Bilgi Hasatlaması Yöntemleri ve Kişisel Bilgi Hasatlaması

Celal Turan Ulus, Eyüp Burak Ceyhan, Şeref Sağıroğlu

***Özet*—Bu makalede bilgi hasatlamasına dair genel bir bakış ortaya konmaktadır. Bununla birlikte bilgi hasatlamasının daha detaylı bir başlığı olan kişisel bilgi hasatlaması ile ilişkilendirme yapılmıştır. Bilgi hasatlaması bilgi çıkarma ve bilgi çekme olmak üzere iki temel başlıkta incelenmiştir. Bilgi çıkarma işleminde sonuç belli bir sorgu yardımıyla oluşmaktayken bilgi çekme işleminde hedeflenen belli alanlar doğrudan ya da belli örüntüler ile çekilmektedir. Bir sonraki bölümde, yapılan hasatlama çalışmaları olarak bilgi hasatlamasında yapılan bazı çalışmalardan detaylıca bahsedilmiştir. Bilgi hasatlaması terörist organizasyonların takip edilmesinden Wikipedia üzerinden bilgi çekilmesine kadar birçok alanda kullanılmaktadır. Altıncı bölümde bilgi hasatlaması internet ağı üzerinden kişisel bilgi çekilmesi açısından ele alınmıştır. Sosyal ağlarda ve arama motorlarında kişisel bilgilerin çekilmesi ile birlikte ilgili tehlikelerden bahsedilmiştir.**

***Anahtar Kelimeler*—Bilgi hasatlama, internet ağı hasatlaması, bilgi çekme, bilgi çıkarma, kişisel bilgi**

***Abstract*— This paper provides an overview of information harvesting. Besides the overview, information harvesting is associated to personal information harvestiong. Information harvesting is reviewed under two main topics: information retrieval and information extraction. The result occurs by the help of a query in the information extraction whereas the targeted fields of information extraction is harvested directly or by the help of patterns. In the following topic, some studies are mentioned as previous information studies in detail. Information harvesting is used for tracking terrorist organizations, Wikipedia harvesting, etc. In the sixth topic, information harvesting is handled according to personal information harvesting. Personal information harvesting from social networks, search engines and dangers related with them are mentioned.**

***Index Terms*— Information harvesting, web harvesting, information extraction, information retrieval, personal information**

I. GİRİŞ

İnternet ağının hızlı büyümesiyle birlikte yüksek miktarda veri, internet kullanıcıları tarafından erişilebilir halde bulunmaktadır. Düşük maliyet, yüksek erişilebilirlik ve özgürce yayın yapabilmek internet ağının karakteristikleri arasında yer almaktadır. Bu durum internet ağının popüleritesini arttırmaktadır. İnternet sayfaları aslında karmaşık metinlerden oluşmaktadır. Metinin ve multimedya bileşenlerinin yanında bağlantılar, HTML etiketleri, tanımlayıcı veri (meta-data) gibi özellikler barındırırlar. Birçok araştırmada internet sayfalarının metin bileşenleri internet ağı hasatlamada en önemli bilgiyi sağladığı varsayılır. Metin olmayan diğer bileşenlerin hasatlama performansını iyileştirdiği varsayılır [1].

İnternet ağı üzerinde veriler yapısal ve yapısal olmayan bir şekilde bulunur. Yapısal veriler alanları, başlıkları, etiketleri belli ve düzenli bir halde bulunur. Bunun sonucunda bilgisayar tarafından kolaylıkla kullanılabilirler. Fakat yapısal olmayan veriler belli bir düzeni olmayan verilerdir. Yapısal olmayan verileri çekmek, okumak ve işlemek daha zordur. Bilgisayardaki ve internetteki bilgilerin çoğu yapısal olmayan veridir [2].

İnternet ağı hasatlamasında, bir ya da iki siteden olan internet sayfaları içeriklerine göre daha önceden tanımlanmış kategorilere atanır. İnternet sayfaları düz metin dokümanlarından daha fazlası olduğu için internet ağı hasatlama metotları diğer içeriklerin niteliğini kullanmayı dikkate almalıdır. Yapılan bir araştırmada internet sayfalarını hasatlamak için düz metin haricinde internet sayfasının başlığı ve sayfa içinde bulunan bağlantıyı belirten sözcük dikkate alınmıştır [1].

