Journal of Engineering and Natural Sciences
Mühendislik ve Fen Bilimleri Dergisi

Sigma 29, 156-169, 2011

Research Article / Araştırma Makalesi

AUTOMATIC TAG CLOUD GENERATION FROM TURKISH CONTENTS

Serdar SAVAŞAN*[1], Banu DİRİ[2]

[1]*Yıldız Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı, Yıldız-İSTANBUL*
[2]*Yıldız Teknik Üniversitesi, Elektrik-Elektronik Fakültesi, Bilgisayar Mühendisliği Bölümü, Yıldız-İSTANBUL*

Received/Geliş: 16.02.2011 Revised/Düzeltme: 27.06.2011 Accepted/Kabul: 28.06.2011

ABSTRACT

Every day we are faced with an intense traffic of information that is caused by the communication medium, shaping by the growth and development of information technologies and the Internet. Users can access to information in a more comfortable way by the help of text summarization, information extraction and document tagging operations. Tag clouds are visual tools that direct users to the information provided on the internet by taking advantage of text mining techniques. Generally the users prefer to follow the agenda from a single newspaper web site. But a large number of unimportant news items that are not published in newspaper printed version, find place for themselves on internet sites and cause information pollution. Considering the needs posed by these and other problems, a research on information retrieval systems, text mining and tag clouds, has been made in this study. With the aid of gathered information, a news compilation and retrieval tool called Agenda Cloud is developed, which makes it more functional to access news on the Internet. The main benefits of the application are; it allows users to access a lot of information in a short amount of time and provides historical information besides daily agenda.
Keywords: Information retrieval systems, text mining, text analysis, tag cloud, collecting news compilation, document tagging.

TÜRKÇE İÇERİKLERDEN OTOMATİK ETİKET BULUTU OLUŞTURMA

ÖZET

Bilgi teknolojileri ve internetin gelişmesiyle şekillenen iletişim ortamı, her geçen gün daha da yoğun bir bilgi trafiği ile karşı karşıya kalmamıza neden olmaktadır. Metin özetlerinin çıkarılması ve dokümanların etiketlenmesi gibi işlemler sayesinde, kullanıcılar bilgiye daha rahat bir şekilde ulaşabilmektedir. Etiket Bulutları, metin madenciliği tekniklerinden faydalanarak, kullanıcıları internet ortamında sunulan bilgilere yönlendiren görsel erişim araçlarıdır. Genel olarak kullanıcılar haberleri, tercih ettikleri tek bir gazete sitesinden takip ederler. Ancak basılı ortamda yayınlanmayan düşük öneme sahip çok sayıda haber, gazetelerin internet sitelerinde kendilerine yer bulmakta ve bilgi kirliliği oluşturmaktadır. Bu ve benzeri sorunların ortaya çıkardığı ihtiyaçlar göz önüne alınarak, bu çalışmada, bilgiye erişim sistemleri, metin madenciliği ve etiket bulutları konularında araştırma yapılmıştır. Elde edilen bilgiler ışığında, internet ortamındaki günlük haberlere erişimi daha işlevsel hale getiren Gündem Bulutu adında bir haber derleme ve erişim aracı geliştirilmiştir. Kullanıcıların kısa sürede çok daha fazla bilgiye erişmesine olanak sağlaması ve anlık gündem dışında geçmişe dönük tarihçe sunuyor olması uygulamanın en önemli faydalarıdır.
Anahtar Sözcükler: Erişim sistemleri, metin madenciliği, metin analizi, etiket bulutu, haber derleme, doküman etiketleme.

* Corresponding Author/Sorumlu Yazar: e-mail/e-ileti: serdars@kets.com, tel: (212) 232 56 66

1. GİRİŞ

Dünya, son yirmi yılda tarihinin hiçbir döneminde olmadığı kadar hızlı bir değişim ve gelişim sürecine girmiştir. Bu sürecin başrolünde teknoloji ve beraberinde getirdiği bilgi erişim kolaylığı vardır.

Yakın zamana kadar bilgiye erişimin en yaygın araçları kütüphane ve kurum arşivleri iken, günümüzde internet ağı ve elektronik arşiv sistemleri büyük oranda bu araçların yerini almıştır. Bu alandaki gelişmeler göz önüne alındığında, bilgi lojistiğinin hemen hemen tamamının, yakın gelecekte sadece elektronik ortamda yapılacağı söylenebilir.

Bu gelişmelere paralel olarak içerik yönetimi sistemlerinin kurumlarda yaygın bir şekilde kullanılmasıyla birlikte elektronik dokümanların sayısı da çok hızlı bir şekilde artmıştır. Bu sistemler, fiziksel arşive göre dokümanlara ulaşmayı kolaylaştırsa da, doküman sayısındaki bu artış, doğru içeriğe kısa bir sürede erişme konusunda aynı başarıyı gösterememektedir. Erişimi daha kolay hale getirebilmek için dokümanları, üst veri bilgileri ile kayıt altına almak, belli kategoriler ve anahtar kelimeler ile sınıflandırmak ve tam metin arama motorlarından faydalanmak gibi çeşitli teknikler kullanılmaktadır. İlk iki teknik, kullanıcıların her doküman için ek veri girişi yapmalarını gerektirdiği için, özellikle çok sayıda doküman üreten yapılarda pratikte kullanılamaz hale gelmektedir. Tam metin arama motorları otomatik çalıştığı için bu sorun yaşanmıyor olsa da, bu teknikte de sorgulama sonuçlarından aranan doğru dokümana erişme konusunda sıkıntılar yaşanmaktadır.

Günümüzün en popüler bilgi kaynağı olan internetteki bilgi miktarı her geçen gün artmakta ve bu bilgi yığını içinde kullanıcıların ihtiyaç duydukları doğru bilgiye ulaşmaları ciddi bir sorun haline gelmektedir. Bu soruna çözüm getirmek üzere ortaya çıkan arama motorlarından da her zaman istenen sonuçlar alınamamaktadır.

Son dönemde internet kullanımı geçmişte yapılan tahminlerin çok daha üzerinde bir hızda yaygınlaşmakta ve iletişim büyük oranda internet üzerinden yapılmaktadır. Yayın kuruluşları açısından internet üzerinden yapılan yayının zaman ve insan kaynağı maliyeti, diğer yöntemlere göre çok daha düşüktür. Bugün için gazetelerin kâğıt ortamındaki yayıncılığı devam etmekle birlikte, gazeteleri sadece elektronik ortamda takip edenlerin sayısı da her geçen gün artmaktadır.

