# CLUEWEB09 VE CLUEWEB12 VERİ KÜMELERİNİN WATERLOO SPAM SIRALAMALARININ RETROSPEKTİF OLARAK DEĞERLENDİRİLMESİ

**İbrahim Barış YILMAZEL**

**YÜKSEK LİSANS TEZİ**

Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Dr. Öğr. Üyesi Ahmet ARSLAN

Eskişehir

Anadolu Üniversitesi

Fen Bilimleri Enstitüsü

Ağustos 2018

## JÜRİ VE ENSTİTÜ ONAYI

**İbrahim Barış YILMAZEL**'in "**ClueWeb09 ve ClueWeb12 veri kümelerinin Waterloo spam sıralamalarının retrospektif olarak değerlendirilmesi**" başlıklı tezi 13/08/2018 tarihinde aşağıdaki jüri tarafından değerlendirilerek "Anadolu Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliği'nin ilgili maddeleri uyarınca, Bilgisayar Mühendisliği Anabilim dalında Yüksek Lisans tezi olarak kabul edilmiştir.

| **Jüri Üyeleri** | **Unvanı Adı Soyadı** | **İmza** |
|---|---|---|
| Üye (Tez Danışmanı) | : Dr. Öğr. Üyesi Ahmet ARSLAN | ..... |
| Üye | : Dr. Öğr. Üyesi Alper Kürşat UYSAL | ..... |
| Üye | : Dr. Öğr. Üyesi Muammer AKÇAY | ..... |

**Prof.Dr. Ersin YÜCEL**
**Fen Bilimleri Enstitüsü Müdürü**

# ÖZET


CLUEWEB09 VE CLUEWEB12 VERİ KÜMELERİNİN WATERLOO SPAM SIRALAMALARININ RETROSPEKTİF OLARAK DEĞERLENDİRİLMESİ

İbrahim Barış YILMAZEL

Bilgisayar Mühendisliği Anabilim Dalı
Anadolu Üniversitesi, Fen Bilimleri Enstitüsü, Ağustos 2018

Danışman: Dr. Öğr. Üyesi Ahmet ARSLAN

ClueWeb09 ve ClueWeb12, Web sayfalarından oluşan en büyük iki veri kümesidir ve 2009'dan 2017'ye kadar birçok TREC görevinde kullanılmıştır. Her yıl yaklaşık 50 yeni sorgu yayınlanmış, bu sorgulara karşılık gelen Web sayfaları havuzu ilgili, ilgisiz ve spam olarak değerlendiriciler tarafından etiketlenmiştir. ClueWeb korpora için önemli miktarda sorgu uygunluk yargısı toplanmıştır. Ticari arama motorlarını kasten kandırmak için tasarlanan spam sayfaları, gerçek Web'in olduğu gibi ClueWeb korporanın da bir parçasıdır ve Web bilgi erişim sistemlerinin spam sayfalarla başa çıkması gerekir. ClueWeb09 veri kümesindeki her sayfanın spam olma değerini belirleyen dört farklı (Fusion, Britney, GroupX, UK2006) spam sıralaması yayınlanmıştır. Bu spam sıralamalarını kullanarak, belirlenen bir eşik değeri için ClueWeb09'daki dokümanları spam ya da non-spam olarak sınıflandırmak mümkündür.

Bu tezde, birçok TREC Web Tracks ve Tasks Tracks sorgu uygunluk yargıları kullanılarak, ClueWeb korpora spam sıralamalarının "intrinsic" ve retrospektif değerlendirmesi sunulmuştur. Herhangi bir sorgu için ilgili olarak etiketlenen Web sayfalarının spam olamayacağı varsayılarak, ikili sınıflandırma için spam ya da ilgili olarak etiketlenen dokümanlardan bir eğitim seti oluşturulmuştur. Evrensel sınıflandırma ölçüm metrikleri kullanılarak yapılan deneylerde elde edilen "intrinsic" değerlendirme sonuçlarının, önceki araştırmacılarca gerçekleştirilen "extrinsic" değerlendirme sonuçlarıyla uyumlu olduğu bulunmuştur. Yapılan analizler, GroupX'in ilgili dokümanlar ile spam dokümanları ayırt etmede en güçlü yöntem olduğunu ortaya koymuştur. Ayrıca, ClueWeb12 spam sıralamasının ClueWeb09 kadar iyi performans göstermediği tespit edilmiştir.



**Anahtar Sözcükler:** Bilgi erişimi, Spam sınıflandırma, ClueWeb, Web spam, TREC.

## ABSTRACT

RETROSPECTIVE EVALUATION OF WATERLOO SPAM RANKINGS OF THE CLUEWEB09 AND CLUEWEB12 DATASETS

İbrahim Barış YILMAZEL

Department of Computer Engineering
Anadolu University, Graduate School of Sciences, August 2018

Supervisor: Dr. Öğr. Üyesi Ahmet ARSLAN

ClueWeb09 and ClueWeb12, are the two largest collection of Web pages that are used in various tracks of TREC ran through 2009 to 2017. For each year, approximately 50 new queries are released and a pool of Web pages are judged against these queries by human assessors as relevant, non-relevant, or spam/junk. Thus, a considerable amount of query relevance judgments is collected for the ClueWeb corpora. Spam pages, which are designed deliberately deceive the commercial search engines, are part of the real Web, so of ClueWeb corpora. Thus, a Web retrieval system has to cope with spam pages. In this direction, four different (Fusion, Britney, GroupX, UK2006) spam rankings that quantify "spamminess" of every page in the ClueWeb09 dataset are released in 2009. For a given threshold, it is possible to classify documents in the ClueWeb09 dataset as spam or non-spam using these spam rankings.

This thesis presents an intrinsic and retrospective evaluation of spam rankings of the ClueWeb corpora using the query relevance judgments of several TREC tracks. A ground truth for binary classification task is created by using documents that are judged as junk/spam or relevant. It is assumed that Web pages judged as relevant for any query cannot be spam. The experimental results of intrinsic evaluation using the universal binary classification evaluation measures are found to be aligned with extrinsic evaluations of spam rankings performed by previous researches. The analysis of the distribution of relevant documents over spam percentile score intervals reveal that GroupX is the most powerful at discriminating relevant documents from spam documents. It is also found that the spam ranking of the ClueWeb12 does not perform as good as ClueWeb09's.



**Keywords:** Information retrieval, Spam classification, ClueWeb, Web spam, TREC.

## TEŞEKKÜRLER

Tez çalışmam boyunca kendisiyle çalışmak büyük onur duyduğum, danışmanım Sayın Dr. Öğr. Üyesi Ahmet ARSLAN'a, çalışmamın her aşamasındaki rehberliği, desteği ve sabrı için çok teşekkür ederim.

Ayrıca, tezim için verdiği fikir ve katkılardan dolayı Arş. Gör. Dr. Burcu YÜREKLİ YILMAZEL'e teşekkür ederim.

İbrahim Barış YILMAZEL

Ağustos, 2018

14/08/2018

## ETİK İLKE VE KURALLARA UYGUNLUK BEYANNAMESİ

Bu tezin bana ait, özgün bir çalışma olduğunu; çalışmamın hazırlık, veri toplama, analiz ve bilgilerin sunumu olmak üzere tüm aşamalarında bilimsel etik ilke ve kurallara uygun davrandığımı; bu çalışma kapsamında elde edilen tüm veri ve bilgiler için kaynak gösterdiğimi ve bu kaynaklara kaynakçada yer verdiğimi; bu çalışmanın Anadolu Üniversitesi tarafından kullanılan "bilimsel intihal tespit programı"yla tarandığını ve hiçbir şekilde "intihal içermediğini" beyan ederim. Herhangi bir zamanda, çalışmamla ilgili yaptığım bu beyana aykırı bir durumun saptanması durumunda, ortaya çıkacak tüm ahlaki ve hukuki sonuçları kabul ettiğimi bildiririm.

İbrahim Barış YILMAZEL

# İÇİNDEKİLER

**Sayfa**


BAŞLIK SAYFASI ……...……i

JÜRİ VE ENSTİTÜ ONAYI ..... ii

ÖZET ..... iii

ABSTRACT ..... iv

TEŞEKKÜRLER ..... v

ETİK İLKE VE KURALLARA UYGUNLUK BEYANNAMESİ ..... vi

İÇİNDEKİLER ……..... vii

ŞEKİLLER DİZİNİ ..... ix

TABLOLAR DİZİNİ ..... x

SİMGELER VE KISALTMALAR DİZİNİ ..... xi

1. GİRİŞ ……..….. 1

2. BİLGİ ERİŞİMİ ..... 4

2.1. Bilgi Erişim Süreci ..... 4

2.1.1. İndeksleme ..... 4

2.1.2. Sorgu formülasyonu ..... 7

2.1.3. Eşleştirme ..... 8

2.1.4. Yeniden sıralama ..... 11

2.2. Web Bilgi Erişiminin Kendine Özgü Özellikleri ..... 11

2.3. Bilgi Erişim Sistemlerinde Test Koleksiyonu Bazlı Değerlendirme ..... 14

2.4. Bilgi Erişim Etkinliği Ölçüm Metrikleri ..... 19

3. ALANYAZIN ..... 21

3.1. Web Spam ..... 21

3.2. Web Spam Üzerine Yapılmış Çalışmalar ..... 22

4. DENEYSEL YÖNTEM ..... 26

5. DENEYSEL SONUÇLAR VE ANALİZ ..... 35
5.1. Değerlendirme Metrikleri ..... 35
5.2. ClueWeb09 Veri Kümesi Sonuçları ..... 38
5.3. ClueWeb12 Veri Kümesi Sonuçları ..... 46
6. TARTIŞMA ..... 51
6.1. Tasks Track Sorgu Uygunluk Formatı ..... 57
6.2. Wikipedia Sayfalarının Spam Sıralama Analizleri ..... 58
6.3. ClueWeb09 Veri Kümesindeki İlgisizi Dokümanların Analizleri ..... 60
6.4. ClueWeb Korpora Sorgu Analizleri ..... 61
7. SONUÇLAR VE ÖNERİLER ..... 67
KAYNAKÇA ..... 69
EKLER
ÖZGEÇMİŞ

# ŞEKİLLER DİZİNİ

**Sayfa**

**Şekil 2.1.** *Bilgi erişim süreci* ..... **5**
**Şekil 2.2.** *TREC ad-hoc gereksinim örneği* ..... **16**
**Şekil 2.3.** *TREC sorgu uygunluk yargı örneği* ..... **17**
**Şekil 2.4.** *TREC gönderim dosyası örneği* ..... **18**
**Şekil 4.1.** *Geçmiş çalışmalarda ve bu çalışmada kullanılan veri kümesi* ..... **33**
**Şekil 5.1.** *Spam sıralamaları bazında ClueWeb09 veri kümesindeki spam ve ilgili doküman dağılımları* ..... **39**
**Şekil 5.2.** *Spam sıralamalarının ClueWeb09 veri kümesi üzerindeki F1 değerleri* ..... **41**
**Şekil 5.3.** *Spam sıralamalarının ClueWeb09 veri kümesi üzerindeki Accuracy değerleri* ..... **41**
**Şekil 5.4.** *(Cormack vd., 2011) tarafından elde edilen sonuç grafiği* ..... **43**
**Şekil 5.5.** *Spam sıralamalarının ClueWeb09 veri kümesi üzerindeki PPV değerleri* ..... **44**
**Şekil 5.6.** *Spam sıralamalarının ClueWeb09 veri kümesi üzerindeki NPV değerleri* ..... **44**
**Şekil 5.7.** *Spam sıralamalarının ClueWeb09 veri kümesi üzerindeki ROC eğrileri* ..... **46**
**Şekil 5.8.** *Fusion spam sıralaması bazında ClueWeb12 veri kümesindeki spam ve ilgili doküman dağılımları* ..... **47**
**Şekil 5.9.** *ClueWeb12 ve ClueWeb09 Fusion sıralamalarının F1, Accuracy, NPV, PPV değerleri* ..... **48**
**Şekil 5.10.** *ClueWeb12 ve ClueWeb09 Fusion sıralamalarının ROC eğrileri* ..... **49**
**Şekil 6.1.** *ClueWeb09 Wikipedia sayfalarının spam sıralamalarına ait puan dağılımları* ..... **59**
**Şekil 6.2.** *ClueWeb09 İlgisiz Dokümanların Yüzdelik Puan Dağılımları* ..... **60**

## TABLOLAR DİZİNİ

**Sayfa**

**Tablo 4.1.** *Sorgu setlerinin istatistikleri* ..... **27**

**Tablo 4.2.** *Altı-puanlı ölçek* ..... **28**

**Tablo 4.3.** *ClueWeb09 veri kümesinde ilgili ve spam işaretlenmiş doküman sayıları* ... **31**

**Tablo 4.4.** *ClueWeb12 veri kümesinde ilgili ve spam işaretlenmiş doküman sayıları* ... **34**

**Tablo 5.1.** *İkili sınıflandırma hata matrisi* ..... **36**

**Tablo 5.2.** *ClueWeb09 veri kümesinde F1 ve Accuracy metriklerini maksimize eden eşik değerleri* ..... **42**

**Tablo 6.1.** *ClueWeb09 sorgu-doküman değerlendirme istatistikleri* ..... **51**

**Tablo 6.2.** *ClueWeb09 veri kümesinde birden çok kez değerlendirilen ve birden çok etiket alan dokümanların istatistikleri* ..... **52**

**Tablo 6.3.** *ClueWeb12 sorgu-doküman değerlendirme istatistikleri* ..... **54**

**Tablo 6.4.** *ClueWeb12 veri kümesinde birden çok kez değerlendirilen ve birden çok etiket alan dokümanların istatistikleri* ..... **55**

**Tablo 6.5.** *Tablo 6.4'ün SNR sembolizasyonu (S=-2 N=0, R=1,2,3,4)* ..... **57**

**Tablo 6.6.** *Tasks Track 2015-2016 uygunluk yargısı değerleri* ..... **58**

**Tablo 6.7.** *ClueWeb korporada en çok ve en az spam doküman döndüren sorgular* ..... **61**

**Tablo 6.8.** *ClueWeb korporada en çok ve en az ilgili doküman döndüren sorgular* ..... **62**

**Tablo 6.9.** *ClueWeb korporada en çok ve en az ilgisiz doküman döndüren sorgular* ... **62**

**Tablo 6.10.** *ClueWeb korporada en çok ve en az doküman döndüren sorgular* ..... **63**

**Tablo 6.11.** *ClueWeb09 veri kümesi için bulunan sorguların karşılaştırmalı analizi* ... **64**

**Tablo 6.12.** *ClueWeb09 veri kümesi için bulunan sorguların karşılaştırmalı analizi* ... **66**

## SİMGELER VE KISALTMALAR DİZİNİ

**AP** : Average Precision (Averaj Duyarlılık)

**BE** : Bilgi Erişim

**BM** : Best Match (En İyi Eşleşme)

**DCG** : Discounted Cumulative Gain (İndirimli Birikimli Kazanç)

**DF** : Document Frequecny (Doküman Frekansı)

**docid** : Document Identity (Doküman Numarası)

**ERR** : Expected Reciprocal Rank (Beklenen Karşılıklı Sıralama)

**FN** : False Negative (Yanlış Negatif)

**FP** : False Positive (Yanlış Pozitif)

**FPR** : False Positive Rate (Yanlış Pozitif Oranı)

**IDF** : Inverse Document Frequency (Ters Doküman Frekansı)

**LM** : Language Model (Dil Modeli)

**MAP** : Mean Average Precision (Ortalama Averaj Duyarlılığı)

**NDCG** : Normalized Discounted Cumulative Gain (Normalleştirilmiş DCG)

**NIST** : National Institute of Standards and Technology

**NPV** : Negative Predictive Value (Negatif Öngörü Değeri)

**PL2** : Poisson model with Laplace after-effect and normalisation 2

**PPV** : Positive Predictive Value (Pozitif Öngörü Değeri)

**qid** : Query Identity (Sorgu Numarası)

**qid/docid** : Sorgu-Doküman Çifti

**qrels** : Sorgu Uygunluk Yargısı

**ROC** : Receiver Operating Characteristic (Alıcı İşletim Karakteristiği)

**TF** : Term Frequeency (Terim Frekansı)

**TN** : True Negative (Doğru Negatif)

**TP** : True Positive (Doğru Positif)

**TPR** : True Pozitif Rate (Doğru Pozitif Oranı)

**TREC** : Text REtrieval Conference

**WebBE** : Web Bilgi Erişim

# 1. GİRİŞ

1990'ların başında internetin doğuşuyla başlayan ve her geçen gün giderek artan web sayfaları günümüze değin devasa boyutlara gelmiş ve web üzerinden ihtiyaç duyulan bilgiye erişim zorlaşmıştır. Kullanıcıların web üzerinde yayınlanan ihtiyaç duydukları yararlı bilgilere erişimini kolaylaştırmak adına Web'e özgü arama motorları doğmuştur. Web arama motorları bilgiye erişimi ne kadar kolaylaştırırsa da arama sonuç listelerinde üst sırada olmanın getirdiği ticari kazanç, uygun olmayan yöntemlerle arama sonuç sıralamalarına müdahaleleri arttırmıştır. Arama motoru algoritmalarını aldatıp arama sonuçlarında daha yüksek sıralama elde etme yöntemlerinden birisi spam içerik ya da spam sayfalar üretmektir. Spam içerikler, arama motorlarının performansını doğrudan etkilemekte ve ticari kayıplar vermelerine sebep olmaktadır. Dolayısıyla, arama motoru şirketleri için önemli bir sorundur.

ClueWeb09 ve ClueWeb12 veri kümeleri, amacı Bilgi Erişim (BE) metodolojilerinin büyük ölçekli değerlendirilmesi için gerekli altyapıyı sağlayarak, alanındaki araştırmaları desteklemek olan ve National Institute of Standards and Technology (NIST) tarafından finanse edilen, Text REtrieval Conference (TREC) serisi kapsamında yayınlanmış en büyük iki veri kümesidir. Web arama teknolojilerinin uygulamadaki artan çeşitliliği, sürekli ölçeği büyüyen, farklı web içerik yelpazesine sahip büyük ölçekli veri kümeleri yaratma gereksinimden dolayı araştırmacılara Web'in temsili niteliğinde sunulan ve Web'den taranan sayfalardan oluşan bu iki veri kümesi, 2009 yılından 2017 yılına kadar birçok TREC görevinde kullanılmıştır.

Ticari arama motorlarını kasten kandırmak için tasarlanan spam sayfaları, gerçek Web'in olduğu gibi ClueWeb korporanın da bir parçasıdır. Web Bilgi Erişim (WebBE) sistem etkinliği açısından spam sayfalarla başa çıkılması gerekir. Bu amaca yönelik, 2009 yılında ClueWeb09 veri kümesindeki her dokümanın spam olma puanını belirleyen dört farklı (Fusion, Britney, GroupX, UK2006) spam sıralaması yayınlanmıştır. Bu spam sıralamalarının, belirlenen bir eşik değeri için ClueWeb09 veri kümesindeki dokümanları spam ya da non-spam olarak sınıflandırma amacıyla kullanılabileceği belirtilmiştir (Cormack vd., 2011). Aslında, belirlenen bir eşik değeri için spam sıralamaları ikili sınıflandırıcıdır, öyle ki spam puanı o eşik değerinden düşük olan dokümanlar spam, yüksek olan dokümanlar ise non-spam olarak sınıflandırılır.

Cormack ve arkadaşları, tarafından yapılan çalışma, spam sıralamalarının tekil başarımına, yani spam sıralamalarının spam ve non-spam ayrımını ne derece yapabildiğine değil, bu sıralamaların tam teşekküllü *(İng.* fully fledged) BE sistemlerine entegre edilmiş haliyle, entegre edilmemiş halini kıyaslar. Bir başka değişle, spam sıralamalarının BE etkinliğine tesiri ölçülür, yani "extrinsic" değerlendirmesi yapılır.

Diğer taraftan, spam sıralamalarını, tekil olarak değerlendirmek de mümkündür. Spam sıralamalarını bir ikili sınıflandırıcı olarak tek başına ele alıp, bu sıralamaların spam ve non-spam dokümanları ne derece ayırt edebildiği incelenebilir. Ancak sınıflandırma başarımının ölçülebilmesi için gerçek doğrular olarak kabul edilebilecek işaretlenmiş bir veri kümesi mevcut değildir.

Bu noktada, TREC konferansı kapsamında yayınlanan *qrels* dosyaları olarak adlandırılan sorgu uygunluk yargıları ile gerçek doğrular taklit edilebilir. TREC konferanslarında her yıl yaklaşık 50 yeni sorgu yayınlanır. Bu sorgulara karşılık gelen Web sayfaları havuzu ilgili, ilgisiz ve spam olarak yetkin kişiler tarafından elle değerlendirilir. Bir dokümanın bir sorguyla ilgili olup olmadığı konusunda insanlar tarafından verilen farklı uygunluk yargıları şöyledir: ilgili dokümanlar için 4 puanlı ölçek (1, 2, 3, ya da 4), ilgisiz dokümanlar için 0 ve spam/önemsiz dokümanlar için -2. Buna TREC metodolojisinde altı puanlı ölçek denir. İlgili ve ilgisiz kavramları sorgu-doküman çiftleri için geçerli olsa da spam/önemsiz dokümanlar sorgudan bağımsızdır. Değerlendirilen tüm sorgu-doküman çiftleri ve atanan uygunluk yargıları her yıl *qrels* dosyalarında yayınlanır.

Şu varsayımlar altında *qrels* yardımıyla spam sıralamalarını, tekil olarak değerlendirmek mümkün olur. (1) Uygunluk yargısı -2 olan dokümanlar spam'dir, (2) İlgili dokümanlar (uygunluk yargısı 1, 2, 3, ya da 4), herhangi bir sorgu için ilgili olarak işaretlendiğinden ötürü spam olamaz, non-spam'dir. Böylece, spam sıralamalarının spam ve non-spam dokümanları ne derece ayırt edebildiğinin "intrinsic" değerlendirmesini yapmak mümkün olur.

2009 yılından günümüze değin her sene 50 yeni sorgu kümeye eklenmiştir. Dolayısıyla, ClueWeb korpora için gerçek doğrular olarak kabul edilebilecek önemli miktarda sorgu uygunluk yargısı toplanmıştır.

Önceki çalışmalarda yapılan "extrinsic" değerlendirmeye karşılık, bu tezde, 2009-2016 yılları arasında yayınlanan sorgu uygunluk yargıları kullanılarak, ClueWeb korpora spam sıralamalarının "intrinsic" ve retrospektif değerlendirmesi hedeflenmiştir.

2009 sonrası işaretlenen sorgu uygunluk yargıları esas alınarak 2009'da yayınlanan spam sıralamaları değerlendirilip inceleneceğinden, yani geçmişe dönük değerlendirme yapılacağından, yapılan çalışma aynı zamanda retrospektif değerlendirme niteliğindedir. Yazarların bildiği kadarıyla böyle bir değerlendirme literatürde daha önce yapılmamıştır.

Bu çalışmanın alana katkısı,

- Sorgu uygunluk yargılarının (*qrels*) kullanımıyla dört spam sıralamasının "intrinsic" değerlendirmesinin yapılması.
- Elde edilen sonuçların geçmişteki "extrinsic" değerlendirme sonuçlarıyla benzer olup olmadığının teyidi.
- Eşik değeri belirlemede alternatif bir yöntem sunulması.
- Dört spam sıralamasından ortalamada hangisinin spam ve non-spam ayrımını daha iyi yapabildiğinin tespiti.
- ClueWeb09 veri kümesi spam konusunda çok araştırılmış olsa da ClueWeb12 veri kümesi henüz hiç araştırılmamış olduğundan, ClueWeb12 veri kümesi için spam sıralamasının nasıl çalıştığının incelenmesi.

Bu çalışmada ele alınan araştırma soruları,

*Araştırma Sorusu 1*: Spam puanlarının "intrinsic" değerlendirmesiyle elde edilen sonuçlar, geçmiş "extrinsic" değerlendirme sonuçları ile uyumlu mu?

*Araştırma Sorusu 2*: Clueweb09 veri kümesi için dokümanların spam ve non-spam olarak işaretlenmesinin hedeflendiği durum için eşik değeri olarak 70 ve dört spam sıralamasından en iyisi istendiğinde Fusion sıralamasının seçilmesi gerektiği sonuçları (Cormack vd., 2011) günümüze değin elde edilen tüm sorgu yargılarının "intrinsic" değerlendirmesinde farklılık gösteriyor mu?

*Araştırma Sorusu 3*: Clueweb12 veri kümesindeki durum ClueWeb09 veri kümesiyle aynı mı?

# 2. BİLGİ ERİŞİMİ

Bu bölümde, genel olarak bilgi erişimin ne olduğu, Web bilgi erişiminin kendine özgü özellikleri, BE sistemlerinde test koleksiyonu bazlı değerlendirme ve BE etkinlik ölçüm metrikleri anlatılacaktır.

## 2.1. Bilgi Erişim Süreci

Bilgi Erişimi (BE, *İng.* Information Retrieval - IR), büyük koleksiyonlar (genellikle bilgisayarlarda saklanan) içerisinden ihtiyaç duyulan bilgi gereksinimine (sorgu) karşılık gelen yapılandırılmamış (genellikle metin) materyalleri (genellikle belgeler) bulma olarak tanımlanır (Manning vd., 2008). BE, büyük miktarda bilgiye hızlı ve etkili içerik-tabanlı erişim sağlayabilen sistemleri modelleme, tasarlama ve uygulama amacı güder (Baeza-Yates ve Ribeiro-Neto, 1999; Salton ve McGill, 1986; Van Rijsbergen, 1979). Nihai hedef, kullanıcının bilgi gereksinimini (*İng.* information need) karşılamaktır. Bunu yapabilmek için; BE sisteminin kullanıcının bilgi gereksinimini karşılayabilecek, bir başka deyişle, kullanıcının ihtiyaç duyduğu bilgileri içerebilecek, belgeleri döndürmesi gerekir. Kullanıcının bilgi gereksinimini karşılayan bu belgeler "ilgili belgeler (*İng.* relevant document)" olarak adlandırılır ve ideal bir BE sisteminin sadece ilgili belgeleri getirmesi, bu ilgili belgeleri de en çok ilgili olandan en az ilgili olana göre sıralayarak kullanıcıya gösterilmesi gerekir.

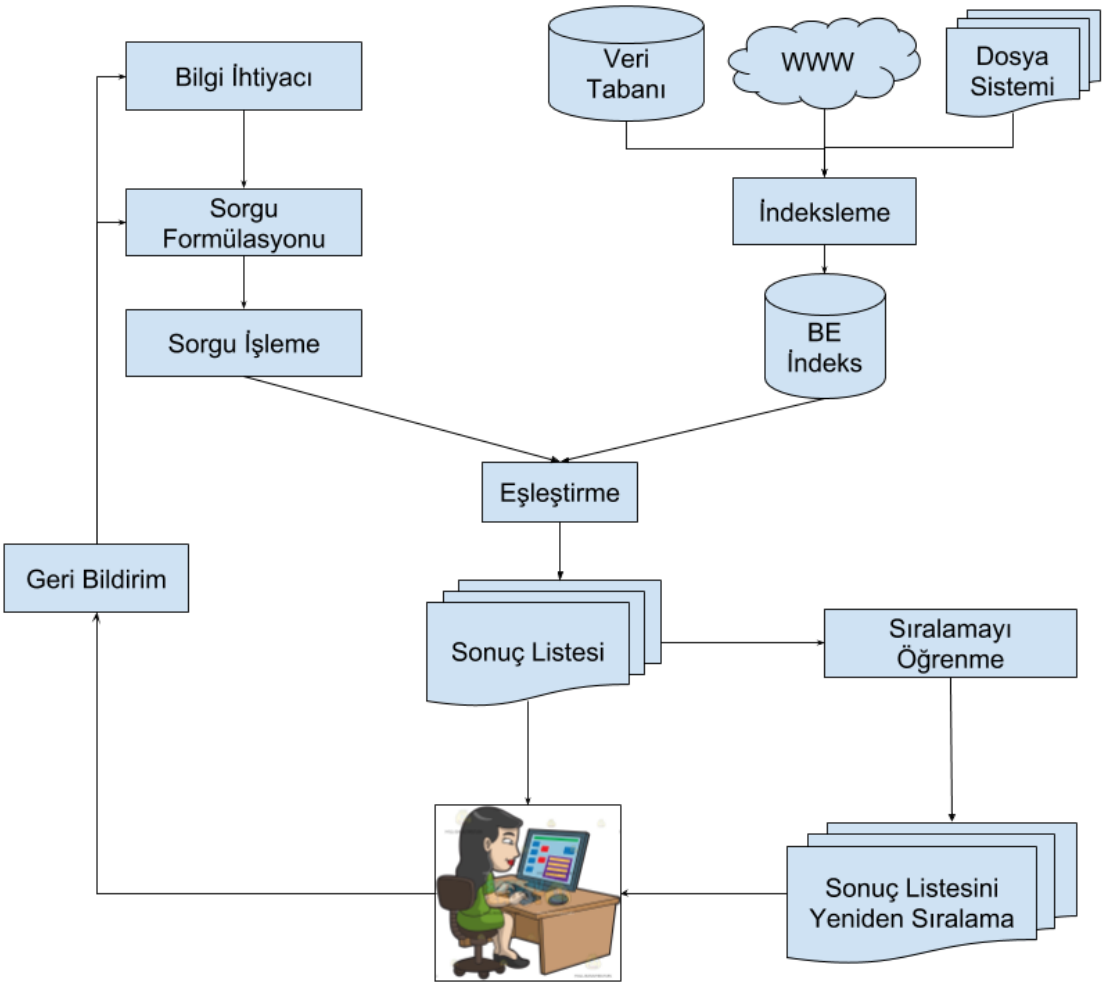
Bir BE sistemi oluşturulurken genellikle dört süreçten taviz verilmez: (i) indeksleme *(İng.* indexing), (ii) sorgu formülasyonu *(İng.* query formulation), (iii) eşleştirme *(İng.* matching), (iv) yeniden sıralama *(İng.* re-ranking). BE sisteminin süreç akış şeması diyagramı Şekil 2.1'de gösterilmektedir.

