Mühendislik ve Fen Bilimleri Dergisi
Journal of Engineering and Natural Sciences

Sigma 5
8-14,
2013

ÖZEL SAYI

Araştırma Makalesi / Research Article

METİNLERİN ANLAMSAL UZAYDAKİ TEMSİL YÖNTEMLERİNİN SINIFLANDIRMA PERFORMANSINA ETKİLERİ

M. Fatih AMASYALI*, Mahmut ÇETİN, Cenk AKBULUT

Yıldız Teknik Üniversitesi, Elektrik-Elektronik Fakültesi, Bilgisayar Müh. Bölümü, Esenler-İSTANBUL

Geliş/Received: 07.02.2013 Revised/Düzeltme: 13.05.2013 Kabul/Accepted: 15.05.2013

ÖZET

Metinlerin sınıflandırılmasında en çok tartışılan sorun metinlerin nasıl temsil edileceğidir. Kelimelerin kendileri, kökleri, karakter ngramları ve anlamsal uzaylar en yoğun olarak kullanılan yöntemlerdir. Bu çalışmada saklı anlam indeksleme ve önerdiğimiz birlikte geçiş matrisi tabanlı anlamsal uzay yöntemleri diğer yöntemlerle 30 sınıflı bir veri kümesi üzerinde karşılaştırılmıştır. Birlikte geçme matrisi tabanlı yöntemin diğer tüm yöntemlere göre çok daha yüksek başarılara ulaştığı görülmüştür. Ayrıca yöntemin diğer tüm yöntemlerin aksine veri kümesindeki sınıf sayısı arttıkça başarısında büyük düşüşler olmadığı görülmüştür.
Anahtar Sözcükler: Metin sınıflandırma, anlamsal uzay.

EFFECTS OF TEXT REPRESENTATION METHODS IN SEMANTIC SPACE ON CLASSIFYING PERFORMANCE

ABSTRACT

The most discussed issue about classification of texts is how to represent them. Words, stem words, character ngrams and semantic spaces are the common methods. In this research latent semantic indexing and semantic space based on co-occurrence matrix methods are compared with other methods on 30 classed data set. According to other methods the semantic space based on co-occurrence matrix method performs higher success. In addition, performance success of this method does not decrease as much as other methods while number of classes increased.
Keywords: Text classification, text categorization, semantic space.

1. GİRİŞ

Metinlerin nasıl temsil edileceği, metin sınıflandırma problemlerinin en temel sorusudur. Bugüne kadar bu soruya birçok cevap verilmiştir. En çok kullanılan metin temsil yöntemleri; kelimelerin kendilerinin, kelimelerin köklerinin, karakter ngram'larının metinlerdeki geçiş sayıları ve metinlerin anlamsal uzayda temsilidir [1]. Metinlerin anlamsal uzayda temsili için en çok kullanılan yöntem ise Saklı Anlam İndekslemedir [2]. Saklı anlam indekslemede metinler kelime uzayından daha düşük boyutlu anlam uzayına tekil değer ayrıştırması yapılarak dönüştürülmektedir. Bu yöntemle hem metinlerin hem de kelimelerin anlamsal koordinatları elde edilmektedir. Kelimelerin anlamsal uzaydaki koordinatları kelimelerin anlamsal benzerlikleriyle uyumludur.

* Corresponding Author/Sorumlu Yazar: e-mail/e-ileti: mfatih@ce.yildiz.edu.tr, tel: (212) 383 57 30

