Akademik Makaleler için Özelleşmiş Yerel Arama Motoru

Ahmet Anıl Müngen, Emre Doğan, Gökhan Yılmaz, Sümeyye Kayaokay

Fırat Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği, Elazığ, Türkiye

(anmungen-emredogan-gokhan-skayaokay)@firat.edu.tr

Özet: Bu makale yerel konferanslar için açık kaynak kodlu yazılımlar ile hızlı ve kolay bütünleşebilen arama motorlu bir arama sistemi modeli önerisidir. Çalışma da birçok kurumsal firmanın kullandığı ElasticSearch arama motoru ile Akademik Bilişim ve Türkiye Internet Konferansının geçmiş tüm senelerinde yayınlanmış tüm makaleleri kapsayan bir arama altyapısı oluşturmaktır. ElasticSearch sayesinde bildiriler indekslenmekte ve kullanıcıların arama yapmasına imkân sağlanmaktadır. Bu sistem sayesinde başka bir yerde indekslenmemiş ve arama yapılamayan konferanslar için kolayca arama motoru destekli bir arama sistemi kurulabilmektedir. Böylece yerel konferanslardaki bildiriler arasında akademisyenlerin başarılı kolay ve hızlı şekilde arama yapması sağlanmaktadır.

Anahtar Sözcükler: Yerel Arama Motoru, Akademik Bildiri, Online Akademik Arama Motoru

Specialized Local Search Engine for Academic Articles

Abstract: This paper presents a model which includes easy integrated, fast and open source search engine for local conference. ElasticSearch search engine which a lot of big corporation used, are used to indexing all published articles from “Akademik Bilişim” and “Turkiye İnternet Konferansı”. All articles are indexed via ElasticSearch and provide a system which users can search on articles. As a result of this system, people can search on articles which are not included by international academic index. Therefore academics can easily and successfully make searching over local conferences.

Keywords: local search engine, academic article, online academic search engine

1. Giriş

Türkiye'de her yıl on binlerce akademik araştırma yapılmakta ve bilim geliştirilmektedir (Şekil 1A). Bunlarla ilişkili olarak Türkiye'de alınan patent sayıları 2011 itibari ile 10.000 sınırını geçmiştir. Türkiye Bilimsel yayım sıralamasında da Dünyada ilk 20'ye giren ülkelerden biri olmuştur.

Yıl	Bilimsel Yayın Sayısı	Milyon Kişi Başına Düşen Bilimsel Yayın Sayısı*
2000	6.977	103
2001	8.379	122
2002	10.807	155
2003	13.156	186
2004	16.300	227
2005	17.438	239
2006	20.092	272
2007	24.093	322
2008	24.846	327
2009	27.786	361
2010	28.194	362

* TÜİK tarafından açıklanan yıl ortası nüfus verileri kullanılmıştır. [1]

Şekil 1A. Türkiye Kaynaklı Bilimsel Yayın Sayıları

Yapılan bu kadar araştırmaların dışında yüksek lisans ve lisans seviyesinde de çalışmalar yapılmakta ve bu çalışmalar da dâhil edildiğinde yapılan çalışma-yayın sayısı her yıl yüz binlere yaklaşmaktadır.

2015 Ocak'a kadar 17 Akademik Bilişim Konferansı 18 tane de Türkiye'de İnternet Konferansı yapılmıştır. Bu konferanslarda binlerce bildiri yayımlanmış bu bildirilerin büyük bir kısmı halen ilgili konferansların sitelerinde mevcuttur.

Türkiye'de bilimin gelişmesine katkı sağlayan bu konferanslar bazı uluslararası makale indekslerinin bir kısmında bulunmasına rağmen uluslararası akademik arama motorlarının birçoğunda yer almamaktadır.

Türkiye'de bu kategoriye giren onlarca konferans mevcuttur, bu konferanslarda yayımlanan bildiriler arasında sağlıklı arama yapacak bir sistem mevcut değildir. Bu problem konferans bazlı yerel arama motorları kurulup tüm bildirilerin indekslenmesi ile çözülebilmektedir. Bu çalışma örnek bir yerel arama motorunun Akademik Bilişim ve Türkiye İnternet Konferansı bildirilerini kapsayacak şekilde özelleştirilmesi ile oluşturulmuştur.

Makale aşağıdaki bölümlerden oluşmaktadır. İlk bölüm giriştir ve Türkiye'deki akademik bildiriler ile ilgili istatistiki bilgi verilir. İkinci bölüm çalışmanın genel yapısını anlatmaktadır. Üçüncü bölüm yapılan uygulama projesinin tasarımı ve uygulamasını ifade eder. Diğer bölüm tartışma bölümüdür çalışma hakkında fikir ve düşünceler tartışılır. Son bölüm ise sonuç bölümüdür.

