Arama Motorları Kullanarak Bulunan Anlamsal Benzerlik Ölçütüne Dayalı Kelime Sınıflandırma

An Approach for Word Categorization Based on Semantic Similarity Measure Obtained from Search Engines

M.Fatih Amasyalı

Bilgisayar Mühendisliği Bölümü
Yıldız Teknik Üniversitesi, İstanbul
mfatih@ce.yildiz.edu.tr

Özetçe

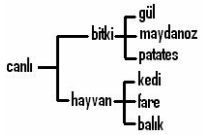
Kelimelerin anlamsal sınıflara ayrılması, birçok doğal dil işleme çalışmasında çözülmesi gereken bir problemdir. Kelimelerin sınıflandırılabilmeleri için aralarındaki benzerliğin bir şeklide ölçülmesi gerekmektedir. Bu çalışmada iki Türkçe kelimenin birbirlerine anlamsal benzerliğinin, kelimelerin Internet'te arka arkaya geçtiği sayfa sayısıyla doğru orantılı olduğu hipotezi ortaya atılmıştır. Sayfa sayısının bulunması içinde Google ve Yahoo adlı arama motorları kullanılmıştır. Hipotezi doğrulamak için yapılan bu ilk çalışmada az sayıda kelime ve sınıfla denemeler yapılmış ve kelimelerin anlamsal sınıfları ortalama %87 doğrulukla belirlenebilmiştir.

Abstract

Word categorization based on semantic similarity is a problem need to be solved for several natural language applications. A similarity measure is need for word categorization. In this study it is proposed that the semantic similarity between two Turkish words is in direct proportion to the number of pages which the words are located next to each other. Google and Yahoo search engines were used to find the number of pages. In the first attempt to verify the proposal, the experiments were done with small datasets. The average success ratio is 87%.

1. Giriş

Anlamsal olarak birbirine benzer kelimeler içeren sınıfların oluşturulması önemli doğal dil işleme konularından biridir. Bu gruplar metin sınıflandırma [1], soru cevaplama [2], makinelerin anlayabildiği sözlüklerin otomatik olarak oluşturulması, kelime anlamını durulaştırma, arama motorlarından daha iyi sonuçlar elde etme [3], otomatik metin özetleme [4] gibi geniş uygulama alanlarında kullanılmaktadır. Bu sınıfların elle oluşturulması oldukça zahmetli bir işlemdir. Otomatik sınıflandırma işlemi için iki kelimenin birbirine benzerliği ya da uzaklığı bir şekilde ölçülmek zorundadır. Araştırmacılar iki kelimenin anlamsal olarak birbirine yakınlığını ölçmek için birçok öneride bulunmuşlardır [5]. Genel olarak 2 tür yaklaşım mevcuttur. İlki bilgisayarları okuyabildikleri sözlüklerin, kavramsal haritaların kullanılmasına, ikincisi ise büyük metin kütüphanelerinden elde edilen istatistiklerin kullanılmasına dayanmaktadır[6]. İlk yaklaşıma örnek olarak kelimelerin birbirlerine altküme ilişkileriyle bağlandıkları haritalar (Ör: İngilizce için Wordnet [7]) verilebilir. Şekil 1'de hayali bir kelime haritası verilmiştir.



Şekil 1: Hayali bir kelime haritası.

Şekil 1'deki gibi bir haritada iki kelime arasındaki benzerliği bulmak için iki kelimeyi birleştiren en kısa yolun uzunluğuna bakılabilir [8]. Örneğin "balık" ile "kedi" arasındaki yol "balık-hayvan-kedi" şeklindeyken, "balık" ile "maydanoz" arasındaki yol "balık- hayvan- canlı- bitki- maydanoz" şeklindedir. Bu durunda kedinin balığa maydanozdan daha çok benzediğini söyleyebiliriz. Peki, böylesi bir ağacın hiç olmadığı ya da yeterli olmadığı dillerde iki kelime arasındaki anlamsal benzerlik nasıl ölçebilir? Bu gibi durumlar için ikinci bir yaklaşım olarak büyük metinlerden elde edilen istatistiklerin kullanılması önerilmektedir. Bu çalışmada da bu yaklaşım benimsenmiş ve Türkçe kelimeler için basit ve etkili bir benzerlik metodu önerilmiş ve bu metoda göre hesaplanan yakınlıklar kullanılarak kelimelere karşılık gelen vektörler bulunmuştur.

