Fen Bilimleri Enstitüsü Dergisi, 20 (2008) 35-49
Marmara Üniversitesi

WEB BELGELERİ KÜMELEMEDE BENZERLİK VE UZAKLIK ÖLÇÜTLERİ BAŞARILARININ KARŞILAŞTIRILMASI

Meltem IŞIK[1] ve Ali Yılmaz ÇAMURCU[2*]

[1] *Şişli Endüstri Meslek Lisesi, Bilgisayar Öğretmeni, Şişli, İstanbul*
[2] *Marmara Üniversitesi, Teknik Eğitim Fakültesi, Elektronik ve Bilgisayar Eğitimi Bölümü, 34722 Göztepe, İstanbul*

Alındığı Tarih: 01 Kasım 2008 *Kabul Tarihi: 23 Şubat 2009*

Özet: İnternetteki web sayfalarının boyutları her geçen gün artmaktadır. Bu sayfalar içerisinde bulunan belgelere erişimde ya da bir belgeyi getirmede yeni teknikler geliştirilmektedir. Bu tekniklerden birisi de web belgelerini kümelemedir. Bu çalışmada, web sayfaları kümelemede belgelerin benzerliklerini bulan tekniklerden Öklid, Kosinüs, Pearson ve Genişletilmiş Jaccard iki ayrı veri setinde test edildi ve başarıları araştırıldı. Web belgelerini kümelemede yapılan testlerde, Öklid uzaklığının yüksek hata oranlarına neden olduğu gözlenmiştir. Benzerlik ölçütlerinde en iyi performansı sağlayan Kosinüs ve Genişletilmiş Jaccard benzerlikleridir. Yapılan deneylerin sonuçlarına göre, web belgeleri kümelemede Kosinüs benzerlik ölçütünün kullanılmasının uygun olduğu bulunmuştur.

Anahtar kelimeler: Benzerlik ölçütleri, belge kümeleme, belge madenciliği, veri madenciliği.

---

[*] Faks: (216) 337 8987 e-posta: camurcu@marmara.edu.tr

COMPARISION OF SIMILARITY AND DISTANCE MEASURES' PERFORMANCES IN WEB DOCUMENTS CLUSTERING

Abstract: The number of web pages in Internet is increased by day by. New techniques are developed to reach or retrieve information from the documents in those web pages. Clustering is one of techniques used on web documents. In this study, the techniques such as Euclidean, Cosine, Pearson and Extended Jaccard used to find document similarities in web pages were tested by two data sets and performances were studied. In the experiments done for web documents clustering, found that Euclidean distance measure has high fault rates. The best performance in the similarity measures are provided by Cosine and Extended Jaccard measures. According to results of experiments that Cosine similarity measure was found suitable to use in the web documents clustering.

Keywords: Similarity measures, document clustering, document mining, data mining

GİRİŞ

İnternette bulunan web sayfalarının içerisindeki belge sayısı sürekli artmaktadır. İnternette aranan bilgiyi getirmek için bu belgelere kolayca erişilmesi gerekmektedir. Erişim işlemini gerçekleştiren yöntemlerden birisi de web belgelerinin kümelenmesidir. Veri madenciliğinde kullanılan yöntemlerden birisi olan kümeleme tekniği ile web belgeleri kümelenerek kolay ve hızlı bir biçimde İnternet kullanıcılarına eriştirilir. Web sayfalarının kümelenmesi bazı benzerlik veya uzaklık ölçütleri kullanılarak gerçeklenir[1,2]. Benzerlik veya uzaklık ölçütleri, web sayfalarını kümelemenin yanında, metin madenciliği[3-5], web madenciliği[6], bilgi getirme[7], yüksek boyutlu veri setlerinde kümeleme ve görselleştirme[8] gibi tekniklerde de kullanılmaktadır.