### 2.1.1. İndeksleme

BE sistemlerinin girdisi kullanıcı sorgusu, çıktısı ise kullanıcı sorgusunda ifade edilen bilgi ihtiyacını karşılayabilecek ilgili dokümanların sıralı listesidir. Bir kullanıcı sorgusu verildiğinde, sorgu terimlerini içeren ilgili dokümanların tespit edilebilmesi için dokümanların sırasıyla taranması (*ör.* "grep") akla gelebilecek ilk seçenektir. Ancak bu yöntem sisteme gelen her bir kullanıcı sorgusu için koleksiyondaki tüm doküman içeriklerinin tek tek taranması anlamına geldiğinden büyük koleksiyonlar için kullanışsızdır. Bu nedenle, BE sistemlerinde koleksiyon içerisindeki dokümanların

verilen bir kullanıcı sorgusuyla eşleşip eşleşmediğinin daha etkin ve hızlı bir şekilde değerlendirilebilmesi için indeksleme denilen çevrimiçi bir süreç uygulanır. Burada amaç, belge koleksiyonundan indeks olarak adlandırılan veri yapılarının oluşturulmasıdır.

Metin verileri için birçok indeks şeması olmasına karşın, en popüleri ters indeks *(İng.* inverted index)'dir (Ceri vd., 2013). Çünkü ters indeks şeması sadece sorgu terimlerini içeren dokümanların etkin erişimini sağlamakla kalmayıp, aynı zamanda çok hızlı ve kolay bir şekilde oluşturulur. Genel olarak, ters indeks yapısı iki bileşen içerir: indekslenen kelimeler *(İng.* vocabulary / indexed terms) ve kayıt listeleri *(İng.* posting list).




**Şekil 2.1.** *Bilgi erişim süreci*

İndekslenen kelimeler, koleksiyonda bulunan dokümanların içerdiği tüm tekil kelimelerden oluşan terimler sözlüğü olarak düşünülebilir. Her bir tekil kelime için o kelimenin hangi dokümanda ve o doküman içerisinde hangi pozisyonda olduğu, terimin frekansı gibi istatistiksel bilgiler ise kayıt listelerinde saklanır.

Bir doğal dil metnini ele aldığımızda, bir dokümanın anlam bilimsel *(İng.* semantic) temsili için o dokümanda geçen tüm kelimelerin eşit derecede etkili olmadığını, bazı kelimelerin, örneğin isim tamlamalarının, daha önemli olduğunu fark ederiz. Dolayısıyla bir dokümanın içeriğinin temsili için dokümanda yer alan kelimelerin tümünden ziyade bir alt kümesinin seçilebilmesi yeterlidir. Bu gözlemden hareketle, ters indekste yer alacak terimlerin önemli terimlerden oluşmasını sağlamak ve ters indeksin boyutunu azaltmak için doküman metinleri indekslenmeden önce bazı önişleme *(İng.* preprocess) aşamalarına tabii tutulur. Sözcüklere ayırma (*İng*. tokenization), gereksiz kelimelerin ayıklanması (*İng.* stop words removal), gövdeleme (*İng.* stemming) bu aşamada uygulanabilecek metin önişleme seçeneklerindendir.

*Sözcüklere Ayırma:* Serbest metnin sözcüklere (*İng.* word) ya da birimlere *(İng.* token) ayrıştırılma işlemidir. Bazı sistemler için boşluklara bakılarak metnin birimlere ayrılması kadar basit bir işlem iken; bazı sistemler için daha karmaşık birimlere ihtiyaç duyulabilir. Örneğin; e-posta adreslerinin bulunması ve bunların bir birim olarak tanınması gibi. Hatta kelimeler arasına boşluk bırakılmadan yazılan Çince, Japonca gibi diller için oldukça zahmetli bir işlemdir. Bu dillerde sözcüklere ayırma işlemi için kelime segmentasyonu kullanılır.

*Gereksiz Kelimelerin Ayıklanması:* Bazı kelimeler, içerisinde bulundukları belgenin içeriğini/anlamını temsil etmez, ancak dil bilgisel gereksinimler için kullanılır. Bu son derece sık kullanılan ve kendi başına bir anlamı bulunmayan sözcükler yapısal sözcükler (*İng.* function words) ya da gereksiz sözcükler (*İng.* Stop words) olarak adlandırılır. Cümle kurmaya yardımcı olmak için kullanılan tanımlıklar (*İng.* articles), edatlar, bağlaçlar ve bazı zamirler bu tür sözcüklere doğal adaydır. Örneğin, İngilizce için “a”, “an”, “at”, “be”, “for”, “to”, “the”, “of”, “on”, “or”, “that” yaygın kullanılan gereksiz kelimelerdendir. Bu sözcüklerin indekslenmemesi gereksiz kelimelerin ayıklanması işlemi olarak adlandırılır. Bu işlem indeks boyutunu azaltır, fakat bazı dezavantajları da vardır. Örneğin, “to be or not to be” gibi bir sorgu için herhangi bir dokümana erişmek mümkün olmayacaktır. Ayrıca “the baby”, “the past”, “the book” gibi iki terimli sorgular için döndürülen sonuçlar hiçbir anlam ifade etmeyecektir.

*Gövdeleme:* Kelimeleri ortak bir temel forma indirgemek için yapım eklerinin korunup, (çekim) eklerinin kaldırılması işlemidir. Örneğin; "walks", "walking", "walker", ve "walked" kelimeleri gövdeleme sonucunda köklerine indirgenerek "walk" olur. Literatürde gövdeleyici (*İng.* stemmer) olarak da adlandırılan birçok gövdeleme algoritması mevcuttur. İngilizce için en yaygın olarak kullanılan gövdeleme algoritması kural tabanlı Porter Stemmer'dır (Porter, 1980). Porter Stemmer'ın daha az agresif alternatifi olan KStemming yaygın kullanılan diğer gövdeleme algoritmasıdır (Krovetz, 1993).

### 2.1.2. Sorgu formülasyonu

BE sistemi kullanıcıları, bilgi gereksinimlerini karşılayacak dokümanları elde etmek için sisteme bir sorgu gönderir. Kullanıcının girdiği bu sorgu doğrultusunda sistem ilgili olarak bulduğu belgeleri kullanıcıya gösterir. Buradaki önemli nokta, kullanıcının bilgi gereksiniminin bir sorgu olarak ifade edildiği, yani sistemin girdisinin bu sorgu olduğudur. Kullanıcının bilgi gereksinimi sorgu olarak ne kadar doğru formüle edebilirse, sistemin sonuçları da o derecede doğru olacaktır. Fakat, kullanıcının bilgi gereksinimi ile BE sistemine gönderilen gerçek sorgu arasında bir ayrım vardır. Örneğin, kullanıcının bilgi gereksinimi şöyle olsun:

"Kolej tenis takımları hakkında bilgi içeren tüm belgeleri bul, öyle ki; (1) ABD'de bir üniversite tarafından sürdürülsün, (2) NCAA tenis turnuvasına katılsın. Ayrıca takımın son üç yıldaki ulusal sıralaması ile takım koçunun e-posta adresi ya da telefon numarası bilgisi belge içinde yer alsın." (Baeza-Yates ve Ribeiro-Neto, 1999).

Açıktır ki, kullanıcının bilgi gereksiniminin bu tam metni, BE sisteminden bilgi istemek için doğrudan kullanılamaz. Bunun yerine, kullanıcı önce bu bilgi gereksinimini BE sistemi tarafından işlenebilecek bir sorguya çevirmelidir. En yaygın haliyle, bu çeviri, kullanıcı bilgi gereksiniminin tam metnini özetleyen bir dizi anahtar kelimeler (*İng.* keyword) ya da indeks terimleri vermelidir. Kullanıcı bilgi gereksiniminin sorgu olarak formüle edilmesi basit bir problem değildir. Kullanıcı bilgi gereksinimini daha iyi formüle edebilmek için anahtar kelime sorguları, boole sorguları, ifade sorguları, yakınlık sorguları, tam belge sorguları ya da doğal dil sorusu gibi farklı sorgu oluşturma yöntemleri kullanabilir (B. Liu, 2007).

Web tabanlı arama motorlarında, son kullanıcılar sorgu oluşturma ile ilgili eğitim almamıştır; zaten böyle bir eğitim almaya ilgileri de yoktur; dolayısıyla yapılan

aramaların genellikle birkaç kelimeyi geçmediği görülür. Ticari arama motorları, kullanıcılarının istedikleri bilgiye daha hızlı ulaşabilmesi ve daha iyi arama deneyimi edinebilmesi için farklı geri bildirim özellikleri sunarak (örneğin, Bunu mu demek istediniz? Benzer aramalar, Resimler içerisinde ara, Haberlerde ara vs …) sorgu oluşturmalarında kullanıcılarına rehberlik eder.

Sorgu formalizasyonu süreci sonunda oluşturulup BE sistemine gönderilen sorgu, indeksleme aşamasında doküman koleksiyonuna uygulanan aynı metin önişleme aşamalarına (sözcüklere ayırma, gereksiz kelimelerin ayıklanması, gövdeleme gibi) tabii tutulduktan sonra ilgili belgelerin elde edilebilmesi için işlenir. Önceden oluşturulmuş ters indeks yapısı, hızlı sorgulama işlemini mümkün kılar. Böylelikle, sorgu terimleri için indekslenmiş belgeler üzerinde eşleştirme yapmak kolay olur.

### 2.1.3. Eşleştirme

İdeal bir BE sisteminin, girilen bir kullanıcı sorgusuna karşılık, sadece ilgili belgeleri, en çok ilgili olandan en az ilgili olana göre sıralı bir biçimde, getirmesi beklenir. Eşleştirme aşamasında, sorgu terimlerini içeren tüm belgeler ters indeks yapısından çekilir. Bir belgenin verilen sorgu ile ilgi (*İng.* relevance) düzeyi, çeşitli BE modelleri tarafından tahmin edilebilir. Literatürde BE modelleri birçok farklı şekilde sınıflandırılmıştır (Hiemstra, 2009). Boole Modeli (*İng.* Boolean Model), Vektör Uzay Modeli (*İng.* Vector Space Model) ve Dil Modeli (*İng.* Language Model) en çok kullanılan standart BE modelleridir. Her ne kadar bu modellerde belgeler ve sorgular farklı şekillerde temsil edilse de aynı yapıyı kullanırlar. Öyle ki, her bir doküman ya da sorgu kelimelerden oluşan birer küme (*İng.* bag of words) olarak ele alınır.

#### *2.1.3.1. Boole modeli*

Boole modeli en eski BE modellerinden biridir (Lancaster ve Gallup, 1973). Modelde dokümanlar ve sorgular terim kümeleri olarak temsil edilir. Sorgu terimlerini birleştirmek için George Boole'nun matematiksel mantığının operatörlerini (VE, VEYA, DEĞİL (*İng*. AND, OR, NOT)) kullanır. Örnek olarak, "masa VE kitap (DEĞİL kalem)" şeklinde yazılmış bir boole sorgusu için, boole modeli hem masa hem de kitap terimlerini içeren ama kalem terimini içermeyen dokümanları getirir.

Boole modeli kesin eşleştirme yaklaşımına dayandığından, kullanıcı sorgusuna yanıt olarak çok az ya da çok fazla belge döndürür. Bu model sıralama

mekanizmasından yoksundur. Belgeler döndürülür ya da döndürülmez, ancak sonuç kümesindeki belgeler sıralanmaz. Bu nedenle sonuç kümesindeki bütün dokümanlar eşit önemde kabul edilir. Modelin bir diğer dezavantajı, sorgu formülasyonu sürecindeki kısıtlamalardır, sorgu iyi tanımlanmamışsa yanıltıcı sonuçlar dönebilir. Bütün bu dezavantajlarından dolayı, Boole modelli, ne modern BE sistemleri, nede web araması için uygun görülmemektedir.

#### *2.1.3.2.Vektör uzay modeli*

Vektör uzay modeli en çok bilinen ve en çok kullanılan BE modelidir. Modelde dokümanlar ve sorgular yüksek boyutlu bir Öklit uzayına *(İng.* Euclidean Space) yerleştirilmiş vektörler olarak kabul edilir (Salton ve McGill, 1986). Her bir terim ayrı bir boyutla temsil edildiğinden, uzayın boyutu indeksteki tekil terimlerin (*N*) toplam sayısına eşittir. *d* vektörü ile temsil edilen bir doküman ile *q* vektörü ile temsil edilen bir sorgu arasındaki benzerlik, Denklem (2.1)'de gösterildiği gibi bu iki vektör arasındaki açısının kosinüsü olarak hesaplanabilir. cos(0°) = 1 ve cos(90°) = 0 olduğuna dikkat edilmelidir.

$$score(\vec{d},\vec{q}) = cos(\theta) = \frac{\vec{d}\cdot\vec{q}}{|\vec{d}|\times|\vec{q}|} = \frac{\sum_{i=1}^{N} d_i \times q_i}{\sqrt{\sum_{i=1}^{N}(d_i)^2}\times\sqrt{\sum_{i=1}^{N}(q_i)^2}} \qquad \textbf{(2.1)}$$

Vektör uzay modelinin ana dezavantajı, vektör bileşenlerinin değerlerinin ne olması gerektiğinin tanımlanmamış olmasıdır. Vektör bileşenlerine uygun ağırlıkların atanması problemi, **terim ağırlıklandırma** *(İng.* term weighting) olarak bilinir. Yapılan deneyler terim ağırlıklandırmanın hiç de önemsiz bir sorun olmadığını, hatta doküman ve sorgu eşleştirmesinde en önemli kriter olduğunu ortaya koymuştur (Salton, 1971; Salton ve Yang, 1973).

Salton ve Yang, terim ağırlıklandırma için Denklem (2.2)'de formülize edilen *tf.idf* terim ağırlıklandırmasını önermiştir (Salton ve Buckley, 1988). Bir terimin bir dokümanda kaç kez geçtiği terim frekansını (*tf*), o terimi içeren doküman sayısı ise doküman frekansını (*df*) gösterir. Bir terim ne kadar çok dokümanda geçiyorsa o kadar az bilgi içeriyor anlamına geldiğinden; *tf* ile ters doküman frekansının (*idf*) kombinasyonu *tf.idf* terim ağırlığıdır.

$$weight(t,D) = tf(t,D) \times log_2 \frac{N}{df(t)} \quad \textbf{(2.2)}$$

BE sistemlerinde kullanılan birçok yeni ağırlıklandırma algoritması *tf.idf* ağırlıklandırmasındaki kavramlara dayanır. BM25 (Robertson ve Zaragoza, 2009), PL2 (Amati ve Van Rijsbergen, 2002), LGD (Clinchant ve Gaussier, 2011) ve DFIC (Kocabaş vd., 2014) bu terim ağırlıklandırma modellerine örnektir. Bu modeller doküman frekansı, terim frekansı ve ortalama doküman uzunluğu gibi çeşitli istatistiki bilgiler kullanır. Dolayısıyla bu modeller makine öğrenmesini kullanmayan statik modellerdir.

#### *2.1.3.3.Dil modeli*

Dil Modelleri (*İng.* Language Model (LM)), otomatik konuşma tanıma sistemlerinde başarıyla kullanılmıştır. Özellikle, akustik modelin oluşturduğu aday metinler içerisinden, en olası metni seçmek için LM kullanılır. Örneğin, aday metinlerin "food born thing", "good corn sing", "mood morning" ve "good morning" olduğunu varsayalım. LM modeli, "good morning" ifadesinin çok daha olası olduğunu saptayacaktır. Çünkü "good morning" ifadesi İngilizcede diğer ifadelerden daha sık görülür (Hiemstra, 2009).

Dil modellerinin BE alanında kullanımı çok yenidir (Hiemstra, 2001; Lafferty ve Zhai, 2001; Ponte ve Croft, 1998). BE'e LM yaklaşımında, her doküman için bir LM oluşturulur. Sorgunun belli bir dokümanın LM'si tarafından oluşturulma olasılığı, her doküman için hesaplanır. Sorgu birden fazla terimden oluşuyor ve hesaplamalarda terimlere tekil olarak bakılıyorsa (unigram modeli), Denklem (2.3)'de gösterildiği gibi her bir terimin olasılığının çarpımıyla bulunacak değer, sorgu sonucu olarak gösterilecek dokümanları sıralamada kullanılır.

$$P(t \in Q|D) = \prod_{t \in Q} P(t_i|D) \;\; where \;\; P(t|D) = \frac{tf_t, D}{|D|} \quad \textbf{(2.3)}$$

Denklem (2.3), sorgu terimlerinin tümünü içermeyen belgelere sıfır olasılık atar. Sıfır olasılığından kaçınmak için yumuşatma *(İng.* smoothing) denilen bir yöntem

uygulanır. Dokümanda olmayan bir terime sıfırdan farklı *(İng.* non-zero) bir olasılık değeri atanır. Jelinek-Mercer, Dirichlet Prior ve Absolute Discount gibi farklı yumuşatma algoritmalarının LM yaklaşımıyla BE üzerine etkileri araştırılmış (Zhai ve Lafferty, 2004), Dirichlet Prior yumuşatma algoritmasının daha etkili olduğu bulunmuştur (Croft vd., 2010).

### 2.1.4. Yeniden sıralama

Yeniden sıralama işlemi, eşleştirme süreci sonrası yapılır. Eşleştirme sürecinde, BE sisteminin kullandığı standart terim ağırlıklandırma modeli (BM25, PL2, LM) verilen bir sorgu için bir belge listesi döndürür. Referans terim-ağırlıklandırma modeli tarafından döndürülen listedeki üst-K *(İng.* top-N) belgenin (örneklem) makine öğrenme teknikleri kullanılarak yeniden sıralanması, Sıralamayı Öğrenme *(İng.* Learning to Rank) olarak adlandırılır (T.-Y. Liu, 2009). Sıralamayı öğrenme, bir belge örneği için çıkarılan çeşitli özelliklerin, etkili öğrenilmiş modellere *(İng.* effective learned mode) yerleştirilmesini ve sonrasında yeniden sıralama *(İng.* re-rank) için kullanılmasını içerir (Macdonald, Santos ve Ounis, 2013). Çok sayıda özellik etkili öğrenilmiş model içinde birleştirilir. Örneğin, hem sorgudan bağımsız belge özellikleri (belgeye gelen bağlantılar, URL derinliği, vb.), hem de sorgu terimleri için farklı terim ağırlık modelleri (BM25, *tf*, *idf*, vb.) tarafından belgenin alanlarına (başlık, gövde, çapa metni, vb.) atanan ağırlık/puan özellikleri gibi sorguya bağlı özellikler dahil edilebilir (Macdonald, Santos, Ounis, vd., 2013).

Sıralamayı öğrenme konusu BE topluluğunda büyük ilgi görmektedir. Nitekim son yıllarda sıralamayı öğrenme yöntemlerinin geliştirilmesi ve karşılaştırılması üzerine birçok çalışma yapılmaktadır (Tax vd., 2015).

## 2.2. Web Bilgi Erişiminin Kendine Özgü Özellikleri

İnternet'in doğuşu ve yaygınlaşmasıyla beraber her geçen gün daha çok web sitesi oluşmaya başladı. Giderek büyüyen web sitesi sayısından dolayı web sitelerine ulaşmak zorlaştı. Bu sorunu çözümlemek için webe özgü arama önemli bir ihtiyaç haline geldi.

Web'in popüler bir iletişim aracı olarak büyümesi, Web Bilgi Erişimi (WebBE, *İng.* Web Information Retrieval) alanının gelişimini teşvik etmiştir. Ayrıca web arama motorları *(İng.* Web search engines), web'in en can alıcı uygulaması olmuş, web kullanıcıları zamanının çoğunu web arama motorlarında harcamaya başlamıştır.

WebBE, BE temellerine dayanır. BE teori ve metodolojilerinin World Wide Web (WWW)'de uygulanması olarak da tanımlanabilir. Ancak "web araması" geleneksel BE modellerinin basit bir uygulaması da değildir. Her ne kadar bazı BE sonuçlarını kullansa da kendine özgü teknikleri vardır ve BE alanında birçok yeni araştırma konusu ortaya çıkarmıştır (Nunes, 2006). Geleneksel BE ile karşılaştırıldığında, WebBE farklı zorluklarla karşı karşıyadır. BE ve WebBE arasındaki temel farklar aşağıda sıralanmıştır.

Geleneksel BE'inde temel birim **doküman**, WebBE'de ise **web sayfasıdır**. Ayrıca, geleneksel BE sistemleri büyük doküman koleksiyonları ile çalışırken; Web'deki sayfa sayısı devasadır. Web sayfaları farklı sunucularda ve farklı coğrafi konumlarda barındırılır. Bu nedenle, geleneksel BE'ne göre veriye erişim fiziksel zorluklar içerir. Arama motorlarındaki içeriklerin güncel tutulabilmesi ve yeni içeriklerin dahil edilebilmesi için "web crawler" olarak adlandırılan yazılımlar kullanılır. Web crawler, periyodik olarak web'i tarayarak, arama motorlarının güncel kalmasını sağlar.

Geleneksel BE sistemleri için ikincil derecede önemli olan **verimlilik** (*İng.* efficiency), WebBE için en önemli konudur. Web kullanıcıları çok hızlı yanıtlar bekler. Bir algoritma ne kadar etkili olursa olsun eğer getirilen sonuçlar yeterince hızlı oluşturulmuyorsa kimse o sistemi kullanmak istemez.

Web sayfaları, geleneksel BE sistemlerinde kullanılan alışılagelmiş metin dokümanlarından oldukça farklıdır (Craswell ve Hawking, 2009). İlk olarak, metin dokümanlarından farklı olarak web sayfalarında üst bağlar (*İng.* hyperlinks) ve çapa metinleri (*İng.* anchor texts) mevcuttur. Üst bağlar arama için çok önemlidir. Ayrıca, arama sıralama algoritmalarında (*İng.* search ranking algorithms) merkezi bir rol oynarlar. Üst bağlarla ilişkilendirilmiş çapa metinler de çok önemlidir, çünkü bir çapa metin parçası genellikle üst bağının işaret ettiği sayfanın daha doğru bir açıklamasıdır. İkinci olarak, web sayfaları yarı-yapılandırılmıştır (*İng.* semi-structured). Bir web sayfası, geleneksel bir dokümandaki gibi sadece birkaç paragraflık metin değildir. Web sayfalarında başlık (*İng.* title), meta veriler, gövde (*İng.* body) gibi farklı alanlar vardır. Belirli alanlarda bulunan bilgiler (örneğin, başlık alanı gibi) diğerlerine kıyasla daha önemlidir. Ayrıca, bir web sayfasındaki içerik tipik olarak çeşitli yapısal bloklarda (dikdörtgen şekiller) düzenlenir ve sunulur. Bu bloklardan bazıları önemliyken; bazıları değildir (örneğin, reklamlar, gizlilik politikası, telif hakkı bildirimleri gibi). Bir web

sayfasının ana içerik bloklarının etkili bir şekilde tespit edilmesi WebBE için faydalıdır, çünkü bu tür ana bloklarda yer alan terimler daha önemlidir. Web sayfaları ile geleneksel metin dokümanları arasındaki bu farklar, ekstra metin önişleme ihtiyacını da beraberinde getirir. Web sayfalarına uygulanabilecek başlıca metin önişleme işlemleri: web sayfalarındaki farklı metin alanlarının belirlenmesi, çapa metinlerin belirlenmesi, HTML etiketlerinin kaldırılması, ana içerik bloklarının bulunmasıdır.

Dokümanların yinelenmesi (*İng.* duplication) geleneksel BE için sorun teşkil etmez. Fakat, web bağlamında önemli bir sorundur. Web sayfalarının ve içeriklerinin yinelenmesinin farklı türleri vardır. Bir sayfanın kopyalanması genellikle yineleme veya kopyalama olarak adlandırılır. Bir sitenin tamamının kopyalanması yansıtma olarak adlandırılır. Yinelenen sayfalar ve yansıtma siteleri, genellikle farklı coğrafi bölgelerdeki sınırlı bant genişliği ve zayıf veya öngörülemeyen ağ performansları nedeniyle dünya genelinde tarama ve dosya indirme verimliliğini artırmak için kullanılır. Tabii ki, bazı yinelenen sayfalar intihal sonuçlarıdır. Bu sayfaları ve siteleri tespit etmek, indeks boyutunu azaltır ve arama sonuçlarını iyileştirir.

Spam, Web'deki önemli bir sorundur; ancak geleneksel BE için bir endişe değildir. Bir arama motorunun döndürdüğü sayfanın sıralamadaki konumu son derece önemlidir. Eğer bir sayfa bir sorguyla ilgili ancak çok düşük bir sırada (Ör. İlk 30. sırada) yer alıyorsa, kullanıcının sayfaya bakması olası değildir. Sayfa bir ürün satış sayfası ise, bu bir ticari kayıptır. Bazı hedef sayfaların arama motoru sıralamasındaki yerini iyileştirmek için "uygunsuz" olarak yapılan çalışmalar "spam" olarak adlandırılır. Spam olarak üretilen düşük kaliteli (hatta alakasız) sayfaların, arama sonuçlarının en üst sıralarında çıkması, arama sonuçlarının kalitesine ve kullanıcı arama deneyimine zarar verir. Bu yüzden web spam'inin algılanması ve mücadele edilmesi kritik bir konudur.

WebBE'de arama sonuçlarının sıralanması geleneksel BE'e göre farklılık gösterir. Web'in üst bağ yapısı WebBE için çok önemlidir. Web sayfalarının birbirlerine yaptığı bağlantılar arama sonuçlarının sıralanmasında kullanılır (Nunes, 2006). Öyle ki; güvenilir olarak işaretlenmiş bir sayfadan diğer bir sayfaya yapılan bağlantı, o hedef sayfanın güvenilirliğini artırır. Diğer taraftan, spam olarak işaretlenmiş bir sayfanın hedef bir sayfaya yaptığı bağlantı, o hedef sayfanın sıralamasını olumsuz etkiler. PageRank (Brin ve Page, 1998) ve HITS (Kleinberg, 1999), web sayfaları arasındaki bağlantı analizinde kullanılan başlıca algoritmalardır.

## 2.3. Bilgi Erişim Sistemlerinde Test Koleksiyonu Bazlı Değerlendirme

Ölçüm, bir BE sisteminin kullanıcılarına yardımcı olma hedefini ne derece başarabildiğinin göstergesi olarak; etkili BE sistemlerinin tasarlanması ve geliştirilmesi için son derece önemlidir. Geliştirilen BE sisteminin başarım hedefi, sonuçları ne kadar hızlı döndürüldüğü, kullanıcı etkileşimini ne kadar desteklediği, kullanıcıların sonuçlar hakkındaki memnuniyeti, sistemin ne derece kullanılabilir olduğu gibi farklı kriterlerden biri olabilir. Alanda yapılan çalışmaların başladığı ilk günden beri "sistem-yönelimli etkinlik" *(İng.* system oriented evaluation) ya da kısaca "sistem etkinliği", başarım ölçümünde birincil yaklaşım olmuştur (Voorhees, 2001). Sistem etkinliği, kullanıcıların bilgi ihtiyacının geliştirilen BE sistemi tarafından ne oranda karşılandığının; bir başka ifadeyle, verilen bir kullanıcı sorgusu için sistemin ilgili dokümanlar ile ilgili olmayan dokümanlar arasında ayrım yapabilme yeteneğinin, ölçümüdür (Voorhees, 2001).

BE araştırmalarının büyük çoğunluğu, etkinlik ölçümü için test koleksiyonlarını kaynak olarak kullanır. Test koleksiyonları, sağladıkları standardizasyon sayesinde farklı BE algoritmalarının etkinliklerinin ölçümünde ve karşılaştırmasında önemli rol oynar (Sanderson, 2010).

Literatüre bakıldığında, BE alanında yapılan ilk çalışmalarda karşılaşılan en önemli problemin araştırmacıların üzerinde çalışabileceği ortak bir veri setinin (test koleksiyonun) olmaması, dolayısıyla da geliştirilen BE sistemlerinin etkinliklerinin karşılaştırılamaması olduğu görülür. Bunu takiben, BE alanında çalışan akademik topluluklar tarafından, farklı BE algoritmalarının etkinliklerinin ölçümünde ve karşılaştırmasında, geliştirilen sistemlerin doğrulanmasında ve tekrar edilebilmesinde standart olacak test koleksiyonlarının oluşturup, paylaşılmasının gerekliliği ortaya konulmuştur (Harman, 2011; Robertson, 2008; Sanderson, 2010). Etkinlik ölçümünde test koleksiyonu tabanlı ölçüm yaklaşımının temelleri, İngiltere'deki Cranfield kütüphanesinde 1958 ve 1966 yılları arasında yapılan deneylere dayanır ve genellikle "Cranfield yaklaşımı" olarak adlandırılır. Bu yaklaşım, farklı indeksleme stratejilerini operasyonel ortamdan soyutlanmış kontrollü koşullar altında ölçebilmek için, ortak bir doküman koleksiyonuna, sorgu görevlerine *(İng.* query tasks) ve esas alınacak gerçeklere *(İng.* Ground truths) ihtiyaç olduğunu ortaya koymuştur. Ayrıca, bu üç bileşenin bir ölçüm kriteri ile birlikte operasyonel bir ortamda bir arama sistemi

kullanıcılarını simule edebileceğini ve BE sistem etkinliği ölçümünde kullanılabileceğini göstermiştir. Dahası, BE sistemlerinin belirtilen yöntem ile ölçümü, farklı arama algoritmalarının karşılaştırılmasını ve algoritma parametrelerinin değişiminin sistematik olarak gözlemlenebilmesini olanaklı hale getirmiştir (Clough ve Sanderson, 2013).

