Akademik Çalışmalara Hakem Ataması için Büyük Veri Altyapısı

Ahmet Yavuz, Süleyman Eken, Ahmet Sayar

Bilgisayar Mühendisliği, Kocaeli Üniversitesi,
41380 İzmit, Türkiye
yavuz1ahmet@gmail.com,
{suleyman.eken, ahmet.sayar}@kocaeli.edu.tr

Özet. Birçok karmaşık işlemin yapılmasını gerektiren akademik çalışma (makale, bildiri vs.) değerlendirme işlemi, çalışmanın uzmanlarca yeterliliğinin, öneminin ve orijinalliğinin değerlendirilmesi için yapılmaktadır. Akademik çalışmalara alanlarına uygun hakemlerin atanması bu işlemlerden biridir. Akademik çalışma yazarları; çalışmalarını, yapılan çalışmanın başlığı, özeti ve konferans konularından hangilerine uygun olduğu gibi bilgiler ile birlikte konferans yönetim aracılığı ile konferans yönetimine bildirirler. Hakemler ise konferans yönetimine uzmanlık alanlarını belirterek kayıt işlemlerini gerçekleştirirler. Konferans başkanları da yazarlar ve hakemler tarafından gönderilen bilgilere göre her akademik çalışmaya manuel olarak hakem ataması yapmaktadırlar. Bu işlem oldukça zaman almakta ve can sıkıcı hale gelebilmektedir. Geliştirilen yazılım ürünü ile birlikte hakem atama işleminin otomatik olarak ve dağıtık mimaride birçok akademik çalışmanın hızlı bir şekilde değerlendirmesi mümkün hale gelmiştir.

Anahtar kelimeler: Otomatik hakem atama, MapReduce, dağıtık yazılım geliştirme, NoSQL, CouchDB

Abstract. Peer review is an evaluation process for the competence, significance and originality of researches by experts. Assigning qualified and relavant reviewers to papers is one of the such diffucult tasks. Authors submit the papers with their paper title, abstract, and keywords. Reviewers are needed to register and declare their expertise on the conference topics. Eventully, the conference chair has to realize assignment process by regarding information provided by both the authors and the reviewers. All these tasks are now done by all actors manually. This process takes quite some time and can become annoying. With developed distributed system, paper-to-review assigment is done automatically and evaluation process of many papers become possible in seconds.

Keywords: Otomatik hakem atama, MapReduce, dağıtık yazılım geliştirme, NoSQL, CouchDB

1 Giriş

Akademik çalışma (makale, bildiri vs.) değerlendirme işlemi çalışmanın uzmanlarca yeterliliğinin, öneminin ve orijinalliğinin değerlendirilmesi için yapılmaktadır [1]. Daha öncesinde yazarlar, akademik çalışma ile ilgili bilgileri (başlık, özet, anahtar kelimeler vs.) sisteme yüklerler. Değerlendirme sürecinde hakemler; araştırmanın bilimsel hatalarını saptama yoluyla geçerliliğini kontrol etme, sonuçların önemini değerlendirerek yapacağı katkıyı araştırma ve özgünlüğünü kontrol etme gibi bir takım işlemler yapmaktadırlar. Yapılan değerlendirmeler sonucunda çalışmalar, kabul edilmekte veya reddedilmektedir. Birçok bildirinin/makalenin kısa sürede değerlendirilmesi gereken durumlarda hakem atama, değerlendirme ve sonuçların alınması çok zaman alan bir süreç olmaktadır. Çalışmanın amacı, efektif bir biçimde uygun hakemlerin hızlı bir biçimde otomatik olarak atanmasını sağlayan dağıtık bir mimarinin geliştirilmesidir.

Hali hazırda var olan uygulamalarda akademik çalışma gönderim süreci başlamadan önce konferans başkanı veya editörler, konferans veya derginin ilgilendiği konuları belirlerler. Daha sonra hakemlerden ilgilendikleri konuları belirlemeleri istenir. Çalışmanın gönderilme sürecinde ise yazarlardan konferans/dergi konularından hangi(leri) ile çalışmanın uyuştuğu bilgisi alınır. Anlatılan bu sistemde yazarların belirlediği konular çalışmanın konusu ile alakasız olabilmekte bu da hakem atama ve değerlendirme sürecinin uzamasına sebep olmaktadır [2]. Bu çalışmada, birçok akademik dokümanın hızlı bir şekilde ilgili konularıyla ilişkilendirilmesi ve indekslenmesi işlemi gerçeklenmiştir. Böylece bilimsel konferans sistemlerine geliştirilen modülle akıllı bir çözüm getirilmiştir. Bu çalışma ile aşağıdaki katkılaı sağlanmıştır:

— Akademik çalışmanın başlık, özet ve metninden doğal dil işleme yöntemleri ile konusunun tespitinin yapılması,
— Belirlenen çalışma konusuna göre alanı/konusu belli, sisteme kayıtlı hakemlerin atanması veya DBLP bibliyografyasındaki çalışma başlıklarına göre hakemlerin bulunması,
— Büyük veri analizi kapsamında birçok çalışmanın aynı anda değerlendirilmesi,

Çalışmanın geri kalanı şu şekilde organize edilmiştir. 2 bölümde, dağıtık mimari tabanlı olarak akademik çalışmalara hakem atanması yazılımı detaylı bir şekilde anlatılacaktır. 3. bölümde, önerilen mimari üzerinde gerçekleştirilen birtakım testler verilip tartışılacaktır. 4. bölümde literatürde hakem ataması ile ilgili var olan çalışmalar irdelenip son kısımda ise sonuçlar sunulacaktır.

2 Metodoloji

Geliştirilen sistem temel olarak iki modülden oluşmaktadır: (1) Sistem kayıt ve yönetim modülü ve (2) Akademik çalışmaların değerlendirilmesi ve hakem ataması modülü. İlerleyen kısımlarda her bir aşama detaylı olarak açıklanacaktır.

2.1 Sisteme Kayıt ve Yönetim Modülü

Sisteme kullanıcılar rollerine gore (editor, hakem ve yazar) kayıt olabilmektedir. Hakem kayıtlarında, kişisel ve iletişim bilgileri ile birlikte hakem çalışma konularının da sisteme girilmesi istenmektedir. Yazar rolündeki kişilerden ise kişisel ve iletişim bilgileri ile birlikte yaptıkları akademik çalışmayı (pdf veya word formatında) sisteme yüklenmesi beklenmaktedir. Tüm bilgiler ve akademik çalışmalar, Javascript Obje Notasyonu (JSON) formatında NoSQL veritabanlarından doküman tabanlı CouchDB veritabanı üzerinde tutulmaktadır. NoSQL kavramı 2009 sonları 2010 başlarında ortaya çıkmıştır. Bu sistemler klasik veritabanlarından büyük verileri kontrol edebilme, ölçeklenebilme, veri formatları, yönetebilme, sorgulara daha hızlı cevap alabilme, eş zamanlı kayıt ve güncelleme işlemlerini gerçekleyebilme, açık kaynak kodlu geliştirme sağlayabilme gibi yönlerden farklılıklar arz etmektedir [3, 4].

2.2 Akademik Çalışmaların Değerlendirilmesi ve Hakem Atanması Modülü

Bu kısım çalışmanın özgün yanını oluşturmaktadır. Konuların tespit edilmesi dağıtık mimari yapısında gerçeklenmiştir. Burada MapReduce programlama paradigmasından yararlanılmıştır [5]. MapReduce uygulaması için Apache açık kaynak Hadoop projesinden faydalanılmıştır.

Çalıştırılacak algoritmanın çalıştığı düğüm ile kullanacağı verinin farklı makinelerde olması dağıtık sistemlerde paralel işlemede yavaşlığa neden olan bir durumdur. HDFS'de [6] işleyicinin çalışacağı düğüm ile işleyeceği veri setinin bulunduğu düğümün yakın olması hedeflenmiştir. Veri yerelliği denen bu durum paralel işleme hızını artırmaktadır. MapReduce, dağıtık dosya sisteminin üstünde olduğundan veri yerelliği avantajından yararlanmaktadır. Bu modüle ait alt adımlar aşağıdaki gibidir:

1. Akademik çalışmadaki konular; direk çalışma başlığından veya özetinden elde edilebildiği gibi metin içinde en çok tekrar eden kelimelerin bulunması ve zarf-edat gibi etkisiz kelimelerin (stop words) elenmesi ile geriye kalanlar şeklinde de elde edilmiştir.
2. Akademik çalışma ataması yapılacak hakemlerin yaptıkları çalışma başlıkları dikkate alınarak tespit edilmesi gerçekleştirilmiştir. Burada bilgisayar bilimleri yayınların bir çoğunun tutulduğu DBLP kütüphanesinden [7] yararlanılmıştır. DBLP verilerine resmi sitesinden erişmek mümkündür. XML formatında verilen bilgileri ayrıştırarak yayın başlıklarından hakem bilgisine erişilmiştir.
3. Bu aşamada elimizde şu şekilde kelime kümeleri vardır: hakem ataması yapılacak yayına ait ayırt edici anahta kelimeler $\{Y_1, Y_2, ...Y_m\}$ ve bir önceki adımda elde edilen herbir çalışmaya ait kelimeler $\{H_1, H_2, ...H_n\}$. Yukarıdaki örnek için B seti şunlardır: {kd-tree, and, quad-tree, decompositions, for, declustering, of, 2D, range, queries, over, uncertain, space,}. Hakem ataması işlemi Eşitlik 1'de verilen benzerlik faktörüne (BF_{YmHn}) göre yapılmaktadır [8].

