ARAMA MOTORLARINDA YENİ KONU TANILAMADA KARAKTER N-GRAM VE YAPAY SİNİR AĞLARI UYGULAMASI

H.Cenk ÖZMUTLU[*]
Burcu ÇAĞLAR[*]

Özet: Günümüzde, arama motorlarının kullanımının artmasıyla beraber kullanıcı davranışlarının tahmini önem kazanmıştır. Bugüne kadar anlam bazlı olmayan pek çok yöntem yeni konu tanılamada kullanılmıştır. Bazı çalışmalardan iyi sonuçlar elde edilmesine rağmen, genelde çalışmaların yazım farklılığı içeren sorgularda hatalı tahminler yaptığı gözlenmiştir. Bu çalışmada, anlam bazlı olmayan, karakter n-gram yöntemi, yeni konu tanılamada kullanılmıştır. Bununla beraber karakter n-gram yöntemiyle önceki anlam bazlı olmayan çalışmaları iyileştirmek hedeflenmiştir. Önceki çalışmalar incelendiğinde yapay sinir ağları yönteminin diğerlerinden daha iyi sonuçlar verdiği gözlenmiştir. Bu yüzden, çalışmada yapay sinir ağları yönteminin tahminleri kullanılmış ve yazım yanlışlarından kaynaklanan hatalı tahminlerin giderilmesi için karakter n-gram yöntemi kullanılmıştır.

Anahtar Kelimeler: İnternet Madenciliği, Arama Motoru Kullanıcı Davranışları, Karakter N-gram Metodolojisi.

Character N-gram and Neural Network Application for New Topic Identification in Search Engines

Abstract: Nowadays, the estimate of web users' behaviors has been important due to the web search engine usage increase. To date, many content-ignorant studies have been performed for automatic new topic identification. Although, some studies performed well, it was observed that they often made mistakes when queries had spelling differences. In this study the character n-gram methodology, which is content ignorant, was used for new topic identification. In addition, it was aimed to improve previous content-ignorant studies. Consideration of previous studies it was observed that the neural network applications gave better results than the other studies. Thus, the neural network method's estimations were used in this study and character n-gram methodology was used in order to eliminate wrong estimations, because of spelling errors.

Key Words: Web Mining, Search Engine Users Behaviors, Character N-gram Methodology.

1. GİRİŞ

Günümüzde, arama motorları kullanım oranındaki hızlı artışla beraber, internet kullanıcılarının davranışlarının tahmini önemli hale gelmiştir. Arama motorları, kullanıcı davranışlarını tahmin ederek istenen bilgiyi en kısa sürede, doğrulukla kullanıcılara sunabilirler. Eğer arama motoru, kullanıcının önceki sorgularını baz alarak aynı konuya devam ettiğini tahmin edebilirse, önceki sorgularda bulduğu sonuçları kullanarak yeni sorguya cevap verme hızını arttırabilir ve daha tutarlı cevaplar verebilir. Bunun yanı sıra, kullanıcının yeni konu aradığı tespit edilirse arama motoru öncekinden farklı, yeni bir kümeden aldığı cevapları ekrana getirebilir. Böylece kullanıcı temelli arama sonuçları elde edilmiş olur. İlerleyen aşamalarda kullanıcıların sorguları baz alınarak ilgi alanları belirlenebilir ve arama motorları kişiselleştirilebilir.

Kullanıcı oturumlarında konu değişikliği tahmini için şimdiye kadar pek çok farklı çalışma gerçekleştirilmiş olup bu çalışmalar, anlam bazlı olanlar ve anlam bazlı olmayanlar olarak iki ana grupta toplanabilir. Anlam bazlı çalışmalar, genel itibariyle sözlüklerle çalışırlar ve bu sözlüklerin başlangıç aşamasında oluşturulması, uygulamalarda kullanılmak üzere depolanması gibi gerekliliklerden dolayı, yüksek maliyet ve emek gerektirirler. Bu sebeple kullanıcı oturumlarında konu değişimi

[*] Uludağ Üniversitesi, Mühendislik-Mimarlık Fakültesi, Endüstri Mühendisliği Bölümü, 16059, Görükle, Bursa.

tespiti için kullanılan anlam bazlı çalışmaların sayısı kısıtlı olup, bu çalışmalardan başarılı sonuçlar elde edilse de, dezavantajlarından dolayı araştırmacılar, gerçek zamanlı uygulamalarda anlam bazlı olmayan istatistiksel metotları tercih etmektedirler. Anlam bazlı olmayan çalışmalar, anlam bazlı olanlara göre daha düşük maliyetli ve daha basit olmalarının yanı sıra, eldeki verileri istatistiksel olarak yorumlayarak gerçekçi sonuçların elde edilmesine yardımcı olmaktadırlar.

Arama motoru sorgularında, konu değişimini tespit edebilmek için pek çok anlam bazlı olmayan istatistiksel yöntem geliştirilmiştir. Anlamı ihmal eden yeni konu tanılama çalışmalarına örnek olarak; Dempster–Shafer yöntemi (He ve diğ. 2002, Özmutlu ve Çavdur, 2005a), genetik algoritmalar (Özmutlu ve diğ. 2006), şartlı olasılık (Özmutlu ve diğ. 2007), yapay sinir ağları (Özmutlu ve diğ. 2004a, Özmutlu ve Çavdur, 2005b, Özmutlu ve diğ. 2008b) ve Monte Carlo simulasyonu (Özmutlu ve diğ. 2008a) gibi yöntemler verilebilir. Geliştirilen yöntemler, Excite (http://www.excite.com) ve FAST (http://www.fast.com) arama motorlarından alınan, arama motorlarını kullanan kullanıcılar hakkında bilgilere sahip veri grupları üzerinde uygulanmıştır. Bu veriler İnternet Protokolü (Internet Protocol – IP) adresi, arama zamanı ve sorgudan oluşmaktadır. Veriler üzerinde bazı işlemler uygulanarak, bu çalışmalarda kullanılabilecek şekle getirilmiştir. Örneğin, aynı IP adresine ait ardışık arama kayıtlarının zamanları arasındaki farklar kullanılarak sorgular 7 farklı zaman aralığına (time interval – *ti*) bölünmüştür. Benzer şekilde, aynı IP adresine ait ardışık aramalarda girilmiş olan sorgular incelenerek, ardışık sorgular arasındaki yapısal ilişkilere bakılarak, arama yapısı sınıflara (search pattern – *sp*) ayrılmıştır (Özmutlu ve Çavdur, 2005a). Eldeki bulgular kullanılarak veriler üzerinde uygulanan yöntemlerin, konu değişikliği ve konu devamı tahminlerinde bulunması sağlanmıştır. Tahminler, uzman tarafından değerlendirilmiş gerçek sonuçlar ile karşılaştırılarak, performans değerlendirmeleri yapılmıştır.

Bugüne kadar yapılan çalışmaların sonuçları incelendiğinde iki önemli bulguyla karşılaşılmıştır. İlk olarak, bu yöntemler anlam bazlı olmadığından eş anlamlı sorguların tahmininde problemlerle karşılaşılmıştır. Örneğin iki ardışık sorgu "hotel" ve "inn" kelimelerinden oluşuyorsa uygulanan yöntemler kelimelerin anlamlarına bakmadığından doğal bir süreç olarak bu iki sorgu sonucu yöntemin cevabı konu değişimi olacaktır. Bu sorunu aşmanın anlama bakılmadığından mümkün olmadığı tespit edilmiştir. İkinci bulgu ise eğer kullanıcılar sorgularında yazım hatası yaparlarsa, yöntemler yanlış yazılmış sorgu ile doğru yazılmış sorguyu farklı kelimeler olarak algılamakta ve yine karar konu değişimi olmaktadır. Örnek olarak, yöntemler FAST arama motorundan alınan ardışık iki sorgu "cybersc@n" ve "cyberscan" ile karşılaştıklarında hatalı karar vermekte ve bu sorguları konu değişikliği olarak algılamaktadırlar. Yapılan incelemeler sonucunda yanlış yazımdan kaynaklanan hatalı tahminlerin karakter n-gram yöntemiyle düzeltilebileceği görülmüştür.

Karakter n-gram yönteminde, aynı oturum içerisindeki ardışık iki sorgunun bütün kelimeleri, varsayılan *n* sayısına göre karakter gruplarına ayrıştırılır. Ardından bu karakter grupları diğer sorgunun bütün kelimeleri için oluşturulan karakter gruplarıyla karşılaştırılır ve aynı olan karakter grubu sayısı, kabul edilen eşik değerini aştığında konu devamı kararı verilir. Sorgular bu şekilde düşünülerek, anlam bazlı olmayan karakter n-gram yöntemiyle yeni konu tanılama gerçekleştirilebilir.

Çalışma iki farklı yaklaşımla gerçekleştirilmiştir. İlk adımda Excite ve FAST arama motorlarından alınan veriler üzerinde konu değişimi tahmini için karakter n-gram uygulaması yapılmıştır. Tahmin sonuçları uzman sonuçları ile karşılaştırılmıştır. Performans değerlendirmesinde baz alınan değişkenler, önceki yöntemlerin sonuçlarıyla karşılaştırıldığında, karakter n-gram ile konu değişikliği tahminlerinin yeterince iyi olmadığı gözlenmiştir. Bu sebeple ikinci yaklaşım olarak, aynı veriler üzerinde şimdiye kadar yapılan çalışmalardan performansı en iyi olan yöntemin sonuçlarına, karakter n-gram yönteminin uygulanması yoluna gidilmiştir. Anlam bazlı olmayan yeni konu tanılama çalışmaları, aynı verileri kullandıkları ve aynı parametrelerle değerlendirildikleri için performansları, Excite verileri için Tablo I ve FAST verileri için Tablo II'deki gibi karşılaştırılabilmiştir. Yapılan performans değerlendirmelerinde yapay sinir ağları yönteminin başarılı tahminlerde bulunduğu görülmüştür. Bu sebeple, yapay sinir ağları yönteminin önceki çalışmalarla elde edilen konu değişimi tahminleri, ikinci yaklaşımda baz alınmış ve yapay sinir ağlarının yazım farklılıklarından dolayı hatalı olarak konu değişimi tahmini yaptığı veriler üzerinde, karakter n-gram metodu uygulanmış, böylece konu değişimi tahminleri güncellenmiştir. Performans değerlendirmesi için karakter n-gram yöntemi ile düzeltilmiş tahminler, uzman tarafından belirlenen sonuçlarla karşılaştırılmış ve başarılı bir iyileştirmenin gerçekleştirildiği gözlenmiştir.

