

**T.C.
İNÖNÜ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

SOSYAL AĞLARDA TOPLULUK YAPILARININ ANALİZİ

ONUR BOY

**YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ
ANABİLİM DALI**

**MALATYA
Ekim 2012**

Tezin Bařlıđı : Sosyal Ađlarda Topluluk Yapılarının Analizi

Tezi Hazırlayan : Onur BOY

Sınav Tarihi : 16 Ekim 2012

Yukarıda adı geen tez, jürimizce deđerlendirilerek Bilgisayar Mühendisliđi Anabilim Dalında Yüksek Lisans Tezi olarak kabul edilmiştir.

Sınav Jürisi Üyeleri

Do Dr. Ali KARCI İnönü Üniversitesi
(Jüri Başkanı)

Yrd. Do. Dr. Metin ERTÜRKLER İnönü Üniversitesi
(Tez Danışmanı)

Yrd. Do. Dr. Ahmet ÇINAR Fırat Üniversitesi
(Üye)

İnönü Üniversitesi Fen Bilimleri Enstitüsü Onayı

Prof. Dr. Mehmet ALPASLAN
Enstitü Müdürü

ONUR SÖZÜ

Yüksek Lisans Tezi olarak sunduğum "Sosyal Ağlarda Topluluk Yapılarının Analizi" başlıklı bu çalışmanın bilimsel ahlak ve geleneklere aykırı düşecek bir yardıma başvurmaksızın tarafımdan yazıldığını ve yararlandığım bütün kaynakların, hem metin içinde hem de kaynakçada yöntemine uygun biçimde gösterilenlerden oluştuğunu belirtir, bunu onurumla doğrularım.

Onur BOY

ÖZET

Yüksek Lisans Tezi

SOSYAL AĞLARDA TOPLULUK YAPILARININ ANALİZİ

Onur BOY

İnönü Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

70 + vii sayfa

2012

Tez Danışmanı: Yrd. Doç. Dr. Metin ERTÜRKLER

Sosyal ağlar, barındırdıkları sosyal aktörlerin ilişkilerine ait oldukça faydalı bilgiler ihtiva etmektedirler. Bu yapı ve ilişkilerin analiz edilmesi aracılığıyla yakınlıklar, benzeşmeler, eğilimler ve etkileşimler gibi verilere ulaşılarak ağdaki ilişkiler hakkında çeşitli yorumlara veya tahminlere varılabilir. Özellikle internet teknolojisinde yaşanan teknik ve kültürel gelişmeler ile sosyal ağlara ilişkin veriler somutlaşmış ve ölçülebilir hale gelmiştir. İnternet üzerinde gerçekleşen sosyal iletişim sonucu oluşan veri yığını çok büyük boyutlara ulaşmaktadır. Söz konusu devasa boyuttaki veri yığını içerisinde anlamlı bilginin çıkarılabilmesi için “Web Madenciliği” teknikleri kullanılmaktadır. Dijital sosyal ağların web madenciliği teknikleri kullanılarak analiz edilmesi ve bu sayede akademik, ticari, sosyolojik vb. alanlarda kıymetli verilere ulaşılması önemli bir çalışma konusu haline gelmiştir.

Bu çalışmada genel hatları ile sosyal ağların web madenciliği teknikleri ile analizine ilişkin yöntemler incelenmiştir. Web madenciliği tekniklerinden biri olan “Web Yapı Madenciliği” ile akademik yayınların içerikten bağımsız olarak benzerliklerinin tahmin edilebilmesi amacıyla literatürde mevcut olan ortak atıf analizi metotlarından farklı bir yaklaşım sunulmuş ve bibliyografik veri seti üzerinden elde edilen sonuçlar mevcut yöntemler ile kıyaslanmıştır.

ANAHTAR KELİMELER: Sosyal Ağ Analizi, Web Madenciliği, Atıf Analizi.

ABSTRACT

Master of Science Thesis

ANALYSING THE COMMUNITY STRUCTURE ON SOCIAL NETWORKS

Onur BOY

İnönü University
Graduate School of Natural and Applied Sciences
Computer Engineering Department

70 + vii pages

2012

Supervisor: Assistant Professor Metin ERTÜRKLER

Social Networks contain useful knowledge about the relationships between actors that they include. By the help of the analysis of these relationships one can determine the relations such as closeness, similarity, trends and interactions and can have some precious estimations and comments about the network. Especially, the technical and cultural developments of internet technologies make the communication data on social networks concrete and computable. The data which is formed as a result of social communication on the Internet reaching huge amounts. Usage of Web Mining Techniques are popular to resolve the useful information from that chaotic data pile. Analyzing digital social networks with the help of web mining techniques and reaching valuable academical, commercial and sociological knowledge has been a very remarkable subject of the study.

In this study, in general terms about the analysis of social networks by the help of web mining techniques and methods are studied. Using "Web Structure Mining", which is one of the web mining techniques, a different approach to the common methods of citation analysis is presented in order to predict the similarity between academical documents regardless of content similarity. The results obtained via the bibliographic data set are compared with existing methods.

KEYWORDS: Social Network Analysis, Web Mining, Citation Analysis.

TEŞEKKÜR

Bu çalışmamda benden desteğini esirgemeyen ve her konuda yardımcı olan tez danışman hocam Yrd. Doç. Dr. Metin ERTÜRKLERE' e ve tüm lisansüstü eğitim süreci boyunca bize her konuda destek olan Bilgisayar Mühendisliği Bölüm Başkanı Doç.Dr. Ali KARCI hocama teşekkürlerimi sunarım.

Ayrıca bu çalışmayı, tüm süreç boyunca sonuna kadar yanımda olan ve motivasyonumu kaybettiğim zamanlarda hep destek sağlayan eşim Pınar'a ve dünyaya gözlerini henüz açmış olan biricik kızımız Derin'e armağan ediyorum.

İÇİNDEKİLER

ÖZET.....	i
ABSTRACT.....	ii
TEŞEKKÜR.....	iii
İÇİNDEKİLER.....	iv
ŞEKİLLER DİZİNİ.....	vi
ÇİZELGELER DİZİNİ	vii
SİMGELER VE KISALTMALAR.....	viii
1. GİRİŞ.....	1
2. KURAMSAL TEMELLER.....	5
2.1. Sosyal Ağın Tanımı ve Sosyal Ağ Analizi Kavramı.....	5
2.1.1. Sosyal ağ nedir?.....	6
2.1.2. Sosyal ağ analizi kavramı.....	7
2.2. Sosyal Ağ Analizi Yöntemleri.....	9
2.2.1. Örüntü bulma yöntemleri.....	10
2.2.2. Çizge teorisi.....	10
2.2.3. Sosyal ağ yapılarına ilişkin ölçütler.....	13
2.3. Çizgelerde Topluluk Tespiti.....	19
2.3.1. Geleneksel yöntemler.....	20
2.3.2. Bölütlemeli algoritmalar.....	22
2.3.3. Modülerite esaslı yöntemler.....	23
2.3.4. Dinamik algoritmalar.....	23
2.3.5. Diğer yöntemler.....	24
2.4. Web Madenciliği Teknikleri.....	24
2.4.1. Web madenciliği tekniklerinin ortak işlem basamakları.....	26
2.4.2. Web madenciliği yöntemleri.....	27
2.5. Bibliyometri.....	30
3. MATERYAL ve YÖNTEM.....	31
3.1. Ortak Atıf Analizi İle İçerikten Bağımsız Olarak Benzerlik Tahmini.....	31
3.2. Mevcut Yöntemler.....	33

3.2.1. Bibliyografik eşleştirme (biblio-coupling).....	34
3.2.2. Müşterek atıf analizi (co-citation).....	35
3.2.3. Amsler Yöntemi.....	37
3.2.4. Diğer Yöntemler.....	37
3.3. Problemin Tanımlanması.....	38
3.4. Önerilen Yöntem.....	43
4. ARAŞTIRMA BULGULARI.....	50
4.1. İkili Benzerlik Tespiti.....	50
4.2. Gruplama Başarısı.....	51
5. TARTIŞMA VE SONUÇ.....	54
6. KAYNAKLAR	56
EKLER	64
ÖZGEÇMİŞ.....	70

ŞEKİLLER DİZİNİ

Şekil 2. 1: A.B.D. İnternet Ana Dağıtım Altyapısı.....	5
Şekil 2. 2: Sosyal Ağ Yapısı Ve Barındırdığı Düğümler	6
Şekil 2. 3 Örnek Bir Çizge ve Komşuluk Matrisi.....	12
Şekil 2. 4 Zachary'nin Karate Kulübü Sosyal Ağı.....	13
Şekil 2. 5 Derece Merkeziliği Örneği	15
Şekil 2. 6 Özvektör Merkeziliği Örneği	16
Şekil 2. 7 Arasındalık Merkeziliği Örneği.....	17
Şekil 2. 8 Yakınlalık Merkeziliği Örneği	18
Şekil 2. 9 Kümelenme Katsayısı Örneği	18
Şekil 2. 10 Çizge Bölütleme	20
Şekil 2. 11 Hiyerarşik Gruplama	21
Şekil 2. 12 K-means Yöntemi	22
Şekil 2. 13 Web Madenciliğinin Veri Kaynakları [Baykal-coskun]	25
Şekil 2. 14 2008 Yılı ABD Ortalama İnternet Trafiği	26
Şekil 2. 15 Web Madenciliği Yöntemleri	27
Şekil 2. 16 Yönlü Çizgeler ve Komşuluk Matrisleri	29
Şekil 3. 1 Makaleler ve Komşuluk Matrisi	33
Şekil 3. 2 Bibliyografik Eşleştirme.....	34
Şekil 3. 3 Müşterek Atıf Analizi	35
Şekil 3. 4 Komşuluk (solda) ve Müşterek Atıf (sağda) Matrisleri Arasındaki İlişki.....	36
Şekil 3. 5 Normalize Edilmiş S Matrisi ($\alpha=60$)	37
Şekil 3. 6 Zamana Bağlı Ortak Atıf Kıyaslaması.....	39
Şekil 3. 7 Bibliyografik Eşleştirmede Zaman Kıyaslaması.....	40
Şekil 3. 8 Müşterek Atıf Analizinde Zaman Kıyaslaması	40
Şekil 3. 9 Ortak Atıflamada Farklı Atıfların Rolü	42
Şekil 3. 10 İki Yayın Arası Doğrudan Atıf İlişkisi	43
Şekil 3. 11 Atıf Eşleme Yöntemi Bileşenleri	44
Şekil 3. 12 Kullanılan Arxiv Veri Setine Ait Meta Data Örneği	46
Şekil 3. 13 Ayrıt Listesi	47
Şekil 3. 14 Akış Diyagramı.....	48
Şekil 4. 1 İkili Eşleştirme Başarısı	50
Şekil 4. 2 Gruba Atayabilme Başarısı (%)	52
Şekil 4. 3 Eşik Değeri Üzerinde Gerçekleşen Doğru Gruplama Başarısı (%).....	53

SİMGELER VE KISALTMALAR

SNA	Social Network Analysis (Sosyal Ağ Analizi)
C++	C Plus Plus (Programlama Dili)
ABD	Amerika Birleşik Devletleri
NP	Non Polynomial (Polinomal Olmayan)
HITS	Hyperlink-Induced Topic Search
OpenMP	Open Multiprocessing (Açık Kaynak Çoklu Programlama)

1. GİRİŞ

Günlük hayatın koşturmacası içerisinde pek farkında olmasak da içinde bulunduğumuz evren üzerinde dört bir yanımız çeşitli ağ yapıları ile çevrilidir. Maddelerin atomları arasındaki bağlardan oluşan yapılar, güneş sisteminin kendisi, vücudumuzdaki çeşitli sistemleri oluşturan ağ yapıları, canlılar ve çevre arasındaki ekolojik ağlar, toplumsal ve kültürel ağ yapıları, karayolu, demiryolu veya havayolu ağları, elektrik dağıtım ağı ve tabii ki en karmaşık ağlardan biri olan internet ağı bu yapılara örnek olarak verilebilir ve bu örnekleri daha da çoğaltmak mümkündür.

Ağ kavramı, kelime anlamı olarak “birçok nokta ile bunlar arasındaki bağlantılarla gösterilebilen bir dizgeye ilişkin yapı” olarak tanımlanmaktadır [1]. Sosyal bir ağ ise, düğümler olarak adlandırılan bireylerden veya elemanlardan oluşan ve bu düğümleri kendi aralarındaki çeşitli ilişki ve etkileşimlere bağlı olarak birbirine bağlayan bir yapı olarak tanımlanabilir. Birbirleri ile etkileşim içerisinde bulunan bu varlıkların sosyal ağ yapısı içerisinde gerçekleştirdikleri her türlü etkileşimin detaylı olarak incelenmesi ve analiz edilmesi sonucunda söz konusu yapıyla ve içerdiği düğümlerle ilgili olarak ilk bakışta fark edilemeyen çok kıymetli bilgiler elde edilebilmektedir. Bu bakış açısından yola çıkılarak ağ yapılarının içerisinde yer alan varlıklar arası ilişkilerin çeşitli bilimsel metodlar aracılığı ile detaylı olarak incelenmesi sonucu elde edilen verilerden anlamlı sonuçlar türetilmesi işi ise “Sosyal Ağ Analizi” olarak tanımlanmaktadır. Sosyal ağ analizi 19. yy. sonlarından itibaren sosyoloji biliminin öncülüğünde gelişim göstermiştir. İletişim teknolojilerinde yaşanan büyük gelişmelerin, bireylerin ve toplulukların etkileşimini çok farklı boyutlara taşıması ile daha da gelişmiştir.

Günümüzde özellikle internet ortamında yaşanan gelişmeler neticesinde bireyler arasındaki etkileşim daha hızlı ve karmaşık bir hal almış durumdadır. İki birey arasındaki iletişim zaman ve mekândan bağımsız bir hale gelmiştir. İnternet üzerindeki sosyal paylaşım siteleri, bloglar, forum ortamları, mesajlaşma yazılımları ve daha bunun gibi pek çok ortam insanlar arasındaki ilişkilerin farklı bir boyut kazanmasına neden olmuştur. Sosyal ağ analizi

günümüzde pek çok alanda kullanılmakta olup bunların başında birey ve sosyal grup yapılarının ve davranışlarının incelenmesi (bileşenlerine ayırma, kümeleme, ilişkilerin belirlenmesi), elektronik ticaret ve çevrimiçi reklamcılık (müşteri profilinin çıkarılması ve eğilim analizi, kişiye özel reklamcılık ve teklif sunma), fiziki yapıların analizi (ulaşım, tesisat, altyapı) ve büyük veri kümelerinin analizi (medya takibi, akademik yayın analizi, genetik araştırmalar) gelmektedir.

Sayısal ortamlarda sağlanan iletişimin yarattığı imkânlardan bir diğeri ise oluşan iletişim verilerinin somut bir biçimde ölçülebilir olmasıdır. Çeşitli ortamlarda saklanan veriler, geliştirilen teknikler sayesinde detaylı olarak analiz edilmekte ve bunun sonucunda bir takım değerli bilgiye ulaşılmaktadır. Ancak, sayısal ortamlar üzerinde oluşan sosyal ağ yapıları içerisindeki aktörlerin birbirleri ile olan ilişkileri, etkileşimleri ve bilgi paylaşımları sonucu ortaya çok büyük ölçekte bir veri kümesi çıkmaktadır. Bu veri yığınının işe yarar verilerin ve ileriye yönelik kuvvetli tahminlerin ortaya çıkarılmasında “Veri Madenciliği Teknikleri” büyük fayda sağlamaktadır. Veri madenciliği tekniklerinin internet üzerinde oluşan sosyal ağlara ait veri kümelerine uygulanmasında ise Web Madenciliği yöntemleri kullanılmaktadır [2-6].

Günümüzde bilgiye ulaşma süreci gelişen sayısal bilgi platformları ile birlikte çok hızlı fakat aynı ölçüde yanıltıcı olabilmektedir. Bu yüzden literatüre kazandırılan büyük miktardaki akademik yayının depolanması, tasniflenmesi ve eşleştirilmesi önemli bir bibliyografik araştırma konusu haline gelmiştir. Benzer bir mantıktan hareketle akademik yayınların ve bu yayınlar arasındaki atıfların oluşturduğu yapının bir sosyal ağ yapısı olarak ele alınması ve söz konusu atıfların dokümanlar arasındaki ilişkiyi temsil etmesiyle, yayınlar arasındaki benzerlik “Atıf Analizi” (citation analysis) yöntemleri aracılığıyla içerikten bağımsız olarak tespit edilebilmektedir [7-9].

Literatürde, dokümanların içeriğine bakılmaksızın sadece aldıkları ve yaptıkları atıfların incelenip analiz edilmesi ile akademik yayınlar arasındaki konu benzerliklerinin tespit edilmesi, dokümanların konu gruplarına göre otomatik olarak ayrılmasının sağlanması ve aynı konu grubu üzerinde çalışan akademisyenlere çalışma arkadaşı tavsiye etme gibi amaçlar çerçevesinde çeşitli yöntemler geliştirilmiştir. Atıf analizine dayalı benzerlik tahminine yönelik olarak

geliştirilen en yaygın yöntemler; Bibliyografik Eşleştirme (Biblio-Coupling) [10], Müşterek Atıf Analizi (Co-Citation) [11], Amsler [12], SimRank [13], rvs-SimRank [14], P-Rank [14] ve Inter-Connection [15] yöntemleridir. Ancak, sunulmuş olan yöntemlerin tamamında yalnızca ortak olarak yapılan veya alınan atıflara odaklanılmış ve söz konusu iki dokümanın birbirlerinden farklı olarak başka (ortak olmayan) dokümanlara yaptıkları atıflar ele alınmamıştır. Daha farklı bir deyişle sadece pozitif ödüle dayalı bir benzerlik yaklaşımında bulunulmuş ve negatif puanlama ile sistemin gerçeğe yakınsanması göz ardı edilmiştir.

Bu tez çalışmasında, sosyal ağ analizi kavramı ve sosyal ağların web madenciliği teknikleri ile analizine ilişkin yöntemler incelenmiştir. Web madenciliğinin üç ana tekniğinden biri olan ve ağ yapısı içerisindeki düğümler arası bağların yapısal bakımdan incelendiği “Web Yapı Madenciliği” yaklaşımı kullanılarak akademik yayınlar arasındaki atıfları (citation) ele alan ve içerikten bağımsız olarak “akademik yayınların aynı konu hakkında yazılmış olma ihtimalini” değerlendiren ve yukarıda bahsi geçen ve literatürde daha önce geliştirilmiş olan yöntemlerden üstün bir yaklaşım sunmak amacıyla içerikten bağımsız bir benzerlik tahmin uygulaması geliştirilmiştir.

Uygulamanın gerçekleştirilmesi amacıyla Arxiv [16] akademik yayın veritabanından 1992 ve 2003 yılları arasında yayımlanmış ve “Yüksek Enerji Fiziği” ana konu başlığı altındaki 27.770 akademik yayını ve bunlara ilişkin 352.807 atıf ilişkisini içeren veri seti [17] kullanılmıştır. Sırasıyla mevcut yöntemler ve önerilen yöntem ile ilgili gerekli hesaplamalar ve incelemeler geliştirilen uygulama aracılığı ile bir sunucu üzerinde gerçekleştirilmiştir.

Bulunan benzerliklerin sağlanmasının ve mevcut yöntemler ile kıyaslanmanın yapılabilmesi için “Yüksek Enerji Fiziği” konusunda yaygın bir şekilde başvurulmuş kaynağın [18] bölüm sonlarında bulunan atıflardaki dokümanlar aynı alt konu grubunda olduklarından referans olarak alınmışlardır. Kaynaktaki beş ana konu başlığı altından ve bunlara uygun olarak Arxiv veritabanından seçilen 504 adet referans dokümanın söz konusu yöntemler tarafından hangi başarı yüzdesinde benzer tespit edildiği saptanmıştır. Yöntemin uygulanması sonucunda veri seti üzerinde oluşan benzerlik değerlerine ilişkin ilgili dokümanların hangi gruba

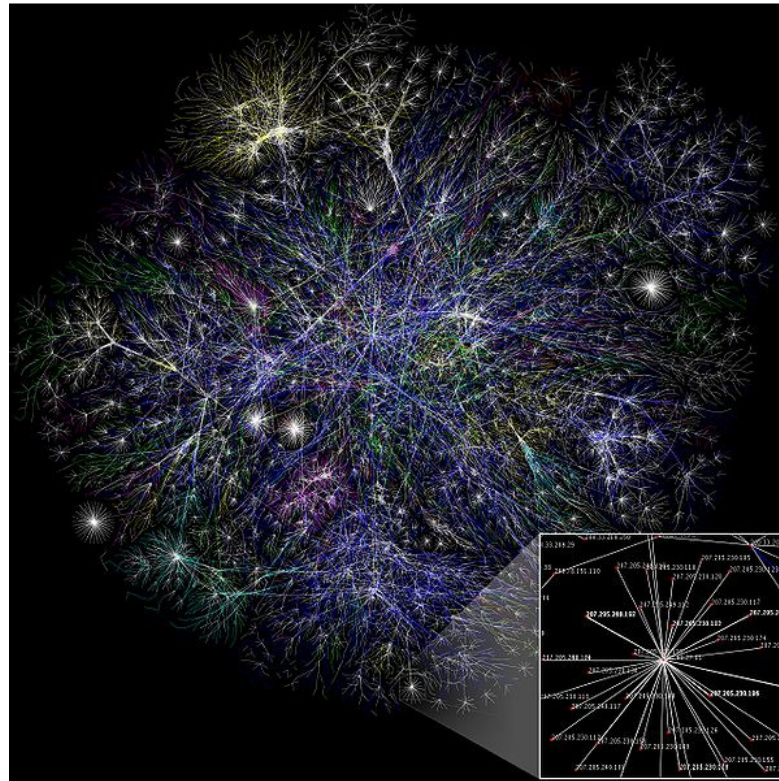
daha benzer olduđuna işaret eden gruplamaların ne derecede başarıya ulaştığı incelenmiş ve sonuçlar kıyaslanmıştır.

Çalışma ile ilgili kavramlara yönelik temel açıklamalar ve tanımlamalar ile mevcut yöntemlerin incelenmesi 2. bölümde, geliştirilen yönteme ve uygulamaya ilişkin detaylar 3. bölümde sunulmuştur. 4. bölümde elde edilen bulgular ve bu bulgular ışığında yapılan yorumlar ele alınmış olup, 5. ve son bölümde ise çalışma ile ilgili sonuçlar sunulmuştur.

2. KURAMSAL TEMELLER

2.1. Sosyal Ağın Tanımı ve Sosyal Ağ Analizi Kavramı

“Ağ” kelimesi Bilişim Terimleri Sözlüğünde; “Birçok nokta ile bunlar arasındaki bağlantılarla gösterilebilen bir dizgeye ilişkin yapı” olarak tanımlanmaktadır [1]. Günlük yaşamımızda çevremiz çeşitli ağ yapıları ile çevrilidir. Bu ağ yapıları arasında Biyolojik Ağlar (Hücre, doku, dolaşım sistemi vb.), Ekolojik Ağlar (Besin zincirleri, Bitki örtüsü vb.), Kültürel Ağlar (Edebiyat, İnançlar vb.), Sosyal Ağlar (İlişkiler, ırklar, toplumlar vb.), Ulaşım Ağları (Karayolları, demiryolları vb.) ve Sayısal Ağlar (İnternet, telefon, uydu sistemleri vb.) sayılabilir. Şekil 2.1 'de bir ağ yapısına örnek olarak internetin ağ yapısı gösterilmiştir [19]. Ağ yapılarının analiz edilmesi ile düğümler olarak adlandırılan ağda bulunan öğeler arasındaki doğrudan göz önünde olmayan ancak anlamlı bilgiler ihtiva eden ilişkiler görünür hale getirilebilir.



