T.C.
ULUDAĞ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

İNTERNET ARAMA MOTORU KULLANICI VERİLERİNİN ANALİZİNDE SİMÜLASYON VE OLASILIKSAL YÖNTEMLERİN UYGULANMASI

Buket BÜYÜK

YÜKSEK LİSANS TEZİ
ENDÜSTRİ MÜHENDİSLİĞİ ANABİLİM DALI

BURSA–2009

T.C.
ULUDAĞ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

İNTERNET ARAMA MOTORU KULLANICI VERİLERİNİN ANALİZİNDE SİMÜLASYON VE OLASILIKSAL YÖNTEMLERİN UYGULANMASI

Buket BÜYÜK

Doç.Dr. Seda ÖZMUTLU
(Danışman)

YÜKSEK LİSANS TEZİ
ENDÜSTRİ MÜHENDİSLİĞİ ANABİLİM DALI

BURSA–2009

T.C.
ULUDAĞ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

İNTERNET ARAMA MOTORU KULLANICI VERİLERİNİN ANALİZİNDE SİMÜLASYON VE OLASILIKSAL YÖNTEMLERİN UYGULANMASI

Buket BÜYÜK

YÜKSEK LİSANS TEZİ
ENDÜSTRİ MÜHENDİSLİĞİ ANABİLİM DALI

Bu Tez ..../...../200... tarihinde aşağıdaki jüri tarafından oybirliği ile kabul edilmiştir.

Doç.Dr.Seda ÖZMUTLU ....
Danışman

ÖZET

Akıllı bir arama motoru geliştirmenin en önemli adımlarından biri yeni konu tanımlanmasıdır. Yeni konu tanımlanması kullanıcının tek bir oturum sırasında bir konudan diğerine geçerken ki konu değişiminin belirlenmesidir. Bu çalışmada arama motoru kullanıcı oturumlarındaki konu değişikliklerini tespit ve tahmin etmek amaçlanmıştır. Bunun içinde Şartlı Olasılık Yaklaşımı ve Monte Carlo Simülasyonu olmak üzere iki farklı yaklaşım Excite ve FAST arama motorlarından alınan veri setlerinde kullanılmıştır. Bu yaklaşımlarda veri setindeki her bir sorgu için 'konu değişikliği var' ve 'konu değişikliği yok' diye atamalar yapılmaktadır. Atamaları yapmak içinse sorgunun zaman aralığı ve arama yapısı olmak üzere iki adet istatistiksel özelliği kullanılmaktadır. Yaklaşımlar tarafından yapılan atamalar uzman kişi tarafından yapılan atamalarla karşılaştırılmıştır. Daha sonra yaklaşımların performans ölçütleri hesaplanmış ve incelenmiştir. Şartlı Olasılık yaklaşımı ve Monte-Carlo Simülasyonu yeni konu tanımlaması için yeterli ve tutarlı sonuçlar sağlamıştır.

Anahtar Kelimeler: Arama Motoru Kullanıcı Davranışlarının İncelenmesi, Yeni Konu Tanımlanması, Şartlı Olasılık, Monte Carlo Simülasyonu

ABSTRACT

One of the main elements in developing a personalized intelligent search engine is new topic identification. New topic identification is discovering when the user has switched from one topic to another during a single search session. In this study, it is aimed to estimate and determine new topic identification. For this reason, two different approaches, Conditional Probabilitity and Monte Carlo Simulation are used in Excite and FAST search engine query logs. Due to application of mentioned methods for automatic new topic identification, each query is assigned as "topic shift" or as "topic continuation". This assignment relies on the statistical characteristics of the queries, such as the time interval between consecutive queries and the search pattern of the consecutive queries. Results of human expert and methods are compared and performance measures are calculated. Conditional Probability Approach and Monte Carlo Simulation Method provided acceptable and adequate results.

Keywords: Investigating of Search Engine User Behaviors, Automatic New Topic Identification, Conditional Probability, Monte Carlo Simulation

İÇİNDEKİLER

ÇİZELGELER DİZİNİ

ŞEKİLLER DİZİNİ

GİRİŞ

Günümüzde İnternet pek çok kişi için bilgiye ulaşmanın en önemli platformlarından birisidir. İnternet'in hızla büyümesiyle beraber, İnternet kullanıcılarının bu çok geniş platformdan aradıkları bilgiye nasıl ulaşacakları sorusu gündeme gelmiştir. Bunun için kullanılan yöntemlerden biri de arama motorlarıdır. Bu nedenle arama motoru kullanıcılarının davranışları üzerinde çalışmak önemli bir konudur

Arama motorlarının sayısındaki artışla beraber, sunulan hizmetin kalitesinin de artması gerekmektedir. Eğer arama motoru, kullanıcının önceki sorgularını baz alarak mevcut konuya devam ettiğini tahmin edebilirse, önceki sorgularda bulduğu sonuçları kullanarak yeni sorguya cevap verme hızını arttırabilir ve daha uygun cevaplar sağlayabilir. Bunun yanı sıra, kullanıcının yeni konu aradığı tespit edilirse arama motoru öncekinden farklı, yeni bir kümeden aldığı cevapları ekrana getirebilir. Böylece kullanıcı bazlı arama sonuçları elde edilmiş olur. Sonraki aşamalarda kullanıcıların sorguları baz alınarak ilgi alanları belirlenebilir ve arama motorları kişiselleştirilebilir.

Arama motoru kullanıcı davranışlarının tahmininde anlam bazlı ve anlam bazlı olmayan metotlar kullanılmaktadır. Bu metotların amacı, konu değişikliklerinin tahminiyle kullanıcı davranışlarını tespit etmektir.

Konu değişimi tespiti için pek çok anlam bazlı olmayan istatistiksel yöntem geliştirilmiştir. Bu yöntemlerden bazıları, Excite ve FAST arama motorlarından alınan, arama motorlarını kullanan kullanıcılar hakkında, İnternet Protokolü (IP) adresi, arama zamanı ve sorgu terimi bilgilerine sahip veri grupları üzerinde uygulanmıştır. Veriler üzerinde bazı işlemler uygulanarak, bu çalışmada kullanılabilecek hale getirilmiştir. Aynı IP adresine ait ardışık arama kayıtlarının zamanları arasındaki farklar kullanılarak sorgular 7 farklı zaman aralığına bölünmüştür. Benzer şekilde, aynı IP adresine ait ardışık aramalarda girilmiş olan sorgular incelenerek, ardışık sorgular arasındaki yapısal ilişkilere bakılmış ve 7 farklı arama yapısı sınıfına ayrılmıştır. Veriler üzerinde

uygulanan yöntemlerin, konu değişikliği ve konu devamı tahminlerinde bulunması sağlanmıştır. Tahminler, uzman tarafından değerlendirilmiş gerçek sonuçlar ile karşılaştırılarak, yöntemlerin performans değerlendirmeleri yapılmıştır.

Çalışmada konu değişimi tespiti için anlam bazlı olmayan istatistiksel yöntemlerden Şartlı Olasılık Yaklaşımı ve Monte Carlo Simülasyonu kullanılmıştır.

Excite ve FAST arama motorundan alınan, arama motorlarını kullanan kullanıcılar hakkında bilgilere sahip veri gruplarındaki sorgular, çalışmayı yapan kişi tarafından, ardışık iki sorgu arasında ‘konu değişimi’ ve ‘konu devamı’ diye sınıflandırılmıştır. Daha sonra bu sınıflandırılan sorguların, belirlenen arama yapısı ve zaman aralıklarına göre ‘konu değişimi’ ve ‘konu devamı’ olanlarının sayıları belirlenmiştir. Her bir zaman aralığı-arama yapısı sınıfı içinse bu sayılardan yola çıkarak, şartlı olasılıklar hesaplanmıştır. Daha sonra bu şartlı olasılıklar kullanılarak yukarıda bahsedilen iki yöntem uygulanmış ve test verisindeki sorguların bu iki yönteme göre konu değişimi mi konu devamı mı olduğu tahmin edilmiştir. Başka bir deyişle her bir sorgu için belirtilen yöntemlerle bir tahmin yapılmış ve o sorgu ya ‘konu değişimi’ ya da ‘konu devamı’ olarak işaretlenmiştir. Bu işlemden sonra belirtilen performans ölçütlerine göre her bir yöntemin performans ölçütleri incelenmiş ve yöntemlerin başarısı yorumlanmıştır.

1. KONU İLE İLGİLİ ÇALIŞMALAR

1.1. Kuramsal Bilgiler

İnternet kullanıcılarının, İnternet üzerinde aradıkları bilgilere ulaşabilmeleri için konulara göre düzenlenmiş başlıkların yer aldığı İnternet siteleri bulunmaktadır. İnternet kullanıcıları bu başlıkları tarayarak aradıkları bilgilere ulaşabilirler. Buna indeks kullanarak arama yapmak denir. Bu şekilde yapılan aramada, direk indeks Web sayfasında bulunan seçenekleri kullanma olanağı bulunmaktadır. Dolayısıyla aranan konuya ulaşmak için çeşitli anahtar sözcükler kullanmaya gerek yoktur. Web indeksleri, başka sayfalara bağlanmayı sağlayan sözcüklerden oluşan listelerdir. İndekste ana konulardan başlayarak ve giderek alt düzeylere inerek, aranılan konuya ulaşılabilir. (Balay ve ark. 2006)

İndeks kullanan kullanıcılar, kendileri için en uygun seçeneği indeks üzerinde kendileri bulmak ve karşılarına çıkacak olan çok sayıda sayfadan oluşan sayfa listesinde hangi sayfaların işlerine yarayacağını kendileri belirlemek zorundadır. Bu nedenle, günümüzde İnternet kullanıcıları İnternet'te aradıkları bilgilere ulaşabilmek için arama motorlarını tercih etmektedirler.

Aramak istenilen bilgiyle ilgili anahtar kelimeleri veya bir cümleyi yazarak ağı tarayan ve ilgili sitelerin adreslerini listeleyen yazılımlara arama motoru denmektedir. İnternet'in genelinde arama yapan motorlar, aslında ağı doğrudan taramaz. Her biri, sunucular üzerinde bulunan milyarlarca Web sayfasından seçilerek oluşturulan ve Web sayfalarının metin biçimlerinin tamamını içeren bir veritabanını tararlar.

Arama motorlarında yapılan aramalar benzer özellikler göstermektedir. Bu özelliklerden kısaca bahsetmek gerekirse (Balay ve ark. 2006) ;

*Belirli terimleri içeren/içermeyen arama yapma

*AND, OR, NOT gibi mantıksal operatörleri kullanarak arama yapma

*Pdf, word, pps gibi çeşitli dosya türlerine göre arama yapma

*Bir sayfaya benzer sayfaları arama seçeneği ile ve bir sayfaya bağlantısı olan sayfaları arama seçeneği ile arama yapma

*Aynı siteden olan Web sayfalarını sınırlandırma seçeneği ile arama yapma

*Sayfanın herhangi bir yerinde; başlığında, içeriğinde, adresinde veya bağlantılarında arama yapma

Google Arama Motoru (http://www.google.com)

Dünyanın en büyük arama motorlarından birinin geliştiricisi olan şirkettir. Stanford'da doktora yapan iki öğrenci Larry Page ve Sergey Brin, Google'ı 1998'de kurdu. Özel şirket, Haziran 1999'da, 25 milyon dolar yasal sermayeye sahip olduğunu duyurdu. Şirketin sermaye ortakları, Kleiner Perkins Caufield & Byers ve Sequoia Capital'i kapsıyor.

Google'ın arama teknolojisi ve kullanıcı arabirim tasarımı Google'ı günümüzün ilk-nesil arama motorlarından farklı kılar. Sadece anahtar kelime veya meta arama teknolojisi kullanmak yerine Google en önemli sonuçları ilk getiren gelişmiş PageRank™ teknolojisine dayanır.

PageRank ağ sayfalarının önemini nesnel bir ölçeğe uyarlar; bu 500 milyon değişken ve 2 milyar terimden oluşan bir denklemin çözülmesiyle hesaplanır. PageRank ağın çok sayıda bağlantılı yapısını düzenleyici bir araç olarak kullanır. Doğal olarak, Google, Sayfa A'dan Sayfa B'ye kurulmuş her bağlantıyı Sayfa A'dan Sayfa B'ye bir "oy" olarak yorumlar. Google bir sayfanın önemini aldığı oylarla belirler. Google ayrıca oyu veren sayfayı da inceler.

Yahoo Arama Motoru (http://search.yahoo.com)

Yahoo! Inc., Stanford Üniversitesi öğrencileri Jerry Yang ve David Filo tarafından kurulmuş bir portaldır.

İlk zamanlarda arama motoru olarak hizmet vermesine rağmen zamanla e-posta, anında mesajlaşma, e-posta grubu ve benzeri hizmetler de sunarak pazarda hâkim olmaya çalışmıştır. Sunduğu hizmetlerden Yahoo Messenger özellikle ABD'de de çok yaygın olarak kullanılmasına rağmen, Türkçe olarak sunulmadığı ve MSN Messenger'ın Windows işletim sistemiyle birlikte gelmesi nedeniyle Türk kullanıcılar arasında pek tercih edilmemiştir(http://tr.wikipedia.org/wiki/Yahoo).

Yahoo! Araması, Web'de mevcut olan çok çeşitli bilgiye kolay ve hızlı erişim sağlar. Yahoo! Arama Motoru, Yahoo! Arama Teknolojisi'ni kullanmaktadır. Her arama yapıldığında, Yahoo! Arama Teknolojisi kendi Web sayfalarının veri tabanlarını tarar. Taradığı sayfaların girilen arama terimleriyle ilgililiğini belirler ve sonuçları sıralanmış olarak listeler. Yahoo! Arama Teknolojisi Web sayfalarını belirli arama terimlerine göre sıralarken, metin, başlık ve tanım doğruluğu, kaynak, ilgili bağlantılar ve o nesneye özgü diğer belge özelliklerini dikkate alır.

Yahoo, Inktomi[1]'nin sahibi olunca arama algoritmalarını geliştirerek kendi arama motoru teknolojisini yenileme yolunda yeni adımlar atmıştır.

[1] 10 milyon ücretli URL'yi veritabanında bulunduran ve bu URL'lerin yarısı tık başına tekrarlayan gelir elde eden, tarayıcı tabanlı algoritmalarda son derece başarılı olan şirket :Sullivan, D. Jan 2003. Yahoo To Buy Inktomi, Search Engine Watch,

MSN Arama Motoru (http://search.msn.com)

1 Temmuz 2004 tarihinde, Microsoft MSN arama motorunun ilk çalışmalarına başladı. Bu arama motoru, Web tarayıcısı robotu olan MSNbot tarafından desteklenmekteydi.

MSNBot bir Web tarayıcı robotu olup, Microsoft tarafından 'Live Search'ü desteklemek için görevlendirilmiştir(http://en.wikipedia.org/wiki/Search_engines). 2004 yılında beta sürümü ile ortaya çıkan MSNBot, MSN arama motoru için aranabilir bir dizin yaratmak amacıyla Webden belge toplamaktaydı. 2005 yılında ise MSN arama motoru hizmete tamamen açılmıştı.

MSNBot için tipik bir kullanıcı aracı dizini: "msnbot/1.1 ise (+ http://search.msn.com/msnbot.htm)" şeklindedir. Bu Web sunucusu kütüğünde, Web yöneticisine kimin dosya istediğini anlatmak için görünür. Her site yöneticisi verilen aracı tanımını, "MSNBot", erişim için veya erişimi kabul etmemek için kullanma yetkisine sahiptir. Eğer site yöneticileri erişim imtiyazının verilmesini istemiyorlarsa, Robot Dışlama Standartını[2] (MSNBot iyi niyetli davrandığına güvenerek) kullanabilirler(http://en.wikipedia.org/wiki/Msnbot).

[2] Web robotlarının sitenin herkes tarafından erişimine izin verilmeyen kısımlarına ulaşmasını engellemek için kullanılan kural : robot exclusion standard, robots exclusion protocol

1.2. Kaynak Araştırması

Arama motoru kullanıcı davranışlarının tahmin edilmesi arama motorlarının kişiselleştirilmesi açısından önem taşımaktadır. Arama motoru kullanıcı davranışlarını belirleyebilmenin bir yolu konu değişimini tahmin edebilmektir. Konu değişimini tahmin edebilmek için yapılan çalışmalar anlam bazlı olanlar ve anlam bazlı olmayanlar diye iki ana grupta toplanabilir.

1.2.1. Arama motoru kullanıcı oturumlarında konu değişimi tespiti için yapılan anlam bazlı çalışmalar

Anlam bazlı çalışmalar sözlüklerle çalışırlar. Bu sözlüklerin başlangıç aşamasında oluşturulması, uygulamalarda kullanılmak üzere depolanması gerektiğinden dolayı, bu çalışmalar yüksek maliyet ve emek gerektirirler. Bu nedenle arama motoru kullanıcı oturumlarında konu değişimi tespiti için yapılan anlam bazlı çalışmaların sayısı kısıtlıdır. Bu bölümde anlam bazlı çalışmalar; konu analizi, çoklu görev yürütümü, sorgu kümeleme modelleri, metin sınıflandırma ve kategorizasyon modelleri gibi birkaç alt bölümde sınıflandırılmıştır.

1.2.1.1. Konu analizi

Silverstein ve arkadaşları (1999) ve Jansen, Spink ve Saracevic (2000) gibi araştırmacılar, terim düzeyinde arama motoru veri kümeleri için, içerik bazlı araştırmalar yapmışlardır. Bu nedenle de sorgularda terim çiftlerini ve terimlerin sıklığını analiz etmeleri sonucunda en çok aranan terimlerin pornografi, eğlence ve eğitim konuları ile ilgili olduğunu bulmuşlardır. Başka bir çalışmada, Spink ve arkadaşları (2002a), Excite verilerini 1997'den 1999 ve 2001'e kadar konu içeriği bakımından analiz ettiklerinde, insanların bilgi ihtiyaçlarının 1997'den 2001'e değişim gösterdiğini belirtmişlerdir. Kullanıcıların ilgi alanları eğlence ve pornografiden; seyahat, ticaret, ekonomi ve kişi adlarına kaymıştır. Özmutlu ve arkadaşları (2004b) Excite ve FAST verisinin (her bir veri yaklaşık 1.000.000 adet sorgudan oluşmaktadır),

saatlere göre istatistiksel ve konusal bir analizini yapmıştırlar. Bu analizler sırasında konuların popülerliğinin gün içinde değiştiğini bulmuşturlar. Finans, ticaret ve eğitim gibi konular günün ilk saatlerinde daha popülerken, eğlence ve pornografi gibi konular akşam saatlerinde daha popüler olmaktadır. Özmutlu ve arkadaşları (2004b) ve Beitzel, Jensen, Chowdhury, Grossman ve Frieder (2004) çalışması konuların popüleritesinin günün saatlerine göre değiştiğini göstermiştir.

1.2.1.2. Çoklu görev yürütümü

Bir arama oturumunda kimi kullanıcılar birden fazla konu ile ilgilenebilirler. Bilgi erişim deyimlerince, çoklu görev yürütümünde bilgi arama ve araştırma prosesleri ya da "birden fazla, muhtemelen değişen, bilişsel, efektif ve koşulsal durumları da içeren bilgi problemleriyle ilgili zaman içindeki aramalar " (Spink ve ark. 2002b) diye adlandırılan aramalar akla gelmektedir. Spink ve arkadaşları (2002) çalışmalarında 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 belirlenmesini hedeflemişlerdir. Birbirinden bağımsız dört farklı çalışma ile farklı bilgi erişim ortamlarında, kullanıcıların davranışlarını inceleyebilmek için veriler toplamışlardır Toplanan verileri birbirinden bağımsız olarak inceleyen yazarlar, dört çalışmadaki çoklu görev yürütümü yaygınlığının değişken olduğunu ve bu değişkenliğin nedenlerinin belirlenmesinin zor olduğunu belirtmişlerdir. Yalnız yazarlar, farklı bilgi erişim ortamlarında, insanların genelde birden fazla bilgi problemiyle çalıştığı sonucuna varmışlardır. Spink ve arkadaşları (1999) Excite arama motoru kullanıcılarının %3,8'inin çoklu görev yürütümü aramaları yaptığını belirtmiştirler. 1999 yılında bir gün içinde Excite arama motorundan toplanan verilerde, kullanıcıların %11,4'ünün çoklu görev yürütümü aramalarını gerçekleştirmiş olduğu ve 2001 yılında bir günde FAST arama motorundan elde edilen verilerde yapılan bir araştırmada kullanıcıların %31,8'inin çoklu görev yürütümü aramalarını gerçekleştirmiş olduğu görülmüştür(Özmutlu ve ark. 2003a).

