



**ÇEVİRİMİÇİ FORUM SİTELERİNDEKİ HASTALIK
BELİRTİLERİNE GÖRE HASTALIKLAR VE BUNA UYGUN
İLAÇLAR ARASINDAKİ İLİŞKİLERİN BULUNMASI**

Esra GÜNDOĞAN

**Yüksek Lisans Tezi
Bilgisayar Mühendisliği Anabilim Dalı
Danışman: Prof. Dr. Mehmet KAYA**

AĞUSTOS-2017

T.C
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

ÇEVİRİMİÇİ FORUM SİTELERİNDEKİ HASTALIK BELİRTİLERİNE GÖRE
HASTALIKLAR VE BUNA UYGUN İLAÇLAR ARASINDAKİ İLİŞKİLERİN BULUNMASI

YÜKSEK LİSANS TEZİ

Esra GÜNDOĞAN

(151129111)

Tezin Enstitüye Verildiği Tarih : 17/07/2017

Tezin Savunulduğu Tarih : 07/08/2017

Tez Danışmanı : Prof. Dr. Mehmet KAYA (F.Ü)



Diğer Jüri Üyeleri : Doç. Dr. Taner TUNCER (F.Ü)



Doç. Dr. Muhammed Fatih TALU (İ.Ü)



AĞUSTOS-2017

ÖNSÖZ

Bu çalışmada sosyal ağlarda bağlantı tahmininin sağlık alanında bir uygulaması gerçekleştirilmiştir. Günümüzde internetin sağladığı ücretsiz, zahmetsiz ve kolay yolun insanların çevrimiçi forum sitelerine sorunları hakkında bilgi almak için yönelimini artırması bu çalışma için bir çıkış noktası olmuştur. Çevrimiçi forum sitelerinden elde edilen hastalık belirtilerine göre hastalıklar ve bu hastalıklara uygun ilaçlar arasındaki ilişkilerin bulunması için kurulan iki parçalı ağlarda bağlantı tahmini işlemi gerçekleştirilmiştir. İnsanların forum sitelerini bu kadar yaygın kullanması, bu sitelerden alınan bilgilerin güvenilirliği konusunu bir araştırma konusu haline getirmiş ve bu konuda yapılan çalışma ile bu sitelerden elde edilen bilgilerin doğruluğu test edilmiştir. Hazırladığım bu tez çalışmasının da sosyal ağlarda bağlantı tahminine sağlık alanında sunduğu yaklaşım ile faydalı bir kaynak olmasını umuyorum.

Bu çalışmada değerli vaktini bana ayırarak çalışmamın bitirilmesinde her türlü desteğini esirgemeyen sayın danışman hocam Prof. Dr. Mehmet KAYA' ya sonsuz teşekkürlerimi sunuyorum.

Ayrıca hayatımın her anında yanımda olan ve her türlü desteğini esirgemeyen aileme çok teşekkür ediyorum.

TEŞEKKÜR

Bu tez çalışması MF.17.04 nolu FÜBAP kapsamında yürütülen Yüksek Lisans Tezi Araştırma Projesi ile desteklenmiştir. Maddi desteklerinden dolayı Fırat Üniversitesi Bilimsel Araştırma Projeleri Birimine teşekkür ederim.

Esra GÜNDOĞAN
ELAZIĞ - 2017

İÇİNDEKİLER

Sayfa No

ÖNSÖZ	II
İÇİNDEKİLER	III
ÖZET	V
SUMMARY	VI
ŞEKİLLER LİSTESİ	VII
TABLolar LİSTESİ	VIII
SEMBOLLER LİSTESİ.....	IX
KISALTMALAR LİSTESİ	X
1. GİRİŞ	1
1.1. Literatür Özeti	4
1.2. Tezin Amacı ve Kapsamı	8
1.3. Tezin Yapısı	8
2. İKİ PARÇALI AĞLARDA BAĞLANTI TAHMİNİ	10
2.1. İki Parçalı Ağlar	10
2.2. Bağlantı Tahmini	11
2.2.1. Matematiksel Tanım	12
2.2.2. Benzerlik Tabanlı Yaklaşımlar	14
2.2.2.1. Komşuluk Tabanlı Ölçütler	14
2.2.2.2. Yol Tabanlı Ölçütler	21
2.2.3. Öğrenme Tabanlı Yaklaşımlar	23
2.3. İki Parçalı Ağlarda Bağlantı Tahmini Yöntemleri	23
2.3.1. Yansıma Tabanlı Bağlantı Tahmini	23
2.3.2. Ağırlıklı Yansıma ve İç Bağlantılar Kullanarak Bağlantı Tahmini	26
2.3.3. Yarı-İki Parçalı Ağ ile Bağlantı Tahmini	28

2.3.4.	Denetimli Makine Öğrenmesi Yaklaşımı ile Bağlantı Tahmini	30
2.3.5.	Yapısal Delikler Yardımıyla Bağlantı Tahmini	32
3.	ÇEVİRİMİÇİ FORUM SİTELERİNDE HASTALIK-BELİRTİ BAĞLANTILARININ TAHMİNİ	36
3.1.	Hastalık ve Belirti Veri Setlerinin Oluşturulması.....	37
3.2.	Çevrimiçi Forum Sitelerinin Analizi	40
3.3.	Hastalık Belirtilerinin Elde Edilmesi	41
3.4.	Soru-Hastalık Ağının Kurulması	43
3.5.	Önerilen Yöntem	43
3.6.	Sonuçlar	45
4.	HASTALIK-İLAÇ İLİŞKİLERİNİN BULUNMASI	48
4.1.	Hastalık-İlaç Verilerinin Elde Edilmesi	48
4.2.	Hastalık-İlaç Ağının Kurulumu	49
4.3.	Hastalık-İlaç Ağında Bağlantı Tahmini	49
4.4.	Sonuçlar	51
5.	ÇEVİRİMİÇİ FORUM SİTELERİNİN GÜVENİLİRLİK ANALİZİ.....	52
5.1.	Hastalık-Belirti Ağının Kurulması.....	52
5.2.	Çevrimiçi Forum Sitelerinde Soru ve Cevapların Analizi	53
5.3.	Çevrimiçi Forum Sitelerinin Güvenilirliğinin Analizi	54
6.	SONUÇLAR	56
	KAYNAKLAR	58
	ÖZGEÇMİŞ	65

ÖZET

Sosyal ağlar, çevremizdeki nesneler, kişiler, kuruluşlar ve topluluklar arasındaki ilişkileri modellemek için kullanılan yapılardan biridir. Farklı alanlarda ve farklı şekillerde karşılaştığımız birçok karmaşık sistem sosyal ağlar ile görselleştirilebilir. Sosyal ağlar karmaşık sistemlerin yapısının, gelişiminin ve ilişkilerinin anlaşılmasında yararlı bir analiz aracı olmuştur. Sosyal ağ analizi ile karmaşık sistemler hakkında önemli bilgiler elde edilir. Bu bilgiler birçok amaçla kullanılabilir. Sosyal ağ analizinin en önemli araştırma konularından biri bağlantı tahminidir. Ağlar zamanla değişir ve gelişir. Bağlantı tahmini ağların dinamikliğini araştırmak için iyi bir yöntemdir ve birçok alanda kullanıldığı gibi sağlık alanında da önemli bir analiz konusu haline gelmiştir.

Bu tez çalışmasında üç temel çalışma sunulmuştur. Birinci çalışma çevrimiçi forum sitelerindeki sorularının analizi ile elde edilen hastalık belirtilerinin hastalıklarla olan ilişkilerinin bulunmasına yöneliktir. Son yıllarda birçok alana uygulanan bağlantı tahmini, internetin hızlı gelişimi ile insanlar için sağlık konusunda önemli bir başvuru kaynağı olan çevrimiçi forum sitelerinin analizinden elde edilen verilerle kurulmuş iki parçalı ağ üzerinde gerçekleştirilmiştir. İkinci çalışma çevrimiçi forum sitelerinde kullanıcı sorularına karşılık bulunan hastalıkların ilaçlarla olan ilişkisini belirlemeye yöneliktir. Bu amaçla hastalıklar ve ilaçlardan oluşan bir iki parçalı ağ modellenmiş ve bu ağ üzerinde bağlantı tahmini işlemi ile hastalıklara uygun ilaçlar arasındaki ilişkiler bulunmuştur. Son çalışmada günümüzde insanların sıklıkla kullandıkları çevrimiçi forum sitelerinin güvenilirliği üzerine bir çalışma yapılmıştır. Bu çalışma ile insanların bu forumlardaki uzman doktorlardan aldıkları cevapların ne derece doğru olduğu test edilmiştir.

Son olarak bu tez çalışması MF.17.04 nolu FÜBAP kapsamında yürütülen Yüksek Lisans Tezi Araştırma Projesi ile desteklenmiş, yapılan çalışmalar çeşitli uluslararası yayınlar ile sonuçlandırılmıştır.

Anahtar Kelimeler: Sosyal ağ analizi, İki parçalı ağlar, Bağlantı tahmini, Çevrimiçi sağlık forumları.

SUMMARY

FINDING RELATIONSHIPS OF SYMPTOM-DISEASE AND DISEASE-DRUG IN ONLINE HEALTH FORUM SITES

Social networks are one of the structures used to model relationships between the objects, individuals, organizations and communities in our environment. Many complex systems that we face in different areas and in different ways can be visualized with social networks. Social network is a useful tool for understanding the structure, development, and relationships of complex systems. Social network analysis provides valuable information about complex systems. This information can be used for many purposes. One of the most important research topics in social network analysis is link prediction. Social networks change and evolve over time. Link prediction is an effective way to investigate dynamics of networks and has become an important analysis topic in health as well as in many areas.

In this thesis, three main studies are presented. The first study is based on finding of associations between diseases and symptoms obtained with analysis of the questions on the online forum sites. Link prediction, which has been applied in many areas over the past few years, has been carried out on a bipartite network constructed with data obtained from analysis of online forum sites, which is an important reference source in health field for people with rapid development of the Internet. The second study aims to determine relation of the diseases that correspond to user questions with the drugs in the online forum sites. For this purpose, a bipartite network consisting of diseases and drugs is modeled, and relations between diseases and drugs suitable to these diseases are found with link prediction. A recent study has focused on the reliability of online forum sites that people often use. With this study, how accurate the answers that people get from health professionals on these forums have been tested.

Finally, this thesis study was supported by Project No:MF.17.04 the Master Thesis Research Project carried out within the scope of FÜBAP, and the studies were concluded with various international publications.

Key Words: Social network analysis, Bipartite networks, Link prediction, Online health forums.

ŞEKİLLER LİSTESİ

Sayfa No

Şekil 1.1	Günümüzde sosyal ağ platformları	1
Şekil 1.2	Sosyal ağ yapısı	2
Şekil 1.3	Denetimli makine öğrenmesi yaklaşımının akış şeması [16]	6
Şekil 1.4	Graf kernel yaklaşımının akış şeması [18]	7
Şekil 2.1	Yazar-makale iki parçalı ağı	10
Şekil 2.2	Bağlantı tahmini problemi	12
Şekil 2.3	Bağlantı tahmininin genel çerçevesi [55]	13
Şekil 2.4	İki parçalı ağ örneği	24
Şekil 2.5	İki parçalı ağın yansımaları	25
Şekil 2.6	Örnek bir iki parçalı ağ ve X - yansıması G_x [15]	26
Şekil 2.7	Örnek bir okur - makale ağı	28
Şekil 2.8	Yapısal delik örneği	33
Şekil 2.9	Yapısal delik görülen örnek bir graf	33
Şekil 2.10	İki parçalı ağın düzenlenmiş hali	34
Şekil 3.1	Örnek ICD-10 hastalık kodları	39
Şekil 3.2	Örnek ICD-10 belirti kodları	39
Şekil 3.3	Elasticsearch ve veri tabanı eşleştirmesi	41
Şekil 3.4	drugs.com web sitesinin yapısı [72]	42
Şekil 3.5	Beş farklı yöntemin kesinlik değerleri karşılaştırması	46
Şekil 3.6	Beş farklı yöntemin duyarlılık değerleri karşılaştırması	46
Şekil 3.7	Beş farklı yöntemin F-ölçütü değerleri karşılaştırması	46
Şekil 4.1	drugs.com sitesinde yer alan ilaç bilgileri [72]	48
Şekil 4.2	Beş farklı yöntemin performans karşılaştırması	50
Şekil 5.1	Hastalık-belirti ilişki ağı	53

TABLÖLAR LİSTESİ

Sayfa No

Tablo 2.1 Karışıklık matrisi	18
Tablo 2.2 Benzerlik tabanlı metotların performans karşılaştırması	20
Tablo 3.1 ICD-10 bölümleri	38
Tablo 4.1 Hastalık-ilaç verilerinden bir kesit	49



SEMBOLLER LİSTESİ

G	: Sosyal ağ
x	: G ağında bir düğüm
y	: G ağında bir düğüm
z	: x ve y düğümlerinin ortak özelliği
G_x	: X -yansıması
$\Gamma(x)$	: x düğümünün komşuları
$\Gamma(y)$	: y düğümünün komşuları
$N(u)$	: u düğümünün komşularının kümesi
$N(v)$	: v düğümünün komşularının kümesi

KISALTMALAR LİSTESİ

SAA	: Sosyal Ağ Analizi
OK	: Ortak Komşular
JK	: Jaccard Katsayısı
AA	: Adamic/Adar Katsayısı
TB	: Tercihli Bağlılık
KT	: Kaynak Tahsisi
SI	: Salton İndeksi
SRI	: Sorenson İndeksi
HP	: Hub Promoted
HD	: Hub Depressed
LHN	: Leicht-Holme-Nerman
YY	: Yerel Yol
FL	: FriendLink
PR	: Köklü PageRank
HT	: Hitting Time
ICD	: International Classification of Diseases
WHO	: World Health Organization
İB	: İç Bağlantılar

1. GİRİŞ

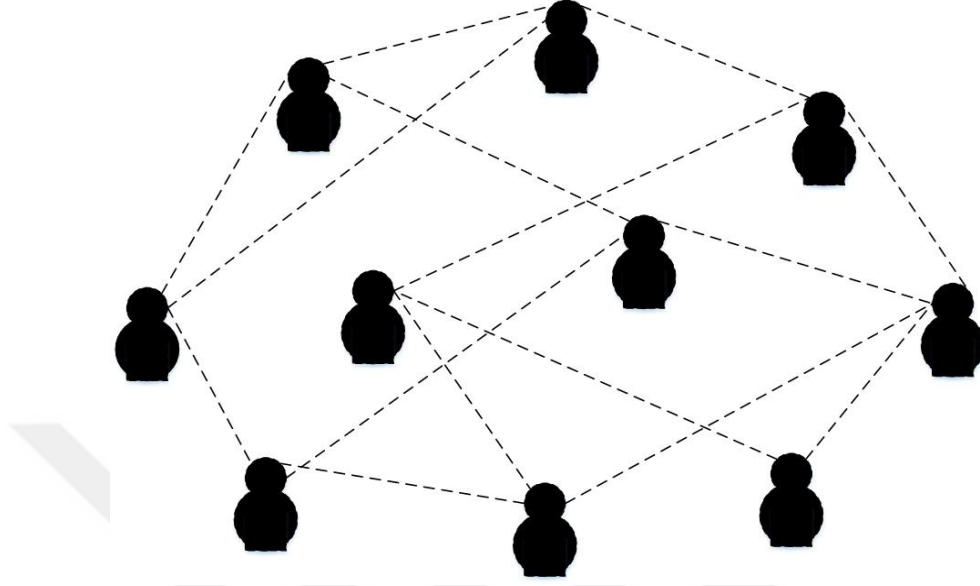
Günümüzde internetin hızlı gelişimi, teknolojiye meydana gelen değişimler insanlar arasında iletişim ve iş birliğini daha yaygın hale getirmiştir. Facebook, Twitter, Instagram gibi sosyal ağlar günlük yaşamın önemli bir parçası haline gelmiş ve insanlara bilgi aktarımı için çeşitli platformlar sunmuştur. Her geçen gün sosyal ağların kullanımı artmakta ve buna bağlı olarak bu ağlar üzerinde yer alan veriler de büyük bir veri havuzunu oluşturmaktadır. Sosyal ağlar üzerinde insanlar hakkında bu kadar çok verinin yer alması, bu ağların incelenerek insanların davranışları, eğilimleri, ilgili oldukları konular vb. özelliklerin elde edilmesiyle birçok konuda bireyler hakkında fikir sahibi olabilmek açısından sosyal ağları önemli bir analiz kaynağı haline getirmiştir. Şekil 1.1 günümüzdeki sosyal ağ platformlarından bazılarını+ göstermektedir.



Şekil 1.1. Günümüzde sosyal ağ platformları

Dünyada canlı ve cansız varlıklar arasında çeşitli şekillerde ilişkiler bulunmaktadır. Sosyal ağlar, etkileşim içinde bulunan bu varlıklar ve aralarındaki ilişkileri incelemek için kullanılan sosyal yapılardır. Bir sosyal ağ bir graf olarak görselleştirilebilir. Graf üzerinde yer alan düğümler varlıkları (organizasyonlar, bireyler, topluluklar, vs), ayrıtlar ise bu varlıklar arasındaki ilişkileri (beğeni, etkileşim, bilgi akışı, vs) gösterir. Birçok bilgi sistemleri, sosyal ve biyolojik sistemler düğümlerin bireyleri, biyolojik unsurları (protein, gen, vb), yazarları ve bağlantıların ise bu düğümler arasındaki ilişkileri temsil ettiği ağlar

olarak tanımlanabilir. Şekil 1.2 bireyler arasındaki birtakım ilişkilerden dolayı oluşmuş bir sosyal ağ yapısını göstermektedir.



Şekil 1.2. Sosyal ağ yapısı

Çevremizde birçok karmaşık sistem ile iç içe yaşamaktayız. Kompleks ağlar bu karmaşık sistemleri modellemek ve analiz etmek için yararlı bir araçtır. Yaşamımızdaki internet, sinir sistemi, protein ağları, ulaşım ağları karmaşık ağlardan bazı örneklerdir. Ağlar karmaşık sistemlerin yapısını, gelişimini, temsil ettiği varlıklar arasındaki ilişkileri analiz etmek için kolaylaştırıcı olmuştur. Karmaşık ağların incelenmesi bu yüzden birçok bilim dalının odak noktası haline gelmiştir.

Sosyal ağlar eğitim, sağlık, akademik, ticaret gibi birçok alanda farklı şekillerde karşımıza çıkmaktadır. Sosyal ağların bu kadar yaygın bir şekilde kullanılması incelenip analiz edilmesini de gerekli kılmıştır. Sosyal Ağ Analizi (SAA), bir sosyal ağda yer alan bireyler, topluluklar, organizasyonlar arasındaki ilişkilerin belirlenmesi, bu ilişkilerin sosyal yapı içindeki yerleşimlerinin ve zaman içindeki değişimlerinin incelenmesidir. SAA, 19. yy sonlarından itibaren yaygınlaşmaya başlamıştır. Teknolojide yaşanan gelişmeler, internetin insanlara farklı ortamlar sunması ve bununla beraber bireyler ve topluluklar arasındaki etkileşimin farklı boyutlara ulaşması SAA' nın daha da gelişmesini sağlamıştır. Mevcut bir sosyal ağın analizi, temsil ettiği karmaşık sistem hakkında işe yarar bilgiler elde edilmesini sağlar. Ağ yapısı incelenerek ağı oluşturan varlıklar arasındaki

ilişkiler, bu ilişkilerin devamlılığı, gelecekte hangi varlıklar arasında ilişkiler oluşabileceği, hangi varlıklar arasındaki ilişkilerin kaybolacağı gibi bilgiler elde edilebilir.

Bir sosyal ağ analiz edilip sosyal ilişkiler ve bilgi akışları görünür hale getirildikten sonra değerlendirmeler, karşılaştırmalar, ölçümler yapılabilir. SAA' nın sonuçları bireyler, organizasyonlar, toplumlar tarafından şu amaçlarla kullanılabilir:

- Ağda merkezi rol oynayan bireylerin, grupların ve birimlerin belirlenmesi,
- Bilgi akışındaki eksik noktaların, darboğazların belirlenmesi,
- Organizasyonel ve fonksiyonel birimler arasındaki bilgi akışının hızlandırılması,
- Mevcut iletişim kanallarının etkinliğinin ve verimliliğinin artırılması,
- İnovasyon ve öğrenmeyi geliştirme,
- Resmi olmayan ağların önemine olan farkındalığı artırma,
- Stratejileri düzeltme.

Sosyal ağ analizi birçok alanda farklı amaçlarla kullanılmaktadır. Sosyal ağı oluşturan varlıklar arasındaki bağlantı tahmini [1-4], tavsiye edici sistemler [5], terörist ağı analizi [6-7], tüketici davranış analizi [8], toplum keşfi [9-11] SAA' nın uygulandığı alanlardan bazılarıdır.

Sosyal ağlar üzerinde yer alan büyük miktardaki veriler bireyler, toplumlar hakkında kaliteli, işe yarar, keşfedilmesi birtakım durumların daha iyi anlaşılmasını sağlayacak, mevcut durumda iyileştirmeler yapılmasına olanak verecek özelliklere sahiptir. Bu yüzden farklı alanlardan araştırmacılar sosyal ağlara büyük ilgi göstermişlerdir. Ancak sosyal ağların analizi kolay bir iş değildir. Analiz edilirken iki zorlukla karşılaşılır. Birincisi, sosyal ağlardan sadece ağın içerdiği bilginin bir kısmı elde edilebilir. Çıkarılan bilgi tam değildir. İkincisi, sosyal ağların dinamik yapısından kaynaklanmaktadır. Zamanla yeni düğümler ve bağlantılar ağa eklenebilir ya da ağda var olan düğümler ve bağlantılar kaybolabilir. Bu yüzden ağın mevcut durumunda gözlemlenmeyen veya eksik bağlantıların tahmini ve ağa gelecekte eklenecek ya da ağdan silinecek bağlantıların tahmini, sadece sosyal ağların gelişimini anlamak için değil aynı zamanda sosyal ağların mevcut durumunu tamamlamak için oldukça önemlidir. Karşılaşılan bu problem bağlantı tahmini olarak bilinir.

Bağlantı tahmini, SAA' nın en önemli araştırma konularından birisidir. Bağlantı tahmini ile gelecek hakkında yorumlar yapmayı sağlayacak, durumda gerekli iyileştirmeler yapılmasına olanak verecek, sorunları önceden tahmin edebilme şansı tanıyacak bilgilere

ulaşılır. Bu yüzden son yıllarda sosyal ağlarda bağlantı tahmini problemini çözmek için farklı alanlardan bilim adamları tarafından birçok çalışma yapılmıştır.

Çevremizdeki birçok karmaşık sistem iki parçalı ağ yapısındadır. Karmaşık sistemlerin büyük oranda iki parçalı ağ yapısında modellenmesi, karmaşık sistemlerin analizinde büyük oranda kullanılan bağlantı tahmininin bu tür ağlara uygulanmasına olan yönelimi artırmıştır. Artık iki parçalı olarak modellenmiş ağlarda bağlantı tahmini üzerinde yoğun çalışmaların yapıldığı bir konu haline gelmiştir.