Şekil 2. 1: İnternetin Ağ Yapısı [19]

2.1.1. Sosyal ağ nedir?

“Sosyal Ağ” kavramı ilk olarak 1954 yılında J. A. Barnes tarafından kullanılmıştır [20,21]. Sosyal bir ağ; düğümler olarak adlandırılan unsurlardan, bireylerden veya organizasyonlardan oluşan ve bu düğümleri aralarındaki çeşitli ilişki ve etkileşimlere bağlı olarak birbirine bağlayan bir yapı olarak tanımlanabilir. David Liben-Nowell’e göre ise sosyal ağ, sosyal bağlamda kişilerin veya diğer çoklukların aralarındaki etkileşimi, yardımlaşmayı, etkileri gösteren bir yapıdır [20,22].

Günümüzde ise sosyal ağ denince akla gelen ilk tanım, bireylerin sayısal platformlar üzerinden duygu, düşünce, resim, müzik, video vb. içerik paylaşımında bulunduğu, yakınları, arkadaşları ve diğer insanlar ile iletişim kurdukları sanal ortamlardır.

Sosyal aktörler olarak adlandırılan bu düğümleri birbirine bağlayan ilişkiler, çeşitli anlamlarda karşımıza çıkmaktadır. Bunlar, tanışıklık, yakınlık, benzerlik, ticari etkileşim, fiziki bağlılık, iletişim, yönlendirme vb. olabilir. Özellikle internet teknolojisi, mesafe ve zaman kısıtını ortadan kaldırarak sosyal ağların oluşması için mükemmel bir altyapı teşkil etmektedir. Son yıllarda, özellikle genç kuşağın, neredeyse günlük yaşamlarının tamamını sosyal ağlar üzerinde yaşar ve paylaşır hale geldikleri gözlemlenmeye başlanmıştır. En yaygın sayısal sosyal ağlara örnek olarak Facebook, Twitter, Myspace örnek gösterilebilir.



Şekil 2. 2: Sosyal ağ yapısı ve barındırdığı düğümler.

2.1.2. Sosyal ağ analizi kavramı

Sosyal ağ analizi (SNA=Social Network Analysis); sosyal ağ yapıları içerisindeki ilişkilerin çeşitli bilimsel metotlar aracılığı ile detaylı olarak incelenmesi sonucu elde edilen verilerden anlamlı sonuçlar türetilmesi olarak tanımlanabilir. Bir başka tanıma göre sosyal ağ analizi, sosyal yapıyı aktörlerden ve aktör çiftlerini birbirine bağlayan ilişki kümelerinden oluşan bir ağ olarak görerek, sosyal yapıyı ve etkilerini inceler [20,23].

Sosyal ağ analizi 19. yy. sonlarından itibaren sosyoloji biliminin öncülüğünde gelişim göstermiştir. Sosyal ağ analizi aslında yıllar boyunca yapılmış olan sosyoloji, sosyal psikoloji ve antropoloji alanlarındaki birçok araştırmanın sonucudur. Birçok ağ analizi fikri birbirinden bağımsız olarak geliştirilip çalışmalarda kullanılmıştır. Örneğin, 1940'lı yıllarda sosyal gruplar kavramı resmi olarak tanımlanmaya başlanmıştır. 1950 yıllarının ortalarında, gerçek hayattaki gözlemlerin ağlar ile temsil edilmesi ve bu ağların incelenmesine yönelik çalışmalar yapılmaya başlanmıştır. Bu gelişmelerin ardından Moreno, birbirine doğrudan bağlı düğümlerden oluşan ağların görsel temsili için sosyal ilişki çizgeleri (sociograms) terimini kullanmıştır. Devam eden birkaç yılda ise ağların analizine yönelik tekniklerin gelişimi devam etmiştir. Günümüze kadar devam eden süreçte, gelişen web teknolojileri ile birlikte sosyal ağ uygulamalarının da popüler olması, ağ analizi çalışmalarını giderek hızlandırmıştır [24]. Çeşitli ağ yapılarının içerisindeki etkileşimler ve bu etkileşimler sonucunda açığa çıkan ikili ilişki verileri her zaman çok sistemli ve düzenli olmayabilmektedir. Çoğu zaman bu ilişkiler karmaşık ve anlaşılabilirliği düşük bir şekilde karşımıza çıkmaktadır. Buna rağmen özellikle son yıllarda teknolojiye yaşanan gelişmeler ve söz konusu etkileşimlerin ve ilişkilerin sayısal platformlar üzerinde gerçekleşiyor olması, ortaya çıkan devasa veri kümelerinin çeşitli teknikler kullanılarak analiz edilebilmesine ve bunun sonucunda da bahse konu ağ yapısına ait çeşitli kıymetli sonuçlara ulaşılmasına ve ileriye yönelik güçlü tahminlerin yapılmasına olanak sağlamaktadır.

Sosyal ağ analizinde, ağdaki aktörlerin niteliklerinden çok aktörler arası ilişkilere odaklanılır. Bireyler arası ilişkilerin sayısallaştırılıp bilimsel hale getirilmesi anlamına gelen sosyal ağ analizi, önemli olaylar karşısında çeşitli organizasyonların, ya da bu organizasyonların oluşturduğu ağların (network) ilişkilerini rakama dökmek için kullanılmaktadır. Bilgisayar programlarına girilen verilere göre alınacak olan çıktının niteliği de değişmekte ve bu esneklik organizasyonel verimliliği test etmek için kullanılabilecek yeni bir olanak sağlamaktadır [25].

Sosyal ağ analizi günümüzde pek çok alanda kullanılmakta olup bunların başında sosyolojik açıdan birey ve grup yapılarının ve davranışlarının incelenmesi (bileşenlerine ayırma, kümeleme, ilişkilerin belirlenmesi), elektronik ticaret ve çevrimiçi reklamcılık konularında müşteri profilinin çıkarılması, satın alma eğilimlerinin ve ilgi alanlarının analizi, kişiye özel reklamlar önerme ve teklifler sunma, ulaşım, tesisat, iletişim altyapıları gibi fiziki yapıların analiz edilerek darboğazların tespiti ve sistemlerin iyileştirilmesi ve sayısal platformlar üzerinde oluşan büyük veri kümelerinin analizi (medya takibi, akademik yayın analizi, genetik araştırmalar) gelmektedir.

Genelde (özellikle Batı Avrupa'da) SNA, ekonomik ilişkileri ortaya koymak amacıyla kar amacı güden organizasyonların bağlantılarını işlemek üzere kullanılırken, ABD'de her türlü ikili ya da daha çoklu ilişkiyi ortaya çıkarmak için kullanılmaktadır. Özellikle sosyal ilişki kurma amacıyla kurulmuş internet sitelerinde bireylerin diğerleriyle kurdukları kontakların bilimsel dilde anlaşılabilmesi için SNA'lar yoğun bir şekilde kullanılmaktadır. Yine bu amaçla bu sitelerde (Örneğin Facebook, MySpace, Linked-in vs.) sosyal ağ grupları kurulmakta ve bilginin bireyler arasında ne yönde taşındığı konusunda araştırmalar ortaya konmaktadır [25].

Sosyal ağ analizine sosyolojide, antropolojide, coğrafyada, sosyal psikolojide, iletişimde, enformasyon biliminde, işletme ve ekonomide ve biyoloji alanında sıklıkla başvurulmaktadır [20,26].

2.2. Sosyal Ağ Analizi Yöntemleri

Mevcut ağların analizi sosyal, biyolojik, bilgisel, teknolojik ve tüm bunların karışımı şekillerinde yapılabilir. Ağın analiz edilmesi aşamasındaki ilk adım ilişkilere bakmaktır. Bu ilişkiler de şu şekilde sıralanabilir [27];

a) Akım ağları: Bu ağ yapısında bağlantı bir akım şeklindedir ve fizikseldir. Örnek olarak para, elektrik veya suyun akışı verilebilir.

b) Etkileşim ağları: Olay temellidir ve bir zamana bağlıdır. Örneğin e-mail atmak, alışveriş yapmak, vb.

c) İfade ağları: Bağlantının ortaya çıkabilmesi için ifadeye ihtiyaç duyulan ağlardır. İfadeye bağlı olduğu için statik değildir. Örneğin arkadaşlıklar arası ağı kurabilmek için kişinin bunu belirtmesi gerekmektedir.

d) Ait olma (Affiliation) ağları: Bir yere ait olmadan kaynaklı oluşmuş ağlardır. İfade ağının aksine ifadeden bağımsızdır ve kategorileri temel alır. Örneğin aynı okulda okuyanlar ya da aynı işte çalışanlar birer ait olma ağı oluşturabilirler.

Sosyal ağ analizinde temel olarak odaklanılan nokta ağ yapısı içerisinde yer alan varlıklar arasındaki ilişkidir. Bu ilişki farklı şekillerde karşımıza çıkabilir. Ağın özelliğine ve analiz edilmek istenen etkileşime göre; bireylerin arkadaş olmaları, aynı iş yerinde çalışmaları, haftada üç defadan fazla cep telefonu görüşmesi yapmaları, elektronik posta aldığı kişinin postalarını sıklıkla yönlendirmesi, aynı konulara yorum eklemeleri, benzer ürünleri satın almaları ve buna benzer birçok ilişki türü tanımlanabilir [20].

Ağ analizi yapılırken ilişkilerin incelenmesinden sonra bu ilişkilerin merkeziyeti, kümeleşmesi, benzeşmesi ve yoğunluğu incelenir. Bu incelemeyle kullanıcılar arasında bir ayrıştırma yapılabilir. Ayrıca bu ayrıştırmadaki önemli aktörleri ya da aktörler arasındaki önemi tespit etmek için de ağ analizi önem taşır. Söz konusu ayrıştırmayı sosyal ağlarda kullanmak firmalar için çok fazla avantaj sağlar. Bu ağ analizi sayesinde bir markayı takip eden bireyler arasındaki etkileşimi ve bu etkileşimin boyutunu tespit etmek mümkündür. Bu tespitten yola çıkarak alışveriş siteleri kullanıcılarına daha önce satın aldıkları ürünlere göre yeni ürünler önerebilmekte, arkadaşlık siteleri tanıma ihtimalinizin

olduđu kiřilerin listenizi size sunabilmektedir. Bu da çevrimiçi pazarlama alanında kiřiye özel reklam imkânını doğurmaktadır [27].

2.2.1. Örüntü bulma yöntemleri

Sosyal ağ analizinde örüntü olarak adlandırılan ve veri içerisinde aranılan bilgiye ait belirli bir ipucuna, oluşuma veya grulařmaya ışık tutan motifler çıkarmayı sađlayan tekniklere ihtiyaç duyulmaktadır. Bu teknikler kısaca şöyle açıklanabilir [28]:

a) İstatistiksel işlemler: Elimizdeki veriyi tanımlamakta ve bilgi elde etmekte kullanılan en güçlü teknikler istatistik metotlardır. Analist tarafından farklı deđişkenleri temel alan tanımlayıcı istatistik analizler yapılabilir.

b) Birliktelik kuralları (Association Rules): Beraber kullanılan veriler ilişkilendirme kuralları uygulanarak bulunup aynı kategoriye konulabilirler. İlişkilendirme kuralları genellikle veri tabanındaki veriler arasındaki ilişkileri tespit etmeye çalışır. (Bunu satın alan řunu da aldı vb. tanımlamalar)

c) Kümeleme (Clustering) : Grulama(kümeleme) analizi veriler arasında benzer karakteristik deđerler taşıyanları bir araya getirerek gruplar oluşturmayı hedefler. Sosyal ağ analizinde çok sık kullanılmaktadır.

d) Sınıflandırma (Classification) : Bu teknikler verileri ait oldukları tanımlı sınıflara koymaya çalışır.

e) Sıralı örüntüler: Zamana yayılan veri kümeleri arasında benzer örüntüler bulmaya çalışılır.

f) Bađımlılık modellemesi: Web deđişkenleri arasındaki bađımlılıkları ortaya çıkaran modeller oluşturmak hedeflenir.

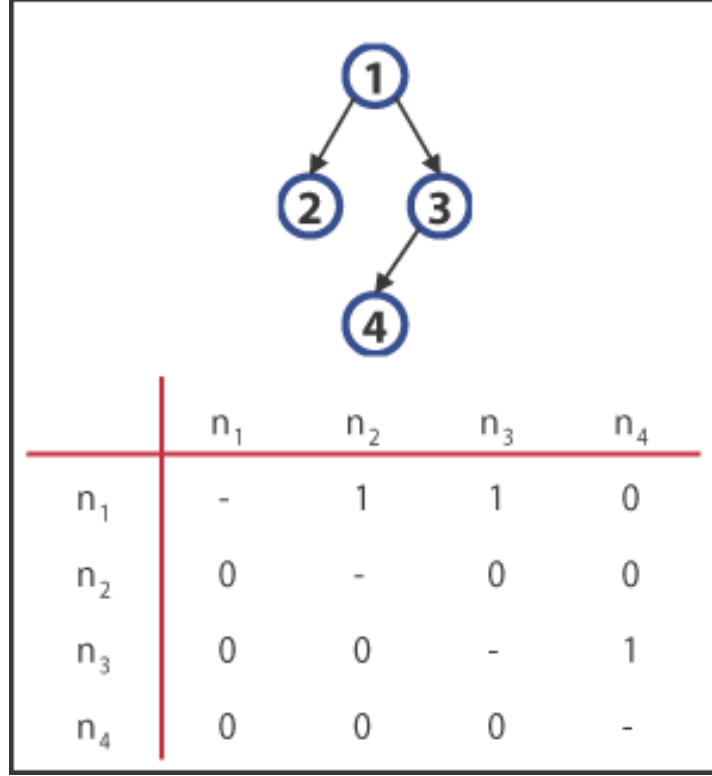
2.2.2. Çizge teorisi

Tanımlanan bu kuralların mevcut ağ yapısı içerisinde incelenerek ilişkilerin analiz edilmesi ve bu sayede ağ üzerinde çeřitli yorumlara varılması konusunda çizge teorisi (graph theory) yöntemlerinden sıkça faydalanılmaktadır [29]. Sosyal yapının kendisi bir çizge, içerdіđi varlıklar bu çizgedeki düğümler ve

varlıklar arası ilişkiler de çizgedeki ayrıtlar olarak ele alınmaktadır. Çizgeler ve bu çizgelerden elde edilen matrisler üzerinde gerçekleştirilen çeşitli matematiksel hesaplamalar ile düğümler arasındaki kümelenme (clustering), merkezilik, benzerlik, uzaklık-yakınlık vb. ilişkiler somut olarak hesaplanmakta ve bu değerlere bağlı olarak ağdaki ilişkiler yorumlanmaktadır [29]. Başlıca SNA araçları UCI-NET [30], NetMiner [31], Pajek [32], ORA, Stat-Net, SocNet-V, InFlow ve Keyhubs'tır. İlk üç yazılım sosyal ağ analizinde yoğun bir şekilde kullanılmaktadır.

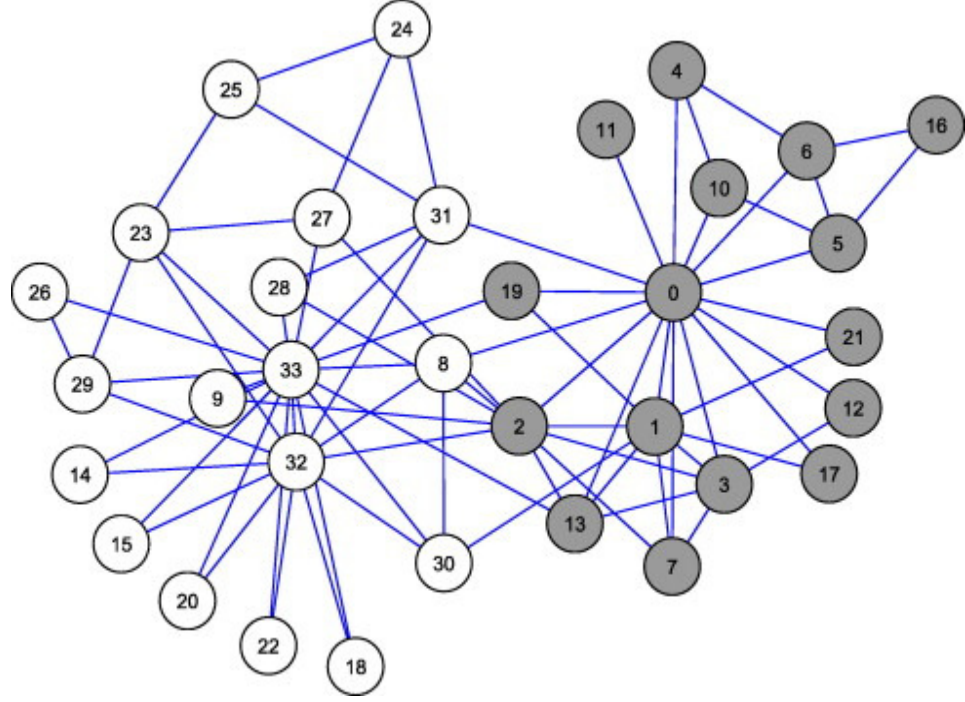
Genel çizge gösterimi $G = (V, E)$ şeklindedir. Bu gösterimdeki V , sonlu düğüm (vertex) kümesini ve E sonlu ayrıt (edge) kümesini belirtir. Her bir ayrıt kendisi ile ilişkilendirmiş bir ya da iki adet düğüm içerir ve bu düğümleri birbirine bağlar. Sosyal ağlardaki kullanıcılar (insan, grup, kurum vb.) çizgede düğüm olarak temsil edilirken, sosyal ilişkiler (arkadaşlık, profesyonel iş ilişkisi vb.) çizgede ayrıt olarak temsil edilirler. Bazı çizgelerde iki düğüm arasında birden fazla ayrıt bulunabilir. Bu tür çizgelere çoklu çizgeler (multigraphs) adı verilmektedir. Bir ayrıt bir düğümü kendisine de bağlayabilir. Bu tür döngüler çizge içerisinde yer alabilir. Hatta ayrıtlar yönlü de olabilir. Yönlü çizgeler (directed graphs) adı verilen bu çizge yapılarında ayrıtların başlangıç ve bitiş noktaları bellidir. Yönlü ve yönsüz ayrıtların birlikte bulunduğu karışık çizgeler (mixed graphs) de mevcuttur. Bu tür çizge özelliklerini daha da sıralayabiliriz. Günlük hayatta pek çok farklı alanda kullanılan çizge modelleri sosyal ağları temsil etmek için kullanılabilir. Çizge yapısının temsil edilmesi için genel olarak iki farklı yöntem tercih edilebilir. Bunlardan ilki, komşuluk listesi (adjacency list) adı verilen, ağdaki her bir düğüm için komşu ya da bitişik olduğu düğümlerin listesinin dizi şeklinde gerçekleştirilmesidir. Bir düğümün komşuları liste dolaşarak tespit edilir. İkinci yöntem ise sık kullanılan komşuluk matrisi (adjacency matrix) yöntemidir. Bu yöntemde çizgedeki n tane düğüm için oluşturulan $n \times n$ boyutundaki matris üzerinde birbirine komşu olmayan düğümlerin girdileri 0, komşu düğümlerin girdileri 1 olarak atanır. Şekil 2.3'te basit bir sosyal ağın çizge olarak temsili ve bu çizgedeki düğümler ve ayrıtlardan elde edilen komşuluk matrisi örneği yer almaktadır. Matris işlenirken de her bir

girdinin deęerine bakılarak, karřılık geldięi satır ve sütünadaki dűęűmler arasında ayırıt(baęlantı) olup olmadıęına karar verilir [24].



řekil 2. 3 rnek Bir izge ve Komřuluk Matrisi [33].

1977 yılında W. W. Zachary tarafından yapılan bir alıřmada sunulan Zachary'nin Karate Kulűbű sosyal aę rneęi sosyal aę analizi konusunda popűler bir rnektir [34]. řekil 2.4'te grűldűęű űzere karate kulűbűnde 34 adet ęrenci bulunmaktadır ve birbiri ile arkadař olan ęrenciler iin izge űzerindeki dűęűmler arasında bir ayırıt mevcuttur.



Şekil 2. 4 Zachary'nin Karate Kulübü Sosyal Ağı [35].

Bu karate kulübünde, kulübün yöneticisi ve öğretmeni arasında bir anlaşmazlık bulunduğu ve öğrencilerin de bu anlaşmazlık çerçevesinde gruplaşarak yönetici ve öğretmen çevresinde yakınlaştıkları gerçeği üzerine kulüp incelenmiş, ilişkiler bir çizge şekline dönüştürülmüş ve matematiksel analizlerin bu gerçeği ne derece yansıttığına bakılmıştır [34]. Şekildeki çizge yapısından da açıkça belli olduğu üzere 0 ve 33 numaralı düğümlerin çok sayıda arkadaşı bulunmakta ancak ortak pek arkadaşları bulunmamaktadır. Çizge bu iki düğüm etrafında iki ayrı gruba ayrılmıştır. Bu iki düğümün kendi gruplarının liderleri olduğu düşünülebilir. Görüldüğü üzere sadece çizge görüntüsüne bakılarak dahi ağın yapısı ile ilgili birtakım sosyal çıkarımlara varılabilmektedir.

2.2.3. Sosyal ağ yapılarına ilişkin ölçütler

2.2.3.1. Yoğunluk (density)

Ağda var olan tüm bağlantıların, olası tüm bağlantılara oranı yoğunluk ölçütünü verir. Tüm aktörlerin birbirine bağlı olduğu ağlarda yoğunluk değeri 1 olacaktır. Yoğunluk değerinin yüksek olması ağdaki aktörlerin birbirlerine daha

güçlü bağlı olduklarını ve birbirlerinden soyutlanmış durumda olmadıklarını gösterir [24,36-38].

2.2.3.2. Büyüklük (size)

Sosyal ağdaki aktörlerin sayısı o ağın büyüklüğünü göstermektedir. Ağdaki aktör sayısı arttıkça ağın büyüklüğü de artmaktadır. Ağdaki düğüm sayısı ile düğümlerin birbirleri ile ilişki kurmaları ihtimali ters orantılıdır. Başka bir deyişle az sayıda düğüm içeren ağ yapılarında düğümlerin birbirleri ile ilişkili olmaları ihtimali daha yüksektir ve buna bağlı olarak düğümler arası ilişkiler daha kuvvetlidir. Düğüm sayısı arttığında ise herhangi iki düğümün birbiri ile ilişki kurma veya etkileşimde bulunma ihtimali düşmektedir. Bazı düğümler arasında ilişki dahi bulunmayabilmektedir. Bu tarz büyük ağ yapılarında tüm düğümlerin birbiriyle etkileşim içerisinde bulunduğu bir sosyal ağ yaratabilmek neredeyse imkânsız olarak kabul edilmektedir [24,36-38].

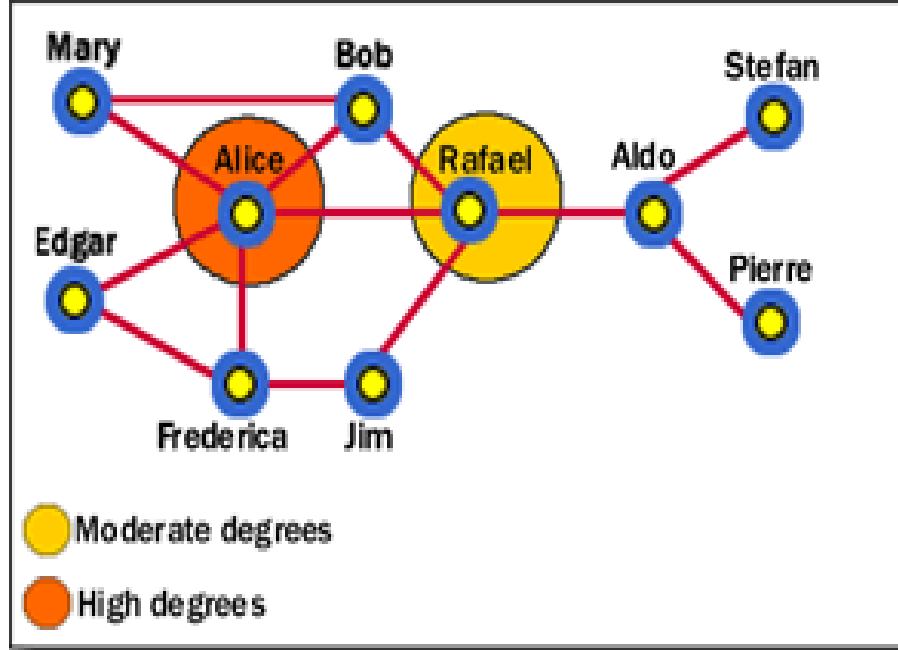
2.2.3.3. Genel kümelenme katsayısı (global clustering coefficient)

Ağdaki üçgen (üç düğüm ve bu üç düğümün birbirlerine tam bağlı olduğu alt ağ) sayısının, açık ya da kapalı tüm üçlülere (üç düğümden oluşan alt ağ, düğümlerin hepsi birbirine tam bağlı olmak zorunda değil) oranı ile hesaplanır. Bazı kaynaklarda geçişkenlik (transitivity) olarak da geçen bu ölçüt, mevcut ağ yapısının kümelenmeye ne kadar eğilimli olduğunu tespit etmek için kullanılmaktadır [24,39,40].

