EGE ÜNİVERSİTESİ

E. Ü. FEN BİLİMLERİ ENSTİTÜSÜ

DOKTORA TEZİ

BİLGİ ERİŞİM SİSTEMLERİNDE İSTATİSTİKSEL BAĞIMSIZLIK ESASINDA İNDEKS TERİM AĞIRLIKLANDIRMA

İlker KOCABAŞ

Tez Danışmanı : Prof Dr. Bahar KARAOĞLAN

İkinci Danışmanı : Yrd. Doç. Dr. Bekir Taner DİNÇER

Uluslararası Bilgisayar Anabilim Dalı

Bilim Dalı Kodu : 619.02.04
Sunuş Tarihi : 11.03.2011

Bornova-İZMİR
2011

EGE ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ

(DOKTORATEZİ)

BİLGİ ERİŞİM SİSTEMLERİNDE İSTATİSTİKSEL BAĞIMSIZLIK ESASINDA İNDEKS TERİM AĞIRLIKLANDIRMA

İlker KOCABAŞ

Tez Danışmanı : Prof. Dr. Bahar KARAOĞLAN

İkinci Danışmanı : Yar. Doç. Dr. Bekir Taner DİNÇER

Uluslararası Bilgisayar Anabilim Dalı

Bilim Dalı Kodu : 619.02.04
Sunuş Tarihi : 11.03.2011

Bornova-İZMİR

2011

Sayın İlker Kocabaş tarafından DOKTORA tezi olarak sunulan "Bilgi Erişim Sistemlerinde İstatistiksel Bağımsızlık esasında İndeks Terim Ağırlıklandırma" başlıklı bu çalışma E.Ü. Lisansüstü Eğitim ve Öğretim Yönetmeliği ile E.Ü. Fen Bilimleri Enstitüsü Eğitim ve Öğretim Yönergesi'nin ilgili hükümleri uyarınca tarafımızdan değerlendirilerek savunmaya değer bulunmuş ve 11.03.2011 tarihinde yapılan tez savunma sınavında aday oybirliği/oyçokluğu ile başarılı bulunmuştur.

Jüri Üyeleri:		**İmza**
Jüri Başkanı	**: .....**	**.....**
Raportör Üye	**: .....**	**.....**
Üye	**: .....**	**.....**
Üye	**: .....**	**.....**
Üye	**: .....**	**.....**

ÖZET

BİLGİ ERİŞİM SİSTEMLERİNDE İSTATİSTİKSEL BAĞIMSIZLIK ESASINDA İNDEKS TERİM AĞIRLIKLANDIRMA

KOCABAŞ, İlker

Doktora Tezi, Uluslararası Bilgisayar Enstitüsü
Tez Yöneticisi: Prof. Dr. Bahar KARAOĞLAN
İkinci Danışmanı: Yrd. Doç. Dr. Bekir Taner DİNÇER

Mart 2011, 120 sayfa

Bu tezde *bilgi erişim* (kıs. BE) sistemlerinde *indeks terim ağırlıklandırma* (kıs. İTA) işlemi için iki farklı yaklaşımda özgün modeller geliştirilmiştir. *Bağımsızlıktan sapma* (İng. Divergence From Independence, kıs. DFI) ve *Luhn-tabanlı* modeller olarak adlandırılan bu iki yaklaşım, sırasıyla: 'istatistiksel bağımsızlık fikri' ve 'Luhn'un kelime frekansı ile kelime önemi ilişkisi hakkındaki iddiasını' temel almaktadır.

Luhn'un iddiasının BE açısından geçerliliği detaylıca incelenmiş ve destekleyici bulgulara ulaşılmıştır. Luhn'un iddiasını nicel olarak gösteren ve 'Terim Frekansı × Ters Belge Frekansı' (İng. Term Frequency × Inverse Document Frequency, kıs. TF×IDF) şemasının temel alan İTA formülleri: *z puanları tabanlı* ve *medyan tabanlı* olmak üzere iki farklı yöntemle oluşturulmuştur. Ortaya konulan modellere uygun İTA formüllerinin BE başarımları TREC (İng. Text Retrieval Conference) 6, 7 ve 8 anlık sorgu izi veri kümelerinde test edilmiştir. Bu testlerde elde edilen BE başarımları ile sunulan istatistiksel yaklaşımların indeks terim ağırlıklandırma probleminin çözümü için kullanılabileceği sonucuna varılmıştır.

Bu tez kapsamında geliştirilen DFI ile ağırlıklandırma kullanan BE sistemi ile aktif olarak TREC-2009 ve TREC-2010'a katılınmıştır. Türkiye'den ilk defa katılınan 2009 yılındaki TREC'te yalnızca ağırlıklandırma ile bile ortalama bir başarım elde edilmiştir. BE işlemindeki temel bazı yöntemlerin DFI üzerine eklenmesi ile gerçekleştirilen yürütümler ile TREC-2010 web izi anlık sorgu görevinde en iyi sistemler arasına girilmiştir.

Anahtar sözcükler: İndeks terim ağırlıklandırma, bilgi erişim, Luhn'un iddiası, bağımsızlıktan sapma, TF×IDF

ABSTRACT

INDEX TERM WEIGHTING BASED ON STATISTICAL INDEPENDENCE FOR INFORMATION RETRIEVAL SYSTEMS

KOCABAŞ, İlker

PhD Thesis, International Computer Institute
Supervisor: Prof. Dr. Bahar KARAOĞLAN
Second Supervisor: Asst. Prof. Dr. Bekir Taner DİNÇER

March 2011, 120 pages

In this thesis, two novel models are developed for index term weighting (ITW) process in information retrieval (IR) systems: one of these is based on statistical independence notion and named as divergence from independence model (DFI) and the other is based on Luhn's claim on the relation between term frequency and term importance.

Luhn's claim's validity in the scope of IR has been investigated elaborately and supporting findings are reached. In order to express Luhn'claim quantitatively, ITW formulas based on Term Frequency × Inverse Document Frequency (TF×IDF) schema are constructed by means of z scores and median approaches. The IR perfomances of ITW formulas of the developed models are tested on TREC (İng. Text Retrieval Conference) 6, 7 and 8 adhoc track datasets. IR perfomance analysis shows that presented statistical approaches can be used in the solution of index term weighting problem.

Active participations in TREC-2009 and TREC-2010 have been carried out with the IR system which was developed around the idea of DFI weighting in the scope of this thesis. In TREC-2009, the IR system developed achieved average performance even it was using only ITW, and was actually the first participation from Turkey. The addition of some fundemantal methods on DFI have raised the performance of the IR system to the level of those ranking at the top in TREC-2010 web track adhoc task.

Keywords: Index term weighting, information retrieval, Luhn's claim, divergence from independence, TF×IDF

TEŞEKKÜR

Bu çalışma süresince deneyimi, bilgisi ve önerileriyle beni araştırma ve geliştirmeye yönlendiren tez danışmanım Bahar Karaoğlan'a ve ikinci tez danışmanım Taner Dinçer'e en içten teşekkürlerimi sunarım. Ayrıca Bilkent Üniversitesinden Fazlı Can'a bu çalışmaya verdiği katkıdan ve harcadığı emekten dolayı teşekkürü bir borç bilirim.

Çalışmalarda desteklerini esirgemeyen, Ulaslararası Bilgisayar Enstitüsünden sevgili hocalarım Mehmet Emin Dalkılıç, Muhammed Cinsdikici ve Cengiz Güngör'e; dostum Geylani Kardaş'a; çalışma arkadaşlarım Müge Sait, Ahmet Bilgili ve Murat Kurt başta olmak üzere diğer tüm arkadaşlarıma ve çalışanlara şükranlarımı sunarım.

Bu çalışma boyunca bana maddi ve manevi her türlü desteği veren, beni yüreklendiren ve bana inanan babama; anneme ve kardeşime teşekkürlerimi sunarken; bu tezi eşim Neylan Kocabaş'a, biricik oğlum Bülent Ege Kocabaş'a ve babam Kadri Kocabaş'a ithaf ediyorum.

İÇİNDEKİLER

Sayfa

İÇİNDEKİLER (devam)

Sayfa

İÇİNDEKİLER (devam)

Sayfa

İÇİNDEKİLER (devam)

Sayfa

ŞEKİLLER DİZİNİ

Şekil Sayfa

ÇİZELGELER DİZİNİ

Çizelge Sayfa

ÇİZELGELER DİZİNİ (devam)

Çizelge Sayfa

ÇİZELGELER DİZİNİ (devam)

Çizelge Sayfa

ÇİZELGELER DİZİNİ (devam)

Çizelge Sayfa

SİMGELER VE KISALTMALAR DİZİNİ

Simgeler	Açıklama
β	Eğim
$B_{(1)}$	Bose-einstein modeli
$B_{(2)}$	İki binom dağılımının oranı
b_j	Belge
C	Y-ekseni kesmesi
IF	Ters terim frekansı modeli
In	Ters belge frekansı modeli
In_exp	Ters beklenen belge frekansı modeli
i	Terim numarası
j	Belge numarası
L	Laplace modeli
M_j	Belge içi medyan
M_j^T	Tahmini belge içi medyan
P	Poisson modeli
s_j	Belge içi standart sapma
t_i	Terim
z	z puanı

Kısaltmalar

AP	Averaj Duyarlık

SİMGELER VE KISALTMALAR DİZİNİ (devam)

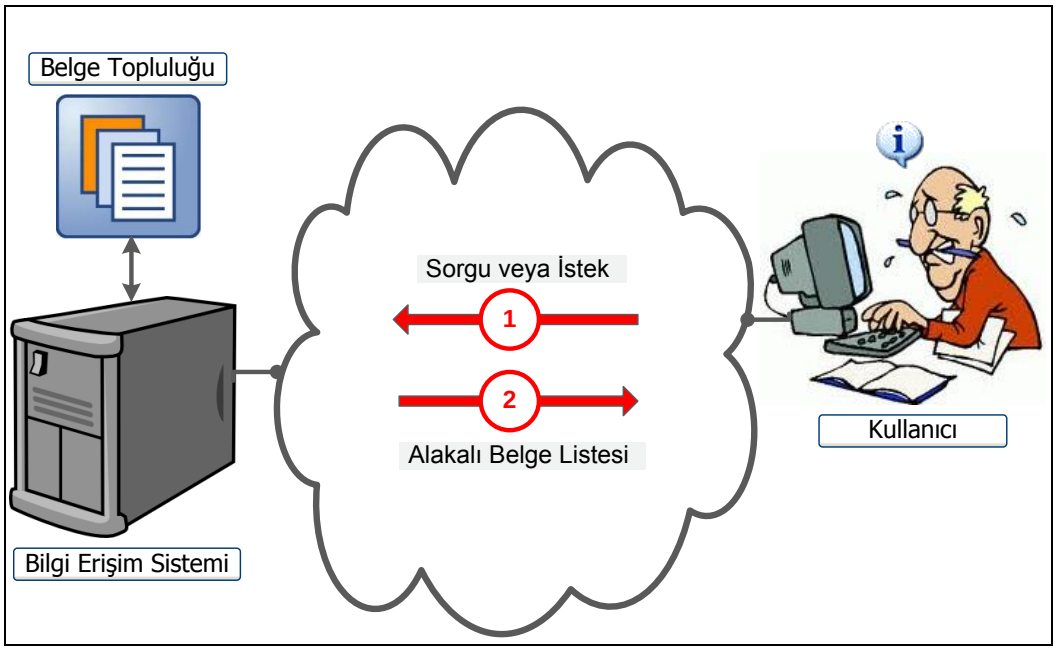
Kısaltmalar	Açıklama
B	Bose-Einstein Modeli
BE	Bilgi Erişim
bpref	İkili Tercih
DFI	Bağımsızlıktan Sapma
ERR	Tahmini Karşıt Sırası
IDF	Ters Belge Frekansı
IEDF	Ters Beklenen Belge Frekansı
IF	Ters Terim Frekansı Modeli
In	Ters Belge Frekansı Modeli
In_exp	Ters Beklenen Belge Frekansı Modeli
ITF	Ters Terim Frekansı
İBF	İstatistiksel Bağımsılık Fikri
LSA	En Küçük Kareler Yaklaşımı
MAP	Ortalama Averaj Duyarlık
MTC	En Düşük Test Topluluğu
nDCG	Normalize-İndirgenmiş Kümülatif Kazanç
NIST	National Institute of Standards and Technology
P	Duyarlık
PCA	Temel Bileşenler Analizi

SİMGELER VE KISALTMALAR DİZİNİ (devam)

Kısaltmalar	Açıklama
Prec-IA	Niyet Duyarlı Duyarlık
R-P	R-Duyarlık
statAP	İstatistiksel Averaj Duyarlık
TDV	Terim Ayırt Etme Değeri
TERRIER	Terabyte Retriever
TF	Terim Frekansı
TF×IDF	Terim Frekansı × Ters Belge Frekansı
TREC	Text Retrieval Conference

1 GİRİŞ

Yirminci yüzyılın ortalarından itibaren bilimsel, sanatsal, siyasal, ekonomik, güncel vb. iş sahalarında yapılan birikimli çalışmalar bizleri yönetilmesi, depolanması ve erişilmesi gereken büyük hacimli bir enformasyon yığını ile karşı karşıya bırakmıştır. Bilgisayar ve iletişim teknolojilerindeki gelişmelere paralel olarak üretilen bilginin; yani enformasyonun %90'dan fazlasının sayısal olduğu tahmin edilmektedir (Varian, 2005). Yine bilgisayar bilimlerindeki gelişmeler doğrultusunda, yönetim ve depolama ihtiyacı kataloglama ve indeksleme açısından -sayısal kütüphanelerde örnekleri gözlemlendiği şekilde- kısmen de olsa çözümlere kavuşturulmuştur. Fakat ihtiyaçlardan biri olan istenilen enformasyona tekrar erişim, daha doğrusu bir kişinin ihtiyaç duyduğu bilgi ile ilişkili/alakalı olan belgelere erişim araştırma için hala açık olan bir konudur. *Bilgi erişim* (İng. *information retrieval*, kıs. BE) işi bir sorgu veya konu başlığı talebine karşılık tanımlı bir yapıya sahip olmayan alakalı kayıtların/belgelerin geri getirimi ile ilgilenen bir disiplin olarak tanımlanabilir. Kullanıcı talebinin nasıl ifade edileceğine dair önceden tanımlı bir yapı olabilir veya olmayabilir. Sorgu, doğal dilde yazılmış bir soru cümlesi olabileceği gibi Boole ifadesi şeklinde de yapılandırılmış olabilir.

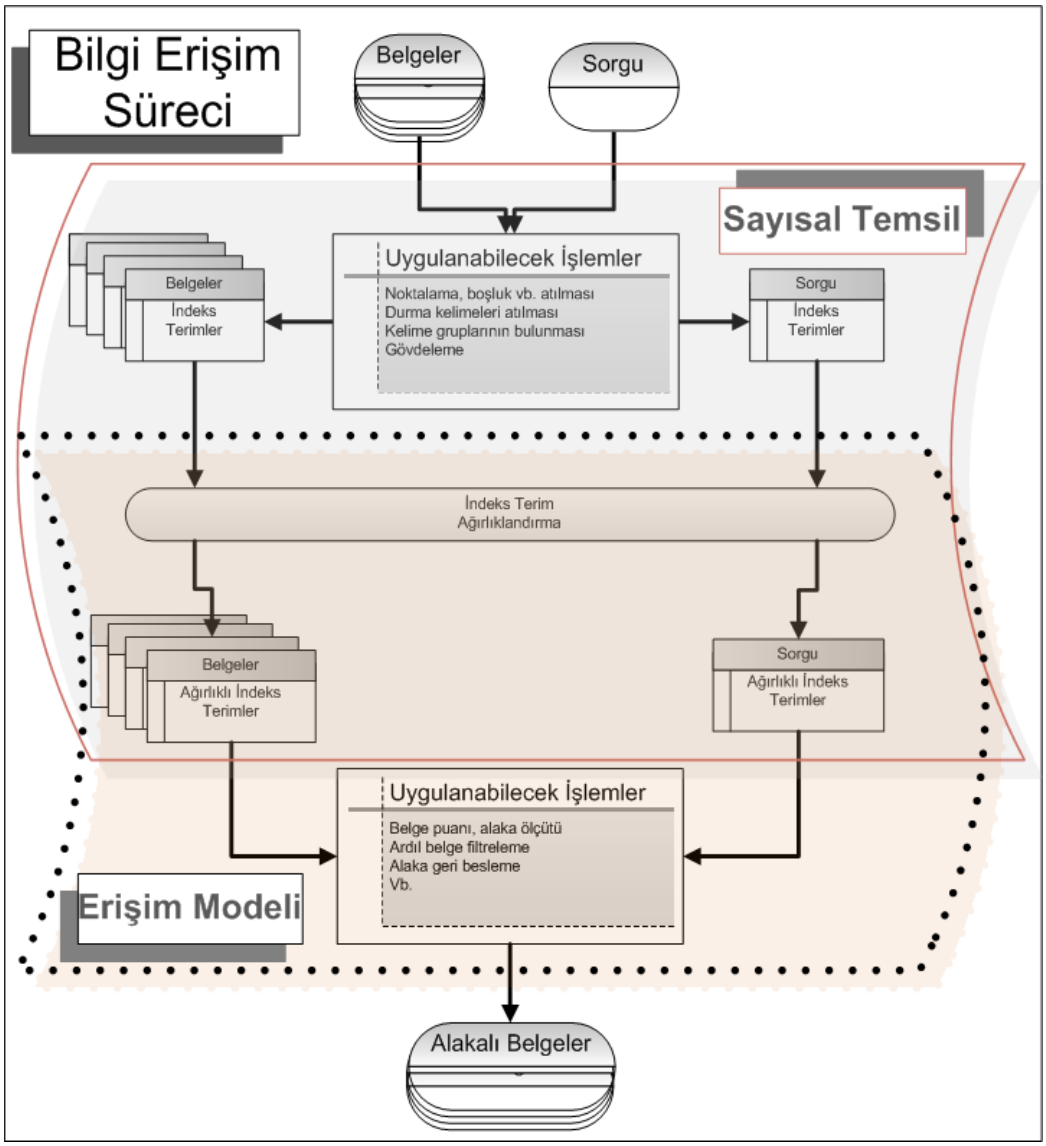


Şekil 1.1 Sayısal ortamda gerçekleşen klasik bilgi erişim işlemi.

Kullanıcının ihtiyaç duyduğu bilgiyi ifade eden sorgu karşılığında sistem tarafından bulunan alakalı belgelerin kullanıcıya getirilmesinden oluşan bilgi erişim işleminin sayısal ortamda **otomatik** olarak gerçekleştiren sistem grafiksel olarak Şekil 1.1'de verilmiştir.

Bilgi erişim sahasındaki ilk çalışmalar "yazılı belge derlemlerinde bilgi erişimi" ile sınırlıydı. Ancak, günümüzde kuramsal alt yapıyı paylaşan ama uygulamada farklılaşan çok sayıda alt başlık bulunmaktadır: ses, müzik, resim, video v.b. çoklu ortam nesnelerinden oluşan durağan derlemlerde anlık-sorgu (İng. adhoc) veya değişen derlemlerde (örn: İnternet) yönlendirme/filtreleme (İng. routing/filtering) sorgu bilgi erişimi; dağıtık derlemlerde (İng. distributed databases) bilgi erişimi; farklı bilgi erişim sistem sonuçlarının katışımı/füzyonu esasında bilgi erişim (İnternetteki kullanımına verilen adı ile meta-search veya kuramsal adı ile data fusion), vb gibi. Bu çalışma kapsamında geliştirilen istatistiksel indeks terim ağırlıklandırma yöntemleri "yazılı belge derlemleri" ile sınırlı tutulmuştur.

Bilgi erişiminin otomatik olarak gerçekleştirilmesine yönelik öncü çalışmalarda ve yakın geçmişe kadar takip eden araştırmalarda bilgi erişim süreci bağımsız iki parça olarak ele alınmıştır: sayısal temsil ve erişim modeli. Sayısal temsil, hem hedef derlemi oluşturan belgelerde hem de sorguda içeriği temsil edecek terimlerin, yani indeks terimlerin tespit edilmesi ve sayısal ortamda temsil edilmesi işini kapsar; erişim modeli de verilen herhangi bir sorgu ile belgeler arasındaki alaka düzeyinin hesaplanması işlemini tanımlar. Doğal dilde yazılmış belge ve sorgular için sayısal temsil sürecinde değişik metin işleme işlemleri: *belgelerin noktalama, boşluk gibi bazı yapılardan temizlenmesi, durma kelimelerinin (İng. stopwords) atılması, kelime gruplarının bulunması, gövdeleme* vb. gibi yöntemler uygulanabilir. Erişim süreci ise belge ve sorgu temsilleri arasında kurulan ilişkileri: *belgeleri puanlayan alaka ölçütü, alaka geri besleme (İng. relevance feedback) işlemi, belgelere öncül/ardıl filtreleme* vb. gibi işlemleri içerir. İndeks terim ağırlıklandırma (İng. index term weighting) ise belge ve sorgu içeriklerini temsil edecek terimlere içeriğe yaptıkları katkıyı nicel olarak gösteren ağırlıklar atanması işlemidir. Aslen indeks terim ağırlıklandırma işlemi sayısal temsil sürecinin bir alt-parçası olsa da, erişim modeline dahil olan indeks terim ağırlıklarının birleştirilme biçimi; yani belge alaka puanının hesaplanması yöntemini doğrudan etkileyebilmektedir. Bu sebepten dolayı, indeks terim ağırlıklandırma işlemi bilgi erişimin her iki alt parçasına dahil edilebilir.



Şekil 1.2 Bilgi erişim sürecinin alt süreçler ve işlemler açısından gösterimi. (Dincer, 2004, uyarlama)

Bilgi erişim sürecinin alt süreçleri ve ilgili temel işlemleri Şekil 1.2'de - Dinçer 'in (2004) gösteriminin uyarlaması- verilmiştir.

Bu tez çalışmasının başlıca amacı, bilgi erişim sahasında indeks terim ağırlıklandırma meselesinin çözümü için yeni ve başarımları kayda değer istatistiksel terim ağırlıklandırma modellerinin geliştirilmesidir. Diğer bir hedef de, Luhn'un (1957) kelimelerin önemi ile ilgili ortaya koyduğu fikrin BE sahasında geçerliliğinin araştırılmasıdır. Bu amaçlar doğrultusunda, tez çalışmasında *"Bağımsızlıktan Sapma" (İng. Divergence From Independence)* ve *"Luhn'un iddiası"* esaslarında olmak üzere iki farklı temel fikir üzerinden çeşitli

istatistiksel indeks terim ağırlıklandırma yöntemleri sunulmuş ve başarımları deneysel olarak incelenmiştir.

Bu çalışma kapsamında geliştirilmiş olan istatistiksel indeks terim ağırlıklandırma yöntemlerinden özgün tasarım ve yüksek başarıma sahip olan "*Bağımsızlıktan Sapma*" modeli üzerine inşa edilmiş bir bilgi erişim sistemi ile uluslararası bir yarışma olan TREC (İng. Text REtrieval Conference) organizasyonuna ulusal düzeyde ilk katılım gerçekleştirilmiştir. TREC derlemleri değişik konularda, farklı amaçlara hizmet eden yazılı belge derlemleridir. Bu derlemler arasında milyar sayıda belge ve terabayt depolama boyutuna kadar ulaşan ölçeklenebilir derlemler bulunmaktadır. Amerika Birleşik Devletleri Ulusal Standartlar ve Teknolojiler Enstitüsü (İng. National Institute of Standards and Technology -kıs. NIST) tarafından her yıl düzenlenen TREC kapsamındaki bilgi erişim çalıştaylarına, farklı alt uğraş dallarında onlarca üniversite (MIT, RMIT, Carnegie Mellon, Massachusetts, Wisconsin, Glasgow, Waterloo, Amsterdam, Geneva, Paris-Sud, Fudan, Tokyo, Chinese Academy of Science, Beijing, Hong Kong Polytechnic, Meiji, vs.) ve pek çok araştırma enstitüsü (National Security Agency, IBM Research, Microsoft Research, Sabir Research vs.) kendi geliştirdikleri bilgi erişim sistemleri ile katılırlar. Bu sebeple, TREC derlemleri bilgi erişim sahasının standart derlemleri olarak kabul edilir. On beşincisi 2006 yılında düzenlenen *TREC* yarışmasına toplam 117 grup katılmıştır. TREC derlemleri, yıllar içinde gelişerek toplam 800 sorgu için alakalı belge taramasından geçirilmiş farklı konulardaki belgelerden oluşmaktadır. Genelde her yıl 50 yeni sorgu hazırlanır ve katılımcılardan belirlenen derlemlerde, her sorgu ile alakalı olduğunu hesapladıkları 1000 belgeyi sonuç kümesi olarak sunmaları istenir; NIST tarafından belirlenen talimatlara göre birden fazla sonuç kümesi de sunabilmektedir. Daha sonra NIST, bu sonuçlara göre başarımları hesaplar ve ilan eder. Her yıl TREC içinde değişik başlıklar altında birden fazla çalışma grubu oluşturulur ve bunların her birine "iz" (İng. track) adı verilir. Bu doktora çalışmasında, geliştirilen indeks terim ağırlıklandırma yöntemlerinin deneysel sınamaları için TREC-6, TREC-7 ve TREC-8 anlık-sorgu izleri materyalleri kullanılmıştır. Yine bu doktora çalışmasında, geliştirilen indeks terim ağırlıklandırma yöntemlerinin kullanıldığı bilgi erişim sistemleri ile TREC-2009 ve TREC-2010 konferanslarına aktif katılım gerçekleştirilmiştir.

Bu tez çalışmasının katkılarını maddeler halinde özetleyecek olursak:

1) Başarımları yüksek olan mevcut yöntemlere alternatif olabilecek özgün indeks terim ağırlıklandırma yöntemleri geliştirilmiştir. Bu modellerin temel aldığı *"Bağımsızlıktan Sapma"* fikrinin bilgi erişim açısından uygun olduğu sonucuna varılmıştır.
2) Luhn'un kelimelerin önemi hakkındaki iddiasını bilgi erişim sahasında **tam** ve **biçimsel** olarak inceleyen ilk çalışmadır. İlgili deneyler ile Luhn'un iddialarını destekleyen bulgular elde edilmiştir. Sonuç olarak, indeks terim ağırlıklandırma yöntemlerinin bu doğrultuda ilerlemesi için bir temel oluşturmuştur.
3) BE sahasında sistemlerin yarıştırıldığı uluslararası standart bir organizasyon olan TREC çalıştayına (2009 yılında) Türkiye'den ilk defa katılım gerçekleştirilmiştir. TREC-2009 organizasyonunda, üzerinde geliştirilen ağırlıklandırma yöntemlerini ek bir iyileştirme olmadan koşturulan BE sistemleri diğer sistemlere göre ortalama başarı yakalamıştır. TREC-2010 organizasyonunda ise, geliştirilen ağırlıklandırma modeli üzerine eklenen bazı temel iyileştirme yöntemlerini kullanan sistemlerin diğer sistemlere göre başarılı olması ileri çalışmalar için teşvik edicidir. BE sahasında farklı fikir ve yaklaşımlar ortaya koyması ile yeni araştırma ufukları açacağı düşünülmektedir.

Bu tez çalışmasının anlatımında izlenen yolu kısaca şu şekilde belirtmek mümkündür: Bölüm 2'de; yani ilgili çalışmalar altında indeks terim ağırlıklandırma problemi verilmekte ve bu problemin çözümüne ilişkin mevcut yöntemler ile bu yöntemlerin değerlendirilmeleri bulunmaktadır. Bölüm 3'te ise tez kapsamında kullanılan materyaller; TREC materyalleri, TERabyte RetrIevER (kıs. TERRIER) bilgi erişim platformu (University of Glasgow, 2010) ile bilgi erişim başarımını ölçmek için kullanılan ölçütler anlatılmaktadır. Bölüm 4'te ise tez kapsamında geliştirilen istatistiksel ağırlıklandırma modelleri; "Bağımsızlıktan Sapma" ve "Luhn'un iddiası" (1957) esasında geliştirilen modellerin temel fikirleri, nicel formülleri ve TERRIER bilgi erişim sistemindeki gerçekleştirimleri açıklanmıştır. Bölüm 5'te ise geliştirilen modellerin deney sonuçları verilmiştir. Bölüm 6'da 2009 ve 2010 yıllarındaki TREC'lerde gerçekleştirilen yürütümlerin başarım sonuçları bulunmaktadır. Son bölümde ise bu doktora tez çalışması kapsamında elde edilen sonuçlar ve ileride yapılması planlanan çalışmalar anlatılmıştır.

2 İLGİLİ ÇALIŞMALAR

2.1 İndeks Terim Ağırlıklandırma Problemi

Yazılı kaynak topluluklarından, yani alanyazındaki[1] adı ile belge derlemlerinden bilgi erişim işinin otomatik şekilde yapılması için öncelikle yazılı belgelerin içeriğini oluşturan *enformasyonun* otomatik bir şekilde belirlenebilmesi gerekmektedir. Bir belgede taşınan enformasyonun belirlenmesi ve sayısal ortamda temsil edilmesi işine, kısaca *belge temsili* adı verilir. Bir belgede taşınan enformasyonun otomatik şekilde belirlenmesi meselesi, aslen o belgenin yazıldığı lisanda otomatik olarak anlaşılması meselesi ile uğraşan bir başka büyük araştırma sahasının kapsamına girer. Eğer, belgelerin yazılı olduğu lisanda otomatik olarak anlaşılması işi çözümlenirse, bilgi erişim işi de çözümlenmiş olacaktır. Van Rijsbergen (1979) bu durumu şöyle özetlemiştir:

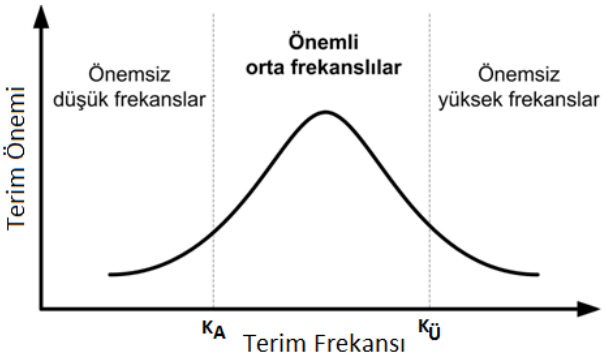
> "Bir bilgi erişim sisteminin belge içeriğini anlamaya çalışması gerektiği bugüne kadar kabul edilmedi.Şu andaki çoğu sistem sadece bibliyografik aramayı hedefliyor. Belgeler yüzeysel tanımlamaları temel alarak alakalı olarak farz ediliyor. Belgeleri anlamak üzere bir bilgisayarı programlamanın kolay bir iş olacağını öngörmüyorum. Öngörüm, sistemdeki her belgenin içeriğinin anahtar kelimler kullanımının ötesinde naïf bir model ile tasarlanması için çalışmalar yapılması gerektiğidir."

Aradan geçen otuz yılı aşkın sürede anlamı yeterli olarak hesaplayabilen bir model ve bu modelin Doğal Dil İşleme (İng. Natural Language Processing) şeklinde alanyazında adlandırılan bu sahaya aktarımı ortaya çıkamamıştır. Sonuç olarak Doğal Dil İşleme sahasındaki meseleler ne yazık ki halen çözülmesi zor problem olma niteliğini sürdürmekte ve ulaşılmış olan noktada bilgi erişim problemlerini çözmedeki başarımı tartışılır durumdadır. Hal böyleyken, belge temsili için ortaya konan mevcut çözümler doğal olarak belgenin yazıldığı lisanın anlaşılması dışındaki yollara yönelmiştir. Bilgi erişim sahasında kabul görmüş çözümlerin neredeyse tamamı istatistik esasında modellere dayanmaktadır. Ancak bu durum, tamamen doğal dil çözümleme, yani dilbilim ekseninden çıkıldığı şeklinde de anlaşılmamalıdır. Belge içerikleri, yazıldıkları lisana ait cümlelerden, cümleler de kelimelerden oluştuğu için dilbilimsel kuramlar, kelime esasındaki istatistiksel yöntemlerin şekillenmesinde büyük rol oynamaktadırlar.

[1] Alanyazın: Herhangi bir bilim kolunda yazılmış olan yazı veya eserlerin bütünü, İng. literature.

Yazılı metinlerde bulunan simge veya simge birlikleri: kelimeler, sayılar, çizimler v.b. insanlar tarafından anlam yüklenmiş her türlü harf, sayı, işaret ve bunların oluşturduğu birliklerdir. Bir metni oluşturan içerik, bu simge veya simge birlikleri ile karşılanan anlamlar bütününden oluşur. Ancak *anlamsal-enformasyonun* yazı ile iletiminde metni oluşturan her simge veya simge birliğinin temsil ettiği anlam, metnin içeriğine, yani anlamlar bütününe her durumda eşit ağırlıkta, yani aynı önemde katkıda bulunmaz hatta bazıları hiç katkıda bulunmaz. Diğer bir deyişle, her terimin içeriğe katkısının ölçüşünü belirten *terim önemi* değeri değişiklik göstermektedir. Eğer anlamsal enformasyonun gösteriminde bu farklılıklar indeks terimlerin ağırlıkları biçiminde kullanılırsa, bir belgenin içeriği daha kesin biçimde karakterize edilebilir (Maron and Kuhns, 1960).

Anlamsal-enformasyon bakış açısından, terim ağırlıklandırmasına yönelik yaklaşım Luhn (1957) tarafından ortaya atılmıştır. Luhn herhangi bir terimin ağırlığını o terimin göreceli *terim sıklığı/frekansı* (İng. term frequency, kıs. tf); yani ilgili terimin gözlenme sayısının verilen bir belgede geçen tüm kelimelere/terimlere göre göreceli değeri biçiminde ifade etmiştir. Luhn bu çalışmasında terim sıklığı bilgisini terim ağırlıklandırması için ilk defa kullanmasına rağmen; 'yazılı bir metinde terimlerin gözlenme sıklığı' ile 'bu terimlerin karşıladıkları anlamların muhtemel enformasyonu oluşturan anlamlar bütünü içindeki önem dereceleri' arasındaki ilişkiyi takip eden çalışmasında açıklamıştır (Luhn, 1958). Luhn'un ortaya koyduğu ilişki/iddia önem derecelerine göre terim frekansları biçiminde Şekil 2.1'de betimlenmiştir.



Şekil 2.1 Bir belge içinde terim önemi ve terim frekansı ilişkisi.

Luhn'un ortaya koyduğu bu ilişkinin önemli yönleri şu şekilde özetlenebilir:

a) Orta frekanslı terimler, düşük ve yüksek frekanslı terimlere göre daha önemlidirler. Belge içinde K_A kesme noktasının altında kalan az görünen

terimler ile $K_Ü$ kesme noktasının üzerinde gözlenen genel terimler belge içeriğine belirgin bir katkı yapmazlar.

b) Terimin çözümleme gücü (İng. resolving power) olarak adlandırılan , yani içeriği ayrıştırmadaki etkinliği orta frekans ile gösterilen aralıkta bir noktada tepe yaparken, bu noktadan her iki kesme yönüne giderken düşer ve kesmelerden sonra ihmal edilebilir olur.

Bu bölüm ve ilerleyen bölümlerde yazılı metinlerde terim olarak *sadece* kelimeler seçilmiştir; bir başka deyişle kelime ile terim aynı anlama gelmektedir.

Taşınan enformasyona katkıları açısından önemsiz kelimelere, dilin yazım kuralları, yani dilbilgisi içinde görevli olan ve yüksek sıklıkta gözlenen kelimeler ile metin içinde çok az görülen kelimeler dahildir. Bu önemsiz kelimeler **işlev kelimeler** olarak adlandırılır, yani yazım veya anlatım içinde işlevleri vardır; ancak nakledilen anlamsal-enformasyon ile ilgili olmadıkları kabul edilir. Bu bakış açısında herhangi bir yazılı belgede bulunan kelimeler belgede taşınan anlamsal-enformasyona yaptıkları katkı açısından eşit ağırlıkta değildir. Bu yüzden, kelimelere içeriğe yaptıkları katkıyı gösteren nicel ağırlıklar verilmesi gerekmektedir. Bir belgede taşınan anlamsal-enformasyona katkıda bulunan kelimelere, **içerik kelimeler** veya **indeks terimler** adı verilir. Belge temsili için kelimelere önem dereceleri açısından ağırlıklar verilmesi işine de alanyazında **indeks terim ağırlıklandırma** adı verilir (Salton, Buckley, 1988).

İdeal bir bilgi erişim sistemi arama yaptığı belge derlemi içinden herhangi bir kullanıcının belirttiği anlamsal-enformasyon ihtiyacı ile alakalı tüm belgeleri sonuç kümesine alırken, alakasız olan belgeleri de dışarıda tutmalıdır. Genel anlamda, derlemdeki belgelerin temsili, yani indeks terimlerin ağırlıklandırması layığı ile yapıldıktan sonra, herhangi bir kullanıcının yazılı olarak tanımladığı enformasyon ihtiyacını karşılayacak sonuç kümesi, enformasyon ihtiyacını tanımlayan kelimelerin geçtiği belgelerin derlemden bir araya getirilmesi ile oluşturulabilir. Aslen bu basit eşleştirme fikri uygulamada bir dereceye kadar iyi sonuçlar vermektedir, fakat idealdeki durumdan da oldukça uzaktır. Yalın kelime eşleştirme karşısındaki tipik mesele şudur: kelimelerin birden fazla anlamının olması alakasız belgelerin sonuç kümesine alınmasına sebep olurken, eş anlamlı ancak farklı imlaya sahip kelimelerle oluşmuş alakalı belgeler de sonuç kümesi dışında kalabilmektedir. Bu mesele, yalın kelime eşleştirme esasında bilgi erişim açısından en meydan okuyucu olandır. Bu mesele belge topluluklarının kümelenmesi (İng. Clustering) yolu ile çözülmeye çalışılmıştır. Kümeleme işlemi;

yani terim vektörleri ile temsil edilen belgelerden otomatik olarak içerik benzerliğine göre kümelenmiş belge toplulukları üretilmesi için bir çok bilinen yöntem mevcuttur (Salton, 1968). Salton çalışmasında (1975a) kümeleme ile geri getirim başarımlarının arttığını göstermiştir. Ancak kaybın bu kadarla sınırlı olmaması da bir başka meseledir. İndeks terimlerin ağırlıklandırılmasında kullanılan yöntemler terimlerin belge içindeki önemlerini layığı ile yansıtamamaktadırlar. Dolayısı ile alakalı belgeler yine sonuç kümesi dışında kalabilmekte, alakasız belgeler de sonuç kümesine alınabilmektedir. Luhn tarafından ortaya konan fikir, metin içinde anlamları karşılayan kelimelerin bir özelliğini daha ziyade hissi olarak belirlemektedir. İndeks terimleri ağırlıklandırma için şimdiye kadar kullanılan bütün yöntemler bu niteliksel tanımı nicel hale getirmeye çalışmakta; bunu da bir dereceye kadar başarmaktadırlar. Luhn'un iddiasında da geçtiği gibi, belge içeriğinde taşınan anlamsal-enformasyonla alakalı olma/olmama durumunun merkezinde **sıklık** fikri bulunmaktadır. Esasen bir indeks terim ağırlıklandırma yönteminin terim sıklığı ile anlamsal-enformasyon arasındaki ilişkiyi ne şekide yorumladığı ve nicel olarak ne şekilde ölçtüğü o ağırlıklandırma yöntemi için belirleyici olan temel özelliktir.

2.2 Mevcut İndeks Terim Ağırlıklandırma Yöntemleri

Bilgi erişim meselesine istatistik kullanımı ile yapılan yaklaşımları iki başlık altında toplamak mümkündür. Bu yaklaşımlardan ilkinde, formal olasılık yoğunluk dağılım fonksiyonları (*binom, poisson, hiper-geometrik, normal* dağılım gibi) kullanılır ve dolayısı ile verilen formüller kesindir. İkincisinde ise hissiyat söz konusudur: bazı akla yatkın, mantıklı formül önerileri yapılır ve başarımları deneysel olarak sınanır, söz gelimi TFxIDF şeması ve türevleri bu yaklaşımın temel örneğidir. Bazı olasılık tabanlı modellerin sezgisel olarak nicel hale getirilmesiyle oluşturulan melez yöntemler de diğer bir ara yaklaşımdır.

2.2.1 TFxIDF şeması ve türevleri

TFxIDF şemasında esas alınan temel fikre göre bir indeks terimin iki işlevi vardır: *temsil* ve *ayırt etme*. *Temsil* işlevi indeks terimin bir belge içeriğini temsildeki ağırlığını; *ayırt etme* işlevi de, indeks terimin belge derlemindeki belgeleri birbirlerinden ne derece ayırt edebildiğini ifade eder. *Temsil* işlevi için, istatistiksel ağırlıklandırma yöntemlerinde esas alınan indeks terim özelliği, terimin belge içindeki gözlenme sıklığıdır. Diğer taraftan, *ayırt etme* gücünün ölçümü için ise çeşitli yöntemler önerilmiştir. Bunlar arasında *ters belge frekansı* (İng. inverse document frequency, kıs.IDF), *sinyal-gürültü oranı* (İng. signal-

noise ratio) ve *terim ayırt etme değeri* (İng. term discrimination value, kıs. TDV) (Salton, 1975b; Can and Özkarahan, 1987) gibi yöntemler en bilinenleridir. Ters belge frekansı ise bunlar arasında en çok tercih edilendir.

