Hissiyat Odaklı Ağ Tarama

Proje No: 112E002

Yürütücü:

Doç.Dr. Pınar KARAGÖZ

NİSAN 2013
ANKARA

ÖNSÖZ

Bu çalışmada günümüzde önemli bir araştırma konusu olan hissiyat ve görüş (sentiment) içeren web sayfalarının hızlı keşfini amaçlayan hissiyat-odaklı ağ tarayıcı tekniğinin geliştirilmesi ele alınmıştır. Bunun yanı sıra Türkçe hissiyat analizi için belli bir seviyede de araştırma yapılmıştır. Çalışma kapsamında bir doktora çalışması yürütülmüş ve iki konferans bildirisi yayınlanmıştır. Bir dergi makalesi hazırlanmıştır. Çalışma 112E002 proje numarası ile TÜBİTAK ARDEB 1002 programı tarafından desteklenmektedir.

İÇİNDEKİLER

ŞEKİL LİSTESİ

ÖZET

Günümüzde bir konu üzerinde hissiyat ve görüş (sentiment) içeren web sayfalarının (örneğin bloglar, köşe yazıları gibi) sayısı hızla artmaktadır. Bu tip sayfalar güncel konulara odaklandığından hızlı olarak bulunmaları önemlidir, ancak mevcut tarayıcılar hissiyat içeren sayfaların keşfedilmesinde ve sorgulanmasında yetersiz kalmaktadır. Genel amaçlı tarayıcılar web kaynaklarına ulaşmada yoğun olarak kullanılmaktadır. Bu tarayıcılar erişilebilen sayfaları gezerek dizinlemekte, bu sayede anahtar kelime tabanlı sorgularda oldukça başarılı çalışmaktadırlar. Ancak genel amaçlı tarayıcılar özellikle hissiyat içeren sayfaların bulunması için kurgulanmamıştır. Odaklı tarayıcı (focused crawler), bir konuya yönelik web sayfalarını normal tarayıcılardan daha hızlı bir şekilde keşfetmek ve indirmek üzere tasarlanmış özel bir tarayıcıdır. Odaklı ağ tarayıcılar, belli bir sayfayı indirmeden önce o sayfaya olan bağlantının verilen konu ile ilgisinin olasılığını kestirerek konu odaklı çalışmaktadır. Bu yaklaşım da hissiyat içeren sayfaların keşfedilmesinde yetersiz kalmaktadır. Bu projede, web üzerinde hissiyat içeren kaynakların daha hızlı keşfi ve toparlanması için hissiyat odaklı web tarayıcı üzerinde çalışılmıştır. Geliştirdiğimiz teknik, sayfayı indirmeden hissiyat içerik değerinin tahminlemesine dayanmaktadır. Yaptığımız deneylerde önerdiğimiz teknikle bilinen ağ tarama stratejilerine göre hissiyat içerikli sayfaların çok daha hızlı keşfedildiği ortaya konmaktadır.

Anahtar Kelimeler: Hissiyat Analizi, Odaklı Web Tarayıcı, Veri Madenciliği, Sınıflandırma, Support Vector Machine (SVM)

ABSTRACT

The number of web pages that have sentimental content on a certain topic (such as blogs and newspaper articles) are increasing tremendiously. Since such pages focus on contemporary events and topics, it is important to discover them fast and in a timely way. However conventional web crawlers do not provide satisfactory results in crawling and querying such content. General web crawlers have high popularity for keyword based search. Such crawlers navigate the web, get accessable web pages and index them in terms of terms/words in the content. Therefore they are successful for keyword based search but not designed to detect and accumulate sentimental content. Conventional focused-crawlers, on the other hand, are designed to detect the pages on a given topic, get such pages and index them. Yet, they do not detect sentimental content, either. In this project, we worked on design of a sentiment-focused web crawler. The technique we propose for detecting sentimental web pages is based on estimation of the sentimant score without downloading the page itself. The experiments conducted within the scope of this study show that the sentimental web pages are detected and accumulated faster than conventional web crawlers.

Keywords: Sentiment Analysis, Focused Web Crawler, Data Mining, Classification, Support Vector Machine (SVM)

1. GİRİŞ

Günümüzde, ortak içerik oluşturma teknolojileri sayesinde kullanıcılar daha fazla web içeriği oluşturabilir hale geldiler (Godbole et al., 2007; Kucuktunc et al., 2012). Film, kitap gibi sanat ürünleri hakkında yorumlar, teknolojik ürünler hakkında yorum ve öneriler, güncel konular hakkında yorumlar web sayflarında sıklıkta yer almakta ve bu içerikler yoğun olarak kişisel görüş, yorum ve duyguları içermektedir. Web'deki metin içeriklerinden duygu ve görüşlerin otomatik olarak çıkartılması sonucu ortaya çıkacak bilgi, öneri sistemleri, arama motorları, reklam sektörü gibi pek çok alan için kullanılabilir.

Bu motivasyonla hissiyat analizi son zamanlarda yoğun olarak ele alınan bir araştırma konusu haline geldi. Web içerikleri üzerinde hissiyat analizi probleminin içerik sınıflandırması, sonuçların sunumu, bilgi çıkarımı gibi çeşitli yönleri üzerinde çalışmalar bulunmaktadır (Abbasi et al., 2008; Bai, 2011; Yi et al., 2003; Dave et al., 2003; Pang et al., 2002; Turney, 2002; Gerani et al., 2009; Zhang et al., 2007; Beineke et al., 2004; Lerman et al., 2009; Gregory et al., 2006). Hissiyat içeren sayfalar genellikle güncel konulara odaklandıkları için bunların hızlı olarak keşfi ve dizinlenmesi, böylece hissiyat odaklı arama yapılabilmesi önem taşımaktadır. Arama motorları günlük hayatta yoğun olarak kullanılan araçlar olmasına karşın, hissiyat odaklı web tarama üzerinde araştırmalarda eksiklik bulunmaktadır.

Bu çalışmada, gözlemlenen bu eksiklikten hareketle, hissiyat odaklı ağ tarama problem ele alınmıştır. Çalışma kapsamında bir çerçeve oluşturularak, çeşitli hissiyat odaklı ağ tarama metotları geliştirildi. Önerdiğimiz metot temel olarak eldeki bir URL için sayfanın içerriğini görmeden hissiyat yönü ve derecesi için tahminleme yaparak arama kuyruğunda uygun şekilde yerleştirilmesine dayanmaktadır. Böylece hissiyat içerikli sayfaların önceliği arttırılarak daha kısa sürede bu tip içeriklerin taranabilmesi amaçlanmıştır.