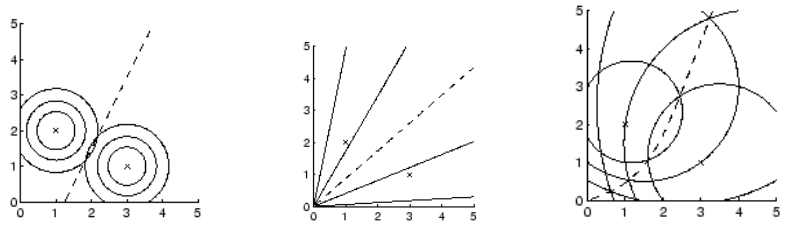
Kümeleme analizi, veri setinin içinde bulunan anlamlı veya kullanışlı grupları ortaya çıkarma sürecidir. Kümeleme, benzer veri nesnelerinin gruplara bölünmesi işlemidir. Kendi içinde benzer, fakat diğer gruplardakine benzemeyen nesneler içeren her gruba küme adı verilir. Veri nesneleri

arasındaki uzaklıkları hesaplamak için birçok farklı uzaklık veya benzerlik ölçüm yöntemi bulunmaktadır.

Web belgelerinin kümelenmesinde, benzer kelimeler aynı küme içerisinde yer almalıdır. Bu nedenle benzerlik ölçütleri seçiminin önemi büyüktür. Kullanılacak benzerlik ölçütü iyi seçilirse oluşacak kümelerin kalitesi de o derecede iyi olacaktır[7,9].

Bu çalışmanın amacı, Öklid uzaklık, Kosinüs benzerlik, Pearson benzerlik ve genişletilmiş Jaccard benzerlik ölçütlerinin gerçek veri seti üzerinde performanslarının değerlendirilmesidir[10]. Gerçek veri seti olarak çok boyutlu vektörlerle temsil edilen web belgeleri seçilmiştir. Bu çalışmanın ikinci bölümünde Öklid uzaklık, Kosinüs benzerlik, Pearson benzerlik ve genişletilmiş Jaccard benzerlik ölçütlerinin yapıları, üçüncü bölümde veri setleri ve kümelemenin değerlendirilmesi için kullanılan saflık (purity), entropi ve ortak bilgi (mutual information) ölçütleri, dördüncü bölümde MATLAB kullanılarak gerçekleştirilen test sonuçları ve son bölümde değerlendirme açıklanmaktadır.

BENZERLİK VE UZAKLIK ÖLÇÜTLERİ



a) *b*) *c*)

Şekil 1. *a*) Öklid uzaklığının kümeleme özelliği, *b*) Kosinüs benzerliğinin kümeleme özelliği, *c*) Genişletilmiş Jaccard benzerliğinin kümeleme özelliği [9,11]

Bu çalışmada kullanılan benzerlik ölçütlerinden Öklid, Kosinüs, Pearson ve genişletilmiş Jaccard benzerlik ölçütlerinin Şekil 1 de görülen benzerlik bulma yapıları aşağıda açıklanmaktadır.

En yaygın kullanılan uzaklık ölçütü olan Öklid Uzaklığı, çok boyutlu uzaydaki nesnelerin birbirlerine geometrik uzaklığıdır. Nesnelerin konumları incelenerek ne kadar farklı oldukları belirlenir. Veri seti yoğun ve birbirlerinden iyi ayrılmış kümeler içeriyor ise iyi sonuç üretir. İki nesne birbirine ne kadar yakın ise Öklid uzaklığı da o kadar sıfıra yaklaşır. Şekil 1.*a*' da görüldüğü gibi Öklid uzaklığı kullanılarak bulunan kümeler küresel bir yapıya sahiptir. Koordinatları belli olan iki nokta arasındaki Öklid uzaklığı Denklem 1 ile hesaplanmaktadır [7,9]:

$$\sqrt{\sum_{i\_1}^{n}(p_i - q_i)^2} = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_n - q_n)^2} \tag{1}$$

Belge kümelemede çok kullanılan vektör tabanlı bir ölçüt olan Kosinüs Benzerliği ile, iki vektör arasındaki açının kosinüs değeri hesaplanarak vektörlerin benzerliği bulunur. Vektör boyutundan etkilenmemesi, kosinüs benzerliğinin güçlü bir özelliğidir. Farklı çok sayıda kelimeler içeren benzer içerikteki belgeleri kolaylıkla tespit eder. Denklem 2 de görüldüğü gibi, vektörlerin skaler çarpımlarının, genliklerine bölünmesiyle iki vektör arasındaki açı elde edilir. İki vektör arasındaki açı ne kadar 0'a yaklaşırsa, açının kosinüs değeri 1'e yaklaşır ve iki vektörün birbirlerine olan benzerlikleri de artar. Şekil 1.*b* 'de Kosinüs Benzerliğinin kümeleme özelliği görülmektedir. Denklem 2 de iki vektör arasındaki açının kosinüs değerinin hesabında, d ve d^* birbirinden farklı iki belgeyi temsil eden çok boyutlu vektörleri ve "•" vektörlerin iç çarpımını, $|d|$ ise vektörün uzunluğunu temsil etmektedir[7,9,11];

$$\cos(\theta) = \frac{d \bullet d^*}{|d||d^*|} = \frac{\sum_{i=1}^{n} d_i\, d_i^*}{\sqrt{\sum_{i=1}^{n}(d_i)^2}\sqrt{\sum_{i=1}^{n}\left(d_i^*\right)^2}} \tag{2}$$