Son yıllarda yaşanan gelişmelerin sosyal ağ analizine olan ilgiyi artırması, yapılan çalışmalardan elde edilen sonuçların farklı alanlardan araştırmacıları sosyal ağlarda bağlantı tahmini konusuna yönlendirmesi, karşılaşılan birçok sistemin iki parçalı ağ yapısında olmasından dolayı bu popüler konunun bu yapıdaki ağlara uygulanmasına olan yönelim, sağladığı imkanlar ve kolaylıklar ile internetin sağlık alanında önemli bir kaynak olması nedenlerinden dolayı bu tez çalışması ile iki parçalı sosyal ağlarda bağlantı tahmini konusuna sağlık alanında bir yaklaşım sunulmuştur.

1.1. Literatür Özeti

İki parçalı ağlar çevremizde farklı alanlarda ve farklı şekillerde karşılaştığımız karmaşık ağların önemli bir türüdür. İki parçalı ağlarda bağlantı tahmini karmaşık ağların yapısını, gelişimini anlamak için önemli bir analiz şeklidir. Son yıllarda iki parçalı ağlarda bağlantı tahmini önemli bir araştırma konusu haline gelmiştir. Bu konuda yapılmış birçok çalışma bulunmaktadır.

Benchettara ve ark. [12] iki düğüm arasında bir bağlantı oluşma olasılığını yansıtabilen yeni topolojik özellikler tanımlamışlar ve iki parçalı ağlarda bağlantı tahmini için denetimli makine öğrenmesi yaklaşımında bu özellikleri kullanmışlardır. Belirli bir düğüm çifti arasında bağlantı oluşma olasılığını ölçmek için bir sınıflandırıcı oluşturulmuştur. Ağın geçmiş zamanlardan anlık görüntüleri çıkarılmış ve bu görüntülerden doğrudan ve dolaylı özellikler bulunmuştur. Bu özellikler sınıflandırıcının eğitilmesi için kullanılmıştır. Eldeki veriler eğitim ve test verisi olarak iki veri setine bölünmüştür. Eğitim verileri ile eğitilen sınıflandırıcı belirli bir zaman periyodunda test verileri üzerinde çalıştırılmıştır. Önerilen yaklaşım DBLP' den çıkarılan ortak yazarlık ağı ve online bir müzik e-ticaret sitesinin 8 yıllık tarihini içeren iki parçalı ağ üzerinde denenmiştir. Tanımlanan yeni ölçütlerin elde edilen sonuçları özellikle tahmin kesinliği açısından geliştirdiği görülmüştür.

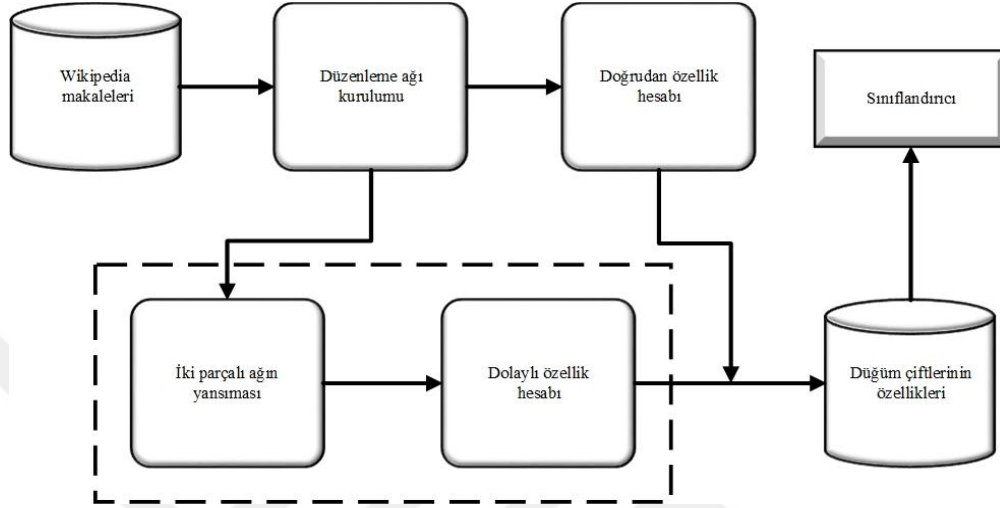
Xia ve ark. [13] bağlantı tahmininde pek dikkat çekmemiş olan yapısal deliklere dayalı bir çalışma yapmışlardır. Bu yaklaşım düğümler arasındaki zayıf bağlara dayanmaktadır. İki düğüm arasında çok fazla ortak düğüm olmaması, yani farklı gruplara ait düğümler olmaları bu düğümler arasında yapısal delik oluşacağını gösterir. Bu yapısal delik ile bu iki düğüm arasında bağlantı oluşturulması bir gruptan diğer gruba bilgi akışını sağlar. Bu çalışmada yapısal delikler için iki yeni ölçüt önerilmiştir. Ayrıca Ortak Komşular, Jaccard Katsayısı, Adamic/Adar metotları iki parçalı ağlarda bağlantı tahmini için genişletilmiştir. Bağlantı tahmini performansı IMDB’ den çıkarılan iki parçalı yönetmen-aktör ağı üzerinde değerlendirilmiştir. Yapısal deliklerin bağlantı tahmininde yararlı olacağı görülmüştür.

Nigam ve ark. [14] kişiler ve ilgili oldukları sağlık konuları arasındaki ilişkilerden oluşan bir iki parçalı ağda bağlantıları tahmin etmek için bir yarı-iki parçalı ağ modeli kullanmışlardır. Kişiler ve ilgili oldukları konulardan oluşan bir iki parçalı ağ tasarlanmıştır. Bu yöntemde bağlantı tahmini için aynı kümede bulunan düğümler arasındaki ortak özellikler kullanılmıştır. Kişilerin daha önceden okumuş oldukları makalelere bakılarak ilgi alanları modellenmiştir. Daha sonra kişilerin demografik bilgisi özellikle yaşadıkları şehir gibi özelliklerine bakılarak o kişilerle ilgili olabilecek diğer konular tahmin edilmeye çalışılmıştır. Denetimli makine öğrenmesi yaklaşımları ile kişilerin demografik özellikleri, okuma alışkanlıkları gibi ortak özelliklerinden yararlanılarak kişilere ilgi alanlarına girecek konular önerilmeye çalışılmıştır.

Allali ve ark. [15] iki parçalı ağlarda iç bağlantılar olarak adlandırılan özel bir tür bağlantılar tanımlamışlardır. Bu bağlantılara dayalı bağlantı tahmini yapılmıştır. İç bağlantılar, ağda o anda mevcut bulunmayan ve ağa eklenmesi ağın yansımını değiştirmeyecek olan bağlantılar olarak tanımlanmıştır. Ağ ilk olarak tek mod yansıma ağa dönüştürülmüştür. Daha sonra ağda o anda bağlantılı olmayan düğüm çiftleri belirlenmiştir. Bu düğüm çiftleri arasında oluşturulacak olan bir bağlantının ağın yansımını değiştirip değiştirmeyeceği kontrol edilmiştir. Değiştirmeyen bağlantılar iç bağlantılar olarak belirlenmiştir. Metot iki gerçek veri seti üzerinde test edilmiştir. Elde edilen sonuçlar tavsiye sistemlerinde işbirlikçi filtreleme yöntemi ile karşılaştırılmıştır. Önerilen yaklaşımın daha iyi bir performansa sahip olduğu ve iç bağlantıların dinamik iki parçalı ağlarda önemli bir rol oynadığı görülmüştür.

Chang ve ark. [16] Wikipedia’ dan elde edilen yazar-makale verilerinden iki parçalı bir ağ oluşturmuşlardır. Yazarlar ve düzenledikleri makaleler arasındaki ilişkilerden yola çıkarak yazarlara düzenleyebilecekleri makaleler önermek için bir bağlantı tahmini işlemi

gerçekleştirmişlerdir. Bunun için denetimli makine öğrenmesi yaklaşımını kullanmışlardır. Oluşturulan iki parçalı ağdan bazı özellikler çıkarılmış ve bu özellikler makine öğrenmesi algoritmasında kullanılmıştır.



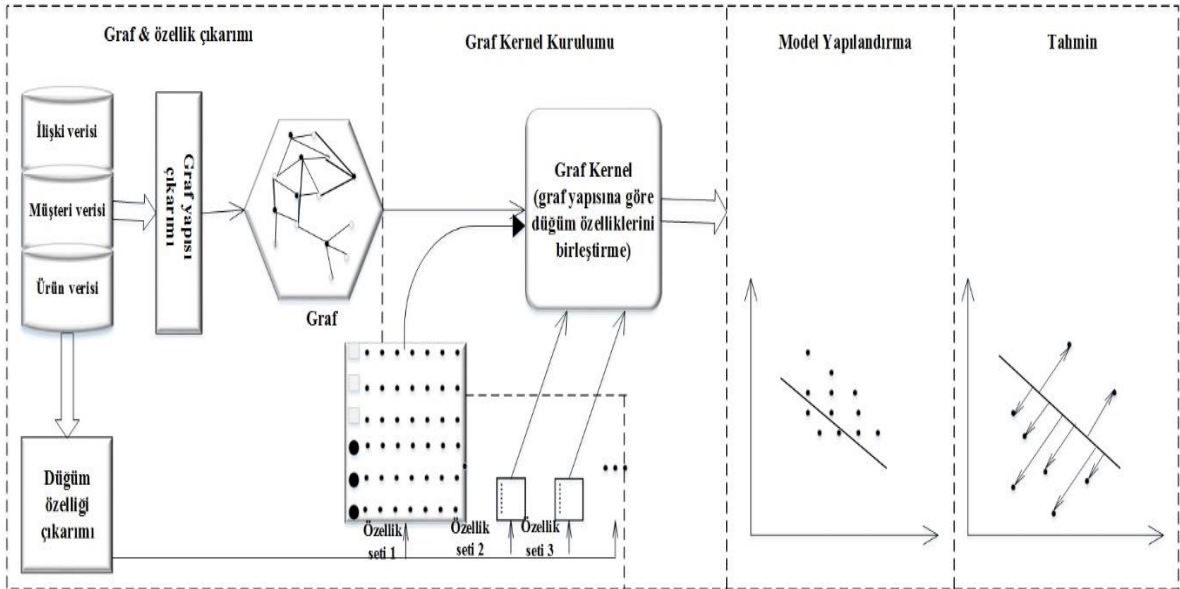
Şekil1.3. Denetimli makine öğrenmesi yaklaşımının akış şeması [16]

Gao ve ark. [17] yansıma tabanlı bir bağlantı tahmini işlemi gerçekleştirmişlerdir. Yapılan çalışmada iki parçalı ağ ilk olarak yansıma olarak adlandırılan ağırlıklı tek mod ağa dönüştürülmüştür. Yansıma ağ üzerinde aday düğüm çiftleri belirlenmiştir. Aday düğüm çifti, ağda o anda bağlı olmayan ve aralarında bir bağlantı oluşturulması ağın yansımasını değiştirmeyen düğüm çiftleridir. Aday düğüm çiftleri belirlendikten sonra bu aday düğüm çiftlerinin kapsadığı örüntüler ve ağırlıkları bulunmuştur. Bir aday düğüm çiftinin kapsadığı örüntülerin ağırlıkları toplamı ne kadar büyükse bu aday düğüm çifti arasına gelecekte bağlantı oluşma ihtimalinin o kadar yüksek olacağı kabulüne göre bağlantı tahmini yapılmıştır. Önerilen yöntem diğer metotlar ile karşılaştırıldığında daha yüksek bir hız ve daha kaliteli bağlantı tahmini sonuçlarına ulaşılmıştır. Çalışmada sunulan aday düğüm çifti kavramı, Allali ve ark. [15] tarafından tanımlanan iç bağlantılar kavramına benzerdir. Ancak bu çalışmada bazı önemli farklar bulunmaktadır. Bu farklar;

- Bir iç bağlantının oluşma olasılığı sadece kapsadığı kenar sayısı ile ilgilidir. Ancak aday düğüm çiftinin gerçek bir bağlantı olma olasılığı kapsadığı örüntülerin sayısı ve bu örüntülerin ağın topolojik bilgisinden yararlanılarak hesaplanan ağırlıklarına göre belirlenir. Daha fazla topolojik bilginin kullanılmasından dolayı bu yöntemde daha doğru tahmin sonuçları elde edilmiştir.

- Bağlantı tahmini sadece aday düğüm çiftleri içinde gerçekleştirildiğinden hesaplama zamanı ve karmaşıklığı bu yöntemde daha azdır.
- Belirli bir eşik değerine eşit veya bu eşik değerinden daha büyük ağırlığa sahip iç bağlantılar potansiyel bağlantılar olarak düşünülür. Bu yüzden daha doğru sonuçlar için doğru bir eşik değeri tanımlanmalıdır. Böyle bir eşik değeri belirlemek zordur. Bu yöntemde ise bir aday düğüm çiftinin bağlılığı ağırlığının daha fazla topolojik bilgisini içeren benzerlik ölçütü olarak belirlenmiştir. Belirlenen bu ölçüt ile daha doğru tahmin sonuçları elde edilmiştir.

Li ve ark. [18] genel bir kernel tabanlı makine öğrenmesi yaklaşımı önermişler ve bu metodu tavsiye sistemlerinde uygulamışlardır. Bir kullanıcı-ürün düğüm çifti seçilmiştir. Bu çiftin etkileşim içinde olma olasılığı yani kullanıcının o ürünü satın alma ya da kullanma olasılığını bulmak için bu düğüm çiftine yakın olan düğümler ve bağlantılar kontrol edilmiştir. Seçilen bu düğüm çiftinin bağlamındaki kullanıcı/ürün özelliklerini ve yapısını çıkaran bir graf kernel oluşturulmuş ve tahmin modelini kurmak için SWM algoritmasına verilmiştir. Graf kernel yapısının geçerlilik ve hesaplama verimliliği incelenmiştir. Bu tavsiye yaklaşımının performansı üç gerçek veri seti üzerinde test edilmiş ve başarılı sonuçlar elde edilmiştir. Şekil 1.4 modelin oluşturulma adımlarını göstermektedir.



Şekil 1.4. Graf kernel yaklaşımının akış şeması [18]

1.2. Tezin Amacı ve Kapsamı

Sosyal ağlar çevremizde farklı alanlarda ve farklı şekillerde karşılaştığımız yapılardır. Etrafımızdaki nesneler, bireyler, topluluklar vb. arasındaki ilişkiler sürekli değiştiğinden dolayı sosyal ağ yapıları da sürekli değişir. Ağların bu dinamik yapısını anlamak ve gelecekteki ilişkileri belirlemek için bağlantı tahmini önemli ve etkili bir çözümdür. Bu tez çalışmasında, bağlantı tahmini konusunun önemine değinilmiş ve şimdiye kadar önerilen tek mod ve iki parçalı ağlara yönelik bağlantı tahmini yöntemleri anlatılmıştır. İki parçalı ağlarda bağlantı tahminini sağlık alanına uygulayan bir çalışma sunulmuştur. Bunun için veri kaynağı olarak çevrimiçi forum siteleri kullanılmıştır. Bu forumların artık günümüzde sıklıkla kullanılması analiz edilerek insanlar hakkında birtakım verilere ulaşmada yararlı olacağı düşünülmüştür. Şimdiye kadar forum sitelerinin analizi ile ilgili pek fazla çalışma yapılmadığından bu çalışma ile sağlık alanında farklı bir sosyal platformun analizi gerçekleştirilmiştir. Bu forumlardan elde edilen hastalık belirtilerine göre hastalıklar ve ilaçlar arasındaki ilişkiler kurulan iki parçalı ağlarda bağlantı tahmini ile bulunmuştur. Ayrıca sıklıkla insanların sorunları hakkında bilgi almak için başvurdukları çevrimiçi forum sitelerinin ne kadar güvenilir bilgi sağladıklarını göstermek amaçlanmıştır.

1.3. Tezin Yapısı

Bu tez çalışması *giriş bölümü* ile birlikte altı bölümden oluşmaktadır. Bu bölümde tez çalışması kapsamında literatür çalışması yapılarak, iki parçalı ağlarda bağlantı tahmini konusunda yapılan çalışmalar incelenmiştir.

İkinci bölümde, iki parçalı ağların yapısı tanımlanmıştır. Sosyal ağ analizinde önemli bir araştırma konusu olan bağlantı tahmini probleminden bahsedilmiş ve bağlantı tahmininin matematiksel tanımı verilmiştir. Bağlantı tahmini konusundaki benzerlik ve öğrenme tabanlı yaklaşımlar ve bu yaklaşımların yapısını oluşturan ölçütler hakkında bilgi verilmiştir. İki parçalı ağlarda bağlantı tahmini yöntemleri açıklanmıştır. Gerçekleştirilen bağlantı tahmini yöntemlerinin performansının değerlendirilmesinde kullanılan ölçütlerden bahsedilmiştir.

Üçüncü bölümde, çevrimiçi sağlık forumlarından elde edilen verilerle kurulmuş bir iki parçalı ağda bağlantı tahmini gerçekleştirilmiştir. Günümüzde çevrimiçi sağlık forumlarına olan yönelimden bahsedilmiş ve bu sitelerde yer alan kullanıcı sorularının analiz işlemi

anlatılmıştır. Uygulamada kullanılan verilerin elde edilme aşamalarından ve bu aşamalarda kullanılan teknolojilerden bahsedilmiştir. Kurulan iki parçalı ağda bağlantı tahmini için iç bağlantılar yöntemi önerilmiştir. Bu yöntemin performansını karşılaştırmak için aynı ağda bağlantı tahmini için benzerlik tabanlı algoritmalarından dört tanesi daha kullanılmıştır. Uygulan beş farklı bağlantı tahmini algoritmasının performansı kesinlik, duyarlılık ve F-ölçütü olmak üzere üç performans değerlendirme ölçütü kullanılarak karşılaştırılmıştır. Elde edilen sonuçlar değerlendirilmiştir.

Dördüncü bölümde, her hastalık için oranlarına göre en çok kullanılan ilaçlardan oluşan bir hastalık-ilaç ağı kurulmuş ve bu ağda bağlantı tahmini işlemi gerçekleştirilmiştir. Yapılan bağlantı tahmini çalışmasının sonuçları değerlendirilmiştir.

Beşinci bölümde, son zamanlarda çevrimiçi sağlık forumlarına olan eğilim düşünülerek bu sitelerde kullanıcı sorularına verilen yanıtların doğruluğunu tespit eden bir uygulama gerçekleştirilmiştir. Bu sitelerde sorulara verilen yanıtlar ile gerçek hastalık ve belirti verileri kullanılarak elde edilen cevaplar karşılaştırılarak bu forum sitelerinde yer alan bilgilerin güvenilirliği ölçülmüştür.

Altıncı bölümde ise tezde yapılan tüm çalışmalar elde edilen sonuçlar doğrultusunda değerlendirilmiştir.

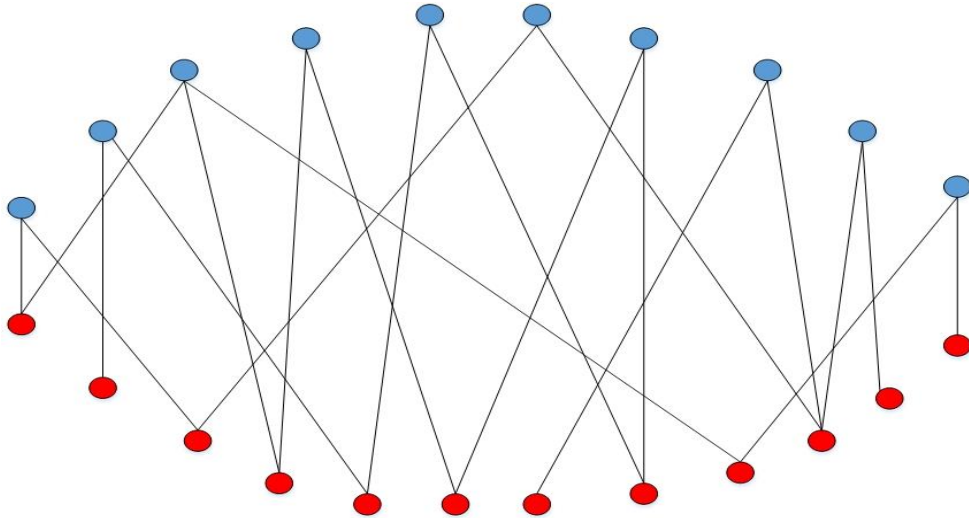
2. İKİ PARÇALI AĞLARDA BAĞLANTI TAHMİNİ

İki parçalı ağlar karmaşık ağların önemli bir türüdür. Çevremizdeki birçok sosyal ağ iki parçalı ağ yapısındadır. Yazar-makale işbirliği ağı [19,20], RNA-protein etkileşim ağı [21], hastalık-gen ağı [22-25], kulüp üyeleri-aktivite ağı [26] iki parçalı ağ yapılarına bazı örneklerdir. Birçok karmaşık sistemin iki parçalı ağ şeklinde modellenmesi, karmaşık sistemlerin analizinde büyük oranda kullanılan bağlantı tahmini konusunun iki parçalı ağlara uygulanması konusundaki çalışmaları artırmıştır. Son yıllarda iki parçalı ağlarda bağlantı tahmini tavsiye [14, 16, 27-29], sosyal ağ analizi [12], topluluk keşfi [30] ve ilaç-yan etki tahmini [31] gibi birçok alanda uygulanmaktadır.

2.1. İki Parçalı Ağlar

İki parçalı ağlar, sosyal ağ yapılarının özel bir türüdür. İki parçalı ağlarda düğümler iki farklı kümede yer alır. Bağlantılar sadece farklı kümelerde bulunan düğümler arasındadır. Aynı kümede yer alan düğümler arasında bağlantı bulunmaz.

İki parçalı ağ $G = (X, Y, E)$ şeklinde gösterilir. Bu gösterimde X ve Y düğüm kümelerini, E ise bu kümelerdeki düğümler arasında yer alan bağlantıları gösterir. Örnek bir bağlantı $x \in X$ ve $y \in Y$ şartını sağlayarak $(x, y) \in E$ şeklinde tanımlanır.



Şekil 2.1. Yazar-makale iki parçalı ağı

Şekil 2.1 örnek bir yazar-makale iki parçalı ağını göstermektedir. Ağda kırmızı ve mavi ile renklendirilen iki farklı düğüm kümesi yer almaktadır. Mavi düğümler yazarları, kırmızı düğümler ise makaleleri göstermektedir. Görüldüğü gibi yazarlara ve makalelere ait düğümlerin kendi içlerinde bağlantıları yoktur. Bağlantılar sadece yazar-makale düğümleri arasında kurulmuştur. Yazar ve makale düğümleri arasındaki ilişkiler yazarların düzenledikleri veya yazdıkları makalelere göre oluşturulabilir.