2.2.3.4. Derece merkeziliği (degree centrality)

Ağdaki bir aktörün derecesi, kendisine bağlı olan bağlantıların sayısı ile hesaplanır. Basit hesaplanmasının yanı sıra, aktörün önemini gösterebilecek önemli bir ölçüttür. Çoğu sosyal ağda, bir bireyin ne kadar çok bağlantısı var ise o kadar önemli ve güçlüdür gibi bir bakış açısı hâkimdir. Hatta derecesi en yüksek olan aktör, ağın en aktif üyesi olarak da yorumlanabilir. Bağlantıların

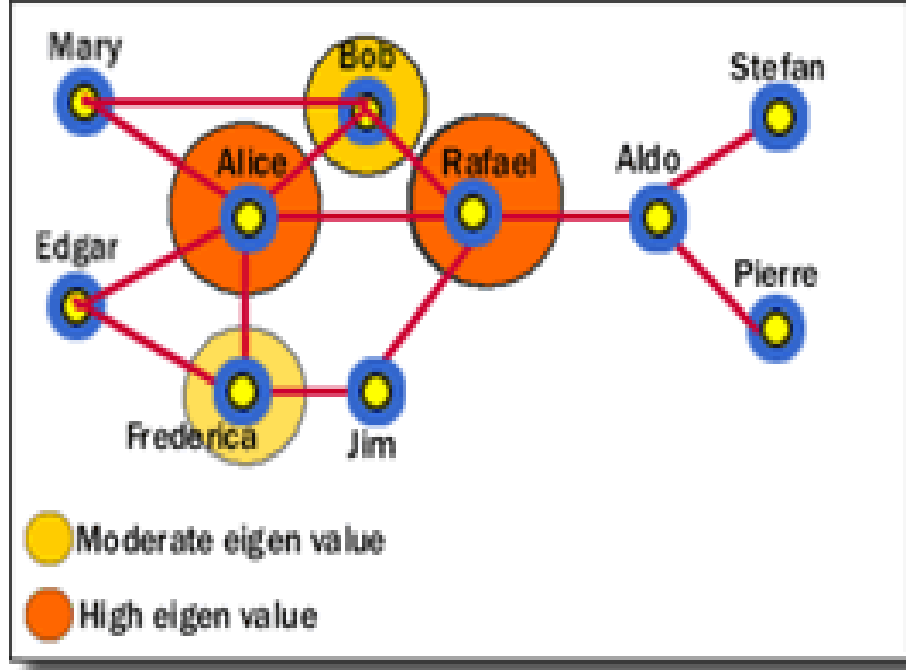
yönlü olduğu durumlarda, giren bağlantıların sayısı (in-degree) ile çıkan bağlantıların sayısı (out-degree) ayrı ayrı hesaplanır. Şekil 2.5'te daha çok ayrıta sahip olan Alice düğümünün derecesinin daha yüksek olduğu görülmektedir [24,41,42].



Şekil 2. 5 Derece Merkeziliği Örneği [43]

2.2.3.5. Özvektör merkeziliği (eigenvector centrality)

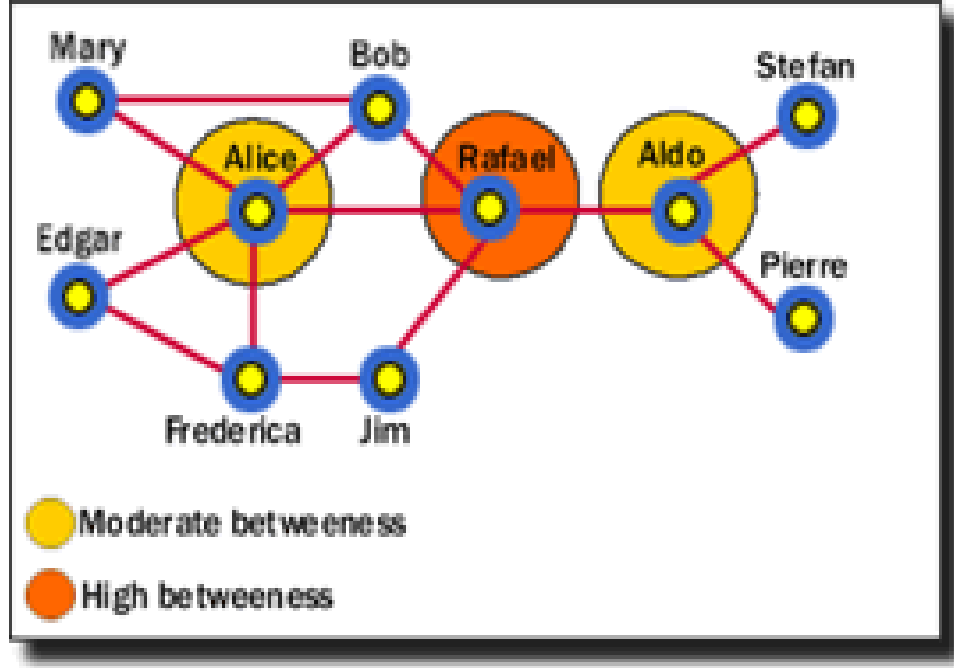
Derece hesaplanırken dikkate alınan ve aktörün sahip olduğu bağlantıların eşit olmadığını gösteren bir ölçüttür. Örneğin, ağdaki bir aktör için önemli aktörlere olan bağlantıların etkisi, diğer sıradan bağlantılardan daha fazla olabilir. Bağlantıda olduğu aktörlerin daha merkezde olması, o aktörün daha merkezi bir konumda olacağını gösterir (Şekil 2.6). Bu ölçüt hesaplanırken komşuların merkezilik derecelerinin toplamı dikkate alınır [24,44,45].



Şekil 2. 6 Özvektör Merkeziliği Örneği [43]

2.2.3.6. Arasındalık merkeziliği (betweenness centrality)

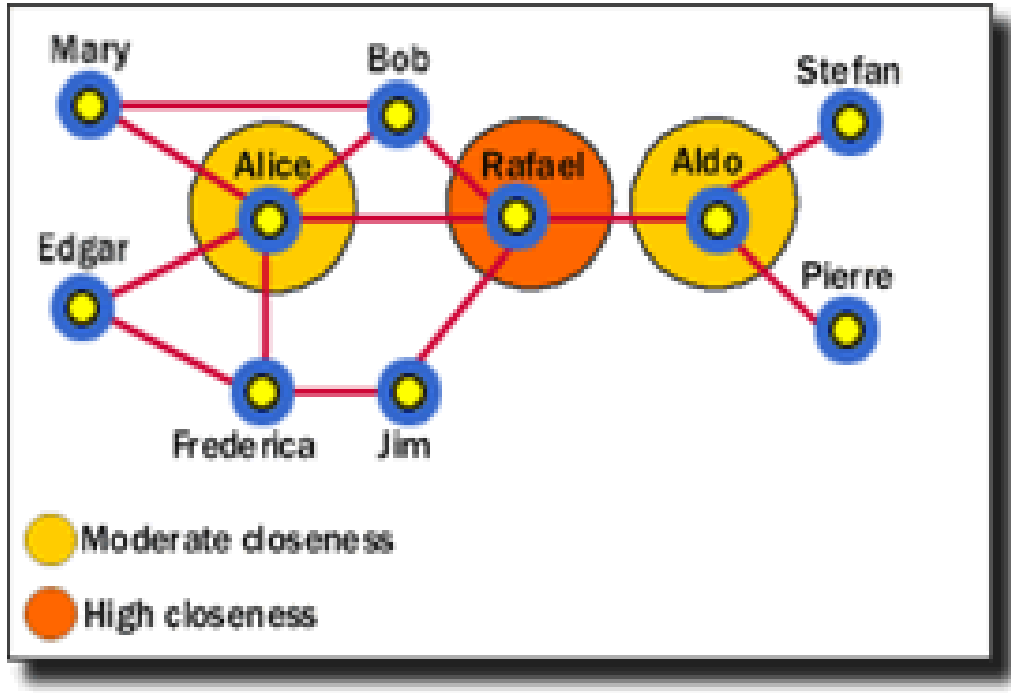
Merkezilik ölçütleri içerisinde hesaplanması en karmaşık ölçüt olan arasındalık ölçütü, ağdaki bir düğümden ya da aktörden geçen en kısa yolların oranı ile bulunur. Öncelikle ağdaki tüm düğüm çiftleri arasındaki en kısa yollar (geodesics) bulunur, ardından bu yolların kaçında o düğümün yer aldığı oranı arasındalık ölçütünü verir. Büyük ağlarda hesaplanması çok maliyetli olabilecek bir ölçüt olduğu için belirli seviyedeki komşulara kadar inilerek de hesaplanabilir. Şekil 2.7’de görüldüğü gibi arasındalık merkeziliği derecesi yüksek olan aktörler diğer aktörlere göre daha önemli bir konumdadırlar ve ağda olup bitenden daha çok haberdar olacakları değerlendirilmektedir [24,41,42].



Şekil 2. 7 Arasındalık Merkeziliği Örneği [43]

2.2.3.7. Yakındalık merkeziliği (closeness centrality)

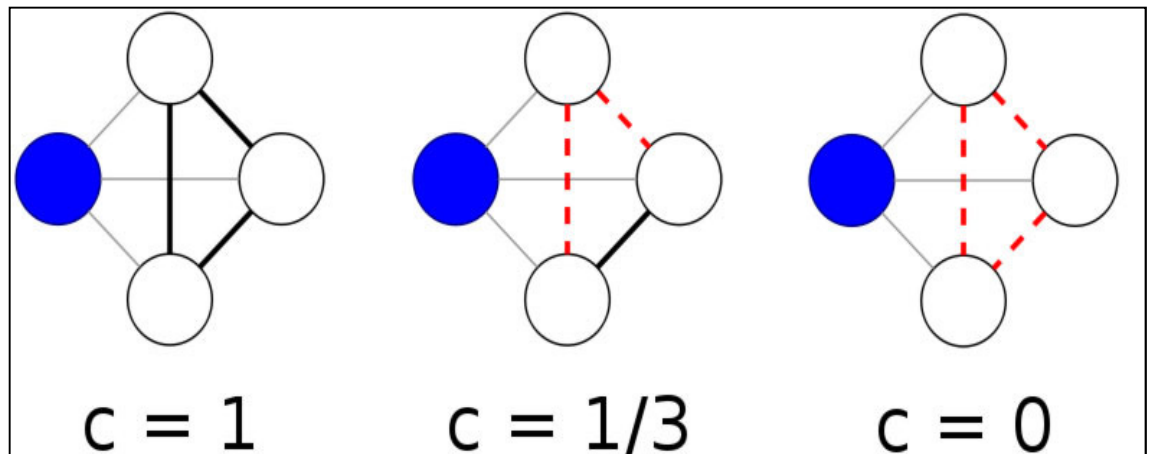
Ağdaki herhangi bir düğümden, diğer tüm düğümlere olan en kısa ortalama uzaklıkların toplamının (geodesic distance) elde edilmesiyle bulunur. Bağlantıların yönlü olması durumunda kısa yolların bulunması aşamasında bu yönler dikkat edilmelidir. Bu durumda da giriş (in-closeness) ve çıkış (out-closeness) olmak üzere iki farklı yakındalık ölçütü hesaplanır. Şekil 2.8’de görüldüğü üzere yakındalık ölçütü daha küçük olan düğümler diğer düğümlere göre ağda daha merkezi bir konumdadırlar ve düğümlere daha hızlı erişim avantajına da sahiptirler [24,42, 45].



Şekil 2. 8 Yakınlık Merkeziliği Örneği [43]

2.2.3.8. Kümelenme katsayısı (clustering coefficient)

Kümelenme katsayısı bir aktörün, komşuları arasındaki bağlantının ne derece güçlü olduğu hakkında bilgi verir. Bir aktörün komşularının kendi aralarındaki gerçek bağlantıların sayısının olası tüm bağlantıların sayısına oranıyla hesaplanır (Şekil 2.9). Tam bağlı (complete) ağlarda, yani tüm aktörlerin birbirleri arasında olası tüm bağlantıların bulunduğu ağlarda, her bir aktör için kümelenme katsayısı ölçütü 1 olacaktır [20, 24, 39].



Şekil 2. 9 Kümelenme Katsayısı Örneği [46]

2.3. Çizgelerde Topluluk Tespiti

Sosyal ağ yapısı içerisindeki bir düğüm kümesi, dışarıya olan bağlantı sayısına göre kendi içinde daha fazla sayıda bağ içeriyor ise bu düğüm kümesi bir topluluk (community) olarak nitelendirilebilir. Bir diğer deyişle içeride çok ancak dışarıda az sayıda bağlantı varsa bu kümeler topluluk adını alırlar [20].

Ağ yapıları içerisindeki topluluklar bize bireylerin ortak ilgi alanları, çalışma konuları, eğilimleri, benzerlikleri vb. hakkında somut bir fikir sunmaktadır.

Gerçek ağlarda ağ yapısı rassal değildir. Yani homojen değil heterojen bir yapı mevcuttur. Belirli bir alanda yoğunlaşan, kümeleşen ve topluluk (community veya cluster) olarak adlandırdığımız yapılar muhtemelen aynı özelliği paylaşan ve/veya benzer rolü bulunan düğüm kümeleridir [29].

Grup tespitinde en önemli nokta düğümlerin bulundukları grup içindeki yapılarına bakarak sınıflandırılabilmesi ve grupların ortaya çıkarılabilmesidir. Grup tespitinde kullanılan yaklaşımların büyük çoğunluğunun temelinde gruplar arası bağların ayrılması fikri yatmaktadır. Gerçek ağ yapılarında karşılaşılan en büyük sorun ise düğümlerin birden fazla gruba ait olabilme ihtimali olarak adlandırılan “overlapping” durumudur. Overlapping faktörünü hesaba katarak yapılan işlemlerin karmaşıklığı çok fazla olduğundan klasik algoritmalar genellikle her bir düğümü bir gruba dahil edecek şekilde çalışırlar ancak bu durum bir kısım bilginin göz ardı edilmesini de beraberinde getirmektedir [29].

Grup tespitinde en sık başvurulan tanımlama, grubun içindeki ayrıt sayısının dışarıya olan bağlantı sayısından fazla olması gerektiği varsayımdır. Bu noktadan hareketle tanımlanmış olan “cut-size” parametresi söz konusu topluluğu çizgenin geri kalanına bağlayan ayrıt sayısı olarak adlandırılmaktadır. İyi bir topluluğun cut-size değerinin düşük olması beklenir.

Bir başka tanımlama olan “Düğüm benzerliği (vertex similarity)” ise düğümlerin bir uzay düzleme yerleştirildiklerinde aralarında kalan mesafenin bir benzerlik ölçütü olarak ele alınmasıdır. Klasik gruplama yöntemleri sıklıkla bu yaklaşımdan faydalanmaktadırlar. Düğümler bir uzay düzleme yerleştirilemiyor ise bu durumda komşuluk matrisi (adjacency matrix) kullanılabilir. Komşuları aynı ise kendileri komşu olmasalar bile benzerdirler denilebilir. Atıf analizi

yaklaşımlarında da bu mantıktan faydalanılmaktadır. Bunun dışında iki düğüm arasındaki bağımsız yolların sayısının ölçülmesi, en kısa yolun mesafesi veya rastgele yürüyüş gibi yöntemlerle de düğümler arası benzerlikler saptanabilir [29].

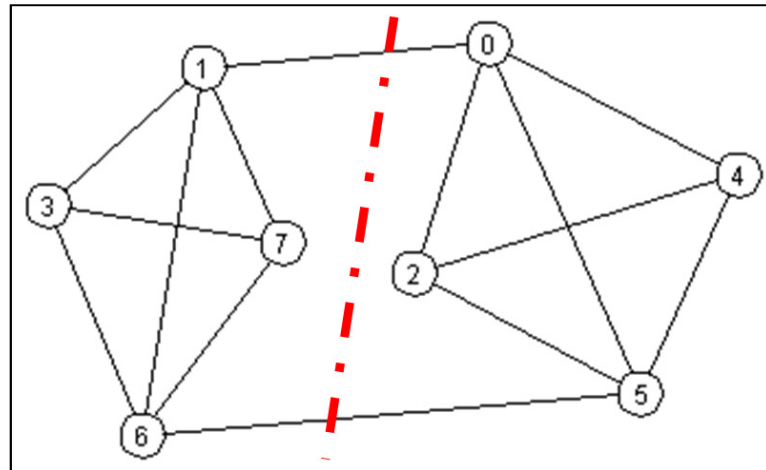
Bir diğer yaygın yaklaşım ise çizge üzerinde yapılan işlemler (bölünme, birleştirme, ayırılma, silme, ekleme vb.) sonucunda bir kalite fonksiyonunun iyileştirilmesine dayanmaktadır. En yaygın kullanılan kalite fonksiyonu modüleritedir.

Her ne kadar ağ yapısı içerisindeki toplulukların tespit edilmesi hesaplama karmaşıklığı açısından zor bir işlem olsa da bu konuda başarı göstermiş yöntemler geliştirilmiştir. Bu yöntemler içerisinde en yaygın ve etkili olarak kullanılanlar aşağıda açıklanmıştır [29].

2.3.1. Geleneksel yöntemler

2.3.1.1. Çizge bölütleme (graph partitioning)

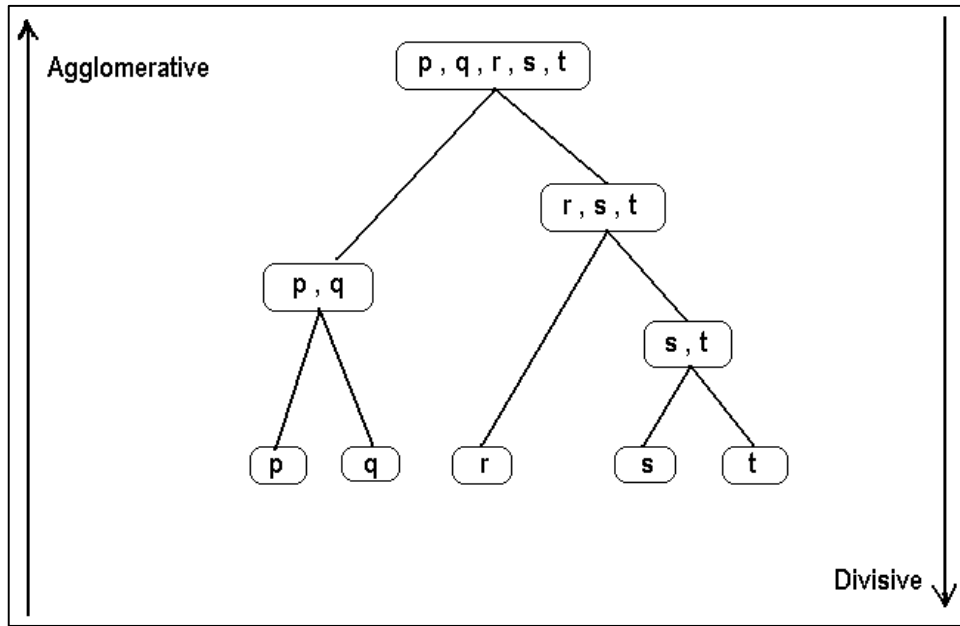
Düğümün, sayısı önceden belirlenmiş olan k adet gruba, gruplar arası ayrılma sayısı minimum olacak şekilde bölünmesidir. Ancak sosyal ağ yapılarında mevcut grup sayısı önceden bilinmediğinde sosyal ağ analizi için uygun bir yaklaşım değildir. En önemli algoritmaları Iterative Bisectioning [58], Kernighan-Lin [59], Max-Flow Min-Cut Theorem' dir [60].



Şekil 2. 10 Çizge Bölütleme [47]

2.3.1.2. Hiyerarşik gruplama

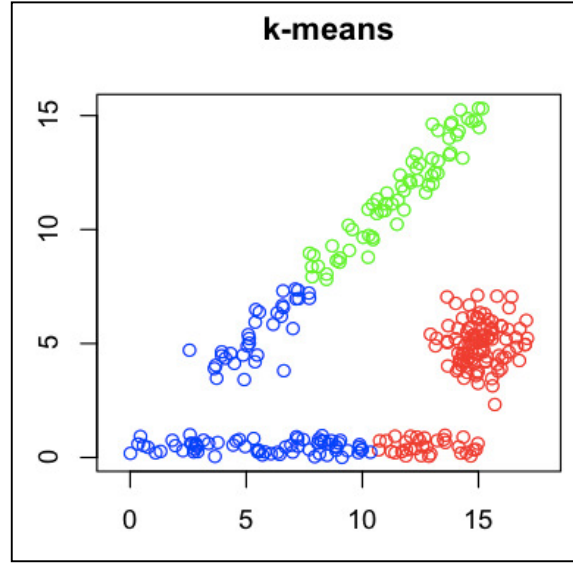
Sosyal ağlar genellikle iç içe hiyerarşik bir yapıda bulunan gruplar içerirler. Benzer olan düğümlerin birleştirilmesi ve grup yapılması ve düşük benzerlikli düğümlerin silinerek grupların bölünmesi fikri üzerine kurulu bir yöntemdir (Şekil 2.11). Sonuçlar tamamen belirlenecek olan benzerlik ölçütüne bağlı olarak değişkenlik gösterecektir [29].



Şekil 2. 11 Hiyerarşik Gruplama [48]

2.3.1.3. Bölütlemeli kümelenme (partitional clustering)

Burada da grup sayısı olan k önceden belirlenir ve her düğüm uzayda bir nokta olarak ele alınır. Amaç, verilen bir fonksiyona göre aralarındaki mesafeye bağlı olarak merkeze olan uzaklıklarına göre noktaları k gruba ayırmaktır. En çok kullanılan fonksiyonlar ise Minimum k -clustering, k -center, k -median, ve k -means [61]' tir (Şekil 2.12). Burada da dezavantaj grup sayısının önceden bilinmesi ihtiyacıdır [29].



Şekil 2. 12 K-means Yöntemi [49]

2.3.1.4. Spektral kümelenme (spectral clustering)

Bu yöntemde önce benzerlik matrisinin özvektörleri alınır ve daha sonra k-means [61] gibi bir fonksiyon ile gruplara ayrılır. En çok kullanılan matris Laplas matrisidir. Bu yaklaşım sayesinde özvektörlerin bileşenlerinden çizgede kaç adet grubun bulunduğu öğrenilebilir. Kullanılan laplas matrisinin normalize edilmiş olup olmamasına göre iki farklı versiyonu mevcuttur [29].

2.3.2. Bölütlemeli algoritmalar

Çizge içindeki grupları birbirine bağlayan ayrıtları bulup silmek ve böylece grupları ayırklaştırıp ortaya çıkarmayı amaçlayan bir yöntemdir. Önemli olan nokta ise bu grupları bağlayan ayrıtların nasıl tespit edileceğidir. En popüler algoritması Girvan-Newman algoritmasıdır [62]. Burada ayrıt merkeziyeti (edge centrality) olarak adlandırılan bir ölçüte bağlı olarak ayrıtlar seçilir. Tüm ayrıtlar için merkezilik değeri hesaplanır. En yüksek merkezilik değerine sahip ayrıt silinir. Tekrar birinci adım gerçekleştirilir ve en yüksek değere sahip ayrıt silinerek bu şekilde işlem devam eder. Ayrıt merkeziyeti kistası dışında ayrıt bitişikliği (edge betweenness), rastgele yürüyüş ayrıt bitişikliği (random walk

edge betweenness) ve akım akışı bitişikliği (current flow betweenness) gibi kıstaslar da kullanılmaktadır [29].