Metnin ikinci bölümünde anlamsal kelime sınıflarının seçildiği Türkçe için hazırlanmış olan Wordnet tanıtılmıştır. Üçüncü bölümde ise kelimelerin sınıflandırılması için önerilen metodun ayrıntıları anlatılmıştır. Sonraki bölümde elde edilen sonuçlar verilmiştir. Son bölümde ise yaklaşımın avantaj ve eksiklikleri ve gelecekte yapılması planlanan çalışmalar verilmiştir.

2. Türkçe Wordnet

Türkçe Wordnet, BalkaNet kapsamında hazırlanan Türkçe kelimeler arasındaki çeşitli ilişkileri içeren bir kavramlar haritasıdır[9]. Orijinal Wordnet'te olduğu gibi aynı anlamı ifade eden kelimelerin oluşturduğu kümelerden (eşküme) ve bu kümeler arasındaki ilişkilerden meydana gelmektedir. Türkçe Wordnet'te Mart 2004 itibariyle 11.628 eşküme ve 17.550 ilişki vardır. Sistemin web ara yüzüne http://www.hlst.sabanciuniv.edu/TL/ adresinden ulaşılabilir. Wordnet'te ilişkiler bilgi tekrarını engellemek için hiyerarşik bir yapı (ağaç) şeklinde ifade edilmiştir. Bu nedenle Wordnet'te aynı üst kavram ilişkisine sahip kelimelerin anlamsal bir sınıf ifade ettiği söylenebilir. Şekil 1'de aynı üst

kavrama (hayvan) sahip kedi, fare ve balık anlamsal bir sınıf oluşturmaktadırlar.

3. Kelime Gruplama Metodu

Hipotezi doğrulamak için Türkçe Wordnet'ten seçilen anlamsal sınıflar içinden ki kelimeler kullanılmıştır.

3.1. Kullanılan Veriler

Türkçe Wordnet'te ortak üst kavram ilişkisine sahip isim türündeki kelimeler seçilmiş ve 6 adet anlamsal sınıf oluşturulmuştur. Sınıflar 3 farklı şekilde bir araya konarak 3 farklı veri kümesi elde edilmiştir. Veri kümelerinde sırasıyla 3, 3 ve 6 sınıftan kelimeler yer almaktadır. Tablo 1'de veri kümelerinin içerdiği gruplar ve gruplardaki kelimeler verilmiştir.

Tablo 1: Anlamsal sınıflar ve veri kümeleri

Veri kümesi-I	*Yer* berber – bakkal- kuaför-hastane- otel- okul-dershane- büfe- ahır-park		*Hayvan* akbaba- arı- baykuş-balina- aslan- at – bizon – akrep - ayı- bit		*Taşıt* otobüs- cip- taksi-ambulans- araba-sedan- tren- limuzin-otomobil- traktör	
Veri kümesi -II	*Ev eşyası* divan - televizyon- avize-merdiven- bilgisayar-buzdolabı- kapı-sandalye- lamba- koltuk		*Hayvan* akbaba- arı- baykuş-balina- aslan- at – bizon – akrep - ayı- bit		*Giyecek* bikini- mayo- bere-süveter- etek-ayakkabı- eldiven-gömlek- şapka- gözlük	
Veri kümesi -III	*Yiyecek* tereyağı kaymak baharat peynir şeker tuz	*Giyecek* ayakkabı eldiven şapka gömlek gözlük etek	*Ev eşyası* avize sandalye lamba koltuk buzdolabı televizyon	*Hayvan* aslan arı at akrep ayı akbaba	*Yer* otel dershane park berber hastane okul	*Taşıt* taksi limuzin ambulans cip otomobil araba

3.2. Anlamsal Benzerlik Ölçümü

İki kelimenin anlamsal olarak birbirine yakınlığı, bu iki kelimenin Internet'te yer alan sayfaların kaçında arka arkaya kullanıldıkları bulunarak belirlenmiştir. Bunun için arama motoruna "kelime1 kelime2" sorgusu gönderilerek gelen sonuç sayfasındaki sonuç sayısı alınmıştır.