Diğer bir yöntem olan Pearson ilişkisinde, Kosinüs benzerliğindeki gibi iki vektörün benzerliğinde vektörlerin aralarındaki açıya göre karşılaştırma yapılır. Kosinüs benzerliğinden farkı iki vektörün iç çarpımı yapılmadan önce her birinin ayrı ayrı ortalama değerleri hesaplanır ve her ortalama değer ait olduğu vektörün tüm elemanlarından çıkarılır. d ve d^* birbirinden farklı iki belgeyi temsil eden çok boyutlu vektör ise aralarındaki Pearson benzerliği Denklem 3 ile hesaplanır [12,14];

$$S^{Pearson}(d,d^*) = \frac{1}{2}\left(\frac{\sum_{i=1}^{n}\left(d_i - \bar{d}_i\right)\left(d_i^* - \bar{d}_i^*\right)}{\sqrt{\sum_{i=1}^{n}\left(d_i - \bar{d}_i\right)^2 \sum_{i=1}^{n}\left(d_i^* - \bar{d}_i^*\right)^2}} + 1 \right) \tag{3}$$

Çıkan sonucu [0,1] aralığında tutmak için normalizasyon işlemi yapılmaktadır. Bunun için hesaplanan değere 1 eklenip daha sonra da 2'ye bölmek yeterli olacaktır. Pearson ilişkisi değeri 1'e yaklaştıkça iki vektörün birbirlerine olan benzerlikleri de artar.

Genişletilmiş Jaccard benzerliği, iki nesnenin paylaşılan parçalarının, nesnelerin tüm parçalarına oranıdır. Nesneler belge vektörleri, parçalar ise kelimeler ile temsil edilir. Formül daha basite indirgendiğinde, iki vektörün kesişim kelimelerinin birleşim kelimelerine oranı olarak ifade edilebilir. Yüksek uzaklık değerlerinde Kosinüs'e, düşük değerlerde ise Öklid'e benzer özellik göstermektedir[13]. Şekil 1.*c*' de Genişletilmiş Jaccard benzerliğinin kümeleme özelliği görülmektedir.

$$S^{(j)}(d,d^*) = \frac{d \cap d^*}{d \cup d^*} \tag{4}$$

d ve d^* birbirinden farklı iki belgeyi temsil eden çok boyutlu vektör ise aralarındaki Genişletilmiş Jaccard Benzerliği aşağıdaki formülle hesaplanır [9,11,15], formüldeki elemanlar Denklem 2 de yapılan açıklamalarda tanımlanmıştır;

$$S^{(j)}(d,d^*) = \frac{d \bullet d^*}{|d|^2 + |d^*|^2 - d \bullet d^*} \quad . \tag{5}$$

VERİ SETLERİ VE KÜMELERİN DEĞERLENDİRİLMESİ

Bu çalışmada, Milliyet gazetesi ve YahooNews (İndirgenmiş) veri setleri kullanıldı. Milliyet gazetesi İnternet arşivlerinden derlenen veri setleri [16] nolu çalışmadan alınmıştır. Milliyet gazetesi veri setinde ekonomi, siyaset ve spor olarak her biri 485'er tane html belgeri içeren üç alt başlık bulunmaktadır. YahooNews indirgenmiş veri seti, içerikleri İngilizce html belgelerini içeren 4 alt başlık oluşmaktadır. Bu veri setinin Business klasöründe 142, Politics klasöründe 114, Health klasöründe 164, Sports klasöründe 141 belge bulunmaktadır.

