YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

TÜRKÇE İÇERİKLERDEN OTOMATİK ETİKET BULUTU OLUŞTURMA

Bilgisayar Müh. Serdar SAVAŞAN

FBE Bilgisayar Mühendisliği Anabilim Dalı Bilgisayar Mühendisliği Programında Hazırlanan

YÜKSEK LİSANS TEZİ

Tez Danışmanı **:** Yrd. Doç. Dr. Banu DİRİ (Y.T.Ü.)

İSTANBUL, 2011

İÇİNDEKİLER

Sayfa

KISALTMA LİSTESİ

ASP.NET	Active Server Pages .NET
DDİ	Doğal Dil İşleme
GZIP	GNU Zip
HTML	HyperText Markup Language
HTTP	Hypertext Transfer Protocol
IIS	Internet Information Services
IMDb	Internet Movie Database
IR	Information Retrieval
RSS	Really Simple Syndication
SQL	Structured Query Language
URL	Uniform Resource Locator
W3C	World Wide Web Consortium
WWW	World Wide Web
XML	eXtensible Markup Language

ŞEKİL LİSTESİ

ÇİZELGE LİSTESİ

Sayfa

ÖNSÖZ

Günlük hayatımızın çok önemli bir parçası haline gelen bilgisayar, akıllı telefon ve benzeri araçlar sayesinde hemen hemen her yerden bilgiye erişmek mümkün hale gelmiştir. Gelişmelerden anında haberdar olmak, yol ve hava durumunu öğrenmek, finans bilgileri gibi anlık verilere hızlı ve kolay bir şekilde ulaşmak, gündelik yaşama çok olumlu katkılar sağlayan etmenlerdir. Bu tür içerikleri sunan Bilgiye Erişim Sistemleri, geliştiriciler tarafından sürekli olarak yenilenmekte ve daha kolay ve etkili bir şekilde kullanılabilmeleri için çalışmalar yapılmaktadır. Bu sistemlerin günümüzde kazandığı önemden yola çıkılarak, bu çalışmada metin madenciliği teknikleri incelenmiş ve Türkçe kaynaklardan güncel haberleri derleyerek otomatik etiket bulutu oluşturan Gündem Bulutu adında bir uygulama geliştirilmiştir.

Tez çalışmam boyunca değerli fikir ve önerileriyle beni yönlendiren ve bilgi ve deneyimlerini benimle paylaşarak büyük katkı sağlayan Yrd. Doç. Dr. Banu Diri'ye çok teşekkür ederim.

Bana her zaman maddi ve manevi destek olan sevgili eşime ve aileme…

ÖZET

Bilgi teknolojileri ve internetin gelişmesiyle şekillenen iletişim ortamı, her geçen gün daha da yoğun bir bilgi trafiği ile karşı karşıya kalmamıza neden olmaktadır. Bilginin zaman ve mekândan bağımsız bir şekilde paylaşılabilmesi de beraberinde bilgi kirliliğini getirmektedir. Bu konularda yapılan çalışmalarda, kişilerin ihtiyaç duydukları, değerli, güvenilir ve güncel bilgiye ulaşmaları hedeflenmektedir. Yapısal olmayan verilerin işlendiği Metin Madenciliğinde, bilgisayar destekli metin analizi yapılarak yüksek kalitede bilgi elde edilmesi sağlanmaktadır. Metin özetlerinin çıkarılması ve dokümanların etiketlenmesi gibi işlemler sayesinde, kullanıcılar bilgiye daha rahat bir şekilde ulaşabilmektedir. Etiket Bulutları, kelimelerin farklı yazı biçimlerinde sunulduğu ağırlıklı listelerdir. Bu görsel erişim araçları metin madenciliği tekniklerinden faydalanarak, kullanıcıları internet ortamında sunulan bilgilere yönlendirir. Gazetelerin internet sitelerinden günlük haberleri okumak, internetin en yaygın kullanım alanlarından birisidir. Genel olarak kullanıcılar gündemi, tercih ettikleri tek bir haber kaynağından 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. Bu sistem yayınlanan yeni içerikleri sürekli takip ederek derlemi anlık olarak güncellemektedir. Toplanan haber içerikleri Türkçe diline özgü analiz edilmekte ve otomatik oluşturulan bir etiket bulutu gösteriminde kullanıcıya sunulmaktadır. Böylece farklı gazetelerde yayınlanan aynı konulu haberlere tek bir noktadan erişmek mümkün olmuş ve gündemin takip edilmesi sağlanmıştır.

Anahtar Kelimeler: Bilgiye Erişim Sistemleri, Metin Madenciliği, Metin Analizi, Etiket Bulutu, Haber Derleme, Doküman Etiketleme

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. Information pollution is posed by the time and location independent sharing of information. Studies on these issues aim to provide the needed, valuable, reliable and updated information to the people. Text Mining works on unstructured data to achieve high-quality information by using computer-aided text analysis. 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 weighted lists composed of words, presented in different fonts and positions. These visual tools direct users to the information provided on the internet by taking advantage of text mining techniques. One of the most common uses of Internet is reading daily news from the newspaper sites. Generally the users prefer to follow the agenda from a single news source. 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 system updates the compilation by tracking the newly published content continuously. The collected news contents are analyzed according to Turkish language and they are presented to the user by means of an automatically generated tag cloud. As an outcome of this study, it has been possible to follow the agenda and to access to similar news published in different newspapers from a single point.

Keywords: Information Retrieval Systems, Text Mining, Text Analysis, Tag Cloud, Collecting News Compilation, Document Tagging

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.

Bilgiye erişimin yakın zamana kadarki en yaygın aracı kütüphaneler ve kurumların arşivleri iken, günümüzde internet ağı ve elektronik arşivleme sistemleri, hızlı bir şekilde bu kütüphane ve arşivlerin yerini almaya başlamışlardır. Bu gelişmeler göz önüne alındığında, bilgi lojistiğinin hemen hemen tamamının, yakın gelecekte sadece elektronik ortamda yapılacağını tahmin etmek çok zor olmayacaktır.

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 bir yapıda pratikte kullanılamaz hale gelmektedir. Tam metin arama motorları otomatik çalıştığı için bu sorun yaşanmıyor olsa da, bu teknikte de kullanıcıların sorgulama sonuçlarından, aradıkları doğru dokümana erişmeleri konusunda sıkıntılar yaşanmaktadır. Çünkü bu tür bir sorgulamada bir dokümanın sonuç listesine gelmesi için aranan kelimenin o doküman içinde bir kez bile geçiyor olması yeterli olmaktadır. Pratik olmaktan uzak bu durum, arama sonuçlarının kalitesinin düşük olmasına dolayısıyla da, zaman ve çaba maliyetinin artmasına neden olmaktadır.

Diğer yandan muazzam bir 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 sorunun çözümü için de en önemli araçlar arama motorlarıdır. Ne var ki arama motorlarından da her zaman istenen sonuçlar alınamamaktadır. Ancak bunun tek nedeni olarak ta sadece arama motorlarının teknik eksikliklerini göstermek doğru olmayacaktır. Arama motorundan sorgu yapan kullanıcının da konu hakkında belli bir ön fikre sahip olması ve doğru anahtar kelimeleri girmesi gerekmektedir.

Son dönemlerde gelecek hakkında yapılan değişik tahminlerden birisi, özellikle yazılı basının

sadece internet aracılığı ile yapılacağı ve kâğıt üzerindeki yayıncılığın sona ereceği şeklindedir. Bu tahminin tamamen gerçekleşmesi çok uzun zaman alacak gibi görünse de yakın zamanda büyük ölçüde gerçekleşeceğini destekleyen ciddi gelişmeler olmaktadır. Örneğin internet kullanımı geçmişte yapılan tahminlerin çok daha üzerinde bir hızda yaygınlaşmaktadır. Yakın bir zamanda iletişimin neredeyse tamamının sadece internet üzerinden yapılacağını söylemek çok yanlış olmayacaktır. Buna ek olarak artan çevre duyarlılığı da herkesin hem fikir olduğu ortak bir görüştür. Yayın kuruluşları açısından da internet üzerinden yapılan yayının zaman ve insan kaynağı maliyeti, diğer yöntemlere göre çok daha düşük olmaktadır. Bugün için gazetelerin kâğıt ortamındaki yayıncılığı devam etmekle birlikte gazeteleri sadece elektronik ortamda takip edenlerin sayısı da her geçen gün artmaktadır.

İnternet ortamında gazete okumak isteyenler ilgili gazetenin sitesine girerek günlük haberlere kolayca ulaşabilirler. Ancak özel bir konuya ait haberlere ulaşılmak istenildiğinde durum biraz daha zorlaşacaktır. Kullanıcı bu amaçla ilgili haberleri bulabilmek için internet arama motorlarına anahtar kelime girerek sorgulama yapabilir. Ne var ki girilen anahtar kelimeyi içeren ilgili ilgisiz birçok sayfa arama sonuçlarında listelenecektir. Bu durumda kullanıcının, listelenen sonuçları teker teker açıp kontrol etmesi gereklidir. Sonuç olarak zaman maliyetinin artmasından dolayı elde edilen fayda azalacaktır. Buna ek olarak, sorgulama sadece gazetelere ait siteler üzerinde yapılmak istendiğinde, işlem biraz daha zorlaşacak ve arama motorlarının yeterli olmadığı durumlarda, gazetelerin siteleri tek tek açılıp, site içi arama yapılması gerekecektir. Bu tür ihtiyaçlardan yola çıkarak internet üzerindeki günlük haberlere erişim konusu problem olarak ele alınmış ve üzerinde çalışılmaya değer bulunmuştur.

Bu çalışmada, teknolojik gelişmelerin nihai hedefi olan zaman tasarrufu ve kolaylık amacına hizmet etmek üzere, internet ortamındaki günlük haberlere erişimi daha işlevsel hale getiren ve farklı gazetelerde yayınlanan aynı konulu haberleri kullanıcıya bir arada sunan zeki bir haber derleme ve erişim aracı geliştirilmiştir.

Tez sürecinde doğal dil işleme ve makine öğrenmesi üzerine yapılmış önceki çalışmalar incelenmiş ve Türkçe içeriklere ait etiket bulutu (tag cloud) oluşturmaya yönelik araştırma yapılmıştır. Yapılan bu çalışmalar sonucunda elde edilen bilgilerden faydalanılarak otomatik etiket bulutu oluşturan bir uygulama geliştirilmiştir. 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 verilere etiket bulutu şeklinde erişim sağlayan bir

internet sitesi olarak tasarlanmıştır. Türkçe desteği sayesinde güncel haberlere ortak terimler üzerinden hızlı bir erişim olanağı sunan ve sürekli kendisini güncelleyen özel bir **Bilgiye Erişim Sistemi (Information Retrieval System)** oluşturulmuştur.

Tezin ikinci bölümünde konuyla ilgili genel bilgiler ve kavramlar anlatılmıştır. Üçüncü bölümde ise geliştirilen uygulamanın görsel bileşeni olan Etiket Bulutları hakkında genel bilgi verilmiştir. Bir sonraki bölümde Gündem Bulutu uygulamasının tasarım aşamaları ve mimarisi anlatılmıştır. Beşinci bölümde Gündem Bulutu uygulamasının kullanımı hakkında bilgi verilmiş ve farklı kullanım şekilleri gösterilerek elde edilen bilgiler karşılaştırılmıştır. Son bölümde ise yapılan çalışmada elde edilen sonuçlara ait detaylara yer verilmiştir.

2. BİLGİYE ERİŞİM VE UYGULAMA ALANLARI

Veri ve bilgi erişiminde son derece önemli olan hız ve kolaylık, teknolojik gelişmelerin amaçları arasında en öne çıkanlarıdır. Sürekli yapılan çalışmalar sonucunda bu amaçlara büyük oranda ulaşılmıştır. Araştırmacılar yaptıkları çalışmanın amacı ve kapsamı ne olursa olsun, sonuca varabilmek için bilgiye erişim sürecinden geçmektedir. Araştırmanın sadece deneysel olduğu durumlarda bile, araştırmanın kuramsal temelini oluşturmak ve elde edilen verileri yorumlamak ve karşılaştırmak için yine bilgiye erişim süreci ihtiyaç olarak belirir. Diğer yandan toplumsal hayatta bilgi paylaşımı, dil öğesinin varlığından beri mevcuttur. Bu bağlamda insanlık tarihinde, bilgiye erişim kavramının sözel iletişim ile birlikte var olmaya başladığını söyleyebiliriz.

Bu çalışma sonucunda ortaya çıkan uygulamanın temel malzemesi; veri ve bilgidir.

Veri (Datum), sözcük anlamı, "bir sonuca varabilmek için gerekli olan ilk bilgi" olmakla birlikte, değişik kaynaklarda farklı şekillerde tanımlanmıştır. Bir tanıma göre veri; işlenmemiş bilgi ya da sistemlerin kendi dışından elde ettikleri, gerçek olaylar ve durumlar ile ortaya çıkan değerlerdir (İslam, 2005). Bir başka ifade ile veri henüz "işlenmemiş kanıtlardır" (Karasar, 2003). Bu tanımların bağlamında, verinin işlenmemiş veya yorumlanmamış, nesnel ham bilgiler olduğunu söyleyebiliriz.

Bilgi (Information), elde edilen verilerin sistemin işine yarayacak şekilde işlenmesi, analiz edilmesi, sıralanması, diğer veri ve bilgilerle birleştirilmesi, özetlenmesi, raporlanması gibi faaliyetlerin ortaya çıkardığı bir veri bütünüdür.

Genel kullanımda sürekli birbirinin yerine kullanılan bu iki kavram, tanımlara baktığımızda birbirinden farklıdır. Veri, işlenmemiş, yorumlanmamış ve sadece sayısal değerlerden (olgusal veri) oluşurken, bilgi olgusal verilerin işlenmiş, yorumlanmış ve karar vericinin işine yarayacak anlamlı hale getirilmiş halidir.

Bu çerçevede, yapılan bu çalışmanın en önemli malzemesi veridir. Çünkü etiket bulutu uygulaması, ham kelime frekanslarının görselleştirilmiş halidir. Etiket bulutu, aranılan metnin kolay ve hızlı ulaşımına hizmet edeceğinden, söz konusu görselleştirilmiş verinin (etiket bulutunun) buradaki işlevi **Bilgiye Erişim (Information Retrieval)** olarak tanımlanabilir.

2.1 Bilgiye Erişim

Bu çalışma sonucunda ortaya çıkan uygulamanın temel işlevi toplanan verilerin işlenmesi ve

sunulması ile bilgiye erişim sağlamaktır.

“Bilgiye Erişim” deyince akla çok geniş bir anlam yelpazesi gelebilir. Örneğin, bir telefon numarasını öğrenmek için rehbere başvurulması, ya da bir arkadaşın deneyiminden faydalanmak üzere deneyimi ile ilgili onunla konuşmak, internet ortamında bir ürünü en uygun fiyata satan yeri aramak, ya da sadece ehliyet seri numarasını öğrenmek için ehliyetin cüzdandan çıkarılması bilgiye erişimin anlamı kapsamına girer.

Ancak akademik ortamda “Bilgiye Erişim” bir terim olarak karşımıza çıkmaktadır.

Bilgiye Erişim genellikle bilgisayarlarda depolanan yapılandırılmamış metin içerikli dokümanlardan oluşan büyük miktardaki veri yığınları içinden ihtiyaç duyulan bilgiyi bulmak olarak tanımlanabilir (Manning vd., 2008). Tonta (2009) 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 tanımlar göz önüne alındığında, bilgiye erişim sadece belli mesleklerdeki insanların ilgilendiği bir konuymuş gibi görünebilir. Ancak özellikle son yirmi yılda gerçekleşen hızlı gelişim ile bugün milyonlarca insan internet ortamında ihtiyaç duydukları bilgilere erişmek için arama yapar hale gelmiştir.

Bilgi Erişimi, Sistem, Bilim ve Dil olarak üç temel öğeye dayanmaktadır (Tonta, 2009).

Dilbilimin disiplinler arası kullanımlarından birisi Bilişim alanındaki kullanımıdır. Dilbilim kendi içinde, Konuşma Dilbilgisi, Yazı Dili Bilimi, Sözcük Bilimi, Biçimbilim (Morfoloji), Ses Bilgisi, Ses Bilimi Pragmatik, Anlambilim, Dil Felsefesi, Sözdizimi, Metin Dilbilim olarak alt dallara ayrılır.

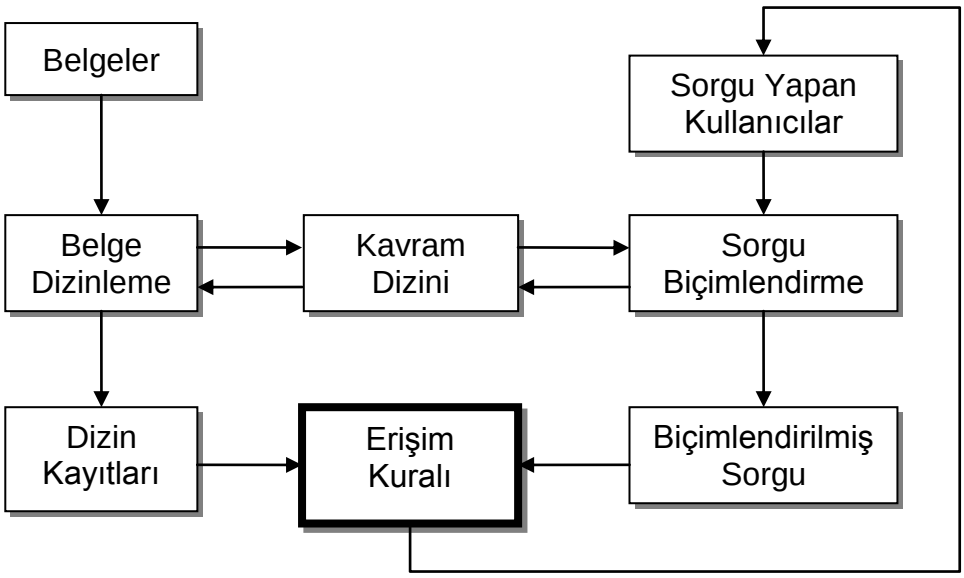
Biçimbilim (Morfoloji), kelime yapısı ve isim, sıfat, fiil çekimi gibi değişken sözcük biçimlerini inceleyen bilim dalıdır [1]. Bu çalışmada Biçimbilim teknikleri kullanılarak Türkçe kelimelerin kök analizi yapılmış ve türleri belirlenmiş, ayrıca metin içindeki özel isimler tespit edilerek etiket bulutunu oluşturan veri seti elde edilmiştir.

2.2 Bilgiye Erişim Sistemleri

İnternetin yaygınlaşmasıyla birlikte sürekli büyüyen bir bilgi havuzundan ihtiyaç duyulan bir bilgiye ulaşmak, günümüzün en revaçta konularından birisi haline gelmiştir. Anahtar kelimelerle yapılan metin aramalarında, arama yapan kişinin isteklerine en uygun metnin veya

metin listesinin çıkarılması nihai amaçtır. Bu işlevleri yerine getiren yapılara Bilgiye Erişim Sistemleri denilmektedir. Kullanıcıların bilgi ihtiyaçlarını karşılamak üzere bir belge derlemindeki ilgili kayıtların tümüne ilgi sırasına göre erişim sağlamak ve ilgili olmayan kayıtları ayıklamak Bilgi Erişim Sistemlerinin temel işlevidir.