Tablo 2: Örnek benzerlik matrisi

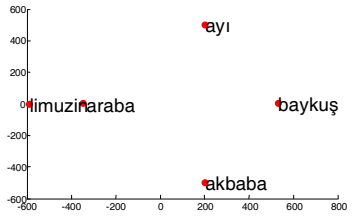
	akbaba	ayı	baykuş	araba	limuzin
akbaba		0	20	1	0
ayı	0		33	4	0
baykuş	20	33		0	0
araba	1	4	0		38
limuzin	0	0	0	38	

Tablo 2'de örnek bir benzerlik matrisi verilmiştir. A kelimesinin B kelimesine benzerliğiyle B kelimesinin a kelimesine benzerliği aynı olduğundan matris simetriktir ve N adet kelime için arama motoruna N*(N–1)/2 adet sorgu gönderilmesi yeterlidir. Matrisin i. satırını j. sütununda yer alan değer i.kelime ve j. kelimenin arka arkaya kaç sayfada geçtiğini göstermektedir.

Tablo 2'de "akbaba" ve "ayı" kelimelerinin arka arkaya geçtiği sayfa bulunmamasına rağmen, "akbaba" ve "baykuş" kelimelerinin geçtiği 20 sayfa vardır. Ayrıca "ayı" ve "baykuş" kelimelerinin geçtiği 33 sayfa vardır. Bu sayede "akbaba" ve "ayı" kelimeleri beraber kullanılmıyor olsalar bile ortak kullanıldıkları kelime olan "baykuş" sayesinde birbirlerine yakın oldukları söylenebilmektedir. Bu örnekten yola çıkılarak her bir sınıf içindeki kelime sayısının fazlalığının böyle ortak kullanılan kelime sayısını arttıracağı ve bu sayede daha başarılı sınıflandırmalar yapılabileceği düşünülmektedir.

3.3. Çok Boyutlu Ölçekleme

Aralarındaki mesafelerin/yakınlıkların bilindiği ancak uzaysal koordinatlarının bilinmediği durumlarda örneklerin mümkün olduğunca az boyutlu bir uzayda orijinal şekle (eldeki mesafelere) yakın bir biçimde ifade edilmesi için Çok Boyutlu Ölçekleme (ÇBÖ - Multi-dimensional Scaling) metodu kullanılmaktadır. [10] Bu çalışmada örnekler kelimelerimizdir. Kelimelerin arasındaki mesafeler/benzerlikler ise arama motorlarıyla oluşturulan benzerlik matrisidir. ÇBÖ sonucunda kelimelerin X boyutlu bir uzaydaki koordinatları bulunmaktadır. Şekil 1'de Tablo 2'deki uzaklık mesafelerinin ÇBÖ sonucunda oluşan 3 boyutlu koordinatlarının 2 boyuta izdüşümü verilmiştir.



Şekil 1: Kelimelerin 2 boyutlu uzayda gösterimi.

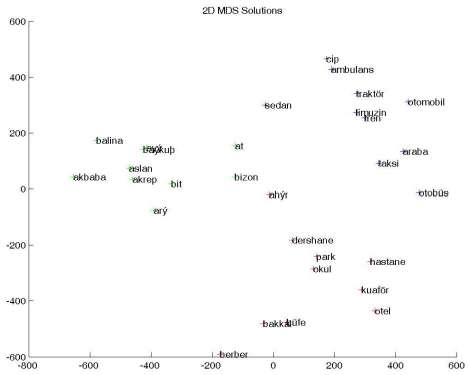
Koordinatların bulunmasında kelimeler hakkında hiçbir sınıf/grup bilgisi kullanılmamasına rağmen anlamsal olarak yakın kelimelerin koordinatlarının da yakın olarak bulunabildiği Şekil 2'de görülmektedir.

4. Deneysel Sonuçlar

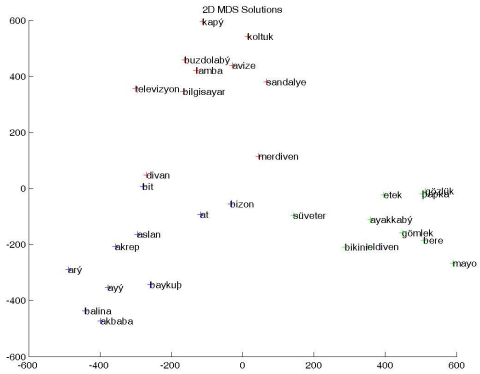
Kelimelerin anlamsal sınıflarının bulunması için Bölüm 3.1'de anlatılan 3 farklı veri kümesi kullanılmıştır. Her bir veri kümesi için önce benzerlik matrisleri bulunmuş daha sonra ÇBÖ ile koordinatlar belirlenmiştir. Son olarak da kelimelerin koordinatları kullanılarak çeşitli sınıflandırma ve gruplandırma algoritmalarıyla kelimelerin Türkçe Wordnet'tekiyle aynı anlamsal sınıflara ayrılabilirliği araştırılmıştır.