Tablo I. Excite verilerine uygulanan yöntemlerin analiz sonuçları

	Analiz edilen Sorgu sayısı	Konu Değişim sayısı	Konu Devamı sayısı	Doğru Tahmin Edilen değişimler	Doğru Tahmin Edilen devamlar	A Tipi Hata	B Tipi Hata	$P_{değişim}$	$R_{değişim}$	P_{devam}	R_{devam}	$F_{ß(değişim)}$	$F_{ß(devam)}$
Uzman Sonuçları	3394	$N_{ger-çekdeği-şim}$ =	$N_{ger-çekde-vam}$ =	-----	-----	----	----	----	----	----	----	----	----
Monte Carlo Simulasyonu	3394	$N_{değişim}$ = 393	N_{devam} = 3001	$N_{değişim\&doğru}$ = 142	$N_{devam\&doğru}$ = 2871	251	130	0.36	0.53	0.96	0.92	0.45	0.94
YSA Sonuçları (2008)	3394	$N_{değişim}$ = 454	N_{devam} = 2940	$N_{değişim\&doğru}$ = 237	$N_{devam\&doğru}$ = 2905	217	35	0.522	0.871	0.988	0.93	0.698	0.95

Tablo II. FAST verilerine uygulanan yöntemlerin analiz sonuçları

	Analiz edilen Sorgu sayısı	Konu Değişim sayısı	Konu Devamı sayısı	Doğru Tahmin Edilen değişimler	Doğru Tahmin Edilen devamlar	A Tipi Hata	B Tipi Hata	$P_{değişim}$	$R_{değişim}$	P_{devam}	R_{devam}	$F_{ß(değişim)}$	$F_{ß(devam)}$
Uzman Sonuçları	4484	$N_{gerçek-değişim}$ =	$N_{gerçek-devam}$ =	-----	-----	----	----	----	----	----	----	----	----
YSA Sonuçları (2005)	4484	$N_{değişim}$ = 865	N_{devam} = 3619	$N_{değişim\&doğru}$ = 305	$N_{devam\&doğru}$ = 3614	560	5	0.353	0.984	0.999	0.866	0,635	0,903
Monte Carlo Simülasyonu	4484	$N_{değişim}$ = 338	N_{devam} = 4146	$N_{değişim\&doğru}$ = 137	$N_{devam\&doğru}$ = 3973	201	173	0.41	0.44	0.96	0.95	0.43	0.95
Şartlı Olasılıklar	4484	$N_{değişim}$ = 276	N_{devam} = 4208	$N_{değişim\&doğru}$ = 146	$N_{devam\&doğru}$ = 4044	130	164	0.529	0.471	0.961	0.968	0.491	0.966
Demster Shafer Teorisi	4484	$N_{değişim}$ = 836	N_{devam} = 3648	$N_{değişim\&doğru}$ = 303	$N_{devam\&doğru}$ = 3641	533	7	0.362	0.977	0.998	0.872	0.642	0.907
YSA Sonuçları (2008)	4484	$N_{değişim}$ = 886	N_{devam} = 3598	$N_{değişim\&doğru}$ = 306	$N_{devam\&doğru}$ = 3594	580	4	0.345	0.987	0.998	0.861	0.583	0.907

2. MATERYAL ve YÖNTEM

2.1. Karakter N-gram Yöntemi

İstatistiksel dil modellemede amaç, sıradaki kelimeyi, daha önce karşılaşılan kelimeler aracılığıyla tahmin etmektir. İlk çalışmalardan biri Shannon'a (1951) aittir ve "Shannon Game" ile bir metindeki sonraki harfi tahmin etmeye çalışmıştır. Bu çalışmayı takip eden birçok farklı çalışma literatürde yer alsa da N-gram modeli, dil modellemede en basit ve en başarılı temeli oluşturmuştur (Huang ve diğ. 2003). N-gram dil modelleri sıradaki kelimenin görülme olasılığının ondan önceki n-1 kelimeye dayandığını varsayar. Karakter n-gram yöntemi aynı yaklaşımı karakterlerin görülme sıraları için yapar.

N-gram yöntemi, dokümanların benzerliklerinin incelenmesinde ve kümeleme çalışmalarında kullanıldığı gibi genelde büyük boyutlu metinlere uygulanır ve metin içinde kullanılan her kelimenin olasılıkları hesaplanarak elde edilen sonuçlar, takip eden kelimelerin görülme olasılıklarına yansıtılır. Fakat arama motoru sorgularında farklı bir durum söz konusudur. Sorgulardaki kelimeler arası benzerlikler veya yazım farklılıkları, karakter düzeyinde bir incelemeyle yakalanabilir. Ek olarak, burada her sorgu sadece bir önceki sorguyla ilişkilidir, dolayısıyla bütün sorguları incelemek ve bunların görülme olasılıklarını hesaplamak mantıklı değildir. Bu sebeple ardışık iki sorgu, karakter bazında incelenmiş

ve karakter n-gram yöntemiyle *n* adet karakterin görülme oranı hesaplanarak, bu iki sorgunun aynı konu ile ilgili olup olmadığı tahminleri gerçekleştirilmiştir.

2.2. Yöntemin Adımları

Karakter n-gram yönteminde, arama motoru sorguları ilk önce kelimelerine ayrılır, daha sonra bu kelimeler karakterlerine ayrılır ve diğer ayrıştırılmış kelimelerin karakterleriyle karşılaştırılır. Kelimelerin karakterlerine ayrılması ile ilgili bir örnek "Uludağ üniversitesi" ifadesi için Tablo III'deki gibi verilebilir.

Tablo III. Kelimelerin karakterlerine ayrılması

	Karakter n-gramlar
2-gram	Ul lu ud da ağ ün ni iv ve er rs si it te es si
3-gram	Ulu lud uda dağ üni niv ive ver ers rsi sit ite tes esi
4-gram	Ulud luda udağ üniv nive iver vers ersi rsit site ites tesi

Yöntemde kelimelerin karakterlerine ayrılması, N-gram yaklaşımından farklı tarafını ortaya koymaktadır. Daha önce de bahsedildiği gibi N-gram yöntemi bir tümcenin kelimelerini belli bir *n* sayısına göre ayrıştırırken, karakter n-gram yöntemi bir kelimenin karakterlerini belli bir *n* sayısına göre ayrıştırmaktadır. Ayrıştırma sonrasında, karakter n-gram yönteminin, karşılaştırılan ifadelerin benzer olup olmadığını belirleyebilmesi için, benzer karakterleri temsil eden bir değişkene ihtiyaç vardır. Benzerlik oranı olarak tanımlanan bu ifade, karşılaştırılan kelimelerde bulunan ortak n-gramları temsil eder. Bunların yanında, karşılaştırılan ifadelerin benzer olduğu kararı bir eşik değerine dayandırılır ve benzerlik oranı eşik değerini aştığında ifadelerin benzer olduğu kararı verilirse, yönteme esneklik kazandırılabilir. Dolayısıyla kullanıcıya bağlı olan iki değişken, *n* sayısı ve eşik değeridir. Kullanıcıya bağlı olan bu değişkenler yardımıyla farklı *n* sayısı ve eşik değerleri için deney sayısı arttırılabilir ve yöntemin performansı değerlendirilirken daha objektif analizler yapılabilir.

Sorguların karşılaştırılması aşamasında, arama motorundan alınan ardışık iki sorgu tek kelime barındırırsa karakter n-gram mevcut kelimeleri, belirlenen *n* sayısına göre karakterlerine ayırır ve aynı olan karakter n-gramları baz alarak benzerlik oranını hesaplar. FAST arama motorundan alınan "cyberscan" ve "cybersc@n" ardışık iki sorgusu, Tablo IV'deki gibi karakterlerine ayrılabilir. Çalışmada kullanılan benzerlik oranı formülü aşağıdaki gibidir:

Benzerlik Oranı = Aynı n-gram sayısı/Toplam n-gram sayısı (2.21)

Sorguların karakter 2-gramları göz önüne alındığında benzerlik oranı 6/8=0,75 olarak hesaplanır. Eğer kullanıcı eşik değerini 0,7 olarak tanımlarsa, benzerlik oranı eşik değerinden büyük olduğu için, karakter n-gram yöntemi bu ardışık sorguların benzer olduğu kararını verir ve konu devamı tahmini yapar. Dolayısıyla, ardışık sorguların benzer olup olmadıkları ve yöntemin tahminleri, kullanıcı tarafından belirlenen *n* sayısına ve eşik değerine bağlıdır.