Bilgi Erişim Sistemlerinin belgelere erişimi doğru bir şekilde gerçekleştirebilmesi için belli koşullar sağlanmalıdır. Belgeler, elle veya otomatik olarak gerçekleştirilen dizinleme işlemleri sırasında çıkarılan temel özelliklerinden faydalanılarak elde edilen içerik belirteçleri ile derleme eklenmelidir. Belge erişimi bu belirteçler aracılığı ile yapıldığı için belgenin tamamını temsil etmeleri büyük önem taşımaktadır. Bilgi erişiminin başarıyla sonuçlanabilmesi bu belirteçlerin doğru bir şekilde kullanıldığı sorguların yapılmasıyla sağlanabilir. Sonuç olarak kullanıcılar bilgi ihtiyaçlarını tanımlarken kullandıkları terimlerin içerik belirteci olarak atandığı belgelere erişim sağlarlar (Tonta vd., 2002). Klasik bir Belge Erişim Sisteminin tasarımı Şekil 2.1'de gösterilmiştir.



Şekil 2.1 Belge erişim sisteminin mantıksal gösterimi (Maron, 1984)

İçerik Belirteçleri, bütünlük gösteren bir metni temsil eden kelime veya kelime gruplarıdır. Bu belirteçler gerek duyulduğunda ilgili belgeye erişim için kullanılacağından metni doğru bir şekilde temsil ediyor olmaları gerekmektedir. Belirteç kümesinin oluşturulması elle veya otomatik olarak gerçekleştirilebilir. Bu çalışmada belirteçler uygulama tarafından otomatik olarak tayin edilmiştir.

2.3 Bilgiye Erişim Sistemlerinden Elde Edilen Sonuçlar ve Değerlendirme Kriterleri

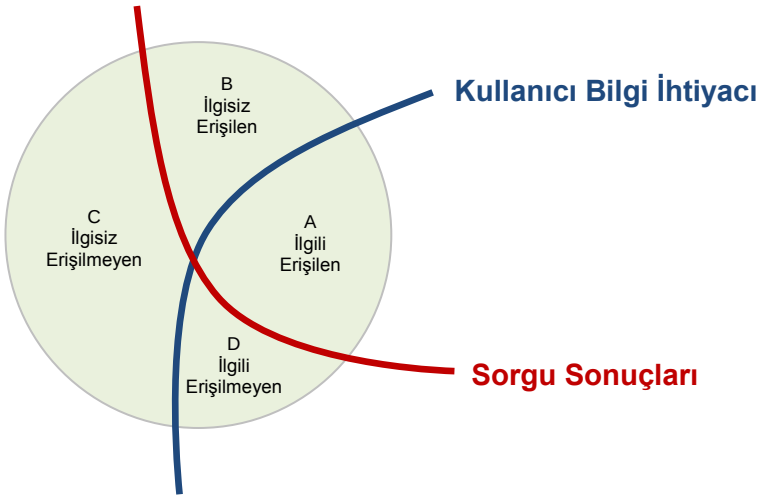
Bilgiye erişim sistemlerinin, arayan kişiyi aradığı bilgiye ulaştırma görevinin başarısı, bulunan sonuçların özellikleriyle değerlendirilse de, arayan kişinin konuya hâkimiyeti, konuyla ilgili anahtar sözcük veya terminoloji bilgisi oldukça önemlidir. Bir başka deyişle, bilgiye erişme eylemi, arayıcının katılımı ile gerçekleşen katılımlı bir eylemdir.

Erişimi gerçekleştiren sistem kullanıcısının bu özelliğinin en uygun seviyede olduğunu varsayarsak, ideal bir bilgi erişim sistemin özelliklerinden şu şekilde söz edebiliriz;

- Sonuçlar hem nesnel hem de öznel olarak ilgili olmalıdır.
- İlgili belgelerin tümüne erişilmeli ve salt ilgili belgelere erişim sağlamalıdır.
- Birbirine benzeyen belgeleri bir araya getirmeli, benzemeyenleri ayıklamalıdır.

(Tonta, 2009)

Bir başka deyişle, mükemmel bir sistemde Şekil 2.2'de D harfi ile gösterilen alan var olmamalıdır. Teknik olarak bu alanın var olmaması pek mümkün olmasa da, bu alanın olabildiğince dar olması başarı açısından önemlidir.



Şekil 2.2 Bilgi erişim sonuçlarının küme gösterimi (Karagedik ve Önal, 2010)

Bilgiye Erişim Sistemlerinin Değerlendirme Ölçütleri

- **İlgililik (Relevance):** Kullanıcının bilgiye erişim sisteminden elde ettiği sorgulama sonuçlarının kendi ihtiyaçlarıyla ne kadar iyi örtüştüğünün göstergesidir.
- **Hassasiyet (Precision):** Erişilen ilgili belge sayısının, erişilen bütün belgelere oranıdır.
- **Anma (Recall):** Erişilen ilgili belge sayısının, bütün ilgili belgelere oranıdır (Singhal, 2001).

- **Posa:** Erişilen ilgisiz belge sayısının, bütün ilgisiz belgelere oranıdır (Tonta, 2009).
- **Genellik (Generality):** İlgili belgelerin tüm derlemdeki belgelere oranıdır (Huijsmans ve Sebe, 2003).

2.4 Bilgiye Erişim Uygulamaları

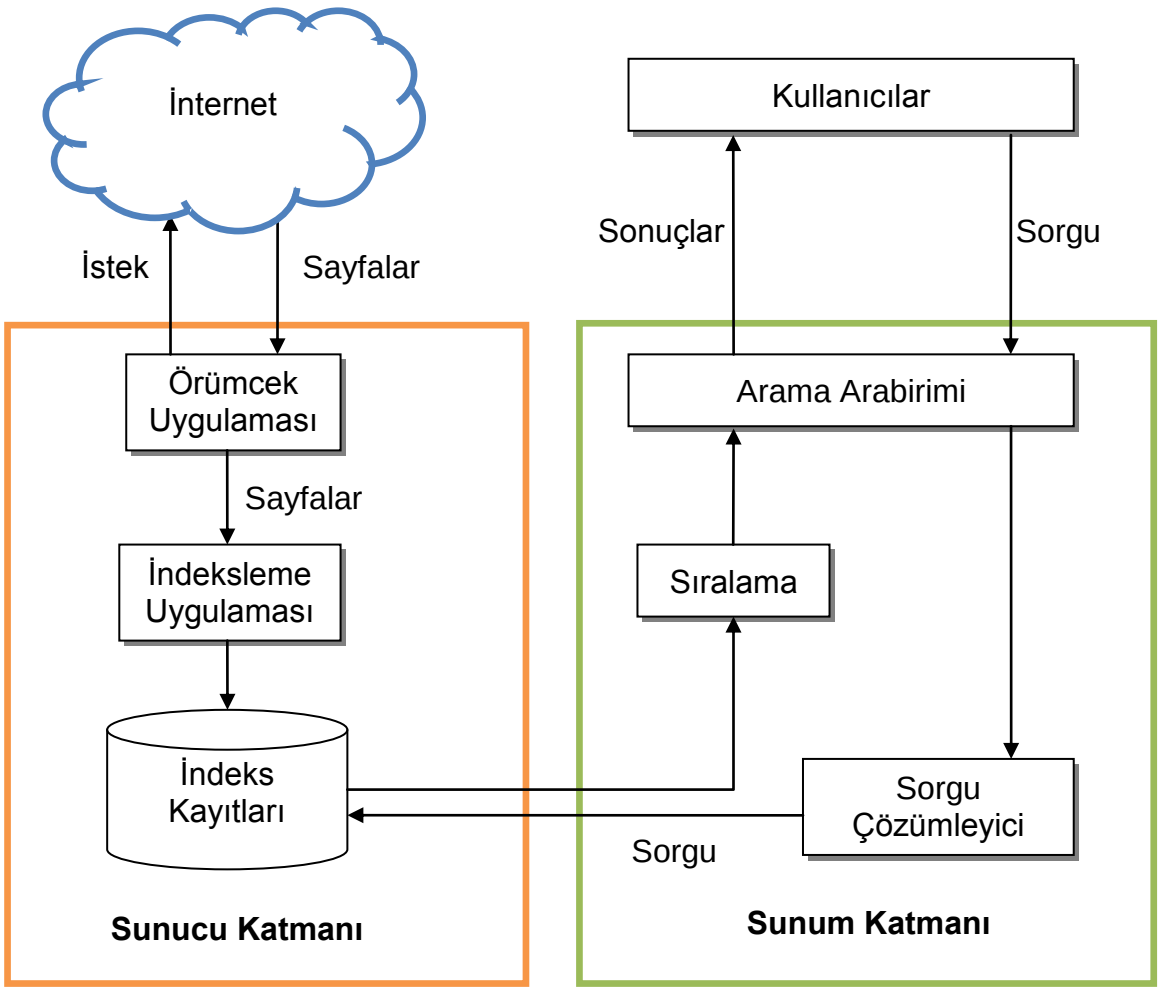
Bilgiye erişim sistemleri, günlük hayatta maruz kaldığımız yoğun bilgi bombardımanının etkisini azaltmaya yönelik hizmet vermektedir. Birçok üniversite ve kütüphane, kitap, dergi ve diğer belgelere erişim sağlamak için bilgi erişim uygulamaları kullanmaktadır. İnternet arama motorları ise günümüzün en yaygın bilgi erişim uygulamalarıdır.

Bilgiye Erişim uygulamaları ana hatlarıyla alttaki şekilde gruplanabilir;

- Elektronik Kütüphaneler (The Universal Digital Library, Google Books)
- Bilgi Filtreleme ve Öneri Sistemleri (Amazon, Netflix, IMDb, Last.fm)
- Medya (Görüntü, ses, resim, haber, günlük) Arama Sistemleri (Google Image Search, Bing Image Search)
- Anahtar kelimelerle arama hizmeti sunan İnternet Arama Motorları (Google, Yahoo, Bing)
- Doğal dil sorguları ile arama hizmeti sunan Soru Cevaplama Sistemleri (Ask.com, Yahoo Answers, Askville, AnswerBus)

2.4.1 Arama Motorları

Katlanarak artan verilere hızlı bir şekilde erişmek her geçen gün daha da önem kazanmaktadır. İnternet ortamında veriye ulaşım için genellikle arama motorları kullanılmaktadır. Arama motorlarının çalışma prensibi internet sayfalarının indekslenmesi ve ardından verilen sorgu ifadesi ile indekslenen sayfa içeriklerinin karşılaştırılmasına dayanmaktadır. Şekil 2.3'te bir arama motoruna ait mimari ana hatlarıyla görülmektedir. Bu sistemde bir kullanıcının gerçekleştireceği arama işlemi sonucunda belli bir sayfaya erişebilmesinin ön koşulu o sayfanın arama motoru tarafından daha önceden indekslenmesidir. Bu koşulu tamamlayan ikinci önemli nokta ise kullanıcının arama sırasında kullandığı anahtar kelimenin indeksleme işleminde o sayfa ile ilişkilendirilmesidir. İndeksleme işlemi sırasında elde edilen ortak anahtar kelimeler kullanılarak arama işlemi sonucunda birbiri ile ilişkili dokümanlar toplu bir şekilde listelenir. Arama motorlarında kayıtlı dokümanların içerdiği kelimelerle, tüm dokümanlardaki kelimelerin istatistikî kullanım örüntüleri, dokümanlar arasındaki ilişkiyi ifade etmektedir.



Şekil 2.3 Arama motoru mimarisi

Arama motoru sistemlerinin kısıtları

Karşımıza çıkan en temel sorun dokümanın içindeki kelimelerle, arama yapan kişinin kullandığı kelimelerin aynı anlama gelmesine rağmen yazılışlarının farklı olmasıdır. Kullanıcının konuya hâkimiyeti, konunun genel kullanımında anıldığı kelimelerin farklı olabilmesi ve hatta aradığı konu ile ilgili tanımlama sıkıntısı ise kullanıcıdan kaynaklanan kısıtlardır.

Bazı kelimelerin tek başına kullanıldığında taşıdığı anlam ile bir kelime grubu içinde kullanıldığında taşıdığı anlam farklı olabilir. Bir kelimenin başka bir kelime ile kullanımı durumunda, sistem kullanıcıyı farklı bir içeriğe yönlendirebilir.

Çok kısıtlı olmakla birlikte internet ortamında profesyonel bir biçimde işlenmiş, organize edilmiş ve yapılandırılmış şekilde sunulan veriler de mevcuttur. Ancak sıradan kullanıcıların aradığı bilgiler genellikle bu tarz verilerden oluşmamaktadır. İnterneti oluşturan içeriğin

büyük bir kısmı yapılandırılmamış metin olarak sunulmaktadır. Yapılandırılmamış veriler arama motorları tarafından doğru bir şekilde analiz edilmeden indekslenirse kalitesiz arama sonuçlarının ortaya çıkması kaçınılmaz olacaktır. Örneğin, doküman uzunluğu dikkate alınmadan yapılan frekans hesaplaması sonucunda oluşan arama veritabanından elde edilen sonuçlar sağlıklı olmayacak ve kullanıcının yaptığı sorguya göre daha üstte listelenmesi gereken dokümanlar alt sıralarda listelenecektir.

Arama sonuçlarının kullanıcıyı yanlış yönlendirebileceği ihtimalinden dolayı bu tür kısıtlar ciddi bir dezavantaj haline gelir ve elde edilen fayda ve kolaylığı önemli ölçüde azaltabilirler. Bundan dolayı kelime odaklı sorgulama tekniğiyle çalışan arama motorlarının bilgiye erişimde kullanılacak en tutarlı araç olmadığı söylenebilir. Arama motorlarının gelişimi sürecinde daha tutarlı arama sonuçları elde etmeyi hedefleyen çalışmalar sonucunda bilgisayar destekli Metin Madenciliği (Text Mining) gibi yeni yaklaşımlar ortaya çıkmıştır.

2.5 Metin Madenciliği

Metin Madenciliği en basit tanımıyla metinlerden yüksek kalitede bilgi elde edilmesi işlemidir. Bilgisayar destekli analiz tekniklerinden faydalanarak metin halindeki verilerden daha önceden bilinmeyen ilginç ve yararlı bilgiler elde eden bir çalışma alanı olarak oldukça yaygın bir hale gelmiştir. 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. Veri Madenciliğinden faydalanılarak bu veriler işlenmekte ve elde edilen bilgiler iş zekâsı uygulamalarında kullanılmaktadır. Aşağıda yapısal olmayan ve yapısal veri için ayrı ayrı örnek verilmiştir.

Yapısal olmayan veri örneği;

25 yaşındayım, bekârım, aylık gelirim 1500 TL, Edirne'de oturmaktayım, SSK'lıyım.

42 yaşındayım, evliyim, aylık gelirim 2500 TL, İstanbul'da ikamet ediyorum ve Bağ-Kur'a prim ödüyorum.

Yapılandırılmış veri örneği;

Çizelge 2.1 Yapılandırılmış veri örneği

Yaş	Gelir	İkamet	Sosyal Güvence	Medeni Durum
25	1500 TL	Edirne	SSK	Bekâr
42	2500 TL	İstanbul	Bağ-Kur	Evli

Yukarıdaki cümlelerin Çizelge 2.1'de gösterilen şekilde yapılandırılmış hale getirilmesiyle elde edilen veriyi, sınıflamak, işleyerek istatistikî analizler yapmak ve düzenleyerek belli bir yapıda saklamak çok daha kolaydır.

Erişilebilir ve kullanılabilir durumdaki verinin önemli bir bölümü metin veritabanlarında veya diğer söylemiyle doküman veritabanlarında bulunmaktadır. Bu nedenle özellikle internet ortamında yapılan bilgi aramalarında, metin madenciliği oldukça önem kazanmaktadır. Metin madenciliğinde, veri madenciliği, yapay zekâ, doğal dil işleme, bilgi yönetimi, bilgi erişimi, makine öğrenmesi ve istatistik gibi farklı disiplinler bir arada kullanılmaktadır.

Metin Madenciliği Aşamaları

- Kaynaklarının belirlenmesi
- Kaynaklara erişilmesi
- Toplanan verilerin ön işlemden geçirilmesi
- Metin analizi yapılması
- Elde edilen sonuçların derlenmesi ve sunulması

Metin Madenciliği Teknikleri

- Bilgi Çıkarma (Information Extraction)
- Özet Çıkarma (Summarization)
- Kümeleme (Clustering)
- Sınıflandırma (Classification)
- Birliktelik Analizi (Association Analysis)

Metin madenciliği süreçlerine geçmeden önce, bu süreçlerde kullanılan ve büyük miktardaki verilerin işlenmesinde insan yerine rol alan yapay zekâ ve istatistik unsurlarını içeren **Makine Öğrenmesi** kavramından bahsetmekte fayda vardır.

Metin veya veri madenciliğinde, uzman kişi ile uzman makine arasında bir işbölümü vardır. Uzman kişi, problemin tanımlanmasında, hedeflerin belirlenmesinde ve veri tabanının hazırlanmasında aktif rol oynar. Uzman makine ise her algoritma için geçerli olmamakla birlikte, hedeflerle uyuşan örüntüleri saptamak amacıyla verileri taramak ve bu örüntülere uygun kuralları öğrenmekten sorumludur. Uzman kişinin bilgi çıkarma sürecindeki bir diğer sorumluluğu ise; uzman makinenin öğrendiği bilginin incelenmesi, sorgulanması, sınanması, tutarsız/anlamsız bilgilerin ayıklanması, sorgulama ve sınama sonucu veri tabanının düzenlenmesi, hatalı bilgi ve verilerin düzeltilmesidir. Tüm bunların çerçevesinde, uzman makinenin asıl amacı, mümkün olduğunca bu işlemleri uzman kişinin yerine yapmaktır. Böylece uzman kişi ve işletmesi için çok değerli olan zaman maliyeti düşürülecek ve arttırılan zamanın daha değerli alanlarda kullanılabilmesi sağlanacaktır (Indurkhaya ve Weiss, 1998).

Çok büyük miktardaki verinin insan unsuru tarafından elle işlenmesi imkânsız olduğundan, bu ihtiyaca cevap verebilmek üzere makine öğrenmesi teknikleri geliştirilmiştir. Bu teknikler genellikle geçmişteki veriyi kullanarak yeni veri için en uygun modeli bulmaya çalışır.

Yapay Zekâ ve Dil Bilimi dallarının bir alt kategorisi olan **Doğal Dil İşleme**, makine öğrenmesi tekniklerinin oldukça yoğun kullanıldığı bir bilim dalıdır. Doğal Dil İşleme, ana işlevi doğal bir dili çözümleme, anlama ve yorumlama olan bilgisayar sistemlerini tasarlamak ve üretmek üzere dilbilimciler ile bilgisayar teknolojisi uzmanlarının ortak çalıştığı bir mühendislik alanıdır.

