

T.C.
AYDIN ADNAN MENDERES ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
YÖNETİM BİLİŞİM SİSTEMLERİ ANABİLİM DALI
YÜKSEK LİSANS PROGRAMI
2022-YL-153

SIK ALT ÇİZGE MADENCİLİĞİ ALGORİTMALARININ
KULLANIM ALANLARI VE UYGULANABİLİRLİĞİ

HAZIRLAYAN
Mehmet Serdar GÜR

TEZ DANIŞMANI
Prof. Dr. Mustafa ÇETİN

AYDIN – 2023

T.C.
AYDIN ADNAN MENDERES ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ MÜDÜRLÜĞÜNE AYDIN

Bu tezde sunulan tüm bilgi ve sonuçların, bilimsel yöntemlerle yürütülen gerçek deney ve gözlemler çerçevesinde tarafımdan elde edildiğini, çalışmada bana ait olmayan tüm veri, düşünce, sonuç ve bilgilere bilimsel etik kuralların gereği olarak eksiksiz şekilde uygun atıf yaptığımı ve kaynak göstererek belirttiğimi beyan ederim.

... / ... / 2023

Mehmet Serdar GÜR

ÖZET

SIK ALT ÇİZGE MADENCİLİĞİ ALGORİTMALARININ KULLANIM ALANLARI VE UYGULANABİLİRLİĞİ

Mehmet Serdar GÜR

Yüksek Lisans Tezi, Yönetim Bilişim Sistemleri Anabilim Dalı

Tez Danışmanı: Prof. Dr. Mustafa ÇETİN

2023 XIV+53 sayfa

Çizge madenciliği, veri madenciliğinin bir alt dalıdır ve veri tabanını çizge olarak ifade eder. Çizgeler, veri madenciliğinin birçok alanında güçlü veri yapıları olarak kullanılır. Özellikle son yıllarda, çizge madenciliği konusuna olan ilgi giderek artmaktadır. Bunun bir nedeni, sosyal medyanın yükselişi ve tıbbi araştırmaların bilgisayar ortamına taşınmasıyla birlikte yapılandırılmış veri yapılarına olan ihtiyacın artması ve araştırmacıların önemli bilgilere olan talepleridir. Yapılandırılmış veri kümelerinde, tekrar eden altyapıların kaldırılmasıyla veriler yorumlanabilir ve genelleştirilebilir. Aynı zamanda, bir bilgi alanında çalışan bir kullanıcı, önceden tanımlanmış bir türde veya bu alana özgü altyapılarda arama yapar. Çizge veri tabanı sayesinde nesneler arasındaki ilişkiler ortaya çıkarılır. Çizge veri tabanı, düğümlerden ve kenarlardan oluşur, düğümler nesneleri temsil ederken, kenarlar nesneler arasındaki ilişkileri ifade eder. Çizge veri tabanı, birçok küçük çizgeden veya tek bir büyük çizgeden oluşabilir. Birçok küçük çizgeden oluşan veri tabanı işlemsel çizge veri tabanı olarak adlandırılırken, tek bir büyük çizgeden oluşan veri tabanı tek bir büyük çizge veri tabanı olarak adlandırılır. Çizge veri tabanları, kimyasal bileşenler, genetik yapılar, bilgisayar ağları, sosyal medya gibi farklı disiplinlerden gelen verilerin ifade edilmesine olanak sağlar.

Sık alt çizge madenciliği, büyük veri kümelerinde tekrar eden desenleri ve önemli ilişkileri keşfetmek amacıyla kullanılan bir veri madenciliği tekniğidir. Bu teknik, veri kümesinde sıklıkla tekrarlanan alt çizgeleri bulmak için özel algoritmalar ve yöntemler kullanır. Sık alt çizgeler, veri kümesinde belirli bir sıklık eşiğini geçen ve anlamlı desenler taşıyan çizgelerdir. Sık alt çizge madenciliği, farklı alanlarda çeşitli uygulama alanlarına sahiptir. Örneğin, veri tabanı yönetimi, sosyal ağ analizi, biyoinformatik ve çizge analizi gibi alanlarda sık alt çizge madenciliği önemli bir rol oynar. Bu teknik, veri kümesinde tekrar eden desenleri belirleyerek işletmelerin daha iyi kararlar almasını sağlar, pazarlama stratejilerini

optimize eder, ağ yapısını analiz eder ve biyolojik verilerde önemli bilgileri ortaya çıkarır. Sık alt çizge madenciliği, terabaytlarca bilgi üzerinde analiz yapabilir ve verinin ilişkilerini, değiştiğinde ve geliştiğinde bile koruyabilir. Bu nedenle, çizge veri tabanları sık alt çizge madenciliği için etkili bir araç sağlar.

Bu çalışmanın amacı, sık alt çizge madenciliğinin nasıl ortaya çıkarılabileceği, sık alt çizgelerin nasıl keşfedilebileceği ve bu çizgelerin kurumlar ve araştırmacılar için nasıl değerli hale getirilebileceğini araştırmaktır.

Çalışma üç aşamadan oluşmaktadır. İlk aşamada, gerçek dünya verilerinin çizge veri tabanına dönüştürülmesi gerçekleştirilmiştir. İkinci aşamada, oluşturulan çizge veri tabanından sık alt çizgelerin keşfedilmesi sağlanmıştır. Üçüncü aşamada ise, bu sık alt çizgelerin bilgisel olarak önem derecelerinin kanıtlanması gerçekleştirilmiştir.

Ayrıca, literatürde yer alan işlemsel çizge veri tabanı için önceden geliştirilmiş güncel bir sık alt çizge algoritması olan GSPAN algoritması, örnek bir vaka üzerinde uygulanır. Örnek vaka olarak, IMDb internet sitesinden 2000-2022 yılları arasında vizyona giren Türk filmleri ve bu filmlerin özellikleri kullanılır. IMDb veri tabanı, işlemsel veri tabanı şeklinde çizge veri tabanına dönüştürülür. Oluşturulan işlemsel veri tabanı üzerinde GSPAN (referans) algoritması çalıştırılmış ve sık alt çizgeler ortaya çıkarılmıştır.

ANAHTAR KELİMELER: Çizge Madencilik, Sık Alt Çizge Madenciliği, Veri Madenciliği.

ABSTRACT

APPLICATION AREAS AND USAGE OF FREQUENT SUBGRAPH MINING ALGORITHMS

Mehmet Serdar GÜR

MSc Thesis at Management Information System

Supervisor: Prof. Dr. Mustafa ÇETİN

2023 XIV+53 pages

Graph mining is a subfield of data mining that represents the database as a graph. Graphs are powerful data structures used in various domains of data mining. In recent years, there has been a growing interest in graph mining. This can be attributed to the rise of social media, the digitization of medical research, and the increasing need for structured data in different disciplines. In structured data sets, the removal of repetitive infrastructure allows for interpretation and generalization of the data. Furthermore, users in a specific knowledge domain search for predefined types or domain-specific infrastructures. Graph databases reveal relationships between objects, where nodes represent objects and edges represent relationships between them. A graph database can consist of many small graphs or a single large graph. A database consisting of many small graphs is referred to as an operational graph database, while a database consisting of a single large graph is referred to as a single large graph database. Graph databases facilitate the representation of data from various disciplines such as chemical compounds, genetic structures, computer networks, and social media.

Frequent subgraph mining is a data mining technique used to discover repeated patterns and significant relationships in large data sets. This technique employs specialized algorithms and methods to identify frequently occurring subgraphs in the data set. Frequent subgraphs are graphs that surpass a certain frequency threshold in the data set and carry specific patterns of meaning. Frequent subgraph mining has diverse application areas. For instance, it plays a crucial role in database management, social network analysis, bioinformatics, and graph analysis. This technique enables businesses to make better decisions by identifying recurring patterns in the data, optimizing marketing strategies, analyzing network structures, and revealing important information in biological data. Frequent subgraph mining can analyze terabytes of information and preserve data

relationships even when they change and evolve. Therefore, graph databases provide an effective tool for frequent subgraph mining.

The objective of this study is to explore how frequent subgraph mining can be applied, how frequent subgraphs can be discovered, and how these subgraphs can be valuable for organizations and researchers. The study consists of three stages. In the first stage, real-world data is transformed into a graph database. In the second stage, frequent subgraphs are discovered from the generated graph database. In the third stage, the informational significance of these frequent subgraphs is validated.

Furthermore, a state-of-the-art frequent subgraph mining algorithm, GSPAN, specifically designed for operational graph databases, is applied to a sample case. The IMDb website is used as the sample case, specifically focusing on Turkish films released between 2000 and 2022 and their properties. The IMDb database is transformed into an operational graph database. The GSPAN algorithm (reference) is then executed on the operational graph database, resulting in the discovery of frequent subgraphs.

KEY WORDS: Data mining, Frequent subgraph mining, Graph mining.

TEŞEKKÜR

Bu araştırmamın yapılmasında, öncelikle yardımlarını esirgemeyen, bana her aşamada yol gösteren ve danışmanlığımı kabul eden değerli danışmanım Prof. Dr. Sayın Mustafa ÇETİN' e en içten teşekkürlerimi sunarım.

Araştırma sürecimdeki tüm aşamalarda bilgi birikimi, tecrübesi ve ilham veren tutumu ile bana büyük bir güven ve destek sağlayan Dr. Öğr. Üyesi Sayın Ahmet Cumhur ÖZTÜRK' e, benden hiçbir zaman desteğini esirgemeyen ve her adımında yanımda olarak yoluma ışık olan değerli babam Necati GÜR' e ve sevgili annem Saniye GÜR' e, kıymetli ablalarım Zahide GÜR KOÇ ve Gamze GÜR ÖZDURAN' a, tüm süreçlerde motivasyon kaynağım olan yol arkadaşım Deniz'ime bu süreçte gösterdikleri sabır ve sevgiden dolayı sonsuz teşekkürlerimi sunarım.

İÇİNDEKİLER

KABUL VE ONAY SAYFASI.....	iii
BİLİMSEL ETİK BİLDİRİM SAYFASI.....	iv
ÖZET.....	v
ABSTRACT	vii
TEŞEKKÜR	ix
İÇİNDEKİLER.....	x
ŞEKİLLER DİZİNİ.....	xii
TABLolar DİZİNİ.....	xiv
GİRİŞ.....	1
1. BÖLÜM.....	4
1. GENEL BİLGİLER.....	4
1.1. Veri Madenciliği.....	4
1.1.1. Veri Tabanında Bilgi Keşif Süreci.....	5
2. BÖLÜM.....	7
2. SIK ALT KÜME MADENCİLİĞİ VE BİRLİKTELİK KURAL MADENCİLİĞİ.....	7
2.1. Sık Alt Küme ve Birliktelik Kural Madenciliği Örneği	10
2.2. Veri Tabanı	11
2.3. İlişkisel Veri Tabanı	11
2.4. İlişkisel Olmayan Veri Tabanı.....	13
2.5. İlişkisel Olan ve İlişkisel Olmayan Veri Tabanı Karşılaştırması	14
2.6. İşlemsel (Transactional) Veri Tabanı	15
3. MATERYAL VE YÖNTEM	18
3. ÇİZGE VERİ TABANI.....	18
3.1. Bağlantılı ve Bağlantısız Çizge Veri Tabanı	21
3.2. Yönlü ve Yönsüz Çizge Veri Tabanı.....	22
3.3. Çizge Veri Tabanı Uygulamaları.....	23
3.4. Tek Bir Büyük Çizge	24
3.5. İzomorfizm	26
3.6. İlişkisel Veri Tabanı ve Çizge Veri Tabanı Farkları	27
4. BULGULAR	29
4. ÇİZGE MADENCİLİĞİ.....	29

4.1. Sık Alt Çizge Madenciliği	31
4.1.1. Sık Alt Çizge Madenciliği Algoritmalarının Kullanım Alanları	34
4.1.2. gSpan Algoritması.....	36
4.1.2. GraMi Algoritması.....	40
4.2. Uygulama.....	42
5. TARTIŞMA VE SONUÇ.....	47
6. KAYNAKLAR.....	49
ÖZGEÇMİŞ.....	53



ŞEKİLLER DİZİNİ

Şekil 1.1. Veri madenciliğinin etkileşimde bulunduğu disiplinler	4
Şekil 1.2. Veri Tabanında Bilgi Keşif Süreci.....	5
Şekil 2.1. İlişkisel Veri Tabanı.....	12
Şekil 3.1. Çizge Örneği	18
Şekil 3.2. Genel çizge veri tabanı modeli	19
Şekil 3.3. Çizge gezinme.....	20
Şekil 3.4. Çizge indeksleme	21
Şekil 3.5. Bağlantılı ve Bağlantısız Çizge Örneği.....	21
Şekil 3.6. Yönlü ve Yönsüz Çizge Örneği	22
Şekil 3.7. İş Birliği Çizgesi	25
Şekil 4.1. 7 kenar ve 5 düğümden oluşan çizge	29
Şekil 4.2. Üç Çizge Örneği.....	32
Şekil 4.3. Sık alt çizge kümesi	33
Şekil 4.4. Sık Alt Çizge Örneği.....	33
Şekil 4.5. Sık alt çizge yapıların örnek kullanım alanları	35
Şekil 4.6. Bileşik içerisinde ortak bileşik gösterimi.	36
Şekil 4.7. Sosyal ağ örneği.....	36
Şekil 4.8. gSpan örnek uygulanması için G1 ve G2 çizgesi.....	38
Şekil 4.9. DFS kod örneği	39
Şekil 4.10. GraMi Örnek Uygulaması için G çizgesi.....	41
Şekil 4.11. Sonuc 1.....	43
Şekil 4.12. Sonuc 2.....	43
Şekil 4.13. Sonuc 3.....	44
Şekil 4.14. Sonuc 4.....	45

Şekil 4.15. Sonuc 5.....	45
Şekil 4.16. Sonuc 6.....	46



TABLÖLAR DİZİNİ

Tablo 2.1. Ürün satış tablosu	10
--	----



GİRİŞ

Veri madenciliği, büyük veri kümelerinden anlamlı bilgilerin çıkarılmasına yardımcı olan bir dizi yöntem ve tekniktir. Bu süreç, karmaşık ve büyük boyutlu veri kümelerini analiz ederek anlamlı desenler, ilişkiler ve eğilimler bulmayı amaçlar. Veri madenciliği teknikleri, işletmelerin müşteri tercihlerini anlamalarına, iş süreçlerini iyileştirmelerine, riskleri yönetmelerine ve yeni fırsatları keşfetmelerine yardımcı olur. Aynı zamanda sağlık sektöründe hastalıkların teşhisi ve tedavi planlaması gibi hayati öneme sahip alanlarda da değerli bilgiler sunar. Veri madenciliği, verilerin daha iyi anlaşılmasını sağlayarak bilgi çağında rekabet avantajı elde etmeyi mümkün kılar. Bu teknikler, verilerdeki desenleri, ilişkileri ve eğilimleri keşfetmek için istatistiksel analizler ve makine öğrenimi algoritmalarını kullanır.

Çizge madenciliği, veri madenciliğinin bir alt konusudur ve veri tabanı çizge olarak ifade edilmektedir. Çizgeler, veri madenciliğinin birçok alanında güçlü veri yapıları olarak kullanılmaktadır. Özellikle son yıllarda çizge madenciliğine olan ilgi giderek artmaktadır. Bunun bir nedeni; sosyal medyanın internetin yükselişi ve tıbbi araştırmaların bilgisayar ortamına taşınmasıyla birlikte yapılandırılmış veri yapılarına olan ihtiyacın artması ve araştırmacıların önemli bilgiler istemesidir. Yapılandırılmış veri kümelerinde, önemli ölçüde tekrar eden altyapıların kaldırılması; bu verileri yorumlamaya ve genelleştirmeye olanak tanımaktadır. Öte yandan, bir bilgi alanında çalışan bir kullanıcı, önceden tanımlanmış bir türde veya bu alana özgü altyapılar aramaktadır. Çizge veri tabanı sayesinde nesneler arasında ilişkiler ortaya çıkarılmaktadır. Çizge veri tabanı düğüm ve kenarlardan oluşmaktadır, düğüm nesneyi, kenar ise nesneler arasındaki ilişkiyi ifade etmektedir. Çizge veri tabanı birçok küçük çizgeden meydana gelebilir ya da tek bir büyük çizgeden meydana gelebilir. Birçok küçük çizgeden meydana gelen veri tabanı işlemsel çizge veri tabanı olarak adlandırılır, tek bir büyük çizgeden meydana gelen veri tabanı tek bir büyük çizge veri tabanı olarak adlandırılır. Çizge veri tabanları sayesinde birçok farklı disiplinlerde yer alan veriler ifade edilebilir; Örneğin Kimyasal bileşenler, genetik yapılar, bilgisayar ağları, sosyal medya vb.

Herhangi bir alt çizge belirli bir eşik parametresinden büyük veya ona eşit destek (support) değerine sahip ise; sık alt çizge olarak adlandırılır. Sık Alt Çizge Madenciliği yöntemleri iki farklı modelde verilere uygulanmaktadır. Birincisi işlemsel çizge veri tabanı,

ikincisi ise tek bir büyük çizge veri tabanıdır.

Sık Alt Çizge Madenciliği, büyük veri kümelerinde tekrarlanan desenleri ve önemli ilişkileri keşfetmek amacıyla kullanılan bir veri madenciliği tekniğidir. Bu teknik, veri kümesinde sıklıkla tekrarlanan alt çizgeleri bulmak için özel algoritmalar ve yöntemler kullanır. Sık alt çizgeler, veri kümesindeki sıklık eşiğini geçen ve belirli bir anlam taşıyan desenlerdir. Sık alt çizge madenciliği, farklı alanlarda çeşitli uygulama alanlarına sahiptir. Örneğin, veri tabanı yönetimi, sosyal ağ analizi, biyoinformatik ve çizge analizi gibi alanlarda sık alt çizge madenciliği önemli bir rol oynar. Bu teknik, veri kümesindeki tekrarlanan desenleri belirleyerek işletmelerin daha iyi kararlar almasını sağlar, pazarlama stratejilerini optimize eder, ağ yapısını analiz eder ve biyolojik verilerde önemli bilgileri ortaya çıkarır.