Tablo IV. Ardışık iki sorgunun karşılaştırılması

2-gram		3-gram		4-gram	
cybersc@n	cyberscan	cybersc@n	cyberscan	cybersc@n	cyberscan
cy	cy	cyb	cyb	cybe	cybe
yb	yb	ybe	ybe	yber	yber
be	be	ber	ber	bers	bers
er	er	ers	ers	ersc	ersc
rs	rs	rsc	rsc	rsc@	rsca
sc	sc	sc@	sca	sc@n	scan
c@	ca	c@n	can		
@n	an				

Arama motorundaki ardışık sorguların tek kelimeden oluştuğu durumlarda, yöntem yukarıda da bahsedildiği gibi sadece bu kelimelerin karakter n-gramlarını karşılaştırır ve aslında kısmen daha basittir. Ancak FAST ve Excite arama motorlarından alınan sorguların çoğu bir kelimeden fazladır,

dolayısıyla kelimelerin ve karakter n-gramların karşılaştırılması da biraz daha karmaşıktır. Bu karmaşıklığı biraz olsun giderebilmek adına karakter n-gramın çok kelimeli sorgulara uygulanması aşamasında, sorgularda en az bir adet aynı kelimenin olması durumunda iki sorgunun benzer olduğu varsayımı yapılmıştır. Aslında bu mantıklı bir yaklaşımdır çünkü daha önceden bahsedilen tahmin yöntemleri, ardışık iki sorguda aynı olan en az bir kelime ile karşılaşırlarsa konu devamı tahmini yapacak şekilde tasarlanmışlardır. Dolayısıyla karakter n-gram yöntemi de aynı uygulamayı devam ettirecek şekilde oluşturulmuştur. Sonuç olarak karakter n-gram yöntemi, belirli bir *n* sayısı ve eşik değerine göre, birden fazla kelime içeren ardışık iki sorguda, kelimeleri karşılaştırarak herhangi bir kelime çifti için benzer kelime sonucuna varıyorsa, bu ardışık iki sorgu için konu devamı tahmini yapmaktadır. Örnek olarak FAST verisinde bulunan aşağıdaki sorgular ele alınsın:

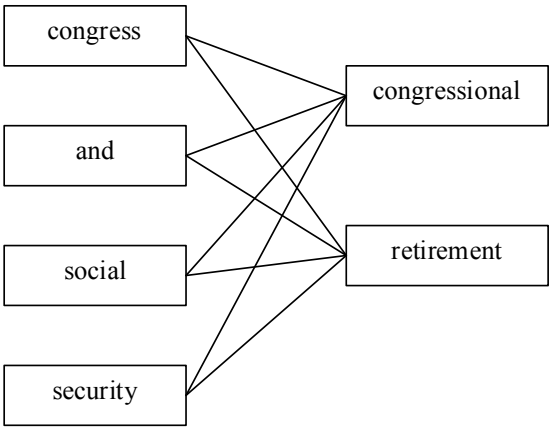
q_i = congress and social security

q_{i+1}= congressional retirement

Karakter n-gram yöntemi, sorguları ilk önce Şekil 1'deki gibi kelimelerine ayırır, daha sonra her kelimenin karakter n-gramını Tablo V'deki gibi bulur. Karşılaştırma aşamasında yine Şekil 1'deki gibi ilk sorgunun ilk kelimesinin karakter n-gramlarını, ikinci sorgunun bütün kelimelerinin karakter n-gramlarıyla sırasıyla karşılaştırır. Bu aşamada karşılaştırılan n-gramların benzerlik oranları eşik değerini geçerse, yöntem benzer kelime kararı verir ve sorguların da benzer olduğu sonucuna varır. Eğer birinci sorgunun ilk kelimesiyle böyle bir sonuç elde edilemezse ikinci kelimesinin karakter n-gramları ele alınır ve yine önceki adıma benzer bir şekilde ikinci sorgunun ilk kelimesinden başlamak üzere bütün kelimelerin n-gramlarıyla karşılaştırmalar yapılır. Bu işlem, sorgulardaki kelimelere ait n-gramların hepsi birbirleriyle karşılaştırılıp benzer sorgu veya aksi bir karar verilene kadar devam eder. Anlatıldığı şekilde yöntem, örnekteki iki sorgunun n-gramlarını karşılaştırmış ve "congress" ile "congressional" için benzer kelime kararı verdiğinden bu sorgular doğru bir şekilde konu devamı olarak işaretlenmiştir. Karşılaştırma 3-gram ve 0,6 eşik değeri için yapılmıştır.

Tablo V. Ardışık iki sorgunun karşılaştırılması

q_i=congress	Con-ong-ngr-gre-res-ess
q_{i+1}= congressional	Con-ong-ngr-gre-res-ess-ssi-sio-ion-ona-nal
q_{i+1}= retirement	Ret-eti-tir-ire-rem-eme-men-ent
q_i=and	and
q_{i+1}= congressional	Con-ong-ngr-gre-res-ess-ssi-sio-ion-ona-nal
q_{i+1}= retirement	Ret-eti-tir-ire-rem-eme-men-ent
q_i= social	Soc-oci-cia-ial
q_{i+1}= congressional	Con-ong-ngr-gre-res-ess-ssi-sio-ion-ona-nal
q_{i+1}= retirement	Ret-eti-tir-ire-rem-eme-men-ent
q_i= security	Sec-ecu-cur-uri-rit-ity
q_{i+1}= congressional	Con-ong-ngr-gre-res-ess-ssi-sio-ion-ona-nal
q_{i+1}= retirement	Ret-eti-tir-ire-rem-eme-men-ent



Şekil 1:
Bir kelimeden fazla sorguların karşılaştırılma yöntemi.

Karakter n-gram yönteminin bir oturum dâhilinde uygulanması ile ilgili algoritma adımları aşağıdaki gibidir:

Adım 0: Bir oturum içindeki sorgular *i* ile gösterilir ve *i*=1 olarak alınır (*i*=1,2,…*I*).

Adım 1: Sorgular sırasıyla "önceki sorgu" (q_i) ve "sonraki sorgu" (q_{i+1}) olarak nitelendirilir. Belirlenen ardışık iki sorgu, kelimelerine ayrılır ve her kelime bir diziye atanır. Birinci sorguda *J* adet ve ikinci sorguda *Z* adet kelime olduğu varsayılır ve $A_i[j]$ ile $A_{i+1}[z]$, ardışık sorguların kelimelerini barındıran diziler olarak tanımlanır (*j*=1,2,….,*J*; *z*=1,2,…,*Z*). *j*=1 ve *z*=1 olarak alınır.

Adım 2: $A_i[j]$ dizisi ve $A_{i+1}[z]$ dizisi karakter n-gramlarına ayrılır ve benzerlik oranı hesaplanır. Eğer benzerlik oranı eşik değerini aşarsa kelimelerin benzer olduğu kararı verilir ve Adım 6'e gidilir.

Adım 3: Eğer *z*<*Z* ise *z*=*z*+1. Diğer bir ifadeyle, ikinci sorguda hala kelime varsa, sonraki kelime seçilir ve Adım 2'ye gidilir.

Adım 4: Eğer *j*<*J* ise *j*=*j*+1, *z*=1 ve Adım 2'ye gidilir. Kısaca ilk sorgunun ilk kelimesi ikinci sorgunun bütün elemanlarıyla karşılaştırılmıştır, bu adımda ilk sorgunun sonraki elemanı alınır ve karşılaştırma işlemleri tekrarlanır.

Adım 5: İki konu benzer değildir ve konu değişimi kararı verilir ve Adım 7'ye gidilir.

Adım 6: İki sorgu benzerdir ve konu devamı kararı verilir.

Adım 7: Eğer *i*<*I* ise *i* sayısının değeri 1 arttırılır, *i*=*i*+*1* ve Adım 1'e gidilir. Diğer bir ifadeyle Adım 3 ve Adım 4'de de kontrolleri yapıldığı gibi iki sorgunun bütün kelimeleri karşılaştırılmıştır. Bu aşamadan sonra bir oturum içindeki sorgu sayısını gösteren indisin değeri bir arttırılarak sonraki iki sorgu incelemeye alınır ve algoritma devam eder.

Adım 8: Eğer (q_i) oturumun son sorgusu ise, bu oturum için algoritma sonlandırılır.

2.3. Kullanılan Veriler

Bu çalışmada kullanılan veriler gerçek arama motoru verileri olup, Excite (http://www.excite.com), ve FAST (http://www.fast.com) arama motorlarından alınmıştır (Özmutlu ve diğ. 2004a).

Excite arama motorundan 1,7 milyon adet sorgu, Mayıs 2001'de toplanmıştır. Excite veri kütüğü yapısında, girişler geliş sırasına göre olmaktadır. Her yeni kullanıcıya tekil bir numara (ID) atanmaktadır ve dolayısıyla yeni kullanıcılar ID'lerinden ayırt edilebilmektedir. Ayrıca Excite her bir sorguya saati, dakikayı ve saniyeyi içeren zaman verisini atamaktadır. Bu zaman verileri daha önceki çalışmalarda yeni konu tanılamada kullanılmıştır (Özmutlu ve diğ. 2004a, Özmutlu ve Çavdur, 2005a, Özmutlu ve diğ. 2008a, Özmutlu ve diğ. 2008b). Tekil kullanıcı listesi hazırlanıp, bu liste üzerinden Poisson örneklemesi yöntemi ile tespit edilen kullanıcıların tüm arama kayıtları seçilerek 10,256 adet sorgu kümesi elde edilmiştir (Özmutlu ve diğ. 2002). Örneklem büyüklüğü çok geniş tutulamamıştır, çünkü çalışma sonuçlarının performans değerlendirmesi için uzman tarafından sorguların gözden geçirilip gerçek bilgilerin elde edilmesi gerekmektedir.

FAST arama motorundan 1,257,891 adet sorgu 6–7 Şubat 2001 tarihlerinde geliş sıraları korunarak toplanmıştır. Bu arama motorunda da Excite arama motorunda olduğu gibi her kullanıcıya bir ID ve zaman verisi atanmaktadır. Toplam veri kümesinden Poisson örneklemesiyle 10,007 adet sorgu seçilmiştir. Yeni konu tanılama amacıyla yapılan ve aynı verileri kullanan önceki çalışmalarda da seçilen veri setleri, eğitim ve test kümesi olarak Tablo VI'da gösterildiği gibi iki eşit parçaya ayrılmıştır.