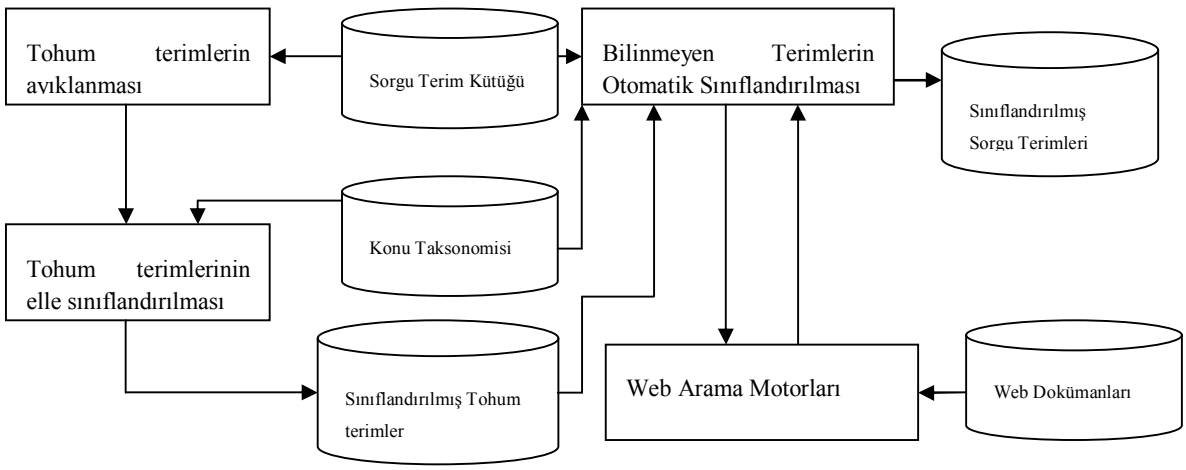
Spink ve arkadaşları (2006) Altavista arama motorunun verilerini incelediklerinde iki sorgulu oturumların %81'inin ve üç veya daha fazla sorgulu oturumların %91'inin

çoklu konuyu içerdiği görülmüştür. Çoklu görev yürütümü oturumlarında geniş bir konu çeşitliliği mevcuttur. Üç ya da daha fazla sorgu oturumu bazen sık konu değişimleri içerebilir.

1.2.1.3. Sorgu kümeleme modelleri

Beeferman ve Berger (2000) ve Wen ve arkadaşları (2002) sorgu kümeleme yöntemini uygulayıp, hangi Web sayfalarının tıklandığı bilgisini içeren, arama motoru sorgu kümelerinde birer çalışma yapmışlardır. Beeferman ve Berger (2000) çalışmalarında sorgu konusunu ihmal etmişlerdir. Ayrıca çalışmalarında kendi geliştirdikleri "konu içeriğini ihmal eden" sorgu kümeleme algoritmasının standart kümeleme tekniklerinden daha verimli olduğunu belirtmektedirler.

Pu ve arkadaşları (2002) Web sorgu terimlerini, otomatik sınıflandırma yöntemi ile geniş konu kapasitesinde sınıflandırabilmişlerdir. Sorgu kütüğünden alınan sorguların içinden en çok sayıda olan terimler ayıklanır. Bu ayıklanan terimlere 'tohum terim' denilmektedir. Uzman kişi daha önce tanımlanan konu taksonomisine göre bu terimleri sınıflandırır ve sınıflandırılmış tohum terimler 'W' setini oluşturur. Konu taksonomisi 'C' ve sınıflandırılmış tohum terimler 'W' kümelerinin yardımıyla üçüncü aşamada T kütüğünde bilinmeyen her bir t terimi uygun bir sınıfta sınıflandırılarak bu işlemin çıktısını oluşturacaktır. Aşağıda Şekil 1.1'de bu otomatik sınıflandırma yöntemi görülebilir.



Şekil 1.1: Önerilen otosınıflandırma yaklaşımının tasarımını gösteren özet bir grafik

Wen, Nie ve Zhang (2002) sorguların ve dokümanların çapraz listeleme yöntemi ile kümelenmesinin, basit anahtar kelime kümelenmesinden daha başarılı olduğunu iddia etmektedirler. Bu çalışmalarında amaçlarının sıkça sorulan soruları belirlemek için benzer sorgu ve soruları kümelemek olduğunu belirten yazarlar, kendi kullandıkları kümeleme yönteminin iki prensibe dayandığını belirtmişlerdir. Bu prensipler sorgu içeriğini ve doküman tıklamasını kullanmaktır.

Zhu, Greiner ve Haubl (2003) link bazlı bir sistem önermişlerdir. Bu sistem daha önce erişilen dokümanın sorgusundaki terimleri içererek yeni sorgular oluşturmaktadır. Bu sistemin algoritması erişilen son dokümanda benzer bir terim olup olmadığını tahmin etme esasına göre çalışmakta ve bunu yaparken de iki araçtan yararlanmaktadır. *IcURLPredictor*; aranan bilgiyi içeren sayfaları tanımlamak için öğrenen bir sınıflandırıcı ve *IcWordFinder;* aranan bilgiyi içeren kelimeleri tanımlamak için öğrenen bir sınıflandırıcı. Buradaki araçlar kullanıcı davranışlarını, geniş miktardaki kullanıcı verisinden oluşan genel kullanıcı modelinden öğrenmektedirler. Alternatif tavsiye sistemlerinin pek çoğu önceden tanımlanmış URL ve kelimeleri kullanmaktadır. Bu sistemler eğitim senaryolarında kullanılan yapı ile değerlendirildiklerinde en verimli halini almaktadırlar. Eğitim seti için başka veri, test seti için başka veri kullanıldığında yeterince yararlı olamamaktadırlar. Araçlar öncelikle o ana kadar taranmış olan kelimelerin veya sayfaların tarama özelliklerini ortaya çıkarmakta ve daha sonra bu özellikleri baz alarak kelimeleri ya da sayfaları mevcut kullanıcı için belirgin kelime ya da sayfa kullanıcı özelliğini öğrenmektedirler. Böylece mevcut durumda kişiye özel öneri yapmak mümkün olabilmektedir.

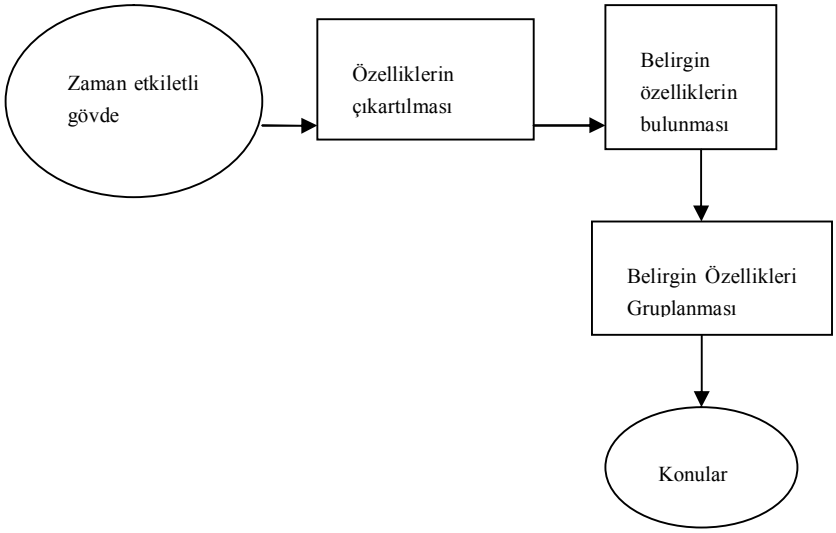
Muresan ve Harper (2004) yönlendirilmiş sorguların geliştirilmesi için konu modelleyen bir sistem önermişlerdir. Kaynak bir derlemede bulunan uygun dokümandaki terimlerin, istatistiksel analizini yapıp, sorgunun sözlük sıralaması bakımından istatistiksel temsilini gerçekleştirmişlerdir. Bu adımdan sonra içerik bazlı araştırmalar gelmiştir. Bu içerik bazlı araştırmalar konuların ilişkisini, terimlerin sıklığını inceler. Özel konulardaki terimlerin benzerliğine dayalı yönlendirilmiş sorgular ise daha sonra gelişmiştir.

Liu ve arkadaşları (2004) sorguları sınıflandırmak için Destek Vektör Makineleri (DVM) yöntemini kullanmıştır. Sorguların bazı sözdizimsel özellikleri: cümlenin uzunluğu, her kelimedeki ortalama karakter sayısı, her bir kelimedeki ortalama hece sayısı bu sorguların sınıflandırılmasında kullanılmıştır. Aynı zamanda anlambilimsel; 2-terimli ve 3-terimli kelime dizilerinin sıklığı gibi özelliklerde de bu yöntemde kullanılmıştır. Sonuçlar DVM yönteminin sorguları tanımada %80 ve hatta %80'den de fazla oranda başarılı olduğunu göstermiştir. Metzler ve Croft'un (2005a) çalışmalarında soru tipindeki sorguların sınıflandırması için kullanılan istatistiksel sınıflandırıcı da DVM yöntemine dayalıdır. Bu sınıflandırıcı soru kelimeleri ve tipleri arasındaki bilginin korelasyon değerini kullanır. Bu istatistiksel çerçevede herhangi bir veri seti, ya da soru ontolojisi kullanılabilir.

1.2.1.4. Metin sınıflandırması ve kategorizasyon modelleri

Metin sınıflandırmasıyla ve veri madenciliğiyle ilgili birçok çalışma bulunmaktadır. Dökümanları belirli bir konu ile ilgili ve ilgili değil şeklinde filtre etmek, ikili bir sınıflandırma problemi oluşturur. İkili sınıflandırma problemini çözmek için DVM yöntemi kapsamlı olarak kullanılmış (Joachims 1998, Yang ve Liu 1999) ve yöntemin uygulamaları metin sınıflandırılmasında başarılı olmuştur.

Veri madenciliğinde önemli çalışmalardan bir tanesi de Swan ve Jensen (2000) Zaman Madeni adlı yöntemleridir. Bu yöntemde Swan ve Jensen, dokümanlardaki tarih etiketlerini ve terimlerin istatistiksel özelliklerini kullanarak, dokümanları farklı kategorilere ayırmayı başarmıştırlar. Metindeki özellikleri keşfetmek içinse aşağıdaki prosesi kullanmışlardır.



Şekil 1.2: Metindeki özellikleri keşfetmek için yapılan sürecin adımları

Lawrie, Croft ve Rosenberg (2001) farklı hiyerarşi modellerini, dokümanları sınıflandırmak için karşılaştırmışlar ve 'dominating set technique' yaklaşımın, diğer konu hiyerarşisi oluşturma yöntemlerine nazaran daha iyi sonuçlar sağladığını belirtmişlerdir. Hu, Bandhakavi ve Zhai (2003) çalışmalarında [3]TREC-8 konularından hangilerinin zor konu olduğunu tanımlamaya çalışmışlardır. Alt konuların konu bütünlüğünü bozduğu durumlarda incelenen konunun 'zor konu' olması durumunun gündeme geldiğini görmüşlerdir. Zhai, Coehn ve Lafferty (2003) alt konulara erişme problemini incelemişlerdir. Bu problemler kullanıcı sorgusunda, ana konunun altındaki alt konular hakkında bilgi içeren dokümanlara ulaşmayı kapsayan problemlerdir. Klasik konu erişim yöntemleri, alt konulara erişme probleminde yetersiz kalmaktadır. Alt konu anma, alt konu duyarlılık ve alt konu ağırlıklı duyarlılık performans ölçütleri tanımlanarak, istatistiksel dil modellerine dayalı, maksimum marjinal ilgililik mertebelendirme stratejisinden de etkilenen alt konu erişim metodları oluşturmuşlardır. Bu metotlarla aradıkları alt konuyla ilgili olan ve olmayan dokümanların mevcut olduğu bir veri setinde çalıştıklarında, aradıkları alt konuyla ilgili dokümanlarda yöntemin daha başarılı sonuçlar verdiğini bulmuşlardır. Mei ve Zhai (2005) istatistiksel dil modellerini, verideki konuların şablonlarını araştıran, özel geçici veri madenciliği görevlerini gerçekleştirmek için kullanmışlardır. Çalışmalarının sonunda kullandıkları metotların,

[3] TREC (Text Retrieval Conference) –8: 8. Metin Edinme Konferansı; National Institute of Standards and Technology (NIST) ve Amerikan hükümetinin yardımıyla düzenlenen konferanstır

herhangi bir metin akım verisine kolayca uygulanabileceğini keşfetmişlerdir. Bu nedenle bu çalışmada kullandıkları yöntemin, özel geçici veri madenciliğinde pek çok uygulaması olabileceğini belirtmişlerdir.

Wang, Mohanty ve McCallum (2005) "grup konu modelini" tanıtmışlardır. Aynı alanda yapılan diğer çalışmalardan farklı olarak, bu model bir metindeki gizli grupları keşfetmekle kalmaz metindeki öğelerin birbiri ile olan etkileşimini belirleyen olayların özelliklerinin kümelenmesini de keşfeder. Li ve McCallum (2005) söz dizimi sırasına göre ve anlamlarına göre kelime gruplarını oluşturmak için "yarı denetimli öğrenme metodunu" kullanmışlardır. Bu çalışmada kelime gruplarını oluşturmak için HMM-LDA modeli kullanılmıştır. Bir kelimenin olasılıksal olarak çoklu kümelere ait olmasına izin verilerek, kelimenin farklı anlamları modellenmiştir. Kullanılan yöntemin "denetimli öğrenme metodu"ndan daha iyi sonuçlar sağladığı görülmüştür.

Shen, Tan ve Zhai (2005) erişim performansını geliştirmek için geribildirim bilgisinden nasıl faydalanılacağını araştırmışlardır. Anlam bazlı erişim için dört dil modeli kullanmışlar; "FixInt, BayesInt, OnlineUp, Batch Up" ve geribildirim bilgisinin özellikle tıklama geçmişi bilgisinin erişim performansını artırdığını göstermişlerdir. Metzler ve Croft (2005a,2005b) "Markov rastgele alanı" modelini verideki terim bağımlılıklarını analiz etmek için kullanmıştır. Terim bağımlılıklarını modellemenin bilgi erişim performansını belirgin bir şekilde artırdığını göstermişlerdir. Bu çalışmada önerilen modelin 3 çeşidi söz konusudur:

1. Tamamıyla Bağımsız Model: Sorgu terimlerinin tamamıyla bağımsız olduğunu varsayar.
2. Sıralı Bağımlı Model: Ardışık sorgu terimleri arasında bazı kesin bağımlılıkların olduğunu varsayar.
3. Tam Bağımlılık Modeli: Bağımsızlığın mevcut olmadığını varsayıp, her bir sorgu terimi alt kümesindeki bağımlılığı yakalamaya çalışır.

1.2.2. Arama motoru kullanıcı oturumlarında konu değişimi tespiti için yapılan anlam bazlı olmayan çalışmalar

Anlam bazlı çalışmalar, sözcüklerin anlamına dayanır. Gerçek zamanlı uygulamalarda karmaşık, yüksek maliyetli ve zahmetlidirler. Anlam bazlı çalışmalardan başarılı sonuçlar elde edilse de, dezavantajları nedeniyle gerçek zamanlı uygulamalarda anlam bazlı olmayan istatistiksel metotlar tercih edilmektedirler. Bu çalışmalar, anlam bazlı çalışmalara göre daha düşük maliyetli ve daha basittirler. Ayrıca verileri istatistiksel olarak yorumlayıp daha gerçekçi sonuçların elde edilmesini sağlamaktadırlar.

He, Goker ve Harper (2002) konu tanımlama algoritmasını öne sürmüşlerdir. Bu algoritma, Dempster-Shafer Teori'sini (Shafer 1976) kullanmıştır. Algoritma Web arama kümelerinin ardışık sorgular arasındaki zaman aralığı ve arama yapısı gibi istatistikî özelliklerini kullanarak konu değişimlerini otomatik olarak tanımaktadır. Dempster-Shafer Teorisi iki ayrı olasılıksal olayı tek bir özellikte kombine edebilmeyi sağlar. Dempster-Shafer Teorisi her bir olayın olma olasılığını gerektirir. Bu ayrı olasılıksal olayların önem derecelerini ve konu devamını tanımlamak için bir başlangıç değerini gerektirir. Veriden yapılan analizle olasılıklar kolayca belirlenebilir. Gerekli ağırlıkları ve başlangıç değerini kullanarak genetik algoritma uygulanır. Konu tanımlama algoritmasının başarısı, duyarlılık ve anma değerleri ve ikisinin bir kombinasyonu olan uygunluk fonksiyonu değerleri ile ölçülür.

Yeni konu tanımlamasını, Özmutlu ve Çavdur (2005) yapay sinir ağlarını kullanarak belirlemeye çalışmışlardır. Bu çalışmada sorguların istatistiksel özellikleri olan zaman aralıkları ve arama yapılarından yararlanılmıştır. Kullanılan yapay sinir ağı 3 katmandan oluşmaktadır: Bir girdi katmanı, bir çıktı katmanı ve bir gizli katman. Girdi katmanında iki nöron mevcuttur. Bu nöronların biri zaman aralıklarını, diğeri ise arama yapısını temsil etmek için kullanılır; zaman aralıkları ve arama yapısını temsil eden nöronlar 1-7 arasında değerler alırlar. Çıktı katmanında sadece tek bir nöron olup, 1 ya da 2 değerini almaktadır ki 1 değeri konu devamını temsil ederken, 2 değeri konu

değişimini temsil etmektedir. Gizli katmanda ise 5 nöron bulunmaktadır. Çalışmada kullanılan yapay sinir ağı ileri beslemeli bir yapay sinir ağıdır ve geriye yayılım kullanılarak eğitilir. Excite 99 verisine ait giriş katmanı ve çıkış katmanı ile ilgili tüm bilgiler yapay sinir ağına verilir. Veri setindeki her bir zaman aralığı- arama yapısı sorgu tipi için uzman tarafından belirlenen gerçek veriler, yapay sinir ağına sağlanır ki yapay sinir ağı kullanılan sinaptik ağırlıkları eğitsin ve çıktının konu değişimi ya da konu devamı değeri doğru olarak belirlenebilsin. Tüm veri seti, eğitim veri seti ve test veri seti olmak üzere iki kısma ayrılmıştır. Eğitim verisi ile yapay sinir ağı eğitilmiş, daha sonra bu eğitilmiş yapay sinir ağı test verisi üzerinde çalıştırılmıştır. Yapay Sinir Ağları yönteminin bulduğu sonuçlarla, gerçek sonuçlar karşılaştırılarak yöntemin başarısı değerlendirilmiştir. Yapay Sinir Ağları yöntemi gerçekte de konu değişimi ve konu devamı olan sorguları tahmin etme de oldukça başarılıdır. Yalnız yöntem konu değişimi sayısını çok fazla tahmin etmiştir. Eşik değerlerinin değiştirilmesi ile bu hatanın giderilebileceği belirtilmiştir.

Özmutlu ve arkadaşları (2007a) DVM metodunu yeni konu tanımlanması problemi için kullanmışlardır. Metin halindeki sorgular, yeni konu veya yeni konu değil şeklinde iki kategoriye ayrılmıştır. Bu çalışmanın sonuçları ise DVM metodunun performansının uygulandığı veri setinin özelliklerine bağlı olduğunu göstermiştir.

