

**T.C.
ERCIYES ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI**

**BÜYÜK SOSYAL MEDYA VERİSİNDEN MEKÂNSAL VE
MEKÂN-ZAMANSAL ÖNEMLİ LOKASYONLAR KEŞFİ VE
BULUT BİLİŞİM SİSTEMLERİNE UYARLANMASI**

**Hazırlayan
Ahmet Şakir DOKUZ**

**Danışman
Yrd. Doç. Dr. Mete ÇELİK**

Doktora Tezi

**Aralık 2017
KAYSERİ**

**T.C.
ERCIYES ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI**

**BÜYÜK SOSYAL MEDYA VERİSİNDEN MEKÂNSAL VE
MEKÂN-ZAMANSAL ÖNEMLİ LOKASYONLAR KEŞFİ
VE BULUT BİLİŞİM SİSTEMLERİNE UYARLANMASI
(Doktora Tezi)**

**Hazırlayan
Ahmet Şakir DOKUZ**

**Danışman
Yrd. Doç. Dr. Mete ÇELİK**

**Bu tez çalışması Erciyes Üniversitesi Bilimsel Araştırma Projeleri
Birimi (Proje No: FDK-2017-7233) ve Microsoft Azure for Research
Ödül Programı (Proje No: 0518113) tarafından desteklenmiştir.**

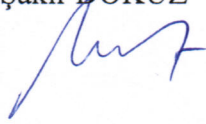
**Aralık 2017
KAYSERİ**

BİLİMSEL ETİĞE UYGUNLUK

Bu çalışmadaki tüm bilgilerin, akademik ve etik kurallara uygun bir şekilde elde edildiğini beyan ederim. Aynı zamanda bu kural ve davranışların gerektirdiği gibi, bu çalışmanın özünde olmayan tüm materyal ve sonuçları tam olarak aktardığımı ve referans gösterdiğimi belirtirim.

Ahmet Şakir DOKUZ

İmza :



YÖNERGEYE UYGUNLUK

“Büyük Sosyal Medya Verisinden Mekânsal Ve Mekân-Zamansal Önemli Lokasyonlar Keşfi Ve Bulut Bilişim Sistemlerine Uyarlanması” adlı Doktora tezi, Erciyes Üniversitesi Lisansüstü Tez Önerisi ve Tez Yazma Yönergesi’ne uygun olarak hazırlanmıştır.

Tezi Hazırlayan

Ahmet Şakir DOKUZ

Danışman

Yrd. Doç. Dr. Mete ÇELİK

Bilgisayar Mühendisliği ABD Başkanı

Prof. Dr. Veysel ASLANTAŞ

Yrd. Doç. Dr. Mete ÇELİK danışmanlığında **Ahmet Şakir DOKUZ** tarafından hazırlanan “**Büyük Sosyal Medya Verisinden Mekânsal ve Mekân-Zamansal Önemli Lokasyonlar Keşfi ve Bulut Bilişim Sistemlerine Uyarlanması**” adlı bu çalışma, jürimiz tarafından Erciyes Üniversitesi Fen Bilimleri Enstitüsü **Bilgisayar Mühendisliği** Anabilim Dalında **Doktora** tezi olarak kabul edilmiştir.

29/12/2017

JÜRİ:

Danışman :Yrd. Doç. Dr. Mete ÇELİK
 Üye :Doç. Dr. Gülay TEZEL
 Üye :Doç. Dr. Alper BAŞTÜRK
 Üye :Doç. Dr. Enis GÜNAY
 Üye :Yrd. Doç. Dr. Humar KAHRAMANLI

(Handwritten signatures of the jury members)

ONAY:

Bu tezin kabulü Enstitü Yönetim Kurulunun 02/01/2018 tarih ve 2018/01-18 sayılı kararı ile onaylanmıştır.

(Handwritten signature of Prof. Dr. Mehmet AKKURT)
 02/01/2018

Prof. Dr. Mehmet AKKURT

Enstitü Müdürü

ÖNSÖZ/TEŞEKKÜR

Tez çalışmalarım süresince beni akademik araştırma yapma, bilimsel düşünce sistematığı kazanma ve mesleki etik başta olmak üzere her konuda yetiştiren değerli danışmanım Yrd. Doç. Dr. Mete ÇELİK hocama teşekkürü bir borç bilirim. Ayrıca tez izleme komitemde bulunup tezimin ilerleyişine katkıda bulunan değerli hocalarım Doç. Dr. Alper BAŞTÜRK ve Doç. Dr. Enis GÜNAY'a da teşekkür ederim.

Üyesi olduğum Niğde Ömer Halisdemir Üniversitesi'nde, doktora çalışmalarım süresince bana her türlü maddi ve manevi yardımlarını esirgemeyen sayın dekanımız Prof. Dr. Kutsi S. ERDURAN hocama ve yardımcıları Doç. Dr. Murat GÖKÇEK ve Doç. Dr. Ersin AYDIN hocalarıma çok teşekkür ederim. Bölüm başkanımız Yrd. Doç. Dr. Bekir Sami TEZEKİCİ hocama, ilgisi ve yardımlarından dolayı teşekkür ederim. Başta Arş. Gör. Yeşim DOKUZ olmak üzere tüm çalışma arkadaşlarıma teşekkür ederim.

Bu tez çalışmasına maddi destek sağlayan Erciyes Üniversitesi BAP Birimi (Proje No: FDK-2017-7233) ve Microsoft'a (Azure for Research Araştırma Ödülü, Proje No: 0518113) teşekkür ederim.

Son olarak kıymetli eşim Yeşim, oğlum Ömer Avni, annem, babam ve tüm aileme, tez çalışmalarım sürecindeki destekleri için en içten teşekkürlerimi sunarım.

Kayseri, 2017

BÜYÜK SOSYAL MEDYA VERİSİNDEN MEKÂNSAL VE MEKÂN-ZAMANSAL ÖNEMLİ LOKASYONLAR KEŞFİ VE BULUT BİLİŞİM SİSTEMLERİNE UYARLANMASI

Ahmet Şakir DOKUZ

Erciyes Üniversitesi, Fen Bilimleri Enstitüsü

Doktora Tezi, Aralık 2017

Danışman: Yrd. Doç. Dr. Mete ÇELİK

ÖZET

Sosyal önemli lokasyonlar, sosyal medya kullanıcılarının sosyal medya geçmişlerinde sıklıkla ziyaret ettikleri yerleri ifade etmektedir. Sosyal önemli lokasyonların keşfi, lokasyon tavsiye sistemleri, odaklanmış reklamcılık ve şehir örüntü keşfi gibi uygulama alanları için önemlidir. Ancak, sosyal medya ağlarından sosyal önemli lokasyonların keşfi, veri büyüklüğü ve boyutu, veri kümelerinin mekânsal ve zamansal boyutları ve hesaplama açısından verimli yaklaşımlar geliştirme gerekliliğinden dolayı zordur. Literatürde, sosyal önemli lokasyonların keşfi ile ilgili çeşitli çalışmalar mevcuttur. Ancak bu çalışmaların pek çoğunda sosyal medya kullanıcılarının geçmişe dayalı verisi dikkate alınmadan sosyal önemli lokasyonların keşfi yapılmıştır. Bu tez çalışması kapsamında, sosyal önemli lokasyonların keşfi için mekânsal ve mekân-zamansal metotlar önerilmiş ve bu metotlar için değerlendirme ölçütleri geliştirilmiştir. Geliştirilmiş olan değerlendirme ölçütlerine dayalı olarak yeni mekânsal ve mekân-zamansal algoritmalar önerilmiştir. Önerilen algoritmalar Türkiye Twitter sosyal medya verisi üzerinde değerlendirilmiştir. Değerlendirme sonuçları, algoritmaların başarılı bir şekilde sosyal önemli lokasyonları keşfedebildiğini göstermektedir.

Bu tez çalışması kapsamında yapılan bir diğer çalışma ise önerilmiş olan mekânsal ve mekân-zamansal sosyal önemli lokasyonlar keşfi algoritmalarının bulut bilişim sistemlerine uyarlanmasıdır. Özel olarak, önerilen algoritmalar Hadoop MapReduce programlama modeline uyarlanmış ve dağıtık kümeler üzerinde sosyal önemli lokasyonların keşfi yapılmıştır. Bulut bilişim sistemleri, önerilen mekânsal ve mekân-zamansal algoritmaların çok kısa sürelerde sonuç üretebilmesini sağlamıştır.

Anahtar Kelimeler: Sosyal medya madenciliği, sosyal önemli lokasyonlar keşfi, büyük veri analizi, bulut bilişim, Apache Hadoop.

**SPATIAL AND SPATIO-TEMPORAL IMPORTANT LOCATIONS
DISCOVERY OF BIG SOCIAL MEDIA DATA AND APPLICATION ON
CLOUD COMPUTING SYSTEMS**

Ahmet Şakir DOKUZ

Erciyes University, Graduate School of Natural and Applied Sciences

Ph.D. Thesis, December 2017

Supervisor: Assist. Prof. Dr. Mete ÇELİK

ABSTRACT

Socially important locations are places which are frequently visited by social media users in their social media lifetime. Discovering socially important locations are important for several applications, such as, location recommendation systems, targeted advertisement, and urban pattern mining. However, discovering socially important locations is challenging due to data size and dimension, spatial and temporal aspects of datasets, and the need for developing computationally efficient approaches. In the literature, several studies are performed for socially important locations discovery. However, most of these studies discovered socially important locations without considering historical data of users. In this thesis, spatial and spatio-temporal methods are proposed to discover socially important locations and developed novel interest measures to quantify socially important locations. Novel spatial and spatio-temporal algorithms are proposed based on these interest measures. Proposed algorithms are evaluated using Turkey Twitter social media dataset. Experimental results present that the algorithms could successfully discover socially important locations.

Another study of this thesis is to apply spatial and spatio-temporal socially important locations discovery algorithms on cloud computing systems. In particular, proposed algorithms are modified using Hadoop MapReduce programming model and socially important locations are discovered on distributed clusters. Discovering socially important locations on cloud computing systems provide algorithms to extract results faster than traditional stand-alone systems.

Keywords: Social media mining, socially important locations discovery, big data analytics, cloud computing, Apache Hadoop.

İÇİNDEKİLER

BÜYÜK SOSYAL MEDYA VERİSİNDEN MEKÂNSAL VE MEKÂN- ZAMANSAL ÖNEMLİ LOKASYONLAR KEŞFİ VE BULUT BİLİŞİM SİSTEMLERİNE UYARLANMASI

BİLİMSEL ETİĞE UYGUNLUK SAYFASI	ii
YÖNERGEYE UYGUNLUK SAYFASI.....	iii
KABUL VE ONAY SAYFASI	iv
ÖNSÖZ/TEŞEKKÜR	v
ÖZET.....	vi
ABSTRACT.....	vii
İÇİNDEKİLER	viii
TABLolar LİSTESİ.....	xiii
ŞEKİLLER LİSTESİ	xiv

GİRİŞ	1
-------------	---

1. BÖLÜM

SOSYAL MEDYA MADENCİLİĞİ

1.1. Sosyal Medya Madenciliği	7
1.2. Sosyal Medya Madenciliğinin Önemi	8
1.3. Sosyal Medya Madenciliğinin Zorlukları.....	9
1.4. Sosyal Medya Madenciliği Metotları.....	10
1.4.1. Sosyal Ağ Analizi.....	10
1.4.2. Duygu Analizi ve Düşünce Madenciliği	11
1.4.3. Topluluk Analizi.....	11
1.4.4. Sosyal Tavsiye Sistemleri.....	11
1.4.5. Mekânsal ve Mekân-Zamansal Analiz	12

2. BÖLÜM

BULUT BİLİŞİM

2.1. Bulut Bilişim Tanımı ve Özellikleri	13
2.1.1. Bulut Bilişim Sistemlerinin Özellikleri	14

2.1.2.	Bulut Bilişimin Avantaj ve Dezavantajları	14
2.2.	Bulut Bilişim Servis Modelleri	15
2.3.	Bulut Bilişim Türleri	17
2.4.	Bulut Bilişimde Veri Türleri	18
2.4.1.	Yapısal Veri	19
2.4.2.	Yapısal Olmayan Veri	20
2.4.3.	Yarı-Yapısal Veri	21
2.4.4.	Bulut Verisinin Karakteristik Özellikleri	21
2.5.	Apache Hadoop.....	21
2.5.1.	Hadoop MapReduce	23

3. BÖLÜM

SOSYAL MEDYA AĞLARINDAN VERİ TOPLAMA

3.1.	Twitter API'leri	29
3.1.1.	REST API	29
3.1.2.	Streaming API	29
3.1.3.	Ads API	29
3.2.	Twitter4J Kütüphanesi	30
3.3.	Veri Toplama Sorguları	30

4. BÖLÜM

MEKÂNSAL VE MEKÂN-ZAMANSAL SOSYAL MEDYA MADENCİLİĞİ VE BULUT BİLİŞİM KONUSUNDA YAPILMIŞ OLAN ÇALIŞMALAR

4.1.	Mekânsal Sosyal Medya Madenciliği	32
4.2.	Mekân-Zamansal Sosyal Medya Madenciliği	37
4.3.	Bulut Bilişim Sistemlerinde Sosyal Medya Madenciliği	39
4.4.	Tez Çalışmasının Literatüre Katkıları	42

5. BÖLÜM

BÜYÜK SOSYAL MEDYA VERİSİNDEN MEKÂNSAL ÖNEMLİ LOKASYONLARIN KEŞFİ

5.1.	Giriş	44
5.2.	Temel Tanımlar ve Problem Tanımlaması	46

5.2.1.	Temel Tanımlar	47
5.2.2.	Problem Tanımlaması.....	48
5.3.	Sosyal Önemli Lokasyonların Keşfi.....	49
5.3.1.	Toy Yaklaşım.....	49
5.3.2.	SS-ILM Algoritması.....	50
5.3.3.	Algoritmanın Örnek Çalışması.....	52
5.4.	Deneyisel Çalışmalar	55
5.4.1.	Veri Kümesi	56
5.4.2.	Önişlem	57
5.4.3.	Deneyisel Sonuçlar	59
5.4.4.	Önerilen SS-ILM Algoritmasının Literatür ile Karşılaştırılması	64
5.5.	Katkılar ve Kapsam	66
5.5.1.	Katkılar	66
5.5.2.	Kapsam	66

6. BÖLÜM

BÜYÜK SOSYAL MEDYA VERİSİNDEN MEKÂN-ZAMANSAL ÖNEMLİ LOKASYONLARIN KEŞFİ

6.1.	Giriş	67
6.2.	Temel Tanımlar ve Problem Tanımlaması	70
6.2.1.	Temel Tanımlar	70
6.2.2.	SMZÖL'lerin Modellenmesi	71
6.2.3.	Problem Tanımlaması.....	74
6.3.	Sosyo-Mekân-Zamansal Önemli Lokasyonların Keşfi	74
6.3.1.	Toy Algoritma	74
6.3.2.	TF-SSTIL Miner Algoritması	78
6.3.3.	SF-SSTIL Miner Algoritması.....	79
6.3.4.	Algoritmanın Örnek Çalışması.....	81
6.4.	Deneyisel Çalışmalar	85
6.4.1.	Veri Kümesi	86
6.4.2.	Önişlem	87
6.4.3.	Deneyisel Sonuçlar	89

6.4.4.	Önerilen TF-SSTIL Miner ve SF-SSTIL Miner Algoritmalarının Literatür ile Karşılaştırılması	95
6.5.	Katkılar ve Kapsam	97
6.5.1.	Katkılar	97
6.5.2.	Kapsam	98

7. BÖLÜM

ÖNERİLEN MEKÂNSAL VE MEKÂN-ZAMANSAL ÖNEMLİ LOKASYONLAR KEŞFİ ALGORİTMALARININ BULUT BİLİŞİM SİSTEMLERİNE UYARLANMASI

7.1.	Giriş	99
7.2.	Bulut Bilişim Tabanlı Sosyal Önemli Lokasyonların Keşfi.....	101
7.2.1.	Bulut SS-ILM algoritması	102
7.2.2.	Bulut TF-SSTIL Miner.....	103
7.2.3.	Bulut SF-SSTIL Miner	105
7.3.	DeneySEL Çalışmalar	106
7.3.1.	Veri Kümesi	107
7.3.2.	Önişlem Aşamaları.....	108
7.3.3.	Bulut Bilişim Ortamı	110
7.3.4.	DeneySEL Sonuçlar	111
7.4.	Katkılar ve Kapsam	117
7.4.1.	Katkılar	117
7.4.2.	Kapsam	118

8. BÖLÜM

BULGULAR

8.1.	Mekânsal Analiz Değerlendirmesi	119
8.2.	Mekân-Zamansal Analiz Değerlendirmesi	120
8.3.	Bulut Bilişim Sistemlerinin Kullanımının Değerlendirmesi.....	121

9. BÖLÜM

TARTIŞMA, SONUÇ VE ÖNERİLER

9.1.	Tartışma ve Sonuç	124
------	-------------------------	-----

9.2. Öneriler	125
KAYNAKÇA	126
ÖZGEÇMİŞ.....	131

TABLOLAR LİSTESİ

Tablo 5.1.	Örnek kullanıcı için lokasyon yoğunluğu ve ziyaret ömrü değerleri ve keşfedilen/elenen lokasyonlar.....	54
Tablo 5.2.	5 sosyal medya kullanıcısı için 20 günlük ziyaret edilen lokasyonlar.....	55
Tablo 5.3.	5 kullanıcı için keşfedilen/elenen lokasyonlar, KY değerleri ve örnek veri kümesinin sosyal önemli lokasyonları (SÖL).....	55
Tablo 5.4.	Keşfedilen mekânsal sosyal önemli lokasyonların açıklaması.....	63
Tablo 5.5.	Khetarpaul vd. [61] tarafından önerilen yaklaşımın keşfettiği sosyal önemli lokasyonlar.....	65
Tablo 6.1.	Örnek çalışma veri kümesi.....	81
Tablo 6.2.	Kullanıcılar için KZLÖ ve KZÖL’lar (Mek. Eledi, lokasyonun mekânsal aşamada elendiğini göstermektedir).....	84
Tablo 6.3.	Zaman pencereleri için ZP_MZKY değerleri, SMZLÖ değerleri ve SMZÖL’lar.....	85
Tablo 6.4.	İstanbul için keşfedilen ilk 10 lokasyon tanımı.....	95
Tablo 6.5.	Bölüm 5’te İstanbul için keşfedilen ilk 10 lokasyon tanımı.....	96
Tablo 6.6.	Khetarpaul vd. [61] tarafından İstanbul için keşfedilen ilk 10 lokasyon tanımı.....	97
Tablo 7.1.	Microsoft Azure işçi boyut seçenekleri.....	114
Tablo 7.2.	Türkiye Twitter kullanıcıları için ilk 10 SÖL’lar.....	117

ŞEKİLLER LİSTESİ

Şekil 1.1.	Sosyal medya madenciliği temel adımları.....	7
Şekil 2.1.	Bulut bilişim servis modelleri.....	16
Şekil 2.2.	Özel, genel ve hibrit bulut ortamları.....	18
Şekil 2.3.	MapReduce çalışma mimarisi [15].....	23
Şekil 2.4.	MapReduce versiyonları [16].....	24
Şekil 3.1.	Örnek bir Twitter API ekranı	27
Şekil 5.1.	a) Bir kullanıcı için sosyal önemli lokasyonlar ve b) Bir sosyal medya kullanıcı grubu için sosyal önemli lokasyonlar	45
Şekil 5.2.	Bir kullanıcının farklı lokasyonlardaki sosyal medya aktivite kayıtları. Şekildeki numaralar, kullanıcının ilgili gün ve lokasyondaki aktivite sayısını göstermektedir.....	52
Şekil 5.3.	Örnek kullanıcının lokasyonlarının haritadaki gösterimi	53
Şekil 5.4.	Deney düzeneği	56
Şekil 5.5.	Veri kümesinin zamansal dağılımı	58
Şekil 5.6.	Kullanıcı sayısının etkisi	60
Şekil 5.7.	Minimum lokasyon yoğunluk eşik değerinin etkisi	61
Şekil 5.8.	Minimum ziyaret ömrü eşik değerinin etkisi	62
Şekil 5.9.	1000 İstanbul kullanıcısı için sosyal önemli lokasyonlar	63
Şekil 6.1.	İki farklı kullanıcının SÖL'ları ve zamansal yaygınlıkları	69
Şekil 6.2.	Kullanıcı 1'in mekânsal sosyal önemli lokasyonları ve zaman dilimlerindeki dağılımları.....	83
Şekil 6.3.	Deney düzeneği.....	85
Şekil 6.4.	Veri kümesinin zamansal dağılımı.....	88
Şekil 6.5.	Kullanıcı sayısının etkisi.....	90

Şekil 6.6.	Minimum lokasyon yoğunluk eşik değerinin etkisi.....	91
Şekil 6.7.	Minimum ziyaret ömrü eşik değerinin etkisi.....	92
Şekil 6.8.	Minimum zaman penceresi oluşumu eşik değerinin etkisi.....	92
Şekil 6.9.	Minimum pencere yaygınlığı eşik değerinin etkisi.....	93
Şekil 6.10.	İstanbul için keşfedilen ilk10 SMZÖL.....	94
Şekil 7.1.	SS-ILM algoritmasının kullanıcı artışına karşılık işlem zamanı değişimi.....	100
Şekil 7.2.	Deney düzeneği.....	107
Şekil 7.3.	Veri kümesinin zamansal dağılımı.....	108
Şekil 7.4.	Kullanıcı sayısının etkisi.....	111
Şekil 7.5.	Veri parçası sayısının etkisi.....	112
Şekil 7.6.	İşçi düğüm sayısının etkisi.....	113
Şekil 7.7.	İşçi boyutunun etkisi.....	115
Şekil 7.8.	Bulut SS-ILM algoritması tarafından keşfedilen SÖL'lar.....	116
Şekil 7.9.	Bulut SF-SSTIL Miner ve Bulut TF-SSTIL Miner algoritmaları tarafından keşfedilen SMZÖL'lar.....	116
Şekil 8.1.	Kullanıcı sayısının artışına karşılık algoritmaların işlem zamanının artışı.....	121

GİRİŞ

Sosyal medya ağları kavramı, geçtiğimiz 10 yıl içerisinde ortaya çıkan ve popülaritesini her geçen gün artırmaya devam eden Web 2.0'ın yaygınlaşmasıyla ortaya çıkmış bir kavramdır. Sosyal medya ağlarının temel işlevi fiziksel olarak bir arada bulunmayan insanları sanal ortamlarda bir araya getirmek, takip ettikleri önemli kişileri bu ağlar aracılığıyla takip etmek ve birbirlerinin düşünce ve duygularını öğrenmektir. Sosyal medya ağlarında, kullanıcılar, birbirlerini takip eder, paylaşımlarını okur ve yorumlarını paylaşırlar.

Sosyal medya ağlarının günümüz insanlarına getirdiği en önemli yenilik, hiç şüphesiz zaman ve mekândan bağımsız olarak kullanıcılarını bir araya getirebilme yeteneğidir. Bu alanın öncüleri arasında Facebook, Twitter, Instagram gibi milyarları bulan kullanıcı sayılarına sahip ağlar sayılabilir. Bu ağlar, Web 2.0'ın yaygınlaşmasıyla birlikte, sıradan web kullanıcılarının da birer bilgi kaynağı olabileceği temeli üzerine kurulmuştur. Bir kullanıcıyı takip etmek, diğer bir bakış açısıyla o kullanıcının fikirlerine değer vermek, herhangi bir konu hakkındaki düşüncelerini öğrenmek amacı taşır.

Sosyal medya ağlarının en önemli özelliği gelişen teknolojileri hızlı ve verimli bir şekilde kendi sistemlerine adapte edebilmeleridir. Bu konudaki en önemli kilometre taşı, mobil akıllı cihazların yaygınlaşması ve bu cihazlara entegre sensörler yardımıyla GPS koordinatlarının alınabilir hale gelmesidir. Sosyal medya ağları, bu teknolojik gelişimi, kullanıcılarının paylaşımlarını yaparken ve birbirleriyle etkileşimde bulunurken mekânsal bilgilerini de toplamakta kullanarak ürettikleri analiz sonuçlarını mekânsal ve zamansal boyuta da yaymışlardır. Kullanıcıların mekânsal ve zamansal olarak analiz edilmesi, verinin boyutunu kat kat artırmakla birlikte, üretilen analiz sonuçlarını çok daha değerli hale getirmektedir.

Sosyal medya madenciliği, sosyal medya ağlarındaki verilerin analizi ile önceden

görülemeyen, anlamlı ve fayda sağlayacak sonuçlar çıkarımı işlemidir. Son istatistiklere göre Google günlük yaklaşık olarak 24 PB'lık veriyi işlemektedir. Facebook günlük yaklaşık olarak 600 TB veri depolamakta ve toplam veri boyutu 300 PB seviyelerinin üzerindedir. Ayrıca Facebook kullanıcıları her saat yaklaşık 10 milyon fotoğraf paylaşmaktadır. Twitter'da günlük ortalama 500 milyon tweet atılmakta ve aylık 320 milyon aktif kullanıcı Twitter'ı kullanmaktadır. 800 milyonluk YouTube kullanıcıları her saniye 1 saatin üzerinde video paylaşımı yapmaktadırlar [1] [2] . Sosyal medya ağlarındaki bu inanılmaz veri çeşitliliği ve boyutu düşünüldüğü zaman sosyal medya madenciliği işleminin zorluğu ve aynı zamanda işlenmesi halinde bu veriden elde edilecek faydaların çokluğu ortaya çıkmaktadır. Bu boyuttaki yarı-yapısal verinin madenciliği işlemi için özel yaklaşımlar ve algoritmalar geliştirilmesi gerekmektedir.

Sosyal medya madenciliği çalışmaları literatürü incelendiği zaman, ilk çalışmalar sadece kullanıcılar arasındaki ilişkilerin bir graf olarak modellenmesiyle başlamıştır. Daha sonra kullanıcıların sosyal medya ağlarındaki paylaşımlarının içerikleri de dâhil edilerek daha doğru sonuçlar üretilmeye çalışılmış ve çalışmaların çeşitliliği artmaya başlamıştır. Kullanıcıların mekânsal bilgilerinin toplanabilmesi ile de mekânsal ve mekân-zamansal sosyal medya madenciliği çalışmaları yaygınlık kazanmış ve bu konudaki çalışmalar çok çeşitli alanlarda kullanılmaya başlanmıştır.

Kullanıcıların mekân bilgilerinin toplanabilmesi ile sosyal medya ağları, büyük bir gelişme göstermişlerdir. Eklenen mekân bilgisi, veri boyutuna ve kullanıcının sosyal ağlardaki bilgi havuzuna bir boyut daha eklemiştir. Ancak bu mekân bilgisinin sağladığı fayda bir boyut eklemekten daha fazlasıdır. Kullanıcıların sosyal medya ağlarını kullanırken toplanan mekân bilgisi aracılığıyla, kullanıcıların fiziksel dünyaları ile sosyal ağların sanal dünyası birleştirilmiş ve artık sosyal medya kullanıcıları sadece sanal bir ağda bulunan herhangi bir soyut varlık olmaktan çıkmış ve fiziksel olarak bulunduğu lokasyonlar takip edilebilen, fiziksel dünyadaki etkileşimleri ve arkadaşlıkları kolaylıkla analiz edilebilen somut kullanıcılar olmuşlardır.

Mekânsal ve mekân-zamansal sosyal medya madenciliği, sosyal medya ağlarındaki kullanıcıların paylaşımlarını yaparken ve birbirleriyle iletişim kurarken mekân ve zaman bilgilerinden de faydalanarak farklı boyutlarda çıkarımlar yapılmasını sağlayan sosyal medya madenciliğinin bir alt dalıdır. Sosyal medya ağlarının, kullanıcılarının mekân

verisini toplamaya başlaması ve bu veriyi uygulama programlama arayüzleri (application programming interfaces – API) aracılığıyla, veri analistleriyle paylaşımlarıyla mekânsal ve mekân-zamansal sosyal medya madenciliği çalışmaları mümkün hale gelmiştir. Kullanıcıların birbirleriyle iletişim kurarken, paylaşımlarda bulunurken hangi lokasyonda bulunduğu bilgisi, enlem-boylam olarak erişilebilmektedir.

Tezin Amacı

Bu tez çalışmasında, mekânsal ve mekân-zamansal sosyal medya madenciliği ve bulut bilişim konuları çalışılmıştır. Sosyal medya ağlarından Twitter kullanılarak, Türkiye kullanıcıları analiz edilerek mekânsal ve mekân-zamansal sonuçlar üretilmiştir. Sosyal medya ağlarından Twitter'ın seçilmesinin nedenlerinden birincisi yaklaşık 11 milyon aktif kullanıcısıyla Türkiye'de oldukça popüler olmasıdır. İkinci sebep de Twitter'dan veri toplama kolaylığı sağlayan API bağlantılarıdır. Twitter'ın API limitleri büyük miktarlarda veri toplamak ve geniş bir aralıkta sorgular çalıştırabilmek için uygundur.

Tez çalışması için mekânsal ve mekân-zamansal sosyal medya madenciliğinin seçilmesinin sebebi ise, dünyada ve Türkiye'de mekânsal ve mekân-zamansal çalışmaların gittikçe popüler hale gelmesidir. Özellikle Türkiye kaynaklı veriler üzerinde mekânsal ve mekân-zamansal analizler yapılmış olan çalışmalar son derece sınırlıdır. Bu konudaki eksikliği tamamlayabilmek ve bu tez çalışması aracılığıyla Türkiye'de mekânsal ve mekân-zamansal sosyal medya madenciliğinin faydalarını gösterebilmek için bu konu seçilmiştir. Yapılmış olan deneysel çalışmalar ile mekânsal ve mekân-zamansal sosyal medya madenciliği çalışmalarının Türkiye kaynaklı veriler üzerinde çalıştırılmasıyla faydalı sonuçlar üretildiği düşünülmektedir.

Ayrıca tez çalışmasında mekânsal ve mekân-zamansal sosyal medya madenciliği algoritmalarının bulut bilişim sistemlerine uyarlanması da gerçekleştirilmiştir. Bulut bilişim, son yıllarda önem kazanan bir diğer bilgisayar bilimleri konusudur. Bulut bilişim sayesinde, bulut kullanıcıları kendi kaynaklarını kullanmadan, herhangi bir ek çaba harcamadan, verilerini analiz edebilmekte, sonuçlar çıkarabilmektedir. Büyük verinin analizi için bulut bilişim sistemleri yaygın bir şekilde kullanılmaya başlanmıştır. Büyük veri kaynaklarından biri olan sosyal medya verisinin analizi için de bulut bilişim faydalar ve geliştirmeler sağlayan bir teknolojidir. Bu tez çalışması kapsamında geliştirilmiş olan mekânsal ve mekân-zamansal sosyal medya madenciliği algoritmaları bulut bilişim

altyapıları üzerinde çalıştırılarak, elde edilen hızlandırmalar ve performans farkları ortaya konulmuştur. Aynı zamanda, bulut bilişim aracılığıyla, herhangi bir ek kaynak gerektirmeden hesaplama maliyetlerinin nasıl düşürüleceği ve büyük verinin nasıl daha verimli bir şekilde analiz edileceği gösterilmiştir.

Tezin Organizasyonu

Tez çalışması Giriş, 8 ana bölüm ve Sonuç bölümlerinden oluşmaktadır. Giriş bölümünde, tez konusunun genel bir tanıtımı yapılmış, hangi alt konuda çalışılacağı ifade edilmiş ve tezin devam eden bölümleri tanıtılmıştır.

1. Bölüm’de sosyal medya madenciliği tanımlanmıştır. Sosyal medya madenciliğinin zorlukları ve faydaları sunulmuş ve uygulama alanları incelenmiştir. Sosyal medya verisinin karakteristik özelliklerine değinilmiş ve veri türleri anlatılmıştır. Son olarak mekânsal sosyal medya madenciliği konusunda özet bilgi sunulmuştur.

2. Bölüm’de bulut bilişim konusunda detaylı bilgiler sunulmuştur. Bulut bilişime ait tanımlamalar yapılmış, bulut bilişimin özellikleri, avantaj ve dezavantajları verilmiştir. Bulut bilişim sistemlerindeki veri türleri, servis modelleri ve bulut bilişim türleri tartışılmıştır. Son olarak da bulut bilişim sistemlerinde büyük veri analizi için yaygın olarak kullanılan Apache Hadoop yazılım çerçevesi ve Hadoop MapReduce programlama modeli anlatılmıştır.

3. Bölüm’de mekânsal ve mekân-zamansal sosyal medya madenciliği konusundaki literatür incelenmiş, hangi çalışmaların bulunduğu, hangi konuların yaygın bir şekilde çalışıldığı tartışılmıştır. Ayrıca, bulut bilişim sistemlerinde sosyal medya madenciliği çalışmaları konusunda literatürde mevcut çalışmalar incelenmiştir. Bu tez çalışmasının literatürdeki konumu incelenmiş ve tez çalışmasının literatüre katkıları sunulmuştur.

4. Bölüm’de tez çalışması için seçilmiş olan sosyal medya ağlarından Twitter üzerinde veri toplama stratejileri anlatılmıştır. Twitter API’lerinin türleri, API bağlantılarının nasıl yapıldığı, Twitter verilerinin toplanabilmesi için ne tür sorguların yapılabileceği ve Türkiye kaynaklı coğrafi sorguların nasıl yapılabileceği anlatılmıştır.

5. Bölüm’de sosyal medya verilerinden mekânsal önemli lokasyonların keşfi konusu ele alınmış, mekânsal önemli lokasyonların keşfi için teorik bir yaklaşım önerilmiş, önerilen

yaklaşımı temel alan iki algoritma önerilmiştir. Algoritmaların kaba kodu ile örnek bir veri üzerindeki çalışma mantığı verilmiştir. Bölümün sonunda, önerilmiş olan mekânsal önemli lokasyonlar keşfi algoritmalarının gerçek Twitter verileri üzerindeki uygulamaları anlatılmıştır. Algoritmaların çıkardığı sonuçlar incelenmiş, hesaplama maliyeti, işlem performansı ve verimliliği karşılaştırmalı olarak test edilmiştir.

6. Bölüm’de sosyal medya verilerinden mekân-zamansal önemli lokasyonların keşfi konusu araştırılmış, mekân-zamansal önemli lokasyonların keşfi için teorik bir yaklaşım önerilmiş ve önerilen yaklaşımı temel alan üç adet mekân-zamansal önemli lokasyonlar madenciliği algoritması önerilmiştir. Algoritmaların kaba kodu ve örnek bir veri üzerindeki çalışma mantığı verilmiştir. Ayrıca, önerilmiş olan mekân-zamansal sosyal medya madenciliği algoritmalarının gerçek Twitter verileri üzerindeki uygulamaları verilmiştir. Algoritmalarından elde edilen sonuçların mekânsal ve zamansal olarak doğruluğu incelenmiş, algoritmaların hesaplama ve işlem maliyeti ve verimliliği karşılaştırmalı olarak test edilmiştir.

7. Bölüm’de, 5. ve 6. Bölümlerde önerilmiş olan mekânsal ve mekân-zamansal önemli lokasyonlar madenciliği algoritmalarının bulut bilişim sistemleri üzerine uyarlanması anlatılmıştır. Algoritmaların bulut bilişim sistemleri üzerinde verimli bir şekilde çalışabilmeleri için gereken düzenlemeler ve geliştirmeler yapılmıştır. Bulut bilişim sistemi olarak Microsoft Azure for Research programı desteğiyle Microsoft Azure bulut ortamı kullanılmıştır. Yapılan geliştirmeler sonucunda elde edilen sonuçların tutarlılığı 5. ve 6. Bölümlerde verilmiş olan analiz sonuçları ile karşılaştırmalı olarak test edilmiş ve aynı zamanda hesaplama maliyeti ve işlem performansı da karşılaştırılmıştır.

8. Bölüm’de tüm tez çalışmaları kapsamında elde edilen deney sonuçlarının toplu şekilde değerlendirmesi yapılmıştır. Elde edilen sonuçlar, algoritmaların işlem zamanı performansı ve sonuçların değerlendirmesi şeklinde yapılmıştır. Özellikle bulut bilişim sistemlerinden elde edilen işlem zamanı performansları, gelecekteki yapılabilecek çalışmalar için teşvik edici niteliktedir. Ayrıca, keşfedilen mekânsal ve mekân-zamansal sosyal önemli lokasyonlar, İstanbul ve Türkiye Twitter kullanıcılarının tercihlerini yansıtmaktadır.

9. Bölüm’de tez çalışmasının literatüre katkıları sunulmuş, önerilen yöntemlerin avantaj ve dezavantajlarından bahsedilmiş, çalışma sonuçları tartışılmıştır. Ayrıca, tez çalışmasını takip edebilecek gelecekteki çalışmalara da yer verilmiştir.

1. BÖLÜM

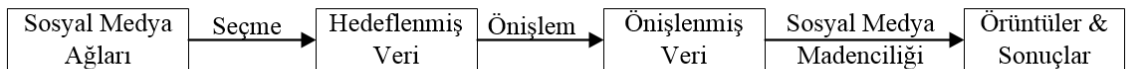
SOSYAL MEDYA MADENCİLİĞİNE GİRİŞ

Bu bölümde, sosyal medya madenciliğinin tanımlaması yapılmıştır. Sosyal medya madenciliğinin zorlukları ve avantajları incelenmiş ve sosyal medya madenciliğinin potansiyel uygulama alanları sunulmuştur. Sosyal medya verisinin karakteristik özelliklerine değinilmiş ve sosyal medya ağlarında bulunan veri türleri anlatılmıştır. Son olarak mekânsal sosyal medya verisi hakkında bilgi verilmiştir.

1.1.Sosyal Medya Madenciliği

Kapsamlı bir tanımlama yapılacak olunursa, sosyal medya madenciliği, kullanıcılar tarafından üretilmiş olan içeriği ve kullanıcıların üretilmiş olan içeriklere yönelik tepkisinin incelenmesi ve değerlendirilmesi ile anlamlı sonuçlar çıkarmayı hedefleyen bir veri madenciliği dalıdır [3] [4]. Sosyal medya madenciliği, büyük oranda veri madenciliği ve bilgisayar bilimine dayalı olsa da diğer alanların da katkı sunduğu disiplinler arası bir alan olarak üzerinde çalışmalar yapılmaktadır. Bu alanların en önemlileri, bilgisayar bilimleri, veri madenciliği, makine öğrenmesi, istatistik, optimizasyon, matematik, sosyoloji, doğal dil işleme ve etnoloji olarak sayılabilir.

Veri madenciliğinin temel adımları olan veri toplama, veri seçme, veri önileme, veri madenciliği ve anlamlandırma adımları sosyal medya madenciliği için de geçerlidir. Şekil 1.1’de sosyal medya madenciliğinin temel adımları sunulmuştur.



Şekil 1.1. Sosyal medya madenciliği temel adımları

Şekil 1.1’de görüldüğü üzere, ilk olarak, sosyal medya ağlarından toplanacak olan veriler

seçilmekte ve bu veriler toplanmaktadır. Toplanan bu hedeflenmiş veri önışlem aşamalarından geçirilmekte ve sosyal medya madencilięi aşaması için hazırlanmaktadır. Sosyal medya madencilięi aşamasında, veri üzerinde yapılacak olan analizler yapılmakta ve örüntüler ve sonuçlar keşfedilmektedir.

1.2.Sosyal Medya Madencilięinin Önemi

Sosyal medya madencilięi, büyük veri kaynaęı olması, içerdii potansiyel bilgilerin çok fazla olması, insanlar arası iletiřimlerin kolaylıkla analiz edilebilmesi, kullanıcıların bir konu hakkındaki duygu ve düşüncelerinin rahatlıkla analiz edilebilmesi, kullanıcıların konumsal bilgilerinden analizler yapılabilmesi gibi pek çok konuda fayda sağlamaktadır. Bu nedenle, arařtırmacılar, yöneticiler ve özel sektör temsilcileri, marka takibi firmaları, anket firmaları başta olmak üzere toplumun neredeyse her kesiminin ilgisini toplamaktadır. Sosyal medya madencilięi ile cevap verilebilecek soruların bir kısmı ařağıdaki gibidir;

- Bir konu hakkındaki sosyal medya kullanıcıları ne düşünüyor?
- Bir olayla ilgili duyguları nelerdir?
- Hangi siyasi parti veya markayı seviyor veya sevmiyorlar?
- Hangi etiketler yaygın hale gelmektedir?
- Kimler, kimlerle arkadaşır?
- Kimler, kimleri takip ediyor ancak takip edilmiyor?
- Hangi hesaplar, sahte hesap olabilir?
- Hangi kullanıcıların paylaşımları daha fazla ilgi uyandırıyor?
- Hangi kullanıcılar, arkadaşlık aęlarında daha popülerdir?
- Hangi kullanıcılar aęlarını yönetmektedir?
- Hangi kullanıcılar benzer paylaşımlar yapıyor?
- Hangi bölgelerde yoğunluk oluşmaktadır?
- Hangi kullanıcılar, bir grup oluşturmaktadır?
- Hangi kullanıcılar, hangi lokasyonlarda bulunuyorlar?
- Hangi lokasyonlar, kullanıcılar arasında ortakır?
- Hangi lokasyonlar, sıklıkla turist kullanıcılar tarafından ziyaret ediliyor?
- Hangi lokasyonlar, bir kullanıcı grubu için önemlidir?

- Hangi lokasyonlar önem kazanmakta, hangi lokasyonlar önemsiz hale gelmektedir?

Yukarıda sayılan sorular, sosyal medya madenciliğinin cevap verebileceği soruların sadece bir kısmıdır. Sosyal medya verisini kullanarak, sosyolojik, metinsel, semantik, ağ, mekân ve zamansal analizler yapılabilmektedir. Bu analiz sonuçları da topluma, yönetimlere ve özel işletmelere fayda sağlamaktadır. Bu yönüyle sosyal medya madenciliği, çıkardığı anlamlı ve önemli örüntülerle son yılların önemli konularından ve araştırma alanlarından biridir.

1.3.Sosyal Medya Madenciliğinin Zorlukları

Sosyal medya madenciliğinin yukarıda sayılan avantajlarının yanında çeşitli zorlukları da vardır. Bu zorluklar aşağıda belirtilmiştir.

Büyük Veri Özelliği: Sosyal medya verileri büyük veri kaynaklarından biridir. Dünya üzerinde sosyal medya ağlarını kullanan ve bu ağlar için içerik üreten milyarlarca insan vardır. Ayrıca, bir kullanıcının birden fazla sosyal medya hesabı olduğu da düşünüldüğü zaman bu veri kaynaklarının doğru yöntemlerle ele alınması gerekmektedir.

Örnek Sayısı: Sosyal medya verileri, sosyal medya ağlarının uygulama programlama arayüzleri (application programming interface – API) aracılığıyla toplanmaktadır. Sosyal medya ağları bünyelerindeki tüm kullanıcı verilerini paylaşamamakta, sadece belirli bir oranının toplanmasına izin vermektedir. Bu durumda, yeterli sayıda örnek toplanmasını zorlaştırmaktadır.

Önişlem Aşamaları: Veri madenciliği metotlarının başarılı sonuçlar üretebilmesi için önişlem aşamalarından geçmesi gerekmektedir. Sosyal medya madenciliği için de aynı durum söz konusudur. Ancak, sosyal medya verisi içerisinde, insan kaynaklı olduğu için gürültü kaynakları fazladır. Örneğin, sahte hesapların tespiti için yenilikçi yöntemlerin geliştirilmesi gerekmektedir. Veri içerisindeki metin içeriklerinin dikkatli analiz edilmesi gerekmektedir. Kullanıcıların sanal özel ağlar (virtual private network – VPN) aracılığıyla konum bilgilerini yanlış girmeleri durumunun incelenmesi gerekmektedir.

Değerlendirme Kriterleri: Veri madenciliğinde örüntülerin değerlendirilmesi için standart yöntem, veriyi eğitim ve test verisi olarak ayırıp, örüntülerin başarısının analiz

edilmesidir. Ancak, sosyal medya madenciliği için genellikle temel doğru örnek bulmak mümkün değildir. Bu nedenle keşfedilen örüntülerin değerlendirilmesi için farklı yöntemler gerekmektedir.

Veri Boyutu: Sosyal medya verisi, içerisinde çok fazla sayıda parametre barındıran bir veri türüdür. Kullanıcılara ait kişisel bilgilerden, kullanıcılar arası ve gruplar arası bilgilere kadar çok geniş bir yelpazede parametresi bulunmaktadır. Bu boyuttaki verinin içerisinde hangi parametrelerin daha önemli, hangi parametrelerin daha az önemli olduğunun tespiti önemli bir problemdir. Örneğin, Twitter, bir *tweete* ait yaklaşık 150 parametre sunmaktadır. Bu parametrelerin bir kısmı tweet içeriğiyle ilgili, bir kısmı da tweeti paylaşan kullanıcıyla ilgilidir. 1 milyon tweetin analiz edildiği bir uygulamada tüm parametrelerin toplanması veri karmaşasına yol açacaktır. Uygulama ihtiyaçlarına en uygun sayıda parametrenin belirlenmesi, sosyal medya madenciliğinin zorluklardandır.

Mekân ve Zaman Bilgisinin Analizi: Sosyal medya verileri içerisinde, mekân ve zaman bilgileri de bulunmaktadır. Bu bilgilerin nasıl kullanılacağı ve hangi hesaplamalardan geçeceği de sosyal medya madenciliğini zorlaştırmaktadır. Mekânsal analiz ve zaman bilgisiyle zenginleştirilmiş mekân-zamansal analizler yoğun hesaplama karmaşıklığı içermekte ve yeni ve özgün değerlendirme ölçütlerinin geliştirilmesi gerekmektedir.

1.4.Sosyal Medya Madenciliği Metotları

Sosyal medya madenciliği, içerdiği veri türlerine göre yapılan analizler için metotlar geliştirilmiştir. Örneğin, sosyal medya kullanıcılarının birbirleriyle olan arkadaşlık ilişkilerinin analizi için ağ analizi metotları, kullanıcıların mekânsal bilgilerinin analizi için mekânsal örüntü madenciliği metotları geliştirilmiştir. Bu metotların temelleri, klasik veri madenciliği, istatistik ve matematik modellere dayalıdır.

Sosyal medya madenciliği metotlarını alt başlıklar halinde inceleyecek olursak; sosyal ağ analizi, duygu analizi ve düşünce madenciliği, topluluk analizi, sosyal tavsiye sistemleri ve mekânsal ve mekân-zamansal analizler olarak sınıflandırabiliriz [5] [2].

1.4.1. Sosyal Ağ Analizi

Sosyal medya kullanıcılarının birbirleriyle olan arkadaşlık ilişkileri, sanal bir ağ oluşturmaktadır. Bu ağların analizi de kullanıcıların tercihleri, birbirleriyle olan

yakınlıkları, potansiyel arkadaşların ve aynı arkadaşla sahip olan kişilerin tespiti, birbirleriyle ilişkili olan grupların tespit edilmesi, grupları birbirlerine bağlayan kritik kullanıcıların belirlenmesi gibi pek çok konuda bilgi sağlamaktadır. Sosyal ağ analizi için kullanılan metotlar, genellikle çizge (graf) madenciliği metotları olmaktadır. Çizge madenciliği, her bir kullanıcıyı düğüm olarak ve kullanıcılar arasındaki ilişkileri de çizgiler olarak görüp, analizleri işletmektedir.

1.4.2. Duygu Analizi ve Düşünce Madenciliği

Duygu analizi ve düşünce madenciliği, kullanıcı içeriklerinden duygu ve düşüncelerin otomatik olarak çıkarılmasını amaçlamaktadır. Sosyal medya ağlarındaki kullanıcıların paylaşımları da içerik olarak görülmekte ve bu analizler yapılabilmektedir. Duygu analizi ve düşünce madenciliği ile kullanıcıların bir marka hakkındaki düşünceleri, bir konuyla ilgili duyguları, bir uygulamaya karşı tepkileri gibi pek çok semantik bilgi çıkarılabilir. Bu analizler için metin madenciliği metotları kullanılmaktadır.

1.4.3. Topluluk Analizi

Topluluklar, içerisindeki kullanıcıların, topluluk dışındaki kullanıcılardan daha fazla etkileşimde bulunduğu gruplar olarak tanımlanabilir. Sosyal medya ağları da doğal olarak birer topluluk olarak görülebilir. Sosyal medya kullanıcılarının oluşturdukları manuel topluluklar, bir etikete destek veren kullanıcılar, bir konuda aynı düşüncede olan kullanıcılar, aynı kullanıcıları takip eden kullanıcılar birer topluluk oluşturmaktadır. Toplulukların analizi de bir konuda topluluğa üye olan kullanıcıların ne düşündükleri ve bir uygulamaya ne tür bir karşılık vereceklerinin analizi için fayda sağlamaktadır.

1.4.4. Sosyal Tavsiye Sistemleri

Geleneksel tavsiye sistemleri, bir kullanıcıya bir ürün tavsiye ederken, eldeki kullanıcı grubunun cevapları veya kullanıcının geçmiş bilgilerini kullanmaktadır. Ancak sosyal medya verisinin kullanıcılar hakkında sağladığı geniş bilgi birikimi kullanılarak tavsiye sistemleri düzenlenebilir. Aynı özelliklere sahip olan diğer kullanıcıların tercihleri veya kullanıcının geçmiş paylaşımları analiz edilerek, kullanıcıya en uygun içerik ve ürünün tavsiye edilmesi mümkün hale gelmiştir. Aynı zamanda, arkadaş tavsiye sistemleri, lokasyon tavsiye sistemleri, reklam görüntüleme sistemleri de sosyal tavsiye sistemlerinin

uygulamalarıdır. Tavsiye sistemleri için birliktelik analizi yöntemleri kullanılabilmektedir.