1960'larda önerilmesine rağmen, Cranfield yaklaşımı, National Institute of Standards and Technology (NIST) tarafından finanse edilen ve 1992 yılında başlayan Text REtrieval Conference (TREC) serisinin büyük ölçekli ölçüm yarışmalarıyla popüler hale geldi. TREC serisi ile hedeflenen, BE metodolojilerinin büyük ölçekli ölçümü için gereken altyapıyı sağlayarak, BE alanındaki araştırmaları desteklemekti. Bu anlamda en geniş tanınırlığa sahip olanı TREC olsa da CLEF, NTCIR ve FIRE gibi daha sonraki yarışmalar sayesinde geçtiğimiz 20 yıl içinde BE alanı çok önemli teknolojik gelişmelere sahne olmuştur (Voorhees ve Harman, 2005).

TREC serisinde de esas alınan, test koleksiyonu tabanlı ölçüm yaklaşımının uygulanması için gerekli olan temel bileşenler şunlardır:

- *Doküman Koleksiyonu*: Bu doküman koleksiyonu içerisinde yer alan her belgenin "*docid*" olarak isimlendirilen, özgün bir doküman numarası vardır.
- *Sorgu ya da Konu Kümesi (İng.* Queries or Topics): Bilgi gereksinimlerini ifade eder. Her sorgunun "*qid*" olarak isimlendirilen bir sorgu numarası vardır.
- *Sorgu Uygunluk Kümesi - qrels* (*İng.* Query relevance set): Hangi dokümanların hangi sorgular ile ilgili olduğunu gösteren sorgu uygunluk yargıları *(İng.* Relevance judgements) ya da doğru yanıtlardır (Voorhees, 2007). Genellikle "*qrels*" olarak isimlendirilen, *qid/docid* çiftlerinden oluşan bir listedir.

TREC Doküman Koleksiyonları: BE alanındaki araştırmaların günümüze gelene kadar değişen öncelik ve ihtiyaçlarını yansıtacak şekilde test koleksiyonlarının içeriklerinde ve kullanılan ölçüm metriklerinde farklılıklar gözlemlenir. Genel olarak üç dönemden bahsedilebilir (Sanderson, 2010).

1. Dönem (1950'lerin başı ile 1990'ların başı arası), test koleksiyonlarının ve ölçüm metriklerinin ilk geliştiği dönemdir. Bu dönem içerisinde, test koleksiyonlarının

içerikleri çoğunlukla akademik makaleler ya da gazete makalelerinin tam metninden oluşuyordu. Başarım ölçümü olarak yüksek recall'a odaklanılmıştı.

2. Dönem (1990'ların başı ile 2000'lerin başı), "TREC ad hoc" dönemi olarak adlandırılır. Ölçümün standardizasyonu ve ölçeklenebilirliği bu dönemin en güçlü temasıdır. BE araştırma toplulukları, çoğunlukla haber makalelerinden oluşan nispeten az sayıda büyük test koleksiyonları oluşturmak için iş birliği yaptı. Yüksek recall bu dönem içinde en baskın başarım ölçütüdür.

3. Dönem (2000'li yılların başından günümüze), "TREC post ad hoc" dönemi olarak da adlandırılan; web'in etkin olduğu dönemdir. Arama teknolojilerinin uygulamadaki artan çeşitliliği, sürekli ölçeği büyüyen, farklı içerik yelpazesine sahip koleksiyonlar yaratma gereksinimini doğurdu. Araştırmacıların tüm web üzerinde çalışabilme olanakları yoktu. BE araştırmalarında kullanılmak üzere, web'in yapısını yansıtabilecek, büyük ölçekli koleksiyonlar oluşturuldu. Ayrıca, web kullanıcılarının yaygın tercihi az sayıda ilgili doküman bulmak olduğundan bu dönemde ölçüm kriterleri de değişerek çeşitlilik kazanmıştır. ClueWeb09 ve ClueWeb12, Web Track, Million Query Track, Robust Track, Terabyte Track gibi farklı görevleri hedefleyen TREC serisi kapsamında gerçekleştirilen yarışmalarda kullanılmak üzere oluşturulan büyük ölçekli koleksiyonlardır.

TREC Sorgusu (Konu Kümesi): TREC konuları, bir konunun tam olarak anlaşılabilmesini sağlamak amacıyla, sorgunun altında yatan bilgi ihtiyacının ayrıntılı bildirimini verecek şekilde XML-benzeri bir şemada yapılandırılmıştır. Şekil 2.2'de örnek bir bilgi gereksinimi için verilen TREC sorgu gösterilmiştir.

```
<topic number="152" type="faceted">
<query>angular cheilitis</query>
<description>
        What home remedies are there for angular cheilitis?
    </description>
</topic>
```

**Şekil 2.2.** *TREC ad-hoc gereksinim örneği*

TREC Sorgu Uygunluk Yargıları: TREC sorgu uygunluk yargıları, *qid/docid* çiftlerinden oluşan bir listedir. Bir sorgu uygunluk yargısı (*qrels*[1]) genellikle dört kısımdan meydana gelir:

1. TOPIC#: Sorgunun kimliğini belirten numaradır (*qid*).
2. ITERATION: Geri bildirim tekrarlama değeridir (Genelde 0 değerini alır ve kullanılmaz).
3. DOCUMENT#: Dokümanın koleksiyon içindeki kimliğini belirten numaradır (*docid*).
4. RELEVANCY: Son sütunda gösterilen bu tamsayı değeri, 1'den küçük olduğunda gösterilen *qid/docid* çiftinin ilgili olmadığını *(*non-relevant), 0'dan büyük olduğunda ise ilgili *(*relevant) olduğunu belirtir.

Şekil 2.3'de, Şekil 2.2'de verilen örnek gereksinime ilişkin sorgu uygunluk yargıları verilmiştir. Örnek olarak, Şekil 2.3'nin ilk satırına bakıldığında; 152 numaralı sorgunun (*qid* = 152), clueweb09-en0001-76-17979 numaralı doküman (*docid* = clueweb09-en0001-76-17979) ile ilgi düzeyinin 0 (*relevancy*=0) olduğu görülür. Her satırın ikinci sütununda yer alan değer geri bildirim tekrarlama değeridir ve 0'dır (*iteration*=0).

```
152  0  clueweb09-en0001-76-17979   0
152  0  clueweb09-en0002-16-13298   -2
152  0  clueweb09-en0019-27-20755   1
152  0  clueweb09-en0019-27-20761   0
152  0  clueweb09-en0019-27-20762   4
152  0  clueweb09-en0033-02-06291   0
152  0  clueweb09-en0033-32-30325   1
152  0  clueweb09-en0033-32-30326   4
152  0  clueweb09-en0033-32-30327   0
152  0  clueweb09-en0033-32-30329   1
152  0  clueweb09-en0033-32-30331   0
152  0  clueweb09-en0033-32-30332   0
152  0  clueweb09-en0033-32-30345   1
152  0  clueweb09-en0034-45-03577   0
152  0  clueweb09-en0044-69-33939   1
152  0  clueweb09-en0047-19-40711   0
152  0  clueweb09-en0119-85-23133   0
152  0  clueweb09-en0120-57-23674   1
152  0  clueweb09-en0129-86-40148   0
```

**Şekil 2.3.** *TREC sorgu uygunluk yargı örneği*

[1]https://trec.nist.gov/data/qrels

Uygunluk değerleri ikili (0 = uygun değil, 1 = uygun) veya kademeli (2 = çok uygun *(İng.* Highly relevant), 1 = uygun, 0 = uygun değil, -2 = spam/junk) değerler alır. Bu uygunluk değerleri her bir sorgu-doküman çifti (*İng.* query-document pair) için gerçek değerlendiriciler (*İng.* Human assessors) tarafından elle atanmıştır. Anlaşılabileceği gibi, uygunluk değeri belirleme süreci fazlasıyla zaman alan ve maliyetli bir süreçtir. Bu nedenle, TREC havuzlama (*İng.* pooling) adı verilen bir yönteme başvurur. Havuzlama yönteminde, değerlendiriciler sadece TREC katılımcıları tarafından her sorgu için getirilen üst-k (genellikle 100) doküman kümesinin birleşimindeki dokümanları değerlendirir ve uygunluk değerlerini belirler.

TREC Sonuç Gönderim Dosyası: TREC katılımcıları, her bir sorgu için koleksiyonda ilgili olarak belirledikleri dokümanları, ilgi düzeyine göre azalan sırada döndürür. Konu kümesindeki her sorgu için, deneysel olarak üretilen, en üstteki K (genellikle ilk 1000) doküman gönderim dosyasına kaydedilir.

```
5 Q0 clueweb09-enwp02-06-01125 1 32.38 example2012
5 Q0 clueweb09-en0011-25-31331 2 29.73 example2012
5 Q0 clueweb09-en0006-97-08104 3 21.93 example2012
5 Q0 clueweb09-en0009-82-23589 4 21.34 example2012
5 Q0 clueweb09-en0001-51-20258 5 21.06 example2012
5 Q0 clueweb09-en0002-99-12860 6 13.00 example2012
5 Q0 clueweb09-en0003-08-08637 7 12.87 example2012
5 Q0 clueweb09-en0004-79-18096 8 11.13 example2012
5 Q0 clueweb09-en0008-90-04729 9 10.72 example2012
```

**Şekil 2.4.** *TREC gönderim dosyası örneği*

TREC sonuç gönderim dosyasının her bir satırı altı sütundan oluşur. Şekil 2.4'de bir örneği gösterilen bu dosyaların içerdiği sütunlar aşağıdaki bilgileri içerir:

1. Birinci sütun sorgu numarasıdır (TOPIC# ya da *qid*).
2. İkinci sütun şu an kullanılmamakta olup, her zaman "Q0" olmalıdır.
3. Üçüncü sütun resmi doküman adıdır (DOCUMENT# ya da *docid*).
4. Dördüncü sütun dokümanın sonuç sırasıdır.
5. Beşinci sütun dokümanın ilgi puanıdır.

6. Kullanılan erişim metodu hakkında basit bilgi veren altıncı sütuna "run tag" adı verilir.

Her TREC için, NIST bir test seti ve soru seti sağlar. Katılımcılar kendi erişim sistemlerini veriler üzerinde çalıştırır ve NIST'e sonuç listesi olarak üst sırada yer alan belgelerden oluşan listeyi gönderir. NIST bireysel sonuçları bir araya getirir, birleşimdeki belgeleri doğru olarak değerlendirir ve sonuçları belirler. TREC döngüsü, katılımcıların deneyimlerini paylaşmaları için bir forum olan bir çalışma alanı ile sona ermektedir.

## 2.4. Bilgi Erişim Etkinliği Ölçüm Metrikleri

BE etkinliğini ölçmek için birçok metrik geliştirilmiştir. Bunlardan en yaygın ve kabul görmüş olan üç metrik aşağıda sunulmuştur.

**Ortalama Averaj Duyarlılığı- Mean Average Precision (MAP):** Sabit bir k değeri için duyarlılık (P@k), top-k sonuç içerisindeki ilgili dokümanların oranıdır. Örneğin, 10 sıralama için duyarlılık (P@10). Ortalama duyarlılık *(İng.* Average precision, AP) ise hher bir ilgili dokuman erişiminden sonra erişilemeyen ilgili dokümanlar için 0 duyarlılık kabul ederek hesaplanan duyarlılık puanlarının ortalamasıdır (Buckley ve Voorhees, 2017).

$$AP(q) = \frac{1}{|R_q|} \sum_{r \in R_q} P@rank(r) \qquad \textbf{(2.4)}$$

$Rq$, $q$ sorgusu ile ilgili dokümanları gösterir. AP(q) hesaplanırken sıralama genellikle 1000'de bırakılır. Tüm sorgu kümesi (q $\in$Q) üzerinden AP değeri ortalaması ile MAP elde edilir. MAP ikili uygunluk yargıları için standart bir metriktir.

**Normalized Discounted Cumulative Gain (NDCG@k):** NDCG, kademeli ölçümler için oldukça sık kullanılan ve uygun bir metriktir (Järvelin ve Kekäläinen, 2002). k değeri için DCG aşağıdaki formül ile hesaplanır:

$$DCG@k = \sum_{i=1}^{k} \frac{2^{g_i} - 1}{\log_2(1+i)} \qquad \textbf{(2.5)}$$

NDCG ise bu değerin normalize edilmiş halidir, DCG'nin ideal DCG'ye bölümüyle elde edilir.

**Expected Reciprocal Rank (ERR@k):** Bu metrikte çok uygun dokümanların altında kalan dokümanlar göz önüne alınır. Yapılan ölçüm kullanıcının uygun bir dokümanı bulmak için harcayacağı zaman olarak kabul edilir. ERR(@k) hesaplaması aşağıdaki formül ile yapılır (Chapelle vd., 2009):

$$ERR@k = \sum_{i=1}^{k} \frac{R(g_i)}{i} \prod_{j=1}^{i-1} (1 - R(g_j)) \quad \textbf{(2.6)}$$

MAP ve NDCG sistemlerin genel performansını yansıtırken, ERR ise kesinlik önyargılıdır. ERR@20 ve NDCG@20 çoklu derecelendirilmiş uygunluk yargılarını işleyebilir. Bu metrikler TREC Web Tracks alt görevinde şu an kullanılan en etkili metriklerdir (Collins-Thompson vd., 2015).

# 3. ALANYAZIN

## 3.1. Web Spam

Günümüzde insanlar, web üzerinde giderek büyüyen veri dağından ihtiyaç duydukları yararlı bilgilere ulaşmak için arama motorlarına daha fazla güvenir (Hochstotter ve Koch, 2009). Kullanıcılar ulaşmak istedikleri bilgi ile ilgili bir veya daha fazla anahtar kelimeyi bir arama motoruna sorgu olarak gönderir, arama motoru da sorgu ilgi düzeyine göre azalan bir şekilde sıralanmış uzun bir sonuç listesiyle kullanıcılarına yanıt verir. Önceki çalışmalar göstermiştir ki, arama motorlarına gönderilen sorguların %85'i için kullanıcılar sadece ilk sayfada gösterilen sonuçlar ile ilgilenir; hatta bu sonuçlar arasından da ortalama olarak sadece ilk üç ila beş bağlantı tıklanır (Silverstein vd., 1999). Arama motoru sonuçlarında daha yüksek sıralarda yer almak, arama motorlarından daha fazla trafik almakla doğru orantılı olduğundan, daha büyük ekonomik fayda sağlanması anlamına gelir. Arama sonuçlarında daha yüksek sıralama elde etmek isteyen spammer'lar, arama motoru algoritmalarını aldatmak için çeşitli spam teknikleri kullanır (Becchetti vd., 2008; Chellapilla ve Chickering, 2006; Cormack vd., 2011; Ntoulas vd., 2006; Radlinski, 2007). Bir web sayfası, arama motoru sonuç sıralamasını uygun olmayan bir şekilde etkilemek için herhangi bir spam tekniği kullanıyorsa, web spam olarak adlandırılır (Al-Kabi vd., 2012; Gyöngyi ve Garcia-Molina, 2005).

İyi bilinir ki, çok sayıda web sayfası, bir arama motoruna istenmeyen taşıma yükü sunma amacıyla oluşturulmuştur (Gyöngyi ve Garcia-Molina, 2005). Genel olarak bu amaç iki mekanizmayla gerçekleştirilir: kişisel tanıtım *(İng.* Self promotion) ve karşılıklı tanıtım *(İng.* Mutual promotion). Kişisel tanıtımın en yaygın şekli anahtar kelime doldurmadır. Bu yöntem, sayfanın sıralamasını iyileştirmek için gereksiz anahtar kelimelerin (genellikle kullanıcının göremeyeceği bir şekilde) sayfaya eklenmesi ile yapılır. Karşılıklı tanıtımın en yaygın hali ise link gruplarıdır *(İng.* link-farm). Bu yöntemde, PageRank gibi konudan bağımsız olan sayfa kalite puanını iyileştirmek için çok sayıda birbirine benzer sayfa kendi aralarında birbirlerine referans verir. İster kişisel tanıtım, ister karşılıklı tanıtım amaçlı olsun, spam sayfa içeriği çoğunlukla mekanik olarak ya da intihal ile oluşturulur (Cormack vd., 2011).

İlk olarak 1996 yılında (Convey, 1996) ortaya çıkan web spam kavramı, arama motoru şirketleri için önemli bir sorun olarak kabul edildi (Henzinger vd., 2002) ve birçok problemi de beraberinde getirdi (Zhuang vd., 2017). Birinci olarak, web spam, arama sonuçlarının kalitesinde ve verimliliğinde keskin bir düşüşe sebep olduğundan, arama motoru şirketlerinin itibarını azaltır. İkinci olarak, spam siteleri, rakiplerini haksız rekabette maruz bırakarak onları kazançlarından yoksun bırakır (Spirin ve Han, 2012). Üçüncü olarak, web spam, kötü amaçlı yazılımların ve yetişkinlere yönelik içeriklerin yayılması için önemli bir kaynak olduğundan müşterilerin kötü niyetli saldırılara maruz kalmasına sebep olur. Dördüncü olarak, web spam'inden kaynaklanan düşük kaliteli arama sonuçları, kullanıcının aradığı sonuçlara ulaşması için daha fazla zaman ve çaba harcaması gerektiği anlamına geldiğinden kullanıcı deneyimini olumsuz etkiler, hatta arama motoru şirketleri için müşteri kaybı anlamına gelir (Spirin ve Han, 2012). Son olarak, web spam, arama motoru şirketleri için spam sayfaların saklanması, indekslenmesi ve analiz edilmesi gibi ekstra kaynak kullanım maliyeti yarattığından ekonomik kayıplar vermelerine sebep olur.

### 3.2. Web Spam Üzerine Yapılmış Çalışmalar

Web spam üzerine yapılan ilk araştırmalar spam oluşturma teknikleri, spam önleme teknikleri ve web içeriğinin spam ya da non-spam olarak sınıflandırılmasına yönelik yöntemlere odaklanmıştır. WEBSPAM-UK2006 ve WEBSPAM-UK2007[2], Web Spam Challenge serisinde[3] bu amaçlarla kullanılmak üzere Yahoo tarafından hazırlanmış veri kümeleridir. Bu veri kümelerinin her biri, ".uk" web alanının bir kısmının taramasını içeriyor olup, her taramayı temsil eden birkaç bin ana bilgisayar örneği spam ya da non-spam şeklinde etiketlenmiştir. AIRWeb Web Spam yarışma serisi'nde, spam ana bilgisayarlarının *(İng.* host) tespiti için çeşitli yöntemlerin etkinliği ölçülse de bu yöntemlerin sistem erişim etkinliğine genel katkısına bakılmamıştır. WEBSPAM korpora üzerinde erişim etkinliğinin incelendiği tek çalışma olan (Jones vd., 2007)'da kullanıcıların spam filtrelenmiş sonuçları filtrelenmemiş sonuçlara tercih ettiğini ortaya konulmuş, fakat spam'in BE sistem etkinliği üzerindeki etkisi ölçülmemiştir. Aynı yazarlar, bir veri kümesinin spam'i etkisiz hale getirebilme

[2]http://barcelona.research.yahoo.net/webspam/datasets
[3]http://webspam.lip6.fr/wiki/pmwiki.php

etkinliğinin ölçülebilmesi için sahip olması gereken özellikleri araştırmıştır (Jones vd., 2009).

Önceki TREC Web BE ölçüm çabaları da büyük ölçüde spam içermeyen dokümanlardan oluşuyordu. Örneğin, TREC Terabyte Track alt görevinde hükümet web sayfalarından oluşan bir veri kümesi kullanılmıştı. Spam'in önemli bir sorun olduğu, 2006–2008 TREC Blog Tracks alt görevinde tespit edildi. Geliştirilen BE sistemlerinin, veri kümesinin %16'sını temsil eden bilindik spam bloglardan, yaklaşık %10'unu sonuçlar arasında getirdiği, spam'lerin kaldırılmasının bağıl sistem sıralamasını önemli ölçüde etkilemediği raporlandı (Macdonald vd., 2009).

Web'den herhangi bir kısıtlama olmaksızın taranan web sayfalarını içeren ClueWeb09 veri kümesinin kullanımı, spam sorununu merkeze taşıdı. Çünkü, üst sıralarda getirilen sonuçlar arasında spam olması, web arama motorlarının etkinliğine zarar veren bir sorundu. TREC 2009 yarışmasına katılan katılımcılardan en az üçü spam sayfaları filtreleyen post-retrieval stratejileri kullandı (Hauff ve Hiemstra, 2009; Lin vd., 2009). Örneğin, (Hauff ve Hiemstra, 2009), sayfa içeriklerine, sayfa başlık özelliklerine ve URL formuna dayanan bir spam algılama algoritması kullanarak spam dokümanları tespit ettikten sonra post-retrieval işlem aşaması yoluyla bu dokümanlar sıralamadan kaldırıldı. Benzer şekilde, (Lin vd., 2009) Yahoo'nun özel yetişkin, spam ve doküman kalite sınıflandırıcısı yardımıyla sonuçlardaki spam dokümanları tespit edip ayıkladı. Diğer katılımcılar da spam'in geliştirdikleri BE sistemlerinin performansına olumsuz etkilerine değindi (He vd., 2009; Kaptein vd., 2009; McCreadie vd., 2009). Fakat bu çalışmaların hiçbiri spam ve spam filtrelemenin BE sistem etkinliği üzerindeki etkisine yönelik nicel çalışma içermemekteydi.

ClueWeb09 büyüklüğündeki bir web veri kümesi üzerinde ilk spam çalışmasını ve spam filtrelemenin BE etkinliğine etkilerinin ilk sayısal sonuçlarını (Cormack vd., 2011) sunmuştur. ClueWeb09 veri kümesindeki spam web sayfaları saptamak için üç farklı eğitim seti üzerinden, her biri tüm veri kümesini etiketlemek için kullanılan içerik-tabanlı üç spam filtresi oluşturulmuştur. Ek olarak, naive Bayes metaclassifier kullanarak bu üç sonucu birleştiren ve "Fusion" olarak adlandırdıkları bir topluluk filtresi (*İng.* ensemle filter) de oluşturulmuştur. Bu eğitim setleri şunlardır:

- *UK2006:* 767 spam sayfa ve 7474 spam olmayan sayfa içeren küçük bir web veri kümesi için eğitilmiş etiket seti.

- *Britney:* Popüler arama motorlarının istatistikleri tarafından bildirilen popüler arama sorgularından sezgisel olarak oluşturulmuş bir eğitim setidir. Bu yöntemin ardından, popüler sorgulara cevap olarak Indri erişim sistemi tarafında getirilen ilk 10 sıralı sayfa, spam olarak kabul edildi. URI'leri "Open Directory Project1"de de yer alan 10000 doküman non-spam olarak sınıflandırılmıştır.
- *GroupX*: Bu eğitim spam varlığının insanlar tarafından değerlendirildiği, Cormack ve arkadaşları tarafından toplanan 756 dokümandan oluşmaktadır.
- *Fusion*: Sonrasında, bu üç filtre tarafından ayrı ayrı elde edilen puanlar, Naive Bayes kombinasyonu kullanılarak bir araya getirilebilir. ClueWeb veri kümesi üzerinde spam tanıma ve doküman erişiminde en etkili olduğu gösterilmiştir.

Yazarlar, eğitim setinin ismi ile eşleştirdikleri bu dört farklı spam filtresini kullanarak, ClueWeb09 veri kümesindeki İngilizce dokümanların her biri için, o dokümanın kullanılan spam filtresine göre spam olma olasılığını gösteren yüzdelik puanları içeren dört farklı spam puan listesi oluşturdu. Yazarlar, bu spam puanları temelinde, belirlenen bir threshold değerinin altındaki dokümanların spam kabul edilerek sonuç listesinden çıkartılması ve sonuç listesinde döndürülen sonuçların bu spam puanlarına bakılarak yeniden sıralanması üzerine iki farklı deney yapıldı. Yaptıkları karşılaştırmalı analizler sonucunda, ClueWeb09 veri kümesinin önemli bir kısmının spam dokümanlardan oluştuğunu ve sonuçlardan spam filtrelemenin ve yeniden sıralamanın, TREC 2009 Web Track'e katılan sistemlerin çoğunda sistem etkinliklerini (*İng.* Retrieval effectiveness) önemli oranda artırdığını raporladı. Yapılan çalışma, büyük ölçekli veri setlerinde spam dokümanların belirlenmesini için graph-based yöntemlere kıyasla daha kolay ve etkili olduğundan, "baseline/ state-of-the-art" niteliği taşımaktadır. Ayrıca yazarlar sonraki araştırmalarda da geliştirilen yöntemin kullanılabilmesi adına "Waterloo spam puanları[4] (*İng.* Waterloo spam scores)" olarak da adlandırılan "percentile-score clueweb-docid" çiftlerinden oluşan bu spam puan listelerini, herhangi bir kısıtlama olmaksızın web'den erişime açmışlardır. Nitekim, yazarlarca üretilen bu spam puan listeleri, TREC 2010 ve 2011 Ad-hoc ve Diversity alt görevlerinde en üst sıralarda gelen sistemler tarafından spam doküman filtreleme için

[4]https://plg.uwaterloo.ca/~gvcormac/clueweb09spam/

eşik değeri olarak (Bendersky vd., 2010; Kamps vd., 2010; Smucker, 2011), dönen sonuçların post-processinde eşik değeri olarak (Hauff ve Hiemstra, 2009), ve doküman sıralamasında bir özellik olarak kullanıldı (Billerbeck vd., 2011; McCreadie vd., 2011).

Spam ve düşük kaliteli dokümanların BE sonuçlarına olumsuz etkileri olduğunun anlaşılmasıyla, TREC 2010 Web Track serisine "Spam Task" alt görevi eklendi. Bu görevin amacı, ClueWeb09 veri kümesindeki İngilizce dokümanların spam olma olasılığına göre derecelendirilmesiydi. Bu alt görev kapsamında çok geniş bir Spam tanımı kullanıldı: "apaçık aldatıcı veya zararlı sayfalarla birlikte esasen junk olan sayfalardan oluşanlar". Cormack ve arkadaşları tarafından en iyi sonuçların elde edildiği "Fusion" spam puanları "baseline" olarak kabul edildi. Göreve katılan üç gruptan hiçbiri bu baseline değerlerini geçemedi (Clarke vd., 2010).

Spam olmaksızın indeksleme, web koleksiyonlarındaki spam sorununa bir başka bakış açısıdır. Yukarıda anlatılan çalışmalarda, sonuçlar üzerinde post-retrieval işlemi ile spam filtreleme yapılırken; bu yaklaşımda geleneksel BE sistem mimarisindeki indeksleme aşamasına spam filtreleme adımı eklenir. Böylece indeksleme zamanında spam dokümanlar ayıklanarak spamden arınmış bir indeks oluşturulur, dolayısıyla sonuç listelerinde spam yer almaz. (Zuccon vd., 2011), spamden arındırılmış indekslemenin faydalarını ve dezavantajlarını araştırmak için TREC ClueWeb 2009-2010 Web Track veri kümesindeki birkaç geleneksel erişim modeli üzerinde deneyler gerçekleştirmiştir. Spam dokümanları indeksleme öncesinde tespit etmek için (Cormack vd., 2011) spam skorlarını kullanılarak, belirli bir eşik değerindeki web sayfaları spam olarak kabul edilerek indekslenmemiştir. Dokümanın spam skoruna göre gerçekleştirdikleri indeks budama tekniğinin hem indeksleme süresini hem de indeks boyutunu azalttığı görülmüştür. Bununla birlikte, sonuçlar üzerinde spam filtreleme yaklaşımını ile karşılaştırıldığında, erişim performansına zarar vermediği görülmüştür. (Zuccon vd., 2011)'ların ClueWeb09 veri kümesinin İngilizce kısmı için yaptıkları araştırmayı, (Crane ve Trotman, 2012) tüm ClueWeb09 veri kümesi için gerçekleştirerek, indeks boyutunda tutarlı bir düşüşün, daha hızlı indeksleme ve aramayı beraberinde getirdiği şeklinde benzer bulgular elde etmiştir.

# 4. DENEYSEL YÖNTEM

ClueWeb09[5] ve ClueWeb12[6] veri kümeleri, WebBE teknolojileri alanındaki araştırmaları desteklemek ve geliştirilen sistemlerin büyük ölçekli değerlendirilmesine imkân sağlamak için web'in temsili niteliğinde oluşturulmuş iki büyük veri kümesidir. Araştırmacılara üzerinde çalışılabilecek daha gerçekçi web veri kümeleri sunma hedefiyle oluşturulan bu veri kümeleri, BE sistemlerinin test koleksiyonu bazlı değerlendirilmesi yaklaşımını esas alan TREC[7] konferanslarının çeşitli alt görevlerinde kullanılan doküman koleksiyonları olmuştur.

ClueWeb09 veri kümesi, Ocak ve Şubat 2009 arasında taranan, on farklı dilde yaklaşık 1 milyar web sayfasından oluşur. Yarım milyarı aşkın sayfa içeren ClueWeb09 koleksiyonunun İngilizce kısmı, 2009'dan 2012'ye kadar dört defa TREC konferansının "Web Track[8]" alt görevinde kullanılmıştır. ClueWeb09 veri kümesi, alandaki araştırmacılara, web'i olduğu gibi temsil edebilecek bir alt küme yaratılması düşüncesiyle; herhangi bir kısıtlama yapılmadan web'de karşılaşılabilecek her türlü içeriği kapsayacak şekilde oluşturulduğundan spam ve yetişkin içeriği içerir.

