KELİME YÖRÜNGELERİ
WORD TRAJECTORIES

M.Fatih Amasyalı[1], Yunusemre Yener[2],Selim Serkan Kaplan[3]

[1,2,3]Bilgisayar Mühendisliği Bölümü
Yıldız Teknik Üniversitesi
mfatih@ce.yildiz.edu.tr, y.emreyener@gmail.com, serkankaplan@gmail.com

ÖZETÇE

İnsanların düşünce süreçleri birbirine ne kadar benzemektedir? Zekanın mekanizmasını anlamak, insanın zihnindeki düşünce akışını modellemek dünyamızın gizemini hala koruyan sorularından biridir. İnsanların düşünce süreçlerini gözlemlemek, direkt ölçmek mümkün değilse de onu, ürünlerinin bazılarıyla (konuşma ve yazılarıyla) dolaylı olarak gözlemlemek mümkündür. Bu çalışmada çeşitli köşe yazarlarının yazıları kullanılarak, metinlerinin yörüngeleri oluşturulmuştur. Bunun için öncelikle, kelimelerin birbirlerine anlamca yakınlıklarına göre düzenlenmiş koordinatları bulunmuş ve kelimelerin koordinatları, kelimelerin metinlerdeki geçiş sırasına göre birleştirilerek metin yörüngeleri elde edilmiştir. Yörüngelerin kişiye özelliğini incelemek için yörüngelerdeki ardışık kelimeler arasındaki açılar ve mesafeler kullanılmıştır. Bu sayede metinlerin içerikleriyle hiç ilgilenilmeden sadece oluşturdukları yörüngenin şekli kullanılmıştır. Sonuç olarak kişilerin birbirlerinden en çok kavramlar arası mesafe ve açılarının frekanslarına göre ayrıldığı görülmüştür.

ABSTRACT

How much people's thought processes are similar to each other? Modeling the thought processes is one of the questions that still preserves the mystery of the world. Thought process cannot be directly observed. But it is possible to measure indirectly with its products such as speaking, writing. In this study, word paths of several texts were constructed. For this purpose, firstly, the semantic coordinates of the words are calculated by using cooccurance matrix. Then the word trajectories are constructed by combining word coordinates according to the order of words in texts. The trajectories were represented with the histograms of the distances and angles between successsive words. So, the text are represented with only the shapes of the trajectories not its contents. According to the our experiments, authors of the texts can be discriminated by the frequency of the distance and angles between the words.

1. GİRİŞ

İnsanların düşünce süreçlerinin incelenmesi çok disiplinli çalışmaların yürütüldüğü bir çalışma alanıdır. Düşünme süreçleri direkt olarak ölçülemediğinden dolaylı ölçüm yöntemleri geliştirilmiştir. Beyin dalgalarının kullanımı eskiden beri uygulanan bir yöntemdir. Teknolojinin gelişmesiyle beyin görüntüleme tekniklerinin (fMRI) kullanımı bu çalışmalara büyük bir ivme kazandırmıştır. Bu çalışmada ise düşüncelerimizin bir ürünü olan yazılarımız kullanılarak, bu süreçlerin 3 boyutlu bir uzayda yörünge olarak temsil edilmesi sağlanmıştır. Farklı kişilerin yazdıkları yazıların yörüngelerinin birbirlerinden farklı olup olmadıkları, aynı kişinin farklı yazılarının benzer yörünge özellikleri taşıyıp taşımadıkları incelenmiştir.

Metin yörüngelerinin oluşturulması düşüncesi daha önceki çalışmalarımızda [1, 2, 3] yer alan "metinlerin anlamsal uzayda temsil edilmesi" fikrinden üretilmişitir. Metinlerin içerdikleri kelimeleri çok boyutlu anlamsal bir uzaya taşıdıktan sonra, kelimelerin metinde geçme sıraları da göz önüne alındığında metinlerin çok boyutlu bir uzayda birer yörünge oluşturdukları görülmüştür. Ve bu yörüngelerin çeşitli özelliklerinin incelenmesine başlanmıştır.

Yazının 2. bölümünde metinlerin yörüngeye dönüştürülme süreci anlatılmıştır. 3. bölümde elde edilen yörüngelerin hangi özelliklerinin çıkarıldığı belirtilmiştir. 4. bölümde ise farklı yazarların yazıları üzerinde yapılan deneysel çalışmalar yer almaktadır. Son bölümde, bu çalışmadan çıkan sonuçlar ve gelecek çalışmaların olası yönleri üzerinde durulmuştur.

