YBS Ansiklopedi
www.YBSAnsiklopedi.com
Cilt 2, Sayı 2, Haziran 2015

Sosyal Ağlarda Veri Madenciliği (Data Mining on Social Networks)

Sadi Evren SEKER

Istanbul Medeniyet University, Department of Business

ÖZET

Makalenin konu başlığı günümüzde hızla gelişmekte olan sosyal ağlar üzerinde veri madenciliği tekniğinin kullanıldığı bazı problemleri tanıtmak ve geliştirilen bazı çözüm yöntemleri ve algoritmaları açıklamaktır. Genel olarak sosyal ağlar üzerinde kullanılan çizge teoremi, büyük veri işleme ve metin madenciliği teknikleri gibi başlıklar altında incelenebilecek, bölütleme (kümeleme), grup belirleme, duygu analizi veya fikir madenciliği konularına yer verilmiştir. Makalenin genel amacı, Türkçe literatüre kavramlar hakkında temel tanımları kazandırmak olduğu kadar genel bir literatür taramasını da okuyucuya sunmaktır.

Anahtar Kelimeler: Veri Madenciliği, Sosyal Ağ Analizi, Metin Madenciliği, Fikir Madenciliği, Çizge Teoremi, Sosyal Ağ Kümeleme

ABSTRACT

The aim of this study is introducing some well-known problems over the rapidly developing social networks and some basic solution approaches and algorithms for the problems. Some basic approaches like graph theory, big data or text mining techniques will be introduced in a broader view and some problems like clustering, group identification, sentimental analysis or opinion mining will be introduced. One major purpose of the paper is introducing terminology to the Turkish literature and proposing a literature survey to the reader.

Keywords : Data Mining, Social Network Analysis, Text Mining, Opinion Mining, Graph Theory, Clustering on Social Networks

1. Giriş

Bu yazı, her gün kullanımı artmakta olan ve artık hayatın bir parçası haline gelmiş olan Facebook, Twitter, Linked-in veya Youtube gibi sosyal ağların nasıl birer veri kaynağı olarak kullanılabileceği ve bu veri kaynakları üzerinde klasik ve yeni veri madenciliği tekniklerinin nasıl kullanılacağını açıklamaktır.

Sosyal ağların bilimsel olarak çok farklı ifade ve modellemeleri olmakla birlikte, literatürde en çok kabul görmüş gösterim şekli çizge teoremi (graph theory) kullanılarak bireylerin veya varlıkların birer düğüm (node) ve ilişkilerin birer kenar (edge) olarak tasvir edildiği gösterimdir (Seker, Cizge Teorisi (Graph Theory), 2015).

Örnek olarak bir Facebook üzerinde birbirini arkadaş olarak eklemiş olan kişilerin temsili çizge gösterimi aşağıda verilmiştir.

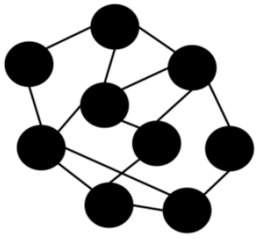


Figure 1 Temsili Sosyal Ağ'ın Çizge Gösterimi

Facebook yapısı itibariyle arkadaşlık ilişkilerini iki yönlü olarak tanımlar. Yani bir kişinin eklediği arkadaş aynı zamanda karşı taraf tarafından da onaylandığı için iki kişi birbirini karşılıklı olarak eklemekte bu yüzden çizge üzerindeki bağlantılar yönsüz olarak gösterilmektedir. Ancak örneğin Facebook ağındaki takip etmek veya Twitter üzerinden takip etme gibi işlemlerin yönü vardır. Yani bir tarafın diğer tarafı takip etmesi bir yön belirtirken aksi yönde takip olabileceği gibi olmak zorunda değildir.

Çizge teorisinin yanında veri madenciliği çalışmaları için en önemli ön bilgilerden birisi, sosyal ağların büyük veri (big data) kavramı ile yakından ilişkili olmasıdır. Yani sosyal ağlarda verinin büyüklüğü (volüme), değişme hızı (velocity), çeşitliliği(variety) ve doğruluğu (veracity) problemleri bulunmaktadır (Seker & Eryarsoy). Veri madenciliği teknikleri, büyük veri çalışmaları için kritik sonuçlar doğurabilecek ve veri kümesinden çeşitli örüntü, kural ve trendleri çıkarabilecek ve bu çıkarımlar sayesinde önemli bilgilerin kullanıma kazanmasını sağlayacak değişik yöntemler içermektedir. Genelde bu işlemler veri ön işlemesi, veri analizi ve veri yorumlaması gibi aşamalardan geçmektedir (Seker, İş Zekası ve Veri Madenciliği, 2013). Bu teknikleri içeren ve sosyal ağa analizi için kullanılan çeşitli yöntemleri gösteren bir tablo, Tablo 1'de verilmiştir.

Yazının başlığı her ne kadar sosyal ağlar ve veri madenciliği ile ilgili olsa da, sosyal ağlar web 2.0 ile gelen kavramlardan

birisidir ve web 2.0 ile gelen blog'lar veya wiki'ler de sosyal ağlara benzer şekilde veri madenciliğinin çalışma alanına girmektedir (Seker, Cankir, & Okur, Strategic Competition of Internet Interfaces for XU30 Quoted Companies, 2014).

Sosyal ağlar genel olarak içerik paylaşımı (Thompson, 2013), (Chelmis & V.K., 2011), kişiselleştirme (Asur & Huberman, 2010) , yorumlama (G. & A, 2005), yaklaşımlar (Kaji & Kitsuregawa, 2006) , değerlendirme (Kaur, 2013), etkileme 8, gözlem (Chou, Y.M., Beckjort, R.P., & Hesse, 2009)u, hisler (Seker & Al-Naami), kanaat ve duygusal ifadeler (B. & L., 2008) için kullanılmaktadır.

Veri madenciliği teknikleri sosyal ağlardan çok daha önce internet üzerinde kullanılmaya başlamıştı. 1990'lı yıllarda henüz sosyal ağlar ortada yokken ve internet hayata yeni yeni girmeye başlamışken sabit web sayfaları üzerinden, ki buna sınıflandırma olarak web 1.0 denilmektedir, insanlar bilgi paylaşımı yapmaktaydı ve bu web sayfaları üzerinde dolaşarak verileri toplayan ve daha sonra bu toplanan verilere göre veri madenciliği teknikleri uygulayarak dolaşma yöntemlerini daha akıllı yapmak da dahil olmak üzere bir dizi bilgi çıkarımı yapan yöntemler bulunmaktaydı. Bu yöntemler daha sonraları literatürde web madenciliği olarak isimlendirildiler (Seker, Al-Naami, & Khan, Author attribution on streaming data, 2013).

Bu makale kapsamında, ikinci bölümde, sosyal ağların arka planı ve sosyal ağ analizini konularına değinilecek, üçüncü bölümde çizge teorisi ile sosyal ağlar için bu kapsamda geliştirilmiş algoritmalar ve yöntemler incelenecek, dördüncü bölümde sosyal ağ analizi için geliştirilen çeşitli çizge teoremi yaklaşımları incelenecek, beşinci bölümde ise sosyal ağlardaki duygu analizi ve fikir madenciliği yöntemleri incelenecektir.

Bahsi geçen konulara başlamadan önce hatırlamakta yarar olan bir konu, büyük veri (big data) işlemenin zorunluluklarından birisinin de, işlemlerin otomatik olarak ele alınması gerekliliği ve dolayısıyla işlemlerin kabul edilebilir sürelerde çalışmasıdır. Bu makalede anlatılacak olan pek çok yöntemden daha iyi çalışan yöntemler literatürde bulunmaktadır. Örneğin duygu analizi için doğal dil işleme (natural language processing) yöntemlerinin çok daha yüksek başarılar elde ettiği söylenebilir. Ancak problem, yüksek başarının yanında işlem süresinin de kabul edilebilir seviyede olması gerekliliğidir.

2. Sosyal Ağ Analizi Yaklaşımları

Sosyal ağlar üzerinde yapılan analiz çalışmaları yaklaşımlara göre iki farklı grupta incelenebilir.

Bağlantı tabanlı ve yapısal analizler. Bu analiz yönteminde sosyal ağda bulunan bağlantılar kullanılarak bir topluluk veya alandaki ilgili düğümlerin, bağlantıların, alt çizgelerin veya ilgi alanlarının çıkarılması hedeflenmektedir.

Hareketli ve Sabit analiz. Ağdaki hareketliliğe bağlı olarak kullanılan sabit (static) analiz veya hareketli (dynamic) analiz yöntemi seçilebilir. Örneğin kitapların atıflarının takip edildiği bir ağda sabit analiz yapılabilirken, verinin hızla aktığı bir ağda hareketli analiz yöntemlerine ihtiyaç duyulmaktadır. Temel olarak sabit analiz yöntemi, bir sosyal ağın değişmemesi veya değişimin sabit boyutlarda olması durumunda kullanılırken, dinamik analiz yöntemi Facebook veya Youtube gibi değişimin ön görülemeyeceği veya üstel olarak değiştiği durumlarda kullanılabilir (Khaled Al-Naami, 2014).