2.2. Bağlantı Tahmini

Ağın gelecekteki durumunun analizi ile ilgili araştırma konularından olan bağlantı tahmini iki düğüm arasında bir bağlantı oluşumunun ağdaki düğümlerin özelliklerine ve diğer gözlemlenen bağlantılara göre tahmin edilmesi problemidir [32]. Bir sosyal ağın gelecekteki durumunun analiz edilmesi ve ağda meydana gelebilecek değişimlerin belirlenmesi ağın geçmiş durumlarından yararlanılarak yapılır.

Zaman içerisinde bireyler, topluluklar, organizasyonlar arasındaki ilişkiler sürekli değişir. Gerçek dünyadaki bu ilişkilerin modellendiği sosyal ağlarda bu değişimlerden dolayı sürekli değişir ve gelişir. Sosyal ağların dinamik yapısında yeni bağlantılar, düğümler ağa eklenebilir ya da var olan bağlantılar, düğümler ağdan kaybolabilir. Bu yüzden bağlantı tahmini bir sosyal ağdaki varlıklar arasında olan dinamik ilişkileri yakalayabilme yeteneğine sahip olmalıdır.

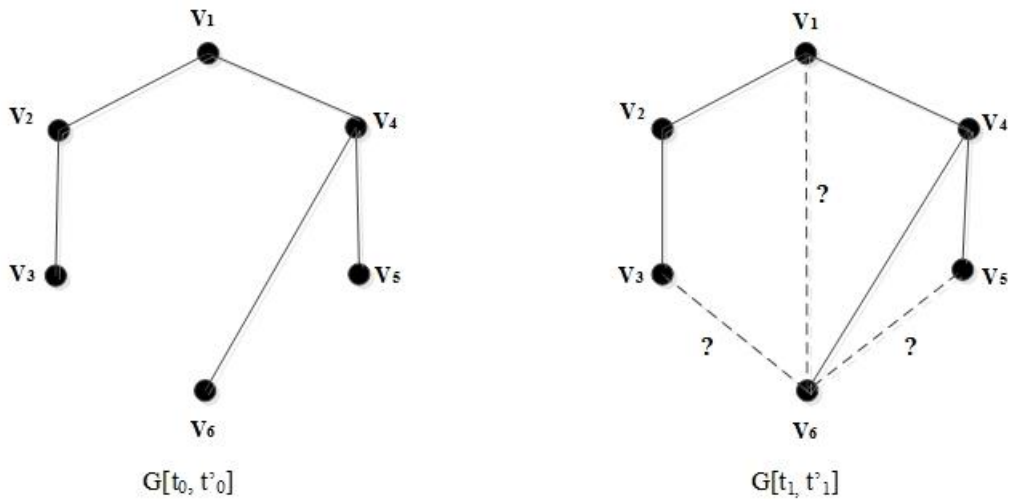
Bağlantı tahmini birçok alanda önemli uygulamalara sahiptir. Bilgi çıkarımı ve e-ticarette tavsiye sistemlerine uygulanabilir. İnsanların yeni arkadaşlar [33] ve potansiyel ortaklar [34,35] bulmalarına yardım edilebilir, online alışverişte ürün tavsiyesinde bulunulabilir [36], kurumsal sosyal ağlarda patent ortakları [35] ve alanlar arası ortaklar [37] önerilebilir, akademik sosyal ağlarda uzmanlar veya ortak yazarlar bulunabilir [38,39], büyük ölçekli iletişim ağlarında cep telefonu bağlantılarının tahmini [40] yapılabilir. Ayrıca kısmen gözlemlenen ağlara dayalı tam ağları çıkarmak [41,42], ağların gelişimini daha iyi anlamak [43-46] ve heterojen sosyal ağlarda köprü görevi gören bağlantıları tahmin etmek [47] için kullanılabilir. Bağlantı tahmini teknikleri biyoinformatik ve biyolojiye uygulanabilir. Sağlık hizmetlerinde [48], gelecekte referans alması daha muhtemel olan uzmanları tahmin etmede ve protein-protein etkileşimlerini bulmada kullanılabilir. Güvenlikle alakalı alanlarda, anormal iletişimlerini tanımlamak [49], terörist gruplarını ve suçluları keşfetmek [50] için uygulanabilir. Sensor ağlarında dinamik

zamansal özellikleri keşfetmek [51], bilgi aktarımı gizliliğini sağlamak [52] ve en uygun yönlendirmeyi gerçekleştirmek [53] amacıyla kullanılabilir. İnsan davranışlarına dayalı oluşturulan ağlarda insanların hareketlerini ve aktivitelerini tanımlamak ve kategorize etmek [54] için kullanılabilir. Ayrıca bağlantı tahmini iletişim ağları ve e-posta ağları gibi sosyal ilişkileri yansıtan ağlarda birçok uygulamalara sahiptir.

2.2.1. Matematiksel Tanım

Bağlantı tahmini Liben-Nowell ve Kleinberg [32] tarafından aşağıdaki gibi ifade edilmiştir:

Bir sosyal ağın $G = \langle V, E \rangle$ şeklinde tanımlandığını düşünelim. Bu sosyal ağda u ve v düğümleri arasında yer alan bir bağlantı belirli bir $t(e)$ zamanında $e = \langle u, v \rangle \in E$ olarak ifade edilir. Ağın düğümleri arasındaki bağlantılar geçmiş zamanlardan itibaren bilinmektedir. Bir $t \leq t'$ zaman aralığı için $G[t, t']$, G ağının t ile t' zaman aralığındaki bağlantıların tanımlı olduğu bir alt ağını gösterir. Bağlantı tahmini için $t_0 < t'_0 < t_1 < t'_1$ olacak şekilde bir zaman periyodu seçilir. Bu zaman periyodu $[t_0, t'_0]$ eğitim aralığı, $[t_1, t'_1]$ test aralığı olacak şekilde bölünür. Bağlantı tahmini işlemi $G[t_0, t'_0]$ ağında olmayan ancak $G[t_1, t'_1]$ ağında oluşacak olan bağlantıların tahmin edilmesidir.

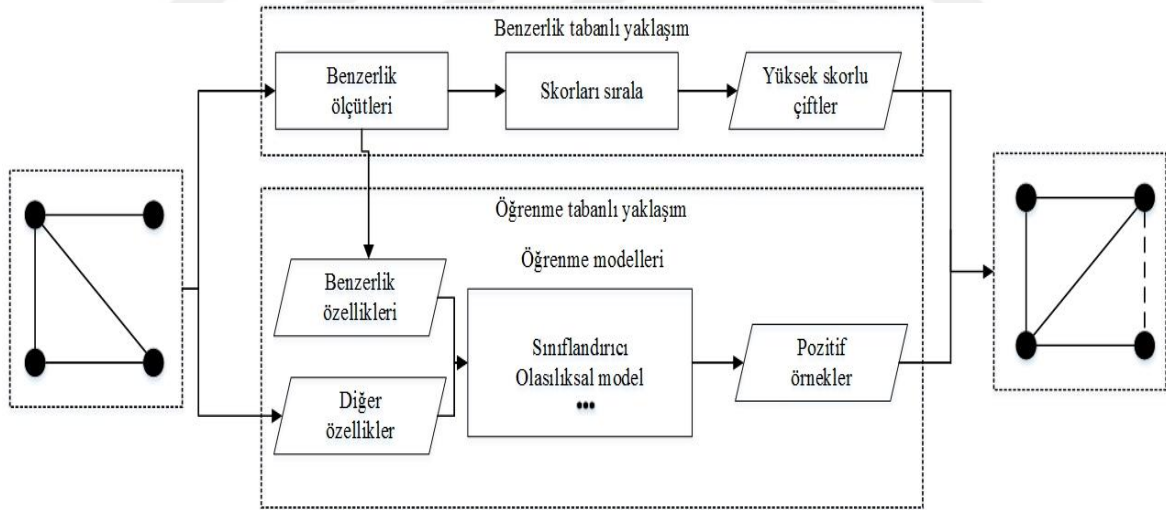


Şekil 2.2. Bağlantı tahmini problemi

Şekil 2.2’ de G ağında gerçekleştirilen bağlantı tahmini gösterilmiştir. Bağlantı tahmini yöntemi $G[t_0, t'_0]$ ağında olmayan ancak $G[t_1, t'_1]$ ağında (V_3, V_6) , (V_1, V_6) ve (V_5, V_6) düğümleri arasında oluşması muhtemel bağlantıları tahmin etmeye çalışır. Tahmin edilen bağlantılar ağda gerçekten oluşabilir ya da oluşmayabilir.

Bağlantı tahmini çevremizde birçok uygulamada gerçekleştirilebilir. Örneğin; günümüzde yaygın olarak kullanılan bir sosyal medya aracı Facebook’ u düşünelim. Facebook’ ta insanlar tanıdıklarıyla arkadaş olurlar. Bu ortak arkadaşlıklara dayalı olarak insanlara bazı kişileri arkadaş olarak eklemeleri önerilir. Arkadaş olarak birbirini eklememiş iki kişinin ne kadar çok ortak arkadaşı varsa bu kişiler arasında bir bağlantı oluşması yani arkadaş olma olasılıkları o kadar yüksektir. Bu yaklaşımla kişiler arasındaki ortak arkadaşlara dayalı bir bağlantı tahmini yapıp kişiler birbirlerine arkadaş olarak eklemeleri için önerilebilir.

Bağlantı tahmini problemini çözmek için tüm düğümler arasındaki bağlantıların oluşma veya kaybolma olasılıklarının belirlenmesi gerekir. Genellikle bu olasılıklar düğüm çiftleri arasındaki benzerlikler veya yakın özellikler ile hesaplanabilir [55].



Şekil 2.3. Bağlantı tahmininin genel çerçevesi [55]

Bağlantı tahmini çözümünün genel bir çerçevesi Şekil 2.3’ te gösterilmiştir. Bir sosyal ağın gelişimini tahmin etmek için benzerlik tabanlı ve öğrenme tabanlı olmak üzere iki yaklaşım vardır. Benzerlik tabanlı yaklaşımlar, bir sosyal ağda bağlı olmayan düğüm çiftlerinin benzerliğini düğümlerin yakınlık ölçütlerine göre ölçer. Aralarında bağlantı

olması muhtemel her düğüm çifti için bir skor hesaplanır. Bir düğüm çifti arasında hesaplanan skor ne kadar yüksek olursa bu düğüm çifti arasında gelecekte bir bağlantı oluşması o kadar yüksek olur. Düğüm çiftleri skorlarına göre büyükten küçüğe doğru sıralanır. En üstte yer alan çiftler arasındaki bağlantılar gelecekte oluşması en muhtemel bağlantılardır. Öğrenme tabanlı yaklaşımlar [56] ise bağlantı tahminini bir ikili sınıflandırma gibi gerçekleştirir. Bu yüzden bu yaklaşımlarda bağlantı tahmini problemini çözmek için sınıflandırıcı gibi bazı makine öğrenmesi modelleri veya olasılıksal model kullanılır. Her bağlı olmayan düğüm çifti, düğümleri tanımlayan özellikler ve sınıf etiketine sahip bir örneğe karşılık gelir. Eğer bir düğüm çiftini bağlayan olası bir bağlantı varsa, bu düğüm çifti pozitif yoksa negatif olarak etiketlenir. Öğrenme tabanlı yaklaşımlar için özellikler iki kısımdan oluşur. Biri benzerlik tabanlı yaklaşımlardan elde edilen özellikler, diğeri ise sosyal ağdan çıkarılan özelliklerdir.

2.2.2. Benzerlik Tabanlı Yaklaşımlar

Bir düğüm çifti arasındaki benzerliği hesaplama bağlantı tahmini için basit bir çözüm oluşturur. Bu yaklaşımlarda temel fikir, bir düğüm çifti ne kadar benzer ise bu düğüm çifti arasında bağlantı oluşması o kadar yüksek olasılıktadır. Bu, kişilerin kendilerine eğitim, bölge, ilgi alanları ve konum gibi benzer olduğu özelliklere sahip insanlar ile ilişkiler kurma eğiliminde olma gerçeği ile tutarlıdır.

Bir ağda düğüm ve bağlantı özellikleri olmasa bile iki düğümün benzerliğini hesaplamak için birçok ölçüt bulunmaktadır. Bu ölçütler komşuluk tabanlı ve yol tabanlı ölçütler olarak sınıflandırılmıştır.

Bir $G(V, E)$ ağında $x, y \in V$ ve $\langle x, y \rangle \in E$ olacak şekilde seçilen bir (x, y) düğüm çifti için $\Gamma(x)$, x düğümünün komşularını, $|\Gamma(x)|$ ise x düğümünün komşularının sayısını göstermektedir.

2.2.2.1. Komşuluk Tabanlı Ölçütler

Bir sosyal ağda insanlar kendilerine yakın olan kişiler ile yeni ilişkiler kurma eğilimindedirler. İnsanlara en yakın olan kişiler komşularıdır. Bu yüzden bağlantı tahmini

için çok sayıda komşuluk tabanlı ölçüt önerilmiştir. Bu ölçütlerden bazıları bu bölümde açıklanmıştır.

- **Ortak Komşular (OK):**

Bir sosyal ağda birbirine bağlı olan birçok insan bulunmaktadır. Ancak her insanın ağdaki tüm insanlara bağlı olması söz konusu değildir. Gelecekte iki kişi birbirine bir ortak komşu vasıtasıyla bağlanabilirler. Ortak komşular ölçütü de iki düğümün gelecekte bağlı olma durumunun ortak komşularının sayısı ile orantılı olduğu varsayımına dayalı olarak hesaplanan bir ölçüttür [57]. İki düğümün ne kadar çok ortak komşusu varsa aralarında bir bağlantı oluşma ihtimalinin o kadar yüksek olacağı kabul edilir. OK ölçütü basitliğinden dolayı bağlantı tahmini probleminde en yaygın olarak kullanılan ölçütlerden birisidir.

$$OK(x, y) = |\Gamma(x) \cap \Gamma(y)| \quad (2.1)$$

- **Jaccard Katsayısı (JK):**

Jaccard katsayısı ortak komşu ölçütünün normalize edilmiş halidir [58]. x ve y düğüm çiftinin komşularının birleşim kümesinden rastgele yapılan bir seçimin x ve y düğüm çiftinin ortak komşularından biri olma olasılığını gösterir. Ortak komşu sayısı ne kadar çok olursa JK ölçütü o kadar büyük olacaktır.

$$JK(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \quad (2.2)$$

- **Adamic/Adar Katsayısı (AA):**

Bu ölçüt Adamic ve Adar tarafından başlangıçta iki web sayfası arasındaki benzerliği hesaplamak için önerildi [59]. z, x ve y' nin ortak özelliklerinin kümesini gösterirken bu ölçüt (2.3)' teki gibi tanımlanmıştır.

$$AA(x, y) = \sum_z \frac{1}{\log(\text{frekans}(z))} \quad (2.3)$$

Bağlantı tahmini için bu formül özellikler ortak komşular olacak şekilde aşağıdaki gibi özelleştirilmiştir [60]. Yeni ölçüt ortak komşuların ağırlığına göre hesaplanır. Ortak özelliklerden nadir olanların ağırlık oranları arttırılmıştır.

$$AA(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log|\Gamma(z)|} \quad (2.4)$$

- **Tercihli Bağlılık (TB):**

Bir x düğümünün bağlantı oluşturma ihtimali komşularının sayısı ile orantılıdır. Yani komşusu çok olan bir düğümün bağlantı oluşturma ihtimali o kadar yüksektir [61]. Bu ölçüt yeni oluşacak bağlantıların, yüksek dereceli düğümlere bağlanmasının düşük dereceli olanlara bağlanmasından daha olası olduğunu gösterir.

$$TB(x, y) = |\Gamma(x)| \cdot |\Gamma(y)| \quad (2.5)$$

- **Kaynak Tahsisi (KT):**

Bu ölçüt Zhou [62] tarafından önerilmiştir. Birbirine doğrudan bağlı olmayan düğüm çiftleri arasında bir bağlantı oluşma ihtimalini ölçer. Her düğüm bir kaynak ünitesine sahiptir ve kaynağını eşit olarak tüm komşularına dağıtır. Aralarından bağlantı olmayan düğümler arasındaki kaynak aktarımı sahip oldukları ortak komşuları aracılığıyla gerçekleşir. Düğümler arasındaki benzerlik aldıkları kaynağa göre belirlenir.

$$KT(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{|\Gamma(z)|} \quad (2.6)$$

- **Salton İndeksi (SI):**

x ve y düğümleri arasındaki benzerliği ölçmek için kullanılan bir kosinüs ölçütüdür. Salton kosinüs benzerliği olarak da bilinir [63].

$$SI(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{\sqrt{|\Gamma(x)| \cdot |\Gamma(y)|}} \quad (2.7)$$

- **Sorenson İndeksi (SRI):**

Jaccard katsayısına [58] benzer bir ölçüt olan Sorenson indeksi [64] türler arasındaki benzerliği ölçmek için önerilmiştir. Ortak komşuların sayısını hesaba katmanın yanı sıra düşük dereceli düğümler arasında bağlantı oluşma ihtimalinin daha yüksek olduğuna dikkat çeker.

$$SRI(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x)| + |\Gamma(y)|} \quad (2.8)$$

- **Hub Promoted (HP):**

Bu ölçüt x ve y düğümlerinin topolojik kesişimini tanımlar. HP değeri düşük dereceli düğümler tarafından belirlenir [65].

$$HP(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{\min(|\Gamma(x)|, |\Gamma(y)|)} \quad (2.9)$$

- **Hub Depressed (HD):**

Zhou tarafından HP ölçütüne benzer bir ölçüt olarak önerilmiştir. HP ölçütünden farklı olarak HD değeri, yüksek dereceli düğümler tarafından belirlenir [62].

$$HD(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{\max(|\Gamma(x)|, |\Gamma(y)|)} \quad (2.10)$$

- **Leicht-Holme-Nerman (LHN):**

Bu ölçüt birçok ortak komşuya sahip olan düğüm çiftine yüksek benzerlik atar [66].

$$LHN(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x)| \cdot |\Gamma(y)|} \quad (2.11)$$

Bağlantı tahmini algoritmaları bir ağa uygulandıktan sonra birçok bağlantı tahmin edilir. Bu bağlantıların bazıları gelecekte ağ üzerinde oluşurken bazıları ise oluşmaz. Tahmin edilen bu bağlantıların gelecekte gerçekten oluşup oluşmayacakları bağlantı tahmini algoritmasının performansını gösterir. Bir bağlantı tahmini algoritmasının performansının daha kolay değerlendirilmesi için ilk olarak Tablo 2.1’de gösterildiği gibi bir karışıklık matrisi oluşturulur. Bu matris üzerinde gelecekte ağda var olacak bağlantıların sınıfı 1, ağda var olmayacak bağlantıların sınıfı ise -1 olarak tanımlanmıştır.

Tablo 2.1. Karışıklık Matrisi

DOĞRU SINIF	TAHMİN EDİLEN SINIF		
		Sınıf= 1	Sınıf= -1
	Sınıf= 1	DP	DN
	Sınıf= -1	YP	YN

DP (Doğru Pozitif): Sınıflandırıcının pozitif olarak sınıflandırdığı ve gerçekte sınıfı pozitif olan test örneklerinin sayısıdır.

YN (Yanlış Negatif): Sınıflandırıcının negatif olarak sınıflandırdığı ve gerçekte sınıfı pozitif olan test örneklerinin sayısıdır.

YP (Yanlış Pozitif): Sınıflandırıcının pozitif olarak sınıflandırdığı ve gerçekte sınıfı negatif olan test örneklerinin sayısıdır.

DN (Doğru Negatif): Sınıflandırıcının negatif olarak sınıflandırdığı ve gerçekte sınıfı negatif olan test örneklerinin sayısıdır.

Performans değerlendirmesinde kullanılan temel ölçütler doğruluk, kesinlik, duyarlılık ve F-ölçütüdür.

▪ **Doğruluk**

Model başarımının ölçülmesinde kullanılan en temel ölçüttür. Doğruluk, doğru sınıflandırılmış örnek sayısının tüm örnek sayısına oranıdır.

$$\text{Doğruluk} = \frac{DP + DN}{DP + YN + DN + YP} \quad (2.12)$$

▪ **Kesinlik**

Doğru sınıflandırılmış pozitif örnek sayısının, pozitif sınıflandırılmış örnek sayısına oranıdır.

$$\text{Kesinlik} = \frac{DP}{DP + YP} \quad (2.13)$$

▪ **Duyarlılık**

Doğru sınıflandırılmış pozitif örnek sayısının, toplam pozitif örnek sayısına oranıdır.

$$\text{Duyarlılık} = \frac{DP}{DP + YN} \quad (2.14)$$

▪ **F-ölçütü**

Kesinlik (K) ve duyarlılık (D) ölçütlerinin harmonik ortalaması alınarak bulunur.

$$\text{F-ölçütü} = \frac{2 * K * D}{K + D} \quad (2.15)$$

Tablo 2.2. Benzerlik tabanlı metotların performans karşılaştırması [67]

Metot	Doğruluk	Kesinlik	Duyarlılık	F-ölçütü
Ortak Komşular (OK)	0,942	0,805	0,819	0,812
Jaccard Katsayısı (JK)	0,919	0,919	0,521	0,665
Adamic-Adar (AA)	0,896	0,807	0,426	0,557
Tercihli Bağlılık (TB)	0,296	0,176	0,979	0,298
Kaynak Tahsisi (KT)	0,902	0,792	0,388	0,521
Salton İndeksi (SI)	0,947	0,810	0,861	0,834
Sorensen İndeksi (SRI)	0,947	0,817	0,843	0,830
Hub Promoted (HP)	0,937	0,744	0,904	0,816
Hub Depressed (HD)	0,059	0,046	0,261	0,078
Leicht-Holme-Nerman (LHN)	0,877	0,637	0,459	0,534

Benzerlik tabanlı bağlantı tahmini algoritmalarının performans karşılaştırması için bir Facebook veri seti kullanılmıştır [67]. Algoritmaların doğruluğunu test etmek için kurulan ağdaki, bağlantıların %10' u çıkarılmış ve daha sonra gelecekteki olası bağlantıları tahmin etmek için benzerlik tabanlı bağlantı tahmini algoritmaları uygulanmıştır. Çıkarılan bağlantı uygulanan benzerlik tahmini algoritmaları ile bulunuyorsa, uygulanan algoritmanın iyi performans gösterdiği söylenebilir. Facebook veri setine uygulanan on tane benzerlik tabanlı bağlantı tahmini algoritmalarının performansı doğruluk, kesinlik, duyarlılık ve F-ölçütü bakımından karşılaştırılmıştır. Elde edilen sonuçlar Tablo 2.1' de gösterilmiştir. Sonuçlara göre HD ve PA algoritmalarının doğruluğu çok düşüktür ve bu algoritmaların bir sosyal ağda bağlantı tahmini için iyi sonuçlar vermediği gözlemlenmiştir. Kesinlik değeri JK için çok yüksek, HD için ise en düşük değere sahiptir. Duyarlılık en yüksek değerine TB algoritması, en düşük değerine ise HD algoritması ile ulaşmıştır. F-ölçütü ise OK, SI, SRI, HP algoritmaları için çok yüksek, HD ve TB algoritmaları için ise çok düşüktür. Tüm ölçütler göz önüne alındığında OK ve JK Facebook sosyal ağında bağlantı tahmini için en doğru ve en iyi sonuçları vermiştir.

