PAMUKKALE ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ

ARAMA MOTORU BİNG'İN GÖVDELEME YÖNTEMLERİNİN ARAŞTIRILMASI

YÜKSEK LİSANS
Fatmana ŞENTÜRK

Anabilim Dalı : Bilgisayar Mühendisliği

Programı : Bilgisayar Mühendisliği

Tez Danışmanı: Yrd. Doç. Dr. Gürhan GÜNDÜZ

TEMMUZ 2012

YÜKSEK LİSANS TEZ ONAY FORMU

Pamukkale Üniversitesi Fen Bilimleri Enstitüsü 081281004 nolu öğrecisi Fatmana ŞENTÜRK tarafından hazırlanan "**ARAMA MOTORU BİNG'İN GÖVDELEME YÖNTEMLERİNİN ARAŞTIRILMASI**" başlıklı tez tarafımızdan okunmuş, kapsamı ve niteliği açısından bir Yüksek Lisans tezi olarak kabul edilmiştir.

Tez Danışmanı : (Jüri Başkanı)	**Yrd. doç. Dr. Gürhan GÜNDÜZ (PAÜ)**
Jüri Üyesi :	**Doç. Dr. Serdar İPLİKÇİ (PAÜ)**
Jüri Üyesi :	**Yrd. Doç. Dr. Emre ÇOMAK (PAÜ)**

Pamukkale Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu'nun 19/07/2012 tarih ve 18/19 sayılı kararıyla onaylanmıştır.

Fen Bilimleri Enstitüsü Müdürü
Prof. Dr. Nuri KOLSUZ

Bu tezin tasarımı, hazırlanması, yürütülmesi, araştırmalarının yapılması ve bulgularının analizlerinde bilimsel etiğe ve akademik kurallara özenle riayet edildiğini; bu çalışmanın doğrudan birincil ürünü olmayan bulguların, verilerin ve materyallerin bilimsel etiğe uygun olarak kaynak gösterildiğini ve alıntı yapılan çalışmalara atfedildiğine beyan ederim.

İmza : 

Öğrenci Adı Soyadı : Fatmana ŞENTÜRK

ÖNSÖZ

Bu çalışmada, arama motoru Bing üzerinde arama yapılırken istenilen bilgiye doğru ve en hızlı bir biçimde nasıl ulaşılabileceği araştırılmıştır.Bu kapsamda öncelikli olarak, İngilizce'de morfolojik açıdan belirli özelliklere sahip benzer kelimeler gruplanmış olup tekil-çoğul, birleşik, son ek alan kelimeler olmak üzere üç grupta toplanmıştır. Daha sonra her bir kelime grubu için farklı algoritmalar üretilerek, oluşturulan sorgular Bing'e gönderilmiş, dönen sonuçlara göre yorumlamalar yapılmıştır. Bu çalışmanın gerçeklenmesinde katkıda bulunan danışman hocam Gürhan GÜNDÜZ'e, kelime seçimlerimde katkısından dolayı Ceyhun Özkal'a ve benden hiçbir zaman desteklerini esirgemeyen arkadaşlarıma ve aileme teşekkür ederim.

Temmuz 2012

Fatmana Şentürk
(Bilgisayar Mühendisi)

İÇİNDEKİLER

Sayfa

KISALTMALAR

SQL	**:** Structured Query Language
FTP	**:** File Transfer Protocol
Veronica	**:** Very Easy Rodent-Oriented Net-wide Indeks to Computerized Archives
Jughead	**:** Jonzy's Universal Gopher Hierarchy Excavation and Display
ÜBS	**:** Üç Basamaklı Sayı

TABLO LİSTESİ

Tablolar

ŞEKİL LİSTESİ

Şekiller

ÖZET

ARAMA MOTORU BİNG'İN GÖVDELEME YÖNTEMLERİNİN ARAŞTIRILMASI

Günümüzde, internet dünyasının gelişimi ile bilgiye ulaşmak hızlı ve basit bir hal almıştır. Bu yoğunluk içerisinde aranılan bilgiye ulaşmak önem kazanmıştır. Bu amaç için arama motorları ortaya çıkmış ve her geçen gün daha da önem kazanmıştır. Birçok insan, arama motorlarını kullanırken istediği bilgiye ulaşmak için yanlış kelime seçimlerinde bulunarak hatalı sonuçlar elde etmekte ve zaman kaybetmektedir. Bu çalışma ile son kullanıcılar açısından doğru bilgiye hızlı ve etkili ulaşması sağlanmakta, araştırmacılar açısından ise arama motorlarının ürettiği sonuçlar yorumlanabilir hale gelmiştir.

Bu çalışmada, arama motoru Bing'in İngilizce kelime gövdeleme algoritması tahminlenmeye çalışılmıştır. Bunun için öncelikli olarak incelenecek örnek kelime grupları belirlenmiş olup, üç grupta toplanmıştır. Bunlar tekil çoğul kelime grubu, bileşik kelime grubu ve son ek alan düzenli fiiller kelime grubudur. Her bir gruba göre sorgular oluşturulup, Bing'e gönderilmiştir. Bing'den dönen sonuçlar analiz edilmiş ve sonuçlar oluşturulmuştur. Bu sonuçlara göre, Bing kelime bazlı ve döküman bazlı indekslemeler yapmaktadır. Bu işlem gereği gövdeleme işlemi de döküman indekslenirken yapılmaktadır. Test sonuçlarına göre, kelimenin tekil çoğul formu gibi güçlü ilişkiler barındıran kelime yapılarında indeksleme oranı daha fazladır.

Anahtar Kelimeler: Bing, Arama Motoru,Kelime Gövdeleme

SUMMARY

INVESTIGATION OF BING STEMMING MECHANISMS

Accessing information has become quick and simple with the development of the internet. Therefore, internet is an important tool to reach desired information. For this purpose, the search engines have emerged and become more important with each passing day. Most people use search engines for this purpose, some of them selected wrong words to reach information and waste their time. This work ways of accessing the right information to users and understood search engine stemming algorithm to developers.

In this work, we tried to estimate Bing's stemming algorithm for English words. For this purpose; we have selected tree different word groups. These are, singular plural words, combined words and regular verbs with postfixes. We have created queries and run them on Bing.We have analysed the returned results from Bing. According to these test results, Bing indexing algortihm based on words and documents. Stemming is processed while document is indexed. Rate of word indexing structures containing strong relationships, such as singular plural form of words, is higher than others.

Key Words: Bing, Search Engine, Stemming

1. GİRİŞ

Giderek bilgi boyutlarının artması, arama motorlarının ortaya çıkmasına sebep olmuştur. Arama motorlarının kullanımı ile birlikte aranılan bilginin çeşitliliği bireylere göre değişmiş, kendi içlerinde bir ayrım yaratma ihtiyacı olmuştur. Günümüzde ise, arama motorları kullanıcıların aradıkları bilgiye hızlı ve kaliteli çözümler üreten arayüzler haline gelmiştir.

Ancak bilginin çokluğu ve kişilerin arayacakları bilgi hakında tutarlı veriler kullanmaması bazı durumlarda yanlış veya eksik sonuçlara yönlendirmektedir. Kullanıcıların zaman kaybının engellenmesi ve kaliteli bilgiye en hızlı sürede erişilmesi önem teşkil etmektedir.

1.1 Tezin Amacı

Bir arama motorunda son kullanıcılar arama yaptığı zaman binlerce farklı sonuçla karşılaşmaktadırlar. Bu gelen sonuçlar içerisinde aranılan bilgiyi kapsamayan, eksik ya da hatalı bilgiler de bulunabilmektedir. Son kullanıcılar açısından, arama kriterlerinin belirlenmesi ve bulunan sonuçların değerlendirilmesi büyük önem taşımaktadır. Ayrıca arama motoru geliştiricek kişiler de bu arama sonuçlarına göre nasıl bir indeksleme yolu çizmesi gerektiğini tespit edebilirler. Bu kapsamda yapılan çalışmada, arama motoru Bing örnek seçilmiş ve arama yaparken girilen ifadenin arama yapılırken hangi işlemlerden geçtiğinin ve döndürülen dökümanların neye göre indekslendiğinin araştırılmasıdır.

1.2 Literatür Özeti

Bugüne kadar, arama motorlarının kapsamı ve içeriği hakkında bir çok çalışma yapılmıştır. Tez kapsamında, literatür kısmı iki gruptan oluşmaktadır.

İlk grup kelime gövdeleme algoritmalarıdır. Kelime gövdeleme, bir kelimenin eklerinin silinmesi ve kelime kökünün bulunmasından oluşmaktadır. Bu algoritmalardan ilki 1968 de Lovins tarafından yayınlanan algoritmadır [1]. Günümüzde ise en çok kullanılanı Porter gövdeleme algoritmasıdır [2]. Kelime gövdeleme algoritmaları ile ilgili detaylı bilgi üçüncü bölümde verilmiştir.

İkinci grup ise arama motorlarının yapısıyla ilgili çalışmalardır. Arama motorları insanların arama alışkanlıkları hakkında bilgi verebilmektedir. Bu yüzden, bir çok çalışmada veri toplamak için kullanılmışlardır [3,4]. Ayrıca kişilerin arama alışkanlıkları farklı bölgelere göre analiz edilmiştir. Örneğin, 2010 yılı içersinde Google arama verileri üzerinden Arap ülkelerinin arama alışkanlıkları incelenmiştir [6].

Arama yaparken bir çok kişi tek bir kelimeye ya da sıralı bir kelime grubuna bağlı kalmakta olup, aranan bilginin aynı olmasına rağmen farklı arama sorguları oluşturulabilmektedir. Bu yüzden, daha önce kullanıcıların oluşturduğu özel sorguları öneren bir uygulama geliştirilmiştir [11]. Ayrıca belirli bir kelime grubu için düzenli sırada arama yapılması ile farklı kombinasyonları ile arama yapılması sonucu dönen dökümanlar farklı olabilmektedir. Bu farkı ortadan kaldırmak adına arama yapılırken oluşturulabilecek tüm kombinasyonlara göre arama yapan bir arayüz geliştirilmiştir [12].

Bir çok çalışmada, arama motorlarının güncelliği ve güvenilirliği ele alınmıştır. Bu çalışmalardan bir tanesinde, arama motorlarının ilk sıralarda gelen sonuçların güncellenme sıklığı ele alınmıştır [13]. Örneğin yapılan bir çalışmada, belirli bir hastalığa yönelik kelimeler aranmış ve sonuçların kalitesi incelenmiştir [15].

Arama motorları arama yapıldığında dönüş hızları açısından da değerlendirmeye alınmıştır. Bunun için büyük arama motorlarına belirli sorgular gönderilmiş ve dönüş süreleri değerlendirilmiştir [14].

L. Vanguhan ve M. Thelwall'ın 2004 yılında yaptığı çalışmada, Google arama motorunun hit sayısı kullanılarak çeşitli web sayfaları için indekslenmiş döküman sayısı bulunmuştur [5]. Bir başka çalışmada ise, dağıtık arama motoru yapısı ile bulut teknolojisi entegre edilmiştir [7].

Bir çok çalışma incelendiğinde tez kapsamında yapılan çalışmaya benzer bir çalışma görülmüştür. Bu çalışmada ise, arama motoru Google'ın kelime gövdeleme algoritması tahminlenmeye çalışılmıştır [8].

Tez kapsamında, ilk olarak çalışılacak kelimeler belirlenmiştir. Daha sonra kelimelere uygun sorgular oluşturulup, arama motoru Bing'e gönderilmiştir. Dönen sonuçlar değerlendirilmiştir.

2. ARAMA MOTORLARI

2.1 Arama Motoru

Arama motoru, dünyadaki hemen hemen tüm web sayfalarının listelendiği, kategorilere ayrılmış, aradığımız bilgilere en kısa yoldan ve hızlı bir şekilde ulaşmamızı sağlayan web sayfalarıdır.

Arama motorları kullanıcılara; web sayfaları, bilgiler, resimler ve diğer dosya türlerini sunmaktadır.