2.3.3. Modülerite esaslı yöntemler

Modülerite çizge analizinde en çok bilinen ve kullanılan kalite fonksiyonudur. Her ne kadar tam olarak ispatlanamamış olsa da yüksek modülerite değerinin iyi grupları işaret ettiği kabul edilmektedir [29]. Eğer bir çizge aynı boyuttaki ve derecedeki bir rastgele çizgeye göre daha yüksek modülerite değerine sahipse o çizgenin grup yapısına sahip olduğu kabul edilir. Ancak modülerite değerinin yüksek çıkması her zaman grup yapısı bulunduğu anlamına gelmeyebilir. Bazı rassal çizgelerde de grup yapısı olmamasına karşın yüksek modülerite değerleri ile karşılaşılabilir. Modülerite fonksiyonunun iyileştirilmesi NP-Complete bir problem olduğundan doğrusal bir zamanda çözümü yoktur. Ancak çeşitli yakınsamalar ile başarılı sonuçlar elde eden algoritmalar geliştirilmiştir [63-67]. Çizge üzerinde yapılan değişiklikler kümesi içerisinde kalite fonksiyonunu en yüksek olarak iyileştiren değişiklik yapılır. Bu işlem bir birleştirme, ayrılma veya ayırma silme olabilir.

2.3.4. Dinamik algoritmalar

Dinamik algoritmalar içerisinde en çok kullanılanlardan biri rastgele yürüyücü algoritmasıdır [68]. Grup yapılarının içerisinde daha çok sayıda (yoğun) düğüm bulunduğu ve dolayısıyla bu düğümler arasında daha fazla miktarda ayırma bulunduğu ve buna bağlı olarak bir rastgele yürüyücünün daha fazla yol kat etme ihtimalinden yola çıkılarak grup yapısı içerisinde daha fazla zaman geçireceği fikrine dayanmaktadır. Bu fikirden hareketle iki düğüm arasındaki mesafe hesaplanmaktadır [29].

2.3.5. Diğer yöntemler

Yukarıda bahsedilen ve sıklıkla kullanılan yöntemlerin dışında istatistiksel çıkarıma dayanan (Bayes vb.) yöntemler [69-75], düğümleri etiketleyen ve her bir iterasyonda komşuları tarafından en çok paylaşılan etiketi alan ve bu şekilde grupları ayıran yöntemler [76], klik filtreleme yöntemleri [77-78], overlapping ile mücadele eden yöntemler [80-85] ve çok çözünürlüklü yöntemler mevcuttur [86-91].

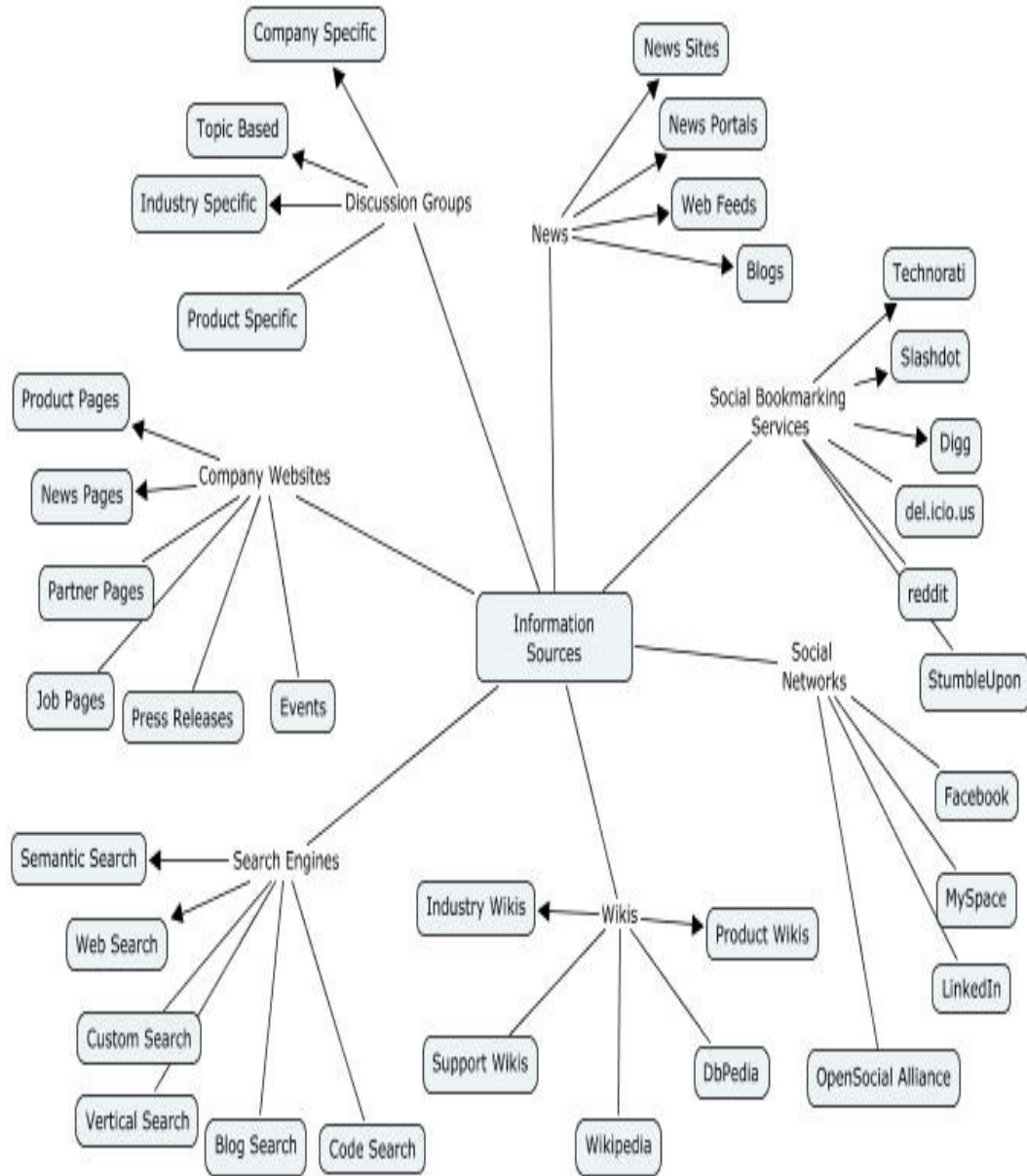
2.4. Web Madenciliği Teknikleri

Sayısal ortamlarda sağlanan iletişimin yarattığı imkânlardan biri de oluşan iletişim verilerinin somut bir biçimde ölçülebilir olmasıdır. Çeşitli ortamlarda saklanan veriler, geliştirilen teknikler sayesinde detaylı olarak analiz edilmekte ve bunun sonucunda bir takım değerli bilgiye ulaşılmaktadır. Ancak, sayısal ortamlar üzerinde oluşan sosyal ağ yapıları içerisindeki aktörlerin birbirleri ile olan ilişkileri, etkileşimleri ve bilgi paylaşımları sonucu ortaya çok büyük ölçekte bir veri kümesi çıkmaktadır. Bu veri yığından işe yarar verilerin ve ileriye yönelik kuvvetli tahminlerin ortaya çıkarılmasında “Veri Madenciliği Teknikleri” büyük fayda sağlamaktadır. Veri Madenciliği Teknikleri, sosyal ağlara ait veri kümelerine uygulandığında ise Web Madenciliği yöntemleri olarak adlandırılmaktadır [2-6].

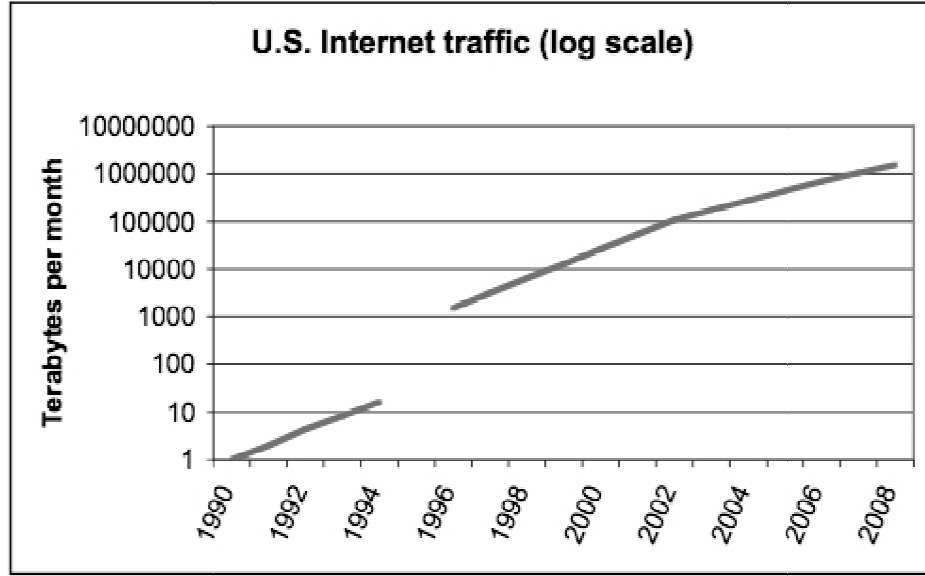
Web madenciliği; çeşitli yapıdaki web sayfalarını, dokümanlarını ve kayıt bilgilerini incelemek ve bunlardaki anlamlı kalıpları keşfetmek için veri madenciliği tekniklerinin kullanılması olarak tanımlanabilir [2].

Literatürde “Web Madenciliği” teriminin ilk kez 1996 yılında Oren Etzioni tarafından ortaya atıldığı belirtilmektedir [50]. Etzioni bu çalışmasında web madenciliğinin, veri madenciliği teknikleri kullanılarak internet ağında bulunan dosya ve servislerden otomatik olarak örüntüler bulunması ve öngörülme bilgiye ulaşılması olduğunu iddia etmektedir. Araştırmacıların büyük çoğunluğu çalışmalarında bu tanımlamayı esas almışlardır [28].

Web madenciliğinin amacı ise veri madenciliği tekniklerinin kullanılarak web belgelerinden ve servislerinden otomatik olarak bilginin ayıklanması, ortaya çıkarılması ve tahlil edilmesidir [5]. Şekil 2.13'te web madenciliğinde kullanılan temel veri kaynakları görülmektedir. 2008 yılında ABD'de yapılan bir çalışmaya göre bir ayda ortalama bir milyon terabyte veri taşınmakta ve bu sayı her geçen yıl artmaktadır (Şekil 2.14) [51].



Şekil 2. 13 Web Madenciliğinin Veri Kaynakları [2]



Şekil 2. 14 2008 Yılı ABD Ortalama İnternet Trafiği [51]

Web madenciliğinde kullanılan veriler, web üzerinde çok geniş bir alandan toplanmaktadır. Bu veriler web sayfaları, log dosyaları, kullanıcı kayıt bilgileri, oturum ve hareket bilgileri, site yapısı ve site içeriğinden oluşmaktadır [2]. Söz konusu büyüklükte ve karmaşıklıkta verilerin etkin bir biçimde analiz edilebilmesi için web madenciliğinde kullanılan tüm yöntemler 4 ana işlem basamağından oluşmaktadır: Kaynak bulma, bilgi çıkarımı ve ön işleme, genelleştirme ve çözümleme [4].

2.4.1. Web madenciliği tekniklerinin ortak işlem basamakları

2.4.1.1. Kaynak bulma

Kaynak bulma çok geniş kapsamlı bir konudur ancak kısaca bilginin elde edilme kısmıdır. Çeşitli verilerin çevrimiçi ya da çevrimdışı olmasına bakılmadan bir veri ambarında toplanmasıyla yapılır. İnternet üzerindeki çeşitli gazeteler, haber grupları vb. gibi yerlerden verinin toplanıp arama amaçlı olarak bir yerde saklanmasıdır [52].

2.4.1.2. Bilgi çıkarımı ve ön işleme

Veri kaynağından toplanan verilerin işlenmesi ve işe yaracak hale getirilmesi, bir bakıma verinin temizlenmesidir. Ayrıca modellenmesi, sınıflandırılması hatta filtrelenmesidir [52].

2.4.1.3. Genelleştirme

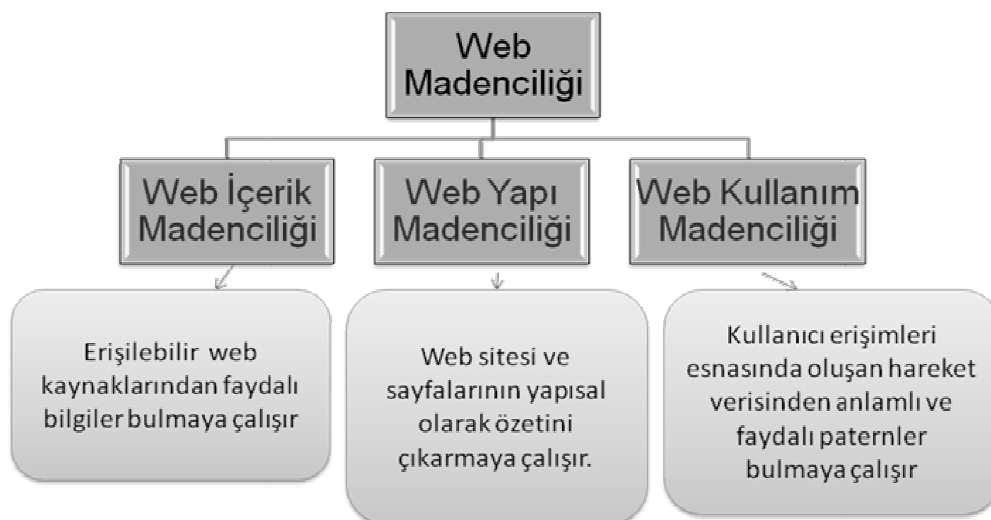
Genelleştirmede, elde ettiğimiz tecrübenin genelleştirilmesi ve bu tecrübeler üzerinden genel-geçer kurallar türetilmesi amaçlanmaktadır.

2.4.1.4. Çözümleme (Analiz)

Çıkarılan kurallar yardımı ile eldeki verilerden anlamlı sonuçlar çıkarılması ve ileriye yönelik tahmin ve değerlendirmelerin yapılmasıdır.

2.4.2. Web madenciliği yöntemleri

Web madenciliğinde kullanılan yöntemler; Şekil 2.15'te görüldüğü üzere, Web İçerik Madenciliği, Web Yapı Madenciliği ve Web Kullanım Madenciliği olarak 3 ana grupta incelenmektedir.



Şekil 2. 15 Web Madenciliği Yöntemleri [53]

2.4.2.1. Web içerik madenciliği

Yapay zekâ, akıllı yazılımlar ve bilgi tarama teknikleri kullanılarak web kaynaklarının içeriklerinden (metin, resim, ses, görüntü, metadata ve hiperlinkler) yararlı bilgiyi elde etmek olarak tanımlanabilir. Bu bilgiler “Web Crawler” olarak adlandırılan çeşitli yazılımlar (robot, örümcek vb.) tarafından özellikle gelişmiş arama motorlarının web içeriğini indekslemesi aracılığıyla toplanır. Web üzerindeki verinin çok farklı yapıda karşımıza çıkması, web içerik madenciliği için geliştirilecek uygulamaların gerçekleşmesini zorlaştırmaktadır. Web içerik madenciliğin amacı, bu kaynaklar arasından istenen bilginin bulunması veya filtrelenmesidir. Web içerik madenciliğinde yaygın olarak kullanılan iki yaklaşım mevcuttur [52,53]:

Bilgi çıkarımı yaklaşımı (information retrieval approach-IR):

Kullanıcı profili temel alınarak kullanıcılara gösterilen bilgileri filtrelemek ve bilgiye erişimi geliştirmek için kullanılan yöntemdir [52,53].

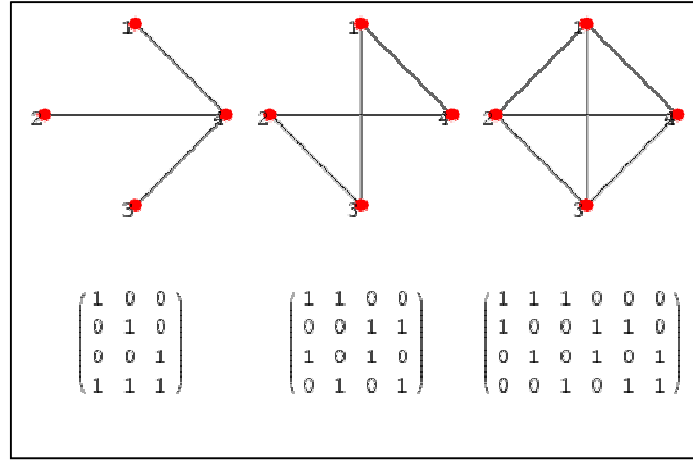
Veritabanı yaklaşımı (database approach):

Web ‘deki veriyi modellemek ve veriyi bütünleştirerek daha karmaşık bir yapıya sokmak için kullanılan yöntemdir. Bu yöntem sayesinde anahtar kelime tabanlı arama yerine daha gelişmiş sorgular çalıştırmak mümkün olmaktadır [52,53].

2.4.2.2. Web yapı madenciliği

Web siteleri ve web sayfaları arasındaki bağlantıların incelenmesiyle gerçekleştirilmektedir. Web yapı madenciliği düğümler arası bağlantılara odaklanır. Düğümler arasındaki bağların durumuna göre farklı düğümler arasındaki benzerlik ve ilişki gibi bilgileri üretir. Web yapı madenciliğinde ağlar yönlü çizge şeklinde ifade edilirler. Ağ üzerindeki her bir eleman bir düğüm, elemanlar arasındaki ilişkiler ise yönlü ve/veya ağırlıklı ayrıtları temsil eder. Çizge yapısının oluşturulmasıyla birlikte çizge teorisindeki bağıntılardan da yararlanılarak özellikle benzerlik ile ilgili olarak komşuluk matrisleri (Şekil 2.16) çıkarılarak çeşitli analizler yapılabilir (Benzer arama sonuçları, atıf analizi vb.)

[2,5,24,53]. Google arama motoru da bir web yapı madenciliği uygulaması olan HITS (Hyperlink-Induced Topic Search) algoritmasını kullanarak benzer sayfaları sonuçlarda göstermektedir [54].



Şekil 2. 16 Yönlü Çizgeler ve Komşuluk Matrisleri [55]

2.4.2.3. Web kullanım madenciliği

Bu yöntemde ihtiyaç olunan veriler internet mecrası üzerindeki çeşitli sunucularda kayıt altına alınmış olan kullanıcılara ait işlem ve oturum bilgilerinin yer aldığı log dosyalarından elde edilir. Burada amaçlanan ziyaretçilerin siteyi kullandıktan sonra geride bıraktığı erişim kayıtlarından veri üretmektir. Bu verilere örnek olarak kullanıcı adı, ip adresi, fiziksel bağdaştırıcı adresi, bağlantı saati, oturum süresi gibi bilgiler verilebilir ve söz konusu veriler kullanıcının isteği dışında ve zorunlu olarak oluşmaktadır. Log dosyalarının analiz edilmesi ile kullanıcılar hakkında detaylı bilgi ve çıkarımlara ulaşılabilir, kişiye özel içerik sunma, teklif gönderme, kişiye özel reklamcılık gibi uygulamalar geliştirilebilir, kullanıcının eğilimleri belirlenebilir ve ileriye yönelik tahminler üretilebilir. Bu sayede özellikle ticari alanlarda firma sahiplerine büyük avantaj sağlanmaktadır. Ayrıca kullanıcı kayıtlarının elde edilen geri beslemeler doğrultusunda tasarımda kişiselleştirmeler ve iyileştirmeler yapılabilir ve kullanılabilirliğin artırılması sağlanabilir [2,5,24,53].

2.5. Bibliyometri

Bibliyometri terimi matematiksel ve istatistiksel yöntemlerin kitaplar ve diğer basılı yayın ortamlarına uygulanması olarak tanımlanmaktadır. Bibliyometrik araştırmalarda belgelerin ya da yayınların belirli özellikleri çözümlenerek bilimsel iletişime ilişkin çeşitli bulgular elde edilmektedir. Araştırmacıların farklı nedenlerle bibliyometrik araştırmalara yöneldiği bilinmektedir. Bibliyometrik araştırmalarla bir yandan herhangi bir konudaki en verimli araştırmacılar belirlenirken, diğer yandan da bunlar arasındaki etkileşimin boyutları gözler önüne serilebilmektedir. Bibliyometrik araştırmalar, benzer bir yaklaşımla çeşitli konularda ülkeler arasında, kurumlar arasında ya da ekoller arasında karşılaştırmalar yapılmasına da olanak sağlamaktadır [56].

Bibliyometrinin ilgilendiği konuların başında yer alan "atıf yapma" (citation) kavramı matbaanın bulunmasından sonra, Rönesans döneminde gelişmiştir. Dipnotların ve referansların kullanım tarihi kesin olarak bilinmemekle birlikte, bu konuda Oxford English Dictionary'de verilen ilk örnek William Savage'ın A Dictionary of the Art of Printing (1841) adlı çalışmasıdır. Dipnota benzeyen örneklerin ilk kullanım tarihi 16. yüzyıla kadar gitmektedir [56].

Koehler, bibliyometri ile uğraşan kişilerin en azından dört gruba ayrılabilceğini belirtmekte ve bu grupları;

- a) Atıf analizi üzerine çalışanlar,
- b) Ortak atıf (co-citation) analizi üzerine odaklananlar,
- c) Kişilerin, kurumların ya da ülkelerin verimliliği ile ilgilenenler,
- d) Kitap, makale, patent gibi bilgi ürünleri ile ilişkili çalışmalar yapanlar olarak belirtmektedir [57].

3. MATERYAL ve YÖNTEM

3.1. Ortak Atıf Analizi İle İçerikten Bağımsız Olarak Benzerlik Tahmini

Akademik yayınların ve bu yayınlar arasındaki atıf alma veya atıf yapma ilişkilerinin oluşturduğu etkileşim kümesi bir sosyal ağ ortamı olarak ele alınabilir. Bu sayede çeşitli veri inceleme ve analiz teknikleri vasıtası ile mevcut veri kaynakları üzerinden yazarlar ve yayınlar hakkında detaylı bilgilere ulaşılabilmektedir.

Bu noktada atıf analizi; akademik yayınların, kaynakçalarının çeşitli teknikler aracılığıyla incelenerek yayının yazarı, konusu, ilgi alanı, etki sahası vb. hakkında çeşitli çıkarımlara ulaşmak amacıyla kıyaslanması olarak tanımlanabilir. Rubin ve Garfield'a göre ise atıf analizi; makale ve kitaplardaki atıflara ait frekans ve örüntülerin çıkarılması ve bu bilgilerle oluşturulan çizgelerin incelenmesidir [93,94].

Literatürde ilk atıf analizi çalışmasının 1927 yılında, "Journal of the American Chemical Society" adlı kimya dergisinde yayımlanan makalelerin kaynakçalarının incelendiği araştırma olduğu belirtilmiştir. Bu araştırmaya dayanılarak, ABD'deki bir kolej kütüphanesi için dergi aboneliği ve eski sayıları satın alma politikası geliştirilmiştir [56].

Günümüzde internet teknolojisinin sağladığı imkânlar ile birlikte literatüre kazandırılan akademik dokümanlar, sayısal kütüphaneler olarak adlandırılan veritabanlarında saklanmaktadır (Otomatik İndeksleme). Bu veritabanlarından elde edilen veri setleri sayesinde, bir dokümana başka yayınlar tarafından yapılan tüm atıflara ve söz konusu dokümanın diğer yayınlara yaptığı tüm atıflara hızla ulaşılabilmektedir. Bu veri setlerinden elde edilen bilgiler ışında ise yayın kümeleri üzerinde matematiksel ve istatistikî hesaplamalar göreceli olarak kısa sürede ve yüksek doğrulukla gerçekleştirilebilmektedir [7,92].