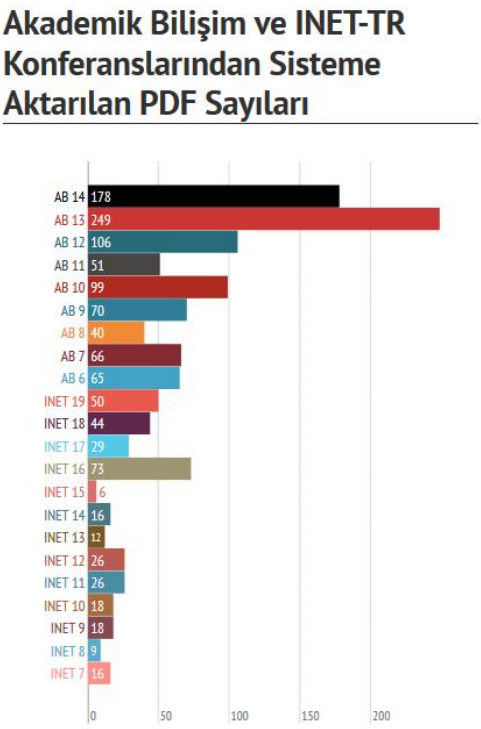
2. Genel Dizayn ve Sistem

Günümüzde piyasada; bazıları bir şirketin tescilli (proprietary), bazıları ise açık kaynak lisanslı olan birçok arama motoru yazılımı bulunmaktadır. Bunların açık kaynak lisanslı olanların en popüleri Solr'dır ve neredeyse hepsi Apache Lucene temellidir.

ElasticSearch Apache Lucene altyapısından geliştirilmiş, hafif, kolay kurulan, açık kaynak kodlu, ücretsiz, ölçeklendirilebilen bir arama motorudur. Restful API üzerinden hizmet veren bu arama motoru bazı 3.part görsel araçları ve güvenlik seçenekleri ile çok hızlı ve kullanışlıdır[2]. ElasticSearch Apache Lucene'den üretilmiş Solr arama motoruna göre daha az karmaşık, daha az ayrıntılıdır ama Solr kadar başarılıdır[3]. ElasticSearch daha önce "Mining Modern Repositories" konusu ile çıkan bir bildiride de kullanılmıştır[4].

Çalışma 'da konferans web sitelerinden veri çekmek ve ElasticSearch'a aktarmak için Java dili ile bir yazılım yazılmıştır. Bu yazılım Apache Tomcat üzerinde çalıştırılmıştır. Yazılım bildirileri indirip ElasticSearch'e eklemektedir. Çalışma ile toplam 480mb boyutundaki 1252

bildiri/dosya indekslenmiştir.



Şekil 2A. Konferans sitelerinden bulunan bildiriler.

3. Tasarım ve Uygulamalar

Çalışmada kullanılan işletim sistemi bir Linux dağıtımı olan Ubuntu Server yazılımıdır ve 512MB RAM, 40GB HDD ve 2.2GHZ Dual Core işlemciye sahiptir.

A. Veri Çekme

Hem Akademik Bilişim hem de Türkiye Internet Konferansında yayımlanan bildiriler konferans sitelerinde http://konferanssiteismi.com/koneransno/bildirino şeklinde bulunur. Sistem ilk konferanstan son konferansa kadar tüm konferanslar için 0-500 bildiri numaraları arasında PDF uzantısı ile arama yapar ve eğer varsa PDF'i dosya sistemine indirir.

B. Verileri Düzenlemek

Çekilen bildiriler dosya ismi, bildirinin yayınlandığı dergi ve bildirinin sıra numarası ile kaydedilir. Örnek olarak "ab14-95.pdf" ismi ile kaydedilir.

C. ElasticSearch'e Aktarım Yapmak

Tüm bildirilerin indirme işlemi bitince indirilen klasördeki tüm dosya isimleri listelenir ve döngü içinde tüm dosyalar pdf-box java kütüphanesi ile text'e dönüştürülür. Eş zamanlı olarak bildirinin meta bilgilerinin bulunduğu link (Örnek: http://ab.org.tr/ab14/ozet/201.html) sistem tarafından üretilir ve ilgili linke sorgu atılır. Eğer linkte bildirinin bilgileri mevcut ise bildirinin özeti, anahtar kelimeleri, yazar isimleri, bildirinin ilgili olduğu başlıklar kısımları ayrıştırılarak elde edilir. Elde edilmiş veriler PDF'den dönüştürülen tam içerik metni, dergi ismi ve bildiri numarası ile birlikte JSON'a çevrilir. Eğer konferans web sitesinde meta bilgi yok ise sadece PDF den dönüştürülen metin JSON'a eklenir. Üretilen JSON Elasticsearch'e gönderilir ve ElasticSearch ilgili veriyi gerçek zamanlı olarak indeksler.