Doğal dil işleme, yapay zekâ (bilgi gösterimi, planlama, akıl yürütme, vb.), biçimsel diller kuramı (dil çözümleme), kuramsal dilbilim, bilgisayar destekli dilbilim ve bilişsel psikoloji gibi çok farklı alanlarda geliştirilmiş kuram, yöntem ve teknolojileri bir araya getirir. Bu alanda yapılan araştırmalardaki temel amaçlar genellikle, doğal dillerin işlev ve yapısını daha iyi anlamak, ara birim olarak doğal dil kullanmak ve bu şekilde bilgisayarlarla insanlar arasındaki iletişimi kolaylaştırmak ve bilgisayar ile dil çevirisi yapmak şeklinde sıralanabilir (Oflazer ve Bozşahin 2006).

Önceleri yapay zekânın küçük bir alt birimi olarak görülen bu konu, yapılan çalışmalarda elde edilen başarılar ve günden güne yeniden belirlenen ve şekillenen ihtiyaçlarla başlı başına bir araştırma alanı haline gelmiştir. Bilişimden tıbba, hukuktan güvenliğe, edebiyattan basına,

yazının ve dilin kullanıldığı her yerde hayati ihtiyaçları karşılamaya devam etmekte ve her geçen gün yeni alanlara girmektedir.

Doğal dil işleme sistemlerinde şu anda yaşanan en büyük problem dile özel geliştirme yapılmasının gerekliliğidir. Her ne kadar dil yapılarından bağımsız bir dil işleme kuramı araştırmacıların ortak amacı olsa da, genellikle araştırmalar dil yapılarına özel olarak devam etmektedir.

Mevcut doğal dil işleme sistemlerinin birçoğu İngilizce gibi yaygın diller için geliştirilmiştir. Bu dillerin yapısı Türkçeden farklı olduğu için bu sistemler Türkçeye uygulandığında hatalı ve geçersiz sonuçlar elde edilmektedir. Türkçe, Ural-Altay dil ailesinin, Türk dilleri kolunun Oğuz öbeğindendir (Ergin, 1998).

Diller kelime yapısına göre de belli gruplara ayrılmaktadır.

Tek Heceli Diller: Kelimeler tek hecelidir. Yapım ve çekim ekleri yoktur. Kelimeler cümledeki kullanım yerlerine göre anlam kazanırlar. Çince bu tür bir dildir.

Çekimli (Bükümlü) Diller: Kelimeler kullanımda değişiklere uğrar. Ön-ek, iç-ek, son-ek kavramları vardır. Bazılarında ünsüzler değişmez, ünlüler değiştirilerek yeni kelimeler yapılır. Yani kökler ünsüzlerden ibarettir. Arapça, Farsça, İngilizce ve Hintçe bu grupta yer alır.

Eklemeli (Agglutinant) Diller: Kelimelerin kökleri değişmez. Kullanımda kelimeye getirilen ekler, kelimelerin anlamlarını ve görevlerini belirler. Türkçe, Fince ve Macarca bu grupta yer alır.

Sözcüklerin kök veya gövdelerine gelerek cümledeki görevlerini belirleyen, onlara değişik anlamlar katan ya da onlardan yeni sözcükler türeten ses veya ses bileşimlerine ek denir. Eklemeli dillerde kelimelerin sonlarına bu ekler gelir, ancak kökler hiçbir zaman değişmez ya da kökün önüne ek gelmez. Ancak aynı durum, ekler için geçerli değildir. Kelimelerin sonuna gelen ekler, sesli uyumu, sesli değişmesi, harf düşmesi gibi nedenlerle değişikliğe uğrayabilir. Ayrıca bu ekler, eklendikleri kök ya da gövdenin cümle içindeki görevini, tek başına anlamını, sözcüğün türünü değiştirebilirler. Türkçe bu özelliği ile diğer dillerden önemli ölçüde ayrılır. Diğer dillerde bir cümle ile anlatılabilecek bir durum, Türkçede eklerin sayesinde tek kelimeyle anlatılabilir. Örneğin İngilizcedeki “I can not stop” cümlesi yerine Türkçede “duramıyorum” demek yeterlidir.

Türkçe, biçimsel özellikleri bakımından sondan eklemeli bir dildir ve Türkçedeki kelime

yapısının en önemli bileşenlerinden birisi kök ve gövdelerin sonuna gelen eklerdir. Bu ekler, yeni kelimeler türettiği gibi, hal, durum, aitlik ve zaman belirtmeye de yararlar. Bunlarla beraber, eklenen ekler, kelimeyi anlam değişikliğine ve dolayısıyla cümle içindeki görev değişikliğine uğrattığı zaman, kelimenin söz dizimindeki yeri de değişir. Tüm bu karışık faktörler bilgisayar destekli dil işlemede belli sorunlara neden olmaktadır.

Örneğin "kalemleri" kelimesinin çözümlenmesi:

Örnek Kullanım	**Çözümleme**
Bunlar Ali'nin kalemleri.	kök + çoğul + 3.tekil-iyelik
Ali'nin kalemlerini geri vereceğim.	kök + 3.çoğul-iyelik
Ali kalemleri geri verdi.	kök + çoğul + durum

Kelime çözümlemedeki bu tür belirsizlikler, üstteki örnekten de görüleceği üzere, bazı durumlarda cümle düzeyinde çözümleme yapmayı gerektirebilir. Ancak cümle düzeyindeki çözümlemenin bile yanılabileceği durumlar olabilmektedir. Örneğin, "Kalemleri kaybolmuş" cümlesinde kalemlerin bir kişiye mi yoksa bir gruba mı ait olduğu belirsizdir.

Çözümlemedeki belirsizlikler, sözcük türlerinin tespit edilmesinde de problemler yaratabilir.

Örneğin "gül" kelimesinin çözümlenmesi:

Örnek Kullanım	**Çözümleme**
Gül lütfen.	Fiil
Bana bir demet gül verdi.	İsim

Bu gibi belirsizliklerde eğer çekim ekleri kullanıldıysa kelimenin türü daha kolay belirlenebilir. "Güllerden" kelimesinde, çekim eki gelmiştir ve burada kökün isim olduğu kolayca görülmektedir.

Tüm bunların yanında Türkçede, sözlüksel belirsizlikler, birbirini tamamlayarak anlam elde eden söz öbekleri, birden fazla kelimenin bir araya gelmesiyle oluşan kelimelerin anlamları, deyimler gibi çözümleme problemleri yaratacak başka özellikler de bulunmaktadır. Bu özelliklerin çözümü ileri araştırma konularıdır.

Sözcük Dizimi

Türkçenin bir diğer önemli özelliği de, bazı durumlarda, sözcüklerin yerlerinin değişmesine rağmen cümlenin anlamının değişmemesi ya da bir anlatım bozukluğu oluşmamasıdır. Aşağıdaki bu duruma bir örnek verilmiştir. Bu örnekte fiilden hemen önce gelen kelimenin önemi vurgulanmaktadır.

Ben Ali'ye kalemlerini verdim. (verilen nesneler önemli)

Kalemlerini Ali'ye ben verdim. (nesnelerin kimin tarafından verildiği önemli)

Ben kalemlerini Ali'ye verdim. (nesnelerin kime verildiği önemli)

Görüldüğü gibi Türkçe sözdizimsel olarak oldukça esnek bir yapıya sahiptir. Diğer yandan İngilizce sözdizimsel olarak oldukça katı bir dildir ve bu unsurlar sözcüğün cümledeki yeri ile belirlenmektedir. Bundan dolayı İngilizce için geliştirilmiş bir çözümleme sisteminden Türkçe için başarılı sonuçlar alınamamaktadır.

2.6 Metin Madenciliği Süreçleri

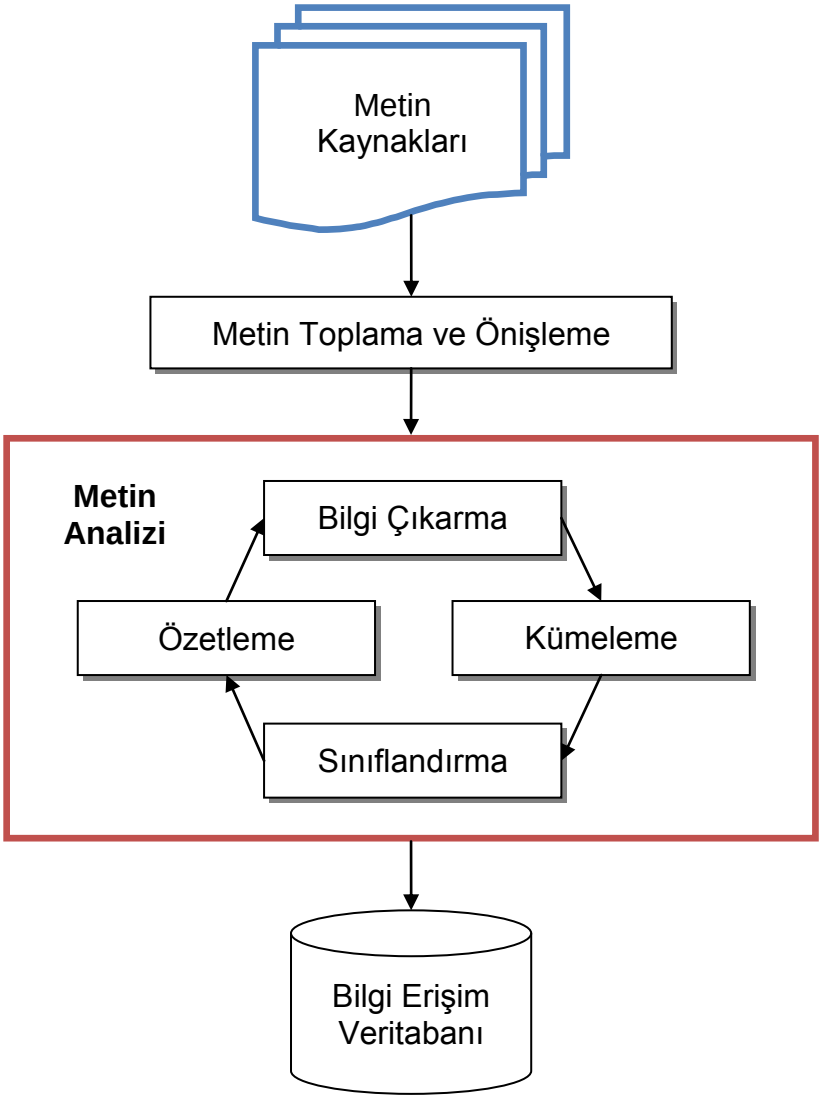
Metin Madenciliği süreçleri ana hatlarıyla üç adımdan oluşmaktadır. İlk safhada ilgili kaynaklardan doküman toplanarak salt metin elde etmek üzere önişlemler yapılır. Sonraki aşamada ise bir önceki adımda elde edilen metin kümesi, belirlenmiş hedeflere yönelik olarak analiz edilerek yorumlanmış bilgiler çıkarılır. Son adımda ise bu bilgilere erişim sağlanır. Bu süreçler Şekil 2.4'te detayları ile gösterilmiştir.

Özellikle son dönemde yapılan Metin Madenciliği çalışmalarında kaynak olarak internet üzerinden toplanan içerik kullanılmaktadır. Sürecin ilk adımını oluşturan bu işlem veri kümesinin çok büyük olmasından dolayı sistemin performansı açısından büyük önem taşımaktadır. Ayrıca ilgili konudaki metinlerin tespit edilmesi de bir başka önemli unsurdur. Metinlerin hangi konuda ve ne ile ilgili olduğunun belirlendiği bu işleme Metin Temsil Değerlemesi (Text Representation) denilmektedir. Bu işlemde **Kelime Torbası** (Bag of Words) ve **Vektör Uzayı** (Vector Space) olarak iki farklı model uygulanmaktadır.

Kelime Torbası modelinde, cümleleri oluşturan kelimelerin dilbilgisi ve hatta kelime sırası göz ardı edilerek, toplanan dokümanlardaki tüm kelimelerin kullanım sıklıkları bulunur ve ağırlıkları hesaplanır. Bu hesaplama işleminde kelimeleri içeren metinlerin hangi konuda oldukları da dikkate alınmaktadır. Çünkü bazı kelimelerin belli konularda yazılmış metinlerde

yer alma olasılığı başka kelimelere göre çok daha yüksektir. Örneğin bilgisayar kelimesinin bilişim konusunda yazılmış metinlerde bulunma sıklığı, sağlık konusunda yazılmış metinlerde bulunma sıklığına göre daha fazladır.

Vektör Uzayı modelinde ise kelimelerin ilgili metinleri temsil etme yüzdeleri hesaplanmaktadır. Ayrıca kosinüs benzerliği gibi teknikler uygulanarak dokümanların birbirleri ile olan benzerlikleri de dikkate alınmaktadır. Yapılan bu analizler sonucunda elde edilen veriler kullanılarak arama sonuç listeleri de derecelendirerek gösterilir.



Şekil 2.4 Metin madenciliği süreçleri

Metinler çok farklı kaynaklardan toplandığı için bu ham veriler üzerinde ciddi bir önişlem safhası uygulanmalıdır. Bu işlemlere geçmeden önce ilk olarak metinlerin ait oldukları dilin kapsamlı bir sözlüğü kullanıma hazır hale getirilir. Seçilecek sözlüğün kalitesi metin

madenciliği süreçlerinde elde edilen başarıyı doğrudan etkilemektedir. Bundan dolayı sözlük seçimine gereken önem verilmelidir.

Metin önişleme adımında, kaynaklardan toplanan dokümanlardaki metin dışı işaret ve etiketlerin temizliği gerçekleştirilir. Özellikle HTML formatındaki dokümanlardan asıl yazı kısmının çıkarılması açısından bu işlem çok önem kazanmaktadır. Çünkü internet sayfaları diğer doküman dosyalarından farklı olarak reklamlar, menü sistemi ve kullanıcı yorumları gibi sayfanın asıl içeriği dışında metin bölümleri de içermektedir. Bu bölümler doğru bir şekilde tespit edilerek ayrıştırılmazsa yapılan metin analizlerinde ciddi yanlışlıklar 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.

İnternet sayfaları dışında yine elektronik ortamda yer alan Acrobat ve Microsoft Word gibi diğer içerik dosyalarından da salt metin çıkarma işlemi yapılması gerekmektedir. Bu ihtiyaca cevap verecek açık kaynak kodlu veya ticari yazılımlar mevcuttur.

Bir önceki adımda elde edilen kelime listesinde hece ve eklere ayrılma işlemi uygulanarak kelime kökü tayini yapılır. Bu işlemden doğru sonuçlar alınabilmesi için ilgili dile özgü bir biçimbilimsel çözümleyici kullanılmalıdır. Örneğin İngilizcede oldukça tatmin edici sonuçlar veren Porter Stemmer yöntemi [2] sondan ekli bir dil olan Türkçe için aynı başarıyı sağlayamamakta ve bazı özel durumları yakalayamamaktadır. Bu konuda Türkçe dili için yapılan çalışmalarda en yaygın olarak kullanılan uygulama Zemberek projesidir [3]. Bu tez çalışmasında geliştirilen uygulamada da Zemberek projesi kullanılmıştır. Kelimelerin kök tespitinin hemen ardından kelime türü tayin edilir. Sözcük Türü Saptama (Part of Speech Tagging) olarak isimlendirilen bu işlem eğitim ve etiketleme olarak iki fazdan oluşmaktadır. Yapılan analizin türüne göre dilin gramer yapısına bakılarak kelime gruplarının tespit edilmesi de gerekebilir.

Analiz ve sınıflandırılma işlemleri yapılmış metin kümeleri içinden ihtiyaç duyulan bilginin çekilmesi ve veritabanına kaydedilmesi işlemleri Bilgi Çıkarım safhasını oluşturur. Bu işlem sonucunda elde edilen bilgiler hedeflenen sonuçların detaylarına göre belli bir yapıda organize edilerek kullanıcıya sunulur.

Metin madenciliği çalışmalarında doğru, geçerli ve kaliteli sonuçlar elde edebilmek için tüm süreçler sağlıklı bir şekilde uygulanmalıdır.

2.7 Metin Madenciliğinin Kullanıldığı Alanlar ve Sektörler

Metin madenciliğinin getirdiği faydalar görüldükçe çok farklı sektörlerde bu tekniklerden faydalanılmaya başlanmıştır.

Metin madenciliğinin uygulama alanlarından bazıları;

Bilişim Sektörü: Bir metnin yazıldığı dilin tespiti, istenmeyen elektronik postaların ayıklanması, internet sitelerine erişimin kısıtlanması, arama motoru sorgu sonuçlarının iyileştirilmesi, vb.

Sağlık ve Biyomedikal Sektörü: Hastalara ait kayıtların ve raporların analiz edilerek hastalık teşhisi ve genetik eğilimlerin tespiti, sektördeki gelişmelerin ve yayınların derlenerek analiz edilmesi, vb.

Finans Sektörü: Borsa gibi anlık gelişmelerden etkilenen alanlar için güncel haberlerin takip edilmesi ve finansal yorumlar üretilmesi, vb.

Pazarlama Sektörü: Müşterilerden gelen yazılı mesajların analiz edilmesi, kullanıcıların internet ortamında yaptığı yorumların değerlendirilmesi, pazardaki rakiplerin analiz edilmesi, risk analizi, vb.

Sigorta Sektörü: Raporların analiz edilerek sahtekârlıkların tespit edilmesi, vb.

Güvenlik Sektörü: Suçların işlenmesini önlemeye yönelik olarak özellikle internet ortamındaki mesajların ve içeriğin analiz edilmesi, olaylar sırasında toplanan metinlerin suçlu tespitine yönelik olarak analiz edilmesi, şifreli metinlerin çözümlenmesi vb.

Basın Yayın Sektörü: Haber metinlerinin analiz edilmesi, sınıflandırılması ve özetlenmesi, vb.

Eğitim Sektörü: Bir metnin yazarının bulunması, akademik çalışmalarda yapılan alıntıların tespit edilmesi, vb.

Gün geçtikçe katlanarak artan veri miktarının getirdiği sıkıntılar bir yana, sadece kullanım alanlarının genişliği bile metin madenciliği alanının neden bu denli revaçta olduğunu açıkça göstermektedir.

2.8 Metin Madenciliği ve Doküman Etiketleme

Etiket (tag) tek bir sözcük veya sözcük grubundan oluşan metin parçasıdır. Etiketler bilişim

sistemleri terminolojisinde bilginin içeriğini işaret eden ve belli bir sıralı düzen içinde olmayan anahtar kelimeler ve terimler olarak tanımlanır. Bu bağlamda etiketler, nitelendirme, sahiplik işareti, doküman içeriğinin çerçevesini çizme ve kimliğini belirleme görevini yerine getirerek, bilginin aranması ve bulunmasını kolaylaştırırlar.

Etiketleme çok eski bir bilgi düzenleme ve bilgiye erişim aracıdır. Etiketlemenin en klasik ve bilinen kullanımı, kütüphane ve arşivlerde karşımıza çıkmaktadır. Günümüzde internet kullanımının yaygınlaşması ve bilgi miktarındaki büyük artış etiketleme gibi bilgiye erişme araçlarının önemini daha da ön plana çıkartmıştır. Etiketlerin, ihtiyaç duyulan bilgiye hızlı ve etkin ulaşma konusunda oynadıkları rol her geçen artmakta, bu küçük kelime öbeklerinin sağladığı zaman ve emek tasarrufu ciddi boyutlara ulaşmaktadır.