Sık Alt Çizge Madenciliği uygulamaları terabaytlarca bilgi üzerinde analiz yapar ve verinin ilişkilerini, hatta değiştiğinde ve geliştiğinde bile korur. Web siteleri sıfırdan milyonlara çıktıkça, geleneksel ilişkisel veri tabanlarının performans seviyeleri düşer. Çizge veri tabanları, uygulama geliştirmeyi basitleştirir. Böylelikle kısa geliştirme süreleri, düşük bakım maliyetleri ve yüksek performans elde edilir. Bu nedenle sosyal olarak yetenekli uygulamalar, milyonlarca kullanıcı arasındaki ilişkileri yönetmek için diğer veri tabanı türlerinin etkili olmamasından dolayı çizge veri tabanlarına yönelmektedir.

Gerçekleştirilen bu çalışma sayesinde araştırmacıların ve kurumların sık alt çizge madenciliği nasıl ortaya çıkarabildiği, sık alt çizgelerin hangi yöntem ve teknikler ile ortaya çıkarılabileceği, ortaya çıkartılan sık alt çizgelerin kurumların ve araştırmacıların nasıl faydalarına sunulabileceği ortaya konacaktır. Çalışma planı üç aşamadan oluşmaktadır. Birinci aşama gündelikte kullanılan gerçek verilerin çizge veri tabanına dönüştürülmesini kapsamaktadır. İkinci aşama oluşturulan çizge veri tabanından sık alt çizgelerin ortaya çıkartılmasını kapsamaktadır. Üçüncü aşama ise bu sık alt çizgelerin bilgisel olarak önemliliklerinin kanıtlanmasıdır.

Bunun yanında literatürde yer alan işlemsel çizge veri tabanı için önceden geliştirilmiş güncel sık alt çizge algoritması (GSPAN algoritması) örnek vaka üzerinde uygulanacaktır. Örnek vaka uygulaması için imdb internet sitesinden (www.imdb.com) 2000 – 2022 yılları arasında vizyona girmiş yerli filmler (Türk filmleri) ve bu filmlerin özellikleri çekilmiştir. Film özellikleri; film adı, film yılı, film süresi, film türü, film yönetmeni, film yazarı, film

oyuncuları, film imdb puanı şeklindedir. Toplam 326 film, 210 farklı yönetmen, 680 farklı oyuncu, 285 farklı yazar ve 13 farklı türe ulaşılmıştır. Imdb veri tabanı, işlemsel çizge veri tabanına dönüştürülmüştür. Oluşturulan işlemsel veri tabanı üzerinde GSPAN (referans) algoritması çalıştırılmış ve sık alt çizgeler ortaya çıkarılmıştır.



1. BÖLÜM

1. GENEL BİLGİLER

1.1. Veri Madenciliği

Bilgi çağında herhangi bir karar alma biriminin en önemli malzemesi veri ve bilgidir. Günümüzde sadece veri elde etmek değil, eldeki verilerden maksimum bilgiyi elde etmek de önemli bir olgu haline gelmiştir. Farklı disiplinler ile bağlantılı olan veri madenciliği, belirli bir amaca yönelik anlamlı ve faydalı bilgiyi elde etmek için büyük veri setleri içerisinde, istatistik, matematik disiplinleri, modelleme teknikleri, veri tabanı teknolojileri ve bilgisayar programları kullanılarak aranması ve analiz edilmesidir (Silahtaroglu, 2008). Veri madenciliği, potansiyel olarak yararlı ve yeni mantıksal bilgileri ortaya çıkarmak için büyük veri tabanlarının çok düzeyli analizini gerçekleştirir. Veri tabanlarında yer alan önceden bilinmeyen faydalı bilgiler, veri madenciliği uygulamaları ile ortaya çıkmaktadır.



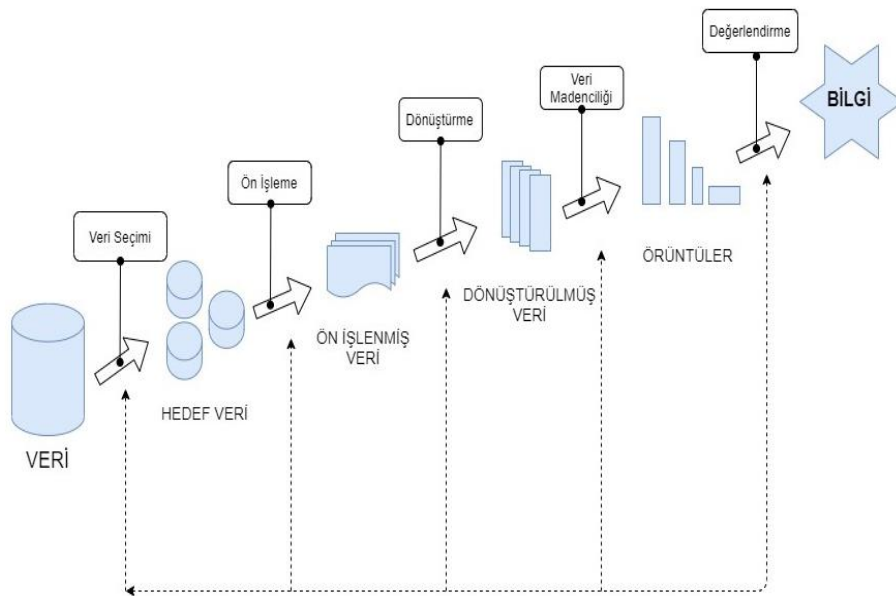
Şekil 1.1. Veri madenciliğinin etkileşimde bulunduğu disiplinler (Han ve diğ., 2012)

Şekil 1.1’de gösterildiği gibi, veri madenciliği istatistik, makine öğrenimi ve veri tabanı yönetimi vb. gibi alanlardan yararlanan çok disiplinli bir alandır. Ham veriden, son kullanıcının anlayabileceği bilgilerin üretilmesine kadar tüm süreci kapsayan ve karar verme sürecine dahil edilebilecek bir yöntemdir, bu nedenle hipotezleri test etmek için değil, yeni gizli örüntüler bulmak için kullanılır. Çeşitli tekniklerin aynı uygulamada kullanılabilmesi, veri madenciliği kullanıcılarına fırsatlar sunmakta ve onu oluşturan makine öğrenimi, istatistik, matematik ve diğer yöntemlerden farklı bir bakış açısı sağlamaktadır (Ay, 2009).

Veri madenciliği, karmaşık veri yapılarının analiz edilmesini sağlar ve tahmin, sınıflandırma, kümeleme ve ilişki kurma gibi analitik görevleri gerçekleştirir. Bu süreçte, veri madenciliği algoritması, veri setindeki desenleri ve ilişkileri belirleyerek öngörücü modeller oluşturur. Veri madenciliği, pazarlama, finans, sağlık, doğal dil işleme ve sosyal ağ analizi gibi birçok alanda uygulanmaktadır (Han ve diğ., 2011).

1.1.1. Veri Tabanında Bilgi Keşif Süreci

Şekil 1.2’de veri tabanındaki bilgi keşif süreci gösterilmektedir. Her işlemi tamamladıktan sonra bir sonraki işleme başlamalı ve veri madenciliği aşamasına gelmelisiniz. (Şen, 2008).



Şekil 1.2. Veri Tabanında Bilgi Keşif Süreci

Veri madenciliği aşamasında, veriye ve istenilen sonuçlara uygun bir veri madenciliği yöntemi seçip uygulanmaktadır (Şen, 2008). Veri tabanında bilgi keşfi süreci genellikle aşağıdaki aşamalardan oluşur:

Veri Seçimi: Bilgi keşfi sürecinin ilk aşaması, analiz için kullanılacak verilerin seçilmesidir. Bu adımda, ilgili veri tabanından veya veri kaynaklarından gerekli veri setleri belirlenir. Veri seçimi, analiz amacına ve hedeflenen sorulara bağlı olarak yapılır.

Veri Ön İşleme: Veri ön işleme aşaması, seçilen veri setinin kalitesini artırmak, gereksiz veya eksik verileri düzeltmek ve veri uyumsuzluklarını gidermek için uygulanan bir dizi işlemi içerir. Bu işlemler arasında veri temizleme (outlier, eksik veya gürültülü verilerin düzeltilmesi), veri entegrasyonu (farklı kaynaklardan gelen verilerin birleştirilmesi) ve veri dönüşümü (veri formatının ve yapısının değiştirilmesi) bulunur.

Veri Dönüştürme: Veri dönüştürme aşamasında, veri seti, analiz için daha uygun bir formata veya yapıya dönüştürülür. Bu dönüşümler, veri setinin analiz algoritması tarafından işlenebilir hale getirilmesini sağlar. Örneğin, kategorik verilerin sayısal değerlere dönüştürülmesi veya veri normalizasyonu gibi işlemler bu aşamada gerçekleştirilir. **Veri Madenciliği:** Veri madenciliği aşamasında, dönüştürülen veri seti üzerinde çeşitli veri madenciliği teknikleri uygulanır. Bu aşamada, desen tanıma, sınıflandırma, kümeleme veya ilişkisel kurallar gibi farklı veri madenciliği yöntemleri kullanılarak veri setindeki gizli bilgiler ve desenler ortaya çıkarılır. Bu aşama, bilgi keşfi sürecinin merkezi noktasını oluşturur.

Değerlendirme: Değerlendirme aşaması, veri madenciliği sonuçlarının analiz edilmesini ve değerlendirilmesini içerir. Bu aşamada, elde edilen sonuçlar ve keşfedilen bilgiler, belirlenen hedeflere ve analiz amacına uygunluğu açısından değerlendirilir. Ayrıca, model performansı, doğruluk oranı, hassasiyet, özgüllük gibi metrikler kullanılarak veri madenciliği modellerinin etkinliği ölçülür. Bu aşamaların sıralaması ve içerikleri, spesifik bir veri tabanı ve analiz sürecine bağlı olarak değişebilir. Ancak genel olarak, veri seçimi, ön işleme, veri dönüştürme, veri madenciliği ve değerlendirme aşamaları, bilgi keşfi sürecinin temel adımlarını oluşturur.

2. BÖLÜM

2. SIK ALT KÜME MADENCİLİĞİ VE BİRLİKTELİK KURAL MADENCİLİĞİ

Sık alt küme madenciliği, büyük veri kümelerinde tekrarlayan veya sık sık birlikte görülen öge kümelerini belirlemek için kullanılan bir veri madenciliği tekniğidir. Bu yöntem, özellikle pazarlama, perakende ve araştırma gibi alanlarda kullanılarak değerli bilgilerin elde edilmesini sağlar. Sık alt küme madenciliği, öğelerin birlikte sıklıkla görülme örüntülerini tespit ederek işletmelere; önemli ilişkilerin belirlenmesi, kişiselleştirilmiş hizmetler, reklam ve pazarlama stratejileri, verimlilik ve maliyet avantajı, yenilikçi ürün ve hizmetlerin geliştirilmesi gibi rekabet avantajları sağlar. Bu yöntemde, veri kümesindeki öğelerin sıklıkları ve frekansları analiz edilir. Sık alt küme madenciliği algoritması, veri kümesindeki frekans ölçümlerini kullanarak sık sık birlikte görülen öğeleri belirler. Örneğin, bir perakende mağazasının satış verilerinde yapılan sık alt küme madenciliği analizi, müşterilerin belirli ürünleri birlikte satın alma eğilimlerini tespit edebilir. Bu bilgi, çapraz satış stratejilerinin geliştirilmesi ve müşteri segmentasyonu için kullanılabilir.

Sık alt küme madenciliği, veri kümesindeki gizli ilişkileri ve kalıpları ortaya çıkarmak için etkili bir araçtır. Bu yöntem, pazarlama kampanyalarının kişiselleştirilmesi, ürün yerleştirme stratejilerinin optimize edilmesi ve işletmelerin müşteri ihtiyaçlarına daha iyi yanıt vermesi gibi birçok alanda değerli bilgiler sağlar (Han ve diğ., 2011).

Sık alt küme madenciliği, veri madenciliği alanında önemli bir tekniktir ve büyük veri kümelerindeki tekrarlayan öğelerin ve öge gruplarının belirlenmesine odaklanır. Bu yöntem, bir veri kümesindeki sık sık birlikte görülen öğeleri ve bunların frekanslarını analiz eder. Sık alt küme madenciliği, pazarlama, perakende, tıp ve sosyal ağ analizi gibi birçok alanda kullanılan etkili bir araçtır. Bu yöntemde, sık alt küme madenciliği algoritması, veri kümesindeki öğelerin frekanslarını hesaplayarak sık sık birlikte görülen öğeleri belirler. Örneğin, bir e-ticaret sitesinin kullanıcı satın alma verilerinde yapılan sık alt küme madenciliği analizi, kullanıcıların birlikte sıkça satın aldıkları ürün gruplarını tespit edebilir. Bu bilgi, öneri sistemlerinin geliştirilmesi, kişiselleştirilmiş kampanyaların oluşturulması ve müşteri sadakati stratejilerinin uygulanması gibi alanlarda değerli bir bilgi kaynağıdır. Sık alt küme madenciliği, büyük veri kümesindeki önemli kalıpları ortaya çıkarmak ve işletmelere rekabet avantajı sağlamak için kullanılan bir yöntemdir. Bu yöntem, veri analizinde

derinlemesine anlayış sağlayarak karar verme süreçlerini iyileştirir ve müşteri ilişkilerinin geliştirilmesine katkıda bulunur (Aggarwal, Zhai, 2012).

Veri tabanlarının sayı ve hacmindeki hızlı artış ile birlikte verilerden anlamlı iç görüler elde etmek, stratejik karar vermeyi desteklemek ve güvenilir tahminler yapmak için giderek daha önemli hale gelmektedir. Bu işlevi yerine getiren araç ise veri madenciliğidir. Başka bir deyişle, veri madenciliği, büyük miktarda depolanmış veride saklı olan kuralları, kalıpları keşfetme sürecidir. Veri madenciliği bu işlemi bilgisayar bilimi, makine öğrenimi, veri tabanı yönetimi, matematiksel algoritmalar ve istatistikleri birleştirerek gerçekleştirir. Veri bilgi yığınlarındaki gizli bilgiler, bir işletmeye stratejik planlar oluşturmak için faydalı ipuçları verebilmektedir. Veri madenciliği, ana veri madenciliği, istatistik, OLAP gibi diğer veri analiz tekniklerini tamamlayıcı bir yaklaşımdır. Veri madenciliği, verileri boyutlara göre gruplandırmanıza, boyutlar arasındaki ilişkileri keşfetmenize, verileri görselleştirmenize ve sonuçları çizgeler ve raporlarda sunmanıza olanak tanımaktadır. Ortaya çıkan kişisel bilgi, özellikle stratejik kararlar alırken işletme yöneticileri için iyi bir yardımcıdır.

Veri madenciliğinde tahmin edici modellerle örüntü tanıma çalışması sınıflandırma, regresyon ve zaman serisi yaklaşımlarını içermektedir. Bu modeller, neyin tahmin edilmesi gerektiğine bağlı olarak farklılık göstermektedir. Veri çıktısının sürekli değerlerini tahmin etmek istiyorsanız, regresyon analizi kullanmanız gereklidir. Zaman serilerinin zaman içindeki değişimlerini incelemek istiyorsanız, az sayıda özel bir veri noktası için kategorik bir değer tahmin etmeniz gerekebilir. Örneğin "iyi" veya "kötü" gibi ayrık kategorilere sınıflandırmak. Bu durumda, kümeleme, ilişkilendirme analizi ve dizi algılama gibi teknikler kullanılabilir. Bu yöntemler, verilerdeki grupları belirlemek, ilişkileri analiz etmek ve dizi tabanlı kurallar türetmek için kullanılır.

Kümeleme, nesneleri özelliklerine göre gruplandırmanın temel amacıdır. Bu çok değişkenli bir analiz yöntemidir. Kümelemede nesneler önceden belirlenmiş bir kritere göre gruplandırılmaktadır. Kümeleme sonuçları, kümeler içinde yüksek derecede homojenlik ve kümeler arasında yüksek derecede heterojenlik göstermelidir. Rastgele değişken kavramı, kümelemenin yapısında önemli bir yer tutmaktadır. Ancak rastgele bir değişkenin kümeleme yapısındaki yeri diğer çok değişkenli yöntemlerden tamamen farklılık göstermektedir. Küme rastgele değişkeni, kümeleme sırasında nesneleri karşılaştırmak için kullanılan özellikleri

temsil eden bir deęişkenler kümesidir. Belirlenen rasgele deęer, yalnızca nesneleri karşılaştırmak için kullanılan deęişkenleri içerdüğinden, nesnelerin özelliğini belirlemektedir. Kümeleme, rasgele deęişkeni deneysel olarak tahmin etmeyen ancak araştırmacı tarafından rasgele deęişkeni belirleyen çok deęişkenli bir analiz yöntemidir. Kümelemenin odak noktası rastgele bir deęişkenin tahmini deęil, nesnelerin rastgele bir deęişkene dayalı olarak karşılaştırılmasıdır.

Tipik olarak, kümeleme kontrolsüz bir şekilde gerçekleşmektedir. Denetimsiz öğrenme yeteneęi ile verilerde gizli olan kalıpları ortaya çıkarabilmektedir. Kümeleme; örüntü tanıma, biyoinformatik ve veri madencilięi gibi çeşitli alanlarda önemli bir araştırma konusu haline gelmiştir, bu nedenle veri madencilięi bu disiplinlerle sıkı bir ilişkiye sahiptir (Taşkın ve Emel, 2010).

İşlemsel veri tabanından ortaya çıkarılan sık alt kümelerden yararlanılarak birliktelik kuralları ortaya çıkarılır. Birliktelik kural madencilięi, büyük veri kümeleri içerisindeki ilişkileri ve eşzamanlılıkları ortaya çıkarmak için kullanılan bir veri madencilięi tekniğidir. Bu teknik, özellikle perakende sektörü ve pazarlama alanında yoğun olarak kullanılmaktadır. Birliktelik kural madencilięi, veri kümesindeki öğeler arasında sık sık birlikte görülen kalıpları ve ilişkileri belirlemek için istatistiksel yöntemler kullanır. Bu yöntemde, veri kümesindeki öğeler arasındaki ilişkileri ifade eden kurallar çıkarılır. Örneğin, "X ürününü satın alan müşteriler genellikle Y ürününü de satın alıyorlar" şeklinde bir kural elde edilebilir. Bu kurallar, perakende işletmeleri için çapraz satış fırsatlarını belirlemek, müşteri davranışını anlamak ve pazarlama stratejilerini optimize etmek için kullanılabilir.