Kümelemenin değerlendirilmesi için kullanılan saflık, entropi ve ortak bilgi ölçütleri, kümelerin sonucuna uygulanmıştır. Bu ölçütlerin uygulamasından önce bütün belgeler etiketlenir. Aynı klasörde bulunan belgeler aynı etiket numarasına sahiptir. Değerlendirme ölçütlerinin hesaplanabilmesi için kümeleme işlemlerinin sonucunda kümelerdeki belgelerin hangi etiket numaralarına sahip olduklarının bilinmesi gerekmektedir. Bu etiketler sayesinde hangi kümede hangi kategorilerden kaçar belge olduğu tespit edilir bir karmaşıklık matrisi (confusion matrix) oluşturulur. Bu matris kullanılarak saflık, entropi ve ortak bilgi hesaplanır.

Saflık, küme elemanları içindeki baskın sınıfın kümedeki eleman sayısına oranını verir. Bir kümedeki elemanların hepsi aynı sınıfa aitse saflık maksimumdur ve Denklem 6 ile hesaplanır [12].

$$\Phi^{(A)}(C_l) = \frac{1}{n_l} \max_h \left(n_l^{(h)} \right) \tag{6}$$

Burada, C_l her bir küme, $n_l^{(h)}$ C_l kümesindeki her bir h kategorisine ait nesne sayısıdır.

Entropi, saflıktan daha kapsamlı bir ölçüttür. Saflık sadece baskın sınıfın içerisinde olan ve olmayan nesne sayılarıyla ilgilenirken, entropi tüm dağılımla ilgilenir. Her bir sınıfa ait belgelerin bir küme içerisinde nasıl dağıldığına bakar. Kümenin içerdiği elemanlarının hepsi aynı sınıfa aitse entropi "0" olur. Denklem 7 ile hesaplanır [12].

$$\Phi^{(B)}(C_l) = -\sum_{h=1}^{g} \frac{n_l^{(h)}}{n_l} \log_g \left(\frac{n_l^{(h)}}{n_l} \right) \tag{7}$$

Saflık ve entropi büyük sayıda kümeleri değerlendirmek üzere kullanılmaktadır. Genel kümelemeyi değil de her bir kümenin kendi içindeki değerlendirilmesini yansıtırlar. Her küme tek bir belgeden oluştuğu zaman optimum değeri üretmektedirler. Bu nedenle kümelemenin genel başarısını gösteremezler.

Ortak Bilgi (Mutual Information), teorik olarak en iyi sonuç veren nitelik ölçütüdür ve tarafsız bir değerdir. [0,1] arasında değerler almaktadır. Sınıflar dengeli olduğu durumlarda kümeleme başarılı ise 1'e doğru bir değer üretir. Saflık ve entropinin etkilendiği olumsuzluklardan etkilenmez. Denklem 8 ile hesaplanır [15].

$$\Phi^{(NMI)}(\lambda, \mathrm{K}) = \frac{2}{n} \sum_{l=1}^{k} \sum_{h=1}^{g} n_l^{(h)} \frac{\left(\frac{n_l^{(h)} n}{\sum_{i=1}^{k} n_i^{(h)} \sum_{i=1}^{g} n_l^{(i)}} \right)}{\log(k \cdot g)} \tag{8}$$

BENZERLİK VE UZAKLIK ÖLÇÜTLERİ TEST SONUÇLARI VE TARTIŞMA

Tablo I, Tablo II, ve Tablo III' de Milliyet veri setinin çeşitli oranlarda kelimeleri rasgele seçilerek elde edilen veriler üzerinde ve Tablo IV'de YahooNews (İndirgenmiş) veri setinde benzerlik ve uzaklık ölçütleri uygulanmış test sonuçları görülmektedir. Tablolarda kullanılan, tohum değeri algoritmaların başlangıçta rasgele sayı üretmek için kullandıkları bir parametredir. Bu parametre ile küme merkezlerinin atanmasında aynı başlangıç koşulları sağlanarak karşılaştırmalar daha doğru yapılır. L değeri ise belgenin içerdiği kelimelerin yüzde kaçının alındığını ifade etmektedir. Öteleme sayısı ise algoritmanın döngü sayısını göstermektedir.