İndeks terim ağırlıklandırma işinde kullanılan yaygın yöntemlerin hemen hepsi TFxIDF (Salton and Buckley, 1988; Harman, 1992a) indeks terim ağırlıklandırma şemasını temel alırlar. TFxIDF, ingilizce **T**erm **F**requency X **I**nverse **D**ocument **F**requency" olarak tanımlanan, Türkçeye "Terim Frekansı X Ters Belge Frekansı" olarak çevrilebilecek indeks terim ağırlıklandırma şemasının kısaltmasıdır. Ters belge frekansının temelindeki fikir, "bir terimin, derlem içinde gözlendiği belge sayısı azaldıkça, gözlendiği belgeler açısından ayırt ediciliğinin artacağı" şeklindedir. Bir *k* indeks terimi için Sparck Jones'un (1972) ters belge frekansı değerinin hesabı Denklem 2.1'de verilmiştir .

$$idf_k = log_2\left(\frac{n}{n_k} + 1\right) \qquad \text{Denklem 2.1}$$

Bu denklemde *n*, derlemdeki toplam belge sayısı; n_k ise, *k* teriminin derlem içinde gözlendiği toplam belge sayısıdır. Ters belge frekansı ölçü olarak bir indeks terimin tüm ağırlığı olarak kabul edilmez, yalnızca bir parçasıdır. Bir terimin bir belge içinde gözlenme sıklığı ile ters belge sıklığı birlikte kullanılmaktadır. Bu birlikte kullanımın yaygın şeması TFxIDF olarak adlandırılır ve Denklem 2.2'de verilmiştir. Bu şemada TF ile terim fekansı, IDF ile de ters belge frekansı temsil edilmektedir. Temelde bahsedilen bu fikri esas alan diğer tüm yöntemler de, bu esasta ilk önerilen yöntem olması sebebi ile alanyazında TFxIDF şeması esasında indeks terim ağırlıklandırma yöntemi olarak anılırlar.

$$w_{i,k} = tf_{i,k} \times log_2\left(\frac{n}{n_k} + 1\right) = TF \times IDF \qquad \text{Denklem 2.2}$$

Bir indeks terim ağırlıklandırma yöntemi belgelerin uzunluklarını da hesaba katmalıdır. Dolayısı ile bir terimin bir belge içeriğine yaptığı katkının ağırlığını hesaplamak için üç temel etmen kullanılmaktadır: terim frekansı (*tf*), ters belge frekansı (*idf*) ve belge uzunluğu (*dl*). Bir belgenin uzunluğu terim (kelime) sayısı cinsinden ölçülmektedir. Dolayısı ile, herhangi bir *i* belgesindeki toplam terim sayısı dl_i ile temsil edilirse, bu *i* belgesi için *normalleştirilmiş* belge uzunluğu da, '$ndl_i = dl_i$' (ortalama belge uzunluğu)" şeklinde hesaplanacaktır. Normalleştirilmiş belge uzunluğunun da hesaba katıldığı TFxIDF ağırlıklandırma şeması Denklem 2.3'de verilmiştir.

$$w_{i,k} = \frac{tf_{i,k} \cdot idf_k \cdot (K_1 + 1)}{K_1 \cdot [1 - b + (b \cdot \text{ndl}_i)] + tf_{i,k}} \qquad \text{Denklem 2.3}$$

Denklemde, K_1 ve b uyarlama sabitleridir. K_1 sabiti ile terim frekansının etkisi genişletilir. Bu sabitin en iyi değeri şu an için ancak verilen bir derlem için yapılan deneylerle tespit edilebilmektedir. Uyarlama sabitlerinden ikincisi olan b ise belge uzunluğunun toplam ağırlık üzerindeki etkisini ayarlamaktadır. Bu sabit 0 ile 1 arasında değişmektedir. Eğer b = 0 seçilirse belgelerin birden fazla konu başlığından oluştuğu; b = 1 seçilirse belgelerin tekrarlar yüzünden uzun olduğu kabul edilmiş olur.

TREC çalışmalarında başarım açısından göze çarpan TFxIDF şeması esasındaki bir diğer indeks terim ağırlıklandırma yöntemi de *ltu.ltu* olarak adlandırılır (Singhal et al., 1996) ve Denklem 2.4'de verilmiştir.

$$w_{i,k} = \frac{log(tf_{i,k} + 1) \cdot idf_k}{0.8 + 0.2 \cdot \text{ndl}_i} \qquad \text{Denklem 2.4}$$

2.2.2 Olasılık kavramını kullanan modeller

İndeks terim ağırlıklandırma konusunda dikkatleri olasılık modellerine çeken ilk kişi Harter (1974) olmuştur. Harter, *2-Poisson* olarak adlandırdığı ve *Poisson* olasılık yoğunluk fonksiyonunu esas alan bir indeks terim ağırlıklandırma yöntemi önermiştir:

$$\begin{aligned} f(x) &= \alpha \cdot pois(x, \lambda) + (1 - \alpha) \cdot pois(x, \mu) \\ &= \alpha \frac{e^{-\lambda} \lambda^{-x}}{x!} + (1 - \alpha) \frac{e^{-\mu} \mu^{-x}}{x!} \end{aligned} \qquad \text{Denklem 2.5}$$

2-Possion modeli şu fikre dayanır: "Belge içeriğine katkıda bulunan terimlerin ne tür bir karakteristiği olduğu, derlemdeki diğer belgelerden göreceli olarak daha yaygın şekilde kullanıldıkları bir grup seçkin/elit belge üzerinde gözlenebilir.". Modelde belge uzayı bir terime göre seçkin olanlar ve olmayanlar diye iki gruba ayrılmaktadır. Verilen formülde, toplamı oluşturan iki *Poisson* dağılımından ilki verilen terimin seçkin belgelerindeki frekans dağılımını modellerken (λ: bir terimin seçkin belgelerde gözlenen ortalama terim sıklığı); ikincisi aynı terimin seçkin olmayan belgelerindeki frekans dağılımını modellemektedir (μ: bir terimin seçkin olmayan belgelerde gözlenen ortalama terim sıklığı). Dolayısı ile herhangi bir terimin verilen bir belge için *2-Poisson*

modeli ile hesaplanan ağırlık değeri (yani *f*(*x*): *x* terimin belge içindeki gözlenme sıklığı), o terim açısından incelenen belgenin seçkin olma olasılığını vermektedir.

Amati and van Rijsbergen (2002) *Rastlantısal Oluştan Sapma* (İng. Divergence From Randomness) adını verdikleri fikri esas alan indeks terim ağırlıklandırma yöntemleri önermişlerdir. Rastlantısal oluştan sapma fikri şudur: "Bir terimin belge içinde gözlenme sıklığı, bu terimin derlemde gözlenme sıklığından ne kadar saparsa, o terimin söz konusu belgedeki içeriğe yaptığı katkı da o kadar artar". Bu fikre göre belirli bir *t* teriminin verilen herhangi bir *b* belgesi için ağırlığı (*w*(*t*|*b*)), o belgede terimin gözlenme sıklığının, belirli bir *M* olasılık dağılım modeli üzerinden hesaplanan olasılığı ile ters orantılı olmaktadır.

$$w(t|b) \propto -log[Pr_M(t \in b|derlen)] \qquad \text{Denklem 2.6}$$

Denklem 2.6'da *M* olasılık dağılımı "rastlantısal oluşun" ölçüsüdür. Dolayısı ile bir terimin bir belge içerisinde rastlantısal oluştan sapmasını ölçebilmek için öncelikle "raslantısal olma" durumunun tanımlanması gerekir. *M* olasılık dağılım modeli için, pek çok temel olasılık dağılımı önerilmiş ve denenmiştir (Amati and van Rijsbergen, 2002). Bunlar arasında, *Binom, Bose-Einstein,Geometrik* dağılım gibi temel olasılık yoğunluk fonksiyonları bulunmaktadır. Örneğin *M* olasılık dağılım modeli *Binom* dağılımı olarak seçilirse, bir terimin verilen bir belgedeki gözlenme sıklığının olasılık hesabı Denklem 2.7'de verildiği gibidir.

$$-log[Pr_M(t \in b|derlen)] = -log\left[\frac{ntf!}{tf!\,(ntf-tf)!}p^{tf}(1-p)^{ntf-tf}\right] \qquad \text{Denklem 2.7}$$

Denklem 2.7'de, *ntf* ifadesi, *t* teriminin derlem genelinde gözlenme sıklığına; *tf* ifadesi verilen belgede gözlenme sıklığına ve *p* de *1/N* değerine eşittir (*N*: derlemdeki toplam belge sayısıdır). Ek olarak, temel olasılık yoğunluk fonksiyonlarından başka, IDF, ITF (ing. Inverse Term Frequency teriminin kısaltması) ve IEDF (ing. Inverse Expected Document Frequency teriminin kısaltması) gibi pek çok akla yatkın istatistik de denenmiştir. Bu bakış açısıyla oluşturulan ağırlıklandırma formülleri Bölüm 3.3.2'de detaylı olarak anlatılmaktadır.

2.2.3 Melez yaklaşımlar

İndeks terim ağırlıklandırma meselesini halletmek için benimsenen TFxIDF ve olasılık dağılımı esasındaki iki temel yaklaşımın ortaklığında bir yaklaşımı benimseyen melez yöntemler de vardır. Bu ağırlıklandırma yöntemlerinde, bir olasılık dağılım modeli alınır ve bu olasılık dağılım modelinin kesin formülü yerine daha anlaşılır ve takip edilebilir bir başka formül oluşturulur. Bu tip melez yaklaşımlara en iyi örnek *BM25* (Robertson et al., 1999) ağırlıklandırma yöntemidir. *BM25* yönteminde esas alınan olasılık dağılım modeli *2-Possion*'dır. *BM25* aslen kuramsal *2-Possion* modeli ile ortaya konan fikrin formüle çevrilmiş pek çok çeşidinden birisidir; *BM* adı altında *2-Poisson* modeli ile ortaya konan fikri uygulamaya geçirmenin birden fazla yolu önerilmiştir: *BM0, BM1, BM11* ve *BM15 gibi* (Robertson, 1994). Söz konusu çalışmalarda ilginç olan yön, *2-Poisson* modeli tam formül olarak geri-getirim için kullanıldığında pek başarılı olmamasına karşın, *BM25*'in başarımı kayda değerdir. *BM25* ağırlıklandırma yöntemi *Okapi* sisteminde kullanılmaktadır (Robertson et al., 1999) ve bu yöntem ile bir terimin ağırlığı Denklem 2.8'de verildiği şekilde hesaplanır.

$$w_{i,k} = \frac{tf_{i,k}}{0.5 + 1.5 \cdot \mathrm{ndl}_i + tf_{i,k}} log\left[\frac{n - idf_k + 0.5}{idf_k + 0.5}\right] \qquad \text{Denklem 2.8}$$

Amati and van Rijsbergen (2002)'in *Rastlantısal Oluştan Sapma* fikrini temel alan melez ağırlıklandırma yöntemleri de mevcuttur. Bu yöntemlerde, temel alınan olasılık dağılım fonksiyonundan elde edilen kesin formüllere Ponte ve Croft'un (1998) dil modelinde kullandığına benzer olan risk bileşeni eklenmiştir. Bu risk bileşeni altında yatan mantık ve oluşturulan melez yöntemler, Bölüm 3.3.2'de detaylı olarak anlatılmaktadır.

2.3 Mevcut Yöntemlerin Değerlendirilmesi

Church (1995) çalışmasında şu yoruma yer vermiştir:

> "*...bazı kelimeler belgelerde neredeyse eşit şansa dayalı şekide gözlemlenirken bazı kelimeler belgelerin oluşumunda alaka, yazar, üslup gibi var olduğu kabul edilen gizli değişkenlere bağlı olarak dağılım gösterirler. Alaka, yazar, üslup gibi gizli değişkenlere bağlı dağılım gösteren kelimeler bilgi erişim amacına uygun olanlardır çünkü merak edilen gizli değişkenlere ışık tutan onlardır...Bu bağlamda,*

belgelerde şanstan uzak şekilde gözlemlenen kelimeler iyi birer indeks terim olmaktadır."

Bu yorumdaki eşit şansa sahip kelimelerin, Luhn'un yorumundaki işlev kelimelerle; alaka, yazar, üslup gibi gizli değişkenlere bağlı dağılım gösteren kelimelerin de Luhn'un yorumundaki içerik kelimelerle örtüştüğü söylenebilir. Bu noktadan hareketle, içerik ve işlev kelimelerin gözlenme sıklılıklarının sırası ile $f_1(w_k)$ ve $f_2(w_k)$ şeklinde (w_k kelime dağarcığındaki *k:* kelime ve $k = 1; 2,.....$ olmak üzere) iki dağılım ile temsil edilebileceği kabul edilebilir ve böylece alanyazında indeks terim ağırlıklandırılması için bu güne kadar önerilmiş olan yöntemler bu iki dağılımı ele alışlarına göre irdelenebilir.

TFxIDF şeması esasında ortaya konan ağırlıklandırma yöntemleri, içerik ve işlev kelimelere ait dağılımları:

$$TF(w_k) \times IDF(w_k) \rightarrow f_1(w_k) \equiv f_2^{-1}(w_k) \qquad \text{Denklem 2.9}$$

şeklinde birbirlerinin tersi biçiminde ifade ederler. TFxIDF şemasında işlev kelimeler, içerik kelime olmama şeklinde temsil edilir, dolayısı ile örtülü olarak içerik ve işlev kelimelere ait dağılımların birbirinin tersi olduğu kabullenmesi yapılır. Bu şemanın TF bileşeni, yani "terim sıklığı" bileşeni, çok açık şekilde Luhn'un ortaya koyduğu "terim sıklığı ile enformasyon arasındaki ilişki" fikri ile çelişmektedir: TF şeması ile bir terimin belge içindeki enformasyona olan katkısı, terimin belge içindeki gözlenme sıklığı arttıkça artmaktadır, oysa Luhn'a göre katkı orta sıklığa doğru daha yüksek değerlere; düşük ve çok yüksek sıklıklara doğru da düşük değerlere ilerlemelidir. Öte yandan IDF bileşeninin de tamamlayıcı bir yaklaşım olduğunu söylemek zordur. IDF bileşeni, bir terimin gözlendiği belgede taşınan enformasyona yaptığı katkıyı, terimin gözlendiği belge sayısı ile (ters) orantılı olarak ele almaktadır: bir terim, derlem içindeki diğer belgelerde ne kadar az gözleniyorsa, gözlendiği belgedeki enformasyona o denli katkı yaptığı kabul edilir. Bir başka söyleyişle, IDF bir terimin belgeler içindeki gözlenme sıklığı ne olursa olsun (yani terimin olasılık dağılımı ne olursa olsun), az belgede gözlenen terimin, gözlendiği belge için daha önemli olduğunu kabul etmektedir. Bu bakış açısı bir dereceye kadar sezgisel olarak doğru gibi gelse de, her şart altında doğru olduğunu söylemek mümkün değildir. Örneğin bilgi erişim işinin belirli bir zaman aralığındaki gazete makaleleri üzerinde gerçekleştirildiğini kabul edelim, böyle bir derlemde söz konusu zaman diliminde önemli haberler

gazetelerin hemen hepsinde yer alacaktır, dolayısı ile haberin önemi artarken IDF tarafından haberle ilişkili kelimelere atanan ağırlık düşecektir.

IDF bileşeninin söz konusu zayıflığı daha önce Church (1995) tarafından zaten ortaya konulmuştur. Söz konusu çalışmada deneysel sonuçlar üzerinden IDF'in indeks terim ağırlıklandırma için çok uygun olmadığı; Artık-IDF (İng. Residual-IDF) yaklaşımının daha uygun olduğunu belirtilmektedir. Herhangi bir terimin gözlendiği belge için Artık-IDF değeri Denklem.2.10'da verildiği gibi hesaplanır.

$$Artık\ IDF = IDF - \widehat{IDF} \qquad \text{Denklem 2.10}$$

bu denklemde $\widehat{IDF}$, ters belge sıklığı olan IDF değerinin, *Poisson* dağılımı altındaki tahminidir:

$$\widehat{IDF} = -log_2(1 - f(\lambda, 0)) = -log_2(1 - e^{-\lambda}) \qquad \text{Denklem 2.11}$$

Denklem 2.12'de ise *Poisson* dağılımı (simge. *P*) gösteren bir *X* rastsal değişkenine ait olasılık yoğunluk fonksiyonu verilmiştir.

$$P(X = x|\lambda) = f(\lambda, x) = \frac{e^{-\lambda}\lambda^{-x}}{x!} \qquad \text{Denklem 2.12}$$

burada X rastsal değişkeni kelimelerin gözlenme sıklığını temsil eder ve λ da bu gözlenme sıklığının ortalamasıdır. Church (1995) çalışmasında, $\widehat{IDF}$ eşitliğindeki λ değerini, verilen bir kelimenin tüm belgeler üzerinden elde edilen ortalama gözlenme sıklığı olarak almıştır. Dolayısı ile, Artık-IDF aslen ters belge sıklığının (IDF) *Poisson* olasılık yoğunluk fonksiyonu ile tahmin edilen ters belge sıklığından ne kadar saptığının bir ölçüsü olmaktadır.

Bu noktada kolayca anlaşılacağı gibi aslen Amati and van Rijsbergen (2002) tarafından 'Rastlantısal Oluştan Sapma Modeli' olarak adlandırılan yaklaşım, Church tarafından ortaya konan 'Şanstan Uzaklaşma' fikrinden başka bir şey değildir. Amati and van Rijsbergen rastlantısal oluş, daha doğrusu 'şans' dağılımı olarak *Binom*, *Bose-Einstein*, *Geometrik* gibi farklı olasılık yoğunluk fonksiyonlarını kullanmışlardır. Rastlantısal oluştan sapma esasında yapılan ağırlıklandırmada, verilen bir kelimenin belirli bir belge için önem derecesi, söz konusu kelimenin o belgedeki gözlenme sıklığının, aynı kelimenin derlem

genelinde eşit sansa dayalı dağılıma sahip olduğu durumdaki beklenen gözlenme sıklığından farkı olarak hesaplanır. Dolayısı ile, rastlantısal oluştan sapma fikrinde asıl mesele herhangi bir kelimenin şans dağılımının ne olduğunun belirlenmesidir. Bir başka anlatımla, rastlantısal oluş ile belirli bir kelimenin verilen bir belge için önemi hesaplanırken, önce işlev kelimelerin *Binom, Bose-Einstein, Geometrik* gibi bir olasılık yoğunluk fonksiyonuna bağlı dağılım gösterdiği kabul edilir ve o kelimenin söz konusu belgedeki gözlenme sıklığı ile işlev kelime olma olasılığı belirlenir; daha sonra bu işlev kelime olma olasılığı 1'den çıkarılır ve verilen kelimenin içerik kelime olma olasılığı elde edilir. Dolayısı ile rastlantısal oluştan sapma esasında indeks terim ağırlıklandırmada temel unsur içerik kelimeler değil işlev kelimelerdir, yani $f_2(w_k)$ dağılımı için önerimde bulunmaktır. Ancak $f_1(w_k)$ dağılımı TFxIDF'te olduğu gibi ters orantılı olarak ele alınmamakta, $f_2(w_k)$ dağılımının tamlayanı olarak kabul edilmektedir:

$$f_1(w_k) = 1 - f_2(w_k) \qquad \text{Denklem 2.13}$$

Aslen bu yaklaşım, akla daha yatkındır ve zaten deneysel sonuçlar da bu yaklaşımın TFxIDF'den daha başarılı olduğunu göstermektedir. Bilindiği gibi işlev kelimelerin yazılı bir metindeki gözlenme sıklığı dil bilgisi kurallarına bağlıdır ve bu koşul her yazılı metin parçası için değişmeksizin geçerlidir, dolayısı ile içerik kelimeler yazılı metinlerde, yazar, alaka, üslup gibi var olduğu kabul edilen farklı değişkenlere bağlı dağılım göstereceği için işlev kelimelerin 'şans' ile açıklanabilecek bir dağılım göstermesi ihtimali içerik kelimelerden daha kuvvetlidir.

2.3.1 Olasılık Dil Modelleri

Dil modellenmesi (İng. language modeling) veya istatistiksel dil modellenmesi genel anlamda bir doğal dilin kullanımındaki istatistiksel düzenliliğini tutan bir olasılık dağılımını tahmin etmeyi içerir. Bilgi erişim sürecine uygulanmasında ise dil modellenmesi sorgu ve belgenin aynı dil modelinden üretilmiş olma olasılığını tahmin etme problemine denk gelir (Liu and Croft, 2005). Bu problemin BE sahasında aktarımı ilk olarak Ponte ve Croft (1998) tarafından yapılmıştır. Daha sonra bir çok farklı grup tarafından yapılan araştırmalar, dil modellenmesi yaklaşımının teorik olarak kuvvetli ve BE problemlerinde çalışılmasına yönelik olasılıksal çatının (İng. framework) potansiyel olarak çok etkin olduğunu doğrulamıştır (Croft and Lafferty, 2003).

Dil modelleri veya özel ismiyle olasılık dil modelleri ile bütünsel bilgi erişimde verilen herhangi bir *Q* sorgusu ile belirli bir *i* belgesi arasındaki alaka $P(Q|M_i)$ şartlı olasılığı ile ölçülmektedir (Manning et al., 2008). Bu şartlı olasılıkta M_i, *i* belgesine ait olasılık dil modeli olmaktadır. Dolayısı ile bu bütünsel yaklaşımda bir sorgunun bir belge ile alakası, sorguyu oluşturan $Q = w_1, w_2,....$ kelime dizisinin *i* belgesine ait M_i dil modeli tarafından üretilmiş olması olasılığı ile doğru orantılı şekilde hesaplanmaktadır. *Q* sorgusunun w_k gibi tek bir kelimeden oluştuğu durum için şartlı olasılık kelimenin belge içindeki göreceli sıklığına eşittir:

$$Pr(w_k|M_i) = \frac{tf_{i,k}}{dl_i} \qquad \text{Denklem 2.14}$$

burada $tf_{i,k}$, w_k kelimesinin *i* belgesindeki gözlenme sıklığı; dl_i de *i* belgesinin kelime cinsinden uzunluğudur. Sorgunun birden fazla kelimeden oluştuğu durumda, sorgunun *i* belgesine şartlı ilişkilik olasılığı, sorguyu oluşturan kelimelerin belgelerde gözlenmelerinin birbirinden bağımsız olduğu kabullenmesi ile;

$$Pr(Q|M_i) = \prod_{w_k \epsilon Q} Pr(w_k|M_i) \qquad \text{Denklem 2.15}$$

şeklinde hesaplanır. Ponte and Croft (1998) çalışmasında *durma kelimeleri* olarak adlandırılan, Luhn'un tanımlaması içinde işlev kelimeler sınıfına giren kelimelerin hesaplamadan çıkarılması gerektiğini belirtmiştir. Geleneksel bilgi erişim işi içinde durma kelimeleri için bir liste vardır. Bu listede yer alan kelimeler söz konusu lisan içinde anlam taşımadığı kabul edilen, 'edat' cinsi kelimelerdir. Lunh'un tanımlamasındaki işlev kelimeler kümesi genelde lisan içinde herhangi bir anlam taşımayan durma kelimelerini kapsar ancak bununla sınırlı olduğunu söylemek doğru değildir. Özelde anlamlı olsalar bile belge ile taşınan enformasyona katkı yapmayan diğer kelimeler Luhn'un işlevsel kelimeler kümesi içine girmektedir. Dolayısı ile işlev kelimelerin, durma kelimelerinde olduğu gibi bir liste haline getirilmesi pek de mümkün gözükmemektedir. İşlev kelimelerin en genel tanımı ile dilbilgisi kurallarına bağlı dağılım gösterdiği ve derlem içindeki her belgede şansa dayalı şekilde gözlendiği kabul edilirse, tüm belgelerde eşit oranlarda gözlenme sıklığına sahip olması gerekeceği için, olasılık dil modelleri ile hesaplanacak şartlı olasılıkları da her belge için eşit olacaktır. Dolayısı ile

hesaplamaya başlamadan önce durma kelimelerinin belgelerden çıkarılması akla yatkın bir yaklaşım olarak gözükmektedir.

Olasılıksal dil modelleri ile yapılan indeks terim ağırlıklandırma işlevsel olarak esasen TFxIDF şemasındaki TF bileşenin görevini yerine getirir, ancak aralarında önemli bir fark vardır. TF bileşeni ile bir kelimenin verilen herhangi bir belge içindeki anlamsal-enformasyona katkısı söz konusu kelimenin o belgedeki gözlenme sıklığı ile doğru orantılı olarak ele alınır. Olasılık dil modelleri de sıklık ile anlamsal-enformasyona katkıyı doğru orantılı olarak ele alır, fakat iki yaklaşım arasındaki temel fark hesaplama esnasında ortaya çıkar. Olasılık dil modelleri ile hesaplanan değerler belge uzunluğuna bağımlı değildir; oysa TF bileşeni ile hesaplanan değerler belge uzunluğuna bağımlıdır. Dolayısı ile, olasılık dil modelleri, TFxIDF şemasını esas alan ancak *normalleştirilmiş* belge uzunluğu kullanan ağırlıklandırma yöntemleri ile kıyaslanmalıdır.

3 TEMEL BAŞARIM ÖLÇÜTLERİ VE KULLANILAN MATERYAL

Bu bölümde tez kapsamında kullanılan başarım ölçütleri ile materyaller anlatılmıştır. Kullanılan materyaller iki bölüm altında ele alınmaktadır. Bunlardan ilki, deneylerde kullanılan ve aktif olarak katılımda bulunulan TREC izlerinde erişim başarımı değerlendirmesi için gerekli olan TREC materyalleri bölüm 3.2'de; diğeri ise bilgi erişim sisteminin gerek duyduğu indeksleme (İng. indexing), erişim (İng. retrieval) ve değerlendirme (İng. evaluation) gibi fonksiyonların Java programlama dilliyle gerçeklendiği TERRIER (TERabyte RetrIEveR) kütüphanesi, bu fonksiyonların içsel bileşenleri ve kütüphane içinde hazır olarak kodlanmış ağırlıklandırma fonksiyonları/modelleri Bölüm 3.3'te verilmiştir.

3.1 Başarım Ölçütleri

Başarım ölçütleri ile BE sistemlerinin başarımlarını değerlendirilmek için kullanılan ölçütler belirtilmektedir. Bu ölçütler, BE görevlerine/amaçlarına göre farklı özelliklere sahiptirler. Ayrıca aynı BE görevi için tasarlanan farklı özelliklerde başarım ölçütleri mevcuttur. Bu bölümde anlatılan ölçütler, tez çalışması kapsamındaki TREC izlerinde NIST tarafından kullanılanlarla sınırlı tutulmuştur.

3.1.1 Duyarlık

Duyarlık (İng. Precision, kıs. P) sistemin başarımını sadece erişilen alakalı belgeler açısından sunan ölçüttür. RR (İng. **R**elevant **R**etrieved) ile erişilen alakalı belge kümesi ve DR (İng. **D**ocument **R**etrieved) ile erişilen belge kümesi gösterilecek olursa, sistemin "duyarlık" değeri Denklem 3.1'deki gibi hesaplanır.[2]

$$Duyarlık = \frac{|RR|}{|DR|} \qquad \text{Denklem 3.1}$$

3.1.2 Anma

Anma (İng. Recall) sistemin başarımını tüm alakalı belgeler üzerinden gösteren ölçüttür. RD (İng. **R**elevant **D**ocuments,) ile derlemde bulunan alakalı

[2] Herhangi bir S kümesinin elaman sayısı |S| biçiminde gösterilmektedir.

belgeler kümesi gösterilecek olursa, sistemin "anma" değeri Denklem 3. 2'deki gibi hesaplanır.

$$Anma = \frac{|RR|}{|RD|} \qquad \text{Denklem 3.2}$$

3.1.3 R-Duyarlık

R-Duyarlık (İng. R-Precision, kıs. R-P) sistemin başarımını derlemde bulunan derlemdeki alakalı belge sayısı kadar ki erişilen belgeler üzerinden gösteren ölçüttür. RR(RD) ile erişilen alakalı belge sayısı –erişim listesindeki ilk alakalı belge sayısı kadarki belgede- gösterilecek olursa, sistemin "R-Duyarlık" değeri Denklem 3.3'teki gibi hesaplanır.

$$R\text{-}Duyarlık = \frac{RR(RD)}{|RD|} \qquad \text{Denklem 3.3}$$

3.1.4 Averaj duyarlık ve ortalama averaj duyarlık

Duyarlık ve anma küme tabanlı ölçütlerdir; yani erişilen belgelerdeki sıralamayı dikkate almadan değerlendirme yaparlar. Sıralı olarak değerlendirme için ise tek bir konuda başarım için "averaj duyarlık" (İng. **A**verage **P**recision, kıs. AP), konu kümesinde başarım için ise konuların AP değerlerinin ortalaması olan "ortalama averaj duyarlık" (İng. **M**ean **A**verage **P**recision, kıs:MAP) ölçütleri (Croft et al., 2009) kullanılır.

TREC derlemlerinde kullanılan AP değeri erişilen her alakalı belgeden sonraki duyarlık değerlerinin ortalamasıdır. Herhangi bir konu için AP değeri Denklem 3.4'de olduğu gibi hesaplanır.

$$AP = \frac{\sum_{i \in DR}^{|DR|} P_i}{|RD|} \qquad \text{Denklem 3.4}$$

MAP ise konu kümesi (K) bulunan her konunun AP değerinin ortalamasıdır ve Denklem 3.5'te hesaplanması verilmiştir.

$$MAP = \frac{1}{|K|} \sum_{j \in K}^{|K|} AP_j \qquad \text{Denklem 3.5}$$

Ancak her konunun alakalı belge sayısı farklı olmasından dolayı konu kümesinin ortalama başarımı hesabında AP ölçütü bu şekliyle yetersiz kalmaktadır. Her konu için ortalama duyarlık değerlerini normalleştiren AP ölçütünde, duyarlık değerleri 11 anma seviyesine ara değerlenir (İng. interpolated); anma seviyeleri 0-1 arasındaki 0,1 aralıklı anma değerleridir {0,0; 0,1;....;0,9; 1,0}. Her anma seviyesindeki duyarlığı kalan anma aralığındaki en büyük değer olarak alınması, ara-değerlenme kuralı olarak kullanılmaktadır (örn. 0,2 seviyesindeki duyarlık değeri [0,2; 1,0], 0,5 seviyesindeki ise [0,5; 1] aralıklarındaki gerçek duyarlık değerlerinin en büyüğüdür). Bir konu için TREC derleminin kullandığı AP ve ara-değerlenmiş AP değerleri hesaplanması aşağıdaki örnekte gösterilmiştir.

Örnek 3.1: Belge kümesinde konuyla alakalı 3 belge {D1, D4, D7} varken; bir sistem tarafından erişilen 10 belge ve her erişilen belgedeki {duyarlık, anma} değerleri Çizelge 3.1'de verilmiştir.

Çizelge 3.1 Örnekteki duyarlık ve anma hesapları (alakalı belgelerin altı çizilmiştir).

Veri Tipi	Sıra									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Belgeler	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Duyarlık (gerçek)	1,00	0,50	0,30	0,50	0,40	0,33	0,43	0,38	0,33	0,30
Anma	0,33	0,33	0,33	0,66	0,66	0,66	1,00	1,00	1,00	1,00

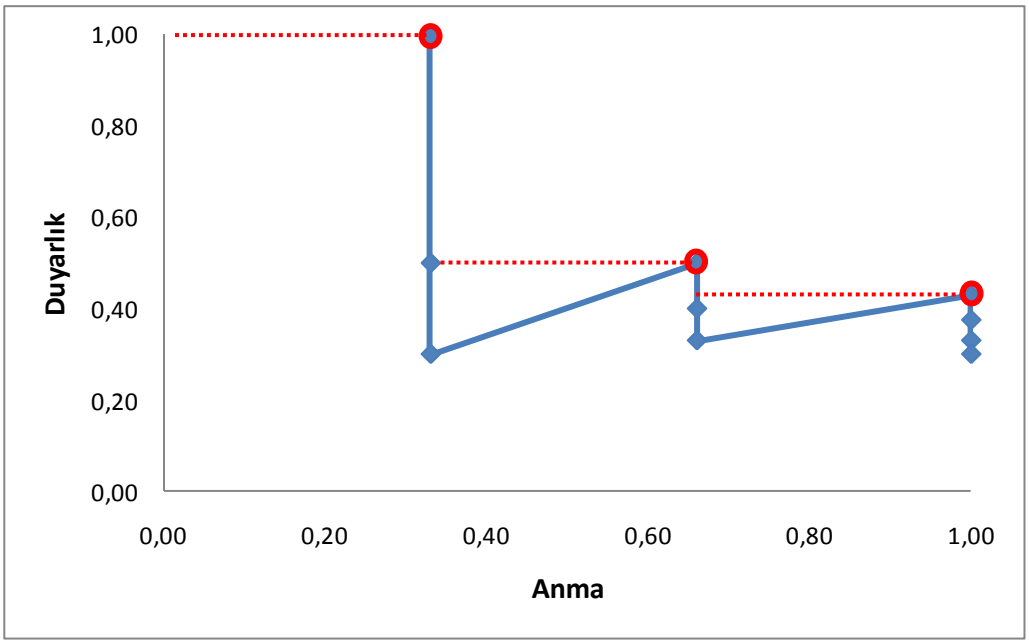
Gerçek duyarlık değerleri kare biçiminde (düz çizgilerle birleştirilerek) 3.1'de verilen duyarlık-anma grafiği üzerinde gösterilmiştir.

Erişilen alakalı belgeler {D1, D4, D7} için sırasıyla {1/1=1; 2/4=0,5; 3/7=0,43} duyarlık değerleri de yuvarlak içinde gösterilmiştir. Bu üç duyarlık değerinin ortalaması alınarak AP hesaplanır:

$AP = (1 + 0{,}5 + 0{,}43) / 3 = 1{,}93 / 5 = 0{,}386$

Ara-değerlenmiş duyarlık ise kesikli çizgilerle belirtilmekte olup, anma seviyelerinde { 1 (0,0); 1 (0,1); 0,5 (0,2); 0,5 (0,3); 0,5 (0,4); 0,43 (0,5); 0,43 (0,6); 0,43 (0,7); 0,375 (0,8); 0,33 (0,9); 0,3 (1,0)} değerlere eşittir. "Ara-değerlenmiş AP" ise aşağıdaki biçimde 0,727 olarak hesaplanmıştır.

$$\textit{Ara-değerlenmiş } AP = (2\times 1 + 3\times 0{,}5 + 3\times 0{,}43 + 0{,}375 + 0{,}33 + 0{,}3)/11 = 0{,}727$$



Şekil 3.1 Örnek 3.1'in Ara-değerlenmiş Duyarlık – Anma Eğrileri

3.1.5 *Normalize-indirgenmiş kümülatif kazanç* ve *beklenen karşıt sırası* ölçütleri

Önceki başarım ölçütleri, belgeleri alakasız (0) ve alakalı (1) olmak üzere iki değerli ele alırlar. Diğer bir deyişle, belgelerin alakalık düzeyleri ikili derecelendirme sistemine dayanır. Belgelerin alaka düzeylerinin gösteriminde çoklu (ikiden fazla) derecelendirmeyi kullanan yapılar için ise daha farklı başarım ölçütleri geliştirilmiştir. *normalize-indirgenmiş kümülatif kazanç* (İng. normalized-discounted cumulative gain, kıs. nDCG) (Järvelin and Kekäläinen, 2002) ve *beklenen karşıt sırası* (İng. expected reciprocal rank, kıs. ERR) (Chapelle et. al., 2009) bu tip başarım ölçütlerinden ikisidir.

"Çoklu derecelendirme" kullanıldığı durumlarda, her belgenin alakalık düzeyi *0* ile *d* (*d* > 1) arasında bir tamsayı ile gösterilir. Belgenin, *0* derecesi ile alakasız olduğu belirtilirken *d* ile en alakalı olduğu belirtilir. Aradaki dereceler ile alakalık düzeyi doğru orantılıdır. nDCG ve ERR'in çoklu derecelendirmede kullanılabilmesinden başka bir diğer özelliği de sorguya karşılık erişilen belgelerin erişim sıralarını yani belge listesindeki pozisyonlarını hesaplamaya katmasıdır. Aslen bu başarım ölçütleri listedeki $k^{\text{ıncı}}$ sıraya kadarki belgeler için hesaplama yapar ve bu ölçütler nDCG@k ve ERR@k biçiminde gösterilir.

Bir belgenin listedeki pozisyonu *r* ile gösterilecek olursa *dr* ($0 \leq dr \leq d$) ile r^{ninci} pozisyondaki belgenin alaka düzeyini gösteren derecedir. DCG@k bu durumda Denklem 3.6'daki gibi hesaplanır.

$$DCG@k = \sum_{r=1}^{k} \frac{2^{dr} - 1}{log_2(r+1)} \qquad \text{Denklem 3.6}$$

Ancak bu formül farklı sorguların karşılaştırılması için yeterli/uygun değildir; Denklem 3.6'dan görüleceği gibi DCG@k ile 1'den büyük bir sayı elde edilmesi mümkündür. DCG@k'nın alabileceği değerin büyüklüğü sorgu ile alakalı olan belgelerin çok olması ile doğru orantılıdır. Ancak bilindiği gibi her sorgunun belirli bir derlem içinde erişebileceği alakalı belgeler kümesi oldukça farklılık gösterir. Sonuç olarak değişken sayıda alakalı belge kümesine sahip sorgulara karşılık gelen belge listelerinde bu tür bir hesaplama kıyaslanabilir olmamaktadır. Bu eksiklik DCG'nin normalleştirilmiş hali olan nDCG ile giderilmiştir:

$$nDCG@k = \frac{DCG@k}{IDCG@k} \qquad \text{Denklem 3.7}$$

Denklem 3.7'de gösterilen IDCG@k, ideal durumda elde edilecek hesaplamayı gösterir. Diğer bir deyişle, ilk *k* belgenin alaka düzeyleri açısından oluşturulabilecek en iyi sırada olması ideal durumdur ve böyle bir listedeki DCG@k değeri ise ideal olan IDCG@k'ya eşittir.

nDCG ölçütü çoklu derecelendirmeye uygun ve erişilen belgelerin sıralarını dikkate alıyor olsa da ERR'nin kullanıcı modeline daha uygun olması sebebi ile daha doğru sonuçlar verdiği öne sürülmektedir (Chapelle et. al., 2009). ERR'ın yarattığı bu farklılık "kullanıcıların alakalı-alakasız yargılarının değişken olduğu" ve "kullanıcıların ilk alakalı yargısına ulaştıkları belgeye kadar erişilen belgelere sırayla baktığı" varsayımlarını hesaba katmasından kaynaklanmaktadır.

Erişilen belge listesindeki r^{ninci} belgenin istekte bulunan kullanıcı açısında alakalı olma ihtimali R_r, *dr* derecesine sahip bu belge için şu de hesaplanır:

$$R_r = \frac{2^{dr} - 1}{2^d} \qquad \text{Denklem 3.8}$$

Diğer bir durumda ise; erişilen belge listesindeki r^{ninci} belgenin istekte bulunan kullanıcı açısında alakasız bulunma ihtimali $1\text{-}R_r$ olmaktadır. İlk k belge için bu iki durumu dikkate alan ERR@k'nın hesaplama yöntemi Denklem 3.9'da verilmiştir.

$$ERR@k = \sum_{r=1}^{k} \frac{1}{r} \left[\prod_{i=1}^{r-1} (1 - R_i) \cdot R_r \right] \qquad \text{Denklem 3.9}$$

Örnek 3.1'te verilmiş problemin çoklu derecelendirilmiş halinin nDCG@10 ve ERR@10 değerlerinin hesaplanması Örnek 3.2'te gösterilmiştir.

Örnek 3.2: 0 ile 3 arasında bir alaka derecelendirmesi kullanıldığında, belge kümesinde konuyla alakalı 3 belge {D1, D4-, D7} ve alaka dereceleri {d1=2, d4=1, d7=3} varken; bir sistem tarafından erişilen 10 belge sırasıyla {D1, D2, D3, D4, D5, D6, D7, D8, D9, D10} olmuştur.

DCG@10 hesaplanmasında Denklem 3.6'dan görüleceği gibi alakasız olan belgeler için toplamadaki bileşenler 0 olmaktadır. Bu durumda DCG@10 sadece r = 1, 4, 7'nin toplamı olacaktır:

$$DCG@10 = \sum_{r=1}^{10} \frac{2^{dr} - 1}{\log_2(r+1)} = \sum^{r=1,4,7} \frac{2^{dr} - 1}{\log_2(r+1)} \cong 17{,}82$$

Bu problem için ideal olan sıralamaları {D7(d1=3),D1(d2=2),D4(d3=1), ...} bu de gösterebiliriz. Diğer belgeler alakasız olduğundan hangi sırayla erişildiğinin önemi yoktur. Böylece IDCG@10 sadece r = 1, 2, 3'teki bileşenlerin toplamı olacaktır.

$$IDCG@10 = \sum_{r=1}^{10} \frac{2^{dr} - 1}{\log_2(r+1)} = \sum^{r=1,2,3} \frac{2^{dr} - 1}{\log_2(r+1)} \cong 27{,}99$$

Böylece $nDCG@10 \cong 17{,}82 \div 27{,}99 \cong 0{,}637$ olmaktadır. ERR@10 hesaplaması içinse kullanılacak her pozisyondaki belgelerin alakalı/alakasız bulunma ihtimalleri Çizelge 3.2'de verilmiştir.

Çizelge 3.2 Örnek 3.2 için belge pozisyonlarına göre alakalı/alakasız bulunma ihtimalleri

Veri Tipi	Sıra									
	1	2	3	4	5	6	7	8	9	10
Belgeler	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Alaka Dereceleri	2	0	0	1	0	0	3	0	0	0
R_r	0,375	0,000	0,000	0,125	0,000	0,000	0,875	0,000	0,000	0,000
$1\text{-}R_r$	0,625	1,000	1,000	0,875	1,000	1,000	0,125	1,000	1,000	1,000

ERR@10 hesaplanmasında Denklem 3.9'dan görüleceği gibi alakasız olan belgeler için toplamadaki bileşenler 0 olmaktadır (R_r = 0). Bu durumda ERR@10 sadece r=1,4,7'nin toplamı olacaktır.

$$ERR@10 = \sum^{r=1,4,7} \frac{1}{r}\left[\prod_{i=1}^{r-1}(1-R_i)\cdot R_r\right]$$

$$= \frac{0{,}375}{1} + \frac{(0{,}625\cdot 1\cdot 1)\cdot 0{,}125}{4} + \frac{(0{,}625\cdot 1\cdot 1\cdot 0{,}875\cdot 1\cdot 1)\cdot 0{,}875}{7} \cong 0{,}463$$

3.1.6 Diğer başarım ölçütleri

Bilgi erişim sürecinde belirli bir konunun/sorgunun farklı algılamaları olabilmektedir. Böyle bir durumda farklı algılamalar; yani alt konulara olan anlamsal yakınlığı hesabının belgenin alaka yargısında kullanılabilmesi mümkündür. Bu alt konulara belirli olasılıklar ile tahmin ederek hesaplayan bilgi erişim başarım ölçütleri bulunmaktadır. Clark et al (2008) tarafından tanımlanan α-nDCG ile Agrawal et al (2009) çalışmalarını temel alan duyarlık ölçütünün "niyet duyarlı" (İng. intent aware") uyarlaması bu tip ihtiyaçlar için kullanılabilmektedir. Niyet duyarlı duyarlık ölçütü için raporun ilerleyen kısımlarında Prec-IA ismi kullanılmıştır.