Birliktelik kural madencilięi algoritması, veri kümesindeki frekans ve destek gibi istatistiksel ölçümleri kullanarak anlamlı kuralları tespit eder. Destek; bir kurallı kapsayan veri öğelerinin yüzdesini ifade ederken güvenilirlik kurallı kapsayan öğelerin yüzdesini ifade etmektedir. Bu ölçümler, kuralların anlamlılığını değerlendirmek ve kullanıcıya en deęerli kuralları sunmak için kullanılır. Birliktelik kural madencilięi, büyük veri kümelerinde gizli ilişkileri ve kalıpları ortaya çıkarmak için etkili bir yöntemdir. Bu yöntem, pazarlama kampanyalarının hedeflenmesi, ürün yerleştirme stratejilerinin geliştirilmesi ve işletmelerin rekabetçi avantaj elde etmesi gibi birçok alanda deęerli bilgiler sağlar (Tan vd., 2006).

Birliktelik kuralları, büyük veri kümeleri arasında ilginç veya ilişkili ilişkiler bulmak için kullanılır. Birliktelik kuralları, belirli bir veri kümesindeki ortak özellik değerlerinin örneklerini tanımlar. Birliktelik kurallarının tipik ve yaygın olarak kullanılan bir alanı market sepeti analizidir (Şen, 2008).

Yapılan bu analizlerde tek seferde birlikte alınan ürünler arasında ilişkiler ele alınır. Bu kural ile müşterilerin satın alma alışkanlıkları analiz edilerek, hangi ürünlerin birlikte alındığı bilgisine ulaşılır. Analiz sonrası market, mağaza yöneticileri rafların yerleşim yerlerini belirleyip, etkin bir satış stratejisi geliştirerek satış oranını arttırabilir (Döşlü, 2008).

2.1. Sık Alt Küme ve Birliktelik Kural Madenciliği Örneği

Destek(support) ve güven(confidence) değerleri, sırasıyla ortaya çıkarılan kuralın alakalılığını ve ilginçliğini ifade etmektedir. Bu değerlerin her biri, tespit edilen kuralın kullanışlılığını (yararlılığını) ve doğruluğunu (kesinliğini) temsil eder. Destek değeri, A ve B nesnelerinin bir arada bulunma olasılığını elde etmemizi sağlar. Güven değeri ise A' yı içeren kayıtların B' yi de içerdiği anlamına gelir (Döşlü, 2008). Destek değeri (1) nolu formül ile hesaplanır, güven değeri (2) nolu formül ile açıklanır.

Tablo 2.1. Ürün satış tablosu

TİD	ÜRÜNLER
1	Bisküvi, Ekmek, Soda, Süt
2	Balık, Ekmek, Bisküvi, Soda
3	Çikolata, Ekmek, Bisküvi, Süt
4	Ekmek, Bisküvi, Süt
5	Çikolata, Bisküvi, Soda, Süt

Tablo 2.1' deki ürün alış hareketlerine göre {Bisküvi, Süt} ile Soda arasındaki ilişki şu şekilde açıklanabilir:

$$\text{Destek(support)} = \frac{\text{{Bisküvi, Süt, Soda}}}{\text{Toplam Hareket}} = \frac{2}{5} = 0.4 \quad (1)$$

$$\text{Güven (confidence)} = \frac{\text{{Bisküvi, Süt, Soda}}}{\text{{Bisküvi, Süt}}} = \frac{2}{4} = 0.5 \quad (2)$$

Bu tablo üzerinde birliktelik kural madenciliği uygulandığında yukarıdaki eşitliklerde {Bisküvi, Süt} $\Rightarrow$ Soda kuralı %40 destek, %50 güven ölçülerine sahip olduğu anlaşılmaktadır. Güven değerinin %50 olmasından çıkan sonuç; Bisküvi ve Süt ürünlerini alanların %50' si Soda ürününü de almışlardır. Güven değerinin %100 olması demek Bisküvi ve Süt ürünlerini alan her kişi Soda ürününü de satın almış anlamına gelmektedir. Bu durum kesin kural olarak adlandırılır. Destek değerinin %40 olmasından çıkan sonuç; yapılan tüm satışların %40' ında Bisküvi, Süt ve Soda birlikte yer almaktadır.

2.2. Veri Tabanı

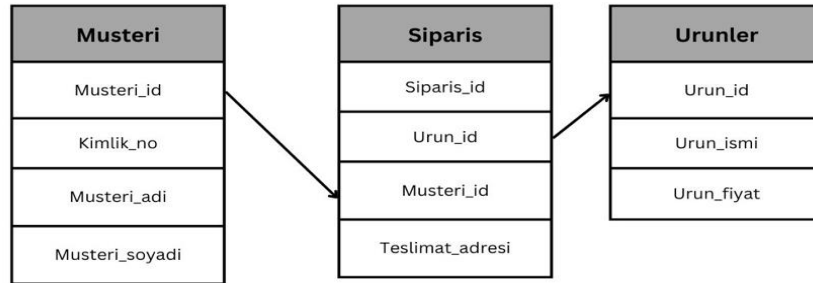
Veri tabanı, belirli bir şekilde düzenlenmiş bir veri topluluğudur. Verileri bir arada tutan yapıdır. Bu yapı, yüklenen veri miktarı ne kadar büyük olursa olsun veya kaç kullanıcı tarafından işleniyor olursa olsun, veri işlemlerine güvenilir ve zamanında cevap vermenin temel görevidir. Veri tabanı yönetim sistemi için ön koşullar; farklı kullanıcıların aynı verileri aynı anda kullanabilmesi, farklı uygulama programları gerektirmesi, veriler üzerindeki işlemlerin tutarlı olması ve veriler üzerindeki etkinin tutarlı olmasıdır (Yorulmaz,2018:10).

2.3. İlişkisel Veri Tabanı

İlişkisel veri tabanları, ilişkisel model adı verilen bir veri tabanı modeline dayanır. Bu modelde, veriler tablolara ve tablolardaki veri öğeleri arasındaki ilişkilere dayalı olarak organize edilir. İlişkisel veri tabanları, verilerin yapısal bütünlüğünü sağlamak için birtakım kısıtlamalar kullanır ve veri tabanı yöneticilerine veri bütünlüğünü koruma ve verileri tutarlı bir şekilde saklama imkanı verir. İlişkisel veri tabanlarının en büyük avantajlarından biri, verilerin farklı tablolar arasında paylaşılabilmesi ve ilişkilendirilebilmesidir. Bu sayede, verilerin tekrarlanması önlenir ve veri tabanı yapısı daha verimli hale gelir (Elmasri ve Navathe, 2016).

İlişkisel veri tabanları, verilerin tablolar halinde düzenlendiği ve tablolardaki veri ögeleri arasındaki ilişkilerin kullanıldığı bir veri tabanı modelidir. İlişkisel modelde, her tablo bir ilişkiyi temsil eder ve tablolar arasındaki ilişkiler, anahtarlar kullanılarak kurulur. İlişkisel veri tabanları, ACID (Atomicity, Consistency, Isolation, Durability) özelliklerine dayanarak veri bütünlüğünü sağlar. Bu özellikler, verilerin tutarlılığını, güvenilirliğini ve işlemlerin güvenli bir şekilde gerçekleştirilmesini garanti eder. İlişkisel veri tabanlarına erişim, SQL (Structured Query Language) adı verilen bir sorgulama dili kullanarak gerçekleştirilir. SQL, verileri sorgulama, ekleme, güncelleme ve silme gibi işlemleri yapmak için kullanılan bir dil olarak bilinmektedir (Connolly ve Begg, 2014).

Bu veri tabanı modeli ilk olarak E.F. Codd tarafından geliştirilmiştir. Bu veri tabanları birçok ilişkinin bir araya gelmesi ile oluşmaktadır. Bu ilişkiler ilişkisel veri tabanı modelinde satırlar ve sütunlardan oluşan tablolar ile gösterilmektedir. Bu tabloda bulunan her bir satır birbirleriyle ilişkili olan veriler topluluğudur. Nitelikler sütunlarda yer alır (Öztürk ve Atmaca, 2017). Tablolar birbirlerine anahtar adı verilen veriler ile bağlanır. Oluşturulan bu ilişki sayesinde birbirlerine bağlı verilerin bir seferde sorgusu yapılabilir (Ünalır vd, 2015). Şekil 2.1.' de ilişkisel veri tabanı örneği verilmiştir.



Şekil 2.1. İlişkisel Veri Tabanı

İlişkisel veri tabanı modelinde yalnız veriler değil, veriler arasındaki ilişkiler de depolanmaktadır (Karacabey, 1997).

İlişkisel veri tabanı bilimsel olarak matematiksel teorilere dayanmaktadır. Veri yapısı farklı olsa da ilişki doğru kurulursa ve tutarlı kısıtlamalar tanımlanırsa, karşılık gelen veri tabanı yönetim sistemi her türlü veri tabanı uygulamasına uyum sağlayabilmektedir.

Tablolar, anahtarlar, ilişkiler, indeksler ilişkisel veri tabanı yönetim sistemi' nin ana unsurlarıdır. İlişkisel veri tabanında verilerin saklandığı birimler tablolardır. Anahtarlar, birçok tabloda tabloları birbirine bağlamak ve veri tekrarını azaltmak için kullanılır, böylece iyi bir veri taşıyıcısı oluşturulmaktadır. İndeksler veri aramak için kullanılmaktadır (Yorulmaz, 2018).

2.4. İlişkisel Olmayan Veri Tabanı

1998'de Carlo Strozzi, NoSQL kavramını tanıtmıştır. Bu, SQL dışı gibi görünse de gerçekten olması gerekmiyordu. Bu kavram veri tabanı dünyasına "sadece SQL" olarak değil, ilişkisel olmayan bir veri tabanı olarak taşınmıştır. Bu kavram iddialı ve çeşitli alanlarda ilişkisel veri tabanı sistemleriyle rekabet halindedir ve etkili bir çözüm olarak kabul edilmektedir. Ancak, ilişkisel bir veri tabanı henüz bazı yönetim avantajlarını içermemektedir.

İlişkisel olmayan veri tabanları, geleneksel ilişkisel veri tabanlarının bazı kısıtlamalarını aşmayı amaçlayan veri yönetimi sistemleridir. Bu tür veri tabanları, büyük ve karmaşık veri kümelerini depolama, yönetme ve işleme yetenekleri ile önemli bir ilerleme sağlamıştır. İlişkisel olmayan veri tabanları, verileri tablolara, sütunlara ve ilişkilere bağımlı olmadan depolama imkânı sunar. Bu, yapısal olmayan veri türleri ve semantik esneklik gerektiren uygulamalar için büyük bir avantajdır. Belge tabanlı, anahtar-değer tabanlı ve çizge tabanlı veri modelleri, ilişkisel olmayan veri tabanlarının popüler seçenekleri arasındadır. İlişkisel olmayan veri tabanları, ölçeklenebilirlik, performans ve dağıtılmış veri işleme gibi alanlarda önemli faydalar sunmaktadır. Bu nedenle, büyük veri analitiği, bulut bilişim ve Nesnelerin İnterneti (Internet of Things-IoT) gibi veri yoğun alanlarda giderek daha fazla kullanılmaktadır (Özgür ve Özsu, 2015).

NoSQL veri tabanlarının genel özelliklerine bakıldığında, dört tür veri tabanı bulunmaktadır:

- Kolon türü veritabanları
- Belge türü veritabanları
- Anahtar değer veritabanları

- Çizge tipi veritabanları

Kolon veri tabanları, verileri Şekil 4.2.'de gösterildiği gibi sütunlarda depolamaktadır. Veriler sütunlara göre sıralanır. Bu, birçok sunucudaki büyük verileri sorgulamanıza olanak tanır. Bu, arama motorları için ideal bir modeldir.

Belge türü veri tabanları ise, bir dizi anahtar-değer çiftinden oluşan verilerin sürmelendirilmesiyle saklanan belge veri tabanlarıdır.

Anahtar Değer veri tabanlarında sağlama tablolar bulunur. Bu tablolarla ilgili değerler (hash'ler) mevcuttur. Günlük girişleri için ideal modeldir.

Çizge veri tabanları, çizge teorisine dayalı bir modeldir. Düğümler arasındaki ilişkiler bir veri tabanında saklanır. Forum istekleri çizge tipi veri tabanlarında saklanabilir (Yorulmaz, 2018).

NoSQL veri tabanları anlık tutarlılığı garanti edememektedir. Ancak, sonunda tutarlı raporlamaya izin vereceğini doğrulamaktadır. Zaten CAP teorisi olarak bilinen tutarlılık, erişilebilirlik, ayrılma toleransları kavramı; herhangi bir dağıtılmış sistemin hem tutarlı hem de kullanılabilir olamayacağı ve parçalanmadan yararlanamayacağı anlamına gelmektedir. Başka bir deyişle, dağıtılmış sistemin bir kısmına erişen tüm sisteme aynı anda yanıt verememektedir. Açıkçası, parçalanmış verileri anlamlandırmak biraz zaman almaktadır.

NoSQL kısaltması "BASE" (Temel Kullanılabilir-Yumuşak Durum-Sürekli Onay), ilişkisel veri tabanlarında, özellikle atomik işlem bilgi işleme (OLTP) çevrimiçi işlemlerinde kullanılan ACID' nin (Atomocity, Consistency, Isolation, Durability) eşdeğeridir (Yorulmaz, 2018).

2.5. İlişkisel Olan ve İlişkisel Olmayan Veri Tabanı Karşılaştırması

İlişkisel veri tabanları, verileri tablolarda organize eden ve ilişkiler aracılığıyla birbirine bağlayan yapılarıyla bilinir. İlişkisel veri tabanları, ACID (Atomicity, Consistency, Isolation, Durability) özelliklerini destekleyerek veri bütünlüğünü sağlar ve SQL (Structured Query Language) adı verilen bir dil aracılığıyla verilere erişim sağlar. İlişkisel veri tabanları, karmaşık ilişkileri ve bağımlılıkları kolayca temsil edebilme yetenekleriyle yaygın

olarak kullanılır. Verilerin yapısını deęiřtirmek veya yeni veri eklemek gerektiğinde, iliřkisel veri tabanlarında řema deęiřiklikleri yapılabilir.

Öte yandan, iliřkisel olmayan veri tabanları, genellikle büyük ve karmařık veri kümelerini depolamak ve iřlemek için kullanılır. Bu tür veri tabanları, hiyerarřık, aę veya belgelendirilmiş veri yapılarını kullanarak verileri organize eder. İliřkisel olmayan veri tabanları, veriye hiyerarřık veya aęsal bir yaklařımla eriřir ve JSON (JavaScript Object Notation) veya XML (eXtensible Markup Language) gibi formatlarda verileri saklar. İliřkisel olmayan veri tabanları, daęıtık ve ölçeklenebilir sistemlerde büyük miktarda veri depolamak ve iřlemek için daha uygundur. Veri yapısı deęiřiklikleri yapmak istendiğinde, iliřkisel olmayan veri tabanlarında esneklik saęlanır ve daha az řema deęiřiklięi gerekebilir.

İliřkisel veri tabanları ve iliřkisel olmayan veri tabanları arasında seęim yaparken, veri yapısı, veri boyutu, iřlem gereksinimleri ve uygulama ihtiyaęları gibi faktörler dikkate alınmalıdır. İliřkisel veri tabanları, iliřkisel modelin avantajlarından yararlanmak ve karmařık sorguları desteklemek için tercih edilirken, iliřkisel olmayan veri tabanları büyük veri setlerinin depolanması ve iřlenmesi için daha uygun olabilir. Her bir veri tabanı türü kendi avantajlarına ve kullanım senaryolarına sahiptir ve doęru seęim, veri yönetimi hedeflerini en iyi řekilde karřılamayı saęlar (Elmasri ve Navathe, 2016).

İliřkisel veri tabanları, iliřkiler yoluyla verilerin organize edildięi ve yapılandırıldıęı popüler bir veri tabanı modelidir. Bu model, tabloların kullanıldıęı ve tablolardaki sütunlar aracılıęıyla verilerin iliřkilendirildięi bir yaklařım sunar. SQL (Structured Query Language) kullanılarak sorgulama ve veri manipölasyonu geręekleřtirilebilir. İliřkisel veri tabanları, karmařık veri yapılarını yönetmek ve karmařık sorguları desteklemek için etkili bir çözümler sunar. Ayrıca, iliřkisel veri tabanları ACID (Atomicity, Consistency, Isolation, Durability) özelliklerini destekler, veri bütünlüęünü saęlar ve veri tutarlılıęını korur. Bununla birlikte, büyük veri kümeleri, ölçeklenebilirlik ve yüksek performans gerektiren durumlarda, iliřkisel veri tabanları bazen sınırlamalar yařayabilir (Rmakrishnan ve Gehrke, 2003).

2.6. İřlemsel (Transactional) Veri Tabanı

İřlemsel veri tabanı, günlük iřlemlerin geręekleřtirildięi ve güncel verilerin depolandıęı bir veri tabanı türüdür. İřlemsel veri tabanları, hızlı ve güvenilir iřlemler için

tasarlanmıştır. Bu tür veri tabanları, işlemleri otomatik, tutarlı, izole edilebilir ve dayanıklı (ACID) şekilde gerçekleştirmek için önemli özelliklere sahiptir. İşlemsel veri tabanları, birçok işlemi aynı anda destekleyebilen çok kullanıcı bir ortamda çalışır. Çoklu kullanıcı erişimi, işlemler arasında çakışmaların önlenmesi ve veri bütünlüğünün korunması için izolasyon düzeyleri kullanılarak yönetilir. İşlemsel veri tabanları aynı zamanda işlem geri alma ve geri dönme yetenekleri sağlayarak veri bütünlüğünün korunmasını sağlar.

İşlemsel veri tabanları, genellikle bankacılık, e-ticaret, envanter yönetimi gibi günlük işlemlerin yoğun olduğu uygulamalarda kullanılır. Bu tür veri tabanları, hızlı veri erişimi, veri bütünlüğü sağlama ve yüksek verimlilik gibi avantajlarıyla işlem performansını artırır.

İşlemsel veri tabanları, işlemlerin paralel olarak gerçekleştirilmesine olanak tanırken, aynı zamanda veri bütünlüğünü ve güvenliğini korumak için gelişmiş yöntemler ve kısıtlamalar sunar. ACID özellikleri sayesinde işlemler güvenli bir şekilde gerçekleştirilir ve veri bütünlüğü sağlanır (Coronel vd., 2016).