İnternet üzerinde yüzlerce hatta binlerce arama motoru bulunmaktadır. Bunların bir kısmı kendi alanlarındaki web sayfalarını listelemekte, bir kısmıda yerel alanlarda hizmet vermekte, bir kısmı da dünya üzerindeki her türlü web sayfasından topladığı bilgileri getirmektedir.

2.2 Arama Motorlarının Tarihçesi

İlk arama motoru 1990 yılında Alan Emtage tarafından Archie adıyla kurulmuştur. Archie, bir üniversitedeki bazı birimlerde bulunan bilgisayarlar arası FTP(File Transfer Protocol) arşivlerini taramakta ve bu arşivlerdeki dosyalardan bir arama listesi oluşturmaktaydı. Arama yapan kişiler, dosya ismini tam olarak yazdıklarında FTP listelerinde bulunan sonuçlar görüntülenmekteydi.

1991 de Gopher'ın yükselişi Veronica ve Jughead adında iki arama motorunun gelişimine sebep oldu. Veronica ve Jughead, Archie gibi dosya adlarında Gopher da indeksli bir şekilde tutulan dosya isimlerinde arama yapılabiliyordu. Veronica(Very Easy Rodent-Oriented Net-wide Indeks to Computerized Archives) Gopher listelerindeki dosya isimlerini, Jughead(Jonzy's Universal Gopher Hierarchy Excavation and Display) ise Gopher sunucularındaki menü bilgilerini sunan bir araçtır.

1993 yazında halen tam anlamıyla bir arama motoru bulunmamaktaydı, ancak el ile yapılmış bir çok özel sistem bulunmaktaydı. Oscar Neirstrasz, periyodik olarak sayfaları kopyalayan ve arama motorlarının temelini oluşturan standart bir format haline getiren bir perl kodu geliştirdi ve ilk arama motoru 2 Eylül 1993 de yayınlandı.

Haziran 1993'de Matthew Gray perl tabanlı world wide web gezginini, muhtemelen ilk web robotunu, üretti ve Wandex adı verilen bir indeks sistemi oluşturup kullandı. Bu gezginin amacı 1995 sonuna kadar world wide web'in boyutlarını ölçmekti.

Kasım 1993'de web'in ikinci arama motoru Aliweb ortaya çıktı. Aliweb web robotu kullanmamıştır, ancak bunun yerine web sayfası yöneticileri tarafından düzenlenen özel formattaki indeks dosyalarına bağlı olmuştur.

Aralık 1993'de, JumpStation web sayfalarını bulmak ve indeks oluşturmak amacıyla bir web robotu ve sorguları çalıştırmak için bir web form arayüzü kullanılmıştır. Bir arama motorunun kullandığı üç temel özelliği (tarama, indeksleme ve arama) birleştiren bir araç olmuştur.

1994'e gelindiğinde, tam metin tarayıcı özelliğe sahip ilk arama motorlarından biri WebCrawler ortaya çıkmıştır. Öncekilerin aksine, her hangi bir web sayfasında herhangi bir kelime aranabilir hale gelmiştir. O zamandan sonra bu özellik bütün büyük arama motorlarında bulunmaktadır.

1996 yılında, Netscape kendi web arayüzü üzerinde arama motoru hizmeti vermek amacıyla anlaşma sağlamak için arayışa girmitir. Büyük arama motorlarıyla yıllık beş milyon dolar ile anlaşınca çok büyük ilgi görmüş ve bu arama motorları Netscape'in arama sayfasında dönüşümlü bir şekilde görüntülenmiştir. Bu arama motorları, Yahoo,Magellan, Lycos, Infoseek ve Excite'dır.

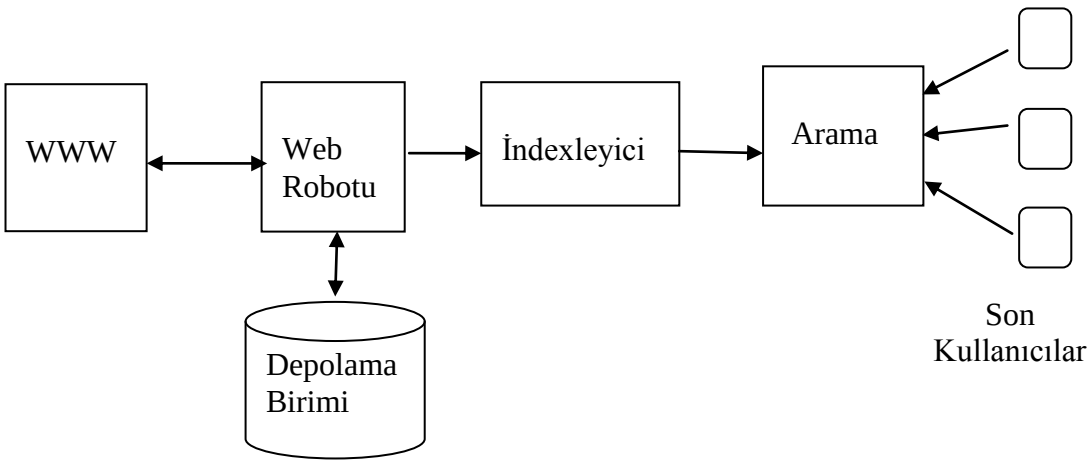
1990'lı yılların sonlarında arama motorları büyük bir yatırım kaynağı haline gelmiştir. Bazı şirketler halka arz edilerek büyük gelirler elde etmiştir. Bazı arama motorları da Northern Light gibi, sadece kurumsal bazda sürümler çıkartmıştır. Bir çok arama motoru 1999 yılında zirve yapmış, bu ekonomik hareketlilik 2001 yılında sona ermiştir.

2000'li yıllara gelindiğinde ise, Google arama motoru göze çarpmıştır. Şirket PageRank adını verdikleri algoritma ile birçok arama için daha iyi sonuçlar elde etmiştir. Bu algoritma, her bir web sayfasına ve bağlantılı linklere iteratif olarak bir PageRank değeri verilmiştir. Dolayısıyla bir arama yapıldığı zaman istenen sayfalardan daha fazla ilgili sayfa listelenmiştir. Google, rakiplerinin gömülü web portallarının aksine kullanıcılara basit bir arama sayfası sunmuştur.

Microsoft 1998 yılı sonbaharında MSN Search adlı arama motorunu ilan etmiştir. 2004 yılında msnbot adı verilen bir web robotu geliştirmiştir. 1 Haziran 2009' da, Microsoft arama motorunun ismini Bing olarak değiştirmiştir.

2.3 Arama Motorlarının Çalışma Şekli

Arama motorları, hizmet verdikleri arayüz üzerinde herhangi bir kelimeyi yazıp search, find, ..vs gibi komutlara tıklandığı zaman istenilen kelimenin geçtiği dökümanları listelemektedirler. Bu işlem arka planda şu şekilde gerçekleşmektedir: web robotları otomatik olarak sunucuları tarayıp bilgileri getirmekte, arama motoruna gelen veriler belirli bir alanda tutulmakta, son kullancılarda tutulan veriler üzerinden arama yapabilmektedirler. Şekil 2.1'de bir arama motorunun genel olarak çalışma şekli verilmiştir [17].



Şekil 2.1 : Bir arama motorunun genel gösterimi.

2.4 Arama Motoru Bileşenleri

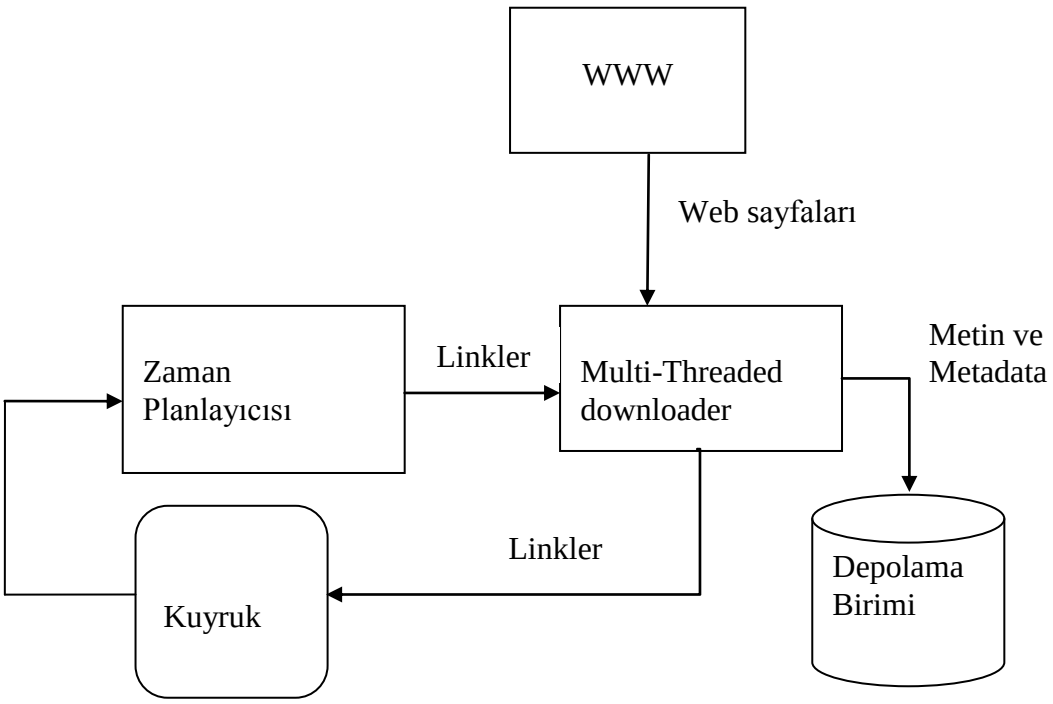
Bir arama motoru web robotu, indeksleyici ve arama olmak üzere üç bileşenden oluşmaktadır[16].

2.4.1 Web robotu

İnternet üzerinde arama yapıp veri toplayan modüllere web robotu (web crawler) denir. Özellikle güncel bilgilere ulaşmak için bir arama motorunun en önemli bileşenlerinden biridir.

Arama motoru üzerinde yapılacak olan arama sonuçlarını hızlı bir şekilde döndürmek için web robotları tarafından getirilen sayfaları kullanmaktadır. Ayrıca arama motoruna daha önce getirilen sayfaları da dolaşarak, hem link kontrolü yapmakta hem de bilgilerin tutarlı olmasını sağlamaktadır.

Web robotlarının; hangi sayfaların indirileceği, hangi sayfaların ne sıklıkla ziyaret edileceği, web sunucularının aşırı yüklenmesinin önlenmesi ve dağıtık sistemlerde nasıl davranması gerektiğinin belirlenmesi gereklidir.



Şekil 2.2 : Standart bir web robotu mimarisi.

İnternet üzerinde her geçen gün bilgi paylaşımının arttığını ve sayfa sayısının arttığı göz önünde bulundurulduğu zaman linklerin belirlenmesi zor bir işlemdir. Sayfaların belirlenmesi için tüm sunucuların gezilmesi ve sayfalar hakkında fikir edinmek için belirli kısımlarının getirilmesi gerekmektedir. Bunun için farklı algoritmalar geliştirilmiştir.

Cho et al. internet tarama politikaları üzerine çalışan ilk kişidir. Standford.edu uzantılı 180.000 link üzerinde farklı tarama teknikleri kullanarak test etmiştir[9]. Najork ve Wiener, breadth-first yöntemini kullanarak 328 milyon sayfa üzerinde tarama işlemi gerçekleştirmişlerdir[10].

Web sayfaları doğası gereği dinamik bir yapıya sahiptir ve bir web robotunun tarama işlemi günlerce ve hatta haftalarca sürebilmektedir. Tarama işlemi bittiğinde veriler değişmiş olabilir veya o link kaldırılmış olabilmektedir. Verilerin güncellenmemiş olması ve sunulan bilgilerin yanlış olması arama motorları açısından olumsuz bir etkidir.

Bir web robotunun sürekli olarak sunuculardan istekte bulunması, sunucuları yavaşlatacak olup, sunucun diğer uygulamalardan gelen isteklere cevap veremez hale gelmesine sebep olacaktır. Dolayısıyla yapılan taleplerin sayısına yönelik kısıtlamalara gidilmesi ve bir çeşit kontrol mekanizması gerekmektedir. Birbirini takip eden iki talep arasına belirli bir süre koyulması çözüm yöntemlerinden biridir. Ancak bu zaman aralığı, performansı düşürmemek için her sunucu için ayrı belirlenmelidir.