Günlük haberlere gazetelerin internet siteleri aracılığı ile kolayca ulaşabilir. Ancak özel bir konuya ait haberlere ulaşmak istendiğinde durum biraz daha zorlaşmaktadır. Bu amaçla ilgili haberleri bulabilmek için internet arama motorlarına anahtar kelimeler girerek sorgulama yapıldığında genellikle girilen anahtar kelimeleri içeren ilgili ilgisiz birçok sayfa arama sonuçlarında listelenmektedir. Bu durumda listelenen sonuçların teker teker açılıp kontrol edilmesi gerekmektedir. Sonuç olarak zaman maliyetinin artmasından dolayı elde edilen fayda azalmaktadır. Buna ek olarak, sorgulama sadece gazetelere ait siteler üzerinde yapılmak istendiğinde, işlem biraz daha zorlaşmakta ve arama motorlarının yeterli olmadığı durumlarda, gazetelerin siteleri tek tek açılıp, site içi arama yapılması gerekmektedir. Bu tür ihtiyaçlardan yola çıkılarak internet üzerindeki günlük haberlere erişim konusu problem olarak ele alınmış ve üzerinde çalışma yapılmıştır.

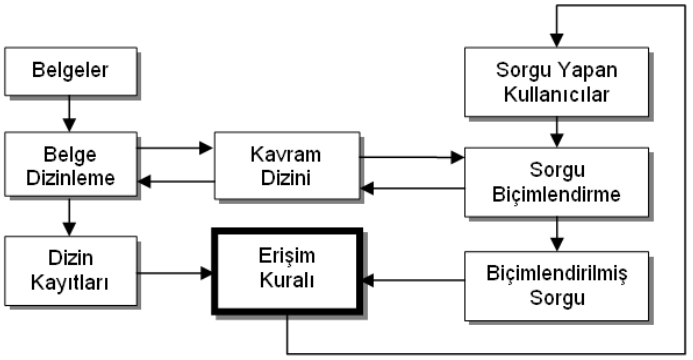
Bu çalışmada, teknolojik gelişmelerin nihai hedefi olan zaman tasarrufu ve kolaylık amacına hizmet etmek üzere, Türkçe desteği ile internet ortamındaki güncel haberlere ortak terimler üzerinden hızlı bir erişim olanağı sunan ve sürekli kendisini güncelleyen özel bir **Bilgiye Erişim Sistemi (Information Retrieval System)** oluşturulmuştur.

1.1. Bilgiye Erişim Sistemleri

Bilgiye Erişim genellikle bilgisayarlarda depolanan yapılandırılmamış metin içerikli dokümanlardan oluşan büyük miktardaki veri yığınları içinden ihtiyaç duyulan bilgiyi bulmak olarak tanımlanabilir [1]. Tonta [2] ise bu terimi "bilgi toplama, sınıflama, kataloglama,

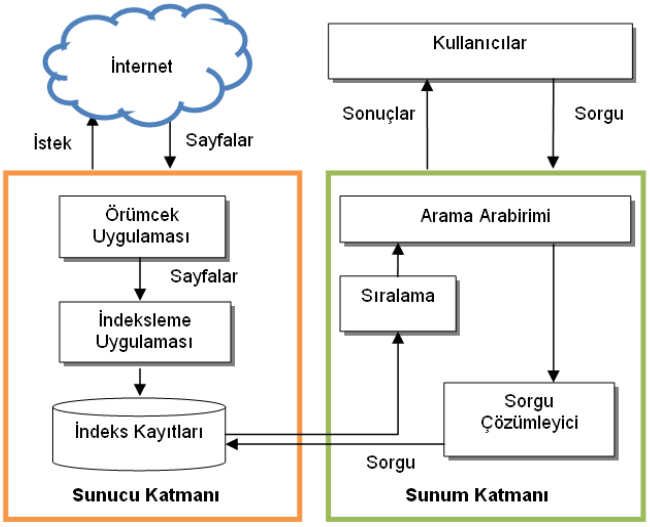
depolama, büyük miktardaki verilerden arama yapma ve bu verilerden istenen bilgiyi üretme (veya gösterme) tekniği ve süreci" şeklinde tanımlamıştır. Bu tanımlardaki işlevleri yerine getiren yapılara Bilgiye Erişim Sistemleri denilmektedir.

Belgelere erişimi sağlayan içerik belirteçleri elle veya otomatik olarak dizinleme işlemleri sonucunda elde edilir [3]. Bu çalışmada belirteçler, uygulama tarafından otomatik olarak tayin edilmiştir. Klasik bir Belge Erişim Sisteminin tasarımı Şekil 1'de görülmektedir.



Şekil 1. Belge erişim sisteminin mantıksal gösterimi [4]

Günümüzde bilgiye erişmek için yaygın olarak kullanılan internet arama motorlarının çalışma prensibi internet sayfalarının indekslenmesi ve girilen sorgular ile sayfa içeriklerinin karşılaştırılmasına dayanmaktadır. Şekil 2'de bir arama motoru mimarisi görülmektedir.



Şekil 2. Arama motoru mimarisi

Bu sistemde belli bir sayfaya erişebilmenin ön koşulu o sayfanın arama motoru tarafından daha önceden indekslenmiş olmasıdır. Bu koşulu tamamlayan ikinci önemli nokta ise kullanıcının arama sırasında kullandığı anahtar kelimenin indeksleme işleminde o sayfa ile ilişkilendirilmesidir.

Arama motoru sistemlerinde önemli kısıtlar da mevcuttur. Dokümanın içindeki kelimelerle, arama yapan kişinin kullandığı kelimelerin aynı anlama gelmesine rağmen yazılışlarının farklı olması, bazı kelimelerin tek başına kullanıldığında taşıdığı anlam ile bir kelime grubu içinde kullanıldığında farklı anlam taşıması gibi durumlar söz konusudur. Benzer şekilde doküman uzunluğu dikkate alınmadan yapılan frekans hesaplaması sonucunda oluşan arama veritabanından elde edilen sonuçlar da sağlıklı olmamaktadır. Arama motorlarının gelişimi sürecinde daha tutarlı arama sonuçları elde etmeyi hedefleyen çalışmalar sonucunda bilgisayar destekli Metin Madenciliği (Text Mining) gibi yeni yaklaşımlar ortaya çıkmıştır.

1.2. Metin Madenciliği ve Doküman Etiketleme

Metin Madenciliği bilgisayar destekli analiz tekniklerinden faydalanarak metinlerden yüksek kalitede bilgi elde edilmesi işlemidir. Metinlerin otomatik olarak özetlenmesi, dokümanı temsil eden terimlerin ve dokümanın içeriğini özetleyen kavramların çıkarılması, benzer dokümanların kümelenmesi ve buna benzer teknikler metin madenciliğinin işlevleri arasında öne çıkanlardır.