Değerlendirme amacıyla genel amaçlı web tarama teknikleri ile simülasyon yolu ile karşılaştırma yapıldı. Deneylerde, akademik olarak benzer bilgi çıkarımı araştırmalarında sıklıkla kullanılan ClueWeb09-B veri seti kullanıldı (ClueWeb09 – TREC 2009 "Category B" dataset, http://lemurproject.org/clueweb09.php). Yaptığımız deneylerde önerdiğimiz yöntemin temel yöntemlere göre hissiyat içeriklerini daha hızlı taradığı ortaya kondu.

Çalışma sırasında hissiyat analizi için elde edinilen birikimle Türkçe metinler için hissiyat analizi için de kısa bir çalışma yapıldı. Hissiyat odaklı ağ tarayıcı kapsamında kullandığımız dil/sözlük tabanlı bir araç olan SentiStrength'in (SentiStrength, http://sentistrength.wlv.ac.uk) sözcük ve kurallarının bir alt kümesi Türkçe için uyarlanarak performansı değerlendirildi.

Çalışma kapsamında bir doktora tezi yürütülmüştür. ODTÜ Bilgisayar Mühendisliği Bölümü Doktora Programı öğrencisi Güral Vural projede bursiyer olarak yer almış ve tez çalışmalarına devam etmiştir. Halen tez çalışması devam etmekle birlikte 2013 içinde tez savunması planlanmaktadır. Çalışma konusunda Yahoo Research'de çalışan Dr. Barla Cambazoğlu ile işbirliği yapılmış, kendisi tez çalışmasında eş-danışman olarak yer almıştır.

Bu çalışmanın literatüre katkılarını şöyle sıralayabiliriz:

— Güncel web sayfası işleme ve hissiyat analizi araçları kullanılarak hissiyat-odaklı ağ tarayıcı tasarlandı.

— Web sayfalarının hissiyat derecelerinin tahminlemesi için çeşitli teknikler geliştirildi. Bu teknikler arama kuyruğunda önceliklendirme için kullanılarak performansları karşılaştırmalı olarak değerlendirildi.

— Önerilen teknikler ClueWeb09 veri kümesi üzerinde simülasyon yoluyla deneysel olarak değerlendirilerek temel ağ tarama tekniklere göre üstünlüğü ortaya kondu.

— Performans değerlendirmesinde referans alınacak web sayfalarının hissiyat puanı standardını oluşturmak için bir anket çalışması yapıldı.

— Türkçe hissiyat analizi için dil/sözlük tabanlı bir çözüm için sözlük ve kural uyarlaması yapılarak Türkçe metinler üzerinde çalışması sağlandı.

Çalışmada elde ettiğimiz bulguları şöyle özetleyebiliriz:

— Hissiyat içerikli sayfalar arası bağlantılar hissiyat odaklı taramayı kolaylaştırmaktadır.

— Sayfaların hissiyat puanını sayfa içeriğini görmeden bağlantı ve bağlantı veren sayfanın özellikleri kullanılarak kabul edilebilir bir başarı oranı ile tahminlemek mümkün olmaktadır.

— Önerdiğimiz hissiyat odaklı web tarama yaklaşımı hissiyat içerikli sayfaların keşfi ve içerik çıkarımı için geleneksel ağ tarama yaklaşımlarına göre çok daha başarılı sonuç vermektedir.

— Dil/sözlük tabanlı bir hissiyat analizi çözümünün sözlük ve kural kümesi üzerinde kısmi bir uyarlama ile film yorumları içeren Türkçe metinler için yaklaşık %80 başarı ile analiz yapmak mümkün olmuştur. Uyarlanan kümenin geliştirilmesi ile başarı oranını yükseltilme potansiyeli bulunmaktadır.

2. GENEL BİLGİLER

2.1. Odaklı Ağ Tarama

Ağ tarama, web sayfalarının keşfedilmesi, indirilmesi ve saklanması işlemlerinin bir bütünü olarak tanımlanmaktadır. (Olston and Najork, 2010). Ağ tarama sistemi ana hatlarıyla şöyle çalışmaktadır. Tarama, belli bir URL kümesi ile başlar. Sırayla her URL'in gösterdiği sayfaya erişilir, içerik ayrıştırılıp analiz edilir. İncelenen sayfadaki yeni URL'ler alınarak, daha önce ziyaret edilmemiş sayfaları gösterenler listeye (kuyruğa) eklenir. Listedeki URL'lerin ziyaret sırası genellikle belli bir önem sırasını takip eder. Daha önemli olarak değerlendirilen sayfaların kuyruktaki önceliği arttırılır. Belli sayıda sayfa taranıncaya kadar ya da listedeki URL'ler tükeninceye kadar işlem devam eder.

Odaklı ağ taramada ise amaç belli bir konu ya da tema ile ilgili sayfaların keşfedilmesi ve saklanmasıdır (Novak, 2004). Ele alınan tema bir sayfa türü (örneğin, blog sayfaları) ya da içerik türü (örneğin, görüntü ya da ses içerikleri) olabilir. Genel amaçlı ağ tarayıcılardan farklı olarak odaklı ağ tarayıcılarda konu ya da tema ile ilgili URL sayfalara öncelik verilir. Böylece ilgili sayfaların daha önce taranması amaçlanır. Bunu sağlamak üzere genellikle sayfayı indirmeden konuyla ilgisini tahminlemek gerekli olur. Bunun için URL'ye bağlantı veren

sayfanın ve URL'nin özellikleri üzerinden bir tahminleme veya sınıflandırma tekniği uygulanır.