Sosyal ağlar, web 2.0 ile hayata girmiştir ve bu makale yazıldığı sıralarda web 3.0'ın tasarımı ile ilgili detaylar bilinmekle birlikte tam ve yaygın bir uygulaması henüz bulunmamaktaydı. Ancak web 3.0 tasarımı itibariyle anlambilimsel ağlardan oluşmakta olup (semantic web) bütün internet erişim noktalarının (sosyal ağlar, web siteleri, wiki'ler vs.) birbiri ile anlambilimsel gösterimler aracılığı ile bilgi alışverişinde bulunmasını önermekteydi. Bu yaklaşım aslına bütün internetin farklı yapılarda bilgi iletimi yaptığı büyük bir çizge olarak düşünülmesi demektir. Örneğin akıllı bir buzdolabının, içerikleri takip ederek buzdolabında süt bitince bu bilgiyi daha önceden belirlenen markete iletmesi ve marketin sütü müşteri için tedarik etmesi gibi bir dizi zincirleme ilişkiyi başlatması ve bütün bu ilişkilerin internet üzerinden akan bir ağda olması mümkündür. Anlambilimsel ağlar, klasik çizge yaklaşımına göre ağdaki aktörlerin birer makine olabileceği yapıları doğurmuştur. Bu anlamda bakıldığında bir aktör ağı teorisi (actor network theory) yaklaşımıdır denilebilir (Seker S. E., Possible Social Impacts of E-Government: A Case Study of Turkey, 2004).

3. Çizge Teorisi Yaklaşımı

Sosyal ağ kavramının çıkışından beri kullanılan ve en çok kabul gören ve muhtemelen ilk yöntemlerden birisi olan çizge teorisi günümüzde de en önemli sosyal ağ analizi yöntemi olarak kabul edilmektedir. Genelde bir ağdaki varlıkları ve bu varlıklar arasındaki ilişkileri göstermek için kullanılırlar. Sosyal bilimler açısından bu gösterime etkileyiciler ve takipçiler ismi verilmektedir. Yani bir sosyal ağda, bir kaynak etkiyi yayma özelliğine sahipken diğer kaynaklar bu etkiyi takip etmektedirler. Özellikle çok büyük ölçekli veri tabanlarında çizge teorisinin kullanımı daha da büyük bir öneme sahiptir. Çizge teorisi her ne kadar kitaplarda ve diğer kaynaklarda görsel olarak bir sosyal ağı göstermek için kullanılsa ve anlatılan bütün konular görselleştiriliyor olsa da, çizge teorisinin çalışması için ağın görsel hale getirilmesi her zaman gerekmez. Örneğin en kısa yol bulma veya seyyar satıcı problemi gibi çok meşhur çizge teoremi problemlerinde bile, problem görselleştirilmeden, doğrudan veriler üzerinden işlem yapmak mümkündür. Kullanılan yöntemler çizge teoremi yöntemleri olsa da matrisler, tablolar veya denklemler şeklinde gösterimler üzerinden de problemler çözülebilmektedir. Örneğin merkezlik ölçümü (centrality measure) bir sosyal ağdaki en etkili yayım gücüne sahip kişi yada kişileri bulmak için kullanılan bir çizge teorisi yaklaşımıdır ve problemin çözümü için kullanılan yöntemler oldukça basit hesaplamalardan ibarettir (Borgatti & Everett, 2006). Merkezlik analizinin daha farklı hesaplama yöntemlerinde bile benzer basitlikte işlemler yapılmaktadır. Örneğin zayıflatılmış yolların sayısını saymak kadar basit bir işlem üzerine kurulu olan α-merkezlik (α-centrality) metriği veya bağlantılık matrisinin oluşturularak üzerinden basit matris işlemleri ile yapılan parametrize merkezlik analizi yöntemleri (Ghosh & Lerman, 2011) hep basit hesaplamalara dayalı, çizge teoreminden beslenen ancak görselleştirmeye ihtiyacı olmayan yöntemlerdir.

4. Bölütleme (kümeleme) ve Topluluk Bulunması

Sosyal ağların çizge olarak gösterildiği durumlarda her düğüm bir bireyi ifade etmektedir. Birden fazla düğümün bir araya geldiği alt çizgeler ise birer topluluk olarak düşünülebilir. Örneğin çizgenin tamamı bir alt çizge olarak kabul edilebilir ve topluluğun tamamını göstermektedir veya çizgeden rasgele olarak seçilen üç düğüm aslında üç kişilik bir topluluğu ifade etmektedir. Bu toplulukların taşımız olduğu özelliklere göre isimlendirilmesi de mümkündür. Örneğin bir sosyal durumu ifade eden ve bir topluluktaki yakın arkadaşları anlatmak için kullanılan klik, çizge üzerinde de birbirine kuvvetli bağlı bir

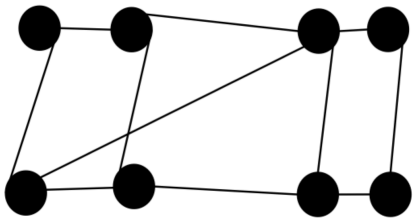
topluluğu ifade etmektedir. Böyle özel bir yapının bulunması için özel bazı algoritmalar geliştirilmiştir.

Topluluğun alt topluluklara bölünmesi ise farklı şekillerde ele alınabilir. Örneğin 4.3 alt bölümünde de anlatılacak olan Girvan-Newman algoritması sosyal yapıları, ilişki yoğunluğuna göre alt sosyal yapılara bölmektedir (Girvan & Newman, 2002). (Fortunato, 2010) ise çalışmasında çok detaylı bir şekilde farklı topluluk çıkarım yöntemlerinden bahsetmektedir. Örneğin çizge parçalama (graph partitioning), düğüm benzerlikleri (vertex similarity), hiyerarşik bölütleme (hierarchical clustering), parçalı bölütleme (partitional clustering) veya tayfla bölütleme (spectral clustering) gibi geleneksel yöntemlerin yanında, üstüste binmiş toplulukların bulunması (overlapping communities), istatistiksel metotlar veya değişken toplulukların yakalanması gibi çeşitli yöntemlerden bahsetmektedir. Literatüre bakıldığında kullanılan çok sayıdaki bölütleme yöntemi arasından en yaygın olanının hiyerarşik bölütleme olduğu söylenebilir (M, 2010). Bu yöntem aslında çok sayıdaki diğer yöntemin birleştirilerek bağımsız grupların gücünü ortaya çıkarmakta ve ağı alt gruplara bölmektedir.

Yöntemlerin detaylarına geçilmeden önce belirtmek gerekir ki bütün yöntemlerde bazı bilgilerin ön tanımlı olması gerekmektedir. Örneğin kaç bölüt bulunmak istediği veya algoritmadaki bazı parametrelerin örneğin eşik değerlerinin tanımlı olması beklenir.

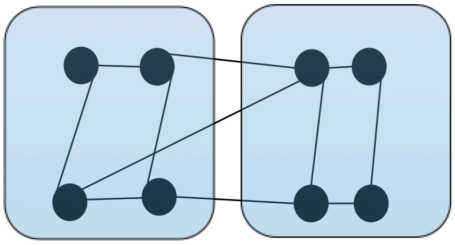
Bölütleme için kullanılan en basit ve ilkel algoritmalardan birisi k-ortalama (k-means) algoritmasıdır. Algoritma k bölütün ortalamasından oluşan bir değere göre bölütlerin çekirdeklerini oluşturmaktadır. Bu algoritma için k sayısının ön tanımlı olarak verilmesi gerekir.

Algoritmaların parametrik olmasının anlamı, aslında algoritmaların problemi hangi seviyede ele alacağının algoritmaya belirtilmesi olarak düşünülebilir. Örneğin aşağıdaki örnekleri ele alalım:

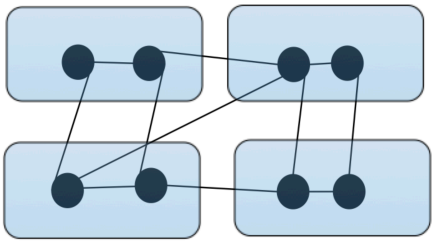


Yukarıdaki çizge için bir çizge bölütlemesi yapmak istiyor olalım ve bu bölütlemenin öklit mesafesine göre çalıştığını kabul edelim. Yani basitçe bütün düğümler ilgili oldukları düğüme yakın durmakta olsun ve düğümler arasındaki mesafe düğümlerin ilişkisini göstersin (düğümler bağlı olmasa bile yakın duruyorlarsa birbirine yakın düğüm olarak kabul edilsin).