Bilgi hasatlama işleminin aşamalarından biri olan bilgi çekme işleminde hasatlanan her bir metin topluluğu genellikle doküman olarak adlandırılır. Dokümanların kendine özgü yapıları vardır. Bir yazılım aracı bu yapıyı belli format işaretlerine ve anahtar sözcüklere göre işaretleyebilir. Fakat tüm bu durumlarda bulunan yapı ilgili kitabın anlamsal içeriğini değil organizasyonel yapısını gösterecektir. Yazılım aracı "bölüm 1", "şekil 1" gibi alanları kolaylıkla bulabilir. Fakat "bilgi çekme" ile ilgili bir başlığı bulmak çok daha zor ve daha belirsiz bir problemdir [3].

Bilgi hasatlama işleminin aşamalarından diğeri bilgi çekme işlemidir. Bilgi çıkarmada belli bir konuda ya da kişi hakkında gereken bilgi doküman içinden çıkartılır. Konu ya da gereken bilgi kullanıcı tarafından belirlenen bir sorgu ile çıkartılır. Belirlenen sorgu tarafından karşılanan dokümanlar kullanıcı tarafından konu ile ilgili, karşılanmayanlar ise ilgisiz olarak nitelenir. Bir bilgi çıkarma motoru dokümanı sınıflandırmak için sorguyu kullanabilir. Sorgu sınıflandırma kriterlerini karşılayan dokümanları sonuç olarak döner [3].

Bu makalenin bundan sonraki bölümünde bilgi hasatlamasından genel itibariyle bahsedilmiştir. Bilgi

Celal Turan Ulus, TÜBİTAK, turanulus@gmail.com
Eyüp Burak Ceyhan, Gazi Üniversitesi, ebceyhan@gazi.edu.tr
Prof. Dr. Şeref Sağıroğlu, Gazi Üniversitesi, ss@gazi.edu.tr

hasatlamasının yöntemlerinden biri olan bilgi çıkarması işlemlerinden üçüncü bölümde bahsedilmiştir. Dördüncü bölümde yine bilgi hasatlamasının diğer bir yöntemi olan bilgi çıkarması ele alınmıştır. Beşinci bölümde bilgi hasatlamasında internet ağı üzerinde yapılan çalışmalardan bahsedilmiş olup altıncı bölümde kişisel bilgi hasatlamasında güvenliğin önemi örneklerle vurgulanmıştır. Son bölümde ise bilgi hasatlamasının aşamaları hakkında özet bilgilerden bahsedilmiş olup kişisel bilgi hasatlamasının bilgi güvenliği ile ilgisi ortaya konmuştur.

II. BİLGİ HASATLAMASI

İnternetin yükselişi ile birlikte sosyal uygulamalar, bloglar, e-postalar ve çeşitli internet uygulamaları hayatımızın içinde yer almaktadır. Bu uygulamalarda kişisel bilgilerimiz bulunmaktadır ve iyi korunmayan bilgiler risk oluşturmaktadır. Bu bilgiler üzerinde bilgi hasatlaması yapılarak kullanıcıların bilgileri tahmin edilmekte ve ele geçirilmektedir.

Birçok sosyal ağ uygulamaları kişilerin ilişkileri hakkında ilginç bilgileri açığa çıkarmaktadır. Örneğin blog yorumcularının gönderilerini ya da yer imi benzerliklerini analiz edilerek insanlar arasındaki bağlantılar ortaya çıkarılabilir. Birçok kaynaktan bu bilgilerin çıkarımını yaparak ve birleştirerek kişisel ve kurumsal sosyal ağların kapsamlı bir resmi ortaya çıkarılabilir. İki kişinin bir makalede yazarlığı varsa aynı zamanda bir sosyal ağda bağlantısının bulunduğunu birçok çalışma göstermektedir. Yeni çalışmalarda bu iki kişinin bağlantısına kanıt olarak e-posta ve internet sayfaları olarak da gösterilmektedir [4].