Etiketler, bilgi ve doküman setlerindeki bilgiyi organize etmek ve kolay ulaşılmasını sağlamak amacıyla kullanılır. Çok farklı görsel şekillere sahip bu etiketler genellikle, yan yana, aralarında ayırıcı işaret olmadan, bazen de altları çizilerek sunulurlar. Son dönemlerde etiketlerin sosyal içerikli kullanımları da ortaya çıkmıştır. Örneğin içerik paylaşım sitelerindeki kullanıcılar yaptıkları etiket atamaları sonucunda hedefledikleri kitleler ile bir araya gelebilmektedir. Etiketlerin uygulamada internet sayfalarına, ağ güncelerine (blog) ve görsel malzemelere eklendiği görülmektedir (Hearst ve Rosner, 2008).

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 (Ames ve Naaman, 2007).

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.

3. ETİKET BULUTLARI

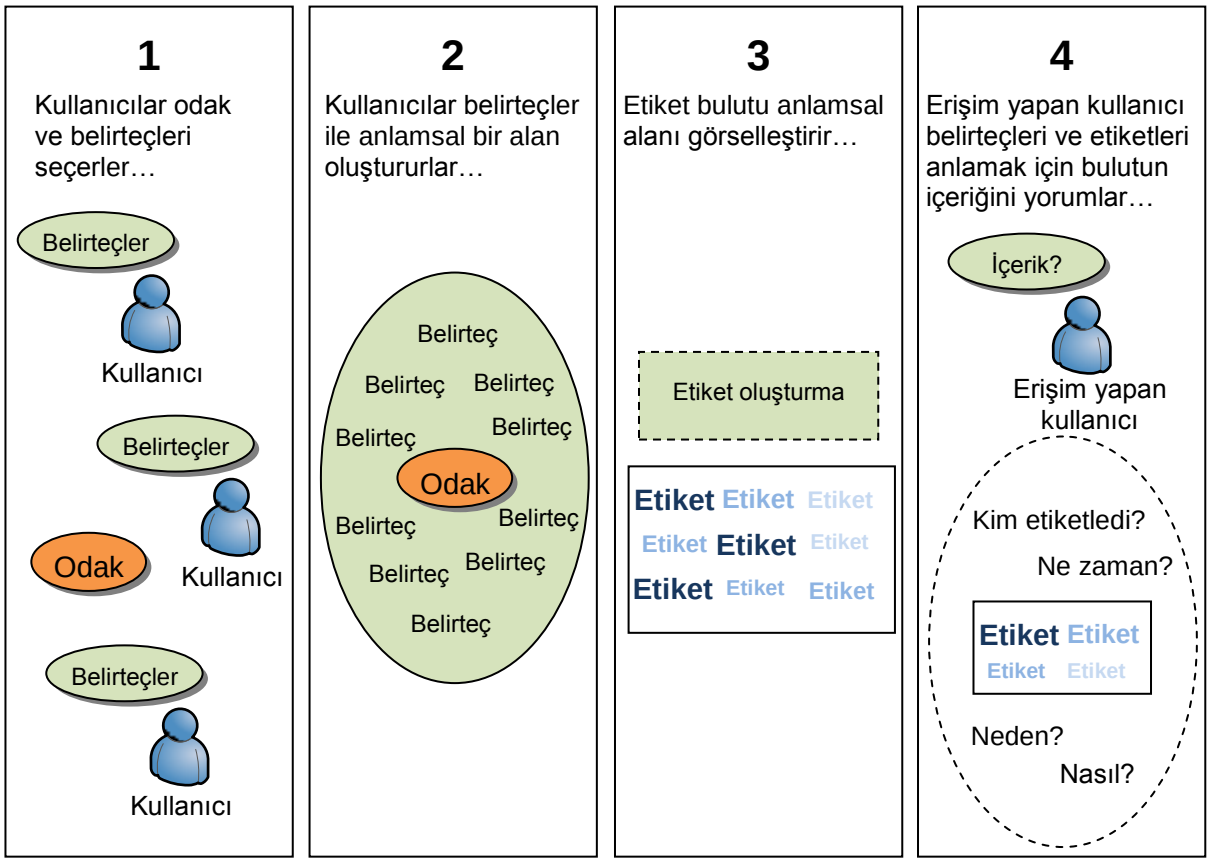
Etiket bulutları, sık kullanılan kelimelerin diğerlerine göre daha büyük bir yazıyüzü boyutunda ve daha belirgin bir renkte sıralandığı bir ağırlıklı liste (weighted list) gösterimidir. Etiket bulutlarının yapısı en temel haliyle Şekil 3.1'de görülmektedir. İlk olarak 2002 yılında Jim Flanagan tarafından bir sunum biçimi olarak kullanılmıştır (Zhou ve Bénel, 2008). 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.

Şekil 3.1 Etiket bulutu yapısı

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. Kullanıcılar tarafından yapılan etiketleme işlemleri sonucunda oluşan etiket bulutlarının

kullanımdaki algılama zinciri Şekil 3.2'de gösterilmiştir.



Şekil 3.2 Etiket bulutları algılama zinciri (Lamantia, 2006)

Bu sürecin ilk adımında bir içeriği etiketleyecek kullanıcılar belirledikleri odağı temsil edecek belirteçleri oluştururlar. Bu işlem sonucunda, aynı içerik, kullanıcıların algılarındaki farklılıktan dolayı değişik belirteçlerin bir araya geldiği bir anlamsal alan ile çevrelenmiş olur. Bazı matematiksel hesaplamalar da yapılarak ortaya çıkan belirteç listesinden etiket bulutu oluşturulur. Son aşamada ise etiketlenen içeriklere erişim yapacak kullanıcılar etiket bulutu görselini inceleyerek ilgilerini çeken etiketler aracılığı ile bilgiye erişim işlemini gerçekleştirirler. Erişim işleminin etkin bir şekilde yapılabilmesi için, oluşturulan etiket bulutundaki belirteçlerin tutarlı bir şekilde seçilmesi ve anlaşılır bir yapıda sunulması büyük önem taşımaktadır. Etiket bulutları ilk kullanıldıkları günden itibaren birçok değişim geçirmiş ve farklı çeşitleri ortaya çıkmıştır.

3.1 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 (Friedman, 2007).

Etiket bulutları içerdikleri verilere göre de gruplandırılmaktadır [4];

Metin Bulutları (Text Clouds)

Etiket bulutlarının en yaygın kullanıldığı şekil bu gruba girmektedir. Metin bulutları, kelimelerin frekanslarına göre sıralandığı bir ağırlıklı liste gösterimidir (Lamantia, 2007). Tek başına bir kitabın metin bulutu çıkarılabileceği gibi, belli bir ortak noktaya sahip birden fazla yazının da ortak etiket bulutu oluşturulabilir.

Veri Bulutları (Data Clouds)

Veri bulutlarında, kullanılan yazıyüzü boyutu ve renk tonları sayısal değerlere karşılık gelir. Veri bulutlarının diğer etiket bulutlarından ayrıldığı en belirgin özellik, kelimelerin frekansları yerine nüfus sayıları ve borsa verileri gibi sayısal bilgilerin kullanılmasıdır.

Eşdizimli Bulutlar (Collocate Clouds)

Eşdizimli bulutlar, bir doküman veya derlemin daha odaklanmış bir gösterimini sunan özel metin bulutlarıdır. Tüm dokümanın özetini göstermek yerine belli kelimelerin kullanımı dikkate alınarak bir gösterim oluşturulur. Bulutu oluşturabilmek için öncelikle bir kelime belirlenir ve seçilen bu kelimeyle ilişkili olan diğer kelimeler bulunarak etiket bulutu oluşturulur. Kelimeler arasındaki ilişkilerin sıklığı, yazıyüzü boyutu, ilişkinin yakınlığı da renk tonu ile gösterilir. Bu gösterim şekli bir dili oluşturan kelimeleri ve aralarındaki ilişkileri

etkileşimli bir şekilde incelemek için kullanılabilir.

Görsel öğeler dikkate alınarak alttaki şekilde bir gruplama yapılabilir;

Dizin Şeklindeki Etiket Bulutları

Kullanıcıların, ön plana çıkarılmış etiketler ile yönlendirilen içeriklere erişmek yerine kendi ilgilendikleri belli terimlere ait içeriklere erişmek isteyebileceği durumlarda kullanılması uygun olan bu gösterimde tüm etiketler herhangi bir özel vurgulama yapılmadan alfabetik olarak sıralanır. Bu kullanıma örnek olan bazı sayfalar Şekil 3.3'te görülmektedir.

Şekil 3.3 Dizin şeklindeki etiket bulutu örnekleri [5]

Yazıyüzü Boyutu İle Ağırlıklandırılmış Etiket Bulutları

Bu tip kullanımda küçükten büyüğe doğru farklı boyutlarda etiketler yer almaktadır. Etiket bulutları genellikle internet sayfalarının sağ veya sol tarafındaki kenar çubuklarının içinde yer alırlar. Bu alanda çoğu zaman kısıtlı yer olduğu için büyük boyutlu yazı tiplerinin dikkatli ve ekonomik kullanılmasında fayda vardır. Bu kullanıma ait bazı örnekler Şekil 3.4'te verilmiştir.

Şekil 3.4 Yazıyüzü boyutuna göre ağırlıklandırılmış etiket bulutu örnekleri [6, 7, 8, 9]

Renk İle Ağırlıklandırılmış Etiket Bulutları

Bazı tasarımcılar, yüksek ağırlıklı etiketleri çok büyük yazıyüzü boyutuyla göstermek yerine farklı renkler kullanmayı tercih etmişlerdir. Bu durumda etiketlerin ağırlıkları yazıyüzü boyutları ve renklerin birleşimleri ile vurgulanmaktadır. Bu birleşime arka plan renklerinin de dâhil olduğu örneklere rastlanmaktadır. Bu durumda en yüksek ağırlıklı etiketlerin ön ve arka plan renkleri çok zıt, düşük ağırlıklı etiketlerin ise birbirine yakındır.

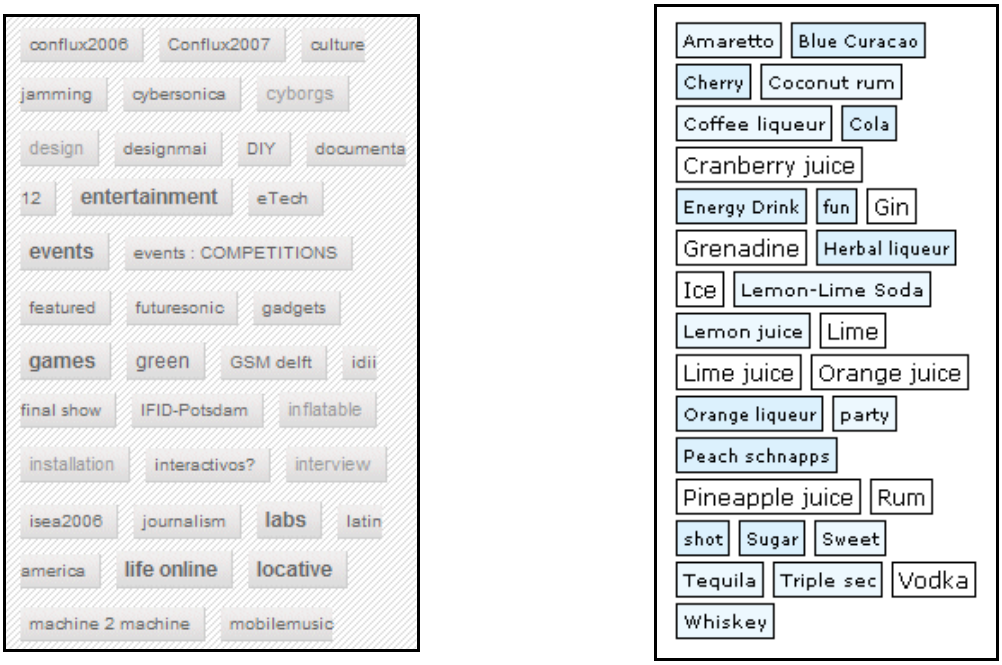
Ancak bu tür bir uygulamada, renklerin az sayıda ve dengeli bir şekilde kullanılmasına dikkat edilmelidir. Renk ile ağırlıklandırılmış etiket bulutlarını kullanan kişiler hangi rengin ne anlama geldiğini bilemeyeceği için çok fazla rengin kullanıldığı durumlarda, kullanıcın kafasının karışması muhtemeldir. Bundan dolayı bu şekilde renklendirilen etiket bulutları birçok tasarımcı ve uygulayıcı tarafından tercih edilmemektedir. Bu kategoriye giren bazı örnek tasarımlar Şekil 3.5'te yer almaktadır. Buradaki ilk örnekte yeşil ile en koyu kahve arasında ortaya çıkan kafa karıştırıcı durum açıkça görülmektedir. Diğer örneklerde ise benzer

renk tonları kullanılarak çok daha kolay algılanabilen etiket bulutları ortaya çıkarılmıştır. Şekil 3.5'teki dördüncü örnekte, etiket bulutunun tüm arka planı zıt bir şekilde renklendirilmiş ve etiketler en yüksekten en düşük ağırlığa doğru sıralanmıştır.

Şekil 3.5 Renk ile ağırlıklandırılmış etiket bulutu örnekleri [5]

Etiket Biçimleri

Etiket bulutlarında, görsel unsur olarak genellikle yazıyüzü boyutu ve renkler kullanılmıştır. Bazı tasarımcılar ise bunlara ek olarak, etiketlerin arka planı için farklı bir renk veya desen kullanarak özel bir vurgulama yapmışlardır (Friedman, 2007). Şekil 3.6'da bu kullanıma örnek tasarımlar görülmektedir.



Şekil 3.6 Etiket biçimi örnekleri [5]

3.2 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 [4].

$t_i > t_{min}$ ise $$s_i = \left\lceil \frac{f_{\max} \cdot (t_i - t_{\min})}{t_{\max} - t_{\min}} \right\rceil$$ diğer durumlarda $s_i = 1$ (3.1)

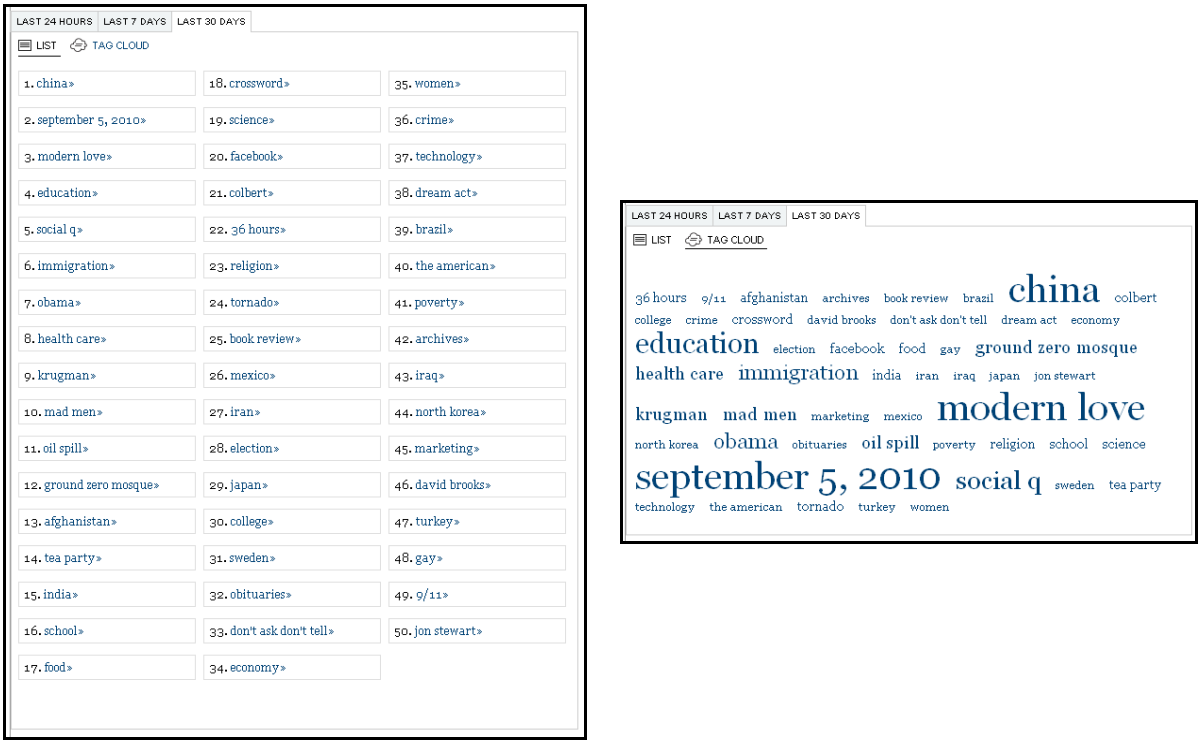
- s_i: yazıyüzü boyutu
- f_{max}: en büyük yazıyüzü boyutu
- t_i: ağırlık
- t_{min}: en küçük ağırlık
- t_{max}: en büyük ağırlık

3.3 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 3.7'de bu iki gösterim şeklinin örnekleri görülmektedir.



Şekil 3.7 Liste gösterimi ile etiket bulutu karşılaştırması [10]

Şekil 3.7'de 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 (2008) 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 (Hearst, 2008).

Bulgularıyla, etiket bulutlarının sosyal iletişimde oynadıkları rolü ön plana çıkaran sadece Hearst değildir. Bir başka araştırmacı (Donath, 2002) etiket bulutlarının sosyal yönünü vurgulamış ve özellikle çevrimiçi iletişimin dinamiklerinin anlaşılmasında sosyal bilimciler için iyi bir yaklaşım olduğunu ifade etmiştir. Ayrıca etiketleme işlemi sırasında kullanıcıların konuşma hissine yakın bir duygu yaşadıklarını ifade etmiştir. Yine başka bir yayında (Hearst ve Rosner, 2008), 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. Olumlu başka bir nokta ise, etiket bulutlarının birbirini destekleyen aşağıdaki üç farklı boyutu aynı anda gösterebilmesidir.

- Kelimelerin kendileri
- Kelimelerin ilişkisel önemleri
- Kelimelerin alfabetik sıralanmaları

Hearst ve Rosner'in (2008) yaptığı çalışmada, internet tasarımcıları ve bilgi görselleştirme uzmanlarından oluşan 20 kişilik bir grup ile yüz yüze görüşme yapılmıştır. Bu görüşmelerin sonucunda elde edilen niteleyici bulgular şu şekildedir;

- Etiket bulutlarının, site içeriğinin keşfedilmesine yardımcı olduğu düşünülmektedir.
- Etiket bulutlarının, dolaşma menüsü olarak kullanışlı olduğunu düşünen katılımcılar ile olmadığını düşünen katılımcılar eşit sayıdadır.
- Etiket bulutlarının eğlenceli, esprili ve davetkâr olduğu düşünülmektedir.