Metin Madenciliği, Veri Madenciliğinin (Data Mining) yapısal olmayan veriler (unstructured data) üzerinde çalışan bir alt dalıdır. Yapısal olmayan veriler bilgisayar ortamında bulunan ancak bilgisayar uygulamaları tarafından kullanılmaya müsait bir veri modeline sahip olmayan bilgi bütünüdür. Buna karşın yapılandırılmış veriler (structured data), sistemli bir şekilde işlenmiş, organize edilmiş ve saklanmış ham bilgilerdir [5]. Veri Madenciliğinden faydalanılarak bu veriler işlenmekte ve elde edilen bilgiler iş zekâsı uygulamalarında kullanılmaktadır [6].

Metin Madenciliği süreçleri ana hatlarıyla üç adımdan oluşmaktadır. İlk safhada ilgili kaynaklardan doküman toplanarak salt metin elde etmek üzere önişlemler yapılır. Sonraki aşamada bir önceki adımda elde edilen metin kümesi, belirlenmiş hedeflere yönelik olarak analiz edilerek yorumlanmış bilgiler çıkarılır. Son adımda ise bu bilgilere erişim sağlanır.

Metin önişleme adımında, kaynaklardan toplanan dokümanlardaki metin dışı işaret ve etiketlerin temizliği gerçekleştirilir. Özellikle internet sayfaları diğer doküman dosyalarından farklı olarak reklam, menü sistemi ve kullanıcı yorumları gibi sayfanın asıl içeriği dışında metin bölümleri de içerdiğinden bu işlem çok önemlidir. Bu bölümler doğru bir şekilde tespit edilerek ayrıştırılmazsa yapılan metin analizlerinde ciddi hatalar ortaya çıkacaktır. Bunlara ek olarak kısaltmalar, satır sonunda ayrılan kelimeler ve noktalama işaretleri de değerlendirilmelidir. Bu adımda yapılan bir diğer işlem ise etkisiz sözcüklerin (stopwords) salt metin içinden ayıklanmasıdır.

Elde edilen bilgi ve doküman setlerindeki bilgiyi organize etmek ve kolay ulaşılmasını sağlamak amacıyla etiketleme işlemi yapılır. Çok farklı görsel şekillere sahip bu etiketler genellikle, yan yana, aralarında ayırıcı işaret olmadan, bazen de altları çizilerek sunulurlar. Son dönemlerde etiketlerin internet sayfaları ve ağ günceleri gibi sosyal içerikli kullanımları da ortaya çıkmıştır [7].

Ana hatlarıyla etiketleme işlemi iki şekilde yapılmaktadır. İlk modelde kullanıcılar kendi belirledikleri anahtar kelimeleri ilgili içerik ile ilişkilendirmektedir. Özellikle metin içermeyen görsel öğelerin etiketlenmesinde bu teknik tercih edilmektedir. Etiketlerin kullanıcılar tarafından atandığı bir internet sitesinde kişilerin ilgilendikleri içeriğe daha kolay eriştiği ve içeriği oluşturanlar ile erişenlerin daha sağlıklı bir iletişime geçtiği görülmüştür. Kişileri etiketleme yapmaya teşvik eden unsurlar sosyallik (sociality) ve kendilik (self) olmak üzere iki ana grup altında toplanan, arama ve erişim için bilgiyi organize etmek, iletişim için bilgiyi işaretlemek ve organize edilen bilgiden faydalanmak şeklinde sıralanabilir [8].

Bu konuda uygulanan ikinci model ise bilgisayar destekli otomatik etiketlemedir. Makine öğrenmesi, doğal dil işleme ve istatistik gibi teknikler kullanılarak geliştirilen algoritmalar aracılığı ile metin içerikleri analiz edilir ve ilgili metni temsil eden anahtar kelimeler belirlenerek etiket ataması yapılır. Bu çalışmanın da konusu olan etiket bulutları, bu şekilde otomatik olarak oluşturulmaktadır.

2. ETİKET BULUTLARI

Etiket bulutları, sık kullanılan kelimelerin diğerlerine göre daha büyük bir yazıyüzü boyutunda ve daha belirgin bir renkte sıralandığı bir ağırlıklı liste (weighted list) gösterimidir. İlk olarak 2002 yılında Jim Flanagan tarafından bir sunum biçimi olarak kullanılmıştır [9]. Bu gösterim şekli Flickr, Technorati, Del.icio.us ve ağ güncelerinden da aldığı büyük destek ile bir moda haline gelmiş ve dünyanın önde gelen medya kuruluşlarının siteleri dahil birçok sitede yaygın olarak kullanılan bir araca dönüşmüştür.

Bir etiket bulutuna ilk kez bakıldığında, bir tasarımcı tarafından rastgele bir araya getirilmiş kelime grubu gibi düşünülebilir. Aslında bu görsel yerleşim anahtar kelimelerin taşıdıkları öneme göre farklı grafik özelliklerde kalın ya da daha büyük bir yazıyüzü boyutu ile gösterilmesi ve konumlandırılması sonucunda ortaya çıkmaktadır. Çalışmalar, etiket bulutu şeklindeki bir sunumun kelimelerin alt alta sıralandığı bir listeye göre çok daha iyi akılda kaldığını göstermektedir.

Klasik kullanımda etiketleme işlemi belli bir içeriğin, ilgili kelimelerle işaretlenmesidir. Etiket bulutları bu süreci daha da geliştirerek, etiketleri, anlamlarına, ağırlıklarına ve kullanım sıklıklarına göre diğer etiketlerle karşılaştırarak elde edilen istatistiklerden faydalanarak özel bir görsellikte sunar. Bulutu oluşturan terimler alfabetik veya kullanım sıklığına göre sıralı bir şekilde listelendiği gibi bazen de tamamen rastgele aralarında herhangi bir hiyerarşi olmadan yerleştirilir. Genel kabul görmüş kullanımda büyük ve kalın yazıyüzleri, en öne çıkarılmak istenen etiketler için, küçük ve ince yazıyüzleri ise en az öneme sahip etiketler için seçilir. Etiket bulutlarını bir erişim aracı olarak kullananlar için pek bir anlam ifade etmese de etiketler için farklı renklerin kullanıldığı etiket bulutu örneklerine de rastlanmaktadır. Etiket bulutları ilk kullanıldıkları günden itibaren birçok değişim geçirmiş ve farklı çeşitleri ortaya çıkmıştır.

2.1. Etiket Bulutu Çeşitleri

Etiket bulutlarının alışılageldik dolaşma menülerine (navigation menus) göre en büyük avantajı, kullanıcıları ilgi çeken konulara etkin bir şekilde yönlendirebilmeleridir. Konuların önemini daha iyi yansıtabilmek için kullanılan farklı görsel uygulamalar aşağıdaki gibi maddeler halinde sıralanabilir;

- Etiketler alfabetik olarak sıralanır. Etiketlerin önemli olması veya sık kullanılmasına göre farklı yazı tipleri kullanılır.
- Etiketler alfabetik olarak sıralanır. Tüm etiketler için aynı yazıyüzü boyutu kullanılır. Önem derecesi yüksek olan etiketler farklı bir renkte yazılır veya arka alanı renklendirilerek vurgulanır.
- Etiketler kullanım sıklıkları ve ağırlıklarına göre sıralanır. Etiketin önemini belirtmek için hem yazıyüzü boyutundan hem de renk tonlarından faydalanılır.
- Etiketler özel bir şekilde sıralanmaz. Önemli etiketlerin belirgin olması, yazıyüzü boyutu, ağırlığı ve renk tonu ile sağlanır.
- Etiketler benzerliklerine göre sıralanır. Yazıyüzü özelliklerinin kullanılmasının yanında birbirine benzer olan etiketlerin yan yana dizili olması sağlanır.