2.2. Hissiyat Analizi

Hissiyat analizi, eldeki metinden olumlu ya da olumsuz olarak yansıtılan ruh hali ve düşünceyi belirleme ve derecelendirme işlemidir (Pang and Lee, 2008). Bu alandaki çalışmalar makina öğrenme tabanlı ya da dil/sözlük tabanlı olarak iki grupta toplanabilir. Makina öğrenme tabanlı yaklaşımda öğrenme için kullanılan metin kümesi hissiyat yönü (olumlu/olumsuz) ve derecesi için notlandırılır. Notlandırılmış bu metin kümesi üzerinden bir model oluşturulur (Pang et al. 2002). Elle notlandırmanın oldukça zahmetli bir işlem olması nedeniyle öğrenme amaçlı büyük metin kolleksiyonları oluşturmada zorluklar yaşanmaktadır. Sözlük tabanlı yöntemler, karşılaşılan bu zorluk sonucu ortaya çıkmıştır (Baccianella and Sebastiani, 2010; Thelwall et al., 2010). Sözlük tabanlı yaklaşımlarda bir kelime kümesi ve kural kümesi oluşturularak eldeki metinin hissiyat yönü ve derecesi bunlar üzerinden belirlenir. Kelime ve kural kümesi uzman bir grup tarafından, genellikle konu ya da alandan bağımsız olarak genel amaçlı hazırlanır.

2.3. Benzer Çalışmalar

Bu kısımda odaklı ağ tarama ve hissiyat analizi konularında yapılan ilgili çalışmalar özetlenerek sunulmuştur. İlk once odaklı ağ tarama, daha sonra hissiyat analizi ve son olarak de her ikisini birleştiren çalışmalar anlatılmıştır.

Odaklı ağ tarayıcılar belli bir konu ya da tema ile ilgili sayfaların taranmasını amaçlar. Konu genellikle önceden notlandırılmış örnek sayfa ya da dokümanlar üzerinden modellenir. Tarama sırasında bu model kullanılarak henüz içeriği indirilmemiş sayfalar için ilgi puanı tahmin edilir. Bu nedenle kurulan modelin konu ya da temayı yeterli olarak yansıtması önem taşımaktadır (Chakrabarti et al., 1999; Johnson et al., 2003). Çeşitli çalışmalar farklı öğrenme modelleri ile performans iyileştirmesi amaçlamışlardır. Hidden Markov Model, Conditional Random Fields (Liu et al., 2004), Support Vector Machines (Choi et al., 2005) gibi makina öğrenme teknikleri kullanan çalışmalar bulunmaktadır. Diğer bir araştırma yönü de etkin arama yöntemlerine yoğunlaşmaktır (Qin et al., 2004). Odaklı tarayıcılar genellikle lokal arama algoritmaları kullanırlar, bu nedenle arama uzayı lokal bir çizgede sınırlı kalır. Arama yöntemini etkinleştirmeye yoğunlaşan çalışmalar lokal çizgelerin birleştirilmesine dayanan teknikler kullanırlar. Bu amaçla ontoloji kullanan (Ehrig and Maedche, 2003) ya da bağlantı semantiğinden faydalanan (Yuvarani et al., 2006) çalışmalar mevcuttur. Bu tip çalışmalarda ontolojik örnekler kullanılarak URL ve bağlantılar üzerinde anlamsal benzerlik değerleri çıkartılır. Yakın zamanlı çalışmalarda odaklı tarayıcıları için konu ya da tema yerine bağlamı odak olarak alan yaklaşımlar da kullanılmaktadır. Örneğin, belli bir konumla ilgili dokümanların çıkartılan uzaysal özellikler yardımıyla keşfedilmesi gibi (Ahlers and Boll 2009).

Kişisel bloglar ve ürün değerlendirme sayfaları gibi öznel içeriklerin web'de artmasıyla bilrlikte hissiyat analizi yoğun çalışılan bir araştırma konusu haline gelmiştir. Hissiyat analizi temel olarak verilen bir metnin öznel içeriğe sahip olup olmadığını, eğer öznelse olumlu mu olumsuz mu olduğunu derecelendirmeyi amaçlar (Pang and Lee, 2008). Kimi çalışmalarda hissiyat analizi belli bir konu etrafında yoğunlaşmaktadır (Nasukawa and Yi, 2003 Kullanılan teknikler genellikle dil/sözlük tabanlı (Nasukawa and Yi, 2003; Thet et al., 2009), veya

makina öğrenme tabanlı olmaktadır (Wang et al., 2011). Bazı çalışmalar alana özel (domain-specific) hissiyat analizi yapmaktadırlar (Choi et al., 2009). Bu yaklaşımlarda genel özellikler yerine alan bağlamında özellikler kullanılarak hissiyat yönü ve derecesi belirlenmektedir. Hissiyat analizi çalışmaları uygulama alanına göre de farklılık göstermektedir. Örneğin gazete haberleri (Nasukawa and Yi, 2003), film yorumları (Thet et al., 2009), ya da sosyal ağlar (Wang et al. 2011) üzerinde hissiyat analizi yapan çalışmalar bulunmaktadır. Hissiyat analizi konusundaki çalışmalar çoğunlukla İngilizce diline yoğunlaşmış durumdadır. Ancak bunun yanı sıra çok dil üzerinde (Wang et al. 2011) ya da farklı diller üzerinde (Abbasi et al., 2008) çalışmalar da mevcuttur. Türkçe üzerinde hissiyat analizi çalışmaları oldukça sınırlıdır. Yakın zamanlı bir tez çalışmasında (Erogul, 2009) Türkçe film yorumları üzerinde makina öğreme ile model geliştirilerek hissiyat analizi yapılmıştır. Bu çalışmada başarı oranı %85 olarak raporlanmıştır.

Hissiyat analizi ve odaklı taramayı birleştirme konusunda projemize yakın sadece tek bir çalışmaya rastlanmıştır (Fu et al., 2012). Bu çalışma konu odaklı tarama ile hissiyat analizini birleştirmektedir. Önce konu odaklı olarak keşfedilen sayfalar için içerik üzerinden hissiyat derecesi bulunmaktadır. Çalışmamız, konu ya da temadan bağımsız olarak hissiyat odaklı tarama yapılması ve hissiyat derecesi için tahminleme kullanılması nedeniyle (Fu et al., 2012)'da sunulan çalışmadan oldukça farklıdır.

