T.C.

ULUDAĞ ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

İNTERNET MADENCİLİĞİNDE İSTATİSTİKSEL METOTLARIN UYGULANMASI

Burcu ÇAĞLAR

YÜKSEK LİSANS TEZİ

ENDÜSTRİ MÜHENDİSLİĞİ ANABİLİM DALI

BURSA 2009

T.C.

ULUDAĞ ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

İNTERNET MADENCİLİĞİNDE İSTATİSTİKSEL METOTLARIN UYGULANMASI

Burcu ÇAĞLAR

Doç. Dr. H.Cenk ÖZMUTLU

(Danışman)

YÜKSEK LİSANS TEZİ

ENDÜSTRİ MÜHENDİSLİĞİ ANABİLİM DALI

BURSA 2009

T.C.

ULUDAĞ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

İNTERNET MADENCİLİĞİNDE İSTATİSTİKSEL METOTLARIN UYGULANMASI

Burcu ÇAĞLAR

YÜKSEK LİSANS TEZİ

ENDÜSTRİ MÜHENDİSLİĞİ ANABİLİM DALI

Bu Tez 31/07/2009 tarihinde aşağıdaki jüri tarafından oybirliği ile kabul edilmiştir.

Doç.Dr. H.Cenk ÖZMUTLU	Prof.Dr. Erdal EMEL	Yrd.Doç.Dr. Figen ERTAŞ
Danışman	Jüri	Jüri

ÖZET

İnternetin yaygın kullanılabilirliği, web sayfalarının sayısında büyük bir artışı da beraberinde getirmiştir. Benzer bilgileri barındıran web sayfalarına ulaşabilmek için kullanılan arama motorları, internet kullanıcıları için vazgeçilmez olmuştur. Arama motorlarının geliştirilmesi için kullanıcıların davranışlarının tahmin edilmesi önemli hale gelmiştir. Geliştirilen arama motorlarıyla kullanıcıların daha kısa sürede aradıkları bilgiye ulaşabilmesi sağlanabileceği gibi, kullanıcı temelli arama motorları da geliştirilebilir. Arama motoru kullanıcı davranışlarının tahmininde anlam bazlı veya anlam bazlı olmayan metotlar kullanılabilir. Bu metotlarda en önemli nokta, konu değişikliklerinin tahminidir.

Şimdiye kadar konu değişikliği tahmini için anlam bazlı olmayan pek çok istatistiksel yöntem, aynı veriler üzerine uygulanmıştır. Excite ve FAST arama motorlarından alınan verilerin kullanıldığı yöntemlerin sonuçları incelendiğinde, sorgulardaki yazım farklılıklarının yöntemlerin hatalı olarak konu değişimi tahmini yapmasına sebep olduğu gözlenmiştir.

Çalışmada yazım farklılıklarından kaynaklanan hatalı tahminlerin azaltılması hedeflenmiştir. Bu amaçla, anlam bazlı olmayan karakter n-gram yöntemi Excite ve FAST verilerine uygulanmıştır. İkinci bir çalışma olarak aynı verilere uygulanan önceki çalışmaların performans değerlendirmeleri yapılmış ve değerlendirme sonucunda bulunan en iyi yöntemin tahmin sonuçlarına karakter n-gram yöntemi uygulanarak tahmin sonuçları güncellenmiş ve her iki çalışmanın performans değerlendirmeleri yapılmıştır.

Anahtar Kelimeler: Veri madenciliği, arama motoru kullanıcı davranışları, Karakter N-gram yöntemi.

ABSTRACT

The widespread availability of Internet has brought about significant increase in the amount of web pages. Search engines that utilize to access web pages which include similar information have become indispensable for internet users. In order to enhance better search engines, determining search engine users' behaviors has become important. Due to developed search engines, users could reach information in a short time, and also user based search engines could be built. Content based or content-ignorant methodologies can be used for determining search engine users' behaviors. The most important thing in these methodologies is to identify the topic changes.

To date, many content-ignorant studies have been performed to same datasets with the aim of automatic new topic identification. Due to performance results of these content-ignorant studies which use Excite and FAST search engines' datasets, it is observed that spelling errors has caused topic shift estimates by mistake.

It is aimed to reduce wrong estimates that are based on spelling errors in this study. For this purpose, a content-ignorant methodology called character n-gram is applied to Excite and FAST datasets. In addition, previous content-ignorant applications that use same datasets are evaluated by their performance, then considering the evaluation results, the topic shift estimations are updated by implementing character n-gram method to the most successful content-ignorant method's estimation results and performance evaluation is performed for both studies.

Key Words: Data mining, web search engine users behaviors, character n-gram methodology.

İÇİNDEKİLER

TEZ ONAY SAYFASI II
ÖZET III
ABSTRACT IV
İÇİNDEKİLER V
KISALTMALAR DİZİNİ VII
ÇİZELGELER DİZİNİ VIII
ŞEKİLLER DİZİNİ X

GİRİŞ 1

1. KAYNAK ARAŞTIRMASI 4
1.1. Arama Motoru Kullanıcı Oturumlarında Konu Değişimi Tespiti için Yapılan Anlam Bazlı Çalışmalar 4
1.2. Arama Motoru Kullanıcı Oturumlarında Konu Değişimi Tespiti için Yapılan Anlam Bazlı Olmayan Çalışmalar 6

2. MATERYAL VE YÖNTEM 21
2.1. Problem Tanımı 21
2.1.1. Boolean Operatörleri ve İnternet Terimlerinden Kaynaklanan Hatalar 24
2.1.2. Kullanıcının Yazım Farklılıklarından Kaynaklanan Hatalar 25
2.2. Çözüm Önerileri 26
2.3. N-Gram Yöntemi 27
2.3.1. Genel Bilgiler 27
2.3.2. Kümeler: Denklik Sınıflarının Oluşturulması 28
2.3.2.1. Güvenilirlik ve Ayrımsama 28
2.3.2.2. N-gram Modelleri 29
2.3.2.3. N-gram Modellerinin Oluşturulması 30
2.3.3. İstatistiksel tahmin yöntemleri 31
2.3.3.1. En Büyük Olabilirlik Tahmini 32
2.3.3.2. Laplace Kuralı, Lidstone Kuralı ve Jeffreys-Perks Kuralı 36
2.3.3.3. Doğrulama Tahmin Yöntemi 40
2.3.3.4. Modellerin Geliştirilmesi ve Test Edilmesi İçin Veriler 41
2.3.3.5. Çapraz Doğrulama Yöntemi 44
2.3.3.6. Good – Turing Tahmini 45
2.3.4. Tahmin Yöntemlerinin Birleştirilmesi 50
2.3.4.1. Basit Doğrusal İnterpolasyon 51
2.3.4.2. Katz Geri Çekilme 52
2.3.4.3. Genel Doğrusal İnterpolasyon 53
2.3.4.4. Austen Verisi İçin Dil Modelleri 55

2.4. Karakter N-Gram Yöntemi 56
2.4.1. Genel Bilgiler 56
2.4.2. Yöntemin Adımları 57
2.4.3. Kullanılan Veriler 61
2.4.4. Verilerin Uzman Tarafından Değerlendirilmesi 64
2.4.5. Verilerin Temizlenmesi 64
2.4.6. Notasyon 65

3. ARAŞTIRMA SONUÇLARI ve TARTIŞMA 67
3.1. Karakter N-gram Yöntemi Sonuçları 67
3.2. İstatistiksel Yöntemin Seçimi 71
3.3. Yapay Sinir Ağları Yöntemi Sonuçları 71
3.4. Karakter n-gram Yönteminin Yapay Sinir Ağları Sonuçlarına Uygulanması 74

SONUÇ 79
KAYNAKLAR 80
EKLER 83
EK 1: Arama yapısı sınıflarının belirlenmesi için kullanılan algoritma 83
EK 2: A Tipi Hataya sebep olan sorgular arama yapısı sınıfları ve zaman aralıkları 84
EK 3: B Tipi Hataya sebep olan sorgular, arama yapısı sınıfları ve zaman aralıkları 86
ÖZGEÇMİŞ 87
TEŞEKKÜR 88

KISALTMALAR DİZİNİ

IP – Internet Protokolü

MLE – En Büyük Olabilirlik Tahmini

ELE – En Büyük Olabilirlik Tahmini

ÇİZELGELER DİZİNİ

Çizelge 1.1: Excite verilerine uygulanan yöntemlerin analiz sonuçları 19
Çizelge 1.2: FAST verilerine uygulanan yöntemlerin analiz sonuçları 19
Çizelge 2.1: N-gram modelleri için parametrelerdeki artış 30
Çizelge 2.2: İstatistiksel Tahmin Yöntemlerinin notasyonu 32
Çizelge 2.3: Persuasion verisindeki bir tümcenin kelimeleri için olasılıklar 35
Çizelge 2.4: Chuch ve Gale(1991) çalışma verisi için tahmin edilen sıklıklar 37
Çizelge 2.5: *was* kelimesini takip eden kelimeler için ELE yöntemi tahminleri 39
Çizelge 2.6: İki sistemin karşılaştırılmasında *t* testinin kullanımı 43
Çizelge 2.7: Austen yığınında 2-gram ve 3-gramlar için frekans dağılımı sıklıkları 48
Çizelge 2.8: 2-gramlar için Good – Turing tahminleri, düzeltilmiş sıklıklar ve olasılıklar 48
Çizelge 2.9: Persuasion verisindeki cümle için Good – Turing 2-gram sıklık tahminleri 49
Çizelge 2.10: Persuasion verisinde test edilen Good – Turing tahminleriyle oluşturulan geri çekilme modelleri 55
Çizelge 2.11: Çeşitli dil modellerine göre test cümleciği olasılık tahminleri 55
Çizelge 2.12: Kelimelerin karakterlerine ayrılması 58
Çizelge 2.13: Ardışık iki sorgunun karşılaştırılması 59
Çizelge 2.14: Ardışık iki sorgunun karşılaştırılması 60
Çizelge 2.15: Sorgu bilgilerinden elde edilen arama yapısı ve zaman aralığı sınıfları 62
Çizelge 2.16: Çalışmada kullanılan verilerin büyüklüğü 63
Çizelge 2.17: Değerlendirilen veri büyüklükleri 64
Çizelge 3.1: Excite verilerinde farklı *n* ve eşik değerleri için Karakter n-gram uygulaması sonuçları 67
Çizelge 3.2: Excite verilerinde farklı n ve eşik değerleri için Karakter n-gram uygulamasının performans analizi 68
Çizelge 3.3: FAST verilerinde farklı n ve eşik değerleri için Karakter n-gram uygulaması sonuçları 69
Çizelge 3.4: FAST verilerinde farklı n ve eşik değerleri için Karakter n-gram performans değerlendirmesi 70
Çizelge 3.5: Excite verilerine uygulanan yöntemlerin analiz sonuçları 71
Çizelge 3.6: FAST verilerine uygulanan yöntemlerin analiz sonuçları 72
Çizelge 3.7: Excite verisinin ikinci yarısı için Yapay sinir ağlarıyla bulunan konu değişim ve konu devamları sayısı 73
Çizelge 3.8: FAST verisinin ikinci yarısı için Yapay sinir ağlarıyla bulunan konu değişim ve konu devamları sayısı 74
Çizelge 3.9: Excite verisi için Yapay Sinir Ağlarıyla birlikte karakter n-gram uygulamasının analizi 75
Çizelge 3.10: FAST verisi için Yapay Sinir Ağlarıyla birlikte karakter n-gram uygulamasının analizi 75
Çizelge 3.11: Yapay sinir ağlarının hatalı tahmin ettiği 38 sorguda doğru

tahmin edilme sayısı 76
Çizelge 3.12: Yapay sinir ağlarının hatalı tahmin ettiği 71 sorguda doğru tahmin edilme sayısı 76
Çizelge 3.13: Excite verisi için Karakter n-gram destekli yöntemin performans değerlendirmesi..... 77
Çizelge 3.14: FAST verisi için Karakter n-gram destekli yöntemin performans değerlendirmesi..... 78

ŞEKİLLER DİZİNİ

Şekil 2.1: Bir kelimeden fazla sorguların karşılaştırılma yöntemi 61

GİRİŞ

Günümüzde, arama motorları kullanım oranındaki hızlı artışla beraber internet kullanıcı davranışlarının tahmini önemli hale gelmiştir. Arama motorları kullanıcı davranışlarını tahmin ederek, istenen bilgiyi en kısa sürede doğrulukla kullanıcılara sunabilirler. Bununla beraber arama motorlarının sayısındaki artış, sunulan hizmetin kalitesinin arttırılmasını da gerektirmektedir. 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.

Arama motoru kullanıcı davranışlarının tahmininde anlam bazlı veya anlam bazlı olmayan metotlar kullanılabilir. Bu metotların temelinde, konu değişikliklerinin tahminiyle kullanıcı davranışlarını tespit etme vardır. Bir kullanıcı oturumu içinde, konu değişiklikleri tespit edilebilirse arama motorlarının kişiselleştirilmesi yönünde önemli adımlar atılabilir. Örneğin, kullanıcının aynı konuya devam ettiği gözlenirse önceki sonuçlara benzer yeni sonuçlar ekrana getirilebilir veya yeni konuya geçtiği gözlenirse önceki sonuçlar elenerek bunların dışında kalan sonuçlar kullanıcıya sunulur. Böylelikle arama motorlarının performansı arttırılabilir.

Bu güne kadar konu değişimi tespiti için pek çok anlam bazlı olmayan istatistiksel yöntem geliştirilmiştir. Bu doğrultuda 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 çalışmada 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 (Bu sınıflar ile ilgili detaylı bilgi ilerleyen bölümlerde verilecektir). 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.

Önceki 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 tahmini konu değişikliği olacaktır. Bu sorunu aşmanın, anlama bakmadan mümkün olmadığı tespit edilmiştir. İkinci bulgu ise eğer kullanıcılar ardışık sorgularında kullandıkları ifadelerde yazım hatası yaparlarsa, anlam bazlı olmayan yöntemler, bu yazım farklılıklarını tespit edemedikleri için, konu değişikliği tahmini yapmaktadırlar. 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. Yazım farklılıklarından dolayı hatalı tahmin edilen ardışık sorgular incelendiğinde, yazım hatası bulunan ifadelerin bir kısmının değişmeden kaldığı gözlenmiş ve ardışık benzerliklere dayalı bir tahmin yöntemi ile bu kısımların belirlenip, konu değişimi tahminlerinde kullanılabileceği tespit edilmiştir. Yapılan araştırmalar sonucunda N-gram yönteminin ardışık benzerliklere dayalı bir yöntem olduğu ve yazım hatalarını tespit edebileceği görülmüştür. Ancak N-gram yöntemi kelime bazında incelemeler yapar ve büyük boyutlu metinlere uygulanır. Metin içinde kullanılan her kelimenin olasılıkları hesaplanarak bu sonuçlar takip eden kelimelerin görülme olasılıklarına yansıtılır. Arama motoru sorgularında ise farklı bir durum söz konusudur. Sorgulardaki yazım hataları ancak kelimeler karakter boyutunda incelendiğinde yakalanabilir. İlave 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 tahmin edilmeye çalışılmıştır.

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. 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.

1. KAYNAK ARAŞTIRMASI

İnsanlar günlük hayatlarında bilgi erişimlerini, arama motorlarını kullanarak gerçekleştirirler. Dolayısıyla arama motoru kullanıcı davranışları incelenerek yapılan tahminlerle, arama motorları kişiselleştirilebilir. Arama motoru kullanıcı davranışlarını belirleyebilmenin bir yolu da konu değişimlerinin tahminidir. 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.

1.1. Arama Motoru Kullanıcı Oturumlarında Konu Değişimi Tespiti için Yapılan Anlam Bazlı Çalışmalar

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 tespiti için kullanılan anlam bazlı çalışmaların sayısı kısıtlı olup, daha çok doğal dil tanılama, sınıflandırma ve kümelemede kullanılmaktadırlar.

Spink ve arkadaşları (2001) Excite arama motorundan alınan bir milyondan fazla sorguyu inceleyerek arama motoru kullanıcı davranışlarını tahmin etmeye çalışmışlardır. Genel olarak, çoğu kullanıcının az kelimeyle arama yaptığını, bazı kelimelerin diğerlerine göre daha fazla sıklıkla kullanıldığını ve sorguların çoğunlukla kendine özgü olduğu gibi sonuçlar elde etmişlerdir. Bu sonuçlara anlam bazlı ve anlam bazlı olmayan istatistiksel yöntemler kullanarak ulaşmışlardır. Excite aramaları, kullanıcıların gerçek sorgularından oluşmaktadır. Sorgulardaki lojik ifadeler ihmal edilmiş ve terimler arası ilişkiyi bulmak için çevrimiçi eş anlamlılar sözlüğü ile kavramlar arası bağlantı kurma metodu olan akıllı kavram ilişkilendirme kullanılmıştır. Bu çalışmada terimler, sorgular, oturumlar ve sonuç sayfaları analiz edilmiştir. Bir oturum, aynı kullanıcı tarafından arama motoruna girilen sorgulardan oluşmakta ve sorgular da birbirinden bağımsız terimlerden oluşmaktadır. Böylece, terim düzeyinde sorgular incelenerek bir oturum hakkında yorum

yapılabilmektedir. Eldeki verilerden toplam kullanıcı sayısı, sorgu sayısı, tekrar eden sorgu sayısı, tekil sorgu sayısı ve bunların ortalaması ile medyanları bulunarak istatistiksel analizler uygulanmış ve kullanıcı başına düşen sorgu sayısı, sorgu başına kullanılan terim sayısı, terimlerin dağılımı ve terimlerin kullanım oranı gibi veriler elde edilmiştir. Bunların yanında anlam bazlı yöntemlerle, kullanıcıların benzer sayfaları kullanma oranı belirlenmiştir. Ayrıca sorgular sınıflandırılmış ve sınıfların dağılımı hesaplanmıştır. Çalışmada sorgular 11 sınıfta toplanmıştır. Çalışma sonuçlarına göre, yapılan aramaların %16,9'u eğlence sınıfına ve %16,8'i cinsel içerik sınıfına dâhildir. Sorguların %10'a yakını sağlık ve bilimle alakalıdır. Sonuç olarak yapılan aramaların çoğu basit içerikli, az kelimelidir. Çok az sorgu karmaşık yapıya sahiptir ve arama motoru kullanıcılarının çoğu ilk ve ikinci sonuç sayfasından öteye gitmemekle beraber benzer sayfaları kullanma oranı düşüktür. Ayrıca az sayıda terimin yüksek sıklıkla kullanıldığı, sorgu dilinin çok zengin olduğu ve oturumların çoğunlukla benzersiz sorgulardan oluştuğu gözlenmiştir.

Kişiselleştirilmiş arama motorları geliştirmek için yapılan çalışmalardan biri de Leung ve arkadaşlarına (2008) aittir. Popüler ve gelişmiş arama motorları, kullanıcı sorgularını geliştirmek ve kullanıcıların daha etkin arama yapmalarına yardım etmek için yapılan sorgu sonrasında kullanıcıya, arama konusuyla ilgili diğer arama seçeneklerini öneri olarak vermektedir. Bu öneriler kullanışlı olsa da arama motorları aynı sorgu için, kullanıcının ilgi alanlarını göz ardı ederek, her kullanıcıya benzer öneriler sunmaktadır. Yazarlar çalışmalarında, kişiselleştirilmiş kavramsal temelli kümeleme tekniklerini baz alan, kullanıcılara kişiselleştirilmiş sorgu önerileri veren bir metot geliştirmişlerdir. Kullanıcının arama sonuçları sonrasında yaptığı tıklamalar, konuyla ilişkili internet sayfalarının özet, URL veya başlık gibi bilgilerini görüp bunlarla ilgilenmesi üzerine gerçekleşir. Tıklama bilgileri çok kolay bir şekilde toplanıp incelenebilmekte ve kullanıcının kavramsal tercihleri hakkında bilgiler elde edilebilmektedir. Bu sebeple yazarlar, her kullanıcıya aynı önerileri veren yöntemlerin aksine, kullanıcıların tercihlerini tahmin etmek ve tercihlerine göre öneriler vermek için tıklama verisinden yararlanmışlardır. Ayrıca anlamsal olarak birbirine yakın sorguları bulabilmek için kullanıcı tercihleri, kavram bazlı kümeleme algoritmasının girdisi olabilmektedir. Araştırmacılar Beeferman ve Berger'in (2000) geliştirdiği grafik-bazlı kümeleme algoritmasının kapsamını genişleterek sorgular, kullanıcılar ve dokümanlar arası ilişkileri tanımlamışlardır. Tıklama verisinin toplanması için kullanıcının tıklama izini takip

eden bir Google arayüzü geliştirmişlerdir. Kullanıcı arama sonuçlarını her tıkladığında arayüz bunları kayıt altına almakta ve ilave her tercihte kayıtların bulunduğu veritabanı güncellenmektedir. Kişiselleştirilmiş kavram-bazlı kümeleme algoritması ile toplanan veriler kullanılarak anlamsal açıdan birbirine en yakın sorgular bulunmaya çalışılmıştır. Çalışma sonuçları, geliştirilen metodun kullanıcının kavramsal ihtiyaçlarını dikkate alarak kişiselleştirilmiş sorgu önerilerini başarılı bir şekilde verebileceğini göstermektedir. Ayrıca geliştirilen kümeleme algoritmasının, önceki algoritmalarla karşılaştırılması ve çeşitli parametreleri baz alınarak, diğerlerinden daha iyi sonuçlar verdiği gözlemlenmiştir.

1.2. Arama Motoru Kullanıcı Oturumlarında Konu Değişimi Tespiti için Yapılan Anlam Bazlı Olmayan Çalışmalar

Anlam bazlı çalışmalar, sözcüklerin anlamına dayandığı için, gerçek uygulamalarda karmaşık olmalarının yanında yüksek maliyetli ve zahmetlidirler. Bu çalışmalardan başarılı sonuçlar elde edilse de, dezavantajlarından dolayı araştırmacılar, gerçek zamanlı uygulamalarda anlam bazlı olmayan istatistiksel metotları tercih etmektedirler. Bu 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.

İnternet kullanıcıları için arama motorları, bilgiye erişim yollarının başında gelir. Bu yüzden internet kullanıcılarının arama davranışlarının incelenmesinde, arama motoru sorgularının analizi kritik öneme sahiptir. Özmutlu ve arkadaşları (2002), Web arama motorlarında kayıt altına alınan sorguların incelenerek, arama motoru kullanıcı davranışlarının analiz edilebileceğini öngörmüşlerdir. Ancak arama motorlarında her gün milyonlarca sorgu kaydedilmekte ve bu sorguların tamamının analiz edilmesi, uğraştırıcı olabilmektedir. Bu sebeple yazarlar, büyük bir veri kümesinin istatistiksel karakteristiğini yansıtacak şekilde örneklem alınması yoluna gidilerek, analizlerin daha etkili yapılabileceği sonucuna varmışlardır. Yapılan çalışmada, Excite arama motoru sorgularından yararlanılmıştır. Bu verilere, Poisson örnekleme yöntemi ve sistematik örnekleme yöntemi uygulanarak, seçilen örneklemlerin ana kütleyi gerçekçi bir şekilde yansıtıp yansıtmadığı incelenmiştir. Yazarlar sistematik örnekleme yönteminin, sorgulardaki düzensiz dağılımı

yansıtamadığını tespit etmişler ve Poisson örneklemiyle alınan verilerin ana kütleyi daha etkin şekilde yansıttığı sonucuna varmışlardır. Ayrıca yazarlar, Poisson örnekleme yönteminin, süreçleriyle ilgili verileri veritabanlarında depolayan şirketler için kullanılabileceğini, toplanan verilerle etkili analizlerin yapılabileceğini ve yaptıkları çalışmaların geliştirilerek, internet kullanıcılarının arama davranışlarının daha iyi anlaşılabileceğini ve daha etkili arama motorları tasarlanmasına katkıda bulunabileceği sonucuna varmışlardır.