ClueWeb12 veri kümesi ise 10 Şubat ve 10 Mayıs 2012 arasında taranan 733.019.372 İngilizce web sayfasından oluşur. TREC konferansının 2013 ve 2014 "Web Track" alt görevinde, 2015'den 2017'ye kadar da "Tasks Track[9]" alt görevinde kullanılmıştır. Ocak 2013'de dağıtımı başlayan ClueWeb12 veri kümesi, ClueWeb09 veri kümesinin devamı niteliğindedir. Bu veri kümesi oluşturulurken web tarayıcıları önceden belirlenen başlangıç sayfalarından başlayarak web'i taramış, kara liste *(İng.* black list) içerisinde yer alan içerikler ile metin gibi görünmeyen (ses, video, sıkıştırılmış dosyalar) URL'ler kapsam dışı bırakılmış ve son-işleme sırasında yetişkin içerik filtreleme uygulanmıştır. Nitekim ClueWeb12 veri kümesi, ClueWeb09'a nazaran daha kontrollü oluşturulduğundan içerdiği spam içeriğin daha az olduğu varsayılır. Araştırmacılara şu an için sunulan en büyük web veri kümelerinden birisidir.

2009 ve 2016 yılları arasında, çeşitli TREC alt görevleri kapsamında, ClueWeb09 ve ClueWeb12 veri kümeleri üzerinde yapılmış olan 9 değerlendirmeye ait istatistiksel

[5]http://boston.lti.cs.cmu.edu/classes/11-742/S10-TREC/TREC-Nov19-09.pdf
[6]http://opensearchlab.otago.ac.nz/SIGIR12-OSIR-callan.pdf
[7]http://trec.nist.gov
[8]https://trec.nist.gov/data/webmain.html
[9]https://trec.nist.gov/data/tasks.html

bilgiler Tablo 4.1'de sunulmuştur. Tablodan görülebildiği gibi, her yıl yaklaşık olarak 50 yeni sorgu ve bu sorgulara ilişkin sorgu uygunluk yargıları yayınlanmıştır. *qrels* dosyası olarak adlandırılan, sorgu uygunluk yargıları, havuzlama yöntemiyle[10] tespit edilen *qid/docid* çiftleri için emekli kütüphaneciler tarafından belirlenen uygunluk yargısı (*İng.* relevance judgement) bilgilerini içerir.

**Tablo 4.1.** *Sorgu setlerinin istatistikleri*

| Veri Kümesi | Track | Yıl | Sorgu Sayısı | Sorgu Başına Ortalama İlgili Doküman Sayısı | Sorgu Başına Ortalama Spam Doküman Sayısı | Uygunluk Yargı Seviyesi Sayısı | Uygunluk Yargı Seviyeleri |
|---|---|---|---|---|---|---|---|
| ClueWeb09 | Millon Query | 2009 | 561 | 15,2 | - | 3 | 0, **1, 2** |
| | Web Track | 2009 | 50 | 139,7 | - | 3 | 0, **1, 2** |
| | | 2010 | 48 | 108,9 | 29,8 | 5 | -2, 0, **1, 2, 3** |
| | | 2011 | 50 | 63,0 | 21,7 | 5 | -2, 0, **1, 2, 3** |
| | | 2012 | 50 | 70,3 | 17,9 | 6 | -2, 0, **1, 2, 3, 4** |
| ClueWeb12 | Web Track | 2013 | 50 | 82,7 | 6,3 | 6 | -2, 0, **1, 2, 3, 4** |
| | | 2014 | 50 | 113,1 | 15,9 | 6 | -2, 0, **1, 2, 3, 4** |
| | Tasks Track | 2015 | 35 | 28,7 | 35,4 | 4 | -2, 0, **1, 2** |
| | | 2016 | 50 | 50,3 | 3,5 | 4 | -2, 0, **1, 2** |

[10]Havuzlama yönteminde, değerlendiriciler sadece TREC katılımcıları tarafından her sorgu için getirilen dokümanların oluşturduğu birleşim kümesindeki üst-k (genellikle 100) dokümanı değerlendirir ve o belgelere ait uygunluk yargı değerlerini manuel olarak belirler.

Önceden görülmemiş sorgular için azalan ilgi düzeyine göre ilgili dokümanların sıralanmasının hedeflendiği ve bu hedef doğrultusunda sistemlerin performanslarının incelendiği TREC ad-hoc alt görevi kapsamında, ClueWeb09 ve ClueWeb12 veri kümeleri üzerinde yapılan son TREC değerlendirmelerinde, *qid/docid* çiftlerine uygunluk yargısı atanırken, Tablo 4.2'de gösterilen altı-puanlı ölçek (*İng.* **six-point scale)** kullanılmıştır (Collins-Thompson vd., 2015).

**Tablo 4.2.** *Altı-puanlı ölçek*

| **Uygunluk Yargısı** | **-2** | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|---|
| **Kısaltma** | ***Junk*** | *Non* | *Rel* | *HRel* | *Key* | *Nav* |
| | **Non-Relevant** | | **Relevant** | | | |

Gerçek kişiler tarafından, her *qid/docid* çifti için açıklama alanı (*İng.* description field) temelinde belirlenen bu uygunluk yargı seviyeleri ve anlamları aşağıdaki gibidir:

- *Uygunluk Yargısı=4*: Sorgu tarafından doğrudan erişilen ana sayfayı gösterir. Kullanıcı bu özel sayfayı arıyor olabilir (*Nav*).
- *Uygunluk Yargısı=3*: Sayfa sorguya adanmıştır, dolayısıyla bir web arama motorunda en iyi sonuç olmaya layıktır (*Key*).
- *Uygunluk Yargısı=2*: Sayfanın içeriği konuyla ilgili önemli bilgiler içerir (*HRel*).
- *Uygunluk Yargısı=1*: Sayfanın içeriği konuyla ilgili minimum düzeyde bilgi içerir. Sadece yararlı bir sayfayı işaret eden umut verici bir çapa metni içermez, ilgili bilgi bu sayfa olmalıdır (*Rel*).
- *Uygunluk Yargısı=0*: Sayfanın içeriği konu hakkında yararlı bilgi içermez, fakat aynı sorgunun diğer yorumları da dahil olmak üzere diğer konularda faydalı bilgi sağlayabilir (*Non*).
- *Uygunluk Yargısı=-2*: Bu sayfa herhangi bir makul amaç için uygun görünmemektedir, spam ya da önemsiz olabilir (*Junk*).

TREC 2012 genel bakış makalesinde de belirtildiği gibi -2 uygunluk yargısı, sadece spam olan *qid/docid* çiftleri için kullanılmış ve değerlendiriciler tarafından spam olduğu düşünülen dokümanlara verilmiştir (Collins-Thompson vd., 2015).

Tablo 4.2'de gösterilen altı-puanlı ölçek üzerinden *gdeval.pl*[11], *statAP.pl*, *trec_eva*l gibi değerlendirme scriptleri ile *estAP@10*, *nDCG@20*, *MAP* gibi BE etkinlik ölçütleri hesaplanıp raporlanabilir. Böylelikle BE sistemleri bu sorgular üzerinden karşılaştırmalı değerlendirilmiş olur.

Spam olgusu WebBE alanına özel bir durumdur. Ticari arama motorlarını bilinçli bir şekilde aldatmak üzere tasarlanmış web sayfaları, spam sayfalardır ve bu sayfalar web'in doğasında vardır. ClueWeb09 veri kümesi, araştırmacılara web'i olduğu gibi sunma amacıyla toplandığından, bünyesinde spam sayfalar barındırır. BE sistem etkinliği açısından bu tür spam dokümanların ya indeksleme sırasında tespit edilip çıkarılarak indekse dahil edilmemesi ya da sorguya karşılık döndürülecek sonuç listesinden çıkartılması ya da sonuç listesinin spam dokümanlar altlarda gelecek şekilde yeniden sıralanması gerekir.

Bu amaca yönelik, Cormack ve arkadaşları, ClueWeb09 İngilizce veri kümesi üzerinde ilk sistematik spam incelemesini gerçekleştirerek, spam filtrelemenin BE etkinliği üzerindeki etkilerinin ilk sayısal sonuçlarını sunmuştur (Cormack vd., 2011). Yazarlar, ClueWeb09 veri kümesinin önemli bir kısmının spam dokümanlardan oluştuğunu, sonuç listesi üzerinde spam filtrelemenin ya da sonuç listesini spam dokümanlar alt sıralarda gelecek şekilde yeniden sıralamanın, TREC 2009 Web Track'e katılan sistemlerin çoğu için *MAP* ve *estP@k* ölçüm metrikleri ile yaptıkları sistem etkinlik *(İng.* retrieval effectiveness) değerlendirmelerini önemli oranda artırdığını raporlamıştır.

Yaptıkları çalışma kapsamında, Cormack ve arkadaşları (Cormack vd., 2011), ClueWeb09 veri kümesindeki İngilizce dokümanlar için "Waterloo spam puanları[12] *(İng.* Waterloo spam scores)" olarak da bilinen, dört grup spam puan listesi ya da spam sıralaması (*UK2006*, *Britney*, *GroupX*, *Fusion*) oluşturmuştur.

Sonraki araştırmalarda da kullanılabilmesi amacıyla herhangi bir kısıtlama olmaksızın web'den erişilip indirilebilen bu dört spam puan listesinin her biri, ClueWeb09 veri kümesindeki her doküman için "yüzdelik puan (*İng.* percentile-score)

[11]https://github.com/trec-web/trec-web-2013/blob/master/src/eval/gdeval.pl
[12]https://plg.uwaterloo.ca/~gvcormac/clueweb09spam/

ve ClueWeb09 doküman numarası (*İng.* clueweb-docid)" çiftlerinden oluşan 503.903.810 satır içerir. Yüzdelik puan, veri kümesindeki "spammier" olan dokümanların yüzdesini gösterir. 0 ile 99 arasında toplam 100 eşit parçaya bölünür. Puanlama algoritmaları kullanılarak sıralanan veri kümesinin her bir yüzdelik dilimi yüzdelik puan olarak anılır. Bu puanlama sisteminde spam olma ihtimali en yüksek belgeler en düşük yüzdelik dilime atanırken; spam olma ihtimali en düşük belgeler en yüksek yüzdelik dilime atanır.

Spam sıralamaları, ClueWeb09 veri kümesindeki dokümanları spam ve non-spam olarak sınıflandırma amacıyla da kullanılabilir. Bu durumda, bir eşik değeri (*İng.* threshold) belirlenir ve yüzdelik puanı bu eşik değerinden düşük olan dokümanlar spam, yüksek olanlar ise non-spam olarak tasnif edilir. Cormack ve arkadaşları (Cormack vd., 2011), amacın sadece dokümanların spam ya da non-spam olarak işaretlemesi olduğu durumda eşik değeri olarak 70 kullanılmasını, yani yüzdelik puanı 70'den küçük olanların spam, geri kalanının non-spam olarak işaretlenebileceğini ve bu dört spam sıralamasından en iyisinin istendiği durumda ise Fusion spam sıralamasının seçilmesi gerektiğini irdelemiştir.

Yukarıda belirtildiği gibi "Yüzdelik Puan<70" şeklinde belirlenen bir eşik değeri aslında ikili sınıflandırıcıdır (*İng.* binary classifier). Bir başka deyişle, belirlenen bir eşik değeri için spam sıralamaları (spam puan listelerindeki yüzdelik puanlar) ikili sınıflandırma yapar.

```
Girdi: ClueWeb docId          // docid = "clueweb12-0209wb-62-29857"
Girdi: threshold              // EşikDeğeri = 70

percentile = Fusion(docId);   // percentile = 24
if(percentile < threshold)
    return Spam;              // 24 < 70 --> Spam
else return Non-Spam;
```

Cormack ve arkadaşları tarafından literatüre kazandırılan bu çalışmada, spam sıralamalarının tekil başarımına değil, bu sıralamaların tam teşekküllü *(İng.* fully fledged) BE sistemlerine bir bileşen olarak entegre edilmesiyle, yani BE süreci içerisine dahil edilmesiyle, elde edilen sonuçlar ile sistemlere bu bileşenin entegre edilmediği durumda, yani spam sıralamalarının BE sürecinde hiç olmadığı yalın süreç ile, elde edilen sonuçların kıyaslamasına bakılır. Spam sıralamalarının BE sistem etkinliğine

tesirinin ölçüldüğü bu tür değerlendirmeye "extrinsic değerlendirme (*İng.* extrinsic evaluation)" adı verilir (Belz ve Gatt, 2008).

Öte yandan, spam sıralamalarını tekil olarak değerlendirmek de mümkündür. Öyle ki, bir ikili sınıflandırıcı olarak tek başına ele alındıklarında, spam sıralamalarının spam ve non-spam ayrımını ne derece yapabildiği incelenebilir. Fakat bu noktada sınıflandırma başarımının değerlendirilebilmesi (ölçülebilmesi) için "gerçek doğrulara *(İng.* ground truth)" ihtiyaç vardır. Doğrudan doğruya manuel olarak spam ve non-spam diye işaretlenmiş dokümanlar mevcut olmasa da aşağıdaki varsayımlar ve *qrels* yardımıyla bu taklit edilebilir.

- Uygunluk yargısı -2 olan *qrels*'lerdeki dokümanlar spam'dir.
- İlgili dokümanlar (yani uygunluk yargısı 1, 2, 3 ya da 4 olan *qrels*'lerdeki dokümanlar), herhangi bir sorgu için ilgili olarak işaretlendiğinden dolayı spam olamaz, non-spam'dir.

Böylelikle, spam sıralamalarının, spam ve non-spam sınıflandırmasının (*İng.* binary classification: spam versus non-spam) "intrinsic" değerlendirmesini yapmak mümkün olur.

Tablo 4.3, ClueWeb09 veri kümesinde yetkin değerlendiriciler tarafından uygunluk yargısı 1, 2, 3 ya da 4 olarak verilen, bir başka deyişle, ilgili (*İng.* relevant) olarak belirlenmiş doküman sayıları ile uygunluk yargısı -2 olarak verilen, yani spam olarak belirlenmiş doküman sayılarını yıl bazında gösterir.

**Tablo 4.3.** *ClueWeb09 veri kümesinde ilgili ve spam işaretlenmiş doküman sayıları*

| | **Sınıf Etiketi:** *spam*<br>**Uygunluk Yargısı:** *-2* | **Sınıf Etiketi:** *non-spam*<br>**Uygunluk Yargısı:** *1, 2, 3, 4* |
|---|---|---|
| 2010 | 1431 | 5233 |
| 2011 | 1019 | 3157 |
| 2012 | 858 | 3523 |
| **Total** | **3308** | **11913** |

Tablo 4.3'de sunulan uygunluk yargıları elle işaretlendiğinden, bir ikili sınıflandırıcı için gerçek doğrular olarak düşünülebilir. Öyle ki, herhangi bir sorgu için gerçek kişiler tarafından elle "ilgili" olarak işaretlenen (yani uygunluk yargısı > 0 olanlar) *qrels*'lerdeki dokümanlar spam olamayacağından, bu dokümanlar non-spam etiketiyle işaretlenebilir. Diğer taraftan, uygunluk yargısı -2 olarak işaretlenen

*qrels*'lerdeki dokümanlar ise sadece spam olabileceğinden ikili bir spam sınıflandırıcısı için "true pozitif" ya da "spam" etiketiyle işaretlenebilir.

Sonuç olarak, önceki çalışmalarda yapılan "extrinsic" değerlendirmeye karşılık, bu çalışmada, Waterloo spam puanlarının "intrinsic ve retrospektif" değerlendirmesi hedeflenmiştir.

Retrospektif değerlendirme (*İng.* retrospective evaluation) ile kastedilen, 2009 yılında Cormak ve arkadaşları tarafından yayınlanan spam sıralamalarının bunu takip eden 4 yıl sürecince (her sene yaklaşık 50 yeni sorgu eklenerek) 2012'ye kadar birikerek toplanan *qrels*'ler üzerinden değerlendirilecek olmasıdır.

Daha açık ifade etmek gerekirse, Waterloo spam puanları 2009 yılında yayınlanmıştır. Diğer taraftan, 2009 yılından 2012 yılına kadar ClueWeb09 veri kümesi üzerine yapılmış TREC yarışmalarında elde edilen tüm sorgu uygunluk yargıları, bir başka ifadeyle BE *qrels*'leri ya da *qid/docid* çiftleri ve bu çiftlere karşılık gerçek değerlendiriciler tarafından elle verilmiş olan uygunluk yargıları *(İng.* relevancy), gerçek doğrular *(İng.* Ground truth") olarak ele alınabilir. Dolayısıyla, 2009 sonrası işaretlenen gerçek doğrular esas alınarak 2009'da yayınlanan Waterloo spam puanları değerlendirilip inceleneceğinden, bu çalışma retrospektif değerlendirme (geçmişe dönük değerlendirme - olaydan sonra durumun incelenmesi) olarak da anılabilir (Greiff, 1998).

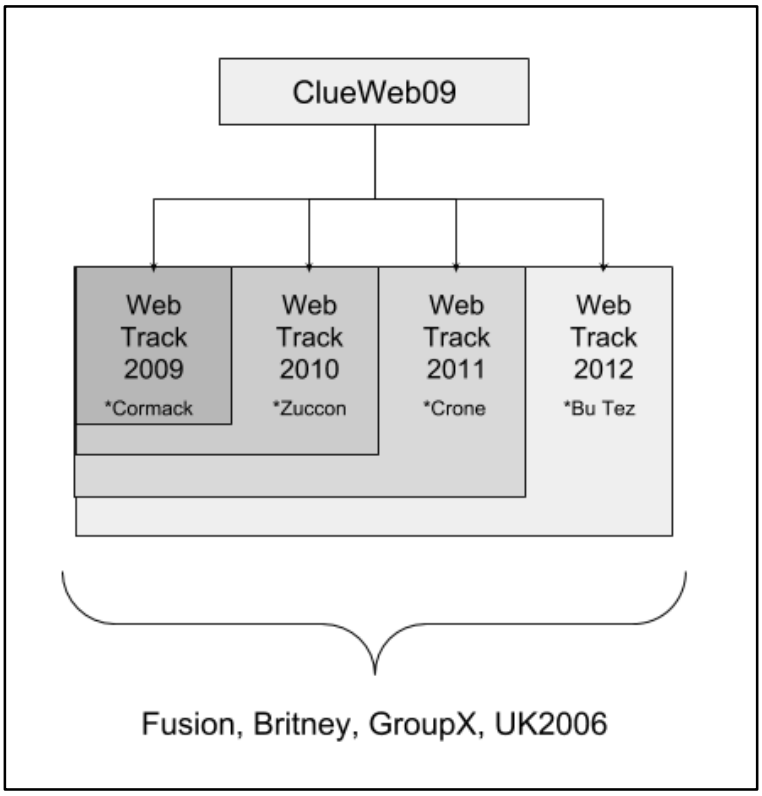
Böyle bir değerlendirme yazarların bildiği kadarıyla *(İng.* to the best of our knowledge) literatürde daha önce yapılmamıştır.

Geçmişteki üç çalışmada da "extrinsic" değerlendirme yapılmış olup, indeksleme sırasında ya da sonuç listesi üzerinde spam filtrelemenin BE sistem etkinliğine tesiri ölçülmüştür (Cormack vd., 2011; Crane ve Trotman, 2012; Zuccon vd., 2011). Ayrıca, Şekil 4.1'de gösterildiği gibi, Cormack ve arkadaşları tarafından yapılan çalışmada sadece 2009 (Cormack vd., 2011), Zuccon ve arkadaşları tarafından yapılan çalışmada 2009-2010 (Zuccon vd., 2011), Crane ve Trotman tarafından yapılan çalışmada ise 2009-2011 (Crane ve Trotman, 2012) Web Track'ler kullanış ve basım yılları sırasıyla 2011, 2011, ve 2012'dir.

Tez kapsamında yazarlar tarafından yapılan çalışmada ise 2009-2012 yılları arasında toplanan *qrels*'ler bazında Waterloo spam puanlarının "intrinsic ve retrospective" değerlendirmesi amaçlanmıştır. Cormack ve arkadaşlarının yapmış olduğu (Cormack vd., 2011) "extrinsic" değerlendirme 2009 yılında yapılmış olup, günümüze değin her sene 50 yeni sorgu daha kümeye eklenmiştir. 2009-2012 yılları

arasında biriken sorgu uygunluk yargıları, ya da *qrels*, gerçek doğrular olarak kabul edilebilir, öyle ki değerlendiriciler tarafından uygunluk yargısı -2 olarak belirlenmiş dokümanlar spam, uygunluk yargısı 1, 2, 3 ya da 4 (uygunluk yargısı >0) olarak belirlenmiş dokümanlar non-spam olarak ele alınabilir. Böylelikle, spam sıralamaları ikili sınıflandırıcı olarak ele alınıp, bu sınıflandırıcıların spam ve non-spam sınıflandırma başarımlarının "intrinsic" değerlendirmesi yapılabilecektir. Bu değerlendirmeyle elde edilen sonuçların geçmişteki "extrinsic" değerlendirme sonuçlarıyla benzerliği, dört spam sıralamasından ortalamada daha iyi olanın tespiti ve eşik değeri belirleme stratejileri incelenebilecektir.

Bunların yanı sıra, spam konusu ve etkileri ClueWeb09 veri kümesi için çok araştırılmış olsa da ClueWeb12 veri kümesi için henüz hiç araştırılmamıştır (Xiong vd., 2016).




**Şekil 4.1.** *Geçmiş çalışmalarda ve bu çalışmada kullanılan veri kümesi*

Makaleleri basılmamış olsa da Cormack ve arkadaşları, ClueWeb12 veri kümesi için sadece Fusion spam sıralamasını içeren bir spam puan listesi de yayınlamıştır. Cormack ve arkadaşları, ClueWeb12 spam sıralaması üzerine kapsamlı bir test

gerçekleştirmediklerini, fakat sonuçların makul göründüğünü belirtmiştir. ClueWeb09 veri kümesi için yayınlanan spam puan listelerine benzer şekilde, ClueWeb12 veri kümesi için yayınlanan Fusion spam sıralaması da veri kümesindeki her doküman için "yüzdelik puan (*İng.* percentile-score) ve ClueWeb12 doküman numarası (*İng.* clueweb-docid)" çiftlerinden oluşmakta ve toplam 733.019.372 satır içermektedir. Araştırmalarda kullanılabilmesi amacıyla ClueWeb12 Fusion sıralaması[13] da herhangi bir kısıtlama olmaksızın web erişimine sunulmuştur.

Tablo 4.4'de ClueWeb12 veri kümesi için 2013-2016 yılları arasındaki 4 yıllık süre içerisinde biriken *qrels* bilgileri, yani gerçek değerlendiriciler tarafından ilgili (uygunluk yargısı 1, 2, 3, 4) ve spam (uygunluk yargısı -2) olarak belirlenmiş doküman sayıları yıl bazında gösterilmiştir.

**Tablo 4.4.** *ClueWeb12 veri kümesinde ilgili ve spam işaretlenmiş doküman sayıları*

| | **Sınıf Etiketi:** *spam*<br>**Uygunluk Yargısı:** *-2* | **Sınıf Etiketi:** *non-spam*<br>**Uygunluk Yargısı:** *1, 2, 3, 4* |
|---|---|---|
| 2013 | 234 | 4150 |
| 2014 | 556 | 5665 |
| 2015 | 1240 | 1003 |
| 2016 | 97 | 2518 |
| **Toplam** | **2127** | **13336** |

Tablo 4.3 ve Tablo 4.4 karşılaştırıldığında, ClueWeb09 veri kümesi için *qid/docid* kapsamında değerlendirilen dokümanların %22 kadarının spam (-2) olarak, ClueWeb12 veri kümesi için ise %14 kadarının spam (-2) olarak işaretlendiği görülür. Bu sonuçlar, ClueWeb09 veri kümesinin, ClueWeb12 veri kümesine nazaran, daha fazla spam içerdiğini göstermektedir.

Tez kapsamında ClueWeb12 Fusion spam sıralamalarının da "intrinsic ve retrospektif" değerlendirmesi amaçlanmıştır.

[13]https://www.mansci.uwaterloo.ca/~msmucker/cw12spam/

# 5. DENEYSEL SONUÇLAR VE ANALİZ

Bu bölümde öncelikle deneylerde kullanılan evrensel sınıflandırma değerlendirme ölçütleri olan doğruluk (*İng.* accuracy), F1-ölçütü (*İng.* F1 measure) ve Receiver Operating Characteristic (ROC) eğrileri anlatılmış, ardından ClueWeb09 ve ClueWeb12 veri kümeleri üzerinde yapılan deneyler ve elde edilen sonuçlar alt başlıklar altında ayrı ayrı sunulmuştur.

Deneylerde ilk olarak hem spam (-2 ilgi ölçeği) hem de non-spam (1, 2, 3 ya da 4 ilgi ölçeği) dokümanların percentile puanlarının üzerine dağılımları grafiksel olarak incelenmiştir.

Tezin deneysel yöntem bölümünde ikili sınıflandırıcıya dönüşümü anlatılan spam sıralamalarının, evrensel sınıflandırma değerlendirme ölçütlerine göre başarımı ölçülmüş ve birbirleriyle karşılaştırılmıştır.

## 5.1. Değerlendirme Metrikleri

İkili sınıflandırıcıya dönüştürülen spam sıralamalarının, spam ve non-spam ayrımını ne derece yapabildiğinin, yani sınıflandırma başarımlarının değerlendirilebilmesi için bir dizi evrensel sınıflandırma değerlendirme metriği mevcuttur. Bu metrikler, gerçek değerler ve tahmin edilen değerlerin yanlış ve doğru sınıflandırma sayılarına dayanmaktadır. Gerçek değerler ve tahmin edilen değerlerin doğru ve hatalı sınıflandırılmasını gösteren tabloya "Hata Matrisi (*İng.* Confusion Matrix)" denilmektedir.

Tablo 5.1'de spam dokümanlar ile non-spam dokümanları tasnif eden bir sınıflandırıcı için hata matrisi sunulmuştur. Burada "pozitif" sınıflandırma dokümanın spam olarak etiketlenmesi anlamına gelirken; "negatif" sınıflandırma dokümanın non-spam olarak sınıflandırması anlamını taşır. Bu doğrultuda, True Positive (TP), gerçekte spam olan ve sınıflandırıcının spam olarak tahmin ettiği doküman sayısı, False Positive (FP) ise gerçekte non-spam olan fakat sınıflandırıcının spam olarak tahmin ettiği doküman sayısıdır. Benzer şekilde, True Negative (TN), gerçekte non-spam olan ve sınıflandırıcının non-spam olarak tahmin ettiği doküman sayısı, False Negative (FN) ise gerçekte spam olan fakat sınıflandırıcının non-spam olarak tahmin ettiği doküman sayısıdır.

**Tablo 5.1.** *İkili sınıflandırma hata matrisi*

| | | Actual Class | | | |
|---|---|---|---|---|---|
| | | Spam *(Positive)* | Non-spam *(Positive)* | | |
| Predicted Class | Spam *(Positive)* | *True Positive (TP)* | *False Positive (FP)* | *Precision*, *Positive Predicted Value* $(PPV) = \frac{TP}{TP + FP}$ | |
| | Non-spam *(Positive)* | *False Negative (FN)* | *True Negative (TN)* | *Negative Predicted Value* $(NPV) = \frac{TN}{TN + FN}$ | |
| | | *Recall, Sensitivity, True Positive Rate* $(TPR) = \frac{TP}{TP + FN}$ | *Fallout, False Positive Rate* $(FPR) = \frac{FP}{TN + FP}$ | *Accuracy* $= \frac{TP + TN}{TP + FP + FN + TN}$ | $F1_{score} = \frac{2 \times Precision \times Recall}{Precision + Recall}$ |
| | | *False Negative Rate* $(FNR) = \frac{FN}{TP + FN}$ | *Specificity, True Negative Rate* $(TNR) = \frac{TN}{TN + FP}$ | | |

Tablo 5.1'deki hata matrisinde gösterilen TP, FP, TN ve FN değerleri yardımıyla ikili sınıflandırıcıya dönüştürülen spam sıralamaları için hesaplanabilecek performans değerlendirme ölçütleri ve tanımları şunlardır:

*Precision / Positive Predictive Value (PPV)*: Doğru sınıflandırılan spam dokümanların spam olarak tahmin edilen toplam dokümanlara oranına Kesinlik (*İng.* Precision) veya Pozitif Öngörü Değeri (*İng.* Positive Predictive Value (PPV)) denir.

*Negative Predictive Value (NPV)*: Doğru sınıflandırılan non-spam dokümanların non-spam olarak tahmin edilen toplam dokümanlara oranına ise Negatif Öngörü Değeri (*İng*. Negative Predictive Value (NPV)) denir.

*Recall / True Pozitif Rate (TPR) / Sensitivity*: Doğru sınıflandırılan spam dokümanların var olan toplam spam doküman sayısına oranıdır. Sınıflandırıcının spam dokümanları doğru tahmin etme oranı olan Recall metriği, Doğru Pozitif Oranı (*İng.* True Pozitif Rate (TPR)) veya Duyarlılık (*İng.* Sensitivity) olarak da adlandırılır.

*F1-score*: Precision ve Recall başarım ölçütlerinin harmonik ortalamasıdır. F1-ölçütü (*İng.* F1 measure) bu iki ölçütün birlikte ele alınmasını sağla

*Accuracy*: Sınıflandırıcının toplam doğru tahmin sayısının (TP ve TN), tüm örneklem sayısına oranı, Doğruluk (İng. Accuracy) olarak adlandırılır. Accuracy metriği sınıflandırmanın doğruluğudur. 1- Accuracy ise sınıflandırmanın Hata Oranını (*İng.* Error Rate (ERR)) verir.

*Specificity / True Negative Rate (TNR)*: Sınıflandırıcının non-spam dokümanları (negatif sınıf etiketlerini) tahmin etmedeki etkililiği Belirleyicilik (*İng.* Specificity) veya Doğru Negatif Oranı (*İng.* True Negative Rate (TNR)) olarak adlandırılır. Doğru sınıflandırılan non-spam dokümanların toplam non-spam doküman sayısına oranıdır.

False Negative Rate (FNR): Gerçekte spam olan ancak non-spam olarak sınıflandırılmış dokümanların, tüm spam etiketli dokümanlara oranı Yanlış Negatif Oranı (*İng*. False Negative Rate (FNR)) olarak adlandırılır. 1-TPR değerine eşittir.