Bilgi erişim başarım ölçütlerinin diğer bir sınıfı ise tahmini ölçütlerdir. Tahmine dayalı başarım ölçütleri yargılama sürecinin belirli bir sorguya karşılık erişilen belgelerin alakalı/alakasız yargısının tam olarak yapılmadığı durumlarda kullanılmaktadır. Böyle durumlar temel olarak alakalı/alakasız saptanması yapılacak belgelerin seçiminde kullanılan modelden kaynaklanmaktadır. Kısacası, bazı özel durumlarda tüm erişilen belgeleri alaka açısından incelemek güçtür ve bazı yöntemler kullanılarak incelenecek belgelerin seçimi sağlanmaktadır.

"En düşük test topluluğu" (İng. Minimal Test Collection, kıs. MTC) (Carterette et al, 2006), ve de rastsal örneklemeye dayanan istatistiksel averaj duyarlık (kıs. statAP) (Aslam et al, 2006, 2007) olarak adlandırılan yöntemler belge seçme işlemi için kullanılmaktadır. Önceki bölümlerde tanımlanan MAP, R-P, P@k ve vb. gibi ölçütlerin verilen formülleri alaka yargısının tam olarak yapılması durumunda geçerlidir. İlgili belge seçim modelinin kullanılması durumunda bu değerlerin tahmini ölçütleri kullanılır. Tahmini bu ölçütlerin açık formülleri MTC için Carterette et al.'ın (2006) ve statAP için Aslam et al.'ın (2006, 2007) çalışmalarında mevcuttur.

Son olarak, *ikili tercih* (İng. binary preference, kıs. bpref) (Croft et al., 2009) başarım ölçütü ise kapsamdaki izlerde NIST tarafından seçilmediğinden dolayı başarım değerlendirilmesinde kullanılmamıştır.

3.2 TREC Materyalleri

TREC (Text Retrieval Conference) çalıştayları NIST (İng. National Institute of Standards and Technology) ve A.B.D. Savunma Bakanlığı desteğinde 1992 yılından beri düzenlenmektedir. Bu çalıştayların amacı, yazılı belge derlemlerinde bilgi erişim yöntemlerinin büyük ölçekli değerlendirildiği bir alt-yapı sağlamaktır. Yazılı belge erişim sahasındaki araştırmaları hızlandırma hedefiyle, TREC kapsamında büyük hacimli yazılı belge derlemleri ve standart bir erişim değerlendirme platformu oluşturulmaktadır. Her yıl düzenlenen TREC kapsamındaki bilgi erişim çalıştaylarına, farklı alt uğraş dallarında onlarca üniversite (MIT, RMIT, Carnegie Mellon, Massachusetts, Wisconsin, Glasgow, Waterloo, Amsterdam, Geneva, Paris-Sud, Fudan, Tokyo, Chinese Academy of Science, Beijing, Hong Kong Polytechnic, Meiji, vs.) ve pek cok araştırma enstitüsü (National Security Agency, IBM Research, Microsoft Research, Sabir Research, vs.) kendi geliştirdikleri bilgi erişim sistemleri ile katılırlar; bu sebeple TREC derlemleri bilgi erişim sahasının standart derlemleri olarak kabul edilirler (National Institute of Standards and Technology, 2010).

Geleneksel bilgi erişimi, yani belirli bir belge derleminde verilen herhangi bir sorguya karşılık alaka sırasına göre belgelere erişim, anlık-sorgu erişimi (İng. adhoc retrieval) olarak adlandırılır. Ancak 'bilgi erişim' ihtiyacı uygulamada farklılıklar gösterir ve sistemlerden beklenen işlevleri 'erişim görevi' (İng. retrieval tasks) tanımıyla çeşitlenir. Yukarıda açıklanan anlık-sorgu erişim görevinden başka alan yazında temel olan diğer bir görev de yönlendirme görevidir (İng. routing task). Yönlendirme görevi sabit sorgular kullanarak yeni

belge derlemlerinde arama yapan sistemlerin performanslarını inceler. İlk iki TREC (Harman, 1992, 1993) yalnızca bu iki temel görevi içermektedir. TREC-1 ve TREC-2'den oluşturulan kuramsal alt-yapının farklı görevler içinde uygun olacağından hareketle, TREC-3'ten (Harman, 1994) itibaren yeni erişim görevleri –etkileşimli görev (İng. interactive task) ve İspanyolca görevi (İng. Spanish task) eklenmiş- oluşturulmaya başlanmıştır. Tanımlanan yeni görevler anlık sorgu ve yönlendirme erişim görevlerinin alt-problemleri olup, bu görevlere odaklanmış alanların her biri "iz" (İng. track) başlığı altında toplanmıştır.

TREC'lerde 2007 yılına kadar toplam 27 farklı iz (Voorhees, 2007) oluşturulmuştur. Her bir TREC izi belirli bir yapıya uygun olarak ilerler. NIST tarafından ilk olarak ilgili izde kullanılacak derlem veya derlemdeki yazılı belge kümesi, sorguları içeren "konular" ve erişim değerlendirme yöntemi belirlenir. TREC derlemlerindeki belgeler ve oluşturulan konular SGML/XML (İng. Standard Generalized Markup Language/Extensible Markup Language) kullanılarak etiketlenmişlerdir. Ayrıca derlemlerdeki belgelerin ve konuların yapısı (etiketli alanlar), sayısı ve boyutu ilgili ize göre değişmektedir. Konu içindeki etiketlenmiş alanlardan hangilerinin sorgulama da kullanılacağı da önceden belirtilir. Erişim değerlendirme yöntemi ise alaka yargısının oluşturulmasına yönelik sistematiği ve başarımı gösterecek değerlendirme ölçütlerini içerir. Katılımcılardan ilgili TREC derleminde verilen sorgulara karşılık sistemlerinin yürütümünden (İng. run) elde ettikleri sonuçları –genellikle ilk 1000 alakalı belgeyi- sunmaları istenir; NIST tarafından sunulan talimatlara göre birden fazla yürütüm sonuçlarını da sunabilmektedirler. Daha sonra NIST oluşturduğu alaka yargılamasına göre alakalı belgeleri işaretler, belirlenen değerlendirme ölçütlerine göre başarımlarını hesaplar ve ilan eder (National Institute of Standards and Technology, 2010).

Deneylerde kullanılan TREC-6, 7 ve 8 anlık sorgu izleri ve materyalleri ile aktif olarak karılım gerçekleştirilen TREC-2009 ve TREC-2010 izleri ve materyalleri ilerleyen bölümlerde anlatılmıştır.

3.2.1 TREC-6, TREC-7 ve TREC-8 anlık sorgu izleri ve materyalleri

Anlık-sorgu izi başlangıçtan itibaren aralıksız olarak TREC-9'a kadar TREC çalıştaylarına dahil edilmiştir. Bu izlerde kullanılan TREC derlemi her biri yaklaşık 1 GB boyutundaki beş ayrı diskte toplanmıştır. Derleme ait istatistikler Çizelge 3.3'de verilmiştir.

Çizelge 3.3 TREC anlık-sorgu izi derlemi özellikleri.

Disk No.	Derlem Tanımı	Boyut (MB)	Belge Sayısı	Ortalama Kelime Sayısı
1	Wall Street Journal, 1987-1989	267	98.732	434,0
	Associated Press , 1989	254	84.678	473,9
	Computer Select makaleler, Ziff-Davis	242	75.180	473,0
	Federal Register , 1989	260	25.960	1315,9
	Abstracts of U.S. DOE publications	184	226.087	120,4
2	Wall Street Journal, 1990-1992	242	74.520	508,4
	Associated Press , 1988	237	79.919	468,7
	Computer Select makaleleri, Ziff-Davis	175	56.920	451,9
	Federal Register , 1988	209	19.860	1378,1
3	San Jose Mercury News,1991	287	90.257	453,0
	Associated Press newswire,1990	237	78.321	478,0
	Computer Select makaleleri, Ziff-Davis	345	161.021	295,4
	U.S. Patent, 1993	243	6.711	5391,0
4	Financial Times, 1991-1994 (FT)	564	210.158	412,7
	Federal Register, 1994 (FR94)	395	55.630	644,7
	Kongre Kayıtları, 1993 (CR)	235	27.922	1373,5
5	Foreign Broadcast Information Service (FBIS)	470	130.471	543,6
	Los Angeles Times (LA)	475	131.896	526,5

Belgeler SGML ile etiketlenmiş olup, örnek olarak Disk-5'teki FBIS belgelerinden biri Ek-1'de bulunmaktadır. Ayrıca gerçekleştirilen tüm anlık-sorgu izleri için toplam 450 tane sorgu hazırlanmış ve bunlar hazırlanış sırasına göre numaralandırılmışlardır.

Deneyler için öngörülen TREC-6 (Voorhees and Harman, 1997) anlık-sorgu izinde TREC derlemindeki disk-4 ve disk-5, TREC-7,8 (Voorhees and Harman, 1998, 1999) anlık sorgu izlerinde ise Kongre Kayıtları(Bkz. Disk No 5) dışındaki disk-4 ve disk-5'teki belgeler kullanılmaktadır. Bunun yanında TREC-6,7 ve 8 için sırasıyla 301-350, 351-400, ve 401-450 numaraları arasındaki 50'şer konu kullanılmıştır. TREC-6 anlık sorgu izinde kullanılan 301 numaralı konu örnek olarak Şekil 3.2'de gösterilmiştir.

Konu; <title> ile etiketlenen TITLE, <desc> ile etiketlenen DESCRIPTION ve <narr> ile etiketlenen NARRATIVE olmak üzere üç alandan oluşmaktadır. TITLE konuyu en iyi tanımlayan en fazla üç kelimeyi içerir. DESCRIPTION bilgi ihtiyacını tanımlayan tek bir cümleden oluşurken, NARRATIVE ise alaka yargısında bulunacak hakemin belgeyi alakalı/alakasız olarak işaretlemesinde kullanacağı kıstasları belirtmektedir.

Sistemler, konu'yu oluşturan alanlara göre 3 tipte sorgu için yürütümler yapabilirler; ilki "çok kısa" olarak adlandırılıp sadece TITLE alanını, ikincisi

"kısa" olarak adlandırılıp TITLE+DESCRIPTION alanlarını ve sonuncusu da "tüm konu" olarak adlandırılıp TITLE+DESCRIPTION+NARRATIVE alanlarını sorgu dizgesi olarak kullanır. Her sorgu tipi için bir katılımcının en fazla 2 tane yürütüm sonucu sunmasına izin verilir.

```
<top>

<num> Number: 301
<title> International Organized Crime

<desc> Description:
Identify organizations that participate in international criminal
activity, the activity, and, if possible, collaborating organizations
and the countries involved.

<narr> Narrative:
A relevant document must as a minimum identify the organization and the
type of illegal activity (e.g., Columbian cartel exporting cocaine).
Vague references to international drug trade without identification of
the organization(s) involved would not be relevant.

</top>
```

Şekil 3.2 301 numaralı konu.

Anlık sorgu izin de başarım gösterimi için değerlendirme ölçütleri olarak erişilen alakalı belge sayısı , ortalama averaj duyarlık ve R-Duyarlık kullanılır.

TREC-6,7 ve 8 anlık sorgu izlerinde kullanılan TREC derleminin 4. ve 5. diskleri yürütmekte olduğumuz TÜBİTAK projesi (Dinçer, 2005) kapsamında NIST'ten ücret karşılığı alınmıştır. İlgili konular ve konulara karşılık gelen belgelerin alaka yargıları da NIST'ten temin edilmiştir.

3.2.2 TREC 2009 ve TREC 2010 izleri

TREC'lerde 2009 yılına kadar toplam otuz farklı iz oluşturulmakla birlikte; blog izi (İng. blog track), kimyasal BE izi (İng. chemical IR track), varlık izi (İng. entity track), yasal iz (İng. legal Track), milyon sorgu izi (İng. million query track), alaka geri-bildirim izi (İng. relevance feedback track) ve web izi olmak üzere toplam yedi alan 2009 yılında düzenlenen konferans kapsamında yer almıştır. İlk defa bu konferansta oluşturulan kimyasal BE ve varlık izleri ile en son 2003 yılında kapsama alınan web izi dışında kalan diğer dört iz 2008 konferansında da yer almıştır.

TREC-2009'daki diğer bir yenilikte web'i daha büyük boyutlu yansıtan ClueWeb09 (Carnegie Mellon University, 2009) derleminin oluşturulmuş

olmasıdır. Web, milyon sorgu, varlık ve alaka geri-bildirim izlerinde kullanılan bu derlemin tümü ,Kategori-A , 10 farklı dilde yaklaşık bir milyar web sayfasından oluşurken, İngilizce ilk 50 milyon belge Kategori-B adı altında toplanmıştır. ClueWeb09 derleminin Kategori-A ve Kategori-B olarak genel istatistikleri Çizelge 3.4'te verilmiştir.

Çizelge 3.4 ClueWeb derlemi Kategori-A ve Kategori-B istatistikleri

Gözlenen Bilgi	Kategori-A	Kategori-B
Belge sayısı	1.040.809.705	50.000.000
Sıkıştırılmamış boyutu	25 TB	> 1TB
Benzersiz URL bağlantı sayısı	4.780.950.903	428.136.613

TREC-2009 ve TREC-2010 kapsamında yer aldığımız; yani yürütümler sunduğumuz web ve milyon sorgu izlerinde: tanımlı görevler, kullanılan derlem, sorgular ve değerlendirme ölçütleri gibi özellikleri bakımından daha ayrıntılı biçimde ilerleyen bölümlerde anlatılmıştır.

3.2.2.1 TREC-2009 milyon sorgu izi

TREC çalıştaylarına 2009 yılında üçüncü kez dahil edilen milyon sorgu izi iki amaç doğrultusunda tasarlanmıştır: (1) büyük belge ve sorgu kümelerinde anlık-sorgu erişiminin araştırılması; (2) sistemlerin değerlendirilmesi için üstünkörü yargılama yapılan çok sayıda sorgu mu, ya da eksiksiz yapılan az sayıda sorgu mu kullanılması daha iyi sorusuna cevap aranması.

TREC-2009'da milyon sorgu izi (Carterette et al, 2009) için kullanılması uygun olan sorgu sayısı 10000'den 40000'e çıkartılmıştır. Ayrıca bu 40000 sorgunun her birine 1'den 4'e kadar olan bir öncelik numarası atanmıştır. Birinci önceliğe sahip 1000 sorgu bulunmaktadır. Katılan gruplardan sistemlerini en az 1000 sorguda çalıştırmaları beklenmektedir. Gruplar bu sistem yürütümlerinden elde ettikleri ilgili sorgu ve bu sorgudaki ilk 1000 sıralı belgeyi NIST'e yollarlar. NIST kendine sunulan yürütüm sonuçlarından alakalı/alakasız yargılama sürecine geçmeden önce yargılama için kullanılacak sorguları seçer. Bu seçim ile kullanılmaya müsait sorgular (maksimum 40000 sorgu) arasından rastlantısal bir

yöntemle daha küçük çaplı bir sorgu kümesi oluşturularak sistemlerin daha kolay ve çabuk bir de değerlendirebilmesine imkan sağlanmaktadır.

TREC-2009 için yeni sorgu kümesi 687 sorgu/konu içermektedir ve normalde değişik TREC izlerinde kullanılan kümenin 50 konudan oluştuğu düşünülecek olursa bu haliyle bile oldukça büyüktür. Bu nedenle yargılama sürecinde yürütüm listelerinin toplandığı havuzdaki tüm belgelerin alakalı/alakasız tespitinin yapılması oldukça güçtür. Havuzdaki belgelerden işleme alınacak belgelerde ayrıca seçilmektedir. Yargılama sürecinde alakalı/alakasız saptanması yapılacak belgelerin seçimi için ise "en düşük test topluluğu" (Carterette et al, 2006), ve de rastsal örneklemeye dayanan istatistiksel averaj duyarlık (Aslam et al, 2006, 2007) yöntemlerinin ikisi de kullanılmıştır.

3.2.2.2 TREC-2009 ve TREC-2010 web izleri

Web izi genel olarak web erişim teknolojilerinin araştırılması ve değerlendirilmesi hedefiyle oluşturulmuştur. TREC-2009 web izi geleneksel anlık-sorgu görevi ve yeni tanımlanan çeşitlilik (İng. diversity) görevi olmak üzere iki değişik görevi kapsamaktadır (Clarke et al, 2009). Çeşitlilik görevi ile erişilen belge listesinin sorguyu tam olarak kapsamasıyla birlikte aşırı tekrarlardan kaçınılması hedeflenmektedir. Yani, bir sorgu ve konuya karşılık erişilen belgelerin oluşturduğu listede konunun/sorgunun değişik bilgi ihtiyaçlarına cevap verebilen belgelerin bulunması istenmektedir.TREC-2010 web izinde ise anlık-sorgu görevi ve çeşitlilik görevine ek olarak ve 'spam görevi' adı altında üçüncü bir görev tanımlanmıştır (Clarke et. al., 2010).

TREC-2009 ve 2010 web izlerindeki tüm görevler içinde ClueWeb09'un Kategori-A ve Kategori-B derlemlerinde iki farklı değerlendirme yapılmaktadır. 2009 yılındaki TREC web izine toplam 26 grup 119 yürütümle ve 2010 yılındaki TREC web izine ise toplam 23 grup 92 yürütümle katılmıştır. İzlerdeki tüm görevlerde farklı derlemler için sunulan yürütümlerin dağılımı ve katılan grupların sayısı Çizelge 3.5'te gösterilmektedir.

Çizelge 3.5 Web izleri görev ve derleme göre katılım istatistikleri

TREC	Görev Türü	Kategori-A Grup (Yürütüm)	Kategori-B Grup (Yürütüm)	Toplam Grup (Yütütüm)
TREC-2009	Anlık-sorgu	*13 (37)*	*14 (34)*	*25 (71)*
	Çeşitlilik	*10(26)*	*10(22)*	*18 (48)*
TREC-2010	Anlık-sorgu	*11 (29)*	*11 (26)*	*20 (55)*
	Çeşitlilik	*9 (22)*	*5 (10)*	*12 (32)*
	Spam	*3 (5)*	*0 (0)*	*3 (5)*

NIST bu iz için her yıl elli (50) yeni konu oluşturmaktadır. Örnek bir sorgu Şekil 3.2'de verilmiştir. Konular *muğlak yani birden fazla "farklı" yoruma sahip* (yukarıdaki konu örneğinde "ambiguous" olarak belirtiyor) ve *birbirleriyle bağlantılı "benzer" alt konulara sahip* (*type="faceted"* biçiminde işaretleniyor) olmak üzere iki tipte olmaktadır. *"query"* alanında verilen sorgunun *"description"* alanında tanımı yapılmaktadır.

```
<topic number="19" type="ambiguous">
<query>the current</query>
      <description>
            I'm looking for the homepage of The
            Current, a program on Minnesota Public Radio.
      </description>
      <subtopic number="1" type="nav">
            Take me to the homepage of The Current, a program on Minnesota
            Public Radio.
      </subtopic>
      <subtopic number="2" type="nav">
            I'm looking for the homepage of The Current newspaper in New
            Jersey.
      </subtopic>
      <subtopic number="3" type="nav">
            I want to find the homepage of The Current newspaper in Hartford.
      </subtopic>
      <subtopic number="4" type="nav">
            I want to find the homepage of The Current magazine in San
            Antonio.
      </subtopic>
</topic>
```

Şekil 3.2 Web izi konu örneği

Çeşitlilik görevinde kullanılmak üzere hazırlanan alt konular, “subtopic” alanında belirtilmiş olup “nav” (İng. Navigational’ın kısaltması) ve “inf” (İng. informational’ın kısaltması) etiketleriyle gösterilen 2 tipte olabilmektedir. Navigasyonel tipteki alt sorgular az sayıda –genellikle 1– alakalı belgeye sahiptirler: belirli bir web sayfasının bulunmasının istenmesi şeklinde varsayılabilir. Enformasyonel tipteki alt sorgular ise çok sayıda alakalı belgeye sahiptir: bilginin kaynağından öte içeriğine odaklanılmış olduğu düşünülebilir. TREC-2009 web izi için oluşturulan elli konunun sahip olduğu alt konuların sayısı üç ila sekiz arasında değişmekle birlikte ortalama olarak konu başına 4.9 alt konu düşmektedir.

Bir belgenin sorguya göre alakalı/alakasız olarak yargılanması işlemi: çeşitlilik görevi için alt-konularda istenilen bilgilere uygunluğuna bakılarak yapılırken, anlık-sorgu görevinde ise öncelikle tanım alanı dikkate alınmak üzere alt-konulara da ek olarak bakılır. Tüm konuyla alakalı olan bir belge, hiçbir alt konuya uygun olmayabilir; yani anlık-sorgu görevinde alakalı olarak işaretlenen bir belge, çeşitlilik görevi için hiçbir alt konuda açıklanan duruma uymayabilir ve alakasız yargısına varılabilinir.

Gruplar her yürütüm ile elde ettikleri sıralı ilk 1000 belgeyi NIST’e sunarlar. Her bir gruba her bir görev için en fazla üç yürütüm sonucu sunma hakkı verilmiştir. Yürütümlerin NIST tarafından alakalı/alakasız yargısının verilmesi sırasında anlık-sorgu görevi için dört seçenekli bir ilişkilendirilme/işaretleme gerçekleştirilir. Belgenin alaka yargısı olan bu seçenekler şu şekildedir: (1) alakalı değil, (2) alakalı değil ama makul, (3) alakalı, (4) çok alakalı. Alakalı değil ama makul ile tanım alanındaki bilgiyle alakalı olmadığı halde makul olan farklı bir yorumlama ile alakalı olduğu belirtilmektedir.

Görev beklentilerindeki farklılığa paralel olarak bu iki görev için değişik başarım değerlendirme ölçütleri kullanılmaktadır. TREC-2009 web izi anlık-sorgu görevinde belgelerin alakalı/alakasız yargılarının yapılması sırasında “en düşük test topluluğu” yöntemi kullanılmıştır. Bu nedende dolayı başarım ölçütleri MTC ile tahmin edilen ölçütlerdir. TREC-2009’un aksine 2010 yılındaki TREC’te yürütümler sonucu elde edilen belge kümesinin tamamında alakalı/alakasız yargısı yapılmıştır. Bu sebepten dolayı yürütümler kesin olan başarım ölçütleri ile değerlendirilmiştir. Çeşitlilik görevinde ise α-nDCG ile Prec-IA başarım ölçütleri kullanılmıştır.

3.3 TERRIER (TERabyte RetrIEveR) Kütüphanesi

TERRIER (University of Glasgow, 2010), Glasgow Üniversitesi Bilgisayar Bilimleri bölümünde yazılı belge erişimi için geliştirilen etkin ve verimli bir arama motorudur. Java programlama diliyle yazılmış olan TERRIER'in açık kaynak kodlu sürümü ise büyük ölçekli yazılı belge derlemleri erişiminde yapılacak araştırmalar ve deneyler için kapsamlı bir platform oluşturmaktadır. TERRIER kütüphanesinde modüler olarak gerçeklenen, indeksleme ve erişim fonksiyonları, yeni kavramların/fikirlerin kolay ve hızlı bir şekilde geliştirilmesini ve değerlendirilmesini sağlamaktadır. Ayrıca TERRIER'de TREC anlık-sorgu izinde belirlenen formata uygun olarak indeksleme, erişim ve değerlendirme yapmaya imkan veren uygulamalar gerçeklenmiştir.

Kütüphanede bulunan indeksleme, erişim ve değerlendirme modüllerinin bileşenleri "terrier.properties" adlı bir özellik dosyası ile konfigüre edilebilmektedir; gövdeleyici (İng. stemmer) veya durma kelimeleri listesi kullanılıp/kullanılmayacağı, kullanılacak indeksleyici/ağırlıklandırma modeli, işlenecek belge alanları (TREC derlemine özel), erişilen belge sayısı gibi bir çok bilgi, 'özellik' olarak tanımlanmıştır.

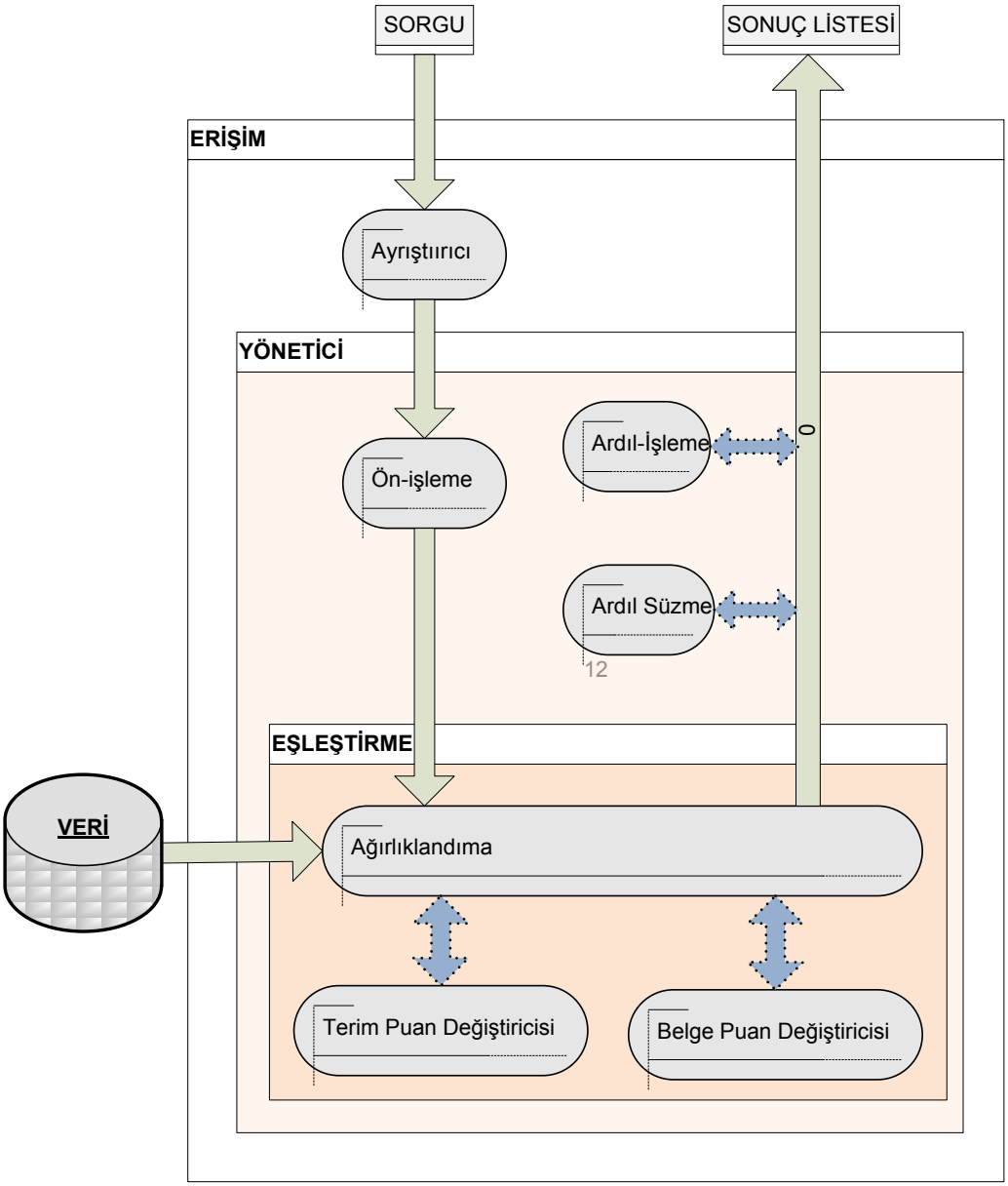
3.3.1 TERRIER içsel fonksiyonları

TERRIER kütüphanesinde indeksleme fonksiyonu için tek geçişli (İng. single-pass) ve klasik iki-geçişli (İng. two-pass) indeksleyiciler mevcuttur. İndeksleyiciler arasındaki temel fark ters indeks liste[3]'nin (İng. inverted index file) oluşturuluş biçimidir. Tek-geçişli indeksleyici belgeleri işleme aşamasında ters index listeyi oluştururken, iki-geçişli işleme safhasında doğrudan indeks listesi[4] (İng. direct index file) ve belge indeksi listesini[5] (İng. document index file) oluşturup ikinci aşamada bu listeleri kullanarak ters indeks listesini yaratır. İki-geçişli indeksleme doğrudan index listesini de oluşturduğu için sorgu genişletme (İng. query expansion) yöntemi kullanılmasını herhangi başka bir işleme gerek duymadan mümkün kılar. Ayrıca erişimde Ponte ve Croft'un (1998) dil modeli kullanılması halinde ihtiyaç duyulan ek göstergeler indeksleyicilerle (yapılandırmada belirtilmesi durumunda) yaratılabilir. İndekslemede bulunan terim-boruhattı (İng. term-pipeline) bileşeniyle terimlerin belirtilen işlemlere göre

[3]Ters indeks listesi temel olarak her terim için gözlendiği belge belirteçlerini tutar. Bu belgelerdeki gözlenme pozisyonları, görünme sıklıkları gibi göstergeler de ayrıca işaretlenir.
[4]Doğrudan indeks listesi temel olarak her belgenin içerdiği kelimeyi tutar.
[5]Belge indeksi listesi ise sabit uzunluktaki girişlerden oluşur ve belge bilgilerini tutar; bunlar toplam terim sayısı, belge belirteci ve doğrudan indeks listesindeki bağıl konumu gibi bilgilerdir.

değiştirilmesi veya kaldırılması mümkündür: gövdeleme/durma kelimeleri listesi kullanımı gibi işlemler.

TERRIER'in erişim fonksiyonunun temel bileşini tüm erişim sürecini yöneten *Manager* modülüdür (nesnesidir). *Manager* modülünün girdisi sorgu terimleri kümesi -önceden sorgunun bir ayrıştırıcı (İng parser) ile terimleri bulunur- çıktısı ise erişilen belge kümesi olup erişim sürecini dört aşamada/işlemde gerçekleştirir (Bkz. 3.3). 3.3'te *Manager* modülü "*Yönetici*" olarak adlandırılmıştır.



Şekil 3.3 TERRIER erişim fonksiyonları bileşenleri.

Erişim sürecinde ilk olarak, sorgu terimlerini terim-boruhattı ile bir ön-işlemden geçirerek terimleri indekslenmiş biçimlerine çevrilir. İkinci adımda, eşleştirme işlemiyle (kütüphanede Matching nesnesi olarak gerçeklenmiş) her belgenin puanını hesaplar ve erişilen sıralı belge listesini oluşturur. Eşleştirme aşamasında ilk olarak her terim için "erişimde tanımlanan ağırlıklandırma modeli" (WeightingModel nesnesi) ile bulunan puan, eğer tanımlanmış bir terim puan değiştiricisi (TermScoreModifier nesnesi) varsa değiştirilerek hesaplanır. Her belgenin puan hesaplanmasında ise vektör uzayı (İng. vector space) erişim modeli kullanmaktadır. Bu modele göre; belge içi terim ağırlıkları vektörü ile sorgu terim ağırlıkları vektörünün *nokta çarpımı* (İng. *dot product*) ile yakınlık; yani belge puanı ölçülür. En son olarak belge puanları tanımlı bir belge puanı değiştiricisi (DocumentScoreModifer nesnesi) varsa değiştirilip, belgeler puanlarına göre sıralanır. *Yönetici* modülüyle gerçekleşen üçüncü işlem ise ardıl-süzmedir. Bu aşama basit olarak belgenin sonuç listesine dahil edilip edilmeyeceğine izin vermektedir; kullanıcının belge erişim alanını kısıtlamak isteyebileceği etkileşimli uygulamalar için uygundur. Son olarak ise sonuç listesinin gerçeklenen herhangi bir yöntemle değiştirilebileceği ardıl-işlem (İng. post-processing) uygulanır. Örnek olarak bir sorgu genişletme yöntemi kullanarak sorguya yeni terimler eklenir ve eşleştirme işlemi yeniden gerçekleştirilir.

3.3.2 TERRIER ile gerçeklenmiş indeks terim ağırlıklandırma fonksiyonları/modelleri

Çeşitli terim ağırlıklandırma modellerinin kaynak kodları TERRIER kütüphanesinde <uk.ac.gla.terrier.matching.models> paketinde bulunmaktadır. Aynı zamanda Amati ve van Rijsbergen'in (2002) rastlantısallıktan sapma fikri temelinde bir DFR (Divergence From Randomness) çatısı gerçeklenmiştir. Bu çatı terim ağırlıklandırma modelini üç bileşene ayırmıştır; bileşenler rastlantısal oluş için ana model, ikincil etki normalizasyonu ve terim gözlenme sıklığı normalizasyonudur. TERRIER ana model olarak: Bose-Einstein modelini (simge. $B^{(1)}$), ters terim frekansı modelini (simge. IF), ters belge frekansı modelini (simge. In), ters beklenen belge frekansı modelini(simge. In_exp) ve Poisson modelini (simge. P) içermektedir. İkincil etki normalizasyonu Ponte ve Croft'un dil modelinde (1998) kullandığına benzer olan olasılıksal risk bileşenidir. Bu bileşen terimin enformasyon kazancının bir ölçüsüdür. Normalleştirmenin temelinde yatan mantık yüksek sıklıkta görünen bir terimin enformasyon içermeme riskinin minimal olması, ancak minimal riskin az bir enformasyon kazancı sağlamasıdır. Riskin olasılığı, yani normalleştirme değeri için kullanılan formüller Laplace

modeli (simge. L, Denklem 3.10) ve iki Bernoulli sürecinin oranıdır (simge. $B^{(2)}$, Denklem 3.11).

$$Prob_{risk} = \frac{1}{tf + 1}$$ Laplace (L) modeli Denklem 3.10

$$Prob_{risk}^{6} = \frac{TF}{df(tf + 1)}$$ İki binom dağılımının oranı ($B^{(2)}$) Denklem 3.11

Diğer normalleştirme ise terim gözlenme sıklığının *(tf)* normalleştirilmesidir. Bu normalleştirilmiş terim gözlenme sıklığı (*tfn*) belge uzunluğunun (*bu*) standart bir uzunluğu (*su*) göre normalizasyonu kullanılarak hesaplanır (**1** numaralı normalleştirme formülü, Denklem 3.12). Normalleştirme formülünün parametreli daha genel hali ise Denklem 3.13'te verilmiştir (**2** numaralı normalleştirme formülü) .

$$tfn = tf \times \log\left(1 + \frac{su}{bu}\right)$$ 1 numaralı normalleştirme Denklem 3.12

$$tfn = tf \times \log\left(1 + c \times \frac{su}{bu}\right)$$ 2 numaralı normalleştirme Denklem 3.13

DFR modelleri bu üç bileşeni ilklendirerek yani, ana rastlantısal oluş modelini seçerek; ilk normalizasyonu uygulayarak ve terim gözlenme sıklıklarını normalize ederek elde edilmektedir. Her bileşende kullanılan seçeneklerin belirteçleri birleştirerek DFR modelinin ismi oluşturulur. Örnek olarak "BB2" modeli: rastlantısal oluş modeli olarak Bose-Einstein'ı, ikincil etki

[6] *TF* terimin derlem genelindeki gözlenme sıklığını, *df* ise terimin geçtiği belgelerin sayısını göstermektedir.

normalizasyonu olarak iki Bernoulli sürecinin oranını ve terim gözlenme sıklığında 2 numaralı normalleştirme formülünün kullanıldığını göstermektedir.

TERRIER kütüphanesinde yer alan DFR tabanlı, TFxIDF şemasındaki ve melez[7] ağırlıklandırma modellerinden önemlileri Çizelge 3.6'da verilmiştir.

Çizelge 3.6 TERRIER kütüphanesindeki önemli ağırlıklandırma modelleri.

Ağılıklandırma Modeli	Açıklama
TF_IDF	Robertson'un (1994) TF ile Sparck Jones'un (1972) IDF'ini kullanır
LemurTF_IDF	Lemur sistemindeki TFxIDF versiyonu (Zhai, 2010)
BM25	Okapi BM25 (Robertson et al., 1999)
DFR_BM25	BM25'in DFR versiyonu
DFRee	Parametrik olmayan DFR versiyonu
PL2	DFR çatısında anlatıldı
IFB2	DFR çatısında anlatıldı
InL2	DFR çatısında anlatıldı
In_expB2	logaritmik işlemlerin 2 tabanında yapıldığı In_expB2
In_expC2	logaritmik işlemlerin e tabanında yapıldığı In_expB2
BB2	DFR çatısında anlatıldı

Bölüm 3.3.1'de anlatılan indeksleme ve erişim fonksiyonlarının yanı sıra, TREC'e özgü değerlendirme fonksiyonunu gerçekleştiren *TrecTerrier* uygulaması TERRIER kapsamında geliştirilmiştir. Bu uygulamanın da dahil olduğu TERRIER Kütüphanesi Java programlama diliyle geliştirme platformu olan Eclipse'te yazılmıştır. Bu nedenden dolayı, tez çalışması kapsamındaki terim ağırlıklandırma modellerinin kodlanması yine bu platformda yapılmıştır.

[7] Melez ağırlıklandırma modelleri olasılıksal dağılımlar ile TFxIDF şemasını birleştiren yaklaşımlarla oluşturulan ağırlıklandırma modelleridir.

4 GELİŞTİRİLEN İNDEKS TERİM AĞIRLIKLANDIRMA MODELLERİ

Bu bölümde, tez kapsamında geliştirilen indeks terim ağırlıklandırma modelleri anlatılmaktadır. Bu modeller ile ortaya konulan ağırlıklandırma fonksiyonları tek bir terim için ağırlık hesaplamasını sabit zamanda, *O(1)*, ve *m* tane terimden oluşan sorgu için ise belge puanı hesaplamasını doğrusal zamanda, *O(m)*, gerçekleştirmektedir.

4.1 Notasyon

Herhangi bir belge derlemi, Şekil 4.1'de verilen *Terim*×*Belge* matrisi **X** şeklinde ele alınabilir. Bu **X** matrisinde satırlar terimleri t_i (i = 1..r), sütunlar da belgeleri b_j (j = 1..c) temsil eder ve x_{ij} hücrelerinde t_i teriminin b_j belgesindeki gözlenme sıklığı vardır.

$$\boldsymbol{X} = \begin{array}{c|cccccccc} & \multicolumn{7}{c}{\textit{Belgeler}} & \\ \textit{Terimler} & b_1 & b_2 & b_3 & \dots & b_j & \dots & b_c & \textit{Toplam} \\ \hline t_1 & x_{11} & x_{12} & x_{13} & \dots & x_{1j} & \dots & x_{1c} & x_{1.} \\ t_2 & x_{21} & x_{22} & x_{23} & \dots & x_{2j} & \dots & x_{2c} & x_{2.} \\ t_3 & x_{31} & x_{32} & x_{33} & \dots & x_{3j} & \dots & x_{3c} & x_{3.} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ t_i & x_{i1} & x_{i2} & x_{i3} & \dots & x_{ij} & \dots & x_{ic} & x_{i.} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ t_r & x_{r1} & x_{r2} & x_{r3} & \dots & x_{rj} & \dots & x_{rc} & x_{r.} \\ \hline \textit{Toplam} & x_{.1} & x_{.2} & x_{.3} & \dots & x_{.j} & \dots & x_{.c} & x_{..} \end{array}$$

Şekil 4.1 *Terim*×*Belge* matrisi

Satır sonlarında '*Toplam*' başlıklı sütun altında görülen '$x_{i.}$' ifadesi t_i terimi için belgeler arası gözlenme sıklıklarının toplamı; sütun sonlarında yine 'Toplam' başlıklı satır içinde listelenen '$x_{.j}$' ifadesi de b_j belgesindeki terimlere ait toplam gözlenme sıklığıdır. deki '$x_{..}$' ise tüm belgelerdeki; yani derlem genelindeki toplam terim sayısını ifade etmektedir. Açıklanan gösterimlerin matematiksel eşitlikleri Denklem 4.1'de verilmiştir.

$$\text{(a)}x_{i.} = \sum_{j=1}^{c} x_{ij} \quad \text{(b)}x_{.j} = \sum_{i=1}^{r} x_{ij} \quad \text{(c)}x_{..} = \sum_{j=1}^{c}\sum_{i=1}^{r} x_{ij} \qquad \text{Denklem 4.1}$$

4.2 İstatistiksel Bağımsızlık Esasında İndeks Terim Ağırlıklandırma

4.2.1 İstatistiksel Bağımsızlık Fikri

Alanyazında şimdiye kadar yapılmış olan çalışmalardan, ideal bir terim ağırlıklandırma yönteminde hem içerik kelimelerin hem de işlev kelimelerin istatistiksel bir bakış açısından hesaba katılması gerektiğini anlamaktayız. Kelimeler ile belgeler arasındaki ilişkinin istatistiksel bir bakış açısından ele alınması için *istatistiksel bağımsızlık fikri* (kıs. İBF) uygun bir seçenektir.

İstatistiksel bağımsızlık fikri *şans oranı* (İng. odds ratio) fikri ile kolayca açıklanabilir. Eğer kelimelerin belgelerde taşınan enformasyona katkıları gözlenme sıklıklarıyla doğru orantılı kabul edilirse, x_{ij}/x_{kj} oranı, b_j belgesinde taşınan enformasyona t_i teriminin katkı yapması şansının t_k terimiyle yapılacak olan katkı nispetindeki ölçüsü olacaktır. Aynı şekilde x_{is}/x_{ks} oranı da, b_s belgesinde taşınan enformasyona t_i teriminin katkı yapması şansının t_k terimi ile yapılacak olan katkı nispetindeki ölçüsü olacaktır. Şans oranı bu iki şansın birbirine oranıdır:

$$\frac{x_{ij}/x_{kj}}{x_{is}/x_{ks}} \qquad \text{Denklem 4.2}$$

Bağımsızlık altında oran 1'e eşit olmalıdır. Bir başka söyleyişle, bağımsızlık altında t_i teriminin enformasyona katkı yapma şansı ile t_k teriminin enformasyona katkı yapma şansı belgeler arasında farklılık göstermeyecektir. Şans oranının kelimeler için kurgulanmış olan eşitlikçi şekilde belgeler içinde kurgulanabilir. x_{ij}/x_{is} oranı t_i teriminin b_j belgesinde taşınan enformasyona katkı yapma şansının b_s belgesinde taşınan enformasyona yapacağı katkı nispetindeki ölçüsüdür. Aynı şekilde x_{kj}/x_{ks} oranı, t_k teriminin b_j belgesinde taşınan enformasyona katkı yapma şansının b_s belgesinde taşınan enformasyona yapacağı katkı nispetindeki ölçüsüdür. Şans oranı yine bu iki şansın oranıdır:

$$\frac{x_{ij}/x_{is}}{x_{kj}/x_{ks}} \qquad \text{Denklem 4.3}$$

Bağımsızlık altında bu oran da 1'e eşit olacaktır. Bir başka söyleyişle, b_j belgesinde taşınan enformasyona katkı yapılması şansı ile b_s belgesinde taşınan enformasyona katkı yapılması şansı terimler arasında farklılık göstermeyecektir.