2. METİN YÖRÜNGELERİNİN OLUŞTURULMASI

Yörüngelerin bulunması için öncelikle yörüngeyi oluşturan nesne/ kavram/ kelimelerin koordinatları bulunmalıdır. Koordinatlar bulunduktan sonra bir metnin yörüngesi içinde geçen kelimelerin koordinatlarının metindeki geçiş sırasına göre birleştirilmesiyle elde edilmektedir. Kelimelerin koordinatları bulunurken Harris'in [4] "Metinlerde birlikte kullanılan kelimeler anlamca birbirlerine yakındır." ifadesinden hareketle kelimelerin birlikte geçiş matrisi elde edilmektedir. Bu matrisin elde edilmesinin ardından kelimelerin bu yakınlıklarını temsil edebilecek uygun koordinatlarının bulunması gerekmektedir. Bunun için çok boyutlu ölçekleme (Multi Dimensional Scaling- MDS) [5] yönteminden yararlanılmıştır.

2.1. Birlikte Geçme Matrisinin Oluşturulması

Birlikte geçme (cooccurance) matrisi Bu matris tüm metinlerde geçen tekil kelime sayısı boyutlu, karesel ve simetrik bir matristir. Matrisin i, j. gözünde i. kelime ile j. kelimenin birlikte kaç kez kullanıldığı yer almaktadır. Bu değerin büyük olması bu iki kelimenin birbirine anlamca

978-1-4673-0056-8/12/$26.00 ©2012 IEEE

yakın olduğunu göstermektedir. Birlikte geçmekten kastın ne olduğu ise tartışmalı bir konudur. Buradan bir pencere içinde geçmekte anlaşılabilir, aynı metnin içinde geçmekte, aynı sınıftaki metinlerin içinde geçmekte. Bununla birlikte metinlerdeki tüm kelimelerin mi yoksa bir alt kümesinin mi kullanılacağı ayrı bir tartışma konusudur. Bu bölümün devamında bu sorulara cevap verebilmek için denediğimiz çeşitli versiyonlar anlatılmıştır.

2.1.1. Pencere Boyutu Ne Olmalı?

Birlikte geçişte pencere boyutunun etkisini görebilmek için 5 farklı yöntem uygulanmıştır. İlk 3'ü 2 kelimenin sırasıyla 2,3, ve 5 boyutlu bir pencere içinde kaç kez birlikte geçtiğini saymaktadır. 4. Yöntem ise 2 kelimenin kaç metin içinde birlikte geçtiğini saymaktadır. Diğer bir deyişle pencere boyutu tüm metin olarak kullanılmıştır. 5. yöntemde ise bir sınıfa ait tüm metinler tek bir metin gibi düşünülmüştür. Bunun sonucu olarak sınıf sayısı tane metin elde edilmiştir. 2 kelimenin bu metinlerden kaçında birlikte geçtiği sayılmıştır. Bu yöntemde 2 kelime en fazla sınıf sayısı kez birlikte geçebilecektir. Bu yöntemde pencere boyutunun çok büyütülmesi olarak algılanabilir.

2.1.2. Hangi Kelimeler Kullanılmalı?

Birlikte geçme matrisi oluşturulurken tüm kelimelerin bir altkümesinin kullanılmasının etkisini görebilmek için 4 yöntem uygulanmıştır. İlk yöntem tüm kelimelerin kullanımıdır. 2. yöntemde metinlerde toplam geçme frekansına göre bir filtreleme yapılmıştır. Bu sayede çok sık ya da çok az kullanılan kelimeler sisteme dahil edilmemektedir. Bu sayede farklı kişilere ait kelime yörüngelerinin aynı yerlerden (çok sık kullanılan kelimeler) ya da çok nadir noktalardan (çok az kullanılan kelimeler) geçmesi engellenebilmektedir. 3. yöntemde ise kelimeler türlerine göre filtrelenmektedir. Tüm kelimeler yerine sadece isim türündeki ya da fiil türündeki kelimeler sisteme dahil edilebilmektedir. Kelimelerin türlerinin bulunmasında Zemberek [6] aracı kullanılmıştır.