*Fallout / False Positive Rate (FPR) / 1-Specificity*: Gerçekte non-spam olan ancak spam olarak sınıflandırılmış dokümanların, tüm non-spam etiketli dokümanlara oranı Yanlış Pozitif Oranı (*İng. False Positive Rate (FPR) ya da Fallout)* olarak adlandırılır. 1-Specificity değerine eşittir.

*ROC Eğrisi*: Alıcı İşlem Karakteristikleri (*İng.* Receiver Operating Characteristic (ROC)) eğrileri, *y* ekseninde TPR, *x* ekseninde FPR olacak şekilde sınıflandırıcının performans grafiğini çizmede kullanılmaktadır. ROC eğrisinin altında kalan alan (İng.

Area Under Curve (AUC), kullanılan modelin doğruluğunu vermektedir. Eğri ne kadar yukarıda ve sol üst köşeye yakın ise, model o kadar başarılı anlamına gelir.

## 5.2. ClueWeb09 Veri Kümesi Sonuçları

ClueWeb09 veri kümesi için Tablo 4.3'de sunulan spam ve non-spam olarak işaretlenmiş dokümanların dört farklı spam sıralamasın göre dağılımları Şekil 5.1'de verişmiştir. Söz konusu grafiklerde, *x* ekseni yüzdelik puan, *y* ekseni ise sorgu doküman çifti sayısını göstermektedir. Kırmızı renkli eğri spam dokümanları, mavi renkli eğri ise non-spam (ilgili) dokümanları göstermektedir.

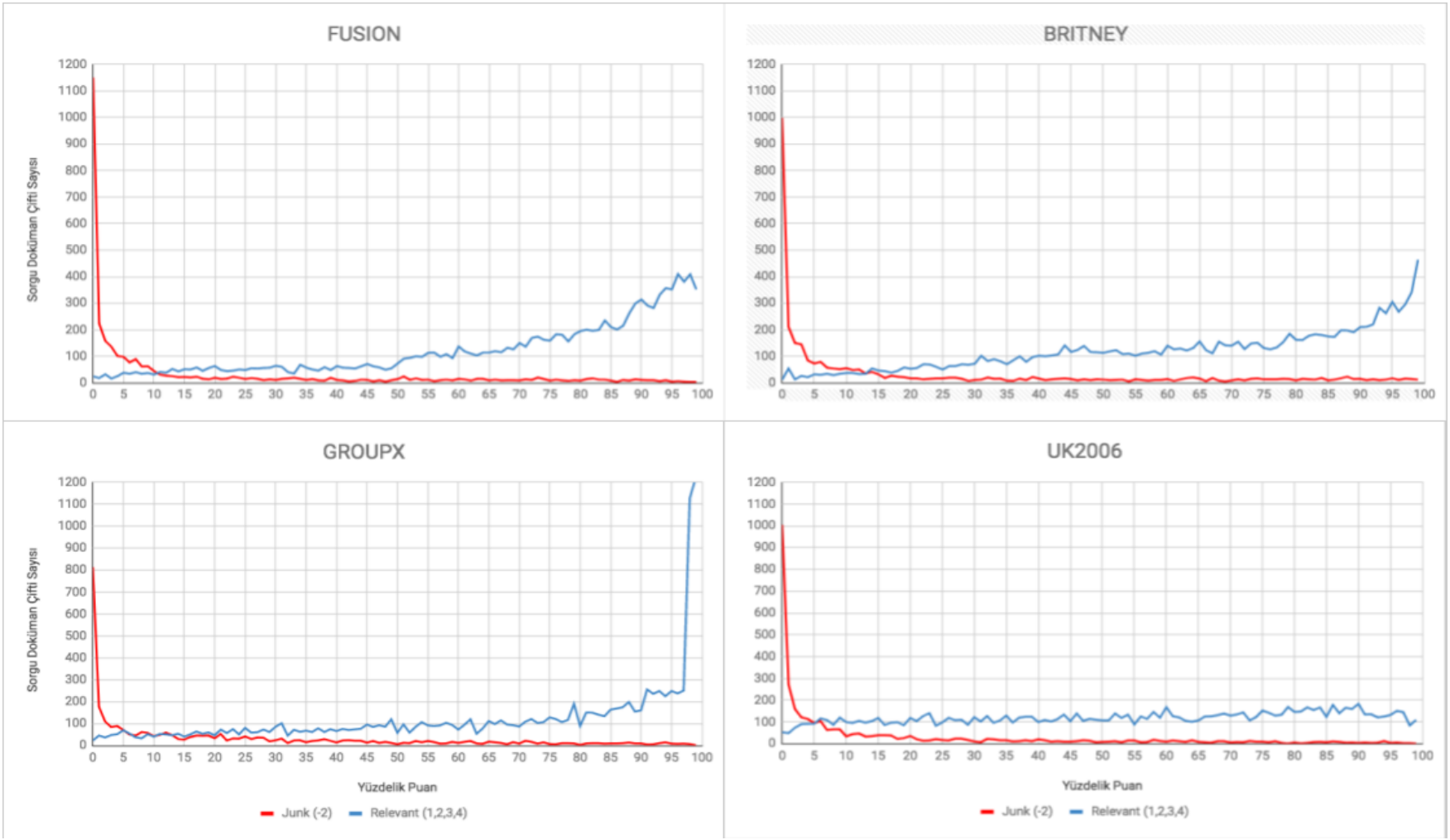
Dört farklı spam sıralaması için ayrı ayrı çizilen şekillerin hepsinde benzer bir eğilim gözlenmektedir. X ekseninin, sağ tarafında (yüzdelik puanın büyük değerleri için) ilgili, sol tarafında ise (yüzdelik puanın küçük değerleri için) spam dokümanların yoğunlaştığı gözlenmektedir. Öte yandan, mavi ve kırmızı eğrilerin bir noktada kesiştiği de dikkat çekicidir. Bu kesişim noktası, *kırılma noktası* ya da bir başka değişle *ideal eşik değeri* olarak nitelendirilebilir. Eşik değeri olarak, bu kesişim noktası kabul edildiğinde aşağıdaki değerlendirmeleri yapmak mümkündür:

Bu kesişim noktasının **sağ** tarafında kalan **kırmızı** eğrinin altındaki alan *False Negative* olur. Benzer şekilde kesişim noktasının **sol** tarafında **mavi** eğrinin altında kalan alan *False Positive* olur. Bu bahsedilen bölgeler puanlama yöntemlerinin hata yaptığı kısımlar olup, bu alanların küçük olması makbuldür.

Aynı doğrultuda, kesişim noktasının **sol** tarafında kalan **kırmızı** eğrinin altındaki alan *True Positive* olur. Benzer şekilde kesişim noktasının **sağ** tarafında **mavi** eğrinin altında kalan alan *True Negative* olur. Bu bahsedilen bölgeler puanlama yöntemlerinin doğru yaptığı kısımlar olup, bu alanların büyük olması makbuldür.

Bu anlatım, Tablo 5.1'de sunulan dört gözlü hata matrisi ile uyumludur. Mavi ve kırmızı renkle gösterilmiş iki eğri ve bu eğrilerin kesişim noktasının sağ ve sol kısımları, *x*-*y* düzlemini dört parçaya bölmektedir.

Dört farklı spam sıralamasının Şekil 5.1'deki farklılıkları incelendiğinde, ilgili dokümanları en belirgin biçimde ayırt edebilen yöntemin GroupX spam sıralaması olduğu görülmektedir. Diğer taraftan kırmızı eğriler incelendiğinde, yani spam tespiti konusunda, grafiklerde gözlenen dikkat çekici bir fark olmamakla birlikte Fusion spam sıralamasının spam dokümanları biraz daha iyi tespit edebildiği söylenebilir.




**Şekil 5.1.** *Spam sıralamaları bazında ClueWeb09 veri kümesindeki spam ve ilgili doküman dağılımları*

İkili sınıflandırıcıya dönüştürülen spam sıralamalarının (Fusion, Britney, GroupX, UK2006), eşik değeri olarak alınan farklı yüzdelik puan değerlerine göre ClueWeb09 veri kümesi üzerindeki spam ve non-spam sınıflandırma başarımlarının değerlendirilebilmesi için

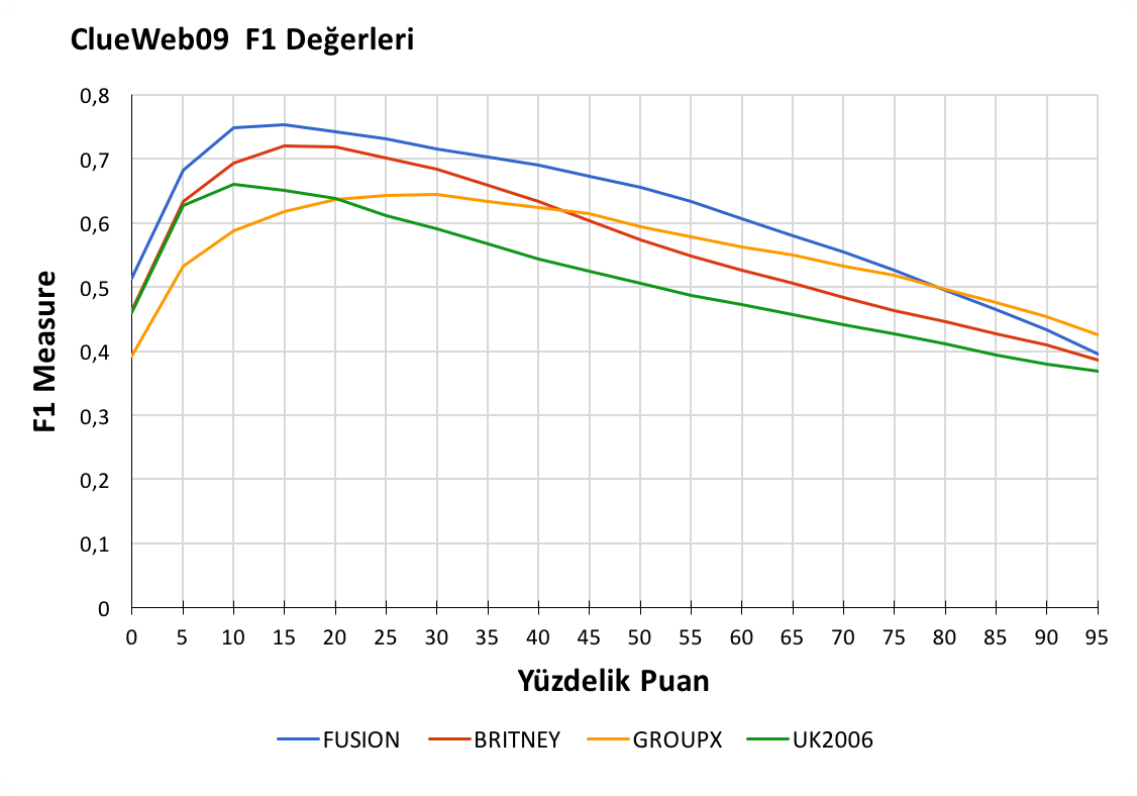
Tablo 5.1'deki hata matrisinden yararlanılarak hesaplanan evrensel sınıflandırma değerlendirme metrikleri aracılığıyla birtakım karşılaştırmalı ölçümler yapılmıştır.

Doğru sınıflandırılan spam dokümanların, spam olarak tahmin edilen toplam dokümanlara oranı olan Precision metriği ve doğru sınıflandırılan spam dokümanların var olan toplam spam doküman sayısına oranı olan Recall metriği, birbiri ile ödünleşen başarım ölçütleridir. Öyle ki; Precision metriği eşik değeri (yüzdelik puan) ile artan bir fonksiyondur. Diğer taraftan, Recall metriği ise eşik değeri (yüzdelik puan) ile azalan bir fonksiyondur. Bazı sistemler için Precision daha önemliyken, bazı sistemler için ise Recall daha önemlidir. Sistem için hangisi daha kritik önemli bir kriter?

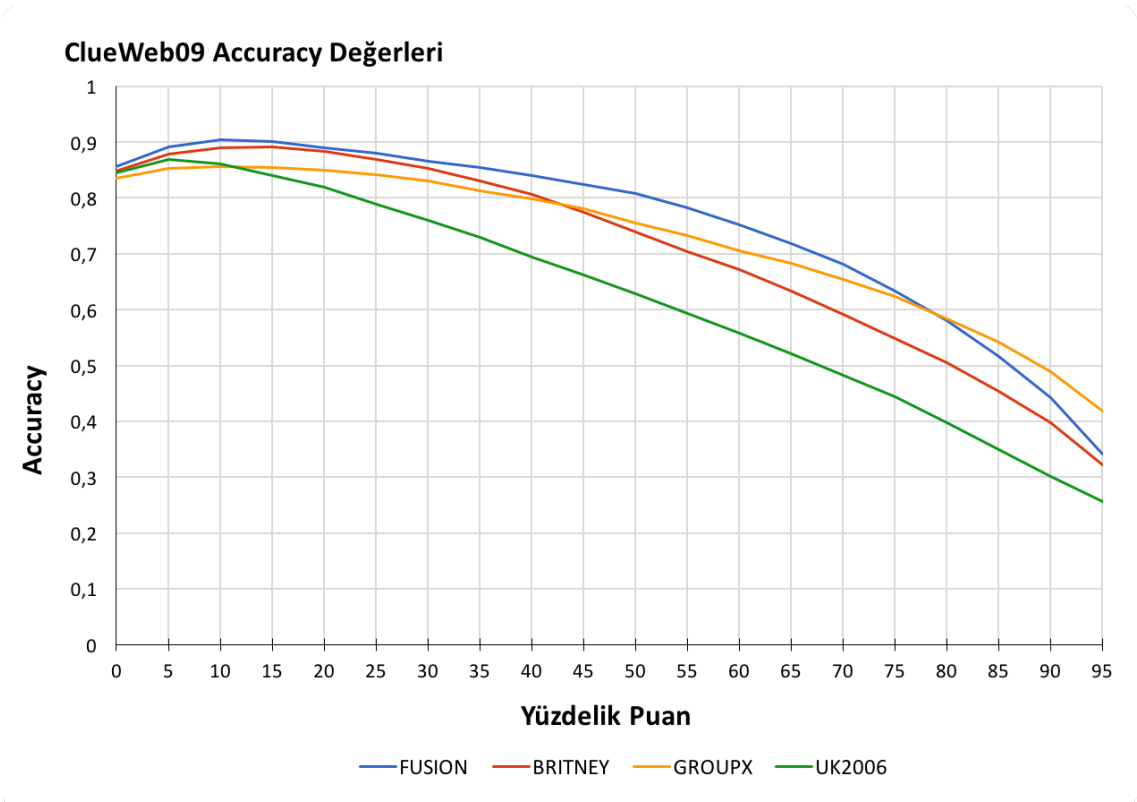
- İlgili dokümanları kaçırmamak mı? (Precision önemli: küçük eşik değeri kullanılmalı)
- Yoksa Spam dokümanları yakalamak mı? (Recall önemli: büyük eşik değeri kullanılmalı)

Harmonik ortalama alınması ile, birbiri ile yarışan bu iki ölçütün her ikisini de aynı anda dikkate almak mümkün olur. Precision ve Recall metriklerinin harmonik ortalaması olan F1-ölçütünün dört farklı spam sıralaması için 0 ile 99 yüzdelik puan aralığındaki grafikleri Şekil 5.2'de sunulurken; sınıflandırıcıların doğruluğunu gösteren Accuracy grafikleri Şekil 5.3'de sunulmuştur.

İkili sınıflandırıcıya dönüştürülen spam sıralamalarının, ClueWeb09 veri kümesi için Şekil 5.2'de verilen F1-ölçütü grafikleri karşılaştırıldığında, mavi renkli eğri ile gösterilen Fusion spam sıralamasının en üstte olduğu gözlenir. 80 eşik değerine kadar en iyi çalışan sınıflandırıcıdır. Eşik değeri 80'in üzerine çıktığında ise turuncu renkli eğri ile gösterilen GroupX spam sıralamasının az farkla birinciliği aldığı gözlenir. Eğrilerin tepe noktalarına bakılarak bir karşılaştırma yapıldığında sıralamanın Fusion > Britney > UK2006 > GroupX şeklinde olduğu görülür. Bu dört sınıflandırıcının Şekil 5.3'de verilen Accuracy grafikleri karşılaştırıldığında da sıralamanın değişmediği ve Fusion spam sıralamasının diğer spam sıralamalarına göre daha yüksek doğruluk değerlerine sahip olduğu görülür.




**Şekil 5.2.** *Spam sıralamalarının ClueWeb09 veri kümesi üzerindeki F1 değerleri*




**Şekil 5.3.** *Spam sıralamalarının ClueWeb09 veri kümesi üzerindeki Accuracy değerleri*

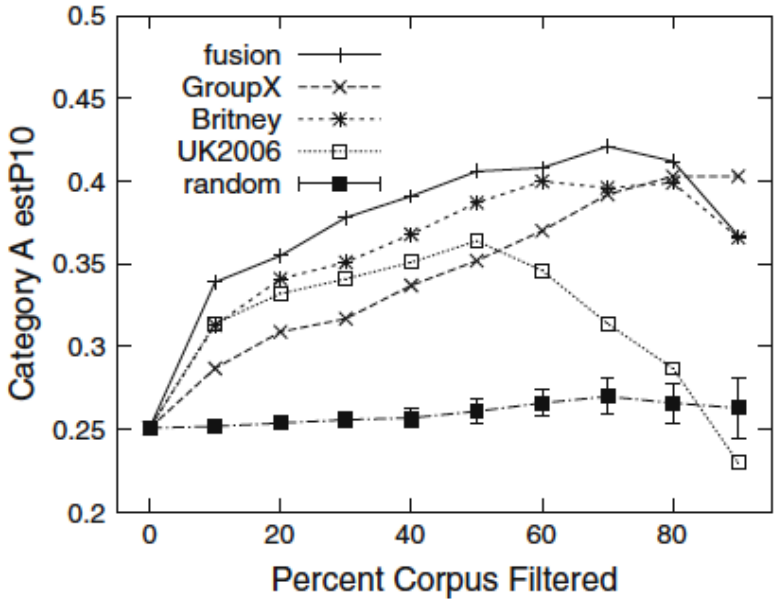
**Tablo 5.2.** *ClueWeb09 veri kümesinde F1 ve Accuracy metriklerini maksimize eden eşik değerleri*

| | **Fusion** | **Britney** | **GroupX** | **UK2006** |
|---|---|---|---|---|
| **F1** | 15 | 15 | 20-30 | 10 |
| **Accuracy** | 10 | 15 | 10 | 5 |

Spam sıralamalarının, F1-ölçütü ve Accuracy metriklerine göre sınıflandırma başarımlarını maksimize eden eşik değerleri, yani yüzdelik puanlar Tablo 5.2'de sunulmuştur. Spam ve non-spam sınıflandırıcı olarak Fusion ve Britney spam sıralamalarının kullanılması durumunda, F1-ölçütü metriğini maksimize eden eşik değerinin 15, UK2006 spam sıralamasının kullanılması durumunda ise bu eşik değerinin 10 olduğu görülür. GroupX spam sıralamasının F1-ölçütü metriğini maksimize eden değer Şekil 5.2'deki grafikten de görülebileceği gibi 30'dur. Fakat eşik değerinin 20 ile 30 aralığında seçilmesi durumunda GroupX spam sıralaması için elde edilen F1-ölçütü değerlerinin birbirlerine çok yakın olduğu, hatta bu aralıkta düze yakın bir çizgi elde edildiği gözlenmiştir. Spam sıralamalarının dördü de göz önünde bulundurulduğunda, F1-ölçütü metriğini maksimize eden eşik değerinin ortalamada 15 olduğu görülür.

Spam sıralamalarının sınıflandırma doğruluğuna bakıldığında, Tablo 5.2'de Accuracy metriği için gösterilen eşik değerleri elde edilir. UK2006 spam sıralamasını kullanan sınıflandırıcının Accuracy metriğini maksimize eden eşik değerinin 5, Fusion ve GroupX spam sıralamalarını kullanan sınıflandırıcıların Accuracy metriğini maksimize eden eşik değerinin 10 ve Britney spam sıralaması için bu eşik değerinin 10 olduğu gözlenir. Spam sıralamalarının dördü de göz önünde bulundurulduğunda, Accuracy metriğini maksimize eden eşik değerinin ortalamada 10 olduğu görülür.

Sonuç olarak, F1-ölçütü ve Accuracy metrikleri için Tablo 5.2'in gösterdiği eşik değeri aralığının (%10-%15 arasında), Cormack ve arkadaşları tarafından önerilen %70 değerinden oldukça küçüktür (Cormack vd., 2011). Aradaki bu farklılığın, "intrinsic" değerlendirme yöntemleri ile "extrinsic" değerlendirme yöntemlerinin farklılığından kaynaklı olduğu düşünülmektedir. Bu nedenle, tam teşekküllü BE sistemlerinin ortalama başarımını maksimize eden ("extrinsic") eşik değerinin, spam sınıflandırmasının doğruluğunu maksimize eden ("intrinsic") eşik değerinden yüksek olduğu gözlemlenmiştir.

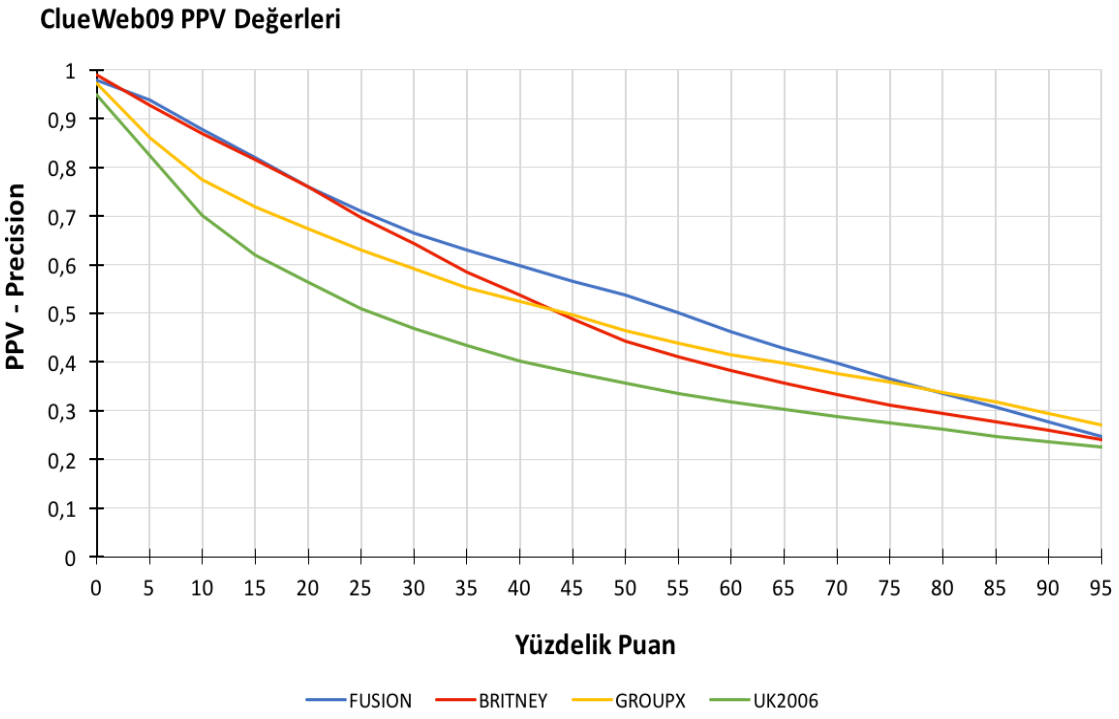



**Şekil 5.4.** *(Cormack vd., 2011) tarafından elde edilen sonuç grafiği*

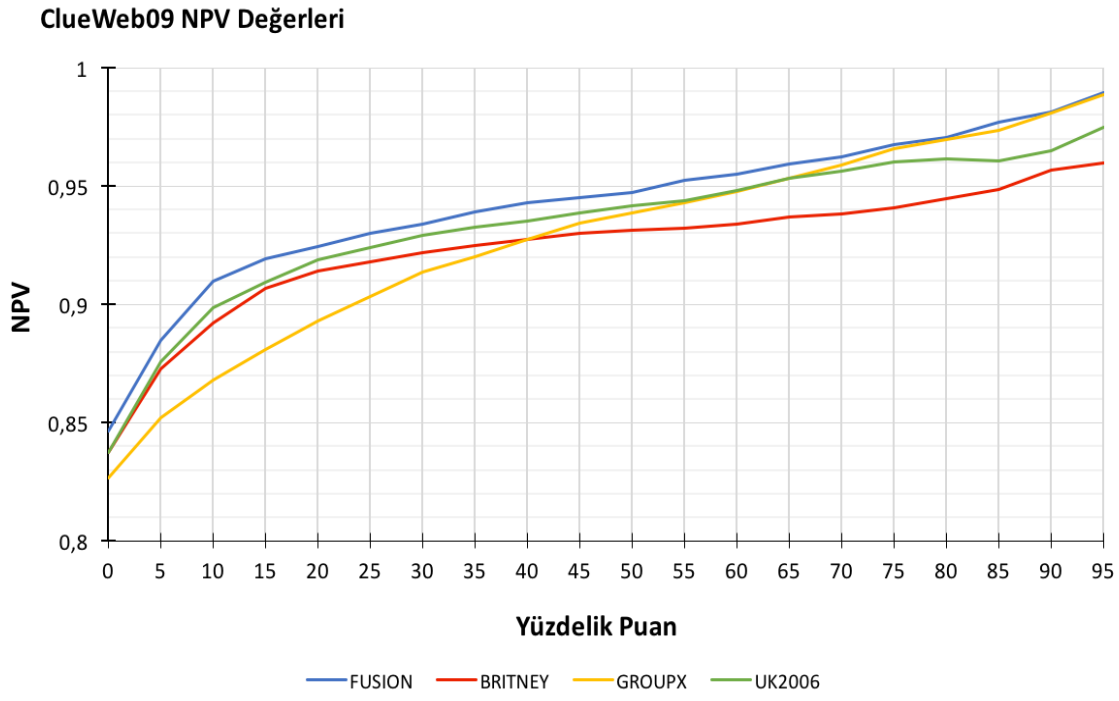
Yapılan çalışmada elde edilen Şekil 5.2 F1-ölçütü grafiğinin, Cormak ve arkadaşları tarafından sunulan sonuç eğrilerine benzer eğilim/davranış gösterdiği tespit edilmiştir (Cormack vd., 2011). Şekil 5.4'de gösterilen söz konusu grafik, tez kapsamında elde edilip Şekil 5.2'de sunulan grafikten farklı olarak, "extrinsic" değerlendirmedir. Bir başka değişle *y* ekseninde BE başarım ölçütü olan *estP@10* verilmiştir (*x* ekseni her iki şekilde de aynı değişkeni, yani eşik değerini, göstermekte olup, tek farklılık *y* eksenin ifade ettiği değişkendir). Bu temel farklılığa rağmen, dört farklı puanlama sisteminin eğrileri dikkat çekici şekilde benzeşmektedir.

Öyle ki, her iki grafikte de Fusion spam sıralamasına ait eğri, yüzdelik puanın bir değerine kadar, diğer eğrileri altında bırakacak şekilde en yukarıda seyir etmektedir. Bu değerden sonra ise, GroupX spam sıralamasına ait eğri, Fusion eğrisinin üstüne çıkmaktadır. Bu değer her iki çalışmada aynı olup 80 olarak ölçülmüştür.

Ayrıca, her iki grafikte de 0-10 yüzdelik puan aralığında Britney ve UK2006 spam sıralamalarına ait eğriler çakışmakta, 10'un üstüne çıkıldığında ise Britney spam sıralama eğrisinin UK2006 spam sıralama eğrisinin oldukça üstüne yer aldığı görülmektedir. UK2006 spam sıralama eğrisi ise belli bir yüzdelik puana ulaştıktan sonra her iki grafikte de hızlı bir düşüş göstererek diğer spam sıralama eğrilerin altında kalır. Diğer yandan, 0'dan itibaren en alt sıralamada yer alan GroupX spam sıralama eğrisi, belli bir yüzdelik puana ulaştıktan sonra her iki grafikte de hızlı bir yükseliş göstermekte, hatta 80'den sonra Fusion spam sıralama eğrisinin üstüne çıkmaktadır.




**Şekil 5.5.** *Spam sıralamalarının ClueWeb09 veri kümesi üzerindeki PPV değerleri*




**Şekil 5.6.** *Spam sıralamalarının ClueWeb09 veri kümesi üzerindeki NPV değerleri*

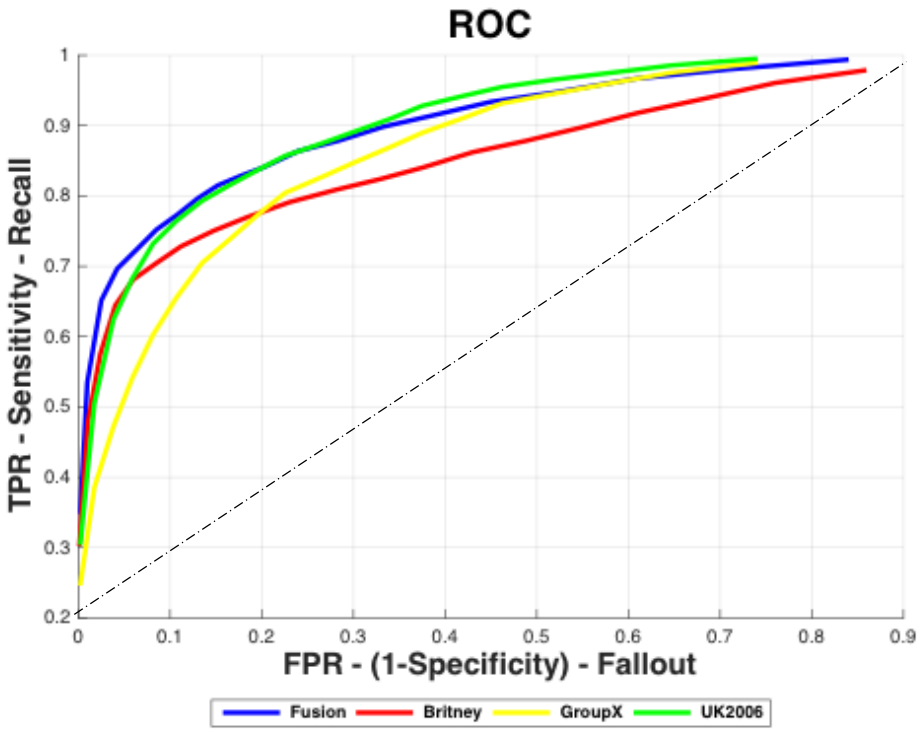
İkili sınıflandırıcıya dönüştürülen Fusion, Britney, GroupX ve UK2006 spam sıralamalarının, farklı yüzdelik puan aralıkları için spam doküman tespitindeki doğrulukları Şekil 5.5'de verilen PPV grafiğinde sunulurken; non-spam doküman tespitindeki doğrulukları Şekil 5.6'daki NPV grafiğinde sunulmuştur.