Arama motorlarının yanında, çevrimiçi kütüphaneler ve diğer bilgi erişim sistemleri de, insanların vazgeçilmez bilgi edinme kaynakları haline gelmiş, dolayısıyla bu sistemlerin geliştirilebilmesi için, kullanıcıların arama davranışları sırasında bilgi erişim teknolojileriyle etkileşimlerini öğrenme ihtiyacı ortaya çıkmıştır. Spink ve arkadaşları (2002), insanların aynı oturum içinde birden fazla farklı konuda arama yapabileceğini öngörmüş ve çoklu konularda arama yapma eğilimini belirlemeye çalışmışlardır. Yazarlara göre, bir arama oturumu sırasında bazı kullanıcılar birden fazla konu ile ilgilenebilirler. Çoklu görev yürütümü (multitasking), aynı anda birden fazla görevin yürütümüdür. Diğer bir deyişle, bilgi arama ve araştırma süreçlerinde birden fazla, muhtemelen değişen, bilişsel, etken ve koşulsal durumlar da içeren bilgi problemleriyle ilgili, zaman içindeki aramalardır. Dolayısıyla bu çalışmanın amacı, farklı bilgi ortamlarındaki farklı çalışmalarla, çoklu görev yürütümünün belirlenmesi, çoklu arama süreçlerinin karakteristiklerinin analizi, çoklu arama yapılan oturumların tekil arama yapılan oturumlarla karşılaştırılması ve kullanıcı oturumlarında konu değişimini gösterebilecek faktörlerin belirlenmesi olarak özetlenebilir. İnsanların bilgiye erişim sürecindeki davranışlarını tespit edebilmek için yazarlar, birbirinden bağımsız dört farklı çalışma ile farklı bilgi erişim ortamlarında, kullanıcıların davranışlarını inceleyebilmek için veriler toplamışlardır. İlk çalışmada, Excite arama motoru ana sayfasına bir anket yerleştirilmiş ve isteğe bağlı olarak kullanıcılardan anketi cevaplamaları istenmiştir. Sorular, kullanıcıların mevcut araştırma konusu, arama terimleri ve bilgi arama aşamalarıyla ilgilidir. Toplamda 480 kişi soruları cevaplamış (Excite ziyaretçilerinin %7,7'si) ve bu cevaplar, ilk çalışmanın verilerini oluşturmuştur. İkinci çalışmada 20 Aralık 1999 tarihinde rassal olarak seçilen, 10,016 sorgudan oluşan 1000 Excite oturumu veri olarak kullanılmıştır. Kullanıcı ID'si (Personal identification) ile yeni kullanıcı oturumları belirlenmiş ve her oturumdaki sorguların zamanları tutulmuştur. Çoklu

arama yapanlar manuel olarak belirlenmiş ve 1000 kullanıcı arasında 114 kullanıcının 1823 sorguyu içeren çoklu arama yaptığı tespit edilmiştir. Üçüncü çalışmada, bir çevrimiçi veritabanı olan DIALOG arama servisini kullanan 198 kişinin arama bilgileri toplanmıştır. Bu bilgiler arasında arama sorgusu, hazırlanan anket sorularına verilen cevaplar, ulaşılan metinler ve birbirleriyle ilişkileri bulunmaktadır. Kullanıcılar disiplinlerine göre sınıflanmış (sosyal bilimler, tıp, mühendislik, fen bilimleri gibi) ve hangi disiplinlerde yoğun olarak araştırma yapıldığı tespit edilmiştir. Dördüncü çalışmada, kütüphane kullanıcılarının arama yöntemleri incelenmiştir. Kütüphane kullanıcılarına isteğe bağlı olarak, çalışmakta oldukları konu başlığı, her başlık için ulaştıkları metin sayısı gibi sorular sorulmuş ve alınan cevaplar veri olarak kullanılmıştır. Toplanan verileri birbirinden bağımsız olarak inceleyen yazarlar, dört çalışmadaki çoklu görev yürütümü yaygınlığının değişken olduğunu ve bu değişkenliğin nedenlerinin belirlenmesinin zor olduğunu belirtmişlerdir. Ancak yazarlar, farklı bilgi erişim ortamlarında, insanların genelde birden fazla bilgi problemiyle çalıştığı sonucuna varmışlardır. Bu çalışma, internet kullanıcılarının bilgi araştırma süreçlerinde çoklu arama kavramına temel oluşturmuştur. Yazarlar ayrıca, çalışmalarla direkt olarak test etmeseler de insanların içinde yaşadıkları karmaşık çalışma ortamlarında gün içinde birden fazla işi aynı anda yapmak zorunda kaldıkları gibi bilgi arama konusunda da birbiriyle alakalı veya alakasız başlıkların araştırmasını da aynı anda yapmak zorunda kaldıklarını ve bu durumun olağan olduğunu vurgulamışlardır. Bu çalışma ile insanların bilgi arama sırasındaki davranışlarına ışık tutulmuş ve ileride gerçekleştirilecek kullanıcı oturumlarının tanımlanması çalışmalarına, temel bir çerçeve oluşturulmuştur.

He ve arkadaşları (2002), arama motoru kullanıcı oturumlarındaki konu değişikliklerini tespit etmede, kullanıcıların IP adreslerinden, arama sürelerinden ve sorgularının bulunduğu kullanıcı kayıtlarından yararlanabileceklerini öngörmüşlerdir. Yazarlar bu verilerden yola çıkarak, olasılık (delil) birleştirme yaklaşımı ile arama motoru kullanıcılarının, arama konularını değiştirmeleri halinde, bunu otomatik olarak tespit edecek bir yaklaşım geliştirmişlerdir. Öncelikle, olasılık birleştirme yönteminde kullanılmak üzere, arama süreleri yardımıyla belirli bir kayda ait zaman aralıkları (time interval - *ti*) ve sorgular yardımıyla da arama yapısı sınıfları (search pattern - *sp*) oluşturulmuştur. Daha sonra Reuters arama motoru verilerinin, zaman aralıklarına göre ve arama yapısı sınıflarına göre, ayrı ayrı konu değişikliği olasılıkları hesaplanmıştır. Hesaplanan olasılıklar, belirsizlik altında

modellemede başarılı olan Dempster-Shafer teorisi yöntemiyle birleştirilmiş ve her kayıt için tek bir olasılık değeri elde edilmiştir. Bir eşik değeri yardımıyla, olasılıklar ikili değişkenlere dönüştürülmüş ve her kayıt için "konu değişikliği yok" ve "konu değişikliği var" şeklinde ikili değişkenlerin atamaları yapılmıştır. Elde edilen atamalar, uzman tarafından yapılan atamalarla karşılaştırılmış ve yaklaşımın performansını değerlendirmek için, Duyarlılık ve Anma performans ölçütlerini birlikte dikkate alan tek bir performans ölçütü kullanılmıştır. Bu performans ölçütünün en büyük değerini alabilmesinde, eşik değeri ve ağırlıklardan oluşan parametrelerin belirleyici olduğu tespit edilmiştir. Bu yüzden yazarlar, incelenen veri örneğindeki kayıtların yaklaşık yarısını, bu parametrelerin en uygun değerlerinin belirlenmesi için kullanmışlardır. Bu amaçla, bir genetik algoritma kullanılarak, en uygun parametre değerleri belirlenmiş ve örnek verisinin ikinci kısmında bu parametre değerleri kullanılarak yaklaşım test edilmiştir. Yazarlar, yaklaşımlarının arama motorları kullanıcı oturumlarındaki konu değişikliklerini belirlemede başarılı olduğunu belirtmişlerdir.

Kullanıcı oturumlarını anlam bazlı olmayan yöntemlerle inceleyen başka bir çalışma, Spink ve arkadaşları (2004) tarafından gerçekleştirilmiştir. Kullanıcı davranışlarının incelenmesinde kullanıcıların ilgi alanları baz alınacak olursa, cinsellikle ilgili konuların, ilgi alanları sıralamasının üst kısımlarında yer aldığı görülür. Bu sebeple yazarlar, internet kullanıcılarının bu konudaki arama davranışlarını inceleme gereği duymuşlardır. Ayrıca yazarlar, diğer konularla karşılaştırma yapabilmek için, cinsellikle ilgili sorguların oranını belirlemenin yanında, cinsellikle ilgili olan ve ilgili olmayan sorguların karakteristiklerini de incelemişlerdir. Çalışmada, Excite arama motoru verilerinden yararlanılmış ve cinsellikle ilgili aramalarda kullanılan sözcük sayısının, diğer aramalara kıyasla daha sınırlı olduğu tespit edilmiştir. Bu sorgularda genelde benzer terimler bulunmakta ve çoğunlukla cinsellikle ilgili terimler tekrarlanmaktadır. Ayrıca bu sorgular, cinsellikle ilgili olmayan sorgulara göre daha uzun olmakla beraber, cinsellikle ilgili bir oturumda büyük bir olasılıkla sorgu adedi 20'den fazla olmaktadır. Ek olarak, cinsellikle ilgili arama yapanlar, diğerlerine göre daha fazla sayıda sayfa görüntülemektedirler ve bu oturumlar görüntü indirmenin zaman alması sebebiyle daha uzun sürebilmektedir. Yazarlar, genel olarak cinsellikle ilgili arama yapanların, diğerlerine göre daha fazla zaman ve çaba harcamaya istekli olduklarını vurgulamışlar ve yaptıkları çalışmanın tek bir arama motoru verisine dayalı olmasından dolayı sınırlı olduğunu da belirtmişlerdir.

Arama motoru kullanıcılarının ilgi alanlarının belirlenmesi, davranışların anlaşılmasında etkili olabileceği gibi davranışların gün içindeki değişiminin incelenmesi de önemli verilerin elde edilmesini sağlayabilir. Bu görüşle yola çıkan Özmutlu ve arkadaşları (2004), Excite ve FAST arama motorlarından alınan verileri inceleyerek kullanıcı davranışlarının gün içindeki değişimini belirlemeye çalışmışlardır. Yazarlar, arama motoru kullanıcılarının sorgularında kullandıkları terim sayısı ve sorguların yeniden düzenlenmesi gibi sorgu karakteristiklerinin gün içinde aynı kaldığını, fakat kullanıcı sayısının ve sorgu sayısının sabah saatlerinde en yüksek düzeyde olduğunu ve bu düzeyin zaman geçtikçe azaldığını tespit etmişler ve Markov analizini sorgulara uygulayarak, günün saatleriyle sorgular arasında spesifik bir eğilim olmadığını gözlemlemişlerdir. Ayrıca yazarlar, çalışmadan elde edilen bulguların, arama motorlarının arama yapılarını geliştirmede kullanışlı olabileceğini vurgulamışlardır.

Kullanıcı davranışlarının anlaşılabilmesinde sorguların içerik veya gün içindeki arama yapılarının yanı sıra oturumlardaki konu değişimlerinin doğru bir şekilde belirlenebilmesi de oldukça önemlidir. Konu değişim tahminleri yapan yöntemlerde aranan en önemli özellik ise, yöntemin farklı veriler üzerinde de başarılı sonuçlar verebilmesidir. Bu amaçla Özmutlu ve Çavdur (2005a), daha önce He ve arkadaşları (2002) tarafından Reuters arama motoru verileri üzerinde yeni konu tanılamada kullanılan olasılık (delil) birleştirme yaklaşımını, yöntemin özelliklerini belirlemek ve geçerliliğini test etmek için kendi çalışmalarında kullanmışlardır. Daha önce de bahsedildiği gibi bu yaklaşım, arama motoru kayıtlarından hesaplanan olasılıkları, Dempster-Shafer teorisini kullanarak birleştirmektedir. Özmutlu ve Çavdur (2005a) çalışmalarında Excite arama motoru verilerini kullanarak, He ve arkadaşları (2002) tarafından geliştirilen yöntemin performansının veri kümesine bağlı olup olmadığını ve parametreler ile girdilerin algoritma performansında ne derece etkili olduğunu belirlemeye çalışmışlardır. Excite arama motorundan alınan verilerde Reuters arama motoru verilerinde olduğu gibi, kullanıcıların IP adresi, arama zamanı ve sorguları vardır. Bu verilerin karar verme aşamasında kullanılabilmesi için düzenlenmesi gerekmektedir. Bu amaçla veri kümesindeki sorgularda iki ardışık sorgu arasındaki arama yapısı (search pattern - *sp*) belirlenmiştir. Arama yapıları; yeni (new), sonraki sayfa (next page), genelleştirme (generalization), özelleştirme (specialization), düzenleme (reformulation), ilgili geri-besleme (relevance feedback) ve diğer (others) olmak üzere yedi sınıfta gruplandırılmıştır. Ayrıca aynı oturum içindeki her ardışık sorgu arasında geçen süre olarak

tanımlanan zaman aralıkları (time interval - *ti*) yedi grup olarak, 0-5 dk, 5-10 dk, 10-15 dk, 15-20 dk, 20-25 dk, 25-30 dk ve 30+ dk şeklinde ayrılmıştır. Her sorgunun arama yapısı PASCAL'da yazılan bir bilgisayar programı tarafından otomatik olarak belirlenmiştir (Programın algoritması EK 1'de verilmiştir). Zaman aralıkları da belirlenen sorguların konu değişim ve konu devamı olup olmadıkları uzman tarafından işaretlenmiştir. Bu aşamalardan sonra sorgular, yöntem tarafından kullanıma hazır hale gelmiş, verilerin hazırlık aşaması tamamlanmıştır. Daha sonra Dempster-Shafer teorisinin kullanacağı parametreler belirlenmiş ve parametrelerdeki çeşitlilik, geliştirilen bir genetik algoritmayla sağlanmıştır. Parametrelerin ve veri grubunun sonuçlar üzerindeki etkisini değerlendirebilmek için altı farklı senaryo geliştirilmiş ve hepsi için sonuçlar elde edilmiştir. Yazarlar bu sonuçlara göre, veri grubunun geliştirilen algoritma üzerinde etkili olduğu sonucuna varmışlardır. Ayrıca çeşitli parametreler için algoritmanın başarılı olduğunu ve parametrelerin performans üzerindeki etkisinin zayıf olduğunu vurgulamışlardır.

Daha önce de bahsedildiği gibi, arama motorları kullanıcı oturumlarındaki konu değişimlerini belirlemek, kullanıcı davranışlarını doğru bir şekilde yorumlayabilmek için gereklidir. Bu sebeple, anlam bazlı olmayan yöntemlerle konu değişimlerini belirlemek için birçok istatistiksel yönteme başvurulmuştur. Bu çalışmalardan biri, Özmutlu ve Çavdur (2005b) tarafından gerçekleştirilmiştir. Yazarlar çalışmalarında, sorguların istatistiksel karakteristikleri olan zaman aralıkları ve arama yapılarından yararlanarak, yapay sinir ağları yöntemi ile kullanıcı oturumlarındaki konu değişimlerini otomatik olarak belirlemişlerdir. Önceki çalışmalarda olduğu gibi veri kümesinden örneklem alınıp, uzman tarafından konu değişimleri belirlenerek hazırlık aşaması tamamlanmıştır (Özmutlu ve Çavdur 2005a). Yazarlar, yapay sinir ağlarının modellenmesinde 1 girdi katmanı, 1 gizli katman ve 1 çıktı katmanından yararlanmışlardır. Girdi katmanında arama yapısı sınıfları ve zaman aralıkları için 2 nöron vardır. Her nöron, arama yapısı sınıflarını ve zaman aralıklarını yansıtacak şekilde 1 – 7 arasında bir değer almaktadır. Çıktı katmanında yer alan bir adet nöron, konu değişimi ve konu devamı şeklinde yorumlanan 1 veya 2 değerini vermektedir. Gizli katmanda bulunan beş adet nöron çeşitli deneyler sonucunda belirlenmiştir. Eğitim ve test kümesi olarak ayrılan verilerden eğitim kümesiyle yapay sinir ağı eğitilmiş ve eğitilmiş sinir ağı, test kümesi üzerinde çalıştırılmıştır. Daha sonra yapay sinir ağlarının konu değişimi veya konu devamı tahminleriyle uzmanın bulduğu sonuçlar karşılaştırılmış ve yöntemin

sonuçları değerlendirilmiştir. Değerlendirmede yapay sinir ağları yönteminin, konu değişimlerinin %98,4'ünü ve konu devamı olan sorguların %86,6'sını doğru tahmin ettiği görülmüştür. Bununla beraber, yapay sinir ağları 865 konu değişimi tahmini yapmış; fakat gerçekte konu değişimlerinin sayısı 310 olarak belirlenmiştir. Bir diğer ifadeyle yöntem, konu değişimlerinde aşırı tahminlerde bulunmuştur. Yazarlar aşırı tahminin, yapay sinir ağlarında kullanılan eşik değerinden kaynaklandığını tespit etmişler ve ilerleyen çalışmalarda eşik değerinin değişimlerini araştıracaklarını belirtmişlerdir. Yazarlar bu çalışmayla, yapay sinir ağlarının yeni konu tanılamada başarılı bir yöntem olduğu sonucuna varmışlar ve ileride daha güçlü, iyileştirilmiş yapay sinir ağları ile daha başarılı tahminler yapılabileceğini vurgulamışlardır.

Yeni konu tanılama amacıyla geliştirilen ve performans analizleriyle başarılı olarak nitelendirilen bir yöntemin, veri kümesinden bağımsız olması, bir başka ifadeyle farklı veri gruplarında da başarılı tahminler yapması beklenir. Bu amaçla Özmutlu ve arkadaşları (2008a) çalışmalarında, yeni konu tanılama amacıyla geliştirdikleri yapay sinir ağları yönteminin (Özmutlu ve Çavdur 2005b) farklı veri gruplarında da başarılı performans gösterip göstermediğini araştırmışlardır. Yazarlar, Excite ve FAST arama motoru verilerini kullanarak, önceki çalışmalarda olduğu gibi hazırlık aşamasını tamamlamışlar (Özmutlu ve Çavdur 2005a) ve her iki veri grubunu eğitim ve test kümesi olarak ikiye ayırmışlardır. Daha sonra 6 deney tasarlamışlar ve bu deneyleri 2 hipotez grubu altında toplamışlardır. İçeriğinde iki farklı deney bulunan 1. grup hipotezlerinde yapay sinir ağları, Excite test kümesiyle test edilmiş ve 1. deneyde Excite eğitim grubuyla eğitilirken 2. deneyde FAST eğitim grubuyla eğitilmiştir. Hipotez grubunda test kümesi aynı fakat eğitim kümeleri farklıdır, dolayısıyla iki çıktı birbirinden bağımsız değildir, sonuç olarak eşleştirilmiş iki grup arasındaki farklılıkların incelenmesine yönelik olan *t* testi (paired t test) bu senaryo için uygundur. *T* testi, iki sonuç arasındaki farkın istatistiksel olarak sıfırdan farklı olup olmadığını araştırır; eğer sonuçlar arası fark istatistiksel olarak sıfırdan farklıysa yapay sinir ağlarının performansı eğitim kümesine bağlı olarak değişir, aksi durumda, yapay sinir ağlarının performansının eğitim kümesine bağlı olmadığı sonucuna varılabilir. 2. grup hipotezlerinde yine iki farklı deney mevcuttur; 1. deney grubu Excite eğitim kümesiyle eğitilmiş, 2. deney grubu FAST eğitim verisiyle eğitilmiş ve her iki deney grubu da FAST test kümesiyle test edilmiştir. 1. grup hipotezlerde olduğu gibi 2. grup hipotezlerde de

deneylerin istatistiksel olarak birbirlerinden farklı sonuçlar üretip üretmediklerini belirlemek için t testi uygulanmıştır. Son olarak, Excite eğitim kümesiyle eğitilmiş yapay sinir ağları Excite test kümesiyle ve FAST eğitim kümesiyle eğitilmiş yapay sinir ağları FAST test kümesiyle test edilerek son iki deney de tamamlanmıştır. T testi sonuçları incelendiğinde, hipotez gruplarında bulunan deneylerin tahmin sonuçları arasındaki farkın, istatistiksel olarak sıfırdan farklı olmadığı hipotezi reddedilememiştir. Dolayısıyla hipotez gruplarındaki deneylerin sonuçlarının, eğitim kümesine bağlı olarak değişmediği sonucuna varılmıştır. Bir başka deyişle, yapay sinir ağları yönteminin tahminlerdeki başarısı, kullanılan eğitim kümesinden bağımsızdır. Fakat bu çalışmada da yöntem konu değişimleri tahminlerini mevcut konu değişimlerinden fazla yapmıştır. Yazarlar bu durumun, daha güçlü ve başarılı yapay sinir ağlarının geliştirilmesiyle düzeltilebileceğini vurgulamışlardır. Sonuç olarak, herhangi bir eğitim kümesiyle eğitilen yapay sinir ağlarının, farklı arama motorları verileri üzerinde test edilebilir ve başarılı tahminler elde edilebilir olduğu ve geliştirilen yapay sinir ağları yönteminin evrensel olarak diğer arama motorlarında da kullanılabileceği kanıtlanmıştır.

Anlam bazlı olmayan metotlar, sorguların istatistiksel karakteristiklerinden yararlanarak konu değişimlerini tahmin etmektedirler. Tahminlerin güvenilir bir şekilde istatistiksel karakteristiklere dayandırılabilmesi için bu karakteristiklerin konu değişimlerini belirlemede etkili olmaları, yani konu değişimlerini yansıtabilmeleri gerekmektedir. Bu sebeple Özmutlu (2006) çalışmasında, çoklu doğrusal regresyon ve çok faktörlü ANOVA tekniklerini Excite arama motoru verilerine uygulayarak zaman aralığı, arama yapısı ve sorgunun oturumdaki sırası gibi karakteristiklerin, yeni konuya geçmedeki etkisini belirlemeye çalışmıştır. Verilere uygulanan çoklu doğrusal regresyonda, bağımlı değişken olarak konu değişimi veya konu devamı kararı alınmıştır. Sorgulardaki konu devamı 1 ile işaretlenirken konu değişimi 2 ile gösterilmiştir. Bağımsız faktörler; arama yapısı (SP), zaman aralığı (TI) ve sorgunun oturum içerisindeki sırasıdır (QN). Ayrıca bu faktörlerin birbirleriyle olan etkileşiminin de bağımlı değişkene etkisi olabileceği göz önünde bulundurularak, SP – TI etkileşimi, SP – QN etkileşimi ve TI – QN etkileşimi de dikkate alınmıştır. ANOVA analizi ise, sorgu karakteristiklerine göre konu değişimlerinin varyansını araştırmada ve bağımlı değişken üzerindeki her regresyonun anlamlılığını test etmede kullanılmıştır. Excite veri grubu hazırlık aşamasından (Özmutlu ve Çavdur 2005a) sonra ikiye bölünerek ilk kısmı regresyon

denklemini oluşturmada ve ikinci kısmı ise önerilen çoklu doğrusal regresyonun performansını test etmede kullanılmıştır. Veri kümesinin ilk kısmıyla çoklu doğrusal regresyon denkleminin katsayıları belirlenmiş ve yine bu gruba ANOVA analizi uygulanmıştır. Veri grubunun ikinci kısmı, denklemin geçerliliği için kullanılmış ve bağımsız faktörler kullanılarak sorguların konu değişimi veya konu devamı tahminleri gerçekleştirilmiştir. Bu tahminler 1 veya 2 şeklinde olduğundan denklem sonucunu ikili (binary) değişkene çevirmek için bir eşik değeri belirlenmiştir. Yazar çalışmasında eşik değerini, 1 ve 2 değerlerinin medyanı olduğu için 1,5 olarak belirlemiş ve regresyon 1 ve 2 arasında cevap verecek şekilde eğitilmiştir. Çoklu doğrusal regresyonun cevabı 1,5 değerinden küçük ise sorgular konu devamı, 1,5 değerinden büyük ise sorgular konu değişimi olarak işaretlenmiştir. Daha sonra yöntemin tahminleri ile uzman sonuçları karşılaştırılmış ve yöntemin performans değerlendirmesi yapılmıştır. Yöntem, başarılı tahminler yapmasına rağmen önceki çalışmalara benzer olarak, gerçekte 152 adet konu değişimi varken 226 adet konu değişimi tahmininde bulunmuş yani fazla tahmin yapmıştır. Fakat çalışma bütün olarak incelendiğinde, oluşturulan çoklu doğrusal regresyon denkleminin geçerli olduğu, konu değişimleri ile sorgu karakteristikleri arasında anlamlı bir ilişki olduğu ve çoklu doğrusal regresyon yönteminin yeni konu tanılamada kullanılabileceği gösterilmiştir. Ayrıca çok faktörlü ANOVA ile anlam bazlı olmayan zaman aralığı, arama yapısı ve sorgunun oturum içindeki sırası gibi faktörlerle beraber, arama yapısı ile zaman aralığı etkileşiminin (SP – TI) de konu değişimlerinde anlamlı bir etkisi olduğu ortaya koyulmuştur.