2.2. Çok Boyutlu Ölçekleme

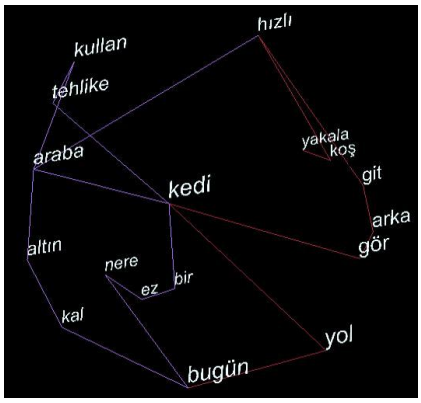
Kelimelerin birbirlerine yakınlıklarını içeren birlikte geçme matrisinin elde edilmesinin ardından kelimelerin koordinatları bulunmaktadır. Birbirine uzaklıkları bilinen nesnelerin, bu uzaklıklara uygun olarak bir koordinat sisteminde karşı geldikleri noktaları bulmak için Çok Boyutlu Ölçekleme yöntemi kullanılmaktadır [5]. Bu yöntemi kullanabilmek için elimizdeki yakınlık matrisini uzaklık matrisine çevirmek gerekmektedir. Bunun için Eşitlik 1'deki formül kullanılmıştır.

$$dis(i,j) = \frac{1}{sim(i,j)} \qquad (1)$$

Eşitlik 1'de, *dis(i,j);* i. ve j. kelimeler arasındaki uzaklığı, *sim(i,j);* i. ve j. kelimelerin birlikte geçtiği (2, 3 ve 5'lik kelime çerçevesi ya da doküman ya da sınıf) sayısını göstermektedir.

3. YÖRÜNGE ÖZELLİKLERİ

Kelimelerin koordinatları belirlendikten sonra bir metnin yörüngesi de bulunmuş olmaktadır. Şekil 1'de iki kısa metnin yörüngeleri verilmiştir. Birinci yazı "Bugün yolda kedi gördüm. Arkasından gittim. Hızlı koşuyordu. Yakalayamadım.", ikinci yazı ise "Hızlı araba kullanmak tehlikelidir. Kediler arabaların altında kalabilirler. Bugün neredeyse eziliyordu bir kedi." şeklindedir.



Şekil 1: Kısa 2 Metnin yörüngeleri.

Birinci yazı kırmızı renkte, ikici yazı mor renktedir. Birinci yazı "bugün" kelimesiyle başlayıp ve "hızlı" kelimesiyle bitmektedir. İkinci yazı ise "hızlı" kelimesiyle başlayıp "kedi" kelimesiyle bitmektedir. İki yazının içinde de geçen kelimeler üç boyutlu uzayda yazıların kesiştiği noktalar olarak göze çarpmaktadır. Bu kelimeler iki kutuba ayrılmış kelime havuzunun da ortasında bulunmaktadır. Çünkü iki taraftaki kelimelerle de ilişki içindedirler.

Kişilerin yazıları arasında yörüngeleri oluşturan kavramların değil de yörüngelerin özelliklerine göre bir ayrım yapmak istediğimizden yörüngeleri ifade etmek için temel 2 özellik kullanılmıştır. İlki kavramlar arası mesafeler, ikincisi ise kavramlar arası açılar. *(n, n+1, n+2, n+3)* arka arkaya gelen 4 koordinat olmak üzere *(n,n+1), (n,n+2), (n,n+3)* arası mesafelerin 10'luk histogram değerleri, frekansları ve *(n,n+1,n+2), (n,n+2,n+3)* arası açıların (PI cinsinden) 10'luk histogram değerleri, frekansları hesaplanmıştır.

d uzunluğundaki bir yörünge de *d-1* adet *(n, n+1)* arası mesafe ölçülmektedir. Bu *d-1* ölçümün eşit aralıklı 10 parçalık histogramı çıkarılmaktadır. Bu histogramın 10 adet değeri ve her değerin bir frekansı bulunmaktadır. Dolayısıyla yörüngenin *(n, n+1)* arası mesafelerini ifade eden 20 adet özellik çıkarılmaktadır. Bu işlem *(n,n+2), (n,n+3)* arası mesafelere ve *(n,n+1,n+2), (n,n+2,n+3)* arası açılara da uygulandığında bir metni ifade eden yörüngeye ait 100 özellik bulunmuş olmaktadır.