2.2.2.2. Yol Tabanlı Ölçütler

Düğüm ve komşularından elde edilen bilgilerin yanı sıra iki düğüm arasındaki yollar düğüm çiftlerinin benzerliğini hesaplamak için kullanılabilir. Böyle metotlar yol tabanlı ölçütler olarak adlandırılır.

- **Yerel Yol (YY):**

Bu ölçüt 2 ve 3 birim uzunluğundaki yerel yolların bilgisini kullanır. Sadece en yakın komşuların bilgisini kullananlardan farklı olarak, mevcut durumda 3 birim uzunluktaki komşuların bilgisini de ek olarak kullanır [68]. 2 birim uzunluklu yollar 3 birim uzunluklu yollardan daha önemlidir. Bu yüzden ölçüte uygulanan bir ayarlayıcı α faktörü vardır. α , sıfıra yakın küçük bir sayı olmalıdır. A^2 ve A^3 , 2 ve 3 birim uzunluğuna sahip bitişiklik matrislerini gösterir. YY ölçütü de aynı zamanda 2 ve 3 birim uzunluktaki düğüm çiftlerini tanımlayan bir bitişiklik matrisidir.

$$YY = A^2 + \alpha A^3 \quad (2.16)$$

- **Friendlink (FL):**

FL ölçütü, sınırlı uzunluktaki tüm yolları kullanarak x ve y düğümleri arasındaki benzerliği ölçmek için tanımlanmıştır [69]. x ve y düğümleri arasındaki benzerlik x düğümünden y düğüme değişen uzunluktaki yolların sayısı olarak tanımlanır. Daha kesin ve daha hızlı bağlantı tahmini sağlar.

$$FL(x, y) = \sum_{i=1}^l \frac{1}{i-1} \cdot \frac{|yollar^i_{x,y}|}{\prod_{j=2}^i (n-j)} \quad (2.17)$$

- **Katz:**

Bu ölçüt düğüm çiftleri arasındaki bütün yollar düşünülerek hesaplanır. x ve y düğümleri arasındaki en kısa yolların sayısının toplamını verir [70]. $\beta > 0$, serbest bir parametre ve $|yollar(x,y)^{<1>}|$ ise x ve y düğümlerini bağlayan 1 birim uzunluklu tüm yolların kümesidir. β çok küçük seçildiği zaman uzun yolların son benzerliğe katkısı çok az olacağı için Katz ölçütünü ortak komşular ölçütüne çok benzer hale getirir.

$$Katz(x,y) = \sum_{l=1}^{\infty} \beta^l \cdot |yollar_{x,y}^l| \quad (2.18)$$

- **Köklü PageRank (PR):**

PageRank bir düğüme ağ üzerinde rastgele bir yürüyüş ile ulaşılabilme olasılığını gösterir. d faktörü orjinal x düğümünü yeniden dolaşma yerine komşuları ziyaret etme olasılığını gösterir [71].

$$PR(x,y) = (1-d) + d \sum_{z \in \Gamma(x)} \frac{P(z)}{|\Gamma(z)|} \quad (2.19)$$

- **Hitting Time (HT):**

Bu ölçüt bir ağ üzerindeki rastgele yollar kavramına dayalıdır. x düğümünden y düğümüne rastgele bir yola başlamak için gerekli adımların sayısıdır. İki düğüm için hesaplanan değer ne kadar kısa ise bu düğüm çifti birbirine o kadar benzerdir ve gelecekte aralarında bağlantı oluşması o kadar yüksektir.

$$HT(x,y) = 1 + \sum_{w \in \Gamma(x)} P_{x,w} HT(w,y) \quad (2.20)$$

2.2.3. Öğrenme Tabanlı Yaklaşımlar

Temel bağlantı tahmini ölçütlerinden sağlanan özellikler, iç özellikler ve dış bilgilere dayalı olarak birçok öğrenme tabanlı bağlantı tahmini yöntemi önerilmiştir. Bu öğrenme tabanlı yöntemler özellik tabanlı sınıflandırma, olasılıksal model ve matris faktörizasyonu olarak sınıflandırılabilir.

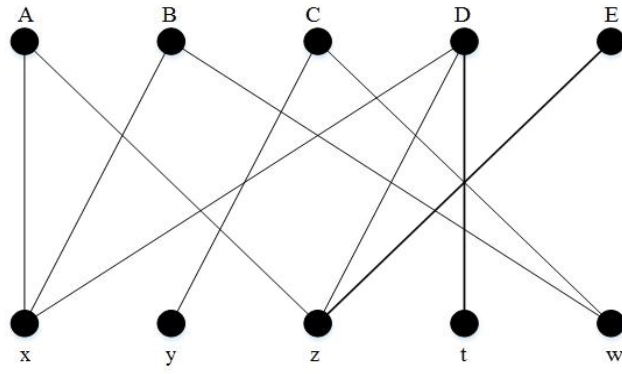
2.3. İki Parçalı Ağlarda Bağlantı Tahmini Yöntemleri

İki parçalı ağlarda bağlantı tahmini için genel bağlantı tahmini yöntemleri doğrudan uygulanamaz. Genel metotlardan farklı olarak iki parçalı ağlarda olası bağlantıları tahmin etmek için birkaç metot önerilmiştir.

2.3.1. Yansıma Tabanlı Bağlantı Tahmini

İki parçalı ağlar iki düğüm kümesinden oluşur. Kümelerdeki düğümler birbirinden farklı tiptedir. Örneğin; ürün - satıcı ilişkisini gösteren bir iki parçalı ağda bir kümedeki düğümler ürünleri, diğer kümedeki düğümler ise satıcıları gösterir. İki parçalı ağlarda farklı kümelerdeki düğümler arasında ilişki bulunur. Genel bağlantı tahmini metotları aynı tipten düğümler arasındaki ilişkileri bulmaya yöneliktir. Bu yüzden bu metotlar doğrudan iki parçalı ağlara uygulanamaz. İki parçalı ağlarda bağlantı tahmini için bu bölümde önerilen yöntemde iki parçalı ağın yansıma olarak adlandırılan tek parçalı ağa dönüştürülmesi, daha sonra genel bağlantı tahmini ölçütlerinin bu tek parçalı ağlara uygulanması gerçekleştirilir.

Bir iki parçalı ağın tek parçalı ağa nasıl dönüştürüldüğünü Şekil 2.4' te yer alan örnek ağ üzerinde inceleyelim.



Şekil 2.4. İki parçalı ağ örneği

Şekil 2.4' teki iki parçalı ağda x, y, z, t, w düğümleri hastalıkları, A, B, C, D, E düğümleri ise ilaçları gösterebilir. İlk olarak ağın yansımasını elde etmek için ağdaki düğümler arasındaki ilişkilere göre komşuluk matrisi oluşturulur.

	x	y	z	t	w
A	1	0	1	0	0
B	1	0	0	0	1
C	0	1	0	0	1
D	1	0	1	1	0
E	0	0	1	0	0

(2.21)

(2.21)' de iki parçalı ağda hastalıkların ilaçlar üzerinden ilişkisini bulmak için oluşturulan komşuluk matrisi devrik matris ile çarpılır.

1	0	1	0	0	*	1	1	0	1	0	=	2	1	0	2	1
1	0	0	0	1		0	0	1	0	0		1	2	1	1	0
0	1	0	0	1		1	0	0	1	1		0	1	2	0	0
1	0	1	1	0		0	0	0	1	0		2	1	0	3	1
0	0	1	0	0		0	1	1	0	0		1	0	0	1	1

(2.22)

(2.22)' de yer alan matriste her hastalığın kendisiyle ilişkisi bulunduğundan köşegen üzerindeki değerler sıfırlanır.

	A	B	C	D	E
A	0	1	0	2	1
B	1	0	1	1	0
C	0	1	0	0	0
D	2	1	0	0	1
E	1	0	0	1	0

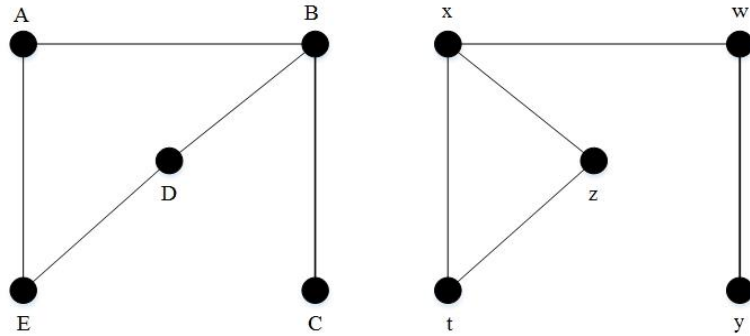
(2.23)

(2.23)' de gösterilen matris hastalıkların ilaçlar üzerinden bağlantısını gösteren tek parçalı ağın komşuluk matrisidir. İşlem tersi şekilde yapıldığında yani devrik matris ile komşuluk matrisi çarpıldığında ise ilaçların hastalıklar üzerinden ilişkisi bulunur. Bu ilişkiyi gösteren komşuluk matrisi (2.24)' te yer almaktadır.

$$\begin{array}{ccccc}
 1 & 1 & 0 & 1 & 0 \\
 0 & 0 & 1 & 0 & 0 \\
 1 & 0 & 0 & 1 & 1 \\
 0 & 0 & 0 & 1 & 0 \\
 0 & 1 & 1 & 0 & 0
 \end{array}
 *
 \begin{array}{ccccc}
 1 & 0 & 1 & 0 & 0 \\
 1 & 0 & 0 & 0 & 1 \\
 0 & 1 & 0 & 0 & 1 \\
 1 & 0 & 1 & 1 & 0 \\
 0 & 0 & 1 & 0 & 0
 \end{array}
 =
 \begin{array}{ccccc}
 3 & 0 & 2 & 1 & 1 \\
 0 & 1 & 0 & 0 & 1 \\
 2 & 0 & 3 & 1 & 0 \\
 1 & 0 & 1 & 1 & 0 \\
 1 & 1 & 0 & 0 & 2
 \end{array}
 \quad (2.24)$$

Köşegenler sıfırlanır ve tek parçalı ilaç ağına ait komşuluk matrisi (2.25)' teki gibi elde edilir.

$$\begin{array}{ccccc}
 & x & y & z & t & w \\
 x & 0 & 0 & 2 & 1 & 1 \\
 y & 0 & 0 & 0 & 0 & 1 \\
 z & 2 & 0 & 0 & 1 & 0 \\
 t & 1 & 0 & 1 & 0 & 0 \\
 w & 1 & 1 & 0 & 0 & 0
 \end{array}
 \quad (2.25)$$



Şekil 2.5. İki parçalı ağın yansımaları **(a)** Hastalıklar üzerinden tek parçalı yansıma
(b) İlaçlar üzerinden tek parçalı yansıma

İki parçalı hastalık - ilaç ağının artık tek parçalı Şekil 2.5 (a) ve (b)' de gösterildiği gibi hastalık ve ilaç olarak iki yansıma grafi oluşturulmuş olur.

Yansıma tabanlı bağlantı tahmini yönteminde, iki parçalı ağlar aynı türden düğümler içeren tek mod ağlara dönüştürüldükten sonra Ortak Komşular, Jaccard Katsayısı,

Adamic/Adar gibi benzerlik tabanlı bağlantı tahmini yöntemleri uygulanarak bağlantı tahmini işlemi gerçekleştirilir.

2.3.2. Ağırlıklı Yansıma ve İç Bağlantılar Kullanarak Bağlantı Tahmini

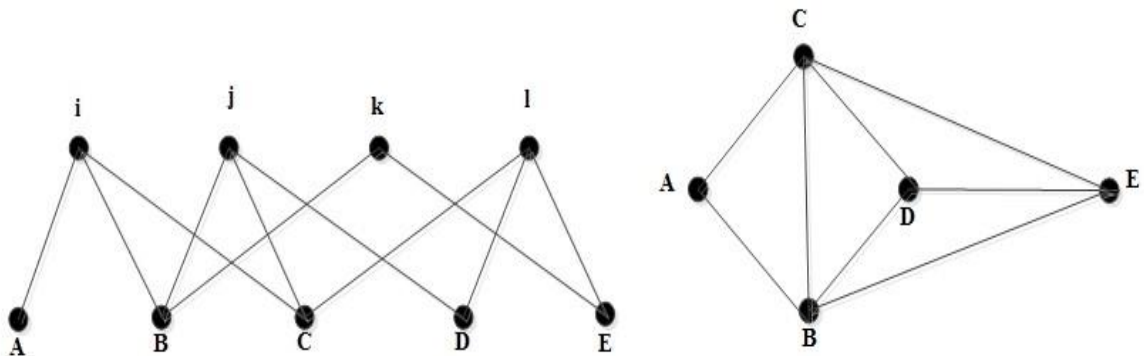
Bu yöntem belirli bir zamanda ağı anlık görüntülerinden gelecekte oluşabilecek bağlantıları tahmin etmeyi içerir. Bağlantı tahmini için bu yöntemde iç bağlantılar adı verilen özel bir tür bağlantı tanımlanmış ve bu bağlantılara dayalı bir yaklaşım önerilmiştir [15]. Bu metot iyi sonuçlara ulaşmış ve bu sonuçlar iç bağlantıların dinamik iki parçalı ağlar üzerinde önemli bir rol oynadığını göstermiştir.

$G=(X, Y, E)$ şeklinde tanımlanmış bir iki parçalı ağ olsun. Bu ağda X alt düğümlerin, Y üst düğümlerin, E ise X ve Y düğümleri arasındaki bağlantıların kümesidir. İki parçalı ağların yapısı gereği bir bağlantı sadece X ve Y kümelerindeki düğümler arasında oluşabilir. G ağında bir u düğümünün komşusu (2.26) 'deki gibi ifade edilir.

$$N(u) = \{v \in (X \cup Y), (u, v) \in E\} \quad (2.26)$$

Eğer $u \in X$ ise $N(u) \subseteq Y$ veya tersi şeklinde yorumlanır.

Bu metotta iç bağlantılar olarak tanımlanan bağlantıları bulmak için kurulan iki parçalı ağın yansımasının alınıp tek mod ağa dönüştürülmesi gerekir. Bu yansıma işlemi Bölüm 2.3.1' de belirtildiği gibi gerçekleştirilir.



Şekil 2.6. Örnek bir iki parçalı ağ ve X-yansıması G_x [15]

G' nin X -yansıması ağı, G' de u ve v düğümleri $(u,v) \in E_x$ ve en azından bir ortak komşuya sahipse $G_x = (X, E_x)$ şeklinde ifade edilir.

G_x ağı iki parçalı G ağından daha az bilgi içerir. Çünkü, u ve v düğümleri G ağında en azından bir ortak komşuya sahip ise G_x yansıma ağında bağlı olarak görülürler. Ancak bu düğümlerin ortak komşularının sayısı net olarak bilinemez. Bu bilgiyi elde etmek için bir yol ağırlıklı yansıma kullanmaktır. X -yansımasında bağlantıları ağırlıklandırmak için ağırlık fonksiyonları (w) kullanılır. Birkaç ağırlık fonksiyonu şunlardır:

- (u,v) bağlantısının ağırlığı G iki parçalı ağdaki ortak komşuların sayısı olarak tanımlanabilir.

$$\delta(u, v) = |N(u) \cap N(v)| \quad (2.27)$$

- (u,v) bağlantısının ağırlığı olarak Jaccard katsayısı kullanılabilir.

$$\gamma(u, v) = \frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|} \quad (2.28)$$

- (u,v) bağlantısının ağırlığını bulmak için delta fonksiyonu kullanılabilir.

$$\sigma(u, v) = \sum_{x \in N(u) \cap N(v)} \frac{2}{|N(x)| \cdot (|N(x)| - 1)} \quad (2.29)$$

Ağın ağırlıklı yansıması bir w ağırlık fonksiyonuna göre belirlendikten sonra iç bağlantılar olarak tanımlanan bağlantılar tahmin edilmeye çalışılır. İç bağlantılar şu şekilde ifade edilir:

$G = (X, Y, E)$ iki parçalı ağ, $(u,v) \notin E$ ve bu ağa $(u,v) \in X \times Y$ bağlantısının eklenmesiyle $G' = (X, Y, E \cup \{(u,v)\})$ elde edilmiş olsun. Eğer $G_x = G'_x$ ise (u,v) bağlantısı iç bağlantıdır. Yani bir G ağına (u,v) bağlantısının eklenmesi onun X -yansımasını değiştirmiyorsa o zaman (u,v) bir iç bağlantıdır. Örneğin; Şekil 2.6' daki ağ için (B,l) bir iç bağlantıdır. Çünkü l' nin G ağındaki tüm komşuları $N(l) = \{C,D,E\}$ zaten G_x yansımasında

B' ye bağlıdır. Yani (B,C), (B,D) ve (B,E) düğüm çiftleri sırasıyla i, j ve k olarak ortak bir komşuya sahip olduklarından yansıma ağda bağlıdırlar. (B,l) bağlantısını G' ye eklemek ortak komşu sayısını arttırır ancak X-yansımasını değiştirmez.

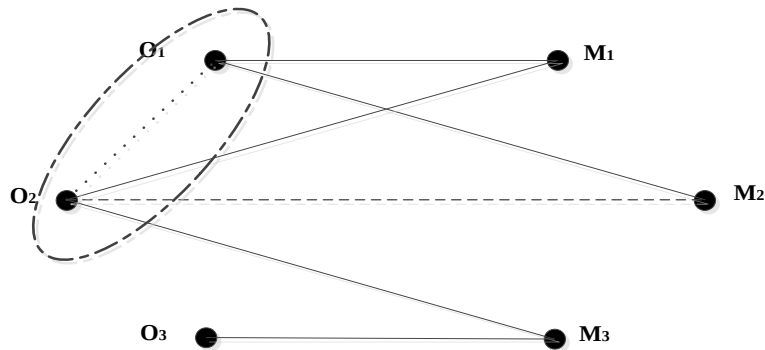
İki parçalı ağ ve yansıması incelenerek iç bağlantılar belirlenir. Ancak bunların tamamı dikkate alınmaz. Bunun için bir ağırlık eşik değeri belirlenir. Bulunan iç bağlantıların ağırlığı eğer bu eşik değerine eşit veya büyükse o zaman iç bağlantı kümesine dahil edilir.

$$E_{x_t} = \{(u, v) \in E_x, w(u, v) \geq \tau\} \quad (2.30)$$

Eşik değeri büyük seçildiğinde gerekli bağlantılar tahmin edilemeyebilir ya da küçük seçildiğinde gereksiz bağlantılar bulunabilir. Bunun için bağlantı tahmininde eşik değeri uygun olarak seçilmelidir.

2.3.3. Yarı-İki Parçalı Ağ ile Bağlantı Tahmini

Bu yöntemde bağlantı tahmininde kullanılan veriler için bir yarı-iki parçalı ağ modellemesi önerilmiştir. Modellenen ağda bağlantı tahmini için aynı kümede bulunan düğümler arasındaki ortak özellikler kullanılmıştır. Bu ortak özelliklerden yararlanarak oluşabilecek bağlantılar tahmin edilmeye çalışılmıştır. Örneğin; kişiler ve okudukları makalelerden oluşan bir iki parçalı ağ düşünelim. Bu ağ üzerinde bağlantı tahmini yapıp okurlara makale tavsiyesinde bulunulacaksa okuyucuların ortak özellikleri bulunmaya çalışılır. Bu özellikler; okurların daha önceden okumuş oldukları makalelerden belirlenen ilgi alanları veya aynı şehirde yaşama gibi özellikler olabilir.



Şekil 2.7. Örnek bir okur - makale grafi

Şekil 2.7' de yer alan okur - makale ağında O_1 ve O_2 okurları arasında benzer ilgi alanları gibi bir ortak özellik varsa O_1 ' in ve O_2 ' nin okuduğu makalelere bakılır. O_1 , M_1 ve M_2 makalelerini okumuşken, O_2 ise O_1 'in okuduğu makalelerden sadece M_1 makalesini okumuştur. O_1 ve O_2 arasında ortak özellik bulunduğundan O_2 'nin de O_1 'nin okuduğu M_2 makalesini okuyabileceği tahmininde bulunulur. Yani O_1 ve O_2 düğümlerinin sahip olduğu ortak özellikten yola çıkılarak O_2 ve M_2 arasında bir bağlantı kurulabileceği tahmini yapılır.

Günümüzde internet insanlar için bilgiye ulaşmada en büyük ve en çok tercih edilen ortam olmuştur. İnsanların bu büyük miktardaki veriler arasından kendileriyle alakalı olan konulara ulaşmaları da bir o kadar zorlaşmıştır. Bunu kolaylaştırmak için Nigam ve ark. [14] yukarıdaki yöntemi kullanarak kişilerin kendileriyle alakalı konulara daha çabuk ulaşmalarını sağlayacak bir çalışma yapmışlardır. Yapılan çalışmada kişiler arasındaki ilişkileri ve ilgili oldukları sağlık konularını bulmak için bir yarı iki parçalı ağ modellenmiş ve denetimli makine öğrenmesiyle bağlantı tahmini yaklaşımı kullanılarak kişilere geçmiş okuma alışkanlıklarına ve yaşadıkları şehir gibi demografik özelliklerine dayalı bir konu tavsiyesi yapılmıştır.

Bir kişinin okuma alışkanlığı ilgi alanı, yaşam tarzı, yaşı veya yaşadığı şehir gibi birçok faktörden etkilenir. Özellikle sağlık konusu birçok değişkenden etkilenir. Kişilerin okuduğu makaleler sahip oldukları bir hastalık ya da yaşadıkları bir problemin belirtisi olabilir, bilgilenmek istedikleri sağlıklı beslenme gibi konularla ilgili olabilir. Bu çeşitli nedenler kişilerin ilgilendikleri alanı belirlemede zorluk oluşturur. Bir makalenin bir kişi için uygunluğunu belirlemek zor bir iştir. Bunun için kullanıcıların ilgi alanları okudukları makaleler üzerinden belirlenmeye çalışılmış ve o alanlarda makaleler tavsiye edilmek istenmiştir.

Her makale bir konuyla ilgilidir. Aynı konudaki makaleler gruplanabilir ve o makaleleri okuyan kişilerin ilgili oldukları konular belirlenebilir. Bu çalışmada da kişilerin önceden okumuş oldukları alanlardaki makalelere göre ilgili oldukları konular belirlenmiş ve daha sonra yaşadıkları şehir gibi kişilerin özelliklerine bakarak onlarla ilişkili olabilecek diğer konular tahmin edilmeye çalışılmıştır.

Konu tavsiyesinde gerekli verileri modellemek için bir yarı iki parçalı ağ oluşturulmuştur. Bu yarı iki parçalı ağ $G = (V_1, V_2, E_1, E_2)$ olarak tanımlanır. V_1 bireyleri, V_2 ise konuları gösteren düğüm kümeleridir. E_1 , V_1 ve V_2 kümelerindeki düğümler

arasındaki bağlantılardır. E_2 ise V_1 kümesindeki düğümler arasında bulunacak ortak özellikleri göstermek için tanımlanacak bağlantılardır.