Yeni konu tanılamada başarılı olan yöntemlerden biri, daha önceki çalışmalarda da bahsedildiği gibi He ve arkadaşları (2002) tarafından geliştirilen Dempster-Shafer teorisi ve genetik algoritma yaklaşımıdır. Yöntemin başarısının veri kümesinden bağımsız olup olmadığını ortaya koyan çalışmalardan biri Özmutlu ve arkadaşları (2006) tarafından FAST arama motoru verileri kullanılarak gerçekleştirilmiştir. Yazarlar yaptıkları çalışmada, arama motoru sorgularındaki konu değişimlerini tahmin etmek için genetik algoritma ve Dempster-Shafer teorisinden yararlanmışlardır. FAST arama motorundan alınan veriler üzerinde hazırlık aşaması tamamlanmış ve veri grubu iki kısma ayrılmıştır (Özmutlu ve Çavdur 2005a). İlk kısım kullanılarak genetik algoritma yardımıyla parametreler belirlenmiş, ikinci kısım kullanılarak Dempster-Shafer teorisi gerçekleştirilmiş ve yöntemin performansı

değerlendirilmiştir. Yöntemin tahminleri ve uzman sonuçları karşılaştırıldığında Dempster-Shafer teorisinin %97,7 doğruluk oranıyla konu değişimlerini ve %87,2 doğruluk oranıyla konu devamlarını tahmin ettiği görülmüştür. Fakat gerçekte 310 adet olan konu değişimi varken, yöntem tarafından 836 adet konu değişimi tahmini yapılmış yani aşırı tahmin gerçekleşmiş ve yazarlar bu durumun, yöntemde kullanılan varsayımlardan kaynaklanabileceğini belirtmişlerdir. Ayrıca yazarlar bu uygulama ile ilk olarak He ve arkadaşlarının (2002) kullandığı Dempster-Shafer teorisinin diğer arama motoru verilerine de uygulanabileceğini ve yöntemin yeni konu tanılamada başarılı tahminler elde edebileceğini göstermişlerdir. Sonuç olarak, anlam bazlı olmayan yeni konu tanılama yöntemlerinden birinin daha başarısının, veri kümesinden bağımsız olduğu tespit edilmiş ve evrenselliği kanıtlanarak farklı arama motoru verileri üzerinde başarılı tahminler gerçekleştireceği ispatlanmıştır.

Arama motorları verilerinden elde edilen sorguların zaman aralığı, arama yapısı ve sorgunun oturumdaki sırası gibi karakteristikleri ile konu değişimleri arasında anlamlı bir ilişki olduğunu, Özmutlu (2006) çoklu doğrusal regresyon ve çok faktörlü ANOVA teknikleri ile ortaya koymuştur. Bu ilişkiden yola çıkan Özmutlu ve arkadaşları (2007), zaman aralığı ve arama yapısından yararlanarak, bu iki istatistiksel verinin kombinasyonlarını içeren sorguların konu değişimi olasılıklarının hesaplanabileceğini ve sonrasında bu olasılıklar yardımıyla konu değişimlerini tahmin edebileceklerini öngörmüşler ve bu amaçla çalışmalarında şartlı olasılık yönteminden yararlanmışlardır. Çalışmada Excite arama motorundan 1999 yılından alınan veriler ile FAST arama motorundan alınan veriler kullanmış ve verilerin hazırlık aşaması tamamlanmıştır (Özmutlu ve Çavdur 2005a). İki kısma ayrılan veri kümelerinin ilk kısmı konu değişim ve konu devamları için şartlı olasılıkların belirlenmesinde, ikinci kısmı ise bu olasılıkların test edilmesinde kullanılmıştır. Sorguların istatistiksel karakteristikleri 7 grup zaman aralığı ve 7 grup arama yapısından oluştuğu için, şartlı olasılıklar toplamda 49 kombinasyon için bulunmuştur. Her kombinasyon için, o kombinasyona sahip toplam sorguların sayısı ve bunların içinden konu değişimi ve konu devamı olarak işaretlenen sorguların sayıları belirlenmiştir. Daha sonra bu sayılar olasılığa dökülerek o kombinasyona ait şartlı olasılıklar hesaplanmıştır. Yazarlar veri kümelerinin ilk kısımlarıyla, bahsedildiği şekilde şartlı olasılıkları elde etmiş ve daha sonra veri kümelerinin ikinci kısımlarındaki sorguları, bu olasılıklara göre tahmin etmişlerdir. Bu

aşamada konu değişimi veya konu devamı tahmini için olasılık değeri büyük olan alınmıştır. Örneğin, zaman aralığı 1 ve arama yapısı 5 olan 1_5 kombinasyonuna sahip sorguların %74'ü konu devamı ve %26'sı konu değişimi olarak işaretlenmiş ise, test kümesinde aynı kombinasyona sahip bir sorgu ile karşılaşıldığında yöntem sorguyu, görülme olasılığı daha büyük olduğu için, konu devamı olarak tahmin etmektedir. Bu şekilde iki test kümesindeki sorguların zaman aralıkları ve arama yapıları baz alınarak tahminler gerçekleştirilmiştir. Daha sonra yöntemin tahminleri ile uzman sonuçları karşılaştırılarak performans değerlendirmesi yapılmıştır. Performans değerlendirmeleri sonucunda şartlı olasılıklar yöntemi, Excite verisinde bulunan konu devamlarının %95,8'ini ve konu değişimlerinin %53,3'ünü ve FAST verisinde bulunan konu devamlarının %96,12'sini ve konu değişimlerinin %47,1'ini doğru tahmin etmiştir. Yazarlar bu çalışma sonuçlarını, Dempster-Shafer teorisi ve yapay sinir ağları çalışmalarıyla karşılaştırmışlar ve şartlı olasılık yönteminin onlar kadar, hatta bazı parametrelerde onlardan daha başarılı olduğunu gözlemlemişlerdir. Ayrıca diğer yöntemlerdeki konu değişimleri tahminlerinde gözlemlenen aşırı tahminlerin, şartlı olasılık yöntemiyle ortadan kalktığı tespit edilmiştir. Sonuç olarak, sorgu karakteristikleri ile konu değişimleri arasında anlamlı bir ilişki olduğu bir kez daha kanıtlanmış ve bu ilişki şartlı olasılık yöntemiyle kullanılarak, yeni konu tanılamada başarılı tahminler üreten bir yöntem daha geliştirilmiştir.

Yeni konu tanılamada kullanılan istatistiksel yöntemlerin çoğu, uygulamada kolay ve tahminlerde başarılıdır. Anlamlı sonuçlar üretebilen ve uygulamada karmaşık olmayan yöntemlerden biri de Monte-Carlo simülasyonudur. Monte-Carlo simülasyonu, olasılık teorisi üzerine kurulu bir sistem olup, istatistiksel ve matematiksel tekniklerle bir deneyi veya çözülmesi gereken bir fiziksel olayı, tesadüfî sayıları defalarca kullanarak çözebilmektedir. Özmutlu ve arkadaşları (2008b), önceki çalışmada başarısı tespit edilen şartlı olasılık yöntemini (Özmutlu ve ark. 2007), Monte-Carlo simülasyonuyla birleştirerek daha güçlü tahminler elde edebileceklerini öngörmüşlerdir. Şartlı olasılık yönteminde konu değişimi tahminleri, 49 kombinasyon için hesaplanan olasılık değerlerinden büyük olanı baz alınarak yapılmaktaydı. Özmutlu ve arkadaşlarının (2008b) çalışmalarında kullandıkları Monte-Carlo simülasyonu yönteminde de benzer bir şekilde, 49 kombinasyon için hesaplanan olasılıklardan yararlanarak konu değişimi tahminleri oluşturulmuştur. Yazarlar çalışmalarında, Excite arama motorundan 1999 yılında ve 2001 yılında ve FAST arama

motorundan 2001 yılında alınan veriler olmak üzere üç farklı veri grubundan yararlanmışlardır. Bütün veri grupları için hazırlık aşaması tamamlanmış ve her bir veri grubu, test ve eğitim kümesi olmak üzere ikiye ayrılmıştır (Özmutlu ve Çavdur 2005a). Önceki çalışmalarda olduğu gibi, eğitim kümeleri kullanılarak 49 kombinasyon içeren sorgular için şartlı olasılıklar bütün veri grupları için belirlenmiştir (Özmutlu ve ark. 2007). Monte – Carlo simülasyonu uygulamasında, başlangıçta düzgün dağılıma uygun (0,1) arasında rastsal bir sayı üretilmiştir. Eğer üretilen rastsal sayı, sıfır ve şartlı olasılığın konu devamı olarak belirlediği olasılık değeri arasındaysa sorgu konu devamı olarak, eğer üretilen sayı şartlı olasılığın konu devamı olasılığını geçiyorsa sorgu konu değişimi olarak işaretlenmiştir. Bu şekilde her üç test kümesi için konu devamı ve konu değişimi tahminleri yapılmış ve güvenilirlik açısından simülasyon 10 defa tekrarlanmış, elde edilen tahminlerin ortalaması performans değerlendirmesinde kullanılmıştır. Monte – Carlo simülasyonu ile Excite 1999 verilerinde konu değişimi olan sorguların %44'ü, konu devamı olan sorguların %94'ü, Excite 2001 verilerinde konu değişimi olan sorguların %53'ü, konu devamı olan sorguların %92'si ve FAST 2001 verilerinde konu değişimi olan sorguların %44'ü, konu devamı olan sorguların %95'i doğru tahmin edilmiştir. Ayrıca önceki yöntemlerde yüksek bulunan fazla tahmin etme parametreleri, bu çalışmada daha düşük bulunmuştur, kısacası Monte-Carlo simülasyonu, konu değişimi tahminlerini diğer yöntemlere kıyasla daha tutarlı bir şekilde yapmıştır. Ancak doğru tahmin etme yüzdeleri önceki yöntemlere göre daha düşüktür. Bununla beraber, çalışmanın hatalı olarak konu değişimi tahmini yaptığı sorgular incelendiğinde, istatistiksel yöntemin uygulanmasından önce, sorguların arama yapısının ve zaman aralıklarının otomatik olarak belirlenmesi aşamasında, arama yapısı "yeni" olarak işaretlenen sorguların tahmini, konu değişimi olarak yapılmaktadır. Bir başka ifadeyle, hatalı olarak tanımlanan "yeni" arama yapısına sahip sorgular, yöntemin aşırı konu değişimi tahmininde bulunmasına yol açmaktadır. Dolayısıyla, arama yapısı "yeni" olan sorguların incelenmesiyle çalışmanın performansının arttırılabileceği öngörülmüştür.

Yeni konu tanılamada şimdiye kadar kullanılan istatistiksel yöntemler, sorguların karakteristiklerini temel almaktadır (Özmutlu ve Çavdur 2005a, Özmutlu ve Çavdur 2005b, Özmutlu 2006, Özmutlu ve ark. 2006, Özmutlu ve ark. 2007, Özmutlu ve ark. 2008a, Özmutlu ve ark. 2008b). Bu çalışmalarda sorguların arama yapısı ve zaman aralığı kategorileri otomatik olarak belirlenir ve bu kategoriler tahmin yöntemlerinin girdisi olarak

kullanılır. Dolayısıyla kategorilerin hatalı olarak belirlenmesi, sonrasında kullanılacak tahmin yönteminin hatalı tahmin üretmesine neden olmakta ve sonuç olarak tahmin yönteminin performansı düşmektedir. Önceki çalışmalarda kullanılan istatistiksel tahmin yöntemlerinin hepsi, arama yapısı "yeni" olan sorguları doğru bir şekilde konu değişimi olarak tahmin etmektedirler. Ancak sorguların, hazırlık aşamasında hatalı olarak "yeni" arama yapısı sınıfına dahil edilmesi, yöntemlerin aşırı konu değişimi tahmini yapmalarına sebep olmaktadır. Özmutlu ve arkadaşları (2008c) çalışmalarında bu hataların sebeplerini araştırmışlar ve sorguya anlam katan fakat konu değişimine neden olmayan "and, or, the, +, -, &, www., http://, .com, .net" gibi ifadelerin, arama yapılarının hatalı olarak belirlenmesine sebep olduğunu gözlemlemişlerdir. Bu yüzden yazarlar tahmin yöntemini uygulamadan önce, sorguları bu ifadelerden temizlemişler ve temizlenmiş sorguların arama yapıları ve zaman aralıklarını otomatik olarak belirlemişlerdir. Sonrasında yazarlar, yapay sinir ağları yönteminden yararlanarak konu değişimlerini tahmin etmişlerdir. Çalışmada kullanılan Excite ve FAST arama motorları verilerinin hazırlık aşaması, verilerin temizlenmesini de içerecek şekilde tamamlanmış ve her iki arama motoru verisi eğitim ve test kümelerine ayrılmıştır. Geliştirilen yapay sinir ağları modeli 3 katmandan oluşmaktadır. İlk katman olan girdi katmanında zaman aralıklarını ve arama yapısı sınıflarını temsilen iki nöron bulunmaktadır. Çıktı katmanı tek nörondan oluşmaktadır. Gizli katman ise çeşitli deneyler sonucunda beş nörondan oluşturulmuştur. Çıktı katmanından elde edilen sonucun, konu devamı için 1 ve konu değişimi için 2 değerini vermesi için yuvarlama değeri olarak 1,5 eşik değeri seçilmiş metodun 1 veya 2 şeklinde çıktı vermesi sağlanmıştır. Eğitim setleriyle yapay sinir ağları eğitilmiş ve sinoptik ağırlıklar belirlenmiştir. Daha sonra bu veriler test kümesinde kullanılmış ve tahminler gerçekleştirilmiştir. Yöntemin tahminleri ile uzman sonuçları karşılaştırılmış ve performans değerlendirmesi için gerekli olan parametreler hesaplanmıştır. Bu çalışma ile Excite verilerinde konu değişimi olan sorguların %87,1'i, konu devamı olan sorguların %93'ü ve FAST verilerinde konu değişimi olan sorguların %98,7'si, konu devamı olan sorguların %86,1'i doğru tahmin edilmiştir. Yazarlar bu çalışma ile elde edilen parametrelerin, önceki çalışmalardan daha iyi olduğu ve bu yöntemin anlam bazlı olmayan yeni konu tanılama yöntemleri içinde en başarılı uygulama olduğu sonucuna ulaşmışlardır. Diğer taraftan, verilerin temizlenmesine rağmen yöntemin

fazla konu değişimi tahmini yapmasını engelleyememişler ve aşırı tahminlere sebep olan sorguların ayrıntılı bir şekilde incelenmesi gerektiğini vurgulamışlardır.

Bugüne kadar yapılan anlam bazlı olmayan yeni konu tanılama çalışmalarının hepsi aynı veriler üzerinde geliştirilmiş ve yöntemlerin karşılaştırılabilmesi için aynı parametrelerle performans değerlendirmeleri yapılmıştır. Bu çalışmaların performans değerlendirmeleri Excite verileri için Çizelge 1.1 ve FAST verileri için Çizelge 1.2' de özetlenmiştir.

Çizelge 1.1: 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}	$F_{\beta(değişim)}$	$F_{\beta(devam)}$
Uzman Sonuçları	3394	$N_{gerçekdeğişim}$ = 272	$N_{gerçekdevam}$ = 3122	-----	-----	----	----	----	----	----	----	----	----
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

Çizelge 1.2: 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	$N_{gerçekdeğişim}$ = 310	$N_{gerçekdevam}$ = 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

Bu çalışmaların, kelimelerin anlamlarını göz ardı ederek konu değişimi tahmini yaptıkları için, hatalı tahminlere sahip olmaları olağandır. Ancak daha etkili yöntemlerin geliştirilmesi için hataların ana nedenleri tespit edilmelidir. Yapılan çalışmaların hata analizlerinin sonuçları, sonraki bölümde ele alınmış ve hataların giderilmesi için alternatif çözüm önerileri geliştirilmiştir.

2. MATERYAL VE YÖNTEM

2.1. Problem Tanımı

Anlam bazlı olmayan yeni konu tanılama çalışmalarında, yöntemlerin hangi noktalarda hata yaptığını bulabilmek için tahmin aşamasında kullanılan karar mekanizmaları incelenmelidir. Arama motorlarından alınan verilerde, arama motorlarını kullanan kullanıcılar hakkında bilgiler vardır. Bu bilgiler IP adresi, arama zamanı ve sorgudan oluşmaktadır. IP değişikliği, kullanıcılar çevirmeli ağ ile internete bağlandıklarında sorguların sona erdiğini göstermeyebilir, çünkü bu durumda kullanıcı İnternet'e her bağlandığında arama motoru farklı bir IP adresi belirlemektedir. Böyle bir durumda arama motoru, kullanıcıları bir çerez (cookie) yardımıyla tanıyabilmekte ve oturum devamını belirleyerek farklı IP adreslerini kaydetmektedir. Dolayısıyla oturum, tek bir kullanıcı tarafından gönderilen sorgu grubunu göstermekle beraber bir oturumdaki IP adresleri farklı olabilmektedir. Özmutlu ve Çavdur (2005a) yaptıkları çalışmada, He ve arkadaşlarının (2002) çalışmalarınıı temel alarak, oturumdaki ardışık sorguların arama yapılarını (search pattern - *sp*) 7 farklı sınıfta aşağıdaki gibi tanımlamışlardır:

- Yeni (New): İkinci sorgu birinci sorguyla ortak terim içermemektedir.

  Örnek: Sorgu $_{j}$: Otomobil

  Sorgu $_{j+1}$: Harry Potter

- Sonraki Sayfa (Next Page): İkinci sorgu birinci sorguyla aynıdır. Yani ikinci sorgu, birinci sorguyla ilgili diğer bir sonuç kümesini istemektedir.

  Örnek: Sorgu $_{j}$: Otomobil

  Sorgu $_{j+1}$: Otomobil

- Genelleştirme (Generalization): İkinci sorgu birinci sorgudan daha az terim içermektedir ve ikinci sorgunun bütün terimleri birinci sorguda yer almaktadır.

  Örnek: Sorgu $_{j}$: Kırmızı otomobil

  Sorgu $_{j+1}$: Otomobil

- Özelleştirme (Specialization): İkinci sorgu birinci sorgudan daha fazla terim içermektedir ve birinci sorgunun bütün terimleri ikinci sorguda yer almaktadır.

  Örnek: Sorgu $_j$: Otomobil

  Sorgu $_{j+1}$: Kırmızı otomobil

- Düzenleme (Reformulation): İkinci sorgu terimlerinin bazıları (tamamı değil) birinci sorguda yer almaktadır ancak birinci sorgu, ikinci sorguda yer almayan bazı terimleri de içermektedir. Bu durum, kullanıcının birinci sorgudan bazı terimleri çıkardığını ve ikinci sorguya yeni terimler eklediğini göstermektedir. Aynı zamanda kullanıcı ilk sorgudaki terimleri, ikinci sorguda farklı bir sırada yazarsa, bu da "düzenleme" olarak düşünülmüştür. Bu durum sorguların aynı olmasını gerektiren "sonraki sayfa" olarak düşünülemez.

  Örnek: Sorgu $_j$: Kırmızı otomobil Toyota

  Sorgu $_{j+1}$: Otomobil Corolla

- İlgili Geri-Besleme (Relevance Feedback): İkinci sorgu hiç terim içermemektedir (boştur) ve kullanıcı ilgili sayfalar seçeneğini seçtiğinde sistem tarafından oluşturulmaktadır.

  Örnek: Sorgu $_j$: Otomobil

  Sorgu $_{j+1}$: "″

- Diğer (Other): İkinci sorgu yukarıdaki kategorilerin hiçbirine uymuyorsa, "diğer" olarak işaretlenir.

  Örnek: Sorgu $_j$: "″

  Sorgu $_{j+1}$: Toyota otomobil

Özmutlu ve Çavdur (2005a) bu çalışmalarında, sorguların arama zamanını da tahmin yöntemlerinde veri olarak kullanmışlardır. Zaman aralığı kategorileri (time interval - *ti*), ardışık sorguların gelişleri arasındaki farkların uzunluğu göz önüne alınarak oluşturulmuş ve 7 kategoride toplanmıştır: 0–5 dk, 5–10 dk, 10–15 dk, 15–20 dk, 20–25 dk, 25–30 dk ve 30+ dk. Çalışmalarda kullanılan sorguların arama yapısı sınıfları, yazarlar tarafından PASCAL'da geliştirilen bilgisayar programıyla otomatik olarak belirlenmiştir (Programın algoritması EK 1'de verilmiştir).

Bugüne kadar yazarlar tarafından Excite ve FAST arama motorları verileri üzerinde, Dempster–Shafer yöntemi (Özmutlu ve Çavdur 2005a), şartlı olasılık yöntemi (Özmutlu ve

ark. 2007), yapay sinir ağları yöntemi (Özmutlu ve Çavdur 2005b, Özmutlu ve ark. 2008a, Özmutlu ve ark. 2008c) ve Monte Carlo simülasyonu (Özmutlu ve ark. 2008b) gibi istatistiksel metotlar kullanılarak yeni konu tanılama çalışmaları gerçekleştirilmiştir. Tahmin etme aşamasında, sorguların anlamını ihmal eden bu yöntemlerin, kullandığı verilerin yanında başka ortak özellikleri de vardır. Her yöntemin uygulamasından önce veri setleri için, Özmutlu ve Çavdur'un (2005a) çalışmalarında olduğu gibi, hazırlık aşaması tamamlanmıştır. Hazırlık aşaması; Poisson örneklemesiyle verilerin belirlenmesi, sorguların arama yapısı sınıflarının ve zaman aralıklarının belirlenmesi, uzman tarafından konu değişimi ve konu devamlarının belirlenmesi adımlarını içermektedir. Hazırlık aşamasından sonra, kullanılacak yöntemin çalışma algoritması göz önüne alınarak, eğitim ve test kümesi olarak ayrılan verilerden yararlanılıp istatistiksel uygulamalar gerçekleştirilmiştir. Sonuçta yöntemler, eğitim kümesindeki arama yapısı sınıfları ve zaman aralıkları kombinasyonlarından yola çıkarak, test kümesi sorgularında konu değişimi ve konu devamı tahminlerini gerçekleştirmişlerdir.