3. YÖNTEM

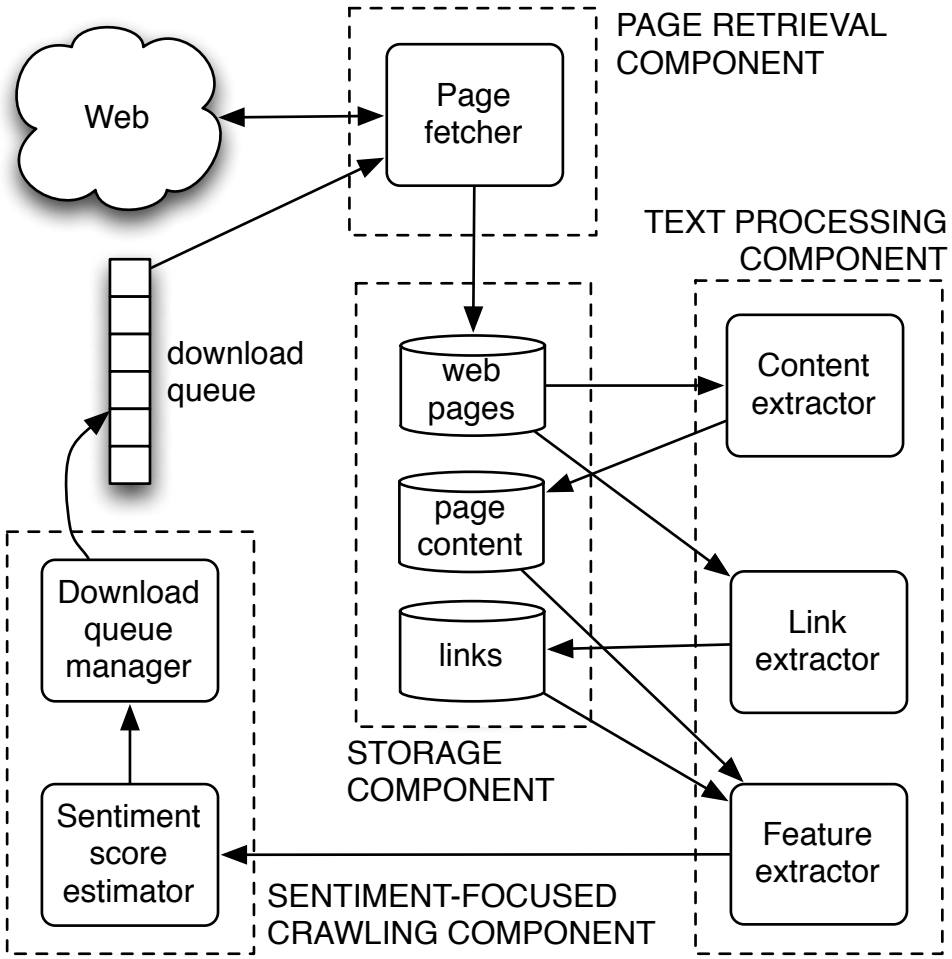
3.1. Genel Yapı

Oluşturduğumuz hissiyat odaklı ağ tarayıcısı çerçevesinin genel mimarisi Şekil 1'de sunulmaktadır. Şekilde sunulduğu gibi mimari dört ana modülden oluşmaktadır: sayfa erişim modülü, saklama modülü, metin işleme modülü ve hissiyat odaklı web tarama modülü.

Sayfa Erişim Modülü (Page Retrieval): Bu modül web sayfalarına eldeki URL üzerinden erişim ve içeriğin HTML olarak saklanmasını gerçekleştirmektedir. Sayfalara erişimde tekrarların önlenmesi, erişim politikalarına uyulması gibi tarama işlemler detayları bu modül tarafından yapılmaktadır. Bu işlemlerin ağ tarayıcının kalbini oluşturduğunu ve ağ tarayıcılarda standart olarak yapıldığı söylenebilir.

Saklama Modülü (Storage): Bu modül erişilen sayfaların veritabanında saklanması ve bağlantıların çıkarımı işlemlerini gerçekleştirir.

Metin İşleme Modülü (Text Processing): Sayfa erişimi sağlandıktan sonra sayfa içindeki metinlerin çıkartılması gereklidir. Bunun için sayfa içindeki etiketler temizlenir. Kullanılan metin işlemcinin performansı sonraki adımlar için kritik olabilir. Çalışmamızda metin işleme amacıya Html Parser (http://htmlparser.sourceforge.net) ve BoilerPipe (http://code.google.com/p/boilerpipe) araçlarını kullandık.



Şekil 1. Hissiyat Odaklı Tarayıcı Çerçevesi Genel Mimarisi

Hissiyat Odaklı Web Tarama Modülü (Sentiment-focused Crawling): Bu modül çerçevenin beynini oluştırmaktadır. Hissiyat puanı tahminleme alt modülü alınan URL için bir tahminleme yaparak sayfayı arama kuyruğunda uygun sıraya yerleştirir. Hissiyat puanı yüksek sayfalar kuyrukta ön sıralarda yer alır ve öncelikli olarak keşfedilir ve erişilir.

3.2. Kullanıcı Anketi Çalışması

Hissiyat odaklı tarayıcının başarı performansını değerlendirebilmek için sayfaların gerçek hissiyat puanlarını bilmek gereklidir. Deneylerde kullandığımız ClueWeb09 veri seti orijinal haliyle sayfaların hissiyat puanlarını içermemektedir. Öte yandan milyonlar mertebesinde web sayfası içeren veri setinin tamamı için makul bir sürede elle değerlendirme yapmak mümkün değildir. Bu nedenle otomatik bir hissiyat puanı çıkarımı yöntemi kullanmamız gerekli oldu. Ancak yöntemin güvenirliğini test etmek için bir kullanıcı anketi çalışması yaptık.

Kullanıcı anketi için, deneyler için kullandığımız ClueWeb09-B veri kümesi içinden rastele seçilmiş 500 web sayfası 5 hakem tarafında “hissiyat içerikli” ya da “hissiyat içerikli değil” şeklinde etiketlendi. Anket çalışmasında hakemler arasında ortalama %85 görüş birliği gözlemlendi. Kappa-kohen analizi yapılarak gözlemlenen görüş birliğinin istatistiksel olarak anlamlı olduğu ortaya konuldu. Kullanıcı değerlendirmelerine göre üç tip “ground-truth” oluşturduk. GT1’de en az

bir hakem tarafından hissiyat içerikli olarak etiketlenmesi halinde sayfanın hissiyat içerikli olduğu kabul edildi. GT2'de en az iki hakemin olumlu görüşü halinde sayfanın hissiyat içerikli olduğu kabul edildi. GT3'de ise üç hakemin olumlu görüşü altında sayfa hissiyat içerikli olarak notlandırıldı.