Tablo VI. Çalışmada kullanılan verilerin büyüklüğü

Arama Motoru	Excite	FAST
Bütün veri kümesi	1,7 milyon	1,257,891
Örneklem Kümesi	10,256	10,007
Örneklem kümesinin ilk kısmı	5128 sorgu	4997 sorgu
Örneklem kümesinin ikinci kısmı	5128 sorgu	5010 sorgu

Karakter n-gram yöntemi, önceki yöntemlerde de olduğu gibi karar verme aşamasında bir oturumdaki sorguları baz alır. Oturum içindeki ardışık sorguları karşılaştırarak konu değişimi veya konu devamı kararı verir. Bu yüzden, her oturumun son sorgusu karar aşamasında değerlendirmeye alınamamıştır. Anlatılanlar ışığında çalışmalarda kullanılan verilerin oturum sayıları belirlenmiş ve her oturum için değerlendirilen sorgu sayısı son sorgu dahil edilmeyecek şekilde düzeltilmiştir. Bu düzeltme sonrası karakter n-gram tarafından iki arama motorunda değerlendirilen sorgu sayıları Tablo VII'deki gibidir. Eğitim kümesi önceden geliştirilen istatistiksel yöntemlerin eğitiminde kullanılmış, daha sonra bu yöntemler test kümesi üzerinde çalıştırılmıştır. Karakter n-gram yönteminin performansının önceki yöntemlerle anlamlı bir şekilde karşılaştırılabilmesi için, yöntem sadece test kümesine uygulanmıştır.

Tablo VII. Değerlendirilen veri büyüklükleri

	Sorgu Sayısı	Oturum Sayısı	Değerlendirilen Sorgu Sayısı
Eğitim Kümesi	5128-Excite 4997-Fast	1858-Excite 437-Fast	3270-Excite 4560-Fast
Test Kümesi	5128-Excite 5010-Fast	1734-Excite 526-Fast	3394-Excite 4484-Fast
Toplam Veri	10256-Excite 10007-Fast	3592-Excite 963-Fast	6664-Excite 9044-Fast

Bu çalışmada sorgu; bir kullanıcının gerçekleştirdiği bir veya daha fazla terimden oluşan arama kümesi ve oturum; bir kullanıcının bütün sorgularının bir kümesidir. Bir oturum tek sorgudan oluşabilir veya birçok sorgu barındırabilir (Spink ve diğ. 2001).

2.4. Verilerin Uzman Tarafından Değerlendirilmesi

Seçilmiş olan Excite ve FAST verileri, bir uzman tarafından manuel olarak değerlendirilmiştir. Bu değerlendirmede uzman sorguları incelemiş ve her bir sorgunun gerçek konu değişimlerini ve devamlılıklarını belirlemiştir. Bu belirleme sayesinde uygulanan tahmin yöntemlerinin performans değerlendirmesi doğrulukla yapılabilmektedir. Karakter n-gramın performans değerlendirmesi için de bu aşama çok önemlidir.

2.5. Verilerin Temizlenmesi

Arama motoru kullanıcıları yaptıkları sorgularda, kelimelerin yanında çeşitli karakterler ve İngilizcede sıklıkla kullanılan, ifadeye anlam katan terimler de kullanmaktadırlar. Kelimeler haricinde kullanılan bu terim ve karakterler, yöntemin yanlış tahminlerde bulunmasına sebep olmaktadır. Örnek olarak arka arkaya gelen www.uludag.edu ve www.uludag.edu.tr sorguları farklı kelime grubu olarak algılanıp konu değişimi kararı verilebilmektedir. Ancak, gerekli temizlemelerden sonra kalan "uludag" ve "uludag" terimlerinin aynı olduğu görülmekte ve konu devamı kararı verilebilmektedir.

Verinin temizlenmesi aşamasında temizleme sırası önem taşımaktadır. Sorgular ".", ",", ":", "+", ":", "%", "&", "[", "]", "(", ")", " ' ", "!", "$", "/", "\", "<", ">" ve "www", "http", "com", "uk", "au", "edu", "and", "or", "on", "of", "at", "in", "a", "an", "for", "to" gibi ifadeye anlam katan; fakat konu değişikliğine neden olmayan ve sıkça kullanılan terimlerden temizlenmiştir. Bu terimlerin ortak olması, anlam bazında konu devamı niteliğinde olmayacağından, olası hataların önlenmesi adına sorgulardan çıkartılmışlardır.

2.6. Notasyon

Çalışmada kullanılan terminoloji aşağıdaki gibidir:

Konu değişimi :Tek kullanıcı oturumunda, sorgular arasında bir konudan diğerine geçiş.

Konu devamı :Tek kullanıcı oturumunda, sorgular arasında aynı konuda kalma.

$N_{değişim}$:Karakter n-gram tarafından konu değişimi olarak tahmin edilen sorgu sayısı.

N_{devam} :Karakter n-gram tarafından konu devamı olarak tahmin edilen sorgu sayısı.

$N_{gerçek\ değişim}$:Uzman tarafından konu değişimi olarak işaretlenen sorgu sayısı.

$N_{gerçek\ devam}$:Uzman tarafından konu devamı olarak işaretlenen sorgu sayısı.

$N_{değişim\&doğru}$:Hem karakter n-gram hem de uzman tarafından konu değişimi olarak işaretlenen sorgu sayısı.

$N_{devam\&doğru}$:Hem karakter n-gram hem de uzman tarafından konu devamı olarak işaretlenen sorgu sayısı.

A tipi Hata :Karakter n-gram tarafından konu değişimi olarak tahmin edilip gerçekte konu devamı olan sorgu sayısı.

B tipi Hata :Karakter n-gram tarafından konu devamı olarak tahmin edilip gerçekte konu değişimi olan sorgu sayısı.

Yukarıdaki notasyonlar arası hesaplamalar aşağıdaki gibidir:

$$N_{gerçek\ değişim} = N_{değişim\&doğru} + \text{B tipi Hata} \quad (2.22)$$

$$N_{gerçek\ devam} = N_{devam\ \&doğru} + \text{A tipi Hata} \quad (2.23)$$

$$N_{değişim} = N_{değişim\&\ doğru} + \text{A tipi Hata} \quad (2.24)$$

$$N_{devam} = N_{devam\ \&doğru} + \text{B tipi Hata} \quad (2.25)$$

Performans değerlendirmeleri için daha önceki anlam bazlı olmayan çalışmalarda da yer alan, Duyarlılık (Precision-*P*) ve Anma (Recall-*R*) ölçütleri kullanılmıştır. Böylelikle karakter n-gram yönteminin aynı performans ölçütleriyle, diğer anlam bazlı olmayan metotlarla karşılaştırılması mümkün olmaktadır. Bu ölçütlerle beraber bir uygunluk fonksiyonu da F_β dikkate alınmıştır. Performans ölçütleri aşağıdaki gibi hesaplanır:

$$P_{değişim} = N_{değişim\&\ doğru} / N_{değişim} \quad (2.26)$$

$$P_{devam} = N_{devam\&\ doğru} / N_{devam} \quad (2.27)$$

$$R_{değişim} = N_{değişim\&\ doğru} / N_{gerçek\ değişim} \quad (2.28)$$

$$R_{devam} = N_{devam\&\ doğru} / N_{gerçek\ devam} \quad (2.29)$$

$$F_{\beta\_değişim} = [(1+\beta^2)\ P_{değişim}.R_{değişim}]/[\beta^2 P_{değişim} + R_{değişim}] \quad (2.30)$$

$$F_{\beta\_devam} = [(1+\beta^2)\ P_{devam}.R_{devam}]/[\beta^2 P_{devam} + R_{devam}] \quad (2.31)$$

Konu değişimlerini yorumlayan duyarlılık değeri ($P_{değişim}$), karakter n-gram yöntemi tarafından doğru bir şekilde konu değişimi olarak işaretlenen sorguların, yöntem tarafından konu değişimi olarak tahmin edilen bütün sorgu sayısına oranıdır. Anma değeri olan ($R_{değişim}$) ise karakter n-gram yöntemi tarafından doğru bir şekilde konu değişimi olarak işaretlenen sorguların, uzman tarafından belirlenen konu değişimlerine oranıdır. Diğer taraftan konu devamlarını yorumlayan duyarlılık değeri (P_{devam}), karakter n-gram yöntemi tarafından doğru bir şekilde konu devamı olarak işaretlenen sorguların, yöntem tarafından konu devamı olarak tahmin edilen sorgu sayısına oranıdır. Anma değeri olan (R_{devam}) ise karakter n-gram yöntemi tarafından doğru bir şekilde konu devamı olarak işaretlenen sorguların, uzman tarafından belirlenen konu devamlarına oranıdır.

Alternatif çözüm algoritmaları genellikle bir performans ölçütünde (*P veya R*) iyileşme sağladığında diğer ölçütte kötüleşmeye neden olur. Bu nedenle, $F_{\beta\_değişim}$ ölçütü, $P_{değişim}$ ve $R_{değişim}$ değerlerini birleştirerek farklı sonuçların sağlıklı karşılaştırmasını sağlamak amacıyla kullanılır. Aynı şekilde F_{β_devam} ölçütü, P_{devam} ve R_{devam} değerleriyle performansı gösteren tek bir değer elde edilmesini sağlar. Bu çalışmada β parametresi konu değişimlerini tahmin etmede çıkan farklı tipteki hataları ölçülendirmek için kullanılmış ve önceki çalışmalarla benzerliği korumak için 1,3 olarak kabul edilmiştir. Böylelikle yeni yöntem ile aynı veriler üzerinde çalışan, aynı ölçütleri kullanan önceki yöntemler arasında sağlıklı karşılaştırmalar yapılabilecektir.

3. SONUÇLAR

3.1. Karakter N-gram Yöntemi Sonuçları

Karakter n-gram yöntemi FAST ve Excite veri kümelerine uygulanmıştır. Bu uygulamada kullanılan sorgular, önceki yöntemlerin test kümesi olarak kullandığı sorgulardır. Böylelikle karakter n-

gram ve diğer yöntemlerin sonuçları gerçekçi bir şekilde karşılaştırılabilir. Her iki arama motoru verilerinde, çeşitli *n* ve eşik değerleri için deneyler tekrarlanmıştır. Sorgulardaki kelime uzunlukları dikkate alınarak *n* değeri 1 – 4 arasında ve eşik değeri 0,5 – 0,7 arasında alınmıştır. Bu deneylerin sonuçları Tablo VIII – Tablo XI'de gösterilmiştir.