1.4.5. Mekânsal ve Mekân-Zamansal Analiz

Mekânsal ve mekân-zamansal sosyal medya madenciliği, sosyal medya kullanıcılarının paylaşımlarının mekân ve zaman bilgisini kullanarak anlamlı örüntülerin keşfedilmesini amaçlamaktadır. Sosyal medya kullanıcılarının sağladığı mekân ve zaman bilgisiyle, kullanıcılar hangi lokasyonları ziyaret etmektedir, sıklıkla ziyaret edilen lokasyonlar nerelerdir, bir kullanıcı grubu için hangi lokasyonlar ortaktır, lokasyonların zamansal olarak ziyaret edilme durumu nedir gibi pek çok soruya cevap bulunabilmektedir. Mekânsal ve mekân-zamansal analizler için örüntü keşfi, eş oluşum madenciliği metotları yaygın olarak kullanılmaktadır.

2. BÖLÜM

BULUT BİLİŞİM

Bu bölümde bulut bilişime ait temel tanım yapılmış, özellikleri ve avantaj ve dezavantajları anlatılmıştır. Daha sonra, bulut bilişim servis modelleri tanıtılmıştır. Ardından, bulut bilişim türleri ve bulut bilişim sistemlerindeki veri türleri incelenmiştir. Son olarak, bulut bilişim sistemlerinde veri analizi için yaygın olarak tercih edilen Apache Hadoop platformu anlatılmıştır.

2.1.Bulut Bilişim Tanımı ve Özellikleri

Bulut bilişim için henüz fikir birliğine varılmış bir tanımlama yoktur. Geniş perspektiften bakıldığında, servis sağlayıcı veri merkezindeki donanım ve yazılım kaynaklarıdır [6]. Bulut bilişim, minimum yönetim gayreti veya servis sağlayıcı etkileşimi ile hızlıca ücretlendirilebilen ve serbest bırakılabilen, ayarlanabilir hesaplama kaynaklarının (örneğin, ağlar, sunucular, depolama, uygulamalar ve servisler) paylaşımlı havuzuna yaygın, pratik ve anlık ağ erişimi sağlayan bir modeldir [7]. Bulut bilişim, servis sağlayıcı ve müşterisi arasında servis antlaşması kurulmuş olan, dinamik olarak ücretlendirilebilen ve bir veya birden fazla birleştirilmiş hesaplama kaynaklarına sahip birbirine bağlı ve sanallaştırılmış bilgisayarlar içeren paralel ve dağıtık hesaplama sistemidir [8].

Daha önce dağıtık ve paralel programlama modelleri ile aynı amaca hizmet etmekte ancak işlevsel olarak bazı farklılıkları bünyesinde barındırmaktadır. Gereken hesaplama gücünün uzak sunucular aracılığıyla sağlanması açısından aynı ortak paydaları bulunmasına karşın bu hesaplama gücünün temini konusunda bulut bilişim diğer modellerden ayrılmaktadır. Bulut bilişimin en önemli özelliği istendiğinde eklenip çıkarılabilen depolama, altyapı ve yazılımın sağlandığı esnek sunucu modeli olmasıdır.

2.1.1. Bulut Bilişim Sistemlerinin Özellikleri

Bulut bilişimin temel bazı karakteristik özellikleri bulunmaktadır. Hesaplama ve depolama için klasik sunucu sistemlerinden farklı yaklaşım sergilemektedir. Bu farklılıklar ve özellikleri [9] [10];

- Sanallaştırılabilirlik
- İstendiğinde, kendi kendine kaynak ayırma
- Kullandığın kadar öde modeli
- Sanal depolama
- Genel bulut ortamı arayüzleri
- Sanal ağlar
- Dinamik kaynak sağlama
- Sanal kümelemeler
- Yüksek seviyede erişilebilirlik ve veri kurtarmadır.

2.1.2. Bulut Bilişimin Avantaj ve Dezavantajları

Bulut bilişim sistemleri, klasik sunucu sistemlerinden çeşitli noktalarda ayrılmakta ve daha büyük faydalar sağlamaktadır. Bulut bilişim sistemlerinin avantajları:

- Düşük Yazılım ve Donanım Maliyeti
- Gelişmiş Performans
- Anlık Güncelleme
- Sonsuz Depolama Kapasitesi
- Artırılmış Veri Güvenliği
- İşletim Sistemleri Arasında Geliştirilmiş Uyum
- Uyumlu Dosya Formatlarıdır.

Bulut altyapısı tarafından donanım ve yazılımlar sağlandığı için bu sistemler için ayrıca herhangi bir ücret ödenmesi söz konusu olmamaktadır. Ayrıca profesyonel ekipler tarafından sağlanan yazılım ve donanım desteği ile daha performanslı sistemler elde etmek mümkün olmaktadır. Bunlara ek olarak, sunuculardaki veri güvenliğini profesyonel ekipler sağladığı için, kişisel sunuculardan daha güvenli bir ortam sunulmaktadır.

Bulut bilişim sistemleri, uzak sunucularda çalışmaları için çeşitli noktalarda dezavantajları bulunmaktadır. Bulut bilişim sistemlerinin dezavantajları:

- Sabit İnternet Bağlantısı
- Düşük Bağlantı Hızında Yavaş Çalışma
- Güvenlik Açıkları
- Sistem Güncellemeleri
- Deneyimsiz Bulut Operatörüdür.

Bulut bilişim sistemlerinde internet üzerinden erişim sağlandığı için sürekli internet bağlantısı gerekmektedir. Bu nedenle düşük internet hızı ile erişim sağlamak uygulamaların yavaş çalışmasına sebep olur. Her ne kadar profesyonel ekipler tarafından güvenlik sağlansa da, küçük bir hata büyük zararlara yol açabilmektedir. Bulut altyapısı güncellendiğinde, bu altyapı üzerindeki uygulama düzgün çalışmayabilir.

2.2.Bulut Bilişim Servis Modelleri

Bulut bilişim servis modelleri [7], bulut sağlayıcısının donanım ve yazılım kaynaklarından kullanıcının ihtiyacına göre yetkilendirilmesi ve erişiminin sağlanması için oluşturulmuş modellerdir. Hangi seviyede bir modele ihtiyaç duyulduğu tespit edildikten sonra bulut sağlayıcıları aracılığıyla amaca en uygun şekilde hizmet edecek olan servis modeli kiralanabilmektedir. Yazılımdan altyapıya geçişte yetki ve özgürlükler arttıkça ücretlendirme de buna paralel olarak artacaktır.

Sayısı konusunda farklı görüşler bulunsa da temel olarak üç tür servis modelinden bahsedilebilir [7]:

- **Servis olarak Yazılım (Software as a Service, SaaS):** Kullanıcıların uygulamalara erişmek için kendi sistemlerine herhangi bir kurulum yapmalarına gerek kalmadan, internet tarayıcıları aracılığıyla bulut bilişim sistemleri üzerindeki uygulamalara erişerek çalışma yapabilmeleridir. Müşteriler alt yapıdaki ağ, sunucu, işletim sistemi ve depolama aygıtları gibi bileşenleri yönetmez veya denetlemez. Sadece kullandıkları uygulamalara özgü ayarlamalar yapabilirler.
- **Servis olarak Platform (Platform as a Service, PaaS):** Servis sağlayıcı, müşteriye kendi uygulamasını geliştirip, çalıştırabileceği bir platform sunar. Bu

platform uygulamanın geliştirileceği, çalıştırılacağı ortamla birlikte, tamamlayıcı servisleri ve gerekli teknolojik altyapıyı da kapsar. Kullanıcının kendi kurduğu uygulama dışında, platform altyapısını oluşturan bileşenler üzerinde herhangi bir kontrolü ve yönetim imkânı yoktur.

- **Servis olarak Altyapı (Infrastructure as a Service, IaaS):** Altyapının bir bulut servisi olarak sunulması modelinde müşteri ihtiyacı olan işlemci, depolama, ağ kaynağı ve diğer temel bilişim kaynaklarını kendisi yapılandırabilmekte ve bunların üzerine ihtiyacı olan işletim sistemi ve uygulamaları kurabilmektedir. Müşterinin ağ yapısı üzerinde yönetimi ve tam bir kontrolü olmamasına rağmen, işletim sistemi seviyesinde sisteme tam bir hâkimiyeti bulunmakta ve bazı ağ bileşenlerini (Firewall benzeri) yönetebilmektedir.



Şekil 2.1. Bulut bilişim servis modelleri

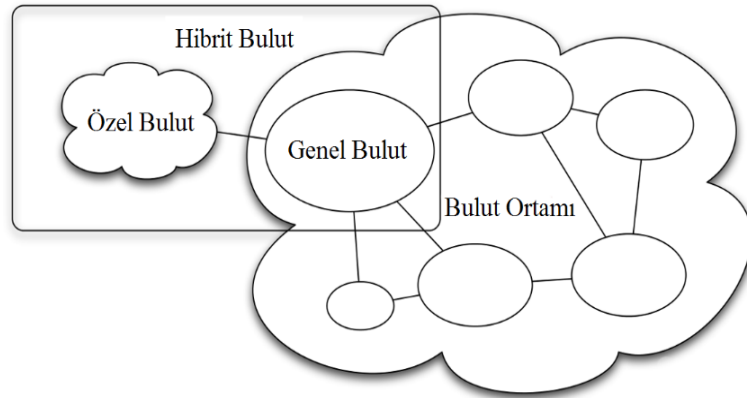
Şekil 2.1’de bulut bilişim servis modellerinin birbirlerine göre konumları ve örnekleri verilmiştir. Örneğin bir servis olarak platform modelinde, veritabanı sistemleri, web servisleri ve geliştirme araçları bulunurken bir servis olarak yazılım modelinde, ofis yazılımları, eposta sistemleri ve sanal masaüstü sistemleri bulunmaktadır. Piramidin en altındaki servis olarak altyapı, kendisinden daha yukarıdaki elemanlar için temel altyapıları sağlamaktadır. Piramidin tabanından tepesine doğru yaklaştıkça ilgilenilmesi gereken teknik detaylar azalmakta ancak bununla beraber özgürlükler de kısıtlanmaktadır. Servis olarak yazılım modeli tarafından sağlanan eposta sistemlerini kullanabilmek için servis olarak altyapı modelindeki ağ altyapısının sağlanması ve ardından servis olarak platform modelindeki web sunucularının ilgili sunucuya kurulmuş olması gerekmektedir. Ardından eposta yazılımına servis olarak yazılım aracılığıyla ulaşılabilir.

2.3.Bulut Bilişim Türleri

Bulut bilişim genel, özel ve hibrit bulut olmak üzere üç farklı tür altında incelenebilir [6] [7].

- **Genel Bulut (Public Cloud):** Genel bulutta servis sağlayıcı, uygulamalar ve depolama alanlarını internet üzerinden açık bir şekilde sağlar. Genel buluta en iyi örnek kullanıcılar tarafından yüklenen Youtube videolarıdır. Video kaynakları ve bilgileri bulut sağlayıcısı tarafından organize edilir ve internet erişimi olan herkes tarafından erişilebilir.
- **Özel Bulut (Private Cloud):** Özel bulutta servis sağlayıcı tarafından sağlanan depolama ve uygulamalar üzerine kullanıcının kendi güvenliğini sağlama yetkisi vardır. Bu sayede özel bilgiler ve veriler, bulut sağlayıcısı dâhil üçüncü şahıslar tarafından erişilemez. Özel buluta örnek Google Drive hizmeti olarak verilebilir. Kişisel dosyalar Google sunucuları tarafından depolanır ve sadece ilgili kullanıcı tarafından erişilebilir ve düzenlenebilir.
- **Hibrit Bulut (Hybrid Cloud):** Hibrit bulutta bulut sağlayıcısından kiralanan alanın tamamının güvenliği sağlanmaz, sadece kullanıcıya ait özel bilgilerin güvenliği sağlanır. Herkes tarafından erişilebilir alanlarda ise özel bilgiler saklanmaz. Örnek olarak Google Drive üzerinden paylaşımına açılan dosyalar verilebilir. Gizli kalması istenen dosyalar paylaşımına açılmayabilir ancak belirli dosyalar Google Drive üzerinden paylaşımına açılabilir.

Bulut bilişim türlerini daha iyi anlayabilmek bir web sitesi örnek olarak verilebilir. Bir web sitesinin giriş sayfası herkes tarafından görülebilir olmakla birlikte kullanıcılara ait bilgiler, giriş yapıldıktan sonra görülebilmektedir. Herkes tarafından görülebilen kısımları genel bulut, kullanıcı girişi yapıldıktan sonra görülebilen kısımları ise özel bulut olarak görülebilir. Eğer web sitesi bu iki türün birleşiminden oluşuyorsa da hibrit bulut yapısına bir örnek olur.



Şekil 2.2. Özel, genel ve hibrit bulut ortamları

Şekil 2.2’de özel, genel ve hibrit bulut ortamları gösterilmiştir. Özel bulut ortamı, bulut sistemlerine entegre olmayan ve kendi içerisinde hesaplama ve depolama altyapılarını barındıran bir bulut türüdür. Genel bulut ise hesaplama ve depolama altyapılarını bir bulut sağlayıcısından sağlayan bir bulut türüdür. Bazı özel durumlarda, hem özel hem genel bulut ortamları birlikte kullanılmak istenebilir. Bu yaklaşıma da hibrit bulut adı verilmektedir. Hibrit bulut, bir yandan özel bulut altyapısına sahip olurken, diğer yandan bazı hesaplama ve depolama ihtiyaçlarını genel bulut sağlayıcılarından temin edebilen bir ortam sunmaktadır.

2.4.Bulut Bilişimde Veri Türleri

Günümüzde gelişen teknolojiler – uydu ve sensör teknolojisi, internet teknolojileri, mobil cihazlardaki gelişmeler – depolanan veri miktarını ve türünü çok büyük ölçüde artırmaktadır. Daha önce sadece metin olarak depolanabilen veri formatları günümüzde fotoğraf, ses veya video olarak da depolanmayı desteklemektedir. Ayrıca çeşitli alanlarda bilgisayar teknolojilerinin etkin kullanılmaya başlanması da veri boyutunu artıran bir etmendir. Örneğin coğrafya verisi harita mühendisliği için önemlidir. Belli bir bölgedeki tüm mekânsal parametrelerin depolanması ve erişiminin maliyeti fazladır. Yeni verilerin depolanması ihtiyacı ortaya çıktığı zaman, eğer eldeki alan yetersiz ise yeni donanım satın almak ve bu donanımın sunucuya entegre edilmesi gerekmektedir. Ancak bulut sistemlerinin sağladığı esnek depolama modeli ile dakikalar içerisinde yeni depolama alanları satın alınarak yeni verilerin kaydı kolayca gerçekleştirilebilir.

Bulut bilişim altyapıları, üzerinde her türden veriyi depolama ve işleme yeteneğine sahiptir. Klasik sunucularda, veriyi saklamak için sabit bir depolama alanı ayrılır ve eğer

yeni veriler depolamak gerekiyor ise fazladan donanım satın almaya ihtiyaç duyulurdu. Bulut bilişim sistemleri sayesinde, yeni depolama alanları dakikalar içerisinde ayrılabilmekte ve kullanılabilir.

Dijital olarak depolanan verilerin bazı karakteristik özellikleri bulunmaktadır. Bu özelliklerine göre dijital veri yapısal (structured), yapısal olmayan (unstructured) ve yarı-yapısal (semi-structured) veri olmak üzere üç kısma ayrılabilir. Bu bölümde yapısal, yapısal olmayan ve yarı-yapısal veri türleri anlatılmış, bulut verisinin karakteristik özellikleri verilmiş, bulut veri türleri ve mahrem ve gizli verilerin güvenliğinin sağlanması tartışılmıştır.

2.4.1. Yapısal Veri

Yapısal veri genellikle öntanımlı bir tür, boyut ve formata sahip olan veriyi ifade etmektedir [11]. Genellikle veri tabanında tutulan, belirli bir yapıya sahip olan veridir. Yapısal veri öncelikli olarak veri modeline dayanır. Yapısal verinin okunması, işlenmesi ve anlaşılması kolaydır.

Yapısal veri, bilinçli olarak verinin toplanması temeline dayanır. İki tür yapısal veri kaynağı bulunmaktadır; bunlar bilgisayar- veya makine-kaynaklı ve insan-kaynaklı olarak isimlendirilebilir. Makine-kaynaklı veri genellikle, insan müdahalesi olmadan makine veya bilgisayar tarafından üretilen veriyi ifade etmektedir. İnsan-kaynaklı veri ise bilgisayar aracılığıyla insanlar tarafından üretilen veriyi ifade eder.

Makine-kaynaklı yapısal veri türlerinin bazıları aşağıda listelenmiştir:

- Sensör verisi
- Web log verisi
- Alışveriş verisi
- Finans verisi
- ...

İnsan-kaynaklı yapısal veri türlerinin bazıları aşağıda listelenmiştir.

- Tıklama verisi
- Oyun-bağlantılı veri
- ...

2.4.2. Yapısal Olmayan Veri

Özel bir tür, boyut veya formata sahip olmayan veriye yapısal olmayan veri denir [11]. Günümüz dünyasında saklanan verilerin büyük çoğunluğu yapısal olmayan verilerden oluşmaktadır. Bu nedenle yapısal olmayan verinin işlenmesi, içerdiği bilgi potansiyeli nedeniyle önem taşımaktadır. Son yıllara kadar yapısal olmayan veri, üzerinde durulmayan ve analizi için çaba sarf edilmeyen bir veri türü olmuştur. Günümüzde büyük boyutlardaki verinin analizi ihtiyacı, yapısal veriden taviz verilmesini ve hızlı bir şekilde verinin analizini gerçekleştirebilmek için verinin yapısal olmayan bir şekilde depolanmasını gerektirmiştir. İçerdiği potansiyel bilgi kapasitesi ile analiz edilme sebebi olduğu için depolanması ve işlenmesi önem kazanmıştır.

Yapısal olmayan veride, yapısal veriyle aynı veri kaynakları bulunmaktadır; makine-kaynaklı ve insan kaynaklı.

Makine-kaynaklı yapısal olmayan veri türleri:

- Uydu resimleri
- Sismik, atmosferik ve yüksek enerjili fizik verisi
- Fotoğraf ve video verisi
- Radar veya sonar verisi
- ...

İnsan-kaynaklı yapısal olmayan veri türleri:

- Özel metin verileri
- Sosyal medya verisi
- Mobil veri
- Web sitesi içerikleri
- ...

Yapısal olmayan veri, yapısal verinin çok üstünde bir yaygınlığa sahiptir. Pek çok uzman, günümüzde depolanan verinin yaklaşık %80'ini yapısal olmayan verinin oluşturduğu ve yapısal veriden çok daha hızlı artış gösterdiği konusunda fikir birliğine varmıştır [11]. İçerdiği potansiyel bilgi zenginliğiyle yapısal olmayan verinin depolanması ve analiz edilmesi yapısal olmayan veriyi cazip hale getirmektedir.

2.4.3. Yarı-Yapısal Veri

Yarı-yapısal veri (semi-structured data), yapısal ve yapısal olmayan veri türlerinin aynı kayıt içerisinde birlikte kullanılmasıyla ortaya çıkan bir veri türüdür. Yarı-yapısal veride, belli parametreler yapısal veri özelliğinde, belli parametreler ise yapısal olmayan veri özelliğindedir. Kimlik numarası, IP adresi gibi zorunlu alanlar yapısal veri olarak depolanmakta; iş geçmişi, kişisel bilgiler gibi zorunlu olmayan alanlar yapısal olmayan veri olarak depolanmaktadır. Yarı-yapısal veriler, yapısal olmayan veriler içerisinde sayılabileceği gibi ayrıca da sınıflandırılabilir.

2.4.4. Bulut Verisinin Karakteristik Özellikleri

Bulut altyapısı üzerinde tutulan verinin genel bazı karakteristik özellikleri vardır. Karakteristik özelliklerine göre bulut verisi, geleneksel veriden verinin depolanması, boyutu, erişimi gibi konularda farklılıklar göstermektedir. Aşağıda bulut verisinin karakteristik özellikleri incelenmiştir.

1. Bulut üzerinde tutulan veri genellikle yapısal olmayan veriden oluşur.
2. Büyük boyutlu veya hızla artan eğilimdedir.
3. İnternet üzerinden erişilir.
4. Erişimi maliyet gerektirir.
5. Genellikle tek merkezde tutulmaz, farklı sunucularda dağıtılmış haldedir.
6. Erişimi esnasında güvenlik kontrolü gerekir.
7. Veri erişim hızı, internet bant genişliğine bağlıdır.
8. Verinin güvenliği bulut sağlayıcıları aracılığıyla sağlanır.
9. Sürekli olarak bulut sağlayıcısına, depolanan verinin boyutuyla orantılı olarak ödeme yapılması gerekir.
10. Bulut türüne bağlı olarak, veri gizli veya açık tutulur.

2.5. Apache Hadoop

Hadoop, sıradan sunucu kümeleri üzerinde büyük verileri güvenilir, ölçeklenebilir ve dağıtık bir şekilde işlemek için geliştirilmiş açık kaynak kodlu bir projedir [12] [13] [10]. Proje içerisinde dört modül bulunmaktadır;

- **Hadoop Common:** Diğer modüller için ortak hizmetleri sağlar.

- **HDFS (Hadoop Distributed File System):** Uygulama verilerine hızlı erişim için kullanılan dağıtık dosyalama sistemidir.
- **Hadoop YARN:** Görev zamanlama ve kaynak yönetimi için kullanılan çerçevedir.
- **Hadoop MapReduce:** Büyük ölçekli veri kümeleri için YARN tabanlı paralel veri işleme sistemidir.

Bu modüller arasından HDFS ve MapReduce, asıl işlemleri yapan modüller olarak değerlendirilebilir. Common ve YARN diğer iki modül için altyapısal ve çevresel destek sağlamaktadır.

HDFS [14], Hadoop ortamında veri depolama işlemleri için kullanılan dağıtık dosyalama sistemidir. HDFS, çok büyük boyutlu (en az terabayt ve genellikle onlarca terabayt seviyesindeki) veriler için kullanılmaktadır. Bu büyüklükteki veri tek bir kaynakta tutulamayacağı için birbirinden uzak sunucularda paylaşımlı ve yedekli bir şekilde depolanan dağıtık dosya sistemine sahiptir.

HDFS, en önemli veri analiz örüntüsünün bir-defa-yaz, çoklu-oku (write-once, read-many-times) yapısına sahip olduğu düşüncesine göre tasarlanmıştır. Bir veri kümesi, kaynakta bir defa oluşturulur ve birçok defalar erişilerek okunur. Bu şekilde daha geniş bir gruba hitap etmek hedeflenmiştir.

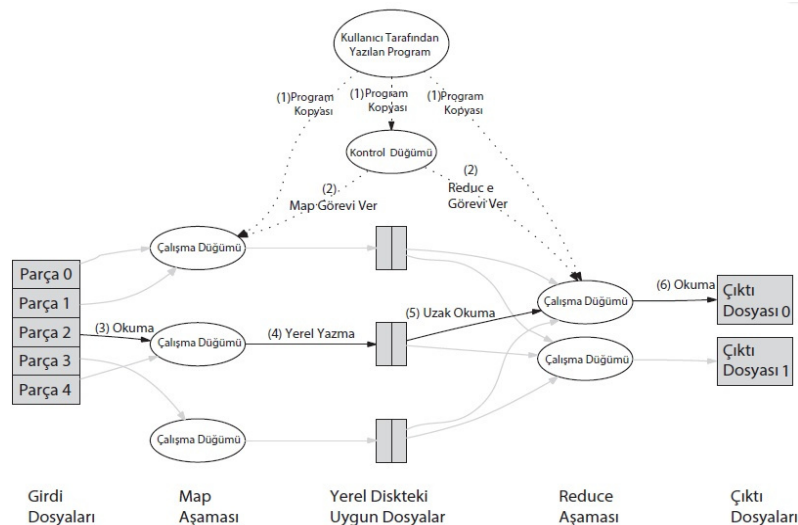
HDFS, pahalı ve yüksek güvenilirlikte donanımlar gerektirmemektedir. Nod bozulmalarının ve erişimlerin kopmasının yaygın olduğu sıradan sunucular kümesi üzerinde çalışmak üzere tasarlanmıştır. HDFS dosya sisteminde hata toleransı sağlamak için çeşitli yaklaşımlar geliştirilmiştir.

Hadoop, Java programı ile geliştirilmiştir ve Hadoop MapReduce işlemi, HDFS üzerinde daha önce anlatıldığı şekilde gerçekleştirilmektedir. Hadoop, *Map* ve *Reduce* görev parçacıklarını kümelerdeki nodlara dağıtarak, sonuçların elde edilmesinden sorumludur. Hadoop, *Map* ve *Reduce* işlemlerinin ikisini de aynı nod içerisinde gerçekleştirebildiği için ağ trafiği gecikmeleri engellenmiş olur ve aynı zamanda yüksek ölçüde ölçeklenebilirlik sağlanır.

2.5.1. Hadoop MapReduce

MapReduce, büyük veri kümeleri üzerinde veri-yoğun (data-intensive) görevleri yerine getirmek için geliştirilmiş olan bir programlama modelidir [15]. Başlangıçta Google tarafından web sayfalarının indekslenmesi için geliştirilmiştir. Anahtar-değer çifti (key-value pair) mantığıyla çalışmaktadır. Map ve Reduce fonksiyonları olmak üzere iki temel alt fonksiyonu bulunmaktadır. Map fonksiyonu $\langle \text{key}, \text{value} \rangle$ çiftlerinden $\langle \text{key}, \text{value} \rangle$ çift dizisi oluştururken, Reduce fonksiyonu $\langle \text{key}, \text{value} \rangle$ çift dizisinden $\langle \text{value} \rangle$ dizisi elde etmektedir.

Şekil 2.3'te görüldüğü üzere parçalanmış dosyanın her bir parçası Usta (Master) düğüm tarafından bir İşçi (Worker) düğümüne atanmakta ve **map** işlemini yapması beklenmektedir. Ardından orta katmandaki verilerin **reduce** işlemi yapılmak üzere Usta düğüm tarafından yine İşçi düğümler görevlendirilmektedir.



Şekil 2.3. MapReduce çalışma mimarisi [15]

MapReduce için yapılan verimlilik hesaplamalarına göre, bu model oldukça verimli çalışmaktadır [15]. İşlem parçacıklarının pek çoğunun işlem dışı kalması durumunda dahi, nispeten daha uzun sürse de işlemi başarılı bir şekilde tamamlanabilmektedir.

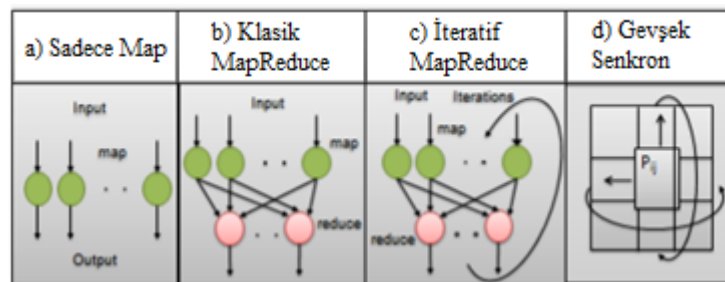
MapReduce mimarisinde hata toleransı iki kısma ayrılmaktadır. İşçi hatası ve Usta hatası. İşçi'lere sürekli ping atılarak hata olup olmadığı ve İşçi'nin halen sisteme dâhil olup olmadığı tespit edilmektedir. Usta hatası için de görevler sürekli loglandığı için eğer Usta hatası olursa sistem son kaldığı yerden çalışmasına devam edebilmektedir. Ancak

MapReduce İşçi için görev ataması yaptıktan sonra, sistemden kopup kopmadığını araştırmamaktadır. Ayrıca yazılımsal sorunlar nedeniyle İşçi'lerden bir kısmı doğru sonuç döndüremeyebilirler. Bu durumları da araştıran bir hata toleransı yaklaşımı geliştirmek daha etkili sonuçlar alınmasını sağlayacaktır.

MapReduce iyi bir kontrol modeline sahiptir. İşçi'lerin bir kısmı ellerindeki görevi bitirmeye yaklaştıklarında, Usta, devam eden görevler için yedek görevler oluşturur ve işlemini bitiren İşçi'lere atar. Birincil veya yedek İşçi, görevi bitirdiğinde ise görev tamamlanmış kabul edilir. Bu sayede, birkaç çok yavaş çalışan cihaz yüzünden tüm sistem yavaşlamamış ve işlem hızlı bir şekilde tamamlanmış olur.

MapReduce veri modelinde dışarıdan gelen girdi dosyası öncelikle parçalara ayrılmaktadır. Daha sonra parçalar <anahtar, değer> çifti haline dönüştürülerek işlemler devam ettirilmektedir. Veri modeli olarak daha etkili ve farklı yöntemler kullanılarak elde edilen sonuçların kalitesi artırılabilir.

MapReduce paralel işlem yapmakla birlikte tamamlanmamış görevleri beklemesi ve Reduce fonksiyonuna geçmek için Map fonksiyonlarının tüm İşçi'lerde tamamlanmasını beklemesi açısından eksiklikler içermektedir. Bu eksiklikleri giderebilmek için geliştirilen MapReduce versiyonları Şekil 2.4'te görülmektedir [16].



Şekil 2.4. MapReduce versiyonları [16]

Şekil 2.4'te görüldüğü üzere, dört temel MapReduce versiyonu bulunmaktadır. Sadece Map versiyonu, Reduce fonksiyonuna ihtiyaç duymayan ve sürekli Map fonksiyonunu çalıştırması gereken uygulamalar için ideal versiyondur. Klasik MapReduce, temel MapReduce modelini ifade etmektedir. İteratif MapReduce, Reduce fonksiyonunun çıktılarını Map fonksiyonuna girdi olarak gönderen ve kendini belirli bir döngü sayısına kadar tekrar eden versiyonu ifade etmektedir. Gevşek senkron MapReduce ise, klasik MapReduce modelinde Map ve Reduce fonksiyonlarının birbirlerinin durumları ve

çıktılarından haberdar olmamasına karşılık geliştirilmiş ve fonksiyonların birbirlerinin durumlarını bilmelerini sağlayan bir versiyondur.

MapReduce, bulut verisinin analizi için oldukça yaygın olarak tercih edilen bir yaklaşımdır. Sağladığı kolay programlanabilir fonksiyonları ve sade programlama çerçevesi sayesinde, hızlı bir şekilde verilerin analizi gerçekleştirilebilmektedir. Pig, Hive, JAQL, SCOPE, Sawzall ve Dryad MapReduce tabanlı veri sorgulama dillerine örnek olarak sayılabilir.

3. BÖLÜM

SOSYAL MEDYA AĞLARINDAN VERİ TOPLAMA

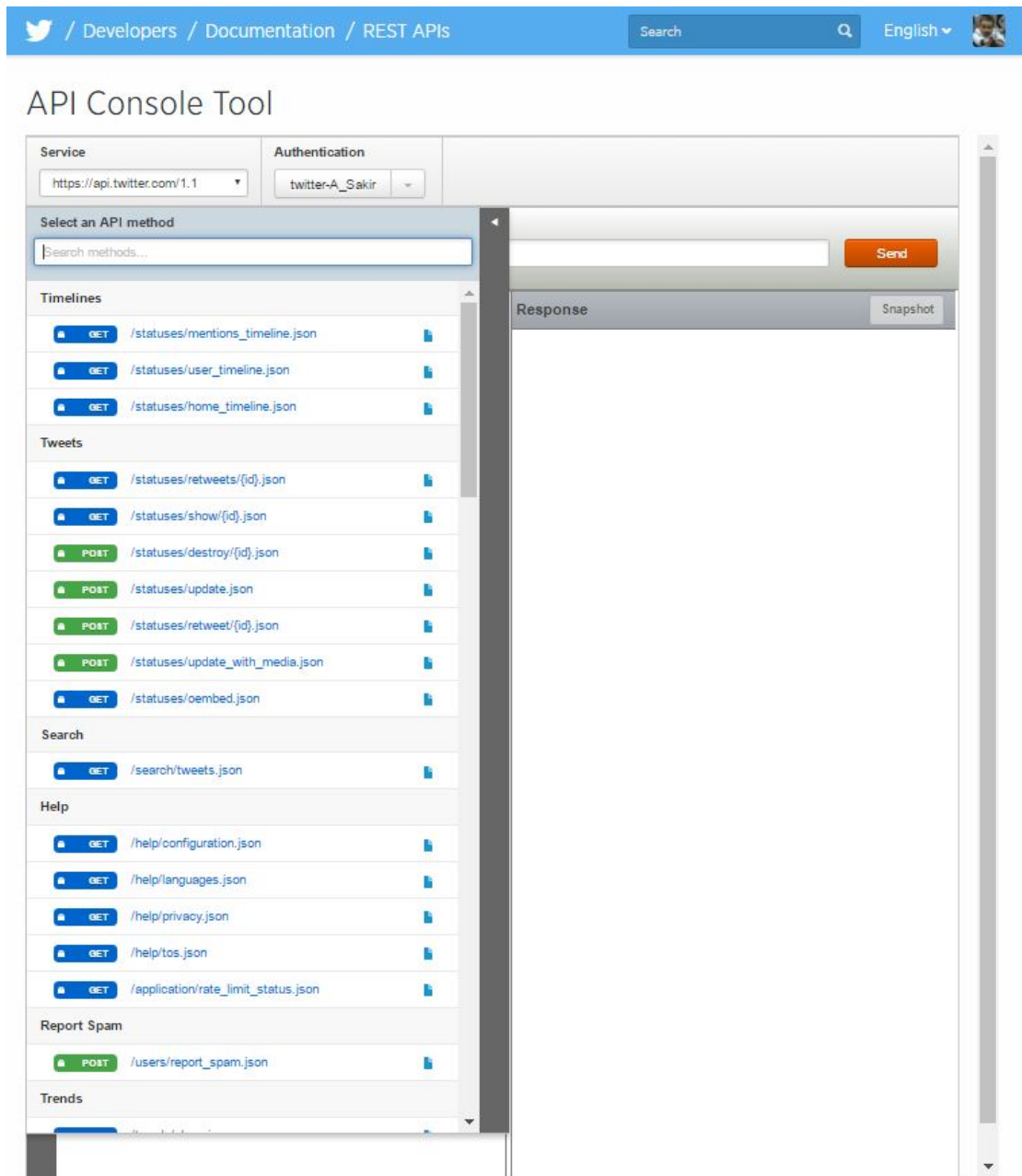
Sosyal medya ağlarından veri toplama işlemi, klasik web sunucularına bağlanarak bulunan verinin tamamını veya bir kısmının kişisel bilgisayarlara kopyalanması kadar kolay ve basit bir işlem değildir. Sosyal medya ağları, kullanıcılarının güvenliği için verilerinin gizliliğini sağlamak ile sosyal medya verisini kullanarak toplumsal, ekonomik ve akademik çeşitli faydalar sağlamak isteyen araştırmacıların bu verilere erişimini sağlamak arasında bir denge gözetmek durumundadırlar. Aynı zamanda, sosyal medya ağları, ellerindeki terabayt ve hatta petabayt boyutundaki verileri verimli bir altyapıyla, araştırmacılar ve verinin kullanıcılarına iletmek durumundadırlar. Bu da belirli miktardaki sunucu gücünü veri sağlama işlemine ayırmayı gerektirmektedir ve ek maliyetler ortaya çıkarmaktadır.

Sosyal medya ağları, hızlı, verimli ve izlenebilir bir şekilde veri iletimi sağlamak amacıyla kendi bünyelerinde oluşturdukları Uygulama Programlama Arayüzleri (Application Programming Interface, API) aracılığıyla ellerindeki verileri araştırmacı ve geliştiricilere iletmektedirler. Neredeyse tüm sosyal medya ağlarının kendi API'leri bulunmaktadır. Bu API'ler aracılığıyla, araştırmacı ve geliştiriciler, istedikleri türden, istedikleri şekilde ve otomatize edilmiş bir şekilde verileri toplayabilmektedirler.

Bu tez çalışması kapsamında sosyal medya veri kaynağı olarak Twitter seçilmiştir. Twitter'ın seçilmesinde 3 temel etmen bulunmaktadır [17];

- Twitter API'leri iyi tasarlanmış, erişimi kolay ve iyi açıklanmıştır.
- Twitter verileri, JSON formatında sunulmaktadır ve analiz için uygun formattadır.
- Veri kullanım koşulları açık ve iyi tanımlanmıştır.

Bu tez çalışması kapsamında sosyal medya ağlarından Twitter analiz edileceği için, Twitter API'lerinden veri toplama işlemi araştırılmıştır. Twitter, web kullanıcılarına herhangi bir ek işlem yapmadan API bağlantısı yapma hakkı vermektedir. Bu bağlantı yardımıyla, herhangi bir sorguyu tweetler arasından arayabilmekte, profili gizli olmayan kullanıcıların profil bilgileri ve paylaşımları çekilebilmekte, bir kullanıcının takipçileri toplanabilmektedir. Twitter'ın sosyal medya araştırmacıları ve geliştiriciler için sunduğu API ekranı Şekil 3.1'de verilmiştir. Bu API sistemine web arayüzleri aracılığıyla erişilebilmektedir ve otomatize edilebilir değildir.



Şekil 3.1. Örnek bir Twitter API ekranı

Şekil 3.1’de görülen, soldaki menüden hangi amaçla sorgu işletileceği seçilmektedir. Twitter’da sorguların işletilebilmesi için bir kimlik doğrulama işlemi yapılması gerekmektedir. Bu işlem menülerin üzerinde “Authentication” başlığı altında bulunmaktadır. Kimlik doğrulama işlemi, Twitter sunucularının aşırı yüklenmesini engellemek için oluşturulmuş bir sistemdir. Bu sistem sayesinde, bir kullanıcı çeşitli limitlerle sorgular yapabilmekte ve otomatize edilmiş sorgular için belirli bir süre (tanımlanan süre 15 dakikadır) beklenmelidir. Kimlik doğrulama işleminden sonra işletilecek olan sorgu seçildikten sonra, sorguya ait parametreler girilerek, sorgu çalıştırılır. Sorgu sonuçları, “Response” başlığı altında görüntülenmektedir. Twitter’dan yapılan sorgulara, JSON formatında sonuç döndürülmektedir.

Ancak bu web arayüzü aracılığıyla otomatize edilmiş sorgular işletilemeyeceği açıktır. Bu amaçla, bir program aracılığıyla istenen sorguların işletilmesi gerekmektedir. Twitter API’leri, yazılım programları aracılığıyla otomatize edilmiş sorguları da desteklemektedir. Ancak işletilecek olan sorguların her bir kullanıcı için 15 dakikalık aralıklarla limiti vardır. Örneğin, kullanıcı araması için 15 dakika içerisinde en fazla 180 sorgu yapılabilmektedir. Bir kullanıcının takipçilerini aramak için 15 sorgu yapılabilmektedir. Bu limitler içerisinde kalacak şekilde programlanmış veri toplama işlemleri yapılabilmektedir. API limitlerini incelemek için [18] numaralı referans kullanılabilir.

Bu tez çalışması kapsamında, Java programlama dili kullanılarak, Twitter API’lerinden veri toplanmıştır. Bu amaçla, Twitter4J Java kütüphanesi [19] kullanılmıştır. API web arayüzünden işletilen tüm sorgular, Twitter4J kütüphanesi aracılığıyla Java programlarından da işletilebilmektedir. Elde edilen sonuçlar, bir veritabanına veya bir metin dosyasına kaydedilebilir veya doğrudan analiz edilebilir formattadır.

Bu bölümde, bu tez çalışması için seçilmiş olan Twitter sosyal medya ağından veri toplama işlemi anlatılacaktır. Bu nedenle öncelikli olarak Twitter API’leri tanıtılacak, bu API’lerin nasıl kullanılacağı anlatılacak, Twitter4J açık kaynak kodlu Java kütüphanesinin veri toplama işlemi esnasında nasıl kullanıldığı gösterilecek ve bu tez çalışması için toplanan verilerin hangi sorgular oluşturularak toplandığı anlatılacaktır.

3.1. Twitter API'leri

Twitter, farklı hedeflere yönelik sorgular için üç farklı API oluşturmıştır; REST API, Streaming API ve Ads API [18]. Twitter sunucuları, gelen API sorgularına göre benzer sorgular işletiyor olsalar da her API, türüne göre farklı hedefler için sorgular içermektedir. Örneğin, REST API geçmişe dayalı bilgileri toplarken, Streaming API güncel sorgular yapılmasını sağlamaktadır. Bir veri toplama işlemi esnasında bu API'ler birlikte kullanılabileceği gibi, bir API'nin sonuçlarıyla diğer bir API'de sorgu işletilebilmek de mümkündür. Bu kısımda Twitter API'leri hakkında genel bilgiler bulunmaktadır.

3.1.1. REST API

REST API, Twitter'dan sorgunun işletildiği andan önceki verileri sağlamaktadır. En geniş sonuç ağına REST API sahiptir. Kullanıcılara ait tweetler, kullanıcıların birbirlerini takip durumlarına, küresel veya yerel ölçekteki arama sonuçlarına, trend olan etiketlerin çekilmesi gibi pek çok veriye REST API aracılığıyla erişilebilmektedir. Geçmişe dayalı veriler toplanacaksa ve kullanıcılara ait bilgilerin toplanması amaçlanıyorsa REST API kullanılabilmektedir. Ayrıca REST API kullanılarak Twitter web sayfasında kullanıcı olarak yapılabilen tüm işlemler programatik olarak yapılabilir.

3.1.2. Streaming API

Streaming API, bir sorgunun sürekli güncel sonuçlarını elde edebilmek için oluşturulmuş olan API'dir. Bu API aracılığıyla belli bir zaman aralığı içerisinde sürekli güncel sonuçlar toplanabilmektedir. Ancak belli bir limit sonra, bağlantının kapatılması gerekmektedir. Twitter üç çeşit yayın (stream) tanımlamıştır; genel yayın (public stream), kullanıcı yayını (user stream) ve grup yayını (site stream). Bu üç yayın arasından, çekilen sorgunun türüne göre seçim yapılmaktadır. Ayrıca Streaming API kullanılarak enlem, boylam ve yarıçapı verilen coğrafi sorgular da işletilebilmektedir. Bu sayede coğrafi bir bölgedeki grup yayınlarına ait sonuçlar toplanabilmektedir.

3.1.3. Ads API

Ads API, Twitter'ın reklamcılık platformunun kullanıcılara ait çözümler için kullanılabilesini sağlamaktadır. Daha çok işletmelerin, kendilerine özgü reklamları kendi sayfalarında Twitter Ads API aracılığıyla göstermesi temeline dayanmaktadır. Ads

API, üç türde kullanılabilmektedir; geliştirici (developer), temel (basic) ve standarttır (standard). Bu üç tür arasındaki temel fark yetkiler ve sorumluluklar temelinde değerlendirilmektedir.

3.2. Twitter4J Kütüphanesi

Twitter4J kütüphanesi [19], Java programlama dili için Twitter API'lerinin Java programları aracılığıyla toplanabilmesini sağlayan açık kaynak kodlu bir kütüphanedir. Bu kütüphane aracılığıyla, Twitter sunucularına bağlanarak, Java programlarına döndürülen sonuçlar elde edilebilmektedir. Twitter4J kütüphanesi sayesinde, temel kodlamalar yapmaya gerek kalmadan, hızlı ve verimli bir şekilde Twitter verileri Java programlarına aktarılabilmektedir.

Twitter4J kütüphanesi, JSON formatındaki verileri kendi içerisinde ayrıştırarak, kullanıcının istediği özelliklerde sonuçları programlara döndürebilmektedir. Bu sayede, verinin de JSON formatında analiz edilmesiyle çok yavaşlayacak algoritmalar önemli bir önışlem aşamasından kurtulmaktadır.

Twitter4J kütüphanesinin sağladığı en büyük avantaj, Twitter API'lerinden yapılan sorguların otomatize edilebilmesini sağlamasıdır. Teker teker sorgular işletmek ve bu sorguların sonuçlarını toplamak gibi zahmetli bir işlem yerine, düzenlenmiş ve sorgu sonuçlarının nasıl toplanacağı belirlenmiş bir program yazılarak Twitter API'lerinden istendiği kadar veri toplanabilmektedir.

3.3. Veri Toplama Sorguları

Bu tez çalışmasında geliştirilen sistemler ve algoritmalarda kullanılacak olan veriler, programatik bir yaklaşım ile toplanmıştır. Veri toplama işleminde kullanılan sorgu yaklaşımı, toplanan parametreler ve sırasıyla hangi aşamalarda ne tür işlemler yapıldığı bu kısımda anlatılacaktır.

Tez çalışması için yapılacak uygulamalar belirli bir coğrafi bölgedeki Twitter kullanıcılarının tespit edilmesi, bu kullanıcılar üzerinde önışlemlerin gerçekleştirilmesi ve bu kullanıcıların geçmiş tüm tweetlerinin toplanması olarak özetlenebilir. Coğrafi bölge içerisindeki kullanıcıların aramasını yapabilmek için Twitter'ın Streaming API'si kullanılmıştır. Bu API aracılığıyla x ve y koordinatları ve yarıçapı verilen bir coğrafi

bölge içerisindeki anlık paylaşımlar ve bu paylaşımları yapan kullanıcıların profil bilgileri toplanmıştır. Ardından toplanan kullanıcıların kimlik profil bilgileri üzerinden bir önışlem aşaması gerçekleştirilmiş ve sahte kullanıcılar ayıklanmıştır. Son olarak eldeki gerçek kullanıcılara ait tüm tweetler REST API aracılığıyla toplanmıştır. Tez çalışmasında geliştirilen veri toplama stratejisi ve sorgularının kaba kodu Algoritma 3.1’de verilmiştir.

Algoritma 3.1’de 1. adımda *coğrafi_arama* fonksiyonu ile istenen coğrafi konum ve yarıçap içerisinde konum bilgisi olan tweetler toplanmıştır. 2-6. adımlar arasında tweetlerinin dili Türkçe olan kullanıcılar seçilmiştir. Daha sonra 7-9. adımlarda *takipçi_bul* fonksiyonu ile seçilen kullanıcıların takipçileri toplanmıştır. 10-14. adımlarda toplanan takipçiler Türkçe tweet atıyor ve Türkiye’den iseler veri kümesine eklenmiştir. Son olarak 15-17. adımlarda *tweet_topla* fonksiyonu ile seçilen kullanıcıların tweetleri toplanmıştır.

Algoritma 3.1. Twitter veri toplama işlemi kaba kodu

```

1. tweetler = coğrafi_arama(koordinat_x, koordinat_y, yarıçap)
2. for tweet in tweetler
3.   if tweet_dili == Türkçe
4.     seçilen_kullanıcılar += tweet_kullanıcısı
5.   end if
6. end for
7. for kullanıcı in seçilen_kullanıcılar
8.   takipçiler += takipçi_bul(kullanıcı)
9. end for
10. for kullanıcı in takipçiler
11. if kullanıcı_dili(kullanıcı) == Türkçe and kullanıcı_lokasyon == Türkiye
12.   seçilen_kullanıcılar += kullanıcı
13. end if
14. end for
15. for kullanıcı in seçilen_kullanıcılar
16.   tweet_topla(kullanıcı)
17. end for

```

4. BÖLÜM

MEKÂNSAL VE MEKÂN-ZAMANSAL SOSYAL MEDYA MADENCİLİĞİ VE BULUT BİLİŞİM KONUSUNDA YAPILMIŞ OLAN ÇALIŞMALAR

Bu bölümde, mekânsal ve mekân-zamansal sosyal medya madenciliği ve bulut bilişim konularında yapılmış olan çalışmalar sunulmuştur. İlk olarak mekânsal sosyal medya madenciliği çalışmaları verilmiştir. Ardından mekân-zamansal sosyal medya madenciliği literatürü incelenmiştir. Daha sonra bulut bilişim sistemlerini kullanan sosyal medya madenciliği çalışmaları tartışılmıştır. Son olarak bu tez çalışması kapsamında yapılmış olan çalışmalar özetlenmiştir.

4.1.Mekânsal Sosyal Medya Madenciliği

Mekânsal sosyal medya madenciliği, Lokasyon Tabanlı Sosyal Ağ (Location Based Social Network – LBSN) verilerinin ortaya çıkışıyla mekânsal ve mekân-zamansal veri madenciliğinde yeni bir araştırma alanı olarak ortaya çıkmıştır [20].

Mekânsal ve mekân-zamansal veri madenciliği, veri madenciliğinin mekânsal ve zamansal veriyi ele almayı amaçlayan önemli bir alanıdır. Mekânsal eş-oluşum madenciliği [21-24] ve mekânsal kümeleme [25] [26], mekânsal ve mekân-zamansal veri madenciliğinin ana başlıklarındandır [27]. Bir eş-oluşum, komşu mekânsal alanlarda bulunan bir grup nesneyi tanımlamaktadır [28-30]. Eş-oluşan nesneler, sosyal medya ağlarındaki kullanıcılar veya lokasyonlar olarak ele alınırsa, eş-oluşum madenciliği yaklaşımları sosyal medya verisetlerini analiz etmek için kullanılabilir. Klasik mekânsal eş-oluşum madenciliği çalışmalarında nesnelerin ikili (boolean) olduğu, yani nesnelerin bir uzay veya zamanda ya varolduğu ya da olmadığı, varsayılır. Ancak bu yaklaşım, bir sosyal medya kullanıcısının bir lokasyonda birden fazla sosyal medya aktivitesine (tweet atmak veya Facebook'ta paylaşımda bulunmak gibi) sahip olabileceğinden dolayı, sosyal

medya verileri için doğru değildir. Bu nedenle, eş-oluşum madenciliği için geliştirilen algoritmalar ve değerlendirme ölçütleri sosyal medya verilerinden sosyal önemli lokasyonların madenciliği için uygun değildir.

Önemli lokasyonlar madenciliği çalışmaları iki kategoriye ayrılabilir: i) kullanıcı-seviyesinde sosyal önemli lokasyonlar madenciliği ve ii) grup-seviyesinde sosyal önemli lokasyonlar madenciliğidir.

Kullanıcı-seviyesinde önemli lokasyonlar madenciliği çalışmaları; lokasyon/rota tavsiyesi, arkadaş tavsiyesi, aktivite/tag tavsiyesi, ilginç/önemli lokasyonlar keşfi, rota madenciliği ve şehir ve hareketlilik örüntü madenciliği çalışmalarını içermektedir.