Şans oranlarıyla açıklanan bu eşitlikçi bağımsızlık fikri indeks terim ağırlıklandırma meselesine kolayca uyarlanabilir. Tanım gereği, belirli bir işlev kelimenin enformasyona yapacağı katkı miktarının bir başka işlev kelimenin yapacağı katkı miktarına oranı belgeler arasında farklılık göstermemelidir (Bkz. Denklem 4.2). Aynı şekilde, belirli bir belge de taşınan enformasyona yapılan katkı miktarının bir başka belge de taşınan enformasyona katkı miktarının oranı işlev kelimeler arasında farklılık göstermemelidir (Bkz. Denklem 4.3). Bu bağlamda, işlev kelimelerin belgelerdeki gözlenme sıklıklarının, bu kelimelerden bağımsızlık altında beklenecek gözlenme sıklıklarına yakın çıkması en olası durumdur. Dolayısı ile içerik kelimelerin de gözlendikleri belgelerde bağımsızlık altında beklenecek gözlenme sıklıklarından farklı çıkması en olası durum olmaktadır.

Kullanılan anlamıyla bağımsızlık aslında kelimeler ile belgeler arasında kategorik bir ilişki olmadığını belirtir. Daha doğrusu, herhangi bir doğal dilde kelimeler gözlendikleri belgelerden bağımsız olarak ancak ve ancak aralarında kategorik bir ilişki olmadığı takdirde kullanılabilir: *işlev kelimelerin içerikten bağımsız olarak dilbilgisi kuralları gereği metne eklenmesi gibi*. Eğer bağımsızlık yoksa; yani bazı kelimeler kasti biçimde belgelerde kullanılıyorsa, bu durum kelimeler ile belgeler arasında kategorik bir ilişki olduğuna dair delil olarak kabul edilebilir: *içerik kelimelerin belge içeriğini oluşturmak için yazar tarafından bilinçli şekilde metne eklenmesi gibi.*

4.2.2 İBF esasındaki modellerinin matematiksel ifadesi

Tanıtılan bu istatistiksel bağımsızlık fikrinin terim ağırlıklandırmada nicel bir ölçü kullanılabilmesi için öncelikle herhangi bir t_i terimi ve b_j belgesi için bağımsızlık altında gözlenme sıklığının beklenen değerinin hesaplanabilmesi gerekmektedir. Herhangi bir t_i teriminin bağımsızlık altında belirli bir b_j belgesinde gözlenme sıklığının beklenen değeri e_{ij}, Denklem 4.4'te verilen eşitlik kullanılarak hesaplanabilir.

$$e_{ij} = x_{i.} \cdot \frac{x_{.j}}{x_{..}} = \frac{x_{i.} \cdot x_{.j}}{x_{..}} \qquad \text{Denklem 4.4}$$

Denklemdeki $x_{.j} / x_{..}$ oranı, b_j belgesinin derlemdeki diğer belgelere göreceli ağırlığını verir. $x_{i.}$ değeri de t_i teriminin derlem genelindeki toplam görünme sıklığı olduğuna göre, e_{ij} değeri t_i teriminin derlem genelindeki toplam

görünme sıklığından bağımsızlık durumunda b_j belgesine düşecek payı vermektedir. Bir başka söyleyişle, bağımsızlık altında her terimin derlem genelindeki gözlenme sıklığı, derlemdeki tüm belgelere uzunlukları oranında paylaştırılmış olmalıdır. Sonuç olarak, bir t_i terimi için teriminin gözlenme sıklığının beklenen gözlenme sıklığından(e_{ij}) farkı, terimin b_j belgesinde taşınan enformasyona yaptığı katkının miktarı ile doğru orantılı olmalıdır. t_i teriminin b_j belgesi için önemi/ağırlığını gösteren bu ilişki, *bağımsızlıktan sapma* (İng. **D**ivergence **F**rom **I**ndependence, kısaca DFI) olarak adlandırılmış ve DFI_{ij} biçiminde Denklem 4.5'te nicel olarak gösterilmiştir.

$$DFI_{ij} \propto fark = x_{ij} - e_{ij} \qquad \text{Denklem 4.5}$$

Denklem 4.5'te görüldüğü gibi, her t_i teriminin b_j belgesi için görünme sıklığının beklenen görünme sıklığından farkı negatif ($fark < 0$) ve pozitif ($fark > 0$) değerler alabilmektedir. $fark = 0$ değeri t_i terimi ile b_j belgesi arasında kategorik bir ilişki olmadığını; yani bağımsız olduklarını gösterir. Tersi durum da; yani $fark \neq 0$ değerleri de t_i terim ile b_j belgesi arasında kategorik bir ilişki olduğunu bildirir. Ancak bu kategorik ilişki için bir de yön söz konusudur. Dolayısı ile $fark \neq 0$ değerleri terimin gözlendiği belgelere bağımlı şekilde dağılım gösterdiğini ortaya koymaktadır. $fark > 0$ durumu; yani pozitif kategorik ilişki terim ağırlıklandırma meselesinde önemli olandır.

Terim ağırlıklandırma ölçüsünün bu fark değerleriyle ifade edilebilmesi için terimlerin fark değerlerinin kıyaslanabilir bir hale dönüştürülmesi gerekmektedir. En basit olarak, değerlerin standart bir ölçüye getirilmesi için *fark*'ın beklenen değere oranı kullanılabilir. Bağımsızlıktan sapma esasındaki bu model Denklem 4.6'da verilen eşitlikteki terim ağırlıklandırma hesabı ile gerçekleşmektedir.

$$DFI_{ij} = \frac{x_{ij} - e_{ij}}{e_{ij}} \qquad \text{Denklem 4.6}$$

Büyük örneklemler için Pearson'ın Ki-Kare (İng. Chi-Square) istatistiği Denklem 4.7'de verilmiştir.

$$G^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} G_{ij}^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(x_{ij} - e_{ij})^2}{e_{ij}} \qquad \text{Denklem 4.7}$$

Pearson'ın Ki-Kare istatistiği, *(r-1)(c-1)* bağımsızlık dereceli $\chi 2$ dağılımıdır (Agresti, 2002). Bu istatistiğin doğal sonucu olarak, herhangi bir t_i (i = 1..r) terimi ve b_j (j = *1..c*) belgesi için bağımsızlıktan sapma miktarı Denklem 4.7'de gösterilen G_{ij}^2'nin karekökü olarak Denklem 4.8'deki gibi hesaplanır.

$$DFI_{ij} = \frac{fark}{\sqrt{e_{ij}}} = \frac{x_{ij} - e_{ij}}{\sqrt{e_{ij}}} \qquad \text{Denklem 4.8}$$

Denklem 4.8'den de görüldüğü üzere, bu ağırlıklandırma hesabı aslen Denklem 4.6'daki Bağımsızlıktan sapma modelindeki standardizasyon parametresini beklenen terim gözlenme sıklığının karekökü biçiminde ($\sqrt{e_{ij}}$) ele alan halidir.

Denklem 4.6 ve 4.8'deki temel modellerden hesaplanan değerleri logaritmik dönüşüm ile kullanmak da mevcut yöntemlerde sıkça rastlanan bir durumdur. Çünkü ana modellerle tahmin edilen terim sıralamaları doğru yapılmış olsa bile, terime atanan nicel değer doğru ifade edilmemişse 1'den fazla terim için (sorgu terimlerinin her birinin toplamı) elde edilecek toplam sıralama yanlış olacaktır. Sonuç olarak ana modellerin değerlerinin logaritmalarını alan, yani terimler arasındaki nicel mesafeleri azaltan modellerin hesaplamada kullandıkları eşitlikler denklem 4.9'da verilmiştir. TERRIER gibi erişim modeli olarak vektör uzayı kullanan sistemlerde belge için gözlenen terimlerin ağırlıklandırma ile hesaplanan puanlarının belge puanı hesabındaki etkileri toplama (+) biçiminde olmaktadır. Erişim modelini vektör uzayı olarak sabit tutup gerçek terim ağrılıklarının logaritmalarının kullanılması bir anlamda, belge puanının terim ağırlıklarının çarpımı olarak hesaplanmasını sağlar.

$$DFI_{ij} = \begin{cases} log_2\left(\frac{x_{ij} - e_{ij}}{e_{ij}} + 1\right) & \text{(a)} \\ log_2\left(\frac{x_{ij} - e_{ij}}{\sqrt{e_{ij}}} + 1\right) & \text{(b)} \end{cases} \qquad \text{Denklem 4.9}$$

Ortaya konulan bağımsızlıktan sapma modelleri ile herhangi bir terimin belge içinde içerik kelime olmasını puanlamaktadır. Bu bağlamda, modellerle hesaplanan terim ağırlıkları terimlerin belgeyi temsil etme gücü, yani alanyazında temsil edildiği şekliyle TF olarak ele alınabilinir. Bağımsızlıktan sapma fikri ile TFxIDF şeması birleştirildiğinde oluşturulan melez modellerin ağırlıklandırma fonksiyonları Denklem 4.10'da verilen eşitlikle hesaplama yaparlar.

$$DFI_{ij} \times IDF_i = \begin{cases} log_2\left(\frac{x_{ij} - e_{ij}}{e_{ij}} + 1\right) \times idf_i & \text{(a)} \\ log_2\left(\frac{x_{ij} - e_{ij}}{\sqrt{e_{ij}}} + 1\right) \times idf_i & \text{(b)} \end{cases} \qquad \text{Denklem 4.10}$$

Denklem 4.10'da gösterilen *idf*, alanyazında terimin belge ayırt ediciliği gücü değeri olarak Sparck Jones'un *idf*'si seçilmiştir; Denklem 2.2'de verildiği gibi $idf_i = log_2\left(\frac{n}{n_i} + 1\right)$ dir.

4.3 Luhn'un İddiası Esasında İndeks Terim Ağırlıklandırma

Luhn'un tarafından ortaya koyulan terim önemi ile terim frekansı arasındaki ilişki Bölüm 2'de detaylı biçimde anlatılmıştır. Aslen Luhn bu iddiasını metin özetleme probleminde cümle seçimi için koymuş olduğu halde, terim ağırlıklandırmasını terim frekansı kavramı üzerinden açıklayan ilk ve halen kabul edilen en önemli kuramsal alt yapıdır. Bu sebepten dolayı indeks terim ağırlıklandırma modelleri oluşturmak için *tam olarak* bilgi erişim sahasına aktarılması mümkündür. Temel olarak Luhn'un iddiasında söylenen veya ileri sürülen ilişki bilgi erişim sahası için TFxIDF şemasına dayalı terim ağırlıklandırma yöntemindeki TF bileşenine denk gelmektedir. Temel TFxIDF ağırlıklandırma şemasındaki TF bileşeni bir terimin belge içeriğine yaptığı katkıyı bu terimin frekans (*tf*) değerinden ölçmektedir. Bir başka deyişle, TF bileşeni belge içeriğine yapılan katkının belge-içi terim frekans (*tf*) değerlerine göre modellenmesidir.

Salton (1970) ve Minker et al. (1973) çalışmalarında TF bileşeninin terim ağırlıklandırma işleminde kullanılmasının, ağırlıklandırılmamış terimlere göre oldukça iyi erişim başarımı sağladığını deneysel olarak ortaya koymuştur. Bunun yanında Robertson et al. (1994) ve Sparck Jones et al. (2000) *TF* bileşeni için değişik hesaplama yöntemleri ileri sürmüşlerdir. Bu çalışmalar temel olarak Luhn'un iddiasına; yani terimin ağırlığı hesaplamasında terim frekansının dikkate alınması fikrinden ilham almasına rağmen, bu *TF* bileşenlerinin hesaplamalarının doğrudan *tf* değeriyle doğrusal olması Luhn'un iddiasındaki "terim önemi belirli bir orta frekans değerinden sonra her iki yönde azalmaktadır" saptamasına uygun değildir. Mevcut yöntemlerde genel olarak temel alınan bu (TF $\propto$ *tf*) varsayım, yüksek frekanslı kelimeleri/terimleri içeren durma kelimeleri listesi kullanılması ile bu tip önemsiz terimleri elenmesiyle belirli bir ölçüde Luhn'un iddiasına yaklaşmaktadır. Ancak iddiada ifade edilen temellerle arasında hala önemli bir boşluk bulunmaktadır.

Bir başka değişik ağırlıklandırma yaklaşımı da belge-içi ve belgeler-arası terim frekanslarını birleştirilmesi temeline dayalı terim ayırt etme değeridir (TDV). TDV'nin (Salton et al., 1976) temel özellikleri Salton ve Yang (1973) tarafından araştırılmış ve Luhn'un bakış açısına ek olarak her terimin derlem genelindeki frekans dağılımını da dikkate alan belirli sonuçlara ulaşılmıştır. Bir anlamda bu sonuçlar Luhn'un iddiasının bir uzantısı veya genişletilmiş hali olarak kabul görse de, aslen belge topluluğunun/derlemin sözlük kelimelerinin ayrıştırmasını hedeflemektedir. Özetle, TDV ölçütü IDF bileşeni ile yapılmaya çalışılan görevin benzerini yerine getirmektedir. Bu sebepten dolayı, Luhn'un *belge sınırları* içinde ortaya koyduğu iddiayla TDV'nin tam olarak aynı özelliklere sahip olduğu bir yargıya ulaşmak mümkün olmamaktadır.

Otomatik indeksleme/ağırlıklandırma için diğer bir yaklaşım da olasılıksal bakış açısını temel almaktadır. Bu yolu izleyen bir çok modelden önemlileri hakkındaki çalışmalar: Harter (1975a, 1975b); Robertson and Sparck-Jones (1976); Cooper and Maron (1978); Croft and Harper (1979); Robertson et al.(1980); Fuhr (1989); Turtle and Croft (1992); Wong and Yao (1995); Ponte and Croft (1998); Hiemstra and de Vries (2000); Amati and Van Rijsbergen (2002). Bu çalışmalar öncelikle dayandıkları modelin parametrelerinin tahminlenmesi ve bu parametrelerin uygun biçimde birleştirilerek ağırlıklandırma formulünün oluşturulmasıyla ilgilenmektedir. Bu yöntemler karmaşık olmalarının yanı sıra, Luhn'un bakış açısını dikkate aldıklarına dair bir sonuca varmak oldukça güçtür. Bu açıdan, Luhn'un iddiasının tam olarak kapsandığına dair herhangi bir kanıt bulunmamaktadır.

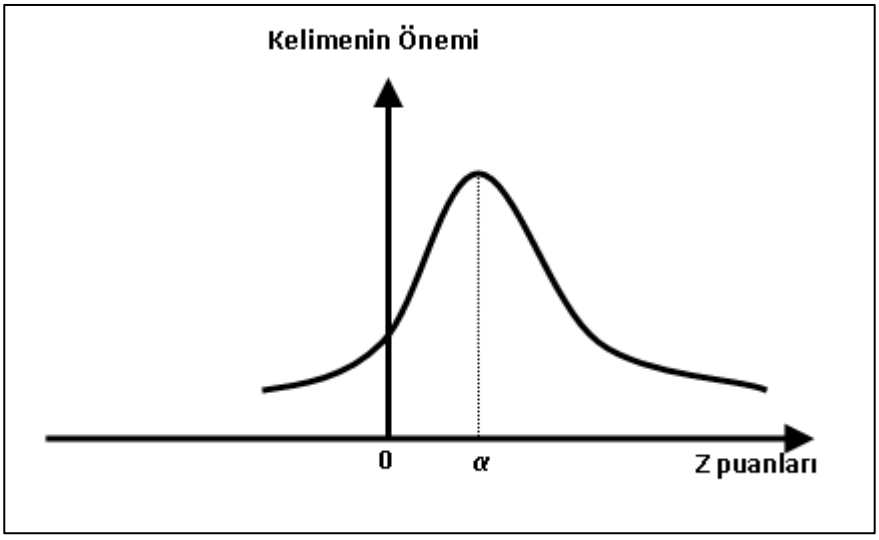
Yukarıda anlatılan analizleri özetleyecek olursak, genel olarak terimin anlamsal enformasyona yaptığı katkı miktarının belge içi terim frekansı açısından gösteren mevcut yöntemler Luhn'un bakış açısından bazı farklılıklar göstermektedir. Luhn'un bakış açısını yansıtan TF bileşenleri ve terim ağırlıklandırma formülleri takip eden bölümlerde anlatılmıştır.

4.3.1 Luhn'un iddiasına göre TF bileşeni ilişkileri

Luhn kelimelerin gözlenme sıklıklarını yalın hali ile değil, gözlendiği belgedeki ortalama kelime sıklığına göreceli olarak değerlendirmektedir. Luhn'un ortaya koyduğu niteliksel bakış açısını nicel olarak ölçmek amacıyla z puanlarından (İng. z scores) yararlanılabilinir (Bkz. Denklem 4.11).

$$z_{ij} = \frac{x_{ij} - \overline{x_j}}{\sqrt{s_j^2}} \qquad \text{Denklem 4.11}$$

Bu denklemde $\overline{x}_j$ ifadesi b_j belgesindeki ortalama kelime sıklığını ifade eder. Herhangi bir t_i teriminin b_j belgesindeki gözlenme sıklığı için hesaplanan z puanı, x_{ij} değerinin $\overline{x}_j$ belge ortalamasından belge içi standart sapması s_j nispetinde uzaklığı vermektedir. Bir başka söyleyişle, t_i teriminin b_j belgesindeki gözlenme sıklığının standartlaştırılmış halidir. Burada standartlaştırılmanın anlamı aynı ölçekte olmaktır. Dolayısı ile herhangi bir teriminin farklı belgeler için hesaplanan z puanları belge uzunluklarından bağımsız bir şekilde birbirleriyle kıyaslanabilir. Mevcut yöntemlerde normalleştirme adı altında yapılmaya çalışılanda zaten budur.



Şekil 4.2 Kelimelerin belgelerde z puanları ile kelime önemi arasındaki ilişki

İfade edilen dönüşüme göre z = α değerine sahip olan terim belgedeki en önemli terim olarak kabul edilmektedir. Bir başka söyleyişle, Şekil 2.1'de en önemli terimi gösteren frekans değerine karşılık gelen z puanı α'ya eşittir. Herhangi bir t_i teriminin b_j belgesindeki z_{ij} puanı ile ilgili terimin önemi, yani terim frekans modellemesi olan TF_{ij} bileşini arasındaki ters yönlü ilişki şu şekilde olmaktadır:

$$z_{ij} \propto \frac{1}{Z_\alpha} \quad , Z_\alpha = |z_{ij} - \alpha| \qquad \text{Denklem 4.12}$$

Z dönüşümüne alternatif olarak Luhn'un iddiasında geçen "orta frekans" tanımını "belge içinde gözlenen frekans değerlerinin medyanı" kabul eden bir yaklaşım ile kelime önemi ve kelime frekansı ilişkisini nicel olarak ifade etmek mümkündür. Bir b_j belgesindeki gözlemlenen frekans değerlerini küçükten büyüğe sıralayacak olursak medyan M_j bu sıralı değerlerin tam ortadaki değerine denk gelmektedir. Medyan seçimindeki temel nokta gözlenen aynı frekans değerlerinden sadece bir tanesinin ele alınmasıdır. Bu bağlamda kelime gözlenme sıklığı ile kelime önemi, yani TF_{ij} arasındaki ters yönlü ilişki şu şekilde ifade edilmiştir:

$$TF_{ij} \propto \frac{1}{|x_{ij} - M_j|} \qquad \text{Denklem 4.13}$$

Görüldüğü gibi herhangi bir kelimenin gözlendiği belgede taşınan anlamsal enformasyona yaptığı katkının miktarı, ilgili kelimenin gözlemlendiği belgedeki sıklığının (x_{ij}) belge medyanı M_j'ye uzaklığı ile ters orantılıdır. Fakat bu haliyle ilişki kelimenin belge cinsinden uzunluğuna bağlıdır. Belge boyutuna bağlılığı bertaraf etmek için ise belge içi frekansların medyan noktasına göre standart sapması (Bkz. Denklem 4.14) kullanılabilir. Böylece aradaki uzaklık belge içi standart sapması s_{mj} nispetinde uzaklık olarak ifade edilerek standart hale getirilmiştir.

$$TF_{ij} \propto \frac{1}{|x_{ij} - M_j| / \sqrt{s_{mj}^2}} \quad , \sqrt{s_{mj}^2} = \sqrt{\frac{1}{r-1}\left(\sum_{i=1}^{r} x_{ij} - M_j\right)} \qquad \text{Denklem 4.14}$$

Belge boyutuna bağlığı ortadan kaldırmak için kullanılabilecek diğer bir yöntem ise kelime sıklığı ile medyan arasındaki uzaklığın medyan nispetinde gösterilmesidir. Herhangi bir kelimenin gözlendiği belgeler içindeki sıklıklarının kıyaslanabilir, yani standart bir hale getirilebilir olması ilgili sıklıkların gözlendiği belgeye ait yapısal bir takım özelliklere (örnek olarak belgenin kelime boyutu) bağımlı olduğu varsayımına dayanmaktadır. Belgelerde gözlenen medyan değerlerine etki eden/bağımlı hale getiren belgeye ait yapısal özelliklerinin de

aynı olduğunu varsaymak mümkündür. Bu bakış açısıyla Denklem 4.13'deki ilişki Denklem 4.15'deki gibi normalleştirilebilir.

$$TF_{ij} \propto \frac{1}{|x_{ij} - M_j|/M_j} \qquad \text{Denklem 4.15}$$

Belgelerin gerçek medyan değerleri (M_j^G) yerine belgenin bir takım özelliklerine göre tahmin edilen medyan değerlerinin-tahmini medyan- (M_j^T) kullanılması da mümkündür. Belgelerin içyapıları yazıldığı doğal dile bağımlı olduğundan dolayı evrensel bir modelin olması gerektiğini varsaymak mümkündür. Böyle bir modelin gösterilebilmesi durumunda belge parametrelerine etki eden *rastsal olaylar* elimine edilmiş olunur (Örnek: yazarın kullandığı dil). Belgedeki toplam terim sayısı *N*, terim kümesinin boyutu/kelime dağarcığı *V* ve medyanı M gösterecek olursa; toplam terim sayısı ile $V \times M$ çarpımı arasında paralel bir ilişki vardır. Aralarındaki bu ilişki Denklem 4.16'daki gibi formüle edilmiştir.

$$N^{\beta} \propto M \times V \qquad \text{Denklem 4.16}$$

İlişkinin genel gösterimi $I = M \times V$ Denklem 4.17'de verilmektedir.

$$N^{\beta}/_{I} = c, c \text{ sabit} \qquad \text{Denklem 4.17}$$

Denklem'de β ile gösterilen üssel parametre verinin logaritmik olarak grafiği çizildiğinde noktalara uyan doğrunun eğimidir. Her iki tarafın logaritması alınırsa Denklem 4.18 elde edilir.

$$\beta log(N) - log(I) = C, C = log(c) \qquad \text{Denklem 4.18}$$

N'nin logaritması (log *N*) apsiste (*x*) ve *I*'nın logaritması ordinatta (*y*) olacak biçimde yukarıdaki denklem Denklem 4.19'daki biçime dönüşür.

$$y = \beta x - C \qquad \text{Denklem 4.19}$$

Denklem 4.19'daki doğrunun β - *eğim*- ve *C –Y ekseni kesmesi*- katsayıları en küçük kareler yaklaşımı (İng. Least Square Approximation, kıs. LSA) kullanılarak elde edilebilir. LSA'ye göre β *ve C* parametreleri, Belge numarası *j* ve toplam belge sayısı *n* ile gösterilecek olursa şu de hesaplanır.

$$\beta = \frac{\left(n\sum_{j=1}^{n} x_j \cdot y_j - \left(\sum_{j=1}^{n} x_j \cdot \sum_{j=1}^{n} x_j\right)\right)}{\left(n\sum_{j=1}^{n} x_j^2 - \left(\sum_{j=1}^{n} x_j\right)^2\right)}$$

Denklem 4.20

$$C = \left(p\sum_{j=1}^{n} x_j - \sum_{j=1}^{n} x_j\right) \Big/ n$$

Bu parametrelerin bulunması ile herhangi bir belgenin medyanı M_j^T, Denklem 4.21'deki hesaplamayla tahmin edilebilir.

$$M_j^T = N^{\beta}/(V \times c) \quad ,c = antilog(c)$$ Denklem 4.21

4.3.2 Luhn Esasında TFxIDF şemasına uygun terim ağırlıklandırma

Önceki bölümde anlatılan terim frekansı ile TF bileşeni arasında 2 farklı yolla (z-puanları ve medyan) kurulan ilişkinin nicel olarak formüle edilmesi üzere standart uzaklık US_{ij} gösterimi kullanılmaktadır. Herhangi bir t_i teriminin b_j belgesindeki standart uzaklık US_{ij} değeri:

a) Z-puanları ile yaklaşımda z_{ij}'nin α değerine uzaklığı olan Z_α'ya (Bkz. Denklem 4.12) eşittir. (Bkz. Denklem 4.22(a)

b) Medyan ile yaklaşımda ise Denklem 4.14 (1) veya Denklem 4.15'de (2) gösterilen biçimde terimin gözlenme sıklığı x_{ij}'nin medyandan olan standart farkına eşittir. (Bkz. Denklem 4.22(b1) ve Denklem 4.22(b2))

$$US_{ij} = \begin{cases} Z_{\alpha} & \text{(a)} \\ \dfrac{|x_{ij} - M_j|}{\sqrt{s_{mj}^2}} & \text{(b1)} \\ \dfrac{|x_{ij} - M_j|}{M_j} & \text{(b2)} \end{cases} \qquad \text{Denklem 4.22}$$

TF bileşenini US_{ij} ile nicel olarak hesaplanabilmesi için kullanılabilecek 2 fonksiyon; TF-1 ve TF-2 olarak adlandırılarak Denklem 4.23'de verilmektedir.

$$TF_{ij} = \begin{cases} \dfrac{1}{US_{ij} + 1} & \text{(a) TF-1} \\ \dfrac{1}{US_{ij}^2 + 1} & \text{(b) TF-2} \end{cases} \qquad \text{Denklem 4.23}$$

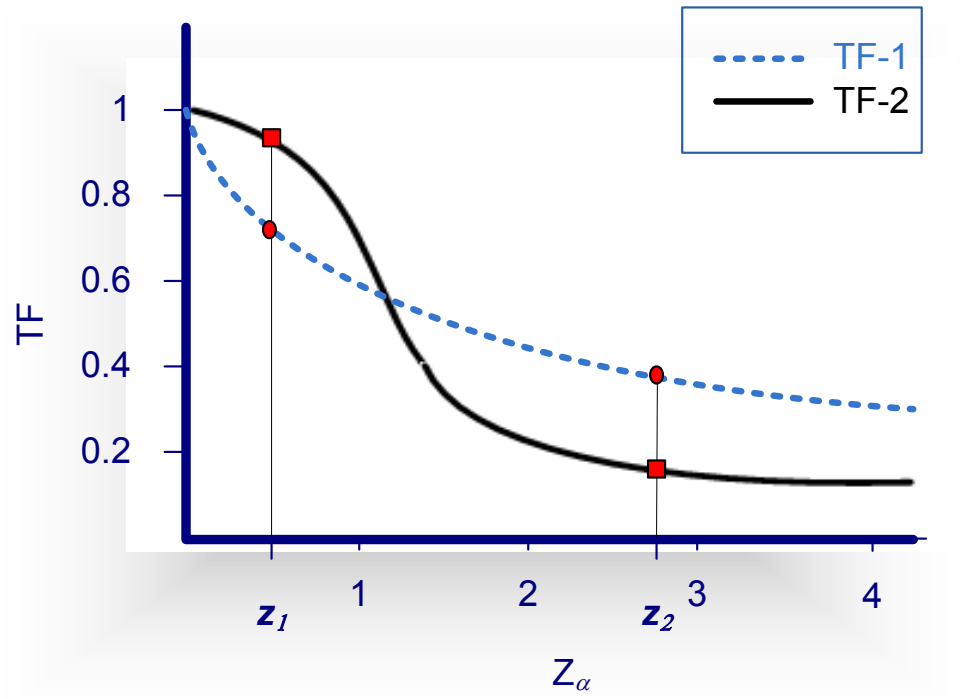
Denklem 4.23'teki fonksiyonlar yerine gerek mevcut ağırlıklandırma işleminde sıkça *kullanılmasından* gerekse belge puanı hesaplanması sırasında vektör uzayı yaklaşımına uygun olmasından dolayı fonksiyonların *logaritmik dönüşümlü* halleri ağırlıklandırma için kullanılmıştır. Bu son fonksiyonlar Denklem 4.24'te verilmiştir.

$$TF_{ij} = \begin{cases} log_2\left(\dfrac{1}{US_{ij} + 1} + 1\right) & \text{(a) TF-1} \\ log_2\left(\dfrac{1}{US_{ij}^2 + 1} + 1\right) & \text{(b) TF-2} \end{cases} \qquad \text{Denklem 4.24}$$

Denklem 4.24'te gösterilen TF-1 ve TF-2 hesaplama yöntemlerinin TFxIDF şemasındaki formüllerinde, IDF bileşeni olarak Sparck Jones'un *idf* hesaplaması (Bkz. Denklem 2.2: $idf_i = log_2\left(\frac{n}{n_i} + 1\right)$) kullanılmış olup sırasıyla WTF-1 ve WTF2 olarak adlandırılmıştır.

$$TF_{ij} \times IDF_i = \begin{cases} log_2\left(\frac{1}{US_{ij}+1}+1\right) \times idf_i & \text{(a) WTF-1} \\ log_2\left(\frac{1}{US_{ij}^2+1}+1\right) \times idf_i & \text{(b) WTF-2} \end{cases} \qquad \text{Denklem 4.25}$$

Denklem 4.24'teki TF-1 ve TF-2 fonksiyonlarının Z_α'ya göre değişimleri Şekil 4.3'te verilmiştir.



Şekil 4.3 TF ile Z_α ilişkisi

Şekil 4.3'te, t_1 ve t_2 terimlerinin Z_α değerleri sırasıyla z_1 ve z_2 olarak gösterilmiştir; t_1 terimi için TF-1 yerine TF-2 ile hesaplamada alaka değeri artarken, t_2 teriminde ters etki yapmaktadır. t_1 terimi t_2 terimine göre orta frekanslı yani önemli durumdadır. Bundan dolayı TF-2 hesaplaması önemli atfedilen orta frekanslı kelimelere oldukça yüksek önem değerleri atarken, kalan terimlere daha düşük değerler atamaktadır. Özetle TF-2 fonksiyonu orta frekanslı terimlere daha duyarlıdır ; yani Luhn'un iddia ettiği orta frekanslı bölgeye daha duyarlıdır.

4.4 Bağımsızlıktan Sapma Modellerinin TERRIER'de Gerçeklenmesi

Bölüm 4.2.2'de açıklanan bağımsızlıktan sapma tabanlı (DFI ve türevleri) modeller ve Bölüm 4.3.2'de açıklanan Luhn tabanlı TFxIDF modelleri, TERRIER kütüphanesi kullanılarak Java programlama diliyle Eclipse platformun'da gerçekleştirilmiştir. İlgili modelleri belirten Java sınıflarının, TERRIER'in eşleştirme modülüne eklenebilmesi için <uk.ac.gla.terrier.matching.models> paketinde bulunması ve *WeightingModel* sınıfından türetilmeleri gerekmektedir.

Denklem 4.6'da verilen bağımsızlıktan sapma modeline göre yaratılan DFI sınıfına ait Java kodu Şekil 4.4'te gösterilmiştir. *termFrequency* ile n_i, *numberOfTokens* ile N temsil edilmekte ve bu değerler *WeightingModel* sınıfından kalıtım yoluyla alınmaktadır. Ayrıca *docLength* ile n_j temsil edilmekte ve bu değişkenin değeri eşleştirme modülü tarafından çağrım sırasında sağlanmaktadır.

```
package uk.ac.gla.terrier.matching.models

public class DFI extends WeightingModel{

    public DFI (){
        super();
    }

    public final String getInfo() {
        return "DFI";
    }

    public final double score(double tf, double docLength){
        double eij = (termFrequency*docLength) /
                                        numberOfTokens;
        double DFIscore= (tf-eij)/eij;
        DFIscore = DFIscore > 0 ? DFIscore : 0;
....... return DFIscore
    }
}
```

Şekil 4.4 Denklem 4.6'da verilen DFI modeline uygun Java Sınıfı

Denklem 4.24.a'da verilen Luhn esasında terim sıklığı/frekansı modellemesine (Denklemde TF-1 ile gösterilen) ait olan Java kodu ise Şekil 4.5'te verilmiştir.

```
package uk.ac.gla.terrier.matching.models

public class TF-1 extends WeightingModel{

    public TF-1 (){
        super();
    }

    public final String getInfo() {
        return "Luhn based TF";
    }

    public final double score(double tf, double docLength){
     //Denklem 4.21.a
      double α = 1.0;
......double USij = Math.abs([(xij-ort_j)/var_j]-α)
     //Denklem 4.21.b
        //double USij = Math.abs((xij-Mj)/var_Mj)
     //Denklem 4.21.c
        //double USij = Math.abs((xij-Mj)/Mj)
        double TF-1score= Math.log( (1/(USij+1))+1);
        return TF-1score;
    }
}
```

Şekil 4.5 Denklem 4.22'de verilen Luhn esasında TF modeline uygun Java Sınıfı

4.5'deki kodda Denklem 4.22'de verilen US_{ij}'nin 3 farklı değerine göre hesaplanması gösterilmiştir. Denklem 4.22.a'ya uygun olan hesaplamada kullanılan belge içi frekans ortalaması *ort_j* ve varyans *var_j* bilgi erişim sürecinden önce bulunmakta olup terim puanı hesaplanırken kullanılır. Ayrıca ilgili kodda bulunan α parametresi için 1 değeri seçilmiştir. Denklem 4.22.b1 ve 4.22.b2'ye uygun olan hesaplamalarda kullanılan belge içi terim frekanslarının medyanları M_j ve varyans(M_j) -M_j'nin varyansı- değerleri de bilgi erişim sürecinden önce hesaplanmaktadır.

TERRIER, eşleştirme modeli olarak vektör uzayı modelini kullanır; sorgu terimleri ağırlıklarının değeri 1 olarak varsayılanda (default) alınmaktadır. Ancak sorgu terimlerinin, sorgu dizgesinde görünme sıklıklarına göre ağırlıklandırılmasını sağlayan bir yapı da mevcuttur. Sorgu terimlerinin görünme sıklıklarının toplam sorgu dizisi uzunluğuna bölünerek normalleştirilmiş değerlerini bulan bu yapı eşleştirme modelinde gömülü olarak bulunmaktadır.

WeightingModel" sınıfında, hesaplanan sorgu terim ağırlıkları anahtar frekansı (af) adı ile "protected double keyFrequency" değişkeninde tutulabilmektedir. TERRIER'e gömülü olarak gelen kodlanmış tüm modeller, ağırlıklandırılmış terim puanlarını son olarak bu anahtar frekansı değişkeni ile çarpmaktadırlar. Anahtar frekans değerinin kullanılması başarımı pozitif yönde etkilemektedir. Yani 4.4'te ve 4.5'te verilen modellerde anahtar frekansı hesaba katılarak elde edilen terim puanları sırasıyla:

```
(1)return DFIscore×keyFrequency;
(2)return Luhn-TFscore×keyFrequency;
```

biçiminde değiştirilmiştir.

Bölüm 4.2.2'te verilen bağımsızlıktan sapma modellerine uygun olarak oluşturulan Java sınıfları/fonksiyon isimleri, temel aldıkları Denklemlere göre Çizelge 4.1'de verilmiştir. Yine raporun ilerleyen bölümlerinde modelin belirteçi olarak Java sınıflarının isimleri kullanılmıştır. Bağımsızlıktan sapma modellerinin ağrılıklandırma hesaplamalarının açık formülleri EK-2'de bulunmaktadır.

Çizelge 4.1 Bağımsızlıktan Sapma modelleri

Temel Model	Model Belirteçi	Denklem
DFI_0	DFI_0_0	Denklem 4.6
	DFI_0_1	Denklem 4.9.a
	DFI_0_2	Denklem 4.10.a
DFI_1	DFI_1_0	Denklem 4.8
	DFI_1_1	Denklem 4.9.b
	DFI_1_2	Denklem 4.10.b

Çizelge 4.1'de gösterilen 6 formül aslen 2 temel modelin; DFI_0 ve DFI_1, nicel olarak 3 değişik ifadesidir. Sırasıyla bu değişik ifadeler: temel modelin direk formülü, temel modelin logaritmik transformasyon ile elde edilen formülü ve temel logaritmik transformasyonuna ek IDF bileşenine sahip melez formülü. Örnek olarak, DFI_0_0 temel model DFI_0'ın doğrudan nicel gösterimi, DFI_0_1 fonksiyonu DFI_0_0'ın logaritmik transformasyonu ve DF_0_2 ise DFI_0_1'in IDF bileşeni eklenen şeklidir.

Ayrıca, Bölüm 4.3.2'de verilen Luhn esasındaki TF ve TFxIDF modellerinin hesaplanması için kullanılan java sınıf/fonksiyon isimleri ilgili modellerin belirteçleri olarak kullanılmaktadır. Belirteçlerle gösterilen Luhn esasındaki modellere karşılık gelen denklemler Çizelge 4.2'de verilmiştir. Ayrıca ağrılıklandırma hesaplamalarının açık formülleri EK-3'de bulunmaktadır.

Çizelge 4.2 Luhn esasındaki modeller

Model Belirteçi	Ağırlıklandırma Denklemi	US_{ij} Denklemi
TF-1(Z)	Denklem 4.24.a	4.22.a
TF-1(M1)	Denklem 4.24.a	4.22.b1
TF-1(M2)	Denklem 4.24.a	4.22.b2
TF-1(β,C)	Denklem 4.24.a	4.22.b2 ile Tahmini medyan için Denklem 4.21
TF-2(Z)	Denklem 4.24.b	4.22.a
TF-2(M1)	Denklem 4.24.b	4.22.b1
TF-2(M2)	Denklem 4.24.b	4.22.b2
TF-2(β,C)	Denklem 4.24.b	4.22.b2 ile Tahmini medyan için Denklem 4.21
WTF-1(Z)	Denklem 4.25.a	4.22.a
WTF-1(M1)	Denklem 4.25.a	4.22.b1
WTF-1(M2)	Denklem 4.25.a	4.22.b2
WTF-1(β,C)	Denklem 4.25.a	4.22.b2 ile Tahmini medyan için Denklem 4.21
WTF-2(Z)	Denklem 4.25.b	4.22.a
WTF-1(M1)	Denklem 4.25.b	4.22.b1
WTF-1(M2)	Denklem 4.25.b	4.22.b2
WTF-2(β,C)	Denklem 4.25.b	4.22.b2 ile Tahmini medyan için Denklem 4.21

5 DENEYLER

5.1 Deney Düzeneği

Deneyler için TREC-6 ve TREC-7,8 anlık sorgu izlerinde kullanılan belge kümesi olmak üzere iki farklı derlem TERRİER kütüphanesinde bulunan *tek-geçişli indeksleyici* kullanılarak indekslenmiştir. İndeksleme sırasında *Porter'ın gövdeleyicisi* (Porter, 1980) ile terimlere gövdeleme işlemi uygulanırken, belgelerin doğal istatistiklerinin bozulmaması için herhangi bir *durma kelimeleri listesi* kullanılmamıştır. İndeksleme sonucunda elde edilen derlem istatistikleri Çizelge 5.1'de gösterilmiştir. TREC-7,8 derlemi TREC-6 derleminden belge uzunlukları büyük olan kongre kayıtlarının (CR, Bkz. Çizelge 3.3) çıkartılması ile oluşmaktadır. CR'ın çıkartılması ortalama belge uzunluğunu 557'den 512'ye düşürerek, yaklaşık %8'lik bir azalmaya yol açmıştır.

Çizelge 5.1 TREC-6 ile TREC-7 ve TREC-8 anlık sorgu izinde kullanılan derlemlerin indeksleme ardından istatistikleri.

Gözlenen Bilgi	TREC-6 Derlemi	TREC-7 ve TREC-8 Derlemi
Belge Sayısı	556.009	528.087
Sözlük Boyutu	769.677	739.679
Kelime Sayısı	310.001.539	270.443.153
İşaretçi Sayısı	125.793.001	118.087.970
Ortalama Kelime/Belge	557	512

Sorgulama sırasında ise TERRIER ile bütünleşik gelen *733* adet kelime içeren *durma kelimeleri listesi* kullanılarak sorgunun içinde enformasyon taşımayan kelimeler elimine edilmiştir. Ayrıca deneylerde "*çok kısa"* (sadece TITLE), "*kısa"* (TITLE+DESC) ve "*tüm konu*" (TITLE+DESC+NARR) olmak üzere üç tipte sorgu dizgisi uygulanmıştır.

Geliştirilen indeks terim ağırlıklandırma formüllerinin; *DFI modeli, yani bağımsızlıktan sapma esasında uygun formüller ve Luhn'un fikri esasına uygun TF bileşeni ve TFxIDF şemasındaki formüller*, kullanıldığı deneyler TREC-6, TREC-7 ve TREC-8 anlık sorgu izlerinde yapılmıştır. Ayrıca TERRIER bilgi erişim platformunda gerçeklenen önemli modeller/formüller (Bkz. Çizelge 3.3) de

yine aynı deney düzeneğinde çalıştırılmıştır. Geliştirilen yöntemleri mevcut yöntemlerle kıyaslamak üzere yürütümleri gerçekleştirilen mevcut tüm modellerin yerine deneylerde öne çıkanlar seçilmiştir.

Sonuçların değerlendirilmesinde erişilen alakalı belge sayısı (*RR*), ortalama averaj duyarlık (*MAP*) ve R-Duyarlık(*R-P*) ölçütlerine (Bkz. Bölüm 3.1) ek olarak ilk X sayıda -X= {1, 5, 10, 30, 100}- belgedeki duyarlık değerleri kullanılmıştır. 'P@X' gösterimi ilgili duyarlık değerini temsil etmektedir. Yöntemlerin değerlendirilmesi/karşılaştırılması ile ortaya konulacak "başarım" sadece bu ölçütler bakımından bağlayıcıdır. Kısacası bu ölçütlerin TREC'in anlık-sorgu izi için değerlendirme ölçütleri olarak kullanılmasından dolayı, bilgi erişim başarımını *doğru olarak* gösterebildiği kabul edilmektedir.

5.2 DFI Tabanlı Modellerin Deney Sonuçları

DFI yani, bağımsızlıktan sapma modellerinin TREC-6, TREC-7 ve TREC-8 anlık sorgu izlerinde "çok kısa" sorgu tipinde yapılan deneylerinin sonuçları sırasıyla Çizelge 5.2, Çizelge 5-3 ve Çizelge 5-4'te verilmiştir.