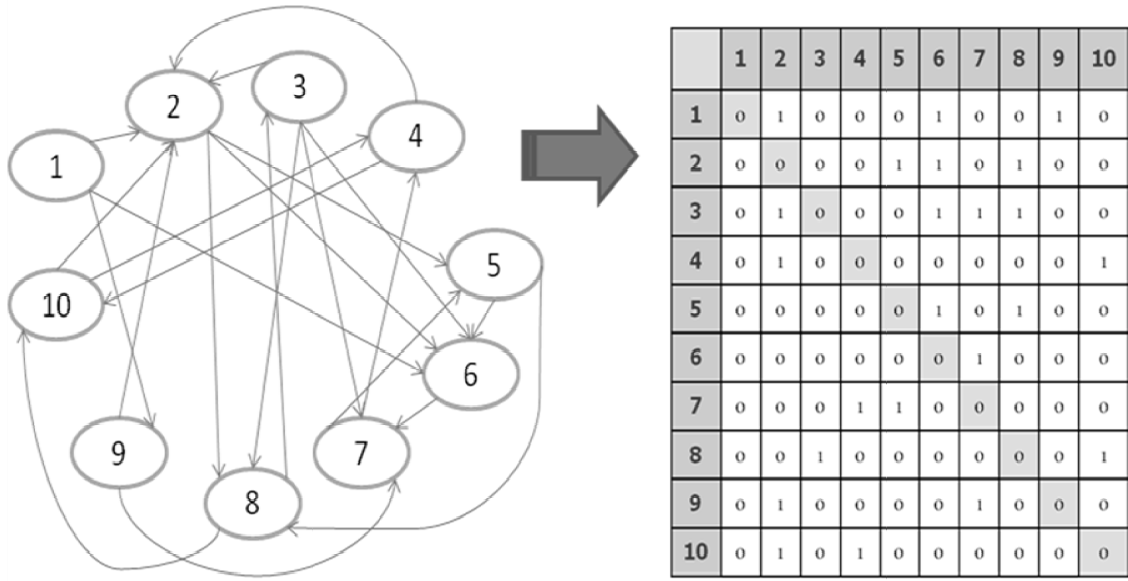
Otomatik atıf indeksleme [95] sistemleri sayesinde milyonlarca sayıdaki akademik doküman üzerinden büyük ölçekli örüntülerin incelenmesi ve bilgi keşfi için analiz yapılabilmektedir. Otomatik atıf indeksleme sistemlerinin ilk

örneđi olan CiteSeer'a [96] ilave olarak Google Scholar [97], Arxiv [16] gibi büyük veritabanları günümüzde sıklıkla kullanılmaktadır.

Atıf ilişkilerinin analizi esnasında, iki farklı akademik doküman arasında somut olarak bir ilişki bulunduğunun tespiti için gerekli en belirgin kriter bahse konu iki dokümanın benzer atıf alma veya benzer atıf yapma eğilimleridir. Yayınların ortak olarak başka dokümanlara yaptıkları veya aldıkları atıflar söz konusu iki dokümanın aynı konu hakkında yazılmış olabileceđi ihtimalini kuvvetlendirmektedir [7,8,92].

Konu benzerliđi kıstası ele alındığında iki akademik doküman arasında benzerlik olup olmadığının anlaşılması için ilgili dokümanların sadece başlık kısımlarının veya metin içeriklerinin karşılaştırılması muhtemelen başarılı sonuçlar vermeyecektir. Çünkü aynı konu, farklı kelime ve ifade şekilleriyle farklı biçimlerde ifade edilebilir. Ayrıca uzun bir metin dosyasını kelime kelime incelemek ve hangi konu hakkında olduğunu anlamak için gerekli sayısal işlemleri yapmak ve bunu başka bir doküman ile kıyaslamak kayda değer derecede hesaplama karmaşıklığını da beraberinde getirecektir. Bunun yerine akademik dokümanların içeriđi ile ilgilenmeden yalnızca bu dokümanlar arasındaki atıf ilişkilerine bakarak farklı iki doküman arasındaki konu benzerliđi tespit edilebilmektedir [7,8,92].

Atıf analizi yöntemlerinde ortak olarak kullanılan matris hesaplama işlemleri, hesaplama karmaşıklığını arttırdığından sayısal ortamda bulunan devasa veri kümelerini analiz etme konusunda zaman zaman donanımsal açıdan yetersiz kalmaktadır. Buna rağmen benzerliklerin içerikten bağımsız olarak sadece dokümanlar üzerindeki atıflara göre bulunması hususu söz konusu yöntemleri içeriđe bağılı benzerlik yaklaşımlarına göre bir adım öne çıkarmaktadır [7,92].



Şekil 3. 1 Makaleler ve Komşuluk Matrisi [7]

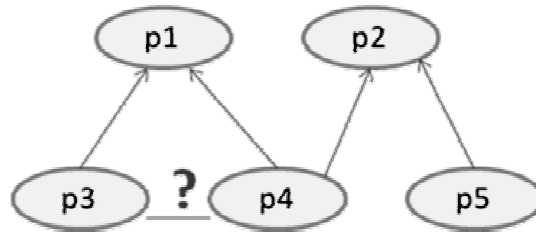
Akademik dokümanlardan oluşan bir makale havuzunu bir sosyal ağ ortamı olarak ele aldığımızda (Şekil 3.1); her bir makale bu ağ içerisindeki düğümleri, makaleler arasındaki atıf ilişkileri ise düğümler arasındaki yönlü ayrıtları temsil etmektedir. Bu sayede söz konusu ağ yapısına ilişkin bir yönlü çizge türetilmekte ve bu çizge üzerinden bir komşuluk matrisi oluşturulabilmektedir. Bu matris üzerinden yapılacak hesaplamalar aracılığıyla ise iki farklı makalenin ne kadar benzer oldukları yönünde bir tahminde bulunabilmektedir [7,8].

3.2. Mevcut Yöntemler

Literatürde atıf analizine dayalı benzerlik tahminine yönelik olarak geliştirilen en yaygın olarak kullanılan yöntemler, Bibliyografik Eşleştirme (Bibliocoupling) [10], Müşterek Atıf Analizi (Co-Citation) [11], Amsler [12], SimRank ve rvs-SimRank [13], P-Rank [14] ve Inter-Connection [15] yöntemleridir.

3.2.1. Bibliyografik eşleştirme (biblio-coupling)

Bibliyografik eşleştirmede temel fikir Şekil 3.2 yardımı ile açıklanabilir. Eğer p3 ve p4 gibi iki doküman, pek çok başka dokümana ortak olarak atıf yapıyorsa, yani p3 ve p4'ün kaynakça kısmında çok sayıda ortak referans bulunuyorsa, p3 ve p4 dokümanları arasında güçlü bir ilişki veya benzerlik olduğu değerlendirilebilir. Ne kadar fazla sayıda dokümana ortak olarak atıf yaparlarsa aralarındaki ilişki de o derece güçlü olacaktır [9].



Şekil 3. 2 Bibliyografik Eşleştirme [8]

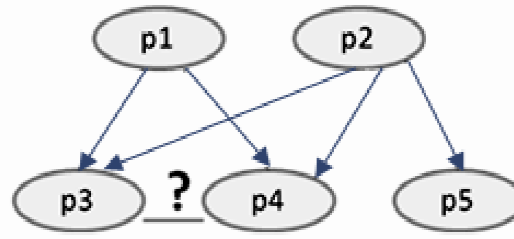
Dokümanlar arasındaki atıf ilişkisine bakılarak eğer i dokümanından j dokümanına bir atıf yapılmışsa $L_{ij}=1$, yapılmamışsa $L_{ij}=0$ olacak şekilde bir "L" atıf matrisi (adjacency-komşuluk matrisi) oluşturulur. Bulunmak istenen ve i ile j dokümanlarının ortak atıf yaptıkları doküman sayısı olan B_{ij} ise;

$$B_{ij} = \sum_{k=1, \dots, n} L_{ik} L_{jk} = (LL^T)_{ij} \quad (3.1)$$

eşitlik (3.1) yardımı ile hesaplanır [8]. Burada n toplam doküman sayısıdır. Elde edilen B kare matrisi ise bibliyografik eşleştirme matrisidir. Eşitlik (3.1)'de görülen matris çarpım işlemi hesaplama açısından zorlayıcı olsa da iç içe yaratılan döngüler ile matris çarpımı yapılmadan sonuç matrisinin hücre değerleri hesaplanabilmektedir.

3.2.2. Müşterek atıf analizi (co-citation)

Müşterek atıf analizinde ise temel fikir bibliyografik eşleştirmeden biraz farklıdır. Burada Şekil 3.3'den de anlaşılacağı üzere p3 ve p4 gibi iki dokümana ne kadar çok sayıda ortak olarak atıf yapan başka dokümanlar mevcut ise p3 ve p4 dokümanları arasındaki benzerlik o derecede kuvvetli olacaktır [11].

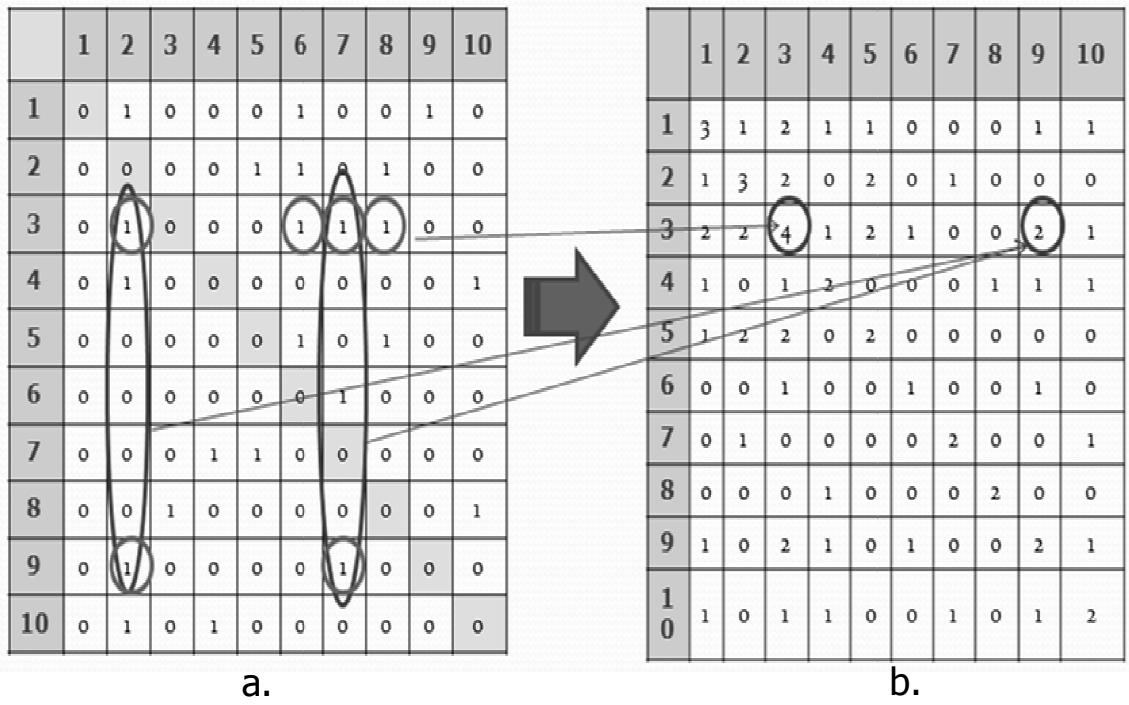


Şekil 3. 3 Müşterek Atıf Analizi [8]

Aynı şekilde “L” atıf matrisi oluşturulduktan sonra aradığımız ve i ile j dokümanlarına ortak atıf yapan doküman sayısı C_{ij} ise;

$$C_{ij} = \sum_{k=1, \dots, n} L_{ki} L_{kj} = (L^T L)_{ij} \quad (3.2)$$

eşitlik (3.2) yardımı ile hesaplanır [8]. Bu hesaplama sonucunda elde edilen C matrisindeki değerler ile komşuluk matrisi arasındaki ilişki Şekil 3.4’de görülmektedir.



Şekil 3. 4 Komşuluk (a.) ve Müşterek Atıf (b.) Matrisleri Arasındaki İlişki [7]

Her iki yöntemde de elde edilen matrislerdeki değerler;

$$S_{ij} = \frac{C_{ij}^2}{\min(C_{ij}, C_{ij}) \times \text{mean}(C_{ij}, C_{ij})} \quad (3.3)$$

eşitlik (3.3) yardımıyla ile 0-1 aralığında normalize edilerek S benzerlik matrisi elde edilir [9]. Burada seçilecek bir α eşik değerinden büyük olan ihtimaller kabul edilebilir. Eşik değeri ne kadar büyük seçilirse eşlenecek doküman sayısı azalacak ancak eşleşmelerin doğru olma ihtimali artacaktır. Aksi durumda ise tam tersi geçerli olacaktır. Şekil 3.5’de bu yöntem ile normalize edilmiş benzerlikleri içeren S benzerlik matrisi görülmektedir.

	1	2	3	4	5	6	7	8	9	10
1	1	0.11	0.38	0.2	0.2	0	0	0	0.2	0.2
2	0.11	1	0.38	0	0.8	0	0.2	0	0	0
3	0.38	0.38	1	0.17	0.67	0.4	0	0	0.67	0.17
4	0.2	0	0.17	1	0	0	0	0.25	0.25	0.25
5	0.2	0.8	0.67	0	1	0	0	0	0	0
6	0	0	0.4	0	0	1	0	0	0.67	0
7	0	0.2	0	0	0	0	1	0	0	0.25
8	0	0	0	0.25	0	0	0	1	0	0
9	0.2	0	0.67	0.25	0	0.67	0	0	1	0.25
10	0.2	0	0.17	0.25	0	0	0.25	0	0.25	1

Şekil 3. 5 Normalize Edilmiş S Matrisi ($\alpha=60$) [7]

3.2.3. Amsler yöntemi

Akademik yayınlar bir zaman cetveli üzerinde kronolojik olarak dizili olduklarından bir makale ancak kendisinden sonra yazılmış bir makale tarafından atıflanabilir. Aynı şekilde bir makalenin atıf yaptığı makaleler de kendinden önce yazılmış makalelerdir. Bu mantıktan hareketle bibliyografik eşleşmede kıyaslanan iki yayından biri eski diğeri yeni ise doğal olarak yeni olanın atıf yapabilecekleri ortak doküman sayısı kısıtlanmaktadır. Bu durumun tersi ise müşterek atıf analizi için geçerli olmaktadır. Bu dezavantajı ortadan kaldırmak amacıyla Robert AMSLER tarafından sunulan bu yöntemde ise hem bibliyografik eşleştirme hem de müşterek atıf analizi yöntemlerinin belirli oranlar dâhilinde ortak olarak hesaba katılması ile benzerlikler bulunmaya çalışılmıştır [12].

3.2.4. Diğer yöntemler

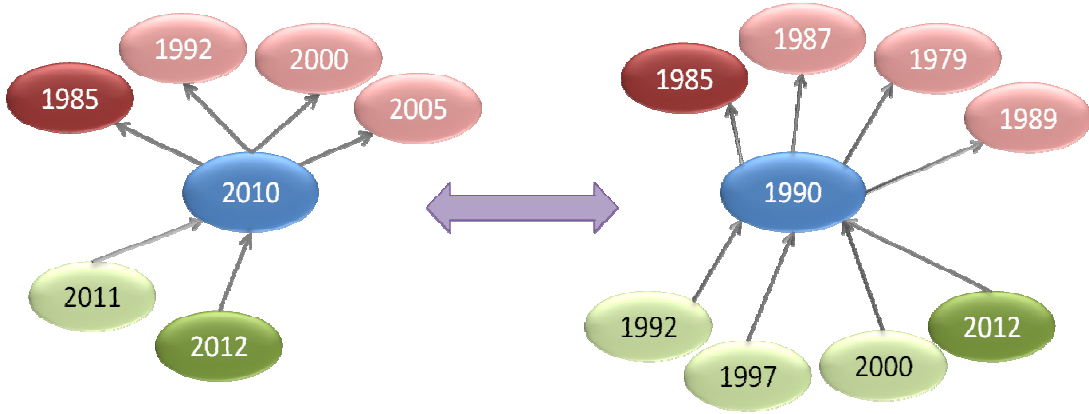
Literatürde karşımıza çıkan ortak atıf analizi yöntemlerinden sıkça kullanılan diğer yöntemlerden SimRank, rvs-SimRank ve P-Rank sırasıyla bibliyografik eşleme, müşterek atıf analizi ve Amsler'in daha hassaslaştırılması

iin benzerliklerin iteratif olarak hesaplaması zerine kurulmuř yntemlerdir [13,14]. Inter-Connection ynteminde klasik iki yaklařıma ek olarak kıyaslanan iki dokmandan birinci dokmanın atıf yaptıėı dokmanların ka tanesinin ikinci dokmana atıf yaptıėı kriteri aranmaktadır [15]. Literatrdeki yntemlerin tamamında temel fikir, ortak olarak alınan ve yapılan atıflar zerine yoėunlařmaktadır. Bu sebeple tm yntemlerin temelinde bibliyografik eřleřtirme ve mřterek atıf analizi yer almaktadır.

3.3. Problemin Tanımlanması

Atıf kavramı doėası gereėi zamana baėlı bir kavramdır. Bir dokmanın diėer bir dokmana atıf yapabilmesi iin o dokmandan daha sonra yazılmıř olması gerekmektedir.

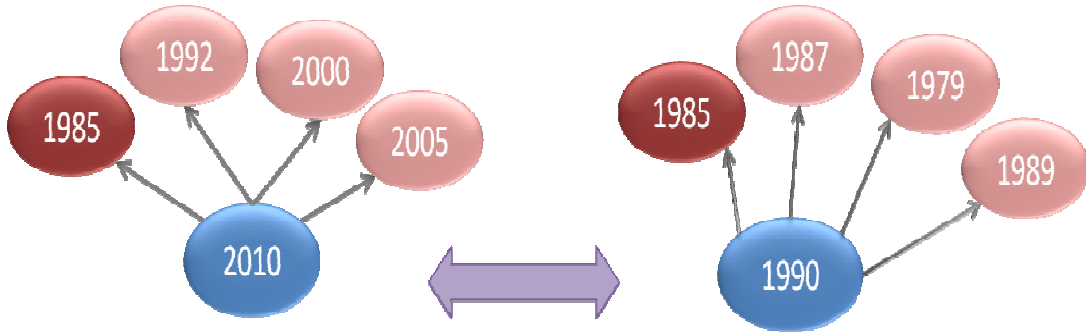
Birlikte ortak olarak atıf yapılan dokmanların sayısı veya ortak olarak atıf alınan dokman sayısı parametrelerine baėlı olarak gerekleřtirilecek analizler aslında tam olarak doėru sonu vermeyecektir. řekil 3.6'da da grldė zere kıyaslanan iki dokmandan biri yakın zamanda, diėeri ise ok daha eski tarihlerde yayımlanmıř olduėunda, doėal olarak yeni yayımlanmıř olan dokmana atıf yapan yayınların sayısı kısıtlı olacaktır. Aynı zamanda eski dokmanın da, yayımlandıėı dnem iinde o konu ile ilgili az sayıda yayın bulunduėundan literatrde atıf yaptıėı dokman sayısı kıyaslandıėı dokmana gre az olacaktır. Dokmanlar arasındaki bu kronolojik fark bibliyografik eřleřtirme ve mřterek atıf analizi yntemlerinin yetersiz kalmasına yol amaktadır.



Şekil 3. 6 Zamana Bağlı Ortak Atıf Kıyaslaması

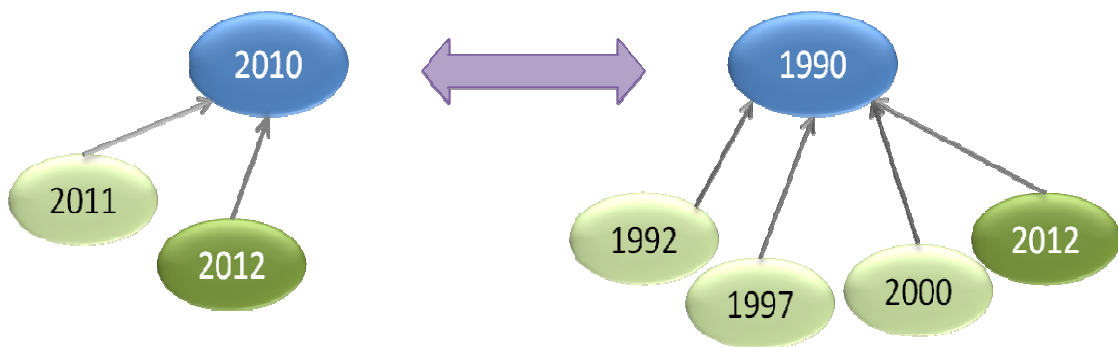
Bir önceki bölümde incelenmiş olan ve literatürde genel olarak kabul görmüş yöntemler ele alındığında söz konusu yetersizliğin bu yöntemler için de geçerli olduğu ve elde edilen sonuçların tam olarak gerçeği yansıtmayacağı anlaşılmaktadır.

Bibliyografik eşleştirme yönteminde temel parametre olarak, başka dokümanlara yapılan ortak atıf sayısı ele alınmaktadır. Doğal olarak Şekil 3.7’de de vurgulandığı üzere bu yaklaşımla sadece, kıyaslanan iki yayından eski olanın yayım tarihinden önceki yıllarda yayımlanan dokümanlara yapılan atıflar hesaba katılabilecektir (2010 ve 1990 yıllarında yayımlanan iki doküman için sadece 1990 yılından önce yayımlanan dokümanlar üzerinden ortak atıf ilişkisi incelenebilecek. Ör:1985). Bu durumda; yeni doküman ile eski doküman arasında ortak nokta bulunuyor olsa dahi, yeni dokümanın daha farklı konulara ait başka daha yakın tarihli yayınlara da atıf yapmış olması durumunda, yöntem sadece birkaç küçük ortak noktaya odaklanacak ve pek de doğru olmayabilecek bir benzerlik tespit edecektir. En azından tespit edilecek olan benzerlik, olması gereken değerin üzerinde tespit edilecektir.



Şekil 3. 7 Bibliyografik Eşleştirmede Zaman Kıyaslaması

Müşterek atıf analizi yönteminde ise kıyaslanan iki dokümana, başka dokümanlar tarafından yapılan ortak atıflar incelenmektedir. Şekil 2.8’de gösterilen bu yaklaşımla da yalnızca, kıyaslanan iki yayından yeni olanın yayım tarihinden sonraki yıllarda yayımlanan dokümanlar tarafından yapılan atıflar hesaba katılabilecektir (2010 ve 1990 yıllarında yayımlanan iki doküman için sadece 2010 yılından sonra yayımlanan dokümanlar üzerinden ortak atıf ilişkisi incelenebilecek. Ör:2012). Benzer bir mantıktan hareketle, yeni doküman ile eski doküman arasında ortak nokta bulunuyor olsa dahi, eski dokümanın daha farklı konulara ait başka eski tarihli yayınlardan da atıf almış olması durumunda, yöntem yine olması gereken değer üzerinde bir benzerlik tespit edecektir. Başka bir deyişle, bulunan benzerliklerde “hatalı pozitiflik oranı” artacaktır.