4. DENEYSEL SONUÇLAR

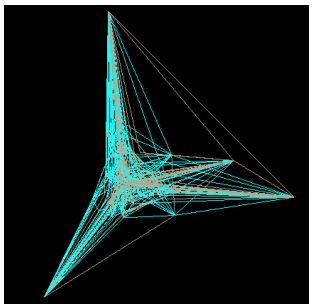
Metinlere ait yörüngeleri çıkarırken pencere boyutu ve kelime filtrelerinin hangi durumda nasıl kullanılacağı ile ilgili araştırmalarımız bu bölümde yer almaktadır. Ayrıca

çeşitli yazar tanıma veri kümeleri üzerindeki çalışmalarımız da bu bölümde sunulmuştur.

Pencere boyutunun ne olması gerektiği ile ilgili soruya cevap vermek için yapılan denemelerde az sayıda örnek içeren veri kümelerinde pencere boyutunun tüm metin ya da tüm sınıf yapıldığında birliktelik matrisinde kelimelerin birbirlerine uzaklıkları arasında pek fark olmadığı ve bu nedenle çok boyutlu ölçeklemenin kelime koordinatlarını belirlemede başarısız olduğu görülmüştür. Bu nedenle örnek sayısı az olan veri kümelerinde çalışırken pencere boyutunu 2, 3 ya da 5 seçmek gerektiği görülmüştür. Ancak örnek sayısının çok olduğu veri kümelerinde ise 2, 3, 5 boyutlu pencereler seçildiğinde birliktelik matrisinin çok seyrek olduğu ve kelime koordinatlarının yine doğru bir şekilde hesaplanamadığı görülmüştür. Bu nedenle eğer örnek sayısı çok ise pencere boyutu olarak tüm metin seçilmelidir. Pencere boyutu olarak tüm sınıfın seçilmesinin faydalı olduğu veri kümeleri ise çok sınıf sayısına sahip veri kümeleridir. Bölüm 2.1.1'de anlatıldığı gibi pencere boyutu tüm sınıf seçildiğinde birliktelik matrisindeki en büyük değer sınıf sayısı kadar olabilmektedir. 2 sınıfa sahip bir veri kümesi için tüm sınıf pencere boyutu olarak seçildiğinde tüm matris 0,1, ve 2'den oluşacak ve kelime koordinatları düzgün olarak belirlenemeyecektir. Bu denemelerden çıkan başlıca sonuçlar aşağıda verilmiştir:

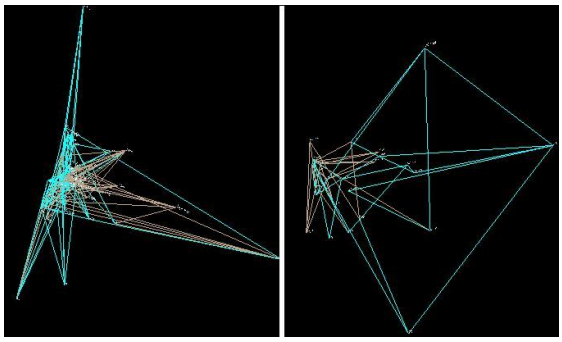
1- Az sayıda sınıf ve metin varsa pencere boyutu 2,3, ya da 5 seçilmeli.
2- Çok sayıda metin varsa, sınıf sayısı azsa pencere boyutu tüm metin seçilmeli
3- Çok sayıda sınıf varsa, metinlerin boyutları kısaysa pencere boyutu tüm sınıf seçilmelidir.

Kelime tür ve frekans filtrelerinin oluşan yörüngelere etkisi, 2 yazara ait birer yazı üzerinde incelenmiştir. Şekil 2'de pencere boyutu 2 seçilip, kelimeler üzerinde hiçbir filtre uygulandığında elde edilen yörüngeler verilmiştir.



Şekil 2: Filtrelenmemiş kelimelerde elde edilen yörüngeler.

Şekil 2 incelendiğinde yazarların kelime havuzlarının birbirine benzediği görülmektedir. Kelimelerden 3'ten az geçenlerin elenmesi sonucunda elde edilen yörüngeler ise Şekil 3 (a)'da, sadece fiil türündeki kelimelerin kullanımıyla elde edilen yörüngeler ise Şekil 3 (b)'de verilmiştir.

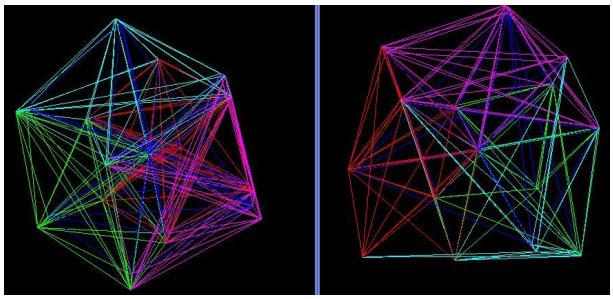


(a) *(b)*

Şekil 3: (a) Kelimelere frekans filtresi uygulandığında (b) sadece fiil türündeki kelimeler kullanıldığında elde edilen yörüngeler.