İşlemsel veri tabanları, günümüzde işletmelerin veri yönetimi ve iş süreçleri için hayati bir öneme sahiptir. Bu veri tabanı türü, büyük miktardaki verilerin etkin bir şekilde depolanması, işlenmesi ve güncellenmesi için özel olarak tasarlanmıştır. İşlemsel veri tabanları, işletmelerin günlük faaliyetlerinde kullanılan verilerin hızlı erişimi, paralel işleme yetenekleri ve yüksek güvenilirlik sağlaması açısından kritik bir rol oynamaktadır" (Lee ve Smith, 2020).

İşlemsel veri tabanlarının birçok uygulaması bulunabilir. Örneğin bir e-ticaret web sitesini ele alalım. Bir e-ticaret web sitesi; müşteri siparişlerini almak, stok yönetimini yapmak, ödeme işlemlerini gerçekleştirmek ve müşteri bilgilerini güncellemek gibi bir dizi işlemi içerir. İşlemsel veri tabanı, bu tür işlemleri gerçekleştirmek için kullanılan bir veri tabanıdır.

Örneğin, bir müşterinin siparişi oluşturulduğunda, işlemsel veri tabanı sipariş bilgilerini kaydedecek, stokta bulunan ürünleri azaltacak ve ödeme işlemlerini gerçekleştirecektir. Aynı şekilde, müşterinin bilgileri güncellediğinde veya yeni bir ürün eklendiğinde, işlemsel veri tabanı bu değişiklikleri hızlı ve güvenli bir şekilde yönetir.

İşlemsel veri tabanları, e-ticaret web sitesinin günlük işlemlerini etkin bir şekilde desteklerken aynı zamanda hızlı erişim, paralel işleme yetenekleri ve veri bütünlüğü gibi avantajlar sunar. Bu sayede işletme müşteri siparişlerini takip etme, envanter yönetimi yapma, finansal işlemleri yönetme gibi kritik faaliyetlerini etkin bir şekilde gerçekleştirebilir.



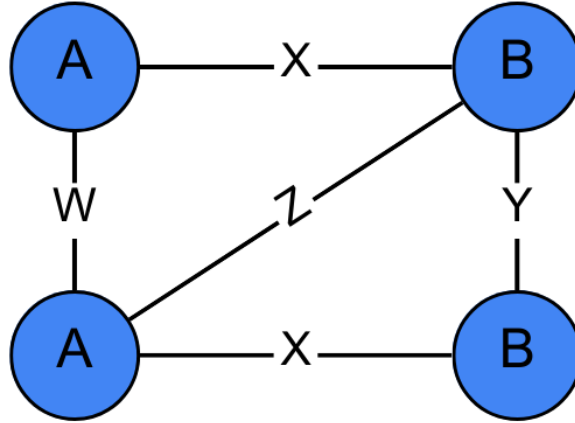
3. MATERYAL VE YÖNTEM

3. ÇİZGE VERİ TABANI

Çizge veri tabanları, günümüzde bilgi yönetimi ve veri analizi alanlarında giderek artan bir ilgiye sahiptir. Çizge veri tabanları, karmaşık ilişkileri ve bağlantıları temsil etmek için etkili bir model sunar ve bu da birçok uygulama ve disiplinde kullanım potansiyelini artırır. Sosyal ağ analizi, bilgi ağı analizi, biyoenformatik, ulaşım ağı optimizasyonu gibi alanlarda çizge veri tabanlarının kullanımı, kritik analitik ve karar verme süreçlerine önemli katkılar sağlamaktadır (Smith ve Jones, 2019).

Çizge veri tabanı, ilişkisel veri tabanlarına alternatif olarak, karmaşık yapıları ve ilişkileri temsil etmek için kullanılan bir veri tabanı modelidir. Çizge veri tabanları, düğümler ve kenarlar arasındaki ilişkileri kullanarak verileri çizge şeklinde organize eder. Bu yaklaşım, karmaşık veri kümelerindeki bağlantıları ve ilişkileri daha etkili bir şekilde analiz etmeyi sağlar (Dong vd.,2014).

Çizge, bazı etiketlere sahip bir dizi köşe ve kenardır.



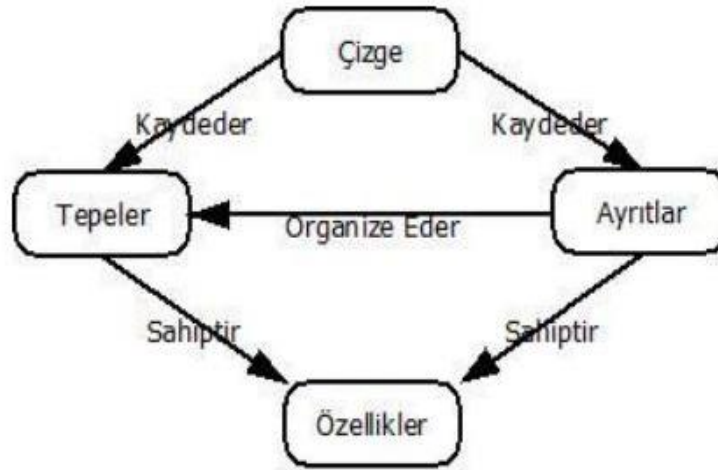
Şekil 3.1. Çizge Örneği

Şekil 3.1’deki çizge dört köşe içerir (mavi daireler olarak gösterilir). Bu köşelerin “A” ve “B” gibi etiketleri vardır. Bu etiketler köşeler hakkında bilgi verir. Örneğin, bu çizgenin bir kimyasal molekül olduğu varsayıldığında A ve B etiketi, sırasıyla Hidrojen ve Oksijenin

kimyasal elementlerini temsil edebilir. Etiketlerin benzersiz olması gerekmez. Başka bir ifadeyle, aynı etiketler, aynı grafikteki birkaç köşeyi tanımlamak için kullanılabilir. Örneğin, yukarıdaki grafik bir kimyasal molekülü temsil ediyorsa, sırasıyla Oksijen ve Hidrojeni temsil eden tüm köşeler için "A" ve "B" etiketleri kullanılabilir.

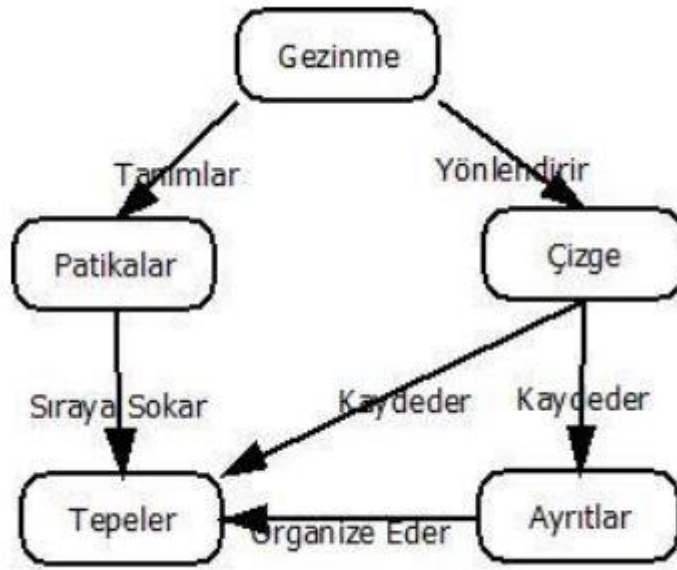
Bir çizge köşelerin yanı sıra kenarları da içerir. Kenarlar, burada kalın siyah çizgilerle temsil edilen köşeler arasındaki çizgilerdir. Kenarların da bazı etiketleri vardır. Şekil 3.1.'deki örnekte, X, W, Y ve Z olmak üzere dört etiket kullanılmıştır. Bu etiketler, köşeler arasındaki farklı ilişki türlerini temsil eder. Kenar etiketlerinin benzersiz olması gerekmez.

Bir çizge veri tabanı, tepe noktası, kenar ve fonksiyon gibi kavramlar arasındaki ilişkileri çizge teorisini kullanarak depolayan ve temsil eden bir veri tabanıdır (Şekil 3.2.). Çizge veri tabanları, yönlendirilmiş çizgeler biçimindeki veri modelini, verileri, kuralları ve kısıtlamaları içermektedir. Genellikle, çizge modelinde veri ve yapılar açık bir şekilde ayrılmış ve farklı katmanlarda temsil edilir. Veri işleme süreçleri, komşuluk, tepe yolları, alt çizgeler, bağımlılık ve çizge dolaşma gibi çizge teorisi yöntemleriyle gerçekleştirilir.



Şekil 3.2. Genel çizge veri tabanı modeli (Yıldırım, 2014)

Çizgede yer alan verileri sorgulamak için seçilen ilk tepe noktasından başlayarak belirli bir kurala göre diğer tepe noktaları arasında navigasyon uygulanmaktadır. Dolaşım sırasında tepe noktaları arasındaki ilişkileri gösteren kenarlar bir navigasyon yolu oluşturmaktadır.

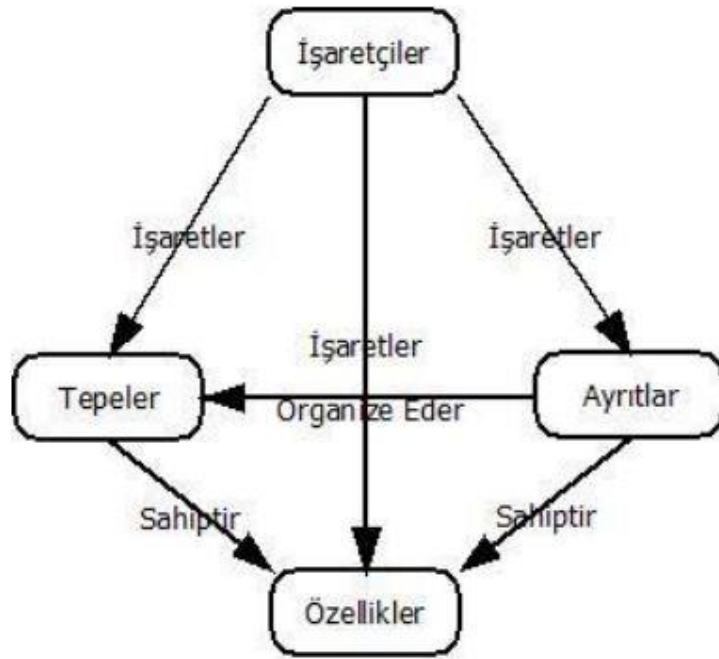


Şekil 3.3. Çizge gezinme (Yıldırım,2014)

Çizge veri tabanların verileri dinamik olarak depolamak için kullanılır. Çizge veri tabanı, verilerin bu yapıda depolandığı ve tüm düğüm ve kenarların aynı yönde özelliklere sahip olabildiği bir veri tabanı sistemidir. İlişkisel veri tabanlarından farklı olarak, çizge veri tabanlarında tablo yoktur. Bu şekilde, çizge veri tabanı, sorgu performansı üzerinde en büyük etkiye sahip olan tablodan tabloya geçiş işlemlerini ortadan kaldırır. Çizge veri tabanları, özellikle veri boyutu büyük olduğunda, ilişkisel veri tabanlarından çok daha hızlı çalışır (Karagöz ve Komesli, 2017).

Bir çizge veri tabanı, köşeler, kenarlar ve özellikler gibi kavramlar arasındaki ilişkileri çizge teorisini uygulayarak depolayan ve ifade eden bir veri tabanıdır. Çizge veri tabanı, yönlendirilmiş çizgeler olarak veri modelleri, veriler, kurallar ve kısıtlamalar içerir. Çizge modelleri ve veriler genellikle net bir şekilde ayrılır ve farklı katmanlarda temsil edilir (Şekil 3.3). Veri işleme işlemleri, komşuluklar, köşe yolları, alt çizgeler, bağımlılıklar gibi çizge teorisi teknikleri kullanılarak gerçekleştirilir (Yıldırım, 2014).

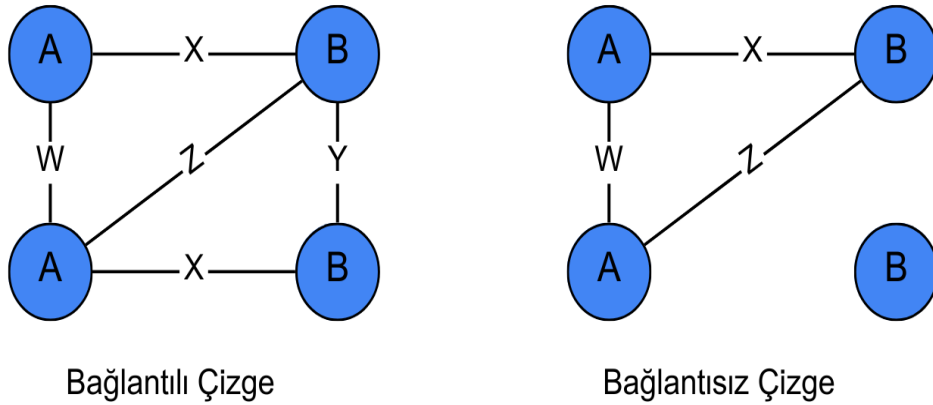
Bir sorgunun belirli bir zirveye ulaşması gerektiğinde, çizgede gezinmek yerine, zirveyi gösteren özel bir işaretçi ile indeksleme kullanılmaktadır (Şekil 3.4). Bu, üste doğrudan erişim sağlar ve navigasyona göre çok daha kısa sürede verilere erişim sağlamaktadır (Yıldırım, 2014).



Şekil 3.4. Çizge indeksleme (Yıldırım, 2014)

3.1. Bağlantılı ve Bağlantısız Çizge Veri Tabanı

Gerçek hayatta birçok farklı çizge türü bulunabilir. Aşağıdaki iki grafiği göz önünde bulunduralım:



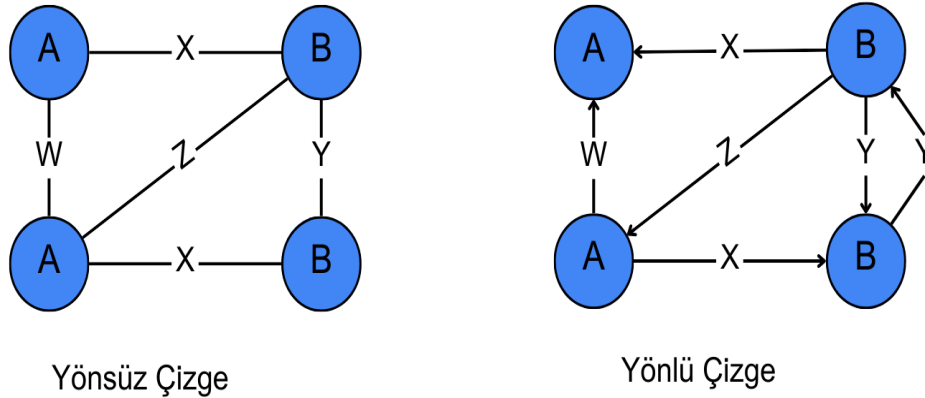
Şekil 3.5. Bağlantılı ve Bağlantısız Çizge Örneği

Şekil 3.5.'de yer alan 2 çizgeyi göz önünde bulundurduğumuzda, soldaki çizge bağlantılı çizge, sağdaki çizge ise bağlantısız çizge olarak adlandırılır. Bağlantılı çizge, kenarları takip ederek herhangi bir tepe noktasından diğer köşelere gitmek mümkün olan çizgedir.

Örneğin Şekil 3.5.'de bağlantılı çizgede, köşelerin şehirleri temsil ettiğini ve kenarların şehirlerarasındaki yollar olduğu düşünüldüğünde, yolları takip ederek herhangi bir şehirden diğer bir şehre gitmek mümkündür. Bir çizge bağlantılı değilse, bağlantısız çizge olduğu söylenir. Örneğin A düğüm kenarlar takip edilerek diğer köşelerden ulaşamadığı için Şekil 6' da sağdaki çizgenin bağlantısı kopmuştur.

3.2. Yönlü ve Yönsüz Çizge Veri Tabanı

Yönsüz bir çizge de kenarlar çift yönlü iken, yönlü bir çizge de kenarlar tek yönlü veya çift yönlü olabilir.



Şekil 3.6. Yönlü ve Yönsüz Çizge Örneği

Şekil 3.6.'da soldaki çizge yönsüz, sağdaki çizge ise yönlü çizgedir. Örneğin, köşelerin konumlar ve kenarların yol olduğu bir çizgeyi ele alalım. Bazı yollar çift yönlü olarak kullanılabilirken, bazı yollar tek yönlü olarak kullanılabilir (bir şehirde “tek yönlü” yollar). Bazı veri madenciliği algoritmaları, yalnızca yönsüz çizgelerle, yönlendirilmiş çizgelerle çalışacak veya her ikisini de destekleyecek şekilde tasarlanmıştır.

Çizge veri tabanları, sosyal ağ analizi, ağ yönetimi, yol bulma, öneri sistemleri ve biyoinformatik gibi birçok alanda yaygın olarak kullanılmaktadır. Bu veri tabanı modeli, düğümleri ve kenarları kullanarak verileri ilişkisel bağlamlarda organize ettiği için karmaşık yapıları daha iyi temsil edebilmektedir. Ayrıca, çizge veri tabanları; veri analitiği ve veri madenciliği gibi disiplinlerde önemli bir araçtır çünkü büyük ve karmaşık veri kümelerindeki desenleri, toplulukları ve yolları keşfetmek için kullanılabilir (Ke vd., 2015).

3.3. Çizge Veri Tabanı Uygulamaları

Çizge veri tabanı modeli, nesneler arasındaki bağımlılıkların önemli olduğu gerçek dünya uygulamalarında tercih edilmektedir. Eğer bu tür uygulamaları iki ana kategoriye ayırmak istenirse, geleneksel uygulamalar ve karmaşık ağ uygulamaları olarak sınıflandırılabilir.

- Veri karmaşıklığına ilişkin ilişkisel veritabanı modellerinin olasılıkları yapılması gereken uygulamalarda çizge veritabanı modeli tercih edilmektedir. Örneğin otobüs, tren, uçak ağları yönetildiğinde çizge veritabanları, uygulamalarda ilişkisel veritabanlarından daha verimlidir.