Tablo I Milliyet veri setinde 25% kelime ve tohum 7 için benzerlik ve uzaklık ölçütlerinin karşılaştırılması

Ölçüt	Ortak Bilgi	Kümelerin Saflığı	Kümelerin Entropisi	Öteleme Sayısı	Geçen Zaman(sn)
Öklid	0.0135	1 1 0.333	0 0 1	3	4.781
Kosinüs	0.9473	0.9876 0.9815 0.9979	0.0610 0.0838 0.0135	9	19
Jaccard	0.9652	0.991 0.987 1	0.043 0.060 0	8	17
Pearson	0.7990	0.9639 0.9059 0.9641	0.1415 0.2926 0.1625	19	162

Tablo I'de Milliyet veri setinin toplam kelimelerinin %25'i alınarak ve tohum 7 için yapılan benzerlik ve uzaklık ölçütlerinin karşılaştırılmasında Öklid

uzaklık ölçütü için genel kümeleme başarısı 0,0135 olarak çok düşük bir değerde bulunmuştur. Ortak bilgi değerinin Jaccard için 0,9652 bulunması ile en iyi küme bulma başarısı gözlenmiştir. Jaccard'dan sonra en başarılı kümeleme sonuçları sırasıyla Kosinüs ve Pearson'a aittir.

Kümeleme sonucunda oluşan her bir kümenin değerlendirilmesi, saflık ve entropi değerlerine bakılarak yapılmıştır. Öklid ölçütüne göre üç ayrı kümenin saflığı incelendiğinde ilk iki kümede sadece tek bir kategoriye ait eleman olması nedeniyle bu kümelerin tamamen saf olduğu, ancak sonuncu kümenin 0,333 olan saflık değeri oldukça düşük olduğu için bu kümenin farklı kategorilerden elemanlar içerdiği görülmüştür. Kosinüs, Jaccard ve Pearson'da ise farklı kategorilere ait nesneler az sayıda bulunduğu için kümelerin saflığı yüksektir. Aralarında çok büyük farklılıklar olmamasına karşın benzerlik ölçütlerini kümelerin saflığına göre sıralandığında Jaccard en iyi, Pearson ise en kötü sonucu üretmiştir. Entropi değeri sıfıra yaklaştıkça o küme içerisindeki elemanların aynı kategoriye ait olduğu görülür. Öklid ölçütü incelendiğinde ilk iki kümede farklı kategorilerden elemanlar olmadığı için bu kümelerin entropisi sıfır ("0"), üçüncü kümede ise farklı kategorilerden çok sayıda eleman olmasından dolayı bu kümenin entropisi ise bir ("1") çıkmıştır. Saflık değerlerinde görüldüğü gibi, Öklid uzaklığı entropi sonucunda da başarısızdır. Çünkü, üçüncü kümede olmaması gereken farklı kategorilere ait birçok eleman bulunmaktadır. Benzerlik ölçütlerinin sonuçları her küme için sıfıra oldukça yakın olduğu için hepsi başarılıdır. Çıkan sonuçlar, sıfıra çok yakın olduğu için en iyi benzerlik ölçütü Jaccard'a, en kötüsü ise Pearson'a aittir.

Öteleme sayısı, amaç fonksiyonun yakınsaklaşmasına bağlı olarak değişmektedir. Az öteleme sayısı geçen zamanı kısalttığı için bir avantajdır, ancak ortak bilgi değerinin de yüksek olması gerekir. Öklid uzaklığının üç kere öteleme sonucunda sonlanması amaç fonksiyonun çok çabuk yakınsaklaştığını gösterir. Fakat, ortak bilgi değerinin çok küçük olması Öklid ile yapılan kümelemenin doğru sonuca ulaşamadan bittiğini göstermektedir. Bu nedenle, Öklid uzaklığının burada da başarısız olduğu görülmüştür. Ayrıca Pearson'ın

Jaccard ve Kosinüs'ten oldukça fazla öteleme sayısına sahip olması ve sonuçta elde edilen ortak bilgi değerinin de bu öteleme sayısına rağmen diğer benzerliklerden düşük olması Kosinüs ve Jaccard'ın Pearson'dan daha başarılı olduğunu göstermektedir. Geçen zaman ise hem öteleme sayısına hem de işlemsel karmaşıklığa bağlı olarak değişir. Öteleme sayısı ve işlemsel karmaşıklık arttıkça geçen zaman artar. Öklid'in hem işlemsel karmaşıklığı hem de öteleme sayısı az olduğu için geçen zaman değeri çok düşük çıkmıştır. Pearson'ın ise hem öteleme sayısı hem de işlemsel karmaşıklığı fazla olduğu için geçen zaman değeri çok yüksek çıkmıştır. Jaccard ve Kosinüs benzerliklerinin Tablo I'de görüldüğü gibi öteleme sayıları ve geçen zaman değerleri birbirine yakındır. Her iki benzerlik ölçütü hem zaman hem de öteleme sayısı açısından avantajlıdır.