Literatürde kelimelerin birbirlerine anlamsal benzerliklerini ölçmek için birçok çalışma yapılmıştır. Li ve arkadaşları, kelimelerin benzerliğini büyük metin kütüphanelerinde kelimelerin aynı metinde birlikte geçme sıklıklarını ve kelimelerin Wordnet [3] hiyerarşisinde birbirlerine olan uzaklıklarını birlikte kullanarak ölçmüşlerdir [4]. Sonuçlarını insan deneklerin verdiği cevaplarla karşılaştırmışlardır. Guihong Cao ve arkadaşları [5] ise kelimelerin aynı metinde birlikte geçme sıklıklarını Reuters külliyatında hesaplamışlardır. Amasyalı ve Beken [6], kelimelerin benzerlikleri büyük bir metin kütüphanesindeki kaç adet metinde birlikte geçtikleriyle ölçülmüşlerdir. Oliva ve arkadaşları, Wordnet'i ve cümlelerin çözümlenme ağaçlarını cümlelerin anlamsal benzerliklerini ölçmede kullanmışlardır [7]. Wenyin ve arkadaşları, kelimelerin ve metinlerin benzerliklerini iteratif bir yöntemle belirlemişlerdir [8].

Literatürde kelimelerin benzerliklerinin genelde kaç tane metinde birlikte geçtikleriyle ölçüldüğü görülmektedir. Bu yöntem metinlerin sınıflarına ihtiyaç duymadığı için eğiticisiz bir yöntem olarak düşünülebilir. "Eğer elimizde metinlerin sınıfları varsa benzerlik hesabı için daha etkin bir yöntem kullanılamaz mı?" sorusu bu çalışmanın çıkış noktalarından biridir. Diğer taraftan kelimelerin birlikte geçtikleri pencere boyutunun artışıyla, metinleri sınıflandırma başarısının da arttığı bilinmektedir. Bu nedenle literatürde pencere boyutunun en büyük hali olan aynı metin içinde geçme kullanılmıştır. "Pencere boyutunu daha da büyütmek mümkün müdür?" sorusu da bu çalışmadaki diğer çıkış noktamızdır.

Çalışmamızda bu iki soruya tek bir cevap verilmiştir. Bir sınıfa ait tüm metinleri tek bir metin gibi düşünürsek hem pencere boyutunu arttırmış hem de sınıf bilgisini kelimelerin benzerlik hesabının içine katmış oluruz. Bu yönteme göre hesaplanan kelime benzerlikleriyle İngilizce köşe yazıları üzerinde yapılan denemeler sonucunda, diğer metin temsil yöntemleri en fazla %86.3'lük bir başarı elde edebilirken, önerdiğimiz yöntemin %98.6'lık başarıya ulaştığı görülmüştür.

2. BİRLİKTE GEÇME MATRİSİ TABANLI ANLAMSAL UZAY(BİGM)

Metinlerin birlikte geçme matrisi tabanlı anlamsal uzayda temsil edilmeleri için önce kelime köklerinin anlamsal uzaydaki koordinatları hesaplanmakta daha sonra bu koordinatlar kullanılarak metinlerin anlamsal uzaydaki koordinatları bulunmaktadır [6].

Kelimelerin anlamsal uzaydaki koordinatları birbirlerine anlamca yakınlıklarına göre oluşturulmaktadır. Harris [9], iki kelimenin birlikte geçtiği doküman / cümle sayısının iki kelimenin benzerliğiyle doğru orantılı olduğunu öne sürmüştür. Bu yaklaşımdan hareketle, kelimelerin anlamsal yakınlıklarını ifade etmek için birlikte geçme matrisi kullanılmıştır. Bu matris, kelimelerin seçilen pencere boyutu kadar aralıkta beraber kaç kez geçtiğini göstermektedir. Birlikte geçme matrisinin [i,j] hücresinin değeri, i. kelime ile j. kelimenin tüm metinler içinde çeşitli boyuttaki çerçevelerde birlikte bulunma sayılarını ifade eder.

Birbirlerine uzaklıkları bilinen ancak koordinatları bilinmeyen nesnelerin bir uzaklık matrisine uygun olan sayısal koordinatlarının bulunması için Çok Boyutlu Ölçekleme (Multi Dimensional Scaling) [10, 11] metodu kullanılmaktadır. Kelimelerin birlikte geçme matrisi anlamsal bir yakınlık matrisi olduğundan ters çevrilerek (uzaklık=1/yakınlık) anlamsal uzaklık matrisi elde edilmekte ve çok boyutlu ölçekleme fonksiyonuna verilmektedir. Bu işlem sonucunda kelimelerin anlamsal uzaydaki koordinatları elde edilmektedir.