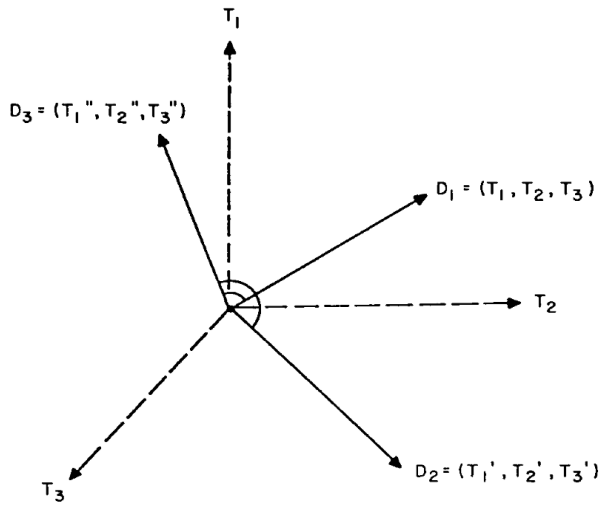
Genellikle bilgi hasatlama teknikleri internet ağı hasatlamasında da kullanılır. İnternet ağı hasatlaması bilgi hasatlamanın genişletilmiş halidir. Bilgi hasatlama çevrimdışı bir işlem iken internet ağı hasatlaması çevrimiçi bir işlemdir. Bilgi Hasatlama verisi çevrimdışı olarak veri ambarında saklanır. İnternet ağı hasatlama verisi sunucu veri tabanında saklanır [5].

III. BİLGİ ÇIKARMA YÖNTEMLERİ

A. *Vektör Uzay Modeli*

Vektör uzay modeli (VSM) 1975'te Salton ve çalışma arkadaşları tarafından SMART sistemi için geliştirilmiştir. Vektör ağırlık modeli bilgi çıkarımı için kullanımının basitliğinden dolayı en çok kullanılan modellerden biridir. Bir VSM'de her bir doküman bir vektör olarak kabul edilir. Her boyut bir terimi ifade eder. Bir terim bir dokümanın içinde yer alıyorsa o terimin vektördeki değeri sıfır değildir [6].

Bir grup D_i dokümanından oluşan doküman uzayı olsun. Her biri bir veya birden fazla indeks T_j terimi ile tanımlansın. Terimler kendi önemlerine göre 0 ve 1 arasından ağırlıklandırılır. Tipik bir üç boyutlu vektör uzayı Şekil 1'de gösterilmiştir. t farklı terim indeksi varsa üç boyutlu örnek t boyuta kadar genişletilebilir [7].



Şekil.1 Doküman uzayının vektörel gösterimi [7]

Terim frekansı (tf) vektör uzay modeli ile birlikte en çok kullanılan yöntemlerden birisidir. Bir çeşit sözcüğün kaç defa bir doküman içinde geçtiğinin sayısıdır. Doküman frekansı (df) ise bir kelimeyi en az bir kere içeren doküman sayıdır. Terim frekansı 0 ve N arasında bir tam sayıdır. Doküman frekansı ise 0 ve D arasında bir tam sayıdır. Bilgi çıkarımında doküman frekansları ters doküman frekanslarına çevrilir. Ters doküman frekansı (IDF) terim ağırlıklandırmada önemli bir rol oynar [8]. IDF(t) t terimini içeren dokümanın sahip olduğu bilgi biti sayısı olarak yorumlanabilir.

$$\mathrm{IDF(t)} = -\log_2 \frac{\mathrm{df(t)}}{\mathrm{D}} \quad (1)$$

Bir dokümandaki her terime sayısal ağırlık belirlenebilir. Böylece ilgili terimin ilgili doküman içindeki yararlılığı sayısal olarak ölçülmüş olur. Yararlılıktan kastedilen durum belli dokümanı diğer dokümanlardan ne kadar ayırt edici olabildiğidir. Belli bir terim farklı dokümanlar içinde farklı ağırlıklara sahip olabilir. Çünkü bir terim, bir doküman için diğer dokümanlara olduğundan daha iyi bir tanımlayıcı ya da ayırt edici olabilir [3].