Dağıtık sistemler üzerinde hizmet veren bir sunucudan bilgi toplamak, karmaşık olabilmektedir. Genel olarak IP adresi ya da alan adı bazında taramalar yapılmaktadır. Her iki yöntemin uygulanması sonucu ortaya çıkan sonuçlar mükemmel olmasa da, sunucular üzerindeki iş yoğunluğunu belli bir seviyede tutabilmektedirler.

Çalışma açısından web robotu türleri şu şekildedir.

2.4.1.1 Geniş kapsamlı web robotları

Web sayfalarından toplanan sayfalarının ve kaynakların bütünlüğü kadar web sitelerinin sayısının da önemli olduğu yüksek bant genişliği gerektiren web robotlarıdır. En geniş kapsamlı tarama; zaman, kaynaklar ve bant genişliğinin yeterli olduğu durumlarda gerçekleşir. Tarayabildikleri kadar veriyi tarayıp kendi depolarına kopyalarlar.

2.4.1.2 Odaklanmış web robotları

Kalite kriterinin önemli olduğu, seçilmiş bazı özel konular ve web sayfalarında tarama yapan küçük ve orta ölçekli web robotlarıdır. Sadece belirli konuları kapsarlar. Veri yoğunluğu diğer arama robotlarına göre azdır.

2.4.1.3 Sürekli web robotları

Geleneksel olak web robotları ilgilenilen konulardaki kaynakları bir defa dolaşıp kaydeder. Daha sonra belirli sürelerde, önceden indirilmiş sayfaları tarayarak değişiklikleri ya da yeni eklemeleri günceller.

2.4.1.4 Deneysel web robotları

Farklı protokoller ile çeşitli algoritmalar kullanarak yapılabilecek bir tarama türüdür.

2.4.2 İndexleyici

Web robotunun topladığı verileri ya da dökümanları bellirli algoritmalar kullanarak, aranabilir indeksler haline dönüştüren modüldür.

Web arama motorları her sayfanın içeriğini indekslemek zorundadırlar. Kopyalanan bir web sayfasındaki içerikleri, başlıklıkları ya da etiketlerdeki bilgilerin tamamını indeksler ve indekslenmiş verileri veritabanında saklamaktadırlar. İndexlemedeki amaç her hangi bir veri arandığı zaman sonucunu en hızlı biçimde kullanıcıya döndürmektir.

Google gibi bazı arama motorları indekslerin yanısıra kopyaladığı sayfaların tamamını saklarken, Alta Vista gibi arama motorları ise sayfalardaki her bir kelimeyi saklamaktadırlar. Ters listeler, vektör uzayları, son ek yapıları ve hibrit yapılar yaygın kullanılan çeşitleridir.

2.4.3 Arama

Arama modülü indekslenmiş veriler üzerinde çalışan son kullanıcı ile etkileşim içerisinde olan modüldür. İndexleyici üzerinde gelen sorguları çalıştırır ve hangi kullanıcı arama yaptıysa o kişiye sonuçları döndürür.

Bir kullanıcı arama motoruna aramak istediği bilgi ile ilgili anahtar kelime yada kelimeleri girdiği zaman, arama motoru indeksleri inceler ve belgenin başlığını, metnin özetini, kendi kriterlerine göre en uygun sıralamayı yapıp web sayfalarının listesini içeren bir metin dönecek şekilde çalışır.

Bir çok arama motoru, arama işlemi yaparken AND, OR ve NOT gibi mantıksal operatörleri desteklemektedir. Bunun yanısıra arama yaparken kullanılan arama parametreleri de mevcuttur. "+" parametresi AND anlamına gelmekte olup, her iki kelimeyi de içeren sayfaların listelenmek istendiği belirtir. Çift tırnak parametresi, içersinde yazılan kelime grubunun aynısını içeren sayfaların listelenmek istendiğini belirtir. "-" parametresi ise, - ifadesinden sonra gelen kelimeleri içermeyen ancak - parametresinden önce gelen kelimeleri içeren sayfaların görüntülenmesi istendiğini belirtmektedir.

2.5 Çalışma Şekillerine Göre Arama Motorları

Temel olarak günümüzdeki arama motorları çalışma mantıklarına göre dört gruba ayrılmaktadır.

2.5.1 Web robotu tabanlı arama motorları

İndexlerini web robotları kullanarak güncel tutan arama motorlarıdır. Alta Vista, Yahoo, Msn gibi arama motorları örnek verilebilir.

2.5.2 İnsan kontrollü güncellemelere sahip arama motorları

İndexleri insanlar tarafından girilen arama motorlarıdır. Bu tür arama motorlarına genel olarak bilgi paylaşım platformları veya forum siteleri örnek gösterilebilir. Bu sitelerdeki bilgiler ve arama kriterleri insanlar tarafından girilmektedir. Arama yapıldığı zaman tekrar kullanıcıların erişimine olanak tanımaktadır.

2.5.3 Hibrit arama motorları

Bu tür arama motorları hem kendi web robotları ile internet sayfalarını tarayıp indekslerini oluşturmakta hem de insan kontrolünde yapılan güncellemelere olanak tanımaktadır.

2.5.4 Dağıtık arama motorları

Dağıtık yapıdaki arama motorları için merkezi bir veritabanı sistemi yoktur. Arama yapıldığı zaman girilen kriterler merkezi bir sunucu üzerinden farklı lokasyonlardaki sunuculara gönderilir. Tüm makinelerden gelen sonuçlar, yine bu merkezi sunucu üzerinden arama yapan kullanıcıya döndürülmektedir. Bu sayede hem veriye farklı kaynaklardan erişilebilmekte, hem de verinin merkezi bir yerde tutulması zorunluluğu ortadan kalkmaktadır.

3. KELİME GÖVDELEME

Gövdeleme, bir kelimeye eklenmiş olan çekim eklerinin çıkartılması ile kelimenin kökünün bulunması işlemine verilen isimdir.

Gövdeleme işlemi dillere göre büyük farklılıklar göstermektedir. Örneğin, İngilizce gibi eklerin kullanımının az olduğu bir dil için yalnızca ekler sözlüğüne bakılarak bir gövdeleme sistemi geliştirmek mümkündür. Bu şekilde geliştirilen gövdeleme algoritmasının Türkçe gibi çok sayıda ek içeren bir dil için kullanılması pek mümkün değildir. Türkçe bitişken bir dil olduğu için, kelimedeki eklerin sayısı ve eklenme şekli daha detaylı bir incelemeyi gerektirmektedir.

3.1 Gövdeleme Algoritmaları

Kelimelerin aldığı ekler, dillere göre farklılık göstermektedir. Ancak her dilin yapısına göre küçük değişiklikler yapılarak o dile uyarlanabilecek kelime gövdeleme algoritmaları oluşturulmuştur. Kelime gövdeleme algoritmalarından birkaçı aşağıda anlatılmıştır.

3.1.1 Brute-force algoritması

Kelime kökü ile türemiş şekli arasındaki ilişkiler tablolarda tutulmaktadır. Bir kelimenin gövdelenebilmesi için, kelime kökü ile türemiş formunun o tabloda bulunması gerekmektedir. Eğer ilişki tablosunda tanımlı ise, gövdeleme işlemi gerçeklenebilir, ancak tanımlı değil ise işlem gerçeklenememektedir.

Bu algoritmada, bir dildeki tüm sözcüklerin ve ilişkilerinin bilinmesi gerekmekte, sözcükler arası ilişkilerin tek tek tanımlanması gerekmektedir. Bu işlem sözcük sayısı bakımından zengin olan dillerde uzun zaman almaktadır. İlişkilerin tamamı tanımlandığı zaman tablolar büyük yer kaplayabilir. Tablolara veri girişi el ile yapıldığı için insan kaynaklı problemler ortaya çıkabilir. Tabloların tasarımları oldukça zor bir işlemdir. Her yeni kelimenin ilişki tablolarına eklenmesi gerekmektedir. Bu algoritmanın getirdiği bazı kolaylıklardan bir tanesi, algoritma hatalarını düzeltmek için ilişkilerin tekrardan tanımlanması ya da yeni bir ilişkinin tanımlanması yeterli olacaktır. Ayrıca algoritma, İngilizce gibi bazı dillerde istisnalara uymayan kelimeler için (düzensiz fiiller) büyük kolaylık sağlamakta ve algoritmanın tekrar tasarlanmasına gerek duyulmamakta, ilişkilerin tanımlanması yeterli olmaktadır.

3.1.2 Suffix-stripping algoritması

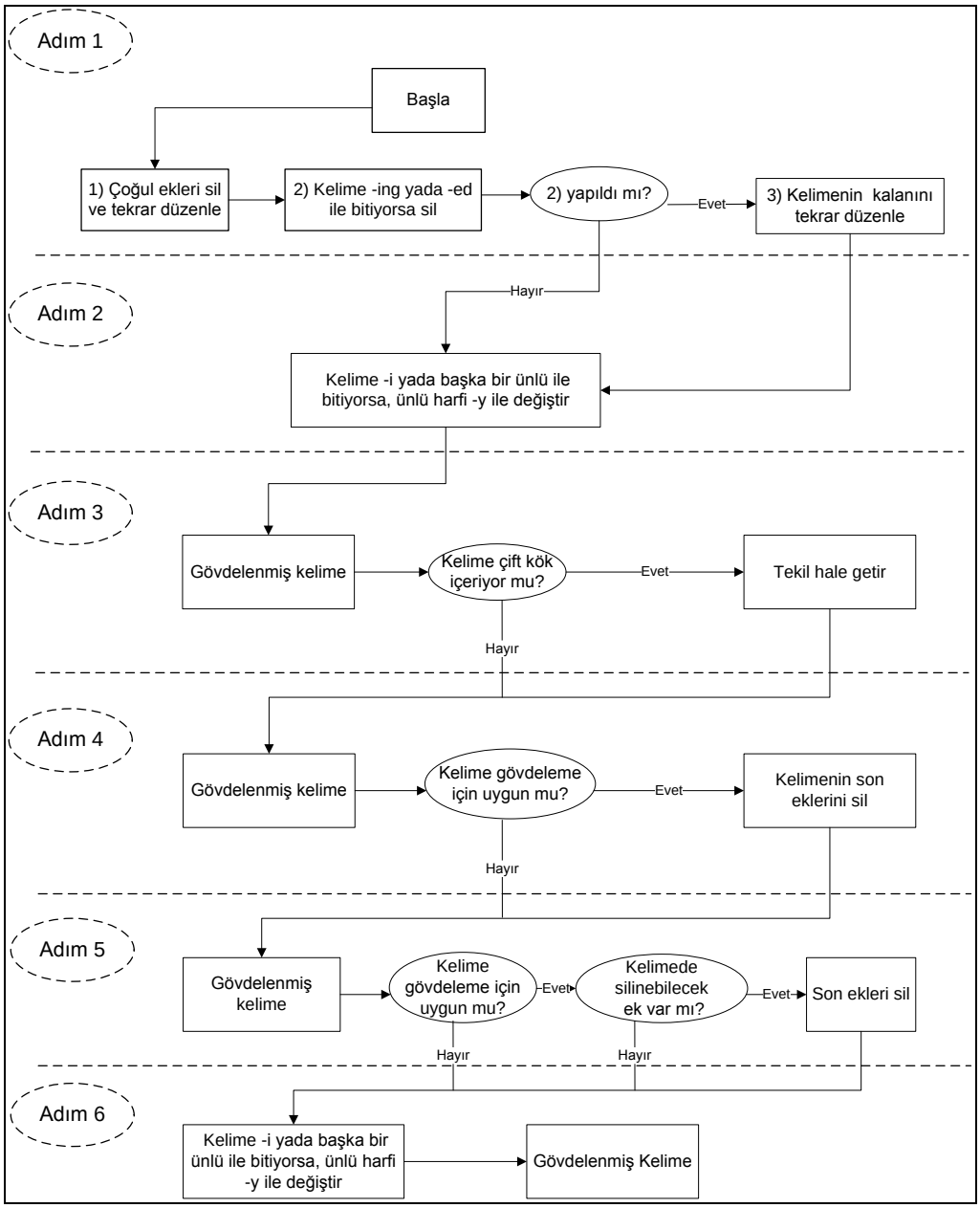
Suffix-stripping algoritmasında, kelime kökünü bulmak için belirli kural listeleri tanımlanmıştır ve bu kurallara göre kelime eklerinin silinmesiyle kök bulunmaktadır. Örneğin; İngilizce dili için, eğer kelime 'ed' ile bitiyorsa, 'ed' eki silinerek kelime kökü bulunabilir.