Bu durumda şekli bölütlere ayırmadan sorulabilecek bir soru, "bu şekilde kaç bölüt vardır?" sorusudur. Örneğin aşağıdaki gibi iki bölüt vardır denilebilir:



Veya aşağıdaki gibi dört bölüt vardır da denilebilir:



Aslında sorunun cevabı, yani kaç bölüt olduğu problemi hangi seviyede ele aldığımızla ilgili bir durumdur.

4.1. Hiyerarşik Bölütleme

Hiyearşik bölütleme, daha önce bahsedildiği gibi parametrelerin belirsiz olduğu durumlarda (ki bu aslında bu durum çoğu zaman yaşanmaktadır) veya kaç bölüt oluşturulacağının bilinmediği veya duruma göre değiştiği durumlarda bölütleme yapmak yerine bölütleme için hiyerarşi oluşturmayı söyler. Hiyerarşik bölütleme yöntemlerini iki grupta incelemek mümkündür:

- Aşağıdan yukarıya (agglomerative)
- Yukarıdan aşağıya (divisive)

Aşağıdan yukarıya yaklaşımda, çizgedeki veya sosyal ağdaki her bir bireyle başlanarak bu bireyler arasındaki ilişkilere göre önce ikili daha sonra daha fazla sayıdaki ilişkileri içeren bağlantılara kurularak topluluklar inşa edilmektedir. Yukarıdan aşağı yaklaşımda ise topluluk bir bütün olarak ele alınmakta ve alt gruplara bölünmektedir. Örneğin önce iki gruba sonra her grup kendi içinde daha alt gruplara bölünerek ilerlenmektedir.

Ne yazık ki bölütleme işlemi her zaman bu örneklerde ele alındaığı kadar basit veri yapıları üzerinde çalışmamaktadır. Veri yapısı karmaşıklaştıkça bölütleme işleminin zorluğu da artmaktadır. Her ne kadar bu konuda çok sayıda çalışma olsa da (KHAN & LUO, 2005) bu konu önemini giderek arttıran ve henüz tam sonuca ulaşmamış çalışma alanlarından birisidir.

Bu yazı kapsamında hiyerarşik bölütleme için farklı yaklaşımların nasıl kullanılacağı anlatılacaktır. Bunlardan ilki klasik k-ortalama algoritmasının nasıl çalıştığıdır.

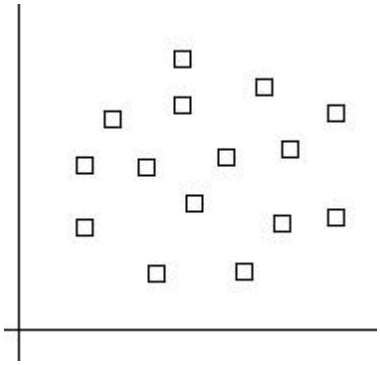
4.2. K-Ortalama Algoritması ve Hiyerarşik K-Ortalama Algoritması

Kullanılan matematiksel yöntem her sınıf için merkez belirlenen noktaya uzaklığa (aynı zamanda bu hata miktarıdır) göre yeni kümelerin yerleştirilmesidir.

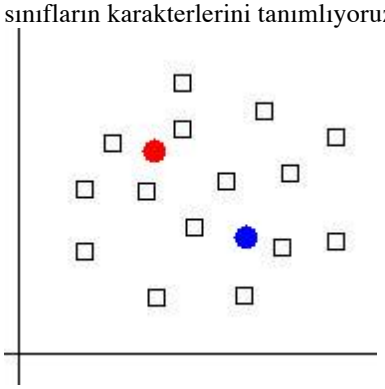
Algoritma temel olarak 4 aşamadan oluşur:

- Küme merkezlerinin belirlenmesi
- Merkez dışındaki örneklerin mesafelerine göre sınıflandırılması
- Yapılan sınıflandırmaya göre yeni merkezlerin belirlenmesi (veya eski merkezlerin yeni merkeze kaydırılması)
- Kararlı hale (stable state) gelinene kadar 2. ve 3. adımların tekrarlanması

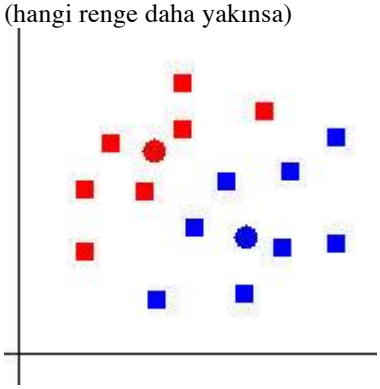
Çalışmayı daha net anlamak için aşağıdaki örnek uzaya dağılmış olan örnekleri inceleyelim:



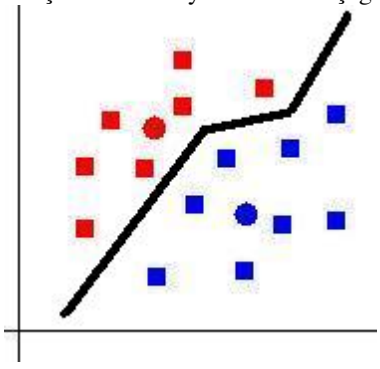
Yukarıda verilen ve uzayda koordinatları kodlanmış olan örnekler için iki adet hedef küme tanımlıyoruz. (iki sınıf ve bu sınıfların karakterlerini tanımlıyoruz)



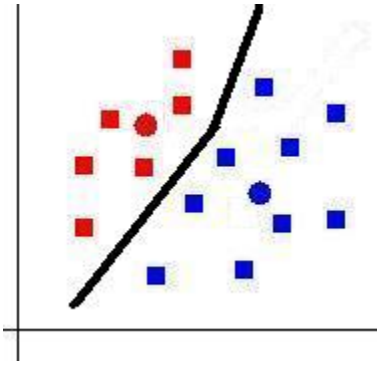
Bu sınıf tanımlarına uzaklıklarına göre (örneğin öklit mesafesi (euclid distance)) bütün örneklerimizi sınıflandırıyoruz. (hangi renge daha yakınsa)



Oluşan sınıfları ayıran bir hat aşağıdaki şekilde çizilebilir:



Daha önceden sınıflandırdığımız örneklerin merkezlerini buluyoruz. (yuvarlak ile gösterilen ve sınıf karakteristiğini temsil eden ilk örneklerin yerini değiştirmek olarak da düşünülebilir)



Merkezleri hareket ettirdikten sonra örneklerden bazıları yeni merkezlere daha yakın olabilir. Buna göre örnek kümelerimizin sınıflandırılmasını güncelliyoruz.

Yukarıda son hali gösterilen k-means algoritmasında yeni merkezler ve her örneğin hangi sınıfa girdiği bulunmuştur.

Hiyerarşik K-Ortalama algoritması ise k-ortalama algoritmasının tek eleman kalana kadar devamlı uygulanması olarak düşünülebilir. Örneğin yukarıdaki böltülemesi tamamlanan ve iki grup bulunan örneklerin daha alt iki kümeye ve daha alt iki kümeye bölünerek sonunda her bölütte tek eleman kalana kadar devam etmesi ile hiyerarşik k-ortalama algoritması çalıştırılmış olur. K-ortalama yöntemi ve hiyerarşik k-ortalama yöntemi, yaklaşım olarak yukarıdan aşağıya (divisve) yaklaşımlardır ve bir sosyal ağı bütün olarak kabul ederek başlar ve her aşamada parçalara bölerek ilerlerler.

4.3. Girvan Newman Algoritması

Girvan-Newman algoritması, karmaşık sistemlerde topluluk yapılarını bulmak için sosyal ağlar dahil olmak üzere pek çok ağ yapısı üzerinde kullanılmaktadır. Algoritma, Michelle Girvan ve Mark Newman tarafından geliştirilmiş olup bu kişilerin soy ismi ile anılmaktadır.

Algoritma oldukça basit bir iddia üzerine kuruludur. Bir toplulukta birden fazla grup olduğunu kabul edelim. Buna göre grupların kendi içlerindeki ilişkileri yoğun, gruplar arası ilişkiler ise daha seyrek olur.

Örneğin telefon etme sıklıklarını ele alalım. Bir ülkedeki grupları tanımlarken aile kavramını kullanırsak, bir toplulukta aile içi telefonlaşmalar, aileler arası telefonlaşmalara göre daha sık olacaktır. Benzer şekilde grup tanımımız şirketler olursa bu durumda Girvan-Newman algoritması bize şirket içindeki bireylerin birbiri ile telefon görüşmelerinin, şirketler arası telefon görüşmelerine göre daha sık olduğunu söyler.

Şimdi soru, bu grupların nasıl bulunacağıdır. Örneğin size bir ülkedeki bütün telefon görüşmelerinin verildiğini ve bu görüşme sıklıklarından grupları çıkarmanız istendiğini düşünelim. Böyle bir problemin çözümünde kullanılacak bazı yollar aşağıdaki şekilde sıralanabilir:

- Hiyerarşik bölütleme (hierarchical clustering)
- k-klik araması (k-clique percolation)
- Şekil parçalaması (graph partitioning)

Yukarıdaki bu yöntemlerin üzerinden geçip Girvan-Newman algoritmasının farkını açıklamaya çalışalım.