Yeni konu tanımlaması için Özmutlu ve arkadaşları (2008a), yapay sinir ağlarının eğitim setinde kullanılandan farklı bir veri seti ile test edildiklerinde sonucun nasıl olacağını araştırmışlardır. Yapılan çalışmada Excite ve FAST veri setleri kullanılmıştır. Her bir verinin eğitim seti ile eğitilen yapay sinir ağı, daha sonra diğer iki verinin test setiyle çalıştırılmıştır. Yapay sinir ağı, eğitim seti ile test setinin hem aynı hem de farklı olduğu durumlarda başarılı sonuçlar vermiştir. Bu çalışmadan, belirli bir arama motorunun eğitim verisi ile eğitilen yapay sinir ağının sinaptik ağırlıklarının, herhangi bir arama motorunun test verisinde kullanılabileceği sonucu çıkarılabilmektedir. Bu da bize yapay sinir ağının evrensel olabileceğini göstermektedir ki, böylece belirli bir veri seti ile eğitilen yapay sinir ağı, başka veri setlerinin testi için kullanılabilir. Yalnız bu tespitin sadece bir bulgu olduğu ve başka veri setleri ile yapılan çalışmalarla desteklenmesi gerektiği göz ardı edilmemelidir.

Özmutlu ve arkadaşları (2008b) yeni konu tanımlaması için çoklu doğrusal regresyonu ve ANOVA'yı kullanmışlardır. Çalışma sonucunda konu değişiminde Web arama sorgularının istatistiksel karakterlerinin önemli olduğu bulunmuştur. Çoklu doğrusal regresyonun konu değişimi ve devamlılığını tanımlamada kullanılabilir bir yöntem olduğu kanıtlanmıştır. Bu çalışma Web arama sorgularının anlamsal olmayan özellikleri ile konu değişimi ve konu devamının oluşması arasında istatistiksel bağ bulunduğunu ispatlamıştır. Çalışmanın yazarların diğer arama motoru kullanıcı oturumlarında konu değişimi tespiti için yapılan anlam bazlı olmayan çalışmalarından farkı, sorgunun istatistiksel özellikleri olarak; arama yapısı, zaman aralığının yanında sorgu numarasının da kullanılmasıdır. Bir sorgunun, oturumdaki sırası o kullanıcı oturumundaki sorgunun pozisyonudur. Tüm veri setinde bir kullanıcı oturumundaki en fazla sorgu sayısı 47 olduğu için sorgular kullanıcı oturumlarında 47 sorguya kadar sınıflandırılmıştır.

Yeni konu tanımlaması için kullanılan istatistiksel yöntemler sorguların istatistiksel özelliklerine dayanmaktadır. (Özmutlu ve Çavdur 2005b, Özmutlu ve ark. 2007a, Özmutlu ve ark. 2008a, Özmutlu ve ark. 2008b). Belirtilen çalışmaların tümünde zaman aralığı ve arama yapısı sınıfları otomatik olarak belirlenmiş ve belirlenen sınıflar kullanılan tahmin yöntemleri için girdi oluşturmuştur. Bu sınıfların hatalı olarak belirlenmesi, daha sonra kullanılacak tahmin yönteminin de hatalı tahmin yapmasına neden olacaktır. Sorguların sınıflandırılma aşamasında hatalı olarak "yeni" arama yapısı sınıfına dâhil edilmesi, kullanılan yöntemlerin aşırı konu değişimi tahmini yapmalarına sebep olmaktadır. Bu hataların sebebi analiz edildiğinde "OR, +, -, &, www, http://, .com, .net, .gov, .mil, on, at, and, or" gibi ifadelerin arama yapılarının yanlış belirlenmesine sebep olduğu görülmüştür. Tahmin yöntemi uygulanmadan önce, kullanılan Excite ve FAST arama motoru verileri bu ifadelerden temizlenmiştir. Temizlenen verilere yapay sinir ağı tahmin metodu uygulanmıştır. Veri temizlenmesinin sonuçlarının karşılaştırılabilmesi amacıyla veri temizlenmeden önce de yapay sinir ağları metoduna tabi tutulmuştur. Veri temizlenmeden ve temizlendikten sonra çıkan yapay sinir ağı uygulaması sonuçları karşılaştırılmıştır. Veri temizlenmesinin sonuçlara

belirgin bir katkısı görülmemiş ve aşırı konu değişimi tahmini yapılması engellenememiştir.

Kullanıcı oturum sürelerinin belirlenmesinde Huang ve arkadaşları (2004) n-gram istatistiksel dil modeli yöntemine dayalı bir yaklaşım önermişlerdir. Çalışma geleneksel zaman aşımı, referans boyutu ve maksimum ilerleme referans metodu yöntemlerine göre daha iyi sonuçlar sağlamıştır.

1.2.3. Şartlı Olasılık

Olasılık belirsizlik ile sistematik bir şekilde uğraşan bir bilim dalıdır. Olasılığın gelişmesi 17.yüzyılda şans oyunlarındaki bazı soruların cevaplanması ile başlamıştır. Olasılığın ilgilendiği belirsizlik durumlarını anlatabilmek için para atışı (bir veya çok kere), oyun kâğıdı destesinden bir veya daha çok kart çekme, zar atışı(bir veya daha fazla zar) gibi deneyler kullanılır.

Bir çift zarın atıldığını düşünün. Bu deneyin örnek uzayı aşağıdaki 36 çıktıdan oluşacaktır.

$$S = \{(i, j),\ i = 1,2,3,4,5,6, \qquad j = 1,2,3,4,5,6\}$$

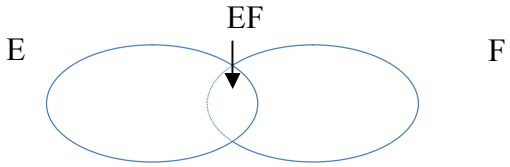
(i,j) çıktısında ise ilk zar *i* tarafını, ikinci zar ise *(j)* tarafını göstermektedir. 36 çıktıdan her birinin meydana gelme olasılığının eşit ve 1/36 olduğu varsayılsın. Ayrıca atılan ilk zarın 3 geldiği bilinmektedir. Bu bilgiler ışığında atılan iki zarın toplamının 8 olması olasılığı nedir?

İlk zarın 3 geldiği biliniyorsa, deneyimizin diğer 6 olası çıktısı (3,1), (3,2), (3,3), (3,4), (3,5), (3,6) şeklinde sıralanacaktır. Bu deneyde her bir çıktının meydana gelme olasılığı eşit olduğuna göre, her bir çıktı eşit olasılığa sahip olmalıdır. İlk zar atışında 3 sayısının geldiği bilgisi verilmektedir. Böylece çıktılar; (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6) olur. Her bir çıktının olasılığı 1/6 olarak hesaplanır. Diğer 30 çıktının örnek uzayda olma olasılığı ise 0 olarak belirlenir. Böylelikle aranılan olasılık 1/6 olarak hesaplanır.

Eğer E ve F sırasıyla, zarların toplamının 8 olması ve ilk zarın 3 olması olarak belirtilirse, elde edilen olasılık F olayı meydana geldikten sonra E olayının şartlı olasılığı olarak adlandırılır. $P(E|F)$ şeklinde gösterilir. $P(E|F)$ için genel bir formül yukarıda tasvir edildiği gibi tüm E ve F olayları için genelleştirilebilir. Eğer F olayı meydana gelirse, E olayının olabilmesi için gerçek meydana geliş noktası hem E, hem de F'de olmalıdır. F olayının meydana geldiği bilindiği için F yeni örnek uzay olmaktadır. Böylece EF olayının meydana gelme olasılığı, F olasılığına bağlı ve EF olasılığına eşit olacaktır.

$$P(E/F) = \frac{P(EF)}{P(F)} \tag{1.1}$$

Bu denklem yalnızca $P(F)>0$ olduğunda tanımlıdır ve böylelikle $P(E|F)$ yalnızca $P(F)>0$ olduğunda tanımlı olacaktır (Ross, S.M. 2000).



Şekil 1.3: Şartlı Olasılık Küme Gösterimi (Ross, S.M, 2000)

Yukarıdaki örnek için E, F ve EF kümeleri sırasıyla belirtilirse; E= {(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)}, F= {(3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6) }, EF= {(3, 5)}iken

$$P(EF) = \frac{1}{36}$$

$$P(F) = \frac{6}{36}$$

$P(E/F) = \frac{1/36}{6/36} = \frac{1}{6}$ olarak bulunur.

Deneyin n kere tekrarladığını varsayalım. F'nin meydana geldiği deneylerde $P(F)$ uzun dönemli orantı olduğundan, F yaklaşık olarak $nP(F)$ kere meydana gelecektir. Benzer şekilde, bu deneylerin yaklaşık olarak $nP(EF)$ adetinde, hem E olayı hem de F

olayı meydana gelecektir. Böylelikle çıktısı *F'*de olan yaklaşık *nP(F)* olayın, *nP(EF)* tanesinin çıktısı E'de de olacaktır. Çıktısı *F*'de olan olaylardan, çıktısı *E*'de de olanların oranı yaklaşık olarak;

$$\frac{nP(EF)}{nP(F)}\text{'dır}$$

Bu yaklaşımda n sayısı büyüdükçe 1.1 nolu denklemde verilen Şartlı Olasılık formülü oluşur.

Örnek.1: Bir kutuda 5 tane tam hasarlı, 10 tane kısmi hasarlı ve 25 tane kabul edilebilir transistör mevcuttur. Kutudan rassal olarak transistör seçilip kullanılmaktadır. Seçilen transistör tam hasarlı değilse, kabul edilebilir olma olasılığı nedir?

Transistör tam hasarlı değilse bu transistörün hasarlı 5 transistörden biri olmadığı bilinmektedir. İstenilen olasılık hasarlı olmayan transistörlerin içinden kabul edilebilir olanını bulma olasılığıdır.

$$P\left(kabul\ edilebilir|\ tam\ hasarlı\ olmayanlar\right)=\frac{P(kabul\ edilebilir)}{P(tam\ hasarlı\ olmayan)}$$

40 transistörden her birinin seçilme olasılığının eşit olduğu düşünülürse,

$$\frac{P(kabul\ edilebilir)}{P(tam\ hasarlı\ olmayan)}=\frac{25/40}{35/40}=\frac{5}{7}$$

*Örnek.2***:** Mrs. Perez, şirketinin Phoenix şehrinde ofis kurması olasılığının %30 olduğunu hesaplamaktadır. Eğer bu ofis kurulursa %60 olasılıkla, Mrs. Perez ofis müdürü olarak atanacaktır. Mrs.Perez'in Phoenix şehrinde ofis müdürü olarak atanması olasılığı nedir?

B olayı Phoenix şehrinde ofis açma olayı ve *M* olayı, Mrs.Perez'in Phoenix şehrindeki ofisin müdürü olması olayı olarak tanımlansın

$$P(BM)= P(B).P(M|B)$$

$$= (0.3).(0.6)$$

$$= 0.18$$

İstenilen olasılık 0.18 olarak bulunur.

1.2.3.1. Şartlı Olasılık uygulamaları

Crestanı ve Rıjsbergen (1998) çalışmalarında farklı bilgi erişim modelleri için erişim zamanındaki olasılıksal terim ağırlığının kinematiğini incelemişlerdir. Olasılıksal erişimi farklı kavramlara dayanan 4 adet model sunmuşlardır. Bunlardan iki adedi klasik olasılıksal teori temeline dayanırken, diğer 2 adedi mantıksal değerlendirme tekniğine dayanmaktadır. Bu son iki teknikten biri şartlı olasılığı değerlendirmeye dayanan mantıksal bir tekniktir. Diğeri ise sanal teknik olarak adlandırılmaktadır.

Jin ve arkadaşları (2002) yaptıkları çalışmada 'başlık dil modeli' isimli yeni bir dil modeli önermişlerdir. Geleneksel dil modellerinden farklı olarak, Şartlı Olasılık tanımlamışlardır. Şartlı Olasılık *P(Q|D)* olarak tanımlanmış olup, *Q* sorgusunu *D* dokümanı için başlık olarak kullanabilme olasılığını yansıtır. *P(Q|D)* olasılığını hesaplamak için koleksiyondaki başlık ve doküman çiftlerinden öğrenilen istatistiksel çeviri modeli adapte edilmiştir. Bu çalışmaya kadar kullanılan yaklaşımlarda doküman dil modelini tahmin edebilmek için sorgu oluşturma yöntemi kullanılırken eksik olan ortak nokta, sorguların ve dokümanların farklı sorgu özellikleri olduğu için farklı stokastik süreçler doğrultusunda oluşturulması gerektiğidir. Bu eksikliği doldurmak içinde dokümanın gözlenmesine dayalı 'sorgu dil modeli' oluşturulmalıdır. Çalışmanın temel noktası, doküman için 'sorgu dil modelini', 'başlık dil modeline' yakınlaştırmaktır. Öncelikle koleksiyondaki tüm başlık-doküman çiftleri kullanılarak 'geçiş modeli' tahmin edilmiştir. Bu 'geçiş modeli' daha sonra kurala uygun bir 'doküman dil modelini', 'başlık dil modeline' çevirmek için kullanılmıştır. Son olarak, sorgu benzerliğini hesaplamak amacıyla her bir doküman için başlık dil modeli kullanılmıştır.

Yao ve arkadaşları (2008) yüksek kaliteli tavsiyeler elde etmek ve tavsiye sistemlerinde veri kıtlığı durumuyla karşı karşıya kalındığında yüksek kaliteli sonuçlar elde etmek için, Sınıf Bazlı Ayarlanmış Şartlı Olasılık Benzerliği ortak filtreleme tekniğini kullanmışlardır. Bu teknikte öncelikle kullanıcı-öğe matrisi analiz edilmiştir. Farklı öğeler arasındaki ilişkiler belirlenmiştir. Bu ilişkiler kullanıcılar için tavsiyeleri

hesaplarken direkt olmayan bir yolla kullanılmıştır. Tavsiyelerde kullanılan ‘derecelendirme’ tahminlerinde, k-en yakın komşuluğu derecelendirme tekniğinde kullanılan ağırlıklı ortalama kullanılmıştır. Daha sonra bu yöntemden elde edilen sonuçlar, korelâsyon benzerliği, kosinüs benzerliği ve ayarlanmış kosinüs benzerliğini kullanan k-en yakın komşuluk yaklaşımına göre daha iyi sonuçlar vermiştir. Deneylerde de bu çalışmada kullanılan algoritmanın diğer öğe bazlı algoritmalardan daha iyi sonuç verdiği ortaya konulmuştur. Çalışmanın temel noktası farklı kategoriler için öğelerin sınıflandırılması ve daha sonra benzerliği hesaplamak için şartlı olasılığın kullanılmasıdır.

1.2.4. Monte Carlo Simülasyonu

Monte Carlo Simülasyonu stokastik ya da deterministik problemleri çözmek için kullanılan $U_{[0,1]}$ düzgün dağılımına göre rassal sayı üretilip bu sayıların simülasyon için kullanıldığı ve zamanın belirgin bir rol oynamadığı bir düzendir.

Monte Carlo Simülasyonu uygulaması aşağıdaki örnekle anlatılmıştır.

$$I = \int_a^b g(x)dx \tag{1.2}$$

1.2 formülünde belirtilen integral değerlendirilmek istensin.

Yukarıdaki *g(x)* fonksiyonu gerçek değerli bir fonksiyon olup integrali alınamamaktadır. Bu deterministik problemin, Monte Carlo simülasyonu yaklaşımı ile nasıl çözüleceğine bakılırsa;

Y rassal bir değişken olsun ve $(b-a)g(X)$ olarak tanımlansın. X, [a,b] aralığında düzgün dağılan sürekli rassal değişkeni temsil etmektedir. $f_x(x) = \frac{1}{(b-a)}$ ’nın beklenen değeri:

$$E(Y) = E[(b-a)g(X)] \tag{1.3}$$

$$= (b-a)E[g(X)] \tag{1.4}$$

$$= (b-a)\int_a^b g(x)f_x(x)dx \tag{1.5}$$

$$=(b-a)\frac{\int_a^b g(x)dx}{(b-a)} \tag{1.6}$$

$$=I \tag{1.7}$$

olarak tanımlanır ve $U_{[a,b]}$ rassal değişkeninin olasılık yoğunluk fonksiyonudur. Böylece integrali değerlendirme problemi $E(Y)$ beklenen değerini tahmin etmeye kadar indirilmiştir. Özellikle $E(Y)=I$ örneklem ortalamasından tahmin edilecektir.

$$\bar{Y}(n)=\frac{\sum_{i=1}^{n} Y_i}{n}=(b-a)\frac{\sum_{i=1}^{n} g(X_i)}{n} \tag{1.8}$$

$X_1,X_2,...,X_n$ ise iid[4] $U_{[a,b]}$ düzgün dağılıma uygun rassal değişkenlerdir.

$\bar{Y}(n)$ değerini bir dikdörtgenin alanı olarak tahmin etmek yol gösterici olabilir. *(b-a)* boyunda ve *I/(b-a)* yüksekliğinde ([a,b] aralığı üzerinde *g(x)*'in devamlı ortalaması). Üstelik $E[\bar{Y}(n)]=I$ olduğunu aşağıdaki şekilde göstermek mümkündür.

$\bar{Y}(n)$, I'nın yansız bir tahmin edicisi ise ve

$$Var\left[\bar{Y}(n)\right]=Var(Y)/n\,. \tag{1.9}$$

Bu durumda *Var(Y)* sonlu ise, $\bar{Y}(n)$ yeterince büyük n için *I*'ya yakın olacaktır.

Yukarıda anlatılanları sayısal olarak göstermek istersek;

$I=\int_0^{\pi} \sin x dx$ integrali hesaplanmak istensin. (1.10)

Bu ifadenin değeri 2 olacaktır. Bu integralin hesaplanması için Monte Carlo simülasyonu uygulanması Çizelge 1.1'de gösterilmiştir (Law ve Kelton,1991).

[4] independent and identically distributed : bağımsız ve özdeş dağılan

Çizelge 1.1: 1.11 formülü ile verilen ifadenin Monte Carlo Simülasyonu ile Bulunması

n	10	20	40	80	160
$\bar{Y}(n)$	2.213	1.951	1.948	1.989	1.993

Monte Carlo Simülasyonu analitik olarak karışık pek çok problemin çözümünde kullanılmıştır. Monte Carlo tekniğinde yapay veri, rassal sayı üretici ve ilginin kümülâtif dağılımı yardımıyla oluşturulmaktadır (Pegden ve ark. 1995). Mantıklı ve kabul edilebilir rassal sayı üreticisinin kullanılması önemlidir. Çünkü üretilen rassal sayılar aslında gerçek rassal değil; pseudo-rassaldır. Yani rassal sayı aralığı gerçekte yeniden üretilebilir (Pegden ve ark. 1995). Yeniden üretilebilirlik lüzumlu olduğunda deneyi tekrar yapmak için gerekebilir. Monte Carlo simülasyonu için genellikle $U_{[0,1]}$ düzgün dağılımından rassal sayılar üretilir ve rassal sayıya göre uygun cevap seçilir.

Madeni paranın havaya atılması olayı incelenirse, öncelikle $U_{[0,1]}$ aralığından bir adet sayı üretilmesi gereklidir. Üretilen sayı 0.5 değerinin altındaysa tura, diğer türlü ise yazı değeri atanacaktır (ya da tam tersi; verilecek cevaba atanan rassal sayı aralığına göre) (Özmutlu ve ark. 2006).

1.2.4.1. Rassal değişken üretimi

Simülasyon metodları rassal değişken üretimine, özellikle bağımsız rassal değişken üretimine bağlıdır. $U_{[0,1]}$ düzgün dağılımı rastsallığın basit olasılıksal gösterimini sağlar. Tüm diğer dağılımlarsa simüle edilecek düzgün değişken aralığı gerektirir. Bu nedenle [0,1] aralığında düzgün dağılıma uyan rassal değişken üretimi üzerinde durulmuştur.

1.2.4.2. Düzgün dağılımla simülasyon

Rassal sayı üretimi ile ilgili mantıksal paradoks [0,1] aralığında deterministik değerler dizisi üretmektir. Bu değerler, iid düzgün dağılıma uygun rassal değişkenlerini $U_{[0,1]}$ taklit eder.