Metinlerin anlamsal uzaydaki koordinatları bulunurken, içerdikleri kelime koordinatlarının ortalaması alınmaktadır.

Bu çalışmada, Amasyalı ve Beken'in [6] önerdiği bu yöntemin benzerlik matrisi oluşturma süreci irdelenmiştir. Amasyalı ve Beken, benzerlik matrisi oluşturulurken sınıflandırılacak metin kümesinden ayrı olarak harici bir doküman kütüphanesi kullanılmışlardır. Kelimelerin birbirlerine benzerliklerini kaç dokümanda birlikte geçtiklerine göre ölçülmüş ve kelime koordinatları bu verilere göre hesaplanmıştır.

Bu çalışmanın başlangıç noktasında, birlikte geçme matrisini oluştururken harici ve

sabit olarak kullanılan kütüphane yerine sınıflandırılacak metinlerin kendilerinin kullanılmasının probleme özgü bir anlamsal uzayın oluşturulmasını sağlayacağı düşünülmüştür.

Birlikte geçme matrisinin oluşturulmasında kullanılan pencere boyutu çok küçük seçildiğinde matrisin içerisindeki çoğu değer 0 olacaktır ve bunun sonucu olarak Çok Boyutlu Ölçekleme yöntemine kelime koordinatlarını hesaplaması için çok az miktarda veri verilebilecektir. Bu sebeple literatürdeki çalışmalarda pencere boyutu tüm metin olarak kullanılmıştır. Pencere boyutunun artışının performans üzerindeki olumlu etkisi düşünüldüğünde pencere boyutunun daha da büyütülmesi fikri doğmuştur. Bunun sonucu olarak bu çalışmada pencere boyutu olarak "aynı sınıf" seçeneği önerilmiştir. Buna göre iki kelimenin benzerliği kaç sınıfta birlikte geçtiklerine göre ölçülür. Buna göre aynı sınıftaki tüm metinler tek bir metinmiş gibi düşünülmektedir. Bu yaklaşım ayrıca, sınıf bilgisini de kullandığı için, "aynı metin"i pencere boyutu olarak kullanan diğer yöntemler eğiticisiz olarak ifade edilebilecekken, eğiticili bir yöntem olarak düşünülebilir. Bu sayede kelimelerin benzerlikleri hem pencere boyutu büyütüldüğünden hem de sınıf bilgisi işin içine katıldığından hem daha doğru hem de eldeki sınıflandırma problemine özgü bir biçimde belirlenebileceği düşünülmüştür. Bu fikirsel altyapının gerçek bir problem üzerinde sınanması için denemelerimiz 30 sınıf içeren bir veri kümesi üzerinde yapılmış ve performansı literatürdeki birçok yöntemle karşılaştırılmıştır.

3. KARŞILAŞTIRMADA KULLANILAN DİĞER METİN TEMSİL YÖNTEMLERİ

Bu bölümde birlikte geçme matrisi tabanlı anlamsal uzayın performansının karşılaştırıldığı diğer metin temsil yöntemlerine yer verilmiştir.

Sayılar: Metinlerdeki toplam kelime ve noktalama işaretlerinin sayılarını, cümle başına oranlarını içerir.

Kelimeler: Bu temsil yönteminde metinler içerdikleri kelimelerin frekanslarıyla ifade edilirler. Bu yönteme göre satırlarında metinlerin, sütunlarında kelimelerin yer aldığı bir matris oluşturulur. Matrisin [i,j] gözünde i.metinde j.kelimenin kaç kere geçtiği bilgisi tutulur. Matrisin satır sayısı metin sayısına, sütun sayısı ise tüm metinlerde en az bir kere geçen farklı kelime sayısına eşittir.