Örneğin hiyerarşik bölütleme (hierarchical clustering) yaklaşımında, bağlantılar öncelikle boş bir ağda tanımlanır ve ilk değer olarak 0 atanır. Ardından ilişkinin derecesine göre bu ilişkilere ağırlık verilir. En yüksek ağırlığa sahipten en zayıf ağırlığa doğru işlem yapılır. Genelde en yüksek ağırlığa sahip olan birey, merkezi olma özelliği gösterir. Ayrıca ilişki tipi zayıfladıkça hata miktarı da artar. Örneğin tek bir bağlantısı olan bireyin o grubun üyesi olduğunu söylemek çok da doğru olmayabilir.

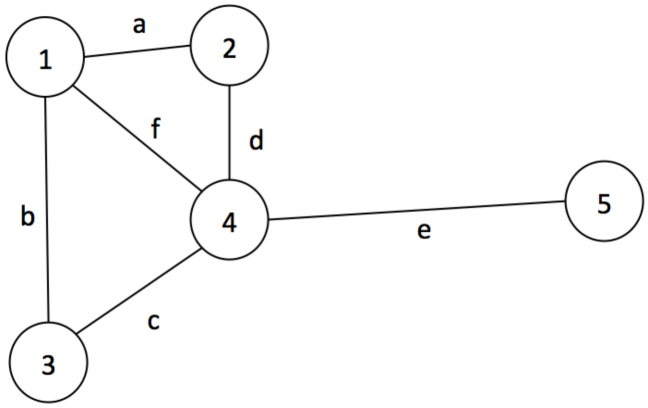
Girvan-Newman algoritması yukarıda anlatılan hiyerarşik

bölütlemenin tam tersini yapar. Merkezi bireylerden başlamak yerine, kenarda kalmış ve düşük ağırlığa sahip bağlantıları bulunan bireylerden başlar. Bu kenarda kalan bireyleri eleyerek çalışmaya devam eder. Bu işlem de hiyerarşik bölütlemenin tam tersidir çünkü hiyerarşik bölütlemede boş bir ağdan başlanarak her adımda yeni bireyler sisteme eklenmektedir, Girvan Newman algoritmasında ise her adımda ağdan bir birey çıkarılır.

Telefon görüşmesi örneğine dönecek olursak, hiyerarşik bölütleme yaklaşımında, ilk başta hiç birey yokken boş bir ağ ile başlanır ve bu ağa en çok görüşme yapan bireyler eklenerek algoritma ilerler. Girvan-Newman algoritmasında ise önce bütün bireylerin kimi aradığı ve kaç görüşme yaptığı gibi bilgilerin tutulduğu bir ağ ile başlanır, bu ağdan her adımda en az görüşme yapan veya en az kişi ile görüşme yapan birey elenerek ilerlenir.

Girvan-Newman algoritmasının getirdiği bir yenilik de orta birey değeri hesabındadır (vertex betweenness). Orta birey (vertex betweenness) kavramı, merkezilik adına uzun süre incelenmiş bir durumdur. Herhangi bir i bireyi için, bu bireyin orta bireylik değeri, bu birey üzerinden geçen en kısa yolların sayısıdır. Yani ağımızda (network) tanımla n adet düğüm için, bu düğümlerin ikili olarak ele alınması halinde, aralarındaki en kısa yollardan (shortest paths) kaç tanesi bu i bireyi üzerinden geçiyorsa, i bireyinin orta bireylik değeri budur.

Örneğin aşağıdaki şekil için durumu inceleyelim:



Şekildeki her düğüm ikilisi için en kısa yolun geçtiği düğümlerin listesini yazıyorum. Burada her kenar uzunluğunu 1 kabul ediyorum. Yani ağırlıklı bir şekil (weighted graph) olarak kabul edilirse, ağırlık değerleri her bağlantıda 1 olacak.

1-2 : {1,2}
1-3 : {1,3}
1-4 : {1,4}
1-5 : {1,4,5}
2-3 : {2,1,3} veya {2,4,3}
2-4 : {2,4}
2-5 : {2,4,5}
3-4 : {3,4}
3-5 : {3,4,5}
4-5 : {4,5}

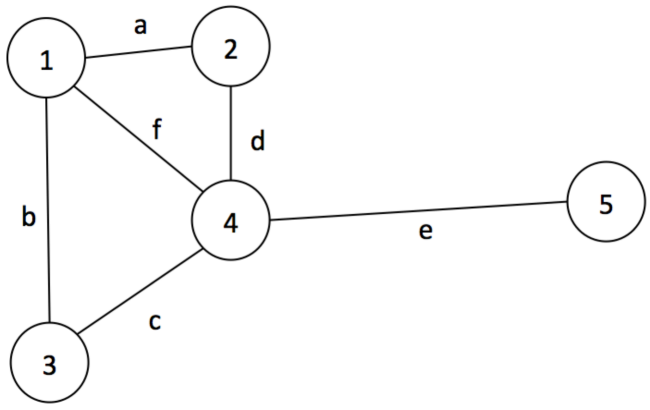
Yukarıdaki listelerde her düğümün (node) kaçar kere geçtiğini listeleyecek olursak:

1 : 5 kere
2 : 4 kere
3 : 4 kere
4 : 7 veya 8 kere
5 : 4 kere

Listeden anlaşılacağı üzere en yüksek orta bireylik değerine sahip düğüm 4 olarak bulunmuştur.

Girvan-Newman algoritması, bu orta bireylik değerini biraz değiştirir ve orta bağlantı (edge betweenness) kavramını getirir. Buna göre iki bireyi bağlayan herhangi bir bağlantı için üzerinden geçen en kısa yol sayısı sayılarak bu bağlantının orta bağlantı değeri hesaplanır.

Bu değerin nasıl hesaplandığını, yine aynı örnek üzerinden kenarlara (edges) isim vererek görelim:



1-2 : {a}
1-3 : {b}
1-4 : {f}
1-5 : {f,e}
2-3 : {a,b} veya {d,c}
2-4 : {d}
2-5 : {d,e}
3-4 : {c}
3-5 : {c,e}
4-5 : {e}

Benzer şekilde her kenarın kaçar kere geçtiğini sayalım:

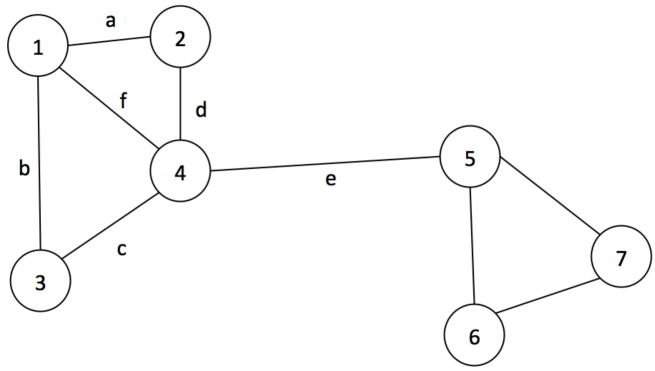
a: 2
b: 1 veya 2
c: 2 veya 3
d: 2 veya 3
e: 4

Görüldüğü üzere en yoğun bağlantımız e kenarıdır.

Şayet birden fazla bağlantı aynı değerlerle en kısa yol sayısına sahipse, bu bağlantılar birbiri ile ilişkilendirilerek daha uzun bir bağlantı elde edilebilir.

Örneğimize dönecek olursak e kenarı olmasaydı, c ve d kenarları eşit yoğunlukta olduğu için bu iki kenarı birleştirerek tek bir kenar gibi düşünebilirdik.