Kullanıcı-seviyesinde çeşitli araştırmacılar tarafından çalışılmıştır. [31-36]. Liu ve Seah [31] mobil kullanıcıların GPS verisine dayalı olarak önemli lokasyon (Point of Interest – POI) tavsiye metodu önermişlerdir. Önerilen metot, yoğunluk tabanlı kümeleme yaklaşımı kullanmaktadır. Bir lokasyonun tavsiye skorunu belirlemek için üç farklı faktör, popülerlik, zamansal ve mekânsal özellikler, kullanılmıştır. Doytsher vd. [32] bir sosyo-mekânsal ağda sıklıkla ziyaret edilen rotaları gösteren bir graf modeli tanımlamıştır ve GPS verisi kullanmıştır. Levandowski vd. ve Bao vd. [33] [34] kullanıcıların lokasyon-tabanlı oylamalarına göre lokasyon-duyarlı tavsiye sistemleri önermişlerdir. Ying vd. [35] POI tavsiyesi için kullanıcı giriş davranışlarını analiz eden UPOI-Mine algoritmasını önermişlerdir ve her bir kullanıcının POI'lerinin benzerlik skorlarını tahmin etmek için regresyon ağacı kullanılmıştır. Liu ve Xiong [36] lokasyon tabanlı sosyal ağların metinsel ve içeriksel bilgilerini bir araya getirerek başlık ve lokasyon duyarlı POI tavsiye sistemi önermişlerdir. Bu çalışmalar genellikle GPS verisi ile çalışmış ve lokasyon/rota tavsiyesi için kullanıcı oylaması kullanmışlardır.

Kullanıcı-seviyesinde arkadaş tavsiye sistemleri çalışmaları genel olarak arkadaş tavsiyesi için kullanıcılar arasındaki benzerlik araştırması üzerine çalışmışlardır [28, 37-39]. Kefalas ve Symeonidis [28] zaman boyutunu tavsiye sistemine entegre ederek arkadaş ve lokasyon tavsiye metodu önermişlerdir. Önerilen metot, zamanla değişen kullanıcı tercihlerini yakalayabilmektedir. Zheng vd. [37] bir kullanıcının önemli lokasyonlarını keşfetmekte ve benzer kullanıcıların önemli lokasyonlarından yeni lokasyonlar tavsiye etmektedirler. Bu çalışmada GPS verisi kullanılmış ve hiyerarşik graf tabanlı benzerlik ölçütü, kullanıcı benzerliği için kullanmışlardır. Ying vd. [38] mobil

kullanıcıların semantik güzergâhlarını arkadaşlık tavsiyesi için kullanmıştır. Hannon vd. [39] takipçi/takip edilen tavsiyesi için çeşitli profillemeye ve tavsiye stratejileri kullanmış ve Twitter veriseti üzerinde çalışmışlardır. Bu çalışma, sosyal medya kullanıcılarının lokasyon ve zaman bilgilerini dikkate almamaktadır.

Aktivite/tag tavsiye sistemleri genellikle kullanıcı taglerini referans almışlardır [40-42]. Feng vd. [40] hiyerarşik mekân-zamansal hashtag kümeleme için veri yapısı yaklaşımı önermişlerdir. Bu veri yapısı, Twitter verisini farklı zaman ve mekân boyutlarında sorgulamayı sağlamaktadır. Budak vd. [41] başlıklar ve lokasyonlar arasında ilişki kurmak için GeoScope'u önermişlerdir. Bu çalışmada lokasyon-başlık ikililerini keşfetmek için Twitter veriseti kullanılmıştır. Sattari vd. [42] verilen bir lokasyon için aktivite öneri sistemi önermişlerdir.

Kullanıcı-seviyesinde ilginç/önemli lokasyonlar keşfi çalışmalarında genellikle GPS ve telefon arama verilerini kullanmışlardır. [43-47]. Cao vd. [43] semantik olarak anlamlı ve belirgin lokasyonların madenciliği için bir çerçeve önermişlerdir. Bu çalışmada, araçlardan GPS verisi toplanmış ve belirgin semantik lokasyonların keşfi için kümeleme tabanlı yaklaşımlar kullanılmıştır. Ying vd. [44] kullanıcıların güzergâhlarının semantik ve coğrafi özelliklerini dikkate alarak kullanıcıların sonraki lokasyonlarını tahmin etmek için kümeleme tabanlı bir strateji önermişlerdir. Zhou vd. [45] kişisel önemli lokasyonların madenciliği için metotlar önermişlerdir. Bu çalışmada GPS verisi kullanılmış ve kümeleme tabanlı yaklaşımlar geliştirilmiştir. Isaacman vd. [46] önemli kişisel lokasyonların madenciliği için kümeleme ve regresyon tabanlı yaklaşımlar önermişlerdir. Bu çalışmada telefon arama kayıtları kullanılmıştır. Pavan vd. [47] kişisel ilginç lokasyonları bulmak için özellik tabanlı bir yaklaşım önermişlerdir.

Kullanıcı-seviyesinde rota madenciliği çalışmaları rotaların keşfine odaklanmıştır [48-50]. Chen vd. [48] semantik anlamı ve iskelet yapısını değiştirmeden, rotaları basitleştiren bir metot önermişlerdir. Cai vd. [49] ilgi odağı bölgelerin tespiti için metotlar önermişlerdir ve önerilen algoritmayı sıralı güzergâh örüntülerini keşfetmek için sıralı örüntü madenciliği ile birleştirmiştir. Cenamor vd. [50] turistler için kişiselleştirilmiş planlar oluşturan bir sistem önermişlerdir. Bu çalışmalar genel olarak rotaların keşfi için güzergâhları kullanmaktadır.

Kullanıcı-seviyesinde şehir ve hareketlilik örüntü madenciliği çalışmaları genellikle kullanıcı lokasyon ve hareket tahminine odaklanmıştır [51-53]. Chen vd. [51] günlük hareketlilik örüntülerinin check-in verisinden çıkarılıp madenciliği ve kümelemesinin yapılması için yaklaşımlar önermiştir. Bu çalışmanın amacı, mekânsal ve zamansal bilgileri kullanarak bir şehirdeki kullanıcıların günlük hareketlilik örüntülerini ortaya yakalamaktır. Cho vd. [52] insan hareketlilik örüntülerini araştırarak insan hareketlilik modelleri önermiş ve gelecekteki hareketlilik örüntülerini tahmin etmişlerdir. Liu vd. [53] geo-etiketli resimler ve check-in verisi kullanarak ilgi odağı bölgelerin keşfi için bir metot önermişlerdir. Bu çalışmalar genel olarak lokasyon ve gelecekteki hareketlerin tahminine odaklanmıştır.

Grup-seviyesinde önemli lokasyonlar madenciliği çalışmaları; lokasyon/rota tavsiyesi, arkadaş tavsiyesi, aktivite/tag tavsiyesi, ilginç/önemli lokasyonlar keşfi, rota madenciliği ve şehir ve hareketlilik örüntü madenciliği çalışmalarını içermektedir.

Grup-seviyesinde lokasyon/rota tavsiye sistemleri güzergâh örüntülerinin keşfine odaklanmıştır [54-56]. Zheng vd. [54] [55] GPS verisini indekslemek, görselleştirmek, sorgulamak ve madenciliğini yapmak için GeoLife çerçevesini önermişlerdir. Bu çalışmalarda, kullanıcı deneyimi ve lokasyon alakasına göre lokasyonlar arasında seyahat sıralaması tespit etmişlerdir. Yin vd. [56] geo-etiketli sosyal medya verisetleri üzerinde yaygın güzergâh örüntülerinin madenciliğine odaklanmışlardır. Bu çalışmada temsil edici güzergâhların bulunması için güzergâh örüntüleri sıralanmıştır.

Grup-seviyesinde arkadaş tavsiye sistemleri sosyal medya verilerine dayalı analizler yapmışlardır [57] [58]. Gaeta-Villegas vd. [57] sosyal ağlarda lokasyon verisini kullanarak kullanıcıları seyahat ortağı olarak eşleştiren bir yaklaşım önermişlerdir. Symeonidis vd. [58] kullanıcı check-in verisinden arkadaşlık, lokasyon ve aktivite tavsiyesi yapan Geo-sosyal çevrimiçi tavsiye sistemi önermişlerdir.

Grup-seviyesinde aktivite/tag tavsiye sistemleri çalışmaları arasında Guy vd. [59] yaptığı çalışma sayılabilir. Guy vd. [59] insanlar ve tag'lere dayalı tavsiye sistemi önermişlerdir. Bu çalışmada tag-tabanlı ve kullanıcı-tabanlı tavsiyeler karşılaştırılmış ve kullanıcı-tag-tabanlı tavsiye sistemi önerilmiştir. Bu çalışma, tavsiye sistemi için tag ve kullanıcı üzerine odaklanmış, lokasyon ve zaman bilgilerini dikkate almamıştır.

Grup-seviyesinde ilginç/önemli lokasyonlar keşfi çalışmaları arasında Zheng vd. [60] ve Khetarpaul vd. [61] tarafından yapılan çalışmalar sayılabilir. Zheng vd. [60] kullanıcıların seyahat deneyimlerini ve lokasyonlar arasındaki ilişkileri dikkate alarak en yüksek n ilginç lokasyon ve en yüksek m seyahat sırasını keşfeden modeller ve yaklaşımlar önermişlerdir. Bu çalışmada GPS verisi kullanılmıştır. Khetarpaul vd. [61] GPS verisini birleştirerek ve çoklu kullanıcı güzergâhlarını analiz ederek ilginç lokasyonları keşfeden bir yaklaşım önermişlerdir.

Grup-seviyesinde rota madenciliği çalışmaları arasında Chen vd. [62] ve Wei vd. [63] tarafından yapılan çalışmalar sayılabilir. Chen vd. [62] kullanıcı seyahat güzergâhlarını dikkate alarak güzergâh verilerinden en popüler yolları keşfeden bir rota madenciliği yaklaşımı önermişlerdir. Wei vd. [63] kesin olmayan güzergâhlardan en yüksek-k rotayı keşfeden bir yaklaşım önermişlerdir.

Grup-seviyesinde şehir ve hareketlilik örüntü madenciliği çalışmaları arasında Ferrari vd. [64] ve Cranshaw vd. [65] tarafından yapılan çalışmalar sayılabilir. Ferrari vd. [64] lokasyon tabanlı sosyal ağlardan şehir örüntülerini keşfeden bir yaklaşım önermişlerdir. Bu çalışmada, Twitter veri kümesi kullanılmıştır ve olasılıksal başlık modeli olan Latent Dirichlet Allocation (LDA) insan hareketlilik örüntülerinin keşfi için kullanılmıştır. Cranshaw vd. [65] şehirlerin farklı bölgelerindeki coğrafi alanların ve insanların şehir içerisindeki hareketlilik örüntülerinin keşfi için kümeleme yaklaşımları önermişlerdir. Bu çalışmada lokasyon tabanlı sosyal ağlardan toplanan check-in veri kümesi kullanılmıştır.

Mekânsal sosyal önemli lokasyonlar keşfi literatüründeki çalışmaların çeşitli eksiklikleri/kısıtlamaları vardır. Birincisi, çalışmaların bazıları GPS veya telefon arama verilerini kullanmıştır. Ancak sosyal medya veri kümeleri GPS veya telefon arama verilerinden daha fazla bilgi içermektedir ve kullanıcılar arasında daha fazla etkileşim barındırmaktadır. İkincisi, çalışmaların bazılarında önemli lokasyonların tespiti için kullanıcı oylamasına ihtiyaç duyulmaktadır. Üçüncüsü, çalışmaların bazıları lokasyon tanımlaması için metin madenciliği metotları ve tag'leri kullanmış ve mekânsal bilgiyi dikkate almamıştır. Son olarak, çalışmaların pek çoğu kullanıcı tercihlerini dikkate alarak sosyal medya gruplarının önemli lokasyonlarının keşfini yapmamıştır.

4.2.Mekân-Zamansal Sosyal Medya Madenciliği

Mekân-zamansal yöntemlerle sosyal önemli lokasyonlar madenciliği ile ilgili çalışmalar, lokasyon/rota tavsiye sistemleri, şehir ve insan hareketlilik örüntü madenciliği ve sosyal önemli lokasyonlar madenciliği konularını içermektedir. Her bir konu, kullanıcı-seviyesinde sosyal önemli lokasyonlar madenciliği ve grup-seviyesinde sosyal önemli lokasyonlar madenciliği olmak üzere iki alt başlığa ayrılabilir.

Mekân-zamansal lokasyon/rota tavsiye sistemlerinde, kullanıcı-seviyesindeki çalışmalar arasında Ying vd. [66], Hu vd. [67] ve Yuan vd. [68] yaptığı çalışmalar sayılabilir. Ying vd. [66] bir kullanıcının bir lokasyonu ziyaret etme olasılığını kestirebilmek için, kullanıcıların coğrafi, zamansal ve semantik etkileşimlerini dikkate alan bir yaklaşım önermişlerdir. Hu vd. [67] bir kullanıcının mekân-zamansal başlıklarını yakalayan olasılık tabanlı bir model önermişlerdir ve mekân-zamansal özelliklere göre lokasyon tavsiyesi yapmışlardır. Yuan vd. [68] zaman-duyarlı POI tavsiyesi için coğrafi ve zamansal etkileri modelleyen bir graf önermişlerdir. Grup-seviyesinde mekân-zamansal lokasyon/rota tavsiye sistemleri çalışmaları arasında Kefalas ve Manolopoulos [69] ve Liu ve Seah [31] tarafından yapılan çalışmalar sayılabilir. Kefalas ve Manolopoulos [69], POI tavsiyesi için hem açık (oylama, yorum gibi) hem de gizli (skor, görüntüleme, kullanıcı etkisi) bilgileri kullanan modeller önermişlerdir. Liu ve Seah [31], mobil kullanıcıların GPS rota verisine dayalı olarak POI tavsiye metodu ve önemli lokasyonların keşfi için yoğunluk tabanlı kümeleme yaklaşımı önermişlerdir.

Mekân-zamansal şehir ve insan hareketlilik örüntü madenciliği konusunda, kullanıcı-seviyesindeki çalışmalar arasında Yuan vd. [70], Huang ve Li [71], Celikten vd. [72], Cho vd. [52] ve Assem vd. [73] tarafından yapılan çalışmalar sayılabilir. Yuan vd. [70] hem coğrafi hem zamansal bilgiyi kullanarak periyodik hareketlilik örüntülerinin keşfedilmesi için parametrik olmayan bir model önermişlerdir. Huang ve Li [71], hem mekân-zamansal hem semantik boyutlara bağlı olarak insan aktivite örüntülerini modellemişlerdir. Semantik boyut insanların bir lokasyonu bir zamandaki ziyaret motivasyonunu ortaya çıkarmaktadır. Celikten vd. [72] çok sayıda şehir içerisinde ziyaret yeri (venue) dağılımını çıkarmak için olasılık tabanlı bir model önermişlerdir. Cho vd. [52] insan hareketlilik örüntülerini analiz etmiş ve insanların gelecekteki hareketlerini tahmin eden modeller önermişlerdir. Assem vd. [73] şehirlerin işlevsel bölgelerinin

zamansal değişimini keşfetmek için kümeleme tabanlı bir yaklaşım geliştirmişlerdir. Ancak, bu çalışmalar önemli lokasyonlar madenciliği konusuna odaklanmamakta ve grup bilgisini kullanmamaktadır. Grup-seviyesinde mekân-zamansal şehir ve insan hareketlilik örüntü madenciliği konusunda yapılmış olan çalışmalar arasında Jiang vd. [74], Zhang vd. [75], Chen vd. [51] ve Ferrari vd. [64] tarafından yapılan çalışmalar sayılabilir. Jiang vd. [74] mekân ve zaman bilgisi kullanarak gündelik örüntülerine göre insanları kümelemişlerdir. Zhang vd. [75] aynı gruptaki kullanıcıların benzer örüntüleri paylaşacağını varsayan grup-seviyesinde hareketlilik modelleme metodu önermişlerdir. Chen vd. [51] kullanıcıların gündelik hareketlerini yakalayabilmek için hareketlilik değişim örüntüsü önermişlerdir. Kullanıcı hareketlilik değişim örüntülerine bağlı olarak, kümeleme yaklaşımıyla temsilci hareketlilik örüntüleri keşfetmişlerdir. Ferrari vd. [64] şehir örüntülerini keşfetmek için olasılıksal başlık modelleme yaklaşımı önermişlerdir. Ancak bu çalışmalar, şehir örüntü madenciliğine odaklanmakta ve önemli lokasyonların çıkarımını yapmamaktadır.

Mekân-zamansal önemli lokasyonlar madenciliği konusunda, kullanıcı-seviyesindeki çalışmalar arasında Khetarpaul vd. [76], Zarezade vd. [77] ve Wei vd. [78] tarafından yapılan çalışmalar sayılabilir. Khetarpaul vd. [76] hareketli nesnelerin (sosyal medya kullanıcılarının) sosyal ağlardaki mekân-zamansal davranışlarını yakalayabilmek için mekân-zamansal sosyal veri modeli önermişlerdir. Zarezade vd. [77] hem kullanıcıların mekân-zamansal örüntüleri hem de arkadaşlarının etkisini dengeleyerek olasılıksal bir model önermişlerdir. Wei vd. [78] kesin olmayan rotalardan popüler rota oluşturan bir çerçeve önermişlerdir. Bu çerçeve, zamana bağlı ve sıralı olarak gözlemlenmiş en yüksek-k rotadan oluşmaktadır. Ancak, bu çalışmalar bir kullanıcı grubunun tercihlerini ve sosyal ağ davranışlarını göz önünde bulundurmamaktadır. Grup-seviyesinde mekân-zamansal sosyal önemli lokasyonlar madenciliği konusunda, Comito vd. [79] insan dinamikleri ve davranışlarının madenciliğini yapan bir metot önermişlerdir. Bu çalışmada, sıralı örüntü madenciliği algoritması kullanarak, yaygın seyahat rotalarının keşfi için günlük rotaların madenciliği yapılmıştır. Çalışmada, rotaların ve lokasyonların keşfi için 6 aylık geo-etiketli Londra veri kümesi kullanılmıştır. Ancak, bu çalışma kullanıcıların geçmişe dayalı verilerini dikkate almamakta ve kullanıcı tercihlerini analiz etmemektedir. Ayrıca, önerilen metot kullanıcı yaygınlığına değil, frekansa dayalıdır.

Mekân-zamansal sosyal önemli lokasyonlar keşfi literatürünün çeşitli eksiklikleri/kısıtlamaları vardır. Birincisi, çalışmaların bir kısmı GPS verisi kullanmaktadır. Ancak, sosyal medya verisi, diğer veri türlerinden farklı ve daha fazla içeriğe sahiptir. İkincisi, sosyal önemli lokasyonlar keşfi için grup bilgisi etkin olarak kullanılmamakta ve grup üyelerinin sosyal önemli lokasyonlara katkıları net değildir. Üçüncüsü, kullanıcıların tüm geçmiş paylaşımları değil, belirli bir zaman aralığı için veriler toplanarak analizler gerçekleştirilmiştir.

4.3.Bulut Bilişim Sistemlerinde Sosyal Medya Madenciliği

Sosyal medya ağları, büyük sosyal medya veri kümelerinden faydalanabilmeleri için, sosyal medya büyük veri kümelerini araştırmacılar ve geliştiricilerle paylaşmaktadır, bu sebeple sosyal medya veri kümeleri büyük verinin yaygın ve önemli kaynaklarından. Neredeyse milyarlarca insan sosyal medya ağlarını günlük olarak kullanmakta ve bu ağlar için içerik oluşturmaktadır. Bu büyüklükteki bir veriyi analiz etmek istediğimiz zaman, geleneksel veri depolama ve analiz metotlarının alternatiflerini gözden geçirmek durumundayız. Bulut bilişim güçlü ve en çok tercih edilen büyük veri analizi alternatifidir.

Bulut bilişim, hesaplama kaynaklarının – sunucu, depolama, uygulama platformu, yazılım, analiz servisi – internet üzerinden ve kullandığın-kadar-öde yaklaşımıyla talep üzerine sağlandığı ve büyük miktardaki hesaplama ve depolama kaynaklarına esnek ve ölçeklenebilir yaklaşımla erişim sağlandığı sistemleri ifade etmektedir [7] [10]. Bulut bilişim sistemlerini kullanarak pek çok hesaplama-yoğun görevler, Apache Hadoop gibi büyük veri analizi çerçeveleri kullanılarak gerçekleştirilebilir. Apache Hadoop, büyük veri kümelerinin dağıtık yaklaşımla depolanması ve analizi için kaynaklar sağlayan bir yazılım çerçevesidir [13] [12]. Ana büyük veri kaynaklarından biri olarak sosyal medya veri kümeleri de Apache Hadoop çerçevesi kullanılarak analiz edilebilmektedir.

Bulut bilişim sistemleri üzerinde sosyal medya madenciliği ile ilgili çalışmalar, semantik sosyal medya madenciliği, sosyal medya çizge (graf) madenciliği ve mekânsal ve mekân-zamansal sosyal medya madenciliği olmak üzere üç ana kategoriye ayrılabilir.

Semantik sosyal medya madenciliği, sosyal medya kullanıcılarının paylaşımlarının temel olarak metin madenciliği metotlarıyla analizini ifade etmektedir. Acil durum yönetimi,

afet koruma ve başlık tespiti bulut bilişim sistemlerinde semantik sosyal medya madenciliğinin ana konularından bazılarıdır. Acil durum yönetimi konusunda, Xu vd. [80] çeşitli sorulara cevap vererek sosyal medya veri kümelerinden şehir acil olayları tespit eden modeller önermişlerdir. Aynı yazarlar, ayrıca, şehir acil olayların semantik ve mekânsal analizi için sosyal medya verisi üzerinde iştirakçi tespit (participatory sensing) yaklaşımı kullanmışlardır. Qiu vd. [81] sosyal medya verisi ve çeşitli veri kaynaklarını kullanarak olay lokasyonundan çıkması için kullanıcıları uyaran akıllı bulut tahliye sistemi önermişlerdir. Afet koruma konusunda, Huang vd. [82] mekânsal ve semantik veri madenciliği metodlarını kullanarak afet olaylarının analizi için sosyal medya ve çeşitli kaynakları kullanan bir çerçeve önermişlerdir. Bu çalışmada, büyük veri hesaplamalarını ölçekleyebilmek için bulut bilişim kaynaklarını kullanmışlardır. Rogstadius vd. [83] yapısal olmayan tweetler ve grup bilgisini kullanarak durumsal farkındalığı (situational awareness) artırmak için gerçek zamanlı analiz ve kitlesel kaynağı (crowdsourcing) birleştiren CrisisTracker sistemini önermişlerdir. Başlık tespiti konusunda, Vakali vd. [84] mikroblog ve blog aktiviteleri üzerinde çok konuşulan başlıkların gerçek zamanlı olarak tespiti için Cloud4Trends çerçevesini önermişlerdir. Moise [85] Twitter veri kümesi üzerinde Hadoop Mahout çerçevesini kullanarak başlık keşfi için bir yaklaşım önermiştir. Bu çalışmada bir başlıktaki duygunun zamansal değişimi de gözlemlenmiştir. Ancak, bu çalışmalar çoğunlukla sosyal medya madenciliğinin semantik boyutuna odaklanmakta ve mekânsal bilgiyi ikincil bilgi olarak dikkate almaktadırlar.

Sosyal medya çizge madenciliği, sosyal medya kullanıcılarının arkadaşlık ağlarının analizini ifade etmektedir. Kullanıcı sıralaması, topluluk ve arkadaşlık analizi ve sahtelik (spam) tespiti bulut bilişim sistemlerinde sosyal medya çizge madenciliğinin ana konuları arasındadır. Kullanıcı sıralama konusunda, Noordhuis vd. [86] büyük ölçekli çizge veri kümelerinde bulut bilişim sistemlerini kullanmayı önermişlerdir. Bu çalışmada, Twitter kullanıcıları PageRank algoritmasıyla sıralanmış ve bulut bilişimin uygun bir çözüm olup olmadığı incelenmiştir. Ching vd. [87] Facebook veri kümesi üzerindeki bir trilyon noktayı analiz etmek için Apache Giraph kullanmışlardır. Bu çalışmada, kullanıcı sıralama, arkadaşların arkadaşları (friend of friend) analizlerini içeren çeşitli çalışmalar bulut bilişim kaynaklarını kullanarak yapılmıştır. Topluluk ve arkadaşlık analizi konusunda, Jin vd. [88] MapReduce modelini kullanarak dağıtık topluluk yapısı

madenciliği çerçevesi önermişlerdir. Gupta vd. [89] Twitter üzerinden çizge işleme metotlarını kullanarak arkadaşlık tavsiye servisi önermişlerdir. Sahtelik tespiti konusunda, Thomas vd. [90] gerçek-zamanlı sahte filtreleme sistemi olan Monarch sistemini önermiş ve Amazon Web Servisleri (Amazon Web Services – AWS) üzerine uyarlamışlardır. Ancak, bu çalışmalar sosyal medya ağlarının, ağ yapısına odaklanmakta ve kullanıcıların paylaşımları dikkate alınmamaktadır.

Mekânsal sosyal medya madenciliği, sosyal medya kullanıcılarını, paylaşımlarının mekânsal bilgisini kullanarak analiz eden çalışmaları ifade etmektedir. Lokasyon/rota tavsiye, şehir örüntü keşfi ve sosyal önemli lokasyonların keşfi, mekânsal sosyal medya madenciliğinin ana konuları arasındadır. Lokasyon/rota tavsiye sistemleri konusunda, Khalid vd. [91] optimal tavsiyeyi üretebilmek için algoritmalar, filtreler ve skorları kullanan bulut tabanlı OmniSuggest çerçevesini önermişlerdir. Wang vd. [92] yeni lokasyonlardaki kullanıcı tercihlerinin madenciliği ve lokasyon tavsiyesinin hassasiyet ve verimliliğini artırmak için işbirlikçi filtreleme (collaborative filtering) algoritmasına dayalı bir metot geliştirmişlerdir. Ayrıca, geliştirilen metot MapReduce modeline dayalı olarak paralelleştirilmiştir. Kosmides vd. [93] sosyal medya ağlarından veri toplayan ve makine öğrenmesi algoritmalarıyla lokasyon tavsiyesi sağlamak için bulut platformlarına gönderen bir sistem mimarisi önermişlerdir. Şehir örüntü keşfi konusunda, Cesario vd. [94] EXPO 2015 Instagram ziyaretçilerinin hareketlilik örüntüleri ve davranışlarını keşfetmek için bulut tabanlı bir metot önermişlerdir. Zhou vd. [95] bulut bilişim sistemlerini ve Flickr resimlerinin mekân-zamansal özelliklerini kullanarak çok sayıda şehrin turistik bölgelerini tespit eden bir metot önermişlerdir. Sosyal önemli lokasyonlar keşfi konusunda, Cao vd. [96] sosyal medya kullanıcılarının mekân-zamansal veri küpü modelini kullanarak büyük ölçekli sosyal medya verilerinden verimli ve sistematik mekân-zamansal analizler yapabilmek için bir çerçeve önermişlerdir. Bu çalışmada büyük ölçekli kullanıcı sayısını ölçekleyebilmek için Hadoop kümeleri kullanılmıştır. Ancak, bu çalışmalar her bir sosyal medya kullanıcısının sosyal önemli lokasyonlarının keşfi ve kullanıcı sosyal önemli lokasyonlarından grup-seviyesinde sosyal önemli lokasyonların keşfine odaklanmamışlardır.

Bulut bilişim sistemlerinde sosyal önemli lokasyonlar keşfi literatürünün çeşitli eksiklikleri/kısıtlamaları vardır. Birincisi, çalışmaların büyük çoğunluğunda bulut bilişim sistemleri uygulamalara detaylı bir şekilde entegre edilmemekte, ikincil yaklaşım olarak

kullanılmaktadır. İkincisi, bulut bilişim sistemlerini kullanarak yapılan çalışmalar azdır. Üçüncüsü, yapılan çalışmalarda, kullanıcı ve grup seviyelerinde mekânsal ve mekân-zamansal sosyal önemli lokasyonlar keşfi konusu çalışılmamıştır.

4.4.Tez Çalışmasının Literatüre Katkıları

Literatürdeki çalışmalar incelendiği zaman, mekânsal ve mekân-zamansal sosyal medya madenciliği konularında çok çeşitli başlıklarda çalışmalar yapıldığı gözlemlenmiştir. Ancak bu çalışmaların pek çoğu, lokasyon/rota tavsiye sistemleri, şehir örüntü keşfi gibi konulara odaklanmakta, sosyal medya ağlarındaki her bir kullanıcının tüm geçmiş verilerini analiz ederek sosyal önemli lokasyonları keşfetmemektedir. Ayrıca, yapılmış olan çalışmaların çok azında bulut bilişim sistemleri, büyük sosyal medya verisini ölçekleyebilmek ve çok sayıda kullanıcıyı analiz etmek için değerlendirilmiştir. Literatürdeki bu boşlukları doldurmak için, bu tez çalışması, oylama veya lokasyon türü gibi herhangi bir açıklayıcı bilgiye ihtiyaç duymadan, kullanıcıların sosyal medya aktivite geçmişlerinden sosyal önemli lokasyonları keşfetmeyi amaçlamıştır. Bu tez çalışmasının katkıları aşağıda listelenmiştir:

- Sosyal önemli lokasyonların keşfi için mekânsal ve mekân-zamansal iki yaklaşım önerilmiştir.
- Önerilen yaklaşımlar kullanıcı ve sosyal grup seviyelerinde sosyal önemli lokasyonları keşfetmektedir.
- Mekânsal sosyal önemli lokasyonların ölçülebilmesi için lokasyon yoğunluğu, ziyaret ömrü ve kullanıcı yaygınlığı değerlendirme ölçütleri önerilmiştir.
- Mekân-zamansal sosyal önemli lokasyonların ölçülebilmesi için zaman penceresi oluşumu, kullanıcı zamansal lokasyon önemi ve sosyo-mekân-zamansal lokasyon önemi değerlendirme ölçütleri önerilmiştir.
- Mekânsal sosyal önemli lokasyonların keşfi için SS-ILM algoritması mekân-zamansal sosyal önemli lokasyonların keşfi için TF-SSTIL Miner ve SF-SSTIL Miner algoritmaları önerilmiştir.
- Önerilen sosyal önemli lokasyonlar madenciliği algoritmaları bulut bilişim sistemlerine de uyarlanmıştır.
- Bulut tabanlı önerilen algoritmalar Microsoft Azure bulut bilişim ortamı ve Apache Hadoop yazılım çerçevesi üzerinde koşturulmuştur.

- Önerilen algoritmalar, İstanbul ve Türkiye Twitter kullanıcılarının büyük veri kümesi üzerinde deneysel olarak değerlendirilmiştir.
- Deneysel sonuçlar, algoritmaların başarılı bir şekilde sonuçları üretebildiğini ve bulut bilişim sistemlerinin kullanılmasının faydalarını ortaya koymaktadır.

5. BÖLÜM

BÜYÜK SOSYAL MEDYA VERİSİNDEN MEKÂNSAL ÖNEMLİ LOKASYONLARIN KEŞFİ

Bu bölümde, büyük sosyal medya verisi kullanılarak, bir sosyal medya kullanıcı grubu için mekânsal önemli lokasyonların keşfi anlatılmış, mekânsal önemli lokasyonların keşfi için yapılan tanımlar verilmiş ve geliştirilen algoritmalar tanıtılmıştır. Öncelikli olarak, önemli lokasyonlar madenciliğinin tanımı, zorlukları ve faydaları açıklanmıştır. Ardından, geliştirilmiş olan mekânsal önemli lokasyonların keşfi yaklaşımı ve geliştirilen algoritmalar tanıtılmıştır. Son olarak deneysel çalışma sonuçları ile önerilen algoritmalar değerlendirilmiş ve sonuç kısmıyla bölüm tamamlanmıştır.

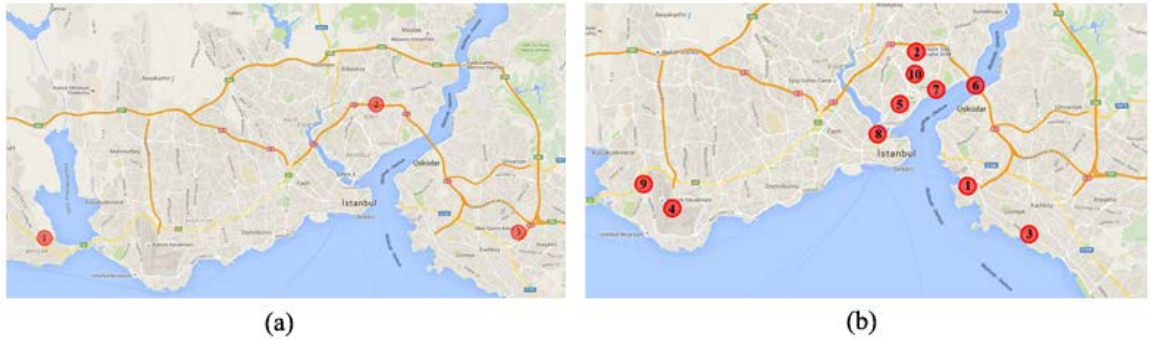
5.1.Giriş

Lokasyon-duyarlı mobil cihazların geliştirilmesiyle birlikte sosyal medya ağları, kullanıcıları çevrimiçi iken kullanıcılarının lokasyon bilgilerini toplamaya başlamışlardır. Lokasyon verisinin sosyal etkileşimle birlikte kullanılmasıyla (lokasyon-lokasyon etkileşimi, kullanıcı-lokasyon etkileşimi ve kullanıcı-kullanıcı etkileşimi gibi) değerli bilgiler ortaya çıkarılabilir. Bunlara örnek olarak, hangi sosyal arkadaşların gerçek arkadaş oldukları, hangi kullanıcının ne tür lokasyonları sevdiği, hangi kullanıcıların mekânsal olarak birbirlerini takip ettiği, hangi kullanıcının tavsiyelerinin daha çok ilgi gördüğü, hangi lokasyonların bir kullanıcı grubu için önemli olduğu sayılabilir. Bu bölümde yapılan çalışmanın amacı, sosyal medya kullanıcıları tarafından sosyal medya geçmişlerinde sıklıkla ziyaret ettikleri sosyal önemli lokasyonların keşfini yapmaktır.

Sosyal önemli lokasyonların keşfi, tavsiye sistemleri, pazarlama (hedeflenmiş reklamcılık uygulamaları) ve şehir planlama gibi çeşitli uygulama alanları için önemlidir. Sosyal önemli lokasyonların keşfi ile sosyal bir grubun mekânsal tercihleri ortaya çıkarılabilecektir.

Ancak sosyal önemli lokasyonların keşfi çeşitli sebeplerden dolayı zordur. Birincisi, mevcut değerlendirme ölçütleri önemli lokasyonların keşfi için yetersizdir. İkincisi, veri boyutu ve ölçeği sürekli artmaktadır. Üçüncüsü, insan davranışları tahmin edilemez ve çeşitli etmenlerden etkilenmektedir.

Önemli lokasyonlar madenciliği pek çok araştırmacı tarafından çalışılmış ve çeşitli çözümler önerilmiştir. Ancak, yapılan çalışmaların bazı kısıtlamaları vardır. Bu çalışmaların bazıları GPS veya telefon arama verileri kullanmışlardır ve önerilen çözümler bu verilere bağımlıdır. Yapılan diğer çalışmalarda sosyal medya veri kümelerini kullanmışlardır. Kullanıcı seviyesinde lokasyon tavsiyesi ve keşfi için çeşitli algoritmalar ve çerçeve programlar geliştirilse de sınırlı sayıdaki çalışmada bir sosyal medya kullanıcı grubunun grup-seviyesinde önemli/ilginç lokasyonlarının keşfi çalışılmıştır.



Şekil 5.1. a) Bir kullanıcı için sosyal önemli lokasyonlar ve b) Bir sosyal medya kullanıcı grubu için sosyal önemli lokasyonlar

Şekil 5.1’de, İstanbul’da yaşayan bir sosyal medya kullanıcısı (Twitter kullanıcısı) ve bir sosyal medya kullanıcı grubu (1000 Twitter kullanıcısı) için sosyal önemli lokasyonlar görülmektedir. Şekil 5.1. a)’da sosyal medya kullanıcısının üç sosyal önemli lokasyonu görülmektedir. Sosyal önemli lokasyonları keşfedilen sosyal medya kullanıcısının Aralık 2012 ve Aralık 2015 tarihleri arasında 1765 sosyal medya aktivitesi (tweetlerinin lokasyon ve zaman bilgisi) bulunmaktadır. Şekil 5.1. a)’da görüldüğü gibi, bu lokasyonlar mekânsal olarak uzak lokasyonlardır ve kullanıcı için farklı anlamları bulunmaktadır. 1 numaralı lokasyonda bir üniversite kampüsü bulunmaktadır, 2 numaralı lokasyonda büyük bir alışveriş merkezi bulunmaktadır ve 3 numaralı lokasyonda büyük bir şirketin merkezi bulunmaktadır. Bu sonuçlarla ilgili çeşitli fikir/açıklamalarda bulunabiliriz. Örneğin, lokasyon 1 kullanıcının üniversitesi olabilir, lokasyon 2 kullanıcının çalışma alanı veya alışveriş mekânı olabilir ve lokasyon 3 kullanıcının çalışma alanı olabilir. Şekil

5.1. b)'de İstanbul'da yaşayan 1000 sosyal medya kullanıcısının 10 adet sosyal önemli lokasyonu görülmektedir. 1000 kullanıcının nasıl seçildiği ile ilgili detaylar Bölüm 5.4.2'de verilmiştir. Gruba ait sosyal önemli lokasyonlar şehrin sosyal hayatının aktif olduğu bölgelerde kümelenmiştir. Gruba ait sosyal önemli lokasyonlarla Şekil 5.1. a)'daki kullanıcıya ait sosyal önemli lokasyonlar karşılaştırıldığı zaman, lokasyon 2 (alışveriş merkezi) sosyal medya kullanıcısı ve sosyal medya kullanıcı grubu için ortak sosyal önemli lokasyon olduğu görülmektedir. Aksine, lokasyon 1 ve 3, sadece kullanıcı için önemli lokasyonlar olurken, sosyal medya grubu için önemli değildir.

Şekil 5.1 a)'daki üç sosyal önemli lokasyon, kullanıcının günlük sosyal medya aktivite kayıtlarına göre keşfedilmiştir. Sosyal medya kullanıcıları bir araya getirildiği zaman, bu kullanıcıların veri kümelerinin analizi, o kullanıcı grubu için önemli lokasyonların keşfini sağlayacaktır. Örneğin, müziksever sosyal medya kullanıcıları bir araya getirilirse, canlı müzik kafeleri, konser alanları ve müzik marketlerin sosyal önemli lokasyonlar olması beklenir. Futbol fanatiklerinin sosyal medya kullanıcıları bir araya getirilirse, stadyumlar ve sporla alakalı lokasyonların sosyal önemli lokasyonlar olması beklenir.

Bu çalışmada, sosyal medya veri kümeleri üzerinde sosyal önemli lokasyonlar madenciliği problemi çalışılmıştır. Bu çalışmada, sosyal önemli lokasyonları ölçmek için lokasyon yoğunluğu, ziyaret ömrü ve kullanıcı yaygınlığı değerlendirme ölçütlerini formülize edilmiştir ve sosyal önemli lokasyonları verimli bir şekilde keşfedebilmek için bir algoritma, SosyoMekansal Önemli Lokasyonlar Madenciliği algoritması (SocioSpatial Important Locations Mining – SS-ILM), önerilmiştir. Önerilen yaklaşım iki adımdan oluşmaktadır. Birinci adımda, lokasyon yoğunluğu ve ziyaret ömrü değerlendirme ölçütleri kullanılarak kullanıcı-seviyesinde önemli lokasyonlar keşfedilmiştir. İkinci adımda, sosyal kullanıcı gruba ait sosyal önemli lokasyonlar keşfedilmiştir. Önerilen algoritma, SS-ILM algoritması, toy yaklaşımla karşılaştırılmıştır. Her iki algoritma İstanbul Twitter kullanıcılarının veri kümesi kullanılarak deneysel olarak test edilmiştir.

5.2. Temel Tanımlar ve Problem Tanımlaması

Bu kısımda öncelikle sosyal medya veri kümelerinden sosyal önemli lokasyonlar madenciliğine ait genel fikir verilmiş ve ardından problem tanımlaması yapılmıştır.

5.2.1. Temel Tanımlar

Bir kullanıcı grubu için sosyal önemli lokasyonların madenciliği problemi iki bileşenden oluşmaktadır. Birinci bileşen, kullanıcı-seviyesinde sosyal önemli lokasyonların tespiti ve ikinci bileşen, grup-seviyesinde sosyal önemli lokasyonların tespiti. Sosyal önemli lokasyonlar madenciliğiyle ilgili tanımlamalar aşağıda verilmiştir. Tanımlar 1-5 kullanıcı-seviyesinde sosyal önemli lokasyonlar keşfi formülize etmektedir. Tanımlar 1-7 grup-seviyesinde sosyal önemli lokasyonlar keşfi problemini formülize etmektedir.

Tanım 1. Sosyal medya aktivite geçmişisi verilmiş olan bir sosyal medya kullanıcısı u ve bir lokasyon l için, l 'nin u için **lokasyon yoğunluğu** değeri, u 'nun l 'deki oluşum sayısının u 'nun toplam oluşum sayısına oranıdır.

$$\text{lokasyon yoğunluğu}_l^u = \frac{u'nun\ l'deki\ oluşum\ sayısı}{u'nun\ toplam\ oluşum\ sayısı}$$

Lokasyon yoğunluğu, lokasyon l 'nin, kullanıcı u 'nun sosyal medya geçmişindeki oluşum sıklığının tespiti için kullanılmaktadır.

Tanım 2. Lokasyon yoğunluğu değeri verilmiş olan bir sosyal medya kullanıcısı u , bir lokasyon l ve minimum lokasyon yoğunluk eşik değeri $min_yoğunluk$ için, eğer l 'nin lokasyon yoğunluk değeri $min_yoğunluk$ eşik değerine eşit veya büyükse, lokasyon l **yaygın lokasyon** olarak tanımlanır; $\text{lokasyon yoğunluğu}_l^u \geq min_yoğunluk$

Tanım 3. Sosyal medya aktivite geçmişisi verilmiş olan bir sosyal medya kullanıcısı u ve bir lokasyon l için, l 'nin u için **ziyaret ömrü** kullanıcı u 'nun lokasyon l 'yi ilk ve son ziyareti arasındaki zaman farkının kullanıcı u 'nun sosyal medya geçmişindeki ilk ve son oluşumunun zaman farkına oranıdır.

$$\begin{aligned} \text{ziyaret ömrü}_l^u &= \frac{u'nun\ l'yi\ ilk\ ve\ son\ ziyaretinin\ farkı}{u'nun\ sosyal\ medya\ ömrü} \\ &= \frac{Son_Ziyaret_{l,u} - İlk_Ziyaret_{l,u}}{Son_Oluşum_u - İlk_Oluşum_u} \end{aligned}$$

Ziyaret ömrü, lokasyon l 'nin kullanıcı u için zamanda yaygınlığını tespit etmek için kullanılmaktadır.

Tanım 4. Ziyaret ömrü verilmiş olan bir sosyal medya kullanıcısı u , bir lokasyon l ve minimum ziyaret ömrü eşik değeri $min_ziyaret$ için, eğer l 'nin ziyaret ömrü değeri $min_ziyaret$ eşik değerine eşit veya büyükse, lokasyon l **zamanda yayılmış lokasyon** olarak tanımlanır; $ziyaret\ ömrü_l^u \geq min_ziyaret$.

Tanım 5. Bir sosyal medya kullanıcısı u ve bir lokasyon l için, eğer lokasyon l , Tanım 2 ve 4'te tanımlandığı gibi yaygın ve zamanda yayılmış lokasyon ise, lokasyon l **Kullanıcı Sosyal Önemli Lokasyon (KSÖL)** olarak tanımlanır.

Tanım 6. Verilen bir lokasyon l için, l 'nin **kullanıcı yaygınlığı** değeri, lokasyon l 'yi KSÖL olarak gören sosyal medya kullanıcılarının, toplam sosyal medya kullanıcı sayısına oranıdır.

$$kullanıcı\ yaygınlığı_l = \frac{l'yi\ KSÖL\ olarak\ gören\ kullanıcı\ sayısı}{toplam\ kullanıcı\ sayısı}$$

Kullanıcı yaygınlığı lokasyon l 'nin sosyal medya kullanıcı grubu içindeki popülaritesini tespit etmek için kullanılmaktadır.

Tanım 7. Kullanıcı yaygınlığı verilmiş olan bir lokasyon l ve minimum kullanıcı yaygınlık eşik değeri min_KY için, eğer lokasyon l 'nin yaygınlığı, min_KY eşik değerine eşit veya büyükse, lokasyon l **Sosyal Önemli Lokasyon (SÖL)** olarak tanımlanır; $kullanıcı\ yaygınlığı_l \geq min_KY$

5.2.2. Problem Tanımlaması

Verilenler:

- Mekânsal bilgi içeren sosyal medya kullanıcı veri kümesi D .
- Minimum lokasyon yoğunluğu eşik değeri, $min_yoğunluk$.
- Minimum ziyaret ömrü eşik değeri, $min_ziyaret$.
- Minimum kullanıcı yaygınlığı eşik değeri, min_KY .

Aranan: $\{L \mid L, \text{ Tanım 7'de tanımlandığı şekilde Mekânsal Sosyal Önemli Lokasyonlar kümesi} \}$

Hedef: Hesaplama maliyetini minimize etmek.

Sabitler: Kriterleri sağlayan mekânsal sosyal önemli lokasyonları keşfetmek

5.3.Sosyal Önemli Lokasyonların Keşfi

Bu bölümde ilk olarak sosyal önemli lokasyonların madenciliği için bir toy yaklaşım tartışılmış ve ardından SS-ILM algoritması önerilmiştir. Daha sonra SS-ILM algoritmasına ait örnek bir çalışma verilmiştir.

Sosyal önemli lokasyonların keşfi için, öncelikle kullanıcı-seviyesinde sosyal önemli lokasyonlar madenciliği, daha sonra grup-seviyesinde sosyal önemli lokasyonlar madenciliği uygulanmıştır. Asıl zaman harcayan kısım kullanıcı-seviyesinde sosyal önemli lokasyonlar madenciliğidir. Bu aşamada iki ana adım bulunmaktadır; lokasyon yoğunluk değerlerinin hesaplanması ve ziyaret ömrü değerlerinin hesaplanmasıdır. Bu bölümde önerilen algoritmalar her iki hesaplamayı yapmaktadır ancak eleme stratejileri farklılık göstermektedir.

5.3.1. Toy Yaklaşım

Toy yaklaşım, lokasyon yoğunluğu ve ziyaret ömrü değerlerini ayrı ayrı hesaplar ve ardından kullanıcı-seviyesinde sosyal önemli lokasyonların keşfi için kullanılan *min_yoğunluk* ve *min_ziyaret* eşik değerlerini sağlayan yaygın ve zamanda yayılmış lokasyonların keşfini gerçekleştirir. Daha sonra *min_KY* eşik değerini sağlayan grup-seviyesinde sosyal önemli lokasyonların keşfi gerçekleştirilir. Bu yaklaşımda yaygın ve zamanda yayılmış lokasyonlar ayrı ayrı keşfedilir ve daha sonra sosyal önemli lokasyonlar keşfedilir. Toy algoritmaya ait kaba kod Algoritma 5.1’de verilmiştir.

Algoritma 5.1’de, 1. adım parametrelerin başlangıç değerlerini ayarlamaktadır, 2-12. adımlar arasında kullanıcı-seviyesinde sosyal önemli lokasyonlar keşfi işlemi gerçekleştirilmektedir ve 13 ve 14. adımlarda grup-seviyesinde sosyal önemli lokasyonlar keşfi yapılmaktadır. 2-12. adımlar arası bütün sosyal medya kullanıcıları için ve 4-10. adımlar arası *u* kullanıcısının tüm lokasyonları için koşturulmaktadır. 5. adımda *l* lokasyonunun lokasyon yoğunluk değeri *lokasyon-yoğunluğu-hesapla* fonksiyonu yardımıyla ve 6. adımda ziyaret ömrü değeri *ziyaret-ömrü-hesapla* fonksiyonu yardımıyla hesaplanmaktadır. 7-9. adımlar arasında *min_yoğunluk* ve *min_ziyaret* eşik değerlerini sağlayan yaygın ve zamanda yayılmış lokasyonlar keşfedilmektedir. 11. adımda *u* sosyal medya kullanıcısına ait yaygın ve zamanda yayılmış lokasyonlar lokasyon setine eklenmektedir. 13. adımda kullanıcı yaygınlığı değeri *KY-hesapla* fonksiyonu yardımıyla

hesaplanmaktadır. 14. adımda min_KY eşik değerini sağlayan lokasyonlar *sosyal-önemli-lokasyonlar-çıkarmak* fonksiyonu yardımıyla çıkarılmaktadır. Son olarak 15. Adımda mekânsal sosyal önemli lokasyonlar (SÖL) algoritma çıktısı olarak verilmektedir.