Kelime Kökleri: Kelimeler yönteminden tek farkı kelimelerin kendilerinin yerine köklerinin kullanımıdır. Kelime kökleri kullanımıyla "visiting" ve "visited" kelimelerinin ikisi de "visit" kökü ile ifade edilmektedir ki bu sayede metinlerin temsil edildiği boyut sayısı azalmaktadır.

Karakter Ngram'ları: Ngramlar N boyutlu karakter çerçeveleridir. 2Gram iki boyutlu, 3Gram ise üç boyutlu karakter kümesidir. Kelimeler yönteminden tek farklı kelimelerin kaç kere geçtikleri yerine Ngramların kaç kere geçtiklerinin kullanılmasıdır.

Saklı Anlam İndeksleme: Literatürde en yaygın kullanıma sahip özelliklerden biridir. Doküman Matrisi (A) üzerinde Tekil Değer Ayrıştırma (Singular Value Decompositon) yaparak metinlerin ifadesinde kullanılan boyut sayısını azaltır. Bunun için A matrisi (m*n'lik bir matris, m→metin sayısı, n→farklı kelime sayısı) aşağıdaki eşitliğe göre öz değer-öz vektör çarpanlarına ayrılır.

$$A_{m\times n} = U_{m\times r}.S_{r\times r}.V_{r\times n} \quad (1)$$

Bu işlem sonunda S matrisinde, öz vektörlerin öz değerleri diyagonalde büyükten küçüğe sıralanır. Kullanıcının seçtiği boyut (r) kadarı işleme alınarak, A matrisi r boyutlu U matrisine dönüşmüş olur. Bu sayede n boyutlu metinler, r boyutlu anlamsal bir uzayda temsil edilmiş olurlar.

Kavram Genelleştirme: Metinlerde geçen kelimelerinin yerine bunların genelleştirilmiş hallerinin kullanılmasıdır. Örneğin "elma", "armut", "kiraz" kelimeleri yerine "meyve" kavramı kullanılacaktır. Diğer bir deyişle kelimeler ait oldukları genelleştirilmiş kavramlara göre kümelenecektir. Bunun için kelimelerin ve genelleştirilmiş hallerinin yer aldığı bir kaynağa ihtiyaç vardır. Bu kaynak için, İngilizce metinlerle çalıştığımızdan Wordnet [3] kullanılmıştır. Wordnet'te yer alan "hypernym" ilişkilerinden yaklaşık 80 bin tane kavram-üst kavram ikilisi

çıkarılmıştır. Bir kelime kavram-üst kavram ikilileri listesindeki kavramlar içinde yer alıyorsa, onun yerine karşılık gelen üst kavram kullanılarak metinlerin ifade edildiği boyut sayısı azaltılmıştır.

4. DENEYSEL SONUÇLAR

Kullanılan veri kümesinde 30 yazarın her birine ait 50'şer yazı bulunmaktadır. Diğer bir deyişle 1500 örneği 30 sınıfı olan bir sınıflandırma problemi çözülmek istenmiştir. Problemde eşit sayıda örnekler içeren 30 sınıfa ait metinler olduğundan rastgele başarı oranı %3,33'tür. Metinler çeşitli yöntemlerle temsil edildikten sonra WEKA [12] kütüphanesinde yer alan karar ağacı (J48), en yakın komşu (IB1), karar destek makinesi (SMO), rastsal ormanlar (RF) ve Naive Bayes (NB) sınıflandırma algoritmalarıyla modellenmiştir. Algoritmaların sınıflandırma performansları 10'lu çapraz geçerleme ile ölçülmüş ve ortalamaları alınmıştır.

Metinlerde her sınıfta çok fazla kullanılan özelliklerin ayırt ediciliği az olmasına karşın, çok az kullanılan özellikler de gereksiz yere veri kümesinin boyutunu arttırmaktadır. Bu nedenle veri kümesinin boyutunu azaltmak için kelimeler, kelime kökleri ve karakter Ngram'ları özelliklerini kullanan tüm temsil yöntemlerinde minimum-maksimum frekans aralığında kalan özellikler kullanılmıştır. Minimum ve maksimum frekans değerlerinin belirlenmesinde birçok deneme yapılmış ve 10-100 aralığının kullanılmasına karar verilmiştir.