Çizelge 5.2 Bağımsızlıktan sapma modellerinin TREC-6 anlık-sorgu izinde "çok kısa" sorgu tipinde başarım sonuçları

Konular: 301-351, Alakalı Belgelerin Sayısı: 4611								
	RR	**MAP**	**R-P**	P@1	P@5	P@10	P@30	P@100
DFI_0_0	*1718*	*0,1307*	*0,1559*	*0,2000*	*0,2080*	*0,1920*	*0,1540*	*0,0998*
DFI_0_1	*2205*	*0,2186*	*0,2661*	*0,4600*	*0,3880*	*0,3480*	*0,2807*	*0,1792*
DFI_0_2	*2178*	*0,2158*	*0,2672*	*0,4600*	*0,3800*	*0,3580*	*0,2760*	*0,1700*
DFI_1_0	*1747*	*0,1411*	*0,1715*	*0,3200*	*0,2520*	*0,2320*	*0,1680*	*0,1044*
DFI_1_1	*2274*	*0,2262*	*0,2691*	*0,4800*	*0,4080*	*0,3660*	*0,2867*	***0,1830***
DFI_1_2	***2323***	***0,2305***	*0,2762*	***0,5200***	*0,4240*	*0,3860*	*0,2900*	*0,1764*
TF_IDF	*2156*	*0,2105*	*0,2544*	*0,5200*	*0,4600*	*0,3960*	*0,2793*	*0,1700*
BM25	*2173*	*0,2061*	*0,2545*	*0,4600*	*0,4040*	*0,3740*	*0,2780*	*0,1682*
IFB2	*2237*	*0,2237*	*0,2753*	*0,5000*	*0,4600*	*0,4100*	*0,2900*	*0,1764*
InL2	*2263*	*0,2270*	***0,2768***	*0,5000*	*0,4440*	***0,4240***	***0,2960***	*0,1798*
In_expB2	*2241*	*0,2250*	*0,2758*	*0,5000*	*0,4640*	*0,4160*	*0,2913*	*0,1794*
In_expC2	*2194*	*0,2197*	*0,2684*	*0,5000*	***0,4760***	*0,4120*	*0,2900*	*0,1750*

TREC-6 "çok kısa" sorgu tipi sonuçlarına göre temel modeller DFI_0 ve DFI_1 için logaritmik transformasyon kullanımı bilgi erişim başarımını %50 civarında arttırmıştır. Ancak benzer başarım artışı IDF bileşeninin hesaba katılması (DFI_0_2 ve DFI_1_2) ile gözlenmemektedir. DFI_1 temel modeli için

çok az da olsa bir başarım bulunurken: DFI_1_1'in 0,1830 olan P@100 değeri dışında kalan ölçütlerde DFI_1_2 daha yüksek değerler elde etmişti; DFI_0_1 R-P, P@1 ve P@10 dışındaki ölçütlerde daha yüksek başarım değerleri göstermiştir.

Çizelge 5.2'de verilen sonuçlara göre, DFI tabanlı modeller arasında DFI_1_2 fonksiyonu P@100 dışındaki tüm ölçütlerde en yüksek başarımı göstermiştir. Başarım ölçütlerinin genelinde mevcut yöntemlerden TF_IDF ve BM25 ile temel DFI modelleri (DFI_0_0 ve DFI_1_0) nispeten daha kötü başarımlara sahiptirler. DFR çatısına uygun olan 4 model ile geriye kalan DFI tabanlı 4 model arasında herhangi bir başarım sıralaması yapabilmek çok güç olmasına rağmen, DFI_1_2'in elde ettiği RR, MAP ve P@1 değerleri <2323 - 0.2305 - 0.52> en yüksek değerlerdir.

Çizelge 5.3 Bağımsızlıktan sapma modellerinin TREC-7 anlık-sorgu izinde "çok kısa" sorgu tipinde başarım sonuçları

Konular: 351-401, Alakalı Belgelerin Sayısı: 4674								
	RR	**MAP**	**R-P**	P@1	P@5	P@10	P@30	P@100
DFI_0_0	*1735*	*0,1047*	*0,1492*	*0,2200*	*0,1960*	*0,1820*	*0,1653*	*0,1214*
DFI_0_1	*2254*	*0,1731*	*0,2253*	*0,4800*	*0,4000*	*0,3260*	*0,2747*	*0,1716*
DFI_0_2	*2228*	*0,1794*	*0,2266*	*0,4600*	*0,4000*	*0,3580*	*0,2747*	*0,1746*
DFI_1_0	*1892*	*0,1167*	*0,1570*	*0,3000*	*0,2560*	*0,2420*	*0,1773*	*0,1216*
DFI_1_1	*2273*	*0,1796*	*0,2334*	*0,4600*	*0,4320*	*0,4240*	***0,3047***	*0,1822*
DFI_1_2	***2355***	***0,1959***	***0,2486***	*0,4600*	*0,4680*	*0,4240*	***0,3047***	***0,1886***
TF_IDF	*2172*	*0,1632*	*0,2161*	*0,4600*	*0,4160*	*0,3660*	*0,2613*	*0,1686*
BM25	*2186*	*0,1641*	*0,2143*	*0,4600*	*0,4200*	*0,3660*	*0,2613*	*0,1692*
IFB2	*2288*	*0,1875*	*0,2406*	*0,5000*	*0,4600*	*0,4240*	*0,2933*	*0,1830*
InL2	*2290*	*0,1848*	*0,2391*	*0,4800*	*0,4600*	*0,4200*	*0,2960*	*0,1806*
In_expB2	*2286*	*0,1877*	*0,2404*	*0,5000*	***0,5100***	***0,4260***	*0,2933*	*0,1834*
In_expC2	*2243*	*0,1814*	*0,2337*	***0,5200***	*0,4560*	*0,4100*	*0,2907*	*0,1796*

TREC-7 "çok kısa" sorgu tipi sonuçlarına göre de temel modeller DFI_0 ve DFI_1 için logaritmik transformasyon kullanımı bilgi erişim başarımını %50'den fazla arttırmıştır. Ancak benzer başarım artışı IDF bileşeninin hesaba katılması (DFI_0_2 ve DFI_1_2) ile gözlenmemektedir. DFI_1_2 için özellikle RR, MAP ve R-P ölçütleri için bir başarım artışı bulunurken: DFI_1_2'in <2355 - 0,1959 - 0.246> olan RR, MAP ve R-P değerleri dışında kalan ölçütlerde DFI_1_1 ile benzer değerler gözlenmektedir; DFI_0_2 ile DFI_0_1 tüm ölçütler açısından çok yakın başarımlar elde etmiştir.

Çizelge 5.3'de verilen sonuçlara göre, DFI tabanlı modeller arasında DFI_1_2 fonksiyonu P@1 dışındaki tüm ölçütlerde en yüksek başarımı göstermiştir. Başarım ölçütlerinin genelinde mevcut yöntemlerden TF_IDF ve BM25 ile temel DFI modelleri (DFI_0_0 ve DFI_1_0) nispeten daha kötü başarımlara sahiptirler. DFR çatısına uygun olan 4 model ile geriye kalan DFI tabanlı 4 model arasında herhangi bir başarım sıralaması yapabilmek güç olmasına rağmen, DFI_1_2'in elde ettiği RR,MAP ve R-P değerleri <2355 - 0,1959 - 0,2486> açısından bilgi erişim başarımı en yüksek olandır. P@1, P@5 ve P@10 duyarlık ölçütlerine bakıldığında ise In_expB2 <0,50 - 0.51- 0.426> ile yüksek/tutarlı bir görüntü vermektedir

Çizelge 5.4 Bağımsızlıktan sapma modellerinin TREC-8 anlık-sorgu izinde "çok kısa" sorgu tipinde başarım sonuçları

Konular: 401-451, Alakalı Belgelerin Sayısı: 4728								
	RR	**MAP**	**R-P**	P@1	P@5	P@10	P@30	P@100
DFI_0_0	*2216*	*0,1402*	*0,1727*	*0,1800*	*0,2040*	*0,2080*	*0,1727*	*0,1338*
DFI_0_1	*2794*	*0,2370*	*0,2923*	*0,3600*	*0,4320*	*0,4340*	*0,3413*	*0,2184*
DFI_0_2	*2674*	*0,2271*	*0,2878*	*0,4200*	*0,3660*	*0,4000*	*0,3147*	*0,2100*
DFI_1_0	*2313*	*0,1469*	*0,1939*	*0,3200*	*0,2480*	*0,2420*	*0,2087*	*0,1436*
DFI_1_1	*2807*	*0,2466*	***0,3017***	***0,5600***	*0,4920*	***0,4660***	***0,3640***	*0,2272*
DFI_1_2	***2854***	***0,2488***	*0,3011*	*0,5400*	*0,4560*	*0,4280*	*0,3467*	*0,2224*
TF_IDF	*2670*	*0,2203*	*0,2804*	*0,4000*	*0,4480*	*0,4240*	*0,3273*	*0,2154*
BM25	*2672*	*0,2198*	*0,2780*	*0,4000*	*0,4360*	*0,4220*	*0,3280*	*0,2142*
IFB2	*2807*	*0,2418*	*0,2935*	*0,4600*	*0,4880*	*0,4580*	*0,3607*	*0,2310*
InL2	*2811*	*0,2415*	*0,2968*	*0,4600*	*0,4640*	*0,4620*	*0,3613*	*0,2300*
In_expB2	*2803*	*0,2424*	*0,2951*	*0,4600*	*0,4800*	*0,4600*	*0,3620*	***0,2326***
In_expC2	*2746*	*0,2330*	*0,2912*	*0,4800*	***0,4960***	*0,4560*	*0,3500*	*0,2264*

TREC-8 "çok kısa" sorgu tipi sonuçlarına göre de diğer anlık sorgu izlerinde olduğu gibi temel modeller DFI_0 ve DFI_1 için logaritmik transformasyon kullanımı bilgi erişim başarımını %50'den fazla arttırmıştır. Ancak benzer de bir başarım artışı IDF bileşeninin hesaba katılması (DFI_0_2 ve DFI_1_2) ile gözlenmemektedir. DFI_1_2 için RR ve MAP ölçütleri için çok az da bir başarım artışı bulunurken: DFI_1_2'in <2854 - 0,2488> olan RR, MAP ve R-P değerleri dışında kalan ölçütlerde DFI_1_1 ile benzer değerler gözlenmektedir; DFI_0_1 ile P@1 dışındaki tüm ölçütler açısından DFI_0_2'den yüksek başarım elde edilmiştir.

Çizelge 5.4'te verilen sonuçlara göre, DFI tabanlı modeller arasında DFI_1_2 fonksiyonu RR ve MAP ölçütlerinde en yüksek başarımı gösterirken;

geriye kalan ölçütlerde DFI_1_1 ile en yüksek başarım elde edilmiştir. Başarım ölçütlerinin genelinde mevcut yöntemlerden TF_IDF ve BM25 ile temel DFI modelleri (DFI_0_0 ve DFI_1_0) nispeten daha kötü başarımlara sahiptirler. DFR çatısına uygun olan dört model ile geriye kalan DFI tabanlı dört model arasında herhangi bir başarım sıralaması yapabilmek güç olmasına rağmen, DFI_0_1 ve DFI_0_2 daha düşük başarımlara sahiptirler. In_expC2 ile elde edilen P@5 değeri 0,4960 ve In_expB2 ile elde edilen P@100 değeri dışında kalan tüm ölçütlerde DFI_1_1 ve DFI_2 en yüksek başarımı göstermiştir.

Çizelge 5.5 Bağımsızlıktan sapma modellerinin TREC-6 anlık-sorgu izinde "kısa" sorgu tipinde başarım sonuçları

Konular: 301-351, Alakalı Belgelerin Sayısı: 4611								
	RR	**MAP**	**R-P**	P@1	P@5	P@10	P@30	P@100
DFI_0_0	*1754*	*0,1479*	*0,1704*	*0,2400*	*0,2600*	*0,2340*	*0,1807*	*0,1088*
DFI_0_1	*2343*	*0,2308*	*0,2795*	*0,4600*	*0,4400*	*0,4060*	*0,3000*	*0,1868*
DFI_0_2	*2506*	*0,2572*	*0,3030*	*0,4400*	*0,4640*	*0,4280*	*0,3133*	***0,1954***
DFI_1_0	*1813*	*0,1444*	*0,1666*	*0,3600*	*0,2880*	*0,2400*	*0,1787*	*0,1126*
DFI_1_1	*2295*	*0,2126*	*0,2538*	*0,4600*	*0,4160*	*0,3560*	*0,2820*	*0,1718*
DFI_1_2	***2535***	***0,2633***	***0,3040***	***0,5800***	*0,4560*	*0,4180*	*0,3153*	*0,1942*
TF_IDF	*2361*	*0,2099*	*0,2637*	*0,4800*	*0,4720*	*0,4300*	***0,3160***	*0,1850*
BM25	*2381*	*0,2121*	*0,2617*	*0,4600*	*0,4560*	*0,4180*	*0,3127*	*0,1856*
IFB2	*2428*	*0,2235*	*0,2702*	*0,5200*	*0,4880*	*0,4120*	*0,3080*	*0,1898*
InL2	*2418*	*0,2293*	*0,2692*	*0,5000*	*0,4920*	***0,4400***	*0,3107*	*0,1880*
In_expB2	*2437*	*0,2239*	*0,2644*	*0,5400*	*0,4800*	*0,4220*	*0,3093*	*0,1916*
In_expC2	*2413*	*0,2173*	*0,2662*	*0,5400*	***0,4960***	*0,4140*	*0,3060*	*0,1898*

TREC-6 "kısa" sorgu tipi sonuçlarına göre de "çok kısa" sorgu tipindeki sonuçlarda gözlenen gibi temel modeller DFI_0 ve DFI_1 için logaritmik transformasyon kullanımı bilgi erişim başarımını %50'den fazla arttırmıştır. Bununla birlikte, IDF bileşeninin hesaba katılması (DFI_0_2 ve DFI_1_2) ile özellikle RR, MAP ve R-P değerlerinde bir başarım artışı (%10-20) gözlenmemektedir. DFI_0_1'in <2343 - 0,2308 - 0,2795> olan RR, MAP ve R-P değerleri DFI_0_2 ile <2506 - 0,2572 - 0.3030> olarak artmış; DFI_1_1'in <2343 - 0,2308 - 0,2795> olan RR, MAP ve R-P değerleri DFI_0_2 ile <2295 - 0,2126 - 0.3040> olarak artmıştır.

Çizelge 5.5'te verilen sonuçlara göre, DFI tabanlı modeller arasında DFI_1_2 fonksiyonu RR, MAP, R-P, P@1 ve P@30 ölçütlerinde en yüksek başarımı gösterirken; geriye kalan ölçütlerde DFI_1_1 ile en yüksek başarım elde edilmiştir. Başarım ölçütlerinin genelinde temel DFI modelleri (DFI_0_0 ve

DFI_1_0) çok daha kötü başarımlara sahiptirler. Kıyaslama için kullanılan mevcut modeller ile DFI tabanlı DFI_0_1 ve DFI_1_1 benzer başarımlar elde ederken, DFI_0_1 ve DFI_0_2 RR, MAP ve R-P açısından belirgin bir yüksek başarıma sahiptirler. In_expC2 ile elde edilen P@5 değeri 0,4960, In_expB2 ile elde edilen P@10 değeri 0,44, ve TF_IDF ile elde edilen P@30 değeri 0.3160 dışında kalan tüm ölçütlerde DFI_1_1 ve DFI_2 en yüksek başarımı göstermiştir.

Çizelge 5.6 Bağımsızlıktan sapma modellerinin TREC-7 anlık-sorgu izinde "kısa" sorgu tipinde başarım sonuçları

Konular: 351-401, Alakalı Belgelerin Sayısı: 4674								
	RR	**MAP**	**R-P**	P@1	P@5	P@10	P@30	P@100
DFI_0_0	*1743*	*0,1072*	*0,1511*	*0,2800*	*0,2000*	*0,1800*	*0,1693*	*0,1214*
DFI_0_1	*2553*	*0,2074*	*0,2593*	*0,5200*	*0,4560*	*0,4400*	*0,3367*	*0,2068*
DFI_0_2	*2503*	*0,2112*	*0,2602*	*0,5400*	*0,4680*	*0,4420*	*0,3253*	*0,2052*
DFI_1_0	*2071*	*0,1224*	*0,1683*	*0,3200*	*0,2800*	*0,2640*	*0,1887*	*0,1300*
DFI_1_1	*2549*	*0,2055*	*0,2600*	*0,5400*	*0,4920*	*0,4520*	*0,3487*	*0,2014*
DFI_1_2	***2647***	***0,2244***	***0,2753***	*0,5200*	*0,5160*	*0,4740*	***0,3520***	***0,2160***
TF_IDF	*2517*	*0,1936*	*0,2482*	*0,4800*	*0,4880*	*0,4440*	*0,3207*	*0,2032*
BM25	*2513*	*0,1936*	*0,2480*	*0,4800*	*0,4680*	*0,4400*	*0,3227*	*0,2038*
IFB2	*2611*	*0,2173*	*0,2675*	***0,6400***	*0,5240*	***0,4900***	*0,3393*	*0,2132*
InL2	*2609*	*0,2150*	*0,2688*	***0,6400***	*0,5040*	*0,4720*	*0,3447*	*0,2116*
In_expB2	*2620*	*0,2177*	*0,2694*	***0,6400***	***0,5280***	*0,4800*	*0,3380*	*0,2146*
In_expC2	*2601*	*0,2116*	*0,2644*	*0,5800*	*0,5320*	*0,4780*	*0,3420*	*0,2118*

TREC-7 "kısa" sorgu tipi sonuçlarına göre de "çok kısa" sorgu tipindeki sonuçlarda gözlenen gibi temel modeller DFI_0 ve DFI_1 için logaritmik transformasyon kullanımı bilgi erişim başarımını %50'den fazla arttırmıştır. Ancak, IDF bileşeninin hesaba katılması (DFI_0_2 ve DFI_1_2) ile belirgin bir başarım artışı gözlenmemektedir. DFI_1_2 ile P@1 dışındaki tüm ölçütler için az da olsa yüksek değerler elde edilmişken, DFI_0_2 ile DFI_0_1 arasında başarım değerlerinin büyüklüğü değişiklik göstertmektedir.

Çizelge 5.6'da verilen sonuçlara göre, DFI tabanlı modeller arasında DFI_1_2 fonksiyonu P@1, P@5 ve P@10 dışındaki ölçütlerde en yüksek başarımı göstermiştir. Başarım ölçütlerinin genelinde; temel DFI modelleri (DFI_0_0 ve DFI_1_0) oldukça düşük başarımlara, mevcut yöntemlerden TF_IDF ve BM25 ise geriye kalan yöntemlere nispeten daha kötü başarımlara sahiptirler. DFR çatısına uygun olan dört model ile geriye kalan DFI tabanlı 4 model arasında herhangi bir başarım sıralaması yapabilmek güç olmasına rağmen, DFI_1_2'in elde ettiği RR,MAP ve R-P değerleri <2647 - 0,2244 - 0,2753> açısından bilgi

erişim başarımı en yüksek olandır. P@1, P@5 duyarlık ölçütlerine bakıldığında ise In_expB2 <0,64 - 0.5280-> ve P@30 için ise IFB2 <0,49> ile en yüksek değerlere ulaşmıştır.

Çizelge 5.7 Bağımsızlıktan sapma modellerinin TREC-8 anlık-sorgu izinde " kısa" sorgu tipinde başarım sonuçları

Konular: 401-451, Alakalı Belgelerin Sayısı: 4728								
	RR	**MAP**	**R-P**	P@1	P@5	P@10	P@30	P@100
DFI_0_0	*2314*	*0,1461*	*0,1737*	*0,2200*	*0,2200*	*0,2160*	*0,1727*	*0,1340*
DFI_0_1	*2926*	*0,2630*	*0,3045*	*0,4800*	*0,5120*	*0,4700*	*0,3580*	*0,2276*
DFI_0_2	*2947*	*0,2532*	*0,3000*	*0,5000*	*0,4600*	*0,4280*	*0,3353*	*0,2158*
DFI_1_0	*2473*	*0,1506*	*0,2033*	*0,3400*	*0,2720*	*0,2580*	*0,2187*	*0,1516*
DFI_1_1	*2856*	*0,2534*	*0,3014*	*0,5800*	*0,5360*	*0,4900*	*0,3600*	*0,2258*
DFI_1_2	***3090***	***0,2718***	***0,3172***	*0,5600*	*0,5120*	*0,4540*	*0,3587*	*0,2272*
TF_IDF	*2825*	*0,2447*	*0,2931*	*0,5000*	*0,5040*	*0,4640*	*0,3593*	*0,2252*
BM25	*2830*	*0,2438*	*0,2920*	*0,5000*	*0,4960*	*0,4480*	*0,3533*	*0,2234*
IFB2	*2977*	*0,2695*	*0,3045*	*0,5800*	*0,5360*	*0,4860*	*0,3687*	*0,2372*
InL2	*2964*	*0,2663*	*0,3087*	*0,5600*	*0,5320*	***0,4920***	***0,3700***	***0,2390***
In_expB2	*2979*	*0,2708*	*0,3074*	*0,6000*	*0,5360*	***0,4920***	*0,3687*	*0,2386*
In_expC2	*2937*	*0,2607*	*0,2982*	***0,6400***	***0,5520***	*0,4880*	*0,3600*	*0,2328*

TREC-8 "kısa" sorgu tipi sonuçlarına göre de "çok kısa" sorgu tipindeki sonuçlarda gözlenen gibi temel modeller DFI_0 ve DFI_1 için logaritmik transformasyon kullanımı bilgi erişim başarımını %50'den fazla arttırmıştır. Ancak, IDF bileşeninin hesaba katılması (DFI_0_2 ve DFI_1_2) ile belirgin bir başarım artışı gözlenmemektedir. DFI_1_2 ile RR, MAP ve R-P açısından yüksek değerler elde edilmişken, şaşırtıcı biçimde P@100 dışındaki tüm duyarlık değerlerinde DFI_1_1 ile daha yüksek başarım değerleri elde edilmiştir. Diğer temel model olan DFI_0 da ise DFI_0_2 ile DFI_0_1 arasında başarım değerlerinin büyüklüğü değişiklik göstertmekle birlikte başarım sonuçları çok yakındır.

Çizelge 5.7'de verilen sonuçlara göre, DFI tabanlı modeller arasında DFI_1_2 fonksiyonu RR, MAP ve R-P açsından en yüksek başarımı göstermiştir. Başarım ölçütlerinin genelinde; temel DFI modelleri (DFI_0_0 ve DFI_1_0) oldukça düşük başarımlara, mevcut yöntemlerden TF_IDF ve BM25 ise geriye kalan yöntemlere nispeten daha kötü başarımlara sahiptirler. DFR çatısına uygun olan dört model ile geriye kalan DFI tabanlı dört model arasında herhangi bir başarım sıralaması yapabilmek güç olmasına rağmen, DFI_1_2'in elde ettiği RR,MAP ve R-P değerleri <3090 - 0,2718 - 0,3172> açısından bilgi erişim

başarımı en yüksek olandır. Duyarlık ölçütlerine bakıldığında ise DFI tabanlı modellerden çok az bir fark olsa da en yüksek değerler DFR çatısındaki modellerden In_expB2,in_ExpC2 ve InL2 arasında dağılım göstermektedir.

Çizelge 5.8 Bağımsızlıktan sapma modellerinin TREC-6 anlık-sorgu izinde "tüm konu" sorgu tipinde başarım sonuçları

Konular: 301-351, Alakalı Belgelerin Sayısı: 4611								
	RR	**MAP**	**R-P**	P@1	P@5	P@10	P@30	P@100
DFI_0_0	*1990*	*0,1641*	*0,1999*	*0,2600*	*0,2760*	*0,2440*	*0,1960*	*0,1256*
DFI_0_1	*2527*	*0,2543*	*0,2977*	*0,5600*	*0,5360*	*0,4620*	*0,3327*	*0,2100*
DFI_0_2	***2698***	***0,2920***	***0,3230***	*0,6000*	*0,5240*	*0,4640*	*0,3380*	***0,2140***
DFI_1_0	*2031*	*0,1487*	*0,1779*	*0,3400*	*0,2880*	*0,2560*	*0,1887*	*0,1276*
DFI_1_1	*2348*	*0,1822*	*0,2261*	*0,5000*	*0,3520*	*0,3080*	*0,2527*	*0,1654*
DFI_1_2	*2594*	*0,2720*	*0,3070*	*0,6400*	*0,4480*	*0,3960*	*0,3127*	*0,1964*
TF_IDF	*2553*	*0,2426*	*0,2851*	*0,5800*	***0,5520***	*0,4600*	***0,3460***	*0,2084*
BM25	*2536*	*0,2418*	*0,2883*	*0,5800*	*0,5360*	*0,4520*	*0,3425*	*0,2076*
IFB2	*2571*	*0,2388*	*0,2774*	*0,6200*	*0,5200*	*0,4340*	*0,3187*	*0,2026*
InL2	*2516*	*0,2830*	*0,2439*	***0,6400***	*0,5280*	***0,4680***	*0,3280*	*0,2040*
In_expB2	*2594*	*0,2389*	*0,2739*	*0,6200*	*0,5320*	*0,4460*	*0,3327*	*0,2068*
In_expC2	*2606*	*0,2363*	*0,2684*	*0,6000*	*0,5360*	*0,4420*	*0,3347*	*0,2076*

TREC-6 "tüm konu" sorgu tipi sonuçlarına göre de "çok kısa" ve "kısa" sorgu tipindeki sonuçlarda gözlenen gibi temel modeller DFI_0 ve DFI_1 için logaritmik transformasyon kullanımı bilgi erişim başarımını %50'den fazla arttırmıştır. Bununla birlikte, IDF bileşeninin hesaba katılması (DFI_0_2 ve DFI_1_2) ile özellikle RR, MAP ve R-P değerlerinde bir başarım artışı (DFI_0_2 için %40-50, DFI_1_2 için %10-20 civarı) gözlenmektedir: DFI_0_1 ile RR, MAP ve R-P açısından <2527 - 0.2543 - 0.2977> elde edilen değerler, DFI_0_2 ile <2698 - 0,2929 - 0,3230> biçiminde artmıştır. Yine DFI_1_1 ile RR, MAP ve R-P açısından <2348 - 0,1822 - 0,2261> elde edilen değerler, DFI_1_2 ile <2594 - 0,2720 - 0,3070> biçiminde artmıştır.

Çizelge 5.8'de verilen sonuçlara göre, DFI tabanlı modeller arasında DFI_0_2 fonksiyonu RR, MAP, R-P ve P@100 açısından en yüksek başarımı göstermiştir. Başarım ölçütlerinin genelinde temel DFI modelleri (DFI_0_0 ve DFI_1_0) çok daha kötü başarımlara sahiptirler. Kıyaslama için kullanılan mevcut modellerle ile DFI tabanlı DFI_0_2 ve DFI_1_2 benzer başarımlar elde ederken, DFI_0_2 ve DFI_1_2 RR, MAP ve R-P açısından belirgin bir yüksek başarıma sahiptirler. InL2 ile elde edilen P@1 ve P@10 değerleri <0,6400 -

0,4680 > ve TF_IDF ile elde edilen P@5 ve P@30 değerleri <0,5520 - 0,3460> dışında kalan tüm ölçütlerde DFI_0_2 en yüksek başarımı göstermiştir.

Çizelge 5.9 Bağımsızlıktan sapma modellerinin TREC-7 anlık-sorgu izinde "tüm konu" sorgu tipinde başarım sonuçları

Konular: 351-401, Alakalı Belgelerin Sayısı: 4674								
	RR	**MAP**	**R-P**	P@1	P@5	P@10	P@30	P@100
DFI_0_0	*1977*	*0,1203*	*0,1597*	*0,2400*	*0,2240*	*0,1920*	*0,1767*	*0,1394*
DFI_0_1	*2665*	*0,2219*	*0,2675*	*0,6000*	*0,4840*	*0,4660*	*0,3613*	*0,2132*
DFI_0_2	*2812*	***0,2463***	*0,2790*	***0,6600***	*0,5400*	*0,4840*	*0,3727*	***0,2298***
DFI_1_0	*2249*	*0,1416*	*0,1866*	*0,3000*	*0,2920*	*0,2720*	*0,2200*	*0,1486*
DFI_1_1	*2528*	*0,1954*	*0,2473*	*0,4600*	*0,4200*	*0,4060*	*0,3193*	*0,2012*
DFI_1_2	***2855***	*0,2432*	*0,2772*	*0,5400*	*0,5040*	*0,4740*	*0,3747*	*0,2272*
TF_IDF	*2670*	*0,2129*	*0,2677*	*0,5800*	*0,4920*	*0,4720*	*0,3507*	*0,2156*
BM25	*2709*	*0,2170*	*0,2634*	*0,6400*	*0,5120*	*0,4780*	*0,3520*	*0,2160*
IFB2	*2783*	*0,2428*	*0,2801*	*0,6000*	***0,5720***	*0,5060*	*0,3780*	*0,2260*
InL2	*2712*	*0,2274*	*0,2709*	*0,6000*	*0,5360*	*0,4760*	*0,3580*	*0,2170*
In_expB2	*2785*	*0,2427*	***0,2803***	*0,6000*	*0,5680*	***0,5080***	***0,3807***	*0,2282*
In_expC2	*2752*	*0,2381*	*0,2746*	***0,6600***	*0,5480*	*0,5060*	*0,3753*	*0,2270*

TREC-7 "tüm konu" sorgu tipi sonuçlarına göre de "çok kısa" ve "kısa" sorgu tipindeki sonuçlarda gözlenen gibi temel modeller DFI_0 ve DFI_1 için logaritmik transformasyon kullanımı bilgi erişim başarımını %50'den fazla arttırmıştır. Bununla birlikte, IDF bileşeninin hesaba katılması (DFI_0_2 ve DFI_1_2) ile özellikle RR, MAP ve R-P değerlerinde bir başarım artışı (özellikle DFI_1_2'de %10-20 civarı) gözlenmektedir: DFI_0_1 ile RR, MAP ve R-P açısından <2665 - 0,2219 - 0,2675> elde edilen değerler, DFI_0_2 ile <2812 - 0,2463 - 0,2790> biçiminde artmıştır. Yine DFI_1_1 ile RR, MAP ve R-P açısından <2528 - 0,1954 - 0,2473> elde edilen değerler, DFI_1_2 ile <2825- 0,2432 - 0,2772> biçiminde artmıştır.

Çizelge 5.9'da verilen sonuçlara göre, DFI tabanlı modeller arasında DFI_0_2 fonksiyonu MAP, R-P, P@1, P@5, P@10 ve P@100 açısından en yüksek başarımı gösterirken DFI_1_2 RR ve P@30 ölçütlerinde en yüksek başarımı göstermiştir. Başarım ölçütlerinin genelinde temel DFI modelleri (DFI_0_0 ve DFI_1_0) çok daha kötü başarımlara sahiptirler. Ayrıca DFI_1_1'de diğerlerinden genel olarak daha az bir bilgi başarımına sahiptir. Kıyaslama için kullanılan mevcut modellerle ile DFI tabanlı DFI_0_2 ve DFI_1_2 benzer başarımlar elde ederken; DFI_0_2 ile MAP, P@1 ve P@100 ölçütlerinde, DFI_1_2 ile RR ölçütünde en yüksek başarım gözlenmiştir. In_Exp2 ile elde

edilen R-P, P@10 ve P@30 değerleri <0,2803 - 0,5080 - 0,3807 > ve IFB2 ile elde edilen P@5 değeri <0,5720> elde edilen en yüksek başarımlardır.

Çizelge 5.10 Bağımsızlıktan sapma modellerinin TREC-8 anlık-sorgu izinde " kısa" sorgu tipinde başarım sonuçları

Konular: 401-451, Alakalı Belgelerin Sayısı: 4728								
	RR	**MAP**	**R-P**	P@1	P@5	P@10	P@30	P@100
DFI_0_0	*2389*	*0,1475*	*0,1869*	*0,2400*	*0,2320*	*0,2260*	*0,1733*	*0,1332*
DFI_0_1	*2821*	*0,2558*	*0,3037*	*0,5600*	*0,5320*	*0,4820*	*0,3653*	*0,2210*
DFI_0_2	*3107*	*0,2746*	*0,3153*	*0,5200*	*0,5120*	*0,4720*	*0,3587*	*0,2314*
DFI_1_0	*2512*	*0,1522*	*0,2030*	*0,3800*	*0,3040*	*0,2580*	*0,2187*	*0,1528*
DFI_1_1	*2681*	*0,2262*	*0,2780*	*0,5200*	*0,4400*	*0,4300*	*0,3233*	*0,2040*
DFI_1_2	***3126***	***0,2785***	***0,3175***	*0,6000*	*0,5200*	*0,4520*	*0,3700*	*0,2322*
TF_IDF	*2915*	*0,2499*	*0,2983*	*0,6800*	*0,5440*	*0,4800*	*0,3587*	*0,2298*
BM25	*2948*	*0,2512*	*0,3023*	*0,6600*	*0,5440*	*0,4800*	*0,3580*	*0,2286*
IFB2	*3153*	*0,2763*	*0,3179*	*0,7200*	*0,5440*	***0,5060***	*0,3767*	***0,2402***
InL2	*2965*	*0,2641*	*0,3110*	*0,6600*	*0,5320*	*0,4940*	*0,3767*	*0,2330*
In_expB2	*3141*	*0,2758*	***0,3180***	***0,7200***	***0,5560***	*0,4980*	***0,3793***	*0,2400*
In_expC2	*3120*	*0,2641*	*0,3128*	***0,7200***	*0,5400*	*0,4980*	*0,3707*	*0,2400*

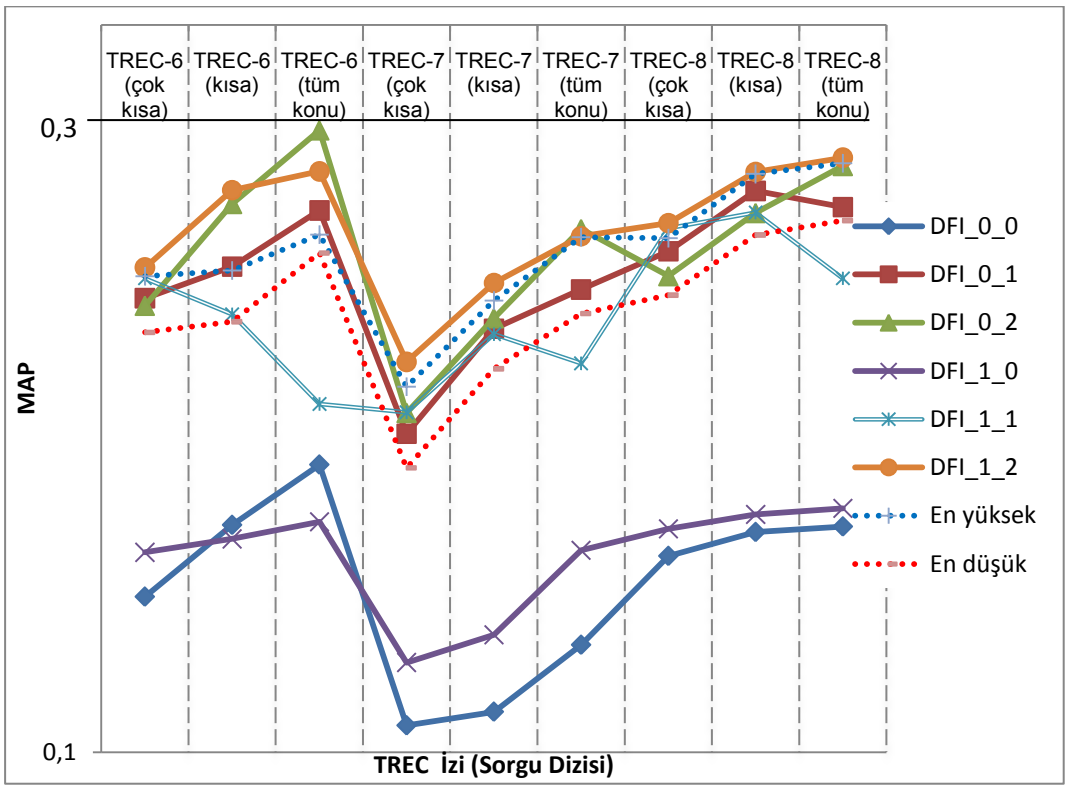
TREC-8 "tüm konu" sorgu tipi sonuçlarına göre de "çok kısa" ve "kısa" sorgu tipindeki sonuçlarda gözlenen gibi temel modeller DFI_0 ve DFI_1 için logaritmik transformasyon kullanımı bilgi erişim başarımını %50'den fazla arttırmıştır. Bununla birlikte, IDF bileşeninin hesaba katılması (DFI_0_2 ve DFI_1_2) ile özellikle RR, MAP ve R-P değerlerinde bir başarım artışı (özellikle DFI_1_2'de %10-20 civarı) gözlenmektedir: DFI_0_1 ile RR, MAP ve R-P açısından <2821 - 0,2558 - 0,3037> elde edilen değerler, DFI_0_2 ile <3107 - 0,2746 - 0,3153> biçiminde artmıştır. Yine DFI_1_1 ile RR, MAP ve R-P açısından <2681 - 0,2262 - 0,2780> elde edilen değerler, DFI_1_2 ile <3126- 0,2785 - 0,3175> biçiminde artmıştır.

Çizelge 5.10'da verilen sonuçlara göre, DFI tabanlı modeller arasında DFI_1_2 fonksiyonu P@5 ve P@10 dışındaki tüm ölçütlerde en yüksek başarımı gösterirken DFI_0_1 sırasıyla bu ölçütlerde <0,5320 - 0,4820> ile en yüksek başarımı göstermiştir. Temel DFI modelleri (DFI_0_0 ve DFI_1_0) başarım ölçütlerinin genelinde çok düşük değerlere sahiptirler. Ayrıca DFI_1_1'de diğerlerinden genel olarak daha az bir bilgi başarımına sahiptir. Kıyaslama için kullanılan mevcut modellerle ile DFI tabanlı DFI_0_2 ve DFI_1_2 benzer başarımlar elde ederken; DFI_1_2 ile RR, MAP, ve R-P ölçütlerinde en yüksek başarım gözlenmiştir. Duyarlık ölçütlerine bakıldığında ise DFI tabanlı

modellerden çok az bir fark olsa da en yüksek değerler DFR çatısındaki modellerden In_expB2,in_ExpC2 ve IFB2 arasında dağılım göstermektedir.

5.2.1 Gözlemlerin özeti

Deneylerde karşılaştırma için kullanılan ağırlıklandırma yöntemlerinin TREC derlemlerinde farklı sorgularda elde ettiği en düşük ve en yüksek MAP başarım değerleri Şekil 5.1'de kesikli çizgiyle gösterilmiştir. Yine aynı şekilde bağımsızlıktan sapma modeline uygun terim ağırlıklandırmalarının elde ettiği MAP başarımları verilmiştir.



Şekil 5.1 TREC izlerinde farklı sorgu tiplerindeki MAP başarımları

Deney sonuçlarına genel olarak bakıldığında ortaya çıkan ilk bulgu temel DFI fonksiyonlarının oldukça düşük bilgi erişim başarımı göstermesidir. DFI modellerinden logaritmik dönüşüm içeren formül ile başarım oldukça artmakla birlikte IDF kullanımının "çok kısa" ve "kısa" sorgu tiplerindeki başarıma katkıları sınırlıdır veya belirsizdir. Ancak "tüm konu" için IDF bileşeni kullanımı özellikle temel model DFI_1 için %10-20 civarında olmaktadır. Yine bu artış temel model DFI_0 için kısıtlıdır ve hatta TREC-8 anlık sorgu izi için yoktur. RR, MAP ve R-P değerlerine bakıldığında deney sonuçlarının genelinde DFI_0_2 ve

DFI_1_2 en yüksek değerleri elde etmiştir. Hatta TREC-6 anlık sorgu izindeki "kısa" ve "tüm konu" sorgu tipleri için bilgi erişim başarımı mevcut yöntemlerden ve diğer DFI tabanlı formüllerden belirgin bir şekide yüksektir. Önceden belirtildiği üzere TREC-6 anlık sorgu izinde kullanılan derlem diğer derlemlerden farklı olarak oldukça büyük boyutlu belgeler içermektedir. Bu yüksek başarımın altında yatan neden büyük boyutlu belgelere sahip derlemlerde geri getirim başarımının diğerlerinden daha iyi olduğu biçimde açıklanabilir. Ancak bu sonuç DFI'ın TREC-6'da bulunan sorgulara daha uyumlu olmasından da kaynaklanabilir.