Suffix-stripping algoritması, ilişki tabloları kadar güvenli değildir. İstisnai durumlar (İngilizce'deki düzensiz fiiler) için kötü performans sergiyelebilmektedir. Ancak dilin özelliklerinin yeterince bilinmesi ile, kelimer arası ilişkilerin tanımlama zorunluluğu ortadan kalkmıştır.

3.1.3 Porter algoritması

Porter algoritması, ilk defa 1980 yılında Martin Porter tarafından geliştirilmiştir. Kelimedeki son ekleri iteratif olarak silmeye dayalı bir gövdeleme algoritmasıdır. Bu algoritma bir çok dil için kullanışlıdır. Algoritma, birbirine bağlı beş ya da altı alt adıma bölünerek, son adımda kelime kökünü bulmayı hedefler.

Algoritmanın ilk adımı, üç parçadan oluşmaktadır. Birinci kısmında, kelimelerin çoğul eklerinin silinmesi işlemi yapılmaktadır. İkinci kısımda, düzenli fiilerin geçmiş zaman eki almış formlarının sonundaki –ed takısının silinmesi işlemi yapılmaktadır. Üçüncü kısım, kelimeyi bir sonraki adım için uygun hale getirir. İkinci adımda, kelimede –ing eki varsa silinir ve bir sonraki adım için kelime uygun formata getirilir. Diğer adımlarda gövdeleme işlemine devam edilir ve son adım, çıktı olarak gövdelenmiş kelimeyi sunar. Porter algoritmasının, iş akış diyagramı şekil 3.1 deki gibidir.

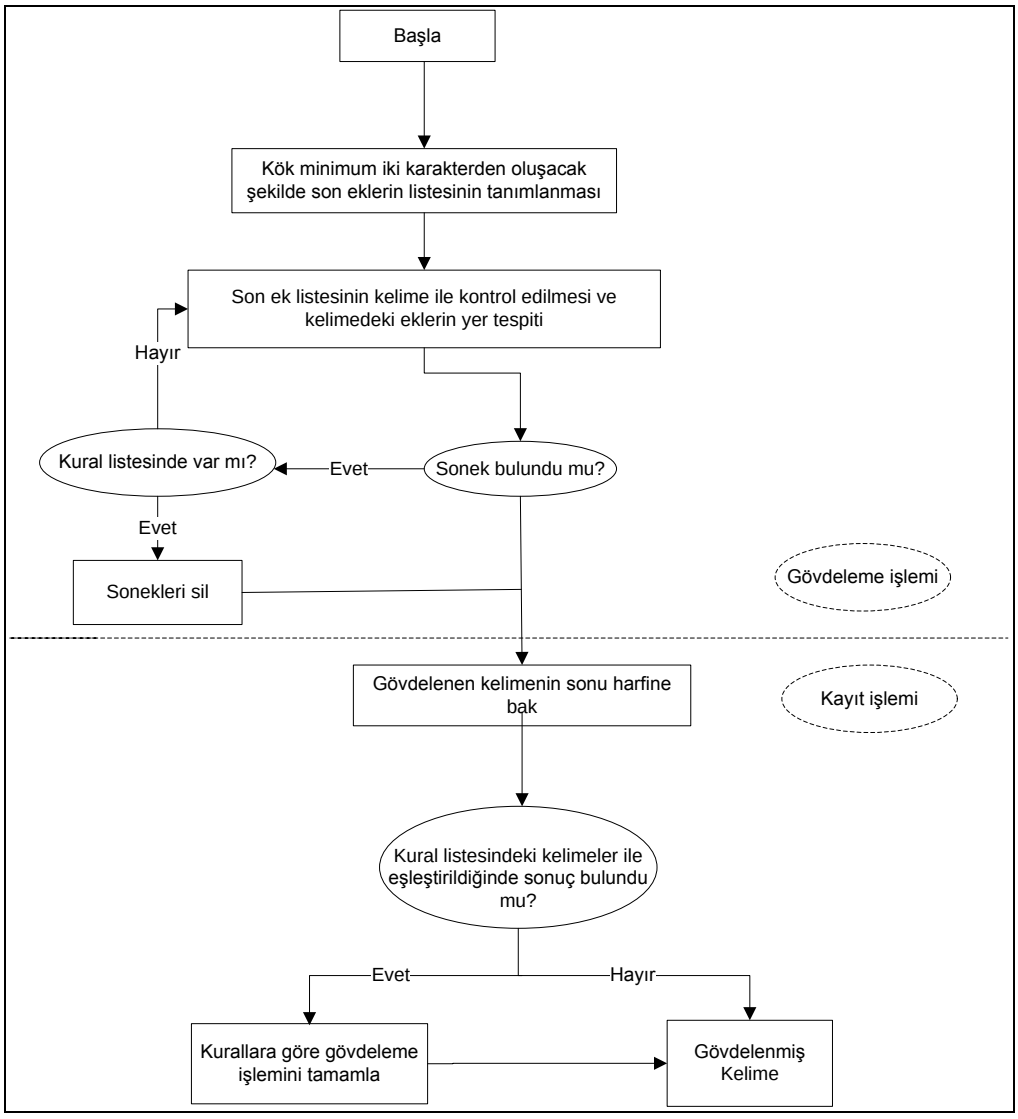


Şekil 3.1 : Porter algoritması iş akış diyagramı.

3.1.4 Lovins algoritması

Lovins kelime gövdeleme algoritması, 1968 yılında Julie Beth Lovins tarafından bulunmuştur. En uzun eşleme algoritmasına göre eklerin kaldırıldığı bir algoritmadır. Bazı uç noktarda hataları engellemek için bir koşul listesi tutulmaktadır. Bu liste en çok karşılaşılan durumlardan yola çıkılarak oluşturulmuştur. Örneğin gövdeleme işlemi yapıldıktan sonra, geriye kalan kök en az iki karakter olmalıdır ve diğer kurallar gövdeleme işlemi sırasında sondaki eklerin silinmesi ile ilgilidir.

Algoritma temel olarak iki adımdan oluşmaktadır. İlk adımda, gövdeleme işlemi, son eklerin silinmesi aşamaları bulunmaktadır. İkinci adımda ise, kelime uygunluğu konrol edilip sonuç çıktısı döndürülür. İş akış diyagramı Şekil 3.2 deki gibidir.



Şekil 3.2 : Lovins algoritması iş akış diyagramı.

4. GERÇEKLENEN SİSTEM

Tez kapsamında, öncelikli olarak tekil çoğul kelime grubu, bileşik kelime grubu ve son ek alan düzenli fiil grubu olmak üzere kelimeler üç gruba ayrılmıştır. Bu üç grup için 25 tane test kelimesi belirlenmiştir. Sonraki aşamada, her bir kelime grubu için ayrı ayrı yöntemler yazılmış ve uzun süren testler yapılmıştır. Sonuç olarak, çıkarımlarda bulunulmuştur. Sistemin gerçekleştirme aşamaları aşağıdaki gibidir.

4.1 Kullanılan Teknolojiler

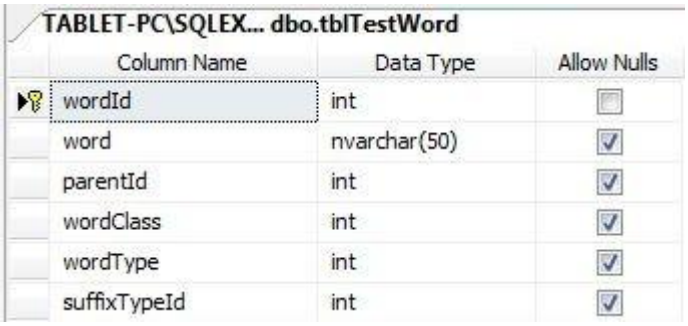
Çalışmada, Bing'den sonuçların daha sonra değerlendirilebilmesi için bir yere kaydedilmesi gerekmektedir. Bu amaç için verilere hızlı ve kolay bir şekilde erişim sağlanması açısından Microsoft SQL (Structured Query Language) server 2008 kullanılmıştır.

Aramaların el ile yapılması, hem zaman kaybına sebebiyet vermekte hemde hatalara sebep olmaktadır. Bu yüzden geliştirilen sistemin otomatik olarak arama yapması ve Bing'den dönen sonuçların otomatik olarak veritabanına kayıt edilmesi için sistem Microsoft Visual Studio 2010 C# kullanılarak otamatik hale getirilmiştir.

4.2 Veri Tabanı

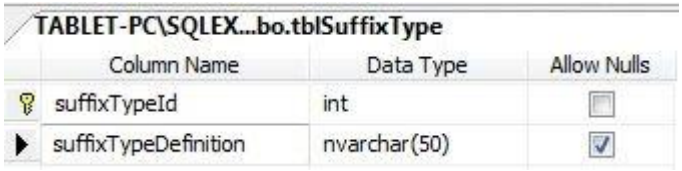
Proje kapsamında, belirlenen kelime gruplarının, bu gruplara dahil olan kelimelerin, bu kelimelerin son eklerinin, Bing'e gönderilen sorguların, Bing'den dönen cevapların ve bu cevaplara göre oluşturulmuş bilgilerin bir yerlere kayıt edilmesi gerekmekteydi. Kullanım kolaylığı ve erişim hızı göz önüne alınarak kayıt bilgilerinin tutulması için veritabanı seçilmiştir. Veritabanı sunucusu olarak MS SQL Server 2008 kullanılmıştır.

Verilerin tutulması için oluşturulan tablolar şu şekildedir.



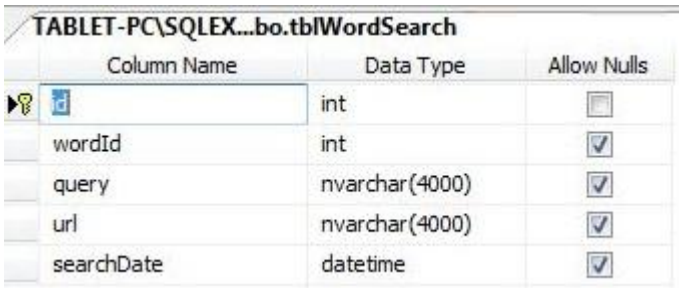
Şekil 4.1 : Kelime tablosu

Kelime tablosunda kelimeler türlerine göre tutulmaktadır. WordClass sütunu kelimenin tipini belirlemektedir. WordType sütunu ise kelimenin hangi gruba dahil olduğunu belirlemektedir. ParentId sutunu o kelimenin hangi kelimeden türediğini tutmaktadır. Eğer kelimenin kendisi kök ise; bu alana değer ataması yapılmamaktadır, eğer kelime türemiş bir kelime ise; kök kelimenin id numarası bu alana eklenmektedir.SuffixTypeId sutununda ise; sonuna aldığı ekin türünü tutmaktadır.



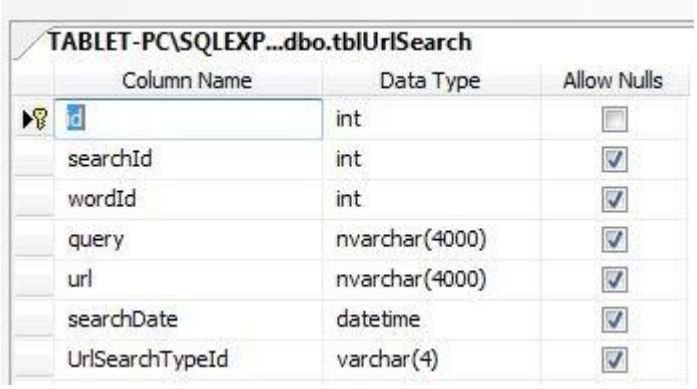
Şekil 4.2 : Son ek tipi tablosu

Son ek tipi tablosunda ise, kelimenin sonuna aldığı -ly, -sive, -tive ,… vb gibi son ek türlerini tutmaktadır.