Algoritma 5.1. Toy Algoritma-Mekânsal Sosyal Önemli Lokasyonlar Madenciliği

Girdiler:

D : Sosyal medya kullanıcı veri kümesi,
 $min_yoğunluk$: Minimum lokasyon yoğunluğu eşik değeri
 $min_ziyaret$: Minimum ziyaret ömrü eşik değeri
 min_KY : Minimum kullanıcı yaygınlığı eşik değeri
 L : Çıkarılmış ve etiketlenmiş lokasyonlar seti

Çıktı: Mekânsal Sosyal Önemli Lokasyonlar kümesi, $SÖL$

1. $tumLokasyonlar = \text{null}$, $L = \text{null}$
2. **for each** kullanıcı veri kümesi u **in** D
3. $KSÖL = \text{null}$
4. **for each** lokasyon l **in** $L[u]$
5. $lokasyon_yoğunluğu \leftarrow \text{lokasyon-yoğunluğu-hesapla}(l, u)$
6. $ziyaret_ömrü \leftarrow \text{ziyaret-ömrü-hesapla}(l, u)$
7. **if** $lokasyon_yoğunluğu \geq min_yoğunluk$ **and** $ziyaret_ömrü \geq min_ziyaret$
8. $KSÖL \leftarrow l$
9. **end if**
10. **end for**
11. $tumLokasyonlar \leftarrow KSÖL$
12. **end for**
13. $lokasyonlar = KY\text{-hesapla}(tumLokasyonlar)$
14. $SÖL = \text{sosyal-önemli-lokasyonlar-çıkarmak}(lokasyonlar, min_KY)$
15. **return** $SÖL$

5.3.2. SS-ILM Algoritması

Toy algoritma sosyal medya kullanıcılarının sosyal önemli lokasyonlarını keşfetmek için en doğrudan yaklaşımdır. Yaygın ve zamanda yayılmış lokasyonları ayrı ayrı keşfetmektedir ve bu durum gereksiz hesaplama maliyetine yol açmaktadır. Toy algoritmanın bu sınırlamasını ortadan kaldırmak ve sosyal önemli lokasyonları keşfetmek için SosyoMekânsal Önemli Lokasyonlar Madenciliği (SocioSpatial Important Locations Mining – SS-ILM) algoritması önerilmiştir. Önerilen algoritma ilk olarak $min_yoğunluk$ eşik değerini sağlayan yaygın lokasyonları keşfetmektedir ve daha sonra yaygın

lokasyonları kullanarak $min_ziyaret$ eşik değerini sağlayan zamanda yayılmış lokasyonları keşfetmektedir. SS-ILM algoritmasında yaygın ve zamanda yayılmış lokasyonların keşfi aşamasında eleme işlemi mümkün olduğu kadar erken uygulanmıştır ve gereksiz hesaplamaların önüne geçilmiştir. SS-ILM algoritmasına ait kaba kod Algoritma 5.2’de verilmiştir.

Algoritma 5.2. SS-ILM Algoritması

Girdiler:

D : Sosyal medya kullanıcı veri kümesi,

$min_yoğunluk$: Minimum lokasyon yoğunluğu eşik değeri

$min_ziyaret$: Minimum ziyaret ömrü eşik değeri

min_KY : Minimum kullanıcı yaygınlığı eşik değeri

L : Çıkarılmış ve etiketlenmiş lokasyonlar seti

Çıktı: Mekânsal Sosyal Önemli Lokasyonlar kümesi, $SÖL$

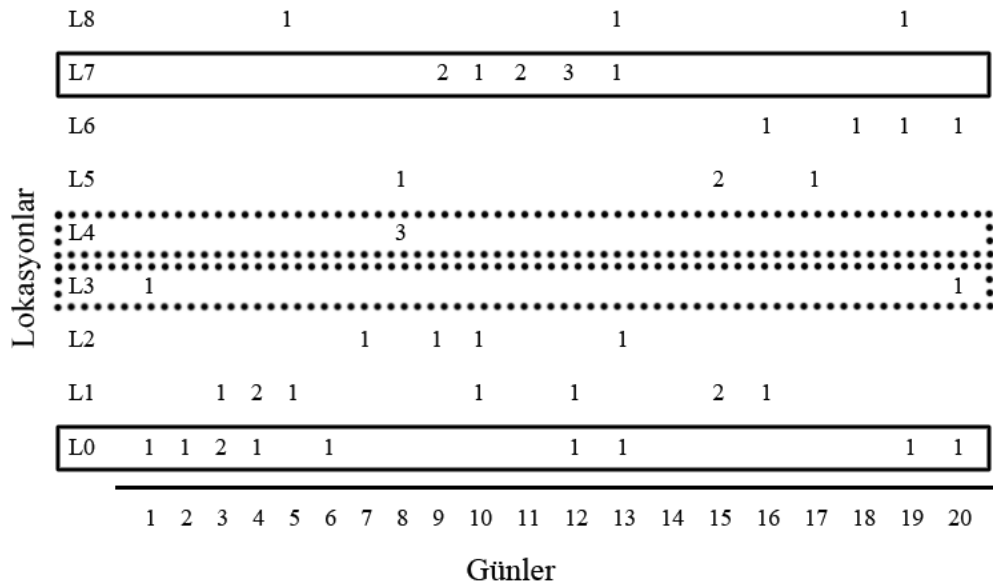
1. $tumLokasyonlar = \text{null}$, $L = \text{null}$
2. **for each** kullanıcı veri kümesi u **in** D
3. $KSÖL = \text{null}$
4. **for each** lokasyon l **in** $L[u]$
5. $lokasyon_yoğunluğu \leftarrow \text{lokasyon-yoğunluğu-hesapla}(l, u)$
6. **if** $lokasyon_yoğunluğu \geq min_yoğunluk$
7. $ziyaret_ömrü \leftarrow \text{ziyaret-ömrü-hesapla}(l, u)$
8. **if** $ziyaret_ömrü \geq min_ziyaret$
9. $KSÖL \leftarrow l$
10. **end if**
11. **end if**
12. **end for**
13. $tumLokasyonlar \leftarrow KSÖL$
14. **end for**
15. $lokasyonlar = KY\text{-hesapla}(tumLokasyonlar)$
16. $SÖL = \text{sosyal-önemli-lokasyonlar-çıkar}(lokasyonlar, min_KY)$
17. **return** $SÖL$

Algoritma 5.2’de 5. adımda l lokasyonunun u kullanıcısı için lokasyon yoğunluk değeri hesaplanmaktadır. 6. adımda l ’nin $min_yoğunluk$ eşik değerini sağlayıp sağlamadığı, yani yaygın lokasyon olup olmadığı, kontrol edilmektedir. Eğer yaygın lokasyon ise zamanda yaygınlığını kontrol etmek için 7-10. adımlar işletilmektedir. 7. adımda l ’nin ziyaret ömrü değeri hesaplanmakta ve 8. adımda eğer ziyaret ömrü değeri $min_ziyaret$ eşik değerini sağlıyorsa kullanıcı sosyal önemli lokasyonlar setine eklenmektedir. Yaygın olmayan

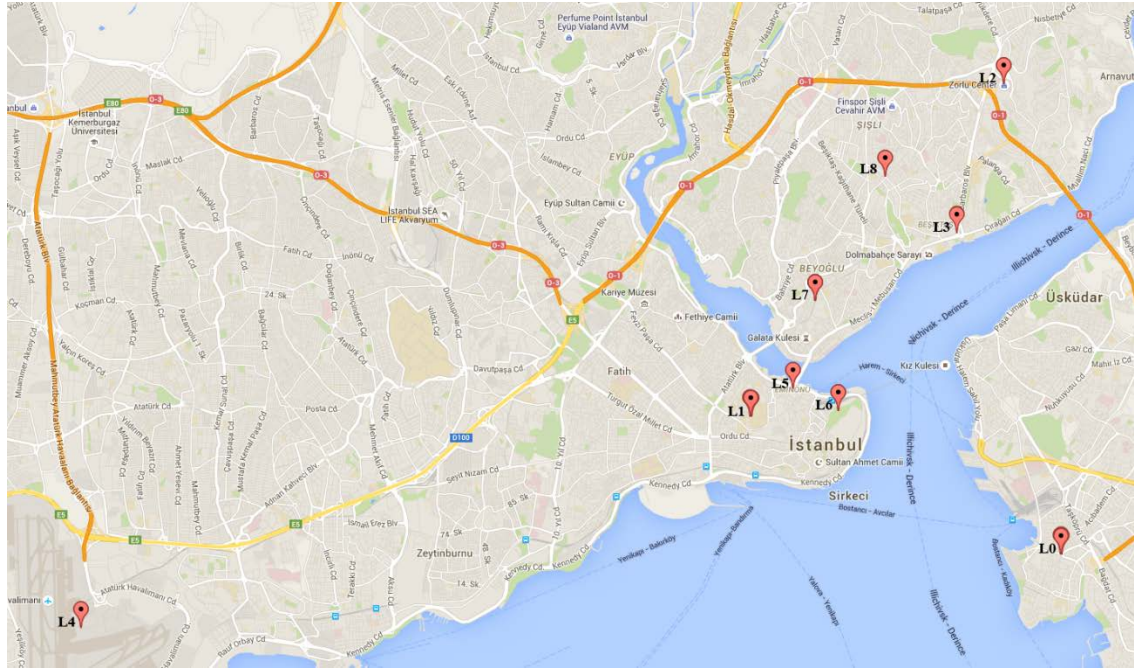
lokasyonların ziyaret ömrü değerlerinin gereksiz hesaplanması SS-ILM algoritmasına işlem hızı sağlamaktadır.

5.3.3. Algoritmanın Örnek Çalışması

Bu bölümde SS-ILM algoritmasına ait bir örnek çalışma verilmiştir. Kullanıcı-seviyesinde mekânsal sosyal önemli lokasyonlar keşfi örnek bir kullanıcının örnek veri kümesi üzerinde ve grup-seviyesinde mekânsal sosyal önemli lokasyonlar keşfi ise 5 örnek kullanıcının örnek veri kümeleri üzerinde gerçekleştirilmiştir. Şekil 5.2’de farklı lokasyonlarda (L0, L1, ..., L8) örnek bir sosyal medya kullanıcısının 20 günlük sosyal medya aktivitesi (Twitter veya Facebook aktivitesi) gösterilmiştir. Günler şeklin altında, lokasyonlar ise şeklin sol tarafında gösterilmiştir. Kullanıcıya ait sosyal medya aktivitesi 9 farklı lokasyonda (lokasyonlar L0...L8) oluşmuştur ve lokasyonların konumları Şekil 5.3’te gösterilmiştir. Şekil 5.2’de görüldüğü üzere, Gün 1’de, L0 lokasyonunda bir sosyal medya aktivitesi (tweet paylaşma) ve Gün 8’de, L4 lokasyonunda üç sosyal medya aktivitesi (üç tweet paylaşma) kaydedilmiştir.



Şekil 5.2. Bir kullanıcının farklı lokasyonlardaki sosyal medya aktivite kayıtları. Şekildeki numaralar, kullanıcının ilgili gün ve lokasyondaki aktivite sayısını göstermektedir.



Şekil 5.3. Örnek kullanıcının lokasyonlarının haritadaki gösterimi

Şekil 5.2’de görüldüğü üzere, kullanıcının sosyal medya aktivitelerinin çoğunluğu düz çizgi ile çerçevelenmiş olan L0 ve L7 lokasyonlarında oluşmuştur. L0, zamanda yayılmış ve önemli lokasyon örneğidir. Kullanıcı, L0 lokasyonunda toplam gün sayısının neredeyse yarısı süresince aktivitesi oluşmaktadır, bununla birlikte günlerdeki sosyal medya aktivite sayısı azdır. L7, yaygın ve önemli lokasyon örneğidir. Kullanıcı L7’de 5 günde sosyal medya aktivitesine sahiptir ve aktivite sayısı fazladır. Şekil 5.2’de görüldüğü üzere, L0 düşük sosyal medya aktivite oranıyla sürekli ziyaret edilen lokasyon iken, L7 kısa süreli ziyaret edilen ancak yüksek sosyal medya aktivite oranına sahip olan bir lokasyondur.

Şekil 5.2’de kesikli çizgilerle çerçevelenmiş olan iki adet kritik lokasyon bulunmaktadır ve bunlar L3 ve L4’tür. L3’te kullanıcı 20 gün boyunca 2 defa sosyal medya aktivitesi yapmıştır. L4’te ise sadece bir günde 3 defa sosyal medya aktivitesi bulunmaktadır. Bu kritik lokasyonlar *min_yoğunluk* ve *min_ziyaret* eşik değerleriyle elenmektedir. Bu örnek çalışma için *min_yoğunluk* ve *min_ziyaret* eşik değerleri, sırasıyla, 4/48 ve 4/20 olarak ayarlanmıştır. Tablo 5.1’de, Şekil 5.2’de görülen lokasyonlar üzerinde eşik değerleri uygulandığı zaman keşfedilen ve elenen lokasyonlar görülmektedir. Tabloda görüldüğü üzere, L3 ve L4, *min_yoğunluk* ve *min_ziyaret* eşik değerlerini sağlamadıkları için elenmişlerdir.

Tablo 5.1. Örnek kullanıcı için lokasyon yoğunluğu ve ziyaret ömrü değerleri ve keşfedilen/elenen lokasyonlar

	L0	L1	L2	L3	L4	L5	L6	L7	L8
Lok. Yoğunluğu	10/48	9/48	4/48	2/48	3/48	4/48	4/48	9/48	3/48
Yaygın mı?	Evet	Evet	Evet	Hayır	Hayır	Evet	Evet	Evet	Hayır
Ziyaret Ömrü	20/20	14/20	7/20	-	-	10/20	5/20	5/20	-
Zamanda Yayılmış mı?	Evet	Evet	Evet	-	-	Evet	Evet	Evet	-
Keşif/Ele	Keşif	Keşif	Keşif	Ele	Ele	Keşif	Keşif	Keşif	Ele

Tablo 5.2. 5 sosyal medya kullanıcısı için 20 günlük ziyaret edilen lokasyonlar

Gün	Kullanıcı 1	Kullanıcı 2	Kullanıcı 3	Kullanıcı 4	Kullanıcı 5
1	L0, L3	L0	L10	L0	L7
2	L0	L1	L0, L1, L3	L2, L10	L5
3	L0, L0, L1	L2, L3	L2, L6	L2, L2	L1, L11
4	L1, L0, L1	L6	L0, L7	L4	L0
5	L1, L8	L0, L1	L1, L9	L6	L2
6	L0	L6, L9	L6	L4, L4	L7, L11
7	L2	L10	L0, L3	L10	L0, L5
8	L4, L4, L5, L4	L2, L6	L4, L12	L4, L6	L9
9	L2, L7, L7	L0, L3	L6	-	L2
10	L1, L7, L2	L0, L1	L1, L3	L0, L4	L9
11	L7, L7	L0, L0, L2	L2, L12	L7, L10	L1, L5
12	L7, L0, L7, L7, L1	L7	L0, L4	L0, L12	L7
13	L0, L8, L2, L7	L2, L3	L4, L12	L10	L1, L1
14	-	L0	L1, L2, L3	L6	L1, L1
15	L5, L1, L1, L5	L1, L6	L6	L10	L1, L11
16	L1, L6	L6, L12	L1, L1, L4	L6	L5
17	L5	L9	L7	L6, L6, L8	L11
18	L6	L1, L3	L9	L8	L6, L7
19	L6, L8, L0	-	L10	L8	-
20	L0, L3, L6	L7, L10	L10, L10	L8	L11

Tablo 5.2, 5 kullanıcının 20 günlük sosyal medya aktivitelerine ait lokasyonları göstermektedir. Bir sosyal medya kullanıcısı bir günde birden fazla sayıda sosyal medya aktivitesine sahip olabileceği için, bu lokasyonlar virgülle ayrılmıştır. Tablo 5.3 her bir kullanıcı için keşfedilen lokasyonları ve lokasyonların kullanıcı yaygınlığı değerlerini

göstermektedir. Bu örnek çalışmada min_KY eşik değeri 3/5 olarak ayarlanmıştır. min_KY eşik değerini sağlayan lokasyonlar sosyal önemli lokasyon olarak keşfedilmiştir. Tablo 5.3'te görüldüğü üzere, L0, L1, L2 ve L6 verilen eşik değerlerini sağlamakta ve algoritma bu lokasyonları sosyal önemli lokasyon olarak vermektedir.

Tablo 5.3. 5 kullanıcı için keşfedilen/elenecek lokasyonlar, KY değerleri ve örnek veri kümesinin sosyal önemli lokasyonları (SÖL)

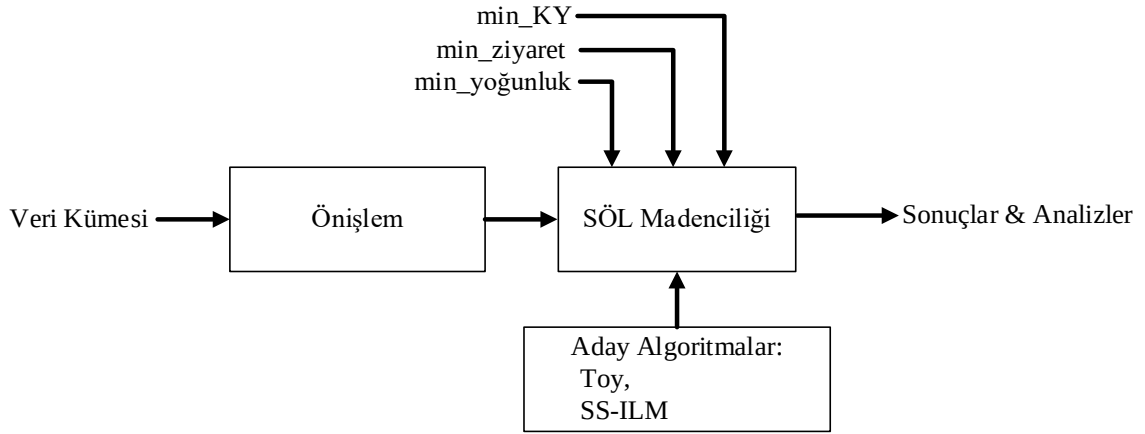
Lokasyonlar	Kul. 1	Kul. 2	Kul. 3	Kul. 4	Kul. 5	KY	SÖL mu?
L0	Keşif	Keşif	Keşif	Keşif	Ele	4/5=0.8	Evet
L1	Keşif	Keşif	Keşif	Ele	Keşif	4/5=0.8	Evet
L2	Keşif	Keşif	Keşif	Ele	Ele	3/5=0.6	Evet
L3	Ele	Keşif	Keşif	Ele	Ele	2/5=0.4	Hayır
L4	Ele	Ele	Keşif	Keşif	Ele	2/5=0.4	Hayır
L5	Keşif	Ele	Ele	Ele	Keşif	2/5=0.4	Hayır
L6	Keşif	Keşif	Keşif	Keşif	Ele	4/5=0.8	Evet
L7	Keşif	Ele	Ele	Ele	Keşif	2/5=0.4	Hayır
L8	Ele	Ele	Ele	Keşif	Ele	1/5=0.2	Hayır
L9	Ele	Ele	Ele	Ele	Ele	0/5=0.0	Hayır
L10	Ele	Ele	Keşif	Keşif	Ele	2/5=0.4	Hayır
L11	Ele	Ele	Ele	Ele	Keşif	1/5=0.2	Hayır
L12	Ele	Ele	Keşif	Ele	Ele	1/5=0.2	Hayır

5.4.Deneysel Çalışmalar

Bu bölümde deneysel çalışmalar sunulmuştur. Deneylerde, sosyal medya aktivite veri kümesi olarak gerçek Twitter veri kümesi kullanılmıştır. Veri kümesi hazırlığı Bölüm 5.4.1 ve 5.4.2'de anlatılmıştır. Şekil 5.4'te bu çalışmaya ait deney düzeneği verilmiştir. İlk olarak veri kümesi sisteme girdi olarak verilmiş ve bir önışlem aşaması gerçekleştirilmiştir (Bölüm 5.4.2). Önışlem aşamasında veri temizleme (Bölüm 5.4.2.1), kullanıcı seçimi (Bölüm 5.4.2.2), zamanda aşırı ağırlıklandırma önleme (Bölüm 5.4.2.3) ve lokasyon etiketleme (Bölüm 5.4.2.4) işlemleri gerçekleştirilmiştir. Önışlemden aşamasından sonra, SS-ILM ve Toy algoritmalar veri kümesi üzerinde koşturulmuştur. Algoritmalar mekânsal sosyal önemli lokasyonları keşfetmek için üç eşik değeri istemektedir; $min_yoğunluk$, $min_ziyaret$ ve min_KY . Lokasyon yoğunluğu eşik değeri $min_yoğunluk$, kullanıcı-seviyesinde yaygın lokasyonların keşfi için, ziyaret ömrü eşik değeri $min_ziyaret$ ise kullanıcı-seviyesinde zamanda yayılmış lokasyonların keşfi için kullanılmaktadır. Kullanıcı yaygınlığı eşik değeri min_KY grup-seviyesinde mekânsal sosyal önemli lokasyonların keşfi için kullanılmaktadır.

SS-ILM ve Toy algoritmaların performansları aşağıdaki sorulara çözüm bulmak için değerlendirilmiştir:

- Sosyal medya kullanıcı sayısının (veri kümesi büyüklüğünün) algoritmalar üzerindeki etkisi nedir?
- Minimum lokasyon yoğunluk eşik değerinin etkisi nedir?
- Minimum ziyaret ömrü eşik değerinin etkisi nedir?
- İstanbul, Türkiye’de yaşayan kullanıcılar için hangi lokasyonlar sosyal olarak önemlidir?



Şekil 5.4. Deney düzeneği

Deneyler 3.40 GHz hızdaki Intel Core i7 işlemcili ve 8 GB RAM’li bir bilgisayarda gerçekleştirilmiştir.

5.4.1. Veri Kümesi

Sosyal medya ağları, Twitter da dâhil olmak üzere, geliştiriciler için Uygulama Programlama Arayüzleri (Application Programming Interfaces – APIs) sağlamaktadır. Bu çalışmada sosyal medya aktivite veri kümesi olarak coğrafi bilgiye sahip olan Twitter veri kümesi seçilmiştir. Twitter’den veri toplamak için, REST API ve Streaming API kullanılmıştır [18]. Buna ek olarak Twitter4j açık kaynak kodlu Java kütüphanesi sorguların iletilmesi ve Twitter API’lerinden sonuçların toplanması için kullanılmıştır [19]. Streaming API akış halinde olan tweetlerden coğrafi arama işlemi sağlamaktadır. İlk olarak İstanbul, Türkiye tabanlı coğrafi arama işlemi gerçekleştirilmiş ve ardından kullanıcılar toplanmıştır. Yaklaşık olarak 2500 kullanıcı bu adımda toplanmıştır. Ardından REST API kullanılarak bu kullanıcılara ait tüm tweetler toplanmıştır. Her bir

tweet için üç parametre toplanmıştır; tarih/saat, enlem ve boylam. Veri kümesi Eylül 2008 ile Aralık 2015 tarihleri aralığındadır.

5.4.2. Önışlem

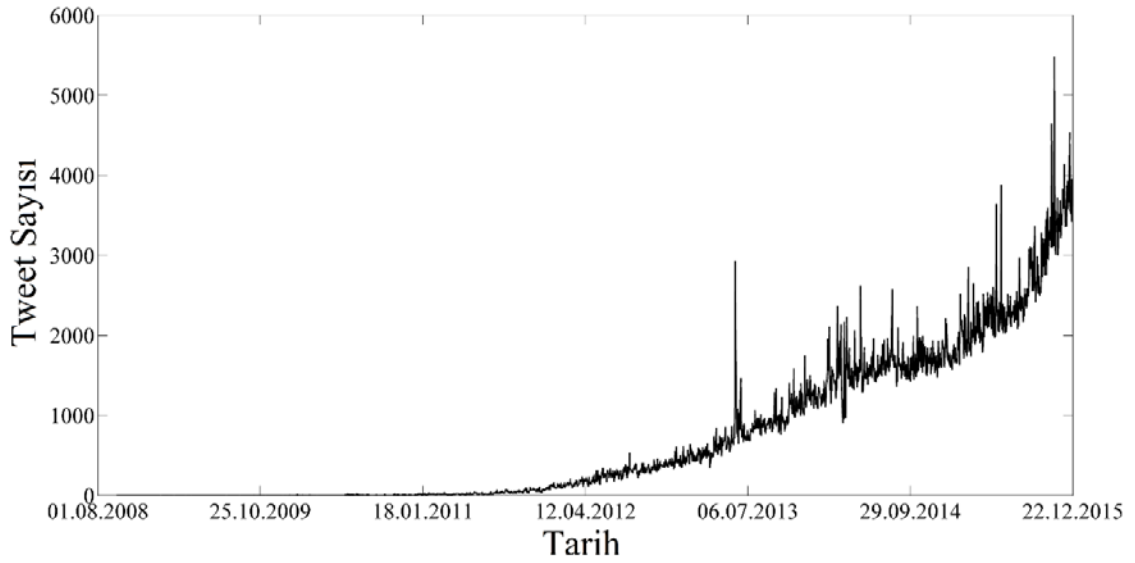
Bu bölümde, veri temizleme, kullanıcı seçimi, zamanda aşırı ağırlıklandırma önleme ve lokasyon etiketleme işlemleri açıklanmıştır.

5.4.2.1. Veri Temizleme

Deneylerde, aktif Twitter kullanıcılarının verisi kullanılmıştır. Bu çalışmada, aktif Twitter kullanıcıları, 2008 ile 2015 arasında 50'den fazla tweet gönderen kullanıcılar olarak tanımlanmıştır. Eğer tweet sayısı az ise, bu kullanıcı ya pasif ya da yeni kullanıcıdır. Ancak, aktif kullanıcı tanımı kullanıcı tercihleri veya uygulamalara göre değişkenlik gösterebilir. Veri kümesindeki kullanıcıların bir bölümü de sahte (spam) kullanıcılardan oluşmaktadır. Sahte kullanıcılardan korunmak için iki kriter kullanılmıştır ve bunlar takipçi sayısı ve takipçi/arkadaş oranıdır. Eğer bir kullanıcının takipçi sayısı 10'un ve takipçi/arkadaş oranı 0.1'in altında ise bu kullanıcı sahte kullanıcı olarak tespit edilmiş ve veri kümesi içerisine alınmamıştır. Bu parametreler için seçilen değerler pek çok sahte kullanıcı tespiti literatürüne göre seçilmiştir ve detaylı bilgiler [97] ve [98]'te bulunabilir.

5.4.2.2. Kullanıcı Seçimi

Bu çalışmada asıl odaklanılan konu İstanbul, Türkiye içerisinde tweet gönderen Twitter kullanıcılarının (İstanbul sakinlerinin) mekânsal sosyal önemli lokasyonlarını keşfetmektir. Kullanıcılar Streaming API aracılığıyla toplandığı için ve bu API geçmişe dayalı verileri döndürmediği için, veri toplama aşamasında kullanıcıların İstanbul içerisinde yeterli sayıda tweete sahip olduğu garanti edilememiştir. Bir kullanıcının İstanbul sakini olarak belirlenebilmesi için, kullanıcının tweetlerinin enlem ve boylamları İstanbul merkez enlem ve boylamından çıkarılmıştır. Eğer sonuçta çıkan fark kullanıcı tweetlerinin yarısından fazlası için anlamlı ise (örneğin İstanbul için en fazla 40 km ise) bu kullanıcı İstanbul sakini olarak seçilmiştir.



Şekil 5.5. Veri kümesinin zamansal dağılımı

Bu işlem uygulanarak deneyler için 1000 kullanıcı gerçek ve aktif İstanbul Twitter kullanıcıları olarak belirlenmiştir. Veri kümesinde 1.886.065 tweet bulunmaktadır ve veri kümesi Eylül 2008 ile Aralık 2015 tarihleri arasındadır. Veri kümesinin zamansal dağılımı Şekil 5.5'te verilmiştir. Şekilde görüldüğü üzere, tweetlerin çoğunluğu 2013-2015 aralığındadır. Bu durumun asıl sebebi, Türkiye sosyal medya kullanıcıları için Twitter'ın 2012 yılından sonra popülerlik kazanmasıdır.

5.4.2.3. Zamanda Aşırı Ağırlıklandırılmanın Önlenmesi

Veri kümesini incelediğimiz zaman, kullanıcıların bir lokasyonda aynı zamanda birden fazla tweet gönderdikleri (sosyal medya aktivitesi gerçekleştirdiklerini) görülmüştür. Bu davranış sürekli hale geldiği zaman, bir lokasyon doğru bulunma sayısından fazla görülmektedir, çünkü kullanıcı lokasyona bir defa gitmiş ancak birden fazla defa tweet göndermiştir. Bu problem bir lokasyonun zamanda aşırı ağırlıklandırılması olarak tanımlanmıştır. Bu problem bazı durumlarda istenmeden, örneğin, bir kullanıcı arkadaşlarıyla sohbet ederken, gerçekleşebilmektedir. Bir lokasyonun zamanda aşırı ağırlıklandırılmasını önlemek için, 60 dakikalık bir eşik değeri tanımlanmıştır. Eğer bir kullanıcı 60 dakika içerisinde aynı lokasyonda birden fazla defa tweet paylaşırsa, bu lokasyon bilgisinin yeni olmadığı varsayılmış ve lokasyonda oluşumu bir defa olarak sayılmıştır. Bu yaklaşım yardımıyla bir lokasyonun zamanda aşırı ağırlıklandırılması önlenmiş olmaktadır. Bunun yanında, zamanda aşırı ağırlıklandırılmayı önlemek için farklı yaklaşımlar/kriterler de geliştirilebilir.

5.4.2.4.Lokasyon Etiketleme

Twitter API'leri kullanıcı tweetlerine ait hassas enlem-boylam ikilileri sağlamaktadır. Bu yaklaşım hassas sonuçlar elde etmek için faydalı olsa da, lokasyon etiketleme işlemi için bir problemdir. Örneğin, bir alışveriş merkezi veya stadyum 1 km² alan içerisinde bulunmaktadır ancak hassas enlem-boylam ikilileri eşleşmediği için çok sayıda farklı lokasyon tanımlanabilmektedir. Bu problemin üstesinden gelebilmek için farklı enlem-boylam ikililerinin aynı lokasyon olarak sayılabilmesi için bir eşik değeri tanımlanmıştır. Daha önce [52] ve [47]'de yapılan çalışmalarda kullanıldığı gibi, bu eşik değeri 100 metre olarak tanımlanmıştır. Eğer iki lokasyon birbirine 100 metreden daha yakınsa, bu iki lokasyon aynı etiketle etiketlenmektedir.

5.4.3. Deneysel Sonuçlar

Bu bölümde, öncelikle SS-ILM ve Toy algoritmalarının performanslarını değerlendirmek için deneyler sunulmuştur. Ardından, İstanbul sakini sosyal medya kullanıcıları için keşfedilen mekânsal sosyal önemli lokasyonlar sunulmuştur.

5.4.3.1.Kullanıcı Sayısının Etkisi

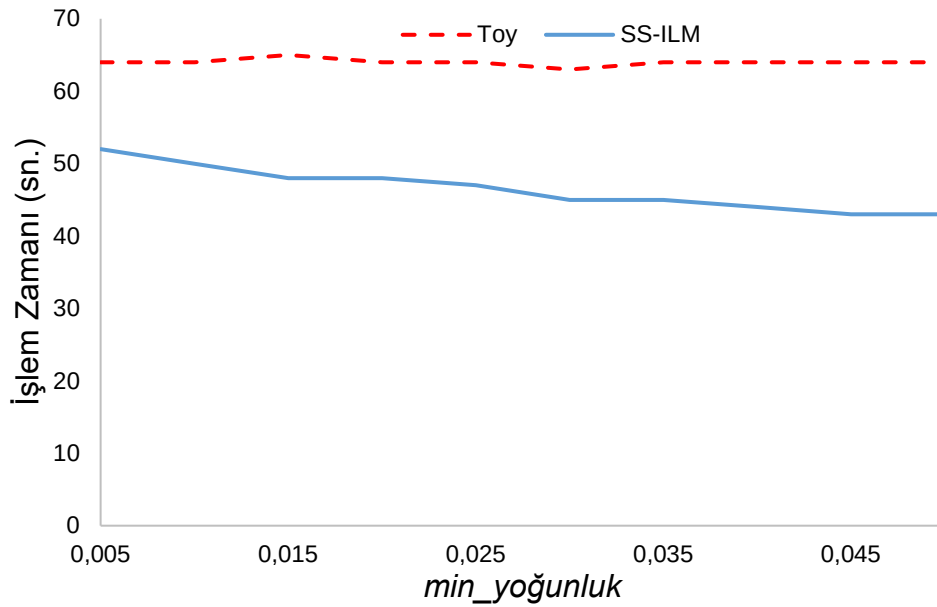
Bu deneyde kullanıcı sayısının algoritmalar üzerindeki etkisi değerlendirilmiştir. *min_yoğunluk*, *min_ziyaret* ve *min_KY* parametreleri, sırasıyla 0.01, 0.05 ve 0.01 olarak seçilmiştir. Kullanıcı sayısı 200'den 1000'e kadar 200'er artırılmıştır. Şekil 5.6'da kullanıcı sayısının algoritmalar üzerindeki etkisi görülmektedir. Şekilde görüldüğü üzere, her iki algoritma da kullanıcı sayısına karşı duyarlıdır ve kullanıcı sayısı arttıkça işlem zamanları artmaktadır. Deneysel sonuçlar SS-ILM algoritmasının Toy algoritmaya göre hesaplama açısından daha etkin olduğunu göstermektedir.



Şekil 5.6. Kullanıcı sayısının etkisi

5.4.3.2. Minimum Lokasyon Yoğunluk Eşik Değerinin Etkisi

Bu deneyde minimum lokasyon yoğunluk eşik değerinin (*min_yoğunluk*) algoritmaların işlem zamanı üzerindeki etkisi değerlendirilmiştir. Kullanıcı sayısı, *min_ziyaret* ve *min_KY* parametreleri, sırasıyla 1000, 0.05 ve 0.01 olarak seçilmiştir. *min_yoğunluk* eşik değeri 0.005'ten 0.05'e kadar 0.05'er artırılmıştır. Şekil 5.7'de *min_yoğunluk* eşik değerinin algoritmalar üzerindeki etkisi görülmektedir. Toy algoritma yaygın ve zamanda yayılmış lokasyonları ayrı ayrı keşfetmektedir. Buna karşılık SS-ILM algoritması öncelikle yaygın lokasyonları keşfetmekte ve yaygın lokasyonlardan zamanda yayılmış lokasyonları keşfetmektedir. Bu sebepten dolayı, şekilde görüldüğü üzere SS-ILM algoritmasının işlem zamanı Toy algoritmadan daha üstündür. SS-ILM algoritması erken eleme stratejisi sayesinde Toy algoritmaya göre yaklaşık %25 daha az zaman harcamaktadır.

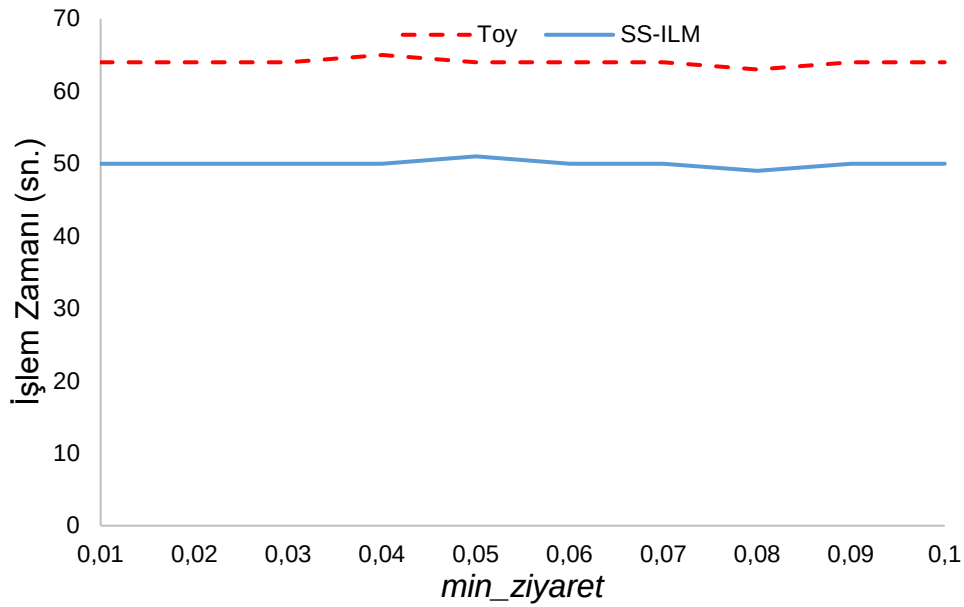


Şekil 5.7. Minimum lokasyon yoğunluk eşik değerinin etkisi

5.4.3.3. Minimum Ziyaret Ömrü Eşik Değerinin Etkisi

Bu deneyde, minimum ziyaret ömrü eşik değerinin ($min_ziyaret$) algoritmalar üzerindeki etkisi değerlendirilmiştir. Kullanıcı sayısı, $min_yoğunluk$ ve min_KY parametreleri, sırasıyla 1000, 0.01 ve 0.01 olarak seçilmiştir. $min_ziyaret$ eşik değeri 0.01'den 0.1'e kadar 0.01'er artırılmıştır. Şekil 5.8'de $min_ziyaret$ eşik değerinin algoritmaların işlem zamanına olan etkisi görülmektedir.

Şekil 5.8'de görüldüğü üzere, işlem zamanı eğilimi, $min_ziyaret$ eşik değeri kullanıcı-seviyesinde önemli lokasyonlar keşfinin son adımı olduğu için her iki algoritma için de aynıdır. Eğer, $min_ziyaret$ eşik değerinden sonra da bir işlem yapılmış olsa idi, algoritmaların işlem zamanlarını yavaş yavaş azaltmaları beklenecekti. Şekil 5.8'den de görülebileceği üzere, SS-ILM algoritması, lokasyonları erken eleme stratejisi sayesinde Toy algoritmaya karşı yaklaşık 10-15 saniye daha kısa sürede sonuç üretirek üstün gelmektedir.



Şekil 5.8. Minimum ziyaret ömrü eşik değerinin etkisi

5.4.3.4. İstanbul için Mekânsal Sosyal Önemli Lokasyonların Keşfi

Bu bölümde İstanbul için mekânsal sosyal önemli lokasyonların keşfi sonuçları sunulmuştur. Kullanıcı sayısı, *min_yoğunluk*, *min_ziyaret* ve *min_KY* parametreleri, sırasıyla 1000, 0.01, 0.05 ve 0.01 olarak seçilmiştir. Gösterim kolaylığı olması için keşfedilen en yüksek 10 sosyal önemli lokasyon sunulmuştur. Sonuçlar Şekil 5.9'da harita üzerinde ve Tablo 5.4'te koordinatları ve açıklamalarıyla verilmiştir. Lokasyon açıklamaları manuel olarak elde edilmiş, bu çalışmanın kapsamı dışında olduğu için isimlendirilmiş varlık tanıma (named entity recognition) gibi herhangi bir metot kullanılmamıştır.

Tablo 5.4'te keşfedilmiş olan 10 mekânsal sosyal önemli lokasyon listelenmiştir. İlk lokasyon, Kadıköy'ün merkezindedir ve İstanbul sakinleri için sosyal olarak önemli bir lokasyondur. İkinci lokasyon, Şişli'deki popüler alışveriş merkezlerinden biridir. Üçüncü lokasyon, Göztepe'deki Bağdat Caddesi'dir. Bağdat Caddesi'nde, pek çok alışveriş mekânı ve restoran bulunmaktadır. Dördüncü lokasyon, günlük çok sayıda yolcu taşıyan Atatürk Havalimanı'dır. Beşinci lokasyon, Beyoğlu'ndaki popüler bir caddedir. Altıncı lokasyon, Anadolu ve Avrupa yakalarını birleştiren köprülerden biridir. Yedinci lokasyon, çok sayıda seyircisi olan bir kültür merkezidir. Sekizinci lokasyon İstanbul Boğazı boyunca devam eden bir caddedir. Dokuzuncu lokasyon, bir üniversite kampüsüdür. Son olarak onuncu lokasyon, Beyoğlu, Şişli ve Beşiktaş'ın kesişimidir.



Şekil 5.9. 1000 İstanbul sakini (kullanıcısı) için mekânsal sosyal önemli lokasyonlar

Tablo 5.4. Keşfedilen mekânsal sosyal önemli lokasyonların açıklaması

Sıra	Keşfedilen Mekânsal Sosyal Önemli Lokasyonlar
1	Kadıköy Boğa Heykeli
2	Şişli Cevahir AVM
3	Bağdat Cd., Göztepe
4	Atatürk Havalimanı
5	İstiklal Cd.
6	15 Temmuz Şehitler Köprüsü
7	Beşiktaş Kültür Merkezi
8	Ragıp Gümüşpala Cd., Eminönü
9	İstanbul Aydın Üniversitesi
10	Vali Konağı Cd., Teşvikiye Cd., Abdi İpekçi Cd.

Elde edilen sonuçlara göre, önerilen algoritmanın başarılı bir şekilde İstanbul sakini Twitter kullanıcılar için sosyal önemli lokasyonları keşfedebildiğini söyleyebiliriz. Ancak, Sultanahmet Meydanı veya Topkapı Sarayı gibi bazı turistik ve popüler lokasyonların algoritmalar tarafından keşfedilmediği de görülmüştür. Bunun sebebi, kullanılan veri kümesinin karakteristiğiyle ilgilidir. Bu çalışmada seçilen kullanıcılar, bu lokasyonlarda önemli oranda tweet göndermemişlerdir. Ayrıca, analizlerde İstanbul sakinlerinin seçilmesinin ve turistlerin seçilmemesinin de etkili olduğunun belirtilmesi gerekir.

5.4.4. Önerilen SS-ILM Algoritmasının Literatür ile Karşılaştırılması

Önerilen mekânsal önemli lokasyonlar keşfi çalışması grup seviyesinde ilginç/önemli lokasyonlar keşfi çalışmaları arasında yer almaktadır. Bildiğimiz kadarıyla grup seviyesinde ilginç/önemli lokasyonlar keşfi çalışmaları Zheng vd. [60] ve Khetarpaul vd. [61] tarafından yapılmıştır.

Zheng vd. [60] tarafından yürütülen çalışmada, önemli lokasyonları belirlemek için kullanıcı derecelendirmeleri (oyları) kullanılmaktadır. Tez çalışması kapsamında kullanılan verilerde kullanıcı derecelendirmesi bulunmadığından, bu çalışma tezde önerilen SS-ILM algoritması ile karşılaştırılamamıştır.

Önerilen SS-ILM algoritması Khetarpaul vd. [61] tarafından yapılan çalışma ile, üç kriter esas alınarak karşılaştırılmıştır. Bunlar keşfedilen ilk 10 sosyal önemli lokasyon, keşfedilen sosyal önemli lokasyonların sayısı ve algoritmaların çalışma zamanlarıdır. Birinci ve ikinci kriter hangi algoritmanın daha makul sonuçları bulduğunu ortaya koymaktadır. Üçüncü kriter ise sosyal önemli lokasyonların keşfinde, hangi algoritmanın daha az zaman harcadığını ortaya koymaktadır. Her iki algoritma tarafından keşfedilen ilk 10 sosyal önemli lokasyonlar karşılaştırıldığında sonuçların farklı olduğu görülmektedir. SS-ILM algoritması tarafından keşfedilen ilk 10 sosyal önemli lokasyon Tablo 5.4'te ve Khetarpaul vd. [61] tarafından önerilen algoritma ile keşfedilen ilk 10 sosyal önemli lokasyon Tablo 5.5'te sunulmaktadır. Sonuçlar karşılaştırıldığında, Kadıköy Boğa Heykeli ve Bağdat Caddesi, Göztepe olmak üzere iki lokasyonun her iki algoritmada aynı olduğu görülmektedir. Ragıp Gümüşpala Caddesi, Eminönü ve İstanbul Aydın Üniversitesi'nin yerleri SS-ILM algoritmasının ilk 10 listesinde sıralanmıştır ve Khetarpaul vd. [61] tarafından önerilen algoritmada listelenmemiştir. Her iki algoritma da Atatürk Havalimanı, İstiklal Caddesi ve 15 Temmuz Şehitler Köprüsü lokasyonlarını keşfetmiştir, ancak ilk 10'daki sıraları farklıdır. Algoritmalar arasındaki sıralama farkının nedeni Khetarpaul vd. [61], ziyaret frekanslarına dayalı olarak lokasyonları sıralarken, SS-ILM algoritması kullanıcıların yaygınlığına dayalı olarak lokasyonları sıralamaktadır. Lokasyonları ziyaret frekans değeri ile sıralama problemi bazı kullanıcıların belirli yerleri çok yoğun olarak ziyaret ederken bazılarının bu yerleri hiç ziyaret etmemeleri durumunda daha çok ortaya çıkmaktadır. Dolayısıyla, kullanıcı grupları için konumların gerçek sıralaması, frekansa dayalı olarak gerçekleştirilmesi uygun değildir. Bu sonuçlar, SS-ILM

algoritmasının hem sosyal önemli lokasyonları keşfetme hem de sıralamada daha başarılı olduğunu göstermektedir. Keşfedilen lokasyon sayısı açısından, Khetarpaul vd. [61] tarafından önerilen algoritma, 1000 kullanıcı için 179 lokasyonun sosyal önemli lokasyon olduğunu keşfetmiştir. Bu algoritmanın sonuçları minimum kullanıcı eşik değerine (ThresCount) bağlıdır. Bu deneyde, minimum kullanıcı sayısı (ThresCount) 80 olarak belirlenmiştir ve bu değer Khetarpaul vd. [61]'nin yürüttüğü çalışmada kullanılan değer ile aynıdır. SS-ILM algoritması, 27 lokasyonu sosyal önemli lokasyon olarak keşfetmiştir ve bu sonuçlar önerilen SS-ILM algoritmasının, önemsiz lokasyonları elemeye daha etkili olduğunu ortaya koymuştur.

Tablo 5.5. Khetarpaul vd. [61] tarafından önerilen yaklaşımın keşfettiği sosyal önemli lokasyonlar

Sıra	Keşfedilen Mekânsal Sosyal Önemli Lokasyonlar
1	Kadıköy Boğa Heykeli
2	Atatürk Havalimanı
3	Bağdat Caddesi, Göztepe
4	Şişli Cevahir AVM
5	15 Temmuz Şehitler Köprüsü
6	Beşiktaş Kültür Merkezi
7	Büyükdere Cd., Şişli
8	Vali Konağı Cd., Teşvikiye Cd., Abdi İpekçi Cd.
9	İstiklal Cd.
10	Ataşehir Cd., Ataşehir

Algoritmalar çalışma zamanları açısından değerlendirildiğinde, Khetarpaul vd. [61] tarafından önerilen algoritma, 1000 kullanıcı için sosyal önemli lokasyonları keşfederken 26 saniye zaman alırken, tez çalışması kapsamında önerilen SS-ILM algoritması 52-44 saniye zaman almaktadır. Elde edilen sonuçlar, Khetarpaul vd. [61] tarafından önerilen algoritmanın SS-ILM algoritmasından daha az zaman harcadığını göstermiştir.

Sonuç olarak, tez çalışması kapsamında önerilen SS-ILM algoritması önemsiz lokasyonların çoğunu eleyerek sosyal önemli lokasyonları keşfetmektedir. Dolayısıyla, önerilen algoritma, Khetarpaul vd. [61] tarafından yapılan çalışmadan daha az ve makul sayıda sosyal önemli lokasyonların keşfinde başarılıdır. Buna karşılık, önerilen SS-ILM algoritması, Khetarpaul vd. [61] tarafından önerilen algorithmadan daha fazla zaman harcamaktadır. Bu durum, her iki yaklaşımın farklılıklarından kaynaklanmaktadır.

Khetarpaul vd. [61] tarafından önerilen algoritma, lokasyonların ziyaret frekanslarını hesaba katarken, önerilen SS-ILM algoritması sosyal önemli lokasyonları kullanıcı yaygınlığına göre keşfetmektedir.

5.5.Katkılar ve Kapsam

Bu kısımda bu bölümde yapılan çalışmanın katkıları ve kapsamından bahsedilmiştir.

5.5.1. Katkılar

Bu bölümde yapılan çalışmanın katkıları aşağıdaki şekilde listelenebilir:

- Mekânsal sosyal önemli lokasyonların ölçülebilmesi için lokasyon yoğunluğu, ziyaret ömrü ve kullanıcı yaygınlığı değerlendirme ölçütleri önerilmiştir.
- Mekânsal sosyal önemli lokasyonların verimli bir şekilde keşfi için SosyoMekânsal Önemli Lokasyonlar Madenciliği (SocioSpatial Important Locations Mining – SS-ILM) algoritması önerilmiştir.
- Önerilen algoritma, İstanbul Twitter kullanıcılarının veri kümesi ile deneysel olarak değerlendirilmiştir.

5.5.2. Kapsam

Bu bölümde yapılan çalışma, sosyal medya veri kümeleri üzerinden mekânsal sosyal önemli lokasyonların madenciliği problemine çözümler üretmeye odaklanmıştır. Aşağıdaki konular bu çalışmanın kapsamı dışındadır:

- Sosyo-mekânsal önemli lokasyonların zamansal analizi,
- Keşfedilen lokasyonların anlamsal olarak isimlendirilmesi,
- Lokasyon etiketleme problemine yenilikçi çözümler.

6. BÖLÜM

BÜYÜK SOSYAL MEDYA VERİSİNDEN MEKÂN-ZAMANSAL ÖNEMLİ LOKASYONLARIN KEŞFİ

Bu bölümde, büyük sosyal medya verisi kullanılarak mekân-zamansal önemli lokasyonların keşfi için geliştirilen yaklaşım ve önerilen algoritmalar anlatılmıştır. Öncelikli olarak, önemli lokasyonlar madenciliğinin tanımı, zorlukları ve faydaları açıklanmıştır. Ardından, geliştirilmiş olan mekân-zamansal önemli lokasyonların keşfi yaklaşımına ait tanımlar ve önerilen algoritmalar verilmiştir. Son olarak deneysel çalışma sonuçları ile önerilen algoritmalar değerlendirilmiş ve sonuç kısmıyla bölüm tamamlanmıştır.

6.1.Giriş

Son yıllarda, GPS takılı cihazlar yaygınlaşmış ve sosyal medya ağları insanların gündelik yaşantılarının önemli bir parçası haline gelmiştir. İnsanlar, bulundukları lokasyonları, sosyal medya ağları aracılığıyla arkadaş ağlarında paylaşmayı ve takip ettikleri arkadaşlarının güncel lokasyonlarını takip etmeyi sevmektedir. Bu paylaşımlar, araştırmacıları sosyal medya verisinin mekânsal ve zamansal bilgisinin analiz edilmesine yöneltmiştir [34] [20]. Lokasyon bilgisine sahip sosyal medya veri kümeleri ve bu sosyal medya paylaşımlarının zamansal bilgisi kullanılarak insanların gündelik hayatları ve alışkanlıkları ile ilgili değerli bilgiler keşfedilebilir. Bu bölümdeki çalışmaların amacı, sosyal medya kullanıcılarının sosyal medya geçmişlerinde sıklıkla ziyaret ettikleri sosyo-mekân-zamansal önemli lokasyonların (SMZÖL) keşfedilmesidir.