4.1. Kelimelerin Koordinatlarının Bulunması

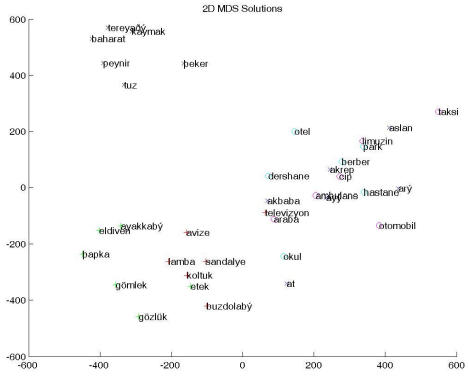
Bölüm 3.1'deki her bir veri kümesi için iki farklı arama motorunun sonuçları kullanılarak benzerlik matrisleri oluşturulmuştur. Benzerlik matrislerindeki değerlere en uygun olduğu düşünülen kelime koordinatları ÇBÖ metoduyla hesaplanmıştır. Şekil 2-5'de 3 farklı veri kümesindeki kelimelerin ÇBÖ ile hesaplanmış çok boyutlu uzaydaki koordinatlarının 2 boyuta izdüşümleri verilmiştir.



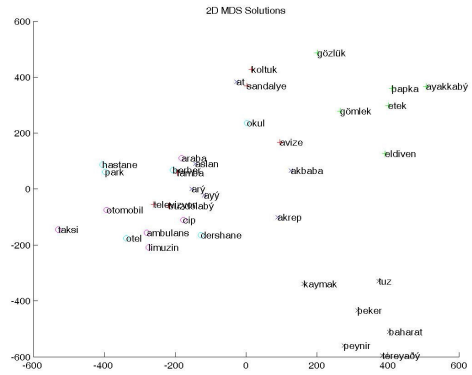
Şekil 2: Veri kümesi–1 için Google ile bulunan benzerlik matrisinden hesaplanmış kelime koordinatları.



Şekil 3: Veri kümesi–2 için Google ile bulunan benzerlik matrisinden hesaplanmış kelime koordinatları.



Şekil 4: Veri kümesi–3 için Yahoo ile bulunan benzerlik matrisinden hesaplanmış kelime koordinatları.



Şekil 5: Veri kümesi–3 için Google ile bulunan benzerlik matrisinden hesaplanmış kelime koordinatları.

Şekiller incelendiğinde kelimeler hakkında herhangi bir sınıf/grup bilgisi kullanılmamasına rağmen Türkçe Wordnet'te aynı anlamsal sınıfta bulunan kelimelerin bulunan koordinatlarının da birbirlerine yakın oldukları görülmektedir. Özellikle 3'er sınıftan 10'ar kelimeden oluşan Veri kümesi–1 ve Veri kümesi–2'deki kelimeler koordinatlarına göre grup içi varyantın düşük, gruplar arası varyantın büyük olduğu için çok iyi gruplandırıldıkları söylenebilir. 6 sınıfa ait 6'şar kelimeden oluşan Veri kümesi–3'te ise yiyecek, giyecek ve ev eşyası isimleri diğerlerinden ayrılabilmiş ancak yer, taşıt ve hayvan isimleri birbirlerinden ayrılamamışlardır. Buna Bölüm3.2'nin son paragrafında anlatılan etkini (sınıf içindeki kelime azlığının) sebep olduğu düşünülmektedir.

Arama motorlarının performansları arasında belirgin bir fark görülmemiştir. Sadece Veri kümesi–3 için ev eşyalarını Yahoo'nun Google'dan daha iyi gruplayabildiği söylenebilir. Her ikisi de oldukça iyi sonuçlar üretmeye katkı sağlamışlardır. Daha büyük ölçekteki verilerle işlem yapıldığında belirgin bir performans farkı belki ortaya çıkabilir.