Kişilerin ilgili oldukları alanlar, çalışmada kullanılacak olan verilerin bulunduğu web sitesine üye olup konular seçtiklerinde veya bir konuyla ilgili bir makale okudukları zaman anlaşılır. Bu iki şekilde elde edilen verilerle hangi kullanıcıların hangi konularla ilgili oldukları belirlenir. Birey ve konu arasındaki bir bağlantı (E_1) o kişinin ilgi alanını gösterir. E_2 , kişiler arasında oluşacak bağlantılardır. Eğer iki kişi aynı şehirdense bunlar arasında bir bağlantı vardır. Aynı şehirden gelen insanlar grupları oluşturur ve grupların tamamı konulara bağlıdır. Aynı şehirden gelen insanların benzer konularla ilgileneceği düşünülerek o şehirde en çok ilgilenilen konu diğer bireyler için de önerilmiştir.

Bir kullanıcı için tüm siteyi dolaşarak kendisiyle ilgili olan konuları bulması zordur. Bu yüzden kullanıcıya en çok ilgileneceği konuyla ilgili bilgiyi sunmak önemlidir. Bu amaçla kişilerin demografik bilgisi ve okuma alışkanlıklarından yararlanılarak kullanıcıların ilgi alanları belirlenmiş ve buna göre konu tavsiyesinde bulunulmuştur.

2.3.4. Denetimli Makine Öğrenmesi Yaklaşımı ile Bağlantı Tahmini

İki parçalı ağlarda bağlantı tahmini için kullanılan yöntemlerden birisi de denetimli makine öğrenmesi yaklaşımıdır. Çok büyük miktardaki verilerin elle işlenmesi, analizinin yapılması mümkün değildir. Bu tür problemlere çözüm bulmak amacıyla makine öğrenmesi metotları geliştirilmiştir. Bu metotlar geçmişteki verileri kullanarak veriye en uygun modeli bulmaya çalışırlar. Yeni gelen verileri de bu modele göre analiz edip sonuç üretirler.

Makine öğrenmesi yaklaşımında denetimli ve denetimsiz olarak iki tür öğrenme yöntemi vardır. Denetimli öğrenmede sınıflar bilinir ve bu sınıflara ait örnekler bulunmaktadır. Var olan sınıfların örnekleri kullanılarak gelecek örneklerin hangi sınıflara ait olabileceği tahmin edilmeye çalışılmaktadır. Denetimsiz öğrenmede ise sınıflar bilinmez, olayların örnekleri bilinir. Yalnız bu verilerin yer aldığı sınıflar belirlenmemiştir. Algoritma var olan veriler üzerinde çalışarak kendi kendine yapılar bulmaya çalışır. İki parçalı ağlarda bağlantı tahmini için önerilen yöntemde denetimli öğrenme kullanılmıştır. Bu yöntemde belirli bir düğüm çifti arasında gelecekte bir bağlantı oluşması olasılığını tahmin etmek için bir sınıflandırıcı oluşturulur ve ağın geçmiş görüntülerinden çıkarılan

özellikler ile eğitilir. Daha sonra bir zaman periyodunda eğitim modeli test verileri üzerinde değerlendirilir [12,16].

Denetimli makine öğrenmesi yaklaşımındaki ana unsur, makine öğrenme algoritmalarını geliştirmek için bilgilendirici olduğu düşünülen özelliklerin çıkarılmasıdır. Özellikler doğrudan ve dolaylı özellikler olarak iki gruba ayrılır. Doğrudan özellikler iki parçalı ağdan, dolaylı özellikler ise ağı yansımalarından çıkarılır.

- **Doğrudan Özellik**

Tek parçalı ağlarda kullanılan ölçütlerin iki parçalı ağa uygulanabilmesi için aşağıdaki denklem kullanılır.

$$G' = \bigcap_{y \in G(x)} G(y) \quad (2.31)$$

Denklem (2.31)' de, bir x düğümü için $G(x)$ x' in komşularının kümesini, $G'(x)$ ise x' in komşularının komşularını ifade eder.

İki parçalı ağlarda seçilen düğümler farklı kümelerle ait olduğundan farklı türden düğümlerdir. Ancak tek mod ağlarda kullanılan topolojik özellikler aynı türden düğümler arasında bulunmaktadır. Örneğin; yazar - makale ağına seçilen düğümlerden biri yazarları temsil ederken, diğer düğüm makaleleri temsil eder ve aynı türden olmadıklarından doğrudan topolojik özellikler çıkarılamaz. Bu yüzden iki parçalı ağlarda topolojik ölçütler hesaplanırken aynı türden düğümleri kullanmak için (2.31)' de gösterildiği gibi x düğümünün komşularının komşuları kullanılır.

Ortak Komşular, Jaccard Katsayısı, Adamic/Adar ölçütleri (2.31)' deki denkleme bağlı olarak doğrudan özellikler olarak hesaplanır. Çünkü bu özellikler orijinal iki parçalı ağdan herhangi bir dönüşüm yapmadan elde edilebilir.

- **Dolaylı Özellik**

Bir iki parçalı ağ $G = \langle X, Y, E \rangle$ şeklinde gösterilmiş olsun. X ve Y düğüm kümelerini, E, ise bu iki kümedeki düğümler arasındaki kenarları ifade eder. Dolaylı özellik, doğrudan

özellik gibi iki parçalı ağdan doğrudan elde edilemez. Bu yüzden iki parçalı ağın tek parçalı ağa dönüştürülmesi gerekir.

$$G_x^n = \langle V_x \subseteq X, E = \{(a, b) : a, b \in X, |\Gamma_{G(a)} \cap \Gamma_{G(b)}| \geq n\} \rangle \quad (2.32)$$

$$G_y^m = \langle V_y \subseteq Y, E = \{(a, b) : a, b \in Y, |\Gamma_{G(a)} \cap \Gamma_{G(b)}| \geq m\} \rangle \quad (2.33)$$

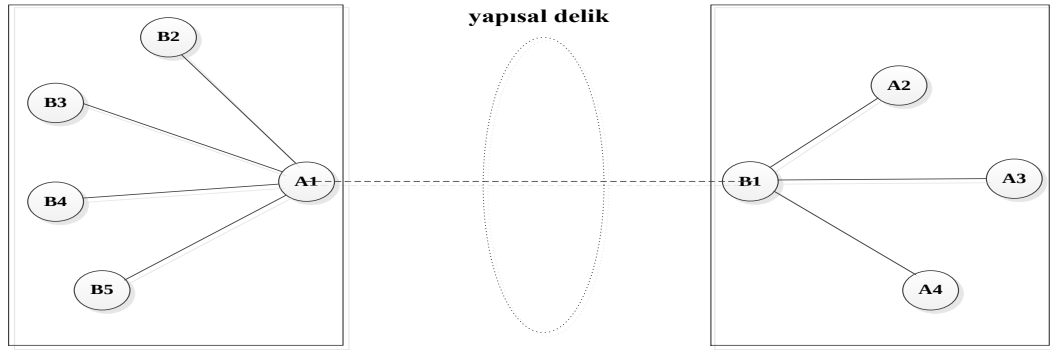
$$\varphi_{u \in \Gamma_G(a)} f_{G_E^n}^i(e, u) \quad (2.34)$$

$$\varphi_{v \in \Gamma_G(e)} f_{G_A^m}^i(v, a) \quad (2.35)$$

Bir iki parçalı ağın X ve Y düğüm kümelerine göre yansımaları kullanılarak (2.34) ve (2.35)' teki özellikler hesaplanır. f özelliği e ve a' nın tüm komşuları arasında hesaplanır. φ fonksiyonu ise özelliğe bağlı olarak $\{\min, \max\}$ kümesinden en uygun değeri seçer. Örneğin; özellik, ortak komşuların sayısı ise max fonksiyonu, en kısa uzaklık ise min fonksiyonu seçilir.

2.3.5. Yapısal Delikler Yardımıyla Bağlantı Tahmini

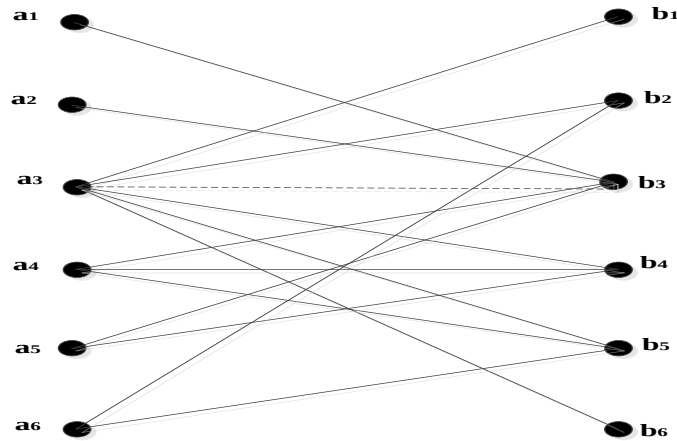
Yapısal delik yaklaşımı düğümler arasındaki zayıf bağlara dayanır. Eğer iki düğüm arasında güçlü bağ varsa, onların gelecekte ilişki kuracağı zaten büyük olasılıktadır. Yapısal deliklerde ise farklı düğüm kümeleri arasındaki zayıf bağlar kullanılır. Örneğin; x ve y kişileri arasındaki bağlantının yüksek bir gereksizlik derecesi varsa x ve y düğümlerini bağlama ne x düğümüne ne de y düğümüne yarar sağlamaz. Çünkü x düğümünün bildiği şeyi y düğümü de bilir ve burada yapısal delik oluşmaz. Eğer x ve y düğümleri arasında çok fazla ortak düğüm yok, yani ayrı gruplardan iseler yapısal delik oluşur. Çünkü x ve y düğümlerini bağlama bir gruptan diğer gruba bilgi akışı sağladığı için yararlı olacaktır [13].



Şekil 2.8. Yapısal delik örneği

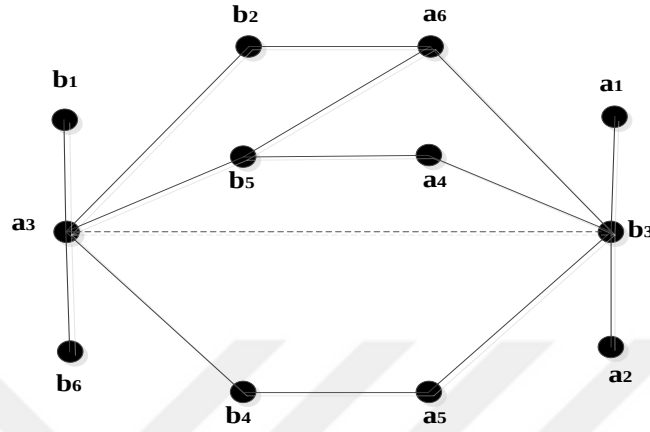
Yapısal delikleri belirlemede zor işlerden biri gereksizliği belirlemektir. Gereksizlik fazla ise orada düğümler arasında yapısal delik oluşmaz, az ise yapısal delik oluşur. Gereksizlik için en yaygın kullanılan ölçüt düğümlerin ortak komşularının sayısıdır. Eğer seçilen düğümlerin ortak komşuları aynı ise yapısal delik oluşmaz. Çünkü düğümler arasında ilişki oluşturmak gereksizdir. Eğer düğümler arasında ortak komşu az ise farklı kümelerle ait bu iki düğüm arasında bilgi akışını sağlamak amacıyla yapısal delik oluşur.

Düğümler arasında yapısal delik oluşumunu göstermek için $G[t,t'] = \langle V, E \rangle$ şeklinde bir ağ tanımlanmış olsun. V düğümler kümesini, E ise t ve t' zaman aralığındaki bağlantıları gösterir. V düğüm kümesi $X \cap Y = \emptyset$ olacak şekilde ayrılır. Her bağlantı t ve t' zaman aralığında (x,y) şeklinde gösterilir. $[t,t']$ zaman aralığı $t_0 < t_1 < t_2 < t_3$ olacak şekilde dört zaman dilimine ayrılır. Ağın bir gözlem ağı $G_0 = G[t_0, t_1]$ ve bir bağlantı oluşum ağı $G_f = [t_2, t_3]$ olacak şekilde iki alt ağı oluşturulur. Bağlantı tahmininin bu ağlar üzerindeki işlevi G_0 ağında yer almayan ancak G_f ağında yer alan bağlantıları bulmaktır.



Şekil 2.9. Yapısal delik görülen örnek bir graf

Yapısal delik yaklaşımını Şekil 2.9’ da verilen ağ üzerinde inceleyelim. Ağda düğümler arasındaki güçlü bağlantılar gösterilmiştir. Yapısal delikleri belirleyebilmek için ağı yeniden düzenleyelim.



Şekil 2.10. İki parçalı ağın düzenlenmiş hali

Bu ağ üzerinde (a_3, b_3) arasında oluşacak bir bağlantının G_f ağında görülüp görülmeyeceği incelenir. (a, b) düğüm çiftinin komşuları ortak düğümler ve ayrık düğümler olarak ikiye ayrılır. Ortak düğümler, a ve b düğümleri arasındaki en kısa yolda bulunan komşu düğümlerdir. Ayrık düğümler ise a ve b düğümleri arasındaki yollarla ulaşılamayan düğümlerdir. Örneğin; Şekil 2.10’ da yer alan ağda (a_3, b_3) düğüm çifti için ortak düğümleri $\{a_4, a_5, a_6\}$ ve $\{b_2, b_4, b_5\}$, ayrık düğümleri ise $\{a_1, a_2\}$ ve $\{b_1, b_6\}$ dır.

Yapısal delikler için en basit ölçüt düğümlerin komşuları arasındaki eksik bağlantıların sayısını hesaplamaktır. Şekil 2.10’ daki ağa bakarak ortak düğümlerden eksik bağlantı sayısı $3 \times 3 = 9$ olarak bulunur. Ancak bunlardan 4 tanesi ağ üzerinde yer alır. Geri kalan 5 bağlantı eksiktir. Bu eksik bağlantılara c-tipi eksik bağlantılar adı verilir. Aynı şekilde ayrık düğümlerden de eksik bağlantıların sayısı hesaplanabilir. Ortak düğümler arasında oluşacak maksimum bağlantı sayısı ise $5 \times 5 = 25$ tir. Buna göre $25 - 9 = 16$ eksik bağlantı bulunur ve ayrık düğümlerden kaynaklı bu bağlantılar s-tipi eksik bağlantılar olarak adlandırılır.

$$AL(a, b) = AL^c(a, b) + \alpha * AL^s(a, b) \quad (2.29)$$

Yapısal delikler c-tipi ve s-tipi eksik bağlantıların doğrusal bir kombinasyonu şeklinde (2.29)'daki gibi oluşturulur. AL^c ve AL^s c-tipi ve s-tipi eksik bağlantıların, α ise bu iki niceliği kontrol eden bir parametreyi gösterir. Örnek verilen (a_3, b_3) için yapısal delik $5+16\alpha$ şeklinde oluşur. $\alpha=0$ olduğunda ayırık düğümlerin eksik bağlantıları, $\alpha=\infty$ ise ortak düğümlerin eksik bağlantıları göz ardı edilir.



3. ÇEVİRİMİÇİ FORUM SİTELERİNDE HASTALIK-BELİRTİ BAĞLANTILARININ TAHMİNİ

Günümüzde bilgisayar ve internet teknolojileri her alanda olduğu gibi sağlık alanında da yaygın olarak kullanılmakta ve kullanım oranları hızla artmaktadır. Türkiye İstatistik Kurumu tarafından gerçekleştirilen Hane Halkı Bilişim Teknolojileri Kullanım Araştırması 2016 sonuçlarına göre, kullanım amaçları dikkate alındığında 2016 yılının ilk üç ayında internet kullanan bireylerin % 82,4' ü sosyal medya üzerinde profil oluşturma, mesaj gönderme veya fotoğraf vb. içerik paylaşırken, bunu % 74,5 ile paylaşım sitelerinden video izleme, % 69,5 ile online haber, gazete ya da dergi okuma, % 65,9 ile sağlıkla ilgili bilgi arama, % 65,5 ile mal ve hizmetler hakkında bilgi arama ve % 63,7 ile internet üzerinden müzik dinleme takip etmiştir. Görüldüğü üzere internet büyük oranda insanlar için sağlık konusunda bir başvuru kaynağı haline gelmiştir.

İnternet insan hayatının vazgeçilmez bir parçası olmuştur. İnsanlar artık merak ettiği sorulara internet sayesinde kolayca cevap bulabilmektedir. İnternetin sağladığı kısa, zahmetsiz ve ücretsiz yol insanlara sağlıklarıyla ilgili merak ettiği konularda fikir sahibi olmaları için son derece cazip gelmektedir. İnsanlar interneti sağlık konusunda da birçok amaçla kullanmaktadır. En aktif kullanım amaçlarından birisi belli hastalık belirtilerine sahip olduklarında bu belirtilerin hangi hastalık ile ilişkili olabileceği hakkında fikir sahibi olmaktır. İnsanlar bazı belirtilere sahip olduklarında doktora gidip birtakım işlemlerden geçerek şikayetlerinin hangi hastalıkla ilgili olduğunu öğrenmeden önce kolay ve daha hızlı yoldan bilgi sahibi olabilecekleri interneti kullanırlar. İnternet üzerinde hasta ve doktorların iletişim halinde oldukları çevrimiçi birçok forum sitesi bulunmaktadır. Birtakım belirtilere sahip olan insanlar belirtilerinin hangi hastalıkla ilişkili olabileceği hakkında bu sitelere şikayetlerini yazıp uzman doktorlardan bilgi almaktadır.

Bu çalışmada günümüzde üzerinde yoğun çalışmaların yapıldığı ve birçok alana uygulanan sosyal ağlarda bağlantı tahmini kullanım oranı giderek artan çevrimiçi forum sitelerinin analizinden elde edilen verilerle kurulmuş olan soru-hastalık ağı üzerinde gerçekleştirilmeye çalışılmıştır.

3.1. Hastalık ve Belirti Veri Setlerinin Oluşturulması

Bu bölümde çevrimiçi forum sitelerine kullanıcıların yapmış oldukları yorumlar analiz edilip sordukları sorulardan hangi hastalık belirtilerine sahip oldukları belirlenmiştir. Forum sitelerinde insanlar yorum yazarken belirtilerini tıp dilinde geçtiği şekilde değil de günlük konuşma dilinde kullandıkları gibi yazarlar. Her insan şikayetlerini farklı cümleler ile ifade ederler. Kişilerin sordukları sorulardan elde edilen belirti verileri kişinin şikayetlerini ifade ettiği şekle göre değişiklik göstermektedir. Bir insan baş ağrısı şikayetini ‘başım ağrıyor’ şeklinde ifade ederken başka bir insan bunu ‘başım çatlıyor’ ya da daha farklı bir şekilde ifade edebilir. Şikayetlerin farklı şekillerde yazılması elde edilecek belirtilerin de farklı farklı olmasına neden olmaktadır. Çalışmada bu farklılığın önüne geçmek ve her kullanıcının farklı şekillerde yazmış olduğu aynı şikayeti tek tip, tam ve doğru olarak elde edebilmek için belirtilerin standart ICD-10 kodları kullanılmıştır. Bu şekilde sorulardaki şikayetlerden elde edilen belirti isimleri tek formda elde edilmiş ve bu da bir sonraki adımda gerçekleştirilecek olan belirtilere karşılık gelebilecek hastalıkların belirlenmesi aşamasını kolaylaştırmıştır. Hastalıklar belirlenirken aynı standart yapının korunması için yine ICD-10 kodları kullanılmıştır. Bu kodlar ile belirtiler ve bu belirtilere karşılık gelen hastalıklar standart bir formda elde edilmiştir.

Hastalıkların istatistiksel sınıflandırması, hastalık verilerinin kullanıcı tarafından kolayca değerlendirilip incelenebileceği bir formda sunulmasına gereksinim duyar. O halde, kullanışlı ve anlaşılır bir enformasyon elde edebilmek için, hastalıkların sistemli ve anlamlı bir şekilde düzenlenmesi gereklidir. Bu amaçla, tamamen teorik bir yaklaşım yerine pratik bir yaklaşımda bulunmak en uygun yoldur ve ICD' nin geliştirilmesinde bu yaklaşım göz önüne alınmıştır.

ICD (International Classification of Diseases) Kodları hastalıkların ve sağlık sorunlarının uluslararası platformda tek bir biçimde temsil edilebilmesini sağlayan istatistik kodlar bütünüdür. 20.000' den fazla bilimsel makalede ve dünya genelinde 100' den fazla ülkede kullanılmaktadır. ICD Kodları, Dünya Sağlık Örgütü (WHO) tarafından yayınlanan ve bilinen hastalıkları ile diğer sağlık sorunlarını belirlemek için benzersiz alfa-nümerik kodları kullanan Uluslararası Hastalık Sınıflamasına dayanır. WHO'ya göre, hekimler, kodlayıcılar, sağlık bilgi yöneticileri, hemşireler ve diğer sağlık uzmanları, teşhis bilgilerini alma ve depolamada işlerini kolaylaştırmak için ICD-10 kodlarını kullanmaktadır.

ICD kod listesinde 68.000' den fazla teşhis kodu bulunmaktadır. ICD-10 yedi karakterden oluşan alfanümerik bir koddur. Her kod bir harf ile başlar. Bu harfin ardından iki sayı gelir. ICD-10' un ilk üç karakteri kategoridir. Kategori, yaralanma veya hastalığın genel türünü tanımlamaktadır. Kategori, bir ondalık basamağı ve alt kategori tarafından izlenir. Bunu, yaralanma veya hastalığın nedenini, belirtisini, yerini, şiddetini ve türünü daha ayrıntılı olarak açıklayan iki alt sınıflandırma takip eder. Son karakter uzantıdır. Uzantı, karşılaşmanın türünü açıklar. Yani, bir sağlık uzmanı hastayı bir durum/yaralanma/hastalık için ilk kez gördüğünde, "ilk karşılaşma" olarak listeler. İlk karşılaşmadan sonraki her karşılaşma "bir sonraki karşılaşma" olarak listelenir.

ICD-10'da toplam 21 bölüm yer almaktadır. Bu bölümler Tablo 3.1'de gösterilmiştir. Görüldüğü üzere R harfi ile başlayan bölüm belirtileri, diğer bölümler ise hastalıkları ifade etmektedir.