Rassal dizi üretmek için tamamıyla deterministik süreç kullanan metotlar mevcuttur. $(X_1, X_2,.....,X_n)$ üretilmiş olsun. X_n değerinin bilinmiş olması ya da $X_1, X_2,.....,X_n$ değerlerinin bilinmiş olması, dönüşüm fonksiyonu mevcut değilse X_{n+1} değerinin bilinmesini sağlamaz. Verilen başlangıç değeri X_0, dönüşüm fonksiyonu ve örneklem $(X_1, X_2,.....,X_n)$ her zaman sabittir. Böylelikle bu tekniklerle üretilen 'pseudo rassallık' sınırlandırılabilir. Çünkü algoritma tarafından üretilen iki örneklem $(X_1, X_2,.....,X_n)$ ve $(Y_1, Y_2,.....,Y_n)$ bağımsız olmayacak, özdeş şekilde dağılmayacak, herhangi bir olasılıksal anlamda karşılaştırılabilir olmayacaktır. Rassal sayı üreticisinin geçerliliği n $+\infty$'a giderken, $X_1, X_2,.....,X_n$ örneklemine dayanmaktadır; n'nin sabit olduğu ve k'nın sonsuza gittiği $(X_{11}, X_{12},.....,X_{1n})$, $(X_{21}, X_{22},.....,X_{2n})$..... $(X_{k1}, X_{k2},.....,X_{kn})$ örneklem tekrarlarına dayanmamaktadır.

Yukarıda belirtilen sınırlama unutulmamalıdır. Aslında n değerlerinin dağılımı, üretilen X_{r1} $(1 \le r \le k)$ başlangıç değerlerinin davranışına bağlıdır.

Belirtilen sınırlama düşünülerek, aşağıdaki tanım önerilebilir:

Düzgün dağılıma uygun pseudo-rassal sayı üretici, bir algoritmadır. Başlangıç değeri u_0 olan, D dönüşümü uygulayıp, [0,1] aralığında $u_i = (D^i(u_0))$ değerler dizisi üreten bir algoritmadır. Tüm n sayıları için, $u_1,.....u_n$ değerleri iid düzgün dağılıma uygun rassal değişken örnekleminin $(V_1,....,V_n)$ davranışını yeniden üretir. $((V_1,....,V_n)$ örnekleminin düzgün dağıldığı testlerle belirlenmiştir.)

Bu tanım deterministik dönüşüm $(u_i=D(u_{i-1}))$, ile bağlantılı olan rassal sayı üretiminin test edilebilir unsurlarıyla sınırlıdır. Algoritmanın geçerliliği, $U_1,.....,U_n$ dizisinin H_0 hipotezinin kabul edilmesine sebep olduğunda doğrulanır.

$$H_0 = U_1,.....,U_n \text{ dizisi iid } U_{[0,1]}$$

Kolmogorov-Smirnov gibi klasik düzgünlük belirleyici testler mevcuttur. Böyle bir kontrol altında pek çok üretici yeterli görünecektir. Ek olarak, zaman serisi metotları U_i ve (U_{i-1},.....,U_{i-k}) arasındaki korelasyonun derecesini belirlemek için kullanılabilir. ARMA *(p,q)* modeli bu metotlara örnek olarak gösterilebilir. Lehmann (1975) veya Randles ve Wolfe (1979) testleri gibi parametrik olmayan testler kullanılabilir.

Böylelikle *Düzgün dağılıma uygun pseudo-rassal sayı üreticisi* tanımı işlevseldir. Test setleri tarafından red edilmediyse, düzgün dağılıma uygun rassal sayılar üreten algoritma kabul edilebilir. Bu metodolojinin de bazı problemleri vardır. Örneğin çok sayıda iterasyon gerektiren bazı uygulamalarda, geniş sapma teorilerinde ya da parçacık fiziğinde standart testlere dirençli olan algoritmalar, ciddi hatalar sergileyebilirler. Özellikle periyodikliği gizleyen ya da daha küçük basamaklar için düzgün olmayan algoritmaların belirlenmesi zor olabilir.

'Deterministik bir sistem rassal bir olguyu taklit edebilir' kavramından yola çıkılarak, rassal sayı üreticisi oluşturmak için kaotik modellerin kullanımı önerilebilir. Karmaşık deterministik yapılarla sonuçlanan bu modeller, dinamik sistem formuna; $X_{n+1}=D(X_n)$ dayalıdır. Bu form, başlangıç koşulu X_0'a karşı oldukça hassastır.

1.2.4.3. Ters dönüşüm

Rassal değişken alanın yapısını tanımlarken, genel olasılık üçlüsünü (Ω, F, P), ([0,1], B, $U_{[0,1]}$) olarak temsil etmek mümkündür. Bu üçlüde Ω tüm alanı göstermektedir, F , Ω 'da σ-*cebirini* temsil eder, P ise olasılık ölçüsüdür. ([0,1], B, $U_{[0,1]}$) üçlüsünde B, [0,1] aralığındaki Borel setini göstermektedir. Böylelikle [0,1] aralığındaki düzgün değişkenle *w'*nin varyasyonuna, $w \in \Omega$ denk gelmektedir. X rassal değişkenleri, [0,1]'den *X'*e fonksiyonlardır. Genelleştirilmiş ters fonksiyon tarafından dönüştürülen, düzgün dağılıma uygun değişkenlerin fonksiyonudur.

R'de artmayan bir F fonksiyonu için, F'nin genelleştirilmiş tersi, F^- fonksiyonu;

$$F^-(u) = \inf \{x: F(x) \geq u\} \qquad (1.11)$$

şeklinde tanımlanır. Daha sonra aşağıdaki kuram oluşturulur. Bu kuram olasılık integral dönüşümü olarak da bilinir. Bu kuram bize herhangi bir rassal değişkenin, düzgün dağılıma uygun rassal değişken dönüşümü olarak temsilini verir.

Kuram: Eğer $U \sim U_{[0,1]}$ ise, rassal değişken $F^{-}(U)$, F dağılımına sahiptir.

Kanıt: Tüm $u \in [0,1]$ ve tüm x $\in F^{-}([0,1])$ için, genelleştirilmiş ters dönüşüm;

$$F(F^{-}(u)) \geq u \text{ ve } F^{-}(F(x)) \leq x \tag{1.12}$$

koşullarını sağlar.

Böylelikle;

$$\{(u,x): F^{-}(u) \leq x\} = \{(u,x): F(x) \geq u\} \tag{1.13}$$

ve

$$P(F^{-}(U) \leq x) = P(U \leq F(x)) = F(x) \tag{1.14}$$

Böylelikle $X \sim F$ rassal değişkeni üretmek için $U_{[0,1]}$ dağılımına göre U üretip, $x = F^{-}(u)$ dönüşümünü yapmak yeterli olacaktır.

1.2.4.4. Kesikli olay simülasyonu uygulamaları

Simülasyon, bilişim teknolojisinde Deshpande ve arkadaşları tarafından (1996) istemci-sunucu modellerini modellemek için kullanılmıştır. Bu çalışmada istemci-sunucu modeli, OSI (Open System Interconnection) modelinin 7 katmanıyla ilişkilendirilerek anlatılmış olup, benzer tabaka bazlı bir yaklaşımın simülasyon çalışmasında da kullanılabileceği belirtilmiştir. Bu yaklaşımın yararları ise istemci-sunucu modelini daha iyi anlamak, ticari olarak uygun simülasyon yazılımlarının kullanılabilirliğinin mümkün olması olarak sıralanabilir.

Simülasyon, internet ve diğer bilgisayar ağlarını modellemek için Breslau ve arkadaşlarının (2000) çalışmalarında kullanılmıştır. Araştırmacılar internet protokollerinin sağlam ve güvenilir olduğunun kanıtlanması için, çeşitli koşullar altında incelenmeleri gerektiğini belirtmişlerdir. VINT: Sanal Ağlararası Test Yatakları Projesi'nin ağ simülatörünün ve ilgili yazılımın geliştirildiğini belirtmişlerdir. Bu

gelişmelerinse, araştırmacıların ağ protokollerini değerlendirdiği koşulları genişletecek pratik yenilikler sağlayabileceğini bildirmişlerdir.

Floyd ve arkadaşları (2001) interneti simüle ederken ki zorlukları incelemişlerdir. Ağın hızlı ve heterojen bir yapısı olduğunu ve bu nedenle simüle etmenin zor olduğunu belirtmişlerdir. Bu zorluklarla yüzleşerek anlamlı simülasyonlar yapmak içinse iki yöntem önermişlerdir. Bu yöntemler sabit parametre için arama yapmak ve parametre uzayını tedbirlice keşfetmektir. Sabit parametreler; internetteki akvitelerin günlük yapısı, toplu internet trafiğinde paket varışlarında görülen uzun dönemli korelâsyonlar, poisson oturum varışları, log-normal bağlantı büyüklükleri, ağır kuyruklu dağılımlar (ağır kuyruk derken $\alpha<2$ olan Pareto dağılımı kastedilmektedir), global topolojinin sabit özellikleri olarak sıralanabilir. İkinci yöntem olan parametre uzayını tedbirlice keşfetmek ise aslında simülasyon senaryosunun hassas olduğu durumları tanımlamaya yaramaktadır.

Kesikli olay simülasyonu bilgisayar ağları simülasyon çalışmalarının çoğunda kullanılmıştır. Bilgi erişim terimlerinde simülasyona, genellikle bilgi erişim sistemlerin performansını test etmek için başvurulmuştur. Burkowski (1990) dağıtılmış bir bilgi erişim sisteminin performansını ölçmek için simülasyonu kullanmıştır. Yine Cahoon ve McKinley (1996) sorgu birleştirmeli bilgi erişim sistemine dayalı dağıtılmış bilgi erişim sistemin performansını incelemek üzere simülasyonu gerçekleştirmişlerdir. Cahoon ve arkadaşları (2000) dağıtılmış bilgi erişim sisteminin performansını; ‘müşteri komut oranı, doküman grubu sayısı, sorgu başına terimler, sorgu terimi sıklığı, geri dönen cevapların sayısı, komut karışımı gibi’ sistem parametrelerinin fonksiyonu olarak ölçmüştürler. Coevreur ve arkadaşları (1994) simülasyonu kullanarak, büyük metin gruplarının arama performansını analiz etmişlerdir. Cacheda ve arkadaşları (2005) kesikli olay simülasyonunu dağıtılmış bilgi erişim sisteminin farklı mimarilerinin performansını incelemek üzere kullanmışlardır. Diğer çalışmalardan farklı olarak simülasyonun kendisi bilgi erişim algoritması olarak kullanılmıştır.

2. MATERYAL VE YÖNTEM

2.1. Materyal

2.1.1. Araştırmada kullanılan veriler

Excite 99, Excite 2001 ve FAST veri grubu olarak adlandırılan 3 gerçek veri grubu kullanılmıştır.

2.1.1.1. Excite 99 veri grubu

Çalışmada kullanılan veriler Excite arama motorundan (http://www.excite.com) edinilmiştir. Veri, 20 Aralık 1999 tarihinde toplanmış olup, 1.025.910 adet arama sorgusundan oluşmaktadır. Excite'daki veri yapısında, veri girişi kaydı, verinin arama motoruna girdiği sıraya göre oluşturulmaktadır. Her yeni kullanıcıya tekil bir numara (ID) atanmaktadır ve yeni kullanıcılar ID'lerinden ayırt edilebilmektedir Yeni bir oturumu, kullanıcının kimlik numarası ile tanımlamak mümkündür. Her bir sorgu 3 bölümden oluşmaktadır:

1) Kimlik (Identification) (IP adresi): Arama motoru tarafından kullanıcıya atanan anonim bir kod

2) Zaman: Veri girişi olan zamanın saat, dakika ve saniye cinsinden değeri

3) Sorgu: Kullanıcı tarafından girilen terimler

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 bir çok sorgu barındırabilmektedir(Spink ve ark. 2001).

Excite 99 verisinde 10003 adet sorgu, 1.025.910 sorgu arasından Poisson örneklemesi kullanılarak seçilmiştir(Özmutlu ve ark. 2002). Her yeni kullanıcıya tekil bir numara (ID) atanmaktadır ve yeni kullanıcılar ID'lerinden ayırt edilebilmektedir. Tekil kullanıcı listesi hazırlanıp, bu liste üzerinden Poisson örneklemesi yöntemi ile tespit edilen kullanıcıların tüm arama kayıtları seçilerek 10003 adet sorgu kümesi elde edilmiştir. Çalışma sonuçlarının performans değerlendirmesi için uzman tarafından sorguların gözden geçirilip gerçek bilgilerin elde edilmesi gerekmektedir. Bu nedenle de örneklem büyüklüğü geniş tutulamamıştır. Yeni konu tanımlama amacıyla yapılan önceki çalışmalarda da kullanılan veri seti, eğitim ve test kümesi olarak Çizelge 2.1'de gösterildiği gibi yaklaşık olarak iki eşit parçaya ayrılmıştır.

2.1.1.2. Excite 2001 veri grubu

Veri 4 Mayıs 2001 tarihinde toplanmış olup 1.700.000 sorgudan oluşmaktadır ve 10256 adet sorgu Poisson örneklemesi kullanarak (Özmutlu ve ark.2002) seçilmiştir. Veri yapısı ve sorgunun özellikleri Excite 99 verisi ile aynıdır. Veri seti, eğitim ve test kümesi olmak üzere Çizelge 2.1'de gösterildiği gibi yaklaşık olarak iki eşit parçaya ayrılmıştır.

2.1.1.3. FAST veri grubu

FAST arama motoru'ndan (http://www.alltheWeb.com) 1.257.891 adet sorgu içeren bir veri grubu edinilmiştir. Bu sorgular 6 Şubat 2001 saat 12.00'den, 7 Şubat 2001 saat 12.00'ye kadar elde edilen sorgulardır. Yine Poisson örneklemesi kullanılarak (Ozmutlu ve ark.2002), 1.257.891 adet sorgudan, 10.007 sorguyu içeren bir örneklem seçilmiştir. Veri yapısı ve sorgunun özellikleri Excite 99 ve Excite 2001 verisiyle aynıdır. Veri seti, eğitim ve test kümesi olmak üzere Çizelge 2.1'de gösterildiği gibi yaklaşık olarak iki eşit parçaya ayrılmıştır.

Çizelge 2.1: Çalışmada kullanılan veriler

Arama Motoru	**Excite 99**	**Excite 2001**	**FAST**
Bütün veri kümesi	1.025.910	1.7 milyon	1.257.891
Örneklem Kümesi	10.003	10.256	10.007
Örneklem kümesinin ilk kısmı	5014 sorgu	5128 sorgu	4997 sorgu
Örneklem kümesinin ikinci kısmı	4989 sorgu	5128 sorgu	5010 sorgu

Çizelge 2.2: Çalışmada kullanılan verilerin ayrıntılı analizi

Veri seti	Sorguların toplam sayısı	Oturum sayısı	Şartlı Olasılık/Monte-Carlo simülasyonu uygulanan sorguların sayısı	Uzman kişi tarafından işaretlenen konu değişimlerinin sayısı	Uzman kişi tarafından işaretlenen konu devamlarının sayısı
Eğitim veri seti	5014-Excite 1999 5128-Excite 2001 4997-FAST	1201-Excite 1999 1858-Excite 2001 437-FAST	3813-Excite 1999 3270-Excite 2001 4560-FAST	269-Excite 1999 391-Excite 2001 386-FAST	3544-Excite 1999 2879-Excite 2001 4174-FAST
Test veri seti	4989-Excite 1999 5128-Excite 2001 5010-FAST	1322-Excite 1999 1734-Excite 2001 526-FAST	3667-Excite 1999 3394-Excite 2001 4484-FAST	152-Excite 1999 272-Excite 2001 310-FAST	3515-Excite 1999 3122-Excite 2001 4174-FAST
Tüm veri seti	10003-Excite1999 10256-Excite2001 10007-FAST	2523-Excite 1999 3592-Excite 2001 963-FAST	7480-Excite 1999 6664-Excite 2001 9044-FAST	421-Excite 1999 663-Excite 2001 696-FAST	7059-Excite 1999 6001-Excite 2001 8348-FAST

2.2. Yöntem

Yukarıda bahsedilen 3 veri seti de, eğitim verisi ve test verisi olarak 2 gruba ayrılmıştır. Eğitim verisinde 2.2.1.3. bölümünde belirtilecek olan zaman aralığı ve arama yapısındaki sorguların sayıları tespit edilmiş ve yine 1.1. formülüne istinaden Şartlı Olasılık değerleri her bir zaman aralığı ve arama yapısındaki sorgu için hesaplanmıştır. Daha sonra test verisindeki sorgunun konu devamı mı konu değişimi mi olduğu, Şartlı Olasılık ve Monte Carlo Simülasyonu yöntemleriyle tahmin edilmiştir. Tahminlerin başarısı ise tanımlanan performans ölçütlerinin hesaplanmasıyla yorumlanmıştır.

2.2.1. Şartlı olasılık yöntemi

Yöntem 7 aşamadan oluşmaktadır.

2.2.1.1. Bütün verinin uzman tarafından değerlendirilmesi

Uzman kişi tüm veriler üzerinde yaptığı değerlendirmede oturum içindeki ardışık sorguları karşılaştırarak , sorgular arasında konu değişimi veya konu devamı olduğu kararını verir. Bu çalışma MS-Excel'de yapılmış olup, konu devamı 1 ile konu değişimi 2 ile sembolize edilmiştir. Her bir oturumun son sorgusu, kendisinden sonra gelen sorgu bulunmadığı, dolayısıyla ardışıklık durumu söz konusu olmadığı için analize alınmamıştır.

2.2.1.2. Verinin iki gruba ayrılması

Yaklaşık olarak verinin ilk yarısı eğitim verisi (Excite 99 verisinde 5014, Excite 2001 verisinde 5128, FAST verisinde 4997), ikinci yarısı ise (Excite'da 4989 adet sorgu, Excite 2001 verisinde 5128,FAST'ta 5010 adet sorgu) test verisi olarak kullanılmıştır. Veri setlerinin ortasındaki sorguyu içeren kullanıcı oturumunun

bütünlüğünü korumak için, veri setleri eşit sayıda sorgu içermemektedir(Özmutlu 2006).

2.2.1.3. Her bir sorgunun zaman aralığının ve arama yapısının belirlenmesi

Özmutlu ve Çavdur (2005a) yaptıkları çalışmada, He ve arkadaşlarının (2002) çalışmalarını baz alarak, oturumdaki ardışık sorguların arama yapılarını (search pattern - *sp*) 7 farklı sınıfta ve zaman aralığı kategorilerini (time interval - *ti*), ardışık sorguların gelişleri arasındaki farkların uzunluğu göz önüne alınarak 7 farklı kategoride toplamıştır: 0–5 dk, 5–10 dk, 10–15 dk, 15–20 dk, 20–25 dk, 25–30 dk ve 30+ dk.

Çizelge 2.3: Sorgu bilgilerinden elde edilen 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

Zaman aralıklarına göre her bir eğitim veri seti için sorguların dağılımı Çizelge 2.4'de görülebilir.

Çizelge 2.4: Excite ve FAST eğitim setinde zaman aralıklarına göre konu devamı/ konu değişiminin dağılımı

	Excite 99 Eğitim Verisi		Excite 2001 verisi		FAST Eğitim Verisi	
Zaman aralığı (Dk)	Konu devamı sayısı	Konu değişimi sayısı	Konu devamı sayısı	Konu değişimi sayısı	Konu devamı sayısı	Konu değişimi sayısı
1	3001	77	2265	137	3464	95
2	218	18	229	38	285	27
3	85	14	109	28	112	24
4	47	7	42	8	56	19
5	22	13	34	13	33	17
6	20	5	25	9	24	10
7	151	135	175	158	200	194
Toplam	3544	269	2879	391	4174	386

Arama yapısının 7 kategorisi ise aşağıda tanımlanmıştır. Arama kategorileri hazırlanan bir PASCAL kodu yardımıyla otomatik olarak belirlenmiştir. PASCAL kodunun mantık akışı Şekil 2.1'de, arama yapısına göre eğitim verilerindeki sorguların dağılımı ise Çizelge 2.4'de verilmiştir.

**Yeni*: İkinci sorgu birinci sorguyla ortak terim içermemektedir.

$Sorgu_j$: Otomobil

$Sorgu_{j+1:}$ Harry Potter

**Sonrakisayfa:* İkinci sorgu birinci sorguyla aynıdır. Yani ikinci sorgu, birinci sorguyla ilgili diğer bir sonuç kümesini istemektedir.