İkinci adım olarak, SentiStrength aracı ile otomatik hissiyat puanı hesaplaması için 8 farklı konfigürasyon oluşturduk:

- HP-SS-All: Metin içeriği HtmlParser ile çıkartıldı, hissiyat analizi için cümle puanları kullanıldı, bütün kelimeler değerlendirildi.
- HP-SS-Adj: Metin içeriği HtmlParser ile çıkartıldı, hissiyat analizi için cümle puanları kullanıldı, sadece sıfatlar değerlendirildi.
- HP-WS-All: Metin içeriği HtmlParser ile çıkartıldı, hissiyat analizi için kelime puanları kullanıldı, bütün kelimeler değerlendirildi.
- HP-WS-Adj: Metin içeriği HtmlParser ile çıkartıldı, hissiyat analizi için kelime puanları kullanıldı, sadece sıfatlar değerlendirildi.
- BP-SS-All: Metin içeriği BoilerPipe ile çıkartıldı, hissiyat analizi için cümle puanları kullanıldı, bütün kelimeler değerlendirildi.
- BP-SS-Adj: Metin içeriği BoilerPipe ile çıkartıldı, hissiyat analizi için cümle puanları kullanıldı, sadece sıfatlar değerlendirildi.
- BP-WS-All: Metin içeriği BoilerPipe ile çıkartıldı, hissiyat analizi için kelime puanları kullanıldı, bütün kelimeler değerlendirildi.
- BP-WS-Adj: Metin içeriği BoilerPipe ile çıkartıldı, hissiyat analizi için kelime puanları kullanıldı, sadece sıfatlar değerlendirildi.

Üçüncü adım olarak, kullanıcı anketinde kullanılan 500 sayfa bu 8 konfigürasyonun her biri altında otomatik olarak puanlandı ve sonuçlar GT1, GT2 ve GT3 ile karşılaştırıldı. Üç hakemin görüş birliğini gerektiren GT3 için bile BP-WS-Adj konfigürasyonun başarılı sonuç verdiği gözlemlendi. Bu nedenle deneylerde kullanılan sayfaların gerçek hissiyat puanları SentiStrength ile BP-WS-Adj konfigürasyonu altında hesaplandı.

3.3. Hissiyat Tahminlemesi

Daha önceki kısımlarda anlatıldığı gibi, önerdiğimiz hissiyat odaklı ağ tarama metodu, sayfa içeriğini görmeden hissiyat puanının tahminlemesine dayanmaktadır. Bu tahminleme için sayfaya bağlantı veren önceki sayfa veya sayfaların, kelime sayısı, bağlantı sayısı, hissiyat puanı gibi özelliklerinden faydalanıldı. Tahminleme için iki temel yöntem geliştirildi:

Bağlantı Veren Sayfanın Özelliklerine Göre Tahminleme: Bu yöntemde, eldeki URL'ye bağlantı veren sayfaların ortalama hissiyat puanı, sayfanın hissiyat puanı olarak tahmin edildi.

Makine Öğrenme Tabanlı Tahminleme: Bu yöntemde, daha önce erişilmiş sayfaların özelliklerinden hareketle bir hissiyat puanı modeli oluşturuldu. Bu model için yukarıda da belirtildiği gibi bağlantı veren sayfaların özellikleri kullanılarak hedef sayfanın hissiyat puanı modellendi. Bu modelleme için SVM algoritması kullanıldı. Tahminleme modeli düzenli aralıklarla o zamana kadar

erişilmiş sayfalar için yeniden oluşturuldu.

4. BULGULAR

4.1. Deney Ortamı

Veri Kümesi. Deneylerde veri kümesi olarak ClueWeb09-B web sayfası kolleksiyonu kullandık. Veri kümesinde yaklaşık 50 milyon web sayfası yer almaktadır. Sayfaların yaklaşık yarısı spam grubuna girmektedir. Yaklaşık %16'sı başka sayfaya bağlantı vermemekte, yaklaşık %26'sı başka sayfadan bağlantı almamaktadır. Deneylerde kullanılan sayfalar için üç tip puan verdik: Hissiyat puanı, Spam Puanı ve PageRank puanı. Hissiyat puanı için daha önce anlatıldığı gibi BP-WS-Adj konfigürasyonu altında puan hesaplandı. PageRank puanları veri seti ile birlikte gelmektedir. Spam puanı ise Waterloo Spam Rankings'ten (Cormack et al., 2010) alındı. Yaptığımız korelasyon analizlerinde hissiyat puanları ve spam puanları arasında ilişki olmadığını gözlemledik. Öte yandan PageRank puanları ve hissiyat puanları arasında zayıf bir ters orantı gözlemlenmiştir.

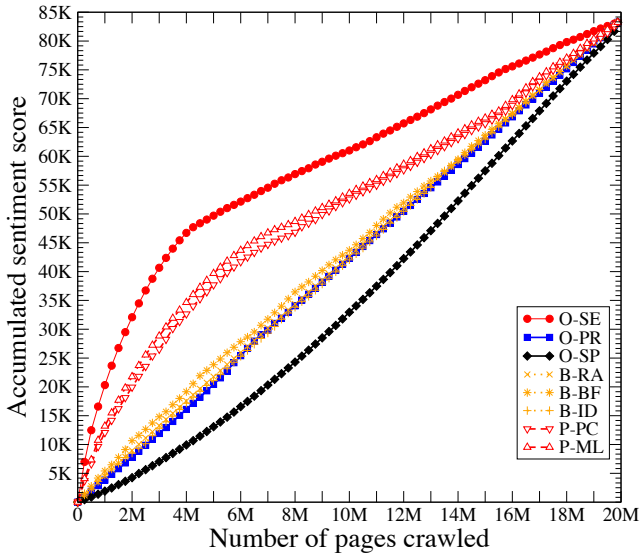
Yazılım ve Donanım. Deneyler için kullanılan simulator Java 6.0 ile kodlandı. Web sayfaları ve çıkartılan sayfa özellikleri MySQL (version 5.1.61) veritabanı sisteminde saklandı. Makina öğrenme için regrasyon modunda LibSVM kullanıldı (Chang and Lin 2011). Kodlar 16-core 48GB RAM konfigürasyonlu bilgisayarda koşturuldu.