Doğru sınıflandırılan spam dokümanların, spam olarak tahmin edilen toplam dokümanlara oranının farklı yüzdelik puan aralıklarına göre değişimini gösteren ve Şekil 5.5'de verilen PPV (ya da Precision) grafiğine göre Fusion spam sıralamasının spam doküman tespitindeki doğruluğunun 80'e kadar en üstte giderken; 80'den sonra GroupX spam sıralama eğrisinin üste geçtiği görülür. Ayrıca, UK2006 spam sıralamasının, spam doküman tespitindeki doğruluğunun diğerlerine göre çok daha düşük olduğu ve grafikte en altta yer aldığı gözlenir.

Doğru sınıflandırılan non-spam dokümanların, non-spam olarak tahmin edilen toplam dokümanlara oranının farklı yüzdelik puan aralıklarına göre değişimini gösteren ve Şekil 5.6'de verilen NPV grafiğine göre de Fusion spam sıralamasının non-spam doküman tespitindeki doğruluğunun diğer spam sıralamalarına kıyasla daha iyi olduğu gözlenir. 65 eşik değerine kadar UK2006 spam sıralaması ikinci sırada yer alırken; 65'dan sonra en sondan gelen GroupX spam sıralamasının UK2006'yı geçtiği; hatta 75 eşik değeri sonrasında neredeyse Fusion spam sıralaması ile aynı başarıma sahip olduğu görülür. 40 eşik değerine kadar Britney spam sıralaması üçüncü sırada gelirken, bu değerden sonra başarımının oldukça düştüğü ve en altta yer aldığı görülür.

Spam sıralaması bazlı sınıflandırıcıların ROC eğrileri aracılığıyla da başarımları karşılaştırılmıştır. $x$ ekseninde FPR, $y$ ekseninde TPR değerleri göz önüne alınarak çizilen spam sıralamalarına ait ROC eğrileri Şekil 5.7'de gösterilmiştir.

ROC eğrisi ne kadar sol üst köşeye yakınsa sınıflandırıcı o kadar başarılı demektir. Grafikteki köşegen sınıf tahminin rasgele yapıldığı durumu gösteren rastgele tahmin çizgisidir. Bu çizginin altında kalan alan ise rastgele tahminden bile kötü olarak nitelendirilir. Buna göre ROC eğrileri incelenip karşılaştırıldığında, Fusion ve UK2006 spam sıralaması bazlı sınıflandırıcıların sol üst köşeye daha yakın olduğu görülür. Ayrıca, bu iki eğri altında kalan alanların da daha büyük olduğu, dolayısıyla daha iyi başarıma sahip oldukları söylenebilir. Fusion ve UK2006 spam sıralamalarının ardından GroupX spam sıralamasının, daha sonra ise Britney spam sıralamasının yer aldığı Şekil 5.7'de görülebilmektedir.



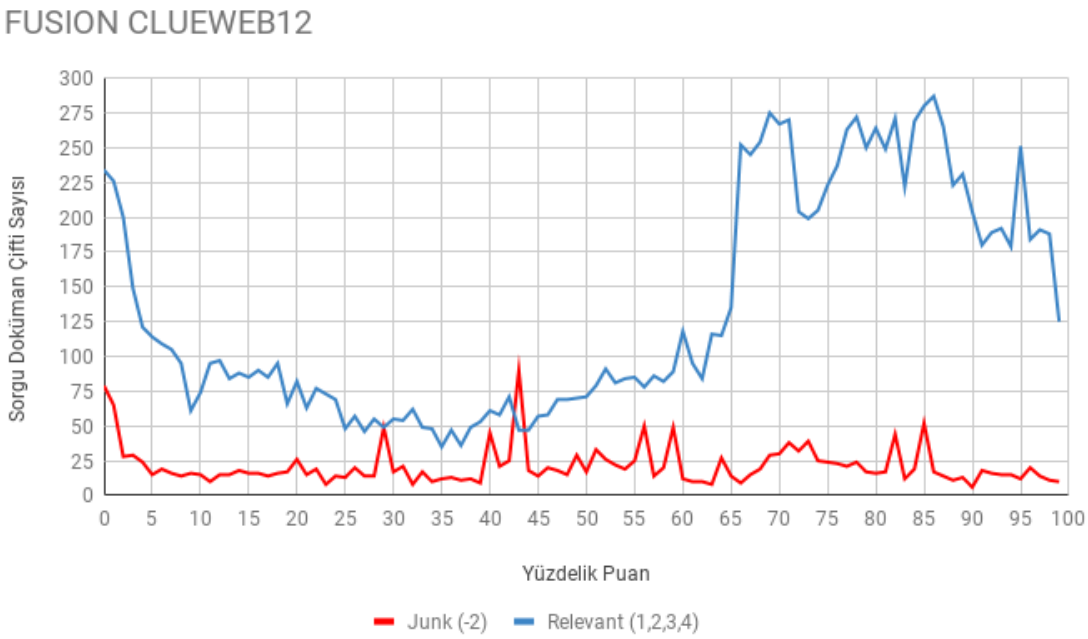

**Şekil 5.7.** *Spam sıralamalarının ClueWeb09 veri kümesi üzerindeki ROC eğrileri*

## 5.3. ClueWeb12 Veri Kümesi Sonuçları

ClueWeb12 veri kümesi için Tablo 4.4'de sunulan spam ve non-spam olarak işaretlenmiş dokümanların Fusion spam sıralamasına göre dağılımı *x* ekseni yüzdelik puan, *y* ekseni ise sorgu doküman çifti sayısını gösterecek biçimde Şekil 5.8'de verilmiştir. Kırmızı renkli çizgi spam dokümanları, mavi renkli çizgi ise non-spam (ilgili) dokümanları göstermektedir.

ClueWeb12 veri kümesi için elde edilen bu sonuçlar incelendiğinde, her yüzdelik puan aralığında spam dokümanların eşit miktarlarda bulunduğu gözlenmektedir. Spam dokümanlar için ClueWeb09 veri kümesinde elde edilen eğilimin, ClueWeb12 veri kümesinde olmadığı, yani spam dokümanların düz çizgiyi andıran, yoğunlaştığı bir nokta bulunmadığı görülmektedir.

Benzer şekilde, Fusion spam sıralaması bazlı ikili sınıflandırıcının non-spam (ilgili) dokümanları ayırt edebilme özelliği de yoktur. Yüzdelik puanının 5'den küçük olduğu kısımda ve 65'den büyük olduğu kısımlarda ilgili doküman sayısının daha çok olduğu gözlenmektedir.



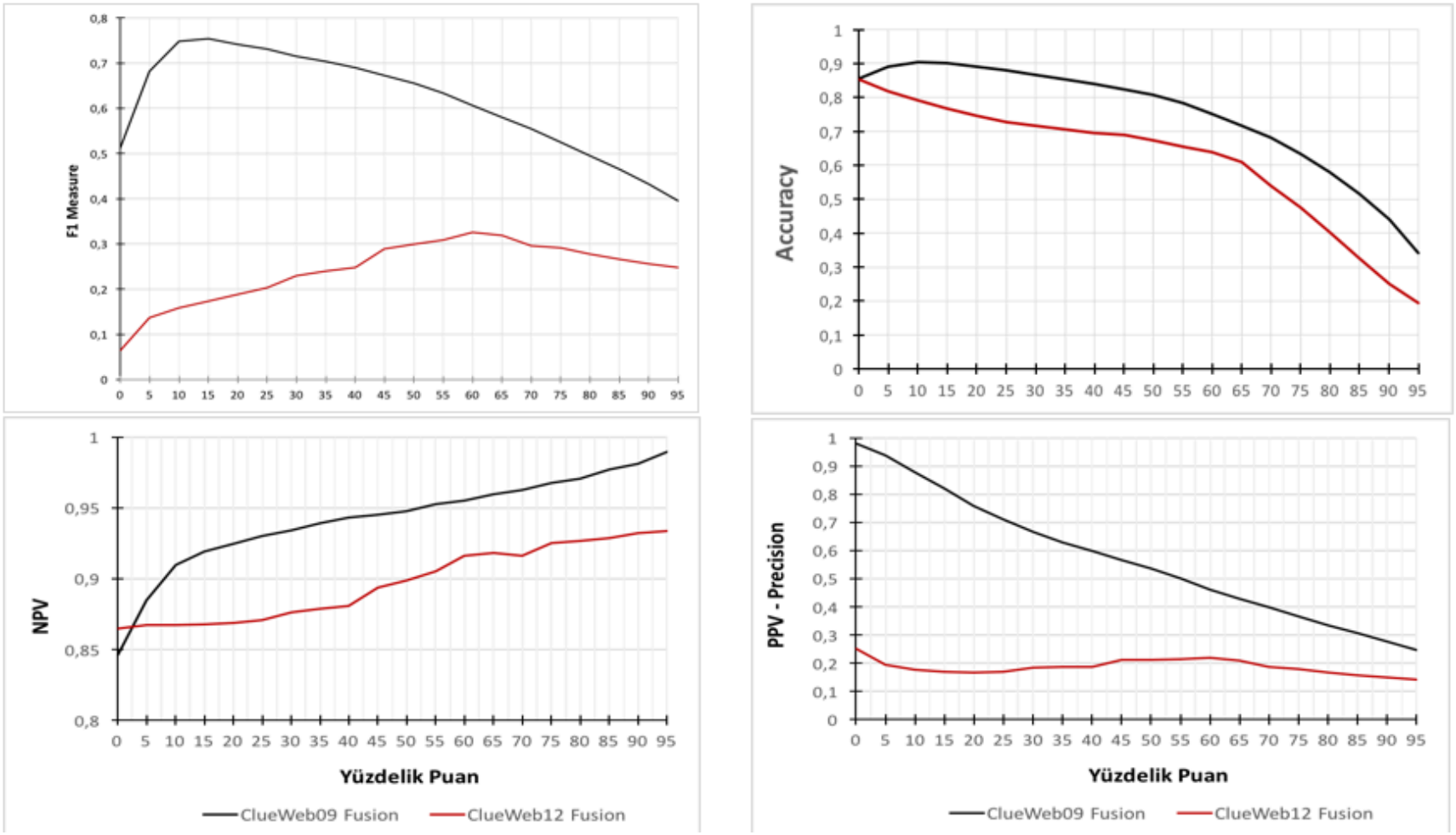

**Şekil 5.8.** *Fusion spam sıralaması bazında ClueWeb12 veri kümesindeki spam ve ilgili doküman dağılımları*

ClueWeb12 veri kümesi için tek spam puan listesi yayınlandığından, bir kıyaslama yapılabilmesi adına, bu ikili sınıflandırıcı ile ClueWeb09 veri kümesindeki Fusion sıralaması bazlı sınıflandırıcının başarısı karşılaştırılmıştır. Eşik değeri olarak alınan farklı yüzdelik puan değerlerine göre ClueWeb12 veri kümesi üzerindeki Fusion sıralamasının, spam ve non-spam sınıflandırma başarımı,

Tablo 5.1'deki hata matrisinden yararlanılarak hesaplanan evrensel sınıflandırma metrikleriyle ölçülmüştür. Şekil 5.9'da CluWeb12 Fusion sıralaması (kırmızı ile gösterilen eğriler) ile ClueWeb09 Fusion sıralaması (siyah ile gösterilen eğriler) bazlı sınıflandırıcıların F1-ölçütü, Accuracy, PPV, NPV başarım ölçütlerine göre karşılaştırmalı grafikleri sunulmuştur.

Şekil 5.9'da gösterilen, Precision ve Recall metriklerinin harmonik ortalaması olan F1-ölçütü grafikleri karşılaştırıldığında ClueWeb12 Fusion sıralamalarının, ClueWeb09 Fusion sıralamalarına göre çok daha düşük olduğu gözlenir.

ClueWeb12 ve ClueWeb09 Fusion sıralaması bazlı sınıflandırıcılar Accuracy metriğine göre kıyaslandığında da aynı durum gözlenir. ClueWeb12 Fusion sıralaması bazlı sınıflandırıcının Accuracy değerinin ClueWeb09 Fusion sıralaması bazlı sınıflandırıcıya göre biraz daha düşük olduğu Şekil 5.9'dan görülebilmektedir.

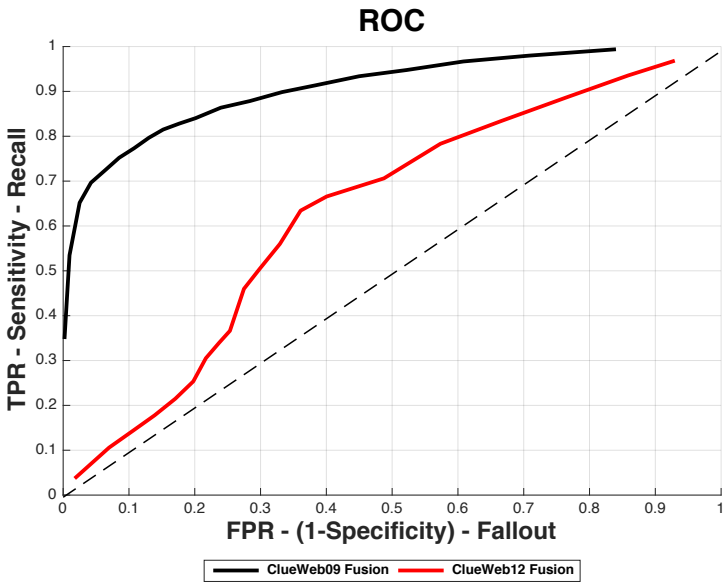



**Şekil 5.9.** *ClueWeb12 ve ClueWeb09 Fusion sıralamalarının F1, Accuracy, NPV, PPV değerleri*

Doğru sınıflandırılan spam dokümanların, spam olarak tahmin edilen toplam dokümanlara oranının farklı yüzdelik puan aralıklarına göre değişimini gösteren PPV grafiğine göre de ClueWeb09 Fusion sıralaması bazlı sınıflandırıcının spam dokümanları daha iyi tespit edebildiği, hatta ClueWeb12 Fusion sıralaması ile bu dokümanların tespitinin oldukça başarısız olduğu Şekil 5.9'dan görülebilmektedir.

Doğru sınıflandırılan non-spam dokümanların, non-spam olarak tahmin edilen toplam dokümanlara oranının farklı yüzdelik puan aralıklarına göre değişimini gösteren NPV grafiğine göre de ClueWeb09 Fusion spam sıralaması bazlı sınıflandırıcının non-spam doküman tespitindeki doğruluğunun ClueWeb12 Fusion spam sıralaması bazlı sınıflandırıcıya göre daha iyi olduğu Şekil 5.9'dan görülebilmektedir.

Sonuç olarak, sınıflandırma başarım ölçütlerine göre çizilen grafiklerin hepsinde ClueWeb09 Fusion sıralaması en üstte yer almaktadır. Dolayısıyla, ClueWeb09 Fusion spam sıralaması bazlı sınıflandırıcı, ClueWeb12 Fusion spam sıralaması bazlı sınıflandırıcıya göre çok daha başarılıdır.




**Şekil 5.10.** *ClueWeb12 ve ClueWeb09 Fusion sıralamalarının ROC eğrileri*

Bu iki sınıflandırıcının ROC eğrisi grafiklerinde de durum aynıdır. Şekil 5.10'da gösterilen ClueWeb12 ve ClueWeb09 Fusion spam sıralamalarına ait ROC eğrileri karşılaştırıldığında, ClueWeb09 Fusion eğrisinin sol üst köşeye daha yakın olduğu, ClueWeb12 Fusion eğrisinin ise çok daha altta kaldığı, hatta spam non-spam sınıf

tahminin rasgele yapıldığı durumu gösteren rastgele tahmin çizgisine (tireler ile gösterilen eğri) çok yakın olduğu gözlenir.

Sonuç olarak, ClueWeb12 Fusion spam puanları, spam ve non-spam ayrımı yapan bir ikili sınıflandırıcı olarak oldukça kötü çalışmaktadır. Bu durum başka araştırmacılar tarafından da gözlemlenmiştir (Xiong vd., 2016).

# 6. TARTIŞMA

ClueWeb09 veri kümesi için 2010, 2011 ve 2012 yıllarında yayınlanan sorgu doküman yargılarını içeren *qrels* dosyaları incelendiğinde, toplam 60765 sorgu doküman çiftinin (*qid/docid*) insanlar tarafından değerlendirildiği görülür. Bu sayı belirtilen üç yıl için yayınlanan *qrels* dosyalarındaki satırların toplam sayısıdır, yani sorgu doküman çiftlerine (*qid/docid*) göre tekil, ama *docid*'ye göre tekil değildir.

Değerlendirilen tekil doküman sayısı analiz edildiğinde, bir başka deyişle, 60765 sorgu doküman çifti (*qid/docid*) içerisindeki tekil doküman (*docid*) sayısının 59769 kadar olduğu görülür. Sadece bir defa değerlendirilen 59443 doküman vardır. Bunun anlamı bazı dokümanların birden fazla sorguya karşı birden çok kez değerlendirilmiş olduğudur. Aradaki farka bakıldığında, toplam 326 dokümanın (59769 - 59443 = 326) birden çok kez değerlendirildiği görülür. Ayrıca yıllar arasında bazı dokümanların farklı yıllarda tekrar değerlendirildiği görülmüştür. Tablo 6.1, ClueWeb09 veri kümesi için yıl bazında sorgu-doküman değerlendirme istatistiklerini içerir.

**Tablo 6.1.** *ClueWeb09 sorgu-doküman değerlendirme istatistikleri*

| Veri Kümesi | Track | Yıl | *qrels* Dosyasındaki Satır Sayısı | Değerlendirilen Tekil Doküman Sayısı | Birden Fazla Defa Değerlendirilen Doküman Sayısı |
|---|---|---|---|---|---|
| ClueWeb09 | Web Track | 2010 | 25329 | 24470 | 206 |
| | | 2011 | 19381 | 19344 | 37 |
| | | 2012 | 16055 | 16036 | 12 |
| | | **Yıllardan Bağımsız** | 60765 | 59769 | 326 |

Birden fazla kez değerlendirilen 326 dokümandan, 277 tanesi tek bir uygunluk yargısı etiketiyle etiketlenmişken, 49 tanesinin ise birden fazla uygunluk yargısı etiketi aldığı görülmüştür. Bu birden fazla etiket alan dokümanların uygunluk yargıları incelendiğinde, yani bu dokümanlar için gerçek kişiler tarafından verilen uygunluk yargısı seviyelerine (-2, 0, 1, 2, 3 ya da 4) bakıldığında, Tablo 6.2 elde edilir.

**Tablo 6.2.** *ClueWeb09 veri kümesinde birden çok kez değerlendirilen ve birden çok etiket alan dokümanların istatistikleri*

| **Farklı Uygunluk Yargısı Etiket Grupları** | **Uygunluk Yargısı Etiket Grubundaki Her bir Etiketin Verilme Sayısı** | **Doküman Sayısı (Değerlendirme Sayısı)** |
|---|---|---|
| "-2, 0" | -2 (32), 0 (79) | 20 (111) |
| "0, 1" | 0 (75), 1 (21) | 20 (96) |
| "0, 2" | 0 (1), 2 (1) | 1 (2) |
| "1, 2" | 1 (6), 2 (6) | 6 (12) |
| "0, 4" | 0 (1), 4 (1) | 1 (2) |
| "-2, 0, 1" | -2 (1), 0 (8), 1 (1) | 1 (10) |

Birden çok kez değerlendirilen dokümanlar, gerçek kişiler tarafından her bir dokümana verilen birbirinden farklı tekil uygunluk yargısı seviyelerine göre gruplandığında, Tablo 6.2'nin ilk sütununda gösterilen uygunluk yargısı etiket grupları elde edilir. Daha açık belirmek gerekirse, birden fazla defa değerlendirilen bir dokümana değerlendiriciler tarafından atanan farklı uygunluk yargısı etiketleri bu sütunda gösterilmiştir. Tablonun ikinci sütununda, birinci sütunda verilen uygunluk yargısı etiket grubundaki dokümanlara her bir etiketin verilme sayısı, uygunluk yargısı etiketi ve parantez içinde o etiketin toplam verilme sayısı (ör. -2(32)) şeklinde sunulmuştur. Tablonun üçüncü sütununda ise birinci sütunda verilen uygunluk yargısı etiket grubuna sahip doküman sayısı, yani frekansı ve parantez içerisinde ise o dokümanın toplam değerlendirilme sayısı verilmiştir. Tablo kendi içinde frekansa göre sıralanmıştır.

Örneğin, ClueWeb09 veri kümesinde birden çok kez değerlendirilen ve birden çok etiket alan dokümanlar içerisinde yer alan "*clueweb09-en0008-18-23941*" doküman numarasına (*docid*) sahip doküman, toplamda 10 farklı sorgu için değerlendirilmiş olup (*qrels* dosyasında 10 farklı satırda geçiyor, yani 10 farklı *qid* için değerlendirilmiş), 8 değerlendirmede uygunluk yargısı etiketi olarak 0, 1 değerlendirmede uygunluk yargısı etiketi olarak 1 ve 1 değerlendirmede de uygunluk yargısı etiketi olarak -2 verilmiştir. Bu doküman sadece Web Track 2010 *qrels* de değerlendirilmiştir. Dolayısıyla bu dokümanın tekil uygunluk etiketi "-2, 0, 1" 'dir ve birden fazla değerlendirilen

dokümanlar kümesi içerisinde bu üç farklı tekil etiket ile değerlendirilmiş tek dokümandır.

Bir başka örnek ise, ClueWeb09 veri kümesinde birden çok kez değerlendirilen ve birden çok etiket alan dokümanlar içerisinde yer alan "*clueweb09-en0002-84-02061*" doküman numarasına (*docid*) sahip doküman, toplamda 15 farklı sorgu için değerlendirilmiş olup, Web Track 2010'da değerlendirildiği 14 farklı sorguda 0 uygunluk yargısı alırken; Web Track 2011'de değerlendirildiği tek sorguda (*qid=133*) ise -2 uygunluk yargısı almıştır. Dolayısıyla, bu dokümanın tekil uygunluk yargısı etiket grubu "-2, 0" 'dır ve birden fazla etiket alan dokümanlar kümesi içerisinde bu iki farklı tekil etiket ile değerlendirilmiş olan 20 dokümandan birisidir.

Sonuç olarak, Tablo 6.2'den de çıkarılabileceği gibi, ClueWeb09 veri kümesinde toplam 49 (20+20+1+6+1+1) doküman birden fazla uygunluk yargısı etiketi ile değerlendirilmiştir.

Bir dokümanın bir sorgu için ilgili olup başka bir sorgu için ilgisiz olması çok doğaldır. Çünkü ilgililik ve ilgisizlik kavramları sorgu-doküman (*qid/docid*) ikilisi için tanımlıdır ve anlamlıdır. Dolayısıyla, birden fazla uygunluk yargısı etiketi almış bir dokümanın aldığı farklı uygunluk yargısı etiketlerinden birisi 0 iken diğerinin 1, 2, 3, ya da 4 olması, yani 0 uygunluk yargısı etiketi ile birlikte 1, 2, 3, 4 uygunluk yargısı etiketlerinden birisinin (ya da birkaçının) birlikte gözlemlenmesi normaldir. Bu bağlamda Tablo 6.2 tekrar incelendiğinde, birden fazla etiket ile işaretlenmiş bir dokümanın, ClueWeb09'daki çoklu değerlendirmeler bazında elde edilen farklı uygunluk yargısı etiket gruplarından "0, 1", "0, 2", ve "0, 4" şeklindekilerden birine sahip olması doğaldır. "1, 2" uygunluk yargısı etiket grubu ise bir sorgu için alakalı olan dokümanın başka bir sorgu için daha alakalı olduğu durumdur. Dolayısıyla, bir doküman farklı sorgular için farklı ilgililik düzeylerine sahip olabileceğinden, çoklu kez değerlendirilmiş bir doküman için 1, 2, 3 ya da 4 uygunluk yargısı etiketlerinden ikisinin ya da daha çoğunun birlikte gözlemlenmesi de doğaldır.

Öte yandan asıl dikkat çekici olan ya da bu çalışmayı ilgilendiren kısım -2 (spam) uygunluk yargısı etiketi ile en çok birlikte kullanılan uygunluk yargısı etiket(ler)inin neler olduğu sorusunun cevabıdır. Tablo 6.2'den gözlemlenen, ClueWeb09 veri kümesi için -2 uygunluk yargısı etiketinin en çok 0 uygunluk yargısı etiketi ile karıştırıldığıdır. Yani değerlendirmeler esnasında değerlendiriciler alakasız (0) ve spam (-2) doküman ayrışımını tam yapamamıştır.

Tablonun sunulmasındaki asıl amaç herhangi bir sorguya karşılık ilgili (1, 2, 3, 4) olarak işaretlenmiş bir dokümanın spam (-2) olamayacağı varsayımını ihlal eden aykırı bir örneğin gözlenip gözlemlenmediğinin kontrol edilmesidir. Dolayısıyla -2 uygunluk yargısı etiketi ile 1, 2, 3 ya da 4 uygunluk yargısı etiketlerinden bir ya da birkaçının birlikte gözlemlenmesi bu çalışmanın varsayımını ihlal eder. ClueWeb09'daki çoklu değerlendirmeler için elde edilen farklı uygunluk etiket gruplarından "-2, 0, 1", yani tablonun son satırı bu tarz bir örnektir. Fakat toplam frekans göz önüne alındığında ihmal edilebilecek bir örnek olduğu görülür, çünkü bu şekilde değerlendirilen sadece tek bir doküman vardır ve 10 sorgu-doküman (*qid/docid*) ikilisine tekabül eder. Bu sayısı 60765 satır dikkate alındığında çok küçüktür.

Bu çalışmada hem spam hem de ilgili olarak birden fazla işaretlenen dokümanlar, talim (gerçek doğrular) verisine birden fazla (çoklu) eklenmiştir.

**Tablo 6.3.** *ClueWeb12 sorgu-doküman değerlendirme istatistikleri*

| Veri Kümesi | Track | Yıl | *qrels* Dosyasındaki Satır Sayısı | Değerlendirilen Tekil Doküman Sayısı | Birden Fazla Defa Değerlendirilen Doküman Sayısı |
|---|---|---|---|---|---|
| ClueWeb12 | Web Track | 2013 | 14474 | 14336 | 94 |
| | | 2014 | 14432 | 14429 | 3 |
| | Tasks Track | 2015 | 3422 | 2714 | 47 |
| | | 2016 | 3671 | 3532 | 90 |
| | | **Yıllardan Bağımsız** | 35999 | 34557 | 681 |

Aynı analizler ClueWeb12 veri kümesine ait sorgu doküman yargıları için de yapılmıştır. ClueWeb012 veri kümesi için 2013, 2014, 2015 ve 2016 yıllarında yayınlanan sorgu doküman yargılarını içeren *qrels* dosyaları incelendiğinde, toplam 35999 sorgu doküman çiftinin (*qid/docid*) insanlar tarafından değerlendirildiği görülür. Bu sayı belirtilen dört yıl için yayınlanan *qrels* dosyalarındaki satırların toplam sayısıdır, yani sorgu doküman çiftlerine (*qid/docid*) göre tekil, ama *docid*'ye göre tekil değildir. 2017 yılında yapılan Tasks Tracks görevi kapsamında da ClueWeb12 veri

kümesi kullanılmıştır, fakat bu yıla ait *qrels* dosyası web'de henüz yayınlanmamış olduğundan çalışmaya dahil edilememiştir.

Bu 35999 sorgu doküman çifti (*qid/docid*) içerisindeki tekil doküman sayısı analiz edildiğinde, 34557 kadar tekil doküman (*docid*) olduğu görülür. Sadece bir defa değerlendirilen 33876 doküman vardır. Aradaki fark kadar doküman birden fazla sorguya karşı birden çok defa değerlendirilmiştir. Dolayısıyla, toplam 681 doküman (34557 - 33876 = 681) birden çok kez değerlendirilmiştir. Tablo 6.3, ClueWeb12 veri kümesi için yıl bazında sorgu-doküman değerlendirme istatistiklerini içerir.