Şekil 3 incelendiğinde ise filtrelerin yazarları birbirinden ayırmaya yardımcı oldukları görülmektedir. Örneğin Şekil 3 (b)'de kahverengi renkli yazarın peş peşe kullandığı fiillerin anlamca birbirine yakın kelimeler olduğunu, yeşil renkli yazarın ise daha dağınık (kavramları arasındaki ortalama uzaklıkların daha fazla) bir stile sahip olduğu söylenebilir.

Sıklıkla kullanılan kelimelerin filtrelenmesinin oluşan yörüngelere etkisi, 5 yazara ait birer yazı üzerinde incelenmiştir. Şekil 4 (a)'da 5 yazının filtresiz yörüngeleri, Şekil 4 (b)'de ise Türkçe'de sık kullanılan 194 kelimenin dışarıda tutulduğunda elde edilen yörüngeleri verilmiştir.



(a) *(b)*

Şekil 4: (a) Kelimelere frekans filtresi uygulandığında (b) sadece fiil türündeki kelimeler kullanıldığında elde edilen yörüngeler.

Şekil 4 incelendiğinde sık geçen kelimeler ayıklanınca orta taraftaki (ortak kullanılan kelimelerde) yoğunlukta bir azalma görülmektedir.

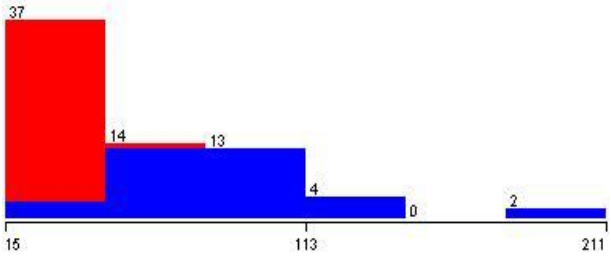
Kelime filtrelerinin ve pencere boyutlarının etkileri incelendikten sonra çalışmanın asıl amacı olan yörünge özelliklerinin yazarları birbirinden ayırmada kullanılıp kullanılamayacağı incelenmiştir. Bunun için önce 2 yazara ait 35'er yazıdan oluşan bir veri kümesi üzerinde çalışılmıştır. Bu 70 yazının yörüngeleri (tüm metin boyutlu pencere ve kelime filtresiz seçenekleriyle) bulunmuş ve her yörüngenin (yazının) Bölüm 3'te anlatılan 100'er özelliği çıkarılmıştır. Oluşturulan bu veri kümesi Weka [7] kütüphanesindeki (C4.5 karar ağacı, En yakın komşu algoritması, Naive Bayes, Destek Vektör Makineleri) metotlar kullanılarak 10'lu çapraz geçerleme ile

sınıflandırılmıştır. Tablo 1'de bu sınıflandırma sonuçları verilmiştir.

Tablo 1: İki yazarın yörüngelerinin ayrılabilirliği

Algoritma	Sınıflandırma Başarısı (%)
C4.5	97.15
Naive Bayes	97.15
En yakın komşu	95.71
Destek Vektör Makineleri	98.57

Tablo 1 incelendiğinde 2 yazarın birbirinden çok başarılı bir şekilde ayrılabildiği görülmüştür. Yörüngelerin hangi özelliklerine göre birbirlerinden ayrılmışlardır sorusuna cevap aramak içinse C4.5 karar ağacının yörüngeleri sınıflandırmak için ürettiği model incelenmiştir. Model tek bir kuraldan oluşmaktadır. Bu kurala göre metinler (n, n+1) arası mesafelerin histogramının ilk parçasının frekansı 43'ten büyükse bir yazara değilse diğer yazara aittir. Şekil 6'da 70 metnin bu özelliğine ait histogram verilmiştir.



Şekil 6: 2 yazara ait 70 yazının (n, n+1) arası mesafelerinin histogramının ilk parçasının frekanslarının histogramı.

Şekil 6 incelendiğinde kırmızılı yazarın ardışık *(n, n+1)* kavramları arası mesafelerinin daha küçük, mavili yazarın mesafelerinin daha büyük oldukları görülmektedir.

