1	Altyapı Tesisi	2	Elektrik
	1	Altyapı Tesisi	2	Elektrik
	1	Altyapı Tesisi	3	Radyo-televizy
	1	Altyapı Tesisi	4	Su
	1	Altyapı Tesisi	4	Su
	1	Altyapı Tesisi	5	Telekom
	1	Altyapı Tesisi	5	Telekom
	1	Altyapı Tesisi	5	Telekom
	1	Altyapı Tesisi	5	Telekom

Şekil 1. Adres Yönetim Şefliği, Adres Bilgi Sistemi Obje Grupları Örneği

Diğer yapılandırılmış veri tabanlarından farklı olarak, web sayfaları yapılandırılmamış özellikte olduğu için verilerin daha öncede vurgulandığı gibi önişlemlerden geçirilmesi gerekmektedir. Bu işlemlerin sonrasında metin içerisindeki terimlerin sayısal olarak ifade edilmesi gerekmektedir. Terim sıklıklarının elde edilebilmesi için çalışmada iki farklı sözlük kullanılmıştır. Birinci sözlük belediye sözlüğü olarak adlandırılmış ve sözlükte eğitim, turizm, sağlık, ticaret vb. temel obje grupları ve bu obje gruplarına ait alt obje gruplarında geçen kelimelere yer verilmiştir.

İkinci sözlük ise, ontolojik olarak desteklenmiş belediye sözlüğüdür. Belediye sözlüğüne ek olarak kelimelerin daha tanıtıcı olması için ek kelimeler kullanılmıştır.

OBJE _KOD tablosu içerisindeki 6 obje grubu 14 farklı obje grubu arasından seçilmiştir ve belediye sözlüğünün oluşturulmasında alt grupları ile birlikte kullanılmıştır. 6 obje grubunun seçilmesinin nedeni adrese ait kullanım fonksiyonlarının dinamik yapıda olmalarıdır (Tablo 2).

Tablo 2. Şişli Çalışma Bölgesinde Seçilen 6 Farklı Grup ve Bunlara Ait Kelime Dağılımları

Obje grubu ismi	Kelime sayısı
Eğitim Kurumu	60
Kültürel Tesis	28
Sağlık Kurumu	83
Ticaret	463
Turizm	41
Yeşil Alan, Spor ve Dinlenme Yerleri	46

Ontolojik olarak desteklenen ikinci belediye sözlüğü ise web sayfalarındaki adres kullanım fonksiyonlarına erişimin etkinliğinin arttırılması için oluşturulmuştur.

Web madenciliği uygulaması için belediye obje alt gruplarının kavramlarının yetmeyeceği açıktır. Bu yüzden bu kavramların ek kelimelerle zenginleştirilmesi gerekmektedir.

Terimlerin anlamları, terimler arasındaki ilişkiler, terimler arasındaki eş anlamlılıklar ve sıradüzen kavramları ontoloji olarak sunulmaktadır. Soyut model oluşturma anlamına gelen ontoloji, bilgisayar alanında "kavramsallaştırmanın açıkça belirtilmesi" olarak açıklanmaktadır (Gruber, 1995).

Ontolojik yaklaşımın kullanılması sayesinde; belediye sözlüğünde oluşan kelime sayısındaki dengesiz dağılımın önüne geçilmiş olup ayrıca web sayfalarının içeriklerinde yer alan adresler ile ilişkilendirilmesi çok daha anlamlı hale gelmesi sağlanmıştır.

Ontolojik yaklaşımın kullanıldığı literatür çalışmaları incelendiğinde kelimenin anlamlı olarak sunulması için farklı seviyeler kullanıldığı tespit edilmiştir (Schuurman ve

Geomatik Dergisi
Journal of Geomatics
2019; 4(1);14-22

Leszczynski, 2006; Machado vd., 2010; Yüksek ve Ünalır, 2012).