**Tablo 6.4.** *ClueWeb12 veri kümesinde birden çok kez değerlendirilen ve birden çok etiket alan dokümanların istatistikleri*

| **Farklı Uygunluk Yargısı Etiket Grupları** | **Uygunluk Yargısı Etiket Grubundaki Her bir Etiketin Verilme Sayısı** | **Doküman Sayısı (Değerlendirme Sayısı)** |
|---|---|---|
| "-2, 0, 1" | -2 (184), 0 (251), 1 (27) | 15 (462) |
| "-2, 0" | -2 (127), 0 (59) | 35 (186) |
| "-2, 1" | -2 (11), 1(11) | 11 (22) |
| "-2, 2" | -2 (1), 2 (1) | 1 (2) |
| "0, 1" | 0 (261), 1 (221) | 220 (482) |
| "0, 2" | 0 (48), 2 (48) | 48 (96) |
| "1, 2" | 1 (51), 2 (51) | 51 (102) |
| "1, 3" | 1 (4), 3 (4) | 4 (8) |

Birden fazla defa değerlendirilen 681 dokümandan, 296 tanesi tek bir uygunluk yargısı etiketiyle etiketlenmişken, 385 tanesinin ise birden fazla uygunluk yargısı etiketi aldığı görülmüştür. Bu birden fazla etiket alan dokümanlar için gerçek kişiler tarafından verilen uygunluk yargısı seviyeleri incelendiğinde, Tablo 6.4 elde edilir. ClueWeb09 için verilen Tablo 6.2'a benzer olarak, Tablo 6.4'ün ilk sütunu birden fazla defa değerlendirilen bir dokümana değerlendiriciler tarafından atanan birbirinden farklı tekil uygunluk yargısı etiket gruplarını, ikinci sütunu o uygunluk yargısı etiket grubundaki dokümanlara her bir etiketin verilme sayısını, üçüncü sütunu ise o uygunluk yargısı etiket grubuyla işaretlenmiş doküman sayısını verir. Tablo kendi içinde uygunluk etiketlerine göre sıralanmıştır.

Dolayısıyla, Tablo 6.4'den de görüldüğü gibi, ClueWeb12 veri kümesinde toplam 385 (15+35+11+1+220+48+51+4) doküman birden fazla uygunluk yargısı etiketi ile değerlendirilmiştir.

Birden çok uygunluk yargısı etiketi ile değerlendirilmiş dokümanlar için ClueWeb12'deki çoklu değerlendirmeler bazında elde edilen farklı uygunluk yargısı etiket gruplarından "0, 1" ve "0, 2" şeklindekiler sorun teşkil etmez. Çünkü daha önce de belirtildiği gibi ilgililik sorgu-doküman çifti ile alakalı bir kavram olduğundan, bir doküman bir sorgu için ilgiliyken başka bir sorgu için ilgisiz olabilir. Aynı şekilde, "1, 2" ve "1, 3" uygunluk yargısı etiket grupları da doğaldır. Bir doküman farklı sorgular için farklı ilgililik düzeylerine sahip olabileceğinden ötürü, yani bir sorgu için alakalı, ama başka bir sorgu için daha da çok alakalı olarak değerlendirilebileceğinden dolayı, bu uygunluk yargısı etiket grupları da sorun oluşturmaz.

ClueWeb12 veri kümesi için -2 (spam) uygunluk yargısı etiketi ile en çok birlikte kullanılan uygunluk yargısı etiket(ler)inin neler olduğu Tablo 6.4'den incelendiğinde, ClueWeb09 analizlerine benzer olarak -2 uygunluk yargısı etiketinin en çok 0 uygunluk yargısı etiketi ile karıştırıldı, dolayısıyla değerlendiricilerin değerlendirmeleri sırasında alakasız (0) ve spam (-2) doküman ayrışımını tam olarak yapamamış oldukları görülür. Tablo 6.4'ün ikinci satırı "-2, 0" uygunluk yargısı etiket grubunu göstermektedir.

Öte yandan, herhangi bir sorguya karşılık ilgili (1, 2, 3, 4) olarak işaretlenmiş bir dokümanın spam (-2) olamayacağı varsayımını ihlal eden aykırı bir örneğin olup olmadığı kontrol edildiğinde, ClueWeb09 sonuçlarından farklı bir durum ortaya çıkar. ClueWeb12 için elde edilen "-2, 0, 1", "-2, 1" ve "-2, 2" uygunluk yargısı etiket grupları, yani Tablo 6.4'in birinci, üçüncü ve dördüncü satırları varsayımını ihlal eden örneklerdir. Çünkü -2 uygunluk yargısı etiketi ile 1, 2, 3 ya da 4 uygunluk yargısı etiketlerinden bir ya da birkaçının birlikte gözlemlenmesi bu çalışmanın varsayımını ihlal eder.

Daha detaylı inceleme için Tablo 6.4'de gösterilen uygunluk yargısı etiket grupları -2 içerenler Spam (S), 0 içerenler non-relevant (N) ve [1, 2, 3 ya da 4] içerenler relevant (R) ile sembolize edilip, kendi içerisinde tekrar gruplanarak, Tablo 6.5'de sunulmuştur.

Tablo 6.5'in birinci ve üçüncü satırları varsayımını bozan "*S, R*" etiket grubunu içermekte olup, bazı sorgular için ilgili bazı sorgular için ise spam olarak işaretlenmiş dokümanları göstermektedir. Etiketlerin verilme sayıları incelendiğinde, 184+27+12+12

defa S ve R etiketlerinin değerlendiriciler tarafından karıştırılmış olduğu görülür. Bir sorguya karşı ilgili olarak işaretlenen bir dokümanın spam olamayacağı varsayımını ihlal eden bu örnekler, genel toplam göz önünde bulundurulduğunda, ihmal edilebilecek kadar azdır.

**Tablo 6.5.** *Tablo 6.4'ün SNR sembolizasyonu (S=-2 N=0, R=1,2,3,4)*

| Farklı Uygunluk Yargısı Etiket Grupları | Uygunluk Yargısı Etiket Grubundaki Her bir Etiketin Verilme Sayısı | Doküman Sayısı | Değerlendirme Sayısı |
|---|---|---|---|
| "S, N, R" | -2 (184), 0 (251), 1 (27) | 15 | 462 |
| "S, N" | -2 (127), 0 (59) | 35 | 186 |
| "S, R" | -2 (12), 1(12) | 12 | 24 |
| "N, R" | 0 (309), 1 (269) | 268 | 578 |
| "R, R" | 1 (55), 2 (55) | 55 | 110 |

## 6.1. Tasks Track Sorgu Uygunluk Formatı

Tasks Track 2015 ve Tasks Track 2016 için düzenlenen sorgu uygunluk yargı dosyalarında, Web Track dosyalarının aksine, bir sorgu-doküman (*qid/docid*) çifti için birden fazla değerlendirme yer alır. Bunun nedeni her ana görev için farklı sayıda alt görevin tanımlanmış olması ve her bir alt görev için o sorgu-doküman çiftinin uygunluk yargısı değerlendirmesinin ayrı ayrı yapılmış olmasıdır. Bu yarışmalara ait genel bakış makalesinde, Tasks Track sorgu-yargılarının Web Track ad-hoc sitiline dönüştürülebilmesi için, yani bir dokümana tek bir uygunluk yargısı düzeyi atanabilmesi için, şu yolun izlenmesi gerektiği belirtilir: bir dokümana olası tüm alt görevlerde verilen uygunluk yargısı değerlerinden maksimum olanı o dokümana ait tek uygunluk yargısı değeri olarak kabul edilir (Yilmaz vd., 2015). Dolaysıyla bu çalışmada da aynı yöntem uygulanarak Tasks Track sorgu doküman yargıları belirlenmiştir.

Bu noktada, bir sorgu-doküman çifti bir alt görevde spam (-2) olarak işaretlenirken; başka bir alt görevde ilgili (1, 2, 3, 4) olarak işaretlenmiş mi ve önerilen yönteme göre maksimum uygunluk değeri alındığından spam (-2) olarak işaretlenmiş bir doküman kaçırılabilir mi sorularından yola çıkarak, her sorgu-doküman çiftinin alt görevlerinde aldığı uygunluk değerleri incelenmiştir.

**Tablo 6.6.** *Tasks Track 2015-2016 uygunluk yargısı değerleri*

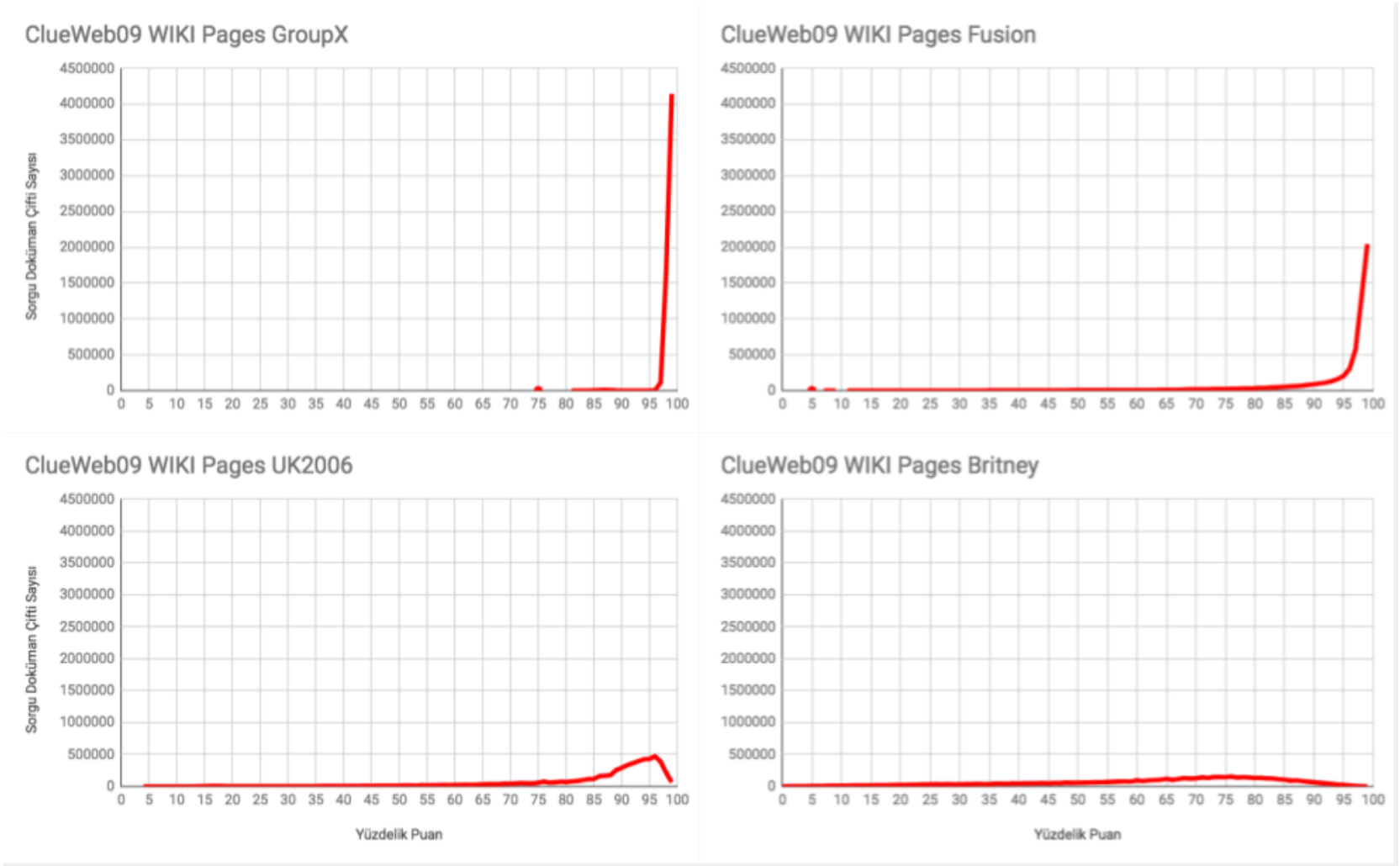
| Tasks Track 2015 | | | Tasks Track 2016 | | |
|---|---|---|---|---|---|
| **En Düşük Uygunluk Değeri (min)** | **En Yüksek Uygunluk Değeri (max)** | **Sorgu Doküman Sayısı** | **En Düşük Uygunluk Değeri (min)** | **En Yüksek Uygunluk Değeri (max)** | **Sorgu Doküman Sayısı** |
| **-2** | **-2** | 1240 | **-2** | **-2** | 97 |
| 0 | 0 | 1179 | 0 | 0 | 1056 |
| 0 | 1 | 863 | 0 | 1 | 2209 |
| 0 | 2 | 140 | 0 | 2 | 309 |

Tablo 6.6 Tasks Tracks'deki sorgu-doküman çiftlerine, alt görevlerinde verilen maksimum ve minimum uygunluk yargısı değerlerini gösterir. Tablodan da görülebildiği gibi, eğer bir doküman belli bir sorgu için spam (-2) olarak işaretlenmişse, o sorguyu kapsayan tüm alt görevlerde de aynı şekilde spam (-2) olarak işaretlenmiştir. Spam dokümanlardaki bu tutarlılığa nazaran ilgisiz (non-relevant) ve ilgili (relevant) dokümanlarda farklı uygunluk yargısı etiketleri gözlenmektedir, fakat bu durum daha önce de belirtildiği gibi doğaldır.

## 6.2. Wikipedia Sayfalarının Spam Sıralama Analizleri

ClueWeb09 veri kümesi, yaklaşık 6 milyon Wikipedia sayfalarını da içinde barındırmaktadır. Wikipedia sayfaları kontrollü oluşturulan web sayfaları olduğundan içerisinde spam olmayacağı düşünülmektedir. Bu kapsamda spam olma olasılığı düşük olan bu tür bir veri kümesi için dört spam sıralamalarının davranışlarının nasıl olacağı incelenmiştir.

Şekil 6.1, ClueWeb09 veri kümesi içerisindeki tüm Wikipedia sayfalarının Fusion, Britney, GroupX ve UK2006 spam sıralamalarına göre dağılımlarını ayrı ayrı göstermektedir. Spam sıralamalarına ait grafikler incelendiğinde, GroupX spam sıralamasının Wikipedia sayfalarına atadığı yüzdelik puanların en düşüğünün 75 olduğu (sadece 1 doküman), 76-81 yüzdelik puan aralığında hiç doküman olmadığı, bu dokümanlara 81-99 aralığında yüzdelik puan atadığı görülür. Hatta Wikipedia sayfalarının %99,5'ini 95 ve üzeri yüzdelik dilimlere almıştır. Dolayısıyla, GroupX spam sıralaması Wikipedia sayfalarına çok yüksek yüzdelik puanlar vererek bu



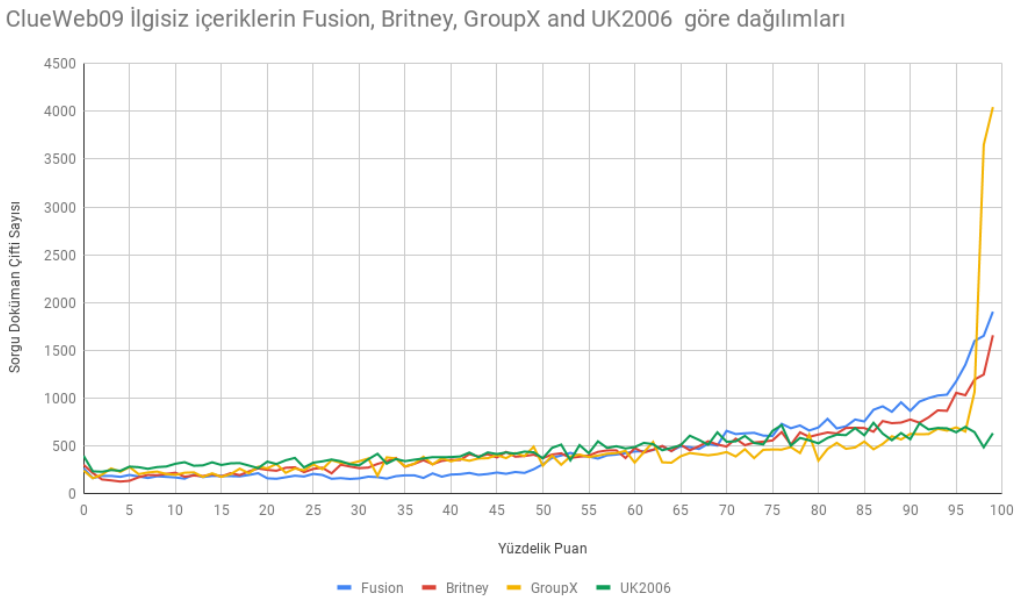

**Şekil 6.1** *ClueWeb09 Wikipedia sayfalarının spam sıralamalarına ait puan dağılımları*

dokümanların spam barındırmadığını göstermektedir. GroupX sıralamasının ardından Fusion, UK2006 ve en son olarak da Britney sıralamasının geldiği gözlenmektedir.

### 6.3. ClueWeb09 Veri Kümesindeki İlgisizi Dokümanların Analizleri

Bu tezde ana olarak ilgili (1, 2, 3, 4) ve spam (-2) olarak işaretlenmiş dokümanlar üzerinden analiz yapılmıştır. İlgisiz (0) olarak işaretlenmiş dokümanların, spam olma ya da spam olmama hakkında bilgi taşıyıp taşımadığı konusu belirsizdir. Eğer işaretleyiciler ilgisiz ve spam ayrımına dikkat etmiş ise ilgisizler non-spam olarak değerlendirebilir. Ancak işaretleyiciler böyle bir ayrıma dikkat etmeyip spam olarak işaretlenmesi gereken dokümanları ilgisiz olarak işaretlemiş olabilirler. Böylesi bir durum nDCG@*k* ve ERR@*k* gibi metriklerini hesabını değiştirmez. Çünkü bu formüller -2 seviyesini 0'a çevirerek işlem yapar. Aslında -2 seviyesi sentetik olup, başarım hesabında 0 yani ilgisiz olarak ele alınırlar. Spam olarak işaretlenen bir belge zaten örtük olarak ilgisiz demektir: spam(-2) $\subseteq$ ilgisiz (0).

Şekil 6.2, ClueWeb09 veri kümesi içerisindeki ilgisiz dokümanların Fusion, Britney, GroupX ve UK2006 spam sıralamalarına göre dağılımlarını göstermektedir. Her bir yüzdelik puan aralığında ilgisiz doküman-sorgu çiftinin gözlemlendiği görülmüştür. GroupX hariç, spam sıralama yöntemlerinin ilgisiz dokümanları spam değil olarak keskin bir şekilde ayrıştırmadığı gözlemlenmiştir.




**Şekil 6.2.** *ClueWeb09 İlgisiz Dokümanların Yüzdelik Puan Dağılımları*

## 6.4. ClueWeb Korpora Sorgu Analizleri

ClueWeb09 ve ClueWeb12 veri kümelerinin birleşimine ClueWeb korpora adı verilir. ClueWeb korpora için yayınlanan *qrels* dosyalarındaki (sorgu uygunluk yargılarındaki) sorgular (*qid*), sorgulara karşılık döndürülen dokümanlara değerlendiriciler tarafından atanan uygunluk yargıları bazında incelenmiştir.

Tablo 6.7, ClueWeb korporada en çok spam doküman döndüren ilk 10 sorgu ile en az spam doküman döndüren ilk 10 sorguyu ClueWeb09 ve ClueWeb12 veri kümeleri bazında ayrı ayrı göstermektedir. ClueWeb09 veri kümesi için "*madam cj walker*" sorgusu, ClueWeb12 veri kümesi için ise "*install an infrared movement detector*" sorgusu en çok spam doküman döndüren sorgulardır. Diğer taraftan, ClueWeb09 veri kümesi için hiç spam döndürmemiş 5 sorgu ile en az spam doküman döndüren ilk 5 sorgu, ClueWeb12 veri kümesi için ise hiç spam doküman döndürmemiş 50 sorgudan rastgele seçilmiş 10 tanesi Tablo 6.7'de gösterilmiştir.

**Tablo 6.7.** *ClueWeb korporada en çok ve en az spam doküman döndüren sorgular*

| | | ***En Çok SPAM* Doküman Döndüren Sorgular** | ***En Az SPAM* Doküman Döndüren Sorgular** |
|---|---|---|---|
| **ClueWeb09** | 1 | madam cj walker | dangers of asbestos |
| | 2 | credit report | rock art |
| | 3 | income tax return online | ralph owen brewster |
| | 4 | dieting | von willebrand disease |
| | 5 | who invented music | septic system design |
| | 6 | satellite | vanuatu |
| | 7 | last supper painting | korean language |
| | 8 | fybromyalgia | dog clean up bags |
| | 9 | 403b | dnr |
| | 10 | er tv show | porterville |
| **ClueWeb12** | 1 | install an infrared movement detector | how to find the mean |
| | 2 | make a small apartment look larger | traverse city |
| | 3 | sit and reach test at home | view my internet history |
| | 4 | paint home interior walls | nasa interplanetary missions |
| | 5 | atypical squamous cells | benefits of running |
| | 6 | septic system options | kids earth day activities |
| | 7 | how to cook pork tenderloin | balding cure |
| | 8 | buy condo in florida | evidence for evolution |
| | 9 | website design hosting | marshall county schools |
| | 10 | lower heart rate | hurricane Irene flooding in manville nj |

**Tablo 6.8.** *ClueWeb korporada en çok ve en az ilgili doküman döndüren sorgular*

| | | ***En Çok İLGİLİ* Doküman Döndüren Sorgular** | ***En Az İLGİLİ* Doküman Döndüren Sorgular** |
|---|---|---|---|
| **ClueWeb09** | 1 | fact on uranus | kenmore gas water heater |
| | 2 | ct jobs | us capitol map |
| | 3 | continental plates | east ridge high school |
| | 4 | neil young | equal opportunity employer |
| | 5 | quit smoking | source of the nile |
| | 6 | septic system design | tn highway patrol |
| | 7 | dangers of asbestos | pacific northwest laboratory |
| | 8 | keyboard reviews | to be or not to be that is the question |
| | 9 | california franchise tax board | grilling |
| | 10 | tangible personal property tax | culpeper national cemetery |
| **ClueWeb12** | 1 | apply for government job in india | uss carl vinson |
| | 2 | morel mushrooms hunting | black and gold |
| | 3 | setup a garage sale | teddy bears |
| | 4 | become vegan | mister rogers |
| | 5 | bird spotting | home theater systems |
| | 6 | get a divorce in texas | transfer money online united states |
| | 7 | windows 10 troubleshoot | halloween activities for middle school |
| | 8 | educational toys for kids | rain man |
| | 9 | build home theater system | roosevelt island |
| | 10 | deliver a baby at home | lump in throat |

**Tablo 6.9.** *ClueWeb korporada en çok ve en az ilgisiz doküman döndüren sorgular*

| | | ***En Çok İLGİSİZ* Doküman Döndüren Sorgular** | ***En Az İLGİSİZ* Doküman Döndüren Sorgular** |
|---|---|---|---|
| **ClueWeb09** | 1 | us capitol map | fact on uranus |
| | 2 | to be or not to be that is the question | credit report |
| | 3 | pacific northwest laboratory | neil young |
| | 4 | east ridge high school | ct jobs |
| | 5 | mayo clinic jacksonville fl | lymphoma in dogs |
| | 6 | hp mini 2140 | madam cj walker |
| | 7 | source of the nile | continental plates |
| | 8 | universal animal cuts reviews | quit smoking |
| | 9 | angular cheilitis | california franchise tax board |
| | 10 | sherwood regional library | fybromyalgia |
| **ClueWeb12** | 1 | black and gold | apply for government job in india |
| | 2 | uss carl vinson | morel mushrooms hunting |
| | 3 | mister rogers | setup a garage sale |
| | 4 | transfer money online united states | become vegan |
| | 5 | home theater systems | bird spotting |
| | 6 | lump in throat | educational toys for kids |
| | 7 | halloween activities for middle school | build home theater system |
| | 8 | land surveyor | make a teddy bear |
| | 9 | folk remedies sore throat | buy Girl Scout cookies |
| | 10 | the american revolutionary | hair dye |

**Tablo 6.10.** *ClueWeb korporada en çok ve en az doküman döndüren sorgular*

| | | ***En Çok* Doküman Döndüren Sorgular** | ***En Az* Doküman Döndüren Sorgular** |
|---|---|---|---|
| **ClueWeb09** | 1 | memory | angular cheilitis |
| | 2 | to be or not to be that is the question | brooks brothers clearance |
| | 3 | who invented music | hip fractures |
| | 4 | dieting | indiana state fairgrounds |
| | 5 | the wall | lipoma |
| | 6 | president of the united states | porterville |
| | 7 | south africa | gs pay rate |
| | 8 | the sun | pork tenderloin |
| | 9 | joints | churchill downs |
| | 10 | tv on computer | sore throat |
| **ClueWeb12** | 1 | what was the name of elvis presley's home | pole vault |
| | 2 | black and gold | recycling lead acid batteries in new jersey |
| | 3 | nba records | become an allergist |
| | 4 | how to find the mean | braid extensions |
| | 5 | home theater systems | prevent a heart attact |
| | 6 | flowering plants | buttons for websites |
| | 7 | what is madagascar known for | cure indigestion |
| | 8 | rain man | morel mushrooms hunting |
| | 9 | wind power | acid stain concrete |
| | 10 | traverse city | fake tan at home |

Tablo 6.8, ClueWeb korporada en çok ilgili doküman döndüren ilk 10 sorgu ile en az ilgili doküman döndüren ilk 10 sorguyu ClueWeb09 ve ClueWeb12 veri kümeleri bazında ayrı ayrı göstermektedir. ClueWeb09 veri kümesi için "*fact on uranus*" sorgusu, ClueWeb12 veri kümesi için ise "*apply for government job in india*" sorgusu en çok ilgili doküman döndüren sorgulardır. Diğer taraftan, ClueWeb09 veri kümesi için "*kenmore gas water heater"* sorgusu, ClueWeb12 veri kümesi için ise "*uss carl vinson*" sorgusu en az ilgili doküman döndüren sorgular olmuştur.

Tablo 6.9, ClueWeb korporada en çok ilgisiz doküman döndüren ilk 10 sorgu ile en az ilgisiz doküman döndüren ilk 10 sorguyu ClueWeb09 ve ClueWeb12 veri kümeleri bazında ayrı ayrı göstermektedir. ClueWeb09 veri kümesi için "*us capitol map*" sorgusu, ClueWeb12 veri kümesi için ise "*black and gold*" sorgusu en çok ilgisiz doküman döndüren sorgulardır. Diğer taraftan, ClueWeb09 veri kümesi için "*fact on uranus"* sorgusu en az ilgisiz doküman döndüren sorgu olmuştur. ClueWeb12 veri kümesi için ise hiç ilgisiz doküman döndürmemiş 11 sorgu bulunmakta olup bu sorgulardan 10 tanesi Tablo 6.9'de gösterilmiştir.

Tablo 6.10, ClueWeb korporada toplamda en çok doküman döndüren ilk 10 sorgu ile toplamda en az doküman döndüren ilk 10 sorguyu ClueWeb09 ve ClueWeb12 veri

kümeleri bazında ayrı ayrı göstermektedir. ClueWeb09 veri kümesi için "*memory*" sorgusu, ClueWeb12 veri kümesi için ise "*what was the name of elvis presley's home*" sorgusu toplamda en çok doküman döndüren sorgulardır. Diğer taraftan, ClueWeb09 veri kümesi için "*angular cheilitis"* sorgusu, ClueWeb12 veri kümesi için ise "*pole vault*" sorgusu toplamda en az doküman döndüren sorgular olmuştur.

Tablo 6.7, Tablo 6.8, Tablo 6.9 ve Tablo 6.10 tablolarında, ClueWeb09 veri kümesi için bulunan sorgular kendi aralarında karşılaştırmalı olarak analiz edildiğinde ortaya çıkarılan bazı bulgular Tablo 6.11'de gösterilmiştir.

**Tablo 6.11.** *ClueWeb09 veri kümesi için bulunan sorguların karşılaştırmalı analizi*

| **ClueWeb09** | ***En Çok Spam*** | ***En Az Spam*** | ***En Çok İlgili*** | ***En Az İlgili*** | ***En Çok İlgisiz*** | ***En Az İlgisiz*** | ***En Çok*** | ***En Az*** |
|---|---|---|---|---|---|---|---|---|
| *madam cj walker* | ■ | | | | | ■ | | |
| *credit report* | ■ | | | | | ■ | | |
| *dieting* | ■ | | | | | | ■ | |
| *who invented music* | ■ | | | | | | ■ | |
| *fybromyalgia* | ■ | | | | | ■ | | |
| *dangers of asbestos* | | ■ | ■ | | | | | |
| *septic system design* | | ■ | ■ | | | | | |
| *porterville* | | ■ | | | | | | ■ |
| *fact on uranus* | | | ■ | | | ■ | | |
| *ct jobs* | | | ■ | | | ■ | | |
| *continental plates* | | | ■ | | | ■ | | |
| *neil young* | | | ■ | | | ■ | | |
| *quit smoking* | | | ■ | | | ■ | | |
| *california franchise tax board* | | | ■ | | | ■ | | |
| *us capitol map* | | | | ■ | ■ | | | |
| *east ridge high school* | | | | ■ | ■ | | | |
| *source of the nile* | | | | ■ | ■ | | | |
| *pacific northwest laboratory* | | | | ■ | ■ | | | |
| *to be or not to be that is the question* | | | | ■ | ■ | | ■ | |
| *angular cheilitis* | | | | | ■ | | | ■ |

ClueWeb09 veri kümesi için en çok spam doküman döndüren sorgular listesinin ilk sırasında yer alan "*madam cj walker*" sorgusu, aynı zamanda en az ilgisiz doküman döndüren ilk 10 sorgu arasında da yer almaktadır. Benzer şekilde, ClueWeb09 veri kümesi için en çok spam doküman döndüren ilk 10 sorgu arasında yer alan "*fybromyalgia*" sorgusu, en az ilgisiz doküman döndüren ilk 10 sorgu arasında da yer almaktadır.

ClueWeb09 veri kümesi için en az spam doküman döndüren sorgular listesinin ilk sırasında yer alan "*dangers of asbestos"* sorgusu, aynı zamanda en çok ilgili doküman döndüren ilk 10 sorgu arasında da yer almaktadır.

ClueWeb09 veri kümesi için en çok ilgili doküman döndüren sorgular listesinin ilk sırasında yer alan "*fact on uranus*" sorgusu, aynı zamanda en az ilgisiz doküman döndüren sorgular listesinin de ilk sırasında yer aldığı gözlemlenmiştir.