SMZÖL'lerin keşfedilmesi, sosyal medya kullanıcılarının ziyaret ettikleri sosyal mekânsal önemli lokasyonların zamansal önem dağılımının keşfedilmesini sağlayacaktır ve tavsiye sistemleri, reklamcılık uygulamaları ve şehir planlamacılığı gibi çeşitli uygulama alanları için önemlidir. Ayrıca bu lokasyonların keşfedilmesi, sosyal medya

kullanıcılarının günlük, haftalık ve uzun dönem örüntülerinin anlamlandırılabilmesine yardımcı olacaktır. Bu örüntülerin keşfi, kullanıcıların hangi lokasyonları sevdikleri, hangi lokasyonlarda hiç bulunmadıkları, hangi lokasyonları sıklıkla ziyaret ettikleri, hangi kullanıcıdan sonra hangi kullanıcıların bir lokasyonu ziyaret ettikleri, kimlerin zamanda sıklıkla birlikte hareket ettikleri ve kimlerin hangi lokasyonları düzenli olarak ziyaret ettiklerinin tespiti için faydalıdır. Aynı zamanda, hangi lokasyonun önemli hale geldiği ve zamanla popülerliğini kaybettiği veya birkaç gün gibi kısa dönemli aralıklarda önemli hale geldiği zamansal boyutun yardımıyla keşfedilebilir.

Ancak, SMZÖL'lerin sosyal medya verisinden keşfedilmesi çeşitli sebeplerden dolayı zordur. Birincisi, mevcut değerlendirme ölçütleri SMZÖL'lerin keşfi için yetersizdir ve yeni değerlendirme ölçütleri geliştirilmelidir. İkincisi, mekân-zamansal sosyal medya veri kümeleri oldukça büyüktür ve boyutu ve ölçeği zamanla artmaktadır ve bu sebepten dolayı hesaplama olarak verimli yöntemler geliştirilmesi gerekmektedir. Üçüncüsü, sosyal medya kullanıcıları çok sayıda lokasyonu birkaç defa ziyaret etmektedir ve aday SMZÖL sayısı gerçek SMZÖL sayısından çok fazladır. Dördüncüsü, insan kaynaklı tutarsızlıklar algoritmalara ekstra yük getirmektedir.

Literatürde sosyal medya veri kümelerini mekânsal ve zamansal boyutta analiz eden ve çeşitli amaçlara ulaşmaya çalışan çok sayıda çalışma mevcuttur. Bu çalışmaların büyük bir kısmı, şehir hesaplaması ve şehir bilişimi, şehirlerin mekânsal yapısının tanımlanması ve şehir sakinlerinin şehir örüntülerinin çıkarılması, konularını çalışmaktadır [65] [52]. Sosyo-mekânsal sorgu işletilmesi sosyal medya madenciliğinin bir diğer önemli konularındandır. Bu alanda, araştırmacılar, büyük sosyal medya veri kümelerinde kullanılabilecek sorgular üzerine çalışmaktadır [99] [100]. Lokasyon tavsiye, mekânsal ve zamansal başlık keşfi, insan hareketlilik örüntü madenciliği ve mekânsal ve zamansal sosyal önemli lokasyonlar madenciliği mekânsal ve zamansal sosyal medya madenciliğinin diğer yaygın uygulama alanlarıdır.



Şekil 6.1. İki farklı kullanıcının SÖL'ları ve zamansal yaygınlıkları

Şekil 6.1'de iki örnek kullanıcının mekânsal sosyal önemli lokasyonları ve bu lokasyonların zamansal yaygınlıklarına ait grafikler verilmiştir. Şekilde, lokasyonlar L1, L2, L3, L4, L5, L6, L20, L22 ve L40 olarak etiketlenmiştir. Bir kullanıcının bir lokasyondaki zamansal yaygınlığı, lokasyona eklenen zaman serisi grafiğiyle verilmiştir. Grafiklerdeki kırmızı kesikli çizgiler eşik değerini göstermektedir. Eğer bir kullanıcının bir lokasyondaki zamansal yaygınlığı belirli bir eşik değeri sağlıyorsa, mekânsal sosyal önemli lokasyon, ilgili kullanıcı için SMZÖL olarak keşfedilir. Örneğin, Şekil 6.1 a)'daki kullanıcıya ait L40 lokasyonu, eşik değeri ile aynı değerlere sahiptir ve bu sebeple SMZÖL olarak keşfedilmiştir. Şekilde görüldüğü üzere, lokasyonlar L1, L2, L5, L22 ve L40 her iki sosyal medya kullanıcısı için ortak mekânsal sosyal önemli lokasyonlardır. Ancak, bu lokasyonların kullanıcılar için zamansal dağılımları farklıdır. Örneğin, L40 Şekil 6.1 a)'daki kullanıcı için zamansal yaygınlığı sağlamaktadır ve SMZÖL olarak keşfedilmiştir, fakat bu lokasyon Şekil 6.1 b)'deki kullanıcı için zamansal olarak yaygın değildir ve SMZÖL olarak keşfedilmemiştir. Lokasyon L1, her iki kullanıcı için de yüksek zamansal yaygınlığa sahiptir, ancak zaman dilimlerindeki yaygınlıklar farklıdır.

Bu çalışma, sosyal medya veri kümeleri üzerinde, sosyal, mekânsal ve zamansal boyutlara dayalı olarak sosyo-mekân-zamansal önemli lokasyonların keşfini amaçlamaktadır. Bu çalışma, bir önceki bölümden farklı olarak, sosyal ve mekânsal boyutlara ek olarak zamansal boyutu da dikkate almakta ve SMZÖL'leri tanımlamaktadır. Kullanıcı seviyesinde mekân-zamansal önemli lokasyonlar mekânsal ve zamansal boyutlar dikkate alınarak keşfedilmekte ve daha sonra, mekânsal sosyal önemli lokasyonlar sosyal gruptaki kullanıcı yaygınlığı ve sosyal medya kullanıcı grubunun zamansal dağılımı ile keşfedilmektedir. Mekânsal ve zamansal boyutların her ikisini de dikkate almakla, önerilen algoritmalar belirli zaman dilimlerinde sosyal medya kullanıcı grubu için önemli olan lokasyonları keşfedebilmektedir.

Bu bölümün devam eden kısımları şu şekilde organize edilmiştir. İkinci kısımda sosyo-mekân-zamansal önemli lokasyonların keşfi için temel tanımlar ve problem tanımlaması yapılmıştır. Üçüncü kısımda önerilen algoritmalar ve örnek algoritma çalışması verilmiştir. Dördüncü kısımda deneysel değerlendirme yapılmıştır. Beşinci kısımda, bu çalışmanın katkıları ve kapsamı verilmiştir.

6.2. Temel Tanımlar ve Problem Tanımlaması

Bu çalışma, sosyal medya veri kümeleri üzerinden SMZÖL'lerin keşfine odaklanmaktadır. Bu lokasyonları, kullanıcı-seviyesinde sosyo-mekân-zamansal önemli lokasyonlar ve grup-seviyesinde sosyo-mekân-zamansal önemli lokasyonlar olarak gruplandırabiliriz. Bu çalışmada önerilen yaklaşım, kullanıcı-seviyesinde SMZÖL'leri kullanarak grup-seviyesinde SMZÖL'lerin keşfini yapmaktadır. Bu nedenle, tanımlar, kullanıcı-seviyesinde mekân-zamansal önemli lokasyonlar keşfi ve grup-seviyesinde sosyo-mekân-zamansal önemli lokasyonlar keşfi olarak iki parça halinde verilmiştir.

Bu kısımda, ilk olarak sosyal medya veri kümeleri üzerinden kullanıcı-seviyesinde sosyal önemli lokasyonların keşfi anlatılmaktadır. Ardından, kullanıcı-seviyesindeki sosyo-mekânsal önemli lokasyonları kullanarak ve zamansal boyut eklenerek SMZÖL'lerin nasıl modellendiği anlatılmaktadır. Son olarak, sosyo-mekân-zamansal önemli lokasyonlar madenciliği problem tanımlaması sunulmuştur.

6.2.1. Temel Tanımlar

Tanım 1. Sosyal medya aktivite geçmişi verilmiş olan bir sosyal medya kullanıcısı u ve bir lokasyon l için, l 'nin u için **lokasyon yoğunluğu** değeri, u 'nun l 'deki oluşum sayısının u 'nun toplam oluşum sayısına oranıdır.

$$\text{lokasyon yoğunluğu}_l^u = \frac{u'nun\ l'deki\ oluşum\ sayısı}{u'nun\ toplam\ oluşum\ sayısı}$$

Lokasyon yoğunluğu, lokasyon l 'nin, kullanıcı u 'nun sosyal medya geçmişindeki oluşum sıklığının tespiti için kullanılmaktadır.

Tanım 2. Lokasyon yoğunluğu değeri verilmiş olan bir sosyal medya kullanıcısı u , bir lokasyon l ve minimum lokasyon yoğunluk eşik değeri $min_yoğunluk$ için, eğer l 'nin

lokasyon yoğunluk değeri $min_yoğunluk$ eşik değerine eşit veya büyükse, lokasyon l **yaygın lokasyon** olarak tanımlanır; $lokasyon\ yoğunluğu_l^u \geq min_yoğunluk$

Tanım 3. Sosyal medya aktivite geçmişisi verilmiş olan bir sosyal medya kullanıcısı u ve bir lokasyon l için, l 'nin u için **ziyaret ömrü** kullanıcı u 'nun lokasyon l 'yi ilk ve son ziyareti arasındaki zaman farkının kullanıcı u 'nun sosyal medya geçmişindeki ilk ve son oluşumunun zaman farkına oranıdır.

$$\begin{aligned} ziyaret\ ömrü_l^u &= \frac{u'nun\ l'yi\ ilk\ ve\ son\ ziyaretinin\ farkı}{u'nun\ sosyal\ medya\ ömrü} \\ &= \frac{Son_Ziyaret_{l,u} - İlk_Ziyaret_{l,u}}{Son_Oluşum_u - İlk_Oluşum_u} \end{aligned}$$

Ziyaret ömrü, lokasyon l 'nin kullanıcı u için zamanda yaygınlığını tespit etmek için kullanılmaktadır.

Tanım 4. Ziyaret ömrü verilmiş olan bir sosyal medya kullanıcısı u , bir lokasyon l ve minimum ziyaret ömrü eşik değeri $min_ziyaret$ için, eğer l 'nin ziyaret ömrü değeri $min_ziyaret$ eşik değerine eşit veya büyükse, lokasyon l **zamanda yayılmış lokasyon** olarak tanımlanır; $ziyaret\ ömrü_l^u \geq min_ziyaret$.

Tanım 5. Bir sosyal medya kullanıcısı u ve bir lokasyon l için, eğer lokasyon l , Tanım 2 ve 4'te tanımlandığı gibi yaygın ve zamanda yayılmış lokasyon ise, lokasyon l **Kullanıcı Mekânsal Önemli Lokasyon (KMÖL)** olarak tanımlanır.

6.2.2. SMZÖL'lerin Modellenmesi

Bu kısımda, kullanıcı-seviyesinde mekânsal sosyal önemli lokasyonlar genişletilerek kullanıcı-seviyesinde ve grup-seviyesinde SMZÖL'ler modellenmiştir.

Tanım 6. Bir sosyal medya kullanıcısı u , bir lokasyon l ve bir zaman penceresi tw için, u 'nun l 'de ve tw 'deki **Zaman Penceresi Oluşumu (ZPO)**, kullanıcı u 'nun, lokasyon l ve zaman penceresi tw 'deki oluşum sayısının u 'nun tw 'deki toplam oluşum sayısına oranıdır. Örneğin, tw 7 ise, l 'nin oluşum sayısı u 'nun sosyal medya verisindeki haftalarına göre hesaplanacaktır.

$$ZPO_l^u = \frac{u'nun\ tw\ zaman\ penceresinde\ l'deki\ oluşum\ sayısı}{u'nun\ tw\ zaman\ penceresinde\ toplam\ oluşum\ sayısı}$$

ZPO değerlendirme ölçütü lokasyon l 'nin tw zaman penceresinde kullanıcı u için önem değerinin belirlenmesinde kullanılmaktadır.

Tanım 7. Bir sosyal medya kullanıcısı u , bir lokasyon l ve bir zaman penceresi tw için, eğer l 'nin ZPO değeri minimum zaman penceresi oluşumu eşik değeri min_ZPO 'yu sağlıyorsa, l **Zaman Penceresinde Zamansal Önemli Lokasyon (ZP_ZÖL)** olarak tanımlanır; $ZPO_l^u \geq min_ZPO$.

Tanım 8. Bir sosyal medya kullanıcısı u , bir lokasyon l için, l 'nin **Kullanıcı Zamansal Lokasyon Önemi (KZLÖ)** l 'nin ZP_ZÖL olduğu zaman penceresi sayısının, kullanıcının toplam zaman penceresi sayısına oranıdır.

$$KZLÖ_l^u = \frac{l'nin\ ZP_ZÖL\ olduğu\ zaman\ penceresi\ sayısı}{u'nun\ toplam\ zaman\ penceresi\ sayısı}$$

KZLÖ lokasyon l 'nin, kullanıcı u 'nun sosyal medya geçmişindeki zamansal önem dağılımını ortaya çıkarmaktadır.

Tanım 9. Bir sosyal medya kullanıcısı u , bir lokasyon l için, eğer l 'nin KZLÖ değeri minimum pencere yaygınlığı eşik değeri $min_PencereYaygınlığı$ 'ni sağlıyorsa, l **Kullanıcı Zamansal Önemli Lokasyon (KZÖL)** olarak tanımlanır; $KZLÖ_l^u \geq min_PencereYaygınlığı$.

Tanım 10. Bir sosyal medya kullanıcısı u 'ya ait mekânsal önemli lokasyonlar (KMÖL) ve zamansal önemli lokasyonlar (KZÖL) listeleri için, eğer bir lokasyon l hem KMÖL hem de KZÖL ise, lokasyon l **Kullanıcı Mekân-Zamansal Önemli Lokasyon (KMZÖL)** olarak tanımlanır.

Kullanıcı-seviyesinde sosyo-mekân-zamansal önemli lokasyonlar Tanım 10'da tanımlanmıştır. Grup-seviyesinde sosyo-mekân-zamansal önemli lokasyonlar Tanımlar 11-14 arasında verilmiştir. Bu tanımlarda, kullanıcı-seviyesindeki tanımlar genişletilmiştir. Kullanıcı-seviyesinde zaman penceresinde zamansal önemli lokasyonlar tanımını (Tanım 6), grup-seviyesinde zaman penceresinde zamansal önemli lokasyonlar tanımını (Tanım 11) sağlayacak şekilde genişletilmiş ve her iki tanım için de tek eşik

değeri kullanılmıştır. Benzer şekilde, kullanıcı-seviyesinde zamansal lokasyon önemi tanımı (Tanım 8), grup-seviyesinde zamansal lokasyon önemi tanımını sağlayacak şekilde genişletilmiş (Tanım13) ve aynı eşik değeri kullanılmıştır.

Tanım 11. Bir lokasyon l ve zaman penceresi tw için, **Zaman Penceresinde Mekân-Zamansal Kullanıcı Yaygınlığı** (ZP_MZKY), tw 'de lokasyon l 'yi $ZP_ZÖL$ olarak gören kullanıcı sayısının, toplam kullanıcı sayısına oranıdır.

$$ZP_MZKY_l^{tw} = \frac{l'nin\ tw'de\ ZP_ZÖL\ olduğu\ kullanıcı\ sayısı}{toplam\ kullanıcı\ sayısı}$$

ZP_MZKY , tw 'de l 'nin sosyal medya kullanıcı grubu içerisinde hangi oranda zamansal önemli lokasyon olduğu bilgisini sağlamaktadır.

Tanım 12. Bir lokasyon l ve zaman penceresi tw için, eğer l 'nin ZP_MZKY değeri minimum zaman penceresi oluşumu eşik değeri min_ZPO 'yu sağlıyorsa, l Zaman Penceresinde Mekân-Zamansal Kullanıcı Yaygın Lokasyon (ZP_MZKYL) olarak tanımlanır; $ZP_MZKY_l^{tw} \geq min_ZPO$.

min_ZPO eşik değeri Tanım 7'de kullanıcı-seviyesinde zaman penceresinde zamansal önemli lokasyonların keşfinde kullanılmıştır. Aynı eşik değeri, Tanım 12'de grup-seviyesinde zaman penceresinde zamansal önemli lokasyonların keşfinde de kullanılmıştır.

Tanım 13. Bir lokasyon l için, l 'nin **Kullanıcı Grubu için Sosyo-Mekân-Zamansal Lokasyon Önemi** ($SMZLÖ$), lokasyon l 'nin ZP_MZKYL olduğu zaman penceresi sayısının toplam zaman penceresi sayısına oranıdır.

$$SMZLÖ_l = \frac{l'nin\ ZP_MZKYL\ olduğu\ zaman\ penceresi\ sayısı}{toplam\ zaman\ penceresi\ sayısı}$$

$SMZLÖ$, l 'nin sosyo-zamansal önemini hesaplayabilmek için l 'nin kullanıcı yaygın olduğu zaman penceresi oranını sağlamaktadır.

Tanım 14. Bir lokasyon l için, eğer $SMZLÖ$ değeri minimum pencere yaygınlığı eşik değeri $min_PencereYaygınlığı$ 'ni sağlıyorsa, lokasyon l **Sosyo-Mekân-Zamansal Önemli Lokasyon** ($SMZÖL$) olarak tanımlanır; $SMZLÖ_l \geq min_PencereYaygınlığı$.

$min_PencereYaygınlığı$ eşik değeri Tanım 9’da kullanıcı seviyesinde zamansal önemli lokasyonların keşfinde kullanılmıştır. Aynı eşik değeri, Tanım 14’de grup-seviyesinde zamansal önemli lokasyonların keşfinde de kullanılmıştır.

6.2.3. Problem Tanımlaması

Verilenler:

- Mekân-zamansal bilgi içeren sosyal medya kullanıcı veri kümesi D .
- Minimum lokasyon yoğunluğu eşik değeri, $min_yoğunluk$.
- Minimum ziyaret ömrü eşik değeri, $min_ziyaret$.
- Zaman penceresi parametresi, tw .
- Minimum zaman penceresi oluşum eşik değeri, min_ZPO .
- Minimum pencere yaygınlığı eşik değeri, $min_PencereYaygınlığı$.

Aranan: $\{L \mid L, \text{ Tanım 14’te tanımlandığı şekilde Sosyo-Mekân-Zamansal Önemli Lokasyonlar kümesi} \}$

Hedef: Hesaplama maliyetini minimize etmek.

Sabitler: Kriterleri sağlayan sosyo-mekân-zamansal önemli lokasyonları keşfetmek

6.3.Sosyo-Mekân-Zamansal Önemli Lokasyonların Keşfi

Bu kısımda sosyo-mekân-zamansal önemli lokasyonların keşfi için önerilen algoritmalar sunulmuştur. SMZÖL keşfi, kullanıcı-seviyesinde mekân-zamansal önemli lokasyonların keşfi ve grup-seviyesinde sosyo-zamansal önemli lokasyonların keşfi olmak üzere iki ana aşamadan oluşmaktadır. Bu kısımda önerilen algoritmalar sosyal önemli lokasyonlar keşfi için her iki aşamayı da kullanmaktadır. Algoritmaların değiştiği kısım, bu adımların işletilmesi noktasındadır.

Bu kısımda Toy, TF-SSTIL Miner ve SF-SSTIL Miner algoritmaları olmak üzere üç algoritma önerilmiştir. Bu kısmın devam eden bölümünde bu algoritmaların tasarımları ve örnek çalışması verilmiştir.

6.3.1. Toy Algoritma

Toy algoritma, kullanıcı-seviyesinde mekânsal önemli lokasyonlar keşfi ve zamansal önemli lokasyonlar keşfi aşamalarını ayrı ayrı işletmekte ve kullanıcı-seviyesinde mekân-

zamansal önemli lokasyonların keşfi için mekânsal ve zamansal önemli lokasyonları birleştirmektedir. Ardından, kullanıcı grubu için SMZÖL'lar, lokasyonların kullanıcı yaygınlığına bağlı olarak keşfedilmektedir. Toy algoritmada herhangi bir erken eleme işlemi bulunmamaktadır ve bu nedenle hesaplama maliyeti yüksektir. Toy algoritmaya ait kaba kod Algoritma 6.1'de verilmiştir.

Algoritma 6.1. Toy Algoritma-Sosyo-Mekân-Zamansal Önemli Lokasyon Madenciliği
Girdiler: <i>D</i>: Sosyal medya kullanıcı veri kümesi, <i>L</i>: Çıkarılmış ve etiketlenmiş lokasyonlar kümesi <i>min_yoğunluk</i>: Minimum lokasyon yoğunluğu eşik değeri <i>min_ziyaret</i>: Minimum ziyaret ömrü eşik değeri <i>tw</i>: Zaman penceresi parametresi <i>min_ZPO</i>: Minimum zaman penceresi oluşumu eşik değeri <i>min_PencereYaygınlığı</i>: Minimum pencere yaygınlığı eşik değeri Çıktı: Sosyo-Mekân-Zamansal Önemli Lokasyonlar kümesi, <i>SMZÖL</i> 1. <i>tumLokasyonlar</i> = null, <i>L</i> = null 2. for each kullanıcı kümesi <i>u</i> in <i>D</i> 3. $I_s = \text{KMÖL-kesfet}(L_u, \text{min_yoğunluk}, \text{min_ziyaret})$ 4. $I_t = \text{KZÖL-kesfet}(L_u, tw, \text{min_ZPO}, \text{min_PencereYaygınlığı})$ 5. $\text{tumLokasyonlar} \leftarrow I_s \cap I_t$ 6. end for 7. $\text{SMZÖL} = \text{SMZÖL-kesfet}(\text{tumLokasyonlar}, tw, \text{min_ZPO}, \text{min_PencereYaygınlığı})$ 8. return <i>SMZÖL</i>

Algoritma 6.1'de, 2-6 arası adımlar kullanıcı-seviyesinde mekânsal ve zamansal önemli lokasyonları keşfetmektedir. Adım 3'te *KMÖL-kesfet* fonksiyonu, kullanıcı *u* için mekânsal önemli lokasyonları, adım 4'te *KZÖL-kesfet* fonksiyonu, kullanıcı *u* için zamansal önemli lokasyonları keşfetmektedir. Mekânsal ve zamansal önemli lokasyonlar keşfi işlemlerine ait detaylı bilgi Fonksiyon 6.1 ve 6.2'de verilmiştir. Ardından, adım 5'te, keşfedilen lokasyonların kesişimi alınarak kullanıcı *u* için hem mekânsal hem de zamansal önemli lokasyonlar çıkarılmıştır. Son olarak adım 7'de, kullanıcı grubu için sosyo-mekân-zamansal önemli lokasyonlar *SMZÖL-kesfet* fonksiyonu ile keşfedilmektedir. *SMZÖL-kesfet* fonksiyonuna ait detaylı bilgi Fonksiyon 6.3'te verilmiştir.

Fonksiyon 6.1. KMÖL-kesfet Fonksiyonu**Girdiler:**

L_u : Kullanıcı u 'ya ait çıkarılmış ve etiketlenmiş lokasyonlar kümesi

$min_yoğunluk$: Minimum lokasyon yoğunluğu eşik değeri

$min_ziyaret$: Minimum ziyaret ömrü eşik değeri

Çıktı: Kullanıcı Mekânsal Önemli Lokasyonlar kümesi, $KMÖL$

1. $KMÖL = \text{null}$
2. **for each** lokasyon l **in** L_u
3. $lokasyon_yoğunluğu \leftarrow \text{lokasyon-yoğunluğu-hesapla}(l, u)$
4. **if** $lokasyon_yoğunluğu \geq min_yoğunluk$
5. $ziyaret_ömrü \leftarrow \text{ziyaret-ömrü-hesapla}(l, u)$
6. **if** $ziyaret_ömrü \geq min_ziyaret$
7. $KMÖL \leftarrow l$
8. **end if**
9. **end if**
10. **end for**
11. **return** $KMÖL$

Fonksiyon 6.1'de $KMÖL$ -kesfet fonksiyonu verilmiştir. $KMÖL$ -kesfet fonksiyonunda ilk olarak adım 3'te, kullanıcı u 'nun her bir lokasyonu l için lokasyon yoğunluğu değeri $lokasyon-yoğunluğu-hesapla$ fonksiyonuyla hesaplanmakta ve $min_yoğunluk$ eşik değerini sağlayan lokasyonların ziyaret ömrü değerleri adım 5'te hesaplanmaktadır. Eğer bir lokasyon hem $min_yoğunluk$ hem de $min_ziyaret$ eşik değerlerini sağlıyorsa, adım 7'de bu lokasyon $KMÖL$ olarak seçilmiştir. Bu fonksiyon, Bölüm 5'te önerilmiş olan SS-ILM algoritmasıdır.

Fonksiyon 6.2'de $KZÖL$ -kesfet fonksiyonu verilmiştir. $KZÖL$ -kesfet fonksiyonunda, her bir zaman penceresi tw 'de zamansal önemli lokasyon çıkarımı (adımlar 2-10) ve son zamansal önemli lokasyonlar keşfi (adımlar 11-16) olmak üzere iki alt işlem gerçekleştirilmektedir. Adım 4'te, kullanıcı u 'nun her bir zaman penceresi tw için lokasyonların oluşumları hesaplanmakta ve adım 5'te, hesaplanan değerin min_ZPO eşik değerini sağlayıp sağlamadığı kontrol edilmektedir. Adım 6'da, min_ZPO eşik değerini sağlayan lokasyonlar $ZP_ZÖL$ dizisine kaydedilmekte ve adım 9'da $ZP_ZÖL$ dizileri $ZP_ZÖL_Tum$ matrisine kaydedilmektedir. Adım 12'de önceki aşamada çıkarılmış olan lokasyonların zaman penceresi sayıları hesaplanmakta ve adım 14'te, zaman penceresi

yaygınlığı $min_PencereYaygınlığı$ eşik değerini sağlayan lokasyonlar KZÖL olarak seçilmektedir.

Fonksiyon 6.2. KZÖL-kesfet Fonksiyonu

Girdiler:

L_u : Kullanıcı u 'ya ait çıkarılmış ve etiketlenmiş lokasyonlar kümesi

tw : Zaman penceresi parametresi

min_ZPO : Minimum zaman penceresi oluşumu eşik değeri

$min_PencereYaygınlığı$: Minimum pencere yaygınlığı eşik değeri

Çıktı: Kullanıcı Zamansal Önemli Lokasyonlar kümesi, $KZÖL$

1. $KZÖL = \text{null}$, $ZP_ZÖL_Tum = \text{null}$
2. **for each** zaman penceresi tw **in** tüm zaman pencereleri
3. **for each** lokasyon l **in** L_u
4. $ZPO \leftarrow \text{pencere-oluşumu-hesapla}(l, tw)$
5. **if** $ZPO \geq min_ZPO$
6. $ZP_ZÖL \leftarrow l$
7. **end if**
8. **end for**
9. $ZP_ZÖL_Tum \leftarrow ZP_ZÖL$
10. **end for**
11. **for each** lokasyon l **in** $ZP_ZÖL_Tum$
12. $KZLÖ = \text{pencere-sayısı-hesapla}(l, ZP_ZÖL_Tum)$
13. **if** $KZLÖ \geq min_PencereYaygınlığı$
14. $KZÖL \leftarrow l$
15. **end if**
16. **end for**
17. **return** $KZÖL$

Fonksiyon 6.3'te $SMZÖL$ -kesfet fonksiyonu verilmiştir. $SMZÖL$ -kesfet fonksiyonunda, her bir zaman penceresi tw 'de sosyo-zamansal önemli lokasyon çıkarımı (adımlar 2-10) ve son sosyo-zamansal önemli lokasyonlar keşfi (adımlar 11-16) olmak üzere iki alt işlem gerçekleştirilmektedir. Adım 4'te, zaman pencerelerine bağlı olarak lokasyonların kullanıcı yaygınlık değerleri hesaplanmakta ve adım 5'te MZKY değerlerinin min_ZPO eşik değerini sağlayıp sağlamadığı kontrol edilmektedir. Eğer değerler, eşik değerini sağlıyorsa, lokasyonlar ZP_MZKYL dizisine kaydedilmektedir. Adım 9'da, ZP_MZKYL dizileri ZM_MZKYL_Tum matrisine kaydedilmektedir. Adım 12'de ZM_MZKYL_Tum matrisindeki lokasyonların pencere oluşumları hesaplanmaktadır. Adım 13'te, zaman penceresi sayısının, $min_PencereYaygınlığı$ eşik değerini sağlayıp sağlamadığı kontrol

edilmektedir. Eğer lokasyonlar eşik değerini sağlıyorsa, adım 14'te verildiği gibi, lokasyonlar SMZÖL olarak kaydedilmektedir.

Fonksiyon 6.3. SMZÖL-kesfet Fonksiyonu

Girdiler:

tumLokasyonlar: Tüm kullanıcılara ait mekân-zamansal önemli lokasyonlar kümesi

tw; Zaman penceresi parametresi

min_ZPO: Minimum zaman penceresi oluşumu eşik değeri

min_PencereYaygınlığı: Minimum pencere yaygınlığı eşik değeri

Çıktı: Sosyo-Mekân-Zamansal Önemli Lokasyonlar kümesi, *SMZÖL*

1. *SMZÖL* = **null**, *ZP_MZKYL_Tum* = **null**
2. **for each** zaman penceresi *tw* **in** tüm zaman pencereleri
3. **for each** lokasyon *l* **in** *tumLokasyonlar*
4. *MZKY* ← mekân-zamansal-kullanıcı-yaygınlığı-hesapla(*l*, *tw*)
5. **if** *MZKY* ≥ *min_ZPO*
6. *ZP_MZKYL* ← *l*
7. **end if**
8. **end for**
9. *ZP_MZKYL_Tum* ← *ZP_MZKYL*
10. **end for**
11. **for each** lokasyon *l* **in** *ZP_MZKYL_Tum*
12. *SMZLÖ* = kullanıcı-yaygın-pencere-sayısı-hesapla(*l*, *ZP_MZKYL_Tum*)
13. **if** *SMZLÖ* ≥ *min_PencereYaygınlığı*
14. *SMZÖL* ← *l*
15. **end if**
16. **end for**
17. **return** *SMZÖL*

6.3.2. TF-SSTIL Miner Algoritması

Toy yaklaşım mekânsal ve zamansal eleme işlemlerini ayrı ayrı gerçekleştirdiği için hesaplama maliyeti yüksektir. TF-SSTIL Miner (Temporal-First Socio-Spatio-Temporal Important Locations Miner – Zamansal-Öncelikli Sosyo-Mekân-Zamansal Önemli Lokasyonlar Madencisi) algoritmasında, ilk olarak zamansal önemli lokasyonlar keşfedilmekte ve ardından zamansal önemli lokasyonlar arasından mekânsal analiz yapılarak SMZÖL'ler keşfedilmektedir. TF-SSTIL Miner algoritması mekânsal analiz için aday önemli lokasyonlar sayısını azaltmakta ve mekânsal önemli lokasyonlar keşfi aşamasında sadece zamansal önemli lokasyonları dikkate almaktadır. Zamansal yaygın

olmayan lokasyonların erken elenmesi algoritmanın hesaplama karmaşıklığını azaltmaktadır. TF-SSTIL Miner algoritmasına ait kaba kod Algoritma 6.2’de verilmiştir.

Algoritma 6.2. TF-SSTIL Miner Algoritması

Girdiler:

D: Sosyal medya kullanıcı veri kümesi,

L: Çıkarılmış ve etiketlenmiş lokasyonlar kümesi

min_yoğunluk: Minimum lokasyon yoğunluğu eşik değeri

min_ziyaret: Minimum ziyaret ömrü eşik değeri

tw; Zaman penceresi parametresi

min_ZPO; Minimum zaman penceresi oluşumu eşik değeri

min_PencereYaygınlığı: Minimum pencere yaygınlığı eşik değeri

Çıktı: Sosyo-Mekân-Zamansal Önemli Lokasyonlar kümesi, *SMZÖL*

1. *tumLokasyonlar* = **null**, *L* = **null**
2. **for each** kullanıcı kümesi *u* **in** *D*
3. $l_t = \text{KZÖL-kesfet}(L_u, tw, min_ZPO, min_PencereYaygınlığı)$
4. $l_s = \text{KMÖL-kesfet}(l_t, min_yoğunluk, min_ziyaret)$
5. $tumLokasyonlar \leftarrow l_s$
6. **end for**
7. $SMZÖL = \text{SMZÖL-kesfet}(tumLokasyonlar, tw, min_ZPO, min_PencereYaygınlığı)$
8. **return** *SMZÖL*

Algoritma 6.2’de görüldüğü üzere, mekânsal, zamansal ve sosyal önemli lokasyonlar çıkarımı fonksiyonları toy algoritmayla aynıdır. Toy algoritmanın aksine, TF-SSTIL Miner algoritması, ilk olarak adım 3’te zamansal önemli lokasyonları (l_t) keşfetmekte ve mekânsal önemli lokasyonlar keşfi fonksiyonu *KMÖL-kesfet* fonksiyonuna girdi olarak vermektedir. Mekânsal önemli lokasyonların çıkarımında zamansal olarak önemli ama mekânsal olarak önemli olmayan lokasyonlar elenmekte ve son olarak kullanıcı-seviyesinde mekân-zamansal önemli lokasyonlar keşfedilmektedir. Adım 7’de, kullanıcı-seviyesindeki mekân-zamansal önemli lokasyonlar, grup-seviyesinde sosyo-mekân-zamansal önemli lokasyonlar keşfinde kullanılmaktadır.

6.3.3. SF-SSTIL Miner Algoritması

SF-SSTIL Miner (Spatial-First Socio-Spatio-Temporal Important Locations Miner – Mekânsal-Öncelikli Sosyo-Mekân-Zamansal Önemli Lokasyonlar Madencisi) algoritmasında, ilk olarak mekânsal önemli lokasyonlar keşfedilmekte ve ardından

mekânsal önemli lokasyonların zamansal analizi yapılmaktadır. SF-SSTIL Miner algoritması mekânsal eleme işlemini mümkün olduğu kadar erken yapmakta ve zamansal önemli lokasyonlar keşfi aşamasında mekânsal önemli lokasyonlar dikkate alınmaktadır. Kullanıcı verisetinin parçalara bölünmesi ve her bir parçada zamansal önemli lokasyonların hesaplanmasından dolayı, zamansal analizin hesaplama maliyeti daha fazla olduğu için, mekânsal eleme işleminin zamansal eleme işleminden daha erken çalıştırılması, SMZÖL'lerin keşfinin daha hızlı yapılabilmesini sağlamaktadır. SF-SSTIL Miner algoritmasına ait kaba kod Algoritma 6.3'te verilmiştir.

Algoritma 6.3. SF-SSTIL Miner Algoritması

Girdiler:

D: Sosyal medya kullanıcı veri kümesi,
L: Çıkarılmış ve etiketlenmiş lokasyonlar kümesi
min_yoğunluk: Minimum lokasyon yoğunluğu eşik değeri
min_ziyaret: Minimum ziyaret ömrü eşik değeri
tw; Zaman penceresi parametresi
min_ZPO; Minimum zaman penceresi oluşumu eşik değeri
min_PencereYaygınlığı: Minimum pencere yaygınlığı eşik değeri

Çıktı: Sosyo-Mekân-Zamansal Önemli Lokasyonlar kümesi, SMZÖL

1. *tumLokasyonlar* = **null**, *L* = **null**
2. **for each** kullanıcı kümesi *u* **in** *D*
3. *l_s* = KMÖL-kesfet(*L_u*, *min_yoğunluk*, *min_ziyaret*)
4. *l_t* = KZÖL-kesfet(*l_s*, *tw*, *min_ZPO*, *min_PencereYaygınlığı*)
5. *tumLokasyonlar* $\leftarrow$ *l_t*
6. **end for**
7. SMZÖL = SMZÖL-kesfet(*tumLokasyonlar*, *tw*, *min_ZPO*, *min_PencereYaygınlığı*)
8. **return** SMZÖL

Algoritma 3'te görüldüğü üzere, temel fonksiyonlar toy algoritma ve TF-SSTIL Miner algoritmasıyla aynıdır. SF-SSTIL Miner algoritmasındaki tek değişiklik mekânsal analizin ilk olarak adım 3'te çalıştırılması ve adım 4'te mekânsal önemli lokasyonlar (*l_s*) üzerinden zamansal analizler yapılmasıdır. Son olarak adım 7'de, sosyo-mekân-zamansal önemli lokasyonlar, kullanıcı-seviyesindeki önemli lokasyonlardan keşfedilmektedir.

6.3.4. Algoritmanın Örnek Çalışması

Tablo 6.1. Örnek çalışma veri kümesi

ZP	Kullanıcı 1	Kullanıcı 2	Kullanıcı 3	Kullanıcı 4	Kullanıcı 5
1	1, 2, 4, 1, 3, 5, 7, 2, 20, 22, 2, 1, 2, 25, 1, 2, 5, 2, 1, 1	5, 6, 9, 2, 20, 2, 20, 25, 2, 2, 8, 6, 2, 10, 6, 2, 22, 6, 6, 5	20, 25, 6, 1, 2, 1, 3, 6, 9, 1, 5, 9, 6, 28, 1, 26, 6, 1, 1, 6	15, 45, 15, 1, 2, 2, 6, 15, 9, 7, 22, 15, 26, 15, 2, 2, 15, 1, 2, 2	20, 6, 4, 1, 9, 1, 1, 2, 4, 20, 6, 20, 20, 20, 20, 20, 8, 1, 1, 1
2	1, 40, 42, 5, 3, 3, 8, 20, 3, 22, 20, 1, 3, 3, 4, 1, 5, 1, 4, 1	1, 2, 6, 40, 40, 1, 1, 8, 1, 20, 22, 40, 1, 10, 40, 9, 40, 6, 40, 8	22, 40, 3, 40, 20, 25, 26, 3, 3, 5, 40, 9, 3, 3, 3, 40, 26, 25, 40	22, 2, 40, 6, 6, 22, 6, 2, 22, 6, 14, 6, 40, 22, 2, 22, 1, 6, 22	1, 1, 22, 1, 22, 6, 1, 4, 6, 22, 1, 22, 6, 8, 6, 56, 22, 1, 6, 6
3	5, 4, 8, 2, 4, 56, 5, 4, 20, 22, 24, 5, 5, 8, 40, 4, 4, 5, 5, 22	1, 20, 5, 60, 5, 10, 8, 40, 5, 1, 8, 5, 1, 5, 22, 6, 1, 10, 20, 1 22	1, 3, 6, 3, 25, 28, 15, 3, 15, 98, 15, 15, 40, 15, 9, 1, 15, 3, 20, 3	2, 20, 20, 15, 20, 15, 10, 2, 20, 15, 20, 1, 22, 5, 15, 40, 15, 2, 22, 20	1, 5, 26, 1, 4, 5, 20, 5, 8, 1, 20, 5, 1, 5, 8, 5, 1, 6, 1
4	1, 40, 9, 5, 40, 24, 4, 5, 1, 1, 5, 1, 4, 5, 2, 5, 1, 22, 4, 5	22, 20, 6, 6, 10, 22, 6, 8, 6, 1, 45, 1, 22, 20, 22, 10, 22, 6	28, 15, 6, 20, 15, 3, 15, 6, 3, 6, 9, 6, 3, 15, 5, 90, 6, 5, 15, 26	2, 60, 22, 2, 22, 40, 2, 22, 22, 5, 1, 2, 14, 2, 10, 22, 22, 40, 1, 2	6, 20, 22, 8, 9, 20, 22, 20, 6, 20, 5, 22, 20, 1, 22, 1, 22, 20, 22
5	1, 4, 40, 4, 4, 1, 22, 1, 4, 10, 4, 1, 1, 3, 40, 4, 40, 40, 1, 40	1, 40, 20, 22, 1, 3, 5, 9, 1, 22, 40, 22, 62, 46, 22, 1, 20, 40, 22, 1	1, 25, 27, 40, 1, 1, 5, 40, 1, 26, 20, 27, 40, 3, 1, 5, 1, 40, 20, 40	22, 20, 22, 2, 2, 30, 22, 25, 10, 2, 40, 2, 15, 22, 2, 22, 2, 20, 22	60, 2, 1, 2, 5, 22, 2, 20, 6, 1, 2, 22, 6, 20, 2, 22, 22, 60, 2, 22
6	4, 4, 1, 3, 50, 40, 1, 4, 2, 1, 5, 20, 3, 4, 2, 2, 2, 2, 1, 4	2, 6, 46, 22, 2, 6, 6, 40, 20, 22, 2, 2, 3, 6, 20, 22, 6, 2, 20	90, 1, 15, 1, 15, 3, 2, 1, 15, 20, 1, 27, 26, 26, 15, 15, 1, 5, 2	15, 50, 15, 6, 40, 1, 3, 6, 15, 6, 6, 15, 1, 15, 15, 20, 20, 6, 15, 6	1, 2, 6, 1, 3, 2, 20, 2, 22, 1, 2, 1, 6, 1, 1, 2, 2, 60, 25, 1
7	3, 1, 3, 40, 3, 5, 22, 20, 1, 3, 3, 2, 40, 1, 2, 2, 22, 3, 2, 2	2, 40, 5, 3, 2, 5, 2, 2, 94, 2, 5, 40, 1, 5, 20, 40, 3, 1, 5	15, 15, 20, 3, 1, 3, 25, 26, 15, 3, 26, 15, 3, 3, 3, 2, 5, 1, 15	20, 15, 15, 3, 40, 15, 15, 26, 20, 20, 1, 2, 3, 15, 15, 1, 20, 40, 20, 3	20, 2, 20, 2, 3, 15, 20, 2, 20, 2, 69, 15, 20, 2, 4, 15, 5, 2, 1, 20
8	2, 1, 5, 5, 3, 3, 3, 5, 22, 3, 40, 3, 1, 3, 5, 5, 1, 2, 2, 53	5, 3, 4, 10, 11, 2, 3, 20, 3, 22, 3, 95, 40, 20, 2, 2, 22, 3, 2, 2	5, 22, 20, 25, 5, 25, 26, 5, 5, 25, 40, 5, 11, 25, 26, 25, 1, 5, 25	6, 2, 40, 2, 6, 6, 1, 3, 8, 9, 2, 6, 40, 2, 1, 6, 2, 20, 40, 2	1, 2, 15, 2, 2, 6, 6, 5, 1, 2, 20, 6, 10, 2, 6, 20, 1, 6, 1, 6
9	1, 40, 6, 40, 9, 40, 25, 1, 1, 3, 5, 40, 5, 40, 25, 1, 22, 5, 6, 1	98, 5, 1, 3, 5, 1, 20, 1, 1, 5, 5, 6, 22, 1, 5, 40, 1, 5, 40	5, 72, 40, 22, 5, 3, 5, 2, 3, 40, 26, 90, 40, 5, 2, 40, 3, 5, 5, 40	3, 6, 20, 20, 15, 1, 6, 20, 1, 40, 22, 20, 20, 20, 10, 20, 20	1, 26, 1, 5, 1, 5, 3, 5, 20, 5, 26, 15, 10, 68, 1, 5, 1, 1, 5
10	40, 40, 40, 6, 6, 3, 1, 6, 52, 50, 90, 98, 1, 3, 6, 6, 40, 22, 40, 3	6, 22, 6, 51, 1, 1, 3, 20, 22, 6, 22, 1, 3, 22, 6, 1, 6, 22, 22, 6	5, 40, 26, 40, 26, 3, 5, 40, 90, 5, 1, 40, 9, 5, 3, 5, 26, 40, 1, 5	20, 10, 20, 22, 20, 22, 22, 5, 22, 40, 20, 22, 20, 9, 20	2, 22, 22, 6, 6, 15, 6, 5, 22, 26, 5, 6, 2, 6, 7, 22, 7, 6, 22

Bu kısımda, önerilen algoritmalara ait örnek çalışma sunulmuştur. İlk olarak Tablo 6.1’de veri kümesi tanıtılmış, ardından kullanıcı-seviyesinde mekân-zamansal önemli lokasyonlar keşfedilmiş ve son olarak kullanıcı grubu için SMZÖL’ler keşfedilmiştir.

Tablo 6.1’de örnek çalışmaya ait veri kümesi bulunmaktadır. Tabloda görüldüğü üzere, veri kümesi 5 kullanıcı ve 10 zaman dilimi içermektedir. Her bir zaman dilimi 20 günden oluşmakta ve veri kümesi toplamda 200 günden oluşmaktadır. Veri kümesinde 50 farklı lokasyon bulunmaktadır. Bir zaman diliminde ziyaret edilen lokasyonlar, numaralarla ifade edilmiş ve virgülle ayrılmıştır. Tablodaki lokasyonların sıralaması kullanıcıların lokasyonları ziyaret sıralamasıyla aynıdır.

6.3.4.1.Mekânsal Analiz

Mekânsal analize ait detaylı bilgi bir önceki bölümde bulunmaktadır. Mekânsal analiz için iki kullanıcı tanımlı değişken bulunmaktadır. Örnek çalışma için *min_yoğunluk* ve *min_ziyaret* eşik değerleri, sırasıyla, 0.05 ve 0.2 olarak belirlenmiştir. Tanımlar 1-5 veri kümesinden KMÖL’ların nasıl çıkarıldığını göstermektedir. Lokasyon yoğunluk değerlerini hesaplamak için, her lokasyonun oluşumu sayılmıştır. Ardından, oluşum sayısı, kullanıcının toplam ziyaret sayısına bölünmüştür (Tanım 1). Örneğin, lokasyon L1’in Kullanıcı 1 için lokasyon yoğunluk değeri aşağıdaki gibi hesaplanacaktır:

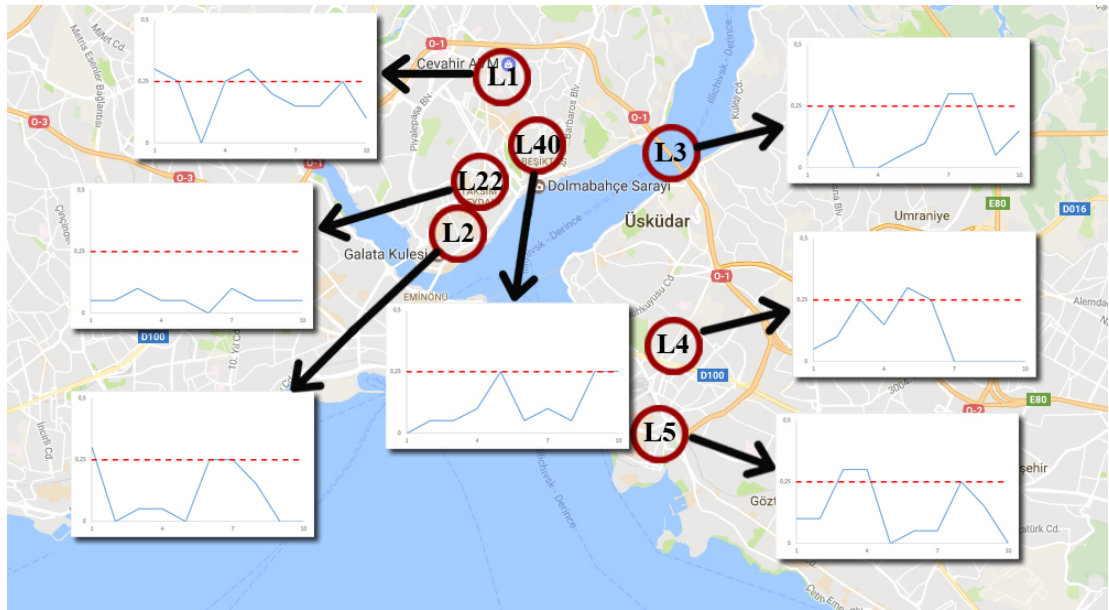
$$lokasyon\ yoğunluğu_{L1}^{Kullanıcı\ 1} = \frac{38}{200} = 0.19$$

Ziyaret ömrü değerini hesaplamak için, lokasyon L1’in Kullanıcı 1 için ilk ve son ziyaretleri arasındaki zaman farkı, Kullanıcı 1’in toplam sosyal medya aktivite gün sayısına bölünmektedir. Lokasyon L1’in Kullanıcı 1 için ziyaret ömrü değeri aşağıdaki gibi hesaplanacaktır:

$$ziyaret\ ömrü_{L1}^{Kullanıcı\ 1} = \frac{191}{200} = 0.955$$

Alan yetersizliğinden dolayı örnek çalışma veri kümesinde kullanıcıların ziyaretlerini günlük olarak veremedik, ancak hesaplama ve mantık kolaydır. Kullanıcı 1 lokasyon L1’i toplam 200 ziyareti arasından 38 defa ziyaret etmiştir. Aynı zamanda, Kullanıcı 1 lokasyon L1’i, toplam 200 günlük sosyal medya ziyaret geçmişi içerisinde 191 gün aralığında ziyaret etmiştir. *min_yoğunluk* ve *min_ziyaret* eşik değerleri uygulandıktan

sonra, 50 farklı lokasyondan 7 tanesi mekânsal olarak önemli lokasyon olarak keşfedilmiştir. Bu lokasyonlar L1, L2, L3, L4, L5, L22 ve L40 olarak etiketlenmiş ve Şekil 6.2’de verilmiştir.