Etiket bulutları içerdikleri verilere göre de gruplandırılmaktadır. Kelimelerin frekanslarına göre sıralandığı bir ağırlıklı liste gösterimi olan Metin Bulutları (Text Clouds) en yaygın kullanım şeklidir. Tek başına bir kitabın metin bulutu çıkarılabileceği gibi, belli bir ortak noktaya sahip birden fazla yazının da ortak etiket bulutu oluşturulabilir. Kullanılan yazıyüzü boyutu ve renk tonlarının sayısal değerlere karşılık geldiği Veri Bulutları (Data Clouds) ise genellikle nüfus sayıları ve borsa verileri gibi bilgilerin sunulmasında kullanılır. Etiketler arasındaki ilişkilerin sıklığının yazıyüzü boyutu, ilişkinin yakınlığının da renk tonu ile gösterildiği

Eşdizimli Bulutlar (Collocate Clouds) bir doküman veya derlemin daha odaklanmış bir gösterimini sunan özel metin bulutlarıdır.

2.2. Etiket Bulutu Oluşturma

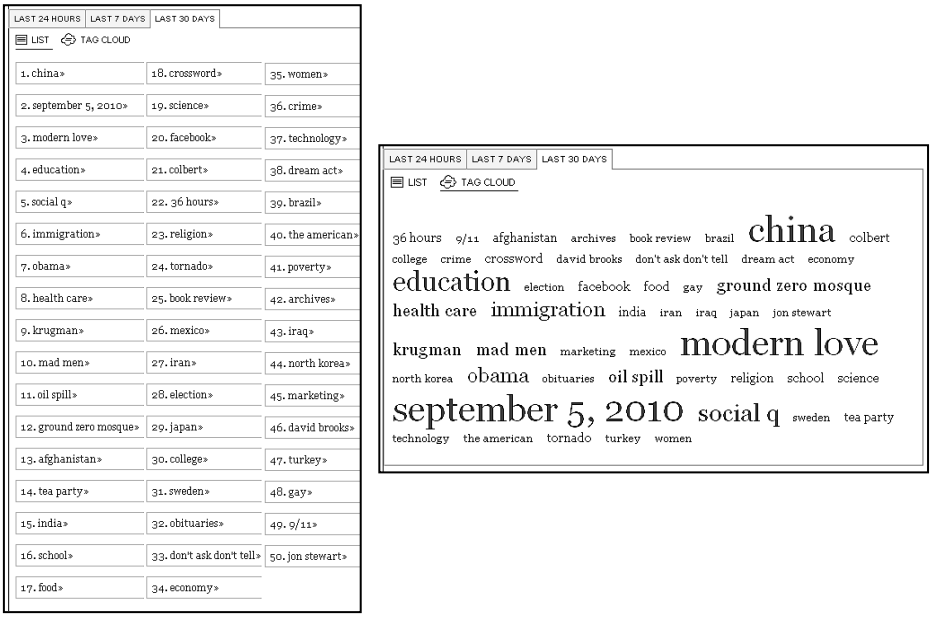
Temel olarak etiket bulutundaki bir kelimenin yazıyüzü boyutu o etiketin, bulut içindeki frekansına göre belirlenmektedir. Küçük frekanslar için yazıyüzü boyutu kelimenin sayısına eşit alınabilir. Ancak daha büyük değerlerde ölçeklendirme yapmak gerekecektir. Doğrusal normalizasyonda, bir etiketin t_i ağırlığı, 1 ile f arasındaki bir boyut ölçeğine karşılık gelir. Burada t_{min} ve t_{max} bulutu oluşturan etiketlerin ağırlık aralığını göstermektedir [10].

$t_i > t_{min}$ ise $$s_i = \left\lceil \frac{f_{\max} \cdot (t_i - t_{\min})}{t_{\max} - t_{\min}} \right\rceil$$ diğer durumlarda $s_i = 1$ (1)

- s_i: yazıyüzü boyutu
- f_{max}: en büyük yazıyüzü boyutu
- t_i: ağırlık
- t_{min}: en küçük ağırlık
- t_{max}: en büyük ağırlık

2.3. Etiket Bulutlarının İşlevi ve Algılanması

Etiket bulutları, metinlerin içeriğini oluşturan kelimelerin, belli bir matematiksel modele uygun olarak görselleştirilmesi ile oluşurlar. Bu gösterim şeklinin, etiket bulutunu kullanan kişiler tarafından algılanmasında, düz listelere göre bir avantaj taşıyıp taşımadığı konusunda farklı yorumlar yapılmıştır. Şekil 3'te bu iki gösterim şeklinin örnekleri görülmektedir.



Şekil 3. Liste gösterimi ile etiket bulutu karşılaştırması [11]

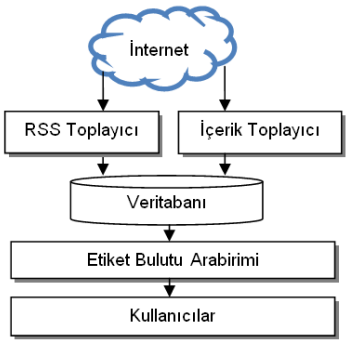
Şekil 3'te soldaki gösterimde, etiketler önem sırasına göre dizilmiş düz bir liste şeklinde sunulmaktadır. Sağdaki gösterimde ise önem derecesine göre yazıyüzü boyutu ve renk tonu farklı bir şekilde kullanılmıştır. İlk bakışta bu iki sunuş arasında sadece görsel bir fark olduğu düşünülür. Ancak bu alanda çalışma yapan kişiler, söz konusu fark hakkında birbiriyle çelişen farklı görüşler ileri sürmüşlerdir.