Birlikte geçme matrisi tabanlı anlamsal uzay yönteminin, birlikte geçme matrisinin oluşturulmasında kullanılan pencere boyutu ve metinlerin anlamsal uzayda temsil edileceği boyut sayısı olmak üzere 2 parametresi bulunmaktadır. Bu parametrelerin sınıflandırma performansına etkileri incelenmiştir. Pencere boyutu parametresinin etkileri Çizelge 1'de gösterilmiştir. 5 farklı pencere boyutu için, 5 farklı sınıflandırıcı ile elde edilen sonuçlar yüzde başarı türünden verilmiştir. Tüm pencere boyutlarında, metinlerin anlamsal uzayda temsil edileceği boyut sayısı 50'dir. Pencere boyutunun yanında o pencere boyutu kullanıldığında birlikte geçme matrisindeki "0"ların yüzdelik oranı verilmiştir.

Çizelge 1. Pencere Boyutunun Başarıya Etkisi

Pencere Boyutu (0 oranı)	**J48**	**IB1**	**SMO**	**RF**	**NB**
2 (99.7)	25.5	29.8	36.4	34.3	49.9
3 (99)	32.1	39.6	45.1	40.9	47.8
5 (98.6)	36.4	42.6	49.5	43.4	52.9
Aynı metin (60)	44.5	58.0	70.9	53.7	68.9
Aynı sınıf (11)	86.3	93.2	98.6	90.5	96.2

Pencere boyutunun bir sayı olması iki kelimenin tüm metinlerde o sayı büyüklüğündeki bir pencere içinde kaç kez birlikte bulunduğunu göstermektedir. Aynı metin ya da sınıf içinde olması ise sırasıyla iki kelimenin birlikte bulundukları metin ya da sınıf sayısını göstermektedir. Buna göre pencere boyutu "aynı metin" ise birlikte geçme matrisindeki sayılar 0 ile toplam metin sayısı aralığında, pencere boyutu "aynı sınıf" ise sayılar 0 ile sınıf sayısı aralığında olacaktır.

Çizelge 1 incelendiğinde, başarı oranlarının değişmesinde birlikte geçme matrisinde bulunan "0" değerlerinin oranının büyük etkisi olduğu görülmektedir. Bir veri kümesinde herhangi iki kelimenin 2, 3 ya da 5 pencere boyutunda birlikte bulunma olasılığı çok düşük olduğundan matris "0" değerleri ile doludur. Bu da kelimelerin dolayısıyla metinlerin koordinatlarının isabetli yerleştirilememesine sebep olmaktadır. Pencere boyutu artıp, "0" ların sayısı azaldıkça kelimeler arasındaki benzerliğe dair daha fazla bilgi kullanılmaktadır. Bu sayede kelime koordinatlarının anlamlılıkları ve dolayısıyla sınıflandırma başarısı yükselmektedir. Bu

sonuçlar beklentilerimizle uyumlu olduğundan pencere boyutu olarak "aynı sınıf"ın kullanılmasına karar verilmiştir.

Birlikte geçme matrisi tabanlı anlamsal uzay yönteminin ikinci parametresi olan metinlerin kaç boyutlu bir uzayda ifade edileceğini belirlemek içinde benzer bir deneme yapılmıştır. Metinler sırasıyla 50, 30 ve 10 boyutlu uzaylarda temsil edilmişler ve boyutun hem saklı anlam indekslemede hem de birlikte geçme matrisi tabanlı anlamsal uzayda etkileri incelenmiştir. Çizelge 2'de boyut sayısının başarıya etkisi görülmektedir. Birlikte geçme matrisi tabanlı anlamsal uzayda pencere boyutu "aynı sınıf" olarak kullanılmıştır.