- Üç kişi, etiket bulutlarının, yaygın eğilimleri gösterdiğini söylemiştir.
- İki kişi, etiket bulutlarının, düz listelerden daha iyi olduğunu belirtmiş, bu iki kişiden birisi de bunun nedeni olarak etiket bulutlarının bütünsel açıdan daha kolay anlaşılır olmasını göstermiştir.
- Değişken ve dinamik bilgilerin gösterilmesinde faydalı olduğu düşünülmektedir.
- İki kişi, etiket bulutlarının, internet sitelerinin özetini görmek için iyi bir araç olduğunu söylemiştir.
- Bir katılımcı, etiket bulutlarının, toplumun ilgi alanlarını gösterdiğini düşünmektedir.
- Bir kişinin şahsi sitesine veya bir yazısına ait etiket bulutuna bakarak, o kişinin ilgi alanları ve yaptıkları hakkında belli bir fikir edinilebildiği söylenmiştir.

Bu konuda çalışmalar yapmış araştırmacılar ve bu tür sistemleri kullanan uzmanların düşünceleri çerçevesinde, etiket bulutlarının düz listelere göre hem olumlu hem de olumsuz yanlarının olduğu söylenebilir. Benzer zıt görüşler, etiket bulutlarının görsel özelliklerinin algılama üzerindeki etkileri konusunda da mevcuttur.

Örneğin, Lohmann ve arkadaşları (2009), yaptıkları çalışmada, etiket bulutlarının kullanıcılar tarafından nasıl algılandığını incelemiş ve aşağıdaki sonuçlara ulaşmışlardır;

Etiket boyutu: Büyük boyutlu etiketler küçük boyutlu olanlara göre, kullanıcıların daha çok dikkatini çekiyor. Bu etki, etiketin harf sayısı, konumu ve komşu etiketlerin özelliklerine göre değişebiliyor.

Tarama (Scanning): Kullanıcılar etiketleri okumak yerine gözleriyle bir tarama işlemi yapıyor.

Merkezlenme (Centering): Etiket bulutunun orta kısmındaki etiketler, kenardakilere göre daha çok dikkat çekiyor. Bu durum etiket bulutunun tasarımından etkilenebiliyor.

Konum: Etiket bulutunun sol üst bölgesi, diğer bölgelere göre daha çok dikkat çekiyor. İngilizce ve Türkçe gibi dillerde, yazının soldan sağa yazılıyor olmasının bu duruma neden olduğu düşünülmüştür.

Arama: Etiket bulutları, kullanıcıların özellikle belli bir etikete ulaşmak istedikleri ve bu etiketin boyutunun ufak olduğu durumlarda yetersiz kalabilirler.

Yapılan bu çalışmanın sonuçları şu şekilde özetlenebilir;

Etiket bulutu tasarımlarının detaylı bir şekilde incelendiği bu çalışmanın sonuçları, ağırlıklandırılmış terimlerin, bulut içine yerleştirilmelerinin sadece tek bir doğru yolu olmadığını ortaya koymuştur. Bu durum, kullanım hedefleri ve amaçlarına göre uygun çözümün farklılık gösterebildiği, etkileşim amaçlı yapılan tasarım çalışmaları ile de benzerlik göstermektedir. Bununla birlikte, yapılan deneysel çalışmalarda incelenen kullanım şekillerine göre en uygun olduğu düşünülen tasarımlar şu şekilde sıralanmıştır;

1. Belirli bir etiketin bulunması: Alfabetik sıralı, ardışık tasarım.
2. En önemli etiketlerin bulunması: Azalan öneme göre dairesel tasarım.
3. Belli bir konuya ait etiketlerin bulunması: Konusal olarak kümelenmiş tasarım.

Bunlara ek olarak, katılımcıların geri bildirim ve değerlendirmeleri, eğlenceli ve estetik özelliklerin, kullanıcıların etiket bulutları ile olan etkileşimini ciddi bir şekilde etkilediğini göstermektedir. Katılımcıların bir kısmının, teknik olarak en iyi verimi sağlamayan tasarımları tercih etmeleri de başka bir ilginç bulgu olmuştur.

Bu çalışmanın sonuçları yorumlanırken, çalışmada etiket bulutlarının sadece internet sitelerindeki kullanımının incelendiğine ve diğer kullanım alanlarının göz ardı edildiğine dikkat edilmesi gereklidir. İnternet sitelerindeki kullanım değerlendirilirken de bu sitelerin işlevi ve kurulum amacı dikkate alınmamıştır. Örneğin bir alışveriş sitesine yapılan hedefli erişimlerde, etiket bulutu iyi bir dolaşma aracı olabilir. Yeni, beğenilen ve çok aranan ürünler, etiket bulutunda belirgin bir şekilde yer aldığında, kullanıcıların ilgili ürünün sayfasına kolay bir şekilde erişmesi sağlanabilir. Bunun yanında, felsefe ve sosyoloji gibi konulardaki internet sitelerinde, belli terimlerin tekrarlarının sıklığından kaynaklanan kullanışsızlıklar ortaya çıkabilir. Bundan dolayı, etiket bulutlarının işlevini sağlıklı bir şekilde değerlendirebilmek için kullanılacakları ortamın dikkate alınması faydalı olacaktır.

3.4 Etiket Bulutlarının Kullanım Alanları

Etiket bulutları en bilinen ve yaygın şekliyle internet sitelerinde kullanılmaktadır. İnternet siteleri, içerikleri, yapıları, amaçları ve tasarımları doğrultusunda kendilerine has özellikler taşırlar. Etiket bulutlarının işlevlerinin değerlendirildiği bir önceki bölümde bahsi geçen çalışmada bu konu göz ardı edilmiştir. Aşağıda etiket bulutlarının, bazı internet sitesi türlerindeki kullanımları ile ilgili yapılan incelemelere yer verilmiştir.

Alışveriş Siteleri

Şekil 3.8'de görülen etiket bulutları bayanlara yönelik ürünler pazarlayan iki farklı siteden alınmıştır. Her iki sitede de cilt bakımı (skincare) terimi en ağırlıklı etiket olarak görülmektedir. Bu durum, cilt bakımını en yoğun ilgi gösterilen ürün grubunu olarak ortaya çıkarmaktadır. Siteye erişen yeni kullanıcıların büyük çoğunluğu da bu gruptaki ürünlere ulaşmak isteyeceklerdir. Bu grubu işaret eden etiket, diğerlerine göre daha büyük boyutlu olduğu için kullanıcılar diğer menülerde dolaşma ihtiyacı duymadan hızlı bir şekilde alt sayfalara erişim sağlayabilecektir. Bu örnekten görüldüğü üzere, etiket bulutları, alışveriş sitelerinde dolaşma aracı olarak kullanılmak için oldukça elverişlidir.

Şekil 3.8 Alışveriş sitelerinden örnek etiket bulutları [11, 12]

Bilgi Paylaşımı Siteleri

Genellikle bu tür siteler, belli bir marka veya ürün grubuna yönelik bilgi paylaşımı yapmak üzere, gönüllü kullanıcılar tarafından oluşturulmaktadır. Şekil 3.9'da buna örnek iki siteden alınan etiket bulutları görülmektedir. Bu internet sitelerinin satış yapmaya yönelik ticari bir amaçları bulunmamaktadır. Bundan dolayı siteyi oluşturan tasarımcılar, etiket bulutunu gelir arttırıcı bir araç olarak değil, siteyi kullanan kişilerin ihtiyaç duyabilecekleri bilgiye kolay

ulaşmasını sağlayacak bir araç olarak sunmuşlardır. Bu şekilde değerlendirildiğinde hedefe başarıyla ulaşıldığı görülmektedir. Bu örnekte dikkat çeken başka bir nokta, sitelerin iki farklı ülkeden kullanıcılar tarafından oluşturulmuş ve içeriğin tamamen birbirinden farklı kişiler tarafından yazılıyor olmasına rağmen oluşan etiket bulutlarının gösterdiği benzerliktir. İki tarafta da özellikle büyük boyuta sahip etiketlerin ağırlıklarının birbirlerine oldukça yakın olduğu görülmektedir.

Şekil 3.9 Bilgi paylaşımı sitelerinden örnek etiket bulutları [13, 14]

Ağ Günceleri

Ağ güncelerinin yayınlandığı siteler alışılageldik bir dolaşım menüsünü oluşturacak yapıya sahip olmadığı için, etiket bulutları bu tür sitelerde ana yönlendirme aracı olarak görev yaparlar. Buna ek olarak, etiket bulutları o güne kadar yayınlanmış tüm yazıların bir özetini de yansıtırlar. Ayrıca bir siteye ilk kez erişen bir kullanıcı, etiket bulutuna bakarak o sitede genel olarak ne tür yazılar olduğunu hızlı bir şekilde görebilir. Örneğin Şekil 3.10'da soldaki etiket bulutu incelendiğinde, bilişim ağırlıklı yazılar içeren bir siteye ait olduğu kolayca görülmektedir. Sağdaki örneğin ise tamamen Microsoft firmasının ürünlerine odaklanmış yazılar içeren bir siteye ait olduğu anlaşılmaktadır.

Ağ günceleri yayınlayan sitelerde yer alan etiket bulutlarını oluşturan etiketler, genellikle yazıları yazan veya okuyan kullanıcılar tarafından atanmaktadır. Bu şekildeki kullanımda etiket bulutunun daha kaliteli olabilmesi için, etiket ataması yapılırken sitede daha önceden

kullanılmış etiketler kontrol edilmeli ve girilmesi düşünülen etiket listede mevcut ise oradan seçilmelidir. Aksi takdirde aynı anlama gelen veya birbirine çok benzeyen etiketler kullanılacağından hem etiket kirliliği ortaya çıkacak hem de oluşan etiket bulutu çok tutarlı olmayacaktır.

Şekil 3.10 Ağ güncesi sitelerinden örnek etiket bulutları [15, 16]

Sosyal Bilimler

Sosyal bilimler alanında etiket bulutu oluşturan kişilerin ilgisini en çok, tarihte yapılmış önemli politik konuşmaların içerikleri çekmiştir. Şekil 3.11'de üç farklı Amerikan başkanının değişik tarihlerde yaptıkları ulusa sesleniş konuşmalarının etiket bulutları görülmektedir. Bu gösterim, gündemin tarihsel olarak nasıl değiştiğini çok açık bir şekilde ortaya koymaktadır.

Şekil 3.11 Amerikan başkanlarının ulusa sesleniş konuşmalarına ait etiket bulutları [17]

Özgeçmişler

İnsan kaynakları uzmanlarının, adaylarla iletişime geçtiklerinde inceledikleri ilk doküman başvuru sahibinin özgeçmişidir. Adayın tüm kariyerinin bir özet halinde sunulduğu bu dokümanı bir bakışta bütünüyle algılamak çok kolay olmamaktadır. Alternatif olarak bu dokümanda yazan bilgilerin Şekil 3.12'de görüldüğü gibi sunulması, o kişi hakkında hızlı bir şekilde fikir edinilmesini sağlamaktadır.

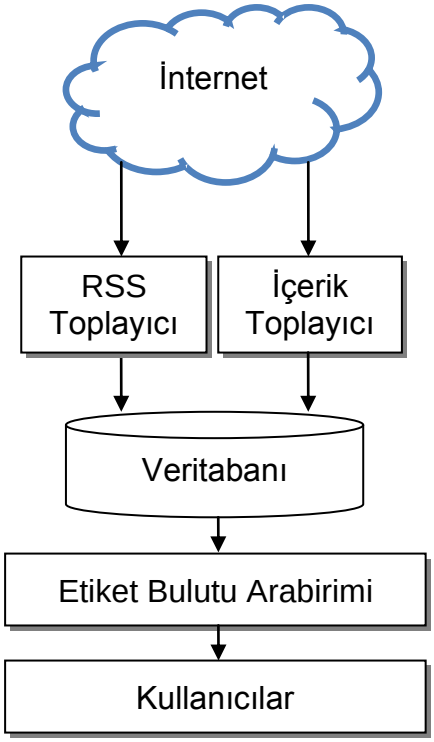
Şekil 3.12 Etiket bulutu haline getirilmiş bir özgeçmiş [18]

4. GÜNDEM BULUTU UYGULAMASININ TASARIM AŞAMALARI

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. 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.

4.1 Uygulamanın Yapısı

Yazılım geliştirme aşamasına geçmeden önce gereksinim analizi ve tasarım çalışmaları yapılarak Şekil 4.1'de görülen yapı oluşturulmuştur.



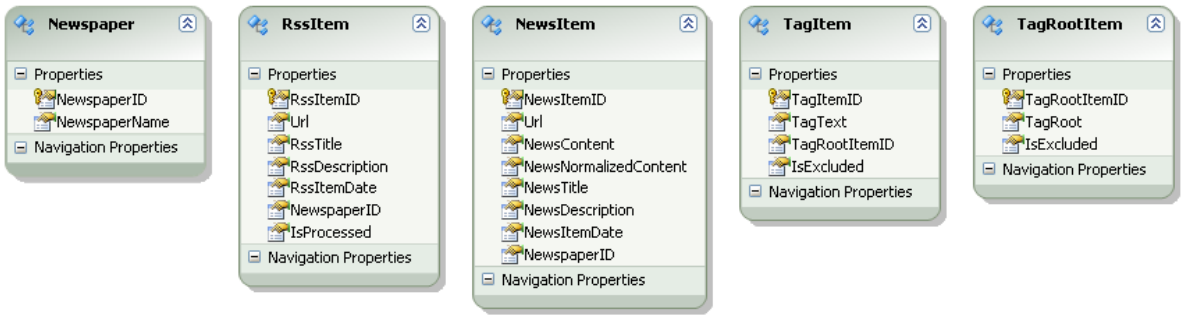
Şekil 4.1 Gündem Bulutu uygulamasının yapısı

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.

4.1.1 Veritabanı

Uygulamanın gereksinimlerine uygun olarak oluşturulan veritabanındaki ana tablolar Şekil 4.2'de görülmektedir.



Şekil 4.2 Veritabanındaki ana tablolar

Tablolara ait detaylar Çizelge 4.1'de listelenmiştir.

Çizelge 4.1 Veritabanı tablo açıklamaları

Tablo Adı	**Açıklama**
Newspaper	Haber kaynaklarının listesi
RssItem	Toplanan RSS kayıtları
NewsItem	RSS kayıtlarındaki erişim adreslerinden toplanan haber içerikleri
TagItem	Analiz sonucunda oluşan etiket kayıtları
TagRootItem	Etiketlerin kökleri

Veritabanındaki tablolar arasında bilgi tutarlılığının sağlanması için gerekli ilişkiler tanımlanmış ve performans açısından tablolar üzerinde indeks tanımları yapılmıştır.

Veritabanı sağlayıcı olarak Microsoft SQL Server uygulaması kullanılmaktadır. Saklanacak içerik Türkçe olduğu için, dil ayarı Turkish_CI_AS seçilerek veritabanı oluşturulmuştur.

4.1.2 RSS Toplayıcı

RSS, genellikle haber ve ağ günceleri gibi içerik sunan sitelerin yayınlarının izlenebilmesine olanak sağlayan bir içerik besleme yöntemidir. Kuralları belirlenmiş özel bir XML (eXtended Markup Language) dosya olarak sunulur.

Bu dosyanın bir örneği Şekil 4.3'te görülmektedir.

```
<?xml version="1.0" encoding="UTF-8" ?>
<rss version="2.0">

      <channel>

            <title>Site Başlığı</title>
            <link>http://www.siteadi.com</link>
            <description>Site Açıklaması</description>
            <lastBuildDate>Mon, 30 Aug 2010 10:22:50 +0200 </lastBuildDate>
            <pubDate>Tue, 31 Aug 2010 11:00:00 +0200</pubDate>
            <language>tr</language>

            <item>
                  <title>İçerik Başlığı</title>
                  <link>http://www.siteadi.com/icerik.html</link>
                  <description>İçerik Açıklaması</description>
                  <guid isPermaLink="false"> 1345261</guid>
                  <pubDate>Tue, 31 Aug 2010 11:00:00 +0200</pubDate>
            </item>

      </channel>
</rss>
```

Şekil 4.3 RSS XML dosya örneği

RSS kısaltmasının açılımı ve zaman içinde gelişimi şöyledir:

- Rich Site Summary (Zengin Site Özeti) (RSS 0.91)
- RDF Site Summary (RDF Site Özeti) (RSS 0.9 and 1.0)
- Really Simple Syndication (Gerçekten Basit Dağıtım) (RSS 2.0.0)

RSS listelerinin toplu hali RSS Beslemeleri veya RSS Kanalları olarak isimlendirilir. Yayınlanan RSS kayıtları, genellikle internet sitesinde sunulan içeriğin başlığı ve özetinden oluşur. İçeriğin kendisine ise RSS listesinde yer alan adres aracılığı ile erişilir.

RSS kaynağı sağlayan internet sitelerinde genellikle şu simgeler bulunur: RSS XML

Projede kullanılmak üzere seçilen RSS kaynakları Çizelge 4.2'de verilmiştir.

Çizelge 4.2 RSS kaynaklarının listesi

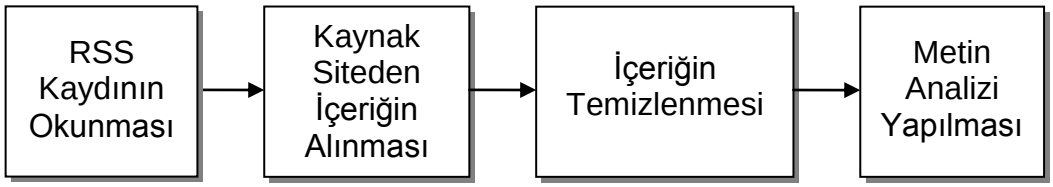
Kaynak Adı	**RSS Adresi**
Hürriyet	http://rss.hurriyet.com.tr/rss.aspx?sectionId=2
Milliyet	http://www.milliyet.com.tr/D/rss/rss/Rss_1.xml
Radikal	http://www.radikal.com.tr/d/rss/Rss_77.xml
Zaman	http://www.zaman.com.tr/anasayfa.rss
Posta	http://www.posta.com.tr/xml/rss/rss_1_0.xml
Star	http://www.stargazete.com/rss.xml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

Geliştirilen uygulamanın hedefi, Türkiye'nin gündemini yansıtmak olduğu için mümkün olduğunca farklı bakış açılarına sahip kaynaklar belirlenerek listenin zenginleşmesi sağlanmıştır. Buna ek olarak kaynakların teknik değerlendirmeleri yapılmış ve bilgilerin sağlıklı toplanıp toplanamadığı kontrol edilmiştir.

RSS yayınlarını toplayan bileşen, arka planda sürekli çalışan bir Windows servisi olarak C# dili ile .NET platformunda geliştirilmiştir. Uygulama, düzenli aralıklarla RSS kaynaklarını kontrol eder ve tespit ettiği yeni içerikleri, adres, başlık, açıklama, tarih ve haber kaynağı bilgileri ile veritabanına kaydeder.

4.1.3 İçerik Toplayıcı

RSS toplayıcı servis tarafından düzenli aralıklarla veritabanına eklenen kayıtları işleyerek, haber içeriklerini kaynak sitelerden alan ve analiz eden bu uygulama, önceki bileşen gibi C# dili ile .NET platformunda geliştirilmiş bir Windows servisidir.