Örneğin, Hearst [7] yayınladığı bir makalesinde, etiket bulutlarının ilk kullanılmaya başlandığı zamanlarda, bu gösterim şeklini çok yadırgadığını ve etiket bulutlarına göre daha basit, açık ve net bir şekilde, öneme göre sıralanan kelime listelerinin daha kullanışlı olduğunu ifade etmiştir. Hearst, etiket bulutlarının algılanmasında ciddi kusurlar ortaya çıkabileceğini belirterek, yapılarında görsel bir akıcılığın olmadığını, iyi bir tasarıma sahip olmadıklarını, buna karşın iyi bir tasarımın, göze rehberlik etmesi ve sezgisel bir başlangıç vermesi gerektiğini savunmuştur. Hearst'e göre etiket bulutlarında, gözler görüntünün üzerinde zikzak çizerek gezinir, büyük etiketi bulur, sonra tekrar başa döner ve görüntüyü tekrar aynı yöntemle tarar. Bu sırada gerçekleşen bu hızlı göz atışta, orta boy etiketler de gözden kaçar. Gözün takip ettiği bu tarama yolunda, küçük etiketler gözün bu hızlı taramasında, orta etiketlerin yanında tamamen göz ardı edilir. Bu savda anlatılanlar dikkate alındığında, etiket bulutlarının kullanışsız olduğu sonucuna varılabilir. Ancak yazar, bu yayınının çalışmasında, etiket bulutlarının bu kusurlarına rağmen neden hızla yaygınlaştıklarını anlamak için 15 internet uzmanından oluşan bir grupla bir çalışma yapmıştır. Bu çalışmada elde ettiği sonuç; etiket bulutları, internet ortamındaki içeriğin insanlar tarafından aktif kullanıldığının, paylaşılan bilgilere yorumlar yapıldığının ve içeriklerin etiketlenerek sınıflandırdığının bir göstergesidir. Bulutların düzensiz görünümleri ve etiketler arasındaki boşluklar, insanların bir topluluktaki yerleşimini ve hareketlerini temsil etmektedir. Bu durum, farklı görünümdeki insanların bir odada oturup sohbet etmesi gibi düşünülebilir. Etiketler böyle bir ortamdaki insanların ne konuştuğunu belirten ufak araçlardır [7].

Bulgularıyla, etiket bulutlarının sosyal iletişimde oynadıkları rolü ön plana çıkaran sadece Hearst değildir. Bir başka araştırmacı [12] etiket bulutlarının sosyal yönünü vurgulamış ve özellikle çevrimiçi iletişimin dinamiklerinin anlaşılmasında sosyal bilimciler için iyi bir yaklaşım olduğunu ifade etmiştir. Yine başka bir yayında [13], etiket bulutlarının, düz listelere göre olumlu yanları sıralanmıştır. Büyük bir bilgi kümesinin küçük bir alanda sunulabilmesi, gözün büyük boyutlu, dolayısıyla en önemli olan etiketlere hızlı bir şekilde yönlenebilmesi gibi özellikler bunlardan bazılarıdır.

3. GÜNDEM BULUTU UYGULAMASI

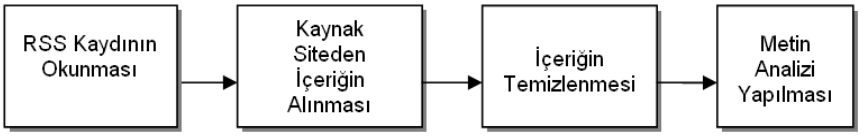
Gündem Bulutu uygulaması günlük haberlere erişimi daha işlevsel hale getiren ve farklı gazetelerde yayınlanan aynı konulu haberleri kullanıcıya bir arada sunan zeki bir haber derleme ve erişim aracıdır. Şekil 4.'te ana yapısı görülen uygulamanın sunucu katmanı, internet üzerinde yayın yapan günlük gazete ve haber sitelerinin yayınladıkları haber metinlerini anlık olarak toplayarak analiz etmekte ve sonuçlarını uygulama veritabanına yazmaktadır. Uygulamanın sunum katmanı ise, bu verilerden otomatik olarak oluşturulan etiket bulutunun sunulduğu bir internet sitesidir. Arka planda sürekli aktif olarak çalışan veri toplama bileşeni, haber kaynaklarında yayınlanan yeni bilgileri, etiket bulutuna çok kısa sürede dâhil etmektedir. Toplanan metinlerin Türkçe diline özgü analiz edilmesi de geliştirilen uygulamanın önemli özelliklerinden birisidir.



Şekil 4. Gündem Bulutu uygulamasının yapısı

Haber kaynaklarından toplanan verilerin yoğunluğu ve sorgulama ihtiyaçlarından dolayı düz yapılı dosyalar yerine bir veritabanı sistemi kullanılması uygun görülmüştür. Uygulamada, veri toplama görevi iki ayrı bileşen tarafından yerine getirilmektedir. İlk bileşen haber içeriklerini anlık yakalayabilmek için ilgili sitelerin RSS (Really Simple Syndication) yayınlarını izlemekte, diğer bileşen ise ilkine göre daha fazla zaman alan, haber içeriklerinin toplanması ve bunların analiz edilerek etiket bulutu oluşturmaya hazır hale getirilmesi işlemlerini gerçekleştirmektedir.

RSS yayınlarını ve içerikleri toplayan bileşenler, arka planda sürekli çalışan Windows servisleri olarak C# dili ile .NET platformunda geliştirilmiştir. RSS toplayıcı düzenli aralıklarla RSS kaynaklarını kontrol eder ve tespit ettiği yeni içerikleri, adres, başlık, açıklama, tarih ve haber kaynağı bilgileri ile veritabanına kaydeder. İçerik toplayıcı bu kayıtları işleyerek, haber içeriklerini kaynak sitelerden alır ve analiz eder. Bu işlemlere ait veri işleme şeması Şekil 5'te görülmektedir.



Şekil 5. İçerik toplayıcı veri işleme şeması

3.1. RSS Kaydının Okunması

Veri işleme sürecinin ilk adımında, uygulama servisi, işlem görmemiş RSS kayıtlarının listesini veritabanından alarak bir döngü içinde işler. Haber içeriğinin yayınlandığı asıl adres RSS kaydının URL sahasında bulunmaktadır. Kayıtlar haber kaynaklarına göre gruplanmış bir şekilde tutulmaktadır. Sitelere geçici olarak erişim sağlanamama durumu da dikkate alınarak ilk olarak sayfaya erişim yapılıp yapılamadığı kontrol edilir. Erişim yapılamayan kayıtlar bir sonraki döngüde işlenmek üzere listede bırakılır. Erişim sağlanan kayıtlar ise işlem sonucunda veritabanında güncellenerek işlem gördü durumuna getirilir.