- Mevcut sorgulama dilleri, sorguları verimli bir şekilde yürütememektedir. Çizge veri tabanı tarafından sağlanan çizge arama algoritmalarına ihtiyaç vardır. Çizge veritabanları, duyuldukları uygulamalarda kullanılmaktadır.

- Görüntü işleme, çizge işleme, multimedya sistemleri ve çizge veritabanlarının bilimsel veri analiz yazılımlarında kullanılması uygundur.

Karmaşık ağ uygulamaları; (Yıldırım, 2014)

- İnsanların, grupların ve aralarındaki ilişkilerin tanımlandığı sosyal yaşam Ağlarda çizge veritabanlarının kullanılması büyük avantajlar sağlamaktadır.

Sosyal ağların ve bu ağlardaki verilerin analizine yönelik artan ihtiyaç, çizge veri tabanlarının önemini artırmaktadır.

- Çizge veritabanları, akademik yayınlar arasındaki referans ağlarının oluşturduğu bilgi ağlarını oluşturmak için kullanılmaktadır.

- Telefon ağıları, coğrafi ağ sistemleri, internet, elektrik ağıları gibi çizge veritabanları şu anda veri depolaması amacıyla çizge veritabanını kullanmaktadırlar.

- Genetik haritalama ve kimyasal ve metabolik yapıların modellenmesi gibi biyolojik ağ uygulamaları, çizge veritabanları için uygun alanlardır (Yıldırım, 2014).

3.4. Tek Bir Büyük Çizge

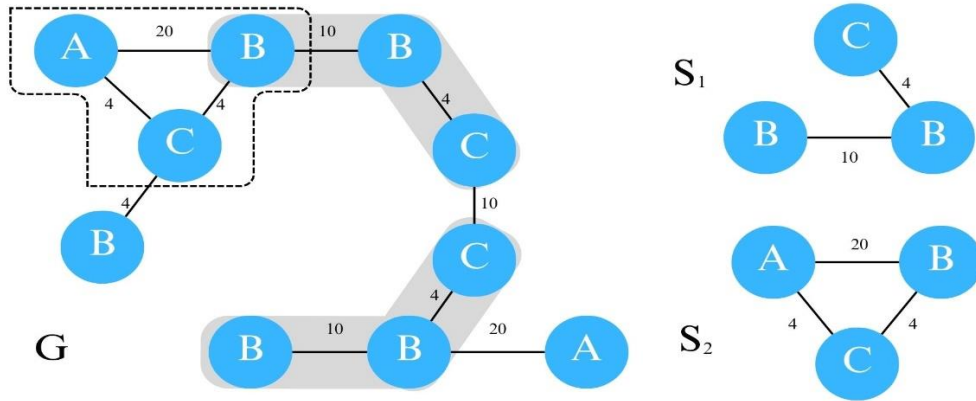
Tek büyük çizge, modern veri analitiği ve ağ analizinde giderek daha da önemli hale gelen bir kavramdır. Bu kavram, büyük ölçekteki veri kümelerini tek bir çizge yapısı altında birleştirme ve analiz etme fikrine dayanır. Tek büyük çizge, birçok farklı veri kaynağından elde edilen verilerin entegrasyonunu sağlar ve karmaşık ilişkileri görselleştirme ve keşfetme imkanı sunar.

Tek büyük çizge (single large graph), büyük ölçekli veri analitiği ve ağ analizi alanında önemli bir araç haline gelmiştir. Bu yaklaşım, çeşitli veri kaynaklarından gelen verileri tek bir çizge yapısı altında birleştirerek karmaşık ilişkileri ortaya çıkarmayı amaçlar. Tek büyük çizge, sosyal ağ analizi, web trafiği analizi, biyoenformatik, finansal ağ analizi gibi birçok alanda kullanılmaktadır. Büyük veri kümesini temsil eden tek bir çizge yapısı, veri keşfi, görselleştirme ve analiz için geniş bir potansiyel sunar (Wu vd., 2014).

Tek büyük çizge analizinde, karmaşık ağ özelliklerinin keşfedilmesi, alt çizgelerin tanımlanması, merkezlilik ölçümlerinin yapılması gibi birçok analitik işlem gerçekleştirilebilir. Bu sayede, büyük veri kümesinin içerisindeki önemli yapısal örüntüler ve ilişkiler ortaya çıkarılabilir. Ayrıca, tek büyük çizge üzerinde gerçek zamanlı sorgular, keşif ve görselleştirme işlemleri yapılabilir. Veri bilimi ve büyük veri analitiği alanında yeni fırsatlar sunan güçlü bir araçtır. Büyük veri setlerinin analizine olanak sağlayan bu yöntem, geniş bir uygulama alanına sahiptir ve bilgi keşfi, tahmin analizi, sosyal ağ analizi gibi birçok alanda değerli bilgilerin elde edilmesine katkı sağlar (Leskovec vd., 2014).

Çizgeler, kimyasal, biyoinformatik, bilgisayarlı görüntü işleme, sosyal ağlar, metin arama ve web analizi gibi çeşitli uygulamalarda nesneler arasındaki karmaşık ilişkileri modellemek için kullanılır. Sık alt çizge madenciliği, çizgelerde merkezi ve iyi çalışılmış bir problemdir ve çizge sınıflandırması, kullanıcı profillerinin modellenmesi, çizge kümeleme,

veri tabanı tasarımı ve indeks seçimi gibi birçok veri madenciliği görevinde kritik bir rol oynar. Sık alt çizge madenciliğinin amacı, görünümleri kullanıcı tanımlı bir eşik değerini aşan alt çizgeleri bulmaktır. Bu, birçok gerçek hayat uygulamasında faydalıdır. Örneğin, protein-protein etkileşimi (PPI) ağları. Bu ağlar, düğümlerin proteinleri temsil ettiği (ve işlevsellikleriyle etiketlendiği) ve kenarların bu proteinler arasındaki etkileşimleri temsil ettiği çizgelerdir. Bu tür çizgeler sürekli olarak yeni proteinleri ve etkileşimlerini içerecek şekilde güncellenir. Biyologlar için önemli bir görev, deneysel test yapmadan yeni bir proteinin işlevselliğini tahmin etmek (ve ilgili etiketi eklemek) tir. Yukarıdaki görev, yeni proteine benzer etkileşimlere sık alt çizge madenciliği yaparak doğru bir şekilde gerçekleştirilebilir.



Şekil 3.7. İş Birliği Çizgesi

Şekil 3.7.'nin iş birliği çizgesi G'si ve yazarlar arasında önemli iş birliklerini madencilik yapmak isteyen bir kullanıcıyı düşünelim. Genellikle, bu tür çizgelerde, sık görülen alt çizgeler, aynı çalışma alanına sahip yazarlar arasındaki iş birliklerini gösterme olasılığı daha yüksektir (yani, veri tabanı araştırmacıları arasındaki iş birlikleri). Daha ilginç alt çizgeleri ortaya çıkarmak için, kullanıcı disiplinler arası iş birliklerin bulunduğu alt çizgeleri keşfedene kadar frekans eşiğini aşama aşama azaltır (yani, A, B ve C araştırmacıları arasında). Frekans eşiğini düşürmek, nitelikli ara sonuçların sayısını artırır ve madencilik sürecinin zaten maliyetli hesaplamalarını yoğunlaştırır.

Bir çizge G 'de bir alt çizge S 'nin frekansını değerlendirmenin en basit yöntemi, G 'de S 'nin izomorfizmalarını aramaktır. İzomorfizmalar, S 'nin G 'deki düğümleri, kenarları ve etiketleriyle tam eşleşmelerdir. Örneğin, Şekil 5.7.'deki iş birliği çizgesi G 'de, alt çizge S_1 'in üç izomorfizması bulunmaktadır.

Tek bir çizgede sık alt çizge madenciliği yapmak için tipik bir yöntem, aşağıdaki adımlarla devam eden büyütme ve depolama yöntemidir:

1. En az τ (minsup) kez görünen tüm düğümleri bulun ve tüm görünümelerini depolayın.
2. Depolanan görünümeleri genişleterek daha büyük potansiyel sık alt çizgeler oluşturun, frekanslarını değerlendirin ve yeni sık alt çizgelerin tüm görünümelerini depolayın.
3. Sık alt çizgeler bulunamayana kadar Adım 2'yi tekrarlayın. SIGRAM gibi mevcut yaklaşımlar, bu büyütme ve depolama yönteminin değişikliklerini kullanır. Bu yaklaşımlar, alt çizgelerin frekansını değerlendirmek için depolanan görünümelerden yararlanır. Bu tür algoritmaların ana sınırlayıcı faktörü, her alt çizge için tüm görünümelerin oluşturulması ve depolanmasıdır. Bu görünümelerin sayısı, çizge ve alt çizgelerin boyutu ve özelliklerine bağlı olarak değişir; hesaplamak ve depolamak çok zorlu olabilir ve büyütme ve depolama çözümlerini pratikte gerçekleştirilemez hale getirebilir.

3.5. İzomorfizm

İzomorfizm, matematikte temel bir kavram olup farklı matematiksel yapılar arasındaki eşdeğerlik ilişkisini ifade eder. İki matematiksel yapı arasında izomorfizm varsa, bu durum bu yapılar arasında bir dönüşüm veya eşleme olduğunu gösterir ve bu dönüşüm, yapıların temel özelliklerini ve ilişkilerini korur. İzomorfizm, matematiksel yapıların sınıflandırılması, karşılaştırılması ve analizinde önemli bir rol oynar. Ayrıca, matematiksel nesneler arasındaki benzerlikleri ve simetriyi keşfetmemizi sağlar ve matematikteki derin bağlantıları ortaya çıkarır (Smith, Johnson, 2018).

Veri madenciliğinde, İzomorfizm, iki çizgenin veya ağın yapısının benzerliğini ifade eden bir kavramdır. İki çizge veya ağ arasındaki izomorfizm, bu yapılar arasında bir eşleme veya dönüşüm olduğunu gösterir, bu dönüşüm, düğümlerin ve kenarların yapısal özelliklerini

ve ilişkilerini korur. İki çizge veya ağ arasında izomorfizm varsa, bu durum, bu yapıların birbirine benzer davranışlar sergilediğini ve bir yapıyı diğerine dönüştürebileceğimizi ifade eder. İzomorfizm, veri madenciliğinde yapısal benzerlikleri belirlemek, çizge ve ağ analizinde kullanılan özellikleri karşılaştırmak ve veri keşfi süreçlerinde önemli bir rol oynamaktadır.

X ve Y matematiksel yapılar olsun. $F: X \rightarrow Y$ bir fonksiyon olsun. F, birbirine tekabül eden öğeleri koruyan bir birebir (1-1) ve sürekli bir eşleme (fonksiyon) ise, F'ye bir izomorfizm denir. Bu durumda, X ve Y izomorf matematiksel yapılar olarak kabul edilir ve F, X ve Y arasındaki izomorfizm ilişkisini temsil eder.

Yukarıdaki tanım, izomorfizmi bir fonksiyon aracılığıyla ifade eder. Fonksiyon, bir matematiksel yapının diğerine dönüştürülmesini sağlar ve bu dönüşüm, yapının temel özelliklerini ve ilişkilerini korur. İzomorf bir fonksiyon, birbirine tekabül eden öğelerin korunmasını sağlar, yani her X elemanı, F tarafından bir Y elemanına eşlenir ve her Y elemanı da F'nin ters fonksiyonu tarafından bir X elemanına eşlenir.

3.6. İlişkisel Veri Tabanı ve Çizge Veri Tabanı Farkları

Çizge veri tabanındaki yapı son derece basittir ve herhangi bir ilişki eklemek istediğinizde yeni düğüm ve uç ile kolayca tepki verebilmektedir. Ancak, ilişkisel veri tabanı şemasına herhangi bir yeni ilişkinin eklenmesi, yeni bir tablo veya tabloların eklenmesini gerektirmektedir. Neo4J'de SQL'deki gibi 'tarih' veri tipi olmadığı için gün ve saat bilgisi 'Unix Time' olarak saklanmaktadır. Unix saatindeki tüm tarihler saniyeye dönüştürülmektedir. Başlangıç tarihi 1 Ocak 1970'dir. Bu tarihten önceki tarihler için negatif sayılar kullanılmaktadır.

Çizge veri tabanı yapısındaki düğüm tipleri 'Kayıt', 'Yazılım Modülü' ve 'Kullanıcı' olarak ayarlanmıştır. Her bir yazılım modülünün hangi kayda ait olduğu bilgisi iki düğüm arasındaki ait ayırtı ile belirtilmiştir. Kullanıcının yazılım modülünü kullandığı tarih ve saat bir öznitelik olarak eklenmektedir. Dolayısıyla ilişkisel veri tabanı tasarımında olduğu gibi yeni bir tablo oluşturmaya gerek yoktur. Bu durumda sorgulama esnasında tüm bu tablolara ulaşabilmek için tabloları birleştirmek gerekmektedir. Bu mod hem yazılım hem de kodlama performansını indirmektedir.

BigLogVis programında bu veri yapılarının dışında gerekli özelliklerin elde edilebilmesi için üç katmanlı bir mimari kullanılmaktadır. Kullanıcı ara yüzü, programın ihtiyaç duyduğu işlemlerin yapıldığı katman ve dosya/veri tabanı işlemleri birbirinden tamamen bağımsızdır.

Ayrıca, ilişkisel bir veri tabanından çizge veri tabanına veri aktarımında performansı artırmak için çeşitli yaklaşımlar vardır. Sık kullanılan verilerin birbirine daha yakın kaydedilmesini öngören bu yaklaşım, performansı artıracaktır (Karagöz ve Komesli, 2017).



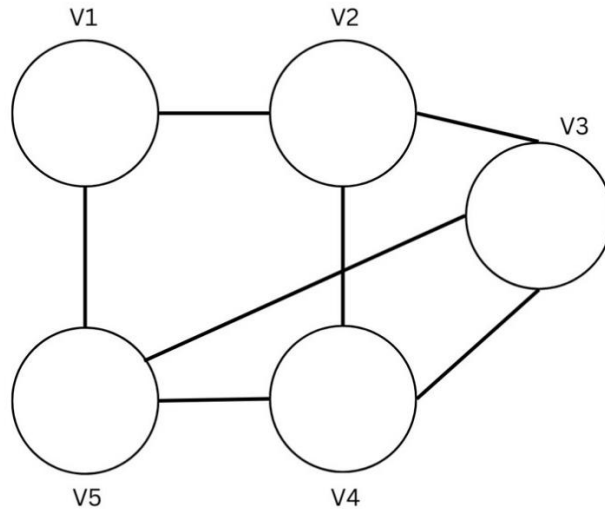
4. BULGULAR

4. ÇİZGE MADENCİLİĞİ

Çizge madenciliği, çizgeler aracılığıyla temsil edilen verilerden yeni bilgilerin çıkarılması sürecidir. Veri madenciliğinin bir alt dalı olan çizge madenciliği, veri madenciliği ve bilgi çıkarma alanında yeni bir alandır. Verilerin çizge tabanlı gösterimi, gerçek veri yapısına daha uygun olup, çözümlerde etkin bir şekilde kullanılmaktadır.

Çizge madenciliği, çeşitli araçlar ve teknikler kullanılarak çizgelere dönüştürülmüş verilerden yararlı bilgiler elde etme tekniğidir. Çizge, nesnelerin birleştirilmesiyle oluşturulan bir yapıdır. Her bir nesne bir düğümü (vertices & node) temsil eder. Düğümlerin arasındaki bağlantı araçları kenar (ayrıt) olarak adlandırılır. Çizge yapısının her bir kenarı iki bağımsız düğümü birbirine bağlar (Oğuz, 2016).

Şekil 4.1., V1, V2, V3, V4 ve V5 düğümlerinden oluşan ve (V1, V2), (V1, V5), (V2, V3), (V2, V4), (V3, V4), (V3, V5), (V4, V5) gibi 7 kenara sahip bir çizgeyi göstermektedir.



Şekil 4.1. 7 kenar ve 5 düğümden oluşan çizge (Oğuz, 2016)

Çizge madenciliği, çizge teorisi ve makine öğrenimi kavramlarını birleştirmektedir. Çizgeler, veri madenciliğinin çeşitli alanlarında etkili veri yapıları olarak kullanılmaktadır.

Çizge tabanlı veri madenciliği teknikleri, çizge yapısına sahip veri kümelerinden önemli kalıpları çıkarmayı hedeflemektedir. Özellikle son yıllarda, çizge madenciliği konusuna ilgi giderek artmaktadır. Bu artışın sebeplerinden biri, sosyal medyanın (örneğin, Facebook gibi platformlar) ve internetin yaygınlaşmasıyla birlikte yapılandırılmış veri yapılarına olan ihtiyacın artması ve araştırmacıların önemli bilgileri elde etme arzusudur. Ayrıca, tıbbi araştırmaların bilgisayar ortamına taşınması da bu ilginin artmasında etkili olmuştur. Yapılandırılmış veri kümelerinde, önemli ölçüde tekrar eden altyapıların kaldırılması; bu verileri yorumlamanıza ve genelleştirmenize olanak tanımaktadır. Öte yandan, bir bilgi alanında çalışan bir kullanıcı, önceden tanımlanmış bir türde veya bu alana özgü altyapılar aramaktadır.

Çizge yapıları, ağaçlardan daha kapsamlı bir şekilde betimleme yeteneğine sahiptir ve matematiksel olarak sağlamdır. Bu nedenle, çizge algoritmaları gürültülü verileri kümelemede daha başarılı bir şekilde kullanılabilir.

Çizge madenciliği, çeşitli alanlarda kullanılabilmektedir. Örneğin market satışları (raf yerleştirme görevi), promosyon kuralları, metin arama, video indeksleme, kimyasal alanlar, ilaç etkileşimi, görüntü işleme ve kablosuz sensör ağları gibi. Örneğin, çizge madenciliği kimyasal yapılar üzerinde kullanıldığında, düğümler ayrı atom türlerini temsil ederken, kenarlar atomik bağları gösterir ve elektronik işlemlerde kullanılabilir. Müşterileri ayırtmak için tepeler kullanıldığında, o müşterinin ödeme yöntemlerini belirtir. İlaç etkileşimlerinde, alt çizgeler kapsamda bulunur ve herhangi bir topolojiye sahip olabilir, ancak küme tespit algoritmaları tarafından tespit edilemez. Özellikle sık alt çizge analizi, verimli bir şekilde bilgi sağlamaktadır. Örneğin, çizge madenciliği ortak protein yapılarını bulma, nesne tanımda ortak örüntüler ve dolandırıcılık tespiti gibi önemli alanlarda uygulamaları vardır.