$$BF_{YmHn} = \frac{adet(Y_m \cap H_n)}{adet(Y_m \cup H_n)} \quad (1)$$

3 Deneysel Sonuçlar ve Tartışma

Sonuç olarak geliştirdiğimiz yazılım sayesinde akademik çalışmalara otomatik olarak hakem ataması yapılabilmektedir. Bu eşleştirme yapılırken akademik çalışma içerisinde en fazla geçen kelimeler, özet ve başlık bölümlerindeki kelimeler dikkate alınmaktadır. Her bir kelime grubu üzerinde ayrı ayrı eşleştirme yapılabildiği gibi bunların kombinasyonları dikkate alınarak atama yapılabilir. Örnek veriseti olarak 2014 IEEE 27th Conference on Software Engineering Education and Training (CSEE&T) konferansında yer alan çalışmalar kullanılmıştır. Bu çalışmalar pdf formatında olup 35 tanedir (60 MB). Her bir çalışmanın konusu bölüm 2.2'de anlatıldığı gibi başlık, özet ve metinden elde edilmiştir. Bu çalışmalara hakem ataması yapmak için hakemler (64 kişi) konferansın web sayfasında yer alan komiteden elde edilmiştir (http://conferences.computer.org/cseet/2014/). Hakemlere ait ilgi alanları DBLP tarafından sunulan XML dosyasından elde edilmiştir. Atama işlemi çalışmanın metni, başığı ve özetine göre ayrı ayrı değerlendirildiği gibi hepsini birden dikkate alarak yapılmıştır. Toplam 35 dokumana hakem atama işleminin ne kadar sürede yapıldığı Tablo 1'de özetlenmiştir. Burada hakem sayısının doküman sayısından fazla olması herhangi bir sorun teşkil etmemektedir. Önemli olan hakemlerin çalışma alanlarına uygun olarak atama işleminin yapılabilmesidir.

Tablo 1. Performans karşılaştırması

Atama Kriteri	Ortalama Atama Süresi (ms)
Özet, Başlık ve En Fazla Geçen Kelimelere Göre Atama	4414
Özet Kısmına Göre Atama	4247
En Fazla Geçen Kelimelere Göre Atama	82
Başlığa Göre Atama	63

En fazla sürede özet, başlık ve metinde geçen kelimelere göre atama yapılmasının sebebi hakem çalışma alanları ile benzerlik araştırmasının yapıldığı kelime sayısının daha fazla olmasındandır. Diğerleri de kelime sayılarının karşılaştırılmasına göre değişik süreler tutmuştur. Bu sistem sayesinde akademik çalışmalara hakem ataması konusunda büyük zaman kazanımı sağlanmıştır. Bu kolaylık ile birlikte bu sistemin geliştirilmesi aşamasında büyük verilerin analizi konusunda bilgi birikimi edinilmiş ve deneyim kazanılmıştır [9, 10].

4 İlgili Çalışmalar

Genel olarak hakem atama yöntemleri literatürde iki kategoride incelenmektedir: (1) tercih tabanlı ve (2) konu tabanlı veya bilgi bulma tabanlı yaklaşımlar. Tercih tabanlı sistemlerde genelde sisteme yüklenen tüm çalışmalar veya hakemlerin ilgi alanlarına kısmen daha yakın olanlar listelenir. Hakemlere bu çalışmalar gönderilir, onlardan ilgilendiklerini seçmeleri beklenir. Bu sistemin zayıf yönü, çalışmalar gön-