Girvan-Newman algoritması, tam bu noktada devreye girerek en yüksek değere sahip kenarları sistemden çıkarmaya başlar (okuyucu, en yüksek değere sahip kenarın aslında sistemin en kenarında kalmış bireyler olduğunu düşünerek bulabilir, nitekim örnekteki 5. düğüm bu şekilde bir düğümdür). Böylelikle şekildeki bazı bireyler veya birey grupları sistemden kopmaya başlar. Örneğimiz aşağıdaki şekilde olsaydı:

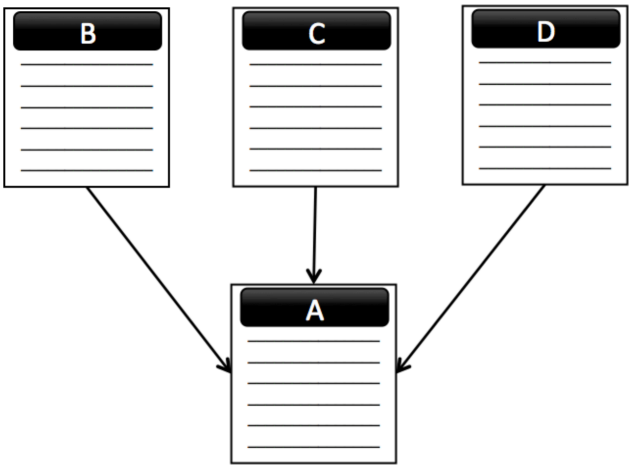


Bahsi geçen kopma, bu şekildeki e kenarının kaldırılması durumunda sistemde iki ayrı grup olmasıdır. İşte Girvan-Newman algoritması da bütün bu ağ yapısında birbiri ile yoğun ilişkisi olan iki grubu bulmayı amaçlamış ve bulmuştur. En baştaki örneğimize dönersek ve bu ağdaki

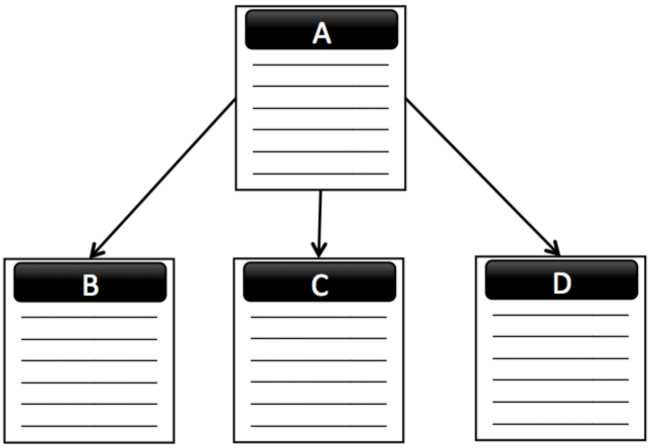
telefon görüşmeleri ise (ve herkes birbirini 1 kere aramış dolayısıyla bütün bağlantılarda 1 ağırlığına sahip olunmuş kabul edileiblir) iki ayrı aileyi veya şirketi ağda bulmuş oluruz.
Algoritmanın adımları aşağıdaki şekilde sıralanabilir:
Bütün kenarlar için orta bağlantı değeri hesaplanır
En yüksek orta bağlantı değerine sahip bağlantı (veya aynı değere sahip birden fazla bağlantı varsa bunların tamamı) kaldırılır
Kalıdrma işleminden sonra bütün orta bağlantı değerleri yeniden hesaplanır.
Yukarıdaki 2. ve 3. adımlar hiçbir kenar kalmayana kadar tekrar edilir.
Girvan-Newman algoritmasına göre, şayet iki grubu bağlayan birden fazla bağlantı varsa, bu bağlantıların aynı anda en yüksek orta bağlantı değerine sahip olması beklenemez. Ancak bunlardan en az bir tanesinin en yüksek orta bağlantı değeri olduğu kesindir. Ayrıca bu en yüksek bağlantı kaldırıldıktan sonra bu iki grubu bağlayan bağlardan geri kalanlarından en az birisi yine en yüksek bağlantı değerini mutlaka taşır.
Girvan-Newman algoritmasının sonucu bir öbekağacı (dendrogram) olarak düşünülebilir. Girvan-Newman algoritması çalıştıkça her dönüşte (iteration) ağacı yukarıdan aşağıya (top-down) doğru oluşturmaktadır ve bu ağırlıkağacının yaprakları aslında ağdaki bireylerden oluşmaktadır.

4.4. HITS ve HUBS Algoritmaları

Sosyal ağlardan çok önce, web içerikleri üzerinden geliştirilen bazı algoritmalar ile web sayfalarının birbirine olan bağlantısını çözerek aslında bir çizge teoremi uygulaması olan HITS ve HUBS algoritmalarından bahsetmekte yarar vardır. Bu algoritmalar günümüzdeki çoğu arama motorunun ve daha sonraları sosyal ağların, wiki'lerin ve blog'ların analiz edilmesinde aktif olarak kullanılmıştır.
Günümüzde ise aktif olarak arama motorlarında, veri madenciliğinde ve metin madenciliği gibi konularda sıkça kullanılmaktadırlar. HITS algoritması, Hyperlink Included Text Search kelimelerinin baş harflerinden oluşmaktaır ve Türkçeye bağlantı dahil metin araması şeklinde çevrilebilir. Anlatılmak istenen metin araması sırasında metinler arasındaki bağlantıların da arama sonucuna etki etmesidir.
Algoritmalar oldukça basit bir şekilde birbirine atıfta bulunan (refere eden) metinlerin skorlanması için geliştirilmiştir.



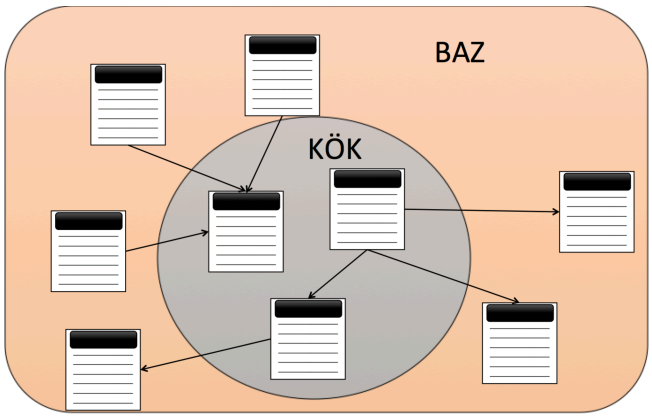
Örneğin, yukarıdaki şekilde B,C ve D dokümanlarının A dokümanına atıfta bulunması hali temsil edilmiştir. Buna karşılık ikinci resimde ise bir A dokümanından atıfta bulunulmuş dokümanlar gösterilmiştir.



İlk resimdeki durumda HITS durumundan bahsedilebilir. HITS algoritması aslında HUBS and Authorities ismi ile de anılmaktadır. Hub kelimesi ingilizcede grafikler üzerindeki bağlantı noktaları için kullanılmaktadır ve İngilizce, Türkçedeki " göbek" kelimesine benzer bir anlam ifade etmektedir. Authorities is Türkçede de otorite olarak geçmektedir ve eski dilde "sulta" veya " kudret" gibi kelimeler ile karşılanmaktadır. Bu anlamda bir metnin sultası veya kudreti, içeriğinin sirayet ettiği metinler ile ölçülebilir. Yukarıdaki ikinci resim de bunu göstermektedir ve A metninin sultasında olan kudret etkisinde olan diğer metinleri göstermektedir.

Gelelim algoritmaya.

Algoritma iki küme üzerinden çalışır. Birincisi kök diğeri ise baz kümeleridir. Kök kümesinde normal metin aramasına göre bulunan sonuçların belirli bir kısmı barındırılır. Örneğin limit olarak 10 belirledik diyelim. Arama sonucunda çıkan yüzlerce sonuçtan en iyi 10 tanesini alıyoruz (buradaki en iyi kavramı herhangi bir skorlamaya göre yapılabilir ve şu anda anlatılan konunun dışındadır). Ardından bu en iyi 10 sonucu kök kümesi ilan edip bu küme ile atıf ilişkisi olan (atıf edilmiş ve aıf edinilen) bütün dokümanları içeren ikinci bir baz kümesi oluşturuyoruz.



Yukarıdaki şekilde bu durumu temsili olarak resmeden iki küme görülmektedir. Buna göre baz kümesi, kök kümesinin üst kümesi (superset) olarak kabul edilmelidir.
Burada yapılacak hesaplama iki adımdan oluşur. İlk adımdan metnin sultasındaki veya kudret etkisindeki diğer metinleri sayıyoruz:

$$S(p) = \sum_{q \in B \mid q \to p} G(q)$$

Burada, metnin sultası (kudreti) hesaplanırken kendisine atıfta bulunan her metin için (göbek) bir değer eklenmektedir. Toplam sembolünde, p metnine atıfta bulunan her q metni için toplamaya bir değer ekleneceği ve bu metinlerin Baz kümesi elemanı olacağı ifade edilmiştir.
Gelelim HITS hesabına:

$$H(p) = \sum_{q \in B \mid p \to q} S(q)$$

Bir önceki adımda hesaplanan sulta değeri için yine aynı metin kümesindeki her bir atıf sayısı sayılır ve kaç tane olduğu HITS değeri olarak tutulur.
Özetlenecek olursa, öncelikle metinler iki kümeye ayrılır. Birincisi metin araması sonucundaki kümenin belirli sayıdaki elemanı (kök küme), ikincisi ise bu metinlerle atıf ilişkisi olan metinlerdir (baz küme). İki adımda HITS değeri hesaplanır. Birinci adımda metnin etki ettiği metinlerin sayısı sayılır ve ikinci adımda metne etki eden metinlerin etki değerleri toplanır.

HITS algoritması, şu anda google tarafından da kullanılan pagerank algoritmasının öncüsü olarak görülebilir ve pagerank algoritmasına göre biraz daha ilkel kabul edilebilir.