Karakter n-gram yönteminin Excite arama motorundan alınan verilere uygulanması sonucu oluşturulan Tablo VIII'de, uzman değerlendirmeleri ve karakter n-gram yöntemi sonuçları karşılaştırma amacıyla bir arada verilmiştir.

Tablo VIII. Excite verilerinde farklı *n* ve eşik değerleri için karakter n-gram uygulaması sonuçları

				Uzman Sonuçları					
Ngram	**Eşik Değeri**	$N_{değişim}$	N_{devam}	$N_{gerçekdeğişim}$	$N_{gerçekdevam}$	$N_{değişim\&doğru}$	$N_{devam\&doğru}$	**A tipi hata**	**B tipi hata**
n=1	**0,5**	200	3194	272	3122	69	2991	131	203
	0,6	339	3055	272	3122	116	2899	223	156
	0,7	495	2899	272	3122	180	2807	315	92
n=2	**0,5**	682	2713	272	3122	247	2688	434	26
	0,6	724	2670	272	3122	261	2659	463	11
	0,7	739	2655	272	3122	263	2646	476	9
n=3	**0,5**	752	2642	272	3122	262	2632	490	10
	0,6	770	2624	272	3122	263	2615	507	9
	0,7	778	2616	272	3122	264	2608	514	8
n=4	**0,5**	855	2539	272	3122	263	2530	592	9
	0,6	866	2528	272	3122	264	2520	602	8
	0,7	876	2518	272	3122	264	2510	612	8

Daha önceden de bahsedildiği gibi, çalışmada kullanılan Excite örnekleminin büyüklüğü 10,256 iken oturumların son sorgularının tahminlere katılamamasından dolayı, toplam sorgu adedi 6,664'e düşmüştür. Karakter n-gram yöntemi, diğer yöntemlerle karşılaştırma yapabilmek amacıyla, test kümesini oluşturan 3,394 adet sorguya uygulanmıştır. Yöntemin konu değişimi tahminleri değişik *n* ve eşik değerleri için farklıdır. Düşük eşik değeri ve *n* sayısı için konu değişimi tahmini daha az olmakla beraber bu tahminler *n* sayısı ve eşik değeri arttıkça artar. Uzman tarafından belirlenen konu değişimi sayısı dikkate alındığında karakter n-gram yönteminin 1-gram ve 0,5 eşik değeri sonuçları hariç, her zaman daha fazla tahminler yaptığı görülmektedir. Konu değişimi tahminlerinin fazla yapılması, A tipi hatanın artmasına sebep olmaktadır. Diğer taraftan konu devamı tahminleri incelendiğinde yöntemin yine 1-gram ve 0,5 eşik değeri sonuçları hariç, mevcut konu devamı kararlarından daha az konu devamı tahmini yaptığı görülmektedir. Hatta bu tahminler *n* sayısı ve eşik değeriyle ters orantılıdır; *n* sayısı ve eşik değeri arttıkça yöntemin konu devamı tahmin sayısı azalır ve uzman tarafından belirlenen konu devamı sayısının altına düşer. Bu azalmayla beraber konu devamı tahmininde yapılan hataları yansıtan B tipi hatada da azalma olduğu görülmektedir. Excite verisi için Tablo IX'da Duyarlılık ve Anma değerleri ile uygunluk fonksiyonları hesaplanmıştır.

Tablo IX incelendiğinde karakter n-gram yönteminin, konu devamı ve konu değişimi tahminlerinin doğruluk oranlarının, birbirleriyle ters orantılı olduğu görülmektedir. Örneğin yöntemin doğru tahmin ettiği konu devamı oranı 2-gram ve 0,7 eşik değeri için R_{devam}=0,848 iken konu değişimleri ($R_{değişim}$) %96,7 oranında doğru tahmin edilmiştir. Bu değerler kabul edilebilir olmakla beraber farklı n-gram ve eşik değerlerinde yöntemin sonuçları kötüleşmektedir. Aynı şekilde Tablo IX incelenmeye devam edilirse, performans göstergeleri olan $F_{\beta(değişim)}$ ve $F_{\beta(devam)}$ değerlerinde de başarılı artışlar gerçekleştiği görülmektedir. Örneğin 2-gram ve 0,7 eşik değerinde $F_{\beta(değişim)}$ değeri 0,59 olarak bulunurken, $F_{\beta(devam)}$ değeri 0,897 olarak bulunmuştur. Yine aynı şekilde bu performans göstergelerindeki artışlar n-grama ve eşik değerine göre farklılıklar göstermektedir.

Tablo IX. Excite verilerinde farklı *n* ve eşik değerleri için karakter n-gram uygulamasının performans analizi

Ngram	Eşik Değeri	$P_{değişim}$	P_{devam}	$R_{değişim}$	R_{devam}	$F_{ß(değişim)}$	$F_{ß(devam)}$
n=1	**0,5**	0,345	0,936	0,254	0,958	0,281	0,950
	0,6	0,342	0,949	0,426	0,929	0,391	0,936
	0,7	0,364	0,968	0,662	0,899	0,507	0,924
n=2	**0,5**	0,362	0,991	0,908	0,861	0,582	0,905
	0,6	0,360	0,996	0,960	0,852	0,593	0,900
	0,7	0,356	0,997	0,967	0,848	0,590	0,897
n=3	**0,5**	0,348	0,996	0,963	0,843	0,582	0,894
	0,6	0,342	0,997	0,967	0,838	0,575	0,890
	0,7	0,339	0,997	0,971	0,835	0,574	0,889
n=4	**0,5**	0,308	0,996	0,967	0,810	0,538	0,871
	0,6	0,305	0,997	0,971	0,807	0,536	0,869
	0,7	0,301	0,997	0,971	0,804	0,532	0,866

Tablo X. FAST verilerinde farklı *n* ve eşik değerleri için karakter n-gram uygulaması sonuçları

				Uzman Sonuçları					
Ngram	Eşik Değeri	$N_{değişim}$	N_{devam}	$N_{gerçekdeğişim}$	$N_{gerçekdevam}$	$N_{değişim\&doğru}$	$N_{devam\&doğru}$	A tipi hata	B tipi hata
n=1	**0,5**	180	4304	310	4174	55	4049	125	255
	0,6	302	4182	310	4174	103	3975	199	207
	0,7	474	4010	310	4174	171	3871	303	139
n=2	**0,5**	710	3774	310	4174	277	3741	433	33
	0,6	751	3733	310	4174	289	3712	462	21
	0,7	770	3714	310	4174	295	3699	475	15
n=3	**0,5**	783	3701	310	4174	297	3688	486	13
	0,6	800	3684	310	4174	303	3677	497	7
	0,7	803	3681	310	4174	303	3674	500	7
n=4	**0,5**	917	3567	310	4174	303	3560	614	7
	0,6	921	3563	310	4174	303	3556	618	7
	0,7	924	3560	310	4174	303	3553	621	7

Karakter n-gram yönteminin FAST verilerine uygulanması sonucu elde edilen tahminler Tablo X'da verilmiştir. FAST örnekleminin toplam büyüklüğü 10,007 iken önceden de bahsedildiği gibi oturumların son sorgularının tahminlere katılamamasından dolayı toplam sorgu adedi 9,044'e düşmüştür. Karakter n-gram yöntemi, diğer yöntemlerle karşılaştırılabilmek amacıyla, test kümesini oluşturan 4,484 adet sorguya uygulanmıştır.

Karakter n-gram yönteminin konu değişimi tahminleri, farklı *n* ve eşik değerleri için değişmektedir. Düşük eşik değeri ve *n* sayısı için konu değişimi tahminleri adedi, *n* sayısı ve eşik değeri arttıkça artmaktadır. Uzman tarafından belirlenen konu değişimi sayısı dikkate alındığında karakter n-gramın tahminlerinin fazla olduğu görülmekte ve bu durum da A tipi hatadaki artışı açıklamaktadır. Bununla beraber, konu devamı tahminleri incelendiğinde elde edilen sonuçlar uzman sonuçlarıyla karşılaştırıldığında, yöntemin daha az sayıda konu devamı tahmini yaptığı görülmektedir. Ayrıca bu tahminler *n* sayısı ve eşik değeriyle ters orantılıdır; *n* sayısı ve eşik değeri arttıkça yöntemin konu devamı tahmin sayısı azalmakta ve uzman tarafından belirlenen konu devamı sayısının altına düşmektedir, dolayısıyla B tipi hatada da azalma gözlenmektedir. Tablo XI'de FAST verisi için karakter n-gram yönteminin performans ölçütleri yer almaktadır.

Tablo XI. FAST verilerinde farklı *n* ve eşik değerleri için karakter n-gram performans değerlendirmesi

Ngram	Eşik Değeri	$P_{değişim}$	P_{devam}	$R_{değişim}$	R_{devam}	$F_{ß(değişim)}$	$F_{ß(devam)}$
n=1	**0,5**	0,306	0,941	0,177	0,970	0,210	0,959
	0,6	0,341	0,951	0,332	0,952	0,335	0,952
	0,7	0,361	0,965	0,552	0,927	0,461	0,941
n=2	**0,5**	0,390	0,991	0,894	0,896	0,604	0,929
	0,6	0,385	0,994	0,932	0,889	0,610	0,926
	0,7	0,383	0,996	0,952	0,886	0,613	0,924
n=3	**0,5**	0,379	0,996	0,958	0,884	0,611	0,922
	0,6	0,379	0,998	0,977	0,881	0,616	0,921
	0,7	0,377	0,998	0,977	0,880	0,614	0,921
n=4	**0,5**	0,330	0,998	0,977	0,853	0,566	0,902
	0,6	0,329	0,998	0,977	0,852	0,564	0,901
	0,7	0,328	0,998	0,977	0,851	0,563	0,900

Karakter n-gram yönteminin FAST verileri için gerçekleştirdiği tahminlerde, Excite verilerinde de olduğu gibi, konu devamı ile konu değişimi tahminlerinin doğruluk oranlarında ters orantı söz konusudur, doğru tahmin edilen konu değişimi sorgularının yüzdesi artarken, konu devamı sorguların doğru tahmin edilme yüzdesi düşmektedir. Örneğin yöntemin doğru tahmin ettiği konu devamı sorgu sayısı oranı 2-gram ve 0,7 eşik değeri için R_{devam}=0,886 iken konu değişimi sorgu sayısı oranı $R_{değişim}$=0,952 olarak bulunmuştur. Bu değerler kabul edilebilir olmakla beraber farklı n-gram ve eşik değerlerinde yöntemin sonuçları kötüleşmektedir. Aynı şekilde Tablo XI incelenmeye devam edilirse, performans göstergeleri olan $F_{ß(değişim)}$ ve $F_{ß(devam)}$ değerlerinde de başarılı artışlar gerçekleştiği görülmektedir. Örneğin 2-gram ve 0,7 eşik değerinde $F_{ß(değişim)}$ değeri 0,613 olarak bulunurken, $F_{ß(devam)}$ değeri 0,924 olarak bulunmuştur.