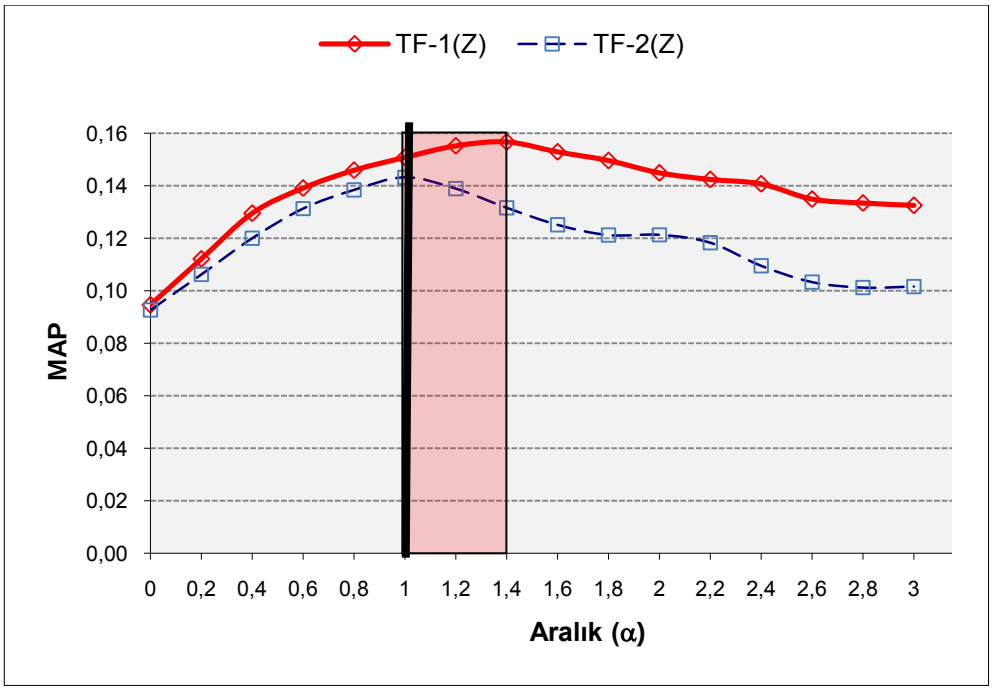
5.3 Luhn Esasında Geliştirilen Modellerin Deney Sonuçları

5.3.1 Z-Puanları için en uygun α değeri

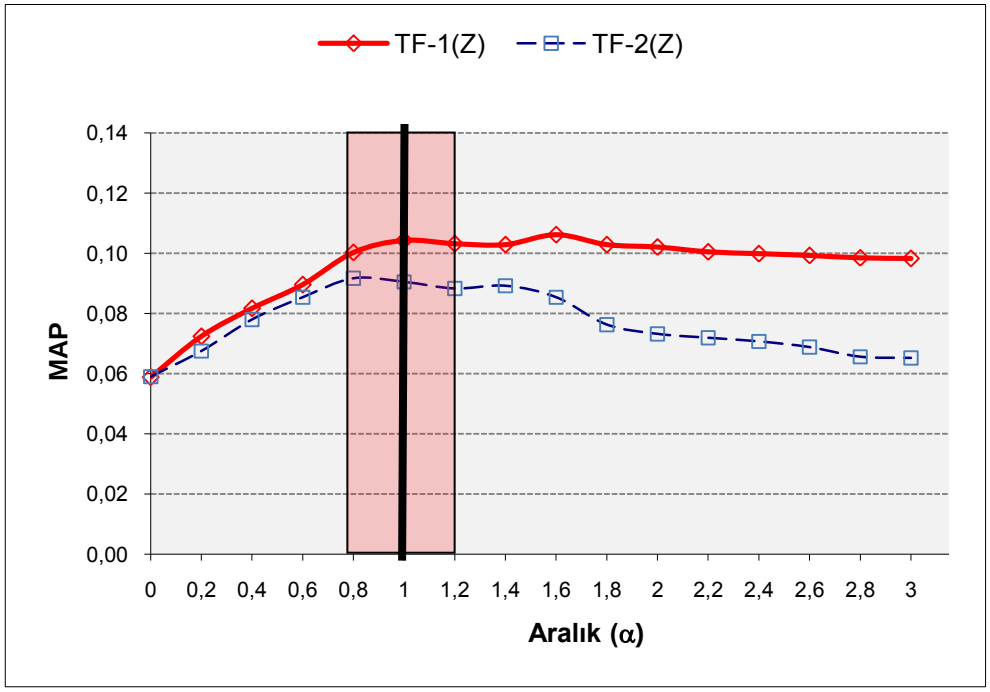
Erişimi istenilen bilgileri en iyi tarif eden kelimelerden oluşan “çok kısa” sorgu tipinde, sorgudaki her kelimenin anlamsal olarak etkisinin eşit olduğu varsaymak mümkün olabilir. Bu yüzden en uygun α değerinin bulunması veya var olup/olmadığının ortaya çıkartılması açısından Luhn'un fikri esasındaki TF değerini yalnız kullanmak uygun olmaktadır.

Z-puanları kullanılarak gerçeklenen ağırlıklandırma fonksiyonlarının; yani TF-1(Z) ve TF-2(Z)'in en yüksek başarımı sağladığı α değerinin bulunması için TREC-6, TREC-7 ve TREC-8 derlemlerinde “çok kısa” sorgu tipindeki deneyler temel alınmıştır. Bu deneylerde farklı α değerleri – α={0,0 - 0,2 - 0,4 - ….- 2,8 - 3,0} olmak üzere toplam 15 değer- için ilgili fonksiyonların başarımları hesaplanmıştır. Deney sonuçları TREC-6, TREC-7 ve TREC-8 için sırasıyla Şekil 5.2, Şekil 5.3 ve Şekil 5.4’te grafiksel olarak gösterilmiştir.

TREC-6 anlık sorgu izindeki her iki ağırlıklandırma modelinin tepe noktaları –en yüksek başarımın elde edildiği α değeri- alındığında arada kalan alan [1,0 1,2] aralığında olmaktadır. TREC-7 ve TREC-8’te ise aynı yaklaşımla tepe noktaları arasındaki aralık [0,8 1,2] olmaktadır. TREC-6 derleminden büyük boyutlu belgelerin çıkartılmasıyla oluşturulan TREC-7 ve TREC-8 derlemlerinde gözlemlenen bu aralık değişimi Z-puanlarının standartlaştırmadaki etkisinin tam olarak yeterli olmaması olarak ifade edilebilir. Ancak her üç derlemde de şekillerde koyu çizgi olarak gösterilen α=1,0 değeri her iki model içinde yaklaşık tepe noktası olarak kabul edebilir.

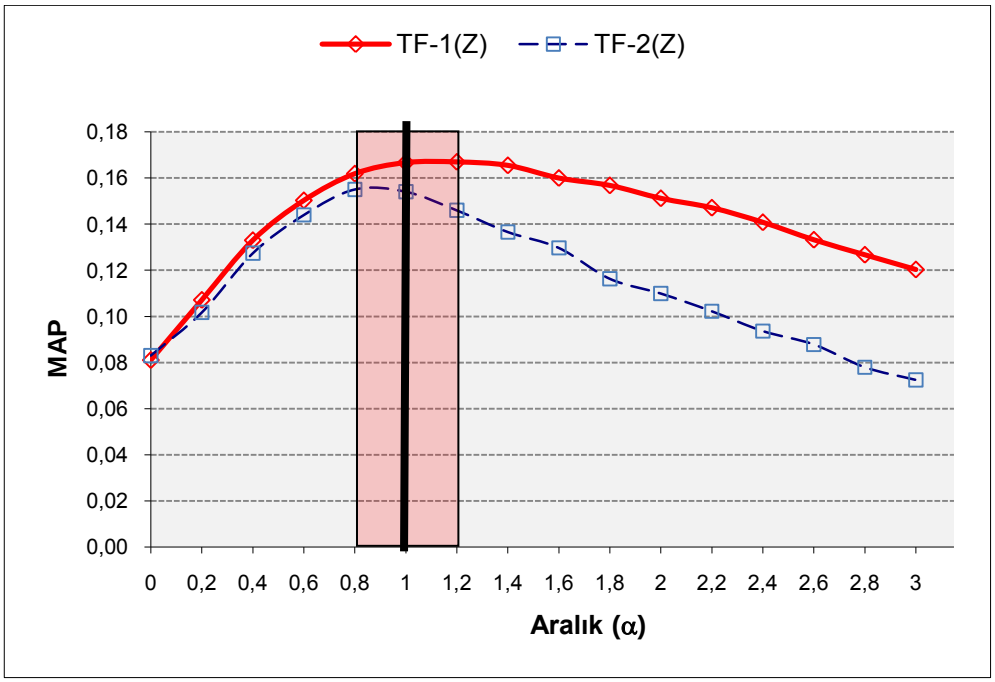


Şekil 5.2 TREC-6 anlık sorgu izinde "çok kısa" sorgu tipi için α ile MAP ilişkisi



Şekil 5.3 TREC-7 anlık sorgu izinde "çok kısa" sorgu tipi için α ile MAP ilişkisi

Tüm derlemlerde elde edilen sonuçlarda gözlemlenen: "Z-puanları için en uygun α değeri, yani en yüksek başarımın sağlandığı değerin bulunması ve bu değerden her iki yöne uzaklaştıkça başarımın düşmesi", Luhn'un iddiasını destekleyen bir bulgudur. Diğer bir önemli nokta ise yukarıda belirtilen bulgunun en açık biçimde ortaya çıktığı TREC-8 derlemi olmak ile birlikte (Bkz. Şekil 5.4) bunu sırasıyla TREC-6 ve TREC-7 derlemleri takip etmektedir. Aynı sıralamanın mevcut ağırlıklandırma yöntemlerinde elde edilen başarım sıralamasına denk gelmesi, erişilmesi istenilen bilgiyi en iyi tarif eden kelimelerin seçimindeki başarının, yani eşit anlam etkisine sahip kelimelerin seçimi olarak açıklanabilir.



Şekil 5.4 TREC-8 anlık sorgu izinde "çok kısa" sorgu tipi için α ile MAP ilişkisi

5.3.2 Modellerin başarımları ve mevcut yöntemlerle karşılaştırılması

Medyan değerinin tahmini için ortaya konulan hesaplama fonksiyonu için gerekli β ve C parametreleri TREC-6 derlemi için β = 1,2659 ve C = 1,4678 (hataların karelerinin ortalaması = 0,8289) olarak hesaplanmıştır. Yine TREC-7 ve TREC-8 anlık sorgu izlerinde kullanılan derlem için β = 1,2725 ve C = 1,5240 (hataların karelerinin ortalaması = 0,8986) olarak hesaplanmıştır. Bu parametreler ışığında iki farklı β ve C değerlerinin alınmasıyla medyan değeri tahmin edilmiştir. Medyan tahminine bağlı olan indeks terim ağırlıklandırılması hesaplamaları için iki farklı (β, C) kullanılmıştır: ilkinde (β, C) değerleri (1,27; 1,5) olarak ikincisinde ise (1,25; 1,0) olarak alınmıştır. Yine Z-puanları temel alan

yaklaşımlarda α değeri en uygun olarak tahmin edilen 1,0 değeri olarak kabul edilmiştir.

Genel olarak sonuçlara bakıldığında, Z-puanları ve medyanların standart sapmaları ile normalleştirilmiş modeller diğerlerine göre başarısız görünmektedir. Ayrıca bunun yanında, "çok kısa" sorgu tipi dışında potansiyel anlamı hesaplamaya katmayan modeller başarısız olduklarından, bunlarla ilgili deney sonuçlarına yer verilmemiştir.

Çizelge 5.11 Luhn tabanlı TF modellerin "çok kısa" sorgu tipindeki TREC-6 başarım sonuçları

Konular: 301-350, Alakalı Belgelerin Sayısı: 4611								
	RR	**MAP**	**R-P**	**P@1**	**P@5**	**P@10**	**P@30**	**P@100**
TF-1(Z)	*1782*	*0,1508*	*0,1903*	*0,3400*	*0,3200*	*0,2940*	*0,2240*	*0,1394*
TF-2(Z)	*1653*	*0,1432*	*0,1847*	*0,3800*	*0,3400*	*0,2880*	*0,2247*	*0,1276*
TF-1(M1)	*1964*	*0,1590*	*0,2014*	*0,3400*	*0,2880*	*0,2640*	*0,2200*	*0,1450*
TF-2(M1)	*1932*	*0,1540*	*0,1971*	*0,3000*	*0,2720*	*0,2560*	*0,2140*	*0,1452*
TF-1(M2)	*2072*	*0,1883*	*0,2442*	*0,3800*	*0,3720*	*0,3580*	*0,2587*	*0,1636*
TF-2(M2)	*2054*	*0,1954*	*0,2445*	*0,5000*	*0,4200*	*0,3580*	*0,2693*	*0,1682*
TF-1(β=1,27,C=1,15)	*2076*	*0,1953*	*0,2375*	*0,5200*	*0,4080*	*0,3440*	*0,2693*	*0,1670*
TF-2(β=1,27,C=1,15)	*2077*	*0,2011*	*0,2454*	***0,5400***	*0,4360*	*0,3680*	*0,2813*	*0,1730*
TF-1(β=1,25,C=1,0)	*2088*	*0,2048*	*0,2464*	*0,4800*	*0,4360*	*0,3880*	*0,2687*	*0,1690*
TF-2(β=1,25,C=1,0)	*2094*	*0,2105*	*0,2547*	*0,4600*	*0,4560*	*0,3980*	*0,2760*	*0,1752*
TF_IDF	*2156*	*0,2105*	*0,2544*	*0,5200*	*0,4600*	*0,3960*	*0,2793*	*0,1700*
BM25	*2173*	*0,2061*	*0,2545*	*0,4600*	*0,4040*	*0,3740*	*0,2780*	*0,1682*
InL2	***2263***	***0,2270***	***0,2768***	*0,5000*	*0,4440*	***0,4240***	***0,2960***	***0,1798***
In_expB2	*2241*	*0,2250*	*0,2758*	*0,5000*	***0,4640***	*0,4160*	*0,2913*	*0,1794*

Modellerin TREC-6 anlık-sorgu izindeki "çok kısa" sorgu tipi için başarım sonuçları Çizelge 5.11 ve Çizelge 5.12'de verilmiştir. Modellerin TFxIDF şemasına uygun tiplerindeki başarımları Çizelge 5.12'den görüleceği gibi yükselmiştir. Genel olarak bakıldığında uzunluğu medyan cinsiden alan modeller mevcut yöntemlere benzer başarım değerlerine sahiptir. WTF-1(β=1,27, C=1,15) ve WTF-2(β=1,27, C=1,15) ile P@1 ölçütünde 0,56 ve WTF-2(β=1,25, C=1,0) ile RR ve P@100 ölçütlerinde sırasıyla <2329, 0,1836> değerlerinde en yüksek başarım sonuçları gözlenmiştir.

Çizelge 5.12 Luhn tabanlı TFxIDF modellerin "çok kısa" sorgu tipindeki TREC-6 başarım sonuçları

Konular: 301-350, Alakalı Belgelerin Sayısı: 4611								
	RR	**MAP**	**R-P**	**P@1**	**P@5**	**P@10**	**P@30**	**P@100**
WTF-1(Z)	*1940*	*0,1709*	*0,2094*	*0,4000*	*0,3560*	*0,3220*	*0,2387*	*0,1450*
WTF-2(Z)	*2030*	*0,1757*	*0,2175*	*0,4000*	*0,3520*	*0,3340*	*0,2287*	*0,1482*
WTF-1(M1)	*2173*	*0,1728*	*0,2147*	*0,3200*	*0,3000*	*0,2740*	*0,2187*	*0,1508*
WTF-2(M1)	*2161*	*0,1703*	*0,2123*	*0,2800*	*0,2880*	*0,2700*	*0,2153*	*0,1502*
WTF-1(M2)	*2292*	*0,2046*	*0,2584*	*0,4200*	*0,4000*	*0,3620*	*0,2620*	*0,1700*
WTF-2(M2)	*2300*	*0,2139*	*0,2639*	*0,5000*	*0,4280*	*0,3700*	*0,2720*	*0,1752*
WTF-1(β=1,27,C=1,15)	*2278*	*0,2109*	*0,2527*	***0,5600***	*0,4200*	*0,3620*	*0,2687*	*0,1732*
WTF-2(β=1,27,C=1,15)	*2290*	*0,2184*	*0,2585*	***0,5600***	*0,4360*	*0,3780*	*0,2847*	*0,1782*
WTF-1(β=1,25,C=1,0)	*2314*	*0,2199*	*0,2617*	*0,5000*	*0,4320*	*0,3960*	*0,2800*	*0,1780*
WTF-2(β=1,25,C=1,0)	***2329***	*0,2254*	*0,2666*	*0,4500*	*0,4560*	*0,4180*	*0,2900*	***0,1836***
TF_IDF	*2156*	*0,2105*	*0,2544*	*0,5200*	*0,4600*	*0,3960*	*0,2793*	*0,1700*
BM25	*2173*	*0,2061*	*0,2545*	*0,4600*	*0,4040*	*0,3740*	*0,2780*	*0,1682*
InL2	*2263*	***0,2270***	***0,2768***	*0,5000*	*0,4440*	***0,4240***	***0,2960***	*0,1798*
In_expB2	*2241*	*0,2250*	*0,2758*	*0,5000*	***0,4640***	*0,4160*	*0,2913*	*0,1794*

Çizelge 5.13 Luhn tabanlı TF modellerin "çok kısa" sorgu tipindeki TREC-7 başarım sonuçları

Konular: 351-400, Alakalı Belgelerin Sayısı: 4674								
	RR	**MAP**	**R-P**	**P@1**	**P@5**	**P@10**	**P@30**	**P@100**
TF-1(Z)	*1607*	*0,1043*	*0,1513*	*0,3600*	*0,3040*	*0,2620*	*0,2053*	*0,1238*
TF-2(Z)	*1530*	*0,0905*	*0,1326*	*0,3400*	*0,3320*	*0,2620*	*0,1880*	*0,1100*
TF-1(M1)	*1791*	*0,1231*	*0,1835*	*0,4400*	*0,3240*	*0,2860*	*0,2313*	*0,1340*
TF-2(M1)	*1791*	*0,1204*	*0,1772*	*0,3800*	*0,3000*	*0,2920*	*0,2240*	*0,1350*
TF-1(M2)	*1786*	*0,1357*	*0,1918*	*0,4000*	*0,3680*	*0,3220*	*0,2467*	*0,1386*
TF-2(M2)	*1806*	*0,1375*	*0,1906*	*0,3800*	*0,3600*	*0,3280*	*0,2507*	*0,1388*
TF-1(β=1,27,C=1,15)	*1803*	*0,1367*	*0,1934*	*0,4400*	*0,3800*	*0,3320*	*0,2487*	*0,1400*
TF-2(β=1,27,C=1,15)	*1806*	*0,1383*	*0,1916*	*0,4800*	*0,3760*	*0,3480*	*0,2500*	*0,1410*
TF-1(β=1,25,C=1,0)	*1871*	*0,1499*	*0,2046*	*0,4400*	*0,4000*	*0,3640*	*0,2560*	*0,1458*
TF-2(β=1,25,C=1,0)	*1891*	*0,1499*	*0,2011*	***0,5000***	*0,4000*	*0,3700*	*0,2607*	*0,1474*
TF_IDF	*2172*	*0,1632*	*0,2161*	*0,4600*	*0,4160*	*0,3660*	*0,2613*	*0,1686*
BM25	*2186*	*0,1641*	*0,2143*	*0,4600*	*0,4200*	*0,3660*	*0,2613*	*0,1692*
InL2	***2290***	*0,1848*	*0,2391*	*0,4800*	*0,4600*	*0,4200*	***0,2960***	*0,1806*
In_expB2	*2286*	***0,1877***	***0,2404***	***0,5000***	***0,5100***	***0,4260***	*0,2933*	***0,1834***

Modellerin TREC-7 anlık-sorgu izindeki "çok kısa" sorgu tipi için başarım sonuçları Çizelge 5.13 ve Çizelge 5.14'te verilmiştir. Modellerin TFxIDF şemasına uygun tiplerindeki başarımları Çizelge 5.14'ten görüleceği gibi yükselmiştir. Çizelge 5.13'e bakıldığında uzunluğu medyan cinsiden alan yalın TF bileşenlerinin başarımı mevcut yöntemlerin altında gerçekleşirken, TFxIDF şemasına uygun tiplerindeki başarımları Çizelge 5.14'ten görüleceği gibi benzer başarım değerlerine sahiptir. Bu izde sadece WTF-2(β=1,27, C=1,15) ile P@1 ölçütünde 0,52 değerinde en yüksek başarım gözlenmiştir.

Çizelge 5.14 Luhn tabanlı TFxIDF modellerin "çok kısa" sorgu tipindeki TREC-7 başarım sonuçları

Konular: 351-400, Alakalı Belgelerin Sayısı: 4674								
	RR	**MAP**	**R-P**	**P@1**	**P@5**	**P@10**	**P@30**	**P@100**
WTF-1(Z)	*1959*	*0,1380*	*0,1821*	*0,3400*	*0,3400*	*0,3040*	*0,2420*	*0,1504*
WTF-2(Z)	*1903*	*0,1309*	*0,1779*	*0,3200*	*0,3800*	*0,3100*	*0,2380*	*0,1458*
WTF-1(M1)	*2134*	*0,1538*	*0,2125*	*0,4600*	*0,3240*	*0,3060*	*0,2627*	*0,1562*
WTF-2(M1)	*2143*	*0,1519*	*0,2074*	*0,3800*	*0,3200*	*0,3100*	*0,2587*	*0,1576*
WTF-1(M2)	*2161*	*0,1689*	*0,2239*	*0,4400*	*0,3720*	*0,3440*	*0,2807*	*0,1642*
WTF-2(M2)	*2170*	*0,1708*	*0,2223*	*0,3600*	*0,3640*	*0,3480*	*0,2820*	*0,1672*
WTF-1(β=1,27,C=1,15)	*2156*	*0,1699*	*0,2253*	*0,5000*	*0,3880*	*0,3500*	*0,2747*	*0,1668*
WTF-2(β=1,27,C=1,15)	*2178*	*0,1728*	*0,2246*	***0,5200***	*0,3920*	*0,3680*	*0,2860*	*0,1696*
WTF-1(β=1,25,C=1,0)	*2224*	*0,1828*	*0,2345*	*0,4800*	*0,4120*	*0,3680*	*0,2853*	*0,1696*
WTF-2(β=1,25,C=1,0)	*2268*	*0,1843*	*0,2332*	*0,4800*	*0,4480*	*0,3880*	*0,2907*	*0,1718*
TF_IDF	*2172*	*0,1632*	*0,2161*	*0,4600*	*0,4160*	*0,3660*	*0,2613*	*0,1686*
BM25	*2186*	*0,1641*	*0,2143*	*0,4600*	*0,4200*	*0,3660*	*0,2613*	*0,1692*
InL2	***2290***	***0,1848***	***0,2391***	*0,4800*	*0,4600*	*0,4200*	*0,2960*	*0,1806*
In_expB2	*2286*	*0,1877*	*0,2404*	*0,5000*	***0,5100***	***0,4260***	***0,2933***	***0,1834***

Çizelge 5.15 Luhn tabanlı TF modellerin "çok kısa" sorgu tipindeki TREC-8 başarım sonuçları

Konular: 401-450, Alakalı Belgelerin Sayısı: 4728								
	RR	**MAP**	**R-P**	**P@1**	**P@5**	**P@10**	**P@30**	**P@100**
TF-1(Z)	*2151*	*0,1667*	*0,2275*	*0,4600*	*0,4200*	*0,3640*	*0,2813*	*0,1736*
TF-2(Z)	*1986*	*0,1541*	*0,2130*	*0,4200*	*0,4240*	*0,3880*	*0,2773*	*0,1650*
TF-1(M1)	*2346*	*0,1900*	*0,2466*	*0,4000*	*0,3720*	*0,3520*	*0,2973*	*0,1864*
TF-2(M1)	*2342*	*0,1870*	*0,2487*	*0,4200*	*0,3680*	*0,3600*	*0,3013*	*0,1884*
TF-1(M2)	*2426*	*0,2073*	*0,2664*	*0,4400*	*0,4000*	*0,3940*	*0,3247*	*0,2022*
TF-2(M2)	*2416*	*0,2102*	*0,2694*	*0,4800*	*0,4320*	*0,4200*	*0,3420*	*0,2090*
TF-1(β=1,27,C=1,15)	*2429*	*0,2073*	*0,2671*	*0,5000*	*0,4480*	*0,4280*	*0,3280*	*0,2050*
TF-2(β=1,27,C=1,15)	*2422*	*0,2118*	*0,2709*	*0,4800*	*0,4480*	*0,4300*	*0,3413*	*0,2122*
TF-1(β=1,25,C=1,0)	*2473*	*0,2094*	*0,2679*	*0,5000*	*0,4600*	*0,4360*	*0,3273*	*0,2088*
TF-2(β=1,25,C=1,0)	*2489*	*0,2166*	*0,2725*	***0,5600***	*0,4760*	*0,4460*	*0,3460*	*0,2150*
TF_IDF	*2670*	*0,2203*	*0,2804*	*0,4000*	*0,4480*	*0,4240*	*0,3273*	*0,2154*
BM25	*2672*	*0,2198*	*0,2780*	*0,4000*	*0,4360*	*0,4220*	*0,3280*	*0,2142*
InL2	***2811***	*0,2415*	***0,2968***	*0,4600*	*0,4640*	***0,4620***	*0,3613*	*0,2300*
In_expB2	*2803*	***0,2424***	*0,2951*	*0,4600*	***0,4800***	*0,4600*	***0,3620***	***0,2326***

Modellerin TREC-8 anlık-sorgu izindeki "çok kısa" sorgu tipi için başarım sonuçları Çizelge 5.15 ve Çizelge 5.16'da verilmiştir. Modellerin potansiyel anlam ifadesi olarak *idf* kullanan tiplerindeki başarımları Çizelge 5.16'dan görüleceği gibi yükselmiştir. Genel olarak bakıldığında uzunluğu medyan cinsiden alan modeller mevcut yöntemlere benzer başarım değerlerine sahiptir. WTF-2(β=1,25, C=1,0) ile RR, MAP, R-P, P@1 ve P@5 ölçütlerinde sırasıyla <2821 - 0,2465 - 0,2981 - 0,62 - 0,4920> değerlerinde en yüksek başarım sonuçları gözlenmiştir.

Çizelge 5.16 Luhn tabanlı TFxIDF modellerin "çok kısa" sorgu tipindeki TREC-8 başarım sonuçları

Konular: 401-450, Alakalı Belgelerin Sayısı: 4728								
	RR	**MAP**	**R-P**	**P@1**	**P@5**	**P@10**	**P@30**	**P@100**
WTF-1(Z)	*2549*	*0,1978*	*0,2474*	*0,4400*	*0,4160*	*0,3740*	*0,2933*	*0,1886*
WTF-2(Z)	*2470*	*0,1902*	*0,2401*	*0,3800*	*0,4500*	*0,4080*	*0,2933*	*0,1858*
WTF-1(M1)	*2706*	*0,2203*	*0,2640*	*0,4200*	*0,3880*	*0,3640*	*0,3093*	*0,2016*
WTF-2(M1)	*2743*	*0,2178*	*0,2649*	*0,4400*	*0,3960*	*0,3700*	*0,3193*	*0,2058*
WTF-1(M2)	*2784*	*0,2341*	*0,2849*	*0,3800*	*0,4120*	*0,3860*	*0,3347*	*0,2164*
WTF-2(M2)	*2812*	*0,2398*	*0,2901*	*0,4600*	*0,4320*	*0,4180*	*0,3487*	*0,2240*
WTF-1(β=1,27,C=1,15)	*2773*	*0,2336*	*0,2837*	*0,4400*	*0,4440*	*0,4140*	*0,3380*	*0,2196*
WTF-2(β=1,27,C=1,15)	*2789*	*0,2390*	*0,2889*	*0,4200*	*0,4440*	*0,4320*	*0,3493*	*0,2260*
WTF-1(β=1,25,C=1,0)	*2808*	*0,2396*	*0,2932*	*0,5200*	*0,4560*	*0,4400*	*0,3440*	*0,2216*
WTF-2(β=1,25,C=1,0)	***2821***	***0,2465***	***0,2981***	***0,6200***	***0,4920***	*0,4540*	*0,3527*	*0,2286*
TF_IDF	*2670*	*0,2203*	*0,2804*	*0,4000*	*0,4480*	*0,4240*	*0,3273*	*0,2154*
BM25	*2672*	*0,2198*	*0,2780*	*0,4000*	*0,4360*	*0,4220*	*0,3280*	*0,2142*
InL2	*2811*	*0,2415*	*0,2968*	*0,4600*	*0,4640*	***0,4620***	*0,3613*	*0,2300*
In_expB2	*2803*	*0,2424*	*0,2951*	*0,4600*	*0,4800*	*0,4600*	***0,3620***	***0,2326***

Çizelge 5.17 Luhn tabanlı TFxIDF modellerin " kısa" sorgu tipindeki TREC-6 başarım sonuçları

Konular: 301-350, Alakalı Belgelerin Sayısı: 4611								
	RR	**MAP**	**R-P**	**P@1**	**P@5**	**P@10**	**P@30**	**P@100**
WTF-1(Z)	*1572*	*0,1182*	*0,1457*	*0,3200*	*0,2600*	*0,2080*	*0,1627*	*0,1068*
WTF-2(Z)	*1634*	*0,1288*	*0,1571*	*0,2400*	*0,2280*	*0,2160*	*0,1673*	*0,1160*
WTF-1(M1)	*1680*	*0,1053*	*0,1290*	*0,1600*	*0,1680*	*0,1520*	*0,1427*	*0,1020*
WTF-2(M1)	*1726*	*0,1221*	*0,1457*	*0,2000*	*0,1760*	*0,1780*	*0,1440*	*0,1066*
WTF-1(M2)	*1959*	*0,1689*	*0,1978*	*0,3400*	*0,2760*	*0,2560*	*0,1980*	*0,1308*
WTF-2(M2)	*1882*	*0,1521*	*0,1839*	*0,2000*	*0,2765*	*0,2340*	*0,1880*	*0,1248*
WTF-1(β=1,27,C=1,15)	*2002*	*0,1792*	*0,2087*	*0,4200*	*0,3300*	*0,2820*	*0,2120*	*0,1382*
WTF-2(β=1,27,C=1,15)	*1894*	*0,1587*	*0,1891*	*0,2800*	*0,2760*	*0,2400*	*0,1940*	*0,1284*
WTF-1(β=1,25,C=1,0)	*1974*	*0,1770*	*0,2093*	*0,3800*	*0,3200*	*0,2720*	*0,2073*	*0,1344*
WTF-2(β=1,25,C=1,0)	*1875*	*0,1592*	*0,1941*	*0,2600*	*0,2960*	*0,2560*	*0,1867*	*0,1268*
TF_IDF	*2361*	*0,2099*	*0,2637*	*0,4800*	*0,4720*	*0,4300*	***0,3160***	*0,1850*
BM25	*2381*	*0,2121*	*0,2617*	*0,4600*	*0,4560*	*0,4180*	*0,3127*	*0,1856*
InL2	*2418*	***0,2293***	***0,2692***	*0,5000*	***0,4920***	***0,4400***	*0,3107*	*0,1880*
In_expB2	***2437***	*0,2239*	*0,2644*	***0,5400***	*0,4800*	*0,4220*	*0,3093*	***0,1916***

TFxIDF şemasına uygun Luhn tabanlı modellerin TREC-6 anlık-sorgu izindeki "kısa" sorgu tipi için başarım sonuçları Çizelge 5.17'de verilmiştir. Ayrıca bu modellerden *idf*'nin etkisinin arttırılmış biçimlerinin başarım sonuçları Çizelge 5.18'de verilmiştir; burada IDF bileşeni olarak Sparck Jones'un *idf* (Bkz. Denklem 2.1) değerinin karesi kullanılmıştır. *idf*'nin karesinin kullanımı MAP değerlerinde yaklaşık %15-20 başarımı yükseltmiştir. Her iki duruma da bakıldığında uzunluğu medyan cinsiden alan modeller de mevcut yöntemlerden tüm duyarlık seviyelerinde daha kötü değerlere sahiptirler. Ancak *idf*'nin karesini

kullanan modellerde MAP ölçütü açısından diğerlerine yakın sonuçlar elde edilmiştir.

Çizelge 5.18 Luhn tabanlı TFxIDF ($IDF=idf^2$) modellerin " kısa" sorgu tipindeki TREC-6 başarım sonuçları

Konular: 301-350, Alakalı Belgelerin Sayısı: 4611								
	RR	**MAP**	**R-P**	**P@1**	**P@5**	**P@10**	**P@30**	**P@100**
WTF-1*(Z)	*1869*	*0,1763*	*0,2047*	*0,3000*	*0,3000*	*0,2740*	*0,1987*	*0,1300*
WTF-2*(Z)	*1937*	*0,1810*	*0,2105*	*0,2800*	*0,2920*	*0,2760*	*0,1953*	*0,1322*
WTF-1*(M1)	*1518*	*0,0893*	*0,1236*	*0,1800*	*0,1760*	*0,1580*	*0,1267*	*0,0870*
WTF-2*(M1)	*1793*	*0,1287*	*0,1657*	*0,1800*	*0,1920*	*0,1740*	*0,1657*	*0,1072*
WTF-1*(M2)	*2245*	*0,2149*	*0,2377*	*0,3600*	*0,3280*	*0,3060*	*0,2260*	*0,1508*
WTF-2*(M2)	*2155*	*0,2013*	*0,2312*	*0,2800*	*0,3040*	*0,2920*	*0,2107*	*0,1428*
WTF-1*(β=1,27,C=1,15)	*2256*	*0,2149*	*0,2499*	*0,3400*	*0,3440*	*0,3220*	*0,2313*	*0,1556*
WTF-2*(β=1,27,C=1,15)	*2163*	*0,2022*	*0,2338*	*0,3000*	*0,3120*	*0,2940*	*0,2180*	*0,1454*
WTF-1*(β=1,25,C=1,0)	*2251*	*0,2160*	*0,2470*	*0,3600*	*0,3600*	*0,3120*	*0,2260*	*0,1552*
WTF-2*(β=1,25,C=1,0)	*2125*	*0,2015*	*0,2376*	*0,3200*	*0,3250*	*0,2920*	*0,2100*	*0,1462*
TF_IDF	*2361*	*0,2099*	*0,2637*	*0,4800*	*0,4720*	*0,4300*	***0,3160***	*0,1850*
BM25	*2381*	*0,2121*	*0,2617*	*0,4600*	*0,4560*	*0,4180*	*0,3127*	*0,1856*
InL2	*2418*	***0,2293***	***0,2692***	*0,5000*	***0,4920***	***0,4400***	*0,3107*	*0,1880*
In_expB2	***2437***	*0,2239*	*0,2644*	***0,5400***	*0,4800*	*0,4220*	*0,3093*	*0,1916*

Çizelge 5.19 Luhn tabanlı TFxIDF modellerin " kısa" sorgu tipindeki TREC-7 başarım sonuçları

Konular: 351-400, Alakalı Belgelerin Sayısı: 4674								
	RR	**MAP**	**R-P**	**P@1**	**P@5**	**P@10**	**P@30**	**P@100**
WTF-1(Z)	*1928*	*0,1275*	*0,1755*	*0,4400*	*0,3720*	*0,3100*	*0,2320*	*0,1482*
WTF-2(Z)	*2009*	*0,1329*	*0,1797*	*0,4000*	*0,3200*	*0,2860*	*0,2273*	*0,1526*
WTF-1(M1)	*2089*	*0,1306*	*0,1767*	*0,2000*	*0,2960*	*0,2640*	*0,2047*	*0,1462*
WTF-2(M1)	*2095*	*0,1359*	*0,1842*	*0,2800*	*0,2920*	*0,2520*	*0,2133*	*0,1456*
WTF-1(M2)	*2237*	*0,1691*	*0,2202*	*0,4800*	*0,3960*	*0,3380*	*0,2660*	*0,1678*
WTF-2(M2)	*2197*	*0,1586*	*0,2106*	*0,4200*	*0,3720*	*0,2960*	*0,2533*	*0,1616*
WTF-1(β=1,27,C=1,15)	*2256*	*0,1741*	*0,2289*	*0,5400*	*0,4200*	*0,3520*	*0,2827*	*0,1714*
WTF-2(β=1,27,C=1,15)	*2198*	*0,1587*	*0,2138*	***0,4200***	*0,3680*	*0,3180*	*0,2593*	*0,1648*
WTF-1(β=1,25,C=1,0)	*2290*	*0,1772*	*0,2314*	*0,5200*	*0,4210*	*0,3580*	*0,2813*	*0,1722*
WTF-2(β=1,25,C=1,0)	*2225*	*0,1637*	*0,2125*	*0,4400*	*0,3600*	*0,3160*	*0,2587*	*0,1662*
TF_IDF	*2517*	*0,1936*	*0,2482*	*0,4800*	*0,4880*	*0,4440*	*0,3207*	*0,2032*
BM25	*2513*	*0,1936*	*0,2480*	*0,4800*	*0,4680*	*0,4400*	*0,3227*	*0,2038*
InL2	***2609***	***0,2150***	***0,2688***	*0,6400*	*0,5040*	***0,4720***	***0,3447***	***0,2116***
In_expB2	*2620*	*0,2177*	*0,2694*	*0,6400*	***0,5280***	*0,4800*	*0,3380*	*0,2146*

TFxIDF şemasına uygun Luhn tabanlı modellerin TREC-7 anlık-sorgu izindeki "kısa" sorgu tipi için başarım sonuçları Çizelge 5.19'da verilmiştir. Ayrıca bu modellerden *idf*'nin etkisinin arttırılmış biçimlerinin başarım sonuçları Çizelge 5.20'de verilmiştir; burada IDF bileşeni olarak Sparck Jones'un *idf* (Bkz. Denklem 2.1) değerinin karesi kullanılmıştır. *idf*'nin karesinin kullanımı MAP değerlerinde yaklaşık %15-20 başarımı yükseltmiştir. Her iki duruma da

bakıldığında uzunluğu medyan cinsiden alan modeller de DFR tabanlı modellerden tüm duyarlık seviyelerinde daha kötü değerlere sahiptirler. Özellikle *idf*'nin karesini kullanan modellerde MAP ve diğer ölçütler açısından TF_IDF ve BM25'e yakın sonuçlar elde edilmiştir.

Çizelge 5.20 Luhn tabanlı TFxIDF ($IDF=idf^2$) modellerin " kısa" sorgu tipindeki TREC-7 başarım sonuçları

Konular: 351-400, Alakalı Belgelerin Sayısı: 4674								
	RR	**MAP**	**R-P**	**P@1**	**P@5**	**P@10**	**P@30**	**P@100**
WTF-1*(Z)	*2090*	*0,1455*	*0,1914*	*0,4200*	*0,4000*	*0,3160*	*0,2407*	*0,1640*
WTF-2*(Z)	*2062*	*0,1468*	*0,1964*	*0,4200*	*0,3520*	*0,3080*	*0,2467*	*0,1656*
WTF-1*(M1)	*1924*	*0,1058*	*0,1549*	*0,2000*	*0,2240*	*0,2100*	*0,1807*	*0,1224*
WTF-2*(M1)	*2116*	*0,1303*	*0,1786*	*0,3000*	*0,2240*	*0,2320*	*0,2020*	*0,1470*
WTF-1*(M2)	*2283*	*0,1862*	*0,2380*	*0,4000*	*0,4480*	*0,3720*	*0,2920*	*0,1796*
WTF-2*(M2)	*2218*	*0,1739*	*0,2247*	*0,4400*	*0,3840*	*0,3360*	*0,2813*	*0,1758*
WTF-1*(β=1,27,C=1,15)	*2296*	*0,1903*	*0,2403*	*0,5000*	*0,4480*	*0,3900*	*0,3013*	*0,1800*
WTF-2*(β=1,27,C=1,15)	*2227*	*0,1789*	*0,2299*	***0,4800***	*0,4160*	*0,3120*	*0,2880*	*0,1768*
WTF-1*(β=1,25,C=1,0)	*2329*	*0,1983*	*0,2461*	*0,4800*	*0,4560*	*0,3880*	*0,3033*	*0,1844*
WTF-2*(β=1,25,C=1,0)	*2234*	*0,1848*	*0,2276*	*0,4600*	*0,4160*	*0,3600*	*0,2867*	*0,1772*
TF_IDF	*2517*	*0,1936*	*0,2482*	*0,4800*	*0,4880*	*0,4440*	*0,3207*	*0,2032*
BM25	*2513*	*0,1936*	*0,2480*	*0,4800*	*0,4680*	*0,4400*	*0,3227*	*0,2038*
InL2	***2609***	***0,2150***	***0,2688***	*0,6400*	*0,5040*	***0,4720***	***0,3447***	***0,2116***
In_expB2	*2620*	*0,2177*	*0,2694*	*0,6400*	***0,5280***	*0,4800*	*0,3380*	*0,2146*

Çizelge 5.21 Luhn tabanlı TFxIDF modellerin " kısa" sorgu tipindeki TREC-8 başarım sonuçları

Konular: 401-450, Alakalı Belgelerin Sayısı: 4728								
	RR	**MAP**	**R-P**	**P@1**	**P@5**	**P@10**	**P@30**	**P@100**
WTF-1(Z)	*2550*	*0,1896*	*0,2361*	*0,4600*	*0,4600*	*0,4080*	*0,2980*	*0,1830*
WTF-2(Z)	*2625*	*0,1936*	*0,2346*	*0,4800*	*0,4360*	*0,3820*	*0,2860*	*0,1808*
WTF-1(M1)	*2725*	*0,1910*	*0,2345*	*0,3400*	*0,3240*	*0,3160*	*0,2707*	*0,1828*
WTF-2(M1)	*2726*	*0,1927*	*0,2341*	*0,3200*	*0,3520*	*0,3200*	*0,2667*	*0,1784*
WTF-1(M2)	*2818*	*0,2324*	*0,2730*	*0,4400*	*0,4640*	*0,4200*	*0,3287*	*0,2122*
WTF-2(M2)	*2787*	*0,2205*	*0,2624*	*0,4400*	*0,4200*	*0,3880*	*0,3067*	*0,1992*
WTF-1(β=1,27,C=1,15)	*2816*	*0,2367*	*0,2757*	*0,4400*	*0,4720*	*0,4280*	*0,3360*	*0,2158*
WTF-2(β=1,27,C=1,15)	*2786*	*0,2230*	*0,2653*	*0,4800*	*0,4520*	*0,3980*	*0,3080*	*0,2004*
WTF-1(β=1,25,C=1,0)	*2851*	*0,2373*	*0,2763*	***0,6200***	*0,4920*	*0,4440*	*0,3287*	*0,2098*
WTF-2(β=1,25,C=1,0)	*2784*	*0,2192*	*0,2642*	*0,5400*	*0,4440*	*0,3980*	*0,3073*	*0,1938*
TF_IDF	*2825*	*0,2447*	*0,2931*	*0,5000*	*0,5040*	*0,4640*	*0,3593*	*0,2252*
BM25	*2830*	*0,2438*	*0,2920*	*0,5000*	*0,4960*	*0,4480*	*0,3533*	*0,2234*
InL2	*2964*	*0,2663*	***0,3087***	*0,5600*	*0,5320*	***0,4920***	***0,3700***	***0,2390***
In_expB2	***2979***	***0,2708***	*0,3074*	*0,6000*	***0,5360***	***0,4920***	*0,3687*	*0,2386*

TFxIDF şemasına uygun Luhn tabanlı modellerin TREC- anlık-sorgu izindeki "kısa" sorgu tipi için başarım sonuçları Çizelge 5.21'de verilmiştir. Ayrıca bu modellerden *idf*'nin etkisinin arttırılmış biçimlerinin başarım sonuçları Çizelge 5.22'de verilmiştir; burada IDF bileşeni olarak Sparck Jones'un *idf* (Bkz.

Denklem 2.1) değerinin karesi kullanılmıştır. Çizelge 5.22'den görüleceği gibi *idf*'nin karesinin kullanımı MAP değerlerinde belirgin bir başarım artışı göstermemektedir. DFR tabanlı modeller InL2 ve IN_expB2'nin başarımları diğer tüm modellerden yüksek olmasına rağmen, uzunluğu medyan cinsiden alan modellerin tümü MAP ve diğer ölçütler açısından TF_IDF ve BM25'e yakın sonuçlar elde edilmiştir.