Yöntemlerin test kümesindeki tahminleri incelendiğinde benzer noktalarda hata yaptıkları gözlenmiştir. Her yöntem için aynı terminoloji kullanılmış olup hata tipleri belirlenmiştir. A tipi hata, yöntem tarafından konu değişimi olarak tahmin edilip gerçekte konu devamı olan sorgu sayısı ve B tipi hata, yöntem tarafından konu devamı olarak tahmin edilip gerçekte konu değişimi olan sorgu sayısı olarak tanımlanmıştır. Geliştirilen yöntemlerin amacı, konu değişimlerinin doğru şekilde tahmin edilmesi olarak belirlendiği için A tipi hatalar önem kazanmaktadır. Çünkü A tipi hata, yöntem konu değişimi tahmininde bulunup gerçekte sorgular konu devamı olduğunda ortaya çıkar, dolayısıyla yanlış tahmin etme durumu söz konusudur. B tipi hata ise, yöntem konu devamı kararı verdiğinde gerçekte sorgular konu değişimi gösteriyorlarsa ortaya çıkar. Dolayısıyla konu değişimlerinin doğru bir şekilde belirlenmesi her iki hata tipini de olumlu yönde etkiler. Bununla beraber A tipi hatanın yakalanmasının, literatürde daha önemli olduğu düşünüldüğünden, her yöntem için A tipi hatanın azaltılması hedeflenmiştir.

A tipi hataların meydana geliş sebepleri, her yöntemin çalışma algoritmasına göre farklılık göstermektedir. Örneğin şartlı olasılıklar yönteminde A tipi hataların 30 dk'dan fazla süren sorgularda ortaya çıktığı görülmüştür. Bir diğer ifadeyle, 30 dk'dan fazla süren bir sorgu için, şartlı olasılık yöntemi konu değişimi tahmini yapmış; fakat bazı sorgu gruplarının

bu süre zarfında aynı konuda kaldıkları gözlemlenmiştir. Ayrıca, arama yapılarına göre hatalar incelendiğinde, "yeni" arama yapısı için yapılan tahminlerden bir kısmının hatalı olduğu gözlenmiştir. Diğer bir çalışma olan Monte – Carlo simülasyonunda da benzer bir durum söz konusudur. Yöntemin arama yapısı "yeni" olan sorgularda konu değişimi için bulduğu olasılık değeri, konu devamı olasılık değerinden daha fazla olduğundan sorgular konu devamı bile olsa konu değişimi tahmini yapılmaktadır. Bu durum, arama yapısı hatalı olarak "yeni" şeklinde belirlenen sorguların da yöntem tarafından konu değişimi olarak tahmin edilmesine yol açmakta ve konu değişimi tahminleri, mevcut konu değişimi sayısından daha fazla olmakta; dolayısıyla A tipi hata da artmaktadır. Ayrıca zaman aralığı 1 ve 7 olan sorgularda da yapılan tahminlerden bir kısmının hatalı olduğu gözlenmiştir.

Yöntemlerin tahminleri detaylı bir şekilde incelendiğinde A tipi hataların, arama yapısı "yeni" olan sorgularda yoğunlaştığı tespit edilmiştir. Anlam bazlı olmayan istatistiksel yöntemlerin tahmin aşamasında baz aldıkları girdiler, sorguların arama yapıları ve zaman aralıklarından oluşmaktadır. Dolayısıyla sorguların arama yapılarının sınıflandırılmasında yapılan hatalar, yöntemlerin tahminlerini olumsuz yönde etkilemektedir. Bir diğer ifadeyle, yöntemlerin sonuçlarının iyileştirilmesi, arama yapılarının doğru bir şekilde belirlenmesine bağlıdır. Bu da ancak arama yapılarını belirleyen algoritmanın kullanacağı verilerin iyileştirilmesi ile gerçekleştirilebilir. Verilerin hangi açılardan iyileştirilmesi gerektiğinin belirlenmesi aşamasında, arama motorlarından alınan veriler ile arama yapısı sınıflarını belirleyen algoritma dikkate alınmış ve arama yapısı sınıflarının yanlış belirlenmesine sebep olan hatalar tespit edilerek bir sonraki bölümde olduğu gibi gruplandırılmıştır.

2.1.1. Boolean operatörleri ve internet terimlerinden kaynaklanan hatalar

Boolean operatörleri (+,-, and, or… gibi) ve genel internet terimleri (www, http://, .com… gibi), takip eden sorgunun "yeni" olarak işaretlenmesine neden olabilir ve bu şekilde yapılan hatalı belirlemeler de A tipi hataların ortaya çıkmasına sebep olabilir.

Örnek 1: Sorgu $_j$: EDUCATION+Desert

Sorgu $_{j+1}$: Desert

Örnek 2: Sorgu $_j$: www.wal-mart.com

Sorgu $_{j+1}$: walmart.com

Örnek 1'de verilen ardışık sorgularda ortak terim bulunmasına rağmen "+" operatörü, sorgu $_{j+1}$'in arama yapısının "yeni" olarak tanımlanmasına sebep olmaktadır. Benzer bir şekilde Örnek 2'deki sorgularda ortak terimler bulunmasına rağmen, sorgu$_j$'de bulunan "-" ve "www." ifadeleri sorgu$_{j+1}$'de yer almadığı için sorgu$_{j+1}$'in arama yapısı "yeni" olarak tanımlanmaktadır. Excite 2001 ve FAST 2001 verileri incelendiğinde, Boolean operatörleri ve genel internet terimleri içeren 19 adet A tipi hatanın, bu ifadeler temizlendiğinde ortadan kaldırılabileceği belirlenmiştir (EK 2).

Boolean operatörleri ve internet terimleri, arama yapılarının "yeni" olarak işaretlenmesine neden olabileceği gibi, bu ifadelerin yokluğunda birbirine benzemeyen sorgular için, sonra gelen sorgunun arama yapısının "düzenleme" olarak işaretlenmesine de sebep olabilmektedir.

Örnek: Sorgu $_j$: solar AND cell

Sorgu $_{j+1}$: free AND project AND management AND software

Bu örnekteki sorgular gerçekte konu değişimi olmasına ve "yeni" arama yapısına sahip olmasına rağmen, arama yapısı "AND" ifadesinden dolayı "düzenleme" olarak tanımlanmakta ve bu durum istatistiksel yöntemlerin B tipi hata yapmasına neden olmaktadır. Benzer şekilde B tipi hataya neden olan sorgu çiftleri örnekleri arama motoru sorguları incelenerek genişletilebilir (EK 3).

2.1.2. Kullanıcının yazım farklılıklarından kaynaklanan hatalar

Yazım farklılıklarından kaynaklanan hatalarda, sorgular çok benzer kelimeler içerseler de arama yapıları "yeni" olarak algılanabilir ve bu durum A tipi hatalara neden olabilir.

Örnek: Sorgu $_j$: hotels in istambul

Sorgu $_{j+1}$: istanbul

İlk sorgudaki "istambul" ifadesinin yanlış yazımı arama yapısının "genelleştirme" yerine "yeni" olarak tanımlanmasına sebep olmakta ve dolayısıyla istatistiksel yöntemlerin A tipi hata yapmasına neden olmaktadır. Bununla beraber ardışık sorgularda bulunan çoğul ifadeler de sonraki sorgunun arama yapısının "yeni" olarak tanımlanmasına sebep olabilir.

Örnek: Sorgu $_j$: Scholarships

Sorgu $_{j+1}$: scholarship news

Birinci sorgudaki çoğul ifade, ikinci sorgunun arama yapısının "özelleştirme" yerine "yeni" olarak işaretlenmesine sebep olmakta ve sorgunun istatistiksel yöntem tarafından konu devamı yerine konu değişimi olarak işaretlenmesine yani A tipi hataya neden olmaktadır.

Kelime kısaltmalarının kullanılmasıyla oluşan A tipi hatalarda farklı terimlerle ifade edilen benzer kavramlar, sorguların arama yapılarının "yeni" olarak işaretlenmesine neden olabilir.

Örnek: Sorgu $_j$: DX8

Sorgu $_{j+1}$: DirectX8

Ayrıca ardışık sorgulardaki ifadelerin büyük-küçük harfle yazımından kaynaklanan farklılıklar da A tipi hataların oluşmasına neden olmaktadır. Aşağıdaki örnekte yer alan ardışık iki sorgu, aynı ifadeyi barındırmaktadır; ancak ikinci sorgunun arama yapısı, büyük harfler kullanıldığından "yeni" olarak tanımlanmakta ve sonrasında A tipi hata meydana gelmektedir.

Örnek: Sorgu $_j$: eniac

Sorgu $_{j+1}$: ENIAC

2.2. Çözüm Önerileri

Boolean operatörleri ve internet terimlerinden kaynaklanan hataların yapısı incelendiğinde, bu hataların verilerin filtrelenmesiyle giderilebileceği tespit edilmiştir. Özmutlu ve arkadaşları (2008d) yaptıkları çalışmada, konu değişimlerinin yapay sinir ağları yöntemiyle tahmin edilmesinden önce filtreleme işlemi kapsamında, veri kümelerindeki bütün sorguları küçük harfe dönüştürmüşler ve "and", "or", "the", "a", "+", "-", ",", "&", "www.", "http://", ".com", ".net", ".gov", ".mil", "on", "at" gibi ifadeleri sorgulardan temizlemişlerdir. Bu işlemden sonra yapay sinir ağları yöntemiyle sorguların konu değişim tahminleri gerçekleştirilmiş ve bir önceki çalışmadan (Özmutlu ve ark. 2008c) daha iyi sonuçlar elde edilmiştir. Bununla beraber, iki çalışmada da eğitim kümesi olarak kullanılan sorguların arama yapıları karşılaştırılmış, arama yapısı "sonraki sayfa", "genelleştirme" ve "özelleştirme" olarak işaretlenen sorguların sayısının arttığı ve arama yapısı "düzenleme" ve "yeni" olarak belirlenen sorguların sayısının azaldığı tespit edilmiştir. Sonuç olarak birinci

gruba dâhil olan hataların, verilere uygulanan basit filtreleme işlemleriyle azaltılabileceği görülmüştür.

Filtreleme çalışması sonrasında, hatalı tahminler incelenmiş ve bunlar iki grupta toplanmıştır. Birinci grup hatalar; ortadan kaldırılması, anlam bazlı işlemlere bağlı olan hatalardır. Bu tip hatalara örnek olarak, eş anlamlı kelimeler barındıran ardışık sorgular verilebilir.

Örnek: Sorgu $_j$: hotel

Sorgu $_{j+1}$: inn

Bu örnekte kullanıcı ardışık iki sorgusunda aynı konuya devam etmektedir. Ancak kelimelerin anlamlarına bakmayan istatistiksel yöntemlerle, örnekteki konu devamlılığının tespit edilebilmesi mümkün değildir.

İkinci grup hatalar; hatalı yazımdan dolayı veya yazım kuralları gereği oluşan hatalardır. Aşağıdaki örnekte de görüldüğü gibi kullanıcının aynı konuya devam etmesine rağmen yanlış yazılan bir harf, ikinci sorgunun konu değişimi olarak tanımlanmasına sebep olmakta ve devamında hatalı tahmin meydana gelmektedir.

Örnek: Sorgu $_j$: cybersc@n

Sorgu $_{j+1}$: cyberscan

İkinci grup hataların, doğal dil işleme yöntemleriyle azaltılabileceği öngörülmüştür. Ancak bugüne kadar kullanılan yöntemler karar aşamasında, sorguların istatistiksel karakteristikleri olan arama yapılarını ve zaman aralıklarını dikkate aldıklarından, doğal dil işleme amacıyla yeni bir yöntem geliştirilmelidir. Yapılan araştırmalar sonucu N-gram yaklaşımının bir türevi olan karakter n-gram yöntemi ile ikinci grup hataların azaltılabileceği görülmüştür.

2.3. N-Gram Yöntemi

2.3.1. Genel bilgiler

İstatistiksel doğal dil işleme yöntemlerinin amacı, doğal dil alanında istatistiksel çıkarımlar yapmaktır. Genel anlamıyla istatistiksel çıkarım, belli bir verinin ele alınması ve dağılımı hakkında öngörüde bulunulmasıdır. İstatistiksel çıkarımın adımları, eğitim verisinin

denklik sınıflarına ayrılması, her denklik sınıfı için iyi bir istatistiksel tahmin edici bulunması ve çoklu tahmin edicilerin kombine edilmesi olarak tanımlanabilir.

İstatistiksel tahmin yöntemlerini kullanan dil modellemenin klasik görevi, sıradaki kelimeyi, daha önce karşılaşılan kelimeler aracılığıyla tahmin etmektir ve N-gram yaklaşımı, istatistiksel dil modellemede geniş bir uygulama alanına sahiptir (Zitouni 2007, Huang ve ark. 2004). Bu alanda gerçekleştirilen ilk çalışmalardan biri Shannon'a (1951) aittir ve geliştirilen "Shannon Game" ile bir metindeki sıradaki harf tahmin edilmeye çalışılmıştır (Manning ve Schütze 1999). Bu çalışmayı takip eden birçok farklı uygulama literatürde yer alsa da N-gram modeli, dil modellemede en basit ve en başarılı temeli oluşturmuştur (Huang ve ark. 2003).

Shannon (1951) ile başlayan N-gram yöntemi uygulamaları gün geçtikçe farklı alanlarda kullanılmaya başlanmıştır. Örnek olarak, Damashek (1995) yaptığı çalışmada, kısıtlandırılmamış metinde konular arası benzerlikleri ölçmek için n-gram yaklaşımından yararlanmışken, Huang ve arkadaşları (2003), n-gram dil modelleriyle Livelink veri grubundaki oturum sınırlarını belirlemeye çalışmışlardır. Ayrıca yöntem Canvar ve Trenkle (1994) tarafından metinsel hataların tespitinde de kullanılmıştır. Yazarlar çalışmalarında, elektronik dokümanları araştırmışlar ve n-gram sıklıklarını hesaplayarak dokümanlardaki yazım ve dil bilgisi gibi metinsel hataları tespit etmişlerdir. Diğer taraftan yöntem, dil tanımlama amacıyla Roark ve arkadaşları (2007) tarafından kullanılmıştır. Kısaca N-gram yaklaşımı dil veya görsel karakter (speech or optical recognition) tanıma çalışmalarına temel olmakla beraber yazım düzeltme (spelling correction), el yazısı tanıma (handwriting recognition) ve istatistiksel makine çevirilerinde (statistical machine translation) sıklıkla kullanılmaktadır.

2.3.2. Kümeler: denklik sınıflarının oluşturulması (Bins)

2.3.2.1. Güvenilirlik ve ayrımsama

Normalde bir özellik hakkında çıkarım yapmak için, tahmin edilen modelin diğer özellikleri bulunmak istenir. Burada geçmişteki davranışların gelecekteki olaylara rehberlik ettiği varsayılır (modelin durağan olduğu kabul edilir). Bu varsayım, sınıflandırma görevini

doğurur ve çeşitli sınıf özellikleri temelinde hedef özellik tahmin edilmeye çalışılır. Bunu yaparken veri kümesi, sınıf özelliklerini paylaşacak şekilde eşdeğerlik sınıflarına bölünür ve bu sınıflama kullanılarak verideki yeni parçaların hedef özellik değeri tahmin edilmeye çalışılır. Burada bağımsız varsayımlar yapmak gerekir; veri diğer özelliklere bağlı değildir veya bağımlılık ihmal edilebilecek kadar küçük boyuttadır. Daha fazla sınıf özelliği, hedef özelliğin bilinmeyen olasılık dağılımının daha iyi belirlenmesini sağlar. Diğer bir deyişle, veriyi daha fazla kümeye ayırmak daha güçlü ayrım yapılmasını sağlar. Buna karşılık çok fazla küme kullanılırsa belirli bir küme çok az sayıda eğitim örneği barındırabilir ve bu sebeple, o kümenin hedef özelliği hakkında istatistiksel olarak güvenilir tahminler yapmak zorlaşır. Bu iki kısıt altında eşdeğerlik kümelerinin bulunması ilk hedefi oluşturur.

2.3.2.2. N-gram modelleri

Sonraki kelimenin tahmin edilmesi P ($w_n|w_1,...., w_{n-1}$) olasılık fonksiyonunun tahmin edilmesi olarak ifade edilebilir. Böyle bir stokastik problemde, sonraki kelimeyi tahmin etmek için geçmişin, yani önceki kelimelerin sınıflaması kullanılır. Birçok metne bakılarak hangi kelimenin diğer kelimeyi takip etme eğiliminde olduğu bilinebilir.

Burada her metne ait geçmişin ayrı ayrı göz önüne alınması mümkün değildir. Çoğu zaman daha önce duyulmayan bir cümle dinlenebilir ve bu gibi durumlarda tahminlerin baz aldığı, önceden tanımlı metne ait geçmiş bulunmamaktadır, hatta cümlenin bir kısmı önceden duyulmuş olsa bile sonu farklı bitebilir. Bu yüzden, bazı yönlerden benzer olan geçmiş verileri gruplama yöntemlerine ihtiyaç duyulur. Böylelikle sonraki kelimeyi tahmin etmede geçerli veriler elde edilebilir. Gruplama yöntemlerinden biri olan Markov varsayımı, son birkaç kelimenin sıradaki kelimeyi etkilediği temeline dayanır. Bütün geçmişlerin aynı eşdeğerlik sınıflarında, aynı son n-1 kelimenin yerleştirildiği bir model yapılırsa buna, (n-1). sıra Markov modeli veya n-gram modeli denir (n-gram'ın son kelimesi tahmin edilen kelimedir). N-gram dil modelleri, sıradaki kelimenin görülme olasılığının ondan önceki n-1 kelimeye dayandığını varsayar.

Prensipte n-gram modelleri oldukça geniş olabilir. Örnek olarak;

Sue swallowed the large green ____.

Burada *swallowed* fiili, sonraki kelimenin ne olabileceği hakkında gerçekçi tahminler yapılmasını sağlar. Bu fiilden sonra *pills* gelebilir ancak *large green___.* ifadesini doğal olarak takip edebilecek *a tree, car, mountain* ifadelerinin *swallowed* fiilini takiben kullanılması muhtemel değildir. Bu örnekten de anlaşılacağı gibi verinin çok fazla kümeye ayrılması, daha fazla sayıda parametrenin tahmin edilmesini gerektireceğinden, problem teşkil etmektedir. Örnek olarak, bir konuşmacının kelime dağarcığının 20,000 kelime olduğu varsayılırsa parametre sayıları için tahminler aşağıdaki gibi olur:

Çizelge 2.1: N-gram modelleri için parametrelerdeki artış

Model	Parametreler
(1. sırada) 2-gram modeli	20,000x19,999=400 milyon
(2. sırada) 3-gram modeli	$20,000^2$x19,999=8 trilyon
(3. sırada) 4-gram modeli	$20,000^3$x19,999=$1.6x10^{17}$

Yararlı olabileceği düşünülen 5-gram modelinin büyük bir ana kütle oluşturacağı için, pratikte kullanışlı olmadığı yukarıdaki örnekten rahatlıkla görülebilir. Bu yüzden n-gram modelleri, genelde küçük verilerde 2-gramlar veya 3-gramlar kullanılarak oluşturulur (Manning ve Schütze 1999).

2.3.2.3. N-gram modellerinin oluşturulması

N-gram modellerinin oluşturulmasında ilk aşama verilerin seçimidir. Metin boyutunun büyük seçilmesi haline, çok fazla parametrenin oluşması ve modelin büyük eğitim setiyle eğitilmesi durumları avantaj gibi gözükse de, bu çalışma çok fazla bilgisayar alanı ve CPU süresi gerektirir. Bu yüzden kabul edilebilir boyutlarda metinler, çalışmalarda tercih edilmelidir.

İlk aşama, veri yığının yeniden işlenmesidir. Örneğin metinlerde bulunan noktalama işaretleri hatalara sebep olabilir ve temizlenmesi gerekir. Bu aşamada metinler, eğitim kümesi ve test kümesi olarak ayrıştırılır. Manning ve Schütze (1999) yaptıkları çalışmada Jane Austen'in romanlarından yararlanmışlardır. Gutenberg Projesi kapsamında kullanılan metinler basit ASCII dosyalarıdır. Bu romanlardan *Emma, Mansfield Park, Northanger*

Abbey, Pride and Prejudice ve Sense and Sensibility metinleri eğitim kümesi için kullanılmış, *Persuasion* metni test kümesi olarak ayrılmıştır. Böylece N= 617,091 kelime ve V=14,585 sözcük tipinden oluşan eğitim kümesi elde edilmiştir. Yazarlar, verilerini noktalama işaretlerinden arındırdıklarında çok uzun sıralı kelime grupları elde etmişlerdir; ancak bu durum genelde insanların kullandığı sistemi yansıtamamaktadır. Bu yüzden, cümleler SGML kodları olan <s>, </s> karakterleriyle birbirlerinden ayrılmıştır. Böylece cümlenin başındaki kelimelerin olasılık hesaplamaları, önceki cümlenin son kelimelerine değil de cümlenin başındaki kelimelere bağlı olmuştur. Ayrıca harf boyutlarını değiştirmedikleri için cümle başlangıçlarını kusurlu da olsa tespit edebilmişlerdir.

2.3.3. İstatistiksel tahmin yöntemleri

Belli bir kümeye dâhil olan belli sayıdaki eğitim verisinin elde edilmesinden sonra ikinci amaç, bu verilere dayanarak hedef özellik için nasıl iyi olasılık tahmininin türetileceğidir. N-gramların belirlenmesinde, P (w_1,...., w_n) olasılığı ve P (w_n|w_1,...., w_{n-1}) olasılık fonksiyonu ile ilgilenilir.

$$P(w_n \mid w_1, \ldots, w_{n-1}) = \frac{P(w_1, \ldots, w_n)}{P(w_1, \ldots, w_{n-1})} \qquad (2.1)$$

Koşullu olasılık dağılımlarının iyi tahmin edilmesi, n-gramların bilinmeyen olasılıklarının dağılımının tahmin edilmesini kolaylaştırır. Ancak düzgünleştirmede (smoothing), n-gram olasılık tahminlerinin düzeltilmesi bir seçenek olabileceği gibi, doğrudan koşullu olasılık dağılımlarının düzeltilmesi de söz konusu olabilir. Çok sayıda koşullu olasılık dağılımı ayrı ayrı düzeltileceğinden ki bunların bir şekilde sınıflanması gerekmektedir, iki yöntem aynı sonucu vermez.