Tablo 3.1. ICD-10 bölümleri

Bölüm	Konu
A00-B99	Bazı enfeksiyonlar ve paraziter hastalıklar
C00-D49	Neoplazmlar
D50-D89	Kan ve kan yapıcı organ hastalıkları ve bağışıklık mekanizmasını içeren bazı bozukluklar
E00-E89	Endokrin, beslenme ve metabolizma hastalıkları
F01-F99	Mental ve davranışsal bozukluklar
G00-G99	Sinir sistemi hastalıkları
H00-H59	Göz ve gözle bağlantılı doku hastalıkları
H60-H95	Kulak ve mastoid çıkıntı hastalıkları
I00-I99	Dolaşım sistemi hastalıkları
J00-J99	Solunum sistemi hastalıkları
K00-K95	Sindirim sistemi hastalıkları
L00-L99	Deri ve derialtı dokusu hastalıkları
M00-M99	Kas iskelet sistemi ve bağ dokusu hastalıkları
N00-N99	Genitoüriner sistem hastalıkları
O00- O9A	Gebelik, doğum ve lohusalık
P00-P96	Perinatal dönemden kaynaklanan bazı durumlar
Q00-Q99	Konjenital malformasyonlar, deformasyonlar ve kromozom anomalileri
R00-R99	Semptomlar, belirtiler ve anormal klinik ve laboratuvar bulguları
S00-T88	Yaralanma, zehirlenme ve dış nedenlere bağlı diğer durumlar
V00-Y99	Hastalık ve ölümün dış sebepleri
Z00-Z99	Sağlık durumu ve sağlık hizmetlerinden yararlanmayı etkileyen faktörler

ICD KODU	TANI
A00	Kolera
A00.0	Kolera, <i>Vibrio cholorea</i> 01, biovar kolera'ya bağlı
A00.1	Kolera, <i>Vibrio cholerae</i> 01, biovar eltor'a bağlı
A00.9	Kolera, tanımlanmamış
A01	Tifo ve paratifo
A01.0	Tifo
A01.1	Paratifo A
A01.2	Paratifo B
A01.3	Paratifo C
A01.4	Paratifo, tanımlanmamış
A02	Salmonella enfeksiyonları, diğer
A02.0	Salmonella enteriti
A02.1	Salmonella septisemisi
A02.2	Salmonella enfeksiyonları, lokalize
A02.8	Salmonella enfeksiyonları, diğer tanımlanmış
A02.9	Salmonella enfeksiyonu, tanımlanmamış
A03	Şigelloz
A03.0	Şigelloz, <i>Shigella dysenteriae</i> 'ye bağlı
A03.1	Şigelloz, <i>Shigella flexneri</i> 'ye bağlı
A03.2	Şigelloz, <i>Shigella boydii</i> 'ye bağlı
A03.3	Şigelloz, <i>Shigella sonnei</i> 'ye bağlı
A03.8	Şigellozlar, diğer
A03.9	Şigelloz, tanımlanmamış

Şekil 3.1. Örnek ICD-10 hastalık kodları

ICD KODU	TANI
R00	Kalp atım anormallikleri
R00.0	Taşikardi, tanımlanmamış
R00.1	Bradikardi, tanımlanmamış
R00.2	Çarpıntı
R00.8	Nabız anormallikleri, diğer ve tanımlanmamış
R01	Kardiak üfürüm ve diğer kardiak sesler
R01.0	Benign ve zararsız kalp üfürümleri
R01.1	Kardiyak üfürüm, tanımlanmamış
R01.2	Kalp sesleri, diğer
R02	Gangren, başka yerde sınıflanmamış
R03	Kan basıncının anormal olması, tanısı olmayan
R03.0	Kan basıncının yüksek olması, hipertansiyon tanısı olmayan
R03.1	Özellik göstermeyen düşük tansiyon bulgusu
R04	Solunum yollarından kaynaklanan kanama
R04.0	Epistaksis
R04.1	Boğazdan kan gelmesi
R04.2	Hemoptizi
R04.8	Solunum yollarında herhangi bir yerde hemoraji
R04.9	Solunum yollarından hemoraji, tanımlanmamış

Şekil 3.2. Örnek ICD-10 belirti kodları

Hastalık ve belirtiler örnek ICD-10 kodları Şekil 3.1 ve 3.2' de verilmiştir. Hastalıklar ve belirtilere ait ICD-10 kodları ve isimleri veri tabanına kaydedilerek analiz işleminde kullanılmak üzere bir veri seti oluşturulmuştur.

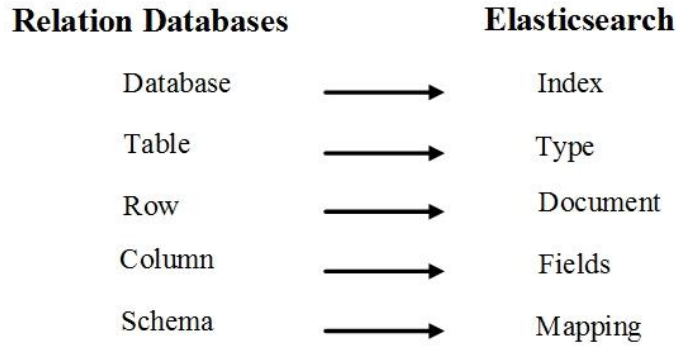
3.2. Çevrimiçi Forum Sitelerinin Analizi

Çevrimiçi forum siteleri insanların birtakım konular hakkında sorular sorduğu, uzman kişiler ya da diğer kullanıcılar tarafından cevaplar aldığı yerlerdir. Sağlık forumları ise tıbbi konular hakkında sorular sorup konusunda uzman kişilerden cevaplar alınan yerlerdir. Sağlık forumları son zamanlarda büyük ilgi görmektedir. Sağladığı kolaylık, ücretsiz olması gibi durumlar göz önüne alındığında artık insanlar doktora gidip birtakım tetkiklerden geçmeden önce bu siteler aracılığıyla şikayetlerinin hangi hastalıklar ile ilgili olabileceği konusunda uzman kişilere sorular sorup durumları hakkında bir fikir sahibi olmaya yönelmektedirler. Bu yönelim sonucunda çevrimiçi sağlık forumlarında binlerce soruyla karşılaşmaktadır. Bu sitelerde sorulan sorular yapılan çalışmada veri kaynağı olarak kullanılmıştır.

İnsanların forumlarda sordukları sorulardan hastalık belirtilerini elde edebilmek için ElasticSearch veri analiz aracı kullanılmıştır. Elasticsearch; platformdan bağımsız, Java ile geliştirilmiş açık kaynak kodlu, geniş dağıtım, kolay ölçeklenebilir, kurumsal düzeyde bir arama motorudur. Lucene tabanlı olarak geliştirilmiş bir tam metin arama teknolojisidir. Kapsamlı ve ayrıntılı bir API aracılığıyla erişilebilen Elasticsearch, büyük veriler içerisinde son derece hızlı aramalar yapmayı destekler.

Günümüzde sosyal medyada, iletişim, sağlık gibi alanlarda büyük miktarda veriler üretilmektedir. Bilişim dünyasında bu veriler big data olarak adlandırılır. Bu veriler başlangıçta dağınıktır ve tek başına bir anlam ifade etmez. Bu nedenle bu büyük miktardaki verilerin analiz edilip işlenmesi gerekir. Elasticsearch big data dünyasında büyük verilerin hızlı bir şekilde analizi için geliştirilmiş bir araçtır. Güçlü, esnek, dağıtık ve gerçek zamanlı olması diğer teknolojilere olan üstünlükleridir.

Çok büyük miktardaki veriler veri tabanlarında saklandığında bu veriler üzerinde gerçekleştirilen sorgularda hız ve performans konusunda sorunlar yaşanmaktadır. Bu sorunu gidermek için metinlerde belirlenen kelimeler indekslenir. Veriler kaydedilirken kelimelerin hangi dokümanda geçtiği indekslenir. Sorgulama işlemlerinde tüm veriler yerine indekslenen kelimeler üzerinde arama yapılır. Böylece veri tabanlarından daha hızlı cevaplar elde edilir.



Şekil 3.3. Elasticsearch ve veri tabanı eşleştirmesi

Şekil 3.3’ te Elasticsearch ve veri tabanlarının yapısı eşleştirilmiştir. Elasticsearch terimlerinin veri tabanlarındaki karşılıkları verilmiştir.

Çevrimiçi forum sitelerinde insanların soru olarak yazdıkları metinleri Elasticsearch ile analiz etmek için ilk olarak standart belirti isimleri ve ICD kodları Elasticsearch’ e indekslenmiştir. Daha sonra sorular analiz edilerek bu sorularda geçen hastalık belirtileri yani kişilerin şikayetleri belirli benzerlik skorları ile belirlenmiştir. Bu skorlar içerisinde yüksek benzerlik skoruna sahip olan belirtiler kişinin sahip olduğu hastalık belirtileri olarak kabul edilmiştir. Her soru bu şekilde analiz edilerek kullanıcının sorusunda ifade ettiği belirtiler bulunmuştur.

3.3. Hastalık Belirtilerinin Elde Edilmesi

Forum sitelerine yazılmış sorular analiz edilip kullanıcı şikayetlerinin belirlenmesinin ardından bu belirtilerin hangi hastalıklar ile ilişkili olabileceğinin bulunabilmesi için hangi hastalıkların hangi belirtilere sahip olduğu bilgisine ihtiyaç duyulmuştur. Bu bilgileri elde edebilmek için www.drugs.com [72] web sitesinden yararlanılmıştır.

Drugs.com; en popüler, kapsamlı ve en güncel çevrimiçi ilaç bilgi kaynağıdır. 24.000'den fazla reçeteli ilaç, doğal ürünler hakkında ücretsiz, bağımsız ve doğru bilgi sağlar. İnternet üzerinde en çok ziyaret edilen ilaç bilgi sitelerindendir. Hem tüketiciler hem de sağlık çalışanları için net, kapsamlı ve güncel bilgileri içerir.

Drugs.com web sitesinde ilaçlar, sınıflarına ve kullanılma koşullarına göre sınıflandırılmıştır. Bu çalışmada ilaçların hastalıklara göre kullanılması sınıflandırmasında hastalıklar hakkında birçok bilgiye ulaşılmaktadır. Hastalıklar hakkında genel bilgi,

hastalık belirtileri, nedenleri, teşhisi, tedavisi ve önlenmesi gibi bilgiler bulunmaktadır. Şekil 3.4’ te bu sitede hastalıklara ait bilgilerin nasıl yer aldığı gösterilmiştir. Çalışmada ihtiyaç duyulan bilgi bu sitede yer alan her hastalığa ait belirti bilgileridir. Siteden elde edilen hastalıklar ve bu hastalıklara ait belirti bilgileri veri tabanına sorulardaki belirtilere karşılık gelen hastalıkları bulma işleminde kullanılmak üzere kaydedilmiştir.

The screenshot shows the Drugs.com website interface. At the top, there is a search bar and a navigation menu with links like 'DRUGS A-Z', 'PILL IDENTIFIER', 'INTERACTIONS CHECKER', 'NEWS', 'Q & A', and 'PRO ED'. The main heading is 'Asthma'. Below it, there are tabs for 'Health Guide', 'Disease Reference', 'Care Notes', 'Medication List', 'Encyclopedia', and 'Q & A'. The 'Encyclopedia' tab is selected. The text describes asthma as a disease causing airway swelling and narrowing, leading to wheezing, shortness of breath, chest tightness, and coughing. It lists common symptoms like cough, wheezing, and shortness of breath, and also mentions emergency symptoms that require medical help, such as bluish color, decreased alertness, extreme difficulty breathing, rapid pulse, severe anxiety, and sweating. Other symptoms like abnormal breathing pattern, temporary breathing stops, chest pain, and chest tightness are also listed.

Asthma

Health Guide Disease Reference Care Notes Medication List Encyclopedia Q & A

Asthma is a disease that causes the airways of the lungs to swell and narrow, leading to wheezing, shortness of breath, chest tightness, and coughing.

Asthma Symptoms

Most people with asthma have attacks separated by symptom-free periods. Some people have long-term shortness of breath with episodes of increased shortness of breath. Either wheezing or a cough may be the main symptom.

Asthma attacks can last for minutes to days, and can become dangerous if the airflow is severely blocked.

Symptoms include:

- **Cough** with or without sputum (phlegm) production
- Pulling in of the skin between the ribs when breathing (**intercostal retractions**)
- **Shortness of breath** that gets worse with exercise or activity
- Wheezing

Emergency symptoms that need prompt medical help:

- **Bluish color** to the lips and face
- Decreased level of alertness, such as severe drowsiness or confusion, during an asthma attack
- Extreme **difficulty breathing**
- **Rapid pulse**
- Severe **anxiety** due to shortness of breath
- Sweating

Other symptoms that may occur:

- Abnormal breathing pattern -- breathing out takes more than twice as long as breathing in
- Breathing temporarily stops
- **Chest pain**
- Tightness in the chest

Şekil 3.4. drugs.com web sitesinin yapısı [72]

3.4. Soru-Hastalık Ağının Kurulması

Bağlantı tahmini konusunda yapılan çalışmaların çoğu tek mod ağlar üzerinedir. Genel bağlantı tahmini yöntemleri de genellikle tek mod ağlar üzerinde gerçekleştirilmiştir. Son zamanlarda iki parçalı ağlarda bağlantı tahmini konusuna bir yönelim başlamıştır. Bu çalışmada araştırmacılar tarafından daha çok ilgi görmeye başlayan iki parçalı ağda bağlantı tahmini yapılmaya çalışılmıştır. Bunun için bağlantı tahmini işleminin gerçekleştirileceği ağ, düğüm kümelerinin sorular ve hastalıklardan oluştuğu bir iki parçalı ağ şeklinde modellenmiştir.

İki parçalı ağda sorular ve hastalıklar olarak iki farklı düğüm kümesi yer almaktadır. Sorular düğüm kümesinde çevrimiçi forum sitelerinden elde edilen her bir soru bir düğüm olarak gösterilmiştir. Hastalıklar düğüm kümesinde ise drugs.com web sitesinden elde edilen verilerle hazırlanmış olan veri tabanındaki her bir hastalık bir düğüm olarak ifade edilmiştir. Ağ kurulurken düğümler soru numaraları ve hastalık ICD kodları ile tanımlanmıştır. Soru ve hastalık düğümleri arasındaki ilişkiler, kullanıcının yazmış olduğu soruda ifade ettiği şikayetlerinin başka bir deyişle sahip olduğu hastalık belirtilerinin hangi hastalıklarla ilişkili olabileceğini göstermektedir. Bu düğümler arasındaki ilişkilerin kurulabilmesi için Bölüm 3.2’ de anlatıldığı gibi forum sitelerinde yer alan sorular analiz edilip kişilerin bahsettikleri hastalık belirtileri elde edilmiştir. Kişinin sahip olduğu bu belirtilerin hangi hastalık ile ilişkili olabileceği Bölüm 3.3’ te anlatılan hastalıklara ait belirtilerden oluşan bir veri tabanı kullanılarak belirlenmiştir. Bu tespit işleminde sorulardan elde edilen belirtiler ile her hastalığın belirtileri karşılaştırılmış ve kişinin sahip olduğu belirtiler en çok hangi hastalıkların belirtileri ile kesişiyorsa bu hastalıkların sorudaki şikayetlerle ilgili olduğu saptanmıştır. Bu karşılaştırma sonucunda elde edilen sorular ve ilişkili hastalık bilgileri ile ağ üzerinde yer alan ilgili soru ve hastalık düğümleri arasında bağlantılar kurulmuştur.

3.5. Önerilen Yöntem

Bu çalışmada sağlığa yönelik çevrimiçi forum sitelerinde kullanıcılar tarafından sorulmuş sorular ve bu sorulardaki şikayetlere göre belirlenmiş hastalıklardan oluşan soru-hastalık iki parçalı ağında bağlantı tahmini yapılmıştır. Kurulan ağ iki parçalı yapıda olduğundan genel tek mod ağlara yönelik olarak oluşturulmuş bağlantı tahmini metotları

bu ağ üzerinde uygulanamaz. Bunun için son zamanlarda iki parçalı ağlara yönelik olarak geliştirilmiş Bölüm 2.3' te anlatılan bağlantı tahmini metotlarının kullanılması gerekmektedir. Bu çalışmada oluşturulan iki parçalı ağda bağlantı tahmini işlemi için iç bağlantılar yöntemi kullanılmıştır. Bu yöntem geliştirilen diğer yöntemler ve genel bağlantı tahmini yöntemleri ile karşılaştırıldığı çalışmalarda daha iyi performans göstermiş ve daha yüksek doğrulukta bağlantı tahmini sağlamıştır.

Kurulan soru-hastalık ağı iki tip düğüm içermektedir. İç bağlantılar yöntemi düğümler aynı türden olmadığından doğrudan iki parçalı ağ üzerine uygulanamaz. Bu yüzden ilk olarak iki parçalı ağın tek tip düğümlerden oluşan yansıma ağ olarak adlandırılan tek mod ağa dönüştürülmesi gerekir. İki parçalı soru-hastalık ağının Bölüm 2.3.1' de anlatıldığı gibi sadece hastalık ve sadece soru düğümlerinden oluşan yansıma ağları oluşturulmuştur. İki parçalı ağlar tek mod ağlara dönüştürüldüğünde oluşturulan yansıma ağ iki parçalı ağ hakkında daha az bilgi içerir. Örneğin, yansıma ağ oluşturulurken iki düğüm arasında en az bir ortak komşu olması bu düğümler arasında yansıma ağda bir bağlantı oluşturulması için yeterlidir. Ancak bağlantı oluşturulan iki düğüm arasında daha fazla ortak komşu olabilir. Bu ortak komşuların sayısı bu düğümlerin ne kadar benzer olduklarını gösterir ve bu benzerlik bağlantı tahmininde kullanılan en önemli ölçütlerden birisidir. Yansıma ağ ağırlıklı olarak oluşturulmadığı zaman bu iki düğüm arasındaki ortak komşu sayısı bilgisine ulaşamaz. Bu yüzden iki parçalı ağ hakkında daha fazla bilgi sağlayabilmesi için yansıma ağ ağırlıklı olarak oluşturulmuştur. Yansıma ağdaki bağlantıların ağırlığını belirlemek için (3.1)' deki ağırlık fonksiyonu kullanılmıştır.

$$\omega(u, v) = |N(u) \cap N(v)| \quad (3.1)$$

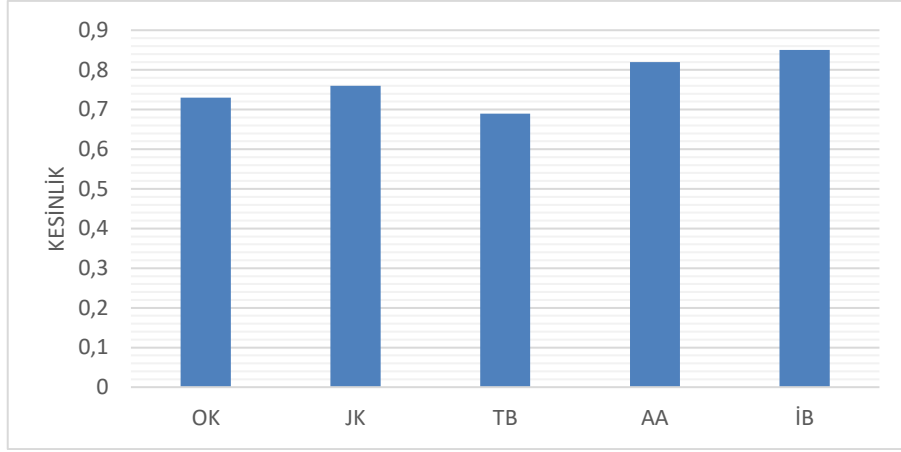
Ağın ağırlıklı yansıması oluşturulduktan sonra iki parçalı ağ üzerinde aralarında bağlantı olmayan düğüm çiftleri belirlenmiştir. Daha sonra bu düğümler arasında oluşturulan bir bağlantının ağın yansımasında bir değişikliğe neden olup olmadığı kontrol edilmiştir. Yansıma ağda herhangi bir değişikliğe neden olmayan düğüm çiftleri arasındaki bağlantılar iç bağlantı adayları olarak belirlenmiştir. Bulunan bağlantılar doğrudan iç bağlantılar olarak kabul edilmemiştir. Bir ağırlık eşik değeri tanımlanmıştır. Yansıma ağ üzerinde bu eşik değerine eşit veya eşik değerinden daha büyük olan bağlantılar iç bağlantılar olarak dikkate alınmıştır. Eşik değerinin daha doğru sonuçlar almak, önemli bağlantıları kaybetmemek ve gereksiz bağlantıların dikkate alınmasını engellemek

amacıyla doğru seçilmesi gerekir. Çalışmada eşik değeri bu şartlar göz önüne alınarak en doğru sonuçları verecek şekilde belirlenmiştir. Eşik değeri belirlendikten sonra ağırlıkları bu eşik değerine eşit veya büyük olan bağlantılar iç bağlantılar olarak tanımlanmıştır.

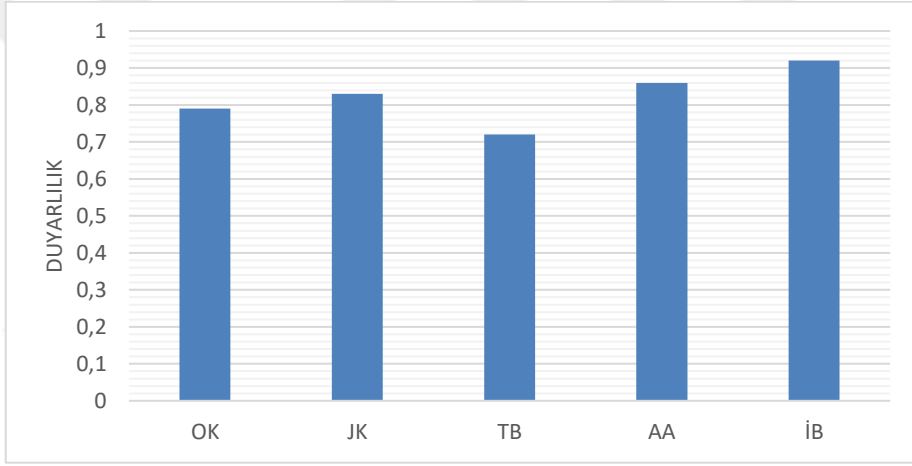
3.6. Sonuçlar

Önerilen iç bağlantılar yöntemi kurulan soru-hastalık ağı üzerinde değerlendirilmiştir. Kurulan ağda soru ve hastalık düğümleri arasında 1,804 bağlantı yer almaktadır. Önerilen yöntemin test edilmesi için sorular ve hastalıklar arasındaki 100 bağlantı çıkarılmıştır. Daha sonra önerilen metot 1,704 bağlantı ile kurulan ağa uygulanmıştır. İç bağlantı olabilecek bağlantılar bulunmuştur. Bu bağlantıların ağırlıkları eşik ağırlık değeri ile karşılaştırılmış ve bu eşik değerine eşit ya da bu değerden büyük olan bağlantılar iç bağlantılar olarak tanımlanmıştır. Bulunan iç bağlantılar ağırlıklarına göre büyükten küçüğe doğru sıralanmıştır. En yüksek ağırlıklı 100 bağlantı seçilmiştir. Seçilen bağlantılar ile ağdan başlangıçta çıkarılan bağlantılar karşılaştırılmıştır. Karşılaştırma sonucunda çıkarılan bağlantılardan kaçının doğru olarak tahmin edildiği bulunmuştur. Bulunan bu değerler performans değerlendirme ölçütlerinden kesinlik, duyarlılık ve F-ölçütünün hesaplanmasında kullanılmıştır.