D. WebSitesinde Gösterim

Web sitesin de anahtar kelime girişi, isteğe bağlı arama alanı seçme radyo butonları ve işlem butonu bulunur. Kullanıcı anahtar kelime giriş alanına yazacağı anahtar kelime ile tam metin içinde kolayca arayabilir. Kullanıcı isterse arama seçeneklerinden arama yapmak istediği (başlık, özet vb) alanı seçip sadece bu alanda da anahtar kelime ile arama yapabilmektedir. Anahtar kelime seçenekler doğrultusunda ElasticSearch'de aranır ve bulunan bildirilerin yayınlandığı konferans ve bildiri no'la ile varsa özeti ile başlığı aynı sayfada sonuçlar olarak sıralanır.

4. Tartışma

Bu çalışma ile yerel konferanslar için kurulumu kolay, hafif ve başarılı bir arama motoru ortaya çıkarılmıştır. Arama motoru ve birlikte çalıştığı tüm sistemlerin açık kaynak kodlu olmasından dolayı hiçbir lisans maliyeti yoktur. Sistem gereksinimleri olarak çok düşük özellikler ile çalışabilen bir sistem olduğu içinde sunucu maliyeti çok düşüktür.

Sistem şuan sadece PDF bildirileri çözebilmektedir. Buna karşın konferansların sitelerinde PDF olmayan bildiriler de mevcuttur.

ElasticSearch standart da herhangi bir güvenlik katmanı ile gelmediği için arama sistemine herkes ulaşabilmektedir, ekstra bir güvenlik katmanı ile arama kısmı yetkilendirilmelidir.

Çalışma Türkiye İnternet Konferansı ve Akademik Bilişim alanında özelleştiği için meta bilgileri de aynı başvuru numarası ile konferans sitelerinden alabilmektedir. Buna karşın AB08'den ve INET16'dan önce bildirilerin konferans web sitesi tarafından meta bilgileri verilmemektedir. Sistem bu bildiriler için sadece tam metin alanını beklendiği gibi doldurup, özet anahtar kelime gibi bilgileri boş olarak ElasticSearch'e yüklemektedir. Kullanıcı başlık ya da özet alanında arama yaptığında bu bildiriler sisteme dâhil olamamaktadır. Bu sorunun çözümü için meta bilgisi gelmeyen bildirilerin başlık ve özet gibi boş olan kısımlarına da tam metni yerleştirip gönderildiğinde bölüm bazlı arama işlemi tam yapılmamış olmakta ve fazladan aynı veriler birden fazla kere sisteme eklenmiş olmaktadır.

Sistem openconf altyapısı ile çalışan konferanslar için uyum gösterebilmektedir buna karşın openconf altyapısı değiştiğinde veya konferans başka bir bildiri/makale yönetimi altyapısı kullandığında sistemin desteği kalkacaktır. Bu probleme karşın sistemde yönetici tarafından elle PDF yükleme gibi bir hizmet geliştirilebilir. Bu hizmet ile yeni altyapıya uyum sağlayana kadar yönetici elle bir konferans bildirileri yükleyebilmektedir.

Sistemin başka bir eksikliği mükerrer veri kabul ediyor olmasıdır. Aynı bildiriyi birden fazla kere indexlenebilir ve sistem bunu engellemez. Sonuç olarak da birden fazla aynı sonucu ayrı ayrı gösterilebilir. Bunların dışında, Jöran Beel, Bela Gipp, Erik Wilde tarafından yapılmış ve "Academic Search Engine Optimization" ismi ile yayınlanmış bildiri, akademik bildiri içindeki verilere dayalı arama algoritması önerisinde birçok önemli tespit ve yöntem içermektedir[5].

5. Sonuç

Sonuç olarak özelleştirilmiş yerel arama motorları ile konferans'a özel arama sayfaları ve/veya indekslenmemiş konferanslar için başarılı arama altyapısı kolayca oluşturulmuştur.

6. Referanslar

[1] Thomson's ISI Web of Science Veritabanı: SCI-EXPANDED, SSCI, A&HCI SCI-EXPANDED: Science Citation Index Expanded SSCI: Social Sciences Citation Index A&HCI: Arts & Humanities Citation Index

[2] Elasticsearch Wikipedia, the free encyclopedia, 25.11.2014, http://en.wikipedia.org/wiki/Elasticsearch

[3] Apache Solr vs ElasticSearch - the Feature Smackdown!, 25.11.2014, http://solr-vs-elasticsearch.com/

[4] Oleksii Kononenko, Olga Baysal, Reid Holmes, and Michael W. Godfrey. Mining Modern Repositories with Elasticsearch. In Proceedings of the Working Conference on Mining Software Repositories (MSR). 2014. 328-331.

[5] Jöran Beel, Bela Gipp, Erik Wilde. Academic Search Engine Optimization (ASEO).Volume 41, Number 2 / January 2010. Journal of Scholarly Publishing. 2009