Şekil 3. 8 Müşterek Atıf Analizinde Zaman Kıyaslaması

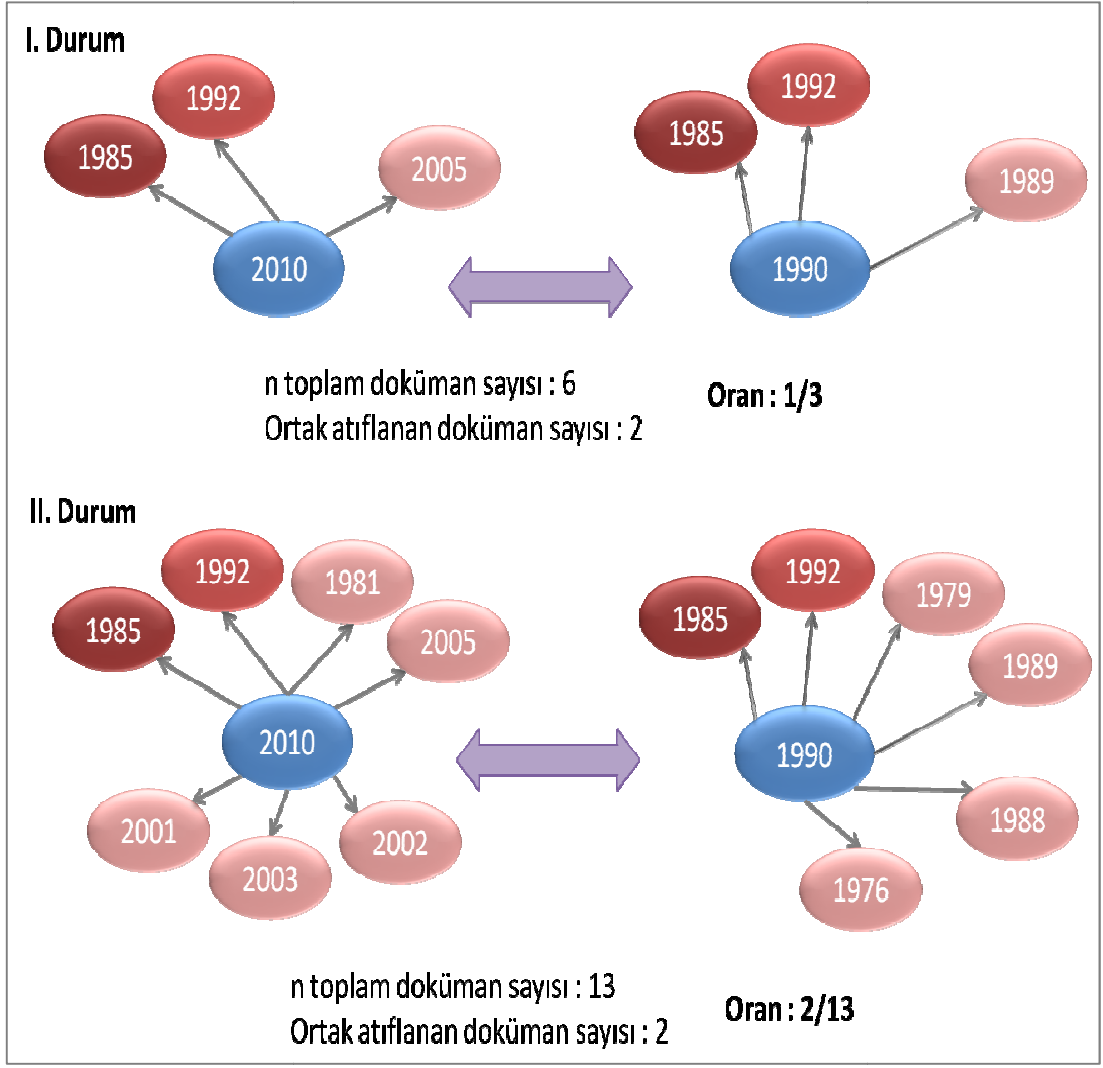
Amsler yöntemi önceki bölümde belirtildiği gibi, bibliyografik eşleştirme ve müşterek atıf analizi yöntemlerinin negatif yönlerinin ortadan kaldırılması amacıyla geliştirilmiştir. Buna rağmen her iki yöntemden elde edilen sonuçların

belirli bir oran dâhilinde hesaba katılması da her ne kadar belirli bir iyileştirme sağlasa da yukarı bahsedilen hatalı pozitiflik oranının düşmesine yeterli katkıyı sağlayamamaktadır. Çünkü temelde elde edilen değerlerde hata bulunduğu için bu değerlerin katsayılarının değiştirilerek hesaba katılması hatayı ortadan kaldırmamaktadır.

Inter-connection yönteminde ise diğer yöntemlerden farklı bir yaklaşım sunulmuş ve birinci dokümanın atıf yaptığı dokümanlardan kaç tanesinin ikinci dokümana atıf yaptığı kriteri hesaba katılmıştır. Bu sayede başarılı sonuçlar elde edilmiştir [15]. Ancak bu yöntemde de sadece ortak atıflar ele alınmış olup, kıyaslanan iki dokümanın farklı olarak atıfladıkları veya atıf aldıkları doküman potansiyeli göz önünde bulundurulmamıştır.

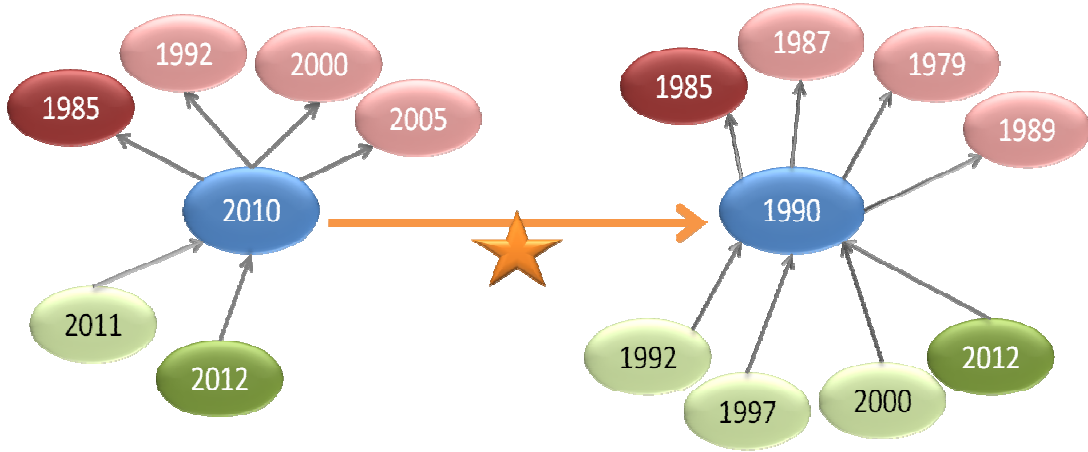
Tüm bu hususlar dikkate alındığında, önceki bölümde değinilen yöntemlerin tamamında yalnızca ortak olarak yapılan veya alınan atıflara odaklanıldığı ve söz konusu iki dokümanın birbirlerinden farklı olarak başka (ortak olmayan) dokümanlara yaptıkları veya aldıkları atıfların ele alınmadığı görülmektedir. Daha farklı bir deyişle sadece pozitif ödüle dayalı bir benzerlik yaklaşımında bulunulmuş ve negatif puanlama ile sistemin gerçeğe yakınsanması göz ardı edilmiştir.

Benzerlik tahmininde ortak olarak yapılan ve/veya alınan atıf miktarındaki artışın benzerliğe olan pozitif katkısı kadar, iki dokümanın birbirlerinden farklı olarak başka dokümanlara yaptıkları atıfların sayısı da söz konusu iki doküman arasındaki benzerliğe negatif etki edecek bir kriter olarak hesaba katılmalıdır. Veri setindeki toplam doküman miktarına kıyasla az bir sayıda ortak dokümana atıf yapmış ve aynı zamanda çok daha fazla sayıda farklı dokümana atıfta bulunmuş iki yayının aynı konu hakkında yazılmış olma ihtimali, farklı atıf sayıları ile ters orantılı bir ilişki içerisinde olacaktır. Şekil 3.9 incelendiğinde, I ve II'nci durumların ikisinde de toplam iki adet ortak olarak atıflanan doküman bulunmasına rağmen, iki durum arasında oluşacak benzerlik farklı çok yüzeysel bir oranlama ile dahi ortaya çıkmaktadır.



Şekil 3. 9 Ortak Atıflamada Farklı Atıfların Rolü

Bulunan benzerliklerin iyileştirilmesinde kullanılabilecek olan bir diğer parametre ise kıyaslanan iki akademik yayın arasında doğrudan atıf ilişkisi bulunması durumudur (Şekil 3.10). Söz konusu iki doküman arasında, herhangi birinden diğerine doğrudan atıf bulunması benzerlik ilişkisini kuvvetlendirecek bir durum olarak değerlendirilebilir. Dolaylı olarak başka dokümanlar üzerinden aradığımız benzerlik ilişkisi göz önünde bulundurulduğunda, dokümanlar arasında doğrudan atıf bulunmasının çok daha güçlü bir benzerlik işareti olacağı açıkça görülmektedir.



Şekil 3. 10 İki Yayın Arası Doğrudan Atıf İlişkisi

3.4. Önerilen Yöntem

Bir önceki bölümde de belirtildiği üzere yayınlar arasındaki ikili benzerliklerin hesaplanmasında mevcut yöntemlerin yetersiz kaldığı değerlendirilmiştir. Yayınlar arasındaki ikili benzerliklerin daha gerçeğe yakın hesaplanabilmesi için en yaygın iki yöntem olan ortak atıf eğilimlerine dayanan bibliyografik eşleştirme ve müşterek atıf analizi yöntemlerine ek olarak farklı atıf yapma eğilimleri ve iki doküman arasında doğrudan atıf ilişkisi bulunması parametreleri hesaba katılacak şekilde yeni bir yöntem oluşturulmuştur. Atıf-eşleme (Cite-match) olarak adlandırdığımız bu yeni yöntemde, karşılaştırılan iki akademik yayının benzerlik tahmininde negatif puanaj parametresi olarak kullanacağımız farklı dokümanlara atıf yapma kriterinin hesaplanması gerekmektedir. Bu hesaplama için veri setinden elde edilecek komşuluk matrisi üzerinden eşitlik (3.1) yardımı ile B bibliyografik eşleştirme matrisi elde edilir. Elde edilen B matrisine eşitlik (3.4) uygulandığında, matristeki her bir doküman ikilisi için bu ikilinin farklı olarak atıfladıkları doküman sayısı bulunabilir [92].

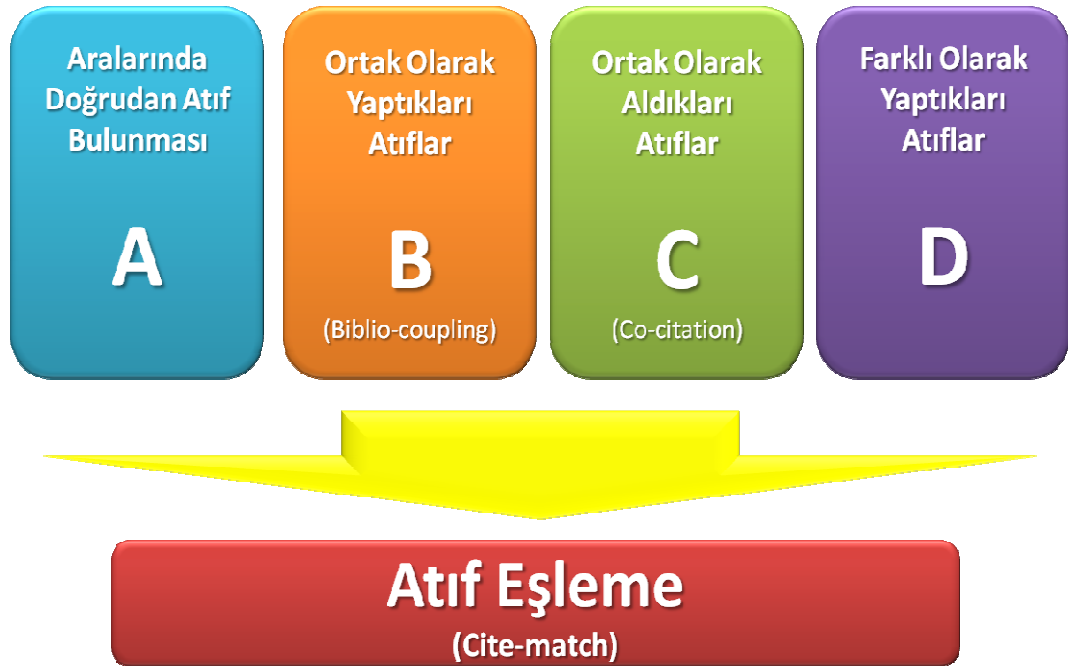
$$D_{ij} = B_{ii} + B_{jj} - 2B_{ij} \quad (3.4)$$

Ancak çalışma esnasında gerçekleştirilen hesaplamalar göstermiştir ki, büyük bir yayın kümesi içerisinde, farklı olarak yapılan atıfların sayısı ortak

olarak yapılan atıflara kıyasla daha fazla miktarda olduğundan söz konusu parametre, hesaplamalara büyük ölçekte negatif olarak etki etmekte ve negatif benzerlik sonuçlarına sebep olmaktadır. Bu problemin ortadan kaldırılması için farklı olarak atıf yapılan doküman sayısı yerine, eşitlik (3.5) yardımı ile farklı olarak atıf yapılan doküman sayısının toplam atıflanan doküman sayısına oranı daha uygun bir kıstas olarak kabul edilmiştir [92]. Böylece, fazla sayıda farklı olarak atıflanan doküman bulunması durumunda yüksek negatif puan çıkma durumu ortadan kalkmaktadır. Bu kıstasın hesaba dâhil edilmesi ile elde edilen benzerliklerde ortaya çıkabilecek hatalı pozitiflik oranının (false positive rate) düşeceği ve daha gerçeğe yakın sonuçlar elde edileceği değerlendirilmiştir.

$$D_{ij} = (B_{ii} + B_{jj} - 2B_{ij}) / (B_{ii} + B_{jj} - B_{ij}) \quad (3.5)$$

Ayrıca bir başka kriter olarak, karşılaştırılan iki doküman arasında doğrudan atıf bulunması hususu da dikkate alınmıştır. İki yayın arasında bulunan doğrudan atfın benzerlik tahmininde güçlendirici bir katkı sağlayacağı değerlendirilmiştir (i'den j'ye atıf varsa $A_{ij}=1$, yoksa $A_{ij}=0$ veya tam tersi).



Şekil 3. 11 Atıf Eşleme Yöntemi Bileşenleri

Bu kapsamda önerilen yöntemin matematiksel olarak formüle edilebilmesi için Şekil 3.11’de de gösterilen 4 ana kıstas hesaba katılmıştır. Bunlar sırası ile; A_{ij} , iki doküman arasında doğrudan atıf bulunup bulunmaması, B_{ij} , ortak atıf yapılan doküman sayısı (biblio-coupling), C_{ij} , ortak atıf alınan doküman sayısı (co-citation) ve D_{ij} , farklı olarak atıf yapılan doküman sayısının toplam atıflanan doküman sayısına oranıdır. İlgili ölçütler belirli bir katsayı dâhilinde hesaba katılmış ve bu uygulama için gerçekleştirilen denemeler sonucunda bu katsayılar sırasıyla toplamaları 1’e eşit olacak şekilde $\alpha=0,2$, $\beta=0,3$, $\gamma=0,3$ ve $\delta=0,2$ olarak belirlenmiştir. Bu kapsamda benzerlikler eşitlik (3.6) yardımı ile hesaplanmıştır [92]. D_{ij} , farklı olarak atıf yapılan doküman sayısının toplam atıflanan doküman sayısına oranı parametresinin doğrudan negatif olarak hesaba katılması yerine $(1 - D_{ij})$ değeri ile elde edilen pozitif ödülün formüle dâhil edilmesi aslında negatif puanaj fikrine hizmet ederken aynı zamanda ortaya çıkabilecek negatif sonuç değerlerini önlemektedir.

$$T_{ij} = \alpha A_{ij} + \beta B_{ij} + \gamma C_{ij} + \delta(1 - D_{ij}) \quad (3.6)$$

Geliştirilen yöntemin gerçekleştirilmesi amacıyla, veri madenciliğine yönelik uygulama geliştirme yarışmalarında [98] kullanılmış bilimsel bir veri seti olan, makalelerin yazarı, yayınlanma tarihi, konusu, özeti gibi bilgileri içeren detay veriye (Şekil 3.1) sahip ve yayınlar arasındaki atıf ilişkisini bir ayrıt listesi (Şekil 3.2) şeklinde bulunduran bir veri setinden faydalanılmıştır [17]. “Arxiv” akademik yayın veritabanından [16] 1992 ve 2003 yılları arasında yayımlanmış ve “Yüksek Enerji Fiziği (High Energy Physics)” ana konu başlığı altındaki 27.770 akademik yayını ve bunlara ilişkin 352.807 atıf ilişkisini içeren bu veri seti üzerinde hesaplamalar ve kıyaslamalar gerçekleştirilmiştir [17].

Bu veri setinin seçilmesindeki en önemli kıstas ise veri setinin oluşturulduğu Arxiv.org akademik indeksleme sisteminin güncel ve sıkça kullanılan bir akademik veritabanı olmasıdır. Ayrıca veri seti içerisinde dokümanları işaret eden kimlik numaraları ile indeksleme sistemi üzerinden dokümana ait detayların sorgulanabilmesi, her bir dokümana ait detay bilginin veri seti içinde bulunması (Şekil 3.12) ve hesaplamaların daha az karmaşıklıkla

yapılabilecek şekilde veri yapısının ayrıt listesi biçiminde olması (Şekil 3.13), hususları da bu noktada dikkate alınmıştır. Özellikle Şekil 3.13’de de görüldüğü gibi detay bilgi içerisinde dokümana ait alt konu (subj-class) bilgisinin bulunması, elde edilen benzerliklerin kıyaslanmasında ve konu grubuna göre sınıflandırma başarısının değerlendirilmesinde oldukça fayda sağlamıştır.

```
-----
\\
Paper: hep-th/9201001 (arxiv.org sitesindeki id numarası)
From: zuber@poseidon.saclay.cea.fr (C. Itzykson)
Date: Tue Dec 31 23:54:17 MET 1991 +0100 (37kb)

Title: Combinatorics of the Modular Group II: the Kontsevich
integrals (Başlık)
Authors: C. Itzykson and J.-B. Zuber (Yazarlar)
Comments: 46 pages
Subj-class: High Energy Physics - Theory; Quantum Algebra (Konu-
Alt Konu)
Journal-ref: Int.J.Mod.Phys. A7 (1992) 5661-5705
\\
We study algebraic aspects of Kontsevich integrals
as generating functions for intersection theory over moduli
space
and review the derivation of Virasoro and KdV constraints.
1. Intersection numbers
2. The Kontsevich integral
2.1. The main theorem
2.2 Expansion of Z on characters and Schur functions
2.3 Proof of the first part of the Theorem
3. From Grassmannians to KdV
4. Matrix Airy equation and Virasoro highest weight conditions
5. Genus expansion
6. Singular behaviour and Painlev'e equation.
7. Generalization to higher degree potentials
\\ (Özeti)
```

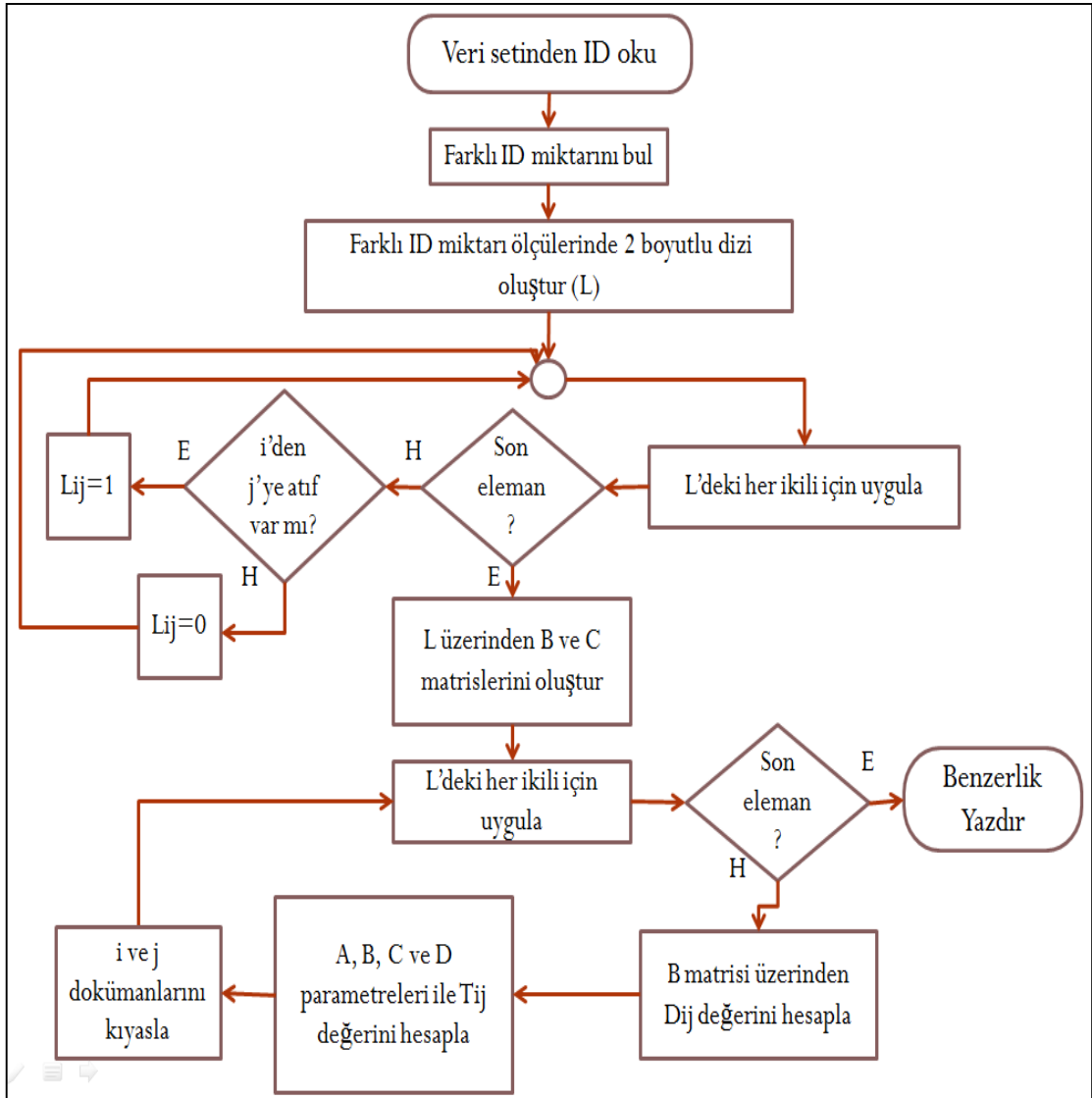
Şekil 3. 12 Kullanılan Arxiv Veri Setine Ait Meta Data Örneği

```
# Directed graph (each unordered pair of nodes is saved once):
Cit-HepTh.txt
# Paper citation network of Arxiv High Energy Physics Theory
category
# Nodes: 27770 Edges: 352807
# FromNodeId    ToNodeId
1001 9304045
1001 9308122
1001 9309097
1001 9311042
1001 9401139
1001 9404151
1001 9407087
1001 9408099
1001 9501030
1001 9503124
1001 9504090
1001 9504145
1001 9505025
1001 9505054
1001 9505105
1001 9505162
1001 9506048
1001 9506112
1001 9506144
1001 9507050
1001 9507158
1001 9508094
1001 9508155
1001 9510142
1001 9510225
```

Şekil 3. 13 Ayrıt Listesi

Bu çalışmada önerilmiş olan yöntemin veri setine uygulanabilmesi ve dokümanlar arasındaki benzerliklerin tespit edilebilmesi için kaynak kodu Ek-1’de sunulmuş olan uygulama geliştirilmiştir. Akış diyagramı Şekil 3.14’de verilen uygulama aracılığı ile öncelikle ayrıt listesinden okunan düğümler arası yönlü ilişkilerden bir komşuluk matrisi oluşturulmuş ve bu matris üzerinden (3.6) yardımı ile T_{ij} iki benzerlikleri hesaplanmıştır. Geliştirilen yöntem ile mevcut yöntemlerin kıyaslanabilmesi amacıyla veri seti üzerinden elde edilen komşuluk matrisine Bibliyografik eşleştirme, Müşterek atıf analizi ve Amsler yöntemleri de uygulanmıştır. Bu yöntemlerin uygulanmasıyla elde edilen benzerlik sonuçları, önerilen yöntem olan “Atıf Eşleme” (Cite-match) ile elde edilen sonuçlarla kıyaslanmıştır.