Şekil 4.4 İçerik toplayıcı veri işleme şeması

Uygulamanın, içerik toplama ve işleme sürecine ait adımları Şekil 4.4'teki şemada görülmektedir.

4.1.3.1 RSS Kaydının Okunması

İlk adımda, işlem görmemiş RSS kayıtlarının listesi veritabanından alınarak bir döngü içinde işlenir. Haber içeriğinin yayınlandığı asıl adres RSS kaydının URL sahasında bulunur. Bu adres sürecin ikinci adımında kullanılır.

4.1.3.2 Kaynak Siteden İçeriğin Alınması

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

4.1.3.3 İçeriğin Temizlenmesi

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. Sayfa içeriklerinde belli bir standart olmadığı için her haber kaynağına özel bir temizleme işlemi yapılmaktadır. Bu aşamada yapılan bir diğer önemli işlem, sayfa içeriğinin metin dışı işaretler, etiketler ve HTML kodlarından arındırılmasıdır. Bu işlem için özel bir yardımcı uygulama sınıfı geliştirilmiştir. Bu sınıftaki ilgili metoda girdi olarak HTML içerik verilir ve çıktı olarak HTML etiketleri temizlenmiş asıl metin alınır.

4.1.3.4 Metin Analizi Yapılması

Bir önceki adımda elde edilen metinlerden etiket listesi oluşturulma işlemi bu aşamada gerçekleştirilir. Salt metin içinden etkisiz sözcüklerin ayıklanması ile metin analizine başlanır. Bu işlem için dinamik olarak yönetilebilen bir etkisiz sözcük dosyası kullanılmaktadır. Zaman içinde yeni bir etkisiz kelime belirlenirse, o kelimeyi bu dosyaya eklemek yeterli olacaktır.

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 [3] ve Snowball [19] dil işleme kütüphanelerinden birisini seçmek üzere çeşitli testler gerçekleştirilmiştir. Seçimi etkileyecek en önemli kıstas kelimelerin köklerinin doğru bir şekilde bulunmasıdır. Bununla ilgili olarak belirli bir kelime listesi alınarak her iki kütüphane ile kökler bulunmuştur. Bu testlerden elde edilen sonuçlar Çizelge 4.3'de verilmiştir.

Çizelge 4.3 Zemberek ve Snowball kütüphaneleri ile kök bulma test sonuçları

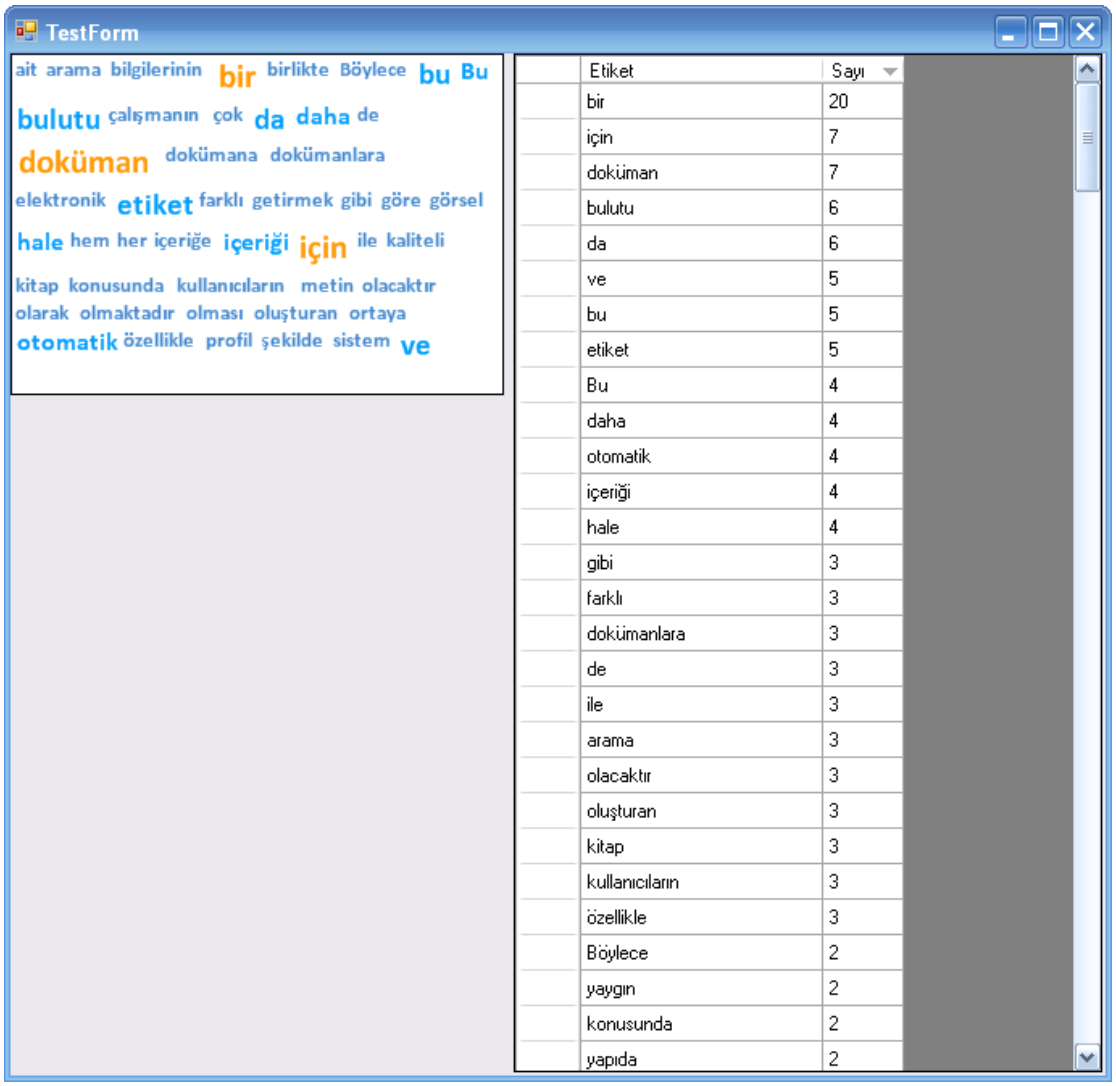
Kelime	**Zemberek**	**Snowball**	**Kelime**	**Zemberek**	**Snowball**
ağacı	ağaç	ağaç	kitabı	kitap	kitap
belirleyen	belirle	belirleye	kokla	kokla	kok
bölünmez	bölünmez	bölünmez	küçücük	küçük	küçücük
burnu	burun	burnu	oğlu	oğul	oğlu
bütünlüğünü	bütün	bütünlük	incelemesine	incele	inceleme
devletinin	devlet	devlet	rengi	renk	rengi
edilmesinde	et	edilme	yapılması	yapılması	yapılmas
gösterilmesi	gösterilmesi	gösterilmes	yükselmek	yükselmek	yükselmek
incelemesine	incele	inceleme	zamanının	zaman	zaman

Sonuçlar incelendiğinde Zemberek ile alınan sonuçların çok daha doğru olduğu görülmektedir. Örneğin *burnu, oğlu* ve *rengi* kelimeleri Zemberek ile *burun, oğul* ve *renk* olarak doğru bir şekilde tespit edilirken Snowball her üç kelime için de yanlış sonuç vermiştir.

Snowball kütüphanesinde yaşanan bir diğer problem ise *gösterilmesi* ve *yapılması* kelimelerinin kökleri olarak *gösterilmes* ve *yapılmas* şeklinde yanlış kelimeler oluşturmasıdır.

Zemberek kütüphanesinin bir diğer önemli üstünlüğü ise kelimelerin türlerini belirleme olanağını sunmasıdır. Haber metinlerinden etiket listesi oluştururken bağlaç, edat, zamir ve zarf türündeki kelimelerin elenebilmesi açısından bu fonksiyon oldukça önem taşımaktadır. Zemberek kütüphanesi paralel olarak hem .NET hem de JAVA üzerinde geliştirilmeye devam etmektedir. Ancak ilk olarak JAVA versiyonu geliştirilmeye başlamasından 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 projede bazı ek geliştirmeler yapılmış ve .NET versiyonundaki sözlükler güncel hale getirilmiştir.

Zemberek kütüphanesi kullanıma hazır hale getirildikten sonra buna ek olarak metin içindeki etkisiz kelimeleri temizleyen bir sınıf oluşturulmuştur. Bu iki bileşenin, etiket listesi oluşturmadaki etkilerini test edebilmek amacıyla, bir metni analiz ederek etiket bulutu oluşturan prototip bir uygulama geliştirilmiştir. Yapılan ilk test işleminde noktalama işaretleri ve boşluklar dikkate alınarak metindeki tüm kelimelerin listesi çıkarılmıştır. Bu aşamada kelimelerin türleri ve kökleri dikkate alınmamıştır. Bu şartlar altında Şekil 4.5'te görülen sonuçlar elde edilmiştir.

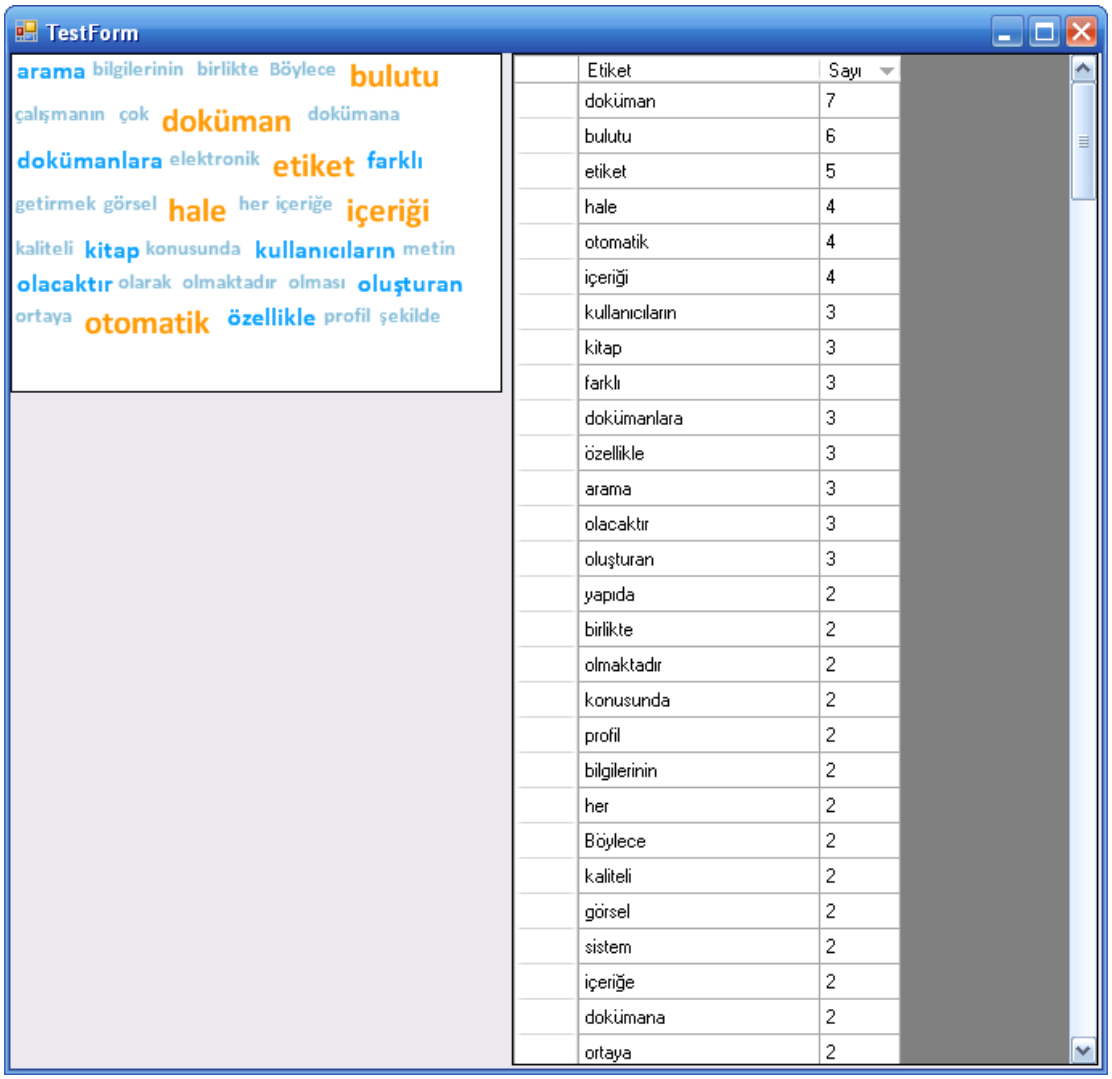


Etiket	Sayı
bir	20
için	7
doküman	7
bulutu	6
da	6
ve	5
bu	5
etiket	5
Bu	4
daha	4
otomatik	4
içeriği	4
hale	4
gibi	3
farklı	3
dokümanlara	3
de	3
ile	3
arama	3
olacaktır	3
oluşturan	3
kitap	3
kullanıcıların	3
özellikle	3
Böylece	2
yaygın	2
konusunda	2
yapıda	2

Şekil 4.5 Prototip uygulamada özel kurallar uygulanmadan oluşturulan etiket bulutu

Sonuçlar incelendiğinde, etkisiz kelimelerin ön sıralarda yer aldığı görülmektedir. Bu şekilde yapılan analizin çok sağlıklı sonuçlar doğurmadığı ortadadır.

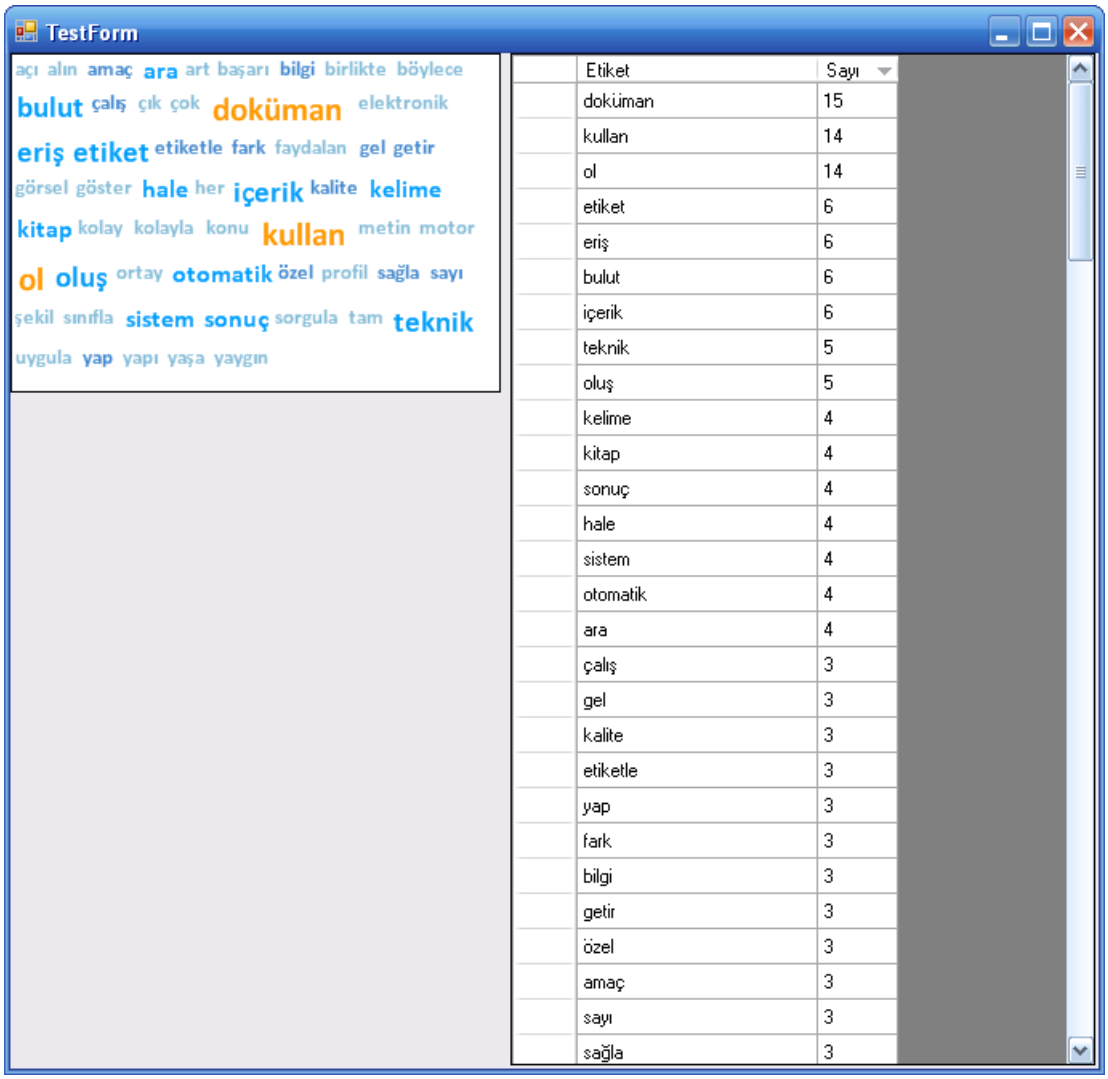
İkinci aşamada etkisiz kelime ayrıştırma işlemi sürece dâhil edilmiştir. Bu durumda alınan sonuçlar Şekil 4.6'daki gibi olmuştur.



Şekil 4.6 Prototip uygulamada etkisiz kelimeler ayrıştırılarak oluşturulan etiket bulutu

Listeye bakıldığında, ilk adımdakine göre çok daha iyi bir sonuç alındığı görülmektedir. Ancak metnin Türkçe diline uygun bir şekilde analiz edilmemesinden dolayı, bazı kelimelerin farklı çekimlerinin ayrı kayıtlar olarak listelenmesi başka bir sorun olarak karşımıza çıkmaktadır.

Üçüncü aşamada Zemberek kütüphanesinden faydalanılarak metindeki kelimeler köklerine göre gruplanmıştır. Örneğin "dokümanları" ve "dokümanı" kelimeleri, önceki örneklerde iki ayrı kelime olarak değerlendirilirken, kelimeler köklerine göre gruplandığında etiket listesinde her ikisi de "doküman" kelimesi altında toplanmıştır. Bu metin analizi işlemleri sonucunda Şekil 4.7'deki gibi bir etiket listesi elde edilmiştir.



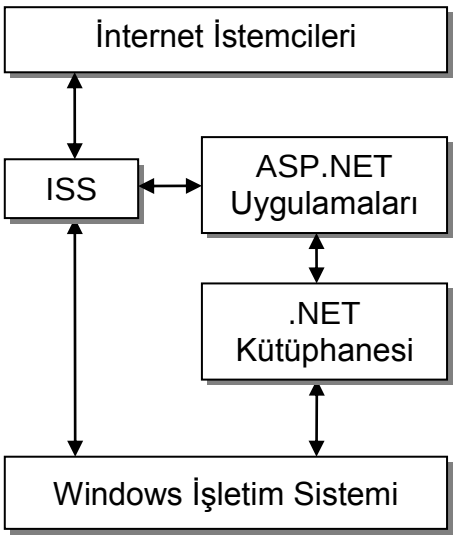
Şekil 4.7 Prototip uygulamada özel kurallar uygulanarak oluşturulan etiket bulutu

Yukarıda anlatılan metin analiz metotları bu projenin asıl veri kaynağı olan haber metinlerine uygulandığında etiket listesi olarak tutarlı sonuçlar elde edilmesini sağlamıştır. Ancak 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 kişilerin, yerlerin ve kurumların etiket bulutunda gösterilmesi sağlanacaktır. Kullanılması uygun görülen ikinci metot ise metinlerde sürekli tekrar eden ikili ve üçlü kelime gruplarının tespit edilmesi ve etiket olarak değerlendirmeye alınmasıdır. Bu işlemin ön koşulu ise bu kelime gruplarının bağlaç, edat, zamir ve zarf türündeki kelimeler ile etkisiz sözcükleri içermiyor

olmasıdır. 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.