5. Fikir Madenciliği (Opinion Mining)

Sosyal ağlardaki problemlerden birisi de fikir veya kanaat madenciliği olarak literatürde geçen problemdir. Problem literatürdeki konumu itibariyle duygu analizi (sentimental analysis) altında geçmektedir. Buna göre bir sosyal medya bilgisinin (mesaj, paylaşım, duvar yazısı, haber v.s.) taşımış olduğu fikri anlambilimsel olarak göstermek için yapılan çalışmaya fikir madenciliği denir. Fikir madenciliği için en önemli kriterlerden birisi, fikir veya kanaat oluşumunun bir topluluk üzerinde inceleniyor olmasıdır. Özel olarak seçilmiş uzmanların fikirleri alınmadığı sürece fikir madenciliğinin ulaşmak istediği sonuç, bir topluluktaki bütün bireylerin fikirlerini anlayabilmektir. Ne yazık ki bütün bireyler ulaşmanın imkansızlığı yüzünden genelde bu işlem örneklemelerle yapılmaktadır. Örneğin anket çalışmaları bu örneklemelerden birisidir.
Fikir madenciliği sırasında unutulmaması gereken konulardan birisi de fikirlerin kişisel olduğudur. Yani doğru veya yanlış bir fikir aranmaz, fikir madenciliği mevcut durumun tespitine çalışır.
Genelde literatürde ilk uygulamaları ve halen üzerinde en çok çalışılan uygulama duygusal kutupsallıktır (sentimental polarity). Bu problemde metinleri duygusal olarak olumlu veya olumsuz şeklinde iki gruba ayrılmaya çalışılır. Örneğin bir siyasi parti, bir futbol takımı veya bir televizyon programı hakkında sosyal medyada yapılan yorumların tamamını bir bilgisayarın inceleyerek doğal dil işleme (Seker S. E., Event Ordering for Turkish Natural Language Texts, 2010) ve metin madenciliği teknikleri (Seker, Mert, Al-Naami, Ozalp, & Ayan, 2014)ile bu konularda yapılan yorumların olumlu veya olumsuz olarak sınıflandırılması mümkündür.
Fikir madenciliği için çok farklı yöntemler geliştirilmiştir. Örneğin kelimelerin olumlu veya olumsuz olarak ayrılması ve yorumlarda geçen kelimelerin sayılarına göre yorumların olumlu veya olumsuz olarak sınıflandırılması ilk ve en basit yöntemlerden birisidir. Ancak alaycı yorumlar düşünüldüğünde bu yöntemin başarı oranının göreceli olarak düşük olduğu görülecektir. Bu yüzden kendi kendini geliştiren ve kelimelerin anlamlarını kendisi bulan daha zeki yöntemler geliştirilmiştir.
Duygu analizi ve fikir madenciliğinin dayandığı ilk sınıflandırma ve metin madenciliği teknikleri aslında iki sınıflı basit problemler olarak görülebilir. Örneğin uzun yıllar çok benzer metin madenciliği teknikleri kullanılarak e-postaların istenmeyen veya zararlı e-postalar veya zararsız e-postalar olarak iki sınıfa ayrılması üzerinde çalışıldı (Masud, Khan, & Al-Shaer, 2006).
Günümüzde problemler daha karmaşık haller almaktadır. Örneğin bir filmi izleyen iki farklı kişinin görüşleri arasındaki farkın ne kadarının filim oyuncularından kaynaklandığı, filmin yönetmeninin veya senaryo yazarının izleyicilerin görüşüne etkisi gibi daha karmaşık çıkarımlar artık sorgulanabilmektedir.
Veya bir internet satış mağazasında ürünler hakkında yapılan yorumların ne kadarı satıcının etkisi ile ne kadarı kargo şirketinden ne kadarı ürünün kendisinden kaynaklanmaktadır gibi sorular artık birer çalışma alanı olarak belirmektedir.
Fikir madenciliği çalışmalarında aşağıdaki sorulara cevap aranabilir:

- Öznellik sınıflandırması (subjectivity classification: verilen bir metnin/cümlenin herhangi bir fikir içerip içermediğine bakılır.
- Duygusal sınıflandırma: verilen bir cümlenin olumlu/olumsuz veya nötr olması durumunu bulmaya çalışır.
- Fikrin yardımcı olma ihtimali: Herhangi bir fikir içeren metnin kişilere ne kadar yardımcı olacağını bulmaya çalışır (daha çok alışveriş sitelerindeki yorumlar için kullanılmaktadır).
- İstenmeyen fikir taraması : Bir fikrin kötü amaçla yazılması durumunu tespit için kullanılır. Örneğin bir yorum yazısının reklam içermesi, okuyucuyu yönlendirici olması gibi durumları bulmaya çalışır.
- Fikir özetleme: Çok sayıdaki fikrin veya uzun bir fikir metninin kısa bir şekilde ifade edilmesi için çalışır. Örneğin anahtar önemdeki cümlelerin çıkarılması, ürün veya bakış açısına göre sınıflandırılması gibi problemleri bulunmaktadır.
- Karşılaştırmalı fikirlerin çıkarılması. Örneğin iki veya daha fazla ürün veya kavramın karşılaştırıldığı fikirlerde, metnin hangi kavramı hangi açıdan karşılaştırdığı ve bu karşılaştırmaya göre kavramların görece pozisyonlarını belirlemekte kullanılırlar.

Fikir madenciliği çalışmaları üç farklı açıdan incelenebilir, bunlar:

- Bakış tabanlı (veya özellik tabanlı) fikir madenciliği teknikleri
- Frekans veya ilişki tabanlı fikir madenciliği teknikleri
- Model tabanlı fikir madenciliği teknikleri

Olarak sayılabilir. Yazının devamında her bir yaklaşım ayrı bir alt başlıkta incelenecektir.

5.1. Bakış Tabanlı / Özellik Tabanlı Fikir Madenciliği

Bu yaklaşımda fikir madenciliği ilgili metin üzerinden hangi bakış açısında fikir beyan edildiğini bulmaya çalışır. Örneğin bir fotoğraf makinesi ile ilgili yapılmış bir yorumda, yorumu yazan kişi amatör bir fotoğrafçı veya profesyonel bir fotoğrafçı olabilir, dolayısıyla kişinin yorumunun kimin işine daha çok yarayacağının anlaşılabilmesi için yorumu yazan kişinin hangi bakış açısıyla yorumu yazdığının anlaşılması önemlidir. Benzer şekilde fotoğraf makinesi örneğinden devam edilirse, fotoğraf makinesinin malzemesinin kalitesi, çektiği resimlerin kalitesi, renk ayarı, fotoğrafçının eline uyumu (tabi bu durumda ince veya şişman parmaklı bir fotoğrafçı oluşu), kasasının rengi, hafifliği, boyutları,

üzerindeki yazılımın kullanım kolaylığı, diğer objektif üreticileri ile uyumluluğu gibi onlarca farklı açıdan incelenmesi mümkündür. Kısaca bir fotoğraf makinesi hakkında olumlu veya olumsuz yorum yapılması duygusal kutupsallık açısından önemli olmakla birlikte hangi bakış açısında göre olumlu veya olumsuz olduğunu inceleyen çalışmalara bakış tabanlı (aspect-based) fikir madenciliği ismi verilmektedir.

Bakış tabanlı fikir madenciliğindeki önemli konulardan birisi de karşılaştırmalı fikirlerin çıkarılmasıdır. Örneğin fotoğraf makinesi hakkında yapılan bir yorumun, aynı fotoğraf makinesinin bir önceki yıl çıkan versiyonuna göre, veya rakip firmanın benzer ürününe göre veya bir üst sınıftaki fotoğraf makinesine göre veya cep telefonunun kamerasına göre olumlu veya olumsuz olmasının yanında hangi açılardan olumlu veya olumsuz olduğunun da incelenmesi gerekir. Örneğin bir fotoğraf makinesini güvenlik kamerasına göre çok daha kolay hareket ettirilebildiğinin söylenmesi ile cep telefonuna göre çok daha kolay hareket ettirilebildiğinin söylenmesi arasında ifade açısından fark vardır.

5.2. Frekans / İlişki Tabanlı Fikir Madenciliği

Metinlerin üzerinden amaca yönelik olarak fikir çıkarımı yapılırken kelime sayısı, isim, sıfat, zarf veya fiil gibi kelimelerin sıklıkları (frekansları) üzerinden fikir madenciliği yapılmasına verilen isimdir. Genelde fikirlerin bu kelimeler ile ifade edildiği kabulüne dayanmaktadır. Örneğin 2011 yılında yapılan bir araştırmada fikirlerin %60-70 gibi önemli bir oranının metindeki isimlere dayandığı bulunmuştur (Liu, 2011). Yine istisnaları olmakla birlikte, bakış tabanlı fikir madenciliğinde de sık tekrar eden isimlerin kişilerin bakışını yansıtmakta olduğu bulunmuştur.