4.2. Kelimelerin Sınıflandırılması

Veri kümeleri için elde edilen koordinatlar kullanılarak kelimeler gruplandırılmış ve sınıflandırılmışlardır. Tüm uygulamalar WEKA [11] aracıyla gerçekleştirilmiştir. Sınıflandırma için oldukça yaygın kullanılan Karar Destek Makineleri (SVM), Karar Ağaçları (C4.5) ve Karar Ormanları (Random Forest) kullanılmıştır. Kümeleme/ gruplama için ise Beklenti Enbüyütme (Expectation Minimization-EM) kullanılmıştır. Tablo 2'de 3 veri kümesinin Google'dan elde edilmiş orijinal boyutlu verilerinde ve Korelasyon Tabanlı Özellik seçimi (CFS) ile seçilmiş boyutlarında yapılan uygulamaların sonuçları görülmektedir. Sınıflandırma uygulamalarında değerlerin rastlantısallığını azaltmak için 5'li çapraz geçerleme (5 fold cross validation) yapılmıştır. Kullanılan algoritmaların ayrıntılarına WEKA aracının dokümantasyon sayfalarından ulaşılabilir.

Tablo 2: Kelimelerin Sınıflandırma ve Kümeleme Başarıları

	v1 10 boyut	v1 2 boyut	v2 10 boyut	v2 2 boyut	v3 10 boyut	v3 4 boyut
SVM	96,6	100	83,3	96,6	88,8	77,7
C4.5	83,3	83,3	90	90	77,7	75
RF	90	90	93,3	93,3	72,2	75
EM	66,6	96,6	66,6	96,6	66,6	83,3

Tablo2'de v1, Veri kümesi1'i; v2 Veri kümesi-2'yi; v3, Veri Kümesi-3'ü göstermektedir. Tablo 2'de elde edilen sonuçlara bakıldığında en başarılı sınıflandırıcının SVM olduğu görülmektedir. Sınıf ayırt ediciliğini belirleyen özelliklerin seçilerek boyut azaltmanın genelde başarıyı yükselttiği söylenebilir. SVM algoritması Veri kümesi-1'i 2 boyutta mükemmel bir şekilde sınıflandırabilmiştir. 3 veri kümesinde sınıflandırıcıların başarılarının ortalaması %87'dir. Ayrıca EM algoritmasıyla hiçbir sınıf bilgisi kullanmadan düşük boyutlarda ulaşılan başarı gerçekten yüksektir. Bununla birlikte yapılan denemeler küçük ölçeklidirler. Daha sağlıklı yorumlara ulaşabilmek için daha fazla sınıf ve kelime içeren veri kümeleriyle çalışmak gerekmektedir.

5. Sonuç ve Gelecek Çalışmalar

Türkçe kelimelerin anlamsal sınıflara ayrılması konusunda yapılan ilk çalışma sunulmuştur. Kelimelerin sınıflandırılabilmesi için bir şeklide kelimelerin benzerliklerinin ölçülmesi gerekir. Türkçe iki kelime arasındaki benzerliğin Internet'te arka arkaya geçtikleri sayfa sayılarıyla ölçülebileceği öne sürülmüş ve bu konuda çeşitli denemeler yapılmıştır. Küçük ölçekte yapılan bu çalışmalar bundan sonraki araştırmalar için bir prototip niteliği taşımaktadır. Aşağıda önerilen metodun avantajları, kısıtları ve gelecek çalışmalar için planlananlar özetlenmiştir.

Avantajlar:

◦ Önerilen metot sadece Türkçe'ye özgü bir metot değildir. Her dil için kullanılabilir.

◦ Denemeler küçük ölçekte yapılmış olsalar da görsel temsilleriyle ve başarılı sınıflandırma sonuçlarıyla (%87) daha geniş çaptaki çalışmalar için umut vermektedir.

◦ Çalışmanın bir ara ürünü olarak kelimeler N boyutlu bir uzaydaki temsili koordinatları bulunmuştur. Bu koordinatlar sayesinde kelimelerin klasik makine öğrenmesi metotlarıyla sınıflandırma, gruplandırma yapılması mümkündür. Örneğin metin sınıflandırmada metinlerin içindeki kelimelerin koordinatları kullanılabilir.