3.2. Karakter n-gram Yöntemi ile Önceki Yöntemlerin Karşılaştırılması

Karakter n-gram yönteminin performans değerlendirmesini sağlıklı bir şekilde yapabilmek için aynı verileri kullanan önceki yöntemlerin sonuçları ile karakter n-gram yönteminin tahminleri karşılaştırılmıştır. Böylece, yapılan çalışmanın önceki araştırmalara oranla hangi alanlarda başarılı olduğu ve nerelerde eksik kaldığı ayrıntılı bir şekilde gözlemlenebilir. Bu amaçla bugüne kadar yapılan, anlam bazlı olmayan yeni konu tanılama çalışmalarının performans parametreleri karşılaştırılmıştır. Bu karşılaştırmalar Excite verisi için Tablo XII'de ve FAST verisi için Tablo XIII'de özetlenmiştir. Bu tablolara karakter n-gram yönteminin başarılı olduğu kombinasyonlar eklenmiştir.

Tablo XII'de başarılı tahmin oranlarını veren R_{devam} ve $R_{değişim}$ değerleri karşılaştırıldığında, en doğru tahminlerin yapay sinir ağları yönteminin uygulanmasıyla elde edildiği görülmektedir. Karakter n-gram yönteminin performansı, diğer yöntemlere göre daha düşük bulunmuştur. Yanlış tahmin edilen konu değişimi adedini veren A tipi hata değeri, diğer yöntemlerin bulgularının iki katından fazladır.

Tablo XII. Excite verilerine uygulanan yöntemlerin analiz sonuçları

	Analiz edilen Sorgu sayısı	Konu Değişim sayısı	Konu Devamı sayısı	Doğru Tahmin Edilen değişimler	Doğru Tahmin Edilen devamlar	A Tipi Hata	B Tipi Hata	$P_{değişim}$	$R_{değişim}$	P_{devam}	R_{devam}	$F_{ß(değişim)}$	$F_{ß(devam)}$
Uzman Sonuçları	3394	$N_{gerçekdeğişim}$ = 272	$N_{gerçekdevam}$ = 3122	-----	-----	----	----	----	----	----	----	----	----
Monte Carlo Simülasyonu	3394	$N_{değişim}$ = 393	N_{devam} = 3001	$N_{değişim\&doğru}$ = 142	$N_{devam\&\ doğru}$ = 2871	251	130	0.36	0.53	0.96	0.92	0.45	0.94
YSA Sonuçları (2008)	3394	$N_{değişim}$ = 454	N_{devam} = 2940	$N_{değişim\&doğru}$ = 237	$N_{devam\&\ doğru}$ = 2905	217	35	0.522	0.871	0.988	0.93	0.698	0.95
Karakter 2-gram Eşik=0,7	3394	$N_{değişim}$ = 739	N_{devam} = 2655	$N_{değişim\&doğru}$ = 263	$N_{devam\&\ doğru}$ = 2646	476	9	0.356	0.967	0.997	0.848	0.590	0.897

Tablo XIII. FAST verilerine uygulanan yöntemlerin analiz sonuçları

	Analiz edilen Sorgu sayısı	Konu Değişim sayısı	Konu Devamı sayısı	Doğru Tahmin Edilen değişimler	Doğru Tahmin Edilen devamlar	A Tipi Hata	B Tipi Hata	$P_{değişim}$	$R_{değişim}$	P_{devam}	R_{devam}	$F_{\beta(değişim)}$	$F_{\beta(devam)}$
Uzman Sonuçları	4484	$N_{gerçekdeğişim}$ = 240	$N_{gerçekdevam}$ = 4174	-----	-----	----	----	----	----	----	----	----	----
YSA Sonuçları (2005)	4484	$N_{değişim}$ = 865	N_{devam} = 3619	$N_{değişim\&doğru}$ = 305	$N_{devam\&\ doğru}$ = 3614	560	5	0.353	0.984	0.999	0.866	0,635	0,903
Monte Carlo Simülasyonu	4484	$N_{değişim}$ = 338	N_{devam} = 4146	$N_{değişim\&doğru}$ = 137	$N_{devam\&\ doğru}$ = 3973	201	173	0.41	0.44	0.96	0.95	0.43	0.95
Şartlı Olasılıklar	4484	$N_{değişim}$ = 276	N_{devam} = 4208	$N_{değişim\&doğru}$ = 146	$N_{devam\&\ doğru}$ = 4044	130	164	0.529	0.471	0.961	0.968	0.491	0.966
Demster Shafer Teorisi	4484	$N_{değişim}$ = 836	N_{devam} = 3648	$N_{değişim\&doğru}$ = 303	$N_{devam\&\ doğru}$ = 3641	533	7	0.362	0.977	0.998	0.872	0.642	0.907
YSA Sonuçları (2008)	4484	$N_{değişim}$ = 886	N_{devam} = 3598	$N_{değişim\&doğru}$ = 306	$N_{devam\&\ doğru}$ = 3594	580	4	0.345	0.987	0.998	0.861	0.583	0.907
Karakter 2-gram Eşik=0,7	4484	$N_{değişim}$ = 770	N_{devam} = 3714	$N_{değişim\&doğru}$ = 295	$N_{devam\&\ doğru}$ = 3699	475	15	0.383	0.952	0.996	0.886	0.613	0.924

Aynı şekilde Tablo XIII'de de görüldüğü gibi, FAST verilerine uygulanan yöntemlerin performans değerlendirmeleri yapıldığında en başarılı tahminler yapay sinir ağları yöntemiyle elde edilmiştir. Karakter n-gram yöntemi ile FAST verisinde daha iyi sonuçlar elde edilmesine rağmen yine de geçmiş çalışmalara göre iyileştirme sağlanamamıştır.

Sonuç olarak karakter n-gram yöntemi, diğer yöntemlerde olmayan hatalar ortaya çıkarmıştır. Gerçekte konu devamı olan birçok sorguya konu değişimi tahmini yaparak doğru tahmin etme yüzdelerini ve dolayısıyla performans ölçütlerini düşürmüştür. Bu noktada karakter n-gram yönteminin tahminlerde bulunurken, sorgulardaki kelimelerin karakterlerinin yan yana gelme sıklıklarını baz alması hatalara sebep olmuştur. Diğer bir ifadeyle, benzer olmayan kelimelerin benzer olan karakter n-gramları, benzerlik oranının değerini arttırmakta ve belirlenen eşik değeriyle karşılaştırıldığında, kelimeler benzer olmasa da konu devamı kararı verilmektedir.

Önceden de bahsedildiği gibi yeni konu tanılamada kullanılan yöntemlerin amacı, konu değişimlerinin doğru şekilde tahmin edilmesi olduğundan, A tipi hatalar önem kazanmaktadır. Ancak karakter n-gram yönteminde A tipi hatalar diğer yöntemlere oranla daha yüksek bulunmuştur. Bu sebeple karakter n-gram yöntemi tek başına yeterli değildir.

Önceki yöntemlerin hatalı olarak konu değişimi tahmini yaptığı sorgular incelendiğinde, bu hataların yazım farklılıklarından kaynaklandığı tespit edilmiş ve karakter n-gram yönteminin, bu hataları giderebilecek çalışma algoritmasına sahip olduğu görülmüştür. Dolayısıyla önceki yöntemlerden en başarılı tahminleri gerçekleştiren çalışmanın konu değişimi tahmini yaptığı sorgulara, karakter n-gram algoritması uygulandığı takdirde, yazım farklılıklarından kaynaklanan hataların giderilebileceği öngörülmüştür. Yukarıda da bahsedildiği gibi önceki yöntemlerden en başarılı tahminleri üreten yapay sinir ağları olduğu için, bu çalışmanın konu değişikliği kararı verdiği sorgulara karakter n-gram yöntemi uygulanmıştır. Burada yapay sinir ağları uygulaması tekrar edilmemiş olup, sadece yöntemin tahmin sonuçları çalışma kapsamında kullanılmıştır.

Yapay sinir ağları yönteminde kullanılan verilerin önceki çalışmalarda hazırlık aşaması tamamlanmış, yöntemin test kümesi verilerine uygulanmasıyla tahminler elde edilmiştir. Tablo XIV ve Tablo XV'de Excite ve FAST verileri için elde edilen sonuçlar görülmektedir. Bu tablolar incelendiğinde A tipi hataların, Excite verisi için 217 ve FAST verisi için 580 adet olduğu görülür.