Tablo II. Milliyet veri setinde 50% kelime ve tohum 7 için benzerlik ve uzaklık ölçütlerinin karşılaştırılması

Ölçüt	Ortak Bilgi	Kümelerin Saflığı	Kümelerin Entropisi	Öteleme Sayısı	Geçen Zaman(sn)
Öklid	0.0135	1.0000 1.0000 0.3338	0 0 1	3	5
Kosinüs	1	1 1 1	0 0 0	5	16
Jaccard	0.9919	1.0000 0.9959 1.0000	0 0.0243 0	5	17
Pearson	0.9488	0.9939 0.9719 1.0000	0.0341 0.1164 0	10	101.6

Tablo II'de Milliyet veri setinin toplam kelimelerinin %50'si alınarak ve tohum "7" için yapılan benzerlik ve uzaklık ölçütlerinin karşılaştırılmasında Öklid uzaklık ölçütü için genel kümeleme başarısı 0,0135 olarak çok düşük bir değerde bulunmuştur. Bu sonuç, Milliyet veri setinin %25 kelimelik değeriyle aynıdır ve kelime sayısı artmasına karşın Öklid uzaklığının iyi sonuç üretemediği tekrar görülmüştür. Milliyet veri setinin %25 kelimelik değerlerinden farklı olarak Kosinüs'ün sonuçları Jaccard'dan daha iyi çıkmış ve bütün benzerlik ölçütlerinin kümeleme başarısının da arttığı görülmüştür. Kosinüs benzerliğinin ortak bilgi değerinin bir ("1") çıkması kümelerin tamamen doğru ayrıldığını göstermektedir. Kosinüs'ten sonra en başarılı kümeleme sonuçları sırasıyla Jaccard ve Pearson'a aittir. Kelime sayısındaki artış benzerlik ölçütlerinin ortak bilgi başarısını olumlu etkilemiştir. Ancak kelime sayısındaki artış her zaman olumlu sonuç vermez. Çünkü %50'lik ve %25'lik kelimeler veri seti içerisinden rasgele seçilmektedir. Eğer belgeyi iyi tanımlayan kelimeler seçilir ve sıra dışılık yaratanlar seçim dışı kalırsa iyi sonuçlar elde edilir. Belgeyi iyi tanımlayan değerler seçilmezse kümeleme başarısı da düşecektir.

Tablo II. deki Öklid ölçütüne göre üç ayrı kümenin saflığı ve entropisi incelendiğinde Tablo I'deki Öklid değerleriyle aynı sonuçlar bulunmuştur. Tablo II de ilk iki kümede sadece tek bir kategoriye ait eleman olması nedeniyle, bu kümelerin tamamen saf olduğu fakat üçüncü kümenin farklı kategorilerden elemanlar içerdiği görülmüştür. Buna bağlı olarak ilk iki kümenin entropi değeri iyi ancak üçüncü kümenin entropi değeri kötüdür. Kosinüs'te kümeler tamamen doğru ayrıldığı için bütün kümeler sadece bir tek kategoriye ait eleman içermektedir. Her küme tamamen saftır ve entropi değerleri en iyi sonuç olan sıfırı ("0") göstermektedir. Jaccard ve Pearson'da ise farklı kategorilere ait nesneler az sayıda bulunmaktadır ve kümelerin saflığı yüksektir. Benzerlik ölçütleri kümelerin saflığına göre sıralandığında Kosinüs en iyi, Pearson ise en kötü sonucu üretmiştir. Entropi değerleri de saflığa paralel sonuçlar üretmiştir. Öklid uzaklığının öteleme sonucunda herhangi bir gelişme görülmemiştir. Ayrıca Kosinüs ve Jaccard'ın öteleme sayısı aynı çıkmış ve

Tablo I'e benzer olarak Tablo II'de de Pearson'ın Jaccard ve Kosinüs'ten oldukça fazla öteleme sayısına sahip olduğu görülmüştür. Tablo II'de Kosinüs ve Jaccard'ın öteleme sayısı aynıdır ve uygun değerlerdedir. Pearson öteleme sayısı açısından en dezavantajlı benzerlik ölçütü olmuştur. Tablo II'de de Tablo I'de olduğu gibi Öklid'in geçen zaman değeri çok düşük çıkmıştır. Pearson'ın ise öteleme sayısında düşüş olduğu için geçen zaman değeri de düşmüştür ancak en yüksek değer olma özelliği değişmemiştir. Jaccard ve Kosinüs benzerliklerinin geçen zaman değerlerinde ise önemsenmeyecek kadar az fark olduğu görülmüştür.