Uygulamanın kodlama aşamasında hesaplama süresinin kısaltılabilmesi amacıyla çok çekirdekli programlamaya olanak sağlayan bir paralel programlama kütüphanesi olan "OpenMP (Open Multiprocessing)" kullanılmış olup çalışmalar dört çekirdekli işlemciye sahip bir sunucu bilgisayar üzerinde gerçekleştirilmiş ve olumlu sonuç alınmıştır.



Şekil 3. 14 Akış Diyagramı

Bu noktada en önemli husus, elde edilen benzerlik değerlerinin sağlanmasının yapılmasıdır. Elde edilen benzerlik değerlerinin tam olarak sağlanmasının yapılabilmesi için en kesin çözüm benzer olarak tespit edilen iki dokümanın konu açısından kıyaslanmasıdır. Ancak şüphesiz ki yaklaşık 28.000

doküman ve bu dokümanlar arasındaki benzerlik ilişkisinin tek tek kontrol edilmesi pek mümkün olmamaktadır. Bu sebeple bulunan benzerliklerin sağlamasının ve mevcut yöntemler ile kıyaslamanın yapılabilmesi için veri setimizdeki dokümanların ana konu başlığı olan “Yüksek Enerji Fiziği” konusunda yaygın bir şekilde başvurulmuş bir ders kitabı olan kaynak referans olarak seçilmiştir [18].

Söz konusu kaynağın bölüm sonlarında bulunan atıflardaki yayınlar aynı alt konu grubunda olduklarından referans olarak kabul edilmişlerdir. Kaynaktan beş ana konu başlığı belirlenmiştir. (1-Standard Model and Weak Interactions, 2-Particle Cosmology, 3-Quantum Chromodynamics, 4-Particle Detectors ve 5-Inelastic Scattering). Belirlenen bu alt konu başlıklarına uygun Arxiv veritabanından referans dokümanlar seçilmiştir. Bu beş ana alt konu başlığı altında seçilen 504 adet referans dokümanın, söz konusu yöntemler tarafından hangi başarı yüzdesinde ikili olarak benzer tespit edildiği incelenmiştir. Dokümanların arasında bulunan benzerliklerden % 25’in üzerinde benzer olarak tespit edilen ikili benzerlikler kabul edilmiştir. % 25 eşik değeri belirlenirken aynı büyüklük ve derecedeki rassal bir çizgede yapılan benzerlik incelemesinde bulunan en büyük benzerlik değerinin % 12 olduğu görülmüş olup bu sebeple rassal benzerlikten daha yüksek bir ihtimale sahip bir kritere sahip olunması amacıyla bu değer seçilmiştir.

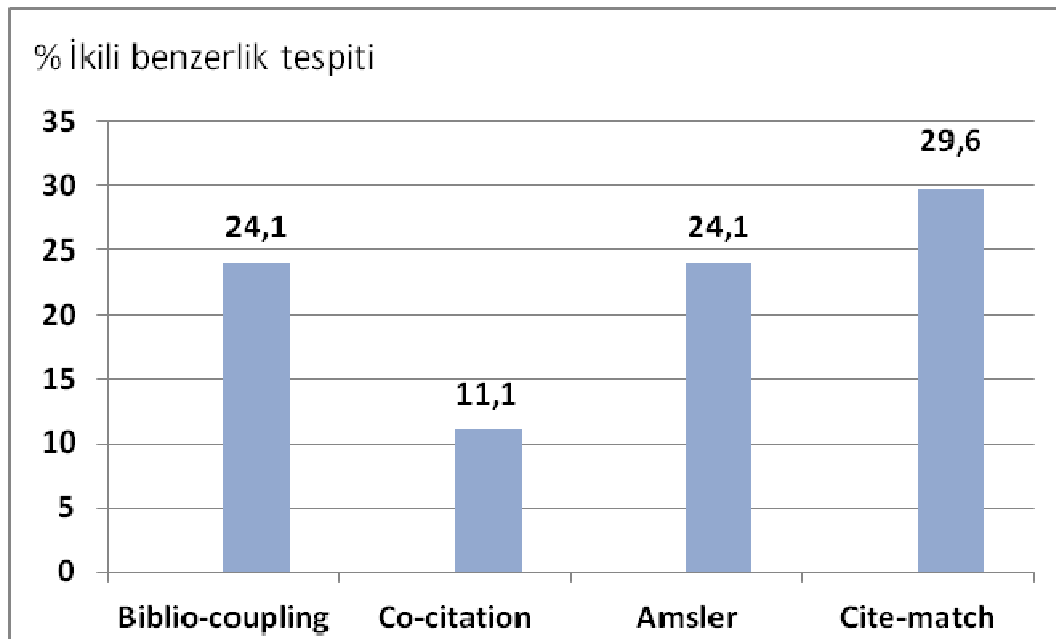
Ayrıca yalnızca ikili benzerlik kıyaslaması ile yetinilmeyip ikincil bir değerlendirme kriteri olan gruplama başarısı değerlendirmeye alınmıştır. Bu sayede yöntemin uygulanması sonucunda veri seti üzerinde oluşan benzerlik değerlerine ilişkin ilgili dokümanların hangi gruba daha benzer olduğuna işaret eden gruplamaların ne derecede başarıya ulaştığı incelenmiş ve sonuçlar kıyaslanmıştır. Daha farklı değerler de seçilerek hesaplamalar yapılabileceğinden söz konusu eşik değeri arttırıldığında ortaya çıkan gruplama başarısı da incelenmiştir.

4. ARAŞTIRMA BULGULARI

Yapılan çalışma sonucunda elde edilen bulgular iki farklı kategoride incelenmiştir. Bunlardan ilki, önerilen yöntemin ikili benzerlikleri (iki doküman arasındaki) tespit edebilme anlamında nasıl bir performans sergilediği, diğeri ise beş farklı doğal konu grubunda bulunan yayınların ne derece doğrulukla doğru konu grubuna dâhil edildiğidir.

4.1. İkili Benzerlik Tespiti

Geliştirilen uygulama yardımı ile ikili benzerlik tahminlerinin başarısının tespit edilebilmesi amacıyla "Arxiv High Energy Physics" [17] veri seti içerisinde referans olarak belirlenen kaynaktan seçilen [18] ve konu grupları daha önceden belirlenmiş olan toplam 504 adet dokümana, karşılaştırılması yapılacak olan yöntemler uygulanmıştır. Eşik değerinin üzerinde benzer tespit edilen doküman ikilileri incelenmiş ve aynı konu grubu içerisinde bulunup da yöntemler tarafından benzer tespit edilenler doğru eşleştirme olarak kabul edilmiştir. İlgili yöntemlerce dokümanların hangi başarı yüzdesi ile benzer olarak eşlendiği kıyaslanmış ve elde edilen sonuçlara ait grafik Şekil 4.1'de sunulmuştur.

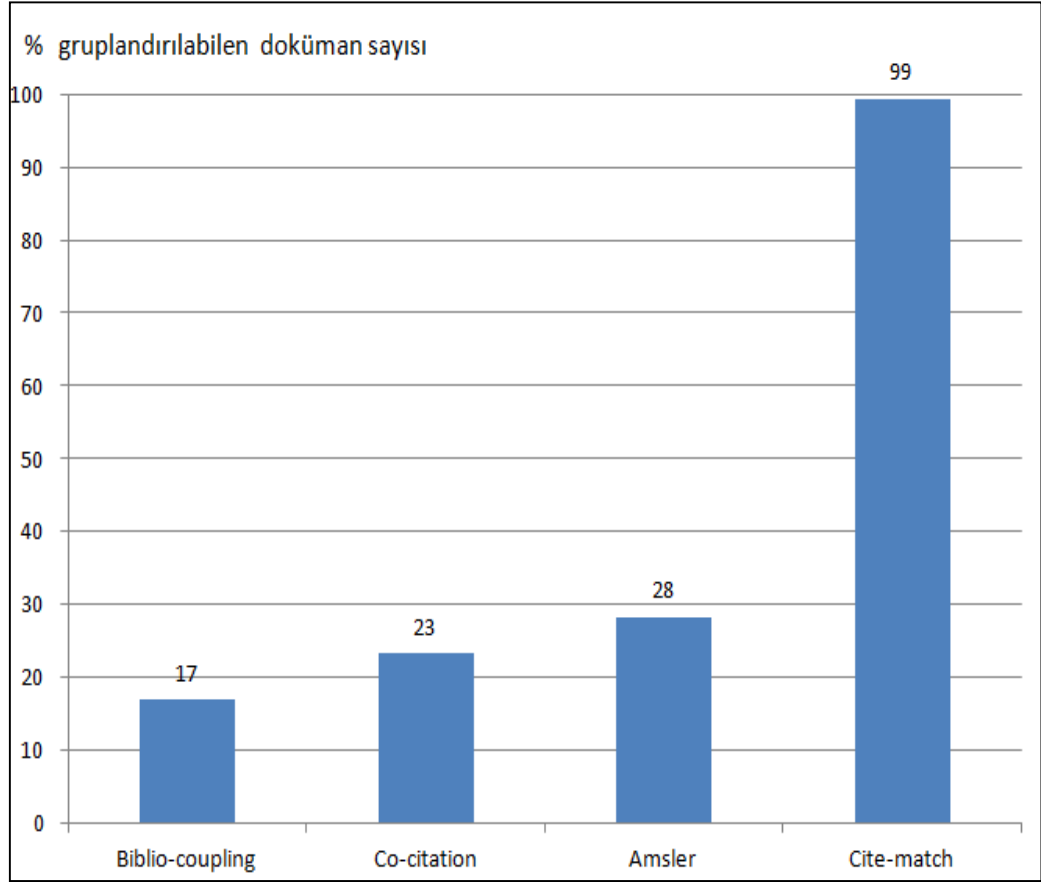


Şekil 4. 1 İkili Eşleştirme Başarısı

Bibliyografik eşleştirme (biblio-coupling) ve Amsler yöntemleri ile konu bakımından mevcut ikili benzerliklerin %24.1'i, Müşterek atıf eşleme (co-citation) ile %11.1'i tespit edilebilirken önerilen yöntem olan Atıf eşleme (Cite-match) yöntemi ile ikili benzerlikler diğer iki yöntemden %5.5'lik daha başarılı bir oranla tespit edilmiştir.

4.2. Gruplama Başarısı

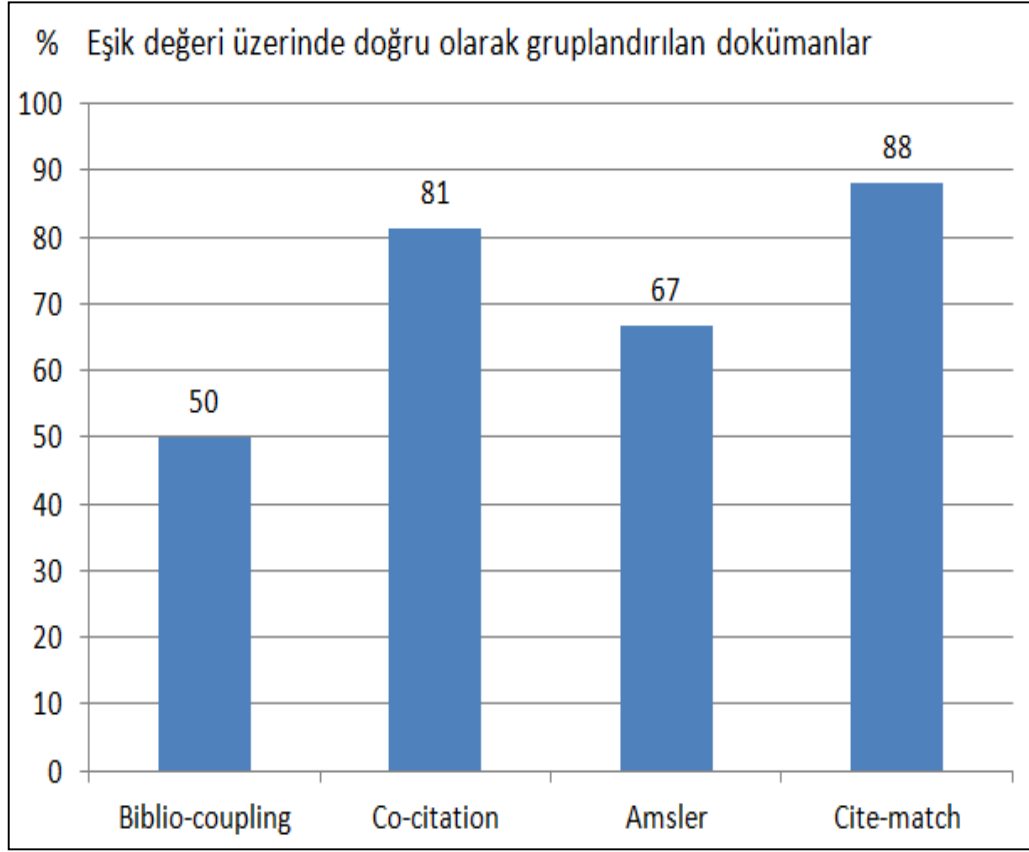
Önerilen yöntemin başarısının ve diğer metotlarla kıyaslamasının ikinci bir yöntemle sınanması için gruplama başarısı incelenmiştir. Referans olarak belirlenen kaynaktan seçilen ve beş ana alt konu başlığı içine giren 504 adet dokümanın önerilen yöntem ve mevcut yöntemler tarafından bulunan benzerlikler sonucunda ne derece doğru grupta tahmin edildiği değerlendirilmiştir. Farklı yöntemlerin uygulanması sonucunda veri seti üzerinde oluşan benzerlik değerlerine ilişkin, ilgili dokümanların hangi gruba daha benzer olduğuna işaret eden gruplamaların ne derecede başarıya ulaştığı araştırılmıştır. Bu kıyaslamaların yapılabilmesi için yöntemlerden elde edilen T_{ij} benzerlik matrisi üzerinde beş farklı konu başlığına göre dokümanlar işaretlenmiştir. Bu konu başlıkları altında bulunan her bir dokümanın diğer dokümanlarla olan benzerlik değerleri incelenmiş ve söz konusu dokümanın hangi gruptaki dokümanlara ait benzerliklerinde daha fazla puan aldığı dikkate alınmıştır. Tespit edilen benzerlik değerleri en yüksek puana sahip olan konu grubu dokümanın atandığı konu grubu olarak kabul edilmiştir. Bu sayede dokümanın farklı yöntemlere göre asli konu grubuna ne derece doğrulukla atandığı karşılaştırması yapılmıştır. Ayrıca gruplama başarısı farklı eşik değerlerine göre de incelenmiştir.



Şekil 4. 2 Gruba Atayabilme Başarısı (%)

Şekil 4.2’de de görüldüğü üzere önerilen yöntem, herhangi bir eşik değeri olmadan %99’luk bir başarı oranı ile neredeyse seçilen dokümanların tamamını bir gruba atayabilmiştir. Bu husus ağdaki düğümlerin her birinin en azından bir gruba dahil edilebilmesi anlamında önem taşımaktadır. Boşta kalan düğüm olmaması ağın fiziki yapısına ilişkin analizler açısından önem taşımaktadır.

Çalışmada esas alınan %25’lik eşik değeri değerlendirmeye katıldığında ise Şekil 4.3’te de görüldüğü üzere %88’lik bir başarı oranı ile doğru grupta yaparak kendisine en yakın yöntemin % 7 önünde performans göstermiştir.



Şekil 4. 3 Eşik Değeri Üzerinde Gerçekleşen Doğru Gruplama Başarısı (%)

5. TARTIŞMA VE SONUÇ

Sosyal ağlar, barındırdıkları sosyal aktörlerin ilişkilerine ait oldukça faydalı bilgiler ihtiva etmektedirler. Bu yapı ve ilişkilerin analiz edilmesi aracılığıyla yakınlıklar, benzeşmeler, eğilimler ve etkileşimler gibi verilere ulaşılarak ağdaki ilişkiler hakkında çeşitli yorumlara veya tahminlere varılabilir. Özellikle internet teknolojisinde yaşanan teknik ve kültürel gelişmeler ile sosyal ağlara ilişkin veriler somutlaşmış ve ölçülebilir hale gelmiştir. İnternet üzerinde gerçekleşen sosyal iletişim sonucu oluşan veri yığını çok büyük boyutlara ulaşmaktadır. Söz konusu devasa boyuttaki veri yığını içerisinde anlamlı bilginin çıkarılabilmesi için “Web Madenciliği” teknikleri kullanılmaktadır.

Gelişen sayısal bilgi platformları ile birlikte özellikle akademik dünya açısından literatüre kazandırılan devasa boyutlarda ve miktardaki yayının depolanması, tasniflenmesi ve eşleştirilmesi önemli bir bibliyografik araştırma konusu haline gelmiştir. Akademik yayınların ve bu yayınlar arasındaki atıfların oluşturduğu yapının bir sosyal ağ yapısı olarak ele alınması ve söz konusu atıfların dokümanlar arasındaki ilişkiyi temsil etmesiyle, yayınlar arasındaki benzerlik “Atıf Analizi” yöntemleri aracılığıyla içerikten bağımsız olarak tespit edilebilmektedir.

Literatürde, dokümanların içeriğine bakılmaksızın sadece aldıkları ve yaptıkları atıfların incelenip analiz edilmesi ile akademik yayınlar arasındaki konu benzerliklerinin tespit edilmesi, dokümanların konu gruplarına göre otomatik olarak ayrılmasının sağlanması ve aynı konu grubu üzerinde çalışan akademisyenlere çalışma arkadaşı tavsiye etme gibi amaçlar çerçevesinde çeşitli yöntemler geliştirilmiştir.

Bu tez çalışmasında sosyal ağ kavramı, sosyal ağ yapılarının web madenciliği teknikleri ile analizine ilişkin yöntemler, çizgelerde grup tespit yöntemleri, atıf analizi kavramı ve atıf analizine ilişkin yöntemler incelenmiş ve ortak atıf analizi ile içerikten bağımsız olarak akademik yayınlar arasındaki benzerliklerin tahmin edilmesine yönelik olarak literatürde mevcut olan yöntemlerden farklı bir yaklaşım olan “Atıf Eşleme” (Cite-match) yöntemi sunulmuş ve bibliyografik veri seti üzerinden yapılan analizler sonucunda

önerilen yöntemin mevcut metotlara kıyasla daha başarılı sonuçlar verdiği gözlemlenmiştir.

Uygulama sonucunda elde edilen veriler incelendiğinde genel olarak uygulamanın dokümanlar arasındaki ikili atıf ilişkisi üzerinden yapılan benzerlik tahmininde nispeten mevcut yöntemlere nazaran daha başarılı sonuçlar verdiği görülmektedir.

Doğru konu grubuna yerleştirme konusunda ise önerilen yöntemin diğer yöntemlere oranla daha yüksek gruplama başarısı gösterdiği tespit edilmiştir.

Yapılacak daha detaylı çalışmalar neticesinde; örneğin hakemli yayınların incelenmesi sürecinde, değerlendirme yapacak hakeme ulaşacak akademik yayının konusunu otomatik olarak tespit eden ve yayını konu hakkında görevlendirilecek olan ilgili hakemlere doğru bir şekilde yönlendiren bir otomasyon sistemi gibi daha kapsamlı uygulamalar geliştirilmesinin mümkün olacağı değerlendirilmektedir. İleriki çalışmalarda önerilen yöntem ile elde edilen benzerlik matrisindeki benzerliklerin çeşitli kümeleme metotlarının kullanılması ile yayın veritabanlarının daha somut grup yapılarına dönüştürülerek yazar eğilimleri, çalışma grupları, uzmanlık alanlarına göre gruplama ve bir alandaki uzmanların saptanması gibi daha detay analizlerin yapılabileceği değerlendirilmektedir.

6. KAYNAKLAR

- [1] A. Köksal, " Bilişim Terimleri Sözlüğü", Türk Dil Kurumu Yayınları, 126 s. 1981.
- [2] A. Baykal, C. Coşkun, "Web Madenciliği Teknikleri", Akademik Bilişim '09 –XI. Akademik Bilişim Konferansı Bildirileri s.797-800, Harran Üniversitesi-Şanlıurfa, Şubat 2009.
- [3] I.H. Ting, "Web Mining Techniques for On-line Social Network Analysis", s.696-700, IEEE 2008.
- [4] N. Tekin, S. Pulat , "Veri Madenciliği"
www.beykent.biz/dokumanlar/data_mining/web_madenciligi.ppt,
Erişim Tarihi 01.06.2011.
- [5] F. Gürcan, C. Köse, "Web İçerik Madenciliği Ve Konu Sınıflandırılması" Akademik Bilişim 2008 Çanakkale Onsekiz Mart Üniversitesi, Çanakkale, 30 Ocak - 01 Şubat 2008.
- [6] C. Ergün, "Web Madenciliği Teknikleri",
<http://www.bilyaz.com/index.php/author/celal-cagin-elgun/>, Erişim Tarihi: 26.05.2011.
- [7] O. Boy, A. Karcı, "Sosyal Ağların Web Madenciliği Teknikleri İle Analizi ve Ortak Atıf Analizi ile Benzerlik Tahmini", Sy:154-161, FEEB-11, Fırat Üniversitesi, Elazığ, 2011.
- [8] G. Xu, Y. Zhang, L. Li, "Web Mining and Social Networking", Web Information Systems Engineering and Internet Technologies Book Series Volume 6, 2011, syf: 89-107.
- [9] D. Greene, J. Freyne, B. Smyth and P. Cunningham, "An Analysis of Research Themes in the CBR Conference Literature", syf:18–43, 2008.
- [10] M. Kessler, "Bibliographic Coupling Between Scientific Papers," **Journal of the American Documentation**, Vol. 14, No. 1, pp. 10-25, 1963.
- [11] H. Small, "Co-citation in the Scientific Literature: A New Measure of the Relationship between Two Documents," **Journal of the American Society for Information Science**, Vol. 24, No. 4, pp. 265-269, 1973.

- [12] R. Amsler, "Application of citation-based automatic classification. Technical report," The University of Texas at Austin Linguistics Research Center, 1972.
- [13] G. Jeh and J. Widom, "SimRank: A Measure of Structural-Context Similarity," In Proc. Int'l. Conf. on Special Interest Group on Knowledge Discovery and Data, syf: 538-543, 2002.
- [14] P. Zhao, J. Han, and Y. Sun, "P-Rank: a Comprehensive Structural Similarity Measure over Information Networks," In Proc. Int'l. Conf. On Information and Knowledge Management, pp. 553-562, 2009.
- [15] S. Yoon, S. Kim, S. Park, "A Link-based Similarity Measure for Scientific Literature" , WWW'10 Proceedings of the 19th international conference on World wide web, syf:1213-1214, 2010.
- [16] <http://www.arxiv.org>, Cornell University, Eriřim Tarihi: 20.01. 2012.
- [17] "Arxiv High-energy Physics Theory Citation Network", <http://snap.stanford.edu/data/cit-HepTh.html>, Eriřim Tarihi: 20.01.2012
- [18] D. Perkins, Introduction to High Energy Physics, 4th Edition, Cambridge University Press, 2000, 440 syf.
- [19] "İnternetin Ağ Yapısı", http://en.wikipedia.org/wiki/Internet_backbone, Eriřim Tarihi: 20.01.2012.
- [20] N. Gürsaka1, Sosyal Ağ Analizi: PAJEK, UNICET ve GMINE Uygulamalı, Dora Basım Yayın: Bilgisayar Kitapları Dizisi, Bursa, 2009, 513 syf.
- [21] http://www.webopedia.com/TERM/S/social_network.html, Eriřim Tarihi: 27.01.2012.
- [22] D. Liben-Nowell. An Algorithmic Approach to Social Networks, Massachusetts Institute of Technology, 2005.
- [23] D. Tindall, B. Wellman, "Canada as social structure: Social network analysis and Canadian Sociology", **Canadian Journal of Sociology**, syf:265-308, 2001.
- [24] H. Kuduğ, "Anlamsal Web Tabanlı Kurumsal İş Ağının Analizi", Ege Üniversitesi Fen Bilimleri Enstitüsü Yüksek Lisans Tezi, 2011.
- [25] <http://www.aliunlu.com/2009/01/28/sosyal-ag-analizi-nedir/>, "Sosyal Ağ Analizi Nedir ?", Eriřim Tarihi: 30.01.2012.