Çizge madenciliğinin birçok uygulamasında araştırmacılar araştırmacılar aşağıdaki gibi birçok probleme odaklanmakta ve çeşitli yaklaşımlar önermektedir. Sorunlar benzer görünse de farklı çözümler gerektirir. Bu problemler;

- Birçok küçük çizge kümesinden sık tekrarlanan alt çizgelerin aranması
- Daha büyük bir çizgeden sık tekrarlanan alt çizgelerin aranması

- Bir çizgedeki iki spesifik tepe noktası arasındaki önemli yollar çıkarılması.
- Çizgedeki en fazla tepe noktasıyla ilişkili alt çizgeyi bulmak

• Farklı zaman aralıklarında topluluk veya nesneler etkileşim sonucunda elde edilen sıralı çizgelerden oluşan bir veri setinden, belirli zaman aralıklarında nasıl davranacağı bulunmalıdır. Veri setindeki sıralı çizgelerin etkileşim sonucunda belirli aralıklarda nasıl davranacağını bulmak

- Çizgelerin kümelenmesi/sınıflandırılması
- Çizgelerde benzerliklerin aranması
- Çizge indeksleme
- Çizge senkronizasyon sorunu

4.1. Sık Alt Çizge Madenciliği

Sık alt çizge madenciliği (frequent subgraph mining), büyük ölçekli çizge verilerinde tekrar eden alt çizge desenlerini belirlemek için kullanılan bir veri madenciliği tekniğidir. Bu yöntem, sosyal ağ analizi, kimyasal bileşik keşfi ve biyoinformatik gibi birçok alanda önemli bir role sahiptir. Sık alt çizge madenciliği, bir çizge içinde belirli bir frekans eşliğini geçen alt çizge desenlerini tespit etmek için özel algoritmalar kullanır. Bu algoritmalar genellikle genişleme-tabanlı (expansion-based) veya düşme-tabanlı (prune-based) stratejileri kullanarak sık alt çizgeleri keşfeder. Sık alt çizge madenciliği, büyük veri setlerindeki yapısal bilgileri analiz ederek, ağların yapısını ve etkileşimlerini anlamamıza yardımcı olur. Bu, sosyal ağlardaki topluluk tespiti, protein-protein etkileşim ağlarındaki önemli alt yapıların tanımlanması gibi alanlarda değerli bilgiler sağlar.

Sık alt çizge madenciliği, büyük veri kümelerinde tekrarlayan desenleri tespit etmek için kullanılan bir veri madenciliği tekniğidir. Bu teknik, çizge veri yapıları üzerinde çalışır ve veri kümesinde sık olarak tekrar eden alt çizgeleri tespit etmeyi amaçlar. Sık alt çizge madenciliği, genellikle kimya, biyoinformatik, sosyal ağ analizi ve moleküler yapı analizi gibi alanlarda uygulanır. Bu yöntem, alt çizgelerin frekansını hesaplamak için genişleyen alt çizge

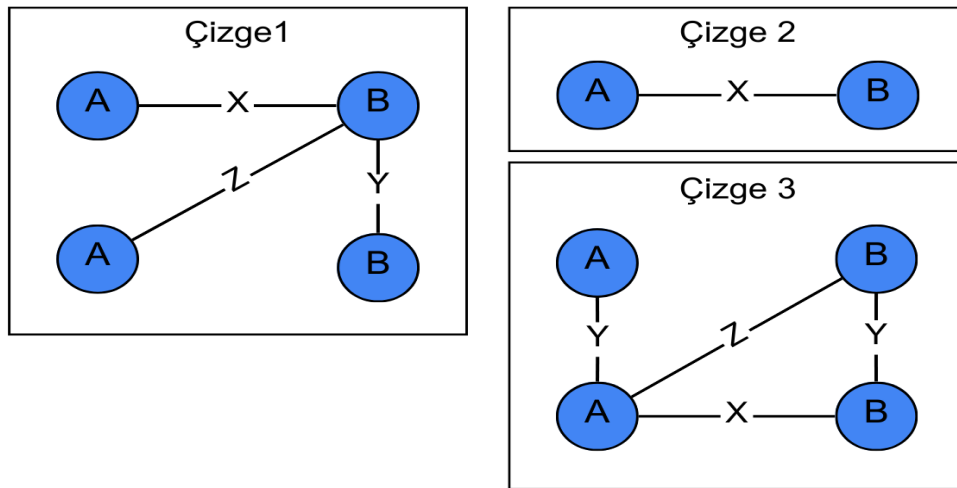
arama yöntemini kullanır. Bu sayede, veri kümesinde sık olarak tekrar eden yapısal motifler ve ilişkiler keşfedilebilir. Sık alt çizge madenciliği, veri tabanlarının analizinde ve çizge verilerindeki önemli kalıpları belirlemede kullanılan etkili bir araçtır (Kuramochi ve Karypis, 2004).

Sık alt çizge madenciliğinin amacı, bir dizi çizgede (bir çizge veri tabanı) görünen ilginç alt çizgeleri keşfetmektir. Bu ilginçlik çeşitli şekillerde tanımlanabilir. Geleneksel olarak, bir sık alt çizge, bir dizi çizgede birden çok kez görünüyorsa ilginç olarak kabul edilir. Başka bir deyişle, çoklu çizgelerde ortak olan sık alt çizgeleri keşfetme sürecidir. Örneğin birkaç kimyasal molekülde ortak olan kimyasal elementler arasındaki ilişkiyi bulmak için kullanılabilir.

Bir dizi çizgede sık alt çizgeleri bulma görevi, sık alt çizge madenciliği olarak adlandırılır. Girdi olarak kullanıcı şunları sağlamalıdır:

- bir çizge veri tabanı (bir dizi çizge)
- minimum destek eşiği (minsup) adı verilen bir parametre.

Ardından, bir çizge madenciliği algoritması, tüm sık alt çizgeleri çıktı olarak sıralayacaktır. Örneğin, üç çizge içeren Şekil 4.2. deki çizge veri tabanını ele alalım.



Şekil 4.2. Üç Çizge Örneği

Şekil 4.2’de, en az üç çizgede görünen tüm alt çizgeleri keşfetmek istediğimizi varsayalım. Böylece minsup parametresini 3 olarak ayarlayacağız. Sık alt çizge madenciliği algoritması uygulayarak en az üç çizgede görünen tüm alt çizgelerin kümesini elde edeceğiz (Şekil 4.3).

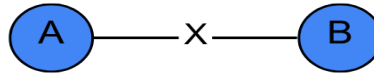
Sık alt çizge 1 :



Sık alt çizge 2 :

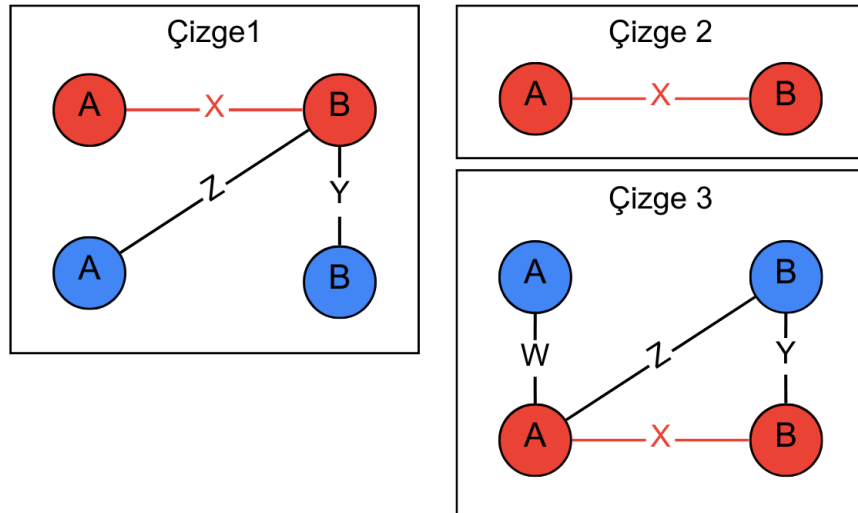


Sık alt çizge 3 :



Şekil 4.3. Sık alt çizge kümesi

Üçüncü alt çizgeyi dikkate alalım ("Sık kullanılan alt bölüm 3"). Bu alt çizgeler sıktır ve girdi çizgelerin üçünde görüldüğü için 3 desteğine (frekansa) sahip olduğu söylenir. Bu oluşumlar aşağıda kırmızıyla Şekil 4.4.’ de vurgulanmıştır:



Şekil 4.4. Sık Alt Çizge Örneği

Pratikte minsup parametresi genellikle deneme yanılma yoluyla belirlenir. Bu parametre çok yüksek ayarlanırsa, birkaç alt çizge bulunurken, çok düşük ayarlanırsa, girdi veri tabanına bağlı olarak yüz milyonlarca alt çizge bulunabilir. Sık alt çizgeleri bulmak için çeşitli sık alt grafik madenciliği algoritmaları mevcuttur. Başlıcaları; GASTON, FSG ve GSPAN' dır.

Alt çizge destek seviyesi (Subgraph support level), G çizgesi için D giriş veri setiyle ilgili olarak destek (GD) olarak adlandırılır ve (3.)' deki formülle hesaplanır.

$$Destek(GS) = \frac{D \text{ içerisindeki } GS \text{ destek seviyesini sağlayan toplam çizge sayısı}}{D \text{ içerisindeki toplam çizge sayısı}} \quad (3.)$$

Çizge hareketleri (Graph motifs), en az destek düzeyi GS olan V veri kümesinde bu en düşük desteği sağlayan tüm alt çizgelerin oluşturduğu çizgelerin listesine verilen isimdir (Oğuz, 2016).

4.1.1. Sık Alt Çizge Madenciliği Algoritmalarının Kullanım Alanları

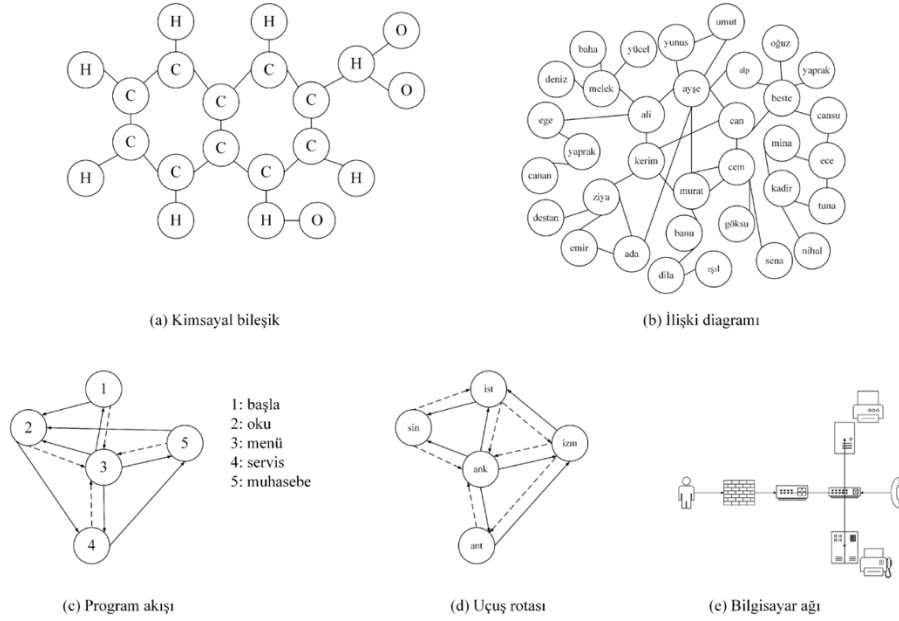
Sık alt çizge madenciliği, birçok uygulama alanında önemli bir veri madenciliği tekniği olarak kullanılmaktadır. Bu teknik, büyük veri kümeleri içinde tekrar eden alt çizgeleri tanımlamak ve keşfetmek için kullanılır. Özellikle sosyal ağ analizi, biyoinformatik, kimya, finans ve pazarlama gibi alanlarda sık alt çizge madenciliği yoğun bir şekilde uygulanmaktadır. Sosyal ağ analizinde, sık alt çizgeler, benzer etkileşim modellerini veya grup yapılarını tespit etmek için kullanılırken, biyoinformatikte protein etkileşim ağları ve genetik verilerde tekrar eden motiflerin keşfedilmesinde kullanılmaktadır. Kimyada, moleküler yapılar ve kimyasal reaksiyonlar arasında tekrar eden yapısal motiflerin belirlenmesinde sık alt çizge madenciliği önemli bir rol oynamaktadır. Finans ve pazarlama alanlarında ise müşteri davranış modellerinin analizi ve benzer pazarlama kampanyalarının keşfi için sık alt çizge madenciliği teknikleri kullanılmaktadır (Johnson ve Smith, 2021).

Özellikle karmaşık biyolojik verilerin analiz edilmesi ve derlenmesine odaklanan Biyoinformatik alanında yer almakta olan veriler kolaylıkla çizge veri tabanı ile ifade edilebilmektedir. Örneğin, Shen ve arkadaşları yaptıkları çalışmada yaptıkları çalışmada 5 farklı protein- protein interaction (PPI) veri tabanını bir araya getirip veri tabanı üzerinde sık alt çizge madenciliği gerçekleştirmiş ve kanser modüllerinin sık alt çizgelerle ortaya

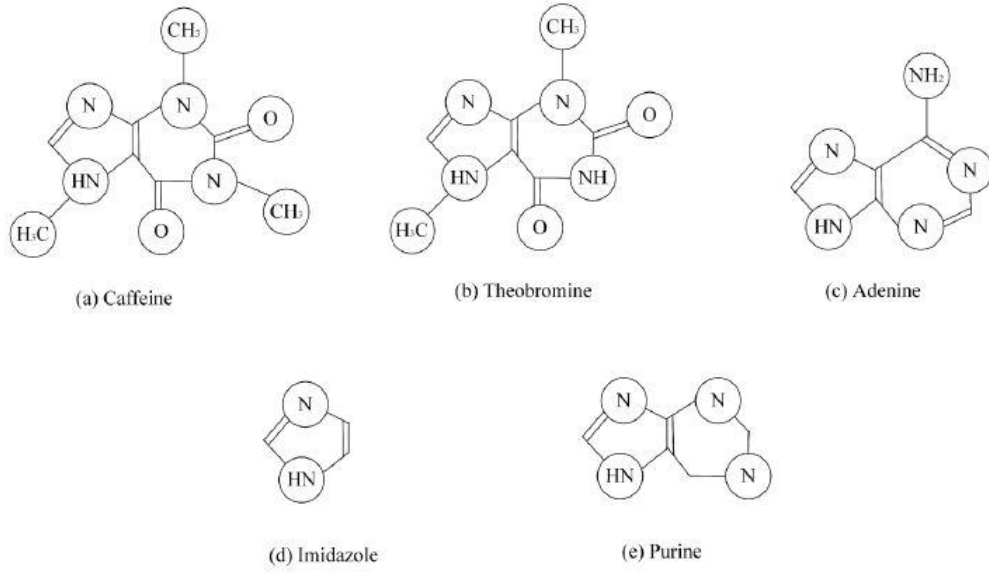
çıkartılmasının mümkün olduğunu göstermiştir (Shen vd., 2012). Bir başka örnek olarak, düğümler amino asitleri, kenarlar ise amino asitler arasındaki etkileşimi belirtebilir. Bu tür bir çizge veri tabanında kendini tekrar eden ya da sık olan alt çizgeler, analiz edilen veriden proteinlerin fonksiyonlarının ortaya çıkarılması sağlanabilmektedir.

Şekil 4.5.'de Sık alt çizge madenciliğinin farklı alanlarda kullanımı, Şekil 4.6.'da kimyasal bileşikler arasındaki sık alt çizgeler gösterilmektedir. Şekil 4.7.' de ise sosyal ağlarda sık alt çizge madenciliğinin kullanım örneği verilmiştir.

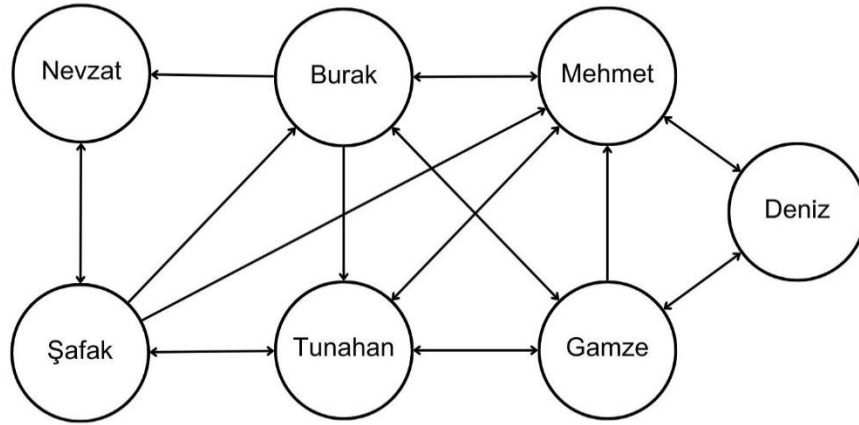
Şekil 4.6.' da (a), (b), (c) bileşikleri içerisindeki ortak bileşiklerin (d) ve (e) olduğu gösterilmektedir.