3.2. Kaynak Siteden İçeriğin Alınması

Bu aşamada ilgili haberin sayfasına ulaşılarak içerik HTML olarak alınır. Bu adımda dikkat edilmesi gereken önemli bir nokta, sayfaların dil ve içerik kodlamasına uygun bir şekilde işlem yapılmasıdır. Aksi takdirde farklı şekilde kodlanmış sayfalardan alınan içerikte yer alan karakterler bozuk bir şekilde gelecektir. Uygulama geliştirme sürecinde haber kaynaklarına ait

sitelerin sayfaları analiz edilerek dil kodlaması dönüşümü uygulanması gerekenler veritabanında işaretlenmiştir. Uygulamada kullanılan kaynakların listesi Çizelge 1'de yer almaktadır. Kullanılan bu sekiz kaynaktan dört tanesinde sayfalar Türkçe, diğer dört tanesinde ise UTF-8 olarak kodlanmıştır. Buna ek olarak bir kaynakta da GZIP içerik kodlaması yapıldığı görülmüştür. Sitelerden içerik okumak için .Net kütüphanesi içindeki *System.Net.WebRequest* ve *System.Net.WebResponse* sınıfları kullanılır. Dil çözümlemesi için *System.Text.Encoding*, içerik çözümlemesi için de *System.IO.Compression.GZipStream* sınıfları kullanılarak farklı kodlanmış sayfaların içeriği Türkçe karakter problemi olmadan veritabanına kaydedilmesi sağlanmıştır.

Çizelge 1. RSS kaynaklarının listesi

Kaynak Adı	**RSS Adresi**
Hürriyet	http://rss.hurriyet.com.tr/rss.aspx?sectionId=2
Milliyet	http://www.milliyet.com.tr/D/rss/rss/Rss_1.xml
Radikal	http://www.radikal.com.tr/d/rss/Rss_77.xml
Zaman	http://www.zaman.com.tr/anasayfa.rss
Posta	http://www.posta.com.tr/xml/rss/rss_1_0.xml
Star	http://www.stargazete.com/rss.xml
Bugün	http://www.bugun.com.tr/rss/gundem.xml
Sabah	http://www.sabah.com.tr/rss/Gundem.xml

3.3. İçeriğin Temizlenmesi

Kaynak siteden alınan içerik üzerinde yapılan ilk işlem, sayfa içeriğinin metin dışı işaretler, etiketler ve HTML kodlarından arındırılmasıdır. Bu kodların "<etiket>içerik</etiket>" şeklinde özel bir yazılışı vardır. Sayfa içeriğini ayıklamak üzere özel bir yardımcı uygulama sınıfı geliştirilmiştir. Bu sınıftaki ilgili metoda girdi olarak HTML içerik verilir ve çıktı olarak HTML etiketlerinden temizlenmiş asıl metin alınır.

Örnek giriş metni:
*<img src="fft13mm192526.jpg" align="left" style="padding:2px 10px"> 23 Nisan Ulusal Egemenlik ve Çocuk Bayramı'nın bu yıl Cuma gününe denk gelmesi, iş yoğunluğundan bunalanlar için mini bir tatil olanağı yaratacak. Bahar havası iyiden iyiye kendisini hissettirirken, "bu mevsimde en güzel neresi gezilir, en güzel tatil nasıl yapılır?" diye düşünmeye pek gerek yok. İşte 2 ya da 3 gün olarak planlanmış turlardan bazıları. *

Örnek çıkış metni:
23 Nisan Ulusal Egemenlik ve Çocuk Bayramı'nın bu yıl Cuma gününe denk gelmesi, iş yoğunluğundan bunalanlar için mini bir tatil olanağı yaratacak. Bahar havası iyiden iyiye kendisini hissettirirken, "bu mevsimde en güzel neresi gezilir, en güzel tatil nasıl yapılır?" diye düşünmeye pek gerek yok. İşte 2 ya da 3 gün olarak planlanmış turlardan bazıları.

Üstteki örnekten de görüleceği üzere bu işlem, etiket temizleme dışında sayfadaki resim kodlarının kaldırılması ve çift tırnak gibi HTML içinde özel kodlama ile yazılan bazı karakterlerin de metin dönüşümünün yapılması adımlarını içermektedir.

İnternet sayfaları genellikle menüler, reklamlar, linkler, okuyucu yorumları gibi asıl metnin dışında birçok ek yazı içermektedir. Uygulamada hedeflenen sağlıklı sonucun alınabilmesi için bu tür ek içeriklerin temizlenerek sadece habere ait metnin veritabanına kaydedilmesi gereklidir. Farklı haber kaynaklarından alınan sayfa içeriklerinde belli bir standart olmadığı için her haber kaynağına özel bir temizleme işlemi yapılmaktadır. Buna uygun olarak haber kaynaklarının sayfaları içerik olarak analiz edilmiş ve haber metninin başlangıç ve bitişini tayin etmek üzere belli ortak terim ve etiketler belirlenmiştir. Genel olarak sayfalar belli şablonlar

kullanılarak oluşturulduğu için aynı haber kaynağı içindeki tüm sayfaların sayfa akışı benzerdir. Örneğin bir kaynakta tüm haber metinleri gazetenin adını takip eden bir slogan ile başlamakta ve günün tarihi ile sonlanmaktadır. Uygulama bu iki bilgiyi tespit ederek arada kalan metni, haber içeriği olarak değerlendirmektedir.

3.4. Metin Analizi Yapılması

Toplanan metinlerden etiket listesi oluşturulma işlemi Metin Analizi aşamasında gerçekleştirilir. İlk adımda salt metin içinden etkisiz sözcükler ayıklanır. Bu işlem için dinamik olarak yönetilebilen bir etkisiz sözcük dosyası kullanılmaktadır [14]. Bu dosya dilimizde çok sık kullanılan ve bundan dolayı kelime frekans analizinde dikkate alınmaması gereken "ve", "veya", "bu", "şu" gibi 339 adet terimden oluşmaktadır. Zaman içinde yeni bir etkisiz kelime belirlenirse, o kelimeyi bu dosyaya eklemek yeterli olacaktır. Kelime analizi öncesinde son olarak metin içinde yer alan *′‘’′<>/:“”„,;.. !?()[]\″»«=* gibi noktalama işaretleri ve diğer özel karakterlerin temizleme işlemi gerçekleştirilmektedir [15].