Doküman uzayında her doküman D, dokümanın içinde yer alan terimlerin ağırlıkları ile tanımlanır. Terim uzayında her doküman bir boyuta karşılık gelir. Terim uzayında vektör bir terimdir. Bir terimin koordinatları, içinde geçtiği dokümanla ilişkili ağırlığı olarak yorumlanır [3]. Burada önemli bir ayrıntı terim ağırlığının nasıl bulunacağıdır. Terim ağırlığını bulurken en çok kullanılan yöntem aşağıdadır.

$$w = tf*IDF \quad (2)$$

Bir terimin bir dokümanın içindeki ağırlığı bulunurken (2) kullanılır. Dokümanın içindeki tf terim frekansını göstermektedir [3]. İki dokümanın benzerliğini bulurken (3) kullanılır. Bu eşitliğe kosinüs benzerliği denir. İki doküman arasındaki kosinüs açısına göre benzerlikleri belirler [9].

$$\mathrm{Sim}(v_1, v_2) = \frac{\sum_{i=1}^{n} v_1 . v_2}{\sqrt{\sum_{i=1}^{n} {v_{1i}}^2} \sqrt{\sum_{i=1}^{n} {v_{2i}}^2}} \quad (3)$$

v metin vektörüdür. v_{1i} , v_1 vektöründe bulunan i. sözcüğün tf-idf ağırlığıdır.

B. İkili İşlem Modeli

İkili işlem modelinde sorgular oluşturulur. Sorgu, ikili işlem operatörlerine dayanarak oluşturulur. Genel bir ikili işlem sorgusu AND, OR ve NOT operatörlerinden oluşur. Örneğin t_1 ve t_2 terimlerini D_1 dokümanı içeriyorsa "t_1 AND t_2" sorgusu D_1 dokümanı tarafından karşılanır. Benzer şekilde t_1 ve t_2 terimlerinden birini D_1 dokümanı içeriyorsa "t_1 OR t_2" sorgusu karşılanır. "t_1 AND NOT t_2" sorgusunda ise t1 terimi varsa ve t_2 terimi D_1 dokümanında yoksa D_1 dokümanı tarafından sorgu karşılanmış olur. Daha karmaşık ikili işlem sorguları oluşturulup bunların sonuçları ilgili doküman açısından sorgulanabilir. Klasik bir ikili işlem modeli sonuç olarak true ya da false değeri döner. Dolayısıyla bir doküman üzerinde işletilen bir ikili işlem sorgusu sonucunda o doküman sorgu içindeki terimle ya alakalı ya da alakasız bir doküman olur. Dokümanlar arasında herhangi bir sıralama yapılmaz. Klasik bir ikili işlem modelinde terim ağırlıkları kullanılmaz. Bir terimin ağırlığı ya birdir (terim vardır) ya da sıfırdır (terim yoktur) [3].

IV. BİLGİ ÇEKME YÖNTEMLERİ

A. İsimlendirilmiş Varlık Tanımlaması

NER isimlendirilmiş varlıkların önceden tanımlanmış türlerinin tespit edilmesi ve sınıflandırılması problemidir. İsimlendirilmiş varlıklara kuruluşlar (Dünya Sağlık Örgütü), kişiler (Muammer Kaddafi), yer isimleri (Baltık Denizi) örnek olarak verilebilir. NER metinden tespit edilen varlıklar hakkında tanımlayıcı bilgi çekilebilir. Örneğin kişi durumu için ünvan, mevki, milliyet, cinsiyet ve kişinin diğer nitelikleri hakkında bilgi çekilebilir [10].

NER'in başarısı, standart varlıklar için %95'e ulaşmaktadır [11]. NER için kullanılan metotlar farklı boyutlara göre sınıflandırılabilirler. Bu boyutlardan birisi manuel-otomatik zıtlığıdır. Bazı yöntemler kaynakları manuel olarak kullanırken diğerleri işaretlenmiş eğitim verisinden otomatik olarak bir model üretmek için öğrenme algoritmalarını kullanır. Diğer boyutta ise varlık modelinin özelliği belirtilir: Bazı modeller sembolik iken diğerleri sayısaldır. Sembolik modellerde kaynaklar açık ve anlaşılabilirdir. Sembolik modellere örnek olarak kural tabanlı sistemler verilebilir. Sayısal modellere ise istatistiksel model örnek verilebilir. Kural tabanlı metotlar ilgili varlığın içeriğini temsil eden kuralları kullanırlar. Örneğin kişinin ismini algılamak için kişinin adının bir ünvandan sonra gelecek şekilde kodlayan aşağıdaki gibi bir kural tanımlanabilir [12] : "*Mr.+ capitalized_word"*