Şişli çalışmasında da üç farklı ontolojik seviye kullanılmıştır. En üst seviyede; sağlık, eğitim, kültür, turizm, spor ve ticaret gibi belediye adres bilgi sisteminde yer alan obje ana grupları yer almaktadır. Çalışmada sağlık obje ana grubunun altında bulunan hastane, sağlık ocağı ve poliklinik, gibi kelimeler orta seviyeyi ifade etmektedir. Alt seviyede ise belediye adres bilgi sisteminde yer almayan fakat kelime olarak sağlık sektörünü çağrıştırabilecek olan doktor, hasta, hemşire, diş, eklem vb. gibi kelimeler kullanılmaktadır. Seviyelendirme sistemi sayesinde ontolojik olarak desteklenen kelimelerin obje gruplarındaki dağılımı artmış ve dengelenmiştir (Tablo 3). Tablo bilgileri incelendiğinde sadece belediye obje tiplerinin kullanılması durumunda eğitim ile ilgili 60, sağlık ile ilgili 83 farklı kelimeye ulaşılabilmektedir. Fakat ontolojik olarak kelimelerin desteklenmesi durumunda eğitim ile ilgili kelimelerin 230'a sağlık ile ilgili kelimelerin ise 295'e kadar zenginleştirilebileceği ortaya çıkmaktadır.

Tablo 3. Obje Grupları ve Kelime Sayıları Karşılaştırılması

Obje grupları	Ontolojik olarak desteklenmiş kelimeler	Belediye obje grupları sözlüğü
Eğitim	230	60
Sağlık	295	83
Ticaret	293	463
Turizm	200	41
Spor	210	46
Kültür	226	28
Diğer	293	575
Toplam	**1747**	**1296**

Çalışmada web sayfalarının içeriklerinin tespit edilmesi için veri ve metin madenciliğinde olduğu gibi web madenciliğinde de kullanılan sınıflandırma yöntem olarak seçilmiştir.

Çalışmanın temelini içeriği belli olmayan web sayfalarının içerik tespiti ve web sayfalarında yer alan adres bilgilerinin de tespit edilen içeriğe dahil edilmesi prensibi oluşturmaktadır.

Web sayfalarının içeriklerinin belirlenmesini sağlamak için eğitici ve eğitilen olmak üzere iki farklı grupta web sayfası kullanılmıştır. Eğitici gruba ait 24 adet web sayfasının içerikleri belirli iken, eğitilen gruptaki 33 adet web sayfasının içerikleri belirsizdir.

3. BULGULAR

Web madenciliği yöntemlerinin yardımı ile adres bilgileri içeren web sayfalarının içerikleri hızlı bir şekilde tespit edilerek, gerekli ilişkilendirmeler ve sınıflandırmalar yapılarak veri tabanlarına aktarımları yapılabilir. Çalışmanın yöntem bölümünde daha önce de vurgulandığı gibi sınıflandırma işlemlerine geçilmeden önce önişlem adımları uygulanmalıdır.

Web sayfalarındaki HTML kodlarının yanı sıra bağlaçların atılması gibi işlemler uygulama bölümünün önişlem aşamasını oluşturmaktadır.

Çalışmada gerek belediye sözlüğü gerekse ontolojik olarak desteklenen belediye sözlükleri kullanılarak 24 adet eğitici ve 33 adet eğitilen web sayfası için ön işlemler yapılmış ve sınıflandırma için hazır hale getirilmiştir. Elde edilen sayısal veri setlerine ayrı ayrı Öklid ve Kosinüs yöntemleri uygulanmış ve 1-10 'a kadar her bir K ayırt edicilik parametresi için sonuçlar analiz edilmiştir.