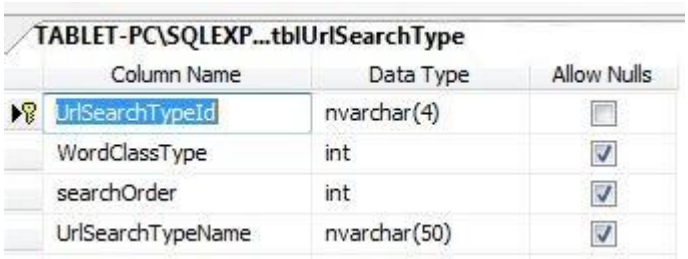
Şekil 4.3 : Kelime arama tablosu

Belirlenen her kelime grubu için farklı algoritmalara göre sorgular oluşturulur.Bu sorgular Bing' gönderilir ve Bing'den dönen sonuçlar kelime arama tablosuna kayıt edilir. Query sutunu Bing'e gönderilen sorguyu, url sutunu Bing'den dönen linkleri tutmaktadır.



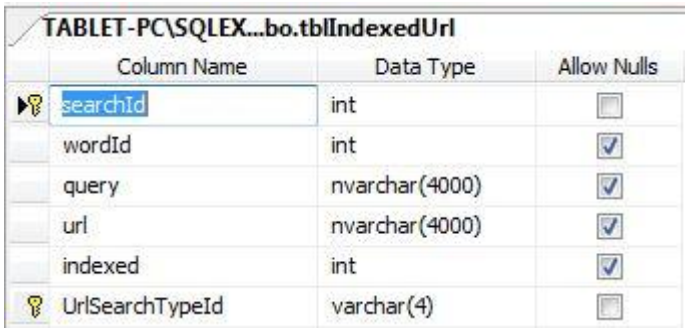
Şekil 4.4 : Link arama tablosu

Bing'den dönen her bir link için, farklı kelime gruplarına göre yeni bir sorgu oluşturulur. Bu oluşturulan sorgu Bing'e tekrar gönderilir ve ilk dönen sonuçlardaki linklerin yeni arama sonuçlarında olup olmadığına bakılır. Bu yeni arama sonuçları link arama tablosunda tutulur. Query sutunu Bing'e gönderilen sorguyu, url sutunu Bing'den dönen yeni linkler tutulmaktadır. urlSearchTypeId ise; her bir kelimeye göre farklı değerler almaktadır.



Şekil 4.5 : Link arama tipi tablosu

Her bir kelime grubu kendi içerisinde farklı yapılarda ekler almaktadır. Bu eklerin arama sonuçları yorumlanırken ayrılması ve arama yaparken belirli bir sıraya göre sorguların oluşturulup Bing'e gönderilmesi gerekliydi. Bu sebeple link arama tipi tablosu oluşturuldu. WordClassType sutunu kelime grubunu, searchOrder sorgular yapılırken oluşturulacak sırayı, urlSearchTypeName ise; her bir satırın açıklamasını tutmaktadır.



Şekil 4.6 : Linklerin indekslenme kontrolünün yapıldığı tablo

Her bir kelime grubu için yapılan aramalara ait linkler ile tekrar sorgu oluşturup aradığımızda, Bing'den yeni linkler arasında sorguyu oluşturan link mevcut ise, aranan link indekslidir. Eğer link sonuçlar arasında yoksa o link indeksli değildir.

Ayrıca veritabanı üzerinde çeşitli amaçlar için bir çok store procedure ve fonksiyon yazılmıştır.

4.3 Yöntemler

4.3.1 Kelime seçimi

Yapılan tez kapsamında incelenen makalelerden yola çıkılarak test için İngilizce dili seçilmiştir. Dil belirlendikten sonra aynı dil yapısına sahip kelimeler belirlenmiş olup ortaya test yapılabilecek üç kelime grubu belirlenmiştir. Bunlar tekil - çoğul, bileşik ve son ek alan kelime grubu olmak üzere üçe ayrılır.

Tekil – çoğul kelime grubundaki her bir kelime sadece tekil ve çoğul olarak bulunur. Bu kelimeler başka ek almamıştır. Bileşik kelimeler, iki farklı kelimeden oluşmuştur. Yine bileşik kelimeler sadece tekil ve çoğul olarak bulunabilirler. Son ek almış kelimeler ise, onbir farklı son ek alan düzenli fillerden oluşmaktadır.

4.3.2 Sorgu oluşturulması

Kelime grupları belirlendikten sonra her bir kelime grubu için ayrı ayrı sorguların oluşturulması gerekir. Her bir kelime için arama yapıldığında sadece o kelime ile ilgili linklerin dönmesi istenmektedir. O kelimeye ek alıp türeyen formatları yada o kelimeden oluşan başka kelimeler istenmez. Bu amaç doğrultusunda her bir kelime grubu için farklı sorgular oluşturulmuştur. Örneğin; tekil – çoğul kelimeler için, kelimenin tekil formu –çoğul formu, "benzene –benzenes", şeklinde sorgu oluşturulurken, bileşik kelimeler için de, kelimenin tekil formu –çoğul formu –alt kelimenin tekil formu –alt kelimenin çoğul formu –ikinci alt kelimenin tekil formu -ikinci alt kelimenin çoğul formu, "brickbath -brickbaths -brick -bricks -bat –bats", şeklinde bir sorgu oluşturulmaktadır. Benzer şekilde son ek almış kelimeler içinde, kelimenin tekil formu son ek almış tüm halleri "define -defines -defined -defining...", şeklinde bir sorgu oluşturulmaktadır.

4.3.3 Link seçimi

Her bir kelime grubu için oluşturulan sorgular Bing'e gönderilir. Bing'den dönen sonuçlar her bir kelime grubu için farklı biçimde ele alınır.

Tekil-çoğul kelime grubu için iki farklı link seçim algoritması kullanılmaktadır. Bunlardan birincisi, dönen sonuçlar içersinde birinci sıradan başlanarak ilk yüz tanesinin seçilmesidir. Örneğin, "benzene –benzenes" sorgusu Bing'e gönderilip dönen ilk 100 link kaydedilmektedir.

İkincisi ise, sıralamada daha gerilerdeki linkleri de ele almak adına sorgu oluşturulurken üç haneli rastgele sayılar seçerek sorgular oluşturulmuştur. Dönen sonuçlar içersinde yine ilk sıradan başlayarak ilk on tanesi seçilmiştir. Örneğin, "benzene 918–benzenes" sorgusu Bing'e gönderilip dönen ilk 10 link kaydedilmektedir. Bu işlem on faklı sayı ile tekrarlanmıştır.

Bileşik kelimeler ve son ek almış kelime grupları için tek algoritma kullanılmıştır. Oluşturulan sorgular Bing'e gönderilmiştir. Bing'den dönen sonuçlar arasından bileşik kelimeler için; ilk sıradan başlayıp ilk yüz tanesi, son ek almış kelimeler için de; ilk otuz tanesi seçilmiştir.

4.3.4 Link kontrolü

Tez kapsamında, tüm sorgular Bing'e gönderildikten sonra dönen cevaplar kontrol edilecek olan linkleri oluşturmaktadır. Bu linkleri kontrol etmekteki amaç o link, arama yapılan kelimelerin farklı formları içinde indekslenip indekslenmediğini belirlemektir. Bu amaç doğrultusunda her bir kelime grubundaki sorgu sonucu gelen linkler ve – parametresi devamına yazılan kelimeler kullanılarak yeni sorgular oluşturulur ve Bing'e gönderilir. Eğer bu yeni sorgu sonucu aranılan link sonuçlar içerisindeyse, o link o kelime için indekslenmiştir.Eğer dönen sonuçlar içerisinde söz konusu link gelmediyse o kelime için indekslenmemiştir.

Tekil – çoğul kelime grubu için oluşturulan link, kelimenin çoğul formu + link1 şeklindedir. Alınan yüz link bu şekilde taranır ve bulunan sonuçlar veritabanına kaydedilir. Benzer şekilde işlemler diğer kelime grupları içinde tekrarlanır.

4.3.5 Bing web servis kullanımı

Bing'e sorguların gönderilip, dönen sonuçların değerlendirilmesi işlemini el ile yapmak zor ve zaman kaybına sebep olmaktadır. Bu işlemin bir program yardımıyla otomatikleştirilmesi gerekmektedir. Bu amaç için Bing'in son kullanıcılara sunduğu web servisi bulunmaktadır.

Web servisini kullanabilmek için geliştirici olarak sisteme kayıt olunması gerekmektedir. Kayıt işlemi sonucunda her kullanıcı için farklı bir key numarası verilmektedir. Kullanılacak olan web servisi projeye eklendikten sonra, her aramada gelecek kayıt sayısı, arama yapılabilmesi için key değerinin girilmesi, arama sonucu dönen verilerin istenilen formata dönüştürülmesi gibi bazı özel ayarların yapılmasını sağlayan bir sınıf yazılmıştır. Bu sınıf içerisinde en çok kullanılan yöntem ise arama fonksiyonudur. Arama işlemi ise şu şekilde yapılmaktadır. Öncelikli olarak key numarası ataması, arama sonucu dönecek kayıt listesi atamaları yapılır. Sonrasında bir tane "SearchRequest" nesnesi oluşturulur. Resim, adres, video, web, ... vs gibi arama tiplerinden uygun olanı seçilir. "BingPortTypeClient" nesnesi yardımı ile Bing'e arama isteği gönderilir. Dönen sonuçlar "WebResult" nesnesine atanır ve sonuçlar üzerinde iyileştirme işlemleri gerçeklenir.

4.4 Tekil Çoğul Kelimeler İçin Gövdeleme

Kelime gövdeleme işleminin en çok yapılan kısmı, çoğul kelimelere gövdeleme işlemi yapılarak tekil formlarının oluşturulmasıdır. Tekil çoğul kelimeler arası ilişkiler çok kuvvetlidir ve istisnası İngilizce kelimeler için çok azdır. Arama motorlarının tekil çoğul kelimelerin gövdeleme işlemine karşı verdikleri cevap, kendi içlerindeki gövdeleme mekanizması hakkında en fazla bilgiyi veren yapıdır. Tez kapsamında ilk incelenen kelime grubudur.

Tekil çoğul kelimeleri incelemek için, öncelikli olarak hangi kelimelerin test için kullanılacağına karar verilmesi gerekmektedir. Çoğul eki almış formları dışında hiç bir ek almamış kelimeler seçildikten sonra içlerinden rastgele 25 tanesi test için seçilmiştir. Test işlemleri dört adımdan oluşmaktadır.

Birinci test adımında, seçilen kelimelerin her biri için "tekil –çoğul" şeklinde sorgular oluşturulmuş olup, Bing'de sorgulama yapılmıştır. Dönen sonuçlar içersinden, Bing'de dönen sorgu sonuçlarına göre ilk sıradan başlayıp ilk yüz tane sonuç sisteme kaydedilmiştir. Daha sonra bu yüz sonuç için, tek tek "çoğul link" şeklinde bir sorgu oluşturulmuş ve Bing'e gönderilmiştir. Bing'den dönen ilk 100 sonuç alınıp veri tabanına kaydedilmiştir. Eğer sorguyu oluşturan link, gelen 100 arama içersinde var ise, bu link kelimenin çoğul formu için indekslenmiştir, eğer yok ise indekslenmemiştir sonucuna ulaşılmaktadır. Ulaşılan bu sonuçlar, veritabanına kaydedilmiştir.

İkinci test adımında, seçilen kelimelerin her biri için "çoğul –tekil" şeklinde sorgular oluşturulur. Birinci adımdakine benzer şekilde Bing'e gönderilip aynı sayıda sonuç veritabanına kaydedilir. Dönen linkler ile tekrardan "tekil link" şeklinde sorgu oluşturulup, Bing'den dönen sonuçlar değerlendirilerek veri tabanına kaydedilir.