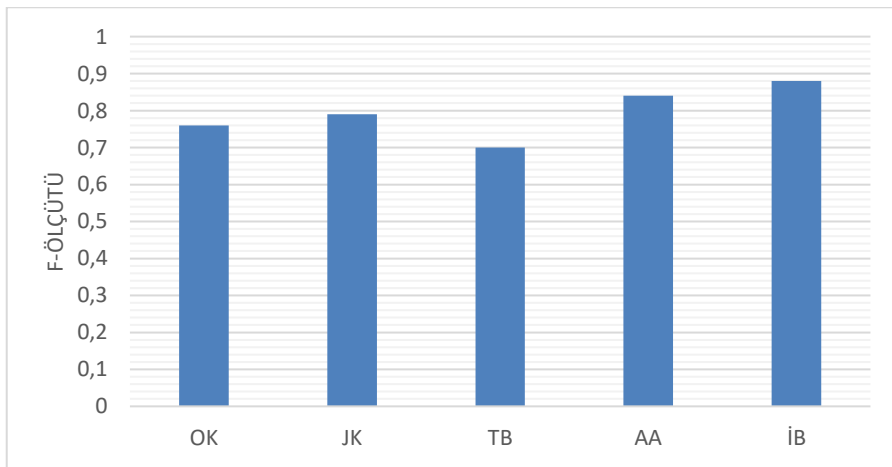
Önerilen yöntem ağa uygulandıktan ve sonuçlar elde edildikten sonra performans karşılaştırması için 1,704 bağlantı ile kurulan ağa benzerlik tabanlı bağlantı tahmini yöntemlerinden Ortak Komşular, Jaccard Katsayısı, Tercihli Bağlılık ve Adamic/Adar yöntemleri uygulanmıştır. Bu yöntemlerin uygulanma şekli detaylı olarak Bölüm 2’ de anlatılmıştır. Bu yöntemlerde iki düğüm arasındaki bir bağlantının önemini göstermek için bir $score(x,y)$ fonksiyonu tanımlanmıştır. Bu fonksiyon x ve y düğümleri arasında bir bağlantı oluşma ihtimalini hesaplamaktadır. Fonksiyon değeri ne kadar büyükse o düğüm çifti arasında gelecekte bir bağlantı oluşma ihtimali o kadar yüksektir. Fonksiyon değerleri Ortak Komşular, Jaccard Katsayısı, Tercihli Bağlılık ve Adamic/Adar yöntemleri için hesaplanmıştır. Tahmin edilen bağlantılar $score(x,y)$ fonksiyonundan elde edilen değerlerine göre büyükten küçüğe doğru sıralanmıştır. En yüksek skorlu 100 bağlantı seçilmiştir. Seçilen bağlantılar ile ağdan başlangıçta çıkarılan bağlantılar karşılaştırılmış ve bu karşılaştırma sonuçlarına göre her yöntem için ayrı ayrı kesinlik, duyarlılık ve F-ölçütü hesaplanmıştır.



Şekil 3.5. Beş farklı yöntemin kesinlik değerleri karşılaştırması



Şekil 3.6. Beş farklı yöntemin duyarlılık değerleri karşılaştırması



Şekil 3.7. Beş farklı yöntemin F-ölçütü değerleri karşılaştırması

Şekil 3.5, 3.6 ve 3.7 sırasıyla beş farklı metot için kesinlik, duyarlılık ve F-ölçütü değerlerinin karşılaştırmasını göstermektedir. Görüldüğü üzere önerilen iç bağlantılar yöntemi kesinlik, duyarlılık ve F-ölçütü bakımından diğer yöntemlerden daha iyi performans göstermiştir. Çıkarılan bağlantılar iç bağlantılar yöntemi ile diğer benzerlik tabanlı dört yöneme kıyasla daha yüksek oranda tahmin edilmiştir. Önerilen yöntemin çıkarılan bağlantıları tahmin etmedeki başarısı, aynı şekilde yüksek bir doğruluk oranı ile gelecekte oluşması muhtemel bağlantıları da tahmin edebileceğini göstermektedir.

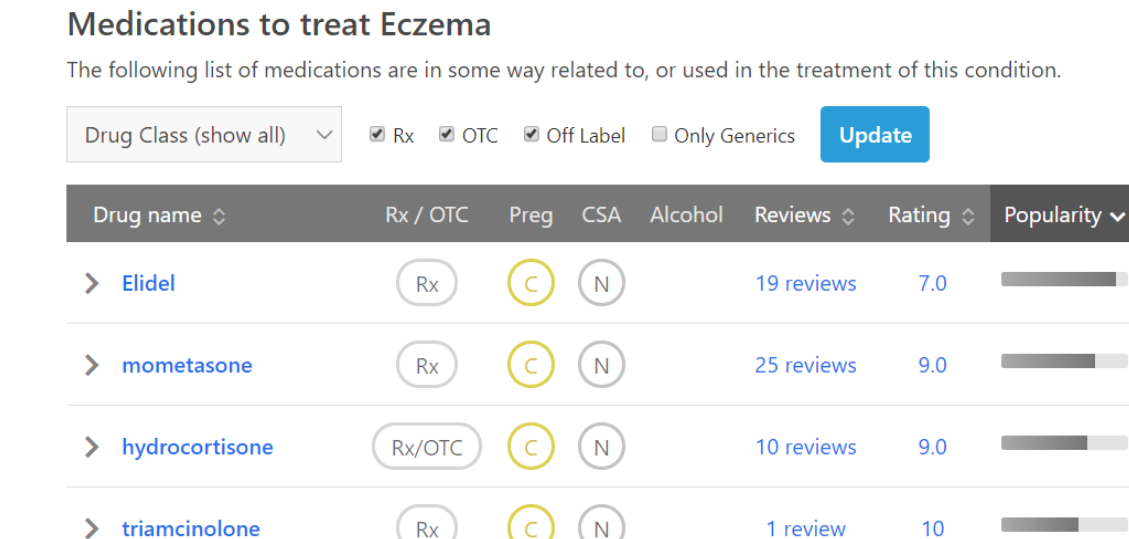


4. HASTALIK-İLAÇ İLİŞKİLERİNİN BULUNMASI

Çevrimiçi forum sitelerinde sorulan soruların analizi, bu sorularda yer alan belirtilerin hastalıklarla olan ilişkisi bulunduğundan sonra bu bölümde gerçekleştirilen uygulamada sorularda belirtilen şikayetler ile ilgili olduğu belirlenen hastalıkların ilaçlarla olan ilişkisi bulunmaya çalışılmıştır.

4.1. Hastalık-İlaç Verilerinin Elde Edilmesi

Bağlantı tahmini işleminde kullanılacak olan ağın kurulumu için gerekli olan hastalıklara ait ilaç verileri drugs.com [72] web sitesinden elde edilmiştir.



Medications to treat Eczema

The following list of medications are in some way related to, or used in the treatment of this condition.

Drug Class (show all) ▾ ☒ Rx ☒ OTC ☒ Off Label ☐ Only Generics [Update](#)

Drug name ▾	Rx / OTC	Preg	CSA	Alcohol	Reviews ▾	Rating ▾	Popularity ▾
> Elidel	Rx	C	N		19 reviews	7.0	
> mometasone	Rx	C	N		25 reviews	9.0	
> hydrocortisone	Rx/OTC	C	N		10 reviews	9.0	
> triamcinolone	Rx	C	N		1 review	10	

Şekil 4.1. drugs.com sitesinde yer alan ilaç bilgileri [72]

Şekil 4.1’ de hastalıklara ait ilaç bilgilerinin sitede [72] nasıl yer aldığı gösterilmiştir. Bu sitede her hastalığa yazılan ilaçlar yer almaktadır. Bu ilaçların sınıf, gebelik türü, oran, popülerlik gibi özellikleri tanımlanmıştır. Bu çalışmada hastalıklara yazılan ilaçlar oranlarına göre belirlenmiştir. En yüksek orana sahip olan ilaçlardan üç tanesi o hastalık için geçerli ilaçlar olarak belirlenmiştir. Oranı düşük olan ilaçlar dikkate alınmamıştır.

4.2. Hastalık-İlaç Ağının Kurulumu

Hastalıklara en yüksek oranda yazılan üç ilaç belirlendikten sonra bu hastalık ve ilaç verilerinden oluşan bir iki parçalı ağ modellenmiştir. Bu ağda hastalıklar bir düğüm kümesini, ilaçlar ise diğer düğüm kümesini oluşturmaktadır. Her hastalık ile o hastalığa yazılan ilaçlar arasında bağlantılar kurulmuştur.

Tablo 4.1. Hastalık-ilaç verilerinden bir kesit

Hastalık	İlaç	Oran
Asthma	Flunisolide	10
	Levalbuterol	10
	Xopenex	9
Bronchiectasis	Doryx	9
	Doxycycline	7
	Ventolin HFA	8
Gastritis/Duodenitis	omeprazole	6
	pantoprazole	5
Sinusitis	Augmentin XR	10
	Deconamine SR	8
	Cedax	9
Migraine	Imitrex	8
	Sumatriptan	8
	Amerge	9
Acne	Aldactone	9
	Minocycline	6
	Tretinoin	7

Tablo 4.1’ de hastalık-ilaç iki parçalı ağının kurulumunda kullanılan verilerden bir kesit yer almaktadır. Görüldüğü üzere her hastalığın ilişkili olduğu ilaçlar ve oranlar bulunmaktadır. En yüksek orana sahip üç ilaç her hastalık için seçilmiştir.

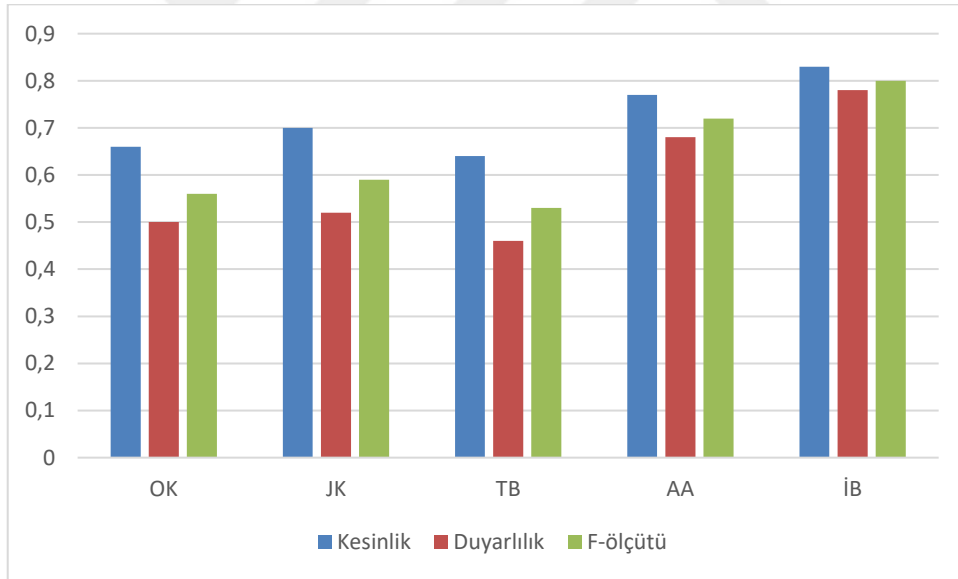
4.3. Hastalık-İlaç Ağında Bağlantı Tahmini

Kurulan iki parçalı hastalık-ilaç ağında bağlantı tahmini yapmak için iç bağlantılar yöntemi ve benzerlik tabanlı bağlantı tahmini yöntemlerinden Ortak Komşular, Jaccard Katsayısı, Tercihli Bağlılık ve Adamic/Adar yöntemleri kullanılmıştır.

Hastalık-ilaç ağı yaklaşık 500 bağlantı içermektedir. Ağda bağlantı tahmini için ilk olarak 500 bağlantıdan 100 tanesi rastgele olarak seçilmiş ve bu bağlantılar ağdan çıkarılmıştır. Daha sonra bağlantı tahmini yöntemlerinin uygulanabilmesi için iki parçalı

hastalık-ilaç ağı tek mod ağı dönüştürülmüştür. Artık elde edilen yansıma ağlar tek tip düğümlerden oluşmaktadır ve bağlantı tahmininin gerçekleştirilmesi için uygundur. Tek mod ağı ilk olarak iç bağlantılar yöntemi Bölüm 3.5’ te anlatıldığı gibi uygulanmıştır. Bu ağda elde edilen iç bağlantılardan en yüksek ağırlığa sahip 100 bağlantı seçilmiştir. Seçilen bağlantılar ile çıkarılan bağlantılar karşılaştırılarak çıkarılan bağlantılardan kaçının uygulanan yöntem ile doğru tahmin edildiği bulunmuştur. Elde edilen değerler ile iç bağlantılar yöntemine ait performans ölçütleri hesaplanmıştır.

İç bağlantılar yöntemi tek mod ağı uygulandıktan sonra benzerlik tabanlı yöntemler uygulanmıştır. Bu yöntemlerden elde edilen düğüm çifti skorları büyükten küçüğe sıralanmış ve en yüksek skora sahip 100 bağlantı her yöntem için seçilmiştir. Seçilen düğüm çiftleri arasındaki bağlantılar ile çıkarılan bağlantılar karşılaştırılmıştır. Karşılaştırma sonucunda elde edilen değerler ile performans değerlendirme ölçütleri hesaplanmıştır.



Şekil 4.2. Beş farklı yöntemin performans karşılaştırması

Beş farklı yöntem için bulunan kesinlik, duyarlılık ve F-ölçütü değerleri Şekil 4.2’ de verilmiştir. Görüldüğü üzere uygulanan yöntemler içerisinde iç bağlantılar yöntemi en iyi tahmin sonuçlarına ulaşmıştır. Gerçekleştirilen bu yöntem ile hastalık-ilaç iki parçalı ağında bağlantı tahmini için en iyi yöntemin iç bağlantılar olduğu görülmüştür.

4.4. Sonular

Bu blmde gerekleřtirilen uygulamada bir hastalıęa yaygın olarak yazılan ilalar modellenen iki paralı aęda baęlantı tahmini ile bulunmuřtur. Blm 3' te ise evrimii forum sitelerinden elde edilen hastalık belirtilerine gre bir baęlantı tahmini yapılıp kullanıcının řikayetleriyle ilgili olabilecek hastalıklar belirlenmiřtir. Son ařamada iki blmde gerekleřtirilen uygulamalardan elde edilen veriler birleřtirilmiřtir. İki blmde ayrı ayrı kurulan aęlarda gerekleřtirilen baęlantı tahmininden elde edilen veriler ile belirti-hastalık-ila řeklinde bir iliřki bulunmuřtur. İlk olarak forum sitelerinden kullanıcı sorularına karřılık hastalıklar elde edilip daha sonra bu hastalıkların iliřkili olduęu ilalar belirlenmiřtir. rneęin; nefes darlıęı, gęs sıkıřması, ksrk gibi hastalık belirtilerine sahip bir insanın hangi hastalıklara sahip olabileceęi, bu hastalıklar iin hangi ilaların nerildięi kurulan aęlarda baęlantı tahmini ile bulunmuřtur.

5. ÇEVİRİMİÇİ FORUM SİTELERİNİN GÜVENİLİRLİK ANALİZİ

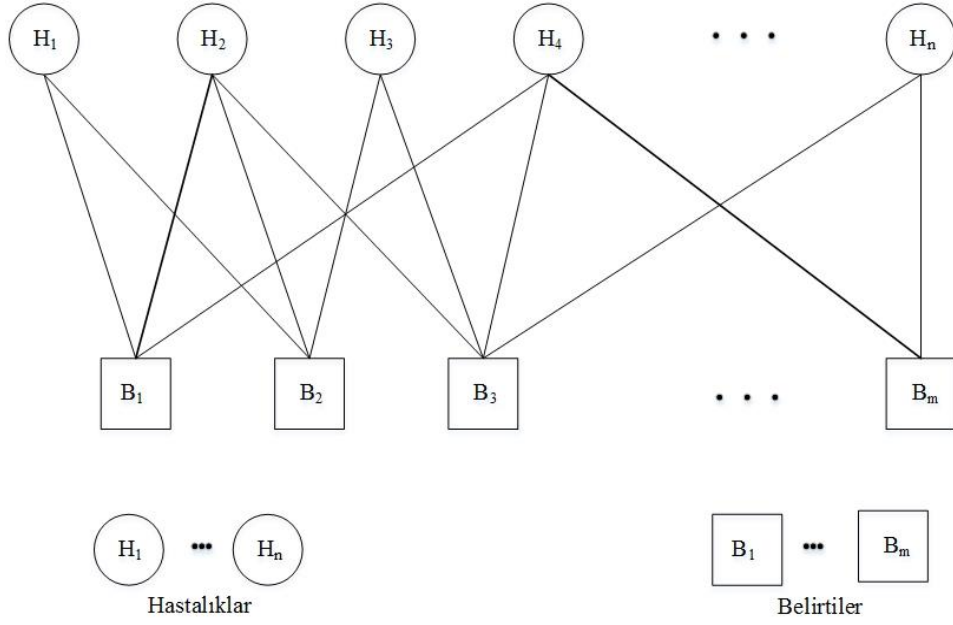
Günümüzde forum sitelerinin sağlık alanındaki yaygın kullanımından Bölüm 3’te bahsedilmiştir. İnsanların çevrimiçi forum sitelerine bu kadar sık başvurmaları bu sitelerdeki uzman doktorlardan şikayetlerine karşılık aldıkları hastalık bilgilerinin ne derece doğru olduğu sorusunu da akla getirmektedir. Bu çalışmada ingilizce forum sitelerindeki doktorların, kullanıcıların sorularına karşılık olarak verdiği yanıtların doğruluğu test edilmeye çalışılmıştır. Yapılan analiz ile forum sitelerinin güvenilirliği ölçülmüştür.

Çevrimiçi forum sitelerinde kullanıcılar sahip olduğu hastalık belirtilerini yazarak sahip olabilecekleri hastalıklar hakkında sorular sorarlar. Siteye kayıtlı uzman doktorlar, bu hastalık belirtilerinin ilişkili olabileceği hastalıkları sorulara cevap olarak yazarlar. Bu cevapların analizi için gerçekleştirilen uygulamamın aşamaları aşağıda verilmiştir.

5.1. Hastalık-Belirti Ağının Kurulması

Çevrimiçi forum sitelerindeki doktor cevaplarının güvenilirliğinin test edilmesi için kullanıcıların sorularında ifade ettikleri belirtilerin hangi hastalıklarla ilgili olduğunun gerçek bilgisine ihtiyaç duyulmaktadır. Bunun için Bölüm 3’te anlatıldığı gibi www.drugs.com [72] web sitesinden elde edilen verilerle oluşturulmuş, hastalıklara ait belirtileri içeren hastalık-belirti veri seti kullanılmıştır. Bu veri setinde her hastalığa ait başlıca belirtiler yer almaktadır.

Hastalık-belirti veri seti kullanılarak bir iki parçalı ağ modellenmiştir. Kullanılan veri seti içerisinde yer alan hastalık ve belirtiler ICD-10 kodları ile saklanmıştır. Bu yüzden iki parçalı ağda yer alan düğüm kümelerinden birini belirtilere ait ICD-10 kodları, diğer düğüm kümesini ise hastalıklara ait ICD-10 kodları oluşturmaktadır. Ağ üzerinde yer alan hastalık ve belirti ICD-10 kodlarını gösteren düğümler arasındaki ilişkiler, her hastalığın göstergesi olan belirti bilgisinden yola çıkılarak oluşturulmuştur. Bu bilgi hastalık-belirti veri seti kullanılarak elde edilmiştir.



Şekil 5.1. Hastalık-belirti ilişki ağı

Şekil 5.1 oluşturulan hastalık-belirti iki parçalı ağından örnek bir kesit göstermektedir. Modellenen ağ üzerinde tanımlı ilişkiler, kullanıcıların sorularında ifade ettikleri belirtilerin gerçekte hangi hastalıkların habercisi olduğu bilgisine ulaşmada kullanılmıştır.

5.2. Çevrimiçi Forum Sitelerinde Soru ve Cevapların Analizi

Belirtilere karşılık gelen hastalıkların belirlenmesinden sonra ilk olarak çevrimiçi forum sitelerinde yer alan kullanıcıların soruları analiz edilmiştir. Analiz işlemi için tam metin arama ve veri analiz aracı olan Elasticsearch kullanılmıştır. Elasticsearch hakkında detaylı bilgi Bölüm 3’te verilmiştir. Kullanıcıların sormuş oldukları sorularda yer alan belirtiler bu analiz aracı kullanılarak Bölüm 3’te anlatıldığı gibi elde edilmiştir.

Çevrimiçi forum sitelerinde insanların sordukları sorular analiz edilip bu sorularda yer alan belirtiler belirlendikten sonra her soruya uzman doktor tarafından verilen cevaplarda aynı şekilde analiz edilmiştir. Doktorlar tarafından verilen, belirtilerin ilişkili olabileceği hastalıkların yer aldığı cevapları Elasticsearch ile analiz etmek için ilk olarak standart hastalık isimleri ve ICD kodları Elasticsearch’e indexlenmiştir. Daha sonra cevaplar analiz edilerek bu cevaplarda geçen hastalık isimleri yani kişilerin şikayetleri belirli benzerlik skorları ile belirlenmiştir. Bu skorlar içerisinde yüksek benzerlik skoruna sahip olan

hastalıklar doktorun kullanıcıların şikayetlerine karşılık cevaplarda vermiş olduğu ilişkili hastalıklar olarak kabul edilmiştir. Bu hastalık bilgileri ICD-10 kodları şeklinde saklanmıştır. Her soruya karşılık gelen cevap bu şekilde analiz edilerek doktorun, kullanıcının sahip olduğu belirtilerin ilişkili olduğunu düşündüğü hastalıklar bulunmuştur. Analiz işlemi sonunda her soruda geçen belirtilere ait ICD-10 kodları ve her soruya karşılık doktorlar tarafından yazılan cevaplarda yer alan hastalık isimlerine ait ICD-10 kodları elde edilmiştir.

5.3. Çevrimiçi Forum Sitelerinin Güvenilirliğinin Analizi

Doktorların vermiş oldukları cevaplardaki ilgili hastalıklar bulunduktan sonra bu cevapların doğruluğunu ölçmek için Bölüm 5.1’ de kurulmuş olan hastalık-belirti ağı üzerinde yer alan ilişkiler kullanılmıştır. Bu test işleminde ilk olarak daha önceden her soruda ifade edilen belirtiler ICD-10 kodlarının hastalık-belirti ağı üzerinde hangi hastalıklar ile bir bağlantıya sahip olduğu belirlenmiştir. Soruda ifade edilen belirtiler en çok hangi hastalıklar ile bağlantıya sahipse o hastalıklar kullanıcının sorusunda ifade ettiği belirtilerin ilişkili olduğu hastalıklar olarak kabul edilmiştir. Bu şekilde her soruda bulunan belirtilerin kurulan hastalık-belirti ağı kullanılarak gerçekte ilişkili olabileceği hastalıklar bulunmuştur.