$Sorgu_j$: Otomobil

$Sorgu_{j+1}$: Otomobil

**Genelleme:* İkinci sorgu birinci sorgudan daha az terim içermektedir ve ikinci sorgunun bütün terimleri birinci sorguda yer almaktadır.

$Sorgu_j$: Kırmızı otomobil

$Sorgu_{j+1}$:Otomobil

**Özelleştirme:* İkinci sorgu birinci sorgudan daha fazla terim içermektedir ve birinci sorgunun bütün terimleri ikinci sorguda yer almaktadır.

$Sorgu_j$: Otomobil

$Sorgu_{j+1:}$ Kırmızı Otomobil

*Düzenleme: İkinci sorgu terimlerinin bazıları (tamamı değil) birinci sorguda yer almaktadır ancak birinci sorgu, ikinci sorguda yer almayan bazı terimleri de içermektedir. 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.

$Sorgu_j$: Kırmızı otomobil Toyota

$Sorgu_{j+1}$: Otomobil Corolla

*İlgili Geri-Besleme: İkinci sorgu hiç terim içermemektedir ve kullanıcı ilgili sayfalar seçeneğini seçtiğinde sistem tarafından oluşturulmaktadır.

$Sorgu_j$: Otomobil

$Sorgu_{j+1}$: ‘‘ ’’

**Diğer*: Eğer ikinci sorgu yukarıdaki sınıfların hiçbirine uymuyorsa o zaman o sorgu diğer diye sınıflandırılır

$Sorgu_j$: ‘‘ ’’

$Sorgu_{j+1}$: Toyota otomobil

Input: Q_{i-1}, Q_i, Q_{i+1}

Local: Q_c: birinci sorgu

Q_n: ikinci sorgu

$B=\{t \mid t \in Q_c \wedge t \in Q_n\}$// her iki sorguda da ortak olan terimler

$C=\{t \mid t \in Q_c \wedge t \notin Q_n\}$// sadece birinci sorguda olan terimler

$D=\{t \mid t \notin Q_c \wedge t \in Q_n\}$// sadece ikinci sorguda olan terimler

Output: sp// arama yapısı(search pattern)

Begin

If ($Q_i == \phi$) Then

If (i==1) Then *sp*=Diğer

Else $Q_c = Q_{i-1}$

End If

Else

$Q_c = Q_i$

$Q_n = Q_{i+1}$

End If

sp=Diğer

If($Q_n == \phi$) Then *sp*=Uygun Geri Bildirim End If

If($B \neq \phi \wedge C \neq \phi \wedge D == \phi$) Then sp=Genelleştirme End If

If($B \neq \phi \wedge C == \phi \wedge D \neq \phi$) Then sp=Özelleştirme End If

If($B \neq \phi \wedge C \neq \phi \wedge D \neq \phi$) Then sp=İlgili Geri Besleme End If

If($Q_n \neq Q_c \wedge B \neq \phi \wedge C == \phi \wedge D == \phi$) Then sp=İlgili Geri Besleme End If

If($Q_c \neq \phi \wedge B == \phi$) Then sp=Yeni End IfEnd

Şekil 2.1: Arama Yapısı Sınıfı Belirleme Algoritması (Özmutlu ve ark. 2005a)

Çizelge 2.5: Excite ve FAST eğitim verisinde arama yapısına göre konu devamı/ konu değişiminin dağılımı

	Excite 99 Eğitim Verisi		Excite 2001 verisi		FAST Eğitim Verisi	
Arama Yapısı	Konu devamı sayısı	Konu değişimi sayısı	Konu devamı sayısı	Konu değişimi sayısı	Konu devamı sayısı	Konu değişimi sayısı
1	2371	0	1627	0	3100	5
2	58	0	88	1	39	0
3	166	0	186	0	136	2
4	327	1	412	21	276	5
5	622	268	566	369	551	370
6	0	0	0	0	70	2
7	0	0	0	0	2	2
Toplam	3544	269	2879	391	4174	386

2.2.1.4. Konu değişimi ve devamı için şartlı olasılıkların hesaplanması

Ek 1'de verilen sorgu zamanına ve arama yapısı kombinasyonlarına göre, 1.1 formülü'ne istinaden şartlı olasılıklar hesaplanır. 1–5 sorgu kombinasyonundaki veri göz önüne alınırsa; Excite 99 veri setindeki 1–5 özelliğine sahip olan sorgulardan 403 adedi konu devamı ve 76 adedi konu değişikliği özelliğine sahiptir. 1-5 kategorisinde toplam 479 adet sorgu vardır. *Olasılık sorgunun 1-5 olması şartına bağlı olduğundan Şartlı Olasılık kullanılmıştır.* Toplam sorgu sayısının olasılığı, *P(F)* kümesini oluşturmaktadır. 1-5 şartına bağlı sorguların konu devamı şartlı olasılığı hesaplanmak istenirse; konu devamı olasılığı 1.1 nolu formülde *P(EF)* olasılığına karşılık gelmektedir.

1-5 kategorisindeki sorguların konu devamı şartlı olasılığı

$$P(E/F) = \frac{P(EF)}{P(F)} = \frac{403/479}{479/479} = \frac{403}{479} = 0.841$$ olarak hesaplanır.

1-5 şartına bağlı sorguların konu değişimi şartlı olasılığı hesaplanmak istenirse; konu değişimi olasılığı yukarıdaki formülde *P(EF)* olasılığına karşılık gelmektedir.

$P(E/F) = \frac{P(EF)}{P(F)} = \frac{76/479}{479/479} = \frac{76}{479} = 0.159$ olarak hesaplanır.

Hesaplanan Şartlı Olasılık değerleri Ek 2'de verilmiştir.

2.2.1.5. Şartlı Olasılıklar kullanılarak, test veri grubundaki sorguların konu değişimi - konu devamının tahmin edilmesi

Eğitim setinden elde edilen şartlı olasılıklar kullanılarak, test veri setindeki her bir sorguda konu değişimi olup olmadığı tahmin edilmeye çalışılmıştır. Arama yapısı-zaman aralığı gibi belli bir özelliğe sahip sorguda şartlı olasılıklardan hangisi daha büyükse o durumun gerçekleşeceği varsayılır.

Excite 99 test verisinde, zaman aralığı bakımından ikinci gruba ve arama yapısı bakımından beşinci gruba ait olan sorgular için konu değişimi olup olmadığı tahmin edilmek istensin. Ek 2'de görüldüğü üzere Excite 99 verisinde 2–5 kategorisinde sınıflandırılan sorguların konu devamı şartlı olasılığı 0,75 iken, konu değişimi şartlı olasılığı 0,25'tir. Bu durumda Şartlı Olasılık yöntemi kullanılarak test verisinde 2–5 sorgusu ile karşılaşıldığında yapılacak tahminde konu devamı kararı verilecektir.

2.2.1.6. Uzman kişinin bulduğu sonuçlarla, Şartlı Olasılıklar kullanılarak bulunan sonuçların karşılaştırması

Uzman kişinin bulduğu gerçek test verisi sonuçları ile Şartlı Olasılık metodu tarafından tahmin edilen sonuçlar her bir performans ölçütü için karşılaştırılmıştır. Bu kısım tezin 3.bölümü olan araştırma sonuçlarında detaylıca verilmiştir.

2.2.1.7. Şartlı Olasılık yöntemi kullanılarak bulunan sonuçlarda hatalı kısımların tespit edilmesi ve hataların sınıflandırılması

Konu değişimleri Şartlı Olasılık metodu ile tahmin edilip, uzman sonuçları ile karşılaştırıldıktan sonra iki tip hata ortaya çıkar.

A tipi hata: Şartlı Olasılık yöntemi tarafından konu değişimi olarak tahmin edilip gerçekte konu devamı olan sorgular

B tipi hata: Şartlı Olasılık yöntemi tarafından konu devamı olarak tahmin edilip gerçekte konu değişimi olan sorgular

A tipi hata ve B tipi hata aşağıdaki sorgu tiplerine göre sınıflandırılmıştır:

- 7 değişik sorgu zamanına göre
- 7 değişik sorgu tipine göre
- 49 değişik sorgu zamanı-sorgu tipi kombinasyonuna göre

3. bölüm olan araştırma sonuçlarında bu sınıflandırma detaylı bir şekilde incelenmiştir.

2.2.1.8. Kullanılan notasyon ve performans ölçütleri

Şartlı Olasılık yönteminde aşağıdaki notasyon ve performans ölçütleri kullanılmıştır:

$N_{değişim}$*:* Şartlı Olasılık yöntemi tarafından konu değişimi olarak tahmin edilen sorgu sayısı

N_{devam}*:* Şartlı Olasılık yöntemi tarafından konu devamı olarak tahmin edilen sorgu sayısı

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

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

$N_{değişim\&doğru}$*:* Hem uzman tarafından hem de Şartlı Olasılık yöntemi tarafından konu değişimi olarak işaretlenen sorgu sayısı

$N_{devam\&doğru}$: Hem uzman tarafından hem de Şartlı Olasılık yöntemi tarafından konu devamı olarak işaretlenen sorgu sayısı

A tipi hata: Şartlı Olasılık yöntemi tarafından konu değişimi olarak tahmin edilip gerçekte konu devamı olan sorgular

B tipi hata: Şartlı Olasılık yöntemi tarafından konu devamı olarak tahmin edilip gerçekte konu değişimi olan sorgular

$$N_{gerçek\ değişim} = N_{değişim\&doğru} + B\ tipi\ Hata$$

$$N_{gerçek\ devam} = N_{devam\ \&doğru} + A\ tipi\ Hata$$

$$N_{değişim} = N_{değişim\&\ doğru} + A\ tipi\ Hata$$

$$N_{devam} = N_{devam\ \&doğru} + B\ tipi\ Hata$$

Yukarıda tanımlanan değerleri kullanarak hesaplanan ve Şartlı Olasılık metodunun başarısını ölçecek olan performans ölçütleri: Duyarlılık (*precision-P*), Anma (*recall-R*) ve uygunluk fonksiyonu (*fitness function -*F_β)'dur. Bu performans ölçütleri hem bilişim teknolojisinde sıklıkla kullanılan hem de konu değişimleri ile ilgili çalışmalarda daha önceden kullanılan performans ölçütleri oldukları için bu çalışmada da kullanılmışlardır. Böylece bu çalışmanın performans ölçütlerinin, diğer anlam bazlı olmayan çalışmaların performans ölçütleri ile karşılaştırılması mümkün olacaktır.

$$P_{değişim} = N_{değişim\&\ doğru} / N_{değişim}$$

$$P_{devam} = N_{devam\&\ doğru} / N_{devam}$$

$$R_{değişim} = N_{değişim\&\ doğru} / N_{gerçek\ değişim}$$

$$R_{devam} = N_{devam\&\ doğru} / N_{gerçek\ devam}$$

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

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

Konu değişimlerini yorumlamak için kullanılan duyarlılık değeri $P_{değişim}$ Şartlı Olasılık yöntemi tarafından doğru bir şekilde konu değişimi olarak işaretlenen sorguların, yöntem tarafından konu değişimi olarak tahmin edilen bütün sorgu sayısına

oranıdır. Konu devamlarını yorumlayan duyarlılık değeri P_{devam}, Şartlı Olasılık yöntemi tarafından doğru bir şekilde konu devamı olarak işaretlenen sorguların, yöntem tarafından konu devamı olarak tahmin edilen sorgu sayısına oranıdır. Anma değeri $R_{değişim}$ ise Şartlı Olasılık yöntemi tarafından doğru bir şekilde belirlenen konu değişimlerinin, uzman tarafından belirlenen konu değişimlerine oranıdır. Anma değeri R_{devam} ise Şartlı Olasılık yöntemi tarafından doğru bir şekilde konu devamı olarak belirlenen 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 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 tutarlı karşılaştırmalar yapılabilecektir (Özmutlu ve ark. 2007).

2.2.2. Monte Carlo Simülasyonu uygulaması

Sorgular belli bir olasılık dağılımı ile ifade edilemedikleri için, düzgün dağılımı kullanarak basit bir şekilde yeni konu tanımlamasının gerçekleştirilmesi hedeflenmiştir. Bu nedenle de Monte Carlo Simülasyonu kullanılmıştır.

2.2.2.1. Araştırmada kullanılan veriler

Şartlı Olasılık yönteminde kullanılan; Excite 99, Excite 2001 ve FAST Veri Grubu bu çalışmada da kullanılmıştır. Veri grubu ile ilgili bilgiler '2.1.1.'numaralı bölümde detaylıca anlatılmıştır.

2.2.2.2. Notasyon ve performans ölçütleri

Şartlı Olasılık yönteminde kullanılan notasyon ve performans ölçütleri Monte Carlo Simülasyonu içinde aynen kullanılmıştır. Sadece yöntem farklı olup notasyon ve performans ölçütleri aynı olduğundan bu bölümde tekrar belirtilmeyecektir.

2.2.3. Monte Carlo Simülasyonu

Yukarıda bahsedilen 3 veri seti de, eğitim verisi ve test verisi olarak 2 gruba ayrılmıştır. Eğitim verisinde belirtilen zaman aralığı ve arama yapısındaki sorguların sayıları tespit edilmiş ve Şartlı Olasılık değerleri her bir zaman aralığı ve arama yapısındaki sorgu için hesaplanmıştır. Daha sonra test verisindeki sorgunun konu değişimi mi konu devamı mı olduğu Monte Carlo Simülasyonu ile tahmin edilmiştir. Yöntemin başarısı tanımlanan performans ölçütlerinin hesaplanmasıyla yorumlanmıştır. Yöntem 7 aşamadan oluşmaktadır ve bu aşamalar, 5 nolu aşama olan 'Monte Carlo Simülasyonu kullanarak, test veri grubundaki sorguların konu değişimi-konu devamının tahmin edilmesi' dışında Şartlı Olasılık yöntemi ile aynıdır. Bu nedenle tezin bu kısmında sadece 5 nolu aşamadan bahsedilecektir. [5]

[5] Diğer aşamalar belirtilmediği için 2.2.3.1 başlığı bu aşama için kullanılmıştır.

2.2.3.1. Monte Carlo Simülasyonu kullanarak, test veri grubundaki sorguların konu değişimi - konu devamının tahmin edilmesi

Eğitim setinden elde edilen olasılıklar kullanılarak, test setinde her bir sorgunun konu değişimi mi konu devamı mı olduğu tahmin edilir. Monte Carlo Simülasyonu MS Excel'de yapılmış olup konu devamı 1 ile sembolize edilirken, konu değişimi 2 ile sembolize edilmiştir. .

Ek-2'deki konu devamı için belirtilen Şartlı Olasılık Değeri, aşağıdaki algoritmada 'ŞOD' olarak gösterilmiştir.

Monte Carlo Simülasyonu Algoritması:

Adım 0: Eğitim seti kullanılarak ŞOD'nin belirlenmesi;

(Her bir zaman aralığı-arama yapısı sorgu sınıfı için konu devamının şartlı olasılığının belirlenmesi)

Adım 1: i=1 ve j=1 olarak ata.

(i= oturum sayısı iken j=sorgu sayısını temsil etmektedir.)

Adım 2: Test verisindeki j sorgusu için zaman aralığı ve arama yapısı değerlerini belirle.

Adım 3: j sorgusu için $U_{[0,1]}$'den R rassal sayısını üret.

Adım 4: Eğer $0 \leq R <$ ŞOD ise sorguyu konu devamı olarak işaretle. Diğer türlü ise sorguyu konu değişimi olarak işaretle.

Adım 5: j'den sonra oturumda herhangi bir sorgu yoksa
$i=i+1$,
j=1,
Eğer $i >$ veri setindeki toplam oturum sayısından büyükse
Algoritmayı sonlandır.
Değilse, $j=j+1$,
Adım 2'ye git.

3. ARAŞTIRMA SONUÇLARI

3.1. Uzman Kişinin Bulduğu Sonuçlarla, Şartlı Olasılıklar Kullanılarak Bulunan Sonuçların Karşılaştırılması

Şartlı Olasılık yöntemi ile test setindeki sorgular için konu devam - konu değişimi tahmini yapıldıktan sonra, uzman tarafından elde edilen gerçek sonuçlarla, yöntemin sonuçları karşılaştırılmıştır. Daha önce notasyon kısmında anlatılmış olan duyarlılık, anma ve uygunluk fonksiyonu performans ölçütleri hesaplanmıştır. Elde edilen sonuçlar Çizelge 3.1'de gösterilmiştir. Çizelge 3.1'de parantez içinde verilmiş olan değerler, gerçek konu değişimi ve konu devam sayılarını ifade etmektedir.

Çizelge 3.1: Çeşitli test verileri üzerinde Şartlı Olasılıklar kullanılarak yapılan konu devam – konu değişiklik tahmin 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}	$F_{\beta(değişim)}$	$F_{\beta(devam)}$
Excite 99 veri seti	3667	228 (152)	3439 (3515)	81	3368	147	71	0,355	0,533	0,979	0,958	0,449	0,966
Excite 2001 veri seti	3394	309 (272)	3085 (3122)	189	3002	120	83	0,612	0,695	0,973	0,961	0,661	0,966
FAST very seti	4484	276 (310)	4208 (4174)	146	4044	130	164	0,529	0,471	0,961	0,969	0,491	0,966

Konu değişimi tespiti için yapılan anlam bazlı olmayan çalışmaların başarısının tespitinde $F_{\beta(değişim)}$ performans ölçütü kritik parametre olup, arama motoru kullanıcı oturumlarında bu parametre değerlendirilmektedir. Çünkü 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.

Çizelge 3.1 incelendiğinde $F_{\beta\_değişim}$ performans ölçütüne göre Şartlı Olasılık yönteminin Excite 2001 verisinde en başarılı sonucu sağladığı gözlemlenmiştir. Diğer veri setleri için hesaplanan $F_{\beta\_değişim}$ değerleri de yeterli ve tutarlı sonuçlar sağlamıştır. P_{devam} ve R_{devam} performans ölçütleri dışında, tüm performans ölçütlerinde Excite 2001 veri setinde en iyi sonuçların elde edildiği görülmektedir. Excite 99 verisindeki $P_{değişim}$ performans ölçütünün düşük çıktığı, yöntemin aşırı konu değişimi tahmini yaptığı, doğru olarak tahmin ettiği konu değişimi sayısının az olduğu ve A tipi hataların arttığı belirlenmiştir. P_{devam} performans ölçütü için Excite 99 verisinin en başarılı sonucu sağladığı gözlemlenmiştir. Yöntemin doğru olarak tahmin ettiği konu devamı sayısının arttığı ve B tipi hataların azaldığı belirlenmiştir. R_{devam} performans ölçütünde ise FAST verisinde en başarılı sonuç elde edilmiştir. Yöntemin gerçek konu devamları içinden doğru olarak tahmin ettiği konu devamı sayısının arttığı ve A tipi hataların azaldığı belirlenmiştir.

3.2. Bulunan Sonuçlarda Hatalı Kısımların Tespit Edilmesi Ve Hataların Sınıflandırılması

Bu bölümde hatalar sorgu zamanlarına, sorgu arama yapılarına ve bu ikisinin kombinasyonuna göre sınıflandırılmıştır.

3.2.1. Sorgu zaman aralıklarına göre hataların sınıflandırılması

Sorgu zaman aralıklarına göre A tipi ve B tipi hataların dağılımı Çizelge 3.2'deki gibidir.

Çizelge incelendiğinde her üç veri setinde de A tipi hataların zaman aralığı 7 olan, 30 dk'dan daha fazla süren sorgularda yoğunlaştığı görülmektedir. Çizelge 3.2 ile Ek-3 birlikte incelendiğinde, 7-5 sınıfına ait olan sorguların konu değişimi olarak işaretlendiği görülmüştür. Bu durumda metodun ürettiği tahminlerde konu değişikliği olarak işaretlenen, fakat gerçekte uzmanın verdiği değerlerde konu devamı olan 7-5 sınıfındaki her sorgu A tipi hata olarak işaretlenmiştir. Excite 2001 verisinde A tipi hata, diğer iki veriye göre daha dağınık bir şekilde kümelenmiştir. 3, 5 ve 7 nolu zaman aralıklarında hatanın yoğunlaştığı gözlenmektedir. Arama yapısı ve zaman aralıklarının kombinasyonuna göre hataların sınıflandırıldığı Ek-3 incelendiğinde bu yoğunlaşmanın 'Yeni' arama yapısından kaynaklandığı söylenebilir.