Eğitim metninin N kelimeden oluştuğu varsayılsın. Eğer metnin başına n-1 gereksiz sembol koyulursa veri kümesinin N adet n-gramdan oluştuğu söylenebilir. Küme veya eşdeğerlik sınıflarının sayısı B ile gösterilsin. V^{n-1}; V sözlük boyutunda sonraki kelimeyi hesaplama işi için ve V^n; farklı n-gramların olasılıklarını tahminleri için kullanılsın. Eğitim verisindeki belirli bir n-gram sıklığı C(w_1,...., w_n) ile gösterilsin ve N_r ise eğitim verisinde r defa görülen n-gram kümesini temsil etsin (ör: N_r =|{ w_1,...., w_n : C(w_1,...., w_n) = r}|). Bu

şekilde tanımlanan sıklıkların sıklığı, tahmin yöntemlerinde oldukça fazla kullanılmaktadır. Buraya kadar anlatılan notasyonlar aşağıdaki gibi özetlenebilir:

Çizelge 2.2: İstatistiksel Tahmin Yöntemlerinin notasyonu

N	Eğitim örneklerinin büyüklüğü
B	Eğitim örneklerinin bölündüğü küme sayısı
w_{1n}	Eğitim metnindeki w_1,...., w_n n-gramı
$C(w_1,...., w_n)$	Eğitim metnindeki w_1,...., w_n n-gramının sıklığı
r	bir n-gramın sıklığı
f(.)	kullanılan yöntemin sıklık tahmini
N_r	r eğitim örneğine sahip kümelerin sayısı
T_r	r sıklığına sahip n-gramların toplam sayısı

2.3.3.1. En büyük olabilirlik tahmini (MLE)

Eşdeğerlik sınıflarının nasıl oluşturulduğu göz ardı edilerek önceki aşamanın, belirli bir sayıda eğitim örneği içeren kümelerin bulunmasıyla tamamlandığı varsayılsın. Sonraki kelimeyi tahmin etmek için, önceki iki kelimeye bakan bir 3-gram modeli ve önceki kelimelerin *come across* olduğu bir küme göz önüne alınsın. Belirli bir yığında *come across*'u barındıran 10 eğitim örneği bulunsun ve bunların sekizinden sonra *as*, birinden sonra *more* ve diğerinden sonra da *a* gelsin. Bu aşamadan sonraki soru, sıradaki kelimenin tahmin edilmesinde hangi olasılık değerlerinin kullanılacağıdır.

İlk akla gelen cevap, göreli frekansların, olasılık değerleri olarak kullanılabileceğidir. Bu durumda olasılıklar aşağıdaki gibi dağılır:

P(as) = 0.8

P(more) = 0.1

P(a) = 0.1

P(x) = 0.0

x, diğer takip eden kelimeler için kullanılmıştır.

Bu tahmin yöntemine, En Büyük Olabilirlik Tahmini (Maximum Likelihood Estimation - MLE) denir ve aşağıdaki formüller yardımıyla olasılıklar hesaplanır:

$$P(w_1,\ldots,w_n) = \frac{C(w_1,\ldots,w_n)}{N} \quad (2.2)$$

$$P_{MLE}(w_n \mid w_1,\ldots,w_{n-1}) = \frac{C(w_1,\ldots,w_n)}{C(w_1,\ldots,w_{n-1})} \quad (2.3)$$

Eğer gözlemlenen veri düzeltilirse ve bütün olası parametre atama uzayı, belirli bir dağılım ile göz önüne alınırsa, benzerlik fonksiyonu elde edilmiş olur. Eğitim yığınındaki en yüksek olasılıkları veren parametre değerleri seçildiği için en büyük benzerlik tahmini de denebilir. Bu yöntem, eğitim yığınında olmayan olaylara olasılık değeri ayırmamakta ve gözlemlenen verilere de yüksek olasılık değerleri atamaktadır.

Genelde MLE yöntemi istatistiksel çıkarım için uygun değildir, çünkü çok kapsamlı sözlükler kullanılsa bile oluşabilen, verilerin seyrekliği, problem oluşturmaktadır. Çoğu kelimenin bilinen, sıradan ve alışıldık olmasının yanında, bunların büyük çoğunluğu seyrek kullanılan, alışılmadık kelimelerdir. MLE, görülmeyen olaylara sıfır olasılık değeri atar ve uzun bir dizinin olasılığı, alt grupların olasılıklarının çarpımından elde edildiği için de, bu sıfırlar yüzünden eğitim metninde görülmeyen n-gramlara, sıfır olasılık değeri atanır. Önceki örnekte MLE yöntemi, *come across*'u takip edebilecek *the* ve *some* kelimelerini dikkate almamaktadır.

Veri eksikliğine bir başka örnek olarak; IBM Lazer Patent Text yığınından alınan 1,5 milyon kelime eğitildikten sonra, aynı yığından alınan başka bir metinde %23 oranında görülmeyen durumla karşılaşılmıştır (Manning ve Schütze 1999). Sonuç olarak, eğitim verisinin büyüklüğü önemli olsa da MLE yöntemi ile başa çıkmak için yetersiz kalmaktadır. Mesela *come across*'u bir sayı takip edebilir ve her sayı eğitim verisinde bulunmayabilir. Özetle, eğitim verisinde görülmeyen olaylara olasılık değerlerinin atanabilmesini sağlayan daha iyi yöntemlere ihtiyaç vardır.

N-gram modellerinde MLE tahmin yönteminin kullanımı

Manning ve Schütze (1999) yaptıkları çalışmada, Austen veri grubunda MLE yöntemini kullanarak n-gram olasılıklarını belirlemişlerdir. Pratik sistemlerde, bütün kelimeler için n-gramlar hesaplanmamaktadır. Bunun yerine *k* kelime için n-gramlar hesaplanır, diğer

kelimeler sözlük dşı (out-of vocabulary-OOV) olarak dikkate alınır ve <UNK> gibi tek belirteçle işaretlenir. Bu işlem, eğitim yığınında sadece bir kere karşılaşılan bütün kelimeler için yapılır. Kelimelerin Zipfian dağılımından dolayı düşük frekanslı öğelerin çıkarılması, parametre uzayını küçültür. Bu aşamada modelin kalitesi düşürülmemelidir.

Yazarlar çalışmalarında, eğitim verisinde buldukları şartlı olasılık değerlerini kullanarak, test yığını olan *Persuasion* verisindeki cümlelerin kelimelerinin olasılıklarını hesaplamışlardır. Bir modeli değerlendirmek için onu farklı verilerde denemek çok önemlidir. Diğer türlü, dilin yapısını tahmin etmede modelin ne kadar iyi çalıştığı gözlemlenemez. Bu olasılık dağılımlarının özeti Çizelge 2.3'de gösterilmiştir. Burada, takip eden kelimenin olasılık dağılımı MLE yöntemi kullanılarak farklı *n* değerleri için hesaplanmıştır. İlk sütunda, farklı kelimelerin beklenen benzerlik sıralaması görülmektedir. 1-gram dağılımı, içeriği tümüyle yok sayarak bütün farklı kelimelerin sıklıklarını kullanmaktadır. Bu, kullanışlı değildir çünkü örnekteki gibi *(in person she was inferior to both sisters)* çoğu cümledeki çoğu kelime, geneldir.

2-gram modeli, sonraki kelimeyi tahmin etmek için önceki kelimeyi kullanır ve genelde daha iyi sonuç verir; fakat olasılık tahminini arttırmayı garantilemez. *She*'nin olasılığı benimsenmiştir çünkü çok fazla kullanılan bir kelimedir. Ama bu örnekte *person*'dan sonra gelmesi beklenmemektedir (P(she | person) = 0,009 olduğundan düşük). Benzer bir şekilde *was*'dan sonra *inferior*'un gelmesinin tahmin edilmesi, veri eksikliğini gösterir çünkü P(was | inferior) =0'dır.

3-gram modeli, bu örnek için mükemmel çalışmaktadır. Örnek olarak *person she*'yi takip eden *was* için 0,5 olasılık tahmini verir; fakat genelde 3-gram modeli kullanışlı değildir. Çünkü eğer öncesinde gelen 2-gram hiç görülmemiş ise takip eden kelime için olasılık dağılımı yoktur veya bu 2-gramı takip eden çok az kelime görülmüşse, veri eksikliğinden dolayı tahminler güvenilir değildir. Örnek olarak 2-gram *to both* eğitim metninde dokuz defa görülmüştür, iki defa *to* tarafından takip edilmiş ve 7 defa diğer kelimelerden sonra gelmiştir. Bu durum, hemen olasılık dağılımı oluşturacak kadar yoğun bir durum değildir. Bunların yanında, 4-gram modeli hiç kullanışlı değildir. Genelde eğitim verisi on milyonlarca veriden oluşmadan kullanılmaz. Çizelge 2.3 incelendiğinde şu sonuca varılmaktadır: Yeterli veri olduğunda yüksek dereceli n-gramlar kullanılır, yeterli veri olmadığında düşük dereceli n-gramlar tercih edilmelidir.

Çizelge 2.3: Persuasion verisindeki bir tümcenin kelimeleri için olasılıklar

Inperson	**she**		**was**		**inferior**		**to**		**both**		**sisters**	
1-gram	**P(.)**		**P(.)**		**P(.)**		**P(.)**		**P(.)**		**P(.)**	
1	the	0,034	the	0,034	the	0,034	the	0,034	the	0,034	the	0,034
2	to	0,032	to	0,032	to	0,032	**to**	**0,032**	to	0,032	to	0,032
3	and	0,030	and	0,030	and	0,030			and	0,030	and	0,030
4	of	0,029	of	0,029	of	0,029			of	0,029	of	0,029
...												
8	was	0,015	**was**	**0,015**	was	0,015			was	0,015	was	0,015
...												
13	**she**	**0,011**			she	0,011			she	0,011	she	0,011
…												
254					both	0,0005			**both**	**0,0005**	both	0,0005
…												
435					sisters	0,0003					**sisters**	**0,0003**
…												
1701					**inferior**	**0,00005**						
2-gram	**P (. | person)**		**P (. | she)**		**P (. | was)**		**P (.|inferior)**		**P (. | to)**		**P (. | inferior)**	
1	and	0,099	had	0,141	not	0,065	to	0,212	be	0,111	of	0,066
2	who	0,099	was	0,122	a	0,052			the	0,057	to	0,041
3	to	0,076			the	0,033			her	0,048	in	0,038
4	in	0,045			to	0,031			have	0,027	and	0,025
...												
23	she	0,009							Mrs	0,006	she	0,009
...												
41									what	0,004	sisters	0,009
…												
293									both	0,0004		
…												
∞												
3-gram	**P(.|in,person)**		**P(.|person,she)**		**P (. |she,was)**		**P(.|was,inf.)**		**P (. |inferior,to)**		**P (. | to, both)**	
1	GORULMEDI		did	0,5	not	0,057	GORULMEDI		the	0,286	to	0,222
2			was	0,5	very	0,038			maria	0,143	chapter	0,111
3					in	0,030			cherries	0,143	hour	0,111
4					to	0,026			her	0,143	twice	0,111
...												
∞					inferior	0			both	0	sisters	0
4-gram	**P(.| u,I,p)**		**P(.| I,p,s)**		**P(.| p,s,w)**		**P(.| s,w,i)**		**P(.| w,i,t)**		**P(.| i,t,b)**	
1	GORULMEDI		GORULMEDI		in	1,0	GORULMEDI		GORULMEDI		GORULMEDI	
…												
∞					inferior	0						

Bu seçim çok yaygın olsa da, n-gram tahmin problemlerinde tek başına yeterli bir çözüm değildir. Örneğin eğitim verisinde *was´* ı takip eden pek çok kelime görülmüştür; ancak *inferior,* bu kelimelerden biri değildir. Sonuç olarak, tahminlerin nasıl kombine edildiği göz ardı edilerek, eğitim verisinde görülmeyen olaylara sıfır olasılık değeri atamayan yöntemlere ihtiyaç vardır.

2.3.3.2. Laplace kuralı, Lidstone kuralı ve Jeffreys-Perks kuralı

Laplace kuralı

MLE yönteminin kötü yönlerini elimine etmek için, daha iyi tahmin ediciler kullanılmalıdır. Bunlardan biri olarak, eski bir yöntem olan Laplace kuralı önerilebilir. Bu kurala göre $P_{LAP}(w_1,\ldots,w_n)$ aşağıdaki gibi hesaplanır;

$$P_{LAP}(w_1,\ldots,w_n) = \frac{C(w_1,\ldots,w_n)+1}{N+B} \qquad (2.4)$$

Bu süreç genelde "bir arttırma" olarak bilinir ve görülmeyen olaylara küçük bir olasılık değeri verir. Laplace kuralından elde edilen olasılıklar sözlük boyutuna bağlıdır ve büyük sözlüklerdeki seyrek veriler için (n-gramlar gibi) Laplace kuralı, görülmeyen olaylara daha çok olasılık oranı ayırır.

Church ve Gale (1991) çalışmalarında, belli bir veri grubunun 2-gramlarını çeşitli tahmin edicilerle belirlemeyi amaçlamışlardır. Çalışmada kullanılan veri yığını 44 milyon kelime ve 400,653 farklı kelimeden oluşmaktadır. Böyle bir sözlük boyutunda $1{,}6\text{x}10^{11}$ olası 2-gram söz konusudur. Bu aynı zamanda P_{LAP} hesabında B>N demektir ve Laplace yöntemi bu durumda yetersiz kalmaktadır. Çalışmada veri yığının yarısı (22 milyon) eğitim kümesi olarak kullanılmıştır. Çizelge 2.4'de çalışmadaki çeşitli tahmin edicilerle ve Laplace kuralıyla elde edilen, beklenen sıklık tahminleri verilmiştir. Burada 5 sütun, eğitim verisinde *r* defa görülen bir 2-gram için, farklı tahmin edicilerle hesaplanan sıklıkları göstermektedir. En büyük olabilirlik yöntemi *r* ile, test kümesinde geçerliliğin kullanımı $f_{empirical}$ ile, bir artırım metodu f_{LAP} ile, eğitim verisini kullanan iki yönlü çapraz geçerleme metodu f_{del} ile Good -

Turing tahminleri de f_{GT} ile gösterilmiştir. Son iki sütunda, sıklıkların sıklıkları ve sonraki metinde belli bir sıklıkta görülen 2-gramların hangi sıklıkta olduğu görülmektedir.

Çizelge 2.4: Chuch ve Gale(1991) çalışma verisi için tahmin edilen sıklıklar

r= f_{MLE}	**$f_{empirical}$**	**f_{LAP}**	**f_{del}**	**f_{GT}**	**N_r**	**T_r**
0	0,000027	0,000137	0,000037	0,000027	74,671,100,000	2,019,187
1	0,448	0,000274	0,396	0,446	2,018,046	903,206
2	1,25	0,000411	1,24	1,26	449,721	564,153
3	2,24	0,000548	2,23	2,24	188,933	424,015
4	3,23	0,000685	3,22	3,24	105,668	341,099
5	4,21	0,000822	4,22	4,22	68,379	287,776
6	5,23	0,000959	5,20	5,19	48,190	251,951
7	6,21	0,00109	6,21	6,21	35,709	221,693
8	7,21	0,00123	7,18	7,24	27,710	199,779
9	8,26	0,00137	8,18	8,25	22,280	183,971

Olasılık tahminleri, sıklık tahminlerinin n-gram sayısına bölünmesiyle (N=22 milyon) elde edilebilir. Laplace kuralında, *r* defa görülen bir n-gram için olasılık tahmini=(r+1)/(N+B)'dir ve sıklık tahmini f_{LAP} =(r+1)N/(N+B) haline gelir. Bu şekilde tahmin edilen sıklıkların yorumlanması, olasılıkların yorumlanmasından daha kolaydır, çünkü azaltmanın (discounting) etkisi daha kolay görülebilir.

Bu yöntemle her görülmeyen olaya atanan olasılık değerlerinin toplam değeri aşağıdaki gibidir:

$$N_0 \times P_{LAP}(.) = \frac{74{,}671{,}100{,}000 \times 0{,}000137}{22{,}000{,}000} = 0{,}465 \qquad (2.5)$$

Görülmeyen olaylara düşük bir olasılık atansa da bunların sayısı fazla olduğundan toplamda %46,5 olasılık değeri, görülmeyen olaylara ayrılmıştır. Bu çok büyük bir pay olduğu gibi sıklıkla görülen olayların tahmini, olasılık değerlerini de düşürmektedir. Bu değerin fazla olduğu Çizelge 2.4'ün ikinci sütunundan da anlaşılabilir. Bu sütunda eğitim metninde görülmeyen olayların sonraki metinde hangi sıklıkta görüldüğünün deneysel

olarak belirlendiği tahminler yer almaktadır. İki sütun karşılaştırıldığında, önceki metinde görülmeyen olayların olasılıklarının, Laplace kuralıyla belirlenen olasılıklardan daha düşük olduğu görülür. Bununla beraber deneysel yöntem, test metnindeki 2-gramların %9,2'sinin önceden görülmeyen olaylar olduğunu hesaplamıştır.

Lidstone kuralı ve Jeffrey – Perk kuralı

Laplace kuralında görülmeyen olaylara atanan aşırı büyük olasılık değerinin önüne geçmek için, istatistiksel uygulamalarda kullanılan Lidstone kuralından yararlanılabilir. Bu kurala göre olasılık değerini bir arttırmak yerine, daha küçük pozitif bir λ değeri eklenir:

$$P_{Lid}(w_1,\ldots,w_n) = \frac{C(w_1,\ldots,w_n)+\lambda}{N+B\lambda} \qquad (2.6)$$

Bu yöntem Hardy ve Lidstone tarafından geliştirilmiştir ve Johnson da, yöntemin MLE yöntemi ile düzgün öncelik arasındaki doğrusal interpolasyon olarak gösterilebileceğini vurgulamıştır. μ=N/($N+B\lambda$) olarak alınırsa $P_{Lid}(w_1,\ldots,w_n)$ değeri aşağıdaki gibi olur.

$$P_{Lid}(w_1,\ldots,w_n) = \mu\frac{C(w_1,\ldots,w_n)+\lambda}{N+B\lambda} + (1-\mu)\frac{1}{B} \qquad (2.7)$$

λ için kullanılan en yaygın değer 0,5'dir ve bu seçim, MLE yöntemi tarafından maksimize edilen değerle aynı sayı beklentisinin olduğu, şeklinde teorik olarak doğrulanabilir ve böylece Jeffrey-Park kuralı veya beklenen benzerlik tahmini (Expected Likelihood Estimation - ELE) olarak yeniden isimlendirilebilir (Box ve Tiao 1973).

Genelde bu yöntemler kullanışlıdır, çünkü küçük seçilen λ değerleri ile görülmeyen olaylara ayrılan olasılık payı azaltılabilmektedir. Ancak burada iki problemle karşılaşılmaktadır: ilk olarak uygun λ değerinin seçimi için başarılı bir yönteme ihtiyaç vardır ve ikinci olarak da Lidstone kuralı kullanılarak yapılan azaltma, her zaman MLE sıklığındaki doğrusal olasılık tahminlerini vermektedir ve bu durum, düşük sıklıktaki deneysel (empirical) dağılım için iyi bir eşleşme değildir.

N-gram modellerinde ELE tahmin yönteminin kullanımı

Yöntemlerde anlatılan tüm problemlere rağmen Manning ve Schütze (1999), ELE yöntemini Austen verisine uygulamışlardır. Test yığını cümlesi olan *she was inferior to both sisters* için türetilebilen tek olasılık tahmini 1-gram tahmini -cümlenin olasılığını verir-, Çizelge 2.3'deki 1-gram kelimelerinin olasılıkları çarpıldığında $3{,}96 \times 10^{-17}$ olarak bulunur ($0{,}011 \times 0{,}015 \times 0{,}00005 \times 0{,}032 \times 0{,}0005 \times 0{,}0003 = 3{,}96 \times 10^{-17}$). Diğer modeller için olasılık tahminleri veri eksikliğinden dolayı sıfır veya tanımsızdır.

Yazarlar *she was inferior to both sisters* cümlesinin olasılığını 2-gram modeli ve ELE yöntemini kullanarak hesaplamışlardır. Çizelge 2.3'de de görüldüğü gibi P (not | was) olasılığı MLE yöntemiyle 0,065 olarak belirlenmiştir. Eğitim yığınında *was* 9409 defa görülmüş, bunu takip eden *not*, 608 defa görülmüş ve kelime çeşidi B=14,589 olarak belirlenmiştir. Bu verilere göre eski tahmin;

$P_{MLE}(\text{not} \mid \text{was}) = \frac{608}{9409} = 0{,}065$ iken, ELE yöntemi tahmini;

$P_{ELE}(\text{not} \mid \text{was}) = \frac{(608 + 0{,}5)}{9409 + 14589 \times 0{,}5} = 0{,}036$ olarak bulunur.

Böylece P (not | was) tahmini neredeyse yarı yarıya azaltılmış olur. Eğer aynı hesaplamalar diğer kelimeler için de yapılırsa Çizelge 2.5'deki son sütun elde edilir. Bu tahminler incelendiğinde, kelimelerin sıklıklarına göre sıralanışı değişmemiştir; ancak eğitim verisinde görülen kelimelerin olasılık tahminleri azaltılmıştır ve görülmeyen kelimelere (inferior gibi) sıfırdan farklı olasılık değerleri verilmiştir.

Çizelge 2.5: *was* kelimesini takip eden kelimeler için ELE yöntemi tahminleri

Sıra	Kelime	MLE	ELE
1	not	0,065	0,036
2	a	0,052	0,030
3	the	0,033	0,019
4	to	0,031	0,017
…			
1482	inferior	0	0,00003

Diğer 2-gram olasılıklarının hesaplanmasına aynı şekilde devam edildiğinde, örnek olarak kullanılan *she was inferior to both sisters* cümlesinin olasılığı $6{,}89 \times 10^{-20}$ olarak bulunur. Ancak bu değer MLE tabanlı 1-gram modelinde bulunan olasılık değerinden daha küçüktür ve aslında ELE yöntemiyle, görülen olayların olasılıklarının nasıl azaltıldığının da bir kanıtıdır.

Olasılık tahminleri düşük olsa da bu yöntem sıralama alternatiflerinde kullanılabilir. Örneğin MLE yöntemi, *she was inferior to both sisters* ve *inferior to was both she sisters* ifadelerine, 1-gram modelinde aynı görülme olasılığını atasa da ELE yöntemi doğru bir şekilde, ilk cümlenin ikinci cümleye oranla, İngilizcede daha sık kullanılan bir cümle olduğu sonucuna varır.

2.3.3.3. Doğrulama tahmin yöntemi

Önceki çalışmalardan MLE yöntemiyle, görülmeyen olaylara atanan %46,5 değerinin gereğinden fazla olduğu, deneysel olarak test edilebilir. Bunun için aynı kaynaktan olan bir başka metin alınır ve eğitim metninde *r* defa görülen 2-gramların yeni metinde ne sıklıkta görüldüğüne bakılır. Bu fikrin gerçekleştirilmesi, doğrulama tahmin edicisi (held out estimator) olarak bilinir (Jelinek ve Mercer 1985).

Doğrulama tahmin yönteminde, her n-gram için $w_1,\ldots., w_n$;

$C_1(w_1,\ldots., w_n)$ = eğitim metnindeki $w_1,\ldots., w_n$ sıklığı,

$C_2(w_1,\ldots., w_n)$ = doğrulama (sonraki) metnindeki $w_1,\ldots., w_n$ sıklığı ve

N_r = eğitim metninde *r* sıklığındaki 2-gramların sayısı olarak tanımlansın. T_r, eğitim metninde *r* defa görülen bütün n-gramların doğrulama (held out) metninde toplam görülme sayısıdır ve aşağıdaki formülle hesaplanır:

$$T_r = \sum_{\{w_1,\ldots., w_n \,:\, C_1(w_1,\ldots., w_n)=r\}} C_2(w_1,\ldots., w_n) \qquad (2.8)$$

$$P_{ho}(w_1,\ldots., w_n) = \frac{T_r}{N_r \times N} \;;\; C(w_1,\ldots., w_n) = r \qquad (2.9)$$

Ayrıca n-gramların ortalama sıklığı T_r / N_r olur ve bu n-gramlardan biri için olasılık tahmini $P_{ho}(w_1,\ldots,w_n)$ değerinin hesaplanmasıyla bulunur.