- [26] R.A. Hanneman, M. Riddle, "Introduction to Social Network Methods", 2005, <http://www.faculty.ucr.edu/~hanneman/nettext/C10Centrality.html>, Erişim Tarihi: 02.03.2012.
- [27] P. Gül, <http://pinargul.wordpress.com/2011/05/01/sosyal-ag-analizi/> "Sosyal Ağ Analizi", Erişim Tarihi: 02.03.2012.
- [28] B. Özakar, H. Püskülcü, "Tıkların Dili", <http://www.teknoturk.org/docking/yazilar/tt000119-yazi.htm>, Erişim Tarihi: 02.03.2012.
- [29] S. Fortunato, Community detection in graphs. arXiv:0906.0612 (2009).
- [30] D. Halgin, "An Introduction to UCINET and NetDraw", NIPS UCINET and NetDraw Workshop 2008, Harvard University, syf:1–47, 2008.
- [31] www.netminer.com, Erişim Tarihi: 03.03.2012.
- [32] V. Batagelj, "Analysis of Large Networks with Pajek", 2008.
- [33] http://cdn.information_management.com/media/editorial/dmreview/200309/200309_068_2.gif, Erişim Tarihi: 12.03.2012.
- [34] W. W. Zachary, An information flow model for conflict and fission in small groups, **Journal of Anthropological Research** 33, 452-473 (1977).
- [35] <http://origin-ars.sciencedirect.com/content/image/1-s2.0-S0167865509003043-gr2.jpg>, Erişim Tarihi:12.03.2012.
- [36] J. Galaskiewicz and S. Wasserman, Advances in Social Network Analysis : Research in the Social and Behavioral Sciences, Introduction: Advances in the Social and Behavioral Sciences from Social Network Analysis, Sage Focus Editions, 171, Thousand Oaks, California, 1994, 299 syf.
- [37] R. Hanneman and M. Riddle, Introduction to social network methods, Riverside, University of California, 2005.
- [38] M. Gencer, C. Gunduz and V.S. Tunalıoğlu, 2009, CL-SNA: Social network analysis with Lisp, SPLASH Systems, Programming, and Applications: Proceedings of the 2007 International Lisp Conference, 17, ACM New York, USA, 9 syf.
- [39] T. Schank and D. Wagner, 2005, Approximating clustering-coefficient and transitivity, **Journal of Graph Algorithms and Applications**, syf: 265–275.

- [40] SONIVIS, "Social Network Analysis Global Clustering Coefficient Metrics", http://sonivis.org/wiki/index.php/Global_Clustering_Coefficient, Erişim tarihi: 24 Nisan 2011.
- [41] J. Scott, Social Network Analysis: A Handbook, London, Sage Publications, 2000, 209 syf.
- [42] P. Mika, 2007, Social Networks and the Semantic Web, Semantic Web and Beyond Computing for Human Experience, Jain,R. and Sheth, A.(Eds.), New York, **Springer Science and Business Media**, syf: 234.
- [43] <http://www.fmsasg.com/SocialNetworkAnalysis/>, Erişim Tarihi: 02.04.2012.
- [44] B. Ruhnau, 2000, "Eigenvector centrality: a node-centrality?", **Social Networks**, 22 (4): 357-365.
- [45] M.E.J. Newman, **Mathematics of Networks**, syf:12., The New Palgrave Dictionary of Economics, Durlauf, S.N. and Blume, L.E. (Eds.), Palgrave Macmillan, 2008.
- [46] http://en.wikipedia.org/wiki/Clustering_coefficient, Erişim Tarihi: 14.04.2012.
- [47] http://pasargad.cse.shirazu.ac.ir/~mhaji/ec2/GA_Graph/graph.gif, Erişim Tarihi: 14.04.2012.
- [48] <http://www.solver.com/xlminer/help/HClst/Clustering1.gif>, Erişim Tarihi: 14.04.2012.
- [49] <http://mikelove.files.wordpress.com/2010/04/spectral1.png>, Erişim Tarihi: 14.04.2012.
- [50] O. Etzioni, "The World Wide Web: Quagmire or Gold Mine", **Communications of the ACM**, 39(11):65–68, 1996.
- [51] <http://www.bretswanson.com/index.php/2009/03/internet-update>, Erişim Tarihi: 02.06.2011.
- [52] Y. Abanoz, "Veri Madenciliği Giriş ve İleri Konular", Bölüm 3, Gebze İleri Teknoloji Enstitüsü, 2010.
- [53] İ. Haberal, "Veri Madenciliği Algoritmaları Kullanılarak Web Günlük Erişimlerinin Analizi ", Başkent Üniversitesi Yüksek Lisans Tezi, 2007.

- [54] S. Brin, L. Page, "The Anatomy Of A Large-Scale Hypertextual Web Search Engine", **Computer Networks** (30):107–117, 1999.
- [55] Yönlü Çizgeler ve Komşuluk Matrisleri, http://mathworld.wolfram.com/images/eps-gif/AdjacencyMatrix_1001.gif, Erişim Tarihi: 17.04.2012.
- [56] U. Al, Türkiye'nin Bilimsel Yayın Politikası: Atıf Dizinlerine Dayalı Bibliyometrik Bir Yaklaşım, Doktora Tezi, Ankara, 2008.
- [57] W. Koehler, 2001, Information science as "Little Science": The implications of a bibliometric analysis of the Journal of the American Society for Information Science. **Scientometrics**, 51(1):117-132.
- [58] A. Pothen, Graph partitioning algorithms with applications to scientific computing, Tech. rep., Norfolk, VA, USA, 1997.
- [59] B.W. Kernighan, S. Lin, An efficient heuristic procedure for partitioning graphs, **Bell Syst. Tech. J.** 49 (1970) 291307.
- [60] L.R. Ford, D.R. Fulkerson, Maximal flow through a network, **Canad. J. Math.** 8 (1956) 399404.
- [61] J.B. MacQueen, Some methods for classification and analysis of multivariate observations, in: L.M.L. Cam, J. Neyman (Eds.), Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, University of California Press, Berkeley, USA, 1967, syf. 281-297.
- [62] M.E.J. Newman, M. Girvan, Finding and evaluating community structure in networks, **Phys. Rev. E** 69 (2) (2004) 026113.
- [63] J. Duch, A. Arenas, Community detection in complex networks using extremal optimization, **Phys. Rev. E** 72 (2) (2005) 027104.
- [64] S. White, P. Smyth, A spectral clustering approach to finding communities in graphs, in: Proceedings of SIAM International Conference on Data Mining, 2005, pp. 76-84.
- [65] M.E.J. Newman, From the cover: Modularity and community structure in networks, **Proc. Natl. Acad. Sci. USA** 103 (2006) 8577-8582.
- [66] S. Lehmann, L.K. Hansen, Deterministic modularity optimization, **Eur. Phys. J. B** 60 (2007) 83-88.

- [67] J. Ruan, W. Zhang, An efficient spectral algorithm for network community discovery and its applications to biological and social networks, in: ICDM'07: Proceedings of the 2007 Seventh IEEE International Conference on Data Mining, IEEE Computer Society, Washington, DC, USA, 2007, syf. 643-648.
- [68] B.D. Hughes, Random Walks and Random Environments: Random Walks, vol. 1, Clarendon Press, Oxford, UK, 1995.
- [69] M.B. Hastings, Community detection as an inference problem, **Phys. Rev. E** 74 (3) (2006) 035102.
- [70] R.G. Gallager, Low Density Parity Check Codes, MIT Press, Cambridge, USA, 1963.
- [71] M.E.J. Newman, E.A. Leicht, Mixture models and exploratory analysis in networks, **Proc. Natl. Acad. Sci. USA** 104 (2007) 9564-9569.
- [72] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, **J. Roy. Stat. Soc. B** 39 (1977) 1-38.
- [73] T.A.B. Snijders, K. Nowicki, Estimation and prediction for stochastic blockmodels for graphs with latent block structure, **J. Classification** 14 (1997) 75100.
- [74] K. Nowicki, T.A.B. Snijders, Estimation and prediction for stochastic blockstructures, **J. Am. Assoc.** 96 (455).
- [75] M. Mungan, J.J. Ramasco, Who is keeping you in that community? eprint arXiv:0809.1398.
- [76] U.N. Raghavan, R. Albert, S. Kumara, Near linear time algorithm to detect community structures in large-scale networks, **Phys. Rev. E** 76 (3) (2007) 036106.
- [77] G. Palla, I. Derényi, I. Farkas, T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, **Nature** 435 (2005) 814-818
- [78] I. Derényi, G. Palla, T. Vicsek, Clique percolation in random networks, **Phys. Rev. Lett.** 94 (16) (2005) 160202.
- [79] B. Adamcsek, G. Palla, I.J. Farkas, I. Derényi, T. Vicsek, Cfinder: Locating cliques and overlapping modules in biological networks, **Bioinformatics** 22 (8) (2006) 1021-1023.

- [80] I. Farkas, D. Ábel, G. Palla, T. Vicsek, Weighted network modules, **New J. Phys.** 9 (2007) 180.
- [81] S. Lehmann, M. Schwartz, L.K. Hansen, Biclique communities, **Phys. Rev. E** 78 (1) (2008) 016108.
- [82] R. Peeters, The maximum edge biclique problem is NP-complete, **Discrete Appl. Math.** 131 (2003) 651-654.
- [83] N. Du, B. Wang, B. Wu, Y. Wang, Overlapping community detection in bipartite networks, in: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, IEEE Computer Society, Los Alamitos, CA, USA, 2008, pp. 176-179.
- [84] J.M. Kumpula, M. Kivelä, K. Kaski, J. Saramäki, Sequential algorithm for fast clique percolation, **Phys. Rev. E** 78 (2) (2008) 026109.
- [85] J. Baumes, M.K. Goldberg, M.S. Krishnamoorthy, M.M. Ismail, N. Preston, Finding communities by clustering a graph into overlapping subgraphs, in: N. Guimaraes, P.T. Isaias (Eds.), **IADIS**, 2005, syf. 97-104.
- [86] T. Heimo, J.M. Kumpula, K. Kaski, J. Saramäki, Detecting modules in dense weighted networks with the Potts method, **J. Stat. Mech.** (2008) P08007.
- [87] P. Pons, Post-processing hierarchical community structures: Quality improvements and multi-scale view, eprint arXiv:cs/0608050.
- [88] F. Glover, Future paths for integer programming and links to artificial intelligence, **Comput. Oper. Res.** 13 (5) (1986) 533-549.
- [89] J.M. Kumpula, J. Saramäki, K. Kaski, J. Kertész, Limited resolution and multiresolution methods in complex network community detection, in: Noise and Stochastics in Complex Systems and Finance, in: **SPIE Conference Series**, vol. 6601, 2007.
- [90] P. Ronhovde, Z. Nussinov, A highly accurate and resolution-limit-free Potts model for community detection, eprint arXiv:0803.2548.
- [91] P. Ronhovde, Z. Nussinov, Multiresolution community detection for megascale networks by information-based replica correlations, **Phys. Rev. E** 80 (1) (2009) 016109.

- [92] O. Boy, A. Avcı, M. Ertr kler, "Ortak Atıf Analizi İle Akademik Yayınlarda İ erikten Bağımsız Olarak Benzerlik Tahminine Y nelik Yeni Bir Yaklaşım", ASYU-2012, Trabzon.
- [93] R. Rubin, (2010). Foundations of library and information science (3rd ed.). New York: Neal-Schuman Publishers. ISBN 978-1-55570-690-6.
- [94] E. Garfield, Citation Indexing - Its Theory and Application in Science, Technology and Humanities Philadelphia:ISI Press, 1983.
- [95] C. Giles, B. Lee, D. Kurt, S. Lawrence, (1998), "CiteSeer: an automatic citation indexing system.", Digital libraries 98 : the Third ACM Conference on Digital Libraries, June 23-26, 1998, Pittsburgh, PA (New York: Association for Computing Machinery): 89–98, DOI:10.1145/276675.276685, ISBN 0-89791-965-3, retrieved July 7, 2011.
- [96] Citeseer, <http://citeseerx.ist.psu.edu/index>, Eriřim Tarihi: 21.06.2012.
- [97] Google Scholar, <http://scholar.google.com.tr/>, Eriřim Tarihi: 21.06.12.
- [98] KDD Cup 2003, <http://www.cs.cornell.edu/projects/kddcup/index.html>, Eriřim Tarihi: 21.06.2012.

EKLER

EK-1 Uygulamaya İlişkin Geliştirilen C++ Kodu

```
#include <cstdio>
#include <cstring>
#include <cstdlib>
#include <fstream>
#include <iostream>
#include <iomanip>
#include <map>
#include <set>
#include <vector>

#include <omp.h>

std::set<int> filter;
std::set<int> keys;
std::vector<int> vkeys;
std::multimap<int, int> links;
std::map<int, int> ids;

template <typename T>
T **newMatrix(int N, T defaultValue) {
    T **matrix;

    matrix = new T*[N];
    for (int i = 0 ; i < N; i++) {
        matrix[i] = new T[N];
        memset(matrix[i], defaultValue, sizeof(T) * N);
    }

    return matrix;
}

template <typename T>
void deleteMatrix(T** dArray, int N) {
    for (int i = 0; i < N; ++i)
        delete[] dArray[i];
    delete [] dArray;
}

template <typename T>
void normalizeMatrix(T** matrix, T** output, int N) {
    for (int i = 0; i < N; ++i) {
        for (int j = 0; j < N; ++j) {
            float denominator = std::min(matrix[i][i], matrix[j][j]) *
(matrix[i][i] + matrix[j][j]) * 0.5f;
            // TODO: check division by zero ?
            if (denominator == 0)
                denominator = 0.1;
            output[i][j] = (matrix[i][j] * matrix[i][j]) / denominator;
        }
    }
}
```

```

template <typename T>
void printMatrix(char *fileName, T** matrix, int N) {
    std::ofstream output(fileName);
    // output << std::endl << "      |";
    // for (int i = 0; i < N; ++i)
    //     output << std::setw(7) << vkeys[i] << "|";
    // output << std::endl;
    // for (int i = 0; i < N; ++i) {
    //     output << std::setw(7) << vkeys[i] << "|";
    //     for (int j = 0; j < N; ++j)
    //         output << std::setw(7) << std::setprecision(3) <<
std::fixed << matrix[i][j] << "|";
        output << std::endl << ",";
        for (int i = 0; i < N; ++i)
            output << vkeys[i] << ",";
        output << std::endl;
        for (int i = 0; i < N; ++i) {
            output << vkeys[i];
            for (int j = 0; j < N; ++j)
                output << "," << std::setprecision(3) << std::fixed <<
matrix[i][j];
            output << std::endl;
        }
        output << std::endl;
        output.close();
    }

template <typename T>
void printSimilarities(char *fileName, T** matrix, int N, float k)
{
    std::ofstream output(fileName);
    for (int i = 0; i < N; ++i) {
        for (int j = i + 1; j < N; ++j) {
            if (matrix[i][j] >= k && filter.count(vkeys[i]) &&
filter.count(vkeys[j]))
                output << std::setw(7) << vkeys[i] << std::setw(10) <<
vkeys[j] << std::setw(7) << std::setprecision(3) << std::fixed <<
matrix[i][j] << std::endl;
        }
    }
    output.close();
}

template <typename T>
T _B(T** matrix, int N, int i, int j) {
    T value = 0;

    for (int k = 0; k < N; ++k)
        value += matrix[i][k] * matrix[j][k];

    return value;
}

template <typename T>
T _C(T** matrix, int N, int i, int j) {
    T value = 0;

```

```

        for (int k = 0; k < N; ++k)
            value += matrix[k][i] * matrix[k][j];

    return value;
}

template <typename T>
void _F(T **F, T** A, T** Bn, T** Cn, T** D, T a, T b, T c, T d,
int N) {
    for (int i = 0; i < N; ++i) {
        // HACK
        Bn[i][i] = Cn[i][i] = 1.0;
        #pragma omp parallel for
        for (int j = 0; j < N; ++j)
            F[i][j] = a * A[i][j] + b * Bn[i][j] + c * Cn[i][j] + d *
D[i][j];
    }
}

int main(int argc, char *argv[]) {
    if (argc < 7) {
        std::cout << "usage: linkanalysis link_file.txt a b c d
threshold" << std::endl;
        return 1;
    }
    // try opening file
    FILE *file = fopen(argv[2], "r");
    // check if opening file failed
    if (file == NULL) {
        std::cout << "error: cannot open file " << argv[2] <<
std::endl;
        return 1;
    }
    // start reading file
    while (!feof(file)) {
        char ch = ' ';
        // skip white space
        while ((ch == ' ' || ch == '\t' || ch == '\r' || ch == '\n') &&
!feof(file))
            ch = fgetc(file);
        // read number
        int number = 0;
        while ((ch >= '0' && ch <= '9') && !feof(file)) {
            number = number * 10 + ch - '0';
            ch = fgetc(file);
        }
        filter.insert(number);
    }
    // close the file
    fclose(file);
    // try opening file
    file = fopen(argv[1], "r");
    // check if opening file failed
    if (file == NULL) {
        std::cout << "error: cannot open file " << argv[1] <<
std::endl;
        return 1;
    }
}

```

```

// data structures for keys and links
// start reading file
while (!feof(file)) {
    char ch = ' ';
    // skip white space
    while ((ch == ' ' || ch == '\t' || ch == '\r' || ch == '\n') &&
!feof(file))
        ch = fgetc(file);
    // read first number
    int number1 = 0;
    while ((ch >= '0' && ch <= '9') && !feof(file)) {
        number1 = number1 * 10 + ch - '0';
        ch = fgetc(file);
    }
    // skip white space
    while ((ch == ' ' || ch == '\t' || ch == '\r' || ch == '\n') &&
!feof(file))
        ch = fgetc(file);
    // read second number
    int number2 = 0;
    while ((ch >= '0' && ch <= '9') && !feof(file)) {
        number2 = number2 * 10 + ch - '0';
        ch = fgetc(file);
    }
    // read until the end of the line
    while ((ch != '\r' && ch != '\n') && !feof(file))
        ch = fgetc(file);
    if (number1 != 0 && number2 != 0 && (filter.count(number1) ||
filter.count(number2))) {
        // print numbers
        // printf("%d %d\n", number1, number2);
        // insert keys into the key set
        keys.insert(number1);
        keys.insert(number2);
        // insert links to the link map
        links.insert(std::pair<int, int>(number1, number2));
    }
}
// close the file
fclose(file);
// print some info
std::cout << "number of nodes: " << keys.size() << std::endl;
std::cout << "number of edges: " << links.size() << std::endl;
// convert keys to ids
int N = keys.size();
int id = 0;
for (std::set<int>::iterator key = keys.begin(); key !=
keys.end(); ++key) {
    ids[*key] = id++;
    vkeys.push_back(*key);
}
// print info
std::cout << "populating the neighbourhood matrix" << std::endl;
// create L matrix
int **L = newMatrix<int>(N, 0);
// fill in the L matrix
for (std::map<int, int>::iterator key = ids.begin(); key !=
ids.end(); ++key) {

```

```

        // std::cout << key->second << ": ";
        for (std::multimap<int, int>::iterator link =
links.equal_range(key->first).first; link != links.equal_range(key-
>first).second; ++link) {
            // std::cout << ids[link->second] << " ";
            L[ids[link->first]][ids[link->second]] = 1;
        }
        // std::cout << std::endl;
    }
    // calculate A
    std::cout << "calculating A... " << std::endl;
    float **A = newMatrix<float>(N, 0);
    #pragma omp parallel for
    for (int i = 0; i < N; ++i) {
        for (int j = 0; j < N; ++j) {
            if (i == j)
                A[i][j] = 1;
            else
                A[i][j] = L[i][j] || L[j][i];
        }
    }
    printMatrix("A.txt", A, N);
    // calculate B
    std::cout << "calculating B... " << std::endl;
    float **B = newMatrix<float>(N, 0);
    #pragma omp parallel for
    for (int i = 0; i < N; ++i) {
        for (int j = 0; j < N; ++j) {
            B[i][j] = _B(L, N, i, j);
        }
    }
    printMatrix("B.txt", B, N);
    // normalize matrix
    std::cout << "normalizing B..." << std::endl;
    float **Bn = newMatrix<float>(N, 0);
    normalizeMatrix(B, Bn, N);
    printMatrix("Bn.txt", Bn, N);
    // calculate C
    std::cout << "calculating C... " << std::endl;
    float **C = newMatrix<float>(N, 0);
    #pragma omp parallel for
    for (int i = 0; i < N; ++i) {
        for (int j = 0; j < N; ++j) {
            C[i][j] = _C(L, N, i, j);
        }
    }
    printMatrix("C.txt", C, N);
    // normalize matrix
    std::cout << "normalizing C..." << std::endl;
    float **Cn = newMatrix<float>(N, 0);
    normalizeMatrix(C, Cn, N);
    printMatrix("Cn.txt", Cn, N);
    // calculate D
    std::cout << "calculating D... " << std::endl;
    float **D = newMatrix<float>(N, 0);
    #pragma omp parallel for
    for (int i = 0; i < N; ++i) {
        for (int j = 0; j < N; ++j) {

```

```

        if (i == j)
            D[i][j] = 1;
        else if (B[i][i] == 0 || B[j][j] == 0)
            D[i][j] = 0;
        else
            D[i][j] = 1 - ((B[i][i] + B[j][j] - 2 * B[i][j]) / (B[i][i]
+ B[j][j] - B[i][j]));
    }
}
printMatrix("D.txt", D, N);
// create result matrix
float **F = newMatrix<float>(N, 0);
// calculate matrices
_F(F, A, Bn, Cn, D, 0.0f, 0.0f, 1.0f, 0.0f, N);
printMatrix("F_0.0_0.0_1.0_0_0.0.csv", F, N);
_F(F, A, Bn, Cn, D, 0.0f, 1.0f, 0.0f, 0.0f, N);
printMatrix("F_0.0_1.0_0.0_0_0.0.csv", F, N);
_F(F, A, Bn, Cn, D, 0.0f, 0.5f, 0.5f, 0.0f, N);
printMatrix("F_0.0_0.5_0.5_0_0.0.csv", F, N);
_F(F, A, Bn, Cn, D, 0.2f, 0.3f, 0.3f, 0.2f, N);
printMatrix("F_0.2_0.3_0.0_3_0.2.csv", F, N);
// print similarities
printSimilarities("similarities.txt", F, N, atof(argv[7]));
// clean up
deleteMatrix(F, N);
deleteMatrix(D, N);
deleteMatrix(Cn, N);
deleteMatrix(C, N);
deleteMatrix(Bn, N);
deleteMatrix(B, N);
deleteMatrix(A, N);
// return
return 0;
}

```

ÖZGEÇMİŞ

1983 yılında Bursa’da doğdu. Işıklar Askeri Lisesinden 2002 yılında, Hava Harp Okulu Bilgisayar Mühendisliği Bölümünden 2006 yılında Teğmen rütbesi ile mezun oldu. 2007-2008 yıllarında arasında Hava Teknik Okullar Komutanlığı Gaziemir/İzmir’de Uçak Bakım Subayı Temel Eğitimi tamamladıktan sonra 2008 yılında 7’nci Ana Jet Üs Komutanlığı Erhaç/Malatya’ya Uçak Bakım Subayı olarak atandı. 2008-2012 yılları arasında 7’nci Ana Jet Üs Uçak Bakım Komutanlığında çeşitli görevlerde bulundu ve halen 7’nci Ana Jet Üs 173’üncü Filo Hat Bakım Komutanı olarak görevine devam etmektedir. Evli ve bir kız çocuk babası olup iyi derecede İngilizce bilmektedir.