4.1.4 Arabirim

Çalışmada kullanılacak verilerin sağlıklı bir şekilde elde edilmesini sağlayacak altyapı sağlandıktan sonra, çalışma sonuçlarının sunulacağı arabirim bileşenine geçilmiştir. Önceki adımlarda verilerin test edilebilmesi için bir prototip Windows uygulaması geliştirilmişti. Ancak uygulamadaki asıl hedef, gündemi yansıtan etiket bulutunun internet üzerinden erişime açılmasıydı. Siteye aynı anda birçok kullanıcı erişeceğinden, performans açısından doğru bir yapının kurgulanması, uygulamanın kullanılabilirliği açısından oldukça büyük önem taşımaktadır. Bu hedeflere uygun bir altyapı sağlayan ve mimarisi Şekil 4.8'de görülen ASP.NET ve IIS (Internet Information Services) internet sunucusunun projede kullanılmasına karar verilmiştir.

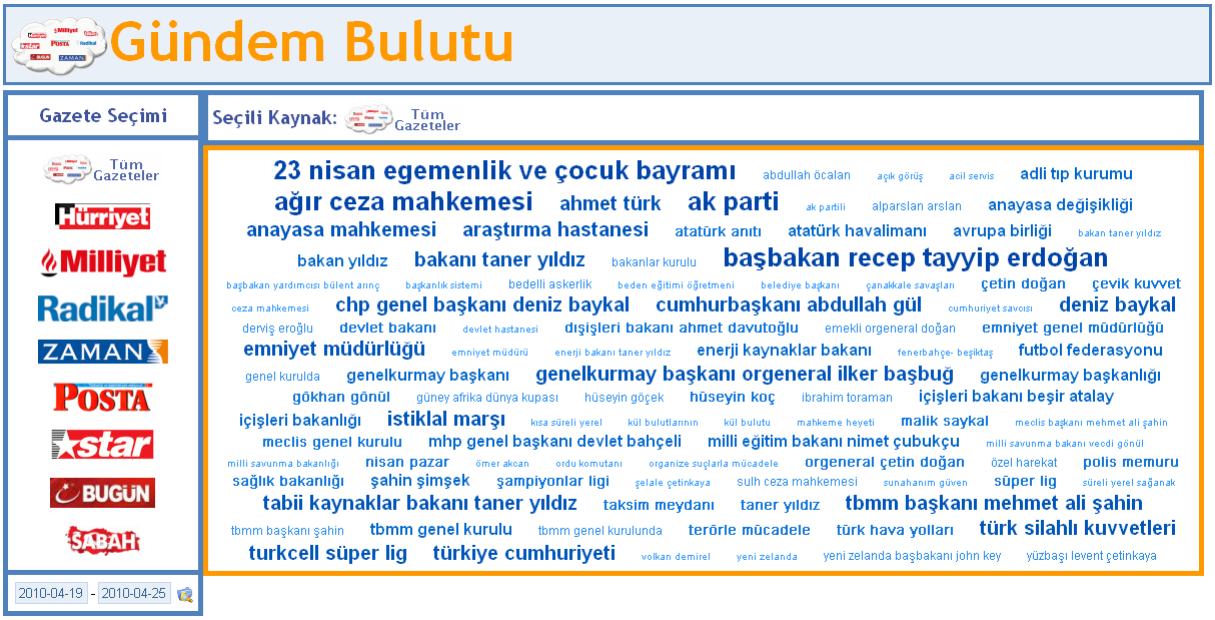


Şekil 4.8 ASP.NET Mimarisi

ASP.NET, internet uygulamaları oluşturmak ve dağıtmak için gereken hizmetleri sağlayan birleştirilmiş uygulama platformudur. Microsoft .NET Kütüphanesinin bir parçası olan ASP.NET, bellek yönetimi, iş parçacığı yönetimi ve kod güvenliği gibi çekirdek hizmetleri sağlayan bir ortam sağlamaktadır (Varallo, 2009).

5. GÜNDEM BULUTU UYGULAMASI

Gündem Bulutu uygulaması bir internet sitesi olarak tasarlanmıştır. Kullanıcılar sitenin yayın yaptığı adresi internet tarayıcısına girerek Şekil 5.1'de görülen uygulamanın ana sayfasına ulaşırlar. Uygulama açıldığında son bir haftaya ait haberlerin içeriğinden oluşan bir etiket bulutu görüntülenir.



Şekil 5.1 Gündem Bulutu ana sayfası

Kullanıcılar, bulut içindeki bir etikete Şekil 5.2'de görüldüğü şekilde tıklayarak o etiketin eşlendiği haberlere ait tarih, başlık ve özet bilgilerinin listelendiği Şekil 5.3'de görülen sayfaya ulaşırlar.

Şekil 5.2 Bir etiketin seçilmesi

Şekil 5.3'de görülen listeden istenen bir haberin linkine tıklayarak haberin yayınlandığı asıl siteye ulaşılıp, haberin tüm metni okunabilir.

Şekil 5.3 Seçilen etikete ait haberlerin listelenmesi

Kullanıcılar, Şekil 5.4'te görülen 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.

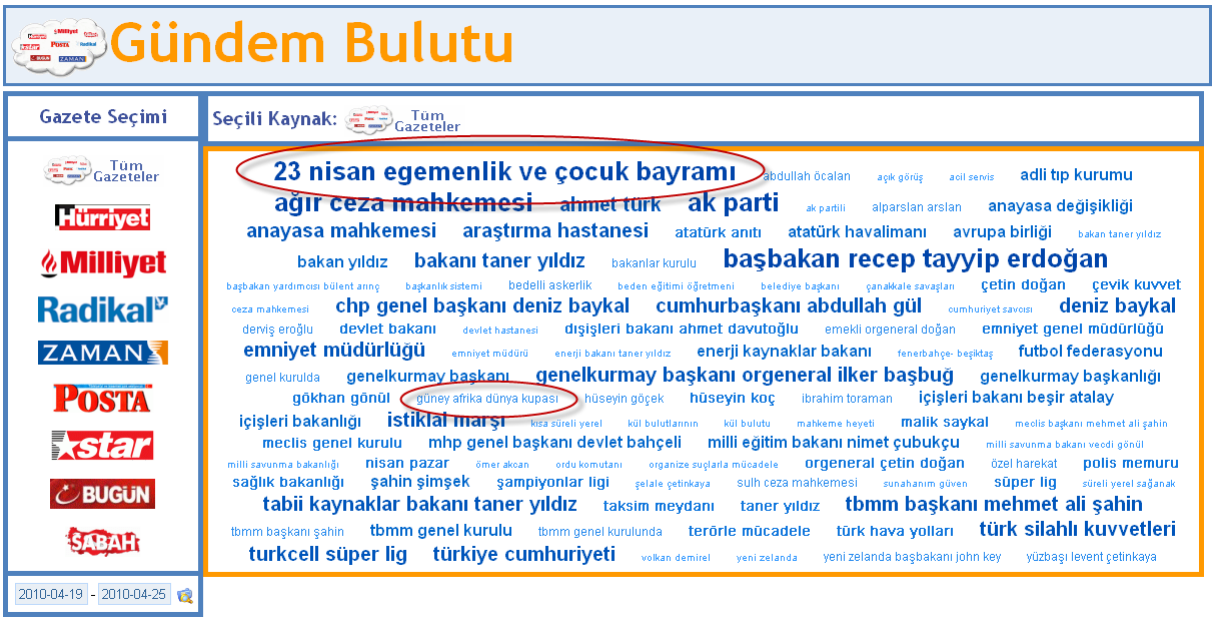
Şekil 5.4 Gazete seçimi

Benzer şekilde kullanıcıların tarih aralığı belirleme şansı da bulunmaktadır. Şekil 5.5'te görülen bölümden başlangıç ve bitiş tarihleri girilerek sadece o tarihler arasındaki haberlere ait etiketlerin değerlendirmeye alınması sağlanır.

Şekil 5.5 Tarih aralığı belirleme

Gündemdeki gelişmelerin, uygulamada oluşan etiket bulutlarına nasıl yansıdığı, aşağıdaki ekran örneklerinde rahatça görülmektedir.

Şekil 5.6'da görüldüğü üzere 23 Nisan haftasına ait haberlerde "23 Nisan Egemenlik ve Çocuk Bayramı" etiketi ön sıralarda gelmektedir. Buna karşın diğer terimlere göre daha küçük olan "Güney Afrika Dünya Kupası" etiketi yaklaşık 2 ay sonra gerçekleşecek dünya kupasına işaret etmektedir.



Şekil 5.6 Gündem Bulutu ekran örneği (19.04.2010 – 25.04.2010)

Şekil 5.7'de görülen etiket bulutunda, Türkiye'de Mayıs ayının 2. Pazar günü kutlanan anneler gününe ait bir etiket olduğu göze çarpmaktadır. Ayrıca "Güney Afrika Dünya Kupası" etiketi de bir önceki aya ait etiket bulutundaki durumuna göre daha belirgin bir hale gelmiştir.

Şekil 5.7 Gündem Bulutu ekran örneği (10.05.2010 – 16.05.2010)

Dünya kupasının gerçekleştiği tarih aralığı seçildiğinde ise "Güney Afrika Dünya Kupası" etiketinin bariz bir şekilde ön plana çıktığı Şekil 5.8'de görülmektedir.

Şekil 5.8 Gündem Bulutu ekran örneği (11.06.2010 – 11.07.2010)

Her yıl Haziran ayının üçüncü Pazar günü kutlanan babalar günü 2010 yılında 20 Haziran'a denk gelmiştir. Şekil 5.9'da görüleceği üzere sadece o güne ait oluşturulan etiket bulutunda ilgili etiket kendini göstermektedir.

Şekil 5.9 Gündem Bulutu ekran örneği (20.06.2010)

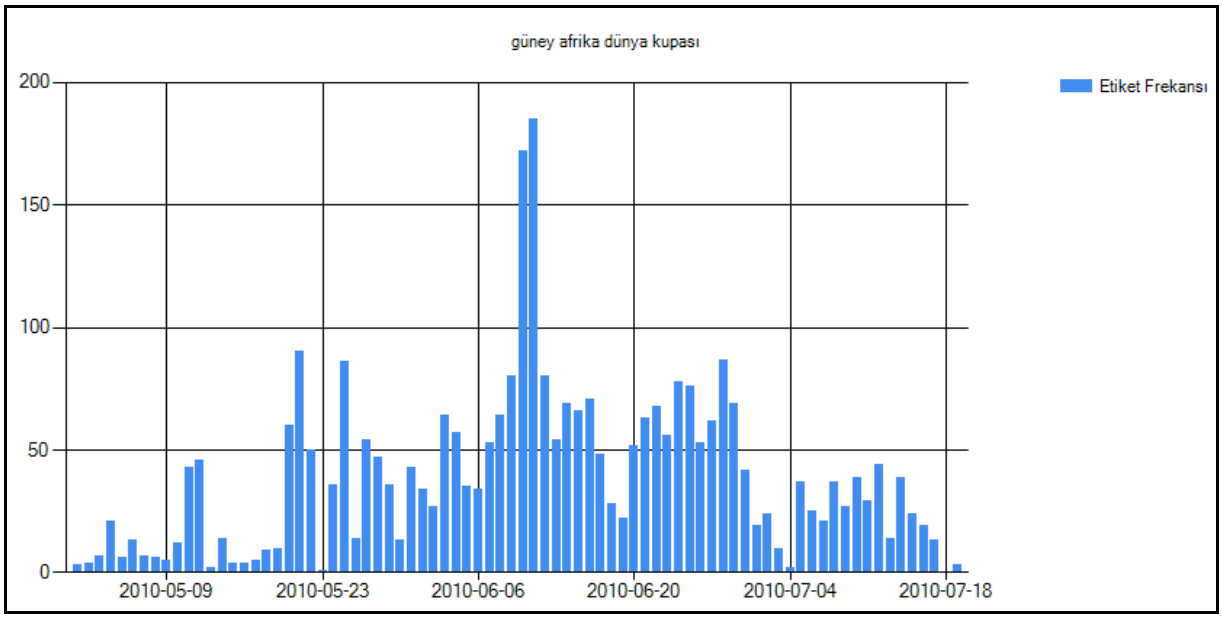
9-14 Ekim 2010 tarihleri arasında gerçekleştirilen Antalya Altın Portakal Film Festivali Şekil 5.10'da görülen etiket bulutunda belirgin bir şekilde listelenmektedir. Bu yıl festivalde jüri üyesi olan Emir Kusturica'nın, hakkında çıkan haberlerden dolayı gündemde önemli bir yer edindiği de görülmektedir. Ayrıca eski Cumhurbaşkanı Turgut Özal, ölümüyle ilgili son dönemde yaşanan tartışmalardan dolayı gündemin önemli bir maddesi haline gelmiştir.

Şekil 5.10 Gündem Bulutu ekran örneği (10.10.2010 – 17.10.2010)

Şekil 5.11'de bir önceki etiket bulutunu takip eden haftaya ait gündem görülmektedir. Etiket listesinden görüleceği üzere Antalya Altın Portakal Film Festivali'ne ait haberler bir anda kesilmiş, buna karşın gündeme yeni konular girmiştir. Almanya Cumhurbaşkanı Christian Wulff'un ülkemizi ziyaretinden dolayı bu konuda yapılan haberler gündemin bir parçası haline gelmiştir. Yine aynı hafta, Galatasaray futbol takımının, teknik direktör Frank Rijkaard ile yollarını ayırması da özellikle spor gündeminin en önemli haberi olarak karşımıza çıkmıştır. Eski Cumhurbaşkanı Turgut Özal ile ilgili haberlerin de halen devam ettiği etiket listesinden görülmektedir.

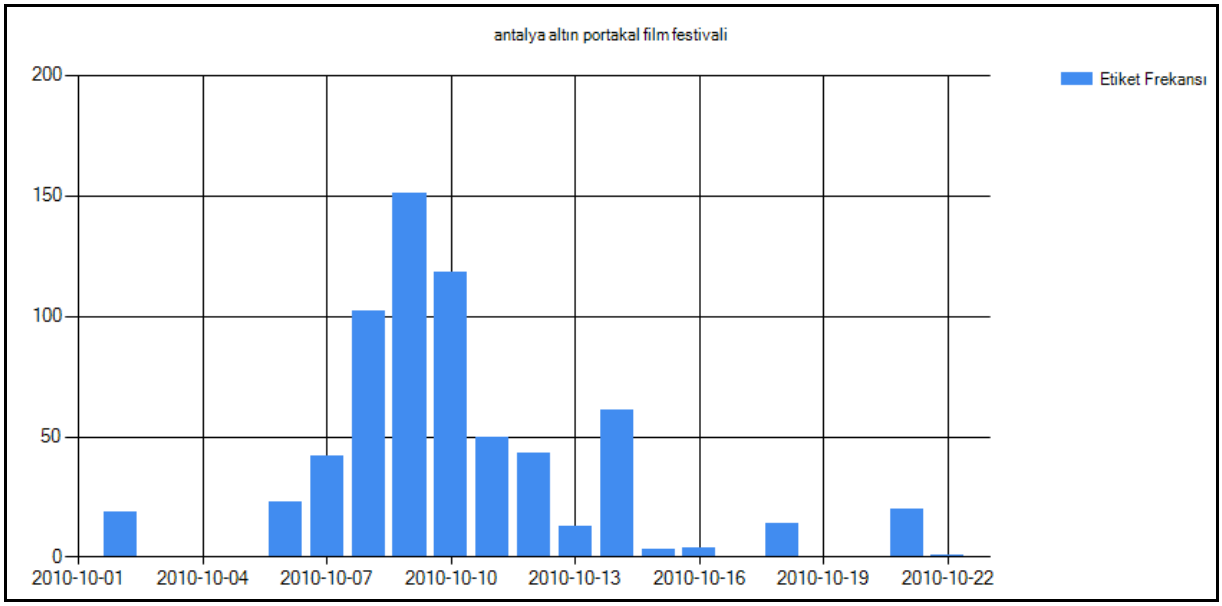
Şekil 5.11 Gündem Bulutu ekran örneği (17.10.2010 – 24.10.2010)

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 5.12'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 5.12 “Güney Afrika Dünya Kupası” etiketinin frekans grafiği

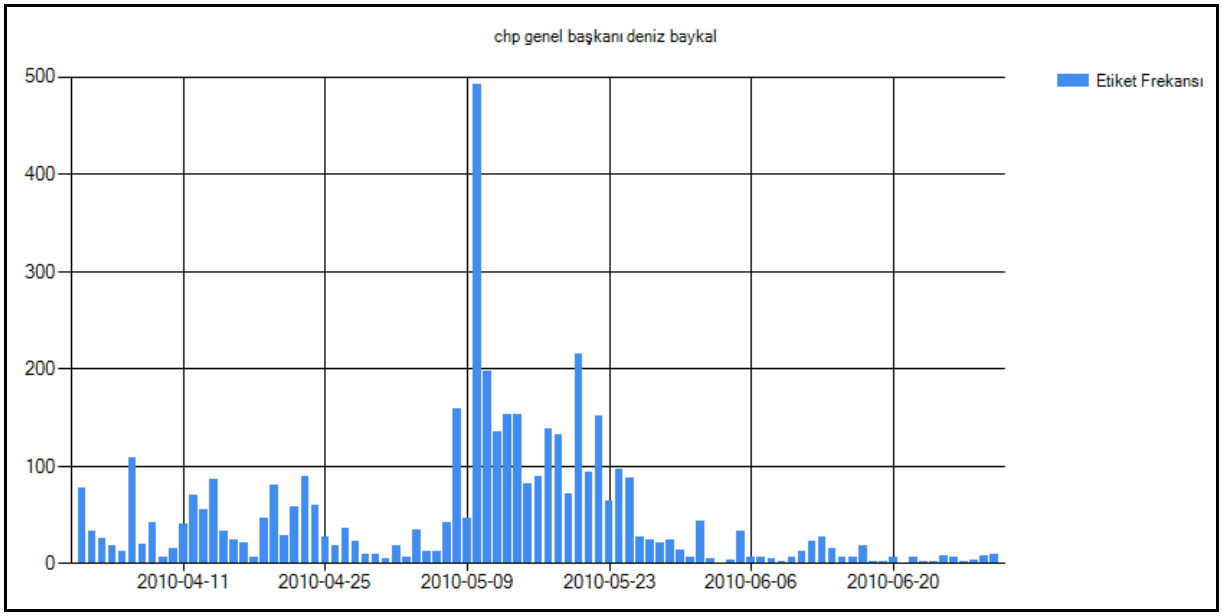
Antalya Altın Portakal Film Festivali’ne ait haberlerde, Şekil 5.13’ten görüleceği üzere sadece festival tarihlerinde bir yoğunluk yaşanmıştır. Grafikteki veriler, bu tür organizasyonların gündemde uzun süreli yer bulmadığını göstermektedir.