Kısıtlar ve öneriler:

◦ N adet kelimenin sınıflandırılabilmesi için arama motoruna N*(N–1)/2 adet sorgu gönderilmesi gerekmektedir. Arama motorları bir kullanıcıdan bir günde yapılabilecek sorgu sayısını yaklaşık 1000'le sınırlamaktadırlar. Bu sebeple N'in örneğin 1000 değeri için benzerlik matrisinin oluşturulabilmek için 500 gün beklemek gerekecektir. Bu sorunu çözebilmek için arama motorlarını kullanmak yerine büyük metin kütüphaneleri kullanılabilir.

◦ En geniş veritabanına sahip arama motorları olan Google ve Yahoo eklemeli dillere yönelik bir arama sistemi kurmamışlardır. Örneğin bir sayfada "arabaları" kelimesi geçip, "araba" kelimesi geçmiyorsa "araba" sorgusunu kullanarak bu sayfaya erişmek mümkün değildir. Bu sebeple Türkçe gibi eklemeli bir dil için büyük metin kütüphanelerinin gövde ve eklerine ayrılmış hallerinin kullanılması başarıyı arttıracaktır. Türkçe kelimelerin morfolojik analizi için bir açık kod projesi olan Zemberek [12] kullanılabilir.

◦ Kelimelerin benzerliğinin ölçümünde sadece arka arkaya geçtikleri sayfa sayıları kullanılmıştır. Bunun yanında Peter D. Turney tarafından eşanlamlı kelimelerin arama motorları kullanılarak bulunması üzerine yaptığı çalışmada [13] olduğu gibi Altavista arama motorunun özel arama kelimeleri de (near-yakınında, not-içermeyen) kullanılabilir.

◦ Kelimelerin benzerlikleri ölçülürken 1.denklemde olduğu gibi sayfa sayıları üzerinde herhangi bir ön işlem yapılmamıştır.

$$benzerlik\ (a,b) = sayfasayis\ i("a\ \&\ b") \quad (1)$$

Bunu yerine Turney'in çalışmasında [13] olduğu gibi 2.denklem kullanılabilir.

$$benzerlik\,(a,b) = \frac{sayfasayis\,i("a\ \&\ b")}{sayfasayis\,i(a)\cdot sayfasayis\,i(b)} \quad (2)$$

1. ve 2. denklemlerde ("a&b"), a ve b kelimelerinin arka arkaya geçtiğini göstermektedir.

6. Kaynakça

[1] Bekkerman R. , El-Yaniv R. , Tishby N. , Winter Y. , "Distributional Word Clusters vs. Words for Text Categorization" ,Journal of Machine Learning Research, 1-48, 2002.

[2] Mann, G., "Fine-grained proper noun ontologies for question answering", SemaNet'02: Building and Using Semantic Networks, 2002.

[3] www.apperceptual.com/ml_text_synonyms_apps.html.

[4] Futrelle R. P., Susan Gauch S.,"The role of automated word classification in the summarization of the contents of sets of documents", CIKM'93, 1993.

[5] Budanitsky A., Hirst G., "Semantic Distance in Wordnet:An Experimental, Application Oriented Evaluation of Five Measures", Proc. Workshop Wordnet and Other Lexical Resources, 2001.

[6] Li Y., Bandar Z. A., McLean D., "An Approach for Measuring Similarity between Words Using Multiple Information Sources", IEEE Trans. Knowledge and Data Eng., (15:4), 2003.

[7] http://wordnet.princeton.edu

[8] Rada R., Milli H., Bichnell E., Blettner M., "Development and Application of a Metric on Semantic Nets", IEEE Trans. Systems, Man, and Cybernetics, (9:1), 1989.

[9] Bilgin O., Çetinoğlu Ö., Oflazer K., "Building a Wordnet for Turkish", Romanian Journal of Information Science and Technology, (7), 2004

[10] Tatlıdil, H., "Uygulamalı Çok Değişkenli İstatistiksel Analiz", Cem Web Ofset, Ankara, 1996. s.353

[11] www.cs.waikato.ac.nz/ml/weka

[12] https://zemberek.dev.java.net/

[13] Turney, P.D., "Mining the Web for synonyms: PMI-IR versus LSA on TOEFL", In Proceedings of the Twelfth Europan Conference on Macine Learning, Springer-Verlag, Berlin, p 491-502, 2001.