ClueWeb09 veri kümesi için en çok ilgisiz doküman döndüren sorgular listesinin ilk sırasında yer alan "*us capitol map*" sorgusu, aynı zamanda en az ilgili doküman döndüren ilk 10 sorgu arasında da yer almaktadır.

ClueWeb09 veri kümesi için toplamda en az doküman döndüren sorgular listesinin ilk sırasında yer alan "*angular cheilitis*" sorgusu, aynı zamanda en çok ilgisiz doküman döndüren ilk 10 sorgu arasında da yer almaktadır.

ClueWeb09 veri kümesi sorgularından biri olan "*to be or not to be that is the question*" sorgusu çoğunlukla stop-word kelimelerden oluştuğundan zor bir sorgudur. Tablolar incelendiğinde, bu sorgunun en az ilgili doküman, en çok ilgisiz doküman ve toplamda en çok doküman döndüren sorgular listelerinin üçünde de yer aldığı görülür.

Benzer şekilde, Tablo 6.7, Tablo 6.8, Tablo 6.9, ve Tablo 6.10 tablolarında, ClueWeb12 veri kümesi için bulunan sorgular kendi aralarında karşılaştırmalı olarak analiz edildiğinde ortaya çıkarılan bazı bulgular Tablo 6.12'de gösterilmiştir.

ClueWeb12 veri kümesi için en az spam doküman döndüren sorgular listesinin ilk sırasında yer alan "*how to find the mean"* sorgusu, toplamda en çok doküman döndüren ilk 10 sorgu arasında da yer almaktadır.

ClueWeb12 veri kümesi için en çok ilgisiz doküman döndüren sorgular listesinin ilk sırasında yer alan "*black and gold*" sorgusunun, aynı zamanda en az ilgili doküman döndüren sorgular listesinde (ikinci sırada) ve toplamda en çok doküman döndüren sorgular listesinde (ikinci sırada) yer alması çok dikkat çekicidir.

ClueWeb12 veri kümesi için en az ilgisiz doküman döndüren sorgular listede yer alan "*apply for government job in india*" sorgusunun en çok ilgili doküman döndüren sorgular listesinde birinci sırada yer aldığı gözlemlenmiştir.

ClueWeb12 veri kümesi sorgularından biri olan "*morel mushrooms hunting*" sorgusunun en çok ilgili doküman, en az ilgisiz doküman ve toplamda en az doküman döndüren sorgular listelerinin üçünde de yer aldığı görülür. ClueWeb09'daki "*to be or not to be that is the question*" sorgusunun tamamen tersi bir durum elde edilmiştir.

**Tablo 6.12.** *ClueWeb09 veri kümesi için bulunan sorguların karşılaştırmalı analizi*

| **ClueWeb12** | ***En Çok Spam*** | ***En Az Spam*** | ***En Çok İlgili*** | ***En Az İlgili*** | ***En Çok İlgisiz*** | ***En Az İlgisiz*** | ***En Çok*** | ***En Az*** |
|---|---|---|---|---|---|---|---|---|
| *how to find the mean* | | ■ | | | | | ■ | |
| *traverse city* | | ■ | | | | | ■ | |
| *apply for government job in india* | | | ■ | | | ■ | | |
| *morel mushrooms hunting* | | | ■ | | | ■ | | ■ |
| *setup a garage sale* | | | ■ | | | ■ | | |
| *become vegan* | | | ■ | | | ■ | | |
| *bird spotting* | | | ■ | | | ■ | | |
| *educational toys for kids* | | | ■ | | | ■ | | |
| *build home theater system* | | | ■ | | | ■ | | |
| *uss carl vinson* | | | | ■ | ■ | | | |
| *black and gold* | | | | ■ | ■ | | ■ | |
| *black and gold* | | | | ■ | ■ | | | |
| *home theater systems* | | | | ■ | ■ | | ■ | |
| *transfer money online united states* | | | | ■ | ■ | | | |
| *halloween activities for middle school* | | | | ■ | ■ | | | |
| *rain man* | | | | ■ | | | ■ | |
| *lump in throat* | | | | ■ | ■ | | | |

# 7. SONUÇLAR VE ÖNERİLER

Bu çalışmada, 2010-2012 yılları arasında yayınlanan sorgu uygunluk yargıları kullanılarak, ClueWeb09 veri kümesi için 3308 adet spam, 11913 adet non-spam doküman işaretlemesi yapılmışçasına elde edilmiş gerçek doğrular üzerinden UK2006, Britney, GroupX ve Fusion spam sıralamalarının çeşitli eşik değerleri için spam sınıflandırıcı başarımları ölçülmüştür.

Spam dokümanların tespit edilmesinde, spam sıralama yöntemlerinin dördünün de iyi performans gösterdiği görülmüştür. Spam ve non-spam ayrışmasında GroupX sıralamasının diğerlerinden farklı bir trend gösterdiği, spam ve ilgili doküman ayrımını en keskin yapabilen yöntem olduğu gözlemlenmiştir. F1 ve Accuracy ölçütleri dikkate alındığında, Fusion sıralamasının diğerlerine kıyasla belirgin bir şekilde daha iyi performans gösterdiği ortaya koyulmuştur. İkili sınıflandırıcıya dönüştürülen dört spam sıralamasının, F1 ve Accuracy ölçütlerini maksimize eden eşik değerlerinin 5 ile 15 arasında değiştiği saptanmıştır.

Waterloo spam sıralamalarının indirilebildiği Web sayfasında[14] tavsiye edilen %70 eşik değeri ve Fusion sıralaması kullanıldığında, ClueWeb09 veri kümesi için ilgili dokümanların %38'inin kaybedildiği gözlemlenmiştir.

Bu tezin işaret ettiği eşik değeri aralığı ise %10-%15 arasında olup, %70 değerinden oldukça küçüktür. Aradaki bu farklılığın, "intrinsic" değerlendirme yöntemleri ile "extrinsic" değerlendirme yöntemlerinin farklılığından kaynaklı olduğu düşünülmektedir. Bu nedenle, tam teşekküllü BE sistemlerinin ortalama başarımını maksimize eden ("extrinsic") eşik değerinin, spam sınıflandırmasının doğruluğunu maksimize eden ("intrinsic") eşik değerinden yüksek olduğu gözlemlenmiştir.

ClueWeb12 Fusion spam sıralamalarının, spam ve non-spam ayrımı yapan bir ikili sınıflandırıcı olarak oldukça kötü çalıştığı saptanmıştır.

ClueWeb09 veri kümesini, Web spam tespiti araştırma konusu yönünde inceleyen çalışmalar alan yazında mevcut olup, başarılı yöntemler literatüre kazandırılmıştır. Ancak, benzer nitelikte çalışmalar ClueWeb12 veri kümesi için yapılmamıştır.

Bilgi erişimi için oluşturulmuş gerçek doğrular verisinin, spam sınıflandırması problemine nasıl devşirilebileceği bu tezde anlatılmıştır. Söz konusu yöntem kullanılarak oluşturulabilecek gerçek doğrular kümesi aracılığıyla ClueWeb12 veri

[14]https://plg.uwaterloo.ca/~gvcormac/clueweb09spam/

kümesi için başarılı bir spam sınıflandırıcısı eğitilebilir. Genel olarak, Bilgi Erişim sorgu uygunluk yargıları (geçmişteki ya da gelecekteki) Web spam sınıflandırması için bir eğitim verisi olarak kullanılabilir.

Tez kapsamında gerçekleştirilen ClueWeb korpora sorgu analizlerinin genişletilip detaylandırılmasıyla elde edilebilecek bulgular, sorgu zorluk tahmini (*İng.* query difficulty estimation) probleminde yol gösterici olabilir.

## KAYNAKÇA


Al-Kabi, M., Wahsheh, H., Alsmadi, I., Al-Shawakfa, E., Wahbeh, A. ve Al-Hmoud, A. (2012). Content-based analysis to detect Arabic web spam. *Journal of Information Science, 38*(3), 284-296.

Amati, G. ve 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.

Baeza-Yates, R. ve Ribeiro-Neto, B. (1999). *Modern information retrieval*. (Vol. 463): ACM press New York.

Becchetti, L., Castillo, C., Donato, D., Leonardi, S. ve Baeza-Yates, R. (2008). *Web spam detection: Link-based and content-based techniques.* Paper presented at the The European integrated project dynamically evolving, large scale information systems (DELIS): Proceedings of the final workshop.

Belz, A. ve Gatt, A. (2008). *Intrinsic vs. extrinsic evaluation measures for referring expression generation.* Paper presented at the Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers.

Bendersky, M., Fisher, D. ve Croft, W. B. (2010). *Umass at trec 2010 web track: Term dependence, spam filtering and quality bias.* Paper presented at the Proceedings of TREC.

Billerbeck, B., Craswell, N., Fetterly, D. ve Najork, M. (2011). *Microsoft Research at TREC 2011 Web Track.* Paper presented at the TREC.

Brin, S. ve Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems, 30*(1-7), 107-117.

Buckley, C. ve Voorhees, E. M. (2017). *Evaluating evaluation measure stability.* Paper presented at the ACM SIGIR Forum.

Ceri, S., Bozzon, A., Brambilla, M., Della Valle, E., Fraternali, P. ve Quarteroni, S. (2013). *Web information retrieval*. Springer Science & Business Media.

Chapelle, O., Metlzer, D., Zhang, Y. ve Grinspan, P. (2009). *Expected reciprocal rank for graded relevance.* Paper presented at the Proceedings of the 18th ACM conference on Information and knowledge management.

Chellapilla, K. ve Chickering, D. M. (2006). *Improving Cloaking Detection using Search Query Popularity and Monetizability.* Paper presented at the AIRWeb.

Clarke, C. L., Craswell, N., Soboroff, I. ve Cormack, G. V. (2010). *Overview of the TREC 2010 Web Track*. Retrieved from

Clinchant, S. ve Gaussier, E. (2011). Retrieval constraints and word frequency distributions a log-logistic model for IR. *Information Retrieval, 14*(1), 5-25.

Clough, P. ve Sanderson, M. (2013). Evaluating the performance of information retrieval systems using test collections. *Information Research, 18*(2), 18-12.

Collins-Thompson, K., Macdonald, C., Bennett, P., Diaz, F. ve Voorhees, E. M. (2015). *TREC 2014 web track overview*.

Convey, E. (1996). Porn sneaks way back on web. *The Boston Herald, 28*.

Cormack, G. V., Smucker, M. D. ve Clarke, C. L. (2011). Efficient and effective spam filtering and re-ranking for large web datasets. *Information Retrieval, 14*(5), 441-465.

Crane, M. ve Trotman, A. (2012). *Effects of spam removal on search engine efficiency and effectiveness.* Paper presented at the Proceedings of the Seventeenth Australasian Document Computing Symposium.

Craswell, N. ve Hawking, D. (2009). Web information retrieval. *Information Retrieval: searching in the 21st Century*, 85-101.

Croft, W. B., Metzler, D. ve Strohman, T. (2010). *Search engines: Information retrieval in practice*. (Vol. 283): Addison-Wesley Reading.

Greiff, W. R. (1998). *A theory of term weighting based on exploratory data analysis.* Paper presented at the Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval.

Gyöngyi, Z. ve Garcia-Molina, H. (2005). Spam: It's not just for inboxes anymore. *IEEE Computer, 38*(10), 28-34.

Harman, D. (2011). Information retrieval evaluation. *Synthesis Lectures on Information Concepts, Retrieval, and Services, 3*(2), 1-119.

Hauff, C. ve Hiemstra, D. (2009). *University of Twente@ TREC 2009: Indexing half a billion web pages*.

He, J., Balog, K., Hofmann, K., Meij, E., Rijke, M. d., Tsagkias, M. ve Weerkamp, W. (2009). *Heuristic ranking and diversification of web documents*.

Henzinger, M. R., Motwani, R. ve Silverstein, C. (2002). *Challenges in web search engines.* Paper presented at the ACM SIGIR Forum.

Hiemstra, D. (2001). *Using language models for information retrieval*. Taaluitgeverij Neslia Paniculata.

Hiemstra, D. (2009). Information retrieval models. *Information Retrieval: searching in the 21st Century*, 2-19.

Hochstotter, N. ve Koch, M. (2009). Standard parameters for searching behaviour in search engines and their empirical evaluation. *Journal of Information Science, 35*(1), 45-65.

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

Jones, T., Hawking, D. ve Sankaranarayana, R. (2007). *A framework for measuring the impact of web spam.* Paper presented at the Proceedings of 12th Australasian Document Computing Symposium (ADCS).

Jones, T., Sankaranarayana, R., Hawking, D. ve Craswell, N. (2009). *Nullification test collections for web spam and SEO.* Paper presented at the Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web.

Kamps, J., Kaptein, R. ve Koolen, M. (2010). *Using Anchor Text, Spam Filtering and Wikipedia for Web Search and Entity Ranking.* Paper presented at the TREC.

Kaptein, R., Koolen, M. ve Kamps, J. (2009). *Result diversity and entity ranking experiments: Anchors, links, text and Wikipedia*.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM), 46*(5), 604-632.

Kocabaş, İ., Dinçer, B. T. ve Karaoğlan, B. (2014). A nonparametric term weighting method for information retrieval based on measuring the divergence from independence. *Information Retrieval, 17*(2), 153-176.

Krovetz, R. (1993). *Viewing morphology as an inference process.* Paper presented at the Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval.

Lafferty, J. ve Zhai, C. (2001). *Document language models, query models, and risk minimization for information retrieval.* Paper presented at the Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval.

Lancaster, F. W. ve Gallup, E. (1973). *Information retrieval on-line*.

Lin, J., Metzler, D., Elsayed, T. ve Wang, L. (2009). *Of Ivory and Smurfs: Loxodontan MapReduce experiments for web search*.

Liu, B. (2007). *Web data mining: exploring hyperlinks, contents, and usage data*. Springer Science & Business Media.

Liu, T.-Y. (2009). Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval, 3*(3), 225-331.

Macdonald, C., Ounis, I. ve Soboroff, I. (2009). *Is spam an issue for opinionated blog post search?* Paper presented at the Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval.

Macdonald, C., Santos, R. L. ve Ounis, I. (2013). The whens and hows of learning to rank for web search. *Information Retrieval, 16*(5), 584-628.

Macdonald, C., Santos, R. L., Ounis, I. ve He, B. (2013). About learning models with multiple query-dependent features. *ACM Transactions on Information Systems (TOIS), 31*(3), 11.

Manning, C. D., Raghavan, P. ve Schütze, H. (2008). *Introduction to information retrieval*. (Vol. 39): Cambridge University Press.

McCreadie, R., Macdonald, C., Ounis, I., Peng, J. ve Santos, R. L. (2009). *University of glasgow at trec 2009: Experiments with terrier*.

McCreadie, R., Macdonald, C., Santos, R. L. ve Ounis, I. (2011). *University of Glasgow at TREC 2011: Experiments with Terrier in Crowdsourcing, Microblog, and Web Tracks.* Paper presented at the TREC.

Ntoulas, A., Najork, M., Manasse, M. ve Fetterly, D. (2006). *Detecting spam web pages through content analysis.* Paper presented at the Proceedings of the 15th international conference on World Wide Web.

Nunes, S. (2006). State of the art in web information retrieval. *academic*, 1-29.

Ponte, J. M. ve Croft, W. B. (1998). *A language modeling approach to information retrieval.* Paper presented at the Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval.

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

Radlinski, F. (2007). *Addressing malicious noise in clickthrough data.* Paper presented at the Learning to rank for information retrieval workshop at SIGIR.

Robertson, S. (2008). On the history of evaluation in IR. *Journal of Information Science, 34*(4), 439-456.

Robertson, S. ve Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval, 3*(4), 333-389.

Salton, G. (1971). The SMART retrieval system-experiments in automatic document processing. *Englewood Cliffs*.

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

Salton, G. ve McGill, M. J. (1986). Introduction to modern information retrieval.

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

Sanderson, M. (2010). Test collection based evaluation of information retrieval systems. *Foundations and Trends® in Information Retrieval, 4*(4), 247-375.

Silverstein, C., Marais, H., Henzinger, M. ve Moricz, M. (1999). *Analysis of a very large web search engine query log.* Paper presented at the ACm SIGIR Forum.

Smucker, M. D. (2011). *Crowdsourcing with a Crowd of One and Other TREC 2011 Crowdsourcing and Web Track Experiments.* Paper presented at the TREC.

Spirin, N. ve Han, J. (2012). Survey on web spam detection: principles and algorithms. *Acm Sigkdd Explorations Newsletter, 13*(2), 50-64.

Tax, N., Bockting, S. ve Hiemstra, D. (2015). A cross-benchmark comparison of 87 learning to rank methods. *Information Processing & Management, 51*(6), 757-772.

Van Rijsbergen, C. J. (1979). *Information Retrieval*. London, England, Butterworths & Co., Ltd.

Voorhees, E. M. (2001). *The philosophy of information retrieval evaluation.* Paper presented at the Workshop of the cross-language evaluation forum for european languages.

Voorhees, E. M. (2007). TREC: Continuing information retrieval's tradition of experimentation. *Communications of the ACM, 50*(11), 51-54.

Voorhees, E. M. ve Harman, D. K. (2005). *TREC: Experiment and evaluation in information retrieval*. (Vol. 1): MIT press Cambridge.

Xiong, C., Callan, J. ve Liu, T.-Y. (2016). *Bag-of-Entities representation for ranking.* Paper presented at the Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval.

Yilmaz, E., Verma, M., Mehrotra, R., Kanoulas, E., Carterette, B. ve Craswell, N. (2015). *Overview of the TREC 2015 Tasks Track.* Paper presented at the TREC.

Zhai, C. ve Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS), 22*(2), 179-214.

Zhuang, X., Zhu, Y., Chang, C.-C., Peng, Q. ve Khurshid, F. (2017). A unified score propagation model for web spam demotion algorithm. *Information Retrieval Journal, 20*(6), 547-574.

Zuccon, G., Nguyen, A., Leelanupab, T. ve Azzopardi, L. (2011). *Indexing without spam.* Paper presented at the Proceedings of the 16th Australasian Document Computing Symposium.

# EKLER

**EK-1** Deneysel Kurulum Ve Verilerin Hazırlanması

İlk olarak bu tez de kullanılması gereken gerçek kişiler tarafından sorgu ilgi değerlendirmeleri yapılmış veri setlerine ihtiyaç vardı. Bu veri setleri 2010-2016 arasında yayınlanmış Tablo 1'de erişim adresleriyle belirtilmiş 7 farklı yıla ait WebTrack ve Tasks Track veri setleridir.

**Tablo 1.** *TREC Web Track veri setleri*

| | |
|---|---|
| TREC 2010 Web Track | https://trec.nist.gov/data/web/10/10.adhoc-qrels.final |
| TREC 2011 Web Track | https://trec.nist.gov/data/web/11/qrels.adhoc |
| TREC 2012 Web Track | https://trec.nist.gov/data/web/12/qrels.adhoc |
| TREC 2013 Web Track | https://trec.nist.gov/data/web/2013/qrels.adhoc.txt |
| TREC 2014 Web Track | https://trec.nist.gov/data/web/2013/qrels.adhoc.txt |
| TREC 2015 Tasks Track | https://trec.nist.gov/data/tasks/qrels-docs.txt |
| TREC 2016 Tasks Track | https://trec.nist.gov/data/tasks/2016-qrels-docs.txt |

Web Track veri seti her bir satırda bir sorgu ve doküman çifti için yapılan değerlendirmeyi barındırır. Tablo 2'de örneği bulunan satır verileri boşlukla ayrılmış olarak listelenmektedir. Sırasıyla satır içerisinde bulunan değerler şunları ifade eder; sorgu numarası (*qid*), 0 kullanılmayan değer, doküman anahtarı (*docid*), ve ilgi yargı değeri (judgment).

**Tablo 2.** *Web Track örnek veri seti*

```
51 0 clueweb09-en0001-74-30054 0
51 0 clueweb09-en0001-80-06127 0
51 0 clueweb09-en0001-80-26227 1
51 0 clueweb09-en0001-84-19365 -2
51 0 clueweb09-en0001-90-05710 0
```

2010 ile 2012 arasında düzenlenen WebTrack yarışmaları ClueWeb09 veri kümesini kullanır. [http://trec.nist.gov/data/webmain.html].

İlki 2015 yılında başlanan Tasks Track yarışmaları ise ClueWeb12 veri kümesini kullanır. [https://trec.nist.gov/data/tasks.html]. Bu Tasks Tracks yarışmalarına ait *qrels* dosyaları WebTrack yarışmalarınınkine benzer bir şekilde her bir satırda bir ilgi değer yargısını vermektedir. Tablo 3'da örneği bulunan satır verileri boşlukla ayrılmış olarak listelenmektedir. Sırasıyla satır içerisinde bulunan değerler şunları ifade eder; görev

numarası (*qid*), alt görev numarası, doküman anahtarı (*docid*), ilgi yargı değeri (judgment), ve son kolondan oluşur.

**Tablo 3.** *Tasks Track örnek veri seti*

```
1 1 clueweb12-0005wb-19-23087 0 0
1 2 clueweb12-0005wb-19-23087 0 0
1 3 clueweb12-0005wb-19-23087 0 0
1 4 clueweb12-0005wb-19-23087 0 0
1 5 clueweb12-0005wb-19-23087 1 1
1 6 clueweb12-0005wb-19-23087 0 0
1 7 clueweb12-0005wb-19-23087 0 0
1 8 clueweb12-0005wb-19-23087 0 0
1 9 clueweb12-0005wb-19-23087 0 0
1 10 clueweb12-0005wb-19-23087 0 0
1 11 clueweb12-0005wb-19-23087 0 0
1 12 clueweb12-0005wb-19-23087 0 0
```

Web Track ve Tasks Track yarışmalarına ait *qrels* dosyalarını daha iyi işleyebilmek için ortak bir veri yapısı kurgulandı ve bu yapı MySQL veri tabanı üzerinde yaratıldı. Aynı veri yapısında ClueWeb09 (Tablo 4) ve ClueWeb12 (

Tablo 5) veri setlerinde yapılan değerlendirmeleri ayırmak için iki tablo oluşturuldu. ClueWeb09 verileri için 4 farklı metot (Fusion, Britney, GroupX, UK2006) ile değerlendirme yapıldığı için her bir metoda ait yüzdelik puanların yer alması deney için daha doğru olduğunu için clueweb09 tablosunda her biri için ayrı alanlar açıldı. ClueWeb12 veri seti için tek bir metot (Fusion) kullanılarak yüzdelik puan oluşturulduğu için bu yüzdelik puanı tutabilmek adına bir adet alan açıldı.

**Tablo 4.** *ClueWeb09 için tablo yapısı*

| **ClueWeb09** | |
|---|---|
| **jyear** | |
| **qid** | PK |
| **docid** | PK |
| judgment | |
| fusion | |
| britney | |
| groupx | |
| uk2006 | |

**Tablo 5.** *ClueWeb12 için tablo yapısı*

| **ClueWeb12** | |
|---|---|
| **jyear** | |
| **qid** | PK |
| **docid** | PK |
| judgment | |
| fusion | |

ClueWeb09 veri seti üzerinde 2010 ve 2012 arasında yapılan WebTrack sorgu ilgi değerlendirmelerini yıllara göre ayrılarak veri tabanına aktarıldı. ClueWeb09 veri seti için sorgu ilgi değerlendirmesi yapılmış toplam 59769 tekil dokümana ait 4 farklı metot ile oluşturulmuş yüzdelik puanlarına ihtiyaç vardı. Bu yüzdelik puanlara erişmek için metotlara ait çıktı dosyalarından faydalanıldı. Dosyaların erişim adresleri aşağıda listelenmiştir.

- ClueWeb09 Fusion değerlendirmesi:
  http://cormack.uwaterloo.ca/clueweb09spam/clueweb09spam.Fusion.bz2
- ClueWeb09 Britney değerlendirmesi:
  http://cormack.uwaterloo.ca/clueweb09spam/clueweb09spam.Britney.bz2
- ClueWeb09 GroupX değerlendirmesi:
  http://cormack.uwaterloo.ca/clueweb09spam/clueweb09spam.GroupX.bz2
- ClueWeb09 UK2006 değerlendirmesi:
  http://cormack.uwaterloo.ca/clueweb09spam/clueweb09spam.UK2006.bz2

"Waterloo Spam Rankings for the ClueWeb09 Dataset" sayfasında (https://plg.uwaterloo.ca/~gvcormac/clueweb09spam/) metotlara ait yüzdelik değerlendirme dosyalarının sıkıştırma ve geri açma yöntemleri açık olarak belirtilmiştir. Bu yöntemleri kullanılarak daha önce sorgu ilgi değerlendirmeleri yapılmış dokümanlara ait yüzdelik puanları veri tabanında ilgili alanlara işlendi.

ClueWeb12 veri seti kullanılarak hazırlanan WebTrack 2013 ve 2014 sorgu ilgi değerlerini veri tabanına aktarıldı. 2015 ve 2016 yıllarına ait Tasks Track yarışmalarına ait *qrels* verileri WebTrack de olduğu gibi tekil sorgu, doküman eşleşmesinden oluşmadığı için Tasks Track verilerini tekil sorgu, doküman eşleşmesi şekline getirilmesi gerekiyordu. Veriler incelediğinde, her bir görev için farklı sayıda alt görev bu görevlere ait ilgi değerlendirmeleri bulunduğu görülmektedir. Herhangi bir ana görevin herhangi bir alt görevinde bir doküman spam (-2) olarak işaretlenmiş ise o dokümanın bütün alt görevlerinde de spam (-2) olarak işaretlendiği görüldü ve bu tip

ana görev doküman eşleşmelerine yargı değeri -2 yani spam olarak veri tabanına aktarıldı. Yargı değeri 0 ve daha büyük olan değerlendirmelerde ana görevin alt görevlerine ait yargı değerlerinden en büyük olanı yargı değeri olarak kabul edildi. Böylece ortak bir sorgu, doküman, ilgi değeri yapısına ulaşıldı. Tablo 6 ve Tablo 7'de spam ve ilgili doküman örnekleri verilmiştir.

**Tablo 6.** *Tasks Track 2015'e ait SPAM doküman örneği*

```
2 1 clueweb12-1703wb-30-15077 -2 0
2 2 clueweb12-1703wb-30-15077 -2 0
2 3 clueweb12-1703wb-30-15077 -2 0
2 4 clueweb12-1703wb-30-15077 -2 0
2 5 clueweb12-1703wb-30-15077 -2 0
2 6 clueweb12-1703wb-30-15077 -2 0
2 7 clueweb12-1703wb-30-15077 -2 0
2 8 clueweb12-1703wb-30-15077 -2 0
2 9 clueweb12-1703wb-30-15077 -2 0
2 10 clueweb12-1703wb-30-15077 -2 0
2 11 clueweb12-1703wb-30-15077 -2 0
2 12 clueweb12-1703wb-30-15077 -2 0
2 13 clueweb12-1703wb-30-15077 -2 0
2 14 clueweb12-1703wb-30-15077 -2 0
```

**Tablo 7.** *Tasks Track 2015'e ait ilgili doküman örneği*

```
16 1 clueweb12-0709wb-57-25504 0 0
16 2 clueweb12-0709wb-57-25504 1 1
16 3 clueweb12-0709wb-57-25504 2 1
16 4 clueweb12-0709wb-57-25504 0 0
16 5 clueweb12-0709wb-57-25504 0 0
16 6 clueweb12-0709wb-57-25504 0 0
```

ClueWeb12 için WebTrack 2013 ve 2014'e ait 28746 ve Tasks Track 2015 ve 2016 için 6241, toplamda 34557 dokümanı veri tabanına aktarıldı. ClueWeb12 veri seti için sadece "Fusion" metodu kullanılarak oluşturulan yüzdelik puan bulunmaktadır. Yüzdelik puan verilerini "Waterloo Spam Rankings for the ClueWeb12 Dataset" (https://www.mansci.uwaterloo.ca/~msmucker/cw12spam/) sayfasından indirildi ve sayfada anlatılan sıkıştırma ve açma metotlarını kullanarak ihtiyaç olan dokümanlara ait yüzdelik puan bilgilerini veri tabanına aktarıldı.

Veri tabanı üzerinde oluşturulan tablolar SQL dili kullanarak yorumlanmaya çalışıldı. İncelemelerde kullanılan SQL sorgu örneklerinden bazıları aşağıda yer almaktadır.

**EK-2** ClueWeb09 veri kümesi kullanılarak yargılanan sorgu doküman çiftlerinin Fusion sıralama puanlarına göre spam ya da ilgili olanlarının sayısını veren SQL cümlesi.

```
SELECT s.fusion,s.spamcount,r.relevantcount
FROM
(
        SELECT fusion, count(1) as spamcount
        FROM clueweb09qrels
        WHERE judgment =-2
        GROUP BY fusion
        ORDER BY fusion
 ) s
 ,
 (
        SELECT fusion, count(1) as relevantcount
        FROM clueweb09qrels
        WHERE judgment >0
        GROUP BY fusion
        ORDER BY fusion
 ) r
WHERE s.fusion = r.fusion
ORDER BY s.fusion;
```

**EK-3** ClueWeb09 veri kümesi kullanılarak yargılanan sorgu doküman çiftleri içerisinde birden çok kez yargılanan dokümanların her bir yargı değeri için yargı değeri ve yargılanma sayısını veren SQL sorgusu.

```
SELECT z.docid,z.judgment,count(*) zcnt
FROM clueweb09qrels z
WHERE z.docid IN
(
    SELECT y.docid
    FROM(
        SELECT x.docid
        FROM clueweb09qrels x
        WHERE docid IN (
                        SELECT t.docid
                        FROM (
                            SELECT  qid,docid,count(1) as say
                            FROM ClueWeb09qrels
                            GROUP BY  qid,docid
                         ) t
                        GROUP BY  t.docid
                        HAVING (count(*) >1)
                        )
        GROUP BY  x.docid,x.judgment
    ) y
    GROUP BY y.docid
    HAVING count(*) >1
)
GROUP BY z.docid,z.judgment;
```