Bu çalışmada yaklaşık 2000 soru ve bu sorulara karşılık verilen cevaplar analiz edilmiştir. Sorularda kullanıcıların yazmış oldukları belirtilerin, hem forum sitelerindeki uzman doktorlar tarafından hem de gerçekte kurulan ağ kullanılarak hangi hastalıkların habercisi olduğu bulunmuştur. Her soru için doktor tarafından ifade edilen hastalıklar ile gerçek ağ kullanılarak bulunan hastalıklar karşılaştırılmıştır. Karşılaştırma sonucunda doktorların verdiği cevapların gerçek ağdan elde edilen verilerle % 69 oranında benzer olduğu bulunmuştur.

Son zamanlarda insanların sağlıklarıyla ilgili sorunlarını ilk olarak internetten araştırma eğilimleri bu çalışmanın yapılma gerekçelerinden biri olmuştur. Doktora gidip birtakım işlemlerden geçerek şikayetlerinin hangi hastalıkların belirtisi olduğunu öğrenmeden önce internetin sunduğu bu kolay yolu kullanma artık çok yaygın bir hale gelmiştir. Birçok Türkçe ve İngilizce çevrimiçi forum sitesinin varlığı, bu sitelerde yer alan binlerce yorum, bu sitelere üye binlerce kişi ve alanında uzman doktorlar çevrimiçi forum sitelerine olan

ilginin göstergesidir. Bu çalışma ile insanların bu kadar ilgi gösterdiği sađlık konusunda bir danisma alaninin guvenilirliđinin analizi yapılmıřtır. Yapılan çalışma insanlar iin bir başvuru kaynađı olan ingilizce forum sitelerinin %69 oranında guvenilirliđe sahip olduđunu gostermiştir.



6. SONUÇLAR

Bu tez çalışmasında son yıllarda önemli bir araştırma konusu haline gelen sosyal ağ yapıları anlatılmıştır. Çevremizde birçok karmaşık sistemin sosyal ağlar ile modellenmesinden dolayı bu yapıların incelenip analiz edilmesi konusuna olan yönelimin artışından bahsedilmiştir. Sosyal ağların dinamik yapısını incelemek için başvurulmuş bağlantı tahminini uygulamalarının önemi vurgulanmıştır. İki parçalı yapıda modellenen ağlarda bağlantı tahmini konusunda yapılan çalışmalar incelenmiştir. Çalışmalarda kullanılan yöntemler ve çalışmaların genel yapısından bahsedilmiştir.

Günümüzde, iki parçalı ağlarda bağlantı tahmini birçok bilim dalı için odak noktası haline gelmiştir. Farklı alanlardan araştırmacıların konuya olan yönelimleri ve yaptıkları çalışmalar artmıştır. Bu nedenlerden dolayı bu tez çalışmasının genel amacını iki parçalı ağlarda bağlantı tahmini oluşturmaktadır. Bu amaç doğrultusunda iki parçalı ağların yapısı anlatılmıştır. Ağların değişimini incelemek için yararlı bir yol olan bağlantı tahmini konusu detaylı olarak incelenmiştir. Bu konunun matematiksel olarak ifadesi verilmiş ve hangi alanlarda hangi uygulamalarda kullanıldığı bilgilerine yer verilmiştir. Bağlantı tahmini konusunda şimdiye kadar önerilen yaklaşımlar incelenmiştir. Bu yaklaşımlarda kullanılan temel ölçütler detaylı olarak açıklanmıştır.

İki parçalı ağlar tek mod ağlardan farklı yapılardır. Günümüzde önerilen bağlantı tahmini yöntemlerinin çoğu ise tek mod ağlara yöneliktir. Bu bağlantı tahmini yöntemleri doğrudan farklı yapıda olan iki parçalı ağlara uygulanamaz. İki parçalı ağlarda bağlantı tahmini için birçok yöntem geliştirilmiştir. Çalışmada bu yöntemler, kullanılan parametreler, yöntemlerin başarısı detaylı olarak incelenmiştir.

İki parçalı ağlarda bağlantı tahmini konusuna sağlık alanında bir uygulama gerçekleştirilmiştir. Bunun için veri kaynağı olarak günümüzde internetin hızlı gelişimi ile insanlar için bir başvuru kaynağı olan çevrimiçi sağlık forumları kullanılmıştır. Bu forumlar son zamanlarda insanlar tarafından büyük ilgi görmektedir. İnsanlar birtakım işlemlerden geçmeden önce bu forumlardaki uzmanlardan sorunları hakkında bilgi almaya çalışırlar. Bu yönelim sonucunda forum siteleri binlerce soru ve cevap barındırmaktadır. İnsanların bu forumları bu kadar sıklıkla kullanması çalışma için bu forum sitelerinin veri kaynağı olarak seçilmesine neden olmuştur. Bu forumlarda kullanıcıların sormuş oldukları sorulardan elde edilen hastalık belirtilerine göre ilgili hastalıklar belirlenmiş ve bir soru-

hastalık iki parçalı ağ modellerlenmiştir. Bu ağ üzerinde bağlantı tahmini için, uygulandığı çalışmalarda iyi performans gösteren iç bağlantılar yöntemi önerilmiştir. Bu yöntem kurulan iki parçalı ağa uygulanmış ve elde edilen sonuçlar benzerlik tabanlı yöntemler ile karşılaştırılmıştır. Karşılaştırma sonuçları değerlendirilmiştir.

Hastalık belirtilerinin ilişkili olabileceği hastalıklar bulunduktan sonra bu hastalıklar ile ilişkili ilaçların belirlenmesi için bir uygulama gerçekleştirilmiştir. Hastalıklar ve yüksek oranda o hastalıklara yazılan ilaçlardan oluşan bir iki parçalı ağ modellerlenmiş ve bu ağ üzerinde bağlantı tahmini gerçekleştirilmiştir. Sonuç olarak forum sitelerinde kullanıcıların yazdıkları bir soruda ifade ettikleri hastalık belirtilerinin hangi hastalıklarla ilişkili olabileceği ve bu hastalıklara en çok yazılan ilaçların neler olduğu bilgisi elde edilmiştir.

Son olarak insanlar tarafından bu kadar yaygın bir şekilde kullanılan çevrimiçi sağlık forumlarının verdikleri bilgiler açısından ne kadar güvenilir olduklarını incelemek amacıyla bir çalışma yapılmıştır. Bu çalışmada uzman doktorlardan alınan cevaplar ile gerçek bilgiler karşılaştırılmış ve alınan cevapların doğruluğu ölçülmüştür.

Bu tez çalışmasında hem günümüzde popüler bir konu haline gelen ve birçok bilim dalı için uygulamaya sahip olan sosyal ağlarda bağlantı tahmini konusu sağlık alanında bir uygulama ile gerçekleştirilmeye çalışılmış hem de son zamanlarda insanların başvuru kaynağı olarak gördükleri forum siteleri analiz edilmiştir. Bu tez çalışması sağlık alanında gerçekleştirilecek olan bağlantı tahmini uygulamaları için bir kaynak olabilir.

KAYNAKLAR

- [1] **Liben-Nowell, D., and Kleinberg, J.,** 2007. The link-prediction problem for social networks. *journal of the Association for Information Science and Technology*, 58(7), 1019-1031.
- [2] **Sun, Y., Barber, R., Gupta, M., Aggarwal, C. C., and Han, J.,** 2011. Co-author relationship prediction in heterogeneous bibliographic networks, In *Advances in Social Networks Analysis and Mining (ASONAM)*, 2011 International Conference on (pp. 121-128). IEEE.
- [3] **Wang, D., Pedreschi, D., Song, C., Giannotti, F., and Barabasi, A. L.,** 2011. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1100-1108). ACM.
- [4] **Han, X., Wang, L., Han, S. N., Chen, C., Crespi, N., and Farahbakhsh, R.,** 2015. Link prediction for new users in social networks. In *Communications (ICC), 2015 IEEE International Conference on* (pp. 1250-1255). IEEE.
- [5] **Benchettara, N., Kanawati, R., and Rouveirol, C.,** 2010. A supervised machine learning link prediction approach for academic collaboration recommendation. In *Proceedings of the fourth ACM conference on Recommender systems* (pp. 253-256). ACM.
- [6] **Anil, A., Kumar, D., Sharma, S., Singha, R., Sarmah, R., Bhattacharya, N., and Singh, S. R.,** 2015. Link Prediction Using Social Network Analysis over Heterogeneous Terrorist Network. In *Smart City/SocialCom/SustainCom (SmartCity)*, 2015 IEEE International Conference on (pp. 267-272). IEEE.
- [7] **Wiil, U. K., Gniadek, J., and Memon, N.,** 2011. Retracted: A Novel Method to Analyze the Importance of Links in Terrorist Networks, In *Counterterrorism and Open Source Intelligence* (pp. 171-187). Springer Vienna.
- [8] **Guo, S., Wang, M., and Leskovec, J.,** 2011. The role of social networks in online shopping: information passing, price of trust, and consumer choice. In *Proceedings of the 12th ACM conference on Electronic commerce* (pp. 157-166). ACM.
- [9] **Fortunato, S.,** 2010. Community detection in graphs. *Physics reports*, 486(3), 75-174.
- [10] **Gregory, S.,** 2009. Finding overlapping communities using disjoint community detection algorithms. *Complex networks*, 47-61.
- [11] **Fortunato, S., and Castellano, C.,** 2007. Community structure in graphs. *arXiv preprint arXiv:0712.2716*.

- [12] **Benchettara, N., Kanawati, R., and Rouveirol, C.,** 2010. Supervised machine learning applied to link prediction in bipartite social networks. In *Advances in Social Networks Analysis and Mining (ASONAM)*, 2010 International Conference on (pp. 326-330). IEEE.
- [13] **Xia, S., Dai, B., Lim, E. P., Zhang, Y., and Xing, C.,** 2012. Link prediction for bipartite social networks: the Role of Structural Holes. In *Advances in Social Networks Analysis and Mining (ASONAM)*, 2012 IEEE/ACM International Conference on (pp. 153-157). IEEE.
- [14] **Nigam, A., and Chawla, N. V.,** 2016. Link prediction in a semi-bipartite network for recommendation. In *Asian Conference on Intelligent Information and Database Systems* (pp. 127-135). Springer, Berlin, Heidelberg.
- [15] **Allali, O., Magnien, C., and Latapy, M.,** 2011. Link prediction in bipartite graphs using internal links and weighted projection. In *Computer Communications Workshops (INFOCOM WKSHPS)*, 2011 IEEE Conference on (pp. 936-941). IEEE.
- [16] **Chang, Y. J., and Kao, H. Y.,** 2012. Link prediction in a bipartite network using Wikipedia revision information. In *Technologies and Applications of Artificial Intelligence (TAAI)*, 2012 Conference on (pp. 50-55). IEEE.
- [17] **Gao, M., Chen, L., Li, B., Li, Y., Liu, W., and Xu, Y. C.,** 2017. Projection-based link prediction in a bipartite network. *Information Sciences*, 376, 158-171.
- [18] **Li, X., and Chen, H.,** 2013. Recommendation as link prediction in bipartite graphs: A graph kernel-based machine learning approach. *Decision Support Systems*, 54(2), 880-890.
- [19] **Li, Y., Wen, A., Lin, Q., Li, R., and Lu, Z.,** 2014. Name disambiguation in scientific cooperation network by exploiting user feedback. *Artificial Intelligence Review*, 1-16.
- [20] **Savic, M., Ivanovic, M., Radovanovic, M., Ognjanovic, Z., Pejovic, A., and Krueger, T. J.,** 2014. The structure and evolution of scientific collaboration in Serbian mathematical journals, *Scientometrics*, 101(3), 1805-1830.
- [21] **Ge, M., Li, A., and Wang, M.,** 2016. A bipartite network-based method for prediction of long non-coding RNA–protein interactions. *Genomics, proteomics & bioinformatics*, 14(1), 62-71.
- [22] **Ferreri, L., Venturino, E., and Giacobini, M.,** 2011. Do diseases spreading on bipartite networks have some evolutionary advantage? In *European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics* (pp. 141-146). Springer, Berlin, Heidelberg.

- [23] **Kaya, B., and Poyraz, M.,** 2016. Unsupervised link prediction in evolving abnormal medical parameter networks, *International Journal of Machine Learning and Cybernetics*, 1(7), 145-155.
- [24] **Kaya, B., and Poyraz, M.,** 2014. Supervised link prediction in symptom networks with evolving case. *Measurement*, 56, 231-238.
- [25] **Zhang, Y., Tao, C., Jiang, G., Nair, A. A., Su, J., Chute, C. G., and Liu, H.,** 2014. Network-based analysis reveals distinct association patterns in a semantic MEDLINE-based drug-disease-gene network. *Journal of biomedical semantics*, 5(1), 33.
- [26] **Shahinpour, S., and Butenko, S.,** 2013. Distance-based clique relaxations in networks: s-clique and s-club. In *Models, algorithms, and technologies for network analysis* (pp. 149-174). Springer, New York, NY.
- [27] **Wang, X., Liu, Y., and Xiong, F.,** 2016. Improved personalized recommendation based on a similarity network. *Physica A: Statistical Mechanics and its Applications*, 456, 271-280.
- [28] **Cui, Y., Zhang, L., Wang, Q., Chen, P., and Xie, C.,** 2016. Heterogeneous Network Linkage-weight Based Link Prediction in Bipartite Graph for Personalized Recommendation. *Procedia Computer Science*, 91, 953-958.
- [29] **Zhao, D., Zhang, L., and Zhao, W.,** 2016. Genre-based link prediction in bipartite graph for music recommendation, *Procedia Computer Science*, 91, 959-965.
- [30] **Xu, Y., Chen, L., and Zou, S.,** 2013. Community detection from bipartite networks. In *Web Information System and Application Conference (WISA)*, 2013 10th (pp. 249-254). IEEE.
- [31] **Luo, Y., Liu, Q., Wu, W., Li, F., and Bo, X.,** 2014. Predicting drug side effects based on link prediction in bipartite network. In *Biomedical Engineering and Informatics (BMEI)*, 2014 7th International Conference on (pp. 729-733). IEEE.
- [32] **Liben-Nowell, D., and Kleinberg, J.,** 2007. The link-prediction problem for social networks. *journal of the Association for Information Science and Technology*, 58(7), 1019-1031.
- [33] **Aiello, L. M., Barrat, A., Schifanella, R., Cattuto, C., Markines, B., and Menczer, F.,** 2012. Friendship prediction and homophily in social media. *ACM Transactions on the Web (TWEB)*, 6(2), 9.
- [34] **Mori, J., Kajikawa, Y., Kashima, H., and Sakata, I.,** 2012. Machine learning approach for finding business partners and building reciprocal relationships. *Expert Systems with Applications*, 39(12), 10402-10407.

- [35] **Wu, S., Sun, J., and Tang, J.,** 2013. Patent partner recommendation in enterprise social networks. In Proceedings of the sixth ACM international conference on Web search and data mining (pp. 43-52). ACM.
- [36] **Akcora, C. G., Carminati, B., and Ferrari, E.,** 2011. Network and profile based measures for user similarities on social networks. In Information Reuse and Integration (IRI), 2011 IEEE International Conference on (pp. 292-298). IEEE.
- [37] **Tang, J., Wu, S., Sun, J., and Su, H.,** 2012. Cross-domain collaboration recommendation. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1285-1293). ACM.
- [38] **Pavlov, M., and Ichise, R.,** 2007. Finding experts by link prediction in co-authorship networks. In Proceedings of the 2nd International Conference on Finding Experts on the Web with Semantics-Volume 290 (pp. 42-55). CEUR-WS.org.
- [39] **Wohlfarth, T., and Ichise, R.,** 2008. Semantic and Event-Based Approach for Link Prediction. In PAKM (pp. 50-61).
- [40] **Raeder, T., Lizardo, O., Hachen, D., and Chawla, N. V.,** 2011. Predictors of short-term decay of cell phone contacts in a large scale communication network. *Social Networks*, 33(4), 245-257.
- [41] **Marchette, D. J., and Priebe, C. E.,** 2008. Predicting unobserved links in incompletely observed networks. *Computational Statistics and Data Analysis*, 52(3), 1373-1386.
- [42] **Kim, M., and Leskovec, J.,** 2011. The network completion problem: Inferring missing nodes and edges in networks. In Proceedings of the 2011 SIAM International Conference on Data Mining (pp. 47-58). Society for Industrial and Applied Mathematics.
- [43] **Barabási, A. L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., and Vicsek, T.,** 2002. Evolution of the social network of scientific collaborations. *Physica A: Statistical mechanics and its applications*, 311(3), 590-614.
- [44] **Juszczyzyn, K., Musial, K., and Budka, M.,** 2011. Link prediction based on subgraph evolution in dynamic social networks. In Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on (pp. 27-34). IEEE.
- [45] **Bringmann, B., Berlingerio, M., Bonchi, F., and Gionis, A.,** 2010. Learning and predicting the evolution of social networks. *IEEE Intelligent Systems*, 25(4), 26-35.

- [46] **Raymond, R., and Kashima, H.,** 2010. Fast and scalable algorithms for semi-supervised link prediction on static and dynamic graphs. *Machine Learning and Knowledge Discovery in Databases*, 131-147.
- [47] **Zhu, J., Hong, J., and Hughes, J. G.,** 2002. Using Markov models for web site link prediction. In *Proceedings of the thirteenth ACM conference on Hypertext and hypermedia* (pp. 169-170). ACM.
- [48] **Almansoori, W., Gao, S., Jarada, T. N., Elsheikh, A. M., Murshed, A. N., Jida, J., and Rokne, J.,** 2012. Link prediction and classification in social networks and its application in healthcare and systems biology. *Network modeling and analysis in health informatics and bioinformatics*, 1(1), 27-36.
- [49] **Huang, Z., and Lin, D. K.,** 2009. The time-series link prediction problem with applications in communication surveillance. *INFORMS Journal on Computing*, 21(2), 286-303.
- [50] **Lü, L., and Zhou, T.,** 2011. Link prediction in complex networks: A survey. *Physica A: statistical mechanics and its applications*, 390(6), 1150-1170.
- [51] **Yang, L., Jiang, H., and Yang, J.,** 2010. The Cyclic Temporal Network Density and Impact on Information Diffusion for Delay Tolerant Network. In *Wireless Communications Networking and Mobile Computing (WiCOM), 2010 6th International Conference on* (pp. 1-5). IEEE.
- [52] **Liu, Z., Ma, J., and Zeng, Y.,** 2013. Secrecy transfer for sensor networks: from random graphs to secure random geometric graphs. *International Journal on Smart Sensing & Intelligent Systems*, 6(1).
- [53] **Jia, X., Xin, F., and Chuan, W. R.,** 2013. Adaptive spray routing for opportunistic networks. *International Journal on Smart Sensing & Intelligent Systems*, 6(1).
- [54] **Sid Ahmed, H., Mohamed Faouzi, B., and Caelen, J.,** 2013. Detection and classification of the behavior of people in an intelligent building by camera. *International Journal on Smart Sensing & Intelligent Systems*, 6(4).
- [55] **Wang, P., Xu, B., Wu, Y., and Zhou, X.,** 2014. Link prediction in social networks: the state-of-the-art. *arXiv preprint arXiv:1411.5118*.
- [56] **Al Hasan, M., Chaoji, V., Salem, S., and Zaki, M.,** 2006. Link prediction using supervised learning. In *SDM06: workshop on link analysis, counter-terrorism and security*.
- [57] **Kossinets, G.,** 2006. Effects of missing data in social networks. *Social networks*, 28(3), 247-268.

- [58] **Jaccard, P.**, 1901. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull Soc Vaudoise Sci Nat*, 37, 547-579.
- [59] **Adamic, L. A., and Adar, E.**, 2003. Friends and neighbors on the web. *Social networks*, 25(3), 211-230.
- [60] **Liben-Nowell, D., and Kleinberg, J.**, 2007. The link-prediction problem for social networks. *journal of the Association for Information Science and Technology*, 58(7), 1019-1031.
- [61] **Barabási, A. L., and Albert, R.**, 1999. Emergence of scaling in random networks. *science*, 286(5439), 509-512.
- [62] **Zhou, T., Lü, L., and Zhang, Y. C.**, 2009. Predicting missing links via local information. *The European Physical Journal B-Condensed Matter and Complex Systems*, 71(4), 623-630.
- [63] **Rodriguez, A., Kim, B., Turkoz, M., Lee, J. M., Coh, B. Y., and Jeong, M. K.**, 2015. New multi-stage similarity measure for calculation of pairwise patent similarity in a patent citation network. *Scientometrics*, 103(2), 565-581.
- [64] **Sørensen, T.**, 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Biol. Skr.*, 5, 1-34.
- [65] **Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabási, A. L.**, 2002. Hierarchical organization of modularity in metabolic networks. *science*, 297(5586), 1551-1555.
- [66] **Leicht, E. A., Holme, P., and Newman, M. E.**, 2006. Vertex similarity in networks. *Physical Review E*, 73(2), 026120.
- [67] **Malviya, V., and Gupta, G. P.**, 2015. Performance evaluation of similarity-based link prediction schemes for social network. In *Next Generation Computing Technologies (NGCT), 2015 1st International Conference on* (pp. 654-659). IEEE.
- [68] **Lü, L., Jin, C. H., and Zhou, T.**, 2009. Similarity index based on local paths for link prediction of complex networks. *Physical Review E*, 80(4), 046122.
- [69] **Papadimitriou, A., Symeonidis, P., and Manolopoulos, Y.**, 2012. Fast and accurate link prediction in social networking systems. *Journal of Systems and Software*, 85(9), 2119-2132.
- [70] **Katz, L.**, 1953. A new status index derived from sociometric analysis. *Psychometrika*, 18(1), 39-43.

[71] **Chung, F., and Zhao, W.**, 2010. PageRank and random walks on graphs. Fete of combinatorics and computer science, 20, 43-62.

[72] www.drugs.com, 5 Haziran 2017.



ÖZGEÇMİŞ

1992 yılında Elazığ’ da doğan Esra GÜNDOĞAN, orta ve lise öğrenimini sırasıyla Nahit Ergene İlköğretim Okulu ve Mehmet Koloğlu Anadolu Lisesinde tamamlamıştır. 2010 yılında kazandığı Fırat Üniversitesi Mühendislik Fakültesi Bilgisayar Mühendisliği Bölümünü 2014 yılında başarıyla bitirmiştir. 2015 yılında yüksek lisans eğitimine Fırat Üniversitesi Fen Bilimleri Enstitüsü Kuramsal Temeller Anabilim Dalında başlamıştır. 2016 yılında Mersin Üniversitesi Bilgisayar Mühendisliği Bölümüne Araştırma Görevlisi olarak atandıktan sonra 35. Madde görevlendirmesi ile Fırat Üniversitesi Bilgisayar Mühendisliğinde çalışmaya başlamıştır. Halen Bilgisayar Mühendisliği Bölümünde Araştırma Görevlisi olarak görevine devam etmektedir.