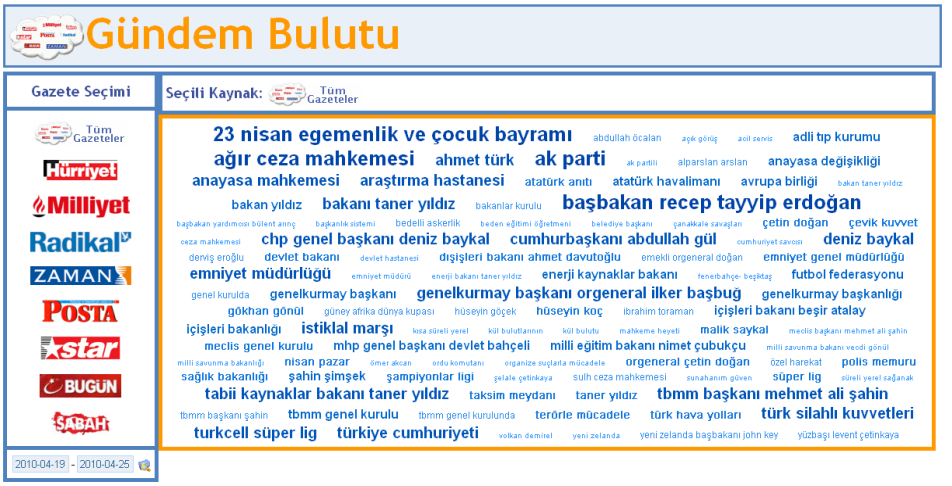
Bu adımda gerçekleştirilen diğer bir işlem ise kelimelerin tiplerinin belirlenmesi ve köklerinin bulunmasıdır. Bu tür işlemleri gerçekleştirecek bir uygulama geliştirmek başlı başına ayrı bir çalışma olacağından, benzer projelerde kullanılmış ve Türkçeyi destekleyen Zemberek [16] kütüphanesinden faydalanılmıştır. Zemberek, paralel olarak hem .NET hem de JAVA üzerinde geliştirilmekte olan bir kütüphanedir. Ancak ilk olarak JAVA versiyonunun geliştirilmeye başlanmasından ötürü .NET versiyonu bazı konularda geriden gelmektedir. Bu proje .NET üzerinde geliştirildiği için Zemberek kütüphanesinin .NET versiyonu kullanılmıştır. Ancak geliştirme sürecinde kodlar incelendiğinde sözlük veritabanının 2006 yılına ait olduğu görülmüştür. JAVA versiyonunda ise 2010 yılına ait çok daha doğru sonuçlar alınabilen sözlükler mevcut olduğu tespit edilmiştir. Buradan yola çıkılarak açık kaynak kodlu proje üzerinde çalışmalar yapılmış ve .NET versiyonundaki sözlükler güncel hale getirilmiştir.

İçerik işleme için gerekli tüm altyapı çalışmaları tamamlandıktan sonra yapılan ilk test çalışmalarında elde edilen etiketlerin sadece tek bir kelime içermesinden dolayı, bu listeden oluşturulan etiket bulutunun, gündemi yansıtmada tatmin edici sonuçlar vermediği görülmüştür. Bu eksikliğin giderilmesi konusunda yapılan araştırmalar sonucunda etiket belirleme işleminde iki yeni metodun uygulanmasına karar verilmiştir. Bunlardan ilki, haber metinlerinde büyük harfle başlayan özel isimlere ait kelime gruplarının belirlenmesidir. Bu şekildeki özel isimlerin grup halinde bir etiket olarak atanmasına karar verilmiştir. Böylece gündemdeki önemli kişi, yer ve kurum isimlerinin etiket bulutunda gösterilmesi sağlanmıştır. Bu adım içerik temizleme işleminden sonra gerçekleşmektedir. Metni oluşturan kelimeler sırayla kontrol edilir ve büyük harfle başlayan ardışık kelimeler özel isim olarak değerlendirilir. Kullanılması uygun görülen ikinci metot ise metinlerde sürekli tekrar eden ikili ve üçlü kelime gruplarının tespit edilmesi ve etiket olarak değerlendirmeye alınmasıdır. Özel isim listesi çıkarıldıktan sonra bu adıma geçilir. Kelime grupları çıkarılmadan önce metin içindeki bağlaç, edat, zamir ve zarf türündeki kelimeler ile etkisiz sözcükler temizlenir. Daha sonra metin içinde ikili ve üçlü gruplar halinde kayan bir pencere ile her kelime grubunun kaç kez geçtiği bulunur. Üçlü gruplar içinde yer alan ikili gruplar ise listeden çıkarılır. Son olarak elde edilen iki liste tek bir listede toplanır ve tekrarlama sayısı birden fazla olanlar etiket listesine dahil edilir. Böylece klasik bir etiket bulutu oluşturma sürecine göre çok daha özel bir filtreleme işlemi yapılmış ve haber gündemini daha iyi yansıtmayı başaran bir etiket listesi elde edilmiştir.

3.5. Arabirim

Uygulamaya aynı anda birden fazla kullanıcının erişeceği göz önüne alınarak altyapıda performans açısından başarılı sonuçlar veren ASP.NET ve IIS (Internet Information Services) internet sunucusu kullanılmıştır.

Ana sayfası Şekil 6'da görülen Gündem Bulutu uygulamasına bir internet gezgini aracılığı ile ulaşılmaktadır. Uygulama ilk açıldığında son bir haftaya ait haberlerin içeriğinden oluşan bir etiket bulutu gösterilir.



Şekil 6. Gündem Bulutu ana sayfası

Kullanıcılar, sayfanın sol tarafındaki gazete seçimi menüsünden bir logoyu seçmediği sürece uygulamada gösterilen etiket bulutları tüm gazetelerden toplanan içerik ile oluşturulur. Eğer bu menüden bir seçim yapılırsa, etiket bulutu oluşturmaya sadece o gazeteden alınan içerikler dâhil edilir. En üstteki "Tüm Gazeteler" linkine tıklanarak etiket bulutu ilk haline döndürülebilir. Buna ek olarak sayfanın sol altındaki bölümden başlangıç ve bitiş tarihleri seçilerek sadece o tarihler arasındaki haberlere ait etiketlerin değerlendirmeye alınması sağlanır.

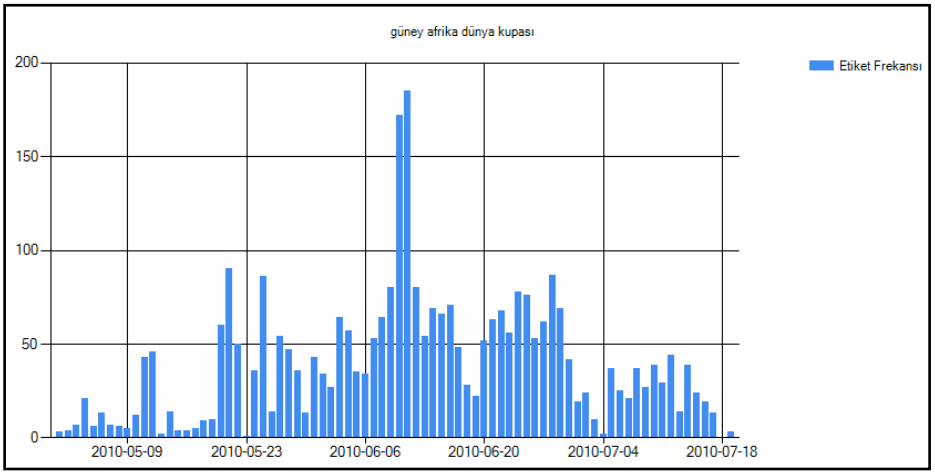
Kullanıcılar, bulut içindeki bir etikete Şekil 7'de gösterildiği gibi tıklayarak o etiketin eşlendiği haberlere ait tarih, başlık ve özet bilgilerinin listelendiği Şekil 8'de görülen sayfaya ulaşırlar.