Frekans tabanlı fikir madenciliğinde, literatürde önemli yer tutan çalışmalardan birisi de özellik tabanlı özetleme çalışmalarıdır. Bu çalışmalarda öncelikle isim kelime grupları bulunarak bunlar uzunluklarına, kullanımdaki gerekliliklerine ve olumlu olumsuz kutupsallığına göre sınıflandırılmaktadır. Ardından bu isim kelime gruplarını tanımlayan sıfatlar kelime gruplarına eklenerek sık tekrar etmeyen duyguların elenmesi sağlanmaktadır (Hu & Liu, 2004).

Diğer bir yaklaşım ise belirli şablonların metin içerisinde aranarak belirli sonuçlara ulaşılmasıdır. Örneğin "harika X", "X özelliği ile gelmektedir", "X özelliği bulunmaktadır" gibi kalıplar aranarak X yerine gelen değerlerin birer bakış olarak aday gösterilmesi ve ardından kutupsallık analizleri ile bakış açısına göre fikir madenciliği yapılması mümkündür (Popescu & Etzioni, 2005).

Frekans tabanlı fikir madenciliği ayrıca kelimelerin dilbilimsel özelliklerine göre etiketlenmeleri ile sağlanabilmektedir. Literatürde POS-Tagging olarak geçen kavram Türkçeye konuşmanın bir kısmının etiketlenmesi olarak çevrilebilir ve basitçe bir metindeki kelimelerin isim, sıfat ve hatta sayı sıfatı, özel isim gibi özelliklerine göre etiketlenmesi esasına dayanır. Bu etiketleme sürecinin fikir beyan edecek şekilde genişletilmesi de mümkündür. Örneğin "harika konumda" kelimeleri etiketlenirken "[güçlü][olumlu] konum" veya "yardımsever çalışanlar" kelimeleri etiketlenirken "[duygusal][olumlu]çalışan" şeklinde etiketlenmesi mümkündür. Daha sonra bu etiketlerin üzerinden yapılan sıklık analizleri ile fikir çıkarımı yapılması çok daha kolay hale gelecektir (Baccianella, Esuli, & Sebastiani, 2009).

Literatürde ayrıca frekans ve ilişkiye dayalı fikir madenciliği çalışmalarının sağladığı karşılaştırmalı madencilik imkanları da bulunmaktadır. Örneğin bir sıfatın kuvvetinin belirli bir ölçeğe oturtulması mümkündür.

Mükemmel -> iyi -> ortalama -> zayıf -> kötü

Gibi bir sıralamada hepsi sıfat olmakla birlikte kuvvet dereceleri değişmekte ve dolayısıyla çıkarılan fikir açısından değişik oranlarda etki etmektedir (Moghaddam & Ester, 2010).

Genel olarak frekans tabanlı yaklaşımların en büyük avantajı, uygulamadaki kolaylık ve çalışma sürelerindeki yüksek başarılarıdır. Bununla birlikte hata miktarları çok yüksek olup genelde elle müdahale edilerek ince ayarlarının yapılması gerekmektedir.

5.3. Model Tabanlı Fikir Madenciliği Teknikleri

Model tabanlı fikir madenciliği yöntemleri daha önceden etiketlenmiş ve dolayısıyla nasıl bir fikir içerdiği bilinen metinlerden oluşturduğu modelleri etiketlenmemiş metinlere uygulamakta ve bu sayede fikir çıkarımı yapılmamış metinlerden fikir çıkartmaya çalışmaktadır.

Aslında şimdiye kadar bahsi geçen diğer yöntemler de bu açıdan ele alındığında birer model tabanlı yaklaşım olarak kabul edilebilir ancak model tabanlı fikir madenciliğinin en belirgin özelliği gizli Markof Modelleri (HMM) veya koşullu rasgele alanlar (CRF) veya daha genel anlamda yapay sinir ağları (ANN) veya Bayez ağları (Bayesian Networks) gibi istatistiksel modellere dayanıyor olmasıdır. Modeller istatistiksel ağırlıklarını (çarpanlarını) etiketli metinlerde öğrenerek bu ağırlık değerlerine göre etiketsiz metinlerde fikir çıkarımı yapmaktadır.

Örneğin (Lakkaraju, Bhattacharyya, Bhattacharya, & Merugu, 2011) çalışmalarında alt küme yaklaşımını kullanarak yeni bir gizli markof modeli geliştirmiş ve bakış ve duygu seviyesi değerlerin cümledeki konumuna göre ilişkileri üzerinden değişen ağırlıklara göre istatistiksel bir model oluşturmuştur. Benzer bir değerlendirme ise fikirlerin çıkarımının yapıldığı web sitelerinde bulunan fikirlere göre gruplamaya dayalı ve bu gruplar arasındaki ağırlıkları hesaplayan ve yine bir gizli markof modeli olan (Wong, Lam, & Wong, 2008)'un çalışmasıdır.

Model tabanlı çalışmaların önemli bir kısmı ise başlıklara odaklanmaktadır. Bu çalışmaların yapmış olduğu kabul, metinlerde içerilen fikrin başlıkta daha net bir şekilde eyr alacağıdır. Bu yaklaşımların en büyük zorluğu başlıklarda fikir ve duygunun aynı anda bulunuyor olmasıdır. Başlık öncelikli, model tabanlı çalışmaların en bilinen uygulamalarından birisi literatürde kısaca LDA olarak geçen (uzun hali Latent Dirichlet Allocation) ve Türkçeye gizli Dirichlet ayrımı olarak çevrilebilecek olan yöntemdir. Bu yaklaşımda metinler gizli başlıkların karışımlarını istatistiksel bir modele oturtulmaktadır ve bu model metinlerdeki kelime dağılımlarından çıkarılmaktadır (Blei, Ng, & Jordan, 2003).

Örneğin aşağıdaki üç cümleyi ele alalım:

- Ben **balık** ve **sebze yerim**
- *Balıklar evcil hayvanlardır*
- Benim *kedim* **balık yer**

Yukarıdaki üç cümlede iki farklı başlıkta fikir çıkarımı yapılabilir. İlk başlık "yemek" olarak ve ikinci başlık "evcil hayvan" olarak çıkarıldıktan sonra, yukarıdaki cümlelerde kalın harflerle yazılı olanların "yemek" başlığında ve yattık olarak yazılanların ise "evcil hayvan" başlığındaki kelimeler olduğu söylenebilir. Buna göre birinci cümle %100 "yemek" başlığında, ikinci cümle %100 "evcil hayvan" başlığındayken üçüncü cümle %33 "evcil hayvan" ve %66 "yemek" balığı altında kabul edilmelidir.

Bu örnekte görüldüğü gibi LDA iki aşamadan oluşmaktadır, öncelikle metinlerden başlıkların çıkarılması ve ardından da başlıklara göre metinlerin sınıflandırılması. Bu iki aşamadan

ikincisi yani metinlerin başlıklara atanması da iki alt aşamadan oluşmaktadır. İlk alt aşama metindeki her kelimenin geçici olarak bir başlığa atanması ardından metindeki kelimelerin yoğunluğuna göre metnin bir başlığa atanması. Ancak bazı durumlarda bir kelime birden fazla başlığa ait olabilir, bu yüzden LDA tekrarlı şekilde (iterative) doğru başlığı bulmak için istatistiksel modelini güncellemektedir. Örneğin yukarıdaki üç cümlede geçen "balık" kelimesi bir evcil hayvan veya bir yemek olarak kabul edilebilir. Biz doğal dili kullanan insanlar olarak "benim kedim balık yer" cümlesini okuyunca buradaki "balık" kelimesinin bir yemek olduğunu anlayabiliyoruz ancak bilgisayarın bu analizi yapabilmesi için "yemek" fiilinin balığı işaret ettiğini çözümlenmesi gerekir. Doğal dil analizi yapılmadan sadece istatistiksel olarak analiz yapılan durumlarda bunun anlaşılması biraz karmaşık ve vakit alıcı olabilir. Basitçe elimizdeki bir sözlükle kelimeleri arayarak bu kelimelerin dahil olacağı sınıfları bulmak aslında çoğu zaman yanlış sonuçlar doğurmaktadır. Bu yüzden ihtimal olarak bir kelimenin birden fazla sınıfa farklı oranlarda dahil olduğu kabulü yapılmakta ve her tekrarda (iterasyon) daha doğru sınıfa atama yapılmaktadır. Örneğin Türkçede balık kelimesi daha yüksek ihtimalle yiyecek olarak kullanılmakta (diyelim ki %80) ve daha düşük oranda bir evcil hayvan olarak kullanılmaktaysa LDA öncelikle kelimeyi bir yemek olarak sınıflandıracak ancak daha sonra cümledeki diğer kelimelerin kullanımına göre sınıfını değiştirerek evcil hayvan sınıfına alabilecektir (veya tam tersi).