Çizelge 3.2: Sorgu zamanına göre Excite 99, Excite 2001 ve FAST test verilerinde A tipi hataların ve B tipi hataların dağılımı

	Excite 99verisi		Excite 2001verisi		FAST verisi	
Zaman Aralığı	A tipi hata adedi	B tipi hata adedi	A tipi hata adedi	B tipi hata adedi	A tipi hata adedi	B tipi hata
1	0	35	0	32	0	78
2	0	14	0	31	0	24
3	0	11	25	0	0	25
4	0	6	0	14	0	12
5	1	1	20	1	0	11
6	0	3	9	2	0	11
7	146	1	66	3	130	3
Toplam	147	71	120	83	130	164

B tipi hatalar üç veri seti için incelendiğinde, A tipi hatalardaki gibi bir nokta üzerinde yoğunlaşma olmadığı, bunun yerine tüm sorgu zamanlarına bir dağılım olduğu gözlemlenmektedir. Yalnız sorgu süresi arttıkça B tipi hatalarda da bir azalma görülmektedir. Sorgu süresi ne kadar kısaysa, Şartlı Olasılık metodu sorguda konunun aynı olduğunu tahmin etmiş ve yanılmıştır. Şartlı Olasılık metodu, büyük ihtimalle eğitim setinin yapısından dolayı kısa süreli sorgulara konu devamı demekte, uzun süreli

sorgularda da konu değişmiştir diye karara varmakta ve yanılmaktadır. Bu aşamada göze çarpan unsur ise Excite 99 ve FAST verisinde sorgu süresi uzadıkça B tipi hataların azalmasıdır.

3.2.2. Sorgu arama yapısına göre hataların sınıflandırılması

Sorgu arama yapısına göre A tipi ve B tipi hataların dağılımı Çizelge 3.3'deki gibidir. Çizelge incelendiğinde ilginç bir durum gözlemlenmektedir. Her 3 veri setinde de hem A, hem de B tipi hatalar 5 nolu arama yapısında yani "yeni" tip arama yapısında ortaya çıkmaktadır. Sorguların arama yapılarının otomatik olarak PASCAL'da yazılan bir kod ile sınıflandırıldığı '2.2.1.3' nolu bölümde belirtilmişti. Sorguların arama yapılarına göre sınıflandırılmasında yapılan hatalar, yöntemin tahminlerini olumsuz yönde etkilemektedir. Yöntemin sonuçlarının iyileştirilmesi arama yapılarının daha doğru bir şekilde belirlenmesi ile sağlanabilir.

Çizelge 3.3: Arama yapısına göre Excite ve FAST test verilerinde A tipi ve B tipi hataların dağılımı

	Excite 99 verisi		Excite 2001verisi		FAST verisi	
Arama Yapısı	A tipi hata adedi	A tipi hata adedi	A tipi hata adedi	B tipi hata adedi	A tipi hata adedi	B tipi hata adedi
1	0	0	0	0	0	0
2	0	0	0	1	0	0
3	0	0	0	0	0	1
4	0	1	0	8	0	6
5	147	70	120	74	130	156
6	0	0	0	0	0	0
7	0	0	0	0	0	1
Toplam	147	71	120	83	130	164

3.2.3. Zaman aralığı ve arama yapısı kombinasyonuna göre hataların sınıflandırılması

Her 3 veri grubu içinde, 49 sorgu sınıfına göre A ve B tipi hataların sınıflandırılması Ek-3'de verilmiştir. Ek-3'den anlaşılacağı üzere A tipi hatalar Excite 99 ve FAST verisinde 7–5 sınıfındaki sorgularda ortaya çıkmaktadır. Excite 2001 verisinde ise 3-5, 5-5 ve 7-5 sınıflarındaki sorgularda yoğunlaşma olmaktadır. Üç veri setinde de hataların arama yapısına göre "yeni" arama yapısında yoğunlaştığı gözlenmektedir. Daha önce de belirtildiği gibi yöntemin sonuçlarının iyileştirilmesi arama yapılarının daha doğru bir şekilde belirlenmesi ile sağlanabilir.

Sorgu zamanına göre oluşan hataların, 7 nolu zaman aralığında yoğunlaştığı görülmüştür. Bunun çözümü 7 sorgu zamanı sınıfının 8, 10 veya daha fazla sorgu zamanı kategorisine çıkarılması olabilir.

B tipi hatalar her üç veri setinde de bir kaç sorgu sınıfında yoğunlaşmışlardır. Bu sorgu sınıfları; 1–5, 2–5, 3–5, 4–5, 5–5, 6–5 ve 7–4 olarak sıralanır. Bu hatalarda zamanın etkisinden ziyade, sorgu sınıfının "yeni" olması etkili görünmektedir.

3.3. Uzman Kişinin Bulduğu Sonuçlarla, Monte Carlo Simülasyonu Kullanılarak Bulunan Sonuçları Karşılaştırması

Monte Carlo Simülasyonu tarafından doğru ve yanlış şekilde işaretlenmiş olan konu devamı ve konu değişimi sonuçları belirlenmiş ve sonuçları değerlendirmek için kullanılan notasyon bölümündeki değerler hesaplanmıştır.

Monte Carlo Simülasyonu'nun uygulanması aşamasında belirtildiği üzere, simülasyonda $U_{[0,1]}$ düzgün dağılımına göre rassal sayı üretilmiştir. Monte Carlo Simülasyonu'nun her tekrarında farklı bir rassal sayı ile çalışıldığından yöntemin performansının daha sağlıklı bir şekilde test edilebilmesi amacıyla her bir veri seti için Monte Carlo Simülasyonu 10 defa tekrar edilmiştir. Monte Carlo Simülasyonu yönteminin performans değerlendirmesinde ise bu 10 tekrarın ortalaması kullanılmıştır. Ek-5, Ek-6 ve Ek-7'de Excite 99, Excite 2001 ve FAST verileri için her bir simülasyon

tekrarının sonuçları belirtilmiştir. Hata analizinin daha sağlıklı olması yapılabilmesi için, yapılan tekrarların ortalamasına en yakın simülasyon tekrarını gösteren satırlar seçilmiş (çizelgelerde koyu renkle işaretli satırlar) ve hata analizleri bu seçilen tekrarlar için yapılmıştır. Çizelge 3.4'de parantez içinde verilmiş olan değerler, gerçek konu değişimi ve konu devam sayılarını ifade etmektedir.

$F_{\beta\_değişim}$ performans ölçütü göz önüne alındığında Monte Carlo Simülasyonu yönteminin Excite 2001 veri seti için en başarılı sonucu verdiği görülmüştür. Excite 99 verisinde $P_{değişim}$ değerinin, diğer iki veriden düşük olduğu göze çarpmaktadır. Yöntem fazla konu değişimi tahmininde bulunmaktadır. Yöntemin doğru olarak tahmin ettiği sorguların sayısının, diğer veri setlerine göre az olduğu ve yöntem tarafından oluşan A tipi hataların fazla olduğu görülmektedir. Dolayısıyla yöntemin $F_{\beta\_değişim}$ değeri diğer veri setlerinden düşük çıkmaktadır. Şartlı Olasılık yönteminde de Excite 99 verisinde $P_{değişim}$ değerinin düşük olduğu göz önüne alınırsa, bu durumun eğitim setinin yapısından kaynaklandığı düşünülebilir. $R_{değişim}$ performans ölçütü göz önüne alındığında Monte Carlo Simülasyonu yönteminin Excite 2001 veri seti için en başarılı sonucu verdiği görülmüştür. Yöntemin doğru olarak tahmin ettiği konu değişimlerinin sayısı diğer veri setlerinden fazladır. P_{devam} performans ölçütü için Excite 99 verisinde en başarılı sonuç elde edilmiştir. Bu veri setinde doğru tahmin edilen konu devamı sayısının arttığı ve B tipi hatanın azaldığı görülmektedir. R_{devam} performans ölçütü için FAST verisinde en başarılı sonuçlar elde edilmiştir. Bu veri setinde doğru tahmin edilen konu devamı sayısının arttığı ve A tipi hatanın azaldığı gözlemlenmektedir.

Çizelge 3.4: Çeşitli test verileri üzerinde Monte Carlo Simülasyonu kullanılarak yapılan konu devam – konu değişiklik tahmin 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}	$F_{\beta(değişim)}$	$F_{\beta(devam)}$
Excite 99 veri seti	3667	283 (152)	3383 (3515)	67	3299	216	85	0,238	0,442	0,975	0,938	0,335	0,952
Excite 2001 veri seti	3394	393 (272)	3001 (3122)	142	2871	251	130	0,362	0,524	0,946	0,923	0,449	0,931
FAST very seti	4484	338 (310)	4146 (4174)	137	3973	201	173	0,405	0,44	0,958	0,952	0,426	0,954

3.4. Bulunan Sonuçlarda Hatalı Kısımların Tespit Edilmesi ve Hataların Sınıflandırılması

Bu bölümde hatalar sorgu zamanlarına, sorgu arama yapılarına ve bu ikisinin kombinasyonuna göre sınıflandırılmıştır.

3.4.1. Sorgu zamanına göre hataların sınıflandırılması

Çizelge 3.5 incelendiğinde A tipi hataların her 3 veri setinde dağınık bir yapılanma sergilediği görülmektedir. Bununla beraber, 1 ve 7 nolu zaman aralıklarında A tipi hataların yoğunlaştığı görülmektedir. Herhangi bir sorgu 5 dk'dan az ve 30 dk'dan fazla sürdüğünde Monte Carlo Simülasyonu tarafından konu değişimi olarak tahmin edilmiştir. Ayrıca her bir veri setindeki sorgularda, 7 nolu zaman aralığı hesaba katılmadığında, süre azaldıkça A tipi hataların arttığı gözlenmektedir.

B tipi hatalar her üç veri setinde tüm sorgu zamanlarına dağılmıştır. 7 nolu zaman aralığı dikkate alınmazsa, sorgu süresi azaldıkça B tipi hatalarda artma olduğu görülmektedir. Sorgu süresi ne kadar kısaysa, Monte Carlo Simülasyonu sorguda konunun aynı olduğunu tahmin etmiş ve yanılmıştır. 7 nolu zaman aralığındaki sorgularda, Monte Carlo Simülasyonu sorgunun konu devamı olacağını düşünmüştür. Gerçekte ise konu değişmiş ve B tipi hata oluşmuştur.

Çizelge 3.5: Zaman aralığına göre FAST ve Excite test verilerinde A ve B tipi hataların dağılımı

	Excite 1999 verisi		Excite 2001 verisi		FAST verisi	
Zaman Aralığı	A tipi hata adedi	B tipi hata adedi	A tipi hata adedi	B tipi hata adedi	A tipi hata adedi	B tipi hata adedi
1	76	29	140	22	68	55
2	24	12	32	22	26	18
3	11	9	16	5	15	19
4	3	5	11	10	12	6
5	1	1	9	4	9	7
6	5	2	4	5	8	7
7	90	29	44	61	70	59
Toplam	210	87	256	129	208	171

3.4.2. Sorgu arama yapısına göre hataların sınıflandırılması

Sorgu arama yapısına göre A tipi ve B tipi hataların dağılımı Çizelge 3.6'daki gibidir. Çizelge incelendiğinde her 3 veri seti için hem A, hem de B tipi hataların 5 nolu arama yapısı sınıfında yani "yeni" sorgu sınıfında ortaya çıktığı görülmüştür. Kullanıcı yeni bir arama yaptığında, aynı konuda kalmış fakat Monte Carlo Simülasyonu bunu yeni bir konu olarak tahmin etmişse, A tipi hata ortaya çıkmaktadır. Kullanıcı yeni bir arama yaptığında konu değiştirdiyse, Monte Carlo Simülasyonu da bu sorguyu konu devamı olarak tahmin ettiyse B tipi hata ortaya çıkmaktadır. Daha önce de belirtildiği gibi yöntemin sonuçlarının iyileştirilmesi arama yapılarının daha doğru bir şekilde belirlenmesi ile sağlanabilir. Ayrıca Excite 2001 ve FAST verisinde A tipi ve B tipi hataların 4 nolu sorgu sınıfı olan "özelleştirme" sınıfında da yoğunlaştığı görülmektedir.

Çizelge 3.6: Sorgu arama yapısına göre Excite 1999, Excite 2001 ve FAST test verilerinde A tipi ve B tipi hataların dağılımı

	Excite 1999 verisi		Excite 2001 verisi		FAST verisi	
Arama Yapısı	A tipi hata adedi	B tipi hata adedi	A tipi hata adedi	B tipi hata adedi	A tipi hata adedi	B tipi hata adedi
1	0	0	0	0	3	0
2	0	0	3	1	0	0
3	0	0	0	0	1	1
4	1	1	23	8	8	4
5	209	86	230	120	194	165
6	0	0	0	0	2	0
7	0	0	0	0	0	1
Toplam	210	87	256	129	208	171

3.4.3. Zaman aralığı ve arama yapısı kombinasyonuna göre hataların sınıflandırılması

Excite 99, Excite 2001 ve FAST verileri için 49 sorgu sınıfına göre A ve B tipi hataların sınıflandırılması Ek-4'deki gibidir.

Ek-4'den de anlaşıldığı üzere her üç veri setinde A tipi ve B tipi hatalar 7-5 tipi başta olmak üzere,1-5, 2-5, 3-5, 4-5, 5-5, 6-5 sorgu sınıflarında ortaya çıkmaktadır. Bu sorgu sınıflarında hataların yoğunlaşmasında zaman aralıklarının etkisinden çok, "yeni" arama yapısının etkili olduğu düşünülmektedir. Daha önce de belirtildiği gibi yöntemin sonuçlarının iyileştirilmesi arama yapılarının daha doğru bir şekilde belirlenmesi ile sağlanabilir. Zaman aralığına göre 7 nolu sorgu sınıfında hataların yoğunlaştığı düşünüldüğünde, 7 sorgu zamanı aralığının 8, 10 veya daha fazla zaman aralığına çıkarılması, yöntemin sonuçlarının iyileştirilmesine yardımcı olabilir.

3.4.4. Uygulanan yöntemin tipine göre karşılaştırmalı analiz sonuçları

Çizelge 3.7: Excite 99 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}	$F_{\beta(değişim)}$	$F_{\beta(devam)}$
Şartlı Olasılık	3667	228 (152)	3439 (3515)	81	3368	147	71	0,355	0,533	0,979	0,958	0,449	0,966
Monte Carlo Simülasyonu	3667	283	3383	67	3299	216	85	0,238	0,442	0,975	0,938	0,335	0,952
YSA Sonuçları (2005)	3667	399	3268	116	3232	283	36	0,291	0,76	0,989	0,919	0,508	0,944

Bugüne kadar uygulanan anlam bazlı olmayan yeni konu tanımlama çalışmaları yöntemlerin birbirleriyle karşılaştırılabilmesi amacıyla aynı parametre ve aynı performans ölçütleriyle ile tanımlanmıştır. Excite 99 veri seti de bu yöntemlerin uygulandığı veri setlerinden biridir. Belirtilen çalışmaların performans değerlendirmeleri incelendiğinde Yapay Sinir Ağları (YSA) yönteminin en başarılı yöntemlerden biri olduğu görülmüş ve bu nedenle çalışmada kullanılan Şartlı Olasılık ve Monte Carlo Simülasyonu metotları ile Yapay Sinir Ağları yönteminin karşılaştırılması hedeflenmiştir. Excite 99 verisine uygulanan bu yöntemlerin karşılaştırılması Çizelge 3.7'de belirtilmiştir. Çizelge 3.7'de parantez içinde verilmiş olan değerler, gerçek konu değişimi ve konu devam sayılarını ifade etmektedir.

$P_{değişim}$ değeri için Şartlı Olasılık yönteminin, bu 3 yöntem arasından en iyi sonucu verdiği gözlemlenmiştir. $R_{değişim}$ değeri için Yapay Sinir Ağları yöntemi en iyi sonucu vermiştir. Bunun sebebi ise yöntemin doğru olarak işaretlediği konu değişimi sayısının diğer iki yöntemden daha fazla olmasıdır.

P_{devam} değeri için Yapay Sinir Ağları yöntemi en iyi sonucu vermiştir. Yöntem tarafından doğru tahmin edilen konu devamı sayısı artarken, B tipi hataların azaldığı söylenebilir. R_{devam} değeri içinse Şartlı Olasılık yöntemi daha iyi sonuçlar vermiştir. Bu

yöntemin gerçek konu devamı sayısına en yakın sayıda konu devamı işaretlediği görülmüştür. $F_{\beta(devam)}$ ise yine Şartlı Olasılık yönteminin en iyi sonucu verdiği bir parametredir. Yöntemin belirlediği R_{devam} ve P_{devam} değerleri yüksek olup, bu iki parametrenin kombinasyonu sonucunda oluşan $F_{\beta(devam)}$ parametresi de yüksek bir değer almaktadır.

Monte Carlo Simülasyonu ve Şartlı Olasılık yöntemi karşılaştırıldığında Şartlı Olasılık yönteminin daha başarılı olduğu görülmektedir. Şartlı Olasılık yönteminde zaman aralığı ve arama yapısına göre belirlenen sorgu tipinin konu devamı ve konu değişimi şartlı olasılıklardan hangisi büyükse, o davranışı göstereceği düşünülüp, olasılık değerlerinde büyük olanına göre tahmin yapılmaktadır. Test veri seti, eğitim veri setinin davranışını direk taklit etmektedir. Monte Carlo Simülasyonu'nda ise çekilen rassal sayı konu devamı şartlı olasılığı ile karşılaştırılarak tahmin yapılmakta, eğitim veri setinin davranışı Şartlı Olasılık Yönteminde olduğu gibi birebir taklit edilmemektedir.

Aynı veri setine uygulanan Yapay Sinir Ağları yöntemi kritik parametre olan $F_{\beta(değişim)}$ parametresine göre bu 3 yöntemden en başarılısıdır. 2005 yılında uygulanan Yapay Sinir Ağları yönteminde β parametresi, Anma (R) değerine öncelik verilmesi amacıyla 1.5 olarak seçilmiştir. Yapay Sinir Ağları yönteminin bulduğu $R_{değişim}$ değeri ise bu 3 yöntem arasındaki en iyi değer olduğundan (yönteminin doğru olarak işaretlediği konu değişimi sorguların sayısının diğer iki yöntemden daha fazladır) $F_{\beta(değişim)}$ değeri en yüksek çıkmaktadır.

Yapay Sinir Ağları yönteminin doğru olarak işaretlediği konu değişimi sayısı diğer iki yöntemden daha fazladır. Yöntem Şartlı Olasılık ya da Monte Carlo Simülasyonu'ndaki gibi olasılıklara dayanmayıp, veri setinin davranışına göre eğitildiğinden, doğru olarak işaretlediği konu değişimi sayısı diğer iki yöntemden daha fazladır.

Çizelge 3.8: Excite 2001 verilerine uygulanan yöntemlerin analiz sonuçları

	Analiz Edilen Sorgu Sayısı	Konu Değişim Sayısı	Konu Devamı Sayısı	Doğru Tahmin Edilen Değişimler	Doğru Tahmin Edilen Devamlar	A Tipi Hata	B Tipi Hata	$P_{değişim}$	$R_{değişim}$	P_{devam}	R_{devam}	$F_{ß(değişim)}$	$F_{ß(devam)}$
Şartlı Olasılık	3394	309 (272)	3085 (3122)	189	3002	120	83	0,612	0,695	0,973	0,961	0,661	0,966
Monte Carlo Simülasyonu	3394	393	3001	142	2871	251	130	0,362	0,524	0,946	0,923	0,449	0,931
YSA Sonuçları (2008)	3394	445	2949	234	2911	211	38	0,525	0,86	0,987	0,932	0,695	0,952

Excite 2001 verisi içinde, Excite 99 verisinde açıklandığı üzere Şartlı Olasılık, Monte Carlo Simülasyonu ve Yapay Sinir Ağları yöntemlerinin karşılaştırılması hedeflenmiştir. Excite 2001 verisine uygulanan bu yöntemlerin karşılaştırılması Çizelge 3.8'de belirtilmiştir. Çizelge 3.8'de parantez içinde verilmiş olan değerler, gerçek konu değişimi ve konu devam sayılarını ifade etmektedir.