N-gram modellerinde doğrulama tahmin yönteminin kullanımı

Sonraki metin sıklığını tahmin etmede tek başına kullanılan $C(w_1,\ldots,w_n)$ n-gram sıklığının yanında, test kümesinin olasılıklarını maksimize etmek amacıyla azaltılmış olasılık tahminlerini bulmak için test kümesinde doğrulama tahmin yöntemi kullanılır. Bu deneysel ölçümler yardımıyla, eğitim verisinde *r* defa görülen n-gramların, test metninde hangi sıklıkta görüldüğü bulunur. Çizelge 2.4'deki $f_{empirical}$ deneysel tahminleri, bütün veri yığınındaki 44 milyon 2-gramın, eşit boyutta ve rassal olarak eğitim ve test kümesine bölünmesiyle bulunmuştur. Daha sonra eğitim verisinde 22 milyon olan 2-gramların sıklıkları bulunmuş ve test kümesiyle doğrulama tahmini gerçekleştirilmiştir.

2.3.3.4. Modellerin geliştirilmesi ve test edilmesi için veriler

İstatistiksel dil modellemede kullanılan verilerin test edilmesi önemli bir konudur. Verilerin test edilmesiyle mevcut yöntemin ne kadar iyi çalıştığı gözlemlenebilir. Bu da yalnızca, daha önce görülmeyen verilerle olur. Genelde modeller fazla eğitildiklerinde ezberleme olayı gözlemlenmekte, beklenen olaylar eğitim verilerindeki gibi olmakta ve başka olaylara olasılık atanmamaktadır, bu yüzden mevcut modeli farklı verilerde test etmek önemli hale gelmektedir. Bu testlerden en yaygını çapraz entropinin (cross entropy) hesaplanmasıdır. Çapraz entropiyi hesaplamak için büyük bir metin örneği alınır ve bu metindeki her kelimenin entropisi model doğrultusunda hesaplanır. Bu hesaplamalar, modelin kalitesi hakkında bilgi ve dilin entropisi için üst sınır vermektedir. Fakat bunların hepsi test verisinin, eğitim verisinden bağımsız olduğu ve dilin karmaşıklığını ortaya koyacak kadar geniş olduğu durumlarda geçerlidir. Eğer eğitim metninde test yapılırsa, çapraz entropi metnin gerçek entropisinden daha düşük çıkar. Çoğu durumda eğitim verisini ezberleyen ve sonraki kelimeyi 1 olasılıkla tahmin eden modeller kurulabilir. Böyle bir durumda gerçekte doğru olmasa da MLE'nin mükemmel bir dil modelleme yöntemi olduğu yanılgısına düşülebilir. Bu yüzden, verilerle çalışırken toplam küme her zaman

eğitim ve test kümesi olarak ayrılmalıdır. Genelde, toplam verinin %5 – 10 gibi düşük bir kısmı test verisi olarak ayrılır, fakat bu veriler sonuçların güvenilir olması için yeterli olmalıdır.

Çoğunlukla, farklı sebeplerden dolayı eğitim ve test kümesi yine kendi içlerinde ikiye ayrılabilir. Doğrulama tahmin yönteminde olduğu gibi, çoğu istatistiksel dil modelleme metotlarında, eğitim kümelerinin bir tanesinden gerekli bilgiler toplanır, diğeriyle doğrulama verisinde meydana gelebilecek olaylar göz önünde bulundurularak, varsayılan modeldeki belli parametrelerin tahmininde düzgünleştirme yapılır. Doğrulama verisi ilk kullanılan eğitim ve test kümelerinden bağımsız olmalıdır. Normalde doğrulama verisinin kullanıldığı aşamada ilk eğitim verisiyle yapılan tahminlerden daha az sayıda parametre tahmini yapıldığı için doğrulama verisi ilk eğitim verisinden daha küçük olabilir. Yine de parametrelerin gerçeği yansıtacak şekilde tahmin edilmesi için yeterli veri olmalıdır, aksi takdirde önemli performans kayıpları meydana gelebilir (Chen ve Goodman 1996: sayfa 317'de gösterilmiştir).

İstatistiksel doğal dil işleme araştırmasında temel yapı; algoritmanın yazılması, eğitilmesi, test edilmesi, hataların bulunması, revize edilmesi ve aynı sürecin birçok defa tekrarlanması şeklindedir. Fakat bu işlemler çok fazla tekrar edilirse test kümesine olan tarafsız bakış açısı sonlanma eğiliminde olacağı gibi, değiştirilen algoritmaların tekrarlı bir şekilde denenmesi ve performanslarına bakılması, test kümesi içeriğinin yanlı olarak gözlemlenmesine sebep olabilir. Kısaca, revize edilen modellerin arka arkaya test edilmesi, aşırı eğitime neden olmaktadır. Bu yüzden doğru yaklaşım, iki test kümesinin kullanılmasıdır. Bunlardan biri, revize edilen modellerin izlendiği geliştirme test kümesi (development test set) ve diğeri ise, algoritmanın performansıyla ilgili final sonuçlarının üretildiği final kümesidir; ancak final kümesinin performansının, geliştirme test kümesinin performansından düşük olacağı göz önüne alınmalıdır.

Test amaçlı kullanılacak verinin nasıl seçileceği ucu açık tartışmalı bir konu olmasına rağmen temelde iki seçenek vardır. Birinci yöntem, bilgilerin (cümleler veya n-gramlar) veriler arasından rassal olarak seçilip test kümesine katılması ve geri kalanların eğitimde kullanılmasıdır. Eğitim verisiyle test verisinin olabildiğince benzer olması, bu seçeneğin avantajıdır. İkinci yöntem, büyük bitişik yığınların test verisi olarak kümelenmesidir. Bu yöntemin avantajı ise önceki yöntemin aksine, eğitim verisinden küçük farklarla ayrılan

verilerin kullanımından vazgeçilmesidir. Böylece eğitim kümesindeki durağanlığı yansıtan test kümesinin kullanımı engellenmiş olur. Yine de eğer doğrulama tahmin yöntemi kullanılıyorsa, doğrulama verisinin test verisinden ayrı seçilmesi, en iyi stratejidir. Bu şekilde doğrulama verisi test verisini en iyi şekilde yansıtabilir.

Test etme süreci incelendiğinde, önceki çalışmalarda yaygın olarak sistem tek test verisi üzerinde çalıştırılır ve tek performans değeri elde edilirdi. Bu yöntem, sistem performansının varyansını göstermediğinden pek sağlıklı bir test etme yöntemi değildir. Daha iyi bir seçenek ise, test verisinin daha küçük (20 gibi) örneklere bölünmesi ve bu küçük örneklerin test sonuçlarının bulunmasıdır. Bu sonuçlardan yararlanarak ortalama performans değerleri bulunabilir ve performansın değişme eğilimini gösteren varyansı hesaplanabilir. Eğer bu yöntem eğitim verisinin devam eden yığınlarıyla kullanılırsa, verinin farklı bölgelerinden alınan ufak test örneklerinin kullanılması, daha başarılı sonuçlar verir.

Bu şekilde işlemlere devam edilirse bir sistem, kazara veya gerçekten ortalamadan yüksek sonuçlar verebilir ve bu durum, değişkenliği yükselttiği gibi fark edilemeyedebilir. Bu yüzden anlamlı bir karşılaştırmada sadece ortalamaların baz alınması yeterli değildir. Bunun yerine ortalamayı ve değişkenliği hesaba katan istatistiksel bir testin uygulanması, güvenilir sonuçların elde edilmesini sağlar. Sadece istatistiksel test, rastlantısal farklılık olasılığını reddederse, güvenilirlik düzeyiyle beraber bir sistemin diğerinden daha iyi olduğu söylenebilir.

Çizelge 2.6'daki örnekte iki sistemin karşılaştırılmasında *t* testinin kullanımı görülmektedir. İki sistemin varyansının aynı olduğu varsayımıyla (609 ve 526 birbirine yakın değerler) örneklem varyansı s^2'nin toplu tahmini (pooled estimate) kullanılmıştır.

Çizelge 2.6: İki sistemin karşılaştırılmasında *t* testinin kullanımı

	Sistem 1	Sistem 2
Değerler	71,61,55,60,68,49 42,72,76,55,64	42,55,75,45,54,51 55,36,58,55,67
Toplam	609	526
n	11	11
Ortalama $\overline{X_i}$	55,4	47,8
$s_i^2 = \sum \left(x_{ij} - \overline{x_i}\right)^2$	1375,4	1228,8
df	10	10

Her veri kümesinin ortalaması hesaplandığında, varyans hesabındaki payda ve serbestlik dereceleri (11-1)+ (11-1) = 20 olarak bulunur. Bu veriler ışığında sistem 1'in sistem 2'den daha iyi olduğu hipotezini reddetmek için $\alpha = 0,05$ güvenilirlik düzeyinde kritik *t* değeri $t = 1,725$ olmalıdır (20 serbestlik dereceli tek taraflı *t* test için). Fakat örnekte $t = 1,56 < 1,725$ olduğundan, anlamlılık testi başarısız olur ve değerlerin yüksek varyasyonundan dolayı, sistem 1'in daha iyi olduğu söylenemez.

Toplu $s^2 = \frac{1375,4 + 1228,8}{10 + 10} \approx 130,2$

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{2s^2}{n}}} = \frac{55,4 - 47,8}{\sqrt{\frac{2x130,2}{11}}} \approx 1,56$$

2.3.3.5. Çapraz doğrulama yöntemi

Daha önce de bahsedildiği gibi $f_{empirical}$ tahminleri, test verisinde gerçekte ne olduğuna bakılarak oluşturulmuştur. Doğrulama tahmin yönteminde aynı etki, eğitim verisinin ikiye bölünmesiyle elde edilebilir. İlk kısımda sayma işlemiyle başlangıç tahminleri bulunur, diğer kısım ise bu tahminlerin arıtılmasında kullanılır. Bu yöntemin tek maliyeti, başlangıçtaki eğitim verisi azaldığından, tahminlerin güvenilirliğinin de azalmasıdır.

Eğitim verisinin bir kısmının sıklıkları saymada, diğer kısmının da olasılık tahminlerini düzgünleştirmede kullanılmasının haricinde, eğitim verisinin her iki kısmının başlangıç eğitim verisi ve doğrulama verisi olarak kullanıldığı daha etkili yöntemler tercih edilebilir. Genelde istatistikte bu yöntemler çapraz doğrulama (cross–validation) adı altındadır.

Jelinek ve Mercer (1985) çalışmalarında, 2 yönlü çapraz doğrulama yöntemini kullanmışlardır. N_r^a, eğitim verisinin *a* kısmında *r* defa görülen n-gram sayısı ve T_r^{ab}, *b* kısmının içinde bulunan, *a* kısmında görülen toplam n-gram sayısı olarak tanımlanır. Hangi kısmın temel eğitim verisi olarak kabul edildiği dikkate alınarak standart doğrulama tahminleri aşağıdaki gibi hesaplanabilir.

$$P_{ho}(w_1, \ldots, w_n) = \frac{T_r^{01}}{N_r^0 \times N} \; veya \; \frac{T_r^{10}}{N_r^1 \times N} \; ; \; C(w_1, \ldots, w_n) = r \qquad (2.10)$$

Daha etkili çapraz doğrulama tahminleri, her iki parçada sayma ve düzgünleştirme işlemlerinin yapılması ve sonrasında N_r^0'a karşı N_r^1'deki kelimelerin oranlarının alınarak, iki kısmın ağırlıklı ortalamasının bulunmasıyla elde edilir.

$$P_{del}(w_1,\ldots,w_n) = \frac{T_r^{01} + T_r^{10}}{N \times (N_r^0 \times N_r^1)};\ C(w_1,\ldots,w_n) = r \qquad (2.11)$$

Geniş eğitim verilerinde, eğitim kümesi üzerinde çapraz doğrulama tahmin yöntemi, Çizelge 2.4'de de görüldüğü gibi, doğrulama tahmin yönteminden daha iyi sonuçlar vermektedir. Ancak bu yöntemde de düşük sıklıklar problem oluşturur. Yöntem, görülmeyen olaylara yüksek olasılık değeri atar ve bunun yanında eğitim kümesinde görülen olayların beklenen sıklıklarını daha düşük hesaplar. Metin ikiye bölündüğünde, N/2 boyutlu bir metinde bir nesnenin hangi sıklıkta görüldüğü, N/2 boyutlu veride *r* defa görülen bir olayın, N boyutlu veride iki katına çıkacağı, varsayımıyla bulunabilir. Buna rağmen, genelde eğitim metninin bölünmesiyle doğrulama verisindeki görülmeyen n-gram yüzdesi ve buna paralel olarak görülmeyen olayların olasılığı düşmektedir. Bu yüzden küçük eğitim kümelerindeki sayma işlemlerinin toplamı, görülmeyen n-gramların olasılıklarının fazla tahmin edilmesinde etkilidir.

2.3.3.6. Good – Turing tahmini

Good (1953) çalışmasında Turing metodundan yararlanmış ve öğelerin binom dağılımına sahip olduğunu varsayarak bunların frekanslarını veya olasılık tahminlerini belirlemiştir. Bu metot, kapsamlı sözlüklerden çekilen fazla sayıda gözlem için uygun olmakla beraber, kelimeler ve n-gramlar binom dağılımına sahip olmasalar da başarılı sonuçlar üretir. Good – Turing yönteminde olasılık tahmini, *r** düzeltilmiş sıklık olarak düşünüldüğünde, P_{GT}= *r** / N şeklindedir. Önceden gözlemlenmiş durumlar için;

$$r^* = (r+1)\frac{E(N_{r+1})}{E(N_r)} \qquad (2.12)$$

şeklinde bulunur. Burada *E*, rassal bir değişkenin beklenen değerini göstermektedir ve $E(N_1) / N$ değeri, görülmeyen olaylar için ayrılan toplam olasılık değeridir.

Deneysel tahminler kullanılarak, gözlemlenen N_r değeri $E(N_r)$ değeri yerine koyulabilir. Ancak r'nin yüksek değerlerinde deneysel tahminler çok güvensiz olacağından, bu değiştirme düzgün şekilde yapılamaz. Tahminlerdeki güvensizliği ortadan kaldırmak için iki farklı çözüm seçeneğinden yararlanılabilir.

İlk alternatif, belli bir *k* sabiti ile sadece r<k sıklıkları için Good – Turing yöntemini kullanarak yeniden tahmin yapmaktır. Düşük frekanslı kelimelerin çok fazla olmasından dolayı, tahminler için sıklıkların gözlenen frekansının alınması doğru sonuç verir, ayrıca yüksek sıklıklı kelimelerin MLE tahminleri de doğru olur, bu yüzden azaltmaya gerek yoktur.

Diğer alternatif ise (r, N_r) gözlem değerleri baz alınarak bir *S* fonksiyonu oluşturmak ve tahminler için S(r) fonksiyonunun düzgünleştirilmiş değerlerini kullanmaktır. Görülmeyen olaylara ayrılan olasılık yığını N_1/N, görülmeyen olaylar arasında düzgünce dağıtılabilir veya başka metotlar tarafından da kullanılabilir (tahmin edicilerin kombine edilmesi). Sonuç olarak, görülmeyen olaylar için düzgün dağılımla olasılıkların tahmin edilmesi aşağıdaki formül yardımıyla gerçekleştirilebilir:

Eğer $C(w_1,\ldots, w_n) = r > 0$;

$$P_{GT}(w_1,\ldots,w_n) = \frac{r^*}{N} \;;\; r^* = \frac{(r+1)S(r+1)}{S(r)} \tag{2.13}$$

Eğer $C(w_1,\ldots, w_n) = 0$;

$$P_{GT}(w_1,\ldots,w_n) = \frac{1-\sum_{r=1}^{\infty} N_r \frac{r^*}{N}}{N_0} \approx \frac{N_1}{N_0 N} \tag{2.14}$$

Gale ve Sampson (1995) yaptıkları çalışmada basit ve etkili bir yaklaşım sergilemişler ve yaklaşımları birleştirerek Basitleştirilmiş Good – Turing metodunu geliştirmişlerdir. Düzgünleştirme eğrisi olarak (smoothing curve) $N_r = ar^b$ eğrisini kullanmışlar (b<-1 ile uygun hiperbolik ilişki verir) ve bu eşitliğin logaritmik formu olan $N_r = a + b\log r$ formülünde basit lineer regresyonla, A ve *b* değerlerini tahmin etmişlerdir. Bu basit eğri,

yalnızca yüksek *r* değerleri için uygun olduğundan, düşük *r* değerlerinde, direkt olarak hesaplanan N_r kullanılabilmektedir. Bu frekanslarla çalışırken direkt tahminler, doğrudan veya düzgünleştirme fonksiyonuyla hesaplanan *r** değerleri arasında anlamlı bir farklılık olmayıncaya kadar kullanılır ve daha sonra düzgünleştirilmiş tahminler, bütün yüksek frekanslar için kullanılır. Basitleştirilmiş Good – Turing metodu ile Çizelge 2.4'de f_{GT} sütununda görüldüğü gibi başarılı tahminler yapılabilmektedir.

Bu yaklaşımlar altında uygun olasılık dağılımlarını sağlamak için, bütün tahminlerin yeniden normalleştirilmesi (renormalize) önemlidir. Yeniden normalleştirme, ya görülmeyen olaylara atanan olasılık miktarının düzeltilmesiyle veya görülmeyen olayların olasılık toplamlarının N_1/N'de tutulup, görülen olayların olasılıklarının yeniden normalleştirilmesiyle gerçekleştirilebilir.

Good – Turing tahmin yönteminin test verisinde kullanımı

Good – Turing yönteminde ilk adım, farklı sıklıkların frekanslarının hesaplanmasıdır. Çizelge 2.7'de Austen verisinden elde edilen 2-gram ve 3-gramların sıklıklarının frekansları verilmiştir (Sayılar Zipfian dağılımına göre benzer gözükebilir ancak yapının detaylarında farklılık vardır ve kelimelerin sıklığı sayıldığından daha abartılıdır). Çizelge 2.8'de de yeniden tahmin edilen *r** değerleri ve bunlara bağlı olarak hesaplanan 2-gram olasılıkları yer almaktadır.

2-gramlar için, görülmeyen 2-gramlara ayrılan olasılık $N_1/N = 138741/617091 = 0{,}2248$ olarak bulunur. 2-gram uzayı, sözlük boyutunun karesidir ve 199,252 adet 2-gram görülür, böylece düzgün dağılımlı tahminler kullanıldığında, her görülmeyen 2-gram için olasılık tahmini $0{,}2248/(14585^2 - 199252) = 1{,}058 \times 10^{-9}$ olarak bulunur. 2-gram modeli için şartlı olasılık tahminlerinin hesaplanmasında; Good – Turing yöntemiyle hesaplanan 2-gram olasılıkları ile MLE yöntemiyle belirlenen 1-gram olasılıkları aşağıdaki gibi kullanılabilir:

$$P(\text{she} \mid \text{person}) = \frac{f_{GT}(\text{person she})}{C(\text{person})} = \frac{1{,}228}{223} = 0{,}0055$$

Çizelge 2.7: Austen yığınında 2-gram ve 3-gramlar için frekans dağılımı sıklıkları

2-gramlar					3-gramlar			
r	N_r	r	N_r		r	N_r	r	N_r
1	138741	28	90		1	404211	28	35
2	25413	29	120		2	32514	29	32
3	10531	30	86		3	10056	30	25
4	5997	31	98		4	4780	31	18
5	3565	32	99		5	2491	32	19
6	2486		…		6	1571		…
7	1754	1264	1		7	1088	189	1
8	1342	1366	1		8	749	202	1
9	1106	1917	1		9	582	214	1
10	896	2233	1		10	432	366	1
	…	2507	1			…	378	1

Çizelge 2.8: 2-gramlar için Good – Turing tahminleri, düzeltilmiş sıklıklar ve olasılıklar

r	**r***	**P_{GT}(.)**	**r**	**r***	**P_{GT}(.)**	**r**	**r***	**P_{GT}(.)**
0	0,007	$1{,}058 \times 10^{-9}$	8	6,942	$1{,}134 \times 10^{-5}$	32	30,84	5035×10^{-5}
1	0,3663	$5{,}982 \times 10^{-7}$	9	7,928	$1{,}294 \times 10^{-5}$	…		
2	1,228	$2{,}004 \times 10^{-6}$	10	8,916	$1{,}456 \times 10^{-5}$	1264	1263	0,002062
3	2,122	$3{,}465 \times 10^{-6}$	…			1366	1365	0,002228
4	3,058	$4{,}993 \times 10^{-6}$	28	26,84	$4{,}383 \times 10^{-5}$	1917	1916	0,003128
5	4,015	$6{,}555 \times 10^{-6}$	29	27,84	$4{,}546 \times 10^{-5}$	2233	2232	0,003644
6	4,984	$8{,}138 \times 10^{-6}$	30	28,84	$4{,}709 \times 10^{-5}$	2507	2506	0,004092
7	5,96	$9{,}733 \times 10^{-6}$	31	29,84	$4{,}872 \times 10^{-5}$	1264	1263	0,002062

Bu şekilde devam edildiğinde Çizelge 2.4 ile karşılaştırılabilen Çizelge 2.9'daki sonuçlar elde edilir.

Çizelge 2.9: Persuasion verisindeki cümle için Good – Turing 2-gram sıklık tahminleri

P (she | person)	0,0055
P (was | she)	0,1217
P (inferior | was)	$6{,}9\text{x}10^{-8}$
P (to | inferior)	0,1806
P (both | to)	0,0003956
P (sisters | both)	0,003874

Bu değerler çarpılarak *she was inferior to both sisters* ve *inferior to was both she sisters* ifadesi için olasılık tahmini $1{,}278\text{x}10^{-17}$ olarak bulunur. Elde edilen bu tahmin değeri ELE yöntemiyle bulunan tahmin değerinden daha yüksek olsa da, görülmeyen 2-gramların düzgün dağılıma uyduğu varsayımı problem oluşturmaktadır.

Ney ve Essen (1993) ile Ney ve arkadaşları (1994) yaptıkları çalışmalarda, iki farklı azaltma modeli önermişlerdir. Birinci model olarak Mutlak azaltma (Absolute discounting) yönteminde, bütün sıfırdan farklı MLE sıklıkları, küçük sabit bir δ değerine indirgenir ve kazanılan frekans, görülmeyen olaylara düzgün dağılıma göre dağıtılır. Mutlak azaltma aşağıdaki gibidir:

Eğer $C(w_1,\ldots., w_n) = r$ ise ve B küme sayısını gösterdiğinde;

$$P_{abs}(w_1,\ldots., w_n) = \begin{cases} (r-\delta)/N & ;r>0 \\ \dfrac{(B-N_0)\delta}{N_0 N} & diger \end{cases} \qquad (2.15)$$

İkinci model olan doğrusal azaltmada (linear discounting) sıfırdan farklı olan MLE sıklıkları, 1'den küçük bir sabitle ölçeklendirilir ve kalan olasılık yığını yeni olaylar arasında dağıtılır. Doğrusal azaltma aşağıdaki gibi hesaplanır:

Eğer $C(w_1,\ldots., w_n) = r$ ise;

$$P(w_1,\ldots., w_n) = \begin{cases} (r-\alpha)/N & ;r>0 \\ \alpha/N_0 & diger \end{cases} \qquad (2.16)$$

Bu tahmin yöntemleriyle, görülmeyen olaylara sıfırdan farklı olarak küçük bir ε sayısı kadar olasılık atanır ve diğer olasılıklar toplamları 1'e eşit olacak şekilde ayarlanır. Genelde iyi bir δ değeri tahmin edebilmek için doğrulama verisinden yararlanılabilir. Mutlak azaltma yaklaşımı ve genişletilmiş uygulamaları başarılı olmasına rağmen, bunun yanında doğrusal azaltma yönteminin uygulanması daha zordur. Genelde, eğitim verisinde sıklıklar yükseldikçe, düzeltilmemiş MLE yöntemi doğrusal azaltma yönteminden daha başarılı sonuçlar üretir.