Şekil 7. Bir etiketin seçilmesi

Şekil 8'de görülen listeden istenen bir haberin linkine tıklayarak haberin yayınlandığı asıl siteye ulaşılıp, haberin tüm metni okunabilir.

Şekil 8. Seçilen etikete ait haberlerin listelenmesi

Uygulamanın bir diğer önemli özelliği de belli bir etiketin gündemde olma frekansının grafik olarak kullanıcılara sunulmasıdır. 2010 yılının en önemli spor olayı olan Güney Afrika Dünya Kupası'na ait etiketin grafiği Şekil 9'de verilmiştir. Etiketin frekansı Mayıs ayının başından itibaren giderek yükselmiş ve kupanın başlangıç tarihi olan 11 Haziran'da en üst seviyeye ulaşmıştır. Bir ay boyunca belli bir değerde devam eden frekans, kupanın bitmesiyle gündemden düşmüştür.



Şekil 9. "Güney Afrika Dünya Kupası" etiketinin frekans grafiği

4. SONUÇLAR

Bu çalışmada, Türkçe gazete sitelerinde yayınlanan haberleri anlık olarak takip ederek, içeriklerini toplayan, analiz eden ve elde edilen bilgilerden otomatik olarak oluşturduğu etiket bulutu aracılığı ile kullanıcıların haberlere erişimini sağlayan Gündem Bulutu adında bir uygulama geliştirilmiştir. Son yirmi yılda gerçekleşen hızlı gelişim ile bugün milyonlarca insan ihtiyaç duydukları bilgilere erişmek için internet ortamını kullanır hale gelmiştir. İnternetin yaygınlaşmasıyla birlikte sürekli büyüyen bir bilgi havuzundan ihtiyaç duyulan bir bilgiye

ulaşmak gün geçtikçe daha zor bir hale gelmektedir. Bu çalışma kapsamında geliştirilen Gündem Bulutu uygulaması kullanıcıların bilgiye ulaşmalarına aracılık eden bir Bilgiye Erişim Sistemidir.

Geliştirilen uygulamada içeriğin toplanması ve analiz edilmesi sonucunda elde edilen bilgiler kullanılarak etiket bulutu oluşturulmaktadır. Etiket bulutlarının genel kullanımında, geçmişe yönelik bir tarihçe tutulması bulunmamaktadır. Buna karşın tasarlanan uygulamada sunulan içeriğin haberlerden oluştuğu göz önüne alınmış ve kullanıcıların istedikleri bir tarih aralığını seçerek o döneme ait gündemi yansıtan etiket bulutunu görebilmelerine olanak sağlayacak bir altyapı oluşturulmuştur. Böylece kullanıcıların hem en güncel gündemi hem de isterlerse eskiye dönük gündemi görebilmeleri sağlanmıştır. Buna ek olarak belli bir etiketin gündemde olma frekansı da görsel olarak kullanıcılara sunulmaktadır.

Çalışma sonucunda ortaya çıkan uygulamanın temel faydası Türkiye'nin gündemini farklı kaynaklardan toplanan bilgiler ile tek bir arayüzden takip edilmesine olanak sağlamasıdır. Bu tür bir günlük kullanımın yanı sıra sosyal bilimler gibi alanlarda yapılan araştırmalarda akademik bir araç olarak da faydalanılabilir. Ayrıca ticari hayatta haber ajansları gibi firmalar çalışmalarında bu altyapıyı kullanabilirler.

Zaman içinde sistemin işleyişinin gözlenmesi ve etiket bulutlarının iyileştirilmesi yönünde çalışmalar yapılması planlanmaktadır. Ayrıca metin madenciliği alanında yapılan yeni çalışmalarda elde edilen bulgular ve geliştirilen yeni tekniklerden faydalanılarak etiket belirleme işleminin daha başarılı bir hale getirilmesi mümkün olabilecektir.

REFERENCES / KAYNAKLAR

[1] Manning, C.D., Raghavan, P., Schütse, H., "Introduction to Information Retrieval", Cambridge University Press, Cambridge, 2008, 1-6.

[2] Bilgi Erişim Sorunu, Available from: http://yunus.hacettepe.edu.tr/~tonta/courses/spring2009/bby220/bby220-bilgi-erisim-sistemleri-2009-1.ppt, [accessed May 25, 2010].

[3] Tonta, Y., Bitirim, Y., Sever, H., Türkçe Arama Motorlarında Performans Değerlendirme, Total Bilişim Ltd. Şti., Ankara, 2002, 7-13.

[4] Maron, M.E., "Probabilistic Retrieval Models", Progress in Communication Sciences Volume 5:145-176, 1984.

[5] Tan, P., Steinbach, M. ve Kumar, V., "Introduction to Data Mining", Pearson Addison Wesley, Boston, 2005, 1-10.

[6] Indurkhaya, N. ve Weiss, S.M., "Predictive Data Mining", Morgan Kaufmann Publishers Inc., San Francisco, 1998, 1-19.

[7] Hearst M.A., "What's Up With Tag Clouds", Visual Business Intelligence Newsletter, Mayıs, 2008.

[8] Ames, M. ve Naaman, M., "Why We Tag: Motivations for Annotation in Mobile and Online Media", Conference on Human Factors in Computing Systems (CHI 2007), San Jose, California, USA, 28 Nisan – 3 Mayıs 2007, 971-980.

[9] Zhou, C. ve Bénel, A., "From the crowd to communities: New interfaces for social tagging", Proceedings of the 8th International Conference on the Design of Cooperative Systems, Carry-le-Rouet, Provence, France, 20-23 Mayıs 2008, 242-250.

[10] Tag Cloud – Wikipedia, Available from: http://en.wikipedia.org/wiki/Tag_cloud, [accessed July 12, 2010].

[11] The New York Times, Available from: http://www.nytimes.com, [accessed October 18, 2010].

[12] Donath, J., "A semantic approach to visualizing online conversations", Communications of the ACM, Volume 45, Issue 4:45-49, 2002.

[13] Hearst, M.A. ve Rosner, D., "Tag Clouds: Data Analysis Tool or Social Signaller?", HICSS 2008, Waikoloa, Big Island, Hawaii, USA, 7-10 Ocak 2008, 160-160.

[14] Manning, C.D., Schütse, H., "Foundations of Statistical Natural Language Processing", MIT Press, Cambridge, 1999, 532-534.

[15] Goyvaerts, J., Levithan, S., "Regular Expressions Cookbook", O'Reilly Media, Inc., Sebastopol, 2009, 25-41.

[16] Nzemberek Project, Available from: http://code.google.com/p/nzemberek, [accessed May14, 2010].