Monte Carlo Simülasyonu ve Şartlı Olasılık yöntemi karşılaştırıldığında Şartlı Olasılık yönteminin daha başarılı olduğu görülmektedir. Aynı veri setine uygulanan Yapay Sinir Ağları yöntemi ise bu 3 yöntemden en başarılısı olarak karşımıza çıkmaktadır.

$P_{değişim}$ değeri için Şartlı Olasılık yönteminin sonucunun 0,612'ye kadar yükseldiği görülmektedir. Bu sonuç, diğer yöntemlerle karşılaştırıldığında iyi bir ilerlemedir. Yöntemin doğru olarak tahmin ettiği sorguların arttığı ve yöntem tarafından oluşan A tipi hataların azaldığı söylenebilir. $R_{değişim}$ değeri için Yapay Sinir Ağları yöntemi en iyi sonucu vermiştir. Bunun sebebi de yönteminin doğru olarak işaretlediği konu değişimlerinin diğer iki yöntemden daha fazla olmasıdır.

P_{devam} değeri için Yapay Sinir Ağları yöntemi en iyi sonucu vermiştir. Yöntem tarafından doğru tahmin edilen konu devamı sayısı artarken, B tipi hataların azaldığı söylenebilir. R_{devam} değeri içinse Şartlı Olasılık yöntemi daha iyi sonuçlar vermiştir. Bu yöntemin gerçek konu devamı sayısına en yakın olacak şekilde konu devamlarını

işaretlediği görülmüştür. $F_{\beta(devam)}$ değeri de yine Şartlı Olasılık yönteminin en iyi sonucu verdiği parametrelerden biridir. R_{devam} ve P_{devam} değerlerinin yüksek olması sonucunda bu iki parametrenin kombinasyonuyla oluşan $F_{\beta(devam)}$ parametresi de yüksek çıkmaktadır.

Monte Carlo Simülasyonu ve Şartlı Olasılık yöntemi karşılaştırıldığında Şartlı Olasılık yönteminin daha başarılı sonuçlar verdiği görülmüştür. Bunun sebepleri Excite 99 verisinde açıklanmıştır.

Yapay Sinir Ağları yöntemi kritik parametre olan $F_{\beta(değişim)}$ parametresine göre bu 3 yöntemden en başarılısıdır. Bu veri setine uygulanan Yapay Sinir Ağları yöntemi 2008 yılında uygulanmıştır. Kullanılan β parametresi, 2006 yılında yapılan Şartlı Olasılık ve 2007 yılında yapılan Monte Carlo Simülasyonu yöntemleri ile uyumlu olmak için 1.3 olarak belirlenmiştir. Yapay Sinir Ağları yönteminin bulduğu $R_{değişim}$ değeri ise bu 3 yöntem arasındaki en iyi değer olduğundan (yönteminin doğru olarak işaretlediği konu değişimlerinin sayısı diğer iki yöntemden daha fazladır.) $F_{\beta(değişim)}$ değeri en yüksek çıkmaktadır.

Yapay Sinir Ağları yönteminin, Monte Carlo Simülasyonu ve Şartlı Olasılık yöntemlerinden daha iyi sonuç vermesinin sebebi Excite 99 veri setinde açıklanmıştır.

Çizelge 3.9: 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}	$F_{\beta(değişim)}$	$F_{\beta(devam)}$
Şartlı Olasılık	4484	276 (310)	4208 (4174)	146	4044	130	164	0,529	0,471	0,961	0,969	0,491	0,966
Monte Carlo Simülasyonu	4484	338	4146	137	3973	201	173	0,405	0,44	0,958	0,952	0,426	0,954
YSA Sonuçları (2008)	4484	887	3597	306	3593	581	4	0,35	0,987	0,998	0,86	0,583	0,84

FAST verisi içinde Excite 2001, Excite 99 verisinde açıklandığı gibi Şartlı Olasılık, Monte Carlo Simülasyonu ve Yapay Sinir Ağları yöntemlerinin karşılaştırılması hedeflenmiş, FAST verisine uygulanan yöntemlerin karşılaştırılması Çizelge 3.9'da belirtilmiştir. Çizelge 3.9'da parantez içinde verilmiş olan değerler, gerçek konu değişimi ve konu devam sayılarını ifade etmektedir.

FAST verisi için Şartlı Olasılık yöntemi, Monte Carlo Simülasyonu yönteminden daha başarılıdır. Yapay sinir ağı yönteminse bu 3 yöntemden en başarılısı olduğu görülmektedir. $P_{değişim}$ değeri için Şartlı Olasılık yönteminin sonucu 0,529'a kadar yükselmiştir. Bu da diğer yöntemlerle karşılaştırıldığında iyi bir ilerlemedir. Yöntemin doğru olarak tahmin ettiği sorguların arttığı ve yöntem tarafından oluşan A tipi hataların azaldığı söylenebilir. $R_{değişim}$ değeri için Yapay Sinir Ağları yöntemi en iyi sonucu vermiştir. Bunun sebebi de yönteminin doğru olarak işaretlediği konu değişimi sorgularının sayısının diğer iki yöntemden daha fazla olmasıdır.

P_{devam} değeri için Yapay Sinir Ağları yöntemi en iyi sonucu vermiştir. Yöntem tarafından doğru tahmin edilen konu devamı sayısı artarken, B tipi hataların azaldığı söylenebilir. R_{devam} değeri içinse Şartlı Olasılık yöntemi daha iyi sonuçlar vermiştir. Bu yöntemin gerçek konu devamı sayısına en yakın sayıda konu devamı işaretlediği görülmüştür. $F_{\beta(devam)}$ ise yine Şartlı Olasılık yönteminin en iyi sonucu verdiği bir parametredir. R_{devam} ve P_{devam} değerlerinin yüksek olması sonucunda bu iki parametrenin kombinasyonuyla oluşan $F_{\beta(devam)}$ parametresi de yüksek çıkmaktadır.

Çizelge 3.7, 3.8 ve 3.9 incelendiğinde üç veri setinin de başta $F_{ß(değişim)}$ performans ölçütü olmak üzere tüm performans ölçütleri için tutarlı sonuçlar sergilediği görülmüştür.

Uygulanan Şartlı Olasılık ve Monte Carlo Simülasyonu yöntemleri, Yapay Sinir Ağları yöntemi kadar iyi sonuçlar vermeseler de bu yönteme yakın sonuçlar vermiştirler.

4. TARTIŞMA VE SONUÇLAR

Çalışmada Şartlı Olasılık ve Monte Carlo Simülasyonu arama motoru kullanıcı oturumlarındaki konu değişikliklerini tespit ve tahmin etmek için kullanılmıştır. Şartlı Olasılık ve Monte Carlo Simülasyonu yöntemlerinin tüm performans ölçütlerinde verdikleri değerler tutarlıdır. Arama motoru kullanıcı oturumlarında konu değişimi tespiti için yapılan anlam bazlı olmayan çalışmalar içinde kullanılan diğer yöntemler olan Demspter- Shafer Teorisi ya da Yapay Sinir Ağları gibi yöntemler, uygulama açısından daha karışık ve zahmetli yöntemlerdir. Şartlı Olasılık ve Monte Carlo Simülasyonu yöntemleri ise uygulama açısından karmaşık yöntemler olmayıp, karmaşık yöntemlere yakın ve tutarlı sonuçlar sunmuşlardır. Bu nedenle bu iki yöntemin arama motoru kullanıcı oturumlarında konu değişimi tespiti için yapılan anlam bazlı olmayan çalışmalar içinde kullanılabileceği ve bu yöntemlerle yeterli ve tutarlı sonuçlara ulaşılabileceği kanıtlanmıştır.

Şartlı Olasılık ve Monte Carlo Simülasyonu yöntemlerinin uygulanması sonucu oluşan hataların sınıflandırılmasında karşımıza çıkan en önemli iki sonuç: her iki yöntemde A tipi hataların sorgu süresi 30 dakikayı geçtiğinde yoğunlaşması ve arama yapısına göre hatalar sınıflandırıldığında “yeni” arama yapısında A ve B tipi hatalarının yoğunlaşmasıdır.

Zaman aralığına göre oluşan hataların, sorgu süresi 30 dakikayı geçtiğinde yoğunlaşması, zaman aralıklarının 8, 10 gibi daha fazla sınıfa çıkarılması ile çözümlenebilir. Sorguların arama yapıları, PASCAL dilinde yazılan bir kod ile otomatik olarak sınıflandırılmıştır. Özellikle Monte Carlo Simülasyonu yönteminde “yeni” arama yapısına sahip sorgular, yöntemin aşırı konu değişimi tahmininde bulunmasına yol açmıştır. Yöntemin sonuçlarının iyileştirilmesi arama yapılarının daha doğru bir şekilde belirlenmesi ile ya da “yeni” arama yapısında sınıflandırılan sorguların içeriğinin ayrıntılı bir şekilde incelenmesi ile sağlanabilir.

KAYNAKLAR

AGICHTEIN, E., E. BRILL ve S. DUMAIS. 2006a. Improving Web Search Ranking by Incorporating User Behavior Information. Proceedings of the Twenty-Ninth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, p. 19–26.

AGICHTEIN, E., E. BRILL, S. DUMAIS ve R. RAGNO. 2006b. Learning User Interaction Models For Predicting Web Search Result Preferences. Proceedings of the Twenty-Ninth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, p.3-10.

BALAY, M., E. TİMUÇİN, E. ÇAĞLAR, A. ŞENTÜRK ve R. ÖZKILIÇ. 2006. Bilgi Teknolojileri II.Ekin Kitabevi, Bursa. p.49-81.

BEEFERMAN, D. ve A. BERGER. 2000. Agglomerative Clustering of A Search Engine Query Log. Proceedings of the Sixth Acm Sigkdd International Conference On Knowledge Discovery And Data Mining**,** Boston, Ma, USA, p**.**407–416.

BEITZEL, S.M., E.C. JENSEN, A. CHOWDHURY, D. GROSSMAN ve O. FRIEDER. 2004. Efficiency and Scaling: Hourly Analysis of a Very Large Topically Categorized Web Query Log. Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval, Sheffield, UK, p.321-328.

BERGER, A. ve J. LAFFETY. 1999. Information Retrieval as Statistical Translation. In Proceedings of SIGIR '99. p.222-229.

BRESLAU, L., D. ESTRIN, K. FALL, S. FLOYD, J. HEIDEMANN, A. HELMY, P. HUANG, S. MCCANNE, K. VARADHAN, X. YA ve Y. HAOBO. 2000. Advances in Network Simulation. Computer, 33(5):59–67.

BURKOWSKI, F.J. 1990. Retrieval Performance of a Distributed Database Utilizing a Parallel Process Document Server. Proceedings of the Second International Symposium on Databases in Parallel and Distributed Systems, ACM Press, New York, p.71–79.

CACHEDA, F., V. PLACHOURAS ve I. OUNIS. 2005. A Case Study of Distributed Information Retrieval Architectures to Index One Terabyte of Text. Information Processing and Management. 41:1141–1161.

CAHOON, B. ve K.S. MCKINLEY. 1996. Performance Evaluation of a Distributed Architecture for Information Retrieval. Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland, p.100–118.

CAHOON, B., K.S. MCKINLEY ve Z. LU. 2000. Evaluating the Performance of Distributed Architectures For Information Retrieval Using A Variety of Workloads. ACM Transactions on Information Systems, 18(1):1–43.

COEVREUR, T.R., R.N. BENZEL, S.F. MILLER, D.N. ZEITLER, D.L. LEE, M. SINGHAL, N. SHIVARATRI ve W.Y.P. WONG. 1994. An Analysis of Performance And Cost Factors In Searching Large Text Databases Using Parallel Search Systems. Journal of the American Society for Information Science 45(7):443–464.

COOLEY, R., B. MOBASHER ve SRIVASTAVA, J. 1999. Data Preparation for Mining World Wide Web Browsing Patterns. Knowledge and Information Systems, 1:5–32.

CRESTANI, F. ve C.J.V. RIJSBERGEN. 1998. A Study of Probability Kinematics in Information Retrieval. ACM Transactions on Information Systems, 16(3):225–255.

ÇAVDUR, F. 2005. Arama Motorları Kullanıcı Oturumlarındaki Konu Değişiklerinin Tespit ve Tahmin Yöntemleri. Uludağ Üniversitesi, Yüksek Lisans Tezi.s.24-29.

DESHPANDE, Y.L., R. JENKINS ve S. TAYLOR. 1996. Use of Simulation To Test Client–Server Models, In Proceedings of the 1996 Winter Simulation Conference, p.1210–121.

FLOYD, S. ve V. PAXSON. 2001. Difficulties In Simulating the Internet, IEEE/ACM Transactions on Networking, 9(4):392–403.

GREMETT, P. 2006. Utilizing a User's Context to Improve Search Results. Journal of the American Society for Information Science and Technology, 57(6):808-812.

HE, D. ve A. GOKER. 2000. Analysing Intranet Logs to Determine Session Boundaries for User-Oriented Learning. Proceedings of AH2000: The International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, Trento, Italy, p.319–322.

HE, D., A. GOKER ve D.J. HARPER. 2002. Combining Evidence for Automatic Web Session Identification. Information Processing and Management, 38:727-742.

HILTY, L.M., P. ARNFALK, L. ERDMANN, J. GOODMAN, M. LEHMANN, ve P.A. WAGER. 2006. The Relevance of Information and Communication Technologies for Environmental Sustainability - A Prospective Simulation Study, Environmental Modeling and Software, 21(11):1618–1629.

HU, X., S. BANDHAKAVI ve C. ZHAI. 2003. Error Analysis of Difficult TREC Topics. Proceedings of 26th Acm International Conference on Research and Development in Information Retrieval (SIGIR'03), Toronto, Canada, p.407–408.

HUANG, X., F. PENG, A. AN ve D. SCHUURMANS. 2004. Dynamic Web Log Session Identification with Statistical Language Models. Journal of The American Society for Information Science and Technology, 55(14):1290-1303.

JOACHIMS, T. 1998. Text Categorization with Support Vector Machines. In Proceedings of the 10th European Conference on Machine Learning (ECML'98), Chemnitz, Germany, p.137–142.

JANSEN, B.J., A. SPINK ve T. SARACEVIC. 2000. Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web. Information Processing and Management, 36:207–227.

JIN, R., G. HAUPTMANN ve C. ZHAI. 2002. Title Language Model for Information Retrieval, *SIGIR '02*, Tempere**,** Finland, p.42-48.

JIN, R., L. SI ve C. ZHAI. 2003. Preference-Based Graphic Models for Collaborative Filtering. Proceedings of the 19th Conference in Uncertainty in Artificial Intelligence, Acapulco, Mexico, p.329-336.

LAWRIE, D., W.B. CROFT, ve A. ROSENBERG. 2001. Finding Topic Words for Hierarchical Summarization. Proceedings Of 24th ACM International Conference On Research And Development In Information Retrieval (SIGIR '01), p.349–357.

KELLY, D., F. DIAZ, N.J. BELKIN ve J. ALLAN. 2004. A User-Centered Approach to Evaluating Topic Models. Lecture Notes in Computer Science, 2997:27–41.

LARKEY, L.S., F. FENG, M. CONNELL ve V. LAVRENKO. 2004. Language-Specific Models in Multilingual Topic Tracking. Proceedings of 27th ACM International Conference on Research and Development in Information Retrieval (SIGIR '04), Sheffield, South Yorkshire, UK, p.402–409.

LAW, A.M. ve W.D. KELTON. 1991. Simulation Modeling and Analysis. Mcgraw-Hill, Newyork. p.113-114

LIAW, S.S. ve H.M. HUANG. 2006. Information Retrieval from the World Wide Web: A User-Focused Approach Based on Individual Experience with Search Engines. Computers in Human Behavior, 22(3):501–517.

LI, W. ve A. MCCALLUM. 2005. Semi-Supervised Sequence Modeling with Syntactic Topic Models, Proceedings of the 12th Conference on Artificial Intelligence, p.813-818.

LIU, X., W.B. CROFT, P. OH ve D. HART. 2004. Automatic Recognition of Reading Levels From User Queries. In: Proceedings of The 27th ACM International Conference on Research and Development in Information Retrieval (SIGIR '04). p.548–549.

LU, T. ve C. HSU. 2007. Mobile Agents for Information Retrieval in Hybrid Simulation Environment. Journal of Network and Computer Applications 30(1):244–264.

MEI, Q. ve C. ZHAI. 2005. Discovering Evolutionary Theme Patterns from Text an Exploration of Temporal Text Mining. The 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, Illinois, USA, p.198–207.

METZLER, D. ve W.B. CROFT. 2005. Analysis of Statistical Question Classification for Fact-Based Questions. Information Retrieval, 8:481–504.

METZLER, D. ve W.B. CROFT. 2005a. A Markov Random Field Model for Term Dependencies. Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05), p.472–479.

METZLER, D. ve W.B. CROFT. 2005b. A Markov Random Field Model for Term Dependencies. Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05), p.472–479.

MURESAN, G. ve D.J. HARPER. 2004. Topic Modeling for Mediated Access to Very Large Document Collections. Journal of the American Society for Information Science and Technology, 55:892–910.

ÖZMUTLU, H.C., A. SPINK ve S. ÖZMUTLU. 2002. Analysis of Large Data Logs: An Application of Poisson Sampling on Excite Web Queries. Information Processing and Management, 38:473–490.

ÖZMUTLU, S., H.C. ÖZMUTLU ve A. SPINK. 2003a. Multitasking Web Searching and Implications for Design. ASIST 2003, Annual Meeting of the American Society for Information Science and Technology, Long Beach, CA, p. 416–421.

ÖZMUTLU, S., H.C. ÖZMUTLU ve A. SPINK, 2003b. Are People Asking Questions of General Web Search Engines, Online Information Review. 27:396–406.

ÖZMUTLU, S., A. SPINK ve H.C. ÖZMUTLU. 2003c. Trends in Multimedia Web Searching: 1997–2001, Information Processing and Management, 39:611–621.

ÖZMUTLU, H.C., F. CAVDUR, S. ÖZMUTLU ve A. SPINK, 2004a. Neural Network Applications for Automatic New Topic Identification on Excite Web Search Engine Datalogs. Proceedings of ASIST, Annual Meeting of the American Society for Information Science and Technology, Providence, RI, p.310–316.

ÖZMUTLU, S., H.C. ÖZMUTLU ve A. SPINK. 2004b. A Day in the Life of Web Searching: An Exploratory Study. Information Processing and Management, 40:319–345.

ÖZMUTLU, H.C. ve F. CAVDUR. 2005a. Application of Automatic Topic İdentification on Excite Web Search Engine Data Logs. Information Processing and Management, 41:1243–1262.

ÖZMUTLU, S. ve F. CAVDUR. 2005b. Neural Network Applications for Automatic New Topic Identification. Online Information Review, 29:35-53.

ÖZMUTLU, H.C., F. CAVDUR ve S. ÖZMUTLU. 2006. Automatic New Topic Identification in Search Engine Data Logs. Internet Research: Electronic Networking Applications and Policy, 16:323–338.

ÖZMUTLU, S., H.C. ÖZMUTLU ve A. SPINK. 2006. Topic Estimation of Web Search Transaction Log Queries Using Monte-Carlo Simulation. The 12th Australasian World Wide Web Conference, Australia. p.118-122

ÖZMUTLU, S. 2006. Automatic New Topic Identification Using Multiple Linear Regression. Information Processing and Management, 42(4):934-950.

ÖZMUTLU, H.C., F. CAVDUR, A. SPINK ve S. ÖZMUTLU. 2006b. Investigating the Performance of Automatic New Topic Identification Across Multiple Datasets. Proceedings of ASIST, Annual Meeting of the American Society for Information Science and Technology, Providence, RI. 43(1):1-27

ÖZMUTLU, S., H.C. ÖZMUTLU ve B. BUYUK. 2007. Using Conditional Probabilities for Automatic New Topic Identification. Online Information Review, 31(4):491-515.