Tarayıcılar. Yöntem kısmında anlatıldığı gibi hissiyat odaklı tarama için iki teknik kullanıldı: bağlantı veren sayfa içeriğine göre tahminleme ve makine öğrenme tabanlı tahminleme. Bu teknikleri kullanan iki ayrı tarayıcı oluşturuldu (P-PC ve P-ML). Önerilen bu iki strateji üç tane genel amaçlı temel tarayıcı stratejisi ile karşılaştırıldı: rastgele seçim (B-RA), indegree-tabanlı (B-ID) ve breadth-first (B-BF). Bunlara ek olarak en iyi performansı belirleme üzere üç tane kahin (oracle) tarayıcı oluşturuldu: hissiyat puanına göre (O-SE), spam puanına göre (O-SP) ve PageRank puanına göre (O-PR). Örneğin O-SE kuyruktaki sayfaların gerçek hissiyat puanlarını bilmekte ve aralarından en yüksek olanı seçmektedir.

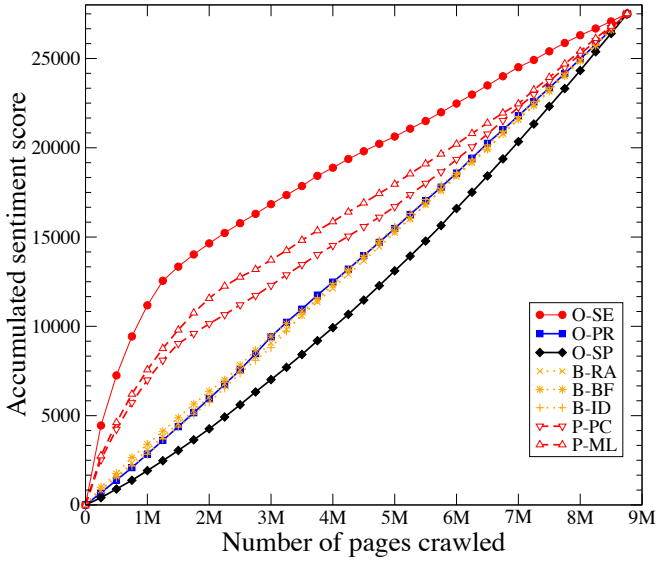
Performans Ölçümleri. Temel performans ölçümü toplanan hissiyat puanı üzerinden yapılmaktadır. Buna ek olarak ortalama PageRank puanları, ortalama spam puanları ve erilişen içerik byte cinsinden kayıt altına alınmaktadır. Taranan her 1000 sayfada bu değerler için ölçüm yapılarak raporlanmaktadır.

4.2. Deney Sonuçları

Tarayıcıların topladıkları hissiyat puanı karşılaştırmaları Şekil 2 ve Şekil 3'te sunulmuştur. Şekil 2'de spam filtresi kullanılmadan yapılan tarama, Şekil 3'te ise spam sayfaların filtrelendiği taramanın sonucu gösterilmektedir. Her iki deneyde de hissiyat puanına göre ilerleyen kahin tarayıcı (O-SE) en başarılı performansı göstermektedir. Önerilen hissiyat odaklı tarama stratejilerinin her ikisi de klasik ağ tarama yaklaşımlardan daha başarılı çalışmaktadır. İki strateji arasında makina öğrenme tabanlı yöntem hissiyat içeriklie sayfalara daha hızlı erişmektedir.

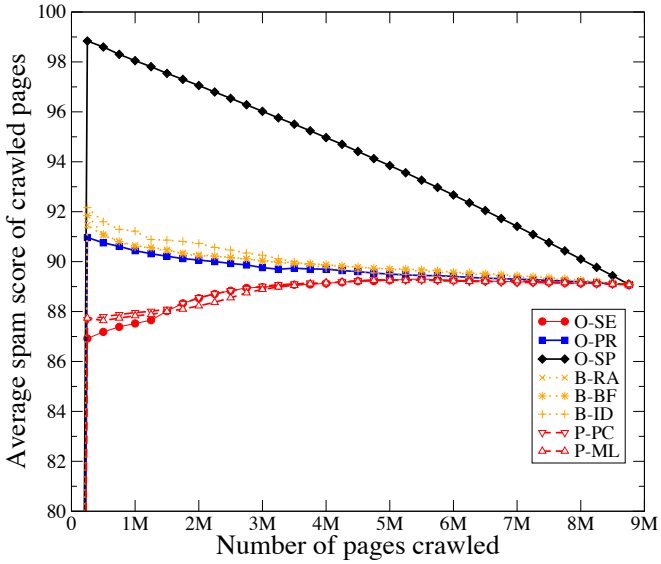


Şekil 2. Toplanan Hissiyat puanı (Spam sayfalar dahil)



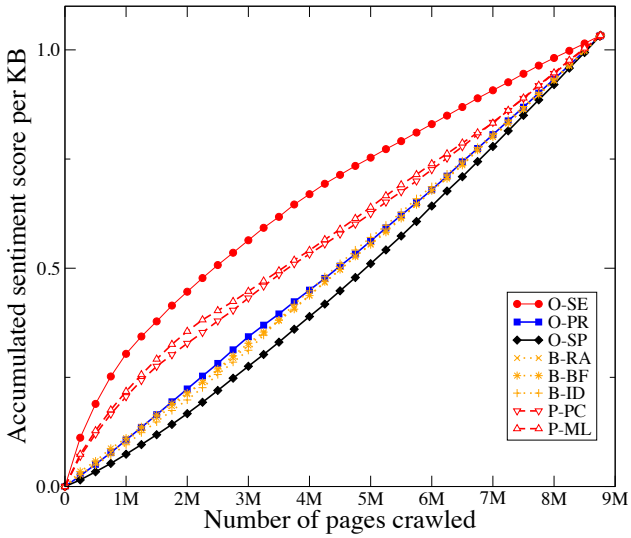
Şekil 3. Toplanan Hissiyat puanı (Spam sayfalar hariç)

Şekil 4'te erişilen sayfaların ortalama spam puanları sunulmaktadır. Beklenen şekilde spam puanına göre ilerleyen kahin tarayıcı (O-SP) yüksek spam puanlı sayfaları hızla toplamaktadır. Şekilde görüldüğü gibi, önerilen sentiment-odaklı tarayıcılar spam içerikli sayfalara yönelmemektedir.