5.4. Kanaat çıkarımında eş görüş kümelemesi

Kanaat çıkarımı (fikir madenciliği) çalışmalarının önemli bir kısmı da şimdiye kadar alt bölümler halinde anlatılan bakış tabanlı, frekans tabanlı veya model tabanlı yaklaşımlar ile çizge teoreminin birlikte kullanılmasından doğar. Buna göre bir sosyal ağda bulunan kişilerin fikirlerinin birbirini etkilemesi veya diğer bir deyişle sosyal ağda fikir yayılımı mümkündür. Aynı fikre sahip kullanıcıların aynı kümede toplandığı ve sosyal ağın aynı fikre sahip kişiler olarak kümelendiği yaklaşımlara eş görüş kümelemesi (homophily) yaklaşımı denilmektedir (McPherson, bM., & Cook, 2001)[1]. Eş görüş çalışmalarını temel kabul ederek farklı amaçlar için kullanan çalışmalar da vardır. Örneğin bir sosyal ağdaki yapının eş demografik özelliklere göre kümelenmesi de mümkündür. Bu çalışmaların amacı bir sosyal ağı, fikir madenciliği yöntemleri ile yaş gruplarına göre veya cinsiyete göre kümelemek olarak görülebilir (Jackson, 2010).

Her ne kadar fikir yayılması ve sosyal ağın farklı şekillerde kümelenmesi çalışmalarında fikir madenciliği teknikleri kullanılsa da problemin en önemli farklılığı hareketli bir ortamda çalışıyor olmasıdır. Örneğin bir kişinin yaptığı basit bir paylaşımla başlayan süreç, çok kısa sürede yüzbinlerce kişiye ulaşmakta ve etkileme süreci başlamaktadır. Bu kişilerden kaçının etkilendiği, kaçının bu fikri beğendiği ve yaymaya başladığının bulunması için problemin hareketli (dinamik) bir yapıda ele alınması gerekir (Kaschesky, Sobkowicz, & Bouchard, 2011).

Problemin dinamik yapısından kaynaklanan özelliği düşünüldüğünde fikir özetleme problemi de farklı bir boyut kazanır. Örneğin bir metinde fikir özetlemesi yapıldıktan sonra elde edilen çok sayıdaki fikrin sadece bir kısmının yayıldığı görülebilir. Bu durumda fikir özetlemesi sadece yayılan fikirlere odaklanarak diğer fikirleri eleme yoluna gidebilir.

6. Sonuç ve Gelecek Çalışmalar

Sosyal ağların artan etkisi ve her geçen gün doğurduğu yeni kullanım alanları, beraberinde bilim insanlarının ilgisini çeken yeni problemler ve yeni imkanlar doğurmaktadır. Bu yeniliklerin takip edilmesi kaçınılmaz olarak sosyal ağ üzerindeki verinin etkili bir şekilde işlenmesi ve takip edilmesi gerekliliğini doğurmakta, aynı zamanda çok farklı arka-planlardan elde edilen yöntemlerin sosyal ağlar üzerinde veri madenciliği çalışmalarına dahil edilmesini sağlamaktadır. Veri madenciliği teknikleri her ne kadar istatistik ve uygulama alanı olarak bilgisayar bilimlerini kapsıyor olsa da gelecekte sosyal ağlar üzerinde kullanılan tekniklerin davranış bilimleri, toplum bilimleri veya işletme gibi çok farklı disiplinlerden besleneceği düşünülmektedir.

7. Kaynaklar

Asur, S., & Huberman, B. (2010). Predicting the future with social network. *Web Intelligence and Intelligent Agent Technology (WIAT)*. IEEE / WIC / ACM.

B., P., & L., L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information Retrieval*, *2*, s. 1-135.

Baccianella, S., Esuli, A., & Sebastiani, F. (2009). Multi-facet rating of product reviews . *Proceeding ECIR '09 Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval* , s. 461 - 472 .

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research* , 993-1022.

Borgatti, S., & Everett, M. (2006). A Graph-Theoretic prespective on centrality. *Social Networks , 28* (4), 466-484.

Chelmis, C., & V.K., P. (2011). Social networking analysis: A stat of the art and the effect of semantics. *Third International Conference on Social Computing, IEEE.* IEEE.

Chou, W., Y.M., H., Beckjort, E., R.P., M., & Hesse, B. (2009). Social media use in the United States: Implications for health communication. *11* (4).

Fortunato, S. (2010). Community detection in Graphs. *Physics Reports , 486* (3), 75-174.

G., A., & A, T. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Knowledge and Data Engineering Transactions* , 734-749.

Ghosh, R., & Lerman, K. (2011). Parameterized centrality metric for network analysis. *PHYSICAL REVIEW E , 83* (6).

Girvan, M., & Newman, M. (2002). Community structure in social and biological network. *Proceedings of the National Academy of Sciences , 99* (12), 7821-7826.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceeding KDD '04 Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* , s. 168-177 .

Jackson, M. O. (2010). *Social and economic networks* . Princeton University Press .

Kaji, N., & Kitsuregawa, M. (2006). Automatic construction of polarity tagged corpus from HTML documents. *COLING/ACL Main Conferense Poster Session.*

[1] Literatürdeki "homophily" kelimesinin tam çevirimi,

Kaschesky, M., Sobkowicz, P., & Bouchard, G. (2011). Opinion Mining in Social network: Modelling, Simulating, and Visualizing Political Opinion Formation in the Web . *The Proceedings of 12th Annual International Conference on Digital Government Research* .

Kaur, G. (2013). Social Netowrk Evaluation criteria and Influence on consumption behaviour. *Young Segment.*

Khaled Al-Naami, S. E. (2014). GISQF: An Efficient Spatial Query Processing System. *Cloud Computing, 2014 IEEE 7th International Conference on* (s. 681-688). IEEE.

KHAN, L., & LUO, F. (2005). HIERARCHICAL CLUSTERING FOR COMPLEX DATA. *International Journal on Artificial Intelligence Tools , 14* (5).

Lakkaraju, H., Bhattacharyya, C., Bhattacharya, I., & Merugu, S. (2011). Exploiting Coherence for the Simultaneous Discovery of Latent Facets and associated Sentiments. *SIAM International Conference on Data Mining (SDM)* , s. 498-509.

Liu, B. (2011). Chapter 11, Opinion Mining and Sentiment Analysis. B. Liu içinde, *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data* (s. 459-526). Springer.

M, N. (2010). *Networks: An Introduction.* Oxford University Press.

Masud, M. M., Khan, L., & Al-Shaer, E. (2006). Email Worm Detection Using Naïve Bayes and Support Vector Machine. *Intelligence and Security Informatics , 3975*, 733-734.

McPherson, bM., S.-L. L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks . *Annual review of sociology* , 415-444.

Moghaddam, S., & Ester, M. (2010). Opinion digger: an unsupervised opinion miner from unstructured product reviews. *Proceeding CIKM '10 Proceedings of the 19th ACM international conference on Information and knowledge management* , s. 1825-1828 .

Popescu, A.-M., & Etzioni, O. (2005). Extracting product features and opinions from reviews . *Proceeding HLT '05 Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing* , s. 339-346 .

Seker, S. E. (2013). *İş Zekası ve Veri Madenciliği.* İstanbul: Cinius.

Seker, S. E. (2015, 6). Cizge Teorisi (Graph Theory). *YBS Ansiklopedi* , 17-29.

Seker, S. E. (2010). Event Ordering for Turkish Natural Language Texts. *CSW-2010 1 st Computer Science Student Workshop* (s. 26-29). Istanbul: Koc University.

Seker, S. E. (2004). *Possible Social Impacts of E-Government: A Case Study of Turkey.* Istanbul Technical University, ESST.

Seker, S. E., & Eryarsoy, E. Generating Digital Reputation Index: A Case Study. *World Conference on Technology, Innovation and Entrepreneurship* . Istanbul.

Seker, S. E., Al-Naami, K., & Khan, L. (2013). Author attribution on streaming data. *Information Reuse and Integration (IRI), 2013 IEEE 14th International Conference on* (s. 497 - 503). San Francisco, CA: IEEE.

Seker, S. E., Cankir, B., & Okur, M. E. (2014). Strategic Competition of Internet Interfaces for XU30 Quoted Companies. *International Journal of Computer and Communication Engineering , 3* (6), 464-470.

Seker, S. E., Mert, C., Al-Naami, K., Ozalp, N., & Ayan, U. (2014). Time Series Analysis on Stock Market for Text Mining Correlation of Economy News. *International Journal of Social Sciences and Humanity Studies , 6* (1), 69 - 91.

Seker, S., & Al-Naami, K. Sentimental analysis on Turkish blogs via ensemble classifier. *Proc. the 2013 International Conference on Data Mining*, (s. 10-16). LA.

Thompson, J. B. (2013). *Media and Modernity: A social Theory of the Media.* John Wiley and Sons.

Wong, T.-L., Lam, W., & Wong, T.-S. (2008). An unsupervised framework for extracting and normalizing product attributes from multiple web sites. *Proceeding SIGIR '08 Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* , s. 35-42 .