Şekil 4.5. Sık alt çizge yapılarının örnek kullanım alanları (Oğuz, 2016)



Şekil 4.6. Bileşik içerisinde ortak bileşik gösterimi (Oğuz, 2016)



Şekil 4.7. Sosyal ağ örneği

4.1.2. gSpan Algoritması

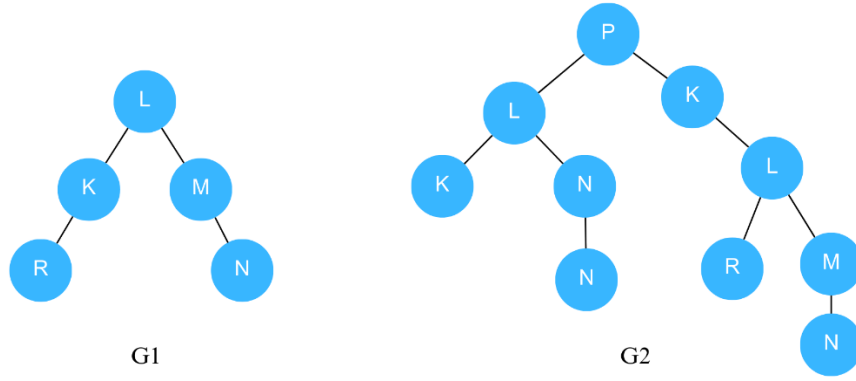
gSpan, veri madenciliğinde sık alt çizgeleri (frequent subgraphs) bulmak için kullanılan etkili bir algoritmadır. Sık alt çizgeler, veri kümesinde sık sık görülen yapılardır ve birçok uygulama alanında önemli bilgiler sağlar. gSpan, büyük ve karmaşık veri kümesi içindeki sık alt çizgeleri keşfetmek için çizge tabanlı bir yaklaşım kullanır. gSpan algoritması,

veri kümesindeki tüm alt çizge adaylarını oluşturur ve bu adayları frekans tabanlı olarak filtreleyerek sık alt çizgeleri belirler. Algoritma, derinlik-öncelikli arama (depth-first search) stratejisini kullanarak alt çizge adaylarını genişletir ve veri kümesi üzerinde dolaşır. Bu sayede, büyük veri kümelerinde dahi etkili bir şekilde sık alt çizgeleri tespit edebilir. gSpan, veri madenciliği alanında sık alt çizge madenciliği problemini çözmek için önde gelen bir yöntemdir. Bu yöntem, moleküler kimyada, biyoenformatikte, sosyal ağ analizinde ve yazılım analizinde kullanılarak değerli bilgilerin ortaya çıkarılmasını sağlar. gSpan algoritması, veri kümesindeki yapısal desenleri tanımlama, benzerlik analizi ve örüntü tanıma gibi birçok uygulama için önemli bir araçtır (Yan ve Han, 2002).

gSpan, sık alt çizge madenciliği için önemli bir çizge temelli algoritmadır. Bu algoritma, veri kümesindeki yapısal desenleri keşfetmek ve frekanslarını belirlemek amacıyla kullanılır. gSpan'in temel özelliği, veri kümesindeki büyük ve karmaşık alt çizgeleri etkili bir şekilde tespit edebilme yeteneğidir. gSpan, alt çizge ağacı yapısını kullanarak alt çizge adaylarını keşfeder. Algoritma, bir başlangıç alt çizgesiyle başlar ve derinlik-öncelikli arama (DFS) stratejisini kullanarak alt çizge adaylarını genişletir. Bununla birlikte, adayları filtrelemek için frekans tabanlı bir yöntem kullanılır, böylece sadece sık alt çizgelerin keşfedilmesi sağlanır. gSpan algoritması, çeşitli uygulama alanlarında etkili bir şekilde kullanılmıştır. Biyoinformatikte, protein etkileşim ağlarındaki motiflerin tespiti ve moleküler yapılarıdaki sık desenlerin belirlenmesi gibi alanlarda önemli bir rol oynamaktadır. Ayrıca, sosyal ağ analizi, kimyasal bileşiklerin keşfi ve veri tabanı analizi gibi birçok veri madenciliği görevinde de kullanılmıştır. gSpan, sık alt çizge madenciliği için etkili ve güçlü bir algoritmadır. Çizge tabanlı yaklaşımı sayesinde büyük veri kümesindeki yapısal desenleri keşfetme yeteneğine sahiptir. Bu da veri analitiği ve keşif çalışmalarında değerli bir araç olarak kullanılmasını sağlar (Liao vd., 2019).

gSpan algoritması, Şekil 4.8.' de ki örnek çizgeler üzerinde açıklanırsa:

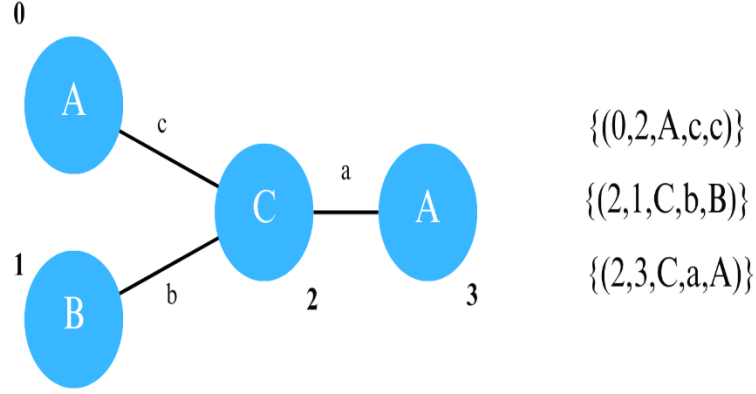
- Her bir tepeye ait frekans değerleri: K-3, L-3, M-3, N-3, P-1, R-2 şeklindedir.
- Sık kullanılmayan tepeler, tepeler listesinden çıkartılır. Buna göre, E tepesi silinmektedir.



Şekil 4.8. gSpan örnek uygulanması için G1 ve G2 çizgesi.

- Listede kalan tepeler: K, L, M, N ve R'dir.
- K tepesi incelenir. R tepesi sadece bir çizgede bulunduğu için alınmamaktadır.
- L tepesi incelendiğinde ise; K ve M tepelerinin her ikisi de sık geçtiği için sık kullanılan tepe listesine K, L ve M tepeleri alınmaktadır.
- Bu işlem tüm tepeler inceleninceye kadar devam eder.
- Tüm tepeler bittiğinde; sık alt çizge olarak K, L, M ve N tepeleri bulunur.

gSpan algoritması, derinlik öncelikli arama stratejisine dayalı olarak çalışır. Bu algoritma, gereksiz adayların üretilmesini ve izomorfizm testlerini önlemek için derinlik öncelikli arama kodu (DFS kodu) ağacı oluşturur. Aynı kodlara sahip etiketli arama alanından çıkarılır. DFS kodları, preorder (önce kök) tepe etiketlemesine göre oluşturulur. Kodun temsil şekli şu şekildedir: (Kök_tepe, Hedef_tepe, kök_label, ayır_tepe, hedef_label).



Şekil 4.9. DFS kod örneği

DFS (Derinlik Öncelikli Arama), çizge teorisi ve algoritma analizinde sıkça kullanılan bir arama algoritmasıdır. DFS, bir çizge yapısı içerisindeki düğümleri derinliğe göre keşfetmek için kullanılır. DFS, yığına (stack) dayalı bir arama algoritmasıdır.

Temel fikir, başlangıç düğümünden başlayarak bir komşu düğüme gidene kadar mümkün olan en derin seviyeye kadar inmektir. Bu süreç, başlangıç düğümüne komşu olan tüm düğümleri keşettikten sonra geri dönerek diğer komşu düğümleri keşfetmeye devam eder. Bu şekilde, çizgenin derinliklerine doğru ilerler ve keşfedilen düğümleri bir yığın veri yapısında saklar. Şekil 4.9.' da bir örneği verilmiştir.

DFS'nin en yaygın kullanım alanlarından biri, bir çizge yapısındaki yol problemlerini çözmektir. Ayrıca, çizge yapısındaki bağlantı veya düğümleri bulmak için de kullanılabilir. DFS algoritması ayrıca topolojik sıralama, labirent bulmacaları ve yapay zeka alanında da birçok problemde kullanılabilir.

Örüntü bulma yöntemleri, DFS kodlarından örüntüleri çıkarmak için dizi tabanlı bir yaklaşım kullanır. Bu süreçte, DFS ağaçları kullanılır ve her bir düğüm adayı temsil eder. Adaylar, DFS kodlarına göre sırasıyla üretilir ve asgari destek için test edilir. Asgari destek, algoritmanın girdisi olarak belirtilen bir destek değeridir. Bir adayın destek değerinden büyük veya eşit sayıda kesişmeyen gömülü düğüme sahip olması, adayın desteklenen bir aday olduğunu gösterir. Yeterli desteği olmayan bir aday ve onun DFS kod ağacındaki alt düğümleri göz ardı edilir. Benzer şekilde, bir adayın DFS kodu daha önceden üretilmişse,

yani ağaçta zaten yer alıyorsa, bu adayın kendisi ve alt düğümleri ileriki aramalarda göz ardı edilir. Bu iki budama tekniği, gSpan algoritmasını rakiplerine göre daha verimli hale getirir ve çeşitli ağlara daha iyi uyarlanabilir (Worlein vd., 2005).

4.1.2. GraMi Algoritması

GraMi (Graph Mining) algoritması, çizge veri yapıları üzerinde desenlerin keşfedilmesi ve çıkarılması için kullanılan bir veri madenciliği yöntemidir. Bu algoritma, çizge veri setlerindeki yapısal özellikleri ve ilişkileri analiz ederek anlamlı desenleri ortaya çıkarmayı hedefler. GraMi algoritması, bir çizge veri setindeki alt çizge desenlerini keşfetmek için çalışır. Bu desenler, çizge yapısındaki düğümlerin ve kenarların belirli kombinasyonlarından oluşur. GraMi, alt çizge desenlerinin frekanslarını ve diğer istatistiksel özelliklerini hesaplayarak önemli ve tekrar eden desenleri belirler.

Algoritma 1: GraMi Algoritması

```
1 Input: A graph  $G$  and frequency of threshold  $t$ 
2 Output: All subgraphs  $S$  of  $G$  such that  $s_G(S) \geq t$ 
3 result  $\leftarrow \emptyset$  ;
4 let  $fEdges$  be set of all frequent edges of  $G$  ;
5 for each edges  $e$  in  $fEdges$  do
6 result  $\leftarrow$  result  $\cup$  SubgraphExtension( $e, G, t, fEdges$ );
7 remove  $e$  from  $G$  and  $fEdges$ ;
8 return result;
```

Algoritma 2: SubgraphExtension Algoritması

```
1 Input: A subgraph  $S$  of a graph data  $G$ , the frequency threshold  $t$  and the
set of frequent edges  $fEdges$  of  $G$ 
```

2 Output: All frequent subgraphs of G that extend S

3 $\text{result} \leftarrow S, \text{candidateSet} \leftarrow \emptyset$;

4 **for** each edges e in $fEdges$ and node u of S **do**

5 **if** e can be used extension u **then**

6 Let ext be the extension of S with e ;

7 **if** ext is not already generated **then**

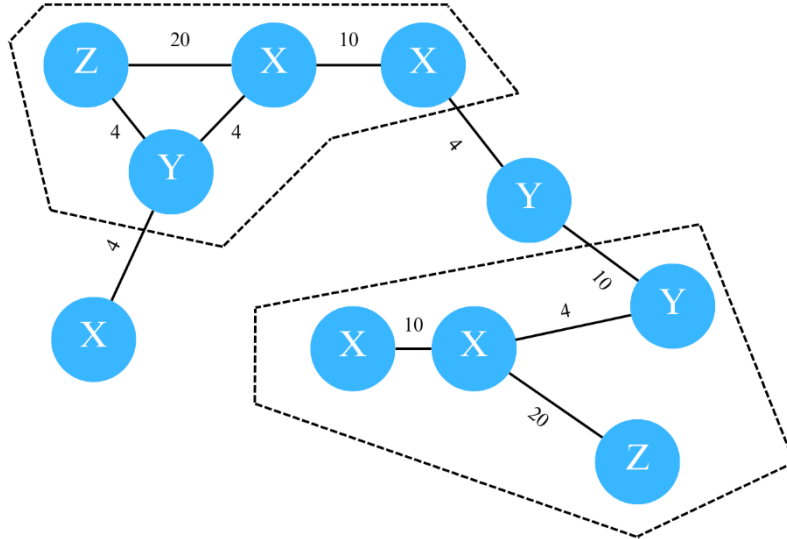
8 $\text{candidateSet} \leftarrow \text{candidateSet} \cup \text{ext}$;

9 **for** each $c \in \text{candidateSet}$ **do**

10 **if** $s_G(c) \geq t$ **then**

11 $\text{result} \leftarrow \cup \text{SubgraphExtension}(e, G, t, fEdges)$;

12 **return** result ;

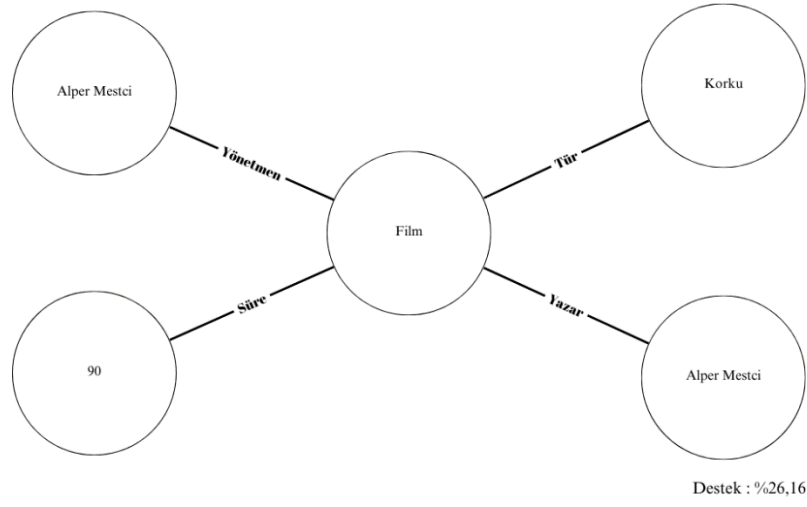


Şekil 4.10. GraMi Örnek Uygulaması için G çizgesi.

- Alt çizgelerin frekansları: $Y-X(4)$, $Y-Z(1)$, $Z-X(2)$, $X-X(2)$, $Y-Y(1)$ hesaplanır.
- Her bir alt çizgelerin tek tek genişletilip genişletilemediği incelenir.
- $Y-X$ çizgesi için, $X-X$ ayrıtı dahil edilerek alt çizge parçası genişletilmeye çalışılır. Örnek çizge üzerinde, $Y-X-X$ yapısı sık geçmediği için genişletilme işlemi yapılamaz ve $A-A$ çizgesi olası frekans listesine eklenmez.
- $Y-X$ çizgesi için, $X-Z$ çizgesi eklenerek genişletilmeye çalışıldığında ise, çizge içerisinde $Y-X-Z$ yapısı sık olarak geçtiği görülmektedir ve frekans listesine eklenir.
- Yeni çizge eklenerek genişletme işlemi tüm çizgeler için yapılır.
- Tüm çizgeler incelendiğinde: $\{X-Y, Z-Z, X-X\}$ çizge kümesi sık kullanılan alt çizge yapısını vermektedir.

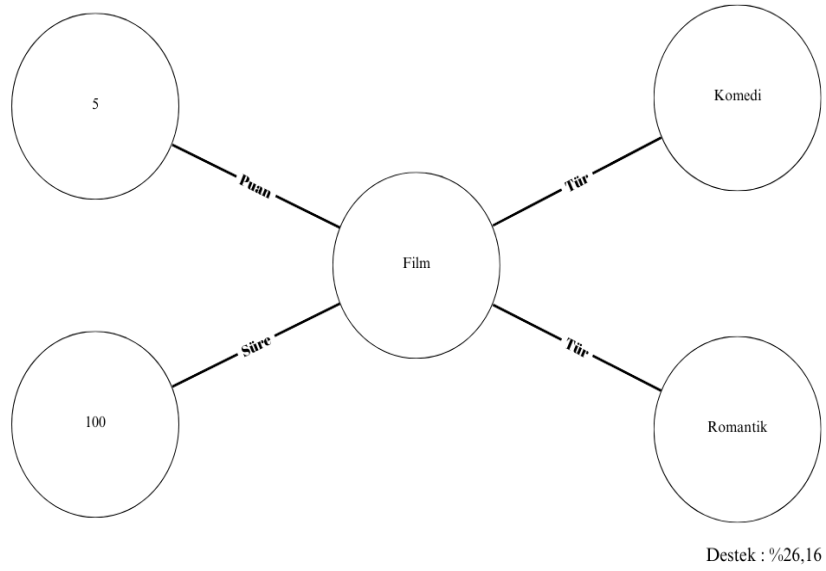
4.2. Uygulama

Örnek vaka uygulaması için imdb internet sitesinden (www.imdb.com) 2000 – 2022 yılları arasında vizyona girmiş yerli filmler (Türk filmleri) ve bu filmlerin özellikleri çekilmiştir. Film özellikleri; film adı, film yılı, film süresi, film türü, film yönetmeni, film yazarı, film oyuncular, film imdb puanı şeklindedir. Toplam 326 film, 210 farklı yönetmen, 680 farklı oyuncu, 285 farklı yazar, 13 farklı türe ulaşılmıştır. Imdb veri tabanı hem işlemsel hem de tek bir büyük çizge olarak ayrı ayrı çizge veri tabanına dönüştürülmüştür. Oluşturulan işlemsel veri tabanı üzerinde GSPAN (referans) algoritması çalıştırılmış ve sık alt çizgeler ortaya çıkarılmıştır.



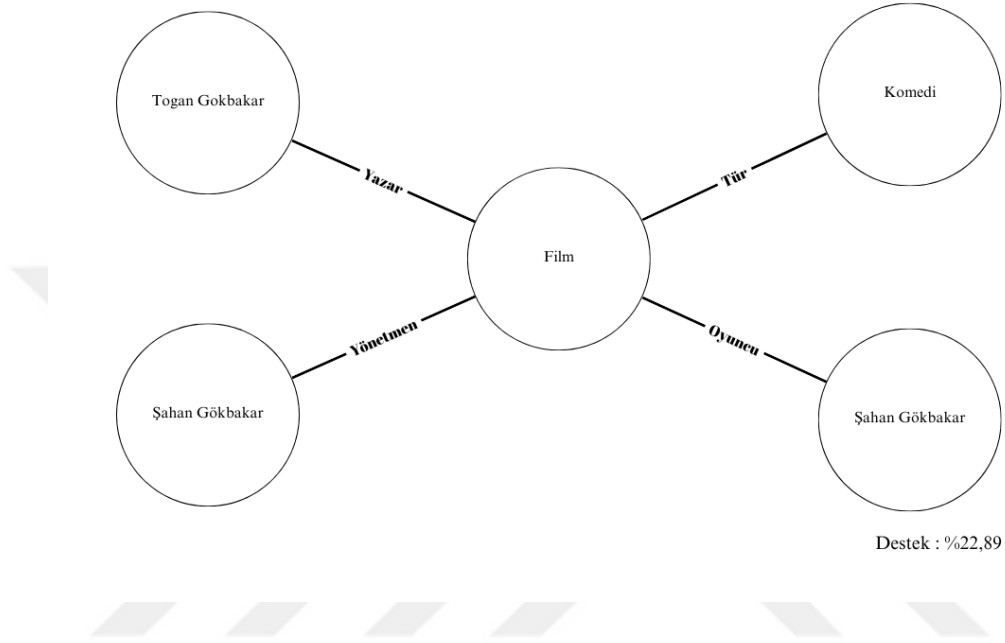
Şekil 4.11. Sonuç 1

Şekil 4.11.'deki sonuçta filmler incelendiğinde, bu filmlerin %26,16' sının yönetmen ve yazarı Alper Mestci, türü Korku ve film süresi 90 dakika olarak saptanmıştır.