Şekil 4. Ortalama spam puanları (Spam sayfalar hariç)

Şekil 5'te erişilen içeriğin büyüklüğü ile toplanan hissiyat puanının normalize edilmesi sonucu oluşan tarayıcı davranışı sunulmaktadır. Bu normalizasyon altında da önceki deneylerde olduğu gibi önerdiğimiz stratejiler başarı göstermektedir.



Şekil 5. Erişilen KB içerik başına düşen toplam hissiyat puanı (Spam sayfalar hariç)

5. SONUÇ ve ÖNERİLER

Bu çalışmamızda, hissiyat odaklı ağ tarama için bir çerçeve oluşturarak, sayfa içeriğine ulaşmadan sayfanın hissiyat puanını tahminlemek amacıyla çeşitli teknikler önerdik. Önerdiğimiz hissiyat odaklı ağ tarama yöntemi, geliştirilen bu

tahminleme tekniklerine dayanmaktadır. Önerdiğimiz yöntemin performansını ağ tarama simülasyonları üzerinden değerlendirerek genel amaçlı tarayıcıların performansları ile karşılaştırdık. Verinin büyüklüğü nedeniyle deneyleri iki aşamalı olarak gerçekleştirdik. Önce yaklaşık 1 milyon web sayfası üzerinde çalıştık, bu sonuçları bir konferans bildirisi olarak sunduk. Daha yaklaşık 50 milyon sayfa içeren kümenin tamamı üzerinde analiz yaptık. Yaptığımız tüm deneyler önerdiğimiz tekniğin temel yöntemlere göre hissiyat içerikli sayfaların daha hızlı bulunmasını sağladığını ortaya koydu.

Hissiyat odaklı web tarayıcı çalışmasına ek olara, hissiyat analizi konusunda edindiğimiz bilgiyi değerlendirerek Türkçe metinler için hissiyat analizine yönelik bir çalışma yaptık. Dil/sözlük tabanlı hissiyat analizi yaklaşımı kapsamında Türkçe bir sözcük ve kural kümesi oluşturularak performansı analiz edildi. Eldeki veriler üzerinde %80 civarında bir başarı oranı sağlandı. Başlatılan bu çalışma sözcük ve kural kümesinin geliştirilmesi ile ilerletilebilir. Bu sayede hissiyat analizi başarısının yükseltilebilmesi için uygun bir potansiyel bulunmaktadır.

Bu çalışmanın devamında çeşitli yönlerde yeni araştırma projeleri oluşturulabilir. Bir araştırma problemi olarak hissiyat puanı tahminlemesi için farklı yöntemlerle performansın daha da iyileştirilmesi üzerinde çalışılabilir. Diğer bir araştırma yönü olarak hissiyat puanının yanı sıra hissiyat türü (olumlu/olumsuz) odaklı ağ tarama yapılabilir. Bu tür bir mekanizma ile olumlu içeriklerin önceliğinin arttığı bir arama motoru mekanizması oluşturulabilir. Başka bir araştırma problemi olarak hissiyat analizi, nüfus analizi ile birleştirilerek örneğin daha özelleşmiş ağ tarayıcılar amaçlanabilir. Buna ek olarak Türkçe hissiyat analizi çalışması ile birleştirilerek Türkçe için hissiyat odaklı bir tarayıcı geliştirilebilir.

Referanslar

ABBASI, A., CHEN, H., AND SALEM, A.. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Trans. Inf. Syst*. 26, 3, 12:1–12:34 (2008).

AHLERS, D. AND BOLL, S.. Adaptive geospatially focused crawling. In *Proc. 18th ACM Int'l Conf. Information and Knowledge Management.* (2009) pp..445–454.

BACCIANELLA, A. E. S. AND SEBASTIANI, F. SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Proc. 7th Conf. Int'l Language Resources and Evaluation.(*2010*)*

BAI, X. 2011. Predicting consumer sentiments from online text. *Decision Support Syst.* 50*, (*2011*)* 732–742.

BEINEKE, P., HASTIE, T., MANNING, C., AND VAITHYANATHAN, S. 2004. Exploring sentiment summarization. In *Proc. AAAI Spring Symp. Exploring Attitude and Affect in Text: Theories and Applications.* (2004) *pp.*1–4.

CHAKRABARTI, S., VAN DEN BERG, M., AND DOM, B. Focused crawling: a new approach to topic-specific web resource discovery. *Computer Networks 31,* 11-16, (1999) pp.1623–1640.

CHANG, C.-C. AND LIN, C.-J. LIBSVM: A library for support vector machines. *ACM Trans. Intelligent Systems and Technology 2,* 27 (2011) pp.1-27.

CHOI, Y., KIM, K., AND KANG, M. A focused crawling for the web resource discovery using a modified proximal support vector machines. In *Proc. 2005 Int'l Conf. Computational Science and its Applications- Volume Part I. (2005)* pp.186–194.

CHOI, Y., KIM, Y., AND MYAENG, S.-H. Domain-specific sentiment analysis using contextual feature generation. In *Proc. 1st Int'l CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion.* (2009) pp.37–44.

CORMACK, G. V., SMUCKER, M. D., AND CLARKE, C. L. A. Efficient and effective spam filtering and re-ranking for large web datasets. CoRR abs/1004.5168 (2010).

DAVE, K., LAWRENCE, S., AND PENNOCK, D. M. 2003. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *Proc. 12th Int'l Conf. World Wide Web.* 519–528.

EHRIG, M. AND MAEDCHE, A. Ontology-focused crawling of web documents. In *Proc. 2003 ACM Symp.Applied Computing.* (2003) pp.1174–1178.

EROGUL, U. *Sentiment Analysis in Turkish*. (Master's thesis), Middle East Technical University, Computer Engineering Department (2009).

FU, T., ABBASI, A., ZENG, D., AND CHEN, H. Sentimental spidering: Leveraging opinion information in focused crawlers. *ACM Transactions on Information Systems 30, (2012) pp.* 4-24.

GERANI, S., CARMAN, M. J., AND CRESTANI, F. Investigating learning approaches for blog post opinion retrieval. In *Proc. 31th European Conf. Information Retrieval. (2009) pp.* 313–324.

GODBOLE, N., SRINIVASAIAH, M., AND SKIENA, S. 2007. Large-scale sentiment analysis for news and blogs. In *Proc. Int'l Conf. Weblogs and Social Media.*

GREGORY, M. L., CHINCHOR, N., WHITNEY, P., CARTER, R., HETZLER, E., AND TURNER, A. User- directed sentiment analysis: visualizing the affective content of documents. In *Proc. Workshop on Sen- timent and Subjectivity in Text.* (2006) pp.23–30.