derildikten sonra hemen hakemler tarafından değerlendirmeye kabul edilmemeleri ve sürecin iyice uzamasıdır. Rigaux [11] işbirlikçi filtrelemeye dayalı tercih tabanlı bir sistem önermiştir. Temel mantık, aynı akademik çalışmaları değerlendirmeyi kabul eden hakemler muhtemelen diğer çalışmalar için de aynı tercihte bulunacaktır şeklindedir. Konu tabanlı yaklaşımda, belirli bir derecede akademik çalışmanın konusundan haberdar olan hakemlere çalışmalar atanmaktadır. Bu yaklaşımda her hakem ilgili çalışmanın konularına göre derecelendirilir ve en yüksek puana/dereceye sahip hakem(ler) çalışmayı incelemek için atanır. Buradaki problemse akademik çalışmanın kapsadığı konuları otomatik olarak sapmaktır. Literatürde ikinci kısımla ilgili çalışmalarda genellikle akademik çalışmanın özetinden ve anahtar kelimelerinden yararlanılmıştır. Dumais ve Nielsen [12] hakem atama işlemini, terim ve metinler arasındaki kalıpları bulmaya çalışan Gizli Anlamsal İndeksleme yöntemi ile gerçeklemişlerdir. Basu ve diğ. [13] çalışmalarında web üzerinden arama motoru yardımıyla potansiyel hakemler tarafından yazılan çalışmalardan özetler çıkarılıp Vektör Uzay Modeli ile eşleştirme işlemi gerçekleştirilmiştir. Literatürde ayrıca her iki yaklaşımı kullanan çalışmalar da mevcuttur [14]. Yaptığımız araştırmalar neticesinde literatürdeki çalışmalarda akademik çalışmaların hakem ataması sıralı olarak yapılmaktadır. Yani sisteme yüklenen her bir çalışma için önerilen yöntemler çalıştırılmaktadır. Bu çalışmada ise konu tabanlı olarak dağıtık sistem mimarisinde birçok çalışma için aynı anda hakem ataması yapılmaktadır.

5 Sonuçlar

Geliştirilen yazılım sayesinde akademik çalışmalara hakem ataması konusunda büyük zaman kazanımı sağlanmıştır. Bu kolaylık ile birlikte bu sistemin geliştirilmesi aşamasında büyük verilerin analizi konusunda bilgi birikimi edinilmiş ve deneyim kazanılmıştır. Benzer şekilde hakem ataması yapılacak yayının anahtar kelimeleri de işin içine katılabilir veya yayından elde edilen kelime setinin anahtar kelimelere olan uygunluğu araştırılabilir. Ayrıca sisteme kayıtlı hakemlere adil veya adile yakın sayıda çalışmayı incelemesi için gönderilmesi gelecek çalışmalar arasındadır. Kullanılan benzerlik faktörü, sadece iki kelime seti içinde geçen ortak kelimelere bakmaktadır. Kelimelerin üst çalışma alanlarında ortaklığı da dikkate alınarak daha gürbüz bir benzerlik indeksi elde edilebilir.

Kaynaklar

1. B.T. Sense, “Peer Review and the Acceptance of New Scientific Ideas”, Sense about Science, ISBN: 0-9547974-0-X, 2004.

2. S. Ferilli, N. Di Mauro, T.M.A. Basile, F. Esposito, and M. Biba, “Automatic topics identication for reviewer assignment”, Lecture Notes in Computer Science, 4031: 721-730, 2006.

3. J. Lennon, Beginning CouchDB. Apress Publisher, 1st edition, NY:U.S.A., 2009.

4. S. Eken, F. Kaya, A. Sayar, A. Kavak, "Doküman Tabanlı NoSQL Veritabanları: MongoDB ve CouchDB yatay ölçeklenebilirlik karşılaştırması," 7. Mühendislik ve Teknoloji Sempozyumu, 2014.

5. Dean, J., Ghemawat, S., MapReduce: A Flexible Data Processing Tool, Communications of ACM, 53(1), 72-77, 2010.

6. Hadoop, http://hadoop.apache.org/ (Erişim Tarihi: 1 Haziran 2016)

7. dblp: DBLP Bibliography. http://www.informatik.uni-trier.de/~ley/db/ (Erişim Tarihi: 1 Eylül 2015)

8. Y. Kalmukov, "An algorithm for automatic assignment of reviewers to papers", International Conference on Computer Systems and Technologies, pp. 1-7, 2006.

9. G. C. Fox, ve diğ. "Algorithms and the Grid", Computing and visualization in science, pp. 115-124, 2009.

10. M. Aktaş, ve diğ. "Implementing Geographical Information System Grid Services to Support Computational Geophysics in a Service-Oriented Environment", in Proc.of NASA Earth-Sun System Technology Conference, pp. 1-9, 2005.

11. P. Rigaux: "An Interactive Rating Method: Application to Web-based Conference Management", in Proc.of ACM Symposium on Applied Computing, pp.1682-1687, 2004.

12. S.T.Dumais and J.Nielsen: "Automating the Assignment of Submitted Manuscripts to Reviewers", in Proc.of 15th ACM International Conference on Research and Development in Information Retrieval, pp.233-244, 1992.

13. C.Basu, H.Hirsh, W.Cohen and C.N.Manning: "Technical Paper Recommendation: A Study in Combining Multiple Information Sources", Journal of Artificial Intelligence Research, pp.231-252, 2001.

14. X. Li, T. Watanabe, "Automatic Paper-to-reviewer Assignment, Based on the Matching Degree of the Reviewers", Procedia Computer Science, 22: 633-642, 2013.