Çizelge 5.22 Luhn tabanlı TFxIDF ($IDF=idf^2$) modellerin " kısa" sorgu tipindeki TREC-8 başarım sonuçları

Konular: 401-450, Alakalı Belgelerin Sayısı: 4728								
	RR	**MAP**	**R-P**	**P@1**	**P@5**	**P@10**	**P@30**	**P@100**
WTF-1*(Z)	*2669*	*0,1990*	*0,2422*	*0,3800*	*0,4280*	*0,4040*	*0,2947*	*0,1886*
WTF-2*(Z)	*2717*	*0,2002*	*0,2464*	*0,3600*	*0,3800*	*0,3740*	*0,2827*	*0,1828*
WTF-1*(M1)	*2423*	*0,1491*	*0,1990*	*0,3400*	*0,2560*	*0,2540*	*0,2233*	*0,1466*
WTF-2*(M1)	*2701*	*0,1875*	*0,2278*	*0,4000*	*0,2640*	*0,2740*	*0,2447*	*0,1662*
WTF-1*(M2)	*2909*	*0,2360*	*0,2851*	*0,4000*	*0,4280*	*0,4080*	*0,3273*	*0,2128*
WTF-2*(M2)	*2873*	*0,2278*	*0,2722*	*0,3400*	*0,4040*	*0,3840*	*0,3053*	*0,1976*
WTF-1*(β=1,27,C=1,15)	*2903*	*0,2347*	*0,2856*	*0,4000*	*0,4320*	*0,4100*	*0,3207*	*0,2122*
WTF-2*(β=1,27,C=1,15)	*2855*	*0,2264*	*0,2759*	*0,3800*	*0,4040*	*0,3860*	*0,3007*	*0,2014*
WTF-1*(β=1,25,C=1,0)	*2931*	*0,2407*	*0,2835*	*0,5000*	*0,4480*	*0,4140*	*0,2760*	*0,2120*
WTF-2*(β=1,25,C=1,0)	*2876*	*0,2291*	*0,2743*	*0,4400*	*0,4040*	*0,3920*	*0,3073*	*0,1970*
TF_IDF	*2825*	*0,2447*	*0,2931*	*0,5000*	*0,5040*	*0,4640*	*0,3593*	*0,2252*
BM25	*2830*	*0,2438*	*0,2920*	*0,5000*	*0,4960*	*0,4480*	*0,3533*	*0,2234*
InL2	*2964*	*0,2663*	***0,3087***	*0,5600*	*0,5320*	***0,4920***	***0,3700***	***0,2390***
In_expB2	***2979***	***0,2708***	*0,3074*	***0,6000***	***0,5360***	***0,4920***	*0,3687*	*0,2386*

TFxIDF şemasına uygun Luhn tabanlı modellerin TREC-6, TREC-7 ve TREC-8 anlık-sorgu izindeki "tüm konu" sorgu tipi için başarım sonuçları sırasıyla Çizelge 5.23, Çizelge 5.24 ve Çizelge 25'te verilmiştir. Ayrıca bu modellerden *idf*'nin etkisinin arttırılmış biçimlerinin başarım sonuçları sırasıyla Çizelge 5.26, Çizelge 5.27 ve Çizelge 28'de verilmiştir; *idf*'nin karesinin kullanımı tüm ölçütlerde belirgin bir başarım artışı göstermesine rağmen, modellerin başarımları diğerlerinden belirgin olarak düşük elde edilmiştir.

Çizelge 5.23 Luhn tabanlı TFxIDF modellerin " tüm konu" sorgu tipindeki TREC-6 başarım sonuçları

Konular: 301-350, Alakalı Belgelerin Sayısı: 4611								
	RR	**MAP**	**R-P**	**P@1**	**P@5**	**P@10**	**P@30**	**P@100**
WTF-1(Z)	*1153*	*0,0359*	*0,0592*	*0,1200*	*0,0920*	*0,1040*	*0,0807*	*0,0556*
WTF-2(Z)	*1224*	*0,0389*	*0,0633*	*0,1000*	*0,1000*	*0,1020*	*0,0744*	*0,0570*
WTF-1(M1)	*1264*	*0,0348*	*0,0521*	*0,0800*	*0,0720*	*0,0660*	*0,0653*	*0,0468*
WTF-2(M1)	*1338*	*0,0468*	*0,0688*	*0,1600*	*0,1200*	*0,1040*	*0,0940*	*0,0816*
WTF-1(M2)	*1456*	*0,0538*	*0,0744*	*0,1400*	*0,1120*	*0,1100*	*0,0987*	*0,0678*
WTF-2(M2)	*1324*	*0,0474*	*0,0711*	*0,1600*	*0,1040*	*0,1060*	*0,0947*	*0,0614*
WTF-1(β=1,27,C=1,15)	*1534*	*0,0610*	*0,0844*	*0,1600*	*0,1280*	*0,1240*	*0,1060*	*0,0720*
WTF-2(β=1,27,C=1,15)	*1210*	*0,0306*	*0,0459*	*0,0800*	*0,0720*	*0,0600*	*0,0567*	*0,0434*
WTF-1(β=1,25,C=1,0)	*1611*	*0,0728*	*0,0947*	*0,2400*	*0,1560*	*0,1560*	*0,1193*	*0,0848*
WTF-2(β=1,25,C=1,0)	*1507*	*0,0597*	*0,0823*	*0,1800*	*0,1280*	*0,1240*	*0,1080*	*0,0748*
TF_IDF	*2553*	*0,2426*	*0,2851*	*0,5800*	***0,5520***	*0,4600*	***0,3460***	***0,2084***
BM25	*2536*	*0,2418*	***0,2883***	*0,5800*	*0,5360*	*0,4520*	*0,3425*	*0,2076*
InL2	*2516*	***0,2439***	*0,2830*	***0,6400***	*0,5280*	***0,4680***	*0,3280*	*0,2040*
In_expB2	***2594***	*0,2389*	*0,2739*	*0,6200*	*0,5320*	*0,4460*	*0,3327*	*0,2068*

Çizelge 5.24 Luhn tabanlı TFxIDF ($IDF=idf^2$) modellerin " tüm konu" sorgu tipindeki TREC-6 başarım sonuçları

Konular: 301-350, Alakalı Belgelerin Sayısı: 4611								
	RR	**MAP**	**R-P**	**P@1**	**P@5**	**P@10**	**P@30**	**P@100**
WTF-1*(Z)	*1649*	*0,1192*	*0,1479*	*0,2400*	*0,2560*	*0,2040*	*0,1607*	*0,1072*
WTF-2*(Z)	*1699*	*0,1295*	*0,1669*	*0,2000*	*0,2480*	*0,2060*	*0,1593*	*0,1130*
WTF-1*(M1)	*1247*	*0,0444*	*0,0680*	*0,1400*	*0,0880*	*0,0860*	*0,0893*	*0,0696*
WTF-2*(M1)	*1518*	*0,0726*	*0,0888*	*0,1400*	*0,0960*	*0,1040*	*0,1053*	*0,0828*
WTF-1*(M2)	*1971*	*0,1776*	*0,2023*	*0,3000*	*0,2960*	*0,2580*	*0,1980*	*0,1312*
WTF-2*(M2)	*1896*	*0,1644*	*0,1992*	*0,2800*	*0,2640*	*0,2420*	*0,1840*	*0,1248*
WTF-1*(β=1,27,C=1,15)	*1921*	*0,1772*	*0,2072*	*0,3200*	*0,3200*	*0,2480*	*0,1887*	*0,1302*
WTF-2*(β=1,27,C=1,15)	*1889*	*0,1746*	*0,1997*	***0,3400***	*0,2880*	*0,2440*	*0,1893*	*0,1282*
WTF-1*(β=1,25,C=1,0)	*2021*	*0,1996*	*0,2247*	*0,4400*	*0,3440*	*0,2740*	*0,2073*	*0,1378*
WTF-2*(β=1,25,C=1,0)	*1984*	*0,1956*	*0,2275*	*0,4200*	*0,3160*	*0,2800*	*0,2033*	*0,1354*
TF_IDF	*2553*	*0,2426*	*0,2851*	*0,5800*	***0,5520***	*0,4600*	***0,3460***	***0,2084***
BM25	*2536*	*0,2418*	***0,2883***	*0,5800*	*0,5360*	*0,4520*	*0,3425*	*0,2076*
InL2	*2516*	***0,2439***	*0,2830*	***0,6400***	*0,5280*	***0,4680***	*0,3280*	*0,2040*
In_expB2	***2594***	*0,2389*	*0,2739*	*0,6200*	*0,5320*	*0,4460*	*0,3327*	*0,2068*

Çizelge 5.25 Luhn tabanlı TFxIDF modellerin " tüm konu" sorgu tipindeki TREC-7 başarım sonuçları

Konular: 351-400, Alakalı Belgelerin Sayısı: 4674								
	RR	**MAP**	**R-P**	**P@1**	**P@5**	**P@10**	**P@30**	**P@100**
WTF-1(Z)	*1782*	*0,0904*	*0,1283*	*0,3200*	*0,2560*	*0,2360*	*0,1687*	*0,1128*
WTF-2(Z)	*1791*	*0,0942*	*0,1308*	*0,2600*	*0,2360*	*0,2100*	*0,1653*	*0,1110*
WTF-1(M1)	*1771*	*0,0747*	*0,1160*	*0,1600*	*0,1440*	*0,1360*	*0,1167*	*0,0976*
WTF-2(M1)	*1809*	*0,0865*	*0,1196*	*0,2200*	*0,1680*	*0,1560*	*0,1360*	*0,1052*
WTF-1(M2)	*2012*	*0,1267*	*0,1632*	*0,2800*	*0,2920*	*0,2640*	*0,1987*	*0,1310*
WTF-2(M2)	*1952*	*0,1179*	*0,1538*	*0,3000*	*0,2640*	*0,2340*	*0,1840*	*0,1210*
WTF-1(β=1,27,C=1,15)	*2048*	*0,1378*	*0,1699*	*0,3600*	*0,3080*	*0,2900*	*0,2173*	*0,1400*
WTF-2(β=1,27,C=1,15)	*1969*	*0,1228*	*0,1604*	*0,3000*	*0,2920*	*0,2420*	*0,1987*	*0,1264*
WTF-1(β=1,25,C=1,0)	*2059*	*0,1365*	*0,1728*	*0,3200*	*0,3160*	*0,2780*	*0,2167*	*0,1354*
WTF-2(β=1,25,C=1,0)	*1960*	*0,1229*	*0,1587*	*0,3000*	*0,2960*	*0,2440*	*0,1940*	*0,1240*
TF_IDF	*2670*	*0,2129*	*0,2677*	*0,5800*	*0,4920*	*0,4720*	*0,3507*	*0,2156*
BM25	*2709*	*0,2170*	*0,2634*	***0,6400***	*0,5120*	*0,4780*	*0,3520*	*0,2160*
InL2	*2712*	***0,2274***	*0,2709*	*0,6000*	*0,5360*	*0,4760*	*0,3580*	*0,2170*
In_expB2	***2785***	*0,2427*	***0,2803***	*0,6000*	***0,5680***	***0,5080***	***0,3807***	***0,2282***

Çizelge 5.26 Luhn tabanlı TFxIDF ($IDF=idf^2$) modellerin " tüm konu" sorgu tipindeki TREC-7 başarım sonuçları

Konular: 351-400, Alakalı Belgelerin Sayısı: 4674								
	RR	**MAP**	**R-P**	**P@1**	**P@5**	**P@10**	**P@30**	**P@100**
WTF-1*(Z)	*2206*	*0,1525*	*0,1885*	*0,5000*	*0,3880*	*0,3380*	*0,2473*	*0,1574*
WTF-2*(Z)	*2190*	*0,1598*	*0,1940*	*0,4400*	*0,3850*	*0,3220*	*0,2453*	*0,1590*
WTF-1*(M1)	*1808*	*0,0852*	*0,1305*	*0,2200*	*0,2000*	*0,1760*	*0,1520*	*0,0974*
WTF-2*(M1)	*2045*	*0,1234*	*0,1601*	*0,2200*	*0,2320*	*0,2160*	*0,1827*	*0,1288*
WTF-1*(M2)	*2352*	*0,1887*	*0,2260*	*0,4400*	*0,4280*	*0,3660*	*0,2847*	*0,1788*
WTF-2*(M2)	*2295*	*0,1794*	*0,2090*	*0,4600*	*0,3720*	*0,3320*	*0,2707*	*0,1722*
WTF-1*(β=1,27,C=1,15)	*2364*	*0,1966*	*0,2294*	*0,5400*	*0,4280*	*0,3920*	*0,2967*	*0,1830*
WTF-2*(β=1,27,C=1,15)	*2306*	*0,1822*	*0,2189*	***0,4800***	*0,3880*	*0,3380*	*0,2793*	*0,1764*
WTF-1*(β=1,25,C=1,0)	*2376*	*0,1976*	*0,2301*	*0,5200*	*0,4400*	*0,3800*	*0,2940*	*0,1844*
WTF-2*(β=1,25,C=1,0)	*2306*	*0,1857*	*0,2210*	*0,4600*	*0,3960*	*0,3440*	*0,2733*	*0,1744*
TF_IDF	*2670*	*0,2129*	*0,2677*	*0,5800*	*0,4920*	*0,4720*	*0,3507*	*0,2156*
BM25	*2709*	*0,2170*	*0,2634*	***0,6400***	*0,5120*	*0,4780*	*0,3520*	*0,2160*
InL2	*2712*	***0,2274***	*0,2709*	*0,6000*	*0,5360*	*0,4760*	*0,3580*	*0,2170*
In_expB2	***2785***	*0,2427*	***0,2803***	*0,6000*	***0,5680***	***0,5080***	***0,3807***	***0,2282***

Çizelge 5.27 Luhn tabanlı TFxIDF modellerin " tüm konu" sorgu tipindeki TREC-8 başarım sonuçları

Konular: 401-450, Alakalı Belgelerin Sayısı: 4728								
	RR	**MAP**	**R-P**	**P@1**	**P@5**	**P@10**	**P@30**	**P@100**
WTF-1(Z)	*2199*	*0,1257*	*0,1711*	*0,3000*	*0,2800*	*0,2480*	*0,1920*	*0,1608*
WTF-2(Z)	*2223*	*0,1219*	*0,1649*	*0,3000*	*0,2280*	*0,2240*	*0,1720*	*0,1528*
WTF-1(M1)	*2247*	*0,1139*	*0,1516*	*0,1800*	*0,1600*	*0,1580*	*0,1467*	*0,1134*
WTF-2(M1)	*2261*	*0,1189*	*0,1587*	*0,1800*	*0,1920*	*0,1760*	*0,1527*	*0,1164*
WTF-1(M2)	*2454*	*0,1616*	*0,2083*	*0,3600*	*0,2840*	*0,2760*	*0,2053*	*0,1458*
WTF-2(M2)	*2398*	*0,1499*	*0,1957*	*0,3200*	*0,2640*	*0,2500*	*0,1833*	*0,1370*
WTF-1(β=1,27,C=1,15)	*2502*	*0,1635*	*0,2144*	*0,4200*	*0,2720*	*0,2900*	*0,2113*	*0,1560*
WTF-2(β=1,27,C=1,15)	*2407*	*0,1516*	*0,1976*	*0,3600*	*0,2640*	*0,2520*	*0,1840*	*0,1386*
WTF-1(β=1,25,C=1,0)	*2469*	*0,1602*	*0,2108*	*0,3800*	*0,2840*	*0,2740*	*0,2020*	*0,1512*
WTF-2(β=1,25,C=1,0)	*2380*	*0,1462*	*0,1884*	*0,3200*	*0,2560*	*0,2500*	*0,1820*	*0,1358*
TF_IDF	*2915*	*0,2499*	*0,2983*	*0,6800*	*0,5440*	*0,4800*	*0,3587*	*0,2298*
BM25	*2948*	*0,2512*	*0,3023*	*0,6600*	*0,5440*	*0,4800*	*0,3580*	*0,2286*
InL2	*2965*	*0,2641*	*0,3110*	*0,6600*	*0,5320*	*0,4940*	*0,3767*	*0,2330*
In_expB2	***3141***	***0,2758***	***0,3180***	***0,7200***	***0,5560***	***0,4980***	***0,3793***	***0,2400***

Çizelge 5.28 Luhn tabanlı TFxIDF ($IDF=idf^2$) modellerin " tüm konu" sorgu tipindeki TREC-8 başarım sonuçları

Konular: 401-450, Alakalı Belgelerin Sayısı: 4728								
	RR	**MAP**	**R-P**	**P@1**	**P@5**	**P@10**	**P@30**	**P@100**
WTF-1*(Z)	*2647*	*0,1867*	*0,2379*	*0,3800*	*0,3600*	*0,3420*	*0,2593*	*0,1740*
WTF-2*(Z)	*2655*	*0,1859*	*0,2281*	*0,3400*	*0,3280*	*0,3140*	*0,2507*	*0,1664*
WTF-1*(M1)	*2194*	*0,1084*	*0,1561*	*0,2400*	*0,1720*	*0,1700*	*0,1507*	*0,1198*
WTF-2*(M1)	*2509*	*0,1489*	*0,1995*	*0,2800*	*0,2040*	*0,2160*	*0,1787*	*0,1394*
WTF-1*(M2)	*2882*	*0,2217*	*0,2669*	*0,4000*	*0,3800*	*0,3520*	*0,2907*	*0,1926*
WTF-2*(M2)	*2789*	*0,2098*	*0,2526*	*0,3200*	*0,3440*	*0,3220*	*0,2633*	*0,1816*
WTF-1*(β=1,27,C=1,15)	*2886*	*0,2249*	*0,2691*	*0,4400*	*0,3720*	*0,3600*	*0,2927*	*0,1958*
WTF-2*(β=1,27,C=1,15)	*2795*	*0,2105*	*0,2556*	*0,4600*	*0,3360*	*0,3340*	*0,2693*	*0,1836*
WTF-1*(β=1,25,C=1,0)	*2866*	*0,2220*	*0,2657*	*0,4800*	*0,3760*	*0,3560*	*0,2853*	*0,1948*
WTF-2*(β=1,25,C=1,0)	*2762*	*0,2075*	*0,2476*	*0,4000*	*0,3450*	*0,3360*	*0,2613*	*0,1798*
TF_IDF	*2915*	*0,2499*	*0,2983*	*0,6800*	*0,5440*	*0,4800*	*0,3587*	*0,2298*
BM25	*2948*	*0,2512*	*0,3023*	*0,6600*	*0,5440*	*0,4800*	*0,3580*	*0,2286*
InL2	*2965*	*0,2641*	*0,3110*	*0,6600*	*0,5320*	*0,4940*	*0,3767*	*0,2330*
In_expB2	***3141***	***0,2758***	***0,3180***	***0,7200***	***0,5560***	***0,4980***	***0,3793***	***0,2400***

5.3.3 Gözlemlerin özeti

Z puanları kullanılarak ortaya konulan TF modellerinin farklı α değerli ile gerçekleştirdikleri başarımlarının incelenmesi sonucunda Luhn'un iddiasını destekleyici bulgular gözlenmiştir. TREC-6, TREC-7 ve TREC-8 anlık-sorgu izlerinde en önemli kelime; yani anlamsal olarak en yüksek kelimenin frekansının

tahmini için α'nın değer aralığı [0,8 1,2] olarak bulunmuş ve α = 1,0 değeri seçilmiştir. Ancak bu modeller –TF1(Z) ve TF-2(Z)- ve karşılık gelen TFxIDF modelleri (WTF-1 ve WTF-2) bilgi erişimde başarısız olmuşlardır. Temel nedenin z puanları ile normalleştimenin yetersizliğinden kaynaklandığı düşünülmektedir.

Luhn esasında ortaya konulan diğer modellerinden; yani medyan tabanlı modellerin medyanlarının standart sapması ile normalleştirilmiş olanları da yine bilgi erişim açısından yetersiz olmuşlardır. Ancak medyan ile normalleştirilen modeller özellikle "çok kısa" sorgu tipinde mevcut yöntemlerle benzer başarımlar göstermişlerdir. Hatta bu modellerin yalnızca TF başarımları (Bkz. Çizelge 5.11, Çizelge 5.13, Çizelge 5.15) BM25 ve TF-IDF ağırlıklandırmalarıyla elde edilen başarımlar ile paralleldir. Başarımı makul bu TF modellerinin belge topluluğu istatistiklerinden bağımsız olması özelliği dolayısıyla, *dağıtık ve içeriği değişebilen belge topluluklarında belirli sayıda anahtar kelime ile bilgi erişim uygulamaları* için uygun bir seçenektir. Medyan tabanlı TFxIDF şemasına uygun modeller ise "çok kısa" ve "kısa" sorgu tipinde başarılı olmuşlar ise de başarım sorgu terim kümesinin büyümesiyle azalmıştır.

6 TREC AKTİF KATILIM BAŞARIM SONUÇLARI

TREC-2009 konferansına 19 ülkeden 67 araştırma grubu ve TREC-2010 konferansına 19 ülkeden 65 araştırma grubu katılmıştır. Ege ve Muğla üniversitesi olarak *irra* grup adıyla katılınan bu konferanslarda: TREC-2009 web izi ile milyon sorgu izine, ve TREC-2010 web izine yürütümler sunulmuştur. TREC-2009[8] ve TREC-2010[9] sonuç raporları web sayfalarından ilgili ize katılan gruplarla sundukları yürütümler hakkındaki detaylı bilgiye ulaşmak mümkündür. Katılınan izlere sunulan yürütümler ve elde edilen başarım değerlendirmeleri ilerleyen bölümlerde anlatılmaktadır.

6.1 Sunulan Yürütümler

6.1.1 TREC-2009 İzlerine Sunulan Yürütümler

TREC-2009 izlerinde 3 farklı ağırlıklandırma fonksiyonu kullanılmıştır. irra1, irra2 ve irra3 olarak adlandırılan bu fonksiyonlar DFIxIDF yapısında olup, Bölüm 4.2'deki denklemler ile tanımlanan DFI bileşeninin IDF bileşeni ile çarpılması yoluyla oluşturulmuştur. Kısaca bu fonsiyonlar:

irra1: DFI bileşeni olarak Dr. Taner Dinçer'in kendi geliştirdiği DFIxIDF modeli.

irra2: DFI bileşeni olarak Denklem 4.9.b'yi kullanan DFIxIDF modeli.

irra3: DFI bileşeni olarak Denklem 4.9.a'yı kullanan DFIxIDF modeli.

Milyon sorgu izi ile web izinde anlık-sorgu ve çeşitlilik görevlerinde temel olarak yukarıda tanımlanan bu üç ağırlıklandırma fonksiyonu kullanılmıştır ve yürütüm adlarının başında kullanılan fonksiyon belirtilmektedir. 2009 TREC web izinde her görev için üç yürütüm sunulmuştur: anlık-sorgu görevi için gerçekleştirilen yürütümler irra1a, irra2a ve irra3a; çeşitlilik görevi için gerçekleştirilen yürütümler irra1d, irra2d ve irra3d ile adlandırılmaktadır. Bu yürütümlerin hiçbirinde temel ağırlıklandırma fonksiyonlardan başka bir bilgi (meta, bağlantı bilgisi gibi) veya yöntem (sorgu genişletme, alaka geri-beslemesi vb gibi.) kullanılmamıştır. Sadece çeşitlilik görevinde temel ağırlıklandırma fonksiyonlarının haricinde aynı sunucudan (URL) gelen birden fazla web sayfası/belge filtrelenmiştir. Milyon sorgu izi için ise toplam beş yürütüm gerçekleştirilmiş/sunulmuştur. Bunlardan irra1mqa ve irra2mqa sadece temel

[8] http://trec.nist.gov/pubs/trec18/t18_proceedings.html
[9] http://trec.nist.gov/pubs/trec19/t19_proceedings.html

fonksiyonları içerirken irra1mqd, irra2mqd ve irra3mqd yürütümleri ek olarak web izindeki çeşitlilik görevi için gerçekleştirilen yürütümlerde kullanılan aynı belge filtrelenmesi tekniğini içermektedir.

6.1.2 TREC-2010 İzlerine Sunulan Yürütümler

TREC 2010 web izinde sadece Kategori-B anlık-sorgu görevine yürütümler sunulmuştur. Bu yürütümlerde kullanılan DFIxIDF ağırlıklandırma modeli: DFI bileşeni olarak Denklem 2.6'daki $(DFI)_{ij}$ ile IDF bileşeni olarak Sparck Jones'un (1972) hesaplamasından oluşmaktadır. Bu ağırlıklandırma modelini kullanan ancak ek yapılar ile farklılaştırılan 3 ayrı bilgi erişim sisteminin yürütümler irra10b, irra10hp, irra10rob olarak adlandırılmıştır.

irra10b'nin gerçekleştiği sistem baz sistem olarak gördüğümüz sistemdir. Bu sistem ağırlıklandırma modelinin yanında Cormack vd. [13] geliştirdiği spam filtreleme yöntemini ve kelime grubu bulma yöntemini içerir. Kelime grubu bulma yöntemi için ise TERRIER (University of Glasgow, 2010) kütüphanesi ile gelen n-gram tabanlı gerçeklenme kullanılmıştır. Bu işlem aslen sorguda geçen kelimelerin belgeler içindeki yakınlıklarına göre puanlama yapmaktadır. Baz yürütümümüz olan irra10b'nin erişim sisteminde başka bir bilgi (meta, bağlantı bilgisi gibi) veya yöntem (sorgu genişletme, alaka geri-beslemesi vb gibi.) kullanılmamıştır. Diğer iki yürütümün (irra10hp ve irra10rob) elde edildiği sistemler ise temel ağırlıklandırma yöntemimiz yanında daha karmaşık teknikler içermektedir.

6.2 Başarım Değerlendirmeleri

Bu bölümde, TREC-2009 milyon sorgu ve web izleri ile TREC-2010 web izine sunulan yürütümlerimizin başarımları ve diğer katılımcıların yürütümleri ile kıyaslanmaları verilmektedir. Bu izlerde elde edilen başarımlar yalnızca ilgili izlerde belirlenen başarım ölçütleri açısından değerlendirilmiştir. Ayrıca TREC-2009 izlerindeki yürütümlerin/sistemlerin; yani temellerindeki modellerin *Temel Bileşenler Analizi* (İng. *P*rinciple *C*omponent *A*nalysis, kıs. PCA) (Dinçer, 2007) tekniğiyle sorgu uzayındaki göreceli değerlendirilmesi gibi sorgu tabanlı incelemeler TREC konferansına sunulan çalışmada (Dinçer et al., 2009) bulunmaktadır.

6.2.1 TREC-2009 Milyon sorgu izi sonuçları

Bu ize uygun olarak gerçekleştirdiğimiz 5 yürütüm de 1000 adet sorgu kullanarak yapılmıştır. Yargılama sürecine katılmak üzere 40000 sorgudan toplam 687 sorgu seçilmiştir. Bu 687 elemanlı temel sorgu kümesinde bizim kullandığımız sorgulardan ancak 310 tanesi bulunmaktadır. Bununla birlikte tüm yürütümlerde ortak olan 146 sorgu bulunmaktadır.

Bizim yürütümlerimizde kullandığımız sorgulardan değerlendirilen 310 sorguya göre: statAP ile tahmin edilen ortalama averaj duyarlılık (gösterim. MAP), R-Duyarlılık (gösterim. R-P) ve sıralı listedeki ilk 10, 30, 50, 100 belge için duyarlılık (gösterim sırasıyla, P@10, P@30, P@50, P@100) ölçümleri Çizelge 6.1'de verilmiştir. StatAP tabanlı tahmin edilen bu ölçütlerin matematiksel ifadeleri J.A. Aslam ve V. Pavlu'nun (2007) çalışmasında bulunmaktadır.

Çizelge 6.1 Değerlendirilmiş 310 sorgudaki statAP başarım tahminleri

Yürütümler	statAP ile Tahmin Edilen					
	MAP	***R-P***	***P@10***	***P@30***	***P@50***	***P@100***
irra1mqa	0,1926	0,2790	0,2731	0,2985	0,2948	0,2699
irra1mqd	0,1443	0,2372	0,2667	0,2684	0,2540	0,2422
irra2mqa	0,1525	0,2316	0,2604	0,2415	0,2223	0,2014
irra2mqd	0,1211	0,2013	0,2553	0,2269	0,2032	0,2045
irra3mqd	0,1508	0,2367	0,2565	0,2767	0,2658	0,2353

NIST tarafından değerlendirilmiş olan 310 sorgu için en iyi başarım sonuçları tüm ölçütler açısından irra1mqa ile elde edilmiştir. irra1mqa ile irra2mqa yürütümlerinin aynı URL'den gelen sayfalarının filtrelenmesi işleminin içeren sürümleri olan irra1mqd ile irra2mqd yürütümlerinde tahmini MAP açısından yaklaşık sırasıyla %33 ve %26'lık bir başarım azalması gözlenmektedir. Ayrıca irra3'ün URL filtrelenmiş hali olan irra3mqd ile irra2mqa yakın başarımlara sahiptir. Bu sebeplerden dolayı, yürütümü sunulmamış olan irra3'ün sadece temel fonksiyonun kullanılması irra2'den (temel fonksiyonun yürütümü olan irra2mqa'dan) daha yüksek başarım sağlayacağı beklenmelidir.

Bu ize katılan gruplardan gelen toplam 35 yürütümün tüm yürütümler için ortak olan 146 sorgu üzerinden MTC ve statAP ile tahmin edilen MAP değerleri – kısaca mtcMAP ve statMAP olarak adlandırılmış- Çizelge 6.2'de bulunmaktadır. Ayrıca aynı tabloda, sistemler/yürütümlerin yine bu mtcMAP ve statMAP değerlerine göre sıralanmasıyla elde edilen sıra numaraları verilmiştir. irra1mqa ile irra2mqa'nın statMAP'a göre genel sıralamada mtcMAP'tan daha yüksek iken (sırasıyla, statMAP'ta 13üncü ve 18inci sırada iken mtcMAP'ta 18inci ve 25inci sırada), URL filtrelemesi kullanılan yürütümler yaklaşık aynı göreceli başarımı göstermiştir. StatMAP açısından sadece temel fonksiyonu kullanan her iki temel yürütümde de ortalama bir başarım elde edildiği gözlenmektedir.

Çizelge 6.2 Ortak olan 146 sorgu üzerinden tüm yürütümlerin MTC ve statAP'ye göre tahmini MAP başarımları

Yürütüm	statMAP	Sıra	mtcMAP	Sıra	Yürütüm	statMAP	Sıra	mtcMAP	Sıra
UDMQAxQEWel	0,227	1	0,124	1	*irra2mqa*	*0,132*	*18*	*0,049*	*25*
uogTRMQdph40	0,198	2	0,089	4	Sab9mq1bf1	0,130	19	0,071	14
uogTRMQdpA10	0,195	3	0,087	5	iiithExpQry	0,130	20	0,055	21
UDMQAxBL	0,192	4	0,079	8	Sab9mq2bf1	0,127	21	0,062	19
uiuc09GProx	0,183	5	0,086	6	udelIndPR	0,123	22	0,054	22
uiuc09Adpt	0,180	6	0,079	9	Sab9mq1bf4	0,122	23	0,067	17
uiuc09MProx	0,179	7	0,082	7	Sab9mqBase1	0,111	24	0,052	23
uiuc09RegQL	0,175	8	0,079	10	Sab9mqBase4	0,111	25	0,052	24
udelIndDM	0,173	9	0,072	13	*irra3mqd*	*0,105*	*26*	*0,048*	*27*
uiuc09KL	0,171	10	0,071	15	*irra1mqd*	*0,103*	*27*	*0,048*	*26*
udelIndSP	0,169	11	0,075	11	iiithAuEQ	0,097	28	0,045	28
udelIndri	0,169	12	0,069	16	*irra2mqd*	*0,097*	*29*	*0,040*	*30*
irra1mqa	*0,156*	*13*	*0,066*	18	iiithAuthPN	0,096	30	0,045	29
udelIndRM	0,155	14	0,073	12	NeuSvmBase	0,079	31	0,024	32
UDMQAxQE	0,149	15	0,099	2	NeuSVMHE	0,078	32	0,023	34
UDMQAxBLlink	0,144	16	0,059	20	NeuSvmStefan	0,077	33	0,034	31
UDMQAxQEWP	0,133	17	0,090	3	NeuSvmPR	0,076	34	0,023	33
					NeuSvmPRHE	0,073	35	0,022	35

6.2.2 TREC-2009 Web izi sonuçları

Web izi sonuçları anlık-sorgu ve çeşitlilik olmak üzere iki görev açısından değerlendirilmiştir.

6.2.2.1 Anlık-sorgu görevi

Gerçekleştirilen/sunulan sistem yürütümlerimizin MTC ile tahmin edilen başarım değerleri Çizelge 6.3'te verilmiştir. MAP ve R-Prec açısından irra1a ve irra3 birbirine yakın ve irra2a'dan daha yüksek başarım göstermelerine rağmen irra2a ilk 5 belgede tahmini olarak daha yüksek duyarlılığa sahiptir.

Yürütümlerden irra1a ve irra3a'nın duyarlılık değerleri belge sayısıyla birlikte artarken, irra2a ise belge sayısı artışına göre belirli bir değişim göstermeyerek sabit bir başarım gözlenmektedir.

Çizelge 6.3 MTC ile tahmin edilen IRRA yürütümleri başarımları

Yürütüm	MTC ile Tahmin Edilen							
	MAP	***R-P***	***P@5***	***P@10***	***P@15***	***P@20***	***P@30***	***P@100***
irra1a	0,0375	0,0968	0,2167	0,2780	0,3017	0,3122	0,3401	0,3366
irra2a	0,0274	0,0860	0,2811	0,2756	0,2865	0,2793	0,2846	0,2865
irra3a	0,0379	0,0971	0,2117	0,2810	0,3027	0,3197	0,3399	0,3420

Web izi anlık-sorgu görevi için sunulan yürütümlerden MAP değerine göre en yüksek başarımı gösteren ilk 6 yürütüm Çizelge 6.4'te verilmiştir. Çizelge'de gösterilen en düşük değerdeki MAP, P@5 ve P@20 sırasıyla (yaklaşık) 0,0425; 0,2772; 0,3314 iken bizim yürütümlerimizde elde edilen en yüksek değerler (yaklaşık) ise 0,0379; 0,2811 ve 0,3197 olmakla birlikte, özellikle MAP cinsinden yaklaşık %10-11 daha az başarıma sahiptir.

Çizelge 6.4 MTC ile tahmin edilen en iyi yürütüm başarımları

Grup	Yürütüm	MTC ile Tahmin Edilen			
		MAP	***P@5***	***P@10***	***P@20***
UMD	UMHOOsd	0,0476	0,3458	0,3999	0,4098
UDel	udelIndDRSP	0,0471	0,2772	0,3561	0,3891
uogTr	uogTrdphCEwP	0,0460	0,5419	0,5282	0,5223
NEU	NeuLMWeb600	0,0442	0,3950	0,4006	0,4065
ICTNET	ICTNETADRun3	0,0433	0,4421	0,4436	0,4424
EceUdel	UDWAxBL	0,0425	0,3340	0,3314	0,3371

Bu görevde her sorgu/konu bazında tüm sistem yürütümlerinden gözlenen en yüksek, ortalama ve en düşük tahmini MAP değerleri EK-4'de verilmiştir. 20 numaralı konuda hiçbir sistem yürütümü alakalı bir belge bulamamıştır. Sistemlerin göreceli olarak başarımları düşükte olsa 12 ve 26 numaralı konularda irra1a, 34 ve 45 numaralı konularda irra2a, 12 numaralı konuda irra3a en yüksek başarımı göstermiştir.

6.2.2.2 Çeşitlilik görevi

Bu görev için sunulan sistem yürütümlerimizin ilk 10 belgedeki α-nDCG ve Prec-IA değerlerine göre başarımları Çizelge 6.5'te verilmiştir. Aynı URL'den gelen web sayfalarına filtreleme işlemi başarımları düşürmüştür. Ayrıca Prec-IA@10'da tüm IRRA yürütümleri için benzer değerler elde edilmesine karşılık irra2d yürütümünü bu görevde α-nDCG@10 açısından en yüksek başarıma ulaşmıştır.

Çizelge 6.5 IRRA yürütümleri başarımları (ilk 10 belgedeki α-nDCG ve Prec-IA değerlerine göre)

Yürütümler	α-nDCG@10	Prec-IA@10
irra1d	0,131	0,063
irra2d	0,161	0,060
irra3d	0,130	0,061

Web izi çeşitlilik görevi için tüm katılımcılar tarafından sunulan yürütümlerden α-nDCG@10 değerine göre en yüksek başarımı gösteren ilk 6 yürütüm Çizelge 6.6'da verilmiştir. Çizelge'de gösterilen en düşük değerdeki α-nDCG@10 ile Prec-IA@10 sırasıyla 0,247 ile 0,079 iken bizim yürütümlerimizde elde edilen en yüksek değerler 0,161 ile 0,063 olmakla birlikte, özellikle α-nDCG@10 cinsinden yaklaşık %55 daha az başarıma sahiptir.

Çizelge 6.6 En iyi yürütüm başarımları (α-nDCG ve Prec-IA değerlerine göre)

Grup	Yürütüm	α-nDCG			Prec-IA		
		@5	*@10*	*@20*	*@5*	*@10*	*@20*
Waterloo	uwgym	0,335	0,369	0,400	0,162	0,144	0,122
uogTr	uogTrDYCcsB	0,253	0,282	0,308	0,142	0,132	0,127
ICTNET	ICTNETDivR3	0,251	0,272	0,301	0,104	0,095	0,092
Amsterdam	UamsDancTFb1	0,232	0,250	0,281	0,086	0,079	0,071
CSIUCD	UCDSIFTdiv	0,212	0,249	0,278	0,112	0,121	0,115
LU_WUME	wume1	0,220	0,247	0,279	0,121	0,113	0,108

6.2.3 TREC-2010 Web izi sonuçları

Bu bölümde, TREC-2010 web izi anlık-sorgu görevinde sunulan baz yürütümümüz irra10b'nin diğer katılımcıların yürütümleri ile başarım kıyaslanmaları verilmektedir. Diğer yürütümlerinde bulunduğu NIST tarafından gönderilen başarım sonuçları EK-5'de verilmiştir. Elde edilen başarımlar toplam 36 sorgu için ve yalnızca ilgili izin görevinde belirlenen başarım ölçütleri açısından değerlendirilmiştir.

Bu görev için değerlendirmede kullanılacak öncelikli başarı ölçütü olarak ERR seçilmiştir. Kategori-B'ye yürütüm sunan her gurubun ERR@20'ye göre en iyi yürütüm sonuçları ile bizim yürütümümüz irra10b'nin erişim başarım sonuçları Çizelge 6.7'de verilmiştir. ERR@20'ye ek olarak nDCG@20, MAP ve P@20 değerleri de tabloda gösterilmiştir.

Çizelge 6.7 TREC 2010 Web izi anlık-sorgu görevi Kategori-B başarım sonuçları

Sıra	Grup	Yürütüm	ERR@20	nDCG@20	MAP	P@20
1	isi	ivoryL2b	**0,1492**	0,2411	0,1467	0,4208
2	irra	irra10b	0,1319	**0,2703**	0,1360	**0,4833**
3	uogTr	uogTrB67LTS	0,1308	0,2157	**0,1499**	0,4083
4	blv79	blv79y00shnk	0,1255	0,2090	0,1498	0,3903
5	UAmsterdam	UAMSA10mSF30	0,1186	0,1599	0,0473	0,2708
6	udel	udelIndriWP	0,1075	0,2225	0,0749	0,3722
7	PKUSEWM	pkusewm1	0,0996	0,1733	0,1128	0,3417
8	UCDSIFT	UCDSIFTSlide	0,0970	0,1707	0,1145	0,3389
9	MediaFutures	MF1	0,0782	0,1381	0,0945	0,2875
10	york	york10wA3	0,0680	0,1323	0,0998	0,2861
11	uottowa	Dfalah2010	0,0643	0,0674	0,0110	0,1194

Yürütümümüz irra10b nDCG@20 ve P@20 ölçütlerinde en yüksek değerlere sahipken, *ERR@20*'de ikinci ve MAP'tada dördüncü en iyi değere sahiptir. irra10b ile IRRA grubu olarak anlık-sorgu görevinde Kategori-B'ye katılan 11 grup arasında en yüksek başarımlı (MAP dışındaki ölçütlerde) yürütüme sahip ilk 2 grup arasına girilmiştir. Ayrıca MAP ölçütü açısından ise en yüksek yürütüme sahip dördüncü grup olunmuştur.

Anlık-sorgu görevine katılan tüm gruplar açısından da irra10b başarılı olmuştur. Kategori-A ve Kategori-B'ye katılan tüm grupların yürütümlerinden ERR@20'ye göre en iyi 10 yürütüm Çizelge 6.8'de verilmiştir.

Çizelge 6.8 TREC 2010 Web izi anlık-sorgu görevi Kategori-A ve B'de ERR@20'ye göre ilk 10 yürütüm başarım sonuçları

Sıra	Grup	Yürütüm	Kategori	ERR@20	nDCG@20	MAP	P@20
1	*Temel**	uwgym	A	**0,1720**	0,2590	0,0720	0,4150
2	msrsv	msrsv3	A	0,1650	0,2590	0,0840	0,3680
3	isi	ivoryL2b	B	0,1492	0,2411	0,1467	0,4208
4	umass	umassSDMW	A	0,1430	**0,3100**	**0,1570**	**0,5320**
5	irra	irra10b	B	0,1319	0,2703	0,1360	0,4833
6	uogTr	uogTrB67LTS	B	0,1308	0,2157	0,1499	0,4083
7	THUIR	THUIR10Str	A	0,1300	0,2110	0,1200	0,3801
8	blv79	blv79y00shnk	B	0,1255	0,2090	0,1498	0,3903
9	Unimelb	UMa10IASF	A	0,1240	0,1910	0,0870	0,3190
10	UAmsterdam	UAMSA10mSF30	B	0,1186	0,1599	0,0473	0,2708

Çizelge 3.8'te gösterilen *Temel** yürütüm sorguların web ortamında ticari bir bilgi erişim sistemi tarafından erişilmesi ile oluşturulmuştur. Ancak doğal olarak erişilen belgeler web izinde kullanılan derlemden farklı belgeler içermektedir. Bu yüzden ClueWeb09 derleminde olmayan belgeler çıkartılarak yürütümün belge listesi oluşturulmuştur. irra10b'nin elde ettiği 0,1319 değerindeki ERR@20 başarımı ile anlık-sorgu görevine katılan toplam 20 grup arasında beşinci olunmuştur. Ayrıca irra10b ile nDCG@20 ve P@20 ölçütleri açısından en yüksek başarıma sahip ikinci grup olunmuştur.