JOHNSON, J., TSIOUTSIOULIKLIS, K., AND GILES, C. L. Evolving strategies for focused web crawling. In *Proc. 20th Int'l Conf. Machine Learning. (2003) pp.*298–305.

KUCUKTUNC, O., CAMBAZOGLU, B. B., WEBER, I., AND FERHATOSMANOGLU, H. A large-scale senti- ment analysis for Yahoo! Answers. In *Proc. 5th ACM Int'l Conf. Web Search and Data Mining. (2012) pp.* 633–642.

LERMAN, K., BLAIR-GOLDENSOHN, S., AND MCDONALD, R. Sentiment summarization: evaluating and learning user preferences. In *Proc. 12th Conf. European Chapter of the Assoc. for Computational Linguistics*. (2009) pp.514–522.

LIU, H., MILIOS, E., AND JANSSEN, J. Probabilistic models for focused web crawling. In *Proc. 6th ACM International Workshop on Web Information and Data Management.* (2004) pp.16–22.

NASUKAWA, T. AND YI, J. Sentiment analysis: capturing favorability using natural language processing. In *Proc. 2nd Int'l Conf. Knowledge Capture.* 70–77. NOVAK, B. 2004. A survey of focused web crawling algorithms. *Proc. SIKDD 5558,* (2003) pp.55–58.

OLSTON, C. AND NAJORK, M. Web crawling. *Found. Trends Inf. Retr. 4,* 3, (2010) pp.175–246.

PANG, B. AND LEE, L. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr. 2*, (2008) pp.1–135.

PANG, B., LEE, L., AND VAITHYANATHAN, S. Thumbs up? sentiment classification using machine learning techniques. In *Proc. ACL-02 Conf. Empirical Methods in Natural Language Processing.* (2002) pp.79–86.

QIN, J., ZHOU, Y., AND CHAU, M. Building domain-specific web collections for scientific digital li- braries: a meta-search enhanced focused crawling method. In *Proc. 4th ACM/IEEE-CS Joint Conf. Digital Libraries.* (2004) pp.135–141.

THELWALL, M., BUCKLEY, K., PALTOGLOU, G., CAI, D., AND KAPPAS, A. Sentiment strength detection in short informal text. *J. Am. Soc. Inf. Sci. Technol. 61,* 12, (2010),pp.2544–2558.

THET, T. T., NA, J.-C., KHOO, C. S., AND SHAKTHIKUMAR, S. Sentiment analysis of movie reviews on discussion boards using a linguistic approach. In *Proc. 1st Int'l CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion.*

*(2009), pp.*81–84.

TURNEY, P. D.. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proc. 40th Annual Meeting on Association for Computational Linguistics. (2002) pp.*417-424.

WANG, X., WEI, F., LIU, X., ZHOU, M., AND ZHANG, M. Topic sentiment analysis in twitter: a graph- based hashtag sentiment classification approach. In *Proc. 20th ACM Int'l Conf. Information and Knowledge Management. (2011) pp.*1031-1040.

YI, J., NASUKAWA, T., BUNESCU, R., AND NIBLACK, W. Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques. In *Proc. 3rd IEEE Int'l Conf. Data Mining. (2003) pp.*427–434.

YUVARANI, M., IYENGAR, N., AND KANNAN, A. LSCrawler: a framework for an enhanced focused web crawler based on link semantics. In *IEEE/WIC/ACM Int'l Conf. Web Intelligence*.(2006) pp.794–800.

ZHANG, W., YU, C., AND MENG, W. Opinion retrieval from blogs. In *Proc. 16th ACM Int'l Conf. Information and Knowledge Management.* (2007) pp.831–840.

TÜBİTAK
PROJE ÖZET BİLGİ FORMU

Proje No: 112E002

Proje Başlığı: Hissiyat Odaklı Ağ Tarama

Proje Yürütücüsü ve Araştırmacılar: Doç. Dr. Pınar KARAGÖZ

Projenin Yürütüldüğü Kuruluş ve Adresi: ODTÜ Bilgisayar Mühendisliği Bölümü A404 06800 Çankaya ANKARA

Destekleyen Kuruluş(ların) Adı ve Adresi: TÜBİTAK ARDEB

Projenin Başlangıç ve Bitiş Tarihleri: 1.05.2012 - 1.05.2013

Öz (en çok 70 kelime)

Günümüzde bir konu üzerinde hissiyat ve görüş içeren web sayfalarının sayısı hızla artmaktadır. Bu tip sayfalar güncel konulara odaklandığından hızlı olarak bulunmaları önemlidir, ancak mevcut tarayıcılar hissiyat içeren sayfaların keşfedilmesinde ve sorgulanmasında yetersiz kalmaktadır. Bu projede, hissiyat içeren web sayfalarının daha hızlı keşfine odaklanan web tarayıcı üzerinde çalışılmıştır. Geliştirdiğimiz teknik, sayfayı indirmeden hissiyat içerik değerinin tahminlemesine dayanmaktadır. Yaptığımız deneylerde önerdiğimiz tekniğin hissiyat içerikli sayfaların çok daha hızlı keşfedildiği ortaya konmaktadır.

Anahtar Kelimeler: Hissiyat Analizi, Odaklı Web Tarayıcı, Veri Madenciliği, Sınıflandırma, Support Vector Machine (SVM)

Fikri Ürün Bildirim Formu **Sunuldu mu?** **Evet** ☐ **Gerekli Değil** ☐

Fikri Ürün Bildirim Formu'nun tesliminden sonra 3 ay içerisinde patent başvurusu yapılmalıdır.

Projeden Yapılan Yayınlar:

- A. G. Vural, B. B. Cambazoglu, P. Karagoz, "Sentiment-Focused Web Crawling", ACM Transactions on The Web (to be submitted).
- A. G. Vural, B. B. Cambazoglu, P. Senkul, "Sentiment-Focused Web Crawling", CIKM 2012, Hawaii, USA, October 2012.
- A. G. Vural, B. B. Cambazoglu, P. Senkul, Z. O. Tokgoz, "A Framework for Sentiment Analysis in Turkish: Application to Polarity Detection of Movie Reviews in Turkish", ISCIS 2012, Paris, France, October 2012.