Tablo XIV. Excite verisinin ikinci yarısı için yapay sinir ağlarıyla bulunan konu değişim ve konu devamları sayısı

	Analiz edilen Sorgu sayısı	Konu Değişim sayısı	Konu Devamı sayısı	Doğru Tahmin Edilen değişimler	Doğru Tahmin Edilen devamlar	A Tipi Hata	B Tipi Hata	$P_{değişim}$	$R_{değişim}$	P_{devam}	R_{devam}	$F_{\beta(değişim)}$	$F_{\beta(devam)}$
YSA Sonuçları	3394	$N_{değişim}$ = 454	N_{devam} = 2940	$N_{değişim\&doğru}$ = 237	$N_{devam\&\ doğru}$ = 2905	217	35	0.522	0.871	0.988	0.93	0.698	0.95
Uzman Sonuçları	3394	$N_{gerçekdeğişim}$ = 272	$N_{gerçekdevam}$ = 3122	-----	-----	----	----	----	----	----	----	----	----

Yapay sinir ağları uygulamasının hata yaptığı sorgular incelendiğinde, bu sorguların yazım hataları veya farlılıkları içerdiği görülür. Hâlbuki yazım hatalarının veya farklılıklarının algılanabilmesi bu sorguların yöntem tarafından farklı değerlendirilmesine yol açar ve büyük ihtimalle, konu devamı tahmini yapılır. Bu sebeple karakter n-gram yöntemi, yapay sinir ağlarının konu değişimi tahmini yaptığı sorgulara uygulanmıştır. Böylelikle hatalı konu değişimi tahminlerini yansıtan A tipi hataların azaltılması hedeflenmiştir. Kısaca, karakter n-gram yönteminin başarılı olduğu noktalar ile yapay sinir ağlarının başarılı olduğu noktaları birleştiren hibrit bir sistem geliştirilmeye çalışılmış ve konu değişimleri tahminlerinin iyileştirilmesi hedeflenmiştir.

Tablo XV. FAST verisinin ikinci yarısı için yapay sinir ağlarıyla bulunan konu değişim ve konu devamları sayısı

	Analiz edilen Sorgu sayısı	Konu Değişim sayısı	Konu Devamı sayısı	Doğru Tahmin Edilen değişimler	Doğru Tahmin Edilen devamlar	A Tipi Hata	B Tipi Hata	$P_{değişim}$	$R_{değişim}$	P_{devam}	R_{devam}	$F_{\beta(değişim)}$	$F_{\beta(devam)}$
YSA Sonuçları	4484	$N_{değişim}$ = 886	N_{devam} = 3598	$N_{değişim\&doğru}$ = 306	$N_{devam\&\ doğru}$ = 3594	580	4	0.345	0.987	0.998	0.861	0.583	0.907
Uzman Sonuçları	4484	$N_{gerçekdeğişim}$ = 310	$N_{gerçekdevam}$ = 4174	-----	-----	----	----	----	----	----	----	----	----

3.3. Karakter n-gram Yönteminin Yapay Sinir Ağları Sonuçlarına Uygulanması

Çalışmada yapay sinir ağlarının konu değişimi tahminleri karakter n-gram yöntemiyle güncellenmiştir. İki yöntemin birleştirilmesiyle elde edilen sonuçlar, uzman sonuçlarıyla karşılaştırılmış ve Excite ve FAST veri grupları için sırasıyla Tablo XVI ve Tablo XVII elde edilmiştir.

Genel olarak tablolar değerlendirildiğinde, farklı n-gramlar ve eşik değerleri için karakter n-gram yöntemi tahminlerinin, yapay sinir ağlarına kıyasla, her zaman daha az A tipi hata içerdiği görülmektedir. Bununla beraber B tipi hatalar *n* sayısı ve eşik değerinin büyümesiyle azalmış, ancak yapay sinir ağları yönteminin B tipi hatalarından daha büyük çıkmışlardır. Fakat daha önce de bahsedildiği gibi B tipi hatalar, yöntemlerin hatalı tahminlerinden değil, konu devamı olan sorguları yakalayamamalarından kaynaklanmaktadır.

Tablo XVI. Excite verisi için yapay sinir ağlarıyla birlikte karakter n-gram uygulamasının analizi

Ngram	Eşik değeri	$N_{değişim}$	N_{devam}	$N_{gerçekdeğişim}$	$N_{gerçekdevam}$	$N_{değişim\&doğru}$	$N_{devam\&doğru}$	A tipi hata	B tipi hata
n=1	**0,5**	114	3280	272	3122	62	3070	52	210
	0,6	187	3207	272	3122	101	3036	86	171
	0,7	280	3114	272	3122	159	3001	121	113
n=2	**0,5**	389	3005	272	3122	220	2953	169	52
	0,6	724	2670	272	3122	261	2659	463	11
	0,7	739	2655	272	3122	263	2646	476	9
n=3	**0,5**	417	2977	272	3122	233	2938	184	39
	0,6	422	2972	272	3122	234	2934	188	38
	0,7	423	2971	272	3122	235	2934	188	37
n=4	**0,5**	423	2971	272	3122	234	2933	189	38
	0,6	425	2969	272	3122	235	2932	190	37
	0,7	427	2967	272	3122	235	2930	192	37
YSA		**454**	**2940**	**272**	**3122**	**237**	**2905**	**217**	**35**

Her iki yöntemde de kullanılan veri kümelerindeki sorgular ayrıntılı olarak incelendiğinde, yapay sinir ağları yönteminin yazım yanlışlarından dolayı konu değişimi tahmini yaptığı sorgu sayısı Excite verisinde 38 adet ve FAST verisinde 71 adet olduğu görülür. Karakter n-gram yönteminin destek olarak kullanılma amacı, bu hataların yakalanmasıdır.

Tablo XVII. FAST verisi için yapay sinir ağlarıyla birlikte karakter n-gram uygulamasının analizi

Ngram	Eşik değeri	$N_{değişim}$	N_{devam}	$N_{gerçekdeğişim}$	$N_{gerçekdevam}$	$N_{değişim\&doğru}$	$N_{devam\&doğru}$	A tipi hata	B tipi hata
n=1	**0,5**	179	4305	310	4174	55	4050	124	255
	0,6	301	4183	310	4174	103	3976	198	207
	0,7	473	4011	310	4174	171	3872	302	139
n=2	**0,5**	707	3777	310	4174	277	3744	430	33
	0,6	748	3736	310	4174	289	3715	459	21
	0,7	767	3717	310	4174	295	3702	472	15
n=3	**0,5**	764	3720	310	4174	297	3707	467	13
	0,6	781	3703	310	4174	303	3696	478	7
	0,7	784	3700	310	4174	303	3693	481	7
n=4	**0,5**	790	3694	310	4174	303	3687	487	7
	0,6	794	3690	310	4174	303	3683	491	7
	0,7	795	3689	310	4174	303	3682	492	7
YSA		**886**	**3598**	**310**	**4174**	**306**	**3593**	**580**	**4**

Farklı eşik değerleri ve n-gramlar için yapılan uygulamaların özetlenmesiyle oluşturulan Tablo XVIII ve Tablo XIX incelendiğinde, karakter n-gram yöntemiyle yazım farklılıklarından kaynaklanan hatalı tahminlerin çoğunun düzeltilmiş olduğu görülür. Karakter n-gram sayesinde bu sorgular konu değişimi yerine konu devamı olarak işaretlenmiştir. A tipi hataların azaltılmasında önemli bir etken de bu hataların ortadan kaldırılabilmesidir.

Tablo XVIII. Yapay sinir ağlarının hatalı tahmin ettiği 38 sorguda doğru tahmin edilme sayısı

Eşik değeri	1-gram	2-gram	3-gram	4-gram
0,5	38	37	33	31
0,6	37	34	30	27
0,7	37	33	29	25

Tablo XIX. Yapay sinir ağlarının hatalı tahmin ettiği 71 sorguda doğru tahmin edilme sayısı

Eşik değeri	1-gram	2-gram	3-gram	4-gram
0,5	71	70	66	59
0,6	71	69	65	56
0,7	70	68	64	57

İki yöntemin birleştirilmesiyle elde edilen sonuçların performans ölçütleri incelendiğinde iyileşmelerin kaydedildiği görülür. Tablo XX incelendiğinde Excite verisi için duyarlılık ölçütleri çeşitli n-gram ve eşik değerleri için, yapay sinir ağları yöntemine göre daha yüksek bulunmuştur. Bu artış karakter n-gram yönteminin uygulanmasıyla tahminlerin kalitesinin arttığını gösterir.

Tablo XX. Excite verisi için karakter n-gram destekli yöntemin performans değerlendirmesi

Ngram	Eşik değeri	$P_{değişim}$	P_{devam}	$R_{değişim}$	R_{devam}	$F_{ß(değişim)}$	$F_{ß(devam)}$	%artış $F_{ß(değişim)}$	%artış $F_{ß(devam)}$
n=1	**0,5**	0,544	0,936	0,228	0,983	0,291	0,965	-58,335	1,480
	0,6	0,540	0,947	0,371	0,972	0,420	0,963	-39,789	1,220
	0,7	0,568	0,964	0,585	0,961	0,578	0,962	-17,130	1,162
n=2	**0,5**	0,566	0,983	0,809	0,946	0,697	0,959	-0,063	0,854
	0,6	0,360	0,996	0,960	0,852	0,593	0,900	-14,994	-5,358
	0,7	0,356	0,997	0,967	0,848	0,590	0,897	-15,414	-5,643
n=3	**0,5**	0,559	0,987	0,857	0,941	0,715	0,958	2,461	0,683
	0,6	0,555	0,987	0,860	0,940	0,714	0,957	2,318	0,606
	0,7	0,556	0,988	0,864	0,940	**0,716**	**0,957**	2,639	0,619
n=4	**0,5**	0,553	0,987	0,860	0,939	0,713	0,957	2,202	0,584
	0,6	0,553	0,988	0,864	0,939	0,715	0,957	2,406	0,574
	0,7	0,550	0,988	0,864	0,939	0,713	0,956	2,175	0,530
YSA		**0,522**	**0,988**	**0,871**	**0,930**	**0,698**	**0,951**	**0,000**	**0,000**

Excite verisi için Duyarlılık ve Anma ölçütlerinden $R_{değişim}$ değeri en yüksek 2-gram ve 0,7 eşik değerinde 0,967 olarak bulunmuştur ve bu değer önceki yöntemlerin doğru tahmin oranından daha fazladır. Bununla beraber R_{devam} değeri, *n* değeri ve eşik değeri arttıkça azalmasına rağmen konu devamı tahmini oranları her zaman yapay sinir ağları yönteminden ve dolayısıyla önceki yöntemlerden daha iyi bulunmuştur. Aynı şekilde Tablo XX incelenmeye devam edilirse, performans göstergeleri olan $F_{\beta(değişim)}$ ve $F_{\beta(devam)}$ değerlerinde de başarılı artışlar gerçekleştiği görülmektedir. Örneğin yapay sinir ağlarında $F_{ß(değişim)}$ değeri 0,698 iken bu ölçüt, 3-gram ve 0,7 eşik değeri için, karakter n-gram yöntemiyle 0,716 olarak elde edilmiş ve konu değişimi tahminlerini gösteren $F_{\beta(değişim)}$ ölçütünde, %2,639'luk bir iyileşme sağlanmıştır. Diğer bir iyileşme ölçütü olan ve konu devamı tahminlerini gösteren $F_{\beta(devam)}$, yapay sinir ağlarında 0,951 iken karakter n-gram ile 0,957 olarak bulunmuş ve %0,619'luk bir artış elde edilmiştir.