Şekil 6.2. Kullanıcı 1'in mekânsal sosyal önemli lokasyonları ve zaman dilimlerindeki dağılımları

6.3.4.2. Zamansal Analiz

Zamansal analiz için kullanıcı tanımlı üç değişkene ihtiyaç vardır. Bu değişkenler zaman penceresi tw , min_ZPO ve $min_PencereYaygınlığı$ eşik değerleridir ve örnek çalışma için, sırasıyla, 20, 0.25 ve 0.3 olarak belirlenmiştir. Tanımlar 6-10 arasında zamansal önemli lokasyonların keşfi anlatılmaktadır. Lokasyonların, her bir zaman penceresinde eşik değerlerini sağlayıp sağlamadıkları değerlendirilmektedir. Lokasyonların zamansal önemli lokasyon sayılabilmeleri için min_ZPO eşik değerini en az 3 zaman penceresi için sağlamaları gerekmektedir. Şekil 6.2’de Kullanıcı 1 için mekânsal önemli lokasyonların zamansal dağılımları verilmiştir. Eğer kullanıcının zamansal dağılımı min_ZPO eşik değerini sağlıyorsa, mekânsal önemli lokasyon kullanıcı için mekân-zamansal sosyal önemli lokasyon olarak seçilmektedir. Şekil 6.2’de görüldüğü üzere, lokasyon L22, Kullanıcı 1’in hiçbir zaman diliminde min_ZPO eşik değerini sağlamamakta ve elenmektedir. Aynı işlem diğer kullanıcılar için de tekrar edilmiştir. Tablo 6.2’de KZLÖ ve KZÖL değerleri verilmiştir.

Tablo 6.2. Kullanıcılar için KZLÖ ve KZÖL’lar
(Mek. Eledi, lokasyonun mekânsal aşamada elendiğini göstermektedir)

Lokasyon	Kullanıcı 1	Kullanıcı 2	Kullanıcı 3	Kullanıcı 4	Kullanıcı 5
L1	5/10	4/10	3/10	0/10 (Eledi)	5/10
L2	3/10	4/10	Mek. Eledi	4/10	4/10
L3	3/10	1/10 (Eledi)	3/10	Mek. Eledi	Mek. Eledi
L4	3/10	Mek. Eledi	Mek. Eledi	Mek. Eledi	Mek. Eledi
L5	3/10	3/10	3/10	Mek. Eledi	2/10 (Eledi)
L6	Mek. Eledi	4/10	2/10 (Eledi)	3/10	3/10
L15	Mek. Eledi	Mek. Eledi	4/10	4/10	Mek. Eledi
L20	Mek. Eledi	0/10 (Eledi)	Mek. Eledi	4/10	3/10
L22	0/10 (Eledi)	3/10	Mek. Eledi	4/10	4/10
L25	Mek. Eledi	Mek. Eledi	1/10 (Eledi)	Mek. Eledi	Mek. Eledi
L26	Mek. Eledi	Mek. Eledi	0/10 (Eledi)	Mek. Eledi	Mek. Eledi
L40	3/10	1/10 (Eledi)	4/10	0/10 (Eledi)	Mek. Eledi

6.3.4.3.Sosyal Analiz

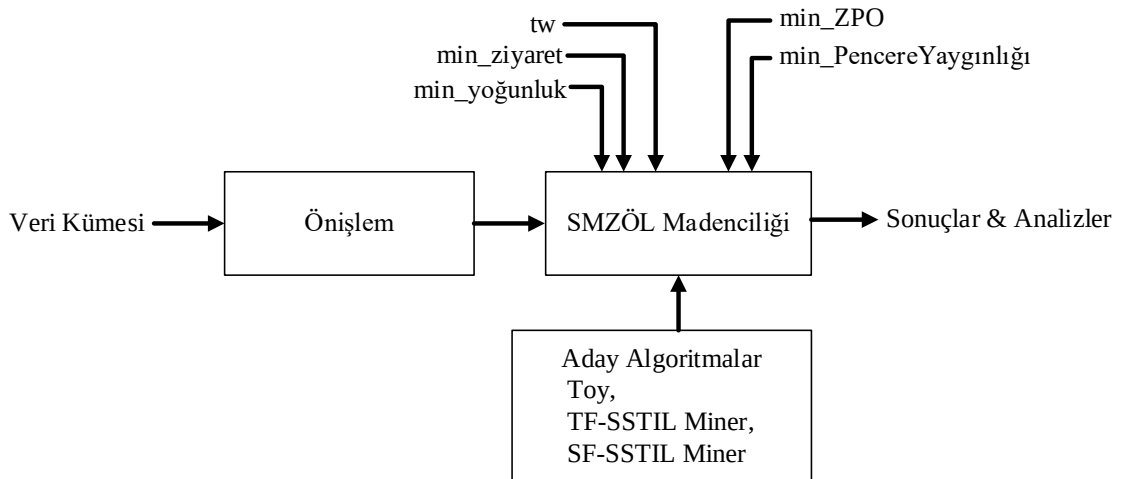
Sosyal-seviyede analiz için veri kümesindeki her bir zaman penceresi için lokasyonların kullanıcı yaygınlığı hesaplanmaktadır. Örnek çalışmada *min_ZPO* ve *min_PencereYaygınlığı* eşik değerleri, sırasıyla, 0.25 ve 0.3 olarak belirlenmişti ve sosyal analizde de aynı eşik değerleri kullanılmıştır. Lokasyonların SMZÖL olarak keşfedilebilmesi için en az 3 zaman diliminde en az 2 kullanıcı tarafından önemli olarak bulunmuş olmaları gerekmektedir. Tablo 6.3’te, her bir zaman penceresi için ZP_MZKY değerleri ve SMZLÖ değerleri ve sosyal grup için SMZÖL’ler verilmiştir. Tabloda görüldüğü üzere, lokasyonlar L1 ve L6, pek çok zaman penceresinde kullanıcı yaygındır ve kullanıcı grubu için SMZÖL olarak keşfedilmiştir. Ancak, lokasyonlar L4 ve L20 yeterli sayıda zaman penceresinde kullanıcı yaygın olmadığı için SMZÖL olarak keşfedilmemiştir.

Tablo 6.3. Zaman pencereleri için ZP_MZKY değerleri, SMZLÖ değerleri ve SMZÖL'lar

Lok.	ZP_MZKY										SMZLÖ	SMZÖL Mu?
	1	2	3	4	5	6	7	8	9	10		
L1	3/5	3/5	2/5	1/5	3/5	2/5	0/5	0/5	3/5	0/5	6/10	Evet
L2	3/5	0/5	0/5	1/5	2/5	3/5	3/5	3/5	0/5	0/5	5/10	Evet
L3	0/5	2/5	1/5	0/5	0/5	0/5	2/5	2/5	0/5	0/5	3/10	Evet
L4	0/5	0/5	1/5	0/5	1/5	1/5	0/5	0/5	0/5	0/5	0/10	Hayır
L5	0/5	0/5	3/5	1/5	0/5	0/5	1/5	2/5	3/5	1/5	3/10	Evet
L6	2/5	2/5	0/5	2/5	0/5	2/5	0/5	2/5	0/5	3/5	6/10	Evet
L15	1/5	0/5	2/5	1/5	0/5	2/5	2/5	0/5	0/5	0/5	3/10	Evet
L20	1/5	0/5	1/5	1/5	0/5	0/5	2/5	0/5	1/5	1/5	1/10	Hayır
L22	0/5	2/5	0/5	3/5	3/5	0/5	0/5	0/5	0/5	3/5	4/10	Evet
L40	0/5	2/5	0/5	0/5	2/5	0/5	0/5	0/5	2/5	2/5	4/10	Evet

6.4.Deneysel Çalışmalar

Bu kısımda, önerilen algoritmaların deneysel değerlendirmeleri yapılmıştır. Sosyal medya veri kümesi olarak gerçek Twitter veri kümesi kullanılmıştır. Veri kümesi ve veri kümesi hazırlama adımları Bölüm 6.4.1 ve 6.4.2’de verilmiştir. Deneysel sonuçlar Bölüm 6.4.3’te verilmiştir. Deneylerde, önerilen algoritmaların performans karşılaştırmaları yapılmış, İstanbul sosyal medya kullanıcıları için SMZÖL’lar keşfedilmiş ve son olarak literatürden bir çalışmayla karşılaştırılmıştır. Şekil 6.3’te çalışmaya ait deney düzeneği verilmiştir.



Şekil 6.3. Deney düzeneği

Performans ile ilgili deneyler 10'ar defa kořturulmuř ve iřlem zamanlarının ortalaması alınmıřtır. Ortalama deęerler ± 1 saniye hassasiyetindedir. Deneyler, ařaęıdaki soruları cevaplamak iin yapılmıřtır:

- Kullanıcı sayısının etkisi nedir?
- Minimum lokasyon yoğunluk eřik deęerinin etkisi nedir?
- Minimum ziyaret ömrü eřik deęerinin etkisi nedir?
- Minimum zaman penceresi oluřumu eřik deęerinin etkisi nedir?
- Minimum pencere yaygınlıęı eřik deęerinin etkisi nedir?
- İstanbul sakini sosyal medya kullanıcıları iin hangi lokasyonlar sosyo-mekân-zamansal olarak önemlidir?
- Önerilen algoritmalar literatür alıřmalarına göre ne kadar başarılıdır?

Deneyler 3.40 GHz hızdaki Intel Core i7 iřlemcili ve 8 GB RAM'li bir bilgisayarda gerekleřtirilmiřtir.

6.4.1. Veri Kümesi

Sosyal medya aęları, Twitter da dâhil olmak üzere, geliřtiriciler iin Uygulama Programlama Arayüzleri (Application Programming Interfaces – APIs) saęlamaktadır. Bu alıřmada sosyal medya aktivite veri kümesi olarak coęrafi bilgiye sahip olan Twitter veri kümesi seilmiřtir. Twitter'dan veri toplamak iin, REST API ve Streaming API kullanılmıřtır [18]. Buna ek olarak Twitter4j açık kaynak kodlu Java kütüphanesi sorguların iřletilmesi ve Twitter API'lerinden sonuçların toplanması iin kullanılmıřtır [19]. Streaming API akıř halinde olan tweetlerden coęrafi arama iřlemi saęlamaktadır. İlk olarak İstanbul, Türkiye tabanlı coęrafi arama iřlemi gerekleřtirilmiř ve ardından kullanıcılar toplanmıřtır. Yaklařık olarak 2500 kullanıcı bu adımda toplanmıřtır. Ardından REST API kullanılarak bu kullanıcılara ait tüm tweetler toplanmıřtır. Her bir tweet iin üç parametre toplanmıřtır; tarih/saat, enlem ve boylam. Veri kümesi Eylül 2008 ile Aralık 2015 tarihleri aralıęındadır. Veri kümesi bir önceki bölümde kullanılan veri kümesinin aynısıdır.

6.4.2. Önışlem

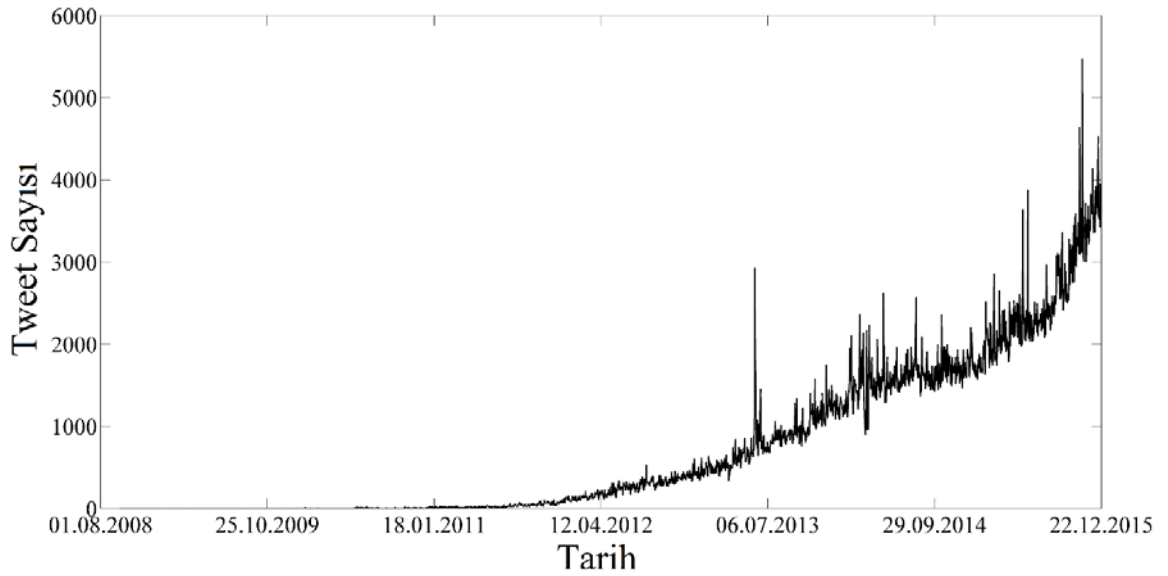
Bu bölümde, veri temizleme, kullanıcı seçimi, zamanda aşırı ağırlıklandırma önleme ve lokasyon etiketleme işlemleri açıklanmıştır.

6.4.2.1. Veri Temizleme

Deneylerde, aktif Twitter kullanıcılarının verisi kullanılmıştır. Bu çalışmada, aktif Twitter kullanıcıları, 2008 ile 2015 arasında 50'den fazla tweet gönderen kullanıcılar olarak tanımlanmıştır. Eğer tweet sayısı az ise, bu kullanıcı ya pasif ya da yeni kullanıcıdır. Ancak, aktif kullanıcı tanımı kullanıcı tercihleri veya uygulamalara göre değişkenlik gösterebilir. Veri kümesindeki kullanıcıların bir bölümü de sahte (spam) kullanıcılardan oluşmaktadır. Sahte kullanıcılardan korunmak için iki kriter kullanılmıştır ve bunlar takipçi sayısı ve takipçi/arkadaş oranıdır. Eğer bir kullanıcının takipçi sayısı 10'un ve takipçi/arkadaş oranı 0.1'in altında ise bu kullanıcı sahte kullanıcı olarak tespit edilmiş ve veri kümesi içerisine alınmamıştır. Bu parametreler için seçilen değerler pek çok sahte kullanıcı tespiti literatürüne göre seçilmiştir ve detaylı bilgiler [97] ve [98]'te bulunabilir.

6.4.2.2. Kullanıcı Seçimi

Bu çalışmada asıl odaklanılan konu İstanbul, Türkiye içerisinde tweet gönderen Twitter kullanıcılarının (İstanbul sakinlerinin) sosyo-mekân-zamansal önemli lokasyonlarını keşfetmektir. Kullanıcılar Streaming API aracılığıyla toplandığı için ve bu API geçmişe dayalı verileri döndürmediği için, veri toplama aşamasında kullanıcıların İstanbul içerisinde yeterli sayıda tweete sahip olduğu garanti edilememiştir. Bir kullanıcının İstanbul sakini olarak belirlenebilmesi için, kullanıcının tweetlerinin enlem ve boylamları İstanbul merkez enlem ve boylamından çıkarılmıştır. Eğer sonuçta çıkan fark kullanıcı tweetlerinin yarısından fazlası için anlamlı ise (örneğin İstanbul için en fazla 40 km ise) bu kullanıcı İstanbul sakini olarak seçilmiştir.



Şekil 6.4. Veri kümesinin zamansal dağılımı

Bu işlem uygulanarak deneyler için 1000 kullanıcı gerçek ve aktif İstanbul Twitter kullanıcıları olarak belirlenmiştir. Veri kümesinde 1.886.065 tweet bulunmaktadır ve veri kümesi Eylül 2008 ile Aralık 2015 tarihleri arasındadır. Veri kümesinin zamansal dağılımı Şekil 6.4'te verilmiştir. Şekilde görüldüğü üzere, tweetlerin çoğunluğu 2013-2015 aralığındadır. Bu durumun asıl sebebi, Türkiye sosyal medya kullanıcıları için Twitter'ın 2012 yılından sonra popülerlik kazanmasıdır.

6.4.2.3. Zamanda Aşırı Ağırlıklandırılmanın Önlenmesi

Veri kümesini incelediğimiz zaman, kullanıcıların bir lokasyonda aynı zamanda birden fazla tweet gönderdikleri (sosyal medya aktivitesi gerçekleştirdiklerini) görülmüştür. Bu davranış sürekli hale geldiği zaman, bir lokasyon doğru bulunma sayısından fazla görülmektedir, çünkü kullanıcı lokasyona bir defa gitmiş ancak birden fazla defa tweet göndermiştir. Bu problem bir lokasyonun zamanda aşırı ağırlıklandırılması olarak tanımlanmıştır. Bu problem bazı durumlarda istenmeden, örneğin, bir kullanıcı arkadaşlarıyla sohbet ederken, gerçekleşebilmektedir. Bir lokasyonun zamanda aşırı ağırlıklandırılmasını önlemek için, 60 dakikalık bir eşik değeri tanımlanmıştır. Eğer bir kullanıcı 60 dakika içerisinde aynı lokasyonda birden fazla defa tweet paylaşırsa, bu lokasyon bilgisinin yeni olmadığı varsayılmış ve lokasyonda oluşum bir defa olarak sayılmıştır. Bu yaklaşım yardımıyla bir lokasyonun zamanda aşırı ağırlıklandırılması önlenmiş olmaktadır. Bunun yanında, zamanda aşırı ağırlıklandırılmayı önlemek için farklı yaklaşımlar/kriterler de geliştirilebilir.

6.4.2.4.Lokasyon Etiketleme

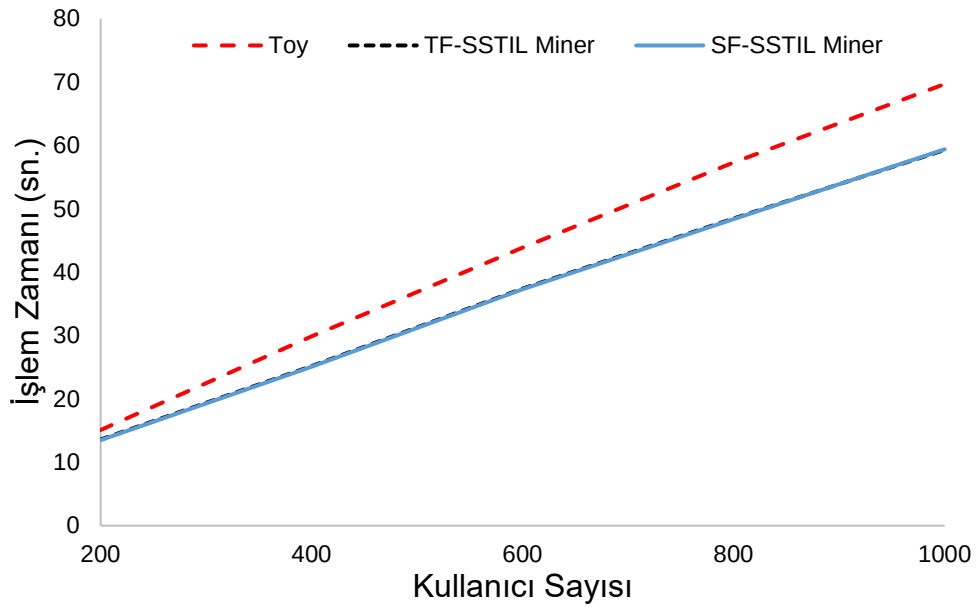
Twitter API'leri kullanıcı tweetlerine ait hassas enlem-boylam ikilileri sağlamaktadır. Bu yaklaşım hassas sonuçlar elde etmek için faydalı olsa da, lokasyon etiketleme işlemi için bir problemdir. Örneğin, bir alışveriş merkezi veya stadyum 1 km² alan içerisinde bulunmaktadır ancak hassas enlem-boylam ikilileri eşleşmediği için çok sayıda farklı lokasyon tanımlanabilmektedir. Bu problemin üstesinden gelebilmek için farklı enlem-boylam ikililerinin aynı lokasyon olarak sayılabilmesi için bir eşik değeri tanımlanmıştır. Daha önce [52] ve [47]'de yapılan çalışmalarda kullanıldığı gibi, bu eşik değerini 100 metre olarak tanımlanmıştır. Eğer iki lokasyon birbirine 100 metreden daha yakınsa, bu iki lokasyon aynı etiketle etiketlenmektedir.

6.4.3. Deneysel Sonuçlar

Bu kısımda, önerilen üç algoritmanın performansları değerlendirilmiş, İstanbul sosyal medya kullanıcılarının SMZÖL'eri keşfedilmiş ve önerilen SMZÖL madenciliği algoritmalarının sonuçları literatürdeki çalışmalarla karşılaştırılmıştır.

6.4.3.1.Kullanıcı Sayısının Etkisi

Bu deneyde, kullanıcı sayısının SMZÖL madenciliği algoritmalarının işlem zamanlarına etkisi incelenmiştir. *min_yoğunluk*, *min_ziyaret*, *tw*, *min_ZPO*, *min_PencereYaygınlığı* değerleri, sırasıyla, 0.01, 0.05, 30, 0.01 ve 0.01 olarak belirlenmiştir. Kullanıcı sayısı, 200'den 1000'e kadar 200'er artırılarak deneyler yapılmıştır. Sonuçlar Şekil 6.5'te verilmiştir.

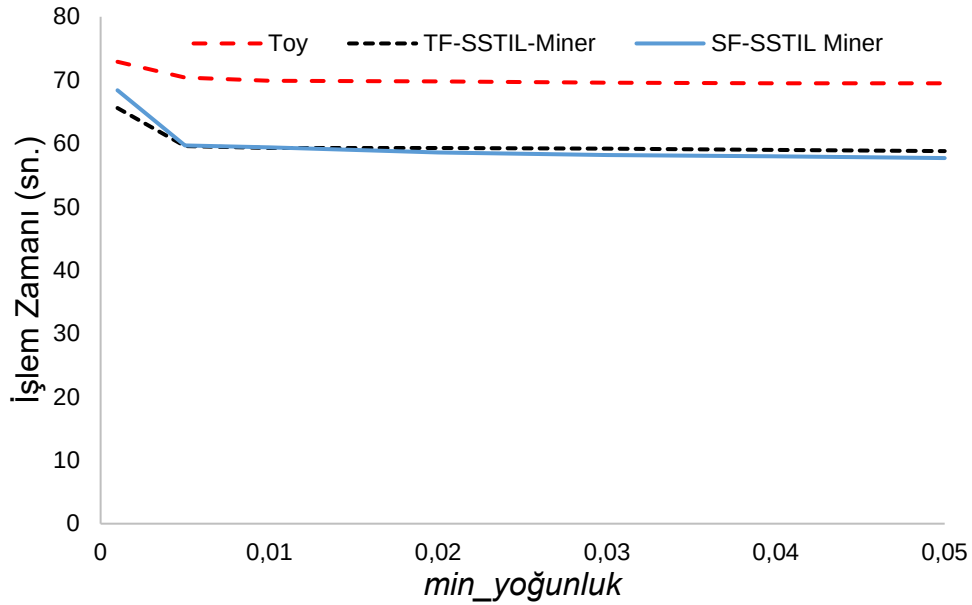


Şekil 6.5. Kullanıcı sayısının etkisi

Şekilde görüldüğü üzere, tüm algoritmaların işlem zamanları, kullanıcı sayısının artışıyla artmaktadır. Ayrıca, yukarı yönlü bir eğilimle algoritmaların kullanıcı sayısının artışı başarılı bir şekilde ele alabildikleri görülmektedir. SF-SSTIL Miner ve TF-SSTIL Miner algoritmaları erken eleme stratejisiyle Toy algoritmadan daha üstündür.

6.4.3.2. Minimum Lokasyon Yoğunluk Eşik Değerinin Etkisi

Bu deneyde, minimum lokasyon yoğunluk eşik değerinin (*min_yoğunluk*) SMZÖL madenciliği algoritmalarının işlem zamanlarına etkisi incelenmiştir. *min_ziyaret*, *tw*, *min_ZPO*, *min_PencereYaygınlığı* ve kullanıcı sayısı değerleri, sırasıyla, 0.05, 30, 0.01, 0.01 ve 1000 olarak belirlenmiştir. *min_yoğunluk* eşik değeri 0.001, 0.005, 0.01, 0.02, 0.03, 0.04 ve 0.05 değerleriyle deneyler yapılmıştır. Sonuçlar Şekil 6.6'da verilmiştir.



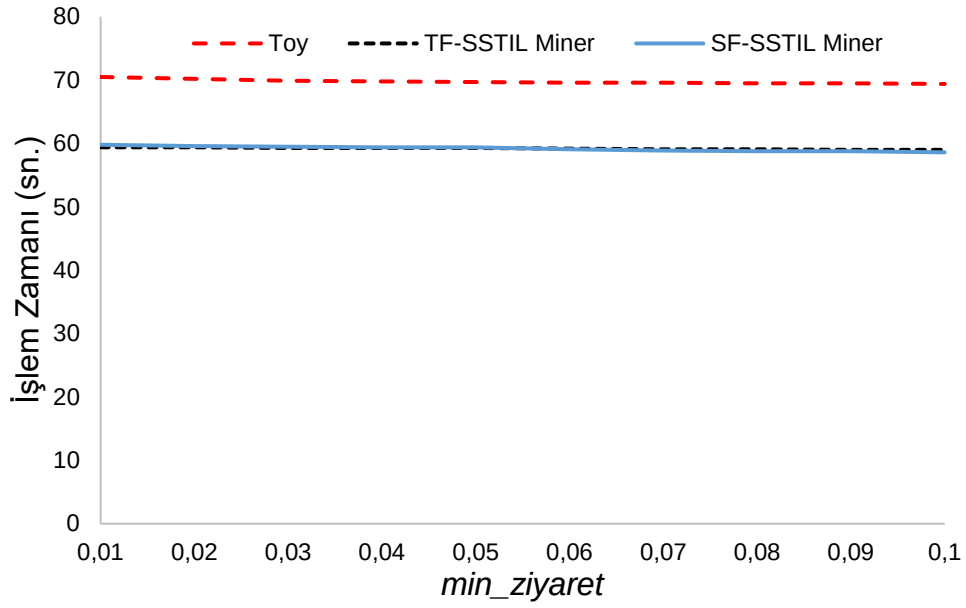
Şekil 6.6. Minimum lokasyon yoğunluğu eşik değerinin etkisi

Şekilde görüldüğü üzere, $min_yoğunluk$ eşik değerinin artışıyla birlikte işlem zamanı düşmektedir. Başlangıçta algoritmalar sert bir düşüş gösterse de $min_yoğunluk$ eşik değerinin artışıyla birlikte düşüş zayıflamaktadır. SF-SSTIL Miner algoritması diğer algoritmalarla göre işlem zamanının azaltmakta daha üstündür, çünkü SF-SSTIL Miner algoritması mekânsal eleme işlemini algoritmanın ilk adımı olarak işletmektedir.

6.4.3.3. Minimum Ziyaret Ömrü Eşik Değerinin Etkisi

Bu deneyde, minimum ziyaret ömrü eşik değerinin ($min_ziyaret$) SMZÖL madenciliği algoritmalarının işlem zamanlarına etkisi incelenmiştir. $min_yoğunluk$, tw , min_ZPO , $min_PencereYaygınlığı$ ve kullanıcı sayısı değerleri, sırasıyla, 0.01, 30, 0.01, 0.01 ve 1000 olarak belirlenmiştir. $min_ziyaret$ eşik değeri 0.01'den 0.1'e kadar 0.01'er artırılarak deneyler yapılmıştır. Sonuçlar Şekil 6.7'de verilmiştir.

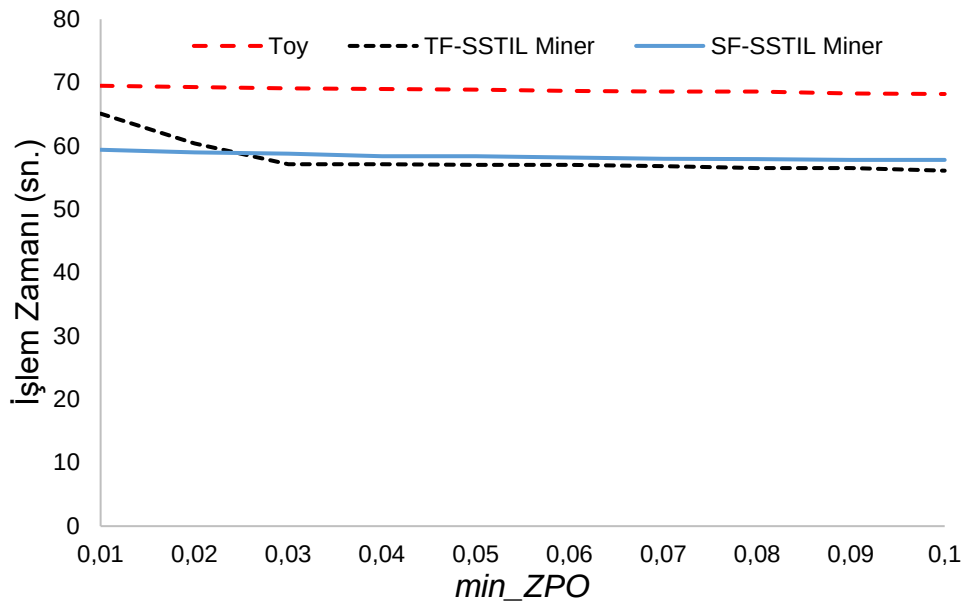
Şekilde görüldüğü üzere, algoritmalar $min_ziyaret$ eşik değerinin artışıyla işlem zamanlarını yavaşça düşürmektedir. Veri karakteristiğinden dolayı, $min_ziyaret$ eşik değeri fazla sayıda lokasyonu eleyememektedir.



Şekil 6.7. Minimum ziyaret ömrü eşik değerinin etkisi

6.4.3.4. Minimum Zaman Penceresi Oluşumu Eşik Değerinin Etkisi

Bu deneyde, minimum zaman penceresi oluşumu eşik değerinin (min_ZPO) SMZÖL madenciliği algoritmalarının işlem zamanlarına etkisi incelenmiştir. $min_yoğunluk$, $min_ziyaret$, tw , $min_PencereYaygınlığı$ ve kullanıcı sayısı değerleri, sırasıyla, 0,01, 0,05, 30, 0,01 ve 1000 olarak belirlenmiştir. min_ZPO eşik değeri 0,01’den 0,1’e kadar 0,01’er artırılarak deneyler yapılmıştır. Sonuçlar Şekil 6.8’de verilmiştir.

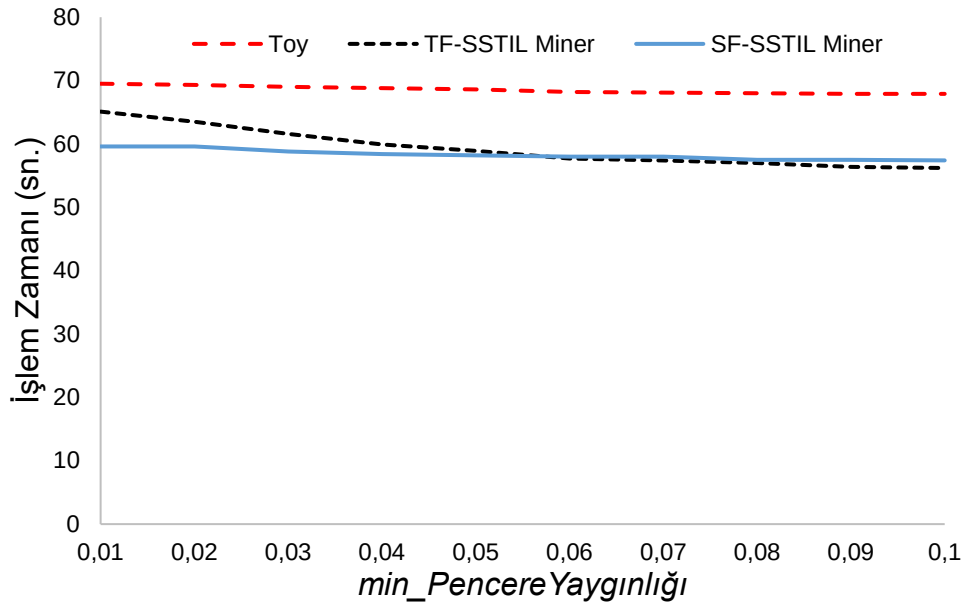


Şekil 6.8. Minimum zaman penceresi oluşumu eşik değerinin etkisi

Şekilde görüldüğü üzere, min_ZPO eşik değerinin artışı algoritmaların işlem zamanlarını düşürmektedir. Zamansal eleme işlemi ilk adım olarak uygulandığı için, TF-SSTIL Miner algoritmasındaki düşüş diğer algoritmalara göre daha fazladır. Bu nedenle TF-SSTIL Miner algoritması diğer algoritmalara göre min_ZPO eşik değerini ele almada daha hızlıdır.

6.4.3.5. Minimum Pencere Yaygınlığı Eşik Değerinin Etkisi

Bu deneyde, minimum pencere yaygınlığı eşik değerinin ($min_PencereYaygınlığı$) SMZÖL madenciliği algoritmalarının işlem zamanlarına etkisi incelenmiştir. $min_yoğunluk$, $min_ziyaret$, tw , min_ZPO ve kullanıcı sayısı değerleri, sırasıyla, 0.01, 0.05, 30, 0.01 ve 1000 olarak belirlenmiştir. $min_PencereYaygınlığı$ eşik değeri 0.01'den 0.1'e kadar 0.01'er artırılarak deneyler yapılmıştır. Sonuçlar Şekil 6.9'da verilmiştir.



Şekil 6.9. Minimum pencere yaygınlığı eşik değerinin etkisi

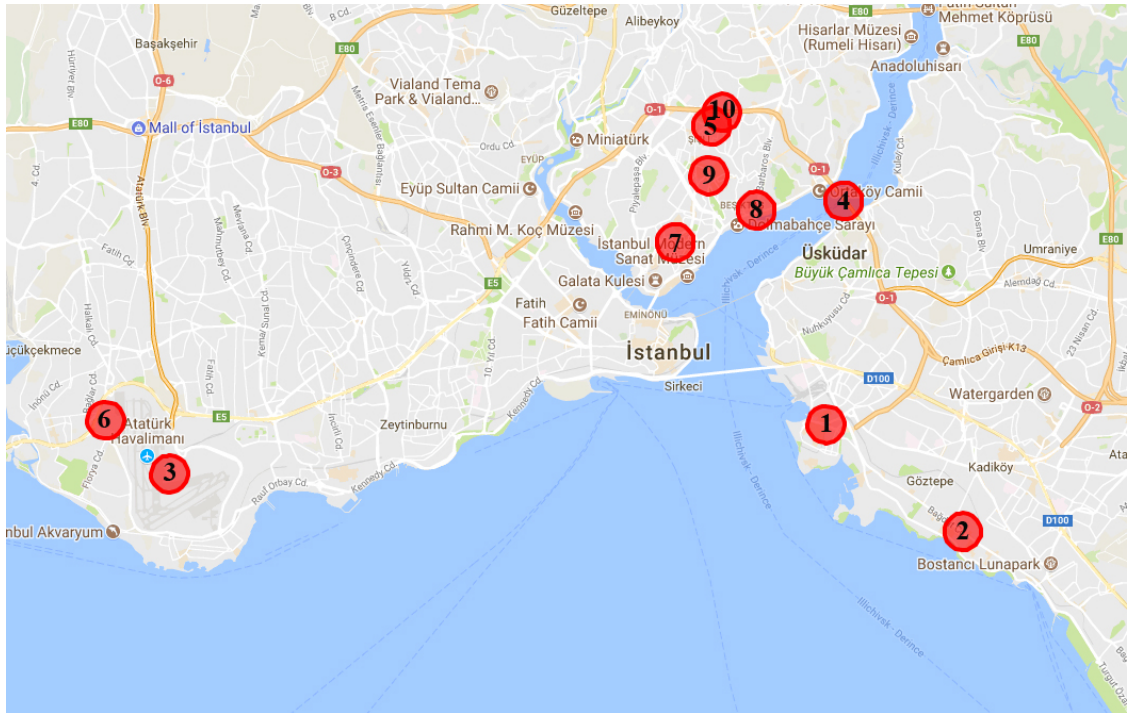
Şekilde görüldüğü üzere, $min_PencereYaygınlığı$ eşik değerinin artışıyla birlikte algoritmaların işlem zamanları düşmektedir. Zamansal elemeyi ilk adım olarak işlettiğinden dolayı, TF-SSTIL Miner algoritmasının diğer algoritmalarından daha hızlı olduğu görülmektedir.

6.4.3.6. İstanbul için Sosyo-Mekân-Zamansal Önemli Lokasyonların Keşfi

Bu deneyde, İstanbul'da yaşayan Twitter sosyal medya kullanıcıları için SMZÖL'lerin keşfi yapılmıştır. $min_yoğunluk$, $min_ziyaret$, tw , min_ZPO , $min_PencereYaygınlığı$ ve

kullanıcı sayısı değerleri, sırasıyla, 0.01, 0.05, 30, 0.01, 0.01 ve 1000 olarak belirlenmiştir. En yüksek 10 lokasyon Şekil 6.10 ve Tablo 6.4’te sunulmuştur. Tabloda verilen lokasyon tanımlamaları manuel olarak çıkarılmış, herhangi bir etiketleme yaklaşımı kullanılmamıştır.

Tablo 6.4’te İstanbul sakinleri için keşfedilen 10 sosyal önemli lokasyon sunulmuştur. Birinci lokasyon, insanların buluşma mekânları olan Kadıköy’ün merkezidir ve İstanbul sakinleri için yıllardır önemli bir lokasyondur. İkinci lokasyon, Göztepe semtinde, çok sayıda alışveriş mağazaları ve restoranların olduğu, oldukça sosyal bir caddedir. Üçüncü lokasyon, günlük çok sayıda yolcunun kullandığı Atatürk Havalimanı’dır. Dördüncü lokasyon, Anadolu ve Avrupa yakalarını birleştiren köprülerden biridir. Beşinci lokasyon, Şişli semtindeki popüler bir alışveriş merkezidir. Altıncı lokasyon, bir üniversite kampüsüdür. Yedinci lokasyon, Beyoğlu semtindeki tarihi bir bulvardır. Sekizinci lokasyon, Beşiktaş semtindeki popüler bir kültür merkezidir. Dokuzuncu lokasyon, Beyoğlu, Beşiktaş ve Şişli semtlerinin kesişimindeki popüler bir lokasyondur. Son olarak onuncu lokasyon, pek çok bankanın genel merkezlerinin, iş merkezlerinin ve gökdelenlerin bulunduğu bir caddedir.



Şekil 6.10. İstanbul için keşfedilen ilk 10 SMZÖL

Önerdiğimiz algoritmalar, İstanbul sosyal medya kullanıcıları için SMZÖL’leri başarılı bir şekilde keşfedebilmiştir. Keşfedilen lokasyonlar, İstanbul sakinleri için popüler ve sıklıkla ziyaret edilen lokasyonlardır. Ayrıca, İstanbul sakini sosyal medya kullanıcılarına odaklandığımız için, turistik lokasyonlar önemli lokasyon olarak keşfedilmemiştir. Bunun sebebi, İstanbul sakinlerinin bu lokasyonları sıklıkla ziyaret etmemeleridir. Bir diğer sebep ise, İstanbul sakinlerinin tercihlerine odaklanmak için turistlerin veri kümesinden çıkarılmasıdır.

Tablo 6.4. İstanbul için keşfedilen ilk 10 lokasyon tanımı

Sıra	Keşfedilen SMZÖL
1	Kadıköy Boğa Heykeli
2	Bağdat Cad., Göztepe
3	Atatürk Havalimanı
4	15 Temmuz Şehitler Köprüsü
5	Şişli Cevahir AVM
6	İstanbul Aydın Üniversitesi
7	İstiklal Cad., Beyoğlu
8	Beşiktaş Kültür Merkezi
9	Vali Konağı Cad., Teşvikiye Cad., Abdi İpekçi Cad.
10	Büyükdere Cad., Şişli

6.4.4. Önerilen TF-SSTIL Miner ve SF-SSTIL Miner Algoritmalarının Literatür ile Karşılaştırılması

Bu bölümdeki çalışma SMZÖL madenciliği problemine odaklanmaktadır. Literatür özetinde belirtildiği üzere, literatürdeki çalışmalar veri kümelerinin farklı boyutlarını dikkate almaktadır. Bilgimiz dâhilinde, literatürde sosyal medya veri kümeleri üzerinden sosyal, mekânsal ve zamansal boyutları kullanarak sosyal önemli lokasyonların ölçen ve SMZÖL’lerin keşfi için kullanıcı yaygınlığını kullanan çalışma bulunmamaktadır. Bu nedenle, bu çalışmada önerilen algoritmaların sonuçlar, bir önceki bölümün sonuçlarıyla ve Khetarpaul vd.’nin [61] önerdiği yaklaşımın sonuçları ile karşılaştırılmıştır. Bu çalışmalar, mekânsal önemli lokasyonların keşfine odaklanmaktadır. Adil bir karşılaştırma yapabilmek için, Khetarpaul vd. [61] tarafından yapılan çalışma, sosyal medya veri kümelerinin zamansal boyutunu da ele alacak şekilde genişletilmiştir.

Bir önceki bölümde keşfedilen mekânsal sosyal önemli lokasyonlar, bu bölümde karşılaştırma yapabilmek adına, Tablo 6.5’te tekrar verilmiştir. Sonuçları

karşılaştırdığımız zaman, zaman boyutunun eklenmesi, lokasyonların zamansal yaygınlıklarına bağlı olarak sıralamayı değiştirmiştir. Ayrıca Tablo 6.5'te bulunan Ragıp Gümüşpala Cad., Eminönü lokasyonu, zamansal yaygınlığı sağlayamadığı için, Tablo 4.4'teki SMZÖL listesinde yer almamaktadır. Bunun yerine, Büyükdere Cad., Şişli sıralamaya girmiştir.

Tablo 6.5. Bölüm 5'te İstanbul için keşfedilen ilk 10 lokasyon

Sıra	Keşfedilen SMZÖL
1	Kadıköy Boğa Heykeli
2	Şişli Cevahir AVM
3	Bağdat Cad., Göztepe
4	Atatürk Havalimanı
5	İstiklal Cad., Beyoğlu
6	15 Temmuz Şehitler Köprüsü
7	Beşiktaş Kültür Merkezi
8	Ragıp Gümüşpala Cad., Eminönü
9	İstanbul Aydın Üniversitesi
10	Vali Konağı Cad., Teşvikiye Cad., Abdi İpekçi Cad.

Bu çalışmanın sonuçları Khetarpaul vd. [61] tarafından önerilmiş olan metot ile, sosyal medya veri kümelerinin hem mekânsal hem de zamansal boyutunu ele alacak şekilde genişletilerek karşılaştırılmıştır. Önerdiğimiz zamansal eleme stratejisi, Khetarpaul vd. [61] tarafından yapılan çalışmaya da eklenip karşılaştırmalar yapılmıştır. Modifiye edilen Khetarpaul algoritması, ilk olarak mekânsal önemli lokasyonları keşfetmekte, ardından zamansal eleme stratejisini kullanarak zamansal önemli lokasyonları keşfetmektedir. Zamansal eleme aşamasında, Bölüm 6.4.3.6'da verilmiş olan zamansal eşik değerleri aynı değerlerle, *tw*, *min_ZPO* ve *min_PencereYaygınlığı* için, sırasıyla, 30, 0.01 ve 0.01, kullanılmıştır. Modifiye edilmiş Khetarpaul algoritmasının sonuçları Tablo 6.6'da verilmiştir. Sonuçları karşılaştırdığımız zaman, Kadıköy Boğa Heykeli, 15 Temmuz Şehitler Köprüsü ve Şişli Cevahir AVM'nin sıralamaları her iki algoritmada da aynıdır. İstanbul Aydın Üniversitesi bu çalışmanın sonuçlarında bulunmakta, ancak Khetarpaul vd. [61] listesinde Ataşehir Bulvarı, Ataşehir bulunmaktadır. İlk 10 lokasyon listesindeki diğer lokasyonlar aynıdır ancak sıralamaları farklıdır. Sonuç olarak, her iki algoritmanın sonuçları farklıdır. Bunun sebebi ise, Khetarpaul vd. [61] SMZÖL'lerin keşfi için lokasyonları ziyaret sıklıklarını dikkate almaktadır. Bu yaklaşımda, az sayıda kullanıcı tarafından çok yüksek ziyaret sıklığıyla ziyaret edilen lokasyonların önem değerleri yüksek olabilmektedir. Tersine, bu bölümde önerilen algoritmalar SMZÖL'lerin keşfi

için kullanıcı yaygınlığını dikkate almaktadır. Bizim yaklaşımımızda, yüksek kullanıcı yaygınlığına sahip lokasyonlar keşfedilmektedir.

Tablo 6.6. Khetarpaul vd. [61] tarafından İstanbul için keşfedilen ilk 10 lokasyon

Sıra	Keşfedilen SMZÖL
1	Kadıköy Boğa Heykeli
2	Atatürk Havalimanı
3	Bağdat Cad., Göztepe
4	15 Temmuz Şehitler Köprüsü
5	Şişli Cevahir AVM
6	Beşiktaş Kültür Merkezi
7	Büyükdere Cad., Şişli
8	Vali Konağı Cad., Teşvikiye Cad., Abdi İpekçi Cad.
9	İstiklal Cad., Beyoğlu
10	Ataşehir Bulvarı, Ataşehir

Keşfedilen SMZÖL sayısını karşılaştırdığımız zaman, 1000 sosyal medya kullanıcısı için, bizim önerdiğimiz algoritmalar 16 SMZÖL keşfetmekte, bunun yanında modifiye edilen Khetarpaul vd. [61] algoritması 158 SMZÖL keşfetmiştir. Modifiye edilen Khetarpaul vd. [61] algoritmasında, minimum kullanıcı sayısı eşik değeri, kendi çalışmalarında da kullandıkları şekilde, 80 olarak belirlenmiştir. Sonuçlar, bizim önerdiğimiz algoritmaların, gereksiz lokasyonları eleme açısından daha verimli olduğu görülmektedir.

Çalışmaları işlem zamanı açısından karşılaştırdığımız zaman, bu çalışma kapsamında önerilen algoritmalar 60 saniye içerisinde SMZÖL'leri keşfedebilmekte, ancak, Khetarpaul vd.'nin [61] algoritması 200 saniyede sonuç üretebilmektedir. Bu sonuçlar da bizim önerdiğimiz SMZÖL madenciliği algoritmalarının, modifiye edilen Khetarpaul vd. [61] algoritmasına karşı üstün olduğunu göstermektedir.

6.5.Katkılar ve Kapsam

Bu kısımda bu bölümde yapılan çalışmanın katkıları ve kapsamından bahsedilmiştir.

6.5.1. Katkılar

Bu bölümde yapılan çalışmanın katkıları aşağıdaki şekilde listelenebilir:

- Sosyo-mekân-zamansal önemli lokasyonlar (SMZÖL)ve SMZÖL madenciliği problemi tanımlanmıştır.

- SMZÖL'lerin keşfi için mekânsal, zamansal ve sosyal değerlendirme ölçütleri önerilmiştir.
- SMZÖL keşfi için yeni algoritmalar önerilmiştir.
- Önerilen algoritmalar, İstanbul Twitter kullanıcılarının veri kümesi ile deneysel olarak değerlendirilmiştir.
- Önerilen algoritmaların sonuçlarıyla benzer çalışmaların sonuçları karşılaştırılmıştır.

6.5.2. Kapsam

Bu bölümde yapılan çalışma, sosyal medya veri kümeleri üzerinden sosyal önemli lokasyonların madenciliği problemine çözümler üretmeye odaklanmıştır. Aşağıdaki konular bu çalışmanın kapsamı dışındadır;

- Keşfedilen lokasyonların anlamsal olarak isimlendirilmesi,
- Lokasyon etiketleme problemine yenilikçi çözümler.

7. BÖLÜM

ÖNERİLEN MEKÂNSAL VE MEKÂN-ZAMANSAL SOSYAL ÖNEMLİ LOKASYONLAR KEŞFİ ALGORİTMALARININ BULUT BİLİŞİM SİSTEMLERİNE UYARLANMASI

Bu bölümde, daha önceki bölümlerde önerilmiş olan mekânsal ve mekân-zamansal sosyal önemli lokasyonların keşfi yaklaşımlarının ve algoritmaların bulut bilişim sistemlerine uyarlanması anlatılmıştır. İlk olarak sosyal önemli lokasyonların keşfi ve bulut bilişim sistemleri tanıtılmıştır. Ardından önemli lokasyonlar madenciliği ve bulut bilişim sistemlerinde sosyal medya madenciliği çalışmalarına ait literatür tartışılmıştır. Daha sonra, sosyal önemli lokasyonların keşfi için önerilmiş olan mekânsal ve mekân-zamansal algoritmalarının bulut bilişim sistemlerine uyarlanması sunulmuştur. Son olarak, deneysel çalışma sonuçları ile bulut bilişim tabanlı olarak önerilmiş olan algoritmalar değerlendirilmiş ve sonuç kısmıyla bölüm tamamlanmıştır.

7.1.Giriş

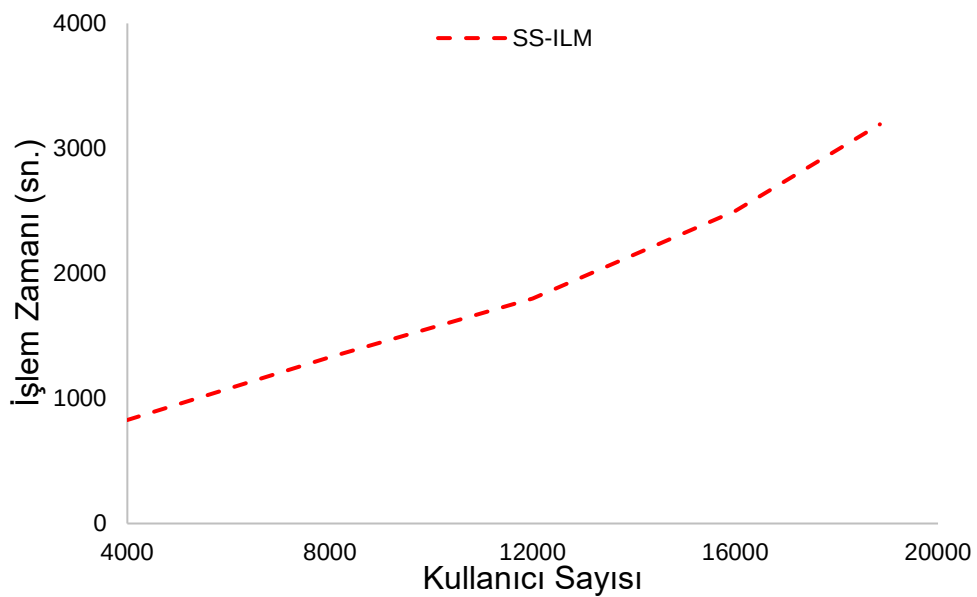
Milyonlarca kullanıcı gündelik olarak sosyal medya ağlarını kullanmakta ve büyük boyutlarda sosyal medya verisi üretmektedirler. Ayrıca, kullanıcıların coğrafi bilgilerinin toplanmaya başlanmasıyla mekânsal sosyal medya madenciliği çalışmaları mümkün hale gelmiştir. Mekânsal sosyal medya verisi kullanılarak, olay tespiti, kullanıcı sıralaması, lokasyon tavsiyesi ve sosyal önemli lokasyonların keşfi gibi pek çok çalışma yapılabilmektedir. Bu bölümdeki çalışmalar, için bulut bilişim sistemleri kullanılarak sosyal önemli lokasyonların keşfi yapılmıştır.