Şekil 5.13 “Antalya Altın Portakal Film Festivali” etiketinin frekans grafiği

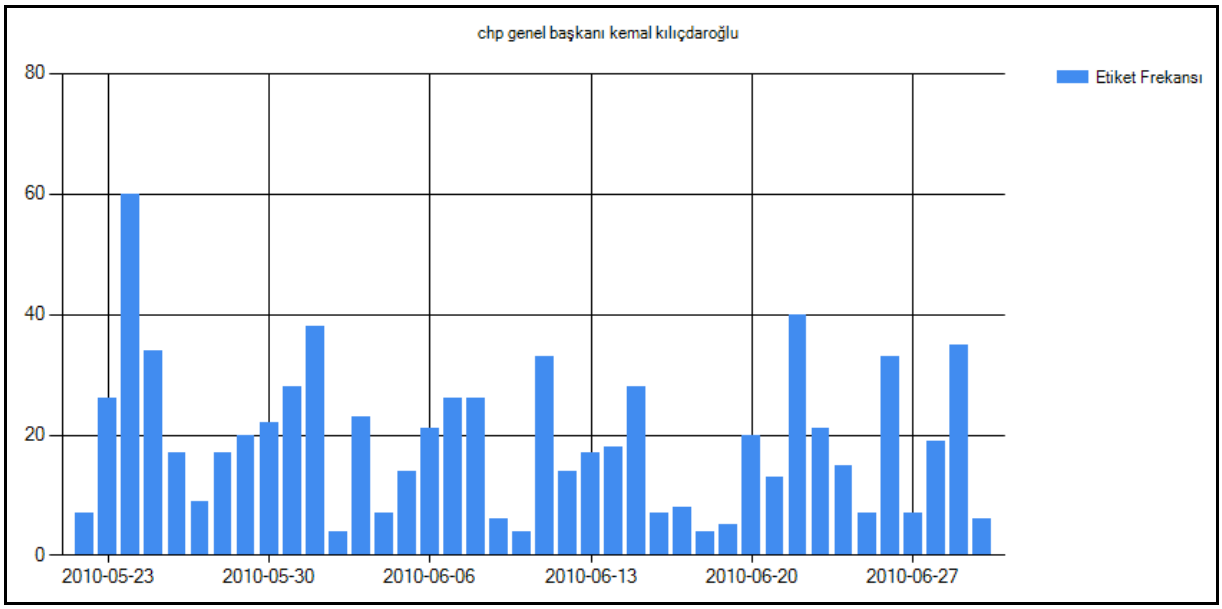
2010 yılının Mayıs ayında Cumhuriyet Halk Partisi’nde genel başkan değişikliği meydana gelmiştir. Bu konunun haberlere nasıl yansıdığını inceleyebilmek adına eski genel başkan Deniz Baykal’ı işaret eden etikete ait Şekil 5.14’teki grafik oluşturulmuştur. Grafikten

görüleceği üzere değişikliğe kadar olan dönemde günlük politik gelişmeleri işleyen haberlerden dolayı belli bir seviyede devam eden frekans, Mayıs ayında bir anda çok yükselmiştir. Değişimin gerçekleşmesi ile de frekans en alt seviyelere düşmüştür.



Şekil 5.14 "CHP Genel Başkanı Deniz Baykal" etiketinin frekans grafiği

Cumhuriyet Halk Partisi'nin yeni genel başkanı için yapılan incelemeye ait grafik ise Şekil 5.15'te sunulmuştur. Başkanlık değişimine kadarki dönemde sıfır olan frekans, değişimle birlikte yüksek seviyelere ulaşmıştır.



Şekil 5.15 "CHP Genel Başkanı Kemal Kılıçdaroğlu" etiketinin frekans grafiği

Uygulamaya ait yukarıda sunulan örneklerden görüleceği üzere geliştirilen sistemden oldukça tutarlı sonuçlar alınabilmektedir. Tez süresince yapılan geliştirmelerde haber metinlerinin en iyi şekilde analiz edilmesine ve içeriği temsil eden etiket listesinin mümkün olduğunca doğru bir şekilde oluşturulmasına çalışılmıştır. Yapılan testler sonucunda, tasarlanan kullanıcı arabiriminin haberlere ulaşmada başarılı olduğu görülmüştür.

6. UYGULAMA ALTYAPISININ FARKLI KULLANIMLARI

Yapılan bu çalışma sonucunda ortaya çıkan altyapı ile özel bir uygulama geliştirilmiş olmakla beraber bu altyapıdan faydalanılarak farklı uygulamalar geliştirilmesi de mümkündür. Bu tür uygulamalara fikir vermesi açısından, tez kapsamında bazı örnek uyarlamalar gerçekleştirilmiştir.

Etiket bulutları, istatistikî verileri sunmak için güzel bir görsel araçtır. Bu sunuma örnek olması açısından Türkiye İstatistik Kurumu'nun sitesinden [20] alınan, 2009 yılına ait nüfus sayımı sonuçları ile Şekil 6.1'de görülen etiket bulutu oluşturulmuştur.

Şekil 6.1 İllerin 2009 yılına ait nüfus sayılarının etiket bulutu

Benzer bir şekilde, yine Türkiye İstatistik Kurumu'nun sitesinden [20] alınan illere ait

yüzölçümü verileri ile Şekil 6.2'de sunulan etiket bulutu oluşturulmuştur.

Şekil 6.2 İllere ait yüzölçümü verileri ile oluşturulan etiket bulutu

Bir diğer örnek uygulama olarak, bu çalışmada yazılan tez metnine ait Şekil 6.3'te görülen etiket bulutu oluşturulmuştur. Çalışmada geliştirilen altyapı ile bu ve benzeri veri ve içerikleri analiz ederek farklı amaçlara yönelik olarak etiket bulutları oluşturmak mümkün olmaktadır.

Şekil 6.3 Tez metninden oluşturulan etiket bulutu

7. 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. Bu hedefe yönelik olarak tez sürecinde doğal dil işleme ve makine öğrenmesi üzerine yapılmış önceki çalışmalar incelenmiş, bilgiye erişim sistemleri, metin madenciliği ve etiket bulutları konularında araştırmalar yapılmıştır. Geliştirilen uygulamanın temel işlevi toplanan verilerin işlenmesi ve sunulması ile bilgiye erişim sağlamaktır. 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. Kullanıcıların bilgiye ulaşmalarını sağlayan aracı uygulamalara Bilgiye Erişim Sistemleri denilmektedir. Tez kapsamında geliştirilen Gündem Bulutu uygulaması da aynı işlevleri yerine getirmektedir. Bundan dolayı bu uygulamanın bir Bilgiye Erişim Sistemi olduğu söylenebilir.

Uygulamanın ilk adımını oluşturan, haber sitelerinden içeriklerin toplanması konusunda yapılan araştırma sonucunda RSS yayınlarından faydalanılması uygun görülmüştür. Bundan dolayı sadece RSS yayını olan haber kaynakları değerlendirmeye alınmıştır. Uygulamanın hedefi gündemi yansıtmak olduğu için, bu yayınlardan en uygun olanları seçilmiştir. Daha sonra haber içeriklerini toplayan robot bileşeni geliştirilmiştir. Bu bileşen tüm haber kaynaklarını takip ederek yeni içerikleri anında derleme dâhil etmektedir. Toplanan haber içeriklerinin analiz edilerek etiket listesi oluşturulması yönünde yapılan araştırma sonucunda metin madenciliğinin bu tür ihtiyaçlara cevap veren çeşitli teknikler sunduğu görülmüştür. Öncelikle metin analizi teknikleri incelenmiş ve Türkçe içerik üzerindeki başarıları araştırılmıştır. Bu işlemlerden sağlıklı sonuçlar alınabilmesi için Türkçeye özgü bir biçimbilimsel çözümleyici kullanılmasına karar verilmiş ve benzer çalışmalarda daha önceden tercih edilen açık kaynak kodlu Zemberek projesi incelenmiştir. Yapılan testler sonucunda .NET sürümündeki sözlük kütüphanesinin JAVA sürümündekine göre çok daha eski olduğu görülmüştür. Bundan dolayı uygulamada daha tutarlı sonuçlar almak amacıyla .NET sürümündeki sözlüğün güncel hale getirilmesi yönünde çalışmalar yapılmıştır.

İçeriğin toplanması ve analiz edilmesi sonucunda elde edilen bilgiler kullanılarak etiket bulutu oluşturma aşamasına geçilmiştir. Bu konuyla ilgili olarak mevcut etiket bulutu uygulamaları incelenmiş ve bu çalışmada sunulacak çözüme yönelik olarak etiket bulutu çeşitleri ve

özellikleri karşılaştırılmıştır. 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 siteyi ziyaret eden kullanıcıların hem en güncel gündemi görebilmeleri hem de isterlerse eskiye dönük gündemi görebilmeleri sağlanmıştır. Kullanıcılar etiket bulutundaki ilgili terimlere tıklayarak o etiketin eşlendiği haberlerin tarih, başlık ve özet bilgilerinden oluşan bir listeye ulaşabilmektedir. Bu listeden istenen bir haberin linkine tıklayarak ta haberin yayınlandığı asıl siteye ulaşılıp, haberin tüm metni okunabilmektedir. Bunlara ek olarak belli bir etiketin gündemde olma frekansı da görsel olarak kullanıcılara sunulmaktadır.

Bu çalışmanın en önemli iki aşaması, haber metinlerinin analiz edilmesi ve içeriği temsil eden etiket listesinin doğru bir şekilde oluşturulmasıdır. Tez sürecinde bu yönde yapılan çalışmalar sonucunda oldukça tatmin edici sonuçlar alınmıştır. Zaman içinde sistemin işleyişinin gözlenmesi ve etiket bulutlarının iyileştirilmesi yönünde çalışmalar yapılması planlanmaktadır. Örneğin özel isim tespitinde kontrollü bir derlem kullanılması yönünde çalışma yapılabilir. Sisteme erişen kullanıcılara daha hızlı hizmet verebilmek amacıyla da bir önbellek yapısı oluşturulabilir. Bunlara ek olarak, metin madenciliği alanında yapılan yeni çalışmalarda elde edilen bulgular ve geliştirilen yeni tekniklerden faydalanılarak etiket belirleme işleminin daha başarılı bir hale getirilmesi mümkün olabilecektir.

KAYNAKLAR

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

Donath, J., (2002), "A semantic approach to visualizing online conversations", Communications of the ACM, 45(4):45-49.

Ergin, M., (1998), Üniversiteler İçin Türk Dili, Bayrak, İstanbul.

Friedman, V., (2007), "Tag Clouds Gallery: Examples And Good Practices", http://www.smashingmagazine.com/2007/11/07/tag-clouds-gallery-examples-and-good-practices.

Hearst M.A., (2008), "What's Up With Tag Clouds", Visual Business Intelligence Newsletter, Mayıs 2008.

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

Huijsmans, D.P. ve Sebe, N., (2003), "Content-based indexing performance: size normalized precision, recall, generality evaluation", Proceedings of the 2003 International Conference on Image Processing, 14-18 Eylül 2003, Barcelona, Catalonia, Spain.

Indurkhaya, N. ve Weiss, S.M., (1998), Predictive Data Mining, Morgan Kaufmann Publishers Inc., San Francisco.

İslam, Y., (2005), Yönlendirilmiş Çalışma I-II, Seçkin, Ankara.

Karagedik, Ö. ve Önal, A., (2010), "Bilgiye Erişim Sistemlerinde Arama Kalitesini İyileştirme: Normalleştirme Etkeninin Önemi", Akademik Bilişim 2010, 10-12 Şubat 2010, Muğla.

Karasar, N., (2003), Bilimsel Araştırma Yöntemi, Nobel, Ankara.

Lamantia, J., (2006), "Tag Clouds Evolve: Understanding Tag Clouds", http://www.joelamantia.com/ideas/tag-clouds-evolve-understanding-tag-clouds.

Lamantia, J., (2007), "Text Clouds: A New Form of Tag Cloud?", http://www.joelamantia.com/tag-clouds/text-clouds-a-new-form-of-tag-cloud.

Lohmann, S., Ziegler, J., Tetzlaff, L., (2009), "Comparison of Tag Cloud Layouts: Task-Related Performance and Visual Exploration", Proceedings of the 12th IFIP TC 13 International Conference on Human-Computer Interaction, 24-28 Ağustos 2009, Uppsala.

Manning, C.D., Raghavan, P., Schütse, H., (2008), Introduction to Information Retrieval, Cambridge University Press, Cambridge.

Maron, M.E., (1984), "Probabilistic Retrieval Models", Progress in Communication Sciences 5:145-176.

Oflazer, K. ve Bozşahin, H.C., (2006), "Türkçe Doğal Dil İşleme", Çukurova Üniversitesi Türkoloji Araştırmaları Merkezi.

Singhal, A., (2001), "Modern Information Retrieval; A Brief Overview", IEEE Data Engineering Bulletin, 24(4):35-43.

Tonta, Y., (2005), "İçerik Analizi ve Karşılaştırmalı Yöntem", http://yunus.hacettepe.edu.tr/~tonta/courses/spring2005/bby208/bby208-4b-icerik-analizi.ppt.

Tonta, Y., (2009), "Bilgi Erişim Sorunu", http://yunus.hacettepe.edu.tr/~tonta/courses/spring2009/bby220/bby220-bilgi-erisim-sistemleri-2009-1.ppt.

Tonta, Y., Bitirim, Y., Sever, H., (2002), Türkçe Arama Motorlarında Performans Değerlendirme, Total Bilişim Ltd. Şti., Ankara.

Varallo, V., (2009), ASP.NET 3.5 Enterprise Application Development with Visual Studio 2008, Wiley Publishing, Inc., Indianapolis.

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

İnternet Kaynakları

[1] http://tr.wikipedia.org/wiki/Dil_bilimi

[2] http://tartarus.org/~martin/PorterStemmer/

[3] http://code.google.com/p/nzemberek

[4] http://en.wikipedia.org/wiki/Tag_cloud

[5] http://www.smashingmagazine.com/2007/11/07/tag-clouds-gallery-examples-and-good-practices

[6] http://www.webdesignerwall.com

[7] http://wordpress.org/support

[8] http://www.last.fm/charts/toptags

[9] http://www.flickr.com/photos/tags

[10] http://www.nytimes.com

[11] http://blog.dermstore.com/index.php/the-difference-between-natural-and-organic-skin-care

[12] http://www.toubeauty.com

[13] http://www.iphoneturkey.biz

[14] http://www.iphoneblog.de

[15] http://www.misjournal.com

[16] http://blogs.msdn.com

[17] http://chir.ag/projects/preztags/

[18] http://twopointoh.co.uk/2007/04/21/curriculum-vitae-as-a-tag-cloud/

[19] http://snowball.tartarus.org

[20] http://www.tuik.gov.tr

[21] http://www.hurriyet.com.tr

[22] http://www.milliyet.com.tr

[23] http://www.radikal.com.tr

[24] http://www.zaman.com.tr

[25] http://www.posta.com.tr

[26] http://www.stargazete.com

[27] http://www.bugun.com.tr

[28] http://www.sabah.com.tr

EKLER

Ek 1 Türkçe Etkisiz Kelime Listesi

a	be	birşeyi	d
aa	belki	biz	da
acaba	ben	bizden	daha
ait	bende	bize	dahi
altı	benden	bizi	dandini
altmış	beni	bizim	de
ama	benim	böyle	defa
amma	beş	böylece	değ
anca	bide	bu	değil
ancağ	bile	buna	değin
ancak	bin	bunda	dek
artık	bir	bundan	demek
asla	birazı	bunu	diğer
aslında	birçoğ	bunun	diğeri
az	birçoğu	burada	diğerleri
b	birçok	bütün	diye
bana	birçokları	c	dk
bari	biri	ç	dha
başkası	birisi	çoğu	doğrusu
bazen	birkaç	çoğuna	doksan
bazı	birkaçı	çoğunu	dokuz
bazıları	birkez	çok	dolayı
bazısı	birşey	çünkü	dört

e
eğer
eh
elbette
elli
en
etkili
f
fakad
fakat
falan
falanca
felan
filan
filanca
g
ğ
gene
gereğ
gerek
gibi
göre
görece
h

hakeza
hakkında
hâlâ
halbuki
hangi
hangisi
hani
hasebiyle
hatime
hatta
hele
hem
henüz
hep
hepsi
hepsine
hepsini
her
her biri
herkes
herkese
herkesi
hiç
hiç kimse

hiçbiri
hiçbirine
hiçbirini
hoş
i
ı
ın
için
içinde
içre
iki
ila
ile
imdi
indinde
intağ
intak
ise
işte
ister
j
k
kaç
kaçı

kadar
kah
karşın
katrilyon
kelli
kendi
kendine
kendini
keşke
keşki
kez
keza
kezaliğ
kezalik
ki
kim
kimden
kime
kimi
kimin
kimisi
kimse
kırk
l

lakin
m
madem
mademki
mamafih
meğer
meğerki
meğerse
mi
mı
milyar
milyon
mu
mü
n
nasıl
nde
ne
ne kadar
ne zaman
neden
nedense
nedir
nerde
nere
nerede
nereden
nereli
neresi
nereye
nesi
neye
neyi
neyse
niçin
ni
nı
nin
nın
nitekim
niye
o
ö
öbürkü
öbürü
on
ön
ona
önce
onda
ondan
onlar
onlara
onlardan
onlari
onların
onu
onun
orada
ötekisi
ötürü
otuz
öyle
oysa
oysaki
p
pad
pat
peki
r
rağmen
s
ş
sakın
sana
sanki
şayet
sekiz
seksen
sen
senden
seni
senin
şey
şeyden
şeye
şeyi
şeyler
şimdi
siz
sizden
size
sizi
sizin
son
sonra

şöyle
şu
şuna
şunda
şundan
şunlar
şunu
şunun
t
ta
tabi
tamam
tl
trilyon
tüm
tümü
u
ü
üç
üsd
üst
uyarınca
üzere
v
var
ve
velev
velhasıl
velhasılıkelam
vesselam
veya
veyahud
veyahut
y
ya
ya da
yani
yazığ
yazık
yedi
yekdiğeri
yerine
yetmiş
yine
yirmi
yoksa
yukarda
yukardan
yukarıda
yukarıdan
yüz
z
zaten
zinhar
zira

ÖZGEÇMİŞ

Doğum tarihi	12.02.1973	
Doğum yeri	Balıkesir	
Lise	1984-1991	Balıkesir Sırrı Yırcalı Anadolu Lisesi
Lisans	1991-1995	Yıldız Teknik Üniversitesi Elektrik Elektronik Fak. Bilgisayar Mühendisliği Bölümü

Çalıştığı Kurumlar

1995-1999	Kets Ltd. Şti.
1999-2003	Invecta, Inc. CA, ABD
2003-Devam ediyor	Kets Ltd. Şti.