Çizelge 2. Birlikte geçme matrisi tabanlı anlamsal uzayda (BiGM) ve saklı anlam indekslemede (SAİ) boyutun başarıya etkisi

Yöntem -boyut	**J48**	**IB1**	**SMO**	**RF**	**NB**
BiGM-50	86.3	93.2	98.6	90.5	96.2
BiGM-30	77.3	85.9	90.6	82.7	90.1
BiGM-10	53.7	59.6	51.9	59.3	62
SAİ-50	40.7	32.1	50.6	40.1	50
SAİ-30	13.5	22.8	8.7	20.2	16.7
SAİ-10	12.6	18.1	5.7	18	12

Çizelge 2 incelendiğinde her iki yöntemde de metinlerin temsil edildikleri boyut sayısı arttıkça başarı oranının arttığı görülmektedir. Bununla birlikte geçme matrisi tabanlı anlamsal uzayın (BiGM) saklı anlam indekslemeye (SAİ) göre çok daha başarılı sonuçlar ürettiği görülmektedir. Bu sonuçlara göre boyut sayısı olarak 50'nin kullanılmasına karar verilmiştir. Muhtemelen daha büyük boyut sayısı kullanımı başarı oranını daha da arttıracaktır.

Birlikte geçme matrisi tabanlı anlamsal uzayın iki parametresinin değerlerinin seçimi işlemi tamamlandıktan sonra yöntemin diğer metin temsili yöntemleriyle karşılaştırılması yapılmıştır. Çizelge 3'te metinlerin temsilinde kullanılan yöntemler, boyut sayıları ve WEKA'da yapılan sınıflandırma denemelerinin sonuçları yüzdelik başarı türünden verilmiştir.

Çizelge 3. Yöntemlerin Başarı Oranları

KullanılanTemsil Yöntemi (özellik sayısı)	**J48**	**IB1**	**SMO**	**RF**	**NB**
Sayılar (10)	41.4	39.9	33.1	47.7	40.4
Kelime Kökleri (4137)	38	6.7	68.9	32.8	58.7
2gramlar (537)	25.8	15.5	37.2	25.9	41.3
3gramlar (5144)	41	8.4	68.9	32.8	58.1
Kelimeler (6323)	44.9	14.13	82.3	34.7	60.9
BiGM (50)	86.3	93.2	98.6	90.5	96.2
SAİ (50)	40.7	32.1	50.6	40.1	50
Kavram Genelleştirme (1487)	38.5	72.3	72.3	47.5	57.3

Çizelge 3 incelendiğinde önerdiğimiz "Birlikte geçme matrisi tabanlı anlamsal uzay"ın diğer yöntemlere göre çok daha başarılı olduğu (%98.6) görülmektedir. Ayrıca en başarılı sınıflandırma algoritmasının destek vektör makineleri (SMO) olduğu, yüksek boyutlu veri kümelerinde en yakın komşu algoritmasının (IB1) başarısız olduğu, karakter 3gramlarının 2gramlardan daha başarılı olduğu, kelimelerin kendilerini kullanmanın köklerini ve

genelleştirilmiş hallerini kullanmaktan daha iyi olduğu da göze çarpmaktadır.

Yüksek başarısının yanında kullanılan sınıflandırıcıya en az duyarlı yönteminde yine birlikte geçme matrisi tabanlı anlamsal uzay olduğu görülmektedir. Örneğin ikinci en başarılı temsil yöntemi olan kelimeler, SMO ile %82.33'lük başarı elde ederken, 1NN ile başarısı 14.13'tür. Bunun sebebi temsil yönteminin içerdiği boyut sayısıdır. Boyut sayısı ne kadar büyükse, sınıflandırıcı performansları arasındaki farklılık o kadar artmaktadır. Önerdiğimiz yöntem, metinleri az sayıda boyutta temsil ettiğinden kullanılan sınıflandırıcıya duyarlılığı azdır.