FAST veri grubu için hesaplanan performans ölçütleri Tablo XXI' de görülmektedir. Duyarlılık ve Anma değerleriyle birlikte uygunluk fonksiyonlarının da yer aldığı değerlendirmeler yapay sinir ağlarının sonuçlarıyla karşılaştırıldığında, yine yapay sinir ağlarıyla birleştirilen karakter n-gram yönteminin, önceki yapay sinir ağları sonuçlarına göre daha başarılı tahminler yaptığı gözlenmiştir. Ayrıca çoğu n-gram ve eşik değeri için uygunluk fonksiyonunun ($F_{\beta(değişim)}$ ve $F_{\beta(devam)}$) hep daha iyi olduğu görülmüş ve bu değerlerle edilen en büyük artış 3-gram ve 0,6 eşik değerleri için konu değişimi tahminlerinde %6,987 ve konu devamı tahminlerinde %1,863 olarak bulunmuştur. Ancak, konu değişimi

sorgularında doğru tahmin oranını gösteren $R_{değişim}$ değerinin en yüksek değeri 0,977 olarak bulunmasına rağmen bu oranın yapay sinir ağları yönteminin doğru tahmin oranından daha düşük olduğu tespit edilmiştir. Bununla birlikte konu devamı tahminlerindeki doğruluk oranını yansıtan R_{devam} değeri, yöntemde kullanılan bütün n-gram ve eşik değeri kombinasyonları için yapay sinir ağları tahminlerinin doğruluk oranından daha fazla bulunmuştur. Uygunluk fonksiyonlarında pozitif yöndeki artışın kaynağı, konu devamı tahminlerinin daha gerçekçi şekilde yapılmasıdır.

Tablo XXI. FAST verisi için karakter n-gram destekli yöntemin performans değerlendirmesi

Ngram	Eşik değeri	$P_{değişim}$	P_{devam}	$R_{değişim}$	R_{devam}	$F_{ß(değişim)}$	$F_{ß(devam)}$	%artış $F_{ß(değişim)}$	%artış $F_{ß(devam)}$
n=1	**0,5**	0,307	0,941	0,177	0,970	0,210	0,959	-63,947	5,703
	0,6	0,342	0,951	0,332	0,953	0,336	0,952	-42,469	4,899
	0,7	0,362	0,965	0,552	0,928	0,461	0,941	-20,966	3,743
n=2	**0,5**	0,392	0,991	0,894	0,897	0,605	0,930	3,687	2,481
	0,6	0,386	0,994	0,932	0,890	0,611	0,926	4,692	2,073
	0,7	0,385	0,996	0,952	0,887	0,615	0,925	5,292	1,895
n=3	**0,5**	0,389	0,997	0,958	0,888	0,620	**0,926**	6,253	2,005
	0,6	0,388	0,998	0,977	0,885	**0,625**	0,924	6,987	1,863
	0,7	0,386	0,998	0,977	0,885	0,623	0,924	6,742	1,808
n=4	**0,5**	0,384	0,998	0,977	0,883	0,620	0,923	6,254	1,700
	0,6	0,382	0,998	0,977	0,882	0,618	0,922	5,932	1,627
	0,7	0,381	0,998	0,977	0,882	0,618	0,922	5,852	1,609
YSA		**0,345**	**0,999**	**0,987**	**0,861**	**0,584**	**0,907**	0,000	0,000

4. DEĞERLENDİRME

Bu çalışma kapsamında, arama motorları kullanıcı oturumlarındaki konu değişikliklerini inceleyen anlam bazlı olmayan metotlar değerlendirmeye alınmış, yöntemlerin konu değişikliği tahmininde kullandıkları karar mekanizmaları detaylı bir şekilde incelenmiştir. Bu yöntemlerin, sorguların eş anlamlı kelimeler veya yazım farklılıkları barındırması halinde, konu devamlılığını tespit edemediği gözlemlenmiştir. Ardışık sorgularda eş anlamlı sözcüklerin yöntemler tarafından tespit edilemeyip sorguların konu değişimi olarak işaretlenmesi hatası anlam bazlı bir sorun olduğundan, baz alınan çalışmalarla giderilememektedir. Ancak sorgulardaki yazım farklılıklarını karakter n-gram yöntemi ile tespit etmek mümkündür. Bu aşamada iki farklı yaklaşım geliştirilmiştir: (i) karakter n-gram yönteminin Excite ve FAST verilerine uygulanması ve (ii) önceki çalışmalarda en iyi performansı sergileyen yöntem ile beraber karakter n-gram yönteminin kullanılması. Bugüne kadar yapılan anlam bazlı olmayan yeni konu tanılama çalışmaları içinde en iyi yöntemin yapay sinir ağları olduğu görülmüştür.

İlk olarak, yazım farklılıklarını kelimelerin anlamına bakmadan ayırt edebilen karakter n-gram uygulaması gerçekleştirilmiş; ancak yeni konu tanılamada anlam bazlı olmayan ve aynı veriler üzerinde aynı performans parametreleriyle çalışılmış yöntemlerden, daha kötü sonuçlar elde edilmiştir. Alternatif olarak, yapay sinir ağlarının konu değişimi tahmini yaptığı sorgulara karakter n-gram yöntemi uygulanmış ve tahminler güncellenmiştir. Bu şekilde destek olarak kullanılan karakter n-gram yöntemi ile çok daha iyi sonuçlar elde edilmiş ve çözüm önerilerine katkıda bulunulmuştur.

Bu çalışma sonunda karakter n-gram yönteminin konu değişiminin tespitinde tek başına yeterli bir yöntem olmadığı ve mevcut başarılı yöntemlerle beraber kullanıldığında, bu yöntemlerin eksik kaldığı noktaları tamamladığı görülmüştür. Bundan sonraki çalışmalarda karakter n-gram yönteminin tek başına yapılan uygulamalarındaki hatalı tahminlerin nedenleri araştırılabilir ve yeni algoritmalar geliştirilebilir, ayrıca anlam bazlı çalışmalar da yöntemlere dahil edilerek eksiklikler tamamlanabilir ve çok daha başarılı tahmin yöntemleri geliştirilebilir.

5. KAYNAKLAR

1. He, D., Goker, A., Harper, D.J., (2002) Combining evidence for automatic Web session identification*, Information Processing and Management* 38,727–742.
2. Huang, X., Peng, F., An, A., Shuurmans, D., Cercone N. (2003) Applying Machine Learning to Text Segmentation for Information Retrieval, *Information Retrieval* 6:333–362.
3. Özmutlu, S, Spink, A., and Özmutlu, H.C. (2002) Analysis of large data logs: an application of Poisson sampling to Excite Web queries. *Information Processing and Management,* 38(3), 473–490.
4. Özmutlu, H.C., Çavdur, F., Spink, A. and Özmutlu, S. (2004a). Neural network applications for automatic new topic identification on excite web search engine data logs, *Proceedings of ASIST 2004*: 67th Annual Meeting of the American Society for Information Science and Technology, Providence, RI, pp. 317-323.
5. Özmutlu, S., Özmutlu, H.C. and Spink, A. (2004b) A day in the life of Web searching: an exploratory study, *Information Processing and Management*, 40, 319-345.
6. Özmutlu, H.C. and Çavdur, F. (2005a). Application of automatic topic identification on excite web search engine data logs*, Information Processing and Management*, 41(5), 1243-1262.
7. Özmutlu, S. ve Çavdur, F., (2005b). Neural Network Applications for Automatic New Topic Identification, *Online Information Review* 29: 34-53.
8. Özmutlu, H.C., Çavdur, F. and Özmutlu, S. (2006). Automatic New Topic Identification in Search Engine Datalogs, *Internet Research: Electronic Networking Applications and Policy*, 16, 323-338.
9. Özmutlu, S., Özmutlu, H.C., Buyuk, B. (2007). Using Conditional Probabilities for Automatic New Topic Identification, *Internet Research: Electronic Networking Applications and Policy*, 37, 491-515.
10. Özmutlu, S., Özmutlu, H.C., Buyuk, B. (2008a). A Monte-Carlo simulation application for automatic new topic identification of search engine transaction logs", Simulation Modeling Practice and Theory, 16, 519-538.
11. Özmutlu, S., Özmutlu, H.C. and Cosar, G.C. (2008b). Neural Network Applications for Automatic New Topic Identification of FAST and Excite search engine transaction logs.(Yayımda)
12. Shannon, C. E. (1951) Prediction and entropy of printed English. *Bell System Technical Journal* 30:50-64.
13. Spink, A., Wolfram, D., Jansen, B.J. and Saracevic, T. (2001) Searching the Web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 53(2), 226–234.

Makale 09.01.2009 tarihinde alınmış, 08.10.2009 tarihinde düzeltilmiş, 16.10.2009 tarihinde kabul edilmiştir. İletişim Yazarı: H. C. Özmutlu (hco@uludag.edu.tr).