Tablo 4.1: Tekil çoğul kelime içeren arama sonuçları.

Tekil Kelime	Çoğul Kelime	Tekil sorgular için döndürülen , çoğul kelimeler içeren döküman yüzdesi	Çoğul sorgular için döndürülen, tekil kelimeler içeren döküman yüzdesi
benzene	benzenes	60	6
brickbat	brickbats	11	3
starling	starlings	83	4
fishmonger	fishmongers	37	37
parvenu	parvenus	26	4
nursery	nurseries	75	64
wellspring	wellsprings	73	43
mistress	mistresses	82	28
quoit	quoits	33	20
ruckus	ruckuses	73	7
aerosol	aerosols	80	36
broom	brooms	69	49
caisson	caissons	50	27
drawback	drawbacks	25	38
fusion	fusions	90	3
genius	geniuses	32	6
heirloom	heirlooms	81	59
mutt	mutts	71	63
nanny	nannies	82	71
pumpkin	pumpkins	72	55
rye	ryes	71	11
satellite	satellites	76	8
scope	scopes	79	47
tadpole	tadpoles	62	40
unction	unctions	27	13

Tablo 4.1'de ilk iki sütun seçilen kelimelerin tekil ve çoğul formlarını göstermektedir. Tablonun üçüncü sütunu tekil kelime içeren dökümanların, çoğul kelimeler için indekslenme oranını göstermektedir. Tablonun son sütünu ise, çoğul kelime içeren dökümanların tekil kelimeler için indekslenme oranını göstermektedir. Tablodan da anlaşıldığı üzere, genel olarak tekil kelime içeren dökümanların, çoğul kelimeler ile indekslenme oranı, çoğul kelime içeren dökümanların tekil kelimeler için indekslenme oranından daha yüksektir. Sorgu bakımından hiç bir fark olmamasına rağmen bazı kelimelerin indekslenme oranı yaklaşık %10 iken,bazı kelimeler için bu oran yaklaşık % 90'dır. Bu oranlar arasındaki farklılıkların hem döküman hem de kelime bazlı olduğu düşünülmektedir.

Üçüncü test adımında ise, bir arama için popüler olan linklerinin etkisinin olup olmadığını anlayabilmek için arama için oluşturulan sorguya 100 ile 999 arasında üç basamaklı bir sayı (ÜBS), eklenip "tekil ÜBS –çoğul" şeklinde bir sorgu oluşturulmuştur. Her bir sorgu için, Bing'den dönen ilk on sonuç veritabanına kaydedilmiştir. Dönen linkler kullanılarak tekrar "çoğul link" şeklinde bir sorgu oluşturulmuş ve yine ilk gelen on sonuç değerlendirmeye alınmıştır. Eğer sorgu oluştururken kullandığımız link, ikinci sorgu sonuçlarında da geliyorsa kullanılan link o kelime için indekslidir, dönmüyorsa indeksli değildir sonucuna ulaşılabilir. Bu işlem her bir kelime için on kez farklı sayılar ile tekrarlanmıştır.

Dördüncü test adımında ise, üçüncü test adımındakine benzer şekilde "çoğul ÜBS – tekil" şeklinde sorgu oluşturulup, Bing'e gönderilir. Yine dönen sonuçlar değerlendirilip sonuçları veritabanına kaydedilir. Her bir kelime için bu işlem on kez farklı sayılar ile tekrarlanır.

Tablo 4.2: ÜBS ve tekil çoğul kelime içeren arama sonuçları.

Tekil Kelime	Çoğul Kelime	ÜBS içeren Tekil kelimeler için döndürülen tekil kelimeler içeren döküman yüzdesi	ÜBS içeren Çoğul kelimeler için döndürülen tekil kelimeler içeren döküman yüzdesi
benzene	benzenes	17	5
brickbat	brickbats	12	5
starling	starlings	26	5
fishmonger	fishmongers	34	22
parvenu	parvenus	16	5
nursery	nurseries	46	24
wellspring	wellsprings	31	31
mistress	mistresses	29	20
quoit	quoits	41	31
ruckus	ruckuses	27	7
aerosol	aerosols	20	11
broom	brooms	30	11
caisson	caissons	31	26
drawback	drawbacks	14	19
fusion	fusions	36	2
genius	geniuses	9	9
heirloom	heirlooms	31	20
mutt	mutts	25	17
nanny	nannies	45	17
pumpkin	pumpkins	37	17
rye	ryes	34	7
satellite	satellites	28	2
scope	scopes	18	13
tadpole	tadpoles	29	33
unction	unctions	42	9

Tablo 4.2'de ilk iki sütun seçilen kelimelerin tekil ve çoğul formlarını göstermektedir. Tablonun üçüncü sütununda, rastgele seçilmiş 100 ile 999 arasındaki üç haneli sayıyı ve kelimenin tekil formunu içeren dökümanların, çoğul kelime formu için indekslenme oranı gösterilmektedir. Tablonun son sütununda ise, rastgele seçilmiş 100-999 arasındaki sayıyı ve kelimenin çoğul formunu içeren dökümanların tekil kelime formu için indekslenme oranı gösterilmektedir. Tablodan da anlaşılacağı gibi ÜBS ve tekil kelime formu içeren döküman sayısının çoğul kelime formu için indekslenme oranı, çoğul kelimelere göre daha fazladır. Bu indeksleme oranı tekil kelime içeren dökümanlar için; %6 ile % 46 arasında değişmekte, çoğul kelime formunu içeren dökümanlar için; %3 ile %35 arasında değişmektedir. Bu oranlar sonucunda dökümanın indeksleme işleminin kelimeye ve dökümana bağlı olduğu düşünülmektedir.

İlk kelime grubu testleri sonucunda her bir kelime için farklı sonuçlar elde edilmiştir. Bazı kelimeler için indeksleme oranları % 80 civarındayken, bazı kelimeler için bu oran yaklaşık %6 dır. Bu oranlar arasındaki farkın ilk sebebi, tekil kelimeler için arama yapılırken arayan kişinin kelimenin çoğul formunu arıyor olabileceği düşünülmüş olabilir. Tablo 4.1 ve Tablo 4.2 incelendiğinde aynı kelime için yapılan aramalar için bulunan indeksleme oranlarının değiştiği gözlemlenmektedir. Buna göre; örnek kelime için arama yapıldığı zaman ilk sıralarda gelen dökümanların, sonraki sıralarda gelen dökümanlara göre indeksleme oranı daha yüksek olabilir sonucu çıkartılabilir. Ayrıca bu oranların değişmesinin diğer bir sebebide her aramada gelen dökümanların farklı olması olabilir. Dolayısıyla dökümanların indekslenip indekslenmediği bilgiside değişmektedir. Bing'in tekil çoğul kelimeler için hem kelime hemde döküman bazlı indeksleme yaptığı tahmin edilmektedir.

Genel olarak, hemen hemen tüm kelimeler için tekil kelime içeren dökümanların çoğul kelime formu içinde indekslenme oranının, çoğul kelime içeren dökümanların tekil kelime formu için indeksleme oranından yüksek olduğu görülmüştür.

4.5 Bileşik Kelimeler İçin Gövdeleme

Birçok dilde olduğu gibi İngilizce'de de bileşik kelimeler yaygın olarak kullanılmaktadır. Bileşik kelimeler birden fazla kelimeden oluşmaktadır. Örneğin, "airway" kelimesi, "air" ve "way" alt kelimelerinden oluşan bir bileşik kelimedir. Bazı durumlarda arama motorları bileşik kelime aranmasına rağmen, alt kelimeleri içeren dökümanları döndürebilmektedir. Bu kısımda Bing'in bileşik kelimeler için nasıl bir yöntem izlediği araştırılmıştır.

Bu kelime grubu için öncelikli olarak, kelimelerin bileşik kelime olup olmadığına karar verilmiştir. Bileşik kelimeler içerisinden, sadece tekil ve çoğul formlarının bulunduğu kelimeler belirlenmiş ve bu bileşik kelimeler içerisinden 25 tanesi test kelimesi olarak seçilmiştir. Bir sonraki işlem adımı olarak, her bir kelimenin alt kelimeleri, bu kelimelere ait tekil ve çoğul formlar bulunmuştur. Bileşik kelime için yapılan testler, kelimenin tekil formu ve çoğul formu için iki farklı grup şeklinde yapılmıştır.

Tablo 4.3'de tez için kullanılan bileşik kelimeler ve bu kelimelere ait kullanılan kelime formları gösterilmiştir.

Tablo 4.3: Bileşik kelime listesi ve alt kelime formları

Tekil Kelime	Çoğul Kelime	Alt Kelime 1	Alt Kelime 1 Çoğul Form	Alt Kelime 2	Alt Kelime 2 Çoğul Form
airway	airways	air	airs	way	ways
beachhead	beachheads	beach	beaches	head	heads
brickbat	brickbats	brick	bricks	bat	bats
fishmonger	fishmongers	fish	fishes	monger	mongers
gumdrop	gumdrops	gum	gums	drop	drops
inflow	inflows	in	ins	flow	flows
oddball	oddballs	odd	odds	ball	balls
penthouse	penthouses	pent	pents	house	houses
starling	starlings	star	stars	ling	lings
wellspring	wellsprings	well	wells	spring	springs
backwater	backwaters	back	backs	water	waters
crossword	crosswords	cross	crosses	word	words
daisywheel	daisywheels	daisy	daisies	wheel	wheels
fairground	fairgrounds	fair	fairs	ground	grounds
grapefruit	grapefruits	grape	grapes	fruit	fruits
headstone	headstones	head	heads	stone	stones
icebox	iceboxes	ice	ices	box	boxes
mainstay	mainstays	main	mains	stay	stays
oilcloth	oilcloths	oil	oils	cloth	cloths
passport	passports	pass	passes	port	ports

roundhouse	roundhouses	round	rounds	house	houses
saddlebag	saddlebags	saddle	saddles	bag	bags
seawall	seawalls	sea	seas	wall	walls
teapot	teapots	tea	teas	pot	pots
waybill	waybills	way	ways	bill	bills

Bileşik kelimelerin ilk grubunu test etmek için, "kelimenintekil formu –çoğulformu – altkelime1 –altkelime2 –altkelimeninçoğulformu1 – altkelimeninçoğulformu2" şeklinde bir sorgu oluşturulup Bing'e gönderilmiştir. Bing'den gelen ilk yüz sonuç veritabanına kaydedilmiştir. Daha sonra her bir alt kelime grubu için, "altkelime1 altkelime2 link", "altkelime1 altkelimeninçoğulformu2 link", "altkelimeninçoğulformu1 altkelime2 link" ve "altkelimeninçoğulformu1 altkelimeninçoğulformu2 link" şeklinde sorgular oluşturulmuştur ve her bir sorgu, ayrı ayrı Bing'e gönderilmiştir. Bing'den dönen ilk yüz sonuç veritabanına kaydedilmiştir ve her bir link için, dönen sonuçlar içinde var olup olmadığı kontrol edilmiştir. Eğer link dönen sonuçlar içerisinde var ise, o link arama yapılan ikinci sorgudaki kelimeler için indekslidir, yok ise indeksli değildir sonucuna ulaşılmaktadır.

Tablo 4.4: Tekil bileşik kelimeler için yapılan aramaların sonuçları.