Buraya kadarki denemelerde 30 sınıflı bir veri kümesi kullanılarak denemeler yapılmıştır. Veri kümesindeki sınıf sayısının artışının problemin zorluğunu arttırdığı bilinen bir olgudur. Sınıf sayısının birlikte geçme matrisi tabanlı anlamsal uzay ve saklı anlam indeksleme yöntemleri üzerindeki etkisini incelemek amacıyla 30 sınıflı veri kümesinden 10 sınıflı rastgele bir alt veri kümesi elde edilmiştir. Çizelge 4'te bu iki yöntemin başarı oranları verilmiştir. Her iki yöntemde de 50 boyutlu uzaylar kullanılmıştır. Birlikte geçme matrisi tabanlı anlamsal uzayın pencere boyutu "aynı sınıf" tır.

Çizelge 4. Birlikte geçme matrisi tabanlı anlamsal uzayda (BiGM) ve saklı anlam indekslemenin (SAİ) sınıf sayısına göre başarıları

Problemdeki sınıf sayısı (yöntem)		**J48**	**IB1**	**SMO**	**RF**	**NB**
30	(BiGM)	86.2	93.2	98.6	90.4	96.2
10	(BiGM)	92	95.4	99.6	96.4	99.2
30	(SAİ)	40.7	32.1	50.6	40.1	50
10	(SAİ)	56	69.8	72.8	64.2	59

Çizelge 4 incelendiğinde sınıf sayısının artışının başarıyı her iki yöntemde de düşürdüğü görülmektedir. Ancak bu düşüş saklı anlam indeksleme de çok fazla iken birlikte geçme matrisi tabanlı anlamsal uzayda çok daha azdır. Bu denemeden çıkan sonuca göre önerdiğimiz yöntemin yüksek sınıf sayısına sahip veri kümelerinde başarıyla kullanılabileceği söylenebilir.

5. SONUÇLAR VE GELECEK ÇALIŞMALAR

Metinlerin ve kelimelerin anlamsal yakınlıklarının bilgisayarlarca belirlenebilmesi; metin sınıflandırma, insan hafızasının modellenmesi, metin özetleme, otomatik soru cevaplama gibi birçok uygulama alanına sahiptir.

Literatürde bu amaçla birçok çalışma yapılmıştır. Bu çalışmalarda, kelimelerin anlamsal benzerlikleri temelde ya metin-kelime matrisi üzerinde tekil değer ayrıştırma yapılarak, ya kaç dokümanda birlikte geçtiklerine bağlı olarak ya da önceden oluşturulmuş bir taksonomi üzerinden yapılmaktadır.

Bu çalışmada, metinlerde geçen kelimelerin benzerlikleri hesaplanırken metinlerin sınıf bilgilerinin de kullanımı önerilmiş ve mevcut yaklaşımlara göre metin sınıflandırma problemleri için çok daha başarılı sonuçlar elde edilmiştir. İki kelimenin benzerlik ölçütü olarak kaç sınıfta birlikte geçtiklerini kullanan yaklaşımımız, hem birlikte geçme matrisinin seyrekliğini azaltıp hem de metin sınıflandırma problemlerinde mevcut olan metin sınıfı bilgisini kelimelerin anlamsal benzerliklerinin ölçümünde kullanmaktadır. Birlikte geçme matrisindeki her bilgi, kelimelerin ve dolayısıyla metinlerin anlamsal uzaydaki koordinatlarının belirlenmesinde bir kısıtı ifade etmektedir. Bu matris ne kadar seyrek olursa o kadar az kısıt olmakta ve kelime koordinatları gürbüz olarak oluşturulamamaktadır. Matrisin seyrekliği önerdiğimiz yolla azaltıldığında ise, çok sayıda kısıta uyan koordinatlar daha gürbüz bir şekilde oluşturulabilmektedir. Birlikte geçme matrisi oluşturulurken doküman sayılarını kullanmak metin sınıflandırma problemlerinde elde olan sınıf bilgisini kullanmamaktadır. Bu değerli bilginin

kelimelerin benzerlik ölçütüne dahil edilmesi kelimelerin anlamsal benzerliklerinin eldeki problem alanına özgü olarak belirlenmesine ve dolayısıyla metinlerin daha başarılı sınıflandırılmasına olanak sağlamaktadır.