Sosyal önemli lokasyonlar, önceki bölümlerde tanımlandığı üzere, sosyal medya kullanıcılarının sosyal medya ömürlerinde sıklıkla ziyaret ettikleri lokasyonları ifade etmektedir. Sosyal önemli lokasyonların keşfi ile hangi lokasyonların bir sosyal medya kullanıcısı için önemli olduğu, hangi lokasyonların bir sosyal medya kullanıcı grubu için

ortak olduğu, hangi lokasyonların bir kullanıcı grubu arasında tercih edildiği, hangi kullanıcıların sosyal ağlarını mekânsal anlamda yönettikleri ve hangi lokasyonların bir şehir kullanıcıları için önemli olduğu ortaya çıkarılabilir. Ayrıca, bulut bilişim sistemleri, bu lokasyonların büyük sosyal medya verilerinden verimli ve hızlı bir şekilde çıkarılması için güçlü bir seçenektir.

Ancak, sosyal medya veri kümelerinden önemli lokasyonların sosyal önemli lokasyonların keşfedilmesi çeşitli sebeplerden dolayı zordur. Birincisi, sosyal medya veri kümelerinin veri büyüklüğü ve boyutu çok büyüktür. İkincisi, mevcut değerlendirme ölçütleri ve algoritmalar büyük sosyal medya veri kümeleri için yetersizdir. Üçüncüsü, sosyal önemli lokasyonların keşfi için büyük çaplı mekânsal ve zamansal hesaplamalar gerekmektedir. Dördüncüsü, insan kaynaklı uyumsuzluklar sosyal medya veri kümelerinde yaygındır.

Literatürde, sosyal önemli lokasyonların keşfi ve bulut bilişim sistemleri üzerinde sosyal medya madenciliği konularında çeşitli çalışmalar yapılmıştır. Çalışmaların çoğunluğunda ilginç, yaygın ve önemli lokasyonların keşfine odaklanmıştır. Çok az çalışma büyük sosyal medya verisini ele alabilmek için bulut bilişim alternatiflerini kullanmıştır. Ancak bu çalışmalarda da sosyal medya kullanıcı grupları üzerinde sosyal önemli lokasyonların keşfi ve her kullanıcının analiz edildiği bir çalışma yapılmamıştır.



Şekil 7.1. SS-ILM algoritmasının kullanıcı artışına karşılık işlem zamanı değişimi

Şekil 7.1, Bölüm 5’te önerilmiş olan SS-ILM algoritmasının kullanıcı sayısının artışına karşılık işlem zamanı değişimini göstermektedir. Şekilde görüldüğü üzere, işlem zamanı lineerden daha fazla bir artış göstermektedir. Aynı zamanda çok sayıda kullanıcının, örneğin 100.000 kullanıcının, ölçeklenebilir bir zamanda analizinin SS-ILM algoritmasıyla yapılamayacağı da görülmektedir. Daha doğru sosyal önemli lokasyonların keşfedilebilmesi ve dolayısıyla daha fazla sayıda kullanıcının analiz edilebilmesi için kullanıcı sayısının artışına göre işlem zamanını ölçekleyebilmemiz gerekmektedir.

Bu bölümdeki çalışmalarda, büyük sosyal medya veri kümelerinden sosyal önemli lokasyonların keşfedilmesi için bulut bilişim sistemleri kullanılmıştır. Sosyal medya kullanıcılarının sosyal önemli lokasyonlarının keşfedilmesi için Hadoop MapReduce programlama modeli ve Microsoft Azure bulut bilişim ortamı kullanılmıştır. Bölüm 5 ve Bölüm 6’da önerilmiş olan mekânsal ve mekân-zamansal sosyal önemli lokasyonlar keşfi algoritmaları MapReduce modeli kullanılarak modifiye edilmiş ve Türkiye Twitter kullanıcılarına ait büyük sosyal medya veri kümesi üzerinde koşturulmuştur. Algoritmaların bulut bilişim sistemlerindeki performansları, deneysel çalışmalarla test edilmiştir. Deneyler, bulut tabanlı algoritmaların klasik algoritmalar göre işlem zamanı anlamında ne kadar başarılı olduğunu ortaya çıkarmak için gerçekleştirilmiştir. Ayrıca, elde edilen sonuçlar, Türkiye Twitter kullanıcılarının mekânsal ve mekân-zamansal sosyal önemli lokasyonlarını keşfetmiştir.

Bu bölümün devam eden kısımları şu şekilde organize edilmiştir. İkinci kısımda literatür özeti verilmiştir. Üçüncü kısımda bulut bilişim tanıtılmıştır. Dördüncü kısımda mekânsal ve mekân-zamansal sosyal önemli lokasyonlar keşfi algoritmalarının bulut bilişim sistemlerine uyarlanması anlatılmıştır. Beşinci kısımda deneysel değerlendirme yapılmıştır. Altıncı kısımda, bu çalışmanın katkıları ve kapsamı verilmiştir. Yedinci kısımda sonuç ve gelecekteki çalışmalar sunulmuştur.

7.2.Bulut Bilişim Tabanlı Sosyal Önemli Lokasyonların Keşfi

Büyük veri kümeleri üzerinde bulut bilişim sistemlerinin kullanılmasının en kolay yöntemlerinden biri Apache Hadoop ve Hadoop MapReduce programlama modelinin kullanılmasıdır. Bölüm 2’de anlatıldığı üzere, uygulamalar Map ve Reduce fonksiyonları kullanılarak dağıtıklaştırılabilen ve bulut bilişim sistemlerine uyarlanabilmektedir.

Bölüm 5 ve Bölüm 6’da önerilen sosyal önemli lokasyonların keşfi metotları da MapReduce programlama modeline uyarlanabilmektedir. Kullanıcı-seviyesinde sosyal önemli lokasyonların keşfi Map fonksiyonu kullanarak, grup-seviyesinde sosyal önemli lokasyonların keşfi ise Reduce fonksiyonu kullanarak gerçekleştirilebilir.

Bu bölümde, Bölüm 5 ve Bölüm 6’da önerilmiş olan mekânsal ve mekân-zamansal sosyal önemli lokasyonların keşfi algoritmaları olan SS-ILM, TF-SSTIL Miner ve SF-SSTIL Miner algoritmalarının MapReduce programlama modeline uyarlanmış versiyonları olan Bulut SS-ILM, Bulut TF-SSTIL Miner ve Bulut SF-SSTIL Miner algoritmaları önerilmiştir.

7.2.1. Bulut SS-ILM algoritması

SS-ILM algoritması, bulut bilişim sistemlerine dağıtık mantıkla paralelleştirilebilmesi ve çok sayıda kullanıcının kısa süreler içerisinde analiz edilebilmesi için MapReduce programlama modeline uyarlanmış ve Bulut SS-ILM algoritması önerilmiştir. SS-ILM algoritmasının kullanıcı-seviyesindeki mekânsal sosyal önemli lokasyonlar keşfi adımları Map fonksiyonuyla, grup-seviyesindeki mekânsal sosyal önemli lokasyonlar keşfi adımları da Reduce fonksiyonuyla gerçekleştirilmiştir. Bir sosyal medya kullanıcısının lokasyon yoğunluğu ve ziyaret ömrü hesaplamaları çok kısa sürdüğü için bir Map işleminde birden çok sayıda kullanıcı analiz edilmiştir. Bulut SS-ILM algoritmasına ait kaba kod Algoritma 7.1’de verilmiştir.

Algoritma 7.1’de görüldüğü üzere, Bulut SS-ILM algoritması, Setup, Map ve Reduce olmak üzere üç fonksiyon kullanmaktadır. Setup fonksiyonunda veri kümesi önışlemden geçirilmekte ve Map fonksiyonuna atanan her bir kullanıcı için Kullanıcı Aktivite (KA) matrisi oluşturulmaktadır. Map fonksiyonunda, her bir kullanıcının kullanıcı sosyal önemli lokasyonları (KSÖL) *KSÖL-kesfet* fonksiyonuyla keşfedilmektedir. *KSÖL-kesfet* fonksiyonunun sonuçları, Map fonksiyonunun çıktısı olarak <key, value> yaklaşımına göre yazdırılmaktadır. Reduce fonksiyonunda, Map çıktıları girdi olarak alınmakta ve her bir lokasyonun oluşumu sayılmaktadır. Eğer bir lokasyonun oluşum sayısı *min_KY* eşik değerini sağlıyorsa Reduce fonksiyonunun çıktısı olarak yazdırılmaktadır.

Algoritma 7.1. BulutSS-ILM Algoritması**Girdiler:**

D: Sosyal medya kullanıcı veri kümesi
L: Çıkarılmış ve etiketlenmiş lokasyonlar kümesi
min_yoğunluk: Minimum lokasyon yoğunluğu eşik değeri
min_ziyaret: Minimum ziyaret ömrü eşik değeri
min_KY: Minimum kullanıcı yaygınlığı eşik değeri

Çıktı: Mekânsal sosyal önemli lokasyonlar kümesi

Algoritma:

Map Setup Aşaması:

for each kullanıcı *k in* map işlemi
 KA $\leftarrow$ önışlem-uyguyla(*D*[*k*], *L*[*k*])
end for

Map Aşaması:

for each kullanıcı *k in* map işlemi
 KSÖL = KSÖL-kesfet(KA[*k*], *min_yoğunluk*, *min_ziyaret*)
 for each *lok in* KSÖL
 map.write(*lok*, 1)
 end for
end for

Reduce Aşaması:

olusumSay = 0;
for each değer *in lok*
 olusumSay += değer
end for

if olusumSay $\geq$ *min_KY*
 reduce.write(*lok*, olusumSay)
end if
return Reduce çıktısı

7.2.2. Bulut TF-SSTIL Miner

TF-SSTIL Miner algoritması, SS-ILM algoritması ile aynı mantıkla Hadoop MapReduce programlama modeline uyarlanmış ve paralelleştirilmiştir. TF-SSTIL Miner algoritması ilk olarak zamansal analiz yapmakta ve kullanıcı zamansal önemli lokasyonları keşfetmekte, ardından kullanıcı zamansal önemli lokasyonları kullanarak kullanıcı sosyo-mekân-zamansal önemli lokasyonları keşfetmektedir. Ardından kullanıcı sosyo-mekân-zamansal önemli lokasyonlar üzerinden grup seviyesinde sosyo-mekân-zamansal önemli

lokasyonlar keşfedilmektedir. TF-SSTIL Miner algoritması MapReduce modeline uyarlanırken, kullanıcı seviyesinde mekân-zamansal önemli lokasyonlar keşfi aşaması Map fonksiyonuyla, grup seviyesinde sosyo-mekân-zamansal önemli lokasyonlar keşfi aşaması Reduce fonksiyonuyla gerçekleştirilmiştir. Geliştirilmiş olan Bulut TF-SSTIL Miner algoritmasına ait kaba kod Algoritma 7.2’de verilmiştir.

Algoritma 7.2. Bulut TF-SSTIL Miner Algoritması

Girdiler:

D: Sosyal medya kullanıcı veri kümesi
L: Çıkarılmış ve etiketlenmiş lokasyonlar kümesi
min_yoğunluk: Minimum lokasyon yoğunluğu eşik değeri
min_ziyaret: Minimum ziyaret ömrü eşik değeri
min_ZPO: Minimum zaman penceresi oluşumu eşik değeri
min_PencereYaygınlığı: Minimum pencere yaygınlığı eşik değeri

Çıktı: Sosyo-mekân-zamansal önemli lokasyonlar kümesi

Algoritma:

Map Setup Aşaması:

for each kullanıcı *k* **in** map işlemi
 KA ← önilem-uygula(*D*[*k*], *L*[*k*])
end for

Map Aşaması:

for each kullanıcı *k* **in** map işlemi
 KZÖL = KZÖL-kesfet(KA[*k*], min_ZPO, min_PencereYaygınlığı)
 KMÖL = KMÖL-kesfet(KA[*k*], KZÖL, min_yoğunluk, min_ziyaret)
 for each *lok* **in** KMÖL
 map.write(*lok*, 1)
 end for
end for

Reduce Setup Aşaması:

ZPÖL ← önilem-uygula()

Reduce Aşaması:

for each *lok* **in** reduce işlemi
 SMZLÖ = SMZLÖ-hesapla(ZPÖL, *lok*, min_ZPO / 10)
 if SMZLÖ ≥ min_PencereYaygınlığı
 reduce.write(*lok*, SMZLÖ)
 end if
end for
return Reduce çıktısı

Algoritma 7.2’de görüldüğü üzere, Map fonksiyonundan önce bir Setup fonksiyonu işletilmiş ve Bulut SS-ILM algoritmasında olduğu gibi, Kullanıcı Aktivite matrisi oluşturulmuştur. Daha sonra Map fonksiyonundaki her bir kullanıcı için, TF-SSTIL Miner algoritmasında olduğu gibi, *KZÖL-kesfet* fonksiyonuyla zamansal analiz yapılmakta ve kullanıcı zamansal önemli lokasyonlar (KZÖL) keşfedilmiştir. Ardından, KZÖL’ler kullanılarak, *KMÖL-kesfet* fonksiyonuyla mekânsal analiz yapılmakta ve kullanıcı sosyo-mekân-zamansal önemli lokasyonlar keşfedilmiştir. Map fonksiyonunun çıktısı olarak KMÖL’ler yazdırılmaktadır. Reduce fonksiyonuna ait Setup fonksiyonunda zaman penceresinde önemli lokasyonlar (ZPÖL) matrisi önışlem aracılığıyla toplanmıştır. Ardından, Reduce fonksiyonundaki *SMZLÖ-hesapla* fonksiyonuyla her bir lokasyon için sosyo-mekân-zamansal lokasyon önemi değeri hesaplanmıştır. Bu fonksiyonun girdilerinden biri olan *min_ZPO* değeri, kullanıcı sayısının artışı ölçekleyebilmek için 10’a bölünerek sosyal seviyeye uyarlanmıştır. Son olarak, eğer lokasyonun SMZLÖ değeri *min_PencereYaygınlığı* eşik değerini sağlıyorsa lokasyon Reduce fonksiyonunun çıktısı olarak verilmiştir.

7.2.3. Bulut SF-SSTIL Miner

SF-SSTIL Miner algoritması da TF-SSTIL Miner algoritmasıyla aynı şekilde MapReduce programlama modeline uyarlanmıştır. İki algoritma arasındaki tek fark, mekânsal ve zamansal analizlerin işletilme sıralamasıdır. Algoritma 7.3’te Bulut SF-SSTIL Miner algoritmasına ait kaba kod sunulmuştur.

Algoritma 7.3’te görüldüğü üzere, Bulut TF-SSTIL Miner ve Bulut SF-SSTIL Miner algoritmalarının tüm adımlar aynı şekilde işletilmekte, sadece KMÖL ve KZÖL’lerin hesaplanması sıralaması değişmektedir.

Algoritma 7.3. Bulut SF-SSTIL Miner Algoritması

Girdiler:

- D*: Sosyal medya kullanıcı veri kümesi
- L*: Çıkarılmış ve etiketlenmiş lokasyonlar kümesi
- min_yoğunluk*: Minimum lokasyon yoğunluğu eşik değeri
- min_ziyaret*: Minimum ziyaret ömrü eşik değeri
- min_ZPO*: Minimum zaman penceresi oluşumu eşik değeri
- min_PencereYaygınlığı*: Minimum pencere yaygınlığı eşik değeri

Çıktı: Sosyo-mekân-zamansal önemli lokasyonlar kümesi

Algoritma:

Map Setup Aşaması:

for each kullanıcı k **in** map işlemi
 $KA \leftarrow \text{önişlem-uygula}(D[k], L[k])$
end for

Map Aşaması:

for each kullanıcı k **in** map işlemi
 $KMÖL = KMÖL\text{-kesfet}(KA[k], \text{min_yoğunluk}, \text{min_ziyaret})$
 $KZÖL = KZÖL\text{-kesfet}(KA[k], KMÖL, \text{min_ZPO}, \text{min_PencereYaygınlığı})$
 for each lok **in** $KZÖL$
 $\text{map.write}(lok, 1)$
 end for
end for

Reduce Setup Aşaması:

$ZPÖL \leftarrow \text{önişlem-uygula}()$

Reduce Aşaması:

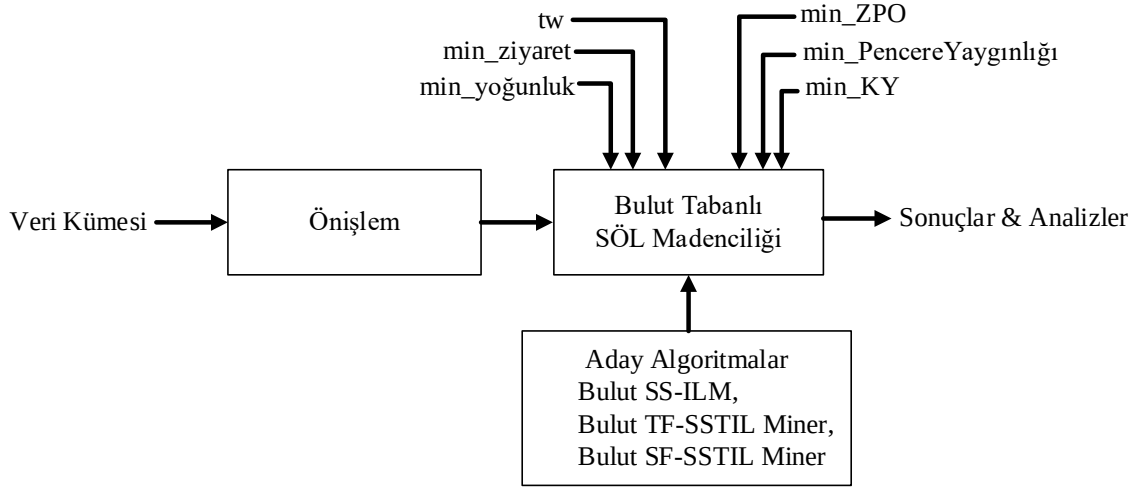
for each lok **in** reduce işlemi
 $SMZLÖ = SMZLÖ\text{-hesapla}(ZPÖL, lok, \text{min_ZPO} / 10)$
 if $SMZLÖ \geq \text{min_PencereYaygınlığı}$
 $\text{reduce.write}(lok, SMZLÖ)$
 end if
end for
return Reduce çıktısı

7.3.Deneysel Çalışmalar

Bu kısımda, ilk olarak veri kümesi tanıtılmış, önişlem adımları anlatılmış, kullanılan bulut bilişim ortamı sunulmuş ve deneyler verilmiştir. İlk dört deney algoritmaların işlem zamanlarının performansı üzerine yapılmış olan deneylerdir. Son deney ise Türkiye Twitter kullanıcıları için keşfedilmiş olan mekânsal ve mekân-zamansal sosyal önemli lokasyonları sunmaktadır. Şekil 7.2’de deney düzeneği verilmiştir. Deneyler aşağıdaki soruları cevaplamak için gerçekleştirilmiştir:

- Kullanıcı sayısının işlem zamanına etkisi nedir?
- Verinin parçalara bölünmesinin işlem zamanına etkisi nedir?
- İşçi düğüm sayısının işlem zamanına etkisi nedir?
- İşçi kaynak boyutunun işlem zamanına etkisi nedir?

- Hangi lokasyonlar Türkiye Twitter kullanıcıları için sosyal önemli lokasyonlardır?

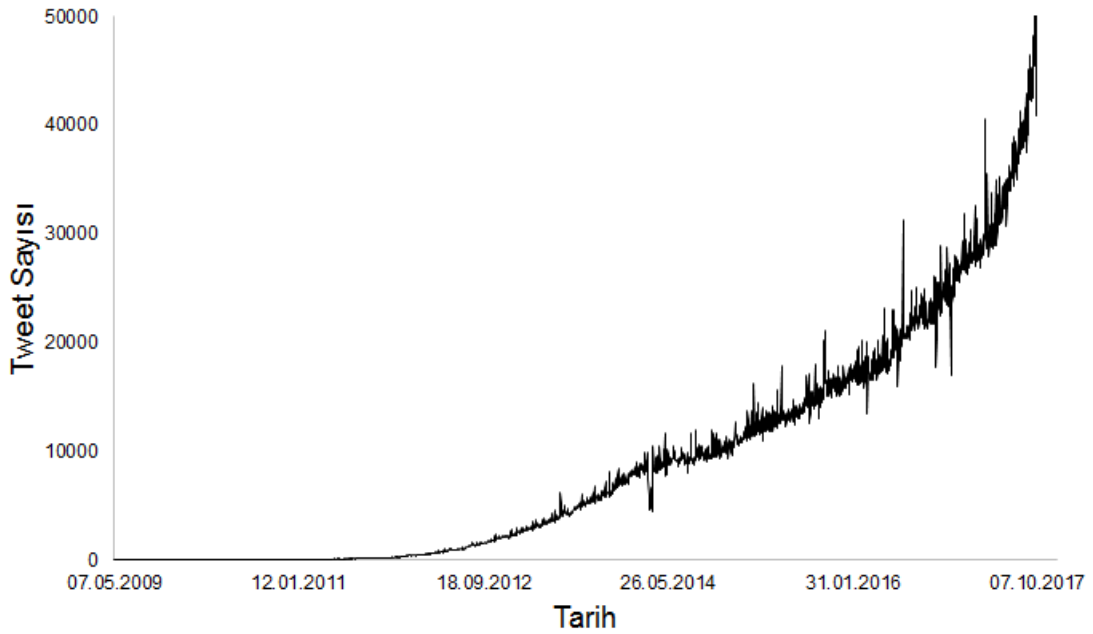


Şekil 7.2. Deney düzeneği

7.3.1. Veri Kümesi

Bu bölümdeki çalışmalarda kullanılan veri kümesi aktif ve gerçek olan Türkiye Twitter kullanıcılarının tüm tweetleridir. Türkiye Twitter kullanıcılarını seçmek ve coğrafi arama yapabilmek için Twitter'ın Streaming API'si kullanılmıştır [18]. Arama yapılırken, Türkiye merkez enlem ve boyları ve Türkiye'yi kapsayacak bir yarıçap belirlenmiş ve sorgu bu mekânsal bilgiler ve ek bilgiler, tweetlerin dili gibi, kullanılarak oluşturulmuştur. Ardından, sorgu Streaming API'ye gönderilmiş ve sonuçlar toplanmıştır. Bu sorgunun defalarca işletilmesiyle veri kümesi oluşturulmuştur. Bu adımda 27501 kullanıcı toplanmıştır.

Bir sosyal medya kullanıcı grubuna ait sosyal önemli lokasyonların keşfedilebilmesi için, bu grubun tüm geçmiş tweetlerinin bulunması gerekmektedir. Bu amaçla, Twitter'dan kullanıcı tweetlerini toplamak için REST API kullanılmıştır [18]. Tweetler için, enlem, boylam ve tarih olmak üzere üç parametre toplanmıştır. Veri kümesinde 27501 kullanıcıya ait 55.586.653 tweet bulunmaktadır. Veri kümesi Mayıs 2009 ve Ekim 2017 tarihleri arasındadır. Veri kümesinin zamansal dağılımı Şekil 7.3'te verilmiştir. Şekilde görüldüğü üzere, kullanıcıların tweetlerinin büyük çoğunluğu son zamanlarda atılmış ve bu eğilim artarak devam etmektedir. Bu durumun asıl sebebi, Twitter'ın 2012'den sonra Türkiye'de yaygın hale gelmeye başlamasıdır.



Şekil 7.3. Veri kümesinin zamansal dağılımı

Twitter API'leriyle ilgili tüm sorgular Twitter4j açık kaynak kodlu Java kütüphanesi kullanılarak gerçekleştirilmiştir [19].

7.3.2. Önilem Aşamaları

Bu kısımda, çalışmaya ait önilem aşamaları anlatılmıştır. İlk olarak veri kümesi temizlenmiş, ikinci olarak kullanıcılar seçilmiş, üçüncü olarak zamanda aşırı ağırlıklandırmanın önlenmesi anlatılmış ve son olarak lokasyon etiketleme prosedürü anlatılmıştır.

7.3.2.1. Veri Temizleme

Sosyal medya veri kümelerinden sosyal önemli lokasyonları başarılı bir şekilde çıkarabilmek için pasif ve sahte kullanıcıların veri kümesinden elenmesi gerekmektedir. Pasif kullanıcı, sosyal medya geçmişinde az sayıda tweet gönderen kullanıcıları ifade etmektedir. Pasif kullanıcıları elemek ve aktif kullanıcıları seçmek için, değeri 50 olan minimum tweet sayısı eşik değeri belirlenmiştir. Eğer bir kullanıcı 50'den az tweet göndermişse, bu kullanıcı pasif olarak etiketlenmiş ve elenmiştir. Ayrıca, sahte davranışlar Twitter'da oldukça yaygındır. Sahte kullanıcı davranışları analiz edildiği zaman, bu kullanıcıların takip ettikleri kişilere göre oldukça az sayıda takipçileri oldukları görülmüştür. Bu nedenle sahte kullanıcıları tespit etmek için iki kriter kullanılmıştır ve

bunlar takipçi sayısı ve takipçi/takip edilen oranıdır. Eğer bir kullanıcının takipçi sayısı 10'un altında ve takipçi/takip edilen oranı 0.1'in altında ise bu kullanıcı sahte olarak tespit edilmiş ve elenmiştir. Sosyal medya ağlarında sahtelik tespiti konusunda daha detaylı bilgi için, [97] ve [98] incelenebilir.

7.3.2.2.Kullanıcı Seçimi

Bu çalışmanın ana odak noktası Türkiye Twitter kullanıcılarının sosyal önemli lokasyonlarının keşfedilmesidir. Türkiye tabanlı kullanıcı toplama işleminde Streaming API kullanıldığı ve bu API geçmiş verileri sağlamadığı için, toplanmış olan kullanıcıların gerçek Türkiye kullanıcısı veya Türkiye'den geçerken tweet paylaşan kullanıcı olduğu, veri toplama işlemi esnasında tespit edilememiştir. Bir kullanıcının gerçek Türkiye kullanıcısı olup olmadığının tespiti için, kullanıcı tweetlerinin enlem ve boylamları Türkiye merkez enlem ve boylamından çıkarılmıştır. Eğer bir kullanıcının tweetlerinin yarısından fazlası Türkiye sınırları içerisinde atılmış olarak tespit edilirse bu kullanıcı gerçek Türkiye kullanıcısı olarak seçilmiştir. Bu işlem sonrasında 18.853 kullanıcı gerçek ve aktif Türkiye Twitter kullanıcısı olarak seçilmiştir. Seçilen veri kümesinde 42.606.566 tweet bulunmakta ve veri kümesi Mayıs 2009 ve Ekim 2017 tarihleri arasındadır.

7.3.2.3.Zamanda Aşırı Ağırlıklandırılmanın Önlenmesi

Veri kümesini incelediğimiz zaman, kullanıcıların bir lokasyonda aynı zamanda birden fazla tweet gönderdikleri (sosyal medya aktivitesi gerçekleştirdiklerini) görülmüştür. Bu davranış sürekli hale geldiği zaman, bir lokasyon doğru bulunma sayısından fazla görülmektedir, çünkü kullanıcı lokasyona bir defa gitmiş ancak birden fazla defa tweet göndermiştir. Bu problemi bir lokasyonun zamanda aşırı ağırlıklandırılması olarak tanımlanmıştır. Bu problem bazı durumlarda istenmeden, örneğin, bir kullanıcı arkadaşlarıyla sohbet ederken, gerçekleşebilmektedir. Bir lokasyonun zamanda aşırı ağırlıklandırılmasını önlemek için, 60 dakikalık bir eşik değeri tanımlanmıştır. Eğer bir kullanıcı 60 dakika içerisinde aynı lokasyonda birden fazla defa tweet paylaşırsa, bu lokasyon bilgisinin yeni olmadığı varsayılmış ve lokasyonda oluşumu bir defa olarak sayılmıştır. Bu yaklaşım yardımıyla bir lokasyonun zamanda aşırı ağırlıklandırılması önlenmiş olmaktadır. Bunun yanında, zamanda aşırı ağırlıklandırılmayı önlemek için farklı yaklaşımlar/kriterler de geliştirilebilir.

7.3.2.4.Lokasyon Etiketleme

Twitter API'leri kullanıcı tweetlerine ait hassas enlem-boylam ikilileri sağlamaktadır. Bu yaklaşım hassas sonuçlar elde etmek için faydalı olsa da, lokasyon etiketleme işlemi için bir problemdir. Örneğin, bir alışveriş merkezi veya stadyum 1 km² alan içerisinde bulunmaktadır ancak hassas enlem-boylam ikilileri eşleşmediği için çok sayıda farklı lokasyon tanımlanabilmektedir. Bu problemin üstesinden gelebilmek için farklı enlem-boylam ikililerinin aynı lokasyon olarak sayılabilmesi için bir eşik değeri tanımlanmıştır. Daha önce [52] ve [47]'de yapılan çalışmalarda kullanıldığı gibi, bu eşik değerini 100 metre olarak tanımlanmıştır. Eğer iki lokasyon birbirine 100 metreden daha yakınsa, bu iki lokasyon aynı etiketle etiketlenmektedir.

7.3.3. Bulut Bilişim Ortamı

Bu çalışmada bulut bilişim ortamı olarak Microsoft Azure platformu kullanılmıştır [101]. Microsoft Azure platformunun temel avantajlarından biri veri ve hesaplama nodlarını birbirinden ayırmasıdır. Hadoop Mapreduce öncelikle hesaplama kümelerin kurulması ve daha sonra verinin bu kümeler içerisine kopyalanması mantığıyla çalıştığı için, bu yaklaşım çok fazla sayıda girdi/çıkı işlemleri gerektirmektedir. Aynı zamanda, bulut bilişim sistemlerinin kullandığın-kadar-öde sistematiğinde, bulut bilişim sistemi özel bulut ortamından oluşmuyorsa, büyük miktarlarda ödeme yapılması gerekliliğine sebep olmaktadır. Ancak, Microsoft Azure Windows Azure Depolama Blobu (Windows Azure Storage Blob – WASB) sağlamakta ve yeni hesaplama kümeleri kurulup silindiği durumlarda dahi veri kümesi WASB'ta güvenli bir şekilde depolanmaktadır. Bu fayda sayesinde deneylerimizi hızlı bir şekilde tamamlama ve maliyeti düşürme imkânına sahip olduk.

Microsoft Azure platformunda hesaplama kümeleme servisi olarak Azure HDInsight servisi kullanılmıştır. Azure HDInsight, ön-kurulmuş ve tüm Hadoop yetenekleri eklenmiş olarak Apache Hadoop kümeleri sağlayan bir bulut servisi. HDInsight kullanılarak, bir makineler kümesi birkaç adımda kolaylıkla kurulabilmektedir. İlk adımda, küme adı, küme türü (Hadoop, Spark, Kafka vb.), işletim sistemi, küme sürümü, küme giriş yetkileri ve kaynak grubu seçilmektedir. İkinci adımda, depolama hesap seçenekleri sağlanmaktadır. Üçüncü adımda, seçimli uygulamalar eklenebilmektedir. Son

olarak dördüncü adımda kaynak kapasitelerini gösteren nod büyüklükleri seçilerek küme dakikalar içerisinde kurulabilmektedir.

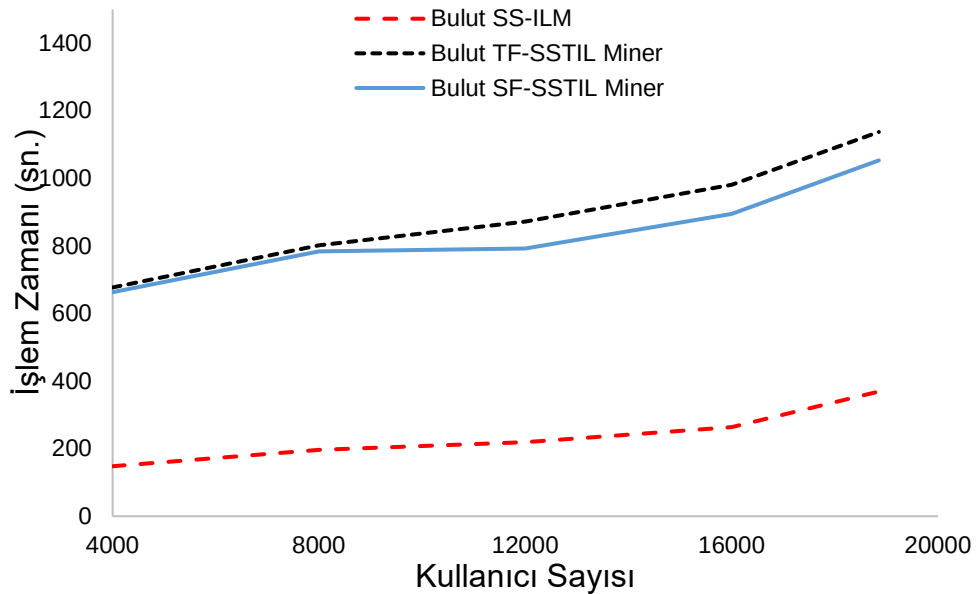
Depolama hesabı olarak, standart WASB depolama seçilmiştir. Veri kümesi ve önışlemden geçirilmiş dosyalar blob olarak hesap içerisinde yüklenmiştir.

7.3.4. Deneysel Sonuçlar

Bu kısımda, bu bölümde önerilmiş olan Bulut SS-ILM, Bulut TF-SSTIL Miner ve Bulut SF-SSTIL Miner algoritmalarının deneysel sonuçları sunulmuştur. İlk dört deney, algoritmaların işlem zamanı performansları değerlendirilmiş, son deneyde ise Türkiye Twitter kullanıcıları için sosyal önemli lokasyonlar değerlendirilmiştir.

7.3.4.1. Kullanıcı Sayısının Etkisi

Bu deneyde, kullanıcı sayısının artışının algoritmaların işlem zamanlarına etkisi analiz edilmiştir. Veri parçası sayısı, işçi düğüm sayısı, işçi boyutu, *min_yoğunluk*, *min_ziyaret*, *min_ZPO*, *min_PencereYaygınlığı* ve *min_KY* değerleri, sırasıyla, 50, 8, A3, 0.005, 0.01, 0.01, 0.01 ve 0.005 olarak belirlenmiştir. Kullanıcı sayısı 4000, 8000, 12000, 16000 ve 18853 olarak deneyler yapılmıştır. Şekil 7.4'te kullanıcı sayısının artışının algoritmaların işlem zamanlarına etkisi sunulmuştur.

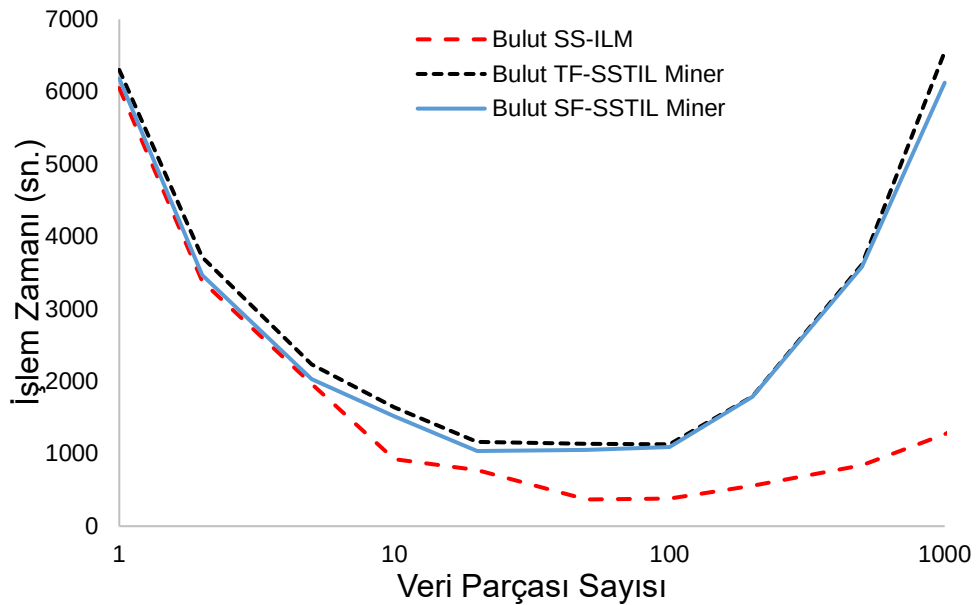


Şekil 7.4. Kullanıcı sayısının etkisi

Şekilde görüldüğü üzere, kullanıcı sayısının artması tüm algoritmaların işlem zamanlarını artırmaktadır. Tüm kullanıcıların kullanıldığı son deneyde ise işlem zamanı lineerden daha fazla artış göstermektedir. Ayrıca, Bulut SS-ILM algoritması, zamansal işlemler yapmaması sebebiyle diğer algoritmalarından daha hızlı sonuç üretmektedir. Bulut TF-SSTIL Miner ve Bulut SF-SSTIL Miner algoritmaları arasında ise Bulut SF-SSTIL Miner algoritması daha hızlı sonuç üretebilmektedir. Bunun sebebi olarak, veri kümesi üzerinde mekânsal eleme işleminin zamansal eleme işleminden daha önce yapılmasının daha fazla aday örüntüyü erken eleyebildiği için işlem zamanını düşürebildiği söylenebilir.

7.3.4.2. Veri Parçası Sayısının Etkisi

Bu deneyde veri kümesinin parçalara bölünmesinin algoritmaların işlem zamanlarına etkisi analiz edilmiştir. Her bir kullanıcı küçük zamanlarda analiz edildiği için, veri kümesinin parçalara bölümlendirilmesinin algoritmaların işlem zamanlarını düşürmesi gerekmektedir. İşçi düğüm sayısı, işçi boyutu, *min_yoğunluk*, *min_ziyaret*, *min_ZPO*, *min_PencereYaygınlığı*, *min_KY* ve kullanıcı sayısı değerleri, sırasıyla, 8, A3, 0.005, 0.01, 0.01, 0.01, 0.005 ve 18853 olarak belirlenmiştir. Şekil 7.5'te veri parçası sayısının algoritmaların işlem zamanlarına etkisi sunulmuştur.



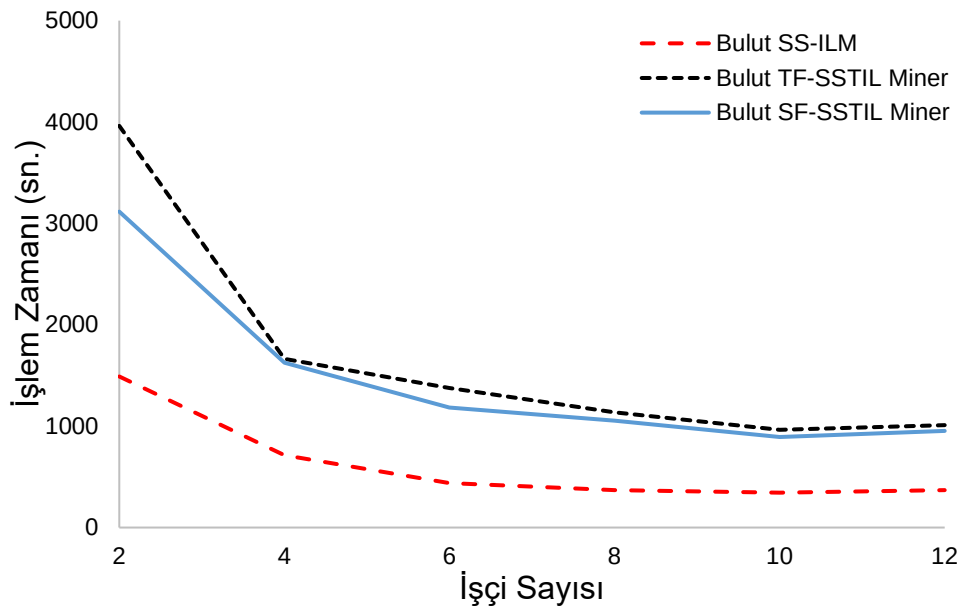
Şekil 7.5. Veri parçası sayısının etkisi

Şekilde görüldüğü üzere, verinin tek parça veya birkaç parça halinde analiz edilmesi işlem zamanlarını oldukça artırmaktadır. Bu durumun sebebi, bulut bilişim sistemlerinin veriyi gruplar halinde analiz etmeye daha yatkın olmasıdır. Aynı zamanda, bir önceki deney

sonucuyla benzer şekilde, Bulut SS-ILM algoritması diğer algoritmalarından daha hızlı sonuç üretebilmektedir. Bulut TF-SSTIL Miner ve Bulut SF-SSTIL Miner algoritmaları ise benzer eğilimi göstermektedirler. Aynı zamanda verinin çok fazla parçaya bölünmesinin de performansı olumsuz etkilediği gözlemlenmiştir. Bulut SS-ILM algoritması için optimum veri parçası sayısı 50 olarak gözlemlenmişken, Bulut TF-SSTIL Miner ve Bulut SF-SSTIL Miner algoritmaları için optimum veri parçası sayısı 20 olarak gözlemlenmiştir.

7.3.4.3.İşçi Düğüm Sayısının Etkisi

Bu deneyde işçi düğüm sayısının algoritmaların işlem zamanlarına etkisi analiz edilmiştir. Veri parçası sayısı, işçi boyutu, *min_yoğunluk*, *min_ziyaret*, *min_ZPO*, *min_PencereYaygınlığı*, *min_KY* ve kullanıcı sayısı değerleri, sırasıyla, 50, A3, 0.005, 0.01, 0.01, 0.01, 0.005 ve 18853 olarak belirlenmiştir. Şekil 7.6’da işçi düğüm sayısının algoritmaların işlem zamanlarına etkisi sunulmuştur.



Şekil 7.6. İşçi düğüm sayısının etkisi

Şekilde görüldüğü üzere, işçi düğüm sayısının artırılması, her üç algoritma için de işlem zamanını düşürebilmektedir. Ancak işlem zamanını düşürebilme oranı eklenen yeni işçi düğümleriyle azalmaktadır. 10 işçi ile yapılan deneyler ile 12 işçi ile yapılan deneyler arasında işlem zamanı açısından çok az bir fark bulunmaktadır. Bu durum da işçi sayısının artırılmasının bir noktadan sonra faydası olmayacağını göstermektedir. Bulut SS-ILM algoritması önceki deneylerde de gözlemlendiği şekilde, zamansal eleme yapmaması

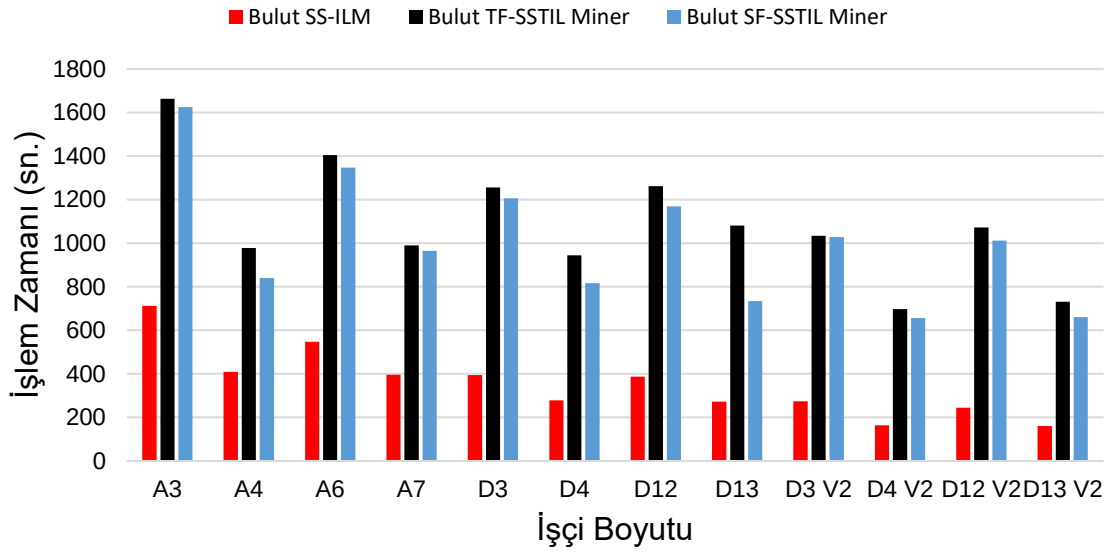
sebebiyle daha hızlı sonuç üretebilirken, Bulut TF-SSTIL Miner algoritması, Bulut TF-SSTIL Miner algoritması ile arasındaki işlem zamanı farkını işçi düğüm sayısının artışıyla birlikte kapatabilmiştir. Bu durumun sebebi ise, işlemin paralelleştirilebilme oranı ne kadar artarsa, algoritmalar arasındaki işlem zamanı farkının da o derece düştüğü olarak ifade edilebilir.

7.3.4.4.İşçi Boyutunun Etkisi

Bu deneyde Microsoft HDInsight kümeleri üzerinde öntanımlı olan işçi boyutunun algoritmaların işlem zamanlarına etkisi analiz edilmiştir. Veri parçası sayısı, işçi sayısı, *min_yoğunluk*, *min_ziyaret*, *min_ZPO*, *min_PencereYaygınlığı*, *min_KY* ve kullanıcı sayısı değerleri, sırasıyla, 50, 4, 0.005, 0.01, 0.01, 0.01, 0.005 ve 18853 olarak belirlenmiştir. Microsoft Azure ortamındaki işçi boyutları ve her bir makinenin donanım kapasitesi Tablo 7.1’de verilmiştir. Ayrıca tabloda, işçi elemanların ek özellikleri de belirtilmiştir. Şekil 7.7’de işçi boyutunun algoritmaların işlem zamanlarına etkisi sunulmuştur.

Tablo 7.1. Microsoft Azure işçi boyut seçenekleri

İşçi Etiket	RAM	Çekirdek Sayısı	Ek özellik
A3	7 GB	4	
A4	14 GB	8	
A6	28 GB	4	
A7	56 GB	8	
D3	14 GB	4	
D4	28 GB	8	
D12	28 GB	4	
D13	56 GB	8	
D3 V2	14 GB	4	35% daha hızlı CPU
D4 V2	28 GB	8	35% daha hızlı CPU
D12 V2	28 GB	4	35% daha hızlı CPU
D13 V2	56 GB	8	35% daha hızlı CPU



Şekil 7.7. İşçi boyutunun etkisi

Şekil ve tablo birlikte incelendiği zaman, algoritmaların işlem zamanlarını etkileyen en önemli faktör, işçi elemanlarındaki işlemci sayısıdır. İkinci önemli etmen de işçi elemanlarındaki RAM miktarlarıdır. İşlemcilerin performansının daha iyi olması da üçüncü derecede işlem zamanlarını etkilemektedir. Ayrıca, Bulut SS-ILM algoritmasının işlem zamanı, diğer algoritmalar göre oldukça düşük olarak gözlemlenebilmiştir. Benzer olarak Bulut SF-SSTIL Miner algoritması da Bulut TF-SSTIL Miner algoritmasına göre daha performanslı çalışmaktadır. Yapılan deney sonuçlarına göre, D4 V2 işçi etiketine sahip elemanların en verimli oldukları da gözlemlenmiştir. En kötü performansı ise A3 işçi etiketli eleman göstermiştir.

7.3.4.5. Türkiye Twitter Kullanıcıları için Sosyal Önemli Lokasyonların Keşfi

Bu deneyde Türkiye Twitter kullanıcıları için keşfedilen sosyal önemli lokasyonlar (SÖL) değerlendirilmiştir. Veri parçası sayısı, işçi sayısı, işçi boyutu, *min_yoğunluk*, *min_ziyaret*, *min_ZPO*, *min_PencereYaygınlığı*, *min_KY* ve kullanıcı sayısı değerleri, sırasıyla, 50, 8, A3, 0.005, 0.01, 0.01, 0.01, 0.005 ve 18853 olarak belirlenmiştir. Mekânsal, Bulut SS-ILM, ve mekân-zamansal, Bulut TF-SSTIL Miner ve Bulut SF-SSTIL Miner, algoritmalar tarafından keşfedilen SÖL'lar Şekil 7.8 ve Şekil 7.9'da sunulmuştur. Ayrıca ilk 10 lokasyon Tablo 7.2'de sunulmuştur.



Şekil 7.8. Bulut SS-ILM algoritması tarafından keşfedilen SÖL'lar



Şekil 7.9. Bulut SF-SSTIL Miner ve Bulut TF-SSTIL Miner algoritmaları tarafından keşfedilen SMZÖL'lar

Şekillerde görüldüğü üzere, keşfedilen sosyal önemli lokasyonlar büyükşehirlerde yoğunlaşmaktadır. Bunun gerekçesi olarak, bu şehirlerdeki insan yoğunluğunun diğer şehirlere oranla daha fazla olması görülebilir. Bunun yanında, bazı büyükşehirler, Konya, Gaziantep ve Adana gibi, insan yoğunluğuna sahip olsalar da önemli lokasyonlar arasında bulunmamaktadırlar. Ayrıca, Ankara'da sadece bir lokasyon sosyal önemli lokasyon olarak keşfedilmiştir. Bunun sebebi olarak ise, verinin toplandığı sırada, Twitter aktivitesi olarak bu şehirlerin aktif olmadığı veya Twitter kullanım istatistiklerinin yüksek olmadığı düşünülebilir. En fazla sayıda sosyal önemli lokasyon, İstanbul ve İzmir için bulunması

da bu şehirlerin sosyal medya aktivitesi olarak oldukça yüksek bir profil çizdikleri söylenebilir.