Tablo III. Milliyet veri setinde 100% kelime ve tohum 13 için benzerlik ve uzaklık ölçütlerinin karşılaştırılması

Ölçüt	Ortak Bilgi	Kümelerin Saflığı	Kümelerin Entropisi	Öteleme Sayısı	Geçen Zaman(sn)
Öklid	0.0170	0.3340 1.0000 1.0000	1 0 0	3	7
Kosinüs	0.7385	0.7273 1.0000 0.5026	0.5334 0 0.6309	6	23
Jaccard	0.6780	1.0000 1.0000 0.5000	0 0 0.6309	16	62
Pearson	0.5854	0.8224 0.9836 0.5091	0.5241 0.0761 0.6308	7	70

Tablo III'de Milliyet veri setinin kelimelerinin tamamı alınarak ve tohum "13" için yapılan benzerlik ve uzaklık ölçütlerinin karşılaştırılmasında Öklid uzaklık ölçütü için genel kümeleme başarısının çok az bir oranda arttığı görülmüştür. Milliyet veri setinin %50'lik değerleriyle benzer olarak, Kosinüs'ün sonuçları Jaccard'dan daha iyi çıkmıştır ve ancak bütün benzerlik ölçütlerinin kümeleme başarısının da azaldığı görülmüştür. Kosinüs

benzerliğinin ortak bilgi değerinin bir (“0.7385”) çıkması kümeleme başarısının orta seviyede olduğunu göstermektedir.

Tablo III’deki Öklid ölçütüne göre üç ayrı kümenin saflığı ve entropisi incelendiğinde, Tablo I ve Tablo II’deki gibi sonuçların başarısız olduğu görülmektedir. Kosinüs’te ikinci küme sadece tek bir kategoriye ait eleman içerirken, diğer kümeler farklı kategorilerden nesneler içermektedir. Jaccard’da ise ilk iki küme tamamen saf ve sadece tek bir kategoriye ait nesneleri içerirken, üçüncü küme diğer kategorilerden nesneler içermektedir. Pearson’da her küme farklı kategorilerden nesneler vardır.

Tablo IV. YahooNews (İndirgenmiş) veri setinde tohum 7 için benzerlik ve uzaklık ölçütlerinin karşılaştırılması

Ölçüt	Ortak Bilgi	Kümelerin Saflığı	Kümelerin Entropisi	Öteleme Sayısı	Geçen Zaman(sn)
Öklid	0.3351	0.9022 1.0000 0.3534 1.0000	0.2310 0 0.9132 0	5	7
Kosinüs	0.6581	0.5745 0.9464 0.9063 0.9776	0.6025 0.1753 0.2709 0.0773	8	21
Jaccard	0.5994	0.5546 0.7590 0.7879 0.9645	0.5656 0.4328 0.4744 0.1105	9	22
Pearson	0.3441	0.6250 0.6170 0.4615 0.7035	0.6501 0.6289 0.8212 0.5734	10	276

Öklid uzaklığının öteleme sonucunda herhangi bir değişim olmazken, Kosinüs'ün öteleme sayısı az oranda Jaccard'ın ise büyük oranda artmıştır.
Tablo IV'de YahooNews (İndirgenmiş) veri setinde tohum "7" ve %100'lük kelime için elde edilen ortak bilgi değerlerinde görüldüğü gibi, Kosinüs en başarılı ölçüt olmuştur. Öklid önceki testlere göre daha iyi sonuç üretmesine rağmen yine de yeterli başarıya ulaşamamıştır. Pearson'da hem genel başarının düşük hem de öteleme sayısının ve geçen zaman değerlerinin yüksek olması, bu ölçütün de başarısız olduğunu göstermektedir.