Kelime	1.Alt Kelime 2.Alt Kelime	1.Alt Kelime 2. Alt Kelimenin Çoğul Formu	3.Alt Kelimenin Çoğul Formu 4.Alt Kelime	1.Ve 2. Alt Kelimelerin Çoğul Formları
airway	35	20	1	10
beachhead	0	17	0	0
brickbat	13	6	0	0
fishmonger	17	15	0	0
gumdrop	25	16	2	0
inflow	8	26	0	0
oddball	50	1	0	0
penthouse	49	43	15	3
starling	32	28	0	0
wellspring	61	51	0	0
backwater	41	49	1	0
crossword	31	25	0	0
daisywheel	0	0	0	0
fairground	56	15	8	5
grapefruit	25	22	1	1
headstone	38	33	0	0
icebox	0	0	0	0
mainstay	57	51	18	12
oilcloth	14	7	2	0
passport	34	30	3	1

roundhouse	45	5	24	2
saddlebag	0	3	0	0
seawall	8	1	0	0
teapot	26	23	0	0
waybill	8	13	0	0

Tablo 4.4'de görüldüğü gibi ilk sütun kelimelerin tekil formlarını içermektedir. Tablonun ikinci sütununda bileşik kelimeye ait alt kelimelerin tekil formları ile yapılan test sonuçlarının yüzdelik oranları gösterilmiştir. Üçüncü sütunda ise, birinci alt kelimenin tekil, ikinci alt kelimenin çoğul formu için yapılmış test sonuçlarının yüzde oranları gösterilmiştir. Dördüncü sütunda, birinci alt kelimenin çoğul ikinci alt kelimenin tekil formu kullanılarak yapılan testlerin yüzdelik oranları görüntülenmiştir. Tablonun son sütununda ise, her iki alt kelimenin de çoğul formları için yapılmış test sonuçları görüntülenmiştir.

Yapılan test sonuçlarına göre, tekil kelime formu içeren dökümanlar genellikle kelimenin her iki alt kelimenin tekil formu için indekslenme oranı, diğerlerine göre daha yüksektir. Her iki alt formunda tekil olduğu grup için bu oran bazı kelimeler için % 0'a yakınken bazı kelimeler için % 60'lara kadar çıkmıştır. Bu grupdan sonra ortalama olarak en yüksek indeksleme oranına sahip ikinci grup birinci alt kelimenin tekil, ikinci alt kelimenin ise çoğul olduğu gruptur. Geriye kalan iki grup için ortalama indeksleme oranı %0'a yakındır.

Bileşik kelimeler ikinci grup test işlemi için, “kelimeninçoğul formu –tekilformu – altkelime1 –altkelime2 –altkelimeninçoğulformu1 – altkelimeninçoğulformu2” şeklinde bir sorgu oluşturulup Bing’e gönderilmiştir. Bir önceki test grubunda olduğu gibi, Bing’den gelen ilk yüz sonuç veritabanına kaydedilmiştir. Daha sonra her bir alt kelime grubu için, “altkelime1 altkelime2 link”, “altkelime1 altkelimeninçoğulformu2 link”, “altkelimeninçoğulformu1 altkelime2 link” ve “altkelimeninçoğulformu1 altkelimeninçoğulformu2 link” şeklinde sorgular oluşturulmuştur ve her bir sorgu, ayrı ayrı Bing’e gönderilmiştir. Bing’den dönen ilk yüz sonuç veritabanına kaydedilmiştir ve ilk arama sonucu gelen her bir link için, dönen sonuçlar içerisinde var olup olmadığı kontrol edilmiştir. Eğer link dönen sonuçlar içerisinde var ise, o link arama yapılan ikinci sorgudaki kelimeler için indekslidir, yok ise indeksli değildir sonucuna ulaşılmaktadır.

Tablo 4.5: Çoğul bileşik kelimeler için yapılan aramaların sonuçları.

Kelime	1.Alt Kelime 2.Alt Kelime	1.Alt Kelime 2. Alt Kelimenin Çoğul Formu	1.Alt Kelimenin Çoğul Formu 2.Alt Kelime	1.Ve 2. Alt Kelimelerin Çoğul Formları
airways	20	27	0	9
beachheads	1	16	0	0
brickbats	2	5	0	0
fishmongers	9	16	0	0
gumdrops	10	10	0	0
inflows	9	17	0	0
oddballs	18	3	0	0
penthouses	31	43	3	1
starlings	12	33	0	0
wellsprings	35	44	0	0
backwaters	19	30	0	0
crosswords	18	32	0	0
daisywheels	0	0	0	0
fairgrounds	19	22	0	5
grapefruits	3	13	0	0
headstones	39	48	0	3
iceboxes	0	0	0	0
mainstays	7	6	0	0
oilcloths	4	15	0	0
passports	21	27	0	0
roundhouses	14	4	5	4
saddlebags	0	11	0	0
seawalls	5	0	0	0
teapots	7	13	0	0
waybills	14	26	0	0

Tablo 4.5'de ilk sütunda bileşik kelimenin çoğul formu bulunmaktadır. Tablonun ikinci sütununda her iki alt kelimenin tekil formu için yapılan test sonuçlarının yüzdelik oranı gösterilmektedir. Üçüncü sütunda ise, ilk alt kelimenin tekil formu, ikinci alt kelimenin ise tekil formu için yapılan test sonuçlarının yüzdelik oranı gösterilmektedir. Son sütunda ise her iki alt kelimeninde çoğul formu için yapılan test sonuçlarının yüzdelik oranı gösterilmiştir.

Tablodan da anlaşıldığı gibi, ilk alt kelimenin tekil ikinci alt kelimenin de çoğul formunun kullanıldığı test sonuçlarının yüzdelik oranı en yüksektir. Bu oranda bu grubun, bileşik kelimenin çoğul formunun bulunduğu dökümanlar için indekslendiğini göstermektedir. En yüksek ikinci indekslenme oranı ise, her iki alt kelimeninde tekil formda olduğu durumdur. Diğer iki grubun indekslenme oranları ortalamaları %0'a yakın değerler almıştır. Burdan o gruptaki kelimelerin, bileşik kelime içeren dökümanlar için indekslenmediğini göstermektedir.

Genel olarak, bileşik kelimenin tekil formunu içeren hemde çoğul fomunu içeren dökümanlar alt kelime gruplarının ilk ikisi için indekslenmiş olup, diğer gruplar için indekslenmemiştir. Bu oranlar test sayısı ve test zamanı değiştikçe çok büyük farklılıklar göstermemektedir. Buna dayanarak dökümanların hem kelime bazlı, hemde içerik bazlı indekslendiği tahmin edilebilir.

4.6 Son ek Alan Kelimeler İçin Gövdeleme

Bu kısıma ise İngilizce'de, son ek alan kelimeler incelenmiş olup test için onbir farklı ek almış düzenli fiiller seçilmiştir. Söz konusu ekler, İngilizce'de en çok karşılaşılan son eklerdir ve örnek bir kelime verilerek Tablo 4.6'da gösterilmiştir. Test için toplamda 25 kelime seçilmiştir.

Tablo 4.6: Son ek alan kelime araması için kullanılan sonekler.

Sonek	Açıklama	Örnek Kelime
-	Kelime kökü	Collect
-s	Geniş zaman eki	Collects
-ed	Geçmiş zaman eki	Collected
-ing	Şimdiki zaman eki	Collecting
-tion/-sion	İsim eki	Collection
-tions/-sions	İsimin çoğul eki	Collections
-or/-er	Fiili gerçekleyen kişi eki	Collector
-ors/-ers	Fiili gerçekleyen kişinin çoğul eki	Collectors
-able	Sıfat eki	Collectable
-tive/sive	Sıfat eki	Collective
-ly	Sıfatı zarf yapan ek	Collectively
-ness	Sıfatı isim yapan ek	Collectiveness

Tablo 4.6'da görüldüğü gibi, tablonun ilk sütununda düzenli fiilin aldığı son ek gösterilmektedir. Tablonun ikici sütunuda, son ekin türü belirtilmektedir. Tablonun son sütununda ise örnek kelime gösterilmiştir. Tablo 4.7'de tez için kullanılan düzenli filler ve bunların sok ek almış formları gösterilmiştir.

Tablo 4.7: Son ek alan düzenli fiiller ve kelime formları.

Kelime	-s	-ed	-ing	-tion/ -sion	-tions/ -sions	-or/-er	-ors/-ers	-able	-tive/ -sive	-ly	-ness
acquire	acquires	acquired	acquiring	acquisition	acquisitions	acquisitor	acquisitors	acquirable	acquisitive	acquisitively	acquisitiveness
appreciate	appreciates	appreciated	appreciating	appreciation	appreciations	appreciator	appreciators	appreciable	appreciative	appreciatively	appreciativeness
define	defines	defined	defining	definition	definitions	definer	definers	definable	definitive	definitively	definitiveness
derive	derives	derived	deriving	derivation	derivation	deriver	derivers	derivable	derivative	derivatively	derivativeness
describe	describes	described	describing	description	description	describer	describers	describable	descriptive	descriptively	descriptiveness
distribute	distributes	distributed	distributing	distribution	distributions	distributor	distributors	distributable	distributive	distributively	distributiveness
elect	elects	elected	electing	election	elections	elector	electors	electable	elective	electively	electiveness
evoke	evokes	evoked	evoking	evocation	evocations	evocator	evocators	evocable	evocative	evocatively	evocativeness
imitate	imitates	Imitated	imitating	imitation	imitations	imitator	imitators	imitable	imitative	imitatively	imitativeness
manipulate	manipulates	manipulated	manipulating	manipulation	manipulations	manipulator	manipulators	manipulatable	manipulative	manipulatively	manipulativeness
act	acts	acted	acting	action	actions	actor	actors	actable	active	actively	activeness
collect	collects	collected	collecting	collection	collections	collector	collectors	collectable	collective	collectively	collectiveness
create	creates	created	creating	creation	creations	creator	creators	creatable	creative	creatively	creativeness
destroy	destroys	destroyed	destroying	destruction	destructions	destroyer	destroyers	destructible	destructive	destructively	destructiveness
determine	determines	determined	determining	determination	determinations	determiner	determiners	determinable	determinative	determinably	determinableness
divide	divides	divided	dividing	division	divisions	divider	dividers	divisible	divisive	divisively	divisiveness
exclude	excludes	excluded	excluding	exclusion	exclusions	excluder	excluders	excludable	exclusive	exclusively	exclusiveness
express	expresses	expressed	expressing	expression	expressions	expresser	expressers	expressible	expressive	expressively	expressiveness
extend	extends	extended	extending	extension	extensions	extender	extenders	extendable	extensive	extensively	extensiveness
prevent	prevents	prevented	preventing	prevention	preventions	preventer	preventers	preventable	preventive	preventively	preventiveness
produce	produces	produced	producing	production	productions	producer	producers	producible	productive	productively	productiveness
protect	protects	protected	protecting	protection	protections	protector	protectors	protectable	protective	protectively	protectiveness
reduce	reduces	reduced	reducing	reduction	reductions	reducer	reducers	reductible	reductive	reductively	reductiveness
reproduce	reproduces	reproduced	reproducing	reproduction	reproductions	reproducer	reproducers	reproducible	reproductive	reproductively	reproductiveness
select	selects	selected	selecting	selection	selections	selector	selectors	selectable	selective	selectively	selectiveness

Sonek alan kelimeleri test etmek için, her bir kelime için "kelimeKökü – ekAlmışFormları" şeklinde sorgular oluşturulmuş ve Bing'e gönderilmiştir. Bing'den dönen ilk otuz sonuç veritabanına kaydedilmiştir. Bing'den dönen sonuçlar için, "kelimeninEkliFormu link" şeklinde bir sorgu oluşturulup Bing'de tekrar arama sorgulama yaptırılır ve dönen sonuçlar veritabanına kaydedilir. Eğer dönen sonuçlar içersinde ilk link var ise, o kelime için link indekslidir, yok ise indeksli değildir sonucuna ulaşılır.

Tablo 4.8: Son ek alan düzenli filler için yapılan arama sonuçları.