Diğer taraftan, Lidstone kuralının kusurlu olmasının sebebi modeldeki küme (bin) sayısına bağlı olmasıdır. Bazı boş kümeler veri eksikliğinden ortaya çıksa da, genelde daha fazlası yöntemin işleyişinden kaynaklanmaktadır. Bunun yanında Good – Turing metodunda önceden görülen olayların tahminleri küme sayısına bağlı değildir. Ristad (1995) yaptığı çalışmada, doğal sıralamaların sadece olası kümelerin bir altkümesini kullandığı hipotezini araştırmıştır. Çalışmasında gözlenen sıklığı $C(w_1,\ldots, w_n) = r$ olan bir n-gram için olasılık tahmini aşağıdaki gibidir:

$$P_{NLS}(w_1,\ldots,w_n) = \begin{cases} \dfrac{r+1}{N+B} & ;N_0 = 1 \\ \dfrac{(r+1)(N+1+N_0-B)}{N^2+N+2(B-N_0)} & ;N_0 > 0 \quad ve \quad r > 0 \\ \dfrac{(B-N_0)(B-N_0+1)}{N_0(N^2+N+2(B-N_0))} & diger \end{cases} \qquad (2.17)$$

Bu kuralın temel özelliklerinden biri Laplace kuralını, bir olayın her kümede görüldüğü durumlarda azaltmasıdır. Ayrıca görülmeyen olaylara ayrılan olasılık büyüklüğü, deneme sayısına bağlı olarak azaldığı gibi küme sayısından bağımsızdır bu yüzden, geniş sözlüklerde doğru bir şekilde kullanılabilir.

2.3.4. Tahmin yöntemlerinin birleştirilmesi

Anlatılan yöntemlerin hepsi, bir n-gramın *r* sıklığını kullanarak gelecek metin içinde en iyi tahminle, görülme olasılığını belirlemeye çalışmaktadır. Şimdiye kadar görülmemiş veya çok nadiren görülme ihtimali olan bütün n-gramlara aynı tahminlerin verilmesi yerine, n-

gramın içinde bulunan (n–1)-gramların sıklığına bakılarak, daha iyi tahminler üretilebilir. Eğer (n–1)-gramlar nadiren görülüyorsa, n-grama düşük olasılık verilir. Eğer (n–1)-gramlar yeterli sıklıktaysa, n-grama yüksek olasılık verilir. Fakat (n–1)-gramlar çok yüksek sıklıktaysa, görülmeyen olaylara da olasılık değeri ayrılabilmesi için olasılık değeri düşürülebilir. Church ve Gale (1991) yaptıkları çalışmada, görülmeyen 2-gramların olasılık tahminlerinin, bunları oluşturan 1-gramların olasılıkları açısından nasıl hesaplandığını göstermişlerdir. Yazarlar, görülmeyen 2-gramlar için $P(w_1)P(w_2)$ ortak bir bağımsız olasılık değeri hesaplamışlar ve buna bağlı olarak 2-gramları kümeler halinde gruplamışlardır. Daha sonra da Good – Turing yöntemini, normalleştirilmiş olasılıkları bulabilmek için her kümede uygulamışlardır.

Daha önceden bahsedilen yöntemler gibi, geçmişe dayanarak gelecek hakkında tahminler yapabilen araçlar birleştirilerek, tahmin sonuçları iyileştirilebilir. Aslında n-gram modelleri için farklı yöntemlerin birleştirilerek kullanılması, daha başarılı sonuçların elde edilmesini sağlar. Örneğin, MLE n-gram modeli ile basit doğrusal interpolasyon yöntemi beraber kullanıldığında iyi bir dil modeli elde edilmektedir (Chen ve Goodman 1996). Sonuç olarak, önceden bahsedilen tahmin yöntemleri basit doğrusal intepolasyon, Katz geri çekilme ve genel doğrusal interpolasyon teknikleri ile beraber kullanıldığında daha başarılı sonuçlar elde edilebilir.

2.3.4.1. Basit doğrusal interpolasyon

3-gram modelindeki veri eksikliği problemini çözmenin yollarından biri, aynı şekilde veri eksikliği bulunan 2-gram ve 1-gram modellerini karıştırmaktır. Çoklu olasılık tahminlerinin olduğu her yerde bunlar arasında bir lineer kombinasyon yapılır ve her birinin katkısı ağırlıklandırılırsa, sonuçta başka bir olasılık fonksiyonu elde edilir. İstatistiksel dil modellemede buna doğrusal interpolasyon veya (sonlu) karışık modeller denir. Fonksiyonların ara değerleri hesaplanırken hepsi en ayırıcı fonksiyonun iyileştirme bilgisinin alt kümesini kullanır (3-gram kombinasyonunda 2-gramlar ve 1-gramlar gibi) ve metot genelde çapraz doğrulama olarak bilinir. N-gram modellerini birleştirmek için en yaygın yol $0 \leq \lambda_i \leq 1$ ve $\sum_i \lambda_i = 1$ olduğunda aşağıdaki gibidir:

$$P_{li}(w_n|w_{n-2},w_{n-1})=\lambda_1 P_1(w_n)+ \lambda_2 P_2(w_n|w_{n-2},w_{n-1})+ \lambda_3 P_3(w_n|w_{n-1},w_{n-2}) \quad (2.18)$$

Burada ağırlıklar elle belirlenebileceği gibi, genelde en iyi sonuçları veren ağırlık kombinasyonlarının kullanılması hedeflenir. Bu belirleme çeşitli sayısal algoritmalarla gerçekleştirilebilir. Örneğin Chen ve Goodman (1996) çalışmalarında Powell algoritmasını kullanmışlar ve bu basit kombinasyonun başarılı sonuçlar verdiğini göstermişlerdir.

2.3.4.2. Katz Geri Çekilme (Backing-off)

Geri çekilme modellerinde, modellerin özellikleri kullanılarak birçok farklı model geliştirilebilir. Model ne kadar detaylandırılırsa, kullanılan metin hakkında o kadar güvenilir bilgi sağladığı varsayılır. Geri çekilme modeli, düzgünleştirme veya bilgi kaynaklarının kombinasyonunda kullanılabilir.

Geri çekilme modeli Katz (1987) tarafından geliştirilmiştir. Bir n-gram için yapılan tahmin, artan bir şekilde kısa geçmişe geri dönüşlere izin verir.

$$P_{bo}(w_i \mid w_{i-n+1}....w_{i-1}) = \begin{cases} \left(1 - d_{w_{i-n+1}....w_{i-1}}\right)\dfrac{C(w_{i-n+1}....w_i)}{C(w_{i-n+1}....w_{i-1})} & ; C(w_{i-n+1}....w_i) \\ \alpha_{w_{i-n+1}....w_{i-1}} P_{bo}(w_i \mid w_{i-n+2}....w_{i-1}) & ; \text{diger} \end{cases} \quad (2.19)$$

Eğer ilgilenilen n-gram *k* (k genelde 0 veya 1 olur) seferden fazla görülüyorsa formülün ilk sırasındaki gibi, n-gram tahminleri kullanılır. Ancak MLE tahmini belli bir değere indirgendiğinde (d fonksiyonuyla gösterilir), geri çekilme yöntemiyle tahmin edilecek olan görülmeyen olayların olasılıkları için olasılık yığını ayrılmış olur. Düşük sıradaki modellere olasılık değeri ayrılması için MLE tahminleri azaltılmalıdır. Azaltma miktarının belirlenmesinde, Good – Turing yönteminden veya Katz metodundan yararlanılabilir. Eğer eğitim verisinde n-gramlar görülmezse ya da *k* defa veya daha az görülürse, kısa n-gramlardan biri tahmin için kullanılır. Ayrıca bu geri çekilme olasılığı *α* normalleştirme fonksiyonuyla çarpılmalıdır, böylece sadece azaltma sürecinden elde edilen olasılık yığını, n-gramlar arasında paylaştırılmış olur. Hemen önce gelen geçmiş, yani (n–1)-gram görülmediğinde formüldeki ilk sıra, w_i'nin herhangi bir değeri için uygulanamaz ve *α* geri

çekilme faktörü 1 değerini alır. Eğer hesaplamalar için formülün ikinci sırası seçilirse, tahminler (n–1)-gram tahminleri yoluyla tekrarlı olarak yapılır. Bu tekrarlar, bir 4-gram modeliyle başlayıp sonraki kelimeyi 1-gram sıklıklarıyla tahmin etmeyle son bulacak şekilde azalarak devam edebilir.

Bu yöntem, veri eksikliğinde geçerli bir metot olmasına rağmen bazı durumlarda çok kötü tahminlere sebep olur. Eğer herhangi bir w_iw_j 2-gramı çok defa görülürse ve w_k genel bir kelime olmasına rağmen $w_iw_jw_k$ 3-gramı hiç görülmezse, bunun önemli olduğu varsayılmalıdır. Hatta rutin olarak geri çekilme yönteminin uygulanıp $P(w_k \mid w_j)$ 2-gram tahminiyle $P(w_k \mid h)$ tahminin gerçekleştirilmesi yerine, "dilbilgisel olarak sıfır" şeklinde temsil edilmesi gerekmektedir. Rosenfeld ve Huang (1992) yaptıkları çalışmada bu durumu düzelten bir geri çekilme modeli geliştirmişlerdir.

Geri çekilme algoritması, tahmine dayalı olarak farklı sırada bir n-gram seçtiğinde veri sayısını arttırdığından, olasılık tahminlerinde ani değişikliklere sebep olur ve bu durum, yöntemin eleştirilen tarafıdır. Yine de yöntem basit ve pratik olduğundan tercih edilebilir.

2.3.4.3. Genel doğrusal interpolasyon

Basit doğrusal interpolasyonda kullanılan ağırlıklar sadece birer sayıdır; fakat bunun yerine ağırlıkların, geçmişin fonksiyonu olarak kullanıldığı daha genel ve güçlü modeller tanımlanabilir. Olasılık fonksiyonu k ve genel formu P_k olan bir doğrusal interpolasyon modeli aşağıdaki gibi tanımlanabilir:

$$P_{li}(w|h) = \sum_{i=1}^{k} \lambda_i(h) P_i(w|h)\,;\ \forall h,\quad 0 \le \lambda_i(h) \le 1 \quad ve \quad \sum_i \lambda_i(h) = 1. \tag{2.20}$$

Doğrusal interpolasyon, modellerin birleştirilmesinde çok genel bir yol olduğu için sıklıkla kullanılır. Fakat bu yöntem, farklı ağırlıklı geçmişlerin ayrılmasında dikkatli olunmadığı takdirde tamamlayıcı modellerin kötü sonuçlar vermesine sebep olur. Örneğin λ_i bir n-gram modelinin interpolasyonundaki katsayılar ise, 3-gram tahmininin iyi veya kötü olduğu göz ardı edilerek, 1-gram tahmini her zaman aynı ağırlıkla kullanılabilir.

Genelde ağırlıklar bireysel geçmişlerine göre kümelenmezler. Her $w_{(i-n+1)(i-1)}$ için ayrık bir $\lambda_{w(i-n+1)(i-1)}$ eğitilmesi genelde uygun değildir, çünkü veri eksikliği problemini kötüleştirebilir. Bunun yerine, geçmişlerin eşdeğerlik sınıfları kullanılabilir. Bahl ve arkadaşları (1983), λ değerinin $C(w_{(i-n+1)(i-1)})$'e bağlı olarak kümelere bölünmesini önermişler ve aynı frekanstaki bütün geçmişlerin parametrelerini bağlamışlardır.

Chen ve Goodman (1996) λ parametrelerini kümelere koymak yerine, sıfır olmayan eleman başına ortalama sayıya göre gruplamanın daha iyi olduğunu göstermişlerdir.

$$\frac{C\left(w_{(i-n+1)(i-1)}\right)}{\left|w_i : C\left(w_{(i-n+1)(i-1)}\right) > 0\right|}$$

Burada w_{i-n+1} ….$w_{i-1}w^x$ n-gramlar için, sıfır olmayan sayılar üzerinden hesaplanan ortalama sayı alınır. Bu yöntemin çalışmasının sebebi, dilin yapısı olabilir; çünkü güçlü yapısal kısıtlamalar, birbirini takip eden kelimeler arasında kurallar oluşturur. Bazı n-gramlar hiç görülmediği gibi, bazıları da dilin gramatik yapısına uymadığından yeni bir tümce oluşturduklarında "dilbilgisel olarak sıfır" adını alırlar. Örneğin Austen eğitim verisinde, *great deal* ve *of that* 2-gramlarının ikisi de 178 defa görülmüştür. Veri grubunda *of that* ifadesi 115 farklı kelime tarafından takip edilmiş ve 1,55 ortalama sayısı (178/115) elde edilmiştir. Bu durum isim tümcesinin sıfat, zarf veya isim tarafından uygun bir şekilde takip edilebildiğinin yanında, herhangi bir kelimeyle yeni cümleye başlanmasının da olası olduğunu yansıtır. Burada genelde fiiller olmak üzere çok az "dilbilgisel olarak sıfır" vardır. Diğer taraftan, *great deal* 36 kelime tarafından takip edilmiş ve ortalama sayısı 4,94 olarak bulunmuştur. Burada yeni bir cümleye başlangıç olası olduğu gibi, dil bilgisi yapıları bağlaçlar ve sıfatların karşılaştırmalı halleri ile kısıtlanmıştır. Genele bakıldığında *of* kelimesinin takip oranı %38'dir. Yüksek çıkan ortalama değeri, bu 2-gramı takip eden "dilbilgisel olarak sıfır" olan kelimelerin fazla sayıda olduğunu gösterir ve bu durumda, yeni görülmemiş kelimelere, bu metinde görülme ihtimallerine düşük olasılıkların verilmesi, doğru bir yaklaşımdır.

2.3.4.4. Austen verisi için dil modelleri

Austen verisi için dil modelleri yapılandırılırken, interpolasyon ve geri çekilme modellerinden yararlanılabilir. Manning ve Schütze (1999), Katz yaklaşımı temelinde Good – Turing tahmin yöntemiyle geri çekilme dil modellerini geliştirmişlerdir. Daha sonra test kümesi olarak *Persuasion* verisini kullanarak, bu modellerin çapraz entropisini ve çapraşıklığını hesaplamışlardır. Çizelge 2.10'da çalışma sonuçları görülmektedir.

Çizelge 2.10: Persuasion verisinde test edilen Good – Turing tahminleriyle oluşturulan geri çekilme modelleri

Model	Çapraz Entropi	Çapraşıklık
2-gram	7.98 bit	252.3
3-gram	7.90 bit	239.1
4-gram	7.95 bit	247.0

Ayrıca örnek tümce için her kelimenin olasılık tahmini ve kullanılan n-gram büyüklüğü Çizelge 2.11'de verilmiştir. Bu çizelgedeki 1-gram tahminleri, önceden hesaplanan MLE tahminleridir. Diğer iki tahmin ise geri çekilme dil modelleri ile hesaplanmıştır. Çizelgedeki son sütunda ise verilen örnek tümce için modellerin hesapladığı tüm olasılık tahminleri yer almaktadır.

Çizelge 2.11: Çeşitli dil modellerine göre test cümleciği olasılık tahminleri

	P(she|h)	P(was|h)	P(inferior|h)	P(to|h)	P(both|h)	P(sisters|h)	Çarpım
1-gram	0.011	0.015	0.0005	0.032	0.0005	0.0003	3.96×10^{-17}
2-gram	0.00529	0.1219	0.0000159	0.183	0.000449	0.00372	3.14×10^{-15}
Kullanılan n	2	2	1	2	2	2	
3-gram	0.00529	0.0741	0.0000162	0.000384	0.000384	0.00323	1.44×10^{-15}
Kullanılan n	2	3	1	2	2	2	

Bu son çalışmayla beraber 1-gram olasılık tahminleri başlangıçtaki tahminlerden daha yüksek çıkmıştır. 3-gram modeli, test verisindeki 2-gram modelinden daha iyi sonuçlar vermesine rağmen, 2-gram modeli daha yüksek olasılıklar atamaktadır. Genele bakıldığında

4-gram modeli, 3-gram modelinden daha düşük performansa sahiptir ve bu durum eğitim verisinin 4-gram tahminleri için yeterli büyüklükte olmamasından dolayı, beklenen bir sonuçtur. Ayrıca geri çekilme modellerinin, basit uzun tümceleri göz ardı ettiği için çok başarılı olduğu söylenemez ve uygun büyüklükte veri olsa dahi uzun n-gramlarda model, kötüleşme eğilimi göstermektedir.

Sonuç olarak, Good – Turing tahmin yöntemi ile doğrusal interpolasyon veya geri çekilme yöntemleri bir arada kullanılarak veri eksikliğinden kaynaklanan hatalar giderilebilir ve başarılı sonuçlar elde edilebilir. Chen ve Goodman (1996, 1998) çalışmalarında farklı düzgünleştirme algoritmalarını değerlendirmişlerdir. Bu değerlendirmeler sonucunda geliştirdikleri Kneser-Ney geri çekilme düzgünleştirmesinin farklı bir versiyonunun, en iyi performansı verdiği sonucuna varmışlardır (Chen ve Goodman 1998). Church ve Gale (1991) yaptıkları çalışmada 2 milyondan fazla kelimeye sahip bir metinde 2-gram modellerini eğitirken kullandıkları Good – Turing düzgünleştirme metoduyla, Kneser-Ney geri çekilme yöntemiyle bulunan sonuçların ötesine geçmişlerdir. Sonuç olarak basit düzgünleştirme teknikleri bu tip çalışmalar için uygun olsa da, optimal performanslı sistemlerin üretilmesinde kullanımından kaçınılmalıdır.

2.4. Karakter N-gram Yöntemi

2.4.1. Genel bilgiler

Daha önce de bahsedildiği gibi N-gram dil modelleri, sıradaki kelimenin görülme olasılığının ondan önceki n–1 kelimeye dayandığını varsayar. Karakter n-gram yöntemi ise aynı yaklaşımı karakterlerin görülme sıraları için yapmaktadır.

Karakter n-gram yönteminin literatür taraması yapıldığında daha çok dil tanımlama çalışmalarında kullanıldığı görülmektedir. Dil modellemede karakter n-gram metodlarının kullanımı Claude Shannon (1951) ile başlamıştır. Yazar, çalışmasında İngilizce için ardışık karakter n-gram ve kelime n-gram tahminlerini tanımlamıştır (McNamee ve Mayfield 2004). Shannon'dan sonra karakter n-gram yöntemi pek çok çalışmada kullanılmıştır. McNamee ve Mayfield (2004) çalışmalarında çok dilli metin erişiminde karakter n-gram yöntemini kullanarak, bu yöntemin diğer dil modellerine kıyasla daha doğru erişim sağladığını ortaya

koymuşlardır. Liu ve Keselj (2007), web sayfalarının içeriklerini karakter n-gram yöntemiyle belirleyerek, web kullanıcı dolaşım yapılarını (navigation pattern) otomatik olarak sınıflandırmışlardır. Kamaris ve Stamatatos (2007) çalışmalarında, arama motorlarının kalitesi arttırmak için web sayfası tiplerini belirlemede karakter n-gram yöntemini kullanmışlardır. Chau ve arkadaşları (2009) Çin arama motorlarında yapılan Çince aramalarda karakter kullanımını araştırmışlardır. Karakter n-gram algoritması dilden bağımsız olduğu için, yazarlar bu yöntemi çalışmalarında rahatlıkla kullanabilmişlerdir. Bu çalışmaların yanında karakter n-gram yöntemi el yazısı tanımlamada da kullanılmış ve başarılı sonuçlar elde edilmiştir. (El-Nasan ve Perrone 2002, Senda ve Yamada 2001).

Karakter n-gramdan farklı olarak 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.

2.4.2. 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 Çizelge 2.12'deki gibi verilebilir.

Çizelge 2.12: 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, Çizelge 2.13'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 değişimi tahmini yapar. Dolayısıyla, ardışık sorguların benzer olup olmadıkları ve yöntemin tahminleri, kullanıcı tarafından belirlenen *n* sayısına ve eşik değerine bağlıdır.

Çizelge 2.13: 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 arama yapılarından "yeni", "ilgili geri besleme" ve "diğer" kategorileri haricinde arama yapısına sahip olan ardışık sorgularda benzer kelimelerin bulunduğu görülür. İstatistiksel yöntemler, hatalı tahminlerini "yeni" arama yapısına sahip sorgularda yaptıkları için, sorgular bu arama yapısından çıkarıldıklarında "sonraki sayfa", "genelleştirme", "özelleştirme" veya "düzenleme" arama yapılarından birine dahil edilecek ve bu değişiklik hatalı tahminleri önlemek adına yeterli olacaktır. Kısaca karakter n-gram yöntemi, belirli bir *n* sayısı ve eşik değerine göre, birden fazla kelime içeren ardışık iki sorguda, kelimeleri karşılaştırarak benzer kelime sonucuna varıyorsa, bu ardışık iki sorgu için konu devamı tahmini yapmaktadır. Örnek olarak FAST verisinde bulunan aşağıdaki sorgular ele alınsın:

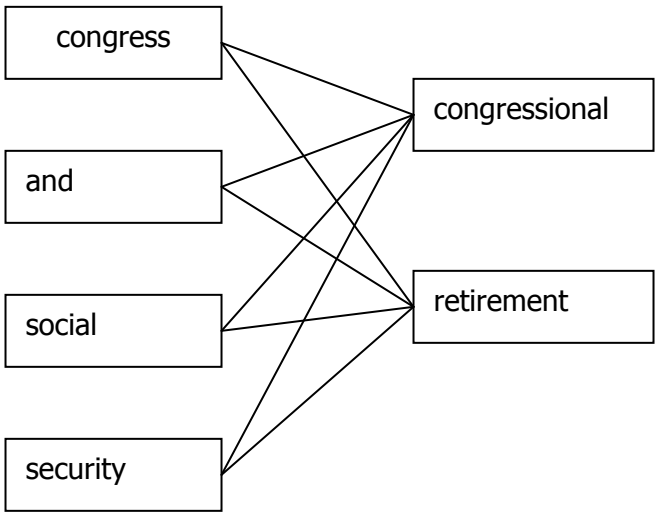
q_i = congress and social security

q_{i+1}= congressional retirement

Karakter n-gram yöntemi, sorguları ilk önce Şekil 2.1'deki gibi kelimelerine ayırır, daha sonra her kelimenin karakter n-gramını Çizelge 2.14'deki gibi bulur. Karşılaştırma aşamasında yine Şekil 2.1'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.

Çizelge 2.14: 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 2.1: Bir kelimeden fazla sorguların karşılaştırılma yöntemi.

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

Adım 0: i=1.