Tablo 4. K Değerleri ve Doğru Olarak Erişilen Sayfa Sayıları

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

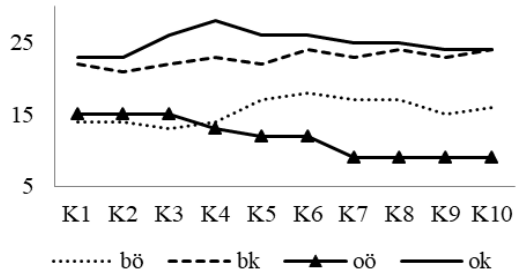
b= belediye o=ontolojik ö=öklid k=kosinüs

K=4 parametresi seçilmesi ve ontolojik kosinüs yönteminin uygulandığı durumda eğitilen olarak kullanılan 33 adet web sayfasının 28 tanesinin içeriğinin tespiti doğru yapılmıştır (Tablo 4). Bu sonuç diğer parametre değerleri ve yöntemler arasındaki en yüksek analiz sonucunu ortaya çıkarmaktadır.

Doğru olarak erişilen web sayfaları ve K parametre arasındaki ilişkiyi ortaya koymak için hazırlanan grafikte ontolojik olarak desteklenen belediye sözlüğünün kosinüs

Geomatik Dergisi
Journal of Geomatics
2019; 4(1);14-22

yöntemi uygulanması sonucunda en yüksek erişim değerine ulaşmaktadır (Şekil 2).



Şekil 2. K Değerlerinin ve Erişilen Sayfaların Karşılaştırılması

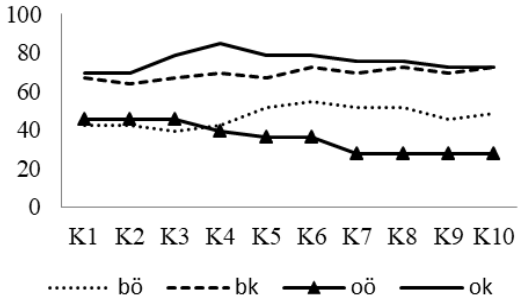
Çalışmanın doğruluğunu tespit etmek için kullanılan duyarlılık değer sonuçları arasında en yüksek değerin parametrenin K=4 olduğu ve ontolojik kosinüs yöntemi uygulaması sonucu elde edildiği tespit edilmiştir (Tablo 5).

Tablo 5. Duyarlılık Değerleri

	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
bö	42	42	39	42	52	55	52	52	46	49
bk	67	64	67	70	67	73	70	73	70	73
oö	46	46	46	39	36	36	27	27	27	27
ok	70	70	79	85	79	79	76	76	73	73

b= belediye o=ontolojik ö=öklid k=kosinüs

Tablo sonuçları değerlendirildiğinde belediye sözlüğü kullanılarak web sayfalarının kullanım fonksiyonlarının tespitinin aynı parametre kullanıldığında %42 ile düşük duyarlılık sonucu elde edilmiştir. Çalışma için seçilen yöntemler değerlendirildiğinde Öklid yönteminin bu tip çalışmalarda kullanılmasının ortaya çıkan düşük duyarlılık sonuçları göz önüne alınarak tercih edilmemesi gerektiği söylenebilir. Bunun yanı sıra belediye sözlüğünün 1. seviyedeki obje grupları değiştirilmeden ontolojik olarak zenginleştirilmesi sonucunda %85 duyarlılık sonucu bulunmuştur.



Şekil 3. K Değerlerinin ve Duyarlılık Yüzdelerinin Karşılaştırılması

Elde edilen analiz sonuçlarına göre oluşturulan grafikte K ayırt edicilik parametresi 4, ontolojik olarak desteklenmiş ve kosinüs yöntemi uygulanması durumunda %85 duyarlılığa ulaşıldığı tespit edilmiştir (Şekil 3).

Yapılan çalışmaya ait analiz sonuçları incelendiğinde web sayfa sayılarının arttırılması ve ontolojik olarak desteklenen sözlükte uygulama sonuçlarında duyarlılık değeri düşük olan turizm ve ticaret gibi obje grupları daha belirgin kavramlar ile desteklenmesi ile %85'lik duyarlılık değerinin daha da artacağı öngörülmektedir.