Kelime	-s	-ed	-ing	-tion/ -sion	-tions/ -sions	-or/ -er	-ors/ -ers	-able	-tive/ -sive	-ly	-ness
acquire	76	73	73	11	0	0	0	0	0	0	0
appreciate	18	36	32	2	2	2	2	2	36	2	2
define	77	70	72	15	10	40	13	6	6	13	2
derive	54	15	58	23	12	11	8	8	16	8	8
describe	53	55	55	2	2	18	7	2	2	2	2
distribute	1	1	1	1	1	7	3	0	1	0	0
elect	96	13	66	0	2	0	0	0	0	0	0
evoke	85	87	87	0	0	0	0	0	0	0	0
imitate	51	52	51	11	5	5	5	5	5	5	5
manipulate	8	31	45	22	8	8	13	8	13	8	8
act	85	91	88	71	7	7	7	7	7	7	7
collect	74	31	30	1	1	1	1	1	3	1	1
create	88	93	93	13	3	8	8	3	3	5	5
destroy	86	86	83	16	5	6	6	5	5	6	5
determine	63	66	70	65	11	3	11	11	11	11	11
divide	74	74	74	2	2	2	2	2	2	2	2
exclude	25	25	25	6	6	6	6	6	6	6	6
express	76	76	76	2	52	6	2	2	2	2	2
extend	58	18	61	7	7	7	7	7	12	7	7
prevent	73	80	78	76	16	12	8	8	8	8	8
produce	77	71	75	25	0	8	75	0	0	0	0
protect	90	90	86	11	2	6	6	2	18	2	2
reduce	71	70	72	7	7	8	7	7	7	7	7
reproduce	40	41	41	7	6	37	6	7	7	6	6
select	97	100	97	21	22	12	12	18	18	10	10

Tablo 4.8'nin ilk sütununda incelenen kelimenin son ek almamış formu gösterilmektedir. Sonraki her sütunda, fiilin aldığı ilgili ek formu için bulunan test sonuçlarının yüzdelik değeri gösterilmektedir. Yani her bir ek için fiilin son ek almamış formuna sahip dökümanların indekslenme yüzdeleri gösterilmektedir. Tablonun geneli incelendiği zaman, son ek almamış fiileri içeren dökümanların fiilin farklı zaman formları için indekslenme oranlarının yüksek olduğu görülmektedir. Bir döküman indekslenirken indekslenecek kelime fiil ise, tüm zaman formlarının düşünüldüğü ve ona göre indeksleme yapıldığı tahmin edilmektedir. Fiilin ek alıp isim olmuş formları içinde indeksleme oranı geri kalan formlarına göre yüksektir. Kullanıcıların kelimenin isim formununda aramalarda kullanacağı düşünülmüş ve bu yüzden indekslendiği tahmin edilmektedir. Geri kalan sıfat ve zarf formlarının fiilin son ek almamış formunu içeren dökümanlar için indeksleme oranı çok düşüktür.

Genel olarak, son ek almış fiiler için yapılan test sonuçlarına göre, dökümanlar kelimelerin çoğu için indekslenmiş olup, örnek alınan kelimeler içerisinde bir kaç tanesinin indekslenme oranının düşük olduğu görülmüştür. Buda önceki kelime gruplarındakine benzer şekilde kelime bazlı indeksleme yapıldığını gösterebilir.

5. SONUÇ VE ÖNERİLER

Tez kapsamında, Bing'in kelime gövdeleme algoritmalarının nasıl davranış gösterdiği anlaşılmaya çalışılmıştır. Gerekli yapıların anlaşılabilmesi için, üç farklı kelime grubu oluşturulmuş ve her gruba uygun 25 adet kelime seçilmiştir. Daha sonra kelime gruplarına özel sorgular hazırlanmıştır.

Birinci kelime grubu tekil çoğul kelime grubudur. Bu gruba ait bir kelimenin sadece tekil ve çoğul kelime formu bulunmaktadır. Bu grubun testleri iki aşamalı gerçeklenmiştir. Birinci aşamanın test sonuçlarına göre, kelimenin tekil formunu içeren dökümanların çoğul form için indeksleme oranı, kelimenin çoğul formunu içeren dökümanların tekil formu için indekslenme oranından yüksektir. Yani kelimenin tekil formunu içeren bir döküman, kelimenin çoğul formu için arama yapıldığında arama sonuçları içersinde gelme ihtimali vardır. Bu oran bazı kelimeler için % 85 civarında iken bazı kelimeler için % 10 civarındadır. Benzer şekilde kelimenin çoğul formunu içeren bir döküman, tekil formu için arama yapıldığında arama sonuçları içerisinde gelebilmektedir. Ancak indekslenme oranı % 60 ile % 6 arasında değişmektedir.

İkinci kelime grubu bileşik kelime grubudur. Bu gruba ait bir kelimenin sadece tekil ve çoğul kelime formu olmalı, aynı zamanda kelime en az 2 alt kelimeden oluşmalıdır. Bu grup için, yapılan testlerde her bir kelimeyi oluşturan alt kelimeler kullanılmıştır. Bileşik kelimelerin tekil ve çoğul formu için ayrı ayrı testler yapılmıştır. Test sonuçlarına göre, kelimeyi içeren dökümanlar, her iki alt kelimenin tekil formları ve birinci alt kelimenin tekil formu, ikinci alt kelimenin çoğul formu için indekslenmiştir. Aynı dökümanların ortalama olarak, birinci kelimenin tekil ikinci kelimenin çoğul formu ve her iki kelimenin çoğul formu için indekslenme oranları % 15 ile % 0 arasında değişmektedir.

Üçüncü kelime grubu son ek alan düzenli fiil kelime grubudur. Bu gruba ait kelimelerin düzenli fiil olması ve tablo 4.6'da yer alan ekleri almış olmaları gerekmektedir. Bu grup için yapılan testler sonucunda, kelimenin geniş zaman, geçmiş zaman ve şimdiki zaman formlarının indekslenme oranı diğer formlara göre daha yüksektir. Yani fiilin geçmiş zaman formu ile oluşturulan bir sorgu arandığında, yalın formunu içeren dökümanların sonuçlar arasında gelme ihtimali vardır. Bu oran bazı kelimeler için % 90 iken, bazı kelimeler için % 1 civarındadır.

Tüm testler göz önüne alındığında son ek alan fiil grubu için fiilin zaman çekimleri, diğer gruplar içinde tekil çoğul formu için indekslenme oranlarının yüksek olduğu görülmüştür. Yani, bir kelime arandığı zaman bu kelime grupları için, diğer formları içeren dökümanların gelme ihtimali yüksektir.

Genel olarak bakıldığında, test oranları her kelime grubuna ve her kelimeye göre farklılıklar göstermiştir. Farklı zaman dilimlerinde yapılan testler ile karşılaştırıldığında oranlar arasında çok fazla değişim olmadığı görülmüştür. Ayrıca arama sonuçlarında ilk sıralarda gelmeyen dökümanlar için indeksleme oranının düşük olduğu görülmüştür. Bu da, kelime gövdeleme algoritmasının kelimelere ve dökümanlara göre farklılık gösterdiği sonucuna ulaşılmasına neden olmuştur. Örneğin, bir döküman bir kelimenin tekil formu için indeksli iken, çoğul formu için indeksli olabilir yada olmayabilir.

Bu çalışma, bileşik kelime grubu ve son ek alan düzenli fiil kelime grubu açısından geliştirilebirir. Bu iki kelime grubu için ilişkili kelimeler kullanılarak tez kapsamında yapılan işlemler tekrarlanıp sonuçlar çıkartılabilir. Örneğin, bir bileşik kelimenin alt kelime formlarını içerip bileşik kelimeyi içermeyen dökümanların, bileşik kelime için indekslenip indekslenmediğine bakılabilir.

Ayrıca farklı diller açısından bu çalışma ele alınıp, Bing'in kelime gövdeleme algoritmasının o dil için nasıl çalıştığı incelenebilir. Daha önce Google için yapılmış kelime gövdeleme çalışması[8] ile karşılaştırılıp sonuçlar elde edilebilir. Özellikle farklı arama motorları açısından kelime gövdeleme işlemleri ile ilgili çalışmalar yapılabilir.

KAYNAKLAR

Manning, C. D., Raghavan, P., and Schütze, H., 2008: Introduction to Informatin Retrieval,Cambridge.

Levitin, A., 2006: Introduction to the Design and Analysis of Algorithms (2nd Edition), Pearson.

Rankins, R., Bertucci, P. T.,Gallelli, C., Silverstein, A. T., 2010 : Microsoft SQL Server 2008 R2 Unleashed, Sams.

Kılıç, A., 2008: Türkçe Dökümanlar için özelleştirilebilir web tabanlı dikey arama motoru, Anadolu Üniversitesi, Yüksek Lisans Tezi.

Kesgin, F., 2007: Türkçe metinler için konu belirleme sistemi, İstanbul Teknik Üniversitesi, Yüksek Lisans Tezi.

Knut, M. R., Michelsen, R. 2002: Search Engines And Web Dynamics, Computer Networks 39.

Url-1 < *http://dev.virtualearth.net/webservices/v1/searchservice/searchservice.svc*>, alındığı tarih 04.05.2012.

Url-1 < *http://en.wikipedia.org/wiki/Web_search_engine*>, alındığı tarih 04.05.2012.

Url-2 < *http://en.wikipedia.org/wiki/WebCrawler*>, alındığı tarih 04.05.2012.

Url-3 < *http://www.comp.lancs.ac.uk/computing/research/stemming/Links/algorithms.htm* >, alındığı tarih 04.05.2012.

Url-4 < *http://en.wikipedia.org/wiki/Stemming*>, alındığı tarih 04.05.2012.

[1] **Lovins, J. B.,** 1968: Delevopment of stemming algorithm, Mechanical Tasnlation and Computational Linguistic 11.

[2] **Porter, M.** 1980: An algorithm for suffix stripping, Program 14**.**

[3] **Vaughan, l. and Thelwall, M.,** 2003*:* Scholarly use of the web: what are the key inducers of links to journal web sites?, Journal of the American Society for Information Science and Technology 54.

[4] **Aguillo, I. F., Granadino, B., Ortega, J. L. and Prieto, J. A.,** 2006: Scientific research activity and communication measured with cybermetrics indicators, Journal of the American Society for Information Science and Technology 57.

[5] **Vaughan, L. and Thelwall, M.,** 2004: Search engine coverage bias: evidence and possible causes, Information Processing and Management 40(4).

[6] **Tawileh, W.** 2011: Exploring web search behavior of Arab internet users, International Conference on Innovations in Information Technology.

[7] **Ichikawa, Y. and Uehara, M.,** 2011: Distributed Search Engine for an IaaS based Cloud, International Conference on Broadband and Wireless Computing, Communication and Applications.

[8] **Uyar, A.** 2009: Google Stemming Mechanisms, Journal of Information Science OnlineFirst.

[9] **Cho, J., Garcia-Molina, H., Page, L.,** 2009: Efficient Crawling Through URL Ordering, Seventh International World-Wide Web Conference, Brisbane, Australia.
[10] **Najork, M. and Wiener, J. L.,** 2001: Breadth-first crawling yields high-quality pages, In Proceedings of the Tenth Conference on World Wide Web.
[11] **Srivastava, V., Kollipara, S. V., Tyagi, V. and Jaiswal, R. K.,** 2010: Share-ken: A Way To Improve Web Search, International Conference on Recent Trends in Information, Telecommunication and Computing.
[12] **Raval, V. and Kumar, P.,** 2011: EGG (Enhanced Guided Google) – A Meta Search Engine for Combinatorial Keyword, International Conference on Current Trends in Technology , IEEE.
[13] **Kim, J. and Carvalho , V. R.,** 2011: An Analysis of Time-Instability in Web Search Result, In Proceedings of the 33rd European Conference on Information Retrieval.
[14] **Edosomwan, J. and Edosomwan , T. O.,** 2010: Comperative Analysis of Some Search Engines, South African Journal of Science.
[15] **Kulasegarah, J. ,Harney, M., Walsh, M., and Walsh, R. M.,** 2011: The Quality of Information on Three Common ENT Procedures on The Internet, Iris Journal of Medical Sicence.
[16] **Odabaşoğlu, C.,** 2009: İnternet Arama Motorları Analizi, Haliç Üniversitesi Yüksek Lisans Tezi.
[17] **Knut, M. R., and Michelsen, R.,** 2002: Search Engines and Web, Computer Networks39.

ÖZGEÇMİŞ

Ad Soyad: Fatmana ŞENTÜRK

Doğum Yeri ve Tarihi: DENİZLİ 16.02.1986

Adres: Hacı Kaplanlar Mah. 1071 Sok. No:20 Kat:3 DENİZLİ

Lisans Üniversite: Pamukkale Üniversitesi Bilgisayar Mühendisliği - 2008