Adım 1: Karakter n-gram yöntemi, kullanıcının aynı oturum içindeki sorgularına uygulanır. Sorgular sırasıyla "önceki sorgu" (q_i) ve "sonraki sorgu" (q_{i+1}) olarak nitelendirilir. Belirlenen ardışık iki sorgu, kelimelerine ayrılır ve her kelime bir bilgisayar programında bir diziye atanır.

Adım 2: Kelimeler seçilen *n* büyüklüğüne göre karakterlerine ayrılarak dizilere atanır.

Adım 3: Sorgulardaki kelimelerin dizi elemanları diğer bir ifadeyle n-gramlar, karşılaştırılır. Kelimelerdeki dizi elemanlarının benzerlik oranı, eşik değerine eşit veya daha büyükse, bu iki sorgu arasında konu devamı tahmini yapılır. Eğer benzerlik oranı eşik değerinin altında kalırsa konu değişimi kararı verilir.

Adım 4: i sayısının değeri 1 arttırılır, i=i+1. Eğer (q_i) oturumun son sorgusu ise, algoritma sonlandırılır, değilse karşılaştırmaya devam edilir ve Adım 1'e geri dönülür.

2.4.3. Kullanılan veriler

Bu çalışmada kullanılan veriler, Excite (http://www.excite.com) ve FAST (http://www.fast.com) arama motorlarından alınmıştır. Arama motorlarından alınan

verilerde arama motorlarını kullanan kullanıcılar hakkında bilgiler vardır. Bu bilgiler, İnternet Protokolü (Internet Protocol – IP) adresi, arama zamanı ve sorgudan oluşmaktadır. Özmutlu ve Çavdur (2005a) yaptıkları çalışmada, arama zamanları arasındaki farklardan yararlanarak, zaman aralıkları (time interval – *ti*) ile aynı IP adresine ait ardışık aramalarda girilmiş olan sorguları inceleyerek bunlar arasındaki yapısal ilişkileri gösteren arama yapısı (search pattern – *sp*) sınıflarını tanımlamışlardır. Tanımlanan sınıflar Çizelge 2.15'de gösterilmektedir. Sorgulara ait bu bilgiler, sonraki çalışmalarda kullanılan yöntemlerin karar verme mekanizmalarında baz aldıkları verileri oluşturmuştur (Özmutlu ve Çavdur 2005b, Özmutlu 2006, Özmutlu ve ark. 2006, Özmutlu ve ark. 2007, Özmutlu ve ark. 2008a, Özmutlu ve ark. 2008b, Özmutlu ve ark. 2008c). Çalışmalarda uygulanan yöntemlerin konu değişim kararları, arama yapılarının ve zaman aralıklarının kombinasyonları temelinde incelenmiş ve hatalı değerlendirmelerin sebepleri daha iyi araştırılabilmiştir.

Çizelge 2.15: Sorgu bilgilerinden elde edilen arama yapısı ve zaman aralığı sınıfları.

Sınıf	**Zaman Aralıkları (dk)**	**Sınıf**	**Arama Yapısı**
1	0 – 5	1	Yeni
2	5 – 10	2	Sonraki Sayfa
3	10 – 15	3	Genelleme
4	15 – 20	4	Özelleştirme
5	20 – 25	5	Düzenleme
6	25 – 30	6	İlgili Geri Besleme
7	30 +	7	Diğer

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. 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 ark. 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 (Özmutlu ve ark. 2002) 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 Çizelge 2.16'da gösterildiği gibi iki eşit parçaya ayrılmıştır.

Çizelge 2.16: 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 Çizelge 2.17'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.

Çizelge 2.17: 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 olarak ve oturum; bir kullanıcının bütün sorgularını içeren küme olarak tanımlanmıştır. Bir oturum tek bir sorgudan oluşabildiği gibi çoğu durumda birçok sorgu barındırabilmektedir (Spink ve ark. 2001).

2.4.4. 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.

2.4.5. 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.

2.4.6. 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:

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

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

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

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

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:

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

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

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

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

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

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

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. ARAŞTIRMA SONUÇLARI ve TARTIŞMA

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 Çizelge 3.1 – Çizelge 3.4'de gösterilmiştir.

Karakter n-gram yönteminin Excite arama motorundan alınan verilerine uygulanması sonucu oluşturulan Çizelge 3.1'de uzman değerlendirmeleri ve karakter n-gram yöntemi sonuçları karşılaştırma amacıyla bir arada verilmiştir.

Çizelge 3.1: 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 değişikliği tahminleri değişik *n* ve eşik değerleri için farklıdır. Düşük eşik değeri ve *n* sayısı için konu değişikliği tahmini daha az olmakla beraber bu tahminler *n* sayısı ve eşik değeri arttıkça artar. Uzman tarafından belirlenen konu değişikliği sayısı dikkate alındığında karakter n-gram yönteminin 1-gram ve 0,5 eşik değeri sonuçları hariç, her zaman daha fazla tahminler yaptığı görülmektedir. Konu değişikliği 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 Çizelge 3.2'de Duyarlılık ve Anma değerleri ile uygunluk fonksiyonları hesaplanmıştır.

Çizelge 3.2: 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

Çizelge 3.2 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 Çizelge 3.2 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 0,59 olarak bulunurken, $F_{ß(devam)}$ değeri 0,897 olarak bulunmuştur. Yine aynı şekilde bu performans göstergelerindeki artışlar n-grama ve eşik değerine göre farklılıklar göstermektedir.

Karakter n-gram yönteminin FAST verilerine uygulanması sonucu elde edilen tahminler Çizelge 3.3'de 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.

Çizelge 3.3: 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 konu değişikliği tahminleri, farklı *n* ve eşik değerleri için değişmektedir. Düşük eşik değeri ve *n* sayısı için konu değişikliği tahminleri adedi, *n* sayısı ve eşik değeri arttıkça artmaktadır. Uzman tarafından belirlenen konu değişikliği 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. Çizelge 3.4'de FAST verisi için karakter n-gram yönteminin performans ölçütleri yer almaktadır.

Çizelge 3.4: 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 Çizelge 3.4 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. İstatistiksel Yöntemin Seçimi

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 Çizelge 3.5'de ve FAST verisi için Çizelge 3.6'da özetlenmiştir. Bu çizelgelere karakter n-gram yönteminin başarılı olduğu kombinasyonlar eklenmiştir.

Çizelge 3.5: 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}	$F_{\beta(değişim)}$	$F_{\beta(devam)}$
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

Çizelge 3.5'de başarılı tahmin oranlarını veren R_{devam} ve $R_{değişim}$ değerleri karşılaştırıldığında, en doğru tahminlerin yapay sinir ağları yönteminin uygulanmasıyla elde edildiği görülmektedir (Özmutlu ve ark. 2008c). 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.

Aynı şekilde Çizelge 3.6'da da görüldüğü gibi, FAST verilerine uygulanan yöntemlerin performans değerlendirmeleri yapıldığında en başarılı tahminler yapay sinir ağları yöntemiyle elde edilmiştir (Özmutlu ve ark. 2008c). 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.

Çizelge 3.6: 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	$N_{gerçekdeğişim}$ = 310	$N_{gerçekdevam}$ = 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

İkinci bölümde 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.

3.3. Yapay Sinir Ağları Yöntemi Sonuçları

Yapay sinir ağları yönteminde kullanılan verilerin hazırlık aşaması tamamlanmış, yöntemin test kümesi verilerine uygulanmasıyla tahminler elde edilmiştir. Çizelge 3.7 ve Çizelge 3.8'de Excite ve FAST verileri için elde edilen sonuçlar görülmektedir (Özmutlu ve ark. 2008c). Bu çizelgeler incelendiğinde A tipi hataların, Excite verisi için 217 ve FAST verisi için 580 adet olduğu görülür.

Çizelge 3.7: 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 arama yapılarının, verilerin hazırlık aşamasında "yeni" olarak sınıflandırıldığı görülür. Yapay sinir ağları yöntemi, sorgunun zaman aralığı ne olursa olsun arama yapısı "yeni" ise konu değişimi tahmini yapar. Hâlbuki yazım hatalarının veya farklılıklarının hazırlık aşamasında algılanması, bu sorguların yöntem tarafından farklı değerlendirilmesine yol açar ve büyük ihtimalle, zaman aralığına da bağlı olarak konu devamı tahmini yapılır. Bu sebeple karakter n-gram yöntemi, yapay sinir ağlarının yaptığı tahminlerden konu değişimi olan sorgulara uygulanmıştır. Böylelikle hatalı konu değişimi tahminlerini yansıtan A tipi hataların azaltılması hedeflenmiştir.

Çizelge 3.8: 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	-----	-----	----	----	----	----	----	----	----	----

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

Yapay sinir ağları yönteminin kullandığı verilerin hazırlığı aşamasında doğru bir şekilde arama yapılarının belirlenmesi, yöntemin tahminlerini iyileştirecektir. Bununla beraber yöntemin konu değişimi tahminlerine, karakter n-gram yönteminin uygulanması, aynı sonuçların elde edilmesini sağladığı için 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 Çizelge 3.9 ve Çizelge 3.10 elde edilmiştir.

Genel olarak çizelgeler 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.

Çizelge 3.9: 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**

Çizelge 3.10: 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**

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.

Farklı eşik değerleri ve n-gramlar için yapılan uygulamaların özetlenmesiyle oluşturulan Çizelge 3.11 ve Çizelge 3.12 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.

Çizelge 3.11: 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

Çizelge 3.12: 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. Çizelge 3.13 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.

Çizelge 3.13: 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 Çizelge 3.13 incelenmeye devam edilirse, performans göstergeleri olan $F_{ß(değişim)}$ ve $F_{ß(devam)}$ 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 $F_{ß(değişim)}$ ölçütünde, %2,639'luk bir iyileşme sağlanmıştır. Diğer bir iyileşme ölçütü olan ve konu devamı tahminlerini gösteren $F_{ß(devam)}$, 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 Çizelge 3.14'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 ($F_{ß(değişim)}$ ve $F_{ß(devam)}$) 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.

Çizelge 3.14: 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

SONUÇ

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.

KAYNAKLAR

BEEFERMAN, D. ve A. BERGER. 2000. Agglomerative clustering of a search engine query log. Proc. Of ACM SIGKDD Conference.

BOX, G. E. P. ve G. C. TIAO. 1973. Bayesian Inference in Statistical Analysis Reading. MA: Addison – Wesley, 34 – 36.

CANVAR, W.B. ve J.M. TRENKLE. 1994. N-Gram-Based Text Categorization. Proceedings of the third annual conference on document analysis and information retrieval (SDAIR), Las Vegas, pp.161–175.

CHAU, M. Y. LU, X. FANG ve C.C. YANG. 2009. Characteristics of character usage in Chinese Web searching. Information Processing and Management, pp. 115–130.

CHEN, S. F. ve J. GOODMAN. 1996. An empirical study of smoothing techniques for language modeling. In ACL 34, pp. 310 – 318.

CHEN, S. F. ve J. GOODMAN. 1998. An empirical study of smoothing techniques for language modeling. Technical Report TR–10–98, Center for Research in Computing Technology, Harvard University.

CHUCH, K.W. ve W. A. GALE. 1991. A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. Computer speech and Language 5: 19 – 54.

DAMASHEK, M. 1995. Gauging similarity with n-Grams: Language-Independent categorization of text. SCIENCE Vol.267.

EL-NASAN A. ve M. PERRONE. 2002. On-Line Handwriting Recognition Using Character BigramMatch Vectors. Proceedings of the Eighth International Workshop on Frontiers in Handwriting Recognition.

GOOD, I.J. 1953. The population frequencies of species and the estimation of population parameters. Biometrika 40: 237 – 264.

He, D. A. GÖKER ve D.J. HARPER. 2002. Combining evidence for automatic Web session identification. Information Processing and Management, 38 (2002), 727–742.

HUANG, X. F. PENG, A. AN, D. SHUURMANS ve N. CERCONE. 2003. Applying Machine Learning to Text Segmentation for Information Retrieval. Information Retrieval 6:333–362.

HUANG, X. F. PENG, A. AN ve D. SHUURMANS. 2004. Dynamic Web Log Session Identification With Statistical Language Models. Journal Of The American Society For Information Science and Technology, 55(14):1290–1303.

JELINEK, F. ve R. MERCER. 1985. Probability distribution estimation from sparse data. IBM Technical Disclosure Bulletin 28: 2591 – 2594.

KAMARIS, I. ve E. STAMATATOS. 2007. Webpage genre identification using variable-length character n-grams. 19th IEEE International Conference on Tools with Artificial Intelligence.

LEUNG, K.W-T. W. NG ve D. L. LEE. 2008. Personalized concept-based clustering of search engine queries. Knowledge and Data Engineering, IEEE Transactions on, Vol.20, Issue.11, 1505 -1518.

LIU, H. ve V. KESELJ. 2007. Combined mining of web server logs and web contents for classifying user navigation patterns and predicting users' future requests. Data & Knowledge Engineering 61, 304–330.

MANNING, C.D. ve H. SCHÜTZE. 1999. Foundations of Statistical Natural Language Processing. The MIT Press Cambridge, Massachusetts London, England. p. 191 – 224.

McNAMEE, P. ve J. MAYFIELD. 2004. Character N-Gram Tokenization for European Language Text Retrieval. Information Retrieval 7, 73 – 97.

ÖZMUTLU, H. C. A. SPINK, S. ÖZMUTLU. 2002. Analysis of large data logs: an application of Poisson sampling on excite web queries. Information Processing and Management, 38 (2002), 473–490.

ÖZMUTLU, S. A. SPINK ve H.C. ÖZMUTLU. 2004. A day in the life of Web searching: an exploratory study. Information Processing and Management, 40 (2004), 319–345.

ÖZMUTLU, H.C. ve F. ÇAVDUR. 2005a. Application of automatic topic identification on excite web search engine data logs. Information Processing and Management, 41(5), 1243–1262.

ÖZMUTLU, S. ve F. ÇAVDUR. 2005b. Neural Network Applications for Automatic New Topic Identification. Online Information Review, 29(1), 34–53.

ÖZMUTLU, S. 2006. Automatic New Topic Identification Using Multiple Linear Regression. Information Processing and Management 42 (2006), 934 – 950.

ÖZMUTLU, H.C. F. ÇAVDUR ve S. ÖZMUTLU. 2006. Automatic New Topic Identification in Search Engine Transaction Logs. Internet Research Vol.16 No.3, 323 – 338.

ÖZMUTLU, S. H.C. ÖZMUTLU ve B. BÜYÜK. 2007. Using Conditional Probabilities for Automatic New Topic Identification. Online Information Review Vol.31 No.4, 491 – 515.

ÖZMUTLU, H.C. F. ÇAVDUR ve S. ÖZMUTLU. 2008a. Cross-Validation of Neural Network Applications for Automatic New Topic Identification. Journal of the American Society for Information Science and Technology, 59(3):339–362.

ÖZMUTLU, S. H.C. ÖZMUTLU ve B. BÜYÜK. 2008b. A Monte-Carlo Simulatıon Application for Automatic New Topic Identification of Search Engine Transaction Logs. Simulation Modelling Practice and Theory, 16 (2008), 519 – 538.

ÖZMUTLU, S. ÖZMUTLU, H.C. ve G. COŞAR. 2008c. Neural Network Applications for Automatic New Topic Identification. (yayım aşamasında).

ÖZMUTLU, S. H.C. ÖZMUTLU ve G. COŞAR. 2008d. Improving the performance of automatic new topic identification of search engine transaction logs. (yayım aşamasında).

ROARK, B. M. SARAÇLAR ve M. COLLINS. 2007. Discriminative n-gram language modeling. Computer Speech and Language 21, 373–392.

ROSENFELD, R. ve X. HUANG. 1992. Improvements in stochastic language modeling. In preceedings of the DARPA speech and Natural Language Workshop, pp.107 – 111.Morgan Kaufman.

RISTAD, E. S. 1995. A Natural Law of Succession. Technical Report CS – TR 495–95, Princeton University.

SENDA, S. ve K. YAMADA. 2001. A maximum likelihood approach to segmentation-based recognition of unconstrained handwriting text. Proceedings of the Sixth International Conference on Document Analysis and Recognition.

SHANNON, C. E. 1951. Prediction and entropy of printed English. Bell System Technical Journal 30: 50–64.

SPINK, A. D. WOLFRAM, B.J. JANSEN ve T. SARACEVIC. 2001. Searching the Web: The public and their queries. Journal of the American Society for Information Science and Technology, 53, 226–234.

SPINK, A. H.C. ÖZMUTLU ve S. ÖZMUTLU. 2002. Multitasking information seeking and searching processes. Journal of the American Society for Information Science and Technology, 53(8), 639–652.

SPINK, A. H.C. ÖZMUTLU ve P. D. LORENCE. 2004. Web searching for sexual information: an exploratory study. Information Processing and Management, 40 (2004), 113–123.

ZITOUNI, I. 2007. Backoff hierarchical class n-gram language models: effectiveness to model unseen events in speech recognition. Computer Speech and Language 21, 88–104.

EKLER

EK 1: Arama yapısı sınıflarının belirlenmesi için kullanılan algoritma

Input: Queries Q_{i-1}, Q_i, Q_{i+1} (set of three subsequent queries)
Local: Q_c, current query (as a string)
Q_n, next query (as a string)
$B = \{t \mid t \in Q_c \text{ and } t \in Q_n\}$, the set of terms (terms determined using "space" as a divider) that are common in both Q_c and Q_n
$C = \{t \mid t \in Q_c \text{ and } t \notin Q_n\}$, the set of terms, which appear in Q_c only
$D = \{t \mid t \notin Q_c \text{ and } t \in Q_n\}$, the set of terms, which appear in Q_n only
Output: Search Pattern, *SP*
begin
if ($Qi == \phi$) **then**
if ($i == 1$) **then** *SP* = *Other*,
else $Q_c = Q_{i-1}$, // if Q_i is empty (relevance feedback) then take the preceding query // (Q_{i-1}) to analyze the relationship
$Q_n = Q_{i+1}$,
endif
else $Q_c = Q_i$,
$Q_n = Q_{i+1}$,
endif
SP = *other* //default value
if ($Q_n == \phi$) **then** *SP* = *Relevance Feedback* **endif** // if the next query is empty then //it is relevance feedback
if ($Q_n == Q_c$) **then** *SP* = *Next Page* **endif**
if ($B \neq \phi$ and $C \neq \phi$ and $D == \phi$) **then** *SP* = *Generalization* **endif**
if ($B \neq \phi$ and $C == \phi$ and $D \neq \phi$) **then** *SP* = *Specialization* **endif**
if ($B \neq \phi$ and $C \neq \phi$ and $D \neq \phi$) **then** *SP* = *Reformulation* **endif**
if ($Q_n \neq Q_c$ and $B \neq \phi$ and $C == \phi$ and $D == \phi$) **then** *SP* = *Reformulation* **endif**
if ($Q_c \neq \phi$ and $B == \phi$) **then** *SP* = *New* **endif**
end

EK 2: A Tipi Hataya sebep olan sorgular, arama yapısı sınıfları ve zaman aralıkları

Sorgu çiftleri	TI	SP
• hard drive format	7	5
• +format +c:	1	5
• Musical Theatre History	3	5
• musical theatre history	1	1
• Scholarships	7	5
• Scholarship news	1	1
• creston iowa	2	5
• (southwestern iowa)	2	5
• +surplus+electronics	3	5
• +capacitor+surplus	1	1
• AEROSMITH	3	5
• Aerosmith	1	1
• EDUCATION+Desert	3	5
• Desert	1	3
• image: dinosaurs	2	5
• dinosaur AND eggs	8	8
• www.wisconsin florists.com	2	5
• florists	8	8
• fechin inn	2	5
• Fechin Inn	1	4
• Screensaver	6	5
• screen saver	1	1
• Hotmail	2	5
• hotmail.com	1	5
• castrator-italy	4	5
• burdizzo-castrator	8	8
• hotels in istambul	3	5
• istanbul	1	3
• vergussm rtel	1	5
• vergussmörtel	8	8
• www.virginblue.com.au	1	5
• virgin blue airline	2	1
• Eniac	7	5
• ENIAC	8	8
• diseño AND de AND planta AND de AND una AND empresa AND industrializadora AND de AND vinos	5	5
• vinos	1	3
• Planaria	2	5
• planaria+eyespot	1	5
• pictures of world war 1	3	5
• world+war+one+information	1	1
• +Independent +brewing +IBCO	4	5
• +Independence +Brewing +Mousehead	1	4
• Wor-Wic Community College	4	5
• wor-wic community college	8	8
• black planet	7	5
• blackplanet	5	1
• www.wal-mart.com	5	5
• walmart.com	8	8

• randolph mantooth • Randolph Mantooth	3 6	5 1
• msnbc.com wsfa.commsnbc.com msnbc.com • wsfa.commsnbc.com	5 5	4 1
• Lab Line incubators • lab-line instruments	4 5	5 8
• atlanta,Georgia • sandy springs,georgia	6 1	5 1
• cards.ocx • cards.tlb • cards AND vc • cards.dll	2 1 7 8	5 5 5 8
• national geographic • nationalgeografic	5 7	5 5
• telefonia • telefon a	1 1	5 1
• skill-network • skill networks	1 8	5 8

EK 3: B Tipi Hataya sebep olan sorgular, arama yapısı sınıfları ve zaman aralıkları

Query pair	TI	SP
• hyman's AND Colorado • cathedral AND spires AND garden AND gods	7 2	4 4
• Cadmus AND Point AND Of AND Purchase • Cap AND Human AND Development	5 8	4 8
• sandy springs,Georgia • georgia chamber of commerce,atlanta,georgia	1 2	5 5
• birmingham news • Birmingham	1 1	2 1
• cotton AND harvester AND picture • agri AND business	7 1	4 1
• farm AND equipment • bobwhite AND quail AND history AND image	6 1	4 1
• solar AND cell • free AND project AND management AND software	6 8	4 8

ÖZGEÇMİŞ

Burcu Çağlar 02.05.1984 tarihinde Bursa'da doğmuştur. İlk ve orta öğrenimini Yalova Saffetçam İlköğretim Okulu ve Şehit Osman Altınkuyu Anadolu Lisesi'nde tamamladıktan sonra, lise öğrenimi için Bursa Şükrü Şankaya Anadolu Lisesi'ne devam etmiştir. 2002 yılında Uludağ Üniversitesi Endüstri Mühendisliği Bölümü'nü kazanan Burcu Çağlar, 2006 yılında ikincilikle lisans derecesini almış ve aynı yıl Uludağ Üniversitesi Fen Bilimleri Enstitüsü Endüstri Mühendisliği Anabilim Dalı'nda yüksek lisans eğitimine başlamıştır. Halen Uludağ Üniversitesi, Mühendislik-Mimarlık Fakültesi, Endüstri Mühendisliği Bölümü'nde araştırma görevlisi olarak çalışmaktadır.

TEŞEKKÜR

Yüksek lisans eğitimim süresince bana yardımcı olan ve yüksek lisans tezini birlikte yaptığım danışmanım sayın Doç. Dr. H. Cenk Özmutlu'ya, eşi sayın Doç. Dr. Seda Özmutlu'ya, lisans ile yüksek lisansım boyunca yardımlarını esirgemeyen Sayın Yrd. Doç. Dr. A. Yurdun Orbak'a ve eğitimim sırasında bana burs vererek maddi anlamda yardımcı olan TÜBİTAK'a teşekkürü bir borç bilirim.

Ayrıca eğitimim boyunca bana her zaman destek olan aileme teşekkür ederim.