DEĞERLENDİRME

Web belgelerinin her biri içerdikleri kelimelerle ifade edildikleri için çok boyutlu vektörlerden oluşurlar. Bu çalışmada, her bir veri setinde yüzlerce belge ve her belgede de yüzlerce kelime olduğundan dolayı çeşitli bellek sorunlarıyla karşılaşıldı. Bir belgede geçen herhangi bir kelime birçok belgede bulunmadığından, diğer belgelerdeki ağırlığı sıfır olarak değerlendirilmektedir. Bellek problemini gidermek için bu sıfırlar indirgenmiştir. Web belgeleri kümelemede, hem uzaklık ölçütü hem de benzerlik ölçütleri kullanılmıştır. Web belgelerini kümelemede yapılan testlerde Öklid uzaklığının yüksek hata oranlarına neden olduğu gözlenmiştir. Benzerlik ölçütlerinde de en iyi performansı sağlayan Kosinüs ve Genişletilmiş Jaccard benzerlikleridir. Web sayfaları ve belgelerinde Kosinüs benzerlik ölçütünün kullanılmasının uygun olduğu yapılan test sonuçlarından görülmektedir.

KAYNAKLAR

[1] Steinbach, M.; Karypis, G.; Kumar, V.: "A Comparison of Document Clustering Techniques". In KDD Workshop on Text Mining, **2000**.
[2] Zamir, O.; Etzioni, O.: "Web Document Clustering: A Feasibility Demonstration," Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 46-54, **1998**.

[3] Bouguettaya, A.: "On-Line Clustering", *IEEE Trans. on Knowledge and DataEngineering,* Vol. 8, No. 2, (**1996**).

[4] Aas, K.; Eikvil, L.: "Text categorisation: A survey", *Technical report,* Norwegian Computing Center, June 1999.

[5] Mendes M.E.S. Rodrigues; Sacks, L.: "A scalable hierarchical fuzzy clustering algorithm for text mining", *In: Proc. of the 4th International Conference on Recent Advances in Soft Computing*, RASC'2004, pp. 269-274, Nottingham, UK, Dec. **2004**

[6] Kosala R.; Blockeel H.: "Web Mining Research: A Survey", *ACM SIGKDD Explorations Newsletter*, 2(1):1–15, **2000**.

[7] Kaufman, L.; Rousueeuw, P.J.: *Finding Groups in Data: an Introduction to Cluster Analysis*, John Wiley & Sons, **1990**.

[8] Bilgin, T.T.; Çamurcu, A. Y.: "A Modified Relationship Based Clustering Framework for Density Based Clustering and Outlier Filtering on High Dimensional Datasets*.", PAKDD 2007: Lecture Notes in Computer Science*, 4426, sayfa 409-416, Springer, **2007**.

[9] Pang-Ning Tan, P.N.; Steinbach, M.; Kumar, V.: *Introduction to Data Mining.*, Addison Wesley, Mart 2006.

[10] Işık M.: "Bölünmeli Kümeleme Yöntemleri İle Veri Madenciliği Uygulamaları", *Yüksek Lisans Tezi*, Marmara Üniversitesi, Fen Bilimleri Enstitüsü, **2006**.

[11] Strehl, A.; Ghosh, J.; Money, R.: "Impact of Similarity Measures on Web-page Clustering", *AAAI Workshop on AI for Web Search*, 58-64, **2000**.

[12] Strehl, S.: "Relationship-based Clustering and Cluster Ensembles for High-Dimensional Data Mining ", *Doktora Tezi*, The University of Texas at Austin, USA, **2002**.

[13] Gover, J.C.: "Discussion of a paper by R.M. Cormack", J. Roy. *Statist. Soc. Ser.* A 134, 360-365, **1971**.

[14] Jain, A.K.; Murty, M.N.; Flynn P.J.: "Data Clustering: A Review", *ACM Computing Surveys*, Vol. 31, No. 3., Eylül 1999.

[15] Schenker, A.; Kandel, A.; Bunke, H.; Last, M.: "Graph-Theoretic Techniques For Web Content Mining"., *World Scientific Publishing Company,* Mayıs 2005.

[16] Veri setlerinin kaynağı : Işık D.; Dolu, O.; Özbek, U.: "Web Sayfalarının Özelliklerini Elde Eden ve Web Sayfaları Benzerlik Ölçütlerini Karşılaştıran Uygulama", *Lisans Tezi*, Bilgisayar Mühendisliği, ITU, **2006**.