4. SONUÇLAR ve TARTIŞMA

Elde edilen %85 gibi yüksek doğruluğa sahip analiz sonuçları değerlendirildiğinde ve günümüzdeki gelişmeler dikkate alındığında internet ortamında sunulan bilgilerin önemi giderek artmaktadır. Bu tip çalışmaların sonuçlarından elde edilecek bilgiler gerekli standartların sağlanması durumunda mevcut sistemlerdeki verilerin güncellenmesinde kullanılabilir. Özellikle dinamik değişimlere sahip İstanbul gibi metropoliten kentlerde adres verilerinin ve mekân kullanım fonksiyonlarının güncellenmesi çok önemlidir. Bu yüzden web sayfalarındaki içerikler göz ardı edilmemeli ve zengin veri potansiyeli barındırdığı ciddi bir şekilde düşünülmelidir.

Web sayfalarından adres ve içerik tespitinin yapıldığı bu çalışma, MAKS gibi doğruluğu yüksek sistemler ile birleştirilmesi durumunda, şirketler ve kurumlar tarafından web sayfalarında sunulan adres ve içerik kontrolünün kolaylaşmasının yanı sıra güncel ve az maliyetli olacağı da öngörülmektedir. Web sayfalarında sunulan adres bilgilerinin ve

içeriklerin kontrolünü sağlayıcı otomatik yazılımlar geliştirilmesi gerekmektedir. Bu sayede sunulan bilgilerin doğruluğu, güvenilirliği ve düzeltilmesi sağlanabilir.

Gelişime açık ve güncel bir alan olan Web madenciliği ve içerik türü tespiti çalışması sonraki çalışmalar için faydalı bir örnek olabilir. Çalışmada kullanılan yöntemlerin haricinde farklı yöntemler en yüksek duyarlılık sonucu elde edilene kadar denenebilir. Çalışmanın kapsamı genişletilerek çok daha fazla web sayfası ile test yapılabilir. Konumsal bilginin çok önemli olduğu günümüzde, bu tür çalışmaların ülkemizde arttırılarak desteklenmesi gerekmektedir.

KAYNAKÇA

Bhatia, MPS., ve Kumar, A., (2008). Information retrieval and machine learning: supporting Technologies for web mining research and practice, *Webology*, **5**, 2.

Boll, S., Jones, C., Kansa, E., Kishor, P., Naaman, M., Purves, R., ... & Wilde, E. (2008, April). Location and the web (LocWeb 2008). *In Proceedings of the 17th international conference on World Wide Web* (pp. 1261-1262). ACM.

Borges, K. A., Laender, A. H., Medeiros, C. B., & Davis Jr, C. A. (2007, November). Discovering geographic locations in web pages using urban addresses. In *Proceedings of the 4th ACM workshop on Geographical information retrieval* (pp. 31-36). ACM.

Borges, K. A., Laender, A. H., Medeiros, C. B., da Silva, A. S., & Davis Jr, C. A. (2003). The Web as a Data Source for Spatial Databases. In *GeoInfo.*

Buyukokkten, O., Cho, J., Garcia-Molina, H., Gravano, L., & Shivakumar, N. (1999). Exploiting geographical location information of web pages. *ACM SIGMOD Workshop on The Web and Databases (WebDB'99)*, Philadelphia, Pennsylvania, USA.

Clough, P. (2005, November). Extracting metadata for spatially-aware information retrieval on the internet. In *Proceedings of the 2005 workshop on Geographic information retrieval* (pp. 25-30). ACM.

Dalli, A. (2006, May). System for spatio-temporal analysis of online news and blogs. In *Proceedings of the 15th international conference on World Wide Web* (pp. 929-930). ACM.

Daş, R., Türkoğlu, İ., & Poyraz, M. (2006). Genetik algoritma yöntemiyle ınternet erişim ka yıtlarından bilgi çıkarılması. *Sakarya Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 10(2), 67-72.

Ding, J., Gravano, L., & Shivakumar, N. (1999). *Computing geographical scopes of web resources*. Stanford.

Etzioni, O. (1996). The World-Wide Web: quagmire or gold mine?. *Communications of the ACM*, 39(11), 65-68.