ÖZMUTLU, S., H.C. ÖZMUTLU ve A. SPINK. 2007a. Using Support Vector Machines for Automatic New Topic Identification. Proceedings of ASIST, 70th Annual Meeting of the American Society for Information Science and Technology, Madison, WI, p.1-5

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

ÖZMUTLU, S., H.C. ÖZMUTLU ve A. SPINK. 2008b. Automatic New Topic Identification in Search Engine Transaction Logs Using Multiple Linear Regression. Proceedings of the 41st Hawai International Conference on System Science, p.1530-1605.

ÖZMUTLU, S. ve G. COŞAR. 2008. Analyzing the Results of Automatic New Topic Identification. Library Hi Tech, 26:466-487.

PEGDEN, C.D., R.E. SHANNON ve R.P. SADOWSKI. 1995. Introduction to Simulation Using Siman, McGraw-Hill, New York.

ROBERT, C.P. 2004. Monte Carlo Statistical Methods. Springer, p.35-39.

ROSS, S.M. 2000. Introduction to Probability and Statistics for Engineers and Scientists. Academic Press, p.71-73.

SİLVERSTEIN, C., M. HENZINGER, H. MARAIS ve M. MORICZ. 1999. Analysis of a Very Large Web Search Engine Query Log Forum, 33:6–12.

SHEN, X., B. TAN ve C. ZHAI. 2005. Context Sensitive Information Retrieval Using Implicit Feedback. Proceedings of the 28th ACM ACM International Conference on Research and Development in Information Retrieval (SIGIR '05), Salvador, Brazil, p.43–50.

SPINK, A., J. BATEMAN ve B.J. JANSEN. 1999. Searching Heterogeneous Collections on the Web: A survey of Excite Users. Internet Research: Electronic Networking. 4(2):1-24.

SPINK, A., B.J. JANSEN, D. WOLFRAM ve T. SARACEVİC. 2002a. From e-Sex to e Commerce: Web Search Changes, IEEE Computer, 35:133–135.

SPINK, A., H.C. ÖZMUTLU ve S. ÖZMUTLU. 2002. Multitasking İnformation Seeking and Searching Processes. Journal of the American Society for Information Science and Technology, 53:639–652.

SPINK, A., M. PARK, B. JANSEN ve J. PEDERSEN. 2006 .Multitasking During Web Search Sessions. Information Processing and Management, 42:264–275.

SWAN, R. ve D. JENSEN. 2000. "TimeMines: Constructing Timelines with Statistical Models of Word Usage", Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, MA, p.73–80.

WANG, X., N. MOHANTY ve A. MCCALLUM. 2005. Group and Topic Discovery from Relations and Text. The 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Workshop on Link Discovery: Issues, Approaches and Applications (LinkKDD–05), Chicago, Illinois, USA, p.28–35.

WEN, J.R., J.Y. NIE ve H.J. ZHANG. 2002. Query Clustering Using User Logs. ACM Transactions on Information Systems, 20:59–81.

YAO, Z. ve F. LAI. 2008. Integrating Item Category Information in Collaborative Filtering Recommender Algorithm. Proceedings of the 2008 Fourth International Conference on Natural Computation, 7: p.33-38.

ZHAI, C., W.W. COEHN ve J. LAFFERTY. 2003. Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval. SIGIR'03, Toronto, Canada. p.10-17.

EKLER

EK 1: Excite 99, Excite 2001 ve FAST eğitim verisindeki konu değişim ve devamlarının sorgu zamanına ve arama yapısına göre dağılımı

		Sorgu adetleri					
		Excite 99 verisi		Excite 2001 verisi		FAST verisi	
Zaman Aralığı	Arama Çeşidi	Konu devamı sayısı	Konu değişimi sayısı	Konu devamı sayısı	Konu değişimi sayısı	Konu devamı sayısı	Konu değişimi sayısı
1	1	2120	0	1358	0	2822	4
1	2	54	0	80	0	31	0
1	3	148	0	160	0	114	2
1	4	276	1	306	9	192	2
1	5	403	76	361	128	244	86
1	6	0	0	0	0	61	1
1	7	0	0	0	0	0	0
2	1	133	0	110	0	168	0
2	2	0	0	3	1	3	0
2	3	10	0	12	0	10	0
2	4	21	0	45	7	38	0
2	5	54	18	59	30	63	27
2	6	0	0	0	0	2	0
2	7	0	0	0	0	1	0
3	1	46	0	56	0	50	0
3	2	1	0	1	0	2	0
3	3	4	0	5	0	6	0
3	4	5	0	23	0	11	0
3	5	29	14	24	28	41	24
3	6	0	0	0	0	2	0
3	7	0	0	0	0	0	0
4	1	20	0	16	0	18	1
4	2	0	0	2	0	2	0
4	3	1	0	4	0	3	0
4	4	6	0	7	0	9	0
4	5	20	7	13	8	23	18
4	6	0	0	0	0	0	0
4	7	0	0	0	0	1	0
5	1	5	0	15	0	10	0

5	2	0	0	1	0	0	0
5	3	1	0	2	0	1	0
5	4	2	0	6	1	5	0
5	5	14	13	10	12	17	17
5	6	0	0	0	0	0	0
5	7	0	0	0	0	0	0
6	1	6	0	12	0	4	0
6	2	1	0	0	0	0	0
6	3	0	0	0	0	0	0
6	4	2	0	5	1	3	0
6	5	11	5	8	8	17	10
6	6	0	0	0	0	0	0
6	7	0	0	0	0	0	0
7	1	41	0	60	0	28	0
7	2	2	0	1	0	1	0
7	3	2	0	3	0	2	0
7	4	15	0	20	3	18	3
7	5	91	135	91	155	146	188
7	6	0	0	0	0	5	1
7	7	0	0	0	0	0	2
Toplam		3544	269	2879	391	4174	386

EK 2: Excite 99, Excite 2001 ve FAST eğitim verisindeki konu değişim ve devamlarının sorgu zamanına ve arama yapısına göre Şartlı Olasılıkları

		Sorgu adetleri					
		Excite 99 verisi		Excite 2001 verisi		FAST verisi	
Zaman Aralığı	Arama Çeşidi	Konu devamı olasılığı	Konu değişimi olasılığı	Konu devamı olasılığı	Konu değişimi sayısı	Konu devamı olasılığı	Konu değişimi olasılığı
1	1	1	0	1	0	0,999	0,001
1	2	1	0	1	0	1	0
1	3	1	0	1	0	0,983	0,017
1	4	0,9964	0,0036	0.971	0.029	0,99	0,01
1	5	0,841	0,159	0.738	0.262	0,74	0,26
1	6	1	0	1	0	0,984	0,016
1	7	1	0	1	0	1	0
2	1	1	0	1	0	1	0
2	2	1	0	0.75	0.25	1	0
2	3	1	0	1	0	1	0
2	4	1	0	0.865	0.135	1	0
2	5	0,75	0,25	0.663	0.337	0,7	0,3
2	6	1	0	1	0	1	0
2	7	1	0	1	0	1	0
3	1	1	0	1	0	1	0
3	2	1	0	1	0	1	0
3	3	1	0	1	0	1	0
3	4	1	0	1	0	1	0
3	5	0,674	0,326	0.462	0.538	0,63	0,37
3	6	1	0	1	0	1	0
3	7	1	0	1	0	1	0
4	1	1	0	1	0	0,95	0,05
4	2	1	0	1	0	1	0
4	3	1	0	1	0	1	0
4	4	1	0	1	0	1	0
4	5	0,74	0,26	0.6190	0.381	0,56	0,44
4	6	1	0	1	0	1	0
4	7	1	0	1	0	1	0

EK 2: Excite 99, Excite 2001 ve FAST eğitim verisindeki konu değişim ve devamlarının sorgu zamanına ve arama yapısına göre Şartlı Olasılıkları (devam)

		Sorgu adetleri					
		Excite 99 verisi		Excite 2001 verisi		FAST verisi	
Zaman Aralığı	Arama Çeşidi	Konu devamı olasılığı	Konu değişimi olasılığı	Konu devamı olasılığı	Konu değişimi sayısı	Konu devamı olasılığı	Konu değişimi olasılığı
5	1	1	0	1	0	1	0
5	2	1	0	1	0	1	0
5	3	1	0	1	0	1	0
5	4	1	0	0.857	0.143	1	0
5	5	0,519	0,481	0.455	0.545	0,5	0,5
5	6	1	0	1	0	1	0
5	7	1	0	1	0	1	0
6	1	1	0	1	0	1	0
6	2	1	0	1	0	1	0
6	3	1	0	1	0	1	0
6	4	1	0	0.8333	0.1667	1	0
6	5	0,6875	0,3125	0.5	0.5	0,63	0,37
6	6	1	0	1	0	1	0
6	7	1	0	1	0	1	0
7	1	1	0	1	0	1	0
7	2	1	0	1	0	1	0
7	3	1	0	1	0	1	0
7	4	1	0	0.8696	0.1304	0,857	0,143
7	5	0,397	0,603	0.3699	0.6301	0,437	0,563
7	6	1	0	1	0	0,833	0,167
7	7	1	0	1	0	0	1

Ek-1'de Excite 99 verisinin eğitim setinde için 2-2 nolu sorgu sınıfında hiç sorgu olmadığına dikkat edilmelidir. Eğitim setinde belli bir zaman aralığı- arama yapısı kombinasyonu için sorgu gözlenmemiş olması, bu sınıfa ait sorguların test verisinde de gözlenmeyeceği anlamına gelmez. Önerilen yöntemin konu değişimi olup olmadığını tahmin etmesi beklendiğinden ve güvenli tahmin yapabilmek amaçlandığından konu devamı şartlı olasılığının bu durumlarda geçerli olduğu kabul edilmiştir.

EK 3: Şartlı Olasılık Yöntemi Sonucu Oluşan A tipi ve B tipi hataların arama yapısı ve zaman aralığı çeşidine göre Excite 99, Excite 2001 ve FAST test verilerinde dağılımı

Zaman Aralığı	Arama Yapısı	Excite 99 verisi		Excite 2001 verisi		FAST veri seti	
		A tipi hata adedi	B tipi hata adedi	A tipi hata adedi	B tipi hata adedi	A tipi hata adedi	B tipi hata adedi
1	1	0	0	0	0	0	0
1	2	0	0	0	1	0	0
1	3	0	0	0	0	0	0
1	4	0	0	0	2	0	1
1	5	0	35	0	29	0	76
1	6	0	0	0	0	0	0
1	7	0	0	0	0	0	1
2	1	0	0	0	0	0	0
2	2	0	0	0	0	0	0
2	3	0	0	0	0	0	0
2	4	0	0	0	0	0	1
2	5	0	14	0	31	0	23
2	6	0	0	0	0	0	0
2	7	0	0	0	0	0	0
3	1	0	0	0	0	0	0
3	2	0	0	0	0	0	0
3	3	0	0	0	0	0	1
3	4	0	0	0	0	0	1
3	5	0	11	25	0	0	23
3	6	0	0	0	0	0	0
3	7	0	0	0	0	0	0
4	1	0	0	0	0	0	0
4	2	0	0	0	0	0	0
4	3	0	0	0	0	0	0
4	4	0	0	0	0	0	0
4	5	0	6	0	14	0	12
4	6	0	0	0	0	0	0
4	7	0	0	0	0	0	0
5	1	0	0	0	0	0	0
5	2	0	0	0	0	0	0

EK 3: Şartlı Olasılık Yöntemi Sonucu Oluşan A tipi ve B tipi hataların arama yapısı ve zaman aralığı çeşidine göre Excite 99, Excite 2001 ve FAST test verilerinde dağılımı (devamı)

Zaman Aralığı	Arama Yapısı	Excite 99 verisi		Excite 2001 verisi		FAST veri seti	
		A tipi hata adedi	B tipi hata adedi	A tipi hata adedi	B tipi hata adedi	A tipi hata adedi	B tipi hata adedi
5	3	0	0	0	0	0	0
5	4	0	0	0	1	0	0
5	5	1	1	20	0	0	11
5	6	0	0	0	0	0	0
5	7	0	0	0	0	0	0
6	1	0	0	0	0	0	0
6	2	0	0	0	0	0	0
6	3	0	0	0	0	0	0
6	4	0	0	0	2	0	0
6	5	0	3	9	0	0	11
6	6	0	0	0	0	0	0
6	7	0	0	0	0	0	0
7	1	0	0	0	0	0	0
7	2	0	0	0	0	0	0
7	3	0	0	0	0	0	0
7	4	0	1	0	3	0	3
7	5	146	0	66	0	130	0
7	6	0	0	0	0	0	0
7	7	0	0	0	0	0	0
Toplam		147	71	120	83	130	164

EK 4: Monte Carlo Simülasyonu Sonucu Oluşan A tipi ve B tipi hataların arama yapısı ve zaman aralığı çeşidine göre Excite 99, Excite 2001 ve FAST test verilerinde dağılımı

Zaman Aralığı	Arama yapısı	Excite 1999 verisi		Excite 2001 verisi		FAST verisi	
		A tipi hata adedi	B tipi hata adedi	A tipi hata adedi	B tipi hata adedi	A tipi hata adedi	B tipi hata adedi
1	1	0	0	0	0	3	0
1	2	0	0	0	1	0	0
1	3	0	0	0	0	1	0
1	4	1	0	6	2	3	0
1	5	75	29	134	19	61	54
1	6	0	0	0	0	0	0
1	7	0	0	0	0	0	1
2	1	0	0	0	0	0	0
2	2	0	0	3	0	0	0
2	3	0	0	0	0	0	0
2	4	0	0	9	0	0	1
2	5	24	12	20	22	26	17
2	6	0	0	0	0	0	0
2	7	0	0	0	0	0	0
3	1	0	0	0	0	0	0
3	2	0	0	0	0	0	0
3	3	0	0	0	0	0	1
3	4	0	0	0	0	0	1
3	5	11	9	16	5	15	17
3	6	0	0	0	0	0	0
3	7	0	0	0	0	0	0
4	1	0	0	0	0	0	0
4	2	0	0	0	0	0	0
4	3	0	0	0	0	0	0
4	4	0	0	0	0	0	0
4	5	3	5	11	10	12	6
4	6	0	0	0	0	0	0
4	7	0	0	0	0	0	0
5	1	0	0	0	0	0	0
5	2	0	0	0	0	0	0

EK 4: Monte Carlo Simülasyonu Sonucu Oluşan A tipi ve B tipi hataların arama yapısı ve zaman aralığı çeşidine göre Excite 99, Excite 2001 ve FAST test verilerinde dağılımı (devamı)

Zaman Aralığı	Arama yapısı	Excite 1999 verisi		Excite 2001 verisi		FAST verisi	
		A tipi hata adedi	B tipi hata adedi	A tipi hata adedi	B tipi hata adedi	A tipi hata adedi	B tipi hata adedi
5	3	0	0	0	0	0	0
5	4	0	0	0	1	0	0
5	5	1	1	9	3	9	7
5	6	0	0	0	0	0	0
5	7	0	0	0	0	0	0
6	1	0	0	0	0	0	0
6	2	0	0	0	0	0	0
6	3	0	0	0	0	0	0
6	4	0	0	0	2	0	0
6	5	5	2	4	3	8	7
6	6	0	0	0	0	0	0
6	7	0	0	0	0	0	0
7	1	0	0	0	0	0	0
7	2	0	0	0	0	0	0
7	3	0	0	0	0	0	0
7	4	0	1	8	3	5	2
7	5	90	28	36	58	63	57
7	6	0	0	0	0	2	0
7	7	0	0	0	0	0	0
TOPLAM		210	87	256	129	208	171

EK 5: Excite 1999 test verisindeki her bir Monte Carlo Simülasyonu tekrarının sonuçları

Tekrar 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}	$F_{\beta(değişim)}$	$F_{\beta(devam)}$
1	297	3370	71	3289	226	81	0,239	0,4671	0,9759	0,9357	0,3448	0,9502
2	267	3397	80	3328	187	72	0,2996	0,5263	0,9796	0,9467	0,4107	0,9587
3	275	3389	65	3305	210	87	0,2363	0,4276	0,9752	0,9402	0,3287	0,9529
4	283	3384	73	3305	210	79	0,2579	0,4802	0,9766	0,9402	0,3637	0,9534
5	279	3385	64	3300	215	88	0,2293	0,421	0,9748	0,9388	0,3212	0,9519
6	272	3395	62	3305	210	90	0,2279	0,407	0,9734	0,9402	0,3153	0,9523
7	287	3380	70	3298	217	82	0,2439	0,4605	0,9757	0,9382	0,3462	0,9518
8	280	3384	55	3290	225	97	0,1964	0,3618	0,9722	0,9359	0,2755	0,9491
9	280	3387	63	3298	217	89	0,225	0,4144	0,9737	0,9382	0,3156	0,9511
10	308	3359	69	3276	239	83	0,224	0,4539	0,9752	0,9320	0,3285	0,9476

EK 6: Excite 2001 test verisindeki her bir Monte Carlo Simülasyonu tekrarının sonuçları

Tekrar 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}	$F_{\beta(değişim)}$	$F_{\beta(devam)}$
1	386	3008	148	2884	238	124	0,3834	0,5441	0,9588	0,9238	0,4708	0,9365
2	400	2994	145	2867	255	127	0,3625	0,533	0,9576	0,9183	0,4537	0,9325
3	397	*2997*	138	2863	259	134	0,3476	0,5074	0,8435	0,917	0,4333	0,8882
4	410	*2984*	155	2867	255	117	0,3780	0,5698	0,9608	0,9183	0,4794	0,9336
5	399	*2995*	143	2866	256	129	0,3584	0,5257	0,9569	0,918	0,4479	0,932
6	378	*3016*	140	2884	238	132	0,3704	0,5147	0,9562	0,9238	0,4496	0,9355
7	386	3008	141	2877	245	131	0,3653	0,5184	0,9564	0,9215	0,4485	0,9342
8	399	2995	139	2862	260	133	0,3484	0,511	0,9556	0,9167	0,4354	0,9308
9	385	3009	138	2875	247	134	0,3584	0,5074	0,9555	0,9555	0,4394	0,9554
10	396	2998	137	2863	259	135	0,346	0,5037	0,955	0,917	0,4307	0,9308

EK 7: FAST test verisindeki her bir Monte Carlo Simülasyonu tekrarının sonuçları

Tekrar 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}	$F_{\beta(değişim)}$	$F_{\beta(devam)}$
1	354	4130	136	3956	218	174	0,3842	0,4387	0,9579	0,9478	0,4167	0,9515
2	317	4167	143	4000	174	167	0,4511	0,4613	0,9599	0,9583	0,4574	0,9589
3	323	4161	133	3984	190	177	0,4118	0,4290	0,9575	0,9545	0,4224	0,9555
4	326	4158	140	3988	186	170	0,4294	0,4516	0,9591	0,9554	0,4431	0,9568
5	333	4151	133	3974	200	177	0,3994	0,4290	0,9574	0,9521	0,4175	0,9540
6	343	4141	139	3970	204	171	0,4052	0,4484	0,9587	0,9511	0,4313	0,9539
7	376	4108	144	3942	232	166	0,383	0,4645	0,9596	0,9444	0,4304	0,9500
8	347	4137	139	3966	208	171	0,4006	0,4484	0,9587	0,9502	0,4293	0,9533
9	318	4166	124	3980	194	186	0,3899	0,4	0,9554	0,9535	0,3962	0,9542
10	343	4141	136	3967	207	174	0,3965	0,4387	0,958	0,9504	0,4220	0,9532

ÖZGEÇMİŞ

Buket Büyük 29.07.1980 tarihinde Bursa'da doğmuştur. İlköğrenimini Gürsu İlköğretim Okulu'nda, orta öğretimini Bursa Kız Lisesi'nde ve lise öğrenimini Bursa Kız Lisesi'nin Süper Lise kısmında tamamlamıştır. 2002 yılında Ege Üniversitesi Tekstil Mühendisliğinden mezun olmuş, 2005–2006 öğretim yılında 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

Bu çalışma ve yüksek lisans eğitimim süresince bana yardımcı olan bölüm başkanım sayın Prof.Dr. Erdal EMEL'e, yüksek lisans tezini birlikte yaptığım danışmanım sayın Doç.Dr. Seda ÖZMUTLU'ya , sayın Doç.Dr. Cenk ÖZMUTLU'ya ve ayrıca sayın Yard.Doç.Dr.Mehmet AKANSEL'e teşekkürü bir borç bilirim. Eğitimim boyunca bana her zaman destek olan aileme teşekkür ederim.