7 KATKI VE İLERDE YAPILMASI PLANLANAN ÇALIŞMALAR

Bu tez çalışmasında iki farklı fikir temelinde özgün istatistiksel indeks terim ağırlıklandırma yöntemleri geliştirilmiştir. Bunlardan ilki *İstatistiksel Bağımsızlık fikri* esasında terim ağırlıklandırmasına uygun *bağımsızlıktan sapma modelleri* olarak adlandırılmıştır. Diğeri ise *Luhn'un iddiası* esasında terim ağırlıklandırmasına uygun *Luhn tabanlı TFxIDF* olarak adlandırılmıştır. Bu modeller, yazılı belge erişim sahasında standart olarak kabul edilen TREC derlemlerinde; TREC-6, TREC-7 ve TREC-8 anlık-sorgu izi kapsamındaki belge ve sorgu kümelerinde deneysel olarak sınanmışlardır.

Bağımsızlıktan sapma modellerinin başarım sonuçlarının aralarında değerlendirmesi ile en yüksek başarım melez bağımsızlıktan sapma modellerinde gözlenmiştir. Ek olarak bir diğer kıyaslama da geliştirilen modeller ile mevcut ağırlıklandırma modelleri arasında gerçekleştirilmiştir. TREC anlık sorgu izleri için standart kabul edilen RR, MAP ve R-P ölçütlerine göre bakıldığında melez modeller genelde en yüksek bilgi erişim başarım değerlerini elde etmiştir. Sadece logaritmik dönüşüm içeren temel DFI modelleri ise yine genel olarak mevcut yöntemler ile benzer başarımlar göstermiştir. Özetle, DFI esasında sunulan bu modellerin mevcut yöntemlere göre ortaya koyduğu paralel (çoğu zaman yüksek) başarım, geliştirilen modellerin günümüzde ağırlıklandırma meselesinin çözümünde kullanılan mevcut modellere alternatif bir yaklaşım olabileceğini göstermektedir.

TF bileşeninin Luhn'un bakış açısına **tam olarak uygun** modellenmesi için Z-puanları kullanan yöntemin, sorguyu en iyi ifade eden kelimelerden oluşan; yani anlamsal olarak eşit ağırlıkta kabul edilebilecek "çok kısa" sorgu tipindeki en uygun alfa değeri bulunmasına yönelik yapılan deneyler: *bilgi erişim başarımının belli bir alfa değerinde en yüksek olduğunu göstererek Luhn'un iddiasını desteklemektedir*. Fakat gerek Z-puanları gerekse medyan temelindeki modellerin deney sonuçları, standart sapma ile normalleştirme işleminin pek başarılı olmadığını işaret etmektedir. Diğer bir deyişle, orta noktaya olan uzaklığın medyan cinsinden gösterilmesi (Medyan değeri ile normalleştirilmesi) ile başarım daha yüksek olmaktadır.

Luhn'un önem/ağırlıklandırma ilişkisin medyan açısından ifade edildiği modeller ile özellikle bunların TFxIDF şemasına uyumlu tipleri, yani erişilmek istenen bilgideki her kelimenin anlama katkısının *Sparck Jones'un idf'si (1972)*

ile temsil edilmesi “çok kısa” sorgularda mevcut yöntemlere alternatif olabilecek ağırlıklandırma fonksiyonları olduğunu göstermiştir. Aynı zamanda orta noktaya olan uzaklıkların karesine dayanan hesaplama yöntemi (TF-2 belirteçli denklem) yine uzaklığın medyan cinsinden ifade edildiği modellerde özellikle MAP ölçütü açısından daha iyi başarım göstermektedir. Ancak *idf*'nin gücü kısa ve tüm konu tiplerindeki sorgularda gittikçe azalmış, özellikle “tüm konu” için başarımı oldukça düşük elde edilmiştir. Bilgi erişimde *idf*'nin karesinin kullanılması ise başarımı biraz daha arttırdıysa da bu artış yeterli seviyede olmamıştır.

Bu çalışmadaki önemli bir kazanım da TREC çalıştaylarına 2009 ve 2010 yıllarında Muğla Üniversitesi ile birlikte katılım gerçekleştirilmesidir. Bilgi erişim sahasında önemli bir yeri olan TREC çalıştaylarına ilk katılan Türk grubu olmak ile birlikte, özgün bir indeks terim ağırlıklandırma yöntemine sahip bilgi erişim sistemiyle katılım ayrı bir kazanımdır. “TREC-2009 milyon sorgu ve web izlerinde” ile “TREC-2010 web izi anlık-sorgu görevinde” sunmuş olduğumuz yürütümlerde DFI tabanlı melez modeller kullanılmıştır. TREC-2009’da sunduğumuz yürütümler sadece indeks terim ağırlıklandırma işlemini gerçekleştiren ve başka ek bilgi kullanmayan bir bilgi erişim sisteminde yapılmıştır. Ve bu izlerde yalın bir bilgi erişim sistemiyle ortalama bir başarım gözlemlenmiştir. TREC-2010’daki BE sistemiz ise geliştirdiğimiz indeks terim ağırlıklandırma yöntemine ek olarak belgelerde spam filtrelemesi ve kelime grupları arama yöntemlerini içermektedir. Diğer sistemler tarafından da kullanılan bu iki genel yöntemi kullanan BE sistemimiz ile TREC-2010’daki en başarılı BE sistemleri arasına girilmiştir. Ayrıca, 2009 yılındaki izlerde başarım değerlendirilmesinin *tahmini* yöntemlerle gerçekleştirilmiş olması bu izlerdeki başarımımızın ölçüsünü doğru yansıtmamış olabilir. TREC-2010’da gerçekleştirilen tüm belgelerin değerlendirilmesi süreci 2009 yılındaki izlerde de uygulanması halinde bilgi erişim başarımlarımızın daha yüksek olacağı düşünülmektedir.

Bu bölümde anlatılanlar doğrultusunda tez çalışmasının katkılarını maddeler halinde özetleyecek olursak:

1) Başarımları yüksek olan mevcut yöntemlere alternatif olabilecek özgün indeks terim ağırlıklandırma yöntemleri geliştirilmiştir. Bu modellerin temel aldığı *"Bağımsızlıktan Sapma"* fikrinin bilgi erişim açısından uygun olduğu sonucuna varılmıştır.
2) Luhn'un kelimelerin önemi hakkındaki iddiasını bilgi erişim sahasında **tam** ve **biçimsel** olarak inceleyen ilk çalışmadır. İlgili deneyler ile Luhn'un iddialarını destekleyen bulgular elde edilmiştir. Sonuç olarak,

indeks terim ağırlıklandırma yöntemlerinin bu doğrultuda ilerlemesi için bir temel oluşturmuştur.

3) BE sahasında sistemlerin yarıştırıldığı uluslararası standart bir organizasyon olan TREC çalıştayına (2009 yılında) Türkiye'den ilk defa katılım gerçekleştirilmiştir. TREC-2009 organizasyonunda, üzerinde geliştirilen ağırlıklandırma yöntemlerini ek bir iyileştirme olmadan koşturulan BE sistemleri diğer sistemlere göre ortalama başarı yakalamıştır. TREC-2010 organizasyonunda ise, geliştirilen ağırlılandırma modeli üzerine eklenen bazı temel iyileştirme yöntemlerini kullanan sistemlerin diğer sistemlere göre başarılı olması ileri çalışmalar için teşvik edicidir. BE sahasında farklı fikir ve yaklaşımlar ortaya koyması ile yeni araştırma ufukları açacağı düşünülmektedir.

İleride yapılması planlanan çalışmalar ise şunlardır:

1) Luhn esasında geliştirilen TF modellerinin Sparck Jones'un *idf*'si (1972) ile uyumundaki problemden ötürü, IDF bileşeninin farklı hesaplamaları için BE başarımlarının sınanması.
2) Luhn'un iddiasını gerçekleştirecek alternatif yöntemlerin geliştirilmesi.
3) İndeks terim ağırlıklandırması dışında bilgi erişim sürecinde kullanılabilecek metin işleme, dilbilimsel, istatistiksel vb. gibi ek yöntemlerin getirdikleri başarım katkısının ileriki TREC'lerde aktif olarak sınanması.
4) Türkçe için geliştirdiğimiz indeks terim ağırlıklandırma hesaplamasını kullanan başarımı yüksek bir BE sistemi tasarlanması ve gerçeklenmesi.

TÜRKÇE-İNGİLİZCE TERİMLER SÖZLÜĞÜ

Terim	İngilizce Karşılığı
Alaka Geri Besleme	Relevance Feedback
Alanyazın	Literature
Anlam	Semantic
Anlık-sorgu	Adhoc
Anma	Recall
Ardıl-işlem	Post-processing
Artık-IDF	Residual-IDF
Averaj Duyarlık	Average Precision
Ayrıştırıcı	Parser
Bağımsızlıktan Sapma	Divergence from Randomness
Beklenen Karşıt Sırası	Expected Reciprocal Rank
Belge İndeks Listesi	Document Index File
Bilgi Erişim	Information Retrieval
Çatı	Framework
Çeşitlilik Görevi	Diversity Task
Çözümleme Gücü	Resolving Power
Değerlendirme	Evaluation
Dil Modellenmesi	Language Modelling
Doğal Dil İşleme	Natural Language Processing

TÜRKÇE-İNGİLİZCE TERİMLER SÖZLÜĞÜ (devam)

Terim	İngilizce Karşılığı
Doğrudan İndeks Listesi	Direct Index File
Durma Kelimeleri	Stopwords
Duyarlık	Precision
En Düşük Test Topluluğu	Minimal Test Collection
En Küçük Kareler Yaklaşımı	Least Square Approximation
Erişim	Retrieval
Erişim Görevi	Retrieval Task
Filitreleme	Filtering
Gövdeleyici	Stemmer
İkili Tercih	Binary Preference
İndeks Terim Ağırlıklandırma	Index Term Weighting
İndeksleme	Indexing
İz	Track
Ki-Kare	Chi-Square
Kümeleme	Clustering
Niyet Duyarlı	Intent Aware
Nokta Çarpımı	Dot Product
Normalize-İndirgenmiş Kümülatif Kazanç	Normalized-Discounted Cumulative Gain

TÜRKÇE-İNGİLİZCE TERİMLER SÖZLÜĞÜ (devam)

Terim	İngilizce Karşılığı
Ortalama Averaj Duyarlık	Mean Average Precision
Rastlantısal Oluştan Sapma	Divergence From Randomness
R-Duyarlık	R-Precision
Sinyal-Gürültü Oranı	Signal-Noise Ratio
Sorgu Genişletme	Query Expansion
Şans Oranı	Odds Ratio
Terim Ayırt Etme Değeri	Term Discrimination Value
Terim Frekansı	Term Frequency
Terim-boruhattı	Term-pipeline
Ters Beklenen Belge Frekansı	Inverse Term Frequency
Ters Belge Frekansı	Inverse Document Frequency
Ters İndeks Liste	Inverted Index File
Ters Terim Frekansı	Inverse Term Frequency
Vektör Uzayı	Vector Space
Yönlendirme	Roughting
Yürütüm	Run

KAYNAKLAR DİZİNİ

Agrawal, R., Gollapudi, S., Halverson, A., and Ieong, S., 2009, Diversifying search results, *In Proceedings of the Second ACM International Conference on Web Search and Data Mining*, 5–14.

Agresti, A., 2002, Categorical Data Analyses, 2nd Edition, Wiley-Interscience, New Jersey, 710p.

Amati, G., and Van RijsBergen, C.J., 2002, Probabilistic models of information retrieval based on measuring the divergence from randomness, *ACM Transactions on Information Systems (TOIS)*, 20(4):357–389.

Aslam, A., Pavlu, P., and Yilmaz, E., 2006, A statistical method for system evaluation using incomplete judgments, *In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 541-548.

Aslam, J.A., and Pavlu., V., 2007, A practical sampling strategy for efficient retrieval evaluation, Technical Report,Northeastern University, 10p.

Can, F., and Özkarahan, E. A., 1987, Computation of term/document discrimination values by use of the cover coefficient concept., *Journal of the American Society for Information Science*, 38(3): 171-183.

Carnegie Mellon University , "ClueWeb09 web sayfası" http://boston.lti.cs.cmu.edu/Data/clueweb09/. (Erişim tarihi: 8 Haziran 2010)

Carterette, B., Allan, J., and Sitaraman, R., 2006, Minimal test collections for retrieval evaluation, *In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 268–275.

Carterette, B., Pavlu, P., Fang, H., and Kanoulas, E., 2009, Million query track 2009 overview, *NIST Special Publication 500-278:The 18th Text Retrieval Conference Proceedings(TREC 2009)*.

KAYNAKLAR DİZİNİ (devam)

Chapelle, O., Metlzer, D., Zhang, Y., and Grinspan, P., 2009, Expected reciprocal rank for graded relevance, *In Proceedings of the 18th ACM Conference on Information and Knowledge Management*, 621-630.

Church, K., W., 1995, One term or two?, *In Proceedings of the 18th Annual Iinternational ACM SIGIR Conference on Research and Development in Information Retrieval*, 310-318.

Clarke, C.L.A., Kolla, M., Cormack, G.V., Vechtomova, O., Ashkann, A., Buttcher, S., and MacKinnon., I., 2008, Novelty and diversity in information retrieval evaluation, *In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 659–666.

Clarke, C.L.A., Craswell., N., and Soboroff, I., 2009, Preliminary report on the TREC 2009 web track, *NIST Special Publication 500-278:The 18th Text Retrieval Conference Proceedings (TREC 2009)*.

Clarke, C.L.A., Craswell, N., Soboroff, I., and Cormack, G.V., 2010, Preliminary overview of the TREC 2010 web track, *The 19th Text REtrieval Conference Proceedings (TREC 2010)*.

Cooper, W.S., and Maron, M.E., 1978, Foundations of probabilistic and utility-theoretic indexing, *Journal of the ACM (JACM)*, 26(1):67-80.

Cormack, G.V., Smucker, M.D., and Clarke, C.L.A., 2010, Efficient and effective spam filtering and re-ranking for large web datasets, Computing Research Repository, abs/1004.5168.

Croft, W.B., and Harper, D.J., 1979, Using probabilistic models of document retrieval without relevance information, *Journal of Documentation*, 35(4): 285-295.

Croft, W.B., and Lafferty, J. (eds.), 2003, Language Modelling for Information Retrieval, Kluwer Academic Publishers, Dordrecht, The Netherlands, 253p.

Croft, W.B., Metzler, D., and Strohman, T., 2009, Search Engines: Information Retrieval in Practice, Addison Wesley, USA, 552p.

KAYNAKLAR DİZİNİ (devam)

Dincer, B.T., 2004, Türkçe için İstatistiksel Bir Bilgi Geri-Getirim Sistemi, Doktora Tezi, Ege Üniversitesi Uluslararası Bilgisayar Enstitüsü, 382s (yayımlanmamış).

Dinçer. B.T., 2005, İstatistiksel bir Bilgi Erişim Sistemi Tasarımı, TÜBİTAK projesi no: 107E192.

Dinçer, B.T., 2007, Statistical components analysis for retrieval experiments, *Journal of the American Society for Information Science and Technology*, 58(4):560-574.

Dinçer, B.T., Kocabaş, İ., and Karaoğlan, B., 2009, IRRA at TREC 2009: Index Term Weighting based on Divergence From Independence Model, *NIST Special Publication 500-278:The 18th Text Retrieval Conference Proceedings (TREC 2009)*.

Fuhr, N., 1989, Models for retrieval with probabilistic indexing, *Information Processing & Management*, 25(1): 55-72.

Harman, D., 1992, Ranking Algorithms, 363-392, Information Retrieval: Data Structures & Algorithms, Frakes, W.B., and Baeza-Yates, R.A. (Eds.), Prentice-Hall, New Jersey, 504p.

Harman, D., 1992, Overview of the First Text REtrieval Conference (TREC-1), *NIST Special Publication 500-207:The First Text Retrieval Conference Proceedings (TREC-1)*, 1-20.

Harman, D., 1993, Overview of the Second Text REtrieval Conference (TREC-2), *NIST Special Publication 500-215:The Second Text Retrieval Conference Proceedings (TREC-2)*, 1-20.

Harman, D., 1994, Overview of the Third Text REtrieval Conference (TREC-3), *NIST Special Publication 500-225:The Third Text Retrieval Conference Proceedings (TREC-3)*, 1-20.

Harter, S.P., 1974, A probabilistic approach to automatic keyword indexing, PhD Thesis, Thesis No. T25146, Graduate Library, The University of Chicago, 117p.

KAYNAKLAR DİZİNİ (devam)

Harter, S.P., 1975a, A probabilistic approach to automatic keyword indexing, Part I: On the distribution of specialty of words in a technical literature, *Journal of the American Society for Information Science (JASIS)*, 26(4):197-216.

Harter, S.P., 1975b, A probabilistic approach to automatic keyword indexing, Part II: An algorithm for probabilistic indexing, *Journal of the American Society for Information Science (JASIS)*, 26(4):280-289.

Hiemstra, D., and de Vries, A.P., 2000, Relating the new language models of information retrieval to the traditional retrieval models, CTIT Technical Report TR-CTIT-00-09, Enschede, The Netherlands: University of Twente, 14p.

Järvelin, K., and Kekäläinen J., 2002, Cumulated gain-based evaluation of IR techniques, *ACM Transactions on Information Systems (TOIS)*, 20(4):422-446.

Liu. X., and Croft. W.B., 2005, Statistical language modelling for information retrieval, *Annual Review of Information Science and Technology*, 39(1):1-31.

Luhn, H.P., 1957, A Statistical approach to mechanized encoding and searching of literary information, *IBM Journal Research and Development*, 1(4):309-317.

Luhn, H.P., 1958, The automatic creation of literature abstracts, *IBM Journal of Research and Development*, 2(2):159-165.

Manning, C.D., Raghavan, P., and Schütze, H., 2008, Introduction to Information Retrieval, Cambridge University Press, New York, 544p.

Maron, M.E., and Kuhns, J.L.; 1960, On relevance, probabilistic indexing and information retrieval; *Journal of the ACM (JACM)*, 25(3):216-244.

Minker, J., Peitola, E., and Wilson, G.A., 1973, Document retrieval experiments using cluster analysis, *Journal of the American Society for Information Science (JASIS)*, 24(4):246:260.

KAYNAKLAR DİZİNİ (devam)

National Institute of Standards and Technology, TREC web sayfası, http://trec.nist.gov/ (Erişim tarihi: 8 Haziran 2010).

Ponte, J. and Croft, B., 1998, A language modeling approach in information retrieval, *In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 275-281.

Porter, M., 1980, An algorithm for suffix stripping, *Program*, 14(3):130-137.

Robertson, S.E., and Sparck Jones, K., 1976, Relevance weighting of search terms, *Journal of the American Society for Information Science (JASIS)*, 27(3):129-146.

Robertson, S.E., Van Rijsbergen, C.J., and Porter, M., 1980, Probabilistic models of indexing and searching, *In Proceedings of the 3rd Annual International ACM Conference on Research and Development in Information Retrieval*, 35-56.

Robertson, S.E, and Walker, S., 1994, Some simple approximations to 2-Poisson model for probabilistic weighted retrieval, *In Proceedings of the 7th Annual International ACMSIGIR Conference on Research and Development in Information Retrieval*, 232-241.

Robertson, S.E, Walker, S., and Beaulieu, M., 1999, Okapi at trec-7: automatic adhoc, filtering, vlc and interactive, *NIST Special Publication 500-242: The Seventh Text Retrieval Conference (TREC-7)*, 253-264.

Salton, G., 1968, Automatic Information Organization and Retrieval, McGraw-Hill, New York, 527p.

Salton, G., 1970, Automatic text analysis, *Science*, 168(929):335-343.

Salton, G., and Yang, C.S., 1973, On the specification of term values in automatic indexing, *Journal of Documentation*, 29(4):351-372.

Salton, G., Wong, A., and Yang, C.S., 1975a, A vector space model for automatic indexing, *Communications of the ACM*, 18(11):613-620.

KAYNAKLAR DİZİNİ (devam)

Salton, G., 1975b, A Theory of Indexing, Society for Industrial and Applied Mathematics, Philadelphia, 55p.

Salton, G., Wong, A., and Yu, C.T., 1976, Automatic indexing using term discrimination and term precision measurements, *Information Processing and Management*, 12(1):43-51.

Salton, G, and Buckley, C, 1988, Term-weighting approaches in automatic text retrieval, *Information Processing and Retrieval*, 24(5):513-523.

Singhal, A., Buckley, C., and Mitra, M, 1996, Pivoted document length normalization, *In SIGIR '96: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 21-29.

Sparck Jones, K., 1972, A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11-21.

Sparck Jones, K., Walker, S., and Robertson, S.E., 2000, A probabilistic model of information retrieval: development and comparative experiments, *Information Processing and Management*, 36(6):779-840.

Turtle, H.R., and Croft, W.B., 1992, A comparison of text retrieval models, *the Computer Journal*, 35(3):279-290.

University of Glasgow, "TERRIER web sayfası", http://terrier.org (Erişim tarihi: 8 Haziran 2010).

Van Rijsbergen, C.J., 1979, Information Retrieval , 2nd Edition, Butterworths, London, 208p.

Varian, H.R., 2005, The digital society: Universal access to information, *Communications of the ACM, 48*(10):65-66.

Voorhees, E.M., and Harman, D., 1997, Overview of the Sixth Text REtrieval Conference (TREC-6), *NIST Special Publication 500-240: The Sixth Text Retrieval Conference Proceedings (TREC-6)*, 1-30.

KAYNAKLAR DİZİNİ (devam)

Voorhees, E.M., and Harman, D., 1998, Overview of the Seventh Text REtrieval Conference (TREC-7), *NIST Special Publication 500-242: The Seventh Text Retrieval Conference Proceedings (TREC-7)*, 1-24.

Voorhees, E.M., and Harman, D., 1999, Overview of the Eighth Text REtrieval Conference (TREC-8), *NIST Special Publication 500-246: The Eighth Text Retrieval Conference Proceedings (TREC-8)*, 1-24.

Voorhees, E.M., 2007, Overview of TREC-2007, *NIST Special Publication 500-274: The 16th Text Retrieval Conference Proceedings (TREC 2007)*, 1-17.

Wong, S. K. M. and Yao, Y. Y., 1995, On modeling information retrieval with probabilistic inference, *ACM Transactions on Information Systems (TOIS)*, 13(1):38-68.

Zhai., C., Notes on the Lemur TFIDF model (Lemur 1.9 dokümantasyonunda), School of CS, Carnegie Mellon University, http://www.cs.cmu.edu/~lemur/1.9/tfidf.ps (Erişim tarihi: 8 Haziran 2010).

EKLER

EK-1 SGML ile Etiketlenmiş TREC Belgesi Örneği

```
<DOC>

<DOCNO> FBIS3-10 </DOCNO>

<HT>    "cr00000011994001" </HT>

<HEADER>

<DATE1>   9 March 1994 </DATE1>

Article Type:FBIS

DUE TO COPYRIGHT OR OTHER RESTRICTIONS THE FOLLOWING ITEM IS
INTENDED FOR USE ONLY BY U.S. GOVERNMENT CONSUMERS. IT IS BASED ON
FOREIGN MEDIA CONTENT AND BEHAVIOR AND IS ISSUED WITHOUT
COORDINATION WITH OTHER U.S. GOVERNMENT COMPONENTS.

Document Type:FBIS TRENDS-07MAR94-VIETNAM

<H3> <TI>      Vietnam-Libya </TI></H3>

</HEADER>

<TEXT>

Hanoi     Finds New Outlet for Surplus Labor

   Judging by a 1 March VNA report, Hanoi has found new
opportunities for employing its surplus labor in Libya. According
to VNA, 100 Vietnamese workers left Hanoi on 28 February to fill
jobs in Libya under a "construction contract" signed by Vietnam's
overseas construction company VINACONEX and South Korea's Dong Ah
Consortium.  These 100 laborers, VNA said, are the "first batch"
of a 2,000-person contingent of construction workers and overseers
who will be sent to Libya in 1994 to work on a man-made river
project.  These workers will join the 1,500 Vietnamese workers
already working in Libya under a contract between VINACONEX and
Dong Ah Consortium.  Hanoi has routinely sent workers overseas in
order to relieve its unemployment problem and to help pay for
needed goods.


(AUTHOR:  HEBBEL.  QUESTIONS AND/OR COMMENTS, PLEASE CALL CHIEF,
ASIA DIVISION ANALYSIS TEAM, (703) 733-6534.)

EAG/BIETZ/ta 07/2051z mar


</TEXT>

</DOC>
```

EK-2 Bağımsızlıktan sapma modellerinin açık formülleri

Temel Model	Model Belirteçi	Ağırlıklandırma Formülü	Sayfa
DFI_0	DFI_0_0	$DFI_{ij} = \frac{x_{ij} - e_{ij}}{e_{ij}}$	44
	DFI_0_1	$DFI_{ij} = log_2\left(\frac{x_{ij} - e_{ij}}{e_{ij}} + 1\right)$	45
	DFI_0_2	$DFI_{ij} \times IDF_i = log_2\left(\frac{x_{ij} - e_{ij}}{e_{ij}} + 1\right) \times log_2\left(\frac{n}{n_i} + 1\right)$	46
DFI_1	DFI_1_0	$DFI_{ij} = \frac{x_{ij} - e_{ij}}{\sqrt{e_{ij}}}$	45
	DFI_1_1	$DFI_{ij} = log_2\left(\frac{x_{ij} - e_{ij}}{\sqrt{e_{ij}}} + 1\right)$	45
	DFI_1_2	$DFI_{ij} \times IDF_i = log_2\left(\frac{x_{ij} - e_{ij}}{\sqrt{e_{ij}}} + 1\right) \times log_2\left(\frac{n}{n_i} + 1\right)$	46

EK-3 Luhn esasındaki modellerin açık formülleri

EK-3.a TF Formülleri

Model Belirteçi	Ağırlıklandırma Formülü	Sayfa
TF-1(Z)	$log_2\left(\frac{1}{Z_\alpha + 1} + 1\right)$	49 ve 52
TF-1(M1)	$log_2\left(\frac{1}{\left(\lvert x_{ij} - M_j \rvert / \sqrt{s_{mj}^2}\right) + 1} + 1\right)$	49 ve 52
TF-1(M2)	$log_2\left(\frac{1}{\left(\lvert x_{ij} - M_j \rvert / M_j\right) + 1} + 1\right)$	50 ve 52
TF-1(β,C)	$log_2\left(\frac{1}{\left(\lvert x_{ij} - M_j^T \rvert / M_j^T\right) + 1} + 1\right)$ $M_j^T = N^\beta / (V \times c) \quad , c = antilog(c)$	51 ve 52
TF-2(Z)	$log_2\left(\frac{1}{Z_\alpha^2 + 1} + 1\right)$	49 ve 52
TF-2(M1)	$log_2\left(\frac{1}{\left(\lvert x_{ij} - M_j \rvert / \sqrt{s_{mj}^2}\right)^2 + 1} + 1\right)$	49 ve 52
TF-2(M2)	$log_2\left(\frac{1}{\left(\lvert x_{ij} - M_j \rvert / M_j\right)^2 + 1} + 1\right)$	50 ve 52
TF-2(β,C)	$log_2\left(\frac{1}{\left(\lvert x_{ij} - M_j^T \rvert / M_j^T\right)^2 + 1} + 1\right)$ $M_j^T = N^\beta / (V \times c) \quad , c = antilog(c)$	51 ve 52

EK-3.b TFxIDF Formülleri

Model Belirteçi	Ağırlıklandırma Formülü	Sayfa
WTF-1(Z)	$\text{TF-1(Z)} \times log_2\left(\frac{n}{n_i}+1\right)$	53
WTF-1(M1)	$\text{TF-1(M1)} \times log_2\left(\frac{n}{n_i}+1\right)$	53
WTF-1(M2)	$\text{TF-1(M2)} \times log_2\left(\frac{n}{n_i}+1\right)$	53
WTF-1(β,C)	$\text{TF-1(β,C)} \times log_2\left(\frac{n}{n_i}+1\right)$	53
WTF-2(Z)	$\text{TF-2(Z)} \times log_2\left(\frac{n}{n_i}+1\right)$	53
WTF-2(M1)	$\text{TF-2(M1)} \times log_2\left(\frac{n}{n_i}+1\right)$	53
WTF-2(M2)	$\text{TF-2(M2)} \times log_2\left(\frac{n}{n_i}+1\right)$	53
WTF-2(β,C)	$\text{TF-2(β,C)} \times log_2\left(\frac{n}{n_i}+1\right)$	53

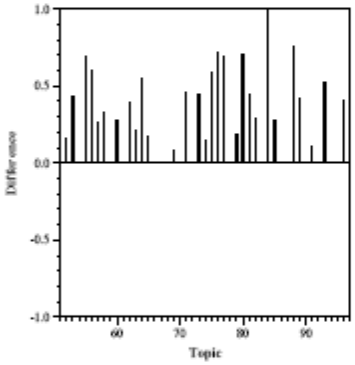
EK-4 TREC'09 Web izi anlık-sorgu görevinde sorgu bazlı en iyi, ortalama ve en kötü başarım değerleri

	Tahmini	AP tahmini			nDCG tahmini		
Konu	#alakalı belge	En iyi	Ort.	En kötü	En iyi	Ort.	En Kötü
1	220.29	0.7643	0.2271	0.0200	0.9478	0.6311	0.0000
2	98.99	0.8214	0.4165	0.0000	0.9063	0.7317	0.0000
3	392.34	0.1886	0.0372	0.0000	0.7177	0.1114	0.0000
4	70.76	0.2177	0.0902	0.0005	0.4098	0.1082	0.0000
5	11.08	0.6338	0.0925	0.0000	0.7601	0.2193	0.0000
6	87.04	0.7953	0.0828	0.0002	0.7487	0.1127	0.0000
7	433.30	0.4826	0.0211	0.0000	0.5019	0.2369	0.0000
8	193.73	0.2507	0.0104	0.0000	0.4155	0.0000	0.0000
9	331.77	0.5975	0.0507	0.0001	0.4804	0.2845	0.0000
10	1206.32	0.8994	0.0395	0.0000	0.5152	0.1228	0.0000
11	292.17	0.2342	0.1525	0.0293	0.7057	0.5701	0.2356
12	3152.99	0.2729	0.1754	0.0000	0.8085	0.4639	0.0000
13	10.00	0.0591	0.0039	0.0000	0.1389	0.0000	0.0000
14	945.61	0.0667	0.0163	0.0000	0.7350	0.0462	0.0000
15	832.59	0.4120	0.1978	0.0264	0.8419	0.5759	0.1380
16	231.14	0.6624	0.3049	0.0145	0.7948	0.5420	0.0649
17	679.59	0.1611	0.0663	0.0001	0.5562	0.1751	0.0000
18	820.64	0.3836	0.0663	0.0000	0.7143	0.4222	0.0000
19	2.00	0.0088	0.0000	0.0000	0.0000	0.0000	0.0000
20	0.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
21	576.04	0.8579	0.3718	0.0043	0.8501	0.4688	0.0602
22	178.84	0.5141	0.4015	0.0084	0.9052	0.6696	0.1562
23	965.39	0.5492	0.0106	0.0000	0.3271	0.0493	0.0000
24	473.17	0.5414	0.1756	0.0000	0.4407	0.2107	0.0000
25	424.13	0.3476	0.1802	0.0207	0.6757	0.2414	0.0235
26	311.68	0.2352	0.0742	0.0004	0.7569	0.2535	0.0000
27	531.71	0.3908	0.2384	0.0018	0.6955	0.2914	0.0000
28	1072.10	0.7384	0.3735	0.0081	0.6450	0.3175	0.0422
29	122.85	0.0918	0.0065	0.0001	0.1782	0.0000	0.0000
30	1091.99	0.4380	0.2012	0.0000	0.7750	0.4037	0.0000
31	971.31	0.6479	0.1900	0.0000	0.8477	0.6413	0.0000
32	2293.04	0.2537	0.0192	0.0002	0.4658	0.2085	0.0000
33	583.90	0.5224	0.4461	0.0415	0.8229	0.6063	0.1487
34	890.97	0.1443	0.0302	0.0000	0.3263	0.0739	0.0000
35	162.63	0.5430	0.2717	0.0229	0.8208	0.5615	0.0273
36	1065.16	0.2621	0.0353	0.0000	0.6847	0.2519	0.0000
37	3.00	0.3346	0.0600	0.0000	0.4693	0.1564	0.0000
38	950.67	0.2470	0.1150	0.0011	0.4970	0.3491	0.0000
39	734.58	0.5115	0.1061	0.0001	0.4292	0.2838	0.0238
40	74.97	0.4842	0.1418	0.0000	0.6742	0.2734	0.0000
41	120.28	0.5567	0.1667	0.0000	0.8587	0.4586	0.0000
42	17.00	0.6846	0.0105	0.0000	0.8247	0.0000	0.0000
43	46.14	0.7122	0.2727	0.0028	0.8805	0.4487	0.0000
44	354.64	0.0765	0.0134	0.0000	0.3016	0.1649	0.0000
45	651.84	0.5421	0.2816	0.0008	0.6403	0.3730	0.0689
46	65.54	0.8228	0.6766	0.0000	0.9042	0.7052	0.0000
47	127.45	0.6056	0.3704	0.0000	0.8333	0.3755	0.0000
48	12.00	0.1754	0.1219	0.0000	0.3786	0.2159	0.0000
49	55.18	0.5709	0.2113	0.0000	0.5936	0.2681	0.0000
50	95.86	0.2089	0.0696	0.0009	0.2536	0.1030	0.0000

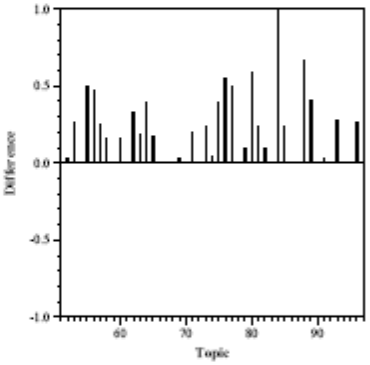
EK-5 NIST Tarafından Gönderilen TREC 2010 İRRA Grup Yürütüm Sonuç Özeti

Summary Statistics	
Run ID:	irra10b
Task :	adhoc
Category:	B
External resources used:	(A) no additional resources
Number of Topics:	36

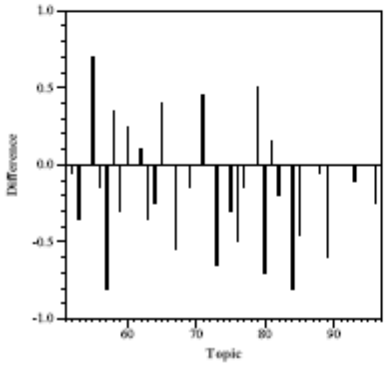
Adhoc measures		Diversity measures	
Retrieved	36000	α-nDCG@10	0.3212
Relevant	4289	α-nDCG@20	0.3653
Relevant retrieved	1506	ERR-IA@10	0.2445
Prec@10	0.4472	P-IA@10	0.1979
Prec@20	0.4833	P-IA@20	0.2229
MAP	0.1360	MAP-IA	0.0859
NDCG@20	0.2703	NRBP	0.2159
ERR@20	0.1319	ERR-IA@20	0.2572



Difference from median alpha-nDCG@20 per topic



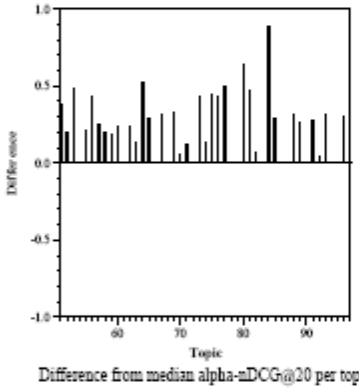
Difference from median ERR-IA@20 per topic



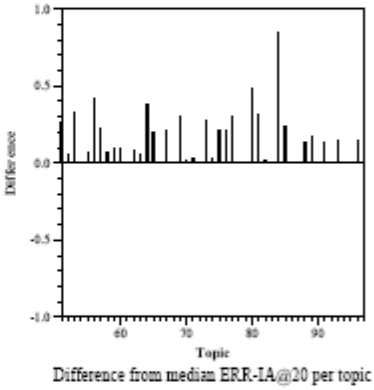
Difference from median P@20 per topic

Summary Statistics	
Run ID:	irra10hp
Task :	adhoc
Category:	B
External resources used:	(A) no additional resources
Number of Topics:	36

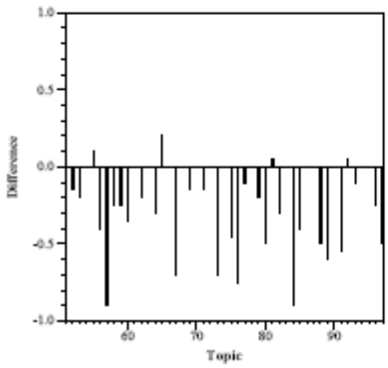
Adhoc measures		Diversity measures	
Retrieved	36000	α-nDCG@10	0.2454
Relevant	4289	α-nDCG@20	0.2877
Relevant retrieved	1847	ERR-IA@10	0.1708
Prec@10	0.3333	P-IA@10	0.1365
Prec@20	0.3236	P-IA@20	0.1439
MAP	0.1271	MAP-IA	0.0687
NDCG@20	0.1579	NRBP	0.1376
ERR@20	0.0825	ERR-IA@20	0.1820



Difference from median alpha-nDCG@20 per topic



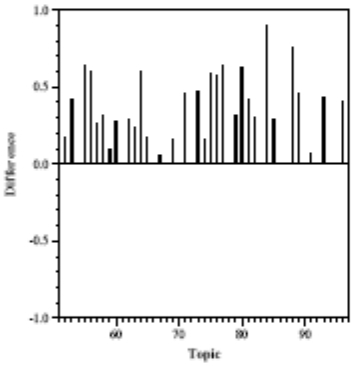
Difference from median ERR-IA@20 per topic



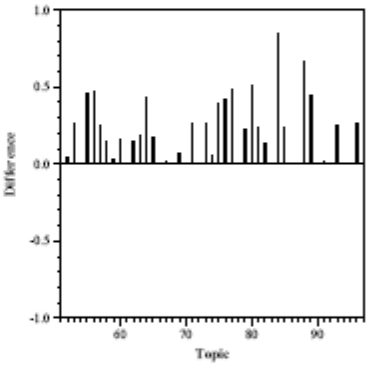
Difference from median P@20 per topic

Summary Statistics	
Run ID:	irra10rob
Task :	adhoc
Category:	B
External resources used:	(A) no additional resources
Number of Topics:	36

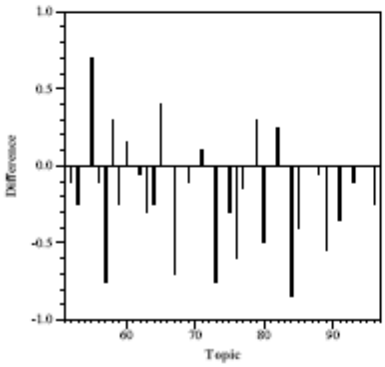
Adhoc measures		Diversity measures	
Retrieved	36000	α-nDCG@10	0.3191
Relevant	4289	α-nDCG@20	0.3589
Relevant retrieved	1756	ERR-IA@10	0.2415
Prec@10	0.4389	P-IA@10	0.1853
Prec@20	0.4528	P-IA@20	0.1963
MAP	0.1416	MAP-IA	0.0804
NDCG@20	0.2284	NRBP	0.2142
ERR@20	0.1273	ERR-IA@20	0.2532



Difference from median alpha-nDCG@20 per topic



Difference from median ERR-IA@20 per topic



Difference from median P@20 per topic

ÖZGEÇMİŞ

İlker Kocabaş

Adres: Uluslararası Bilgisayar Enstitüsü 35100 Bornova/İZMİR

Telefon: (+90 532) 3246905

e-mail: ilker.kocabas@ege.edu.tr

Kişisel Bilgiler

Milliyeti	:	Türkiye Cumhuriyeti
Doğum Yeri ve Tarihi	:	İzmir, 25.04.1978

Eğitim Durumu

Doktora	:	2005 –
		Ege Üniversitesi, Uluslararası Bilgisayar Enstitüsü
Yüksek Lisans	:	2001 – 2005
		Ege Üniversitesi, Uluslararası Bilgisayar Enstitüsü
Lisans	:	1995 – 2000
		Orta Doğu Teknik Üniversitesi, Elektrik ve Elektronik Mühendisliği
Lise	:	1988 – 1995
		Manisa Anadolu Lisesi

Yabancı Dil

İngilizce	:	İyi derecede

Bilgisayar Dilleri

- C/C++, Java, Assembly.

Mesleki İlgi Alanları

- Bilgi Erişim Sistemleri, Doğal Dil İşleme, Dağıtık/Paralel Sistemler ve Algoritmalar.

Projeler

- 2009- , "Anlamsal Servis Keşfinin Mobil Platformlarda Değerlendirilmesi ve Performansının Ölçülmesi" adlı BAP (No:09-UBE-003) projesinde araştırmacı.
- 2008- , "İstatistiksel Bir Bilgi Erişim Sistemi Tasarımı" adlı TUBİTAK (No: 107E192) projesinde bursiyer.
- 2005-2008, "Zipf Kanunları Esasında Güncel Yazılı Türkçe'nin Nicel Dilbilim Ölçütleri" adlı TUBİTAK (No: 104E120) projesinde bursiyer.
- 2005-2007, "Telsiz Duyarga Ağları Laboratuarı kurulması" adlı BAP (No:07-UBE-001) projesinde araştırmacı.

Yayınlar

- **Kocabaş, İ., Karaoğlan, B, ve Dinçer, B.T.,** 2011, Luhn's point of view: Median-based term weighting schemes, *Mathematical and Computational Applications (MCA),* (basımda).
- **Kocabaş, İ., Dinçer, B. T., ve Karaoğlan, B.,** 2011, Investigation of luhn's claim on information retrieval, *Turkish Journal of EE & CS (TJEECS),* (basımda).
- **Dinçer, B.T., Kocabaş, İ., ve Karaoğlan, B**., 2010, IRRA at TREC 2010: Index Term Weighting based on Divergence From Independence Model, NIST Special Publication 500-280:The 19th Text Retrieval Conference Proceedings (TREC 2010).
- **Dinçer, B.T., Kocabaş, İ., ve Karaoğlan, B**., 2009, IRRA at TREC 2009: Index Term Weighting based on Divergence From Independence Model, NIST Special Publication 500-278:The 18th Text Retrieval Conference Proceedings (TREC 2009).
- **Kocabaş, İ., Kışla, T.., ve Karaoğlan, B.,** 2007, Zipf's law of bursiness in Turkish: The length of intervals between repetitions, *In Proceedings of 22nd International Symposium on Computer and Information Sciences (ISCIS'07),* 1-3.