Fake Adres Generator (FAG), (2018), https://www.fakeaddressgenerator.com/, Erişim Tarihi: 04.07.2018

Gelernter, J., & Lesk, M. E. (2008, October). Traditional resources help interpret texts. In *Proceedings of the 2008 ACM workshop on Research advances in large digital book repositories* (pp. 17-20). ACM.

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing?. *International journal of human-computer studies*, 43(5-6), 907-928.

Heuer, J. T., & Dupke, S. (2007). Towards a spatial search engine using geotags. *GI-Days*, 199-204.

Hill, L. L. (2000, September). Core elements of digital gazetteers: placenames, categories, and footprints. In *International Conference on Theory and Practice of Digital Libraries* (pp. 280-290). Springer, Berlin, Heidelberg.

Jones, C. B., Abdelmoty, A. I., Finch, D., Fu, G., & Vaid, S. (2004, October). The SPIRIT spatial search engine: Architecture, ontologies and spatial indexing. In *International Conference on Geographic Information Science* (pp. 125-139). Springer, Berlin, Heidelberg.

Kohavi, R. (1998). Glossary of terms. *Machine Learning*, 30, 271-274.

Kunc, M. (2007). Web Mining Overview. In *Proceedings of the 13th Conference STUDENT EEICT 2007 Volume* (pp. 391-395). Brno University of Technology.

Li, W., Goodchild, M. F., & Raskin, R. (2014). Towards geospatial semantic search: exploiting latent semantic relations in geospatial data. *International Journal of Digital Earth*, 7(1), 17-37.

Li, W. (2018). Lowering the barriers for accessing distributed geospatial big data to advance spatial data science: the PolarHub solution. *Annals of the American Association of Geographers*, 108(3), 773-793.

Machado, I. M., de Alencar, R. O., Junior, R. D. O. C., & Davis Jr, C. A. (2010). An Ontological Gazetter for Geographic Information Retrieval. *In GeoInfo* (pp. 21-32).

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. Y., Suel, T., Long, X., & Seeger, B. (2005, June). Design and Implementation of a Geographic Search Engine. In *WebDB* (Vol. 5, pp. 19-24).

Mekansal Adres Kayıt Sistemi (MAKS), (2018), https://maks.nvi.gov.tr/, Erişim Tarihi: 04.07.2018

McCurley, K. S. (2001, April). Geospatial mapping and navigation of the web. In *Proceedings of the 10th international conference on World Wide Web* (pp. 221-229). ACM.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11), 613-620.

Schuurman, N., & Leszczynski, A. (2006). Ontology-based metadata. *Transactions in GIS*, 10(5), 709-726.

Tezuka, T., Kurashima, T., & Tanaka, K. (2006, May). Toward tighter integration of web search with a geographic information system. In *Proceedings of the 15th international conference on World Wide Web* (pp. 277-286). ACM.

Tri, N. T., & Jung, J. J. (2015). Exploiting geotagged resources to spatial ranking by extending hits algorithm. *Computer Science and Information Systems*, 12(1), 185-201.

Türk Ticaret Kanunu (TTK), (2018), http://www.mevzuat.gov.tr/MevzuatMetin/1.5.6102.pdf , Erişim Tarihi: 04.07.2018

Yüksek, Y., & Ünalır, M. O. (2012). Ulusal Sağlık Veri Sözlüğü için Ontoloji Tabanlı Üst Veri Yönetim Sistemi. *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, 18(2), 133-144.

Waldinger, R., Jarvis, P., & Dungan, J. (2003, May). Pointing to places in a deductive geospatial theory. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references-Volume 1* (pp. 10-17). Association for Computational Linguistics.

Xu, Z., Zhang, H., Sugumaran, V., Choo, K. K. R., Mei, L., & Zhu, Y. (2016). Participatory sensing-based semantic and spatial analysis of urban emergency events using mobile social media. *EURASIP Journal on Wireless Communications and Networking*, 2016(1), 44.