Sınıf sayısının artışının etkilerini incelemek için yazar sayıları arttırılarak çeşitli denemeler (tüm metin pencere boyutu ve kelime filtresiz seçenekleriyle) yapılmıştır. Denemelerde her yazara ait 35'er metin kullanılmıştır. Tablo 2'de 3, 4, ve 8 yazar için alınan sonuçlar görülmektedir.

Tablo 2: 3, 4, ve 8 yazarın yörüngelerinin ayrılabilirliği

Algoritma	3 yazar	4 yazar	8 yazar
C4.5	91.43	90.71	53.21
Naive Bayes	95.24	92.14	62.5
En yakın komşu	80.95	62.14	50
Destek Vektör Makineleri	94.29	85.71	68.57
Zero 0	33.33	25	12.5

Tablo 2'nin son satırında yer alan değerler sınıf değerleri rasgele atandığında elde edilecek başarı oranlarıdır. Tablo 2 incelendiğinde sınıf sayısı arttıkça başarının düştüğü ancak yine de başarılı ayrımlar yapılabildiği görülmektedir. Karar ağaçlarındaki kurallar incelendiğinde karar düğümlerinde en çok geçen yörünge özellikleri mesafe ve açıların frekanslarıdır.

5. SONUÇ VE GELECEK ÇALIŞMALAR

İnsanların düşünce süreçlerinin dolaylı olarak ölçümlenmesi üzerine yapılan bu çalışmada kişilerin yazdıkları metinler çok boyutlu bir uzayda yörüngeler olarak temsil edilmişlerdir. Daha sonra bu yörüngelerin çeşitli özellikleri çıkarılarak kişiler arası yörünge farklılıkları / benzerlikleri incelenmiştir. Bu yöntemle 2 kişiye ait 35'er yazıdan oluşan veri kümesinde bir metnin yazarını tanıma başarısı % 98 olarak ölçülmüştür. yazar sayısı arttıkça başarının düştüğü ancak yine de başarılı ayrımların yapılabildiği görülmüştür. Kişileri ayırmada en çok kullanılan yörünge özelliğinin mesfe ve açı değerleri değil, bunların frekansları olduğu görülmüştür.

Metodun avantajları olarak metinde kullanılan kavramlardan bağımsız olması (sadece yörüngenin özelliklerinin kullanılıyor olması), kişilerin yazdıkları metinlerin sistemin çalışması için yeterli olması (ek bir görüntüleme cihazı gerektirmemesi) söylenebilir.

Önerilen yöntemin olası uygulama alanları olarak psikolojik hastalıkların tespiti, psikolojik hastalıkların düşünce süreçleri üzerindeki etkilerinin araştırılması, cinsiyet, yaş, eğitim farklılıklarının düşünce süreçleri üzerindeki etkilerinin araştırılması verilebilir.

Kelimelerin koordinatlarının bulunmasında kullanılan dönüşüm formülünün ve çok boyutlu ölçeklemenin yerine farklı yöntemlerin kullanılması, yeni yörünge özelliklerinin çıkarılması gelecekte denenebilir. Bununla birlikte, metin yörüngeleri çok boyutlu zaman serileri olarak da görülebilirler. Çeşitli zaman serisi yöntemlerinin ve uygulama alanlarının bu veriler üzerinde uygulanması da gelecek bir çalışma konusu olarak düşünülmektedir.

6. KAYNAKÇA

[1] Amasyalı, M. F., Davletov, F., Torayew, A, Çiftçi, Ü, "text2arff: Türkçe Metinler İçin Özellik Çıkarım Yazılımı", *SİU*, 2010.

[2] Amasyalı, M. F., Beken, A., "Türkçe Kelimelerin Anlamsal Benzerliklerinin Ölçülmesi ve Metin Sınıflandırmada Kullanılması", *SİU*, 2009.

[3] Amasyalı, M. F., "Arama Motorları Kullanarak Bulunan Anlamsal Benzerlik Ölçütüne Dayalı Kelime Sınıflandırma", *SİU*, 2006.

[4] Haris, Z. S., "Mathematical structures of language", *Wiley*, pp.12, 1968.

[5] Multidimensional Scaling for Java, University of Konstanz, Department of Computer & Information Science, *Algorithmics Group*, http://www.inf.uni-konstanz.de/algo/software/mdsj/

[6] http://code.google.com/p/zemberek/

[7] Witten, I. H., Frank, E., "Data Mining: Practical machine learning tools and techniques", 2nd Edition, *Morgan Kaufmann*, San Francisco, 2005.