Tablo 7.2. Türkiye Twitter kullanıcıları için en yüksek 10 SÖL’lar

Sıra	Mekânsal	Mekân-Zamansal
1	Kıbrıs Şehitleri Cad., Alsancak, İzmir	Kıbrıs Şehitleri Cad., Alsancak, İzmir
2	Ragıp Gümüşpala Cad., Eminönü, İstanbul	Ege Üni., İzmir
3	Ege Üni., İzmir	Ragıp Gümüşpala Cad., Eminönü, İstanbul
4	Forum Bornova AVM, İzmir	Forum Bornova AVM, İzmir
5	Talatpaşa Blv., Cumhuriyet Blv., İzmir	Talatpaşa Blv., Cumhuriyet Blv., İzmir
6	Şişli Cevahir AVM, İstanbul	Şişli Cevahir AVM, İstanbul
7	Bağdat Cad., Göztepe, İstanbul	Bağdat Cad., Göztepe, İstanbul
8	Marmara Park AVM, İstanbul	Dokuz Eylül Üni., İzmir
9	Dokuz Eylül Üni., İzmir	Marmara Park AVM, İstanbul
10	Sabiha Gökçen Havalimanı, İstanbul	Boğa Heykeli, Kadıköy, İstanbul

Tablo 7.2’de, Türkiye Twitter kullanıcıları için keşfedilen ilk 10 SÖL verilmiştir. Dikkat çekici bir nokta, İstanbul ve İzmir’in diğer şehirleri domine etmeleridir. İzmir’de iki üniversitenin sosyal önemli lokasyon olarak keşfedilmesi de bu şehirdeki sosyal medya kullanım aktivitesinin nereden kaynaklandığının bir göstergesi olarak görülebilmektedir. Ayrıca, İstanbul için keşfedilen sosyal önemli lokasyonlar, Bölüm 5 ve 6’da keşfedilen lokasyonlarla uyumludur. Bu durum da keşfedilen SÖL’ların doğru olduğunu göstermektedir.

Ayrıca, Tablo 7.2’deki mekânsal ve mekân-zamansal SÖL’lar incelendiği zaman, ufak değişikliklerle birlikte, neredeyse aynı lokasyonların keşfedildiği görülmektedir. Listeler arasındaki tek fark 10 numaralı lokasyonlardır. Diğer lokasyonlarda da yer değişimleri meydana gelmiştir.

7.4.Katkılar ve Kapsam

Bu kısımda bu bölümde yapılan çalışmanın katkıları ve kapsamından bahsedilmiştir.

7.4.1. Katkılar

Bu bölümde yapılan çalışmanın katkıları aşağıdaki şekilde listelenebilir:

- Büyük sosyal medya veri kümelerinden sosyal önemli lokasyonlar keşfi için bulut bilişim sistemleri kullanılmıştır.
- Bölüm 5 ve 6’da önerilen SS-ILM, TF-SSTIL Miner ve SF-SSTIL Miner algoritmalarının bulut tabanlı versiyonları önerilmiştir.
- Önerilen algoritmalar, Hadoop MapReduce programlama modelini kullanmaktadır.
- Önerilen algoritmaların performansları Microsoft Azure bulut bilişim sistemi ve Türkiye büyük sosyal medya veri kümesi üzerinde değerlendirilmiştir.
- Deneysel sonuçlar, algoritmaların başarılı bir şekilde büyük veri kümelerini ele alabildiği ve ölçeklenebilirlik sağladığını göstermektedir.

7.4.2. Kapsam

Bu bölümde yapılan çalışma, sosyal önemli lokasyonlar keşfi için bulut bilişim sistemlerinin kullanımına odaklanmıştır. Aşağıdaki konular bu çalışmanın kapsamı dışındadır:

- Keşfedilen lokasyonların anlamsal olarak isimlendirilmesi,
- Lokasyon etiketleme problemine yenilikçi çözümler,
- Bulut bilişim altyapılarının kurulumu,

8. BÖLÜM

BULGULAR

Bu bölümde, 5, 6 ve 7. Bölümlerde önerilen değerlendirme ölçütleri ve algoritmaların performansı değerlendirilmiş, elde edilen sonuçlar karşılaştırmalı olarak sunulmuştur. Bulut bilişim sistemlerinin kullanılmasının avantajlarına değinilmiş ve algoritmaların performanslarının değişimi analiz edilmiştir.

8.1.Mekânsal Analiz Değerlendirmesi

Mekânsal analizlerin yapıldığı Bölüm 5’te, İstanbul sosyal medya kullanıcısı olan 1000 kişi kullanılarak, İstanbul’da yaşayan kullanıcılar için mekânsal sosyal önemli lokasyonların keşfi yapılmıştır. Bu bölümde, lokasyonların kullanıcılar için önemli olup olmadığının tespiti için iki değerlendirme ölçütü önerilmiştir. Daha sonra, kullanıcı mekânsal önemli lokasyonlarından sosyal medya kullanıcı grubu için sosyal önemli lokasyonlar keşfedilmiştir. Bu aşamada da bir değerlendirme ölçütü önerilmiştir. Bu değerlendirme ölçütlerine dayalı olarak iki algoritma, Toy ve SS-ILM algoritmaları, önerilmiştir. Önerilen algoritmaların başarıları ise çeşitli deneylerle karşılaştırılmıştır. DeneySEL sonuçlar, SS-ILM algoritmasının hesaplama karmaşıklığı açısından Toy algoritmaya göre daha başarılı olduğunu göstermiştir. Ayrıca, *min_yoğunluk* eşik değerinin artışına karşılık SS-ILM algoritması işlem zamanını düşürebilirken, Toy algoritma işlem zamanını sabit tutmaktadır. Yapılan son deneyde ise, İstanbul sosyal medya kullanıcıları için keşfedilen sosyal önemli lokasyonlar değerlendirilmiştir. Elde edilen sonuçlar, algoritmaların İstanbul’daki sosyal olarak aktif ve yaşayan kişilerin yoğun olarak ziyaret ettikleri lokasyonları keşfettiklerini göstermektedir. Bu durumda geliştirilen metodun ve algoritmaların başarılı olduğu söylenebilir.

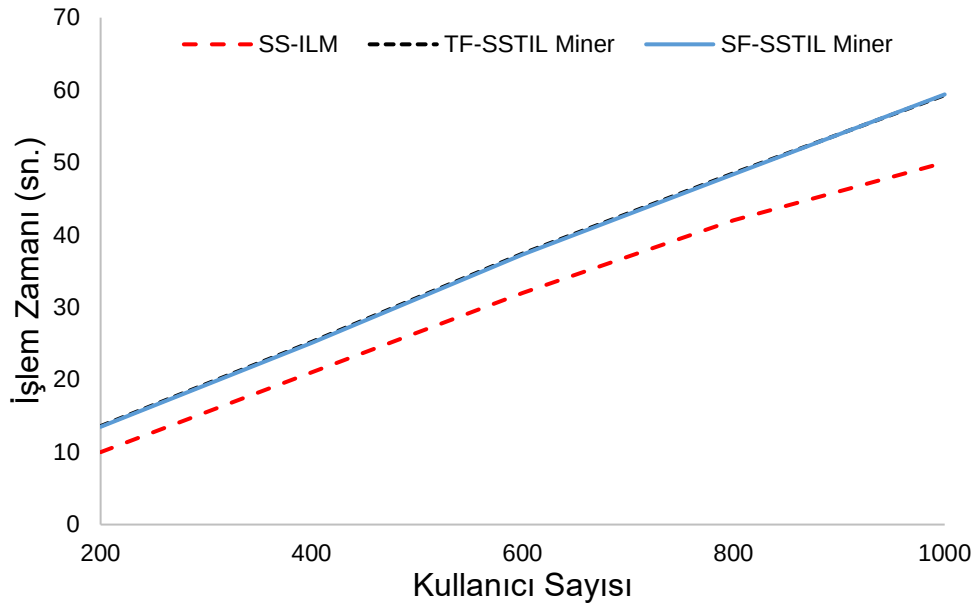
Yapılmış olan mekânsal analizler, 1000 kullanıcılık bir örnek üzerinde yapılmıştır. Bu kullanıcı grubu için SS-ILM algoritması yaklaşık 40-50 saniye arasında zaman harcamaktadır. İşlem zamanı olarak kabul edilebilir bir süre olarak görülse de kullanıcı sayısının artışı, işlem zamanını lineerden daha fazla artırmaya meyillidir. Bu nedenle, SS-ILM algoritmasının bulut bilişim sistemleri üzerinde çalıştırılması, daha fazla sayıda kullanıcının ölçeklenebilir işlem zamanları sunacağı düşünülmektedir.

8.2.Mekân-Zamansal Analiz Değerlendirmesi

Mekân-zamansal analizlerin yapıldığı Bölüm 6'da, Bölüm 5'te kullanılan veri kümesi kullanılmıştır. Sadece mekânsal analiz değil, aynı zamanda lokasyonların kullanıcı grubu için zamansal dağılımı da incelenmiştir. Mekânsal analizler Bölüm 5'te sunulmuş olan değerlendirme ölçütleri kullanılmış, zamansal analizler için ek değerlendirme ölçütleri önerilmiştir. Yapılan ek zamansal analizler, işlem zamanı olarak algoritmaların daha fazla sürede tamamlanmasına sebep olsa da, üretilen sonuçların zamansal derinliği bu dezavantajı ortadan kaldırmaktadır. Sosyo-mekân-zamansal önemli lokasyonların keşfi ile, hem mekânsal olarak kullanıcı grubu için önemli görülen lokasyonlar keşfedilmiş, hem de bu lokasyonların belirli zaman dilimleri boyunca önemli görülmeleri beklenmektedir. Bu durumda, bir lokasyonun kullanıcı grubunda mekân-zamansal önemli lokasyon olması, o lokasyonun uzun bir zaman dilimi için kullanıcı grubu için yaygın olarak ziyaret edilmesi gerekmektedir. Yapılan deneyler de işlem zamanı açısından önerilen TF-SSTIL Miner ve SF-SSTIL Miner algoritmaları değişmeli olarak tercih edilebilir olduklarını göstermektedir. Önerilen her iki algoritmanın da Toy algoritmaya göre hesaplama karmaşıklığı açısından daha başarılı olduğu görülmektedir. Yapılan son deneyde, İstanbul sosyal medya kullanıcı grubu için mekân-zamansal önemli lokasyonlar keşfedilmiştir. Keşfedilen lokasyonların büyük çoğunluğu Bölüm 5'te keşfedilen lokasyonlarla tutarlıdır. Her iki metot arasındaki değişiklikler de mekânsal önemli olarak keşfedilen lokasyonların aslında, zamansal yaygınlığı sağlayamamış olmasıdır. Yapılmış olan bu ek zamansal analizler algoritmaların işlem zamanlarına ek yük getirmektedir. Kullanıcı sayısının artışına karşılık SS-ILM ve TF-SSTIL Miner ve SF-SSTIL Miner algoritmaları arasındaki işlem zamanı artışı Şekil 8.1'de sunulmuştur.

Şekil 8.1'de görüldüğü üzere, SS-ILM algoritması TF-SSTIL Miner ve SF-SSTIL Miner algoritmalarına göre daha kısa sürede sonuç üretebilmektedir. Bu durumun ana sebebi,

SS-ILM algoritmasının zamansal analizler yapmaması, sadece mekânsal analiz ve kullanıcı yaygınlığına göre sonuç üretmesidir. Şekilde de görüldüğü üzere, yapılan ek zamansal analizler toplam işlem zamanına yaklaşık olarak 10 saniyelik artış yükü getirmektedir.



Şekil 8.1. Kullanıcı sayısının artışına karşılık algoritmaların işlem zamanlarının artışı

Mekân-zamansal analizlerde kullanılan veri kümesi, 1000 kullanıcıdan oluşmaktadır. Mekânsal analizlerdeki gibi, mekân-zamansal analizler de 55-65 saniye gibi kabul edilebilir sürelerde sonuç üretse de, veri kümesi büyüdüğü zaman işlem zamanının ölçeklenebilir olması gerekmektedir. Bu durumda, bulut bilişim sistemleri değerlendirilmesi gerekmektedir.

8.3.Bulut Bilişim Sistemlerinin Kullanımının Değerlendirmesi

Bölüm 5 ve 6’da geliştirilmiş olan mekânsal ve mekân-zamansal yaklaşımlar ve önerilen algoritmalar, büyük veri kümeleri söz konusu olduğu zaman işlem zamanını lineer olarak artırmaktadır. Bu durum da çok fazla sayıda kullanıcının bulunduğu veri kümelerinin başarılı bir şekilde analiz edilebilmesine imkân sağlayamamaktadır. Büyük veri analizi için başarılı bir şekilde kullanılan en güçlü alternatif olan bulut bilişim sistemleri de bu tez çalışması kapsamında değerlendirilmiş ve önerilmiş olan mekânsal ve mekân-zamansal algoritmalar bulut bilişim sistemlerine uyarlanmıştır. Veri kümesi olarak Türkiye Twitter kullanıcısı olan yaklaşık 27.000 kullanıcı verisi toplanmış ve bu kullanıcılardan yaklaşık 18.000 kullanıcı seçilerek bulut bilişim sistemleri üzerinde analiz

edilmiştir. Yapılan deneylerde, farklı bulut kaynakları üzerinde algoritmaların işlem zamanlarının analizi yapılmıştır. En optimum sonuçların, veri kümesi 50 parçaya ayrıldığı, 10 işçi kullanıldığı ve D13 V2 işçi boyutu kullanıldığı zaman elde edildiği gözlemlenmiştir. Ancak, edinilen deneyim, veri kümesinin artışı durumunda optimum değeri sağlayan değerlerin değişeceği üzerinedir. Bu bölümde yapılan son deneyde, Türkiye Twitter kullanıcılarının mekânsal ve mekân-zamansal sosyal önemli lokasyonlarının keşfi yapılmıştır. Keşfedilen sosyal önemli lokasyonların büyük çoğunluğu İstanbul'dadır. İstanbul'u İzmir yine yaygın bir lokasyonlar kümesiyle takip etmektedir. Diğer büyükşehirlerin bazılarında da sosyal önemli lokasyonlar keşfedilmiştir. İstanbul'da keşfedilen sosyal önemli lokasyonların yoğun olmasının sebebi, nüfus yoğunluğuyla açıklanabilir. Ancak, İzmir'deki yoğunluk, burada yaşayan kişilerin Twitter'ı oldukça aktif olarak kullandıkları sonucunu ortaya çıkarmaktadır. Diğer büyükşehirlerin de sosyal olarak aktif olmadıkları görülmektedir.

Bölüm 7'de kullanılan veri kümesi, Bölüm 5 ve 6'daki algoritmalar üzerinde çalıştırıldığı zaman yaklaşık 50 ve 58 dakika sürdüğü gözlemlenmiştir. Ancak, optimum süreler dikkate alındığı zaman bu algoritmaların bulut versiyonları yaklaşık 160 ve 190 saniye sürmektedir. Aradaki bu büyük işlem zamanı farkı da, sosyal medya büyük verisi üzerinde bulut bilişim sistemleri kullanılmasının ne kadar avantajlı olduğunu ortaya koymaktadır.

8.4.Tez Çalışmasının Literatüre Katkıları

Bu tez çalışması, oylama veya lokasyon türü gibi herhangi bir açıklayıcı bilgiye ihtiyaç duymadan, kullanıcıların sosyal medya aktivite geçmişlerinden sosyal önemli lokasyonları keşfetmeyi amaçlamıştır. Bu tez çalışmasının katkıları aşağıda listelenmiştir:

- Sosyal önemli lokasyonların keşfi için mekânsal ve mekân-zamansal iki yaklaşım önerilmiştir.
- Önerilen yaklaşımlar kullanıcı ve sosyal grup seviyelerinde sosyal önemli lokasyonları keşfetmektedir.
- Mekânsal sosyal önemli lokasyonların ölçülebilmesi için lokasyon yoğunluğu, ziyaret ömrü ve kullanıcı yaygınlığı değerlendirme ölçütleri önerilmiştir.

- Mekân-zamansal sosyal önemli lokasyonların ölçülebilmesi için zaman penceresi oluşumu, kullanıcı zamansal lokasyon önemi ve sosyo-mekân-zamansal lokasyon önemi değerlendirme ölçütleri önerilmiştir.
- Mekânsal sosyal önemli lokasyonların keşfi için SS-ILM algoritması mekân-zamansal sosyal önemli lokasyonların keşfi için TF-SSTIL Miner ve SF-SSTIL Miner algoritmaları önerilmiştir.
- Önerilen sosyal önemli lokasyonlar madenciliği algoritmaları bulut bilişim sistemlerine de uyarlanmıştır.
- Bulut tabanlı önerilen algoritmalar Microsoft Azure bulut bilişim ortamı ve Apache Hadoop yazılım çerçevesi üzerinde koşturulmuştur.
- Önerilen algoritmalar, İstanbul ve Türkiye Twitter kullanıcılarının büyük veri kümesi üzerinde deneysel olarak değerlendirilmiştir.
- Deneysel sonuçlar, algoritmaların başarılı bir şekilde sonuçları üretebildiğini ve bulut bilişim sistemlerinin kullanılmasının faydalarını ortaya koymaktadır.

9. BÖLÜM

TARTIŞMA, SONUÇ VE ÖNERİLER

9.1.Tartışma ve Sonuç

Bu tez çalışmasında, sosyal medya kullanıcılarının geçmişe dayalı sosyal medya aktivitelerini analiz ederek sosyal önemli lokasyonların keşfine yönelik teknikler geliştirilmiştir. Sosyal önemli lokasyonların keşfi, bir kullanıcı grubunun tercihlerini analiz etmek için gereklidir. Sosyal önemli lokasyonların ölçülebilmesi için lokasyon yoğunluğu, ziyaret ömrü, kullanıcı yaygınlığı, zaman penceresi oluşumu, kullanıcı zamansal lokasyon önemi ve sosyo-mekân-zamansal lokasyon önemi değerlendirme ölçütleri önerilmiştir. Bu değerlendirme ölçütlerini kullanan mekânsal ve mekân-zamansal sosyal önemli lokasyonlar keşfi algoritmaları, SS-ILM, TF-SSTIL Miner ve SF-SSTIL Miner algoritmaları, önerilmiştir. Önerilen algoritmaların performansı İstanbul Twitter kullanıcılarının gerçek veri kümesi üzerinde değerlendirilmiştir. Elde edilen sonuçlara göre önerilen algoritmalar Toy alternatiflerine ve literatürdeki karşılaştırılan algoritmalara karşı üstün gelmiştir. Ayrıca önerilen algoritmalar, İstanbul Türkiye’de yaşayan sosyal medya kullanıcılarına ait sosyal önemli lokasyonları ortaya çıkarmıştır. Önerilen algoritmalar kolaylıkla coğrafi ve zamansal bilgiye sahip olan herhangi bir sosyal medya aktivite verisetine de uyarlanabilir.

Bulut bilişim sistemleri, hesaplama görevlerini hesaplama kaynaklarına dağıtan ve bağımsız hesaplama kaynaklarının (kişisel bilgisayarlar, yerel sunucu sistemleri vb.) güçlü alternatifidir. Sosyal medya veri kümelerinin büyük veri özellikleriyle bu veriler bulut bilişim sistemleri için uygun veri kaynaklarındandır. Bu tez çalışmasındaki bulut bilişim sistemleriyle ilgili çalışmalar, bulut bilişim sistemlerini kullanarak mekânsal ve mekân-zamansal sosyal medya madenciliğine odaklanmaktadır. Özel olarak bu çalışmada, bulut bilişim sistemleri kullanılarak Türkiye sosyal medya kullanıcılarının sosyal önemli lokasyonlarını keşfetmektedir. Önceki bölümlerde önerilmiş olan mekânsal

ve mekân-zamansal sosyal önemli lokasyonlar keşfi algoritmaları Hadoop MapReduce programlama modeli ile modifiye edilmiş ve Türkiye Twitter kullanıcılarının büyük veri kümeleri üzerinde koşturulmuşlardır. Deneysel sonuçlar, bulut bilişim sistemlerinin kullanılmasının bağımsız hesaplama kaynaklarına göre çok daha hızlı sonuçları keşfedebildiklerini göstermektedir. Ayrıca, keşfedilen sosyal önemli lokasyonlar, şehir planlamacıları, yerel yönetimler ve kuruluşlara önemli katkılar sağlayabilecektir.

9.2.Öneriler

Gelecekte, sosyal önemli lokasyonların keşfinin daha verimli yapılacağı yeni algoritmalar geliştirilebilir. Odaklanılmış bir sosyal medya kullanıcı grubu, örneğin bir arkadaş grubu veya marka takipçi grubu, için sosyal önemli lokasyonların keşfinin yapılması algoritmaların performansını daha net ortaya çıkaracağı düşünülmektedir. Mekân-zamansal açıdan bakıldığı zaman da hangi lokasyonların yeni önem kazandığı ve hangilerinin önemsiz hale geldiğinin tespiti, mekân-zamansal sosyal medya madenciliğinin başka bir boyutudur. Ayrıca, zamansal analiz birbirini takip eden zaman dilimlerinin analiziyle genişletilebilir.

Bulut bilişim sistemleri ile ilgili gelecekteki çalışmalarda ise, sosyal medya büyük verisinin farklı problemlerinin bulut bilişim tabanlı olarak çözümlenmesi yapılabilir. Ayrıca, her bir şehir için keşfedilen sonuçlar, şehir bilişimi için farklı bilgiler sağlamaktadır. Sosyal önemli lokasyonlar keşfi çalışmalarının her bir şehir için ayrı ayrı analiz edilmesi ve şehirlerin sosyal önemli lokasyonları arasındaki ilişkinin analiz edilmesi bir diğer gelecekteki yapılabilecek çalışmalar arasındandır.

KAYNAKÇA

1. Mayer-Schonberger, V., Cukier, K., 2013. Big Data: A Revolution That Will Transform How We Live, Work, and Think. Houghton Mifflin Harcourt, Boston 242 pp.
2. Bello-Orgaz, G., Jung, J.J., Camacho, D., 2016. Social big data: Recent achievements and new challenges. **Information Fusion**, **28**(Supplement C): 45-59.
3. Zafarani, R., Abbasi, M.A., Liu, H., 2014. Social Media Mining: An Introduction. Cambridge University Press, Cambridge 332 pp.
4. Baloğlu, A., 2015. Sosyal Medya Madenciliği. Beta Basım Yayım Dağıtım A.Ş., 167 pp.
5. Gundecha, P., Liu, H. Mining Social Media: A Brief Introduction, pp. 1-17 *In* New Directions in Informatics, Optimization, Logistics, and Production,(Eds.
6. Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R.H., Konwinski, A., . . . Zaharia, M., 2009. Above the Clouds: A Berkeley View of Cloud Computing, University of California at Berkeley. 1-23.
7. Mell, P., Grance, T., 2011. The NIST Definition of Cloud Computing, National Institute of Standards and Technology. 1-7.
8. Buyya, R., Yeo, C.S., Venugopal, S., Broberg, J., Brandic, I., 2009. Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. **Future Generation Computer Systems**, **25**(6): 599-616.
9. Buyya, R., Broberg, J., Goscinski, A., 2011. Cloud Computing: Principles and Paradigms. 664 pp.
10. Dokuz, A.S., Celik, M., 2017. Bulut Bilişim Sistemlerinde Verinin Farklı Boyutları Üzerine Derleme. **Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi**, **6**(2): 316 - 338.
11. Hurwitz, J., Nugent, A., Halper, F., Kaufman, M., 2013. Big Data For Dummies. For Dummies, 336 pp.
12. Lam, C., 2010. Hadoop in Action. Manning Publications Co., 325 pp.
13. Hadoop, 2017. <http://hadoop.apache.org/>. (Web sayfası) (Erişim Tarihi: 15.11.2017).

14. Shvachko, K., Kuang, H., Radia, S., Chansler, R., 2010. The Hadoop Distributed File System. *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*. Conference 3-7 May 2010, 1-10.
15. Dean, J., Ghemawat, S., 2008. MapReduce: simplified data processing on large clusters. **Commun. ACM**, **51**(1): 107-113.
16. Fox, G., Gannon, D., 2013. Using Clouds for Technical Computing. IOS Press, 250 pp.
17. Khalifa, M.b., Díaz Redondo, R.P., Vilas, A.F., Rodríguez, S.S., 2017. Identifying urban crowds using geo-located Social media data: a Twitter experiment in New York City. **Journal of Intelligent Information Systems**, **48**(2): 287-308.
18. Documentations, T., 2017. <https://developer.twitter.com/en/docs>. (Web sayfası) (Erişim Tarihi: 15.11.2017).
19. Yamamoto, Y., 2017. Twitter4j Java Library. (Web sayfası: <http://twitter4j.org/en/index.html>) (Erişim Tarihi: 15.11.2017).
20. Kefalas, P., Symeonidis, P., Manolopoulos, Y., 2016. A Graph-Based Taxonomy of Recommendation Algorithms and Systems in LBSNs. **IEEE Transactions on Knowledge and Data Engineering**, **28**(3): 604-622.
21. Celik, M., 2015. Partial spatio-temporal co-occurrence pattern mining. **Knowledge and Information Systems**, **44**(1): 27-49.
22. Celik, M., Shekhar, S., Rogers, J.P., Shine, J.A., 2008. Mixed-Drove Spatiotemporal Co-Occurrence Pattern Mining. **IEEE Transactions on Knowledge and Data Engineering**, **20**(10): 1322-1335.
23. Dadaser-Celik, F., Celik, M., Dokuz, A.S., 2012. Associations between stream flow and climatic variables at Kizilirmak River Basin in Turkey. **Global Nest Journal**, **14**(3): 354-361.
24. Yu, W., 2016. Spatial co-location pattern mining for location-based services in road networks. **Expert Systems with Applications**, **46**(Supplement C): 324-335.
25. Hu, T., Sung, S.Y., 2005. Clustering spatial data with a hybrid EM approach. **Pattern Analysis and Applications**, **8**(1): 139-148.
26. Tung, A.K.H., Hou, J., Jiawei, H., 2001. Spatial clustering in the presence of obstacles. *Proceedings 17th International Conference on Data Engineering*. Conference 2001, 359-367.

27. Shekhar, S., Lu, C.-T., Zhang, P., 2001. Detecting graph-based spatial outliers: algorithms and applications (a summary of results), in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining* ACM: San Francisco, California. 371-376.
28. Kefalas, P., Symeonidis, P., 2015. Recommending Friends and Locations over a Heterogeneous Spatio-Temporal Graph, pp. 271-284 *In Model and Data Engineering: 5th International Conference, MEDI 2015, Rhodes, Greece, September 26-28, 2015, Proceedings*, (Eds L. Bellatreche and Y. Manolopoulos, Editors.), Springer International Publishing, Cham.
29. Vasuki, V., Natarajan, N., Lu, Z., Savas, B., Dhillon, I., 2011. Scalable Affiliation Recommendation using Auxiliary Networks. **ACM Trans. Intell. Syst. Technol.**, **3**(1): 1-20.
30. Weiler, M., Schmid, K.A., Mamoulis, N., Renz, M., 2015. Geo-Social Co-location Mining, in *Second International ACM Workshop on Managing and Mining Enriched Geo-Spatial Data* ACM: Melbourne, VIC, Australia. 19-24.
31. Liu, Y., Seah, H.S., 2015. Points of interest recommendation from GPS trajectories. **International Journal of Geographical Information Science**, **29**(6): 953-979.
32. Doytsher, Y., Galon, B., Kanza, Y., 2011. Storing routes in socio-spatial networks and supporting social-based route recommendation, in *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks* ACM: Chicago, Illinois. 49-56.
33. Levandoski, J.J., Sarwat, M., Eldawy, A., Mokbel, M.F., 2012. LARS: A Location-Aware Recommender System. *2012 IEEE 28th International Conference on Data Engineering*. Conference 1-5 April 2012, 450-461.
34. Bao, J., Zheng, Y., Mokbel, M.F., 2012. Location-based and preference-aware recommendation using sparse geo-social networking data, in *Proceedings of the 20th International Conference on Advances in Geographic Information Systems* ACM: Redondo Beach, California. 199-208.
35. Ying, J.J.-C., Lu, E.H.-C., Kuo, W.-N., Tseng, V.S., 2012. Urban point-of-interest recommendation by mining user check-in behaviors, in *Proceedings of the ACM SIGKDD International Workshop on Urban Computing* ACM: Beijing, China. 63-70.

36. Liu, B., Xiong, H. Point-of-Interest Recommendation in Location Based Social Networks with Topic and Location Awareness, pp. 396-404 *In Proceedings of the 2013 SIAM International Conference on Data Mining*,(Eds.
37. Zheng, Y., Zhang, L., Ma, Z., Xie, X., Ma, W.-Y., 2011. Recommending friends and locations based on individual location history. **ACM Trans. Web**, 5(1): 1-44.
38. Ying, J.J.-C., Lu, E.H.-C., Lee, W.-C., Weng, T.-C., Tseng, V.S., 2010. Mining user similarity from semantic trajectories, in *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks* ACM: San Jose, California. 19-26.
39. Hannon, J., Bennett, M., Smyth, B., 2010. Recommending twitter users to follow using content and collaborative filtering approaches, in *Proceedings of the fourth ACM conference on Recommender systems* ACM: Barcelona, Spain. 199-206.
40. Feng, W., Zhang, C., Zhang, W., Han, J., Wang, J., Aggarwal, C., Huang, J., 2015. STREAMCUBE: Hierarchical spatio-temporal hashtag clustering for event exploration over the Twitter stream. *2015 IEEE 31st International Conference on Data Engineering*. Conference 13-17 April 2015, 1561-1572.
41. Budak, C., Georgiou, T., Agrawal, D., Abbadi, A.E., 2013. GeoScope: online detection of geo-correlated information trends in social networks. **Proc. VLDB Endow.**, 7(4): 229-240.
42. Sattari, M., Manguoglu, M., Toroslu, I.H., Symeonidis, P., Senkul, P., Manolopoulos, Y., 2012. Geo-activity recommendations by using improved feature combination, in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing* ACM: Pittsburgh, Pennsylvania. 996-1003.
43. Cao, X., Cong, G., Jensen, C.S., 2010. Mining significant semantic locations from GPS data. **Proc. VLDB Endow.**, 3(1-2): 1009-1020.
44. Ying, J.J.-C., Lee, W.-C., Weng, T.-C., Tseng, V.S., 2011. Semantic trajectory mining for location prediction, in *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* ACM: Chicago, Illinois. 34-43.
45. Zhou, C., Bhatnagar, N., Shekhar, S., Terveen, L., 2007. Mining Personally Important Places from GPS Tracks. *2007 IEEE 23rd International Conference on Data Engineering Workshop*. Conference 17-20 April 2007, 517-526.

46. Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M., Rowland, J., Varshavsky, A., 2011. Identifying Important Places in People's Lives from Cellular Network Data, pp. 133-151 *In Pervasive Computing: 9th International Conference, Pervasive 2011, San Francisco, USA, June 12-15, 2011. Proceedings*, (Eds K. Lyons, J. Hightower, and E.M. Huang, Editors.), Springer Berlin Heidelberg, Berlin, Heidelberg.
47. Pavan, M., Mizzaro, S., Scagnetto, I., Beggiato, A., 2015. Finding Important Locations: A Feature-Based Approach. *2015 16th IEEE International Conference on Mobile Data Management*. Conference 15-18 June 2015, 110-115.
48. Chen, Y., Jiang, K., Zheng, Y., Li, C., Yu, N., 2009. Trajectory simplification method for location-based social networking services, in *Proceedings of the 2009 International Workshop on Location Based Social Networks ACM: Seattle, Washington*. 33-40.
49. Cai, G., Hio, C., Bermingham, L., Lee, K., Lee, I., 2014. Sequential pattern mining of geo-tagged photos with an arbitrary regions-of-interest detection method. **Expert Systems with Applications**, **41**(7): 3514-3526.
50. Cenamor, I., de la Rosa, T., Núñez, S., Borrajo, D., 2017. Planning for tourism routes using social networks. **Expert Systems with Applications**, **69**(Supplement C): 1-9.
51. Chen, C.-C., Chiang, M.-F., Peng, W.-C., 2016. Mining and clustering mobility evolution patterns from social media for urban informatics. **Knowledge and Information Systems**, **47**(2): 381-403.
52. Cho, E., Myers, S.A., Leskovec, J., 2011. Friendship and mobility: user movement in location-based social networks, in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining ACM: San Diego, California, USA*. 1082-1090.
53. Liu, J., Huang, Z., Chen, L., Shen, H.T., Yan, Z., 2012. Discovering areas of interest with geo-tagged images and check-ins, in *Proceedings of the 20th ACM international conference on Multimedia ACM: Nara, Japan*. 589-598.
54. Zheng, Y., Wang, L., Zhang, R., Xie, X., Ma, W.Y., 2008. GeoLife: Managing and Understanding Your Past Life over Maps. *The Ninth International Conference on Mobile Data Management (mdm 2008)*. Conference 27-30 April 2008, 211-212.

55. Zheng, Y., Xie, X., Ma, W.-y., 2010. GeoLife: A Collaborative Social Networking Service among User, Location and Trajectory. **IEEE Data Engineering Bulletin**, **33**: 32-39.
56. Yin, Z., Cao, L., Han, J., Luo, J., Huang, T., 2011. Diversified Trajectory Pattern Ranking in Geo-Tagged Social Media, pp. 980-991 *In Proceedings of the 2011 SIAM International Conference on Data Mining*,(Eds.
57. Gaete-Villegas, J., Cha, M., Lee, D., Ko, I.-Y., 2012. TraMSNET: a mobile social network application for tourism, in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing* ACM: Pittsburgh, Pennsylvania. 1004-1011.
58. Symeonidis, P., Papadimitriou, A., Manolopoulos, Y., Senkul, P., Toroslu, I., 2011. Geo-social recommendations based on incremental tensor reduction and local path traversal, in *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks* ACM: Chicago, Illinois. 89-96.
59. Guy, I., Zwerdling, N., Ronen, I., Carmel, D., Uziel, E., 2010. Social media recommendation based on people and tags, in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* ACM: Geneva, Switzerland. 194-201.
60. Zheng, Y., Zhang, L., Xie, X., Ma, W.-Y., 2009. Mining interesting locations and travel sequences from GPS trajectories, in *Proceedings of the 18th international conference on World wide web* ACM: Madrid, Spain. 791-800.
61. Khetarpaul, S., Chauhan, R., Gupta, S.K., Subramaniam, L.V., Nambiar, U., 2011. Mining GPS data to determine interesting locations, in *Proceedings of the 8th International Workshop on Information Integration on the Web: in conjunction with WWW 2011* ACM: Hyderabad, India. 1-6.
62. Chen, Z., Shen, H.T., Zhou, X., 2011. Discovering popular routes from trajectories. *2011 IEEE 27th International Conference on Data Engineering*. Conference 11-16 April 2011, 900-911.
63. Wei, L.-Y., Zheng, Y., Peng, W.-C., 2012. Constructing popular routes from uncertain trajectories, in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* ACM: Beijing, China. 195-203.
64. Ferrari, L., Rosi, A., Mamei, M., Zambonelli, F., 2011. Extracting urban patterns from location-based social networks, in *Proceedings of the 3rd ACM*

- SIGSPATIAL International Workshop on Location-Based Social Networks
ACM: Chicago, Illinois. 9-16.
65. Cranshaw, J., Schwartz, R., Hong, J.I., Sadeh, N., 2012. The Livehoods project: Utilizing social media to understand the dynamics of a city. *ICWSM 2012 - Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*. Conference, 58-65.
 66. Ying, J.J.-C., Lee, W.-C., Tseng, V.S., 2014. Mining geographic-temporal-semantic patterns in trajectories for location prediction. **ACM Trans. Intell. Syst. Technol.**, 5(1): 1-33.
 67. Hu, B., Jamali, M., Ester, M., 2013. Spatio-Temporal Topic Modeling in Mobile Social Media for Location Recommendation. *2013 IEEE 13th International Conference on Data Mining*. Conference 7-10 Dec. 2013, 1073-1078.
 68. Yuan, Q., Cong, G., Sun, A., 2014. Graph-based Point-of-interest Recommendation with Geographical and Temporal Influences, in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* ACM: Shanghai, China. 659-668.
 69. Kefalas, P., Manolopoulos, Y., 2017. A time-aware spatio-textual recommender system. **Expert Systems with Applications**, 78(Supplement C): 396-406.
 70. Yuan, Q., Zhang, W., Zhang, C., Geng, X., Cong, G., Han, J., 2017. PRED: Periodic Region Detection for Mobility Modeling of Social Media Users, in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining* ACM: Cambridge, United Kingdom. 263-272.
 71. Huang, W., Li, S., 2016. Understanding human activity patterns based on space-time-semantics. **ISPRS Journal of Photogrammetry and Remote Sensing**, 121(Supplement C): 1-10.
 72. Çelikten, E., Falher, G.L., Mathioudakis, M., 2017. Modeling Urban Behavior by Mining Geotagged Social Data. **IEEE Transactions on Big Data**, 3(2): 220-233.
 73. Assem, H., Xu, L., Buda, T.S., Sullivan, D.O., 2016. Spatio-Temporal Clustering Approach for Detecting Functional Regions in Cities. *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*. Conference 6-8 Nov. 2016, 370-377.
 74. Jiang, S., Joseph Ferreira, J., Gonzalez, M.C., 2012. Discovering urban spatial-temporal structure from human activity patterns, in *Proceedings of the ACM*

- SIGKDD International Workshop on Urban Computing ACM: Beijing, China. 95-102.
75. Zhang, C., Zhang, K., Yuan, Q., Zhang, L., Hanratty, T., Han, J., 2016. GMove: Group-Level Mobility Modeling Using Geo-Tagged Social Media, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining ACM: San Francisco, California, USA. 1305-1314.
 76. Khetarpaul, S., Gupta, S.K., Subramaniam, L.V., 2016. Spatiotemporal social (STS) data model: correlating social networks and spatiotemporal data. **Social Network Analysis and Mining**, 6(1): 81.
 77. Zarezade, A., Jafarzadeh, S., Rabiee, H.R., 2016. Spatio-Temporal Modeling of Users' Check-ins in Location-Based Social Networks. **ACM Transactions on Knowledge Discovery from Data**, 9(4): 1-17.
 78. Wei, L.-Y., Zheng, Y., Peng, W.-C., 2015. Mining Popular Routes from Social Media, pp. 93-116 *In* Multimedia Data Mining and Analytics: Disruptive Innovation,(Eds A.K. Baughman, et al., Editors.), Springer International Publishing, Cham.
 79. Comito, C., Falcone, D., Talia, D., 2016. Mining human mobility patterns from social geo-tagged data. **Pervasive and Mobile Computing**, 33(Supplement C): 91-107.
 80. Xu, Z., Zhang, H., Sugumaran, V., Choo, K.-K.R., Mei, L., Zhu, Y., 2016. Participatory sensing-based semantic and spatial analysis of urban emergency events using mobile social media. **EURASIP Journal on Wireless Communications and Networking**, 2016(1): 44.
 81. Qiu, M., Ming, Z., Wang, J., Yang, L.T., Xiang, Y., 2014. Enabling Cloud Computing in Emergency Management Systems. **IEEE Cloud Computing**, 1(4): 60-67.
 82. Huang, Q., Cervone, G., Zhang, G., 2017. A cloud-enabled automatic disaster analysis system of multi-sourced data streams: An example synthesizing social media, remote sensing and Wikipedia data. **Computers, Environment and Urban Systems**, 66(Supplement C): 23-37.
 83. Rogstadius, J., Vukovic, M., Teixeira, C.A., Kostakos, V., Karapanos, E., Laredo, J.A., 2013. CrisisTracker: Crowdsourced social media curation for disaster awareness. **IBM Journal of Research and Development**, 57(5): 4:1-4:13.

84. Vakali, A., Giatsoglou, M., Antaris, S., 2012. Social networking trends and dynamics detection via a cloud-based framework design, in Proceedings of the 21st International Conference on World Wide Web ACM: Lyon, France. 1213-1220.
85. Moise, I., 2016. The technical hashtag in Twitter data: A hadoop experience. *2016 IEEE International Conference on Big Data (Big Data)*. Conference 5-8 Dec. 2016, 3519-3528.
86. Noordhuis, P., Heijkoop, M., Lazovik, A., 2010. Mining Twitter in the Cloud: A Case Study. *2010 IEEE 3rd International Conference on Cloud Computing*. Conference 5-10 July 2010, 107-114.
87. Ching, A., Edunov, S., Kabiljo, M., Logothetis, D., Muthukrishnan, S., 2015. One trillion edges: graph processing at Facebook-scale. **Proc. VLDB Endow.**, 8(12): 1804-1815.
88. Jin, S., Lin, W., Yin, H., Yang, S., Li, A., Deng, B., 2015. Community structure mining in big data social media networks with MapReduce. **Cluster Computing**, 18(3): 999-1010.
89. Gupta, P., Goel, A., Lin, J., Sharma, A., Wang, D., Zadeh, R., 2013. WTF: the who to follow service at Twitter, in Proceedings of the 22nd international conference on World Wide Web ACM: Rio de Janeiro, Brazil. 505-514.
90. Thomas, K., Grier, C., Ma, J., Paxson, V., Song, D., 2011. Design and Evaluation of a Real-Time URL Spam Filtering Service. *2011 IEEE Symposium on Security and Privacy*. Conference 22-25 May 2011, 447-462.
91. Khalid, O., Khan, M.U.S., Khan, S.U., Zomaya, A.Y., 2014. OmniSuggest: A Ubiquitous Cloud-Based Context-Aware Recommendation System for Mobile Social Networks. **IEEE Transactions on Services Computing**, 7(3): 401-414.
92. Wang, F., Meng, X., Zhang, Y., Zhang, C., 2016. Mining user preferences of new locations on location-based social networks: a multidimensional cloud model approach. **Wireless Networks**.
93. Kosmides, P., Demestichas, K., Adamopoulou, E., Remoundou, C., Loumiotis, I., Theologou, M., Anagnostou, M., 2016. Providing recommendations on location-based social networks. **Journal of Ambient Intelligence and Humanized Computing**, 7(4): 567-578.

94. Cesario, E., Iannazzo, A.R., Marozzo, F., Morello, F., Riotta, G., Spada, A., . . . Trunfio, P., 2016. Analyzing social media data to discover mobility patterns at EXPO 2015: Methodology and results. *2016 International Conference on High Performance Computing & Simulation (HPCS)*. Conference 18-22 July 2016, 230-237.
95. Zhou, X., Xu, C., Kimmons, B., 2015. Detecting tourism destinations using scalable geospatial analysis based on cloud computing platform. **Computers, Environment and Urban Systems**, **54**(Supplement C): 144-153.
96. Cao, G., Wang, S., Hwang, M., Padmanabhan, A., Zhang, Z., Soltani, K., 2015. A scalable framework for spatiotemporal analysis of location-based social media data. **Computers, Environment and Urban Systems**, **51**(Supplement C): 70-82.
97. Benevenuto, F., Magno, G., Rodrigues, T., Almeida, V., 2010. Detecting spammers on twitter, in Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS): Washington, USA. 1-9.
98. Zheng, X., Zeng, Z., Chen, Z., Yu, Y., Rong, C., 2015. Detecting spammers on social networks. **Neurocomputing**, **159**(Supplement C): 27-34.
99. Yang, D.-N., Shen, C.-Y., Lee, W.-C., Chen, M.-S., 2012. On socio-spatial group query for location-based social networks, in Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining ACM: Beijing, China. 949-957.
100. Liu, W., Sun, W., Chen, C., Huang, Y., Jing, Y., Chen, K., 2012. Circle of Friend Query in Geo-Social Networks, pp. 126-137 *In Database Systems for Advanced Applications: 17th International Conference, DASFAA 2012, Busan, South Korea, April 15-19, 2012, Proceedings, Part II*, (Eds S.-g. Lee, et al., Editors.), Springer Berlin Heidelberg, Berlin, Heidelberg.
101. Azure, M., 2017. Microsoft Azure Platform. (Web sayfası: <https://azure.microsoft.com/tr-tr/>) (Erişim Tarihi: 15.11.2017).

ÖZGEÇMİŞ

Adı Soyadı: Ahmet Şakir Dokuz

Uyruğu: Türkiye Cumhuriyeti (T.C.)

Doğum Tarihi ve Yeri: 15.08.1988 - Yozgat

Medeni Durumu: Evli

Tel: 507 533 0256

E-posta: adokuz@ohu.edu.tr, dokuz.a.sakir@gmail.com

Adres: Niğde Ömer Halisdemir Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Merkez/NİĞDE

I. Eğitim Durumu:

2010 – 2013: Erciyes Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Bölümü, Yüksek Lisans Programı (**Derece: 3.88 / 4.00**)

2006 – 2010: Erciyes Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Lisans Programı (**Derece: 2.96 / 4.00**)

2002 – 2006: Nuh Mehmet Baldöktü Anadolu Lisesi / Kayseri (**Derece: 4.96 / 5.00**)

II. İş Tecrübeleri

2011 – Halen: Niğde Ömer Halisdemir Üniversitesi Bilgisayar Mühendisliği Bölümü, Araştırma Görevlisi

2010 – 2013: TÜBİTAK Proje Asistanı (Veri Madenciliği)

III. Tez Çalışması Kapsamındaki Akademik Çalışmalar

SCI, SCI-Expanded, SSCI veya AHCI Kapsamındaki Dergi Yayınları:

1. M. Çelik, A. Ş. Dokuz, "Discovering Socio-Spatio-Temporal Important Locations of Social Media Users", Journal of Computational Science, 22, 85-98, 2017.

2. **A. Ş. Dokuz**, M. Çelik, "Discovering Socially Important Locations of Social Media Users", *Expert Systems with Applications*, 86, 113-124, 2017.

Diğer Dergi Yayınları:

1. **A. Ş. Dokuz**, M. Çelik, "Bulut Bilişim Sistemlerinde Verinin Farklı Boyutları Üzerine Derleme", *Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi*, 6 (2), 316-338, 2017.
2. **A. Ş. Dokuz**, M. Çelik, "Fast SS-ILM: A Computationally Efficient Algorithm to Discover Socially Important Locations", *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4 (4/W4), 197-202, 2017.
3. M. Çelik, **A. Ş. Dokuz**, "Daily and Hourly Mood Pattern Discovery of Turkish Twitter Users", *Global Journal of Computer Sciences*, 5 (2), 90-98, 2015.

Bildiriler:

1. **A. Ş. Dokuz**, M. Çelik, "Bulut Bilişim Sistemleri Kullanarak Sosyal Önemli Lokasyonların Madenciliği", 1. Ulusal Bulut Bilişim ve Büyük Veri Sempozyumu, pp. 31-36, Antalya, Türkiye, 19.10.2017.
2. **A. Ş. Dokuz**, M. Çelik, "Fast SS-ILM: A Computationally Efficient Algorithm to Discover Socially Important Locations", 4th International Workshop on Geoinformation Science, pp. 1-6, Karabük, Türkiye, 14.10.2017.
3. M. Çelik, **A. Ş. Dokuz**, "Daily and Hourly Mood Pattern Discovery of Turkish Twitter Users", 1st World Conference on Big Data, pp. 7-7, İstanbul, Türkiye, 19.11.2015.
4. M. Çelik, **A. Ş. Dokuz**, "Yeni Bir Hesaplama ve Depolama Yaklaşımı: Bulut Bilişim", II. Uluslararası Kop Bölgesel Kalkınma Sempozyumu, pp. -, Niğde, Türkiye, 23.10.2014.

IV. Tez Konusu Dışındaki Diğer Akademik Çalışmalar

SCI, SCI-Expanded, SSCI veya AHCI Kapsamındaki Dergi Yayınları:

1. M. Çelik, F. Dadaşer Çelik ve **A. Ş. Dokuz**, Discovery of Hydro-meteorological Patterns, *Turkish Journal Of Electrical Engineering & Computer Sciences* 22 (4), 840-857, 2014.
2. F. Dadaşer Çelik, M. Çelik ve **A. Ş. Dokuz**, Associations Between Stream Flow and Climatic Variables at Kizilirmak River Basin in Turkey, *Global NEST Journal*, 14 (3), 354-361, 2012.

Bildiriler:

1. **A. Ş. Dokuz**, M. Çelik, Filiz Dadaşer Çelik, "Discovering Emerging or Vanishing Association Patterns in Hydroclimatic Datasets", International Conference on Information Complexity and Statistical Modeling in High Dimensions with Applications (IC-SMHD-2016), pp. 42, Nevşehir, Türkiye, 18.05.2016.
2. **A. Ş. Dokuz**, K. S. Erduran, "Üniversiteler için Akademik Performans Değerlendirme Sistemi: Niğde Üniversitesi Örneği", II. Uluslararası Kop Bölgesel Kalkınma Sempozyumu, pp. -, Niğde, Türkiye, 23.10.2014.
3. M. Çelik, F. Dadaşer-Çelik, A. Ü. Kömüşçü, M. Arslan ve **A. Ş. Dokuz**, "Yağış ve İklim İndeksleri Arasındaki Sıralı Birliktelik Örüntülerinin Tespit Edilmesi", *2012 ASYU - Akıllı Sistemlerde Yenilikler ve Uygulamaları*, pp. 134-138, 3-4 Temmuz 2012 Trabzon Türkiye.
4. F. Dadaşer Çelik, M. Çelik ve **A. Ş. Dokuz**, "Associations Between Stream Flow and Climatic Variables at Kizilirmak River Basin in Turkey", in *2011 CEST - Proceedings of the 12th International Conference on Environmental Science and Technology*, pp. 367-374, 8-10 September 2011, Rhodes Greece.
5. M. Çelik, F. Dadaşer Çelik ve **A. Ş. Dokuz**, "Anomaly Detection in Temperature Data Using DBSCAN Algorithm", in *2011 INISTA - International Symposium on Innovations in Intelligent Systems and Applications*, pp. 91-95, 15-18 June 2011, Istanbul Turkey.