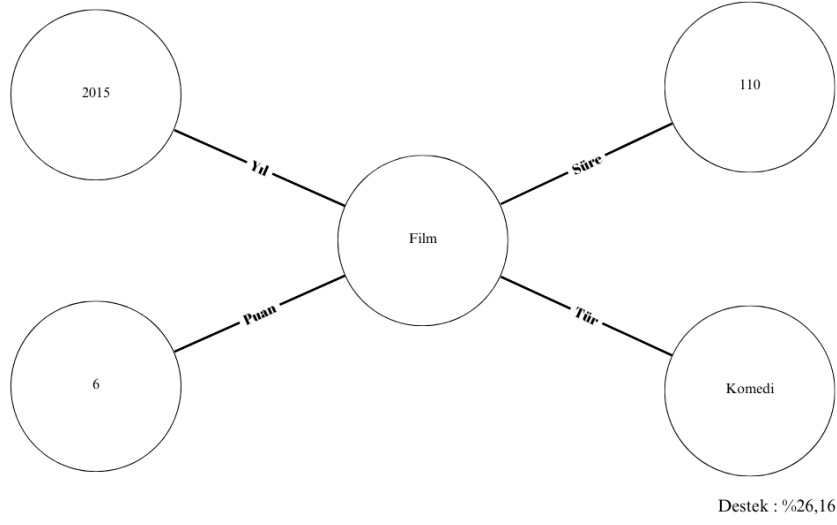
Şekil 4.12. Sonuç 2

Şekil 4.12.' deki sonuçta filmler incelendiğinde, bu filmlerin %26,16' sının, film süresi 100 dakika, imdb puanı 5 olan romantik ve komedi türünde olduğu saptanmıştır.



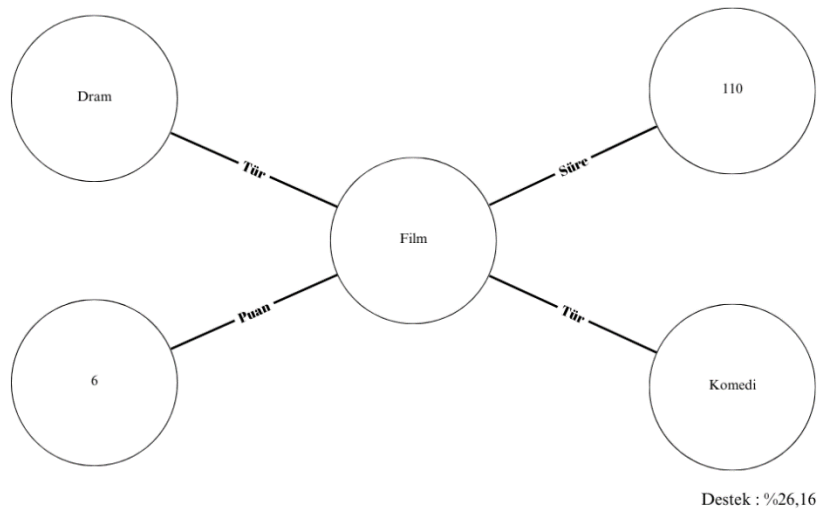
Şekil 4.13. Sonuç 3

Şekil 4.13' deki sonuçta filmler incelendiğinde, bu filmlerin %22,89' unun, oyuncusu ve yönetmeni Şahan Gökbakar, yazarı Togan Gökbakar türü ise komedi olarak saptanmıştır.



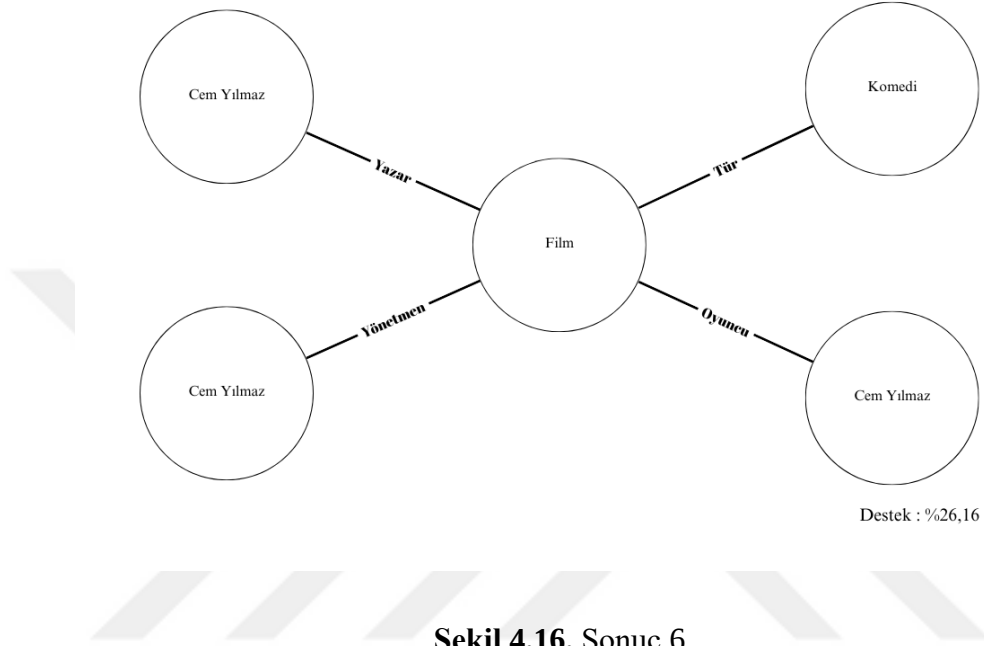
Şekil 4.14. Sonuç 4

Şekil 4.14.' deki sonuçta filmler incelendiğinde, bu filmlerin %26,16' sının çekim yılının 2015 yılı, süresinin 110 dakika, türünün komedi ve imdb puanının da 6 olduğu saptanmıştır.



Şekil 4.15. Sonuç 5

Şekil 4.15.' deki filmler incelendiğinde, bu filmlerin %26,16' sının, film süresi 110 dakika, imdb puanı 6 olan dram ve komedi türünde olduğu saptanmıştır.



Şekil 4.16.' daki filmler incelendiğinde, bu filmlerin %26,16' sının, oyuncusu, yönetmeni ve yazarı Cem Yılmaz, türü ise komedi olarak saptanmıştır.

5. TARTIŞMA VE SONUÇ

Bu tez çalışmasında, sık alt çizge madenciliği algoritmalarının kullanım alanları ve uygulanabilirliği üzerine bir inceleme yapılmıştır. Çalışma, sık alt çizge madenciliği konusunda genel bir bakış sunmuş ve çeşitli algoritmaların kullanım potansiyelini değerlendirmiştir. Araştırmacıların ve kurumların sık alt çizge madenciliğini nasıl ortaya çıkarabildiği, sık alt çizgelerin hangi yöntem ve teknikler ile ortaya çıkarılabileceği, ortaya çıkartılan sık alt çizgelerin kurumların ve araştırmacıların faydasına nasıl sunulabileceği araştırılmıştır. Bunun yanında literatürde yer alan işlemsel çizge veri tabanı için önceden geliştirilmiş güncel sık alt çizge algoritmasına ara yüz tasarlanmış ve sık alt çizge madencilik işlemi kolay bir hale getirilmiştir. Gündelikte yer almakta olan erişilebilir ve paylaşımına açık çizgesel olmayan herhangi bir veri tabanı(imdb.com) çizgesel veri tabanına dönüştürülüp yeni bir veri tabanı yaratılmıştır. Oluşturulan bu yeni veri tabanı üzerinde sık alt çizge madencilik işlemi gerçekleştirilmiş, ortaya çıkartılan sık alt çizgelerin anlamsallığı ve değerleri gösterilmiştir. Çok yüksek işlem kapasitesi ve bellek alanına ihtiyaç duyan sık alt çizge madenciliğinin düşük bellek ve işlem kapasiteli gündelik bir bilgisayar üzerinde gerçekleştirilebilirliği sınanmıştır. Çalışmanın sonuçlarına göre, sık alt çizge madenciliği algoritmaları birçok farklı alanda kullanılabilmektedir. Örneğin, sosyal ağ analizi, biyoinformatik, finansal analiz, internet trafiği analizi, pazarlama stratejileri ve suç analizi gibi birçok alanda bu algoritmaların kullanım potansiyeli vardır. Bu algoritmalar, veri içerisindeki önemli desenleri ve ilişkileri keşfederek bilgi çıkarma süreçlerine katkıda bulunmaktadır.

Ayrıca, çalışmada farklı sık alt çizge madenciliği algoritmalarının uygulanabilirliği de değerlendirilmiştir. Örnek algoritmalar arasında "gSpan" ve "graMi" gibi algoritmalar bulunmaktadır. Her bir algoritmanın farklı avantajları ve dezavantajları vardır ve uygulanacak olan alana bağlı olarak tercih edilmeleri gerekmektedir. Algoritma seçimi uygulama gereksinimlerine ve veri setinin özelliklerine bağlı olarak yapılmalıdır.

Sonuç olarak, sık alt çizge madenciliği algoritmalarının kullanım alanları oldukça geniştir ve birçok farklı sektörde değerli uygulamalara sahiptir. Ancak, bu algoritmaların seçimi ve uygulanması dikkatli bir şekilde yapılmalıdır. Bu tez çalışması, sık alt çizge madenciliği alanında yapılacak çalışmalara katkı sunmak ve ilgili alanda daha fazla araştırma

yapılabilmesi amacıyla yapılmıştır. Gelecekte, daha gelişmiş algoritmaların geliştirilmesi ve daha kapsamlı uygulama senaryolarının keşfedilmesi beklenmektedir.



6. KAYNAKLAR

- Aggarwal, C. C., & Zhai, C. (2012). Mining text data. Springer Science & Business Media.
- Anastasiu, D. C., Iverson, J., Smith, S., & Karypis, G. (2014). Big data frequent pattern mining. *Frequent pattern mining*, 225-259.
- Ay, D. (2009). Veri madenciliği ve apriori algoritması ile süpermarket analizi (Master's thesis, Sakarya Üniversitesi).
- Baykal, A. (2006). Veri madenciliği uygulama alanları. Dicle Üniversitesi Ziya Gökalp Eğitim Fakültesi Dergisi, (7), 95-107.
- Bilgin, T., & Oğuz, M. (2021). A New Approach to Minimize Memory Requirements of Frequent Subgraph Mining Algorithms. *Politeknik Dergisi*, 24(1), 237-246.
- Bodur, S. (2015). Çizge Madenciliği ve Algoritmaları. *Yüksek lisans*. İzmir: Ege Üniversitesi.
- Bonchi, F., Castillo, C., Gionis, A., & Jaimes, A. (2011). Social network analysis and mining for business applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3), 1-37.
- Connolly, T., & Begg, C. (2014). *Database Systems: A Practical Approach to Design, Implementation, and Management*. Pearson Education.
- Dong, X. L., Srivastava, D., & Aggarwal, C. C. (2014). Graph Data Management and Mining: A Survey of Algorithms, Models, and Systems. *IEEE Transactions on Knowledge and Data Engineering*, 26(6), 1458-1476.
- Döşlü, A. (2008). Veri madenciliğinde market sepet analizi ve birliktelik kurallarının belirlenmesi.
- Elmasri, R., & Navathe, S. B. (2016). *Fundamentals of Database Systems*. Pearson Education
- Han, J., Kamber, M. ve Pei, J., 2012, Data mining: concepts and techniques, 3. baskı, Morgan Kaufman Publishers, ABD, ISBN: 978-0-12-381479-1.
- Han, J., Kamber, M., & Pei, J. (2011). The Morgan Kaufmann series in data management systems. *Data mining concepts and techniques*, 5(4), 83-124.

- Inokuchi, A., Washio, T., & Motoda, H. (2000). An apriori-based algorithm for mining frequent substructures from graph data. In *Principles of Data Mining and Knowledge Discovery: 4th European Conference, PKDD 2000 Lyon, France, September 13–16, 2000 Proceedings 4* (pp. 13-23). Springer Berlin Heidelberg.
- Johnson, R., & Smith, J. (2021). The Applications of Frequent Subgraph Mining. *Journal of Data Mining and Knowledge Discovery*, 28(3), 70-90.
- Karacabey, S. (1997). Veritabanı modelleri ve hiyerarşik veritabanı modelinden ilişkisel veritabanı modeline geçiş.
- Karagöz, G., & Komesli, M. (2017). Çizge Veri Tabanı Kullanılarak Geliştirilen Yazılım Lisans Yönetimi Amaçlı Veri Görselleştirme Uygulaması:BigLogVis. *Dokuz Eylül Üniversitesi-Mühendislik Fakültesi Fen ve Mühendislik Dergisi*, 779-789.
- Ke, Y., Li, G., & Feng, J. (2015). Graph Database: A Survey. *Science China Information Sciences*, 58(1), 1-18.
- Kuramochi, M., & Karypis, G. (2004). *Frequent Subgraph Discovery*. Springer.
- Lakshmi, K., & Meyyappan, T. (2012). A comparative study of frequent subgraph mining algorithms. *International Journal of Information Technology Convergence and Services*, 2(2), 23.
- Lee, H., & Smith, J. (2020). The Importance of Operational Databases in Data Management. *Journal of Business Database Management*, 18(2), 20-40.
- Leskovec, J., Rajaraman, A., & Ullman, J. D. (2014). *Mining of Massive Datasets*, Cambridge University Press, Cambridge.
- Liao, S., Zhao, X., & Wei, X. (2019). gSpan-Based Frequent Subgraph Mining Algorithm. *Journal of Electrical and Computer Engineering*, 2019.
- Nijssen, S., & Kok, J. N. (2005). The gaston tool for frequent subgraph mining. *Electronic Notes in Theoretical Computer Science*, 127(1), 77-87.
- Oğuz, M. (2016). Sık Alt Çizge Madenciliği Algoritmalarına Uygulanabilir Alan ve Zaman Verimliliği Arttıran Metotların Geliştirilmesi. *Doktora Tezi*. İstanbul: Maltepe Üniversitesi.

- Özgür, M., & Özsu, M. T. (2015). NoSQL Databases: A Survey and Decision Guidance. *SIGMOD Record*, 44(3), 3-28.
- Öztürk, S. (2017). İlişkisel ve İlişkisel Olmayan Veri Tabanı Yönetim Sistemleri Mimari Performansının Yönetim Bilişim Sistemleri Kapsamında İncelenmesi. *Yüksek lisans*. Ankara: Gazi Üniversitesi.
- Öztürk, S., & Atmaca, H. E. (2017). İlişkisel ve ilişkisel olmayan (NoSQL) veri tabanı sistemleri mimari performansının yönetim bilişim sistemleri kapsamında incelenmesi. *Bilişim Teknolojileri Dergisi*, 10(2), 199-209.
- Ramakrishnan, R., & Gehrke, J. (2003). Database Management Systems. McGraw-Hill
- Silahtaroğlu, G. (2008). Veri madenciliği. *Papatya Yayınları, İstanbul*.
- Smith, J., & Jones, A. (2019). The Importance of Graph Databases in Data Management. *Journal of Information Science*, 15(3), 40-60.
- Şen, F. (2008). Veri madenciliği ile birliktelik kurallarının bulunması (Master's thesis, Sakarya Üniversitesi).
- Tan, P. N., Steinbach, M., & Kumar, V. (2006). Introduction to Data Mining. Pearson Education.
- Tang, J., Liu, H., & Zhang, J. (2009). Mining social media: a brief introduction. *ACM SIGKDD Explorations Newsletter*, 11(1), 9-14.
- Taşkın, Ç., & Emel, G. (2010). Veri Madenciliğinde Kümeleme Yaklaşımları Ve Kohonen Ağları İle Perakendecilik Sektöründe Bir Uygulama. *Süleyman Demirel Üniversitesi*, 395-409.
- Tiryaki, S. (2006). Lojistik Alanında Bir Veri Madenciliği Uygulaması. *Yüksek Lisans Tezi*. İstanbul: İstanbul Teknik Üniversitesi.
- Ünalır, M. O., Inan, E., Yönyül, B., Olca, E., Sentürk, F., Mostafapour, V., ... & Isli, D. (2015). Veri Yogun Bilgi Sistemleri İçin Melez Bir Veri Mimarisi Onerisi.

- Wörlein, M., Meinl, T., Fischer, I., & Philippsen, M. (2005). A quantitative comparison of the subgraph miners MoFa, gSpan, FFSM, and Gaston. In Knowledge Discovery in Databases: PKDD 2005: 9th European Conference on Principles and Practice of Knowledge Discovery in Databases, Porto, Portugal, October 3-7, 2005. Proceedings 9 (pp. 392-403). Springer Berlin Heidelberg.
- Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2014). Data mining with big data. *IEEE transactions on knowledge and data engineering*, 26(1), 97-107.)
- Yan, X., & Han, J. (2002). gSpan: Graph-based substructure pattern mining. *Data Mining and Knowledge Discovery*, 2(2), 141-182.
- Yıldırım, o. (2014). Çizge Veri Tabanlarının İncelenmesi ve Çizge Veri Tabanı Yaratılması. *Yüksek Lisans Tezi*. İzmir: Ege Üniversitesi.
- Yorulmaz, A. (2018). Veri Tabanı Sistemlerinde Yüksek Performans ve Erişilebilirlik ile Güvenliğin Oracle Veri Tabanı Sistemi Özelinde İncelenmesi. *Yüksek Lisans Tezi*. Sakarya: Sakarya Üniversitesi.