Çalışmamızda elde ettiğimiz bulgular aşağıda listelenmiştir:

- Önerilen yöntem, mevcut metin temsili yöntemlerine göre çok daha başarılıdır.
- Yöntem, özellikle sınıf sayısı fazla olan problemler için uygundur. Az sınıf sayısına sahip veri kümelerinde çoğu kelimenin birbirine benzerliği aynı olacağından kelimelerin koordinatları gürbüz bir şekilde belirlenemeyecektir. Sınıf sayısı az olan problemlerde pencere boyutu olarak "aynı metin" kullanımı daha başarılı olacaktır.
- Yöntem harici metin kütüphanelerine ihtiyaç duymamaktadır. Eğitim kümesindeki metinler yeterli olmaktadır.
- Önerilen yöntem herhangi bir dile özgü değildir. Her dil için kullanılabilir.
- Eşanlamlı kelimelerin tek bir kelime olarak değerlendiriliyor oluşu metodun dezavantajıdır.
- Metinlerin anlamsal koordinatlarının hesabı, eğitim işlemi sırasında zaman almakla birlikte, test işlemi metinler oldukça az sayıda boyutla ifade edildiğinden oldukça hızlı yapılabilmektedir.

Bu çalışmada kullanılan veri kümesine www.kemik.yildiz.edu.tr/data/File/30Columnists.z ip adresinden erişilebilir.

Önerdiğimiz metin temsil yönteminin, metin boyutlarının çok kısa olduğu sosyal paylaşım platformlarındaki mesajların sınıflandırılmasında etkin olarak kullanılabileceği düşünülmektedir.

KAYNAKLAR / REFERENCES

[1] L. Ciya, A. Shamim, D. Paul, "Feature Preparation in Text Categorization", Oracle Text Selected Papers and Presentations, 2001.

[2] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, ve R. A. Harshman. "Indexing by latent semantic analysis", Journal of the American Society of Information Science, 41(6), 1990, s.391–407.

[3] C. Fellbaum, WordNet: An Electronic Lexical Database, MIT Press, 1998.

[4] Y. Li, Z.A. Bandar, D. McLean, "An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources", IEEE Transactions on Knowledge and Data Engineering, vol. 15, 2003 s. 871-882.

[5] C. Guihong, S. Dawei, B. Peter, "Fuzzy K-Means Clustering on a High Dimensional Semantic Space", Advanced Web Technologies and Applications (LNCS 3007) - The Sixth Asia Pacific Web Conference (APWeb'04), 2004.

[6] M.F. Amasyalı, A. Beken, "Türkçe Kelimelerin Anlamsal Benzerliklerinin Ölçülmesi ve Metin Sınıflandırmada Kullanılması", IEEE Signal Processing and Communications Applications Conference,SİU-2009.

[7] J. Oliva, J. I. Serrano, M. D. Castillo, A. Iglesias, "SyMSS: A syntax-based measure for short-text semantic similarity", Data & Knowledge Engineering, 70, 2011, s.390–405.

[8] L. Wenyin, X. Quan, M. Feng, B. Qiu, "A short text modeling method combining semantic and statistical information", Information Sciences,180, 2010, s.4031–4041.

[9] Z.S., Haris, "Mathematical structures of language", Wiley, s.12, 1968.

[10] E. Alpaydın, "Introduction to Machine Learning", The MIT Press, s. 121-124, 2004.

[11] Multidimensional Scaling for Java, University of Konstanz, Department of Computer & Information Science, Algorithmics Group, http://www.inf.unikonstanz.de/algo/software/mdsj/ [Erişim Tarihi: 2012, Aralık].

[12] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.