Bu kurallar kural kümesinin sorumluluğunu alacak uzmanlar tarafından geliştirilir. Kural tabanlı sistemler bilgi çekmenin ilk yıllarında üretilmesine rağmen hala gerçek dünya sistemler etkin bir şekilde kullanılmaktadır. Öğrenme tabanlı metotlar varlıkları tanımlayacak modeli öğrenmek için işaretlenmiş veriyi kullanırlar. İşaretlenmiş veri içinde tipleri belirtilmiş şekilde bulunan varlık bulunduran dokümandır. Öğrenilen modeller sembolik ya da istatistiksel olabilir. Sembolik modeller kural öğrenen bir algoritmayı kullanırlar. Kural öğrenen algoritma isimlendirilmiş varlıkların tanımlanması için kural kümesi oluşturur. İstatistiksel modeller, standart istatistiksel makine öğrenme algoritmalarıdır. Makine öğrenme algoritmaları içeriği ya da varlığın niteliğinin gösterimine dayanır. Bu modellerde NER sınıflandırma görevi görür [12].

B. Çoklu Referans Çözümlemesi

Bu yöntemi uygulayabilmek için metin içinde aynı varlığı tanımlayan birden fazla referansın bulunması gerekir. Metin içinde birden fazla bulunma durumu aşağıdaki örneklerle açıklanabilir [10] :

İsimlendirilmiş Varlık: Varlığın isimlendirilmesi durumudur. 'General Electric' ve 'GE' aynı metin içinde aynı varlığı işaret edebilir.

Zamir Durumu: Varlığın zamirle ifade edilmesi durumudur. 'John bought food. But he forgot to buy drinks.' Zamir 'he', 'John'u işaret etmektedir.

Temsili Durum: Varlığın temsili bir sözcük ile ifade edilmesi durumudur. 'Microsoft revealed its earnings. The company also unveiled future plans.' 'The company' sözcüğü 'Microsoft'u temsil etmektedir.

C. İlişki Çekimi

Metinde bulunan varlıklar arasında önceden tanımlanmış ilişkileri tespit etme ve sınıflandırma yöntemidir [13]. Aşağıda ilişki çekimi ile ilgi bazı örnekler verilmiştir:

Çalışan (Bill Gates, Microsoft): Bir kişi ve bir kurum arasındaki ilişkidir. 'Bill Gates works for Microsoft' cümlesinden çıkarılmıştır.

Konum (Uslu, Washington): Bir kişi ve bir konum arasındaki bir ilişkidir. Bu ilişkide kişinin hangi konumda olduğu ilişkisi belirtilir.'Mr. Uslu talked at the conference in Washington' cümlesinden çıkarılmıştır.

AltŞirket (ARK, Seyhan Holding): İki şirket arasındaki ilişkiyi gösterir. Bir şirket diğer şirketin alt şirketidir. 'Listed broadcaster ARK said its parent company, Seyhan Holding, is considering various options for the potential sale'. Genellikle çıkartılabilecek ilişki sayısı limitsiz olmasına rağmen sonuca yönelik çıkartılan ilişki kümesi önceden tanımlanmış, sınırlı ve sabit olmalıdır.

D. Olay Çekimi

Metin içinden olayları tespit ettikten sonra bilgileri detaylı ve yapısal bir biçimde kullanıcıya sunma işlemidir. Yapısal bilgi içerisinde 'kimin kime ne yaptığı', ne zaman, nerde, neler kullanarak ve nasıl yaptığı gibi bilgiler tespit edilebilir. Genelde olay çekimi birkaç varlık ve bu varlıkların arasındaki ilişkilerin çekimini kapsar. Örneğin bir metinden terörist saldırısı ile ilgili bilgi çekme yapılabilir. 'Masked gunmen armed with assault rifles and grenades attacked a wedding party in US, killing at least 44 people.' Belirtilen cümleden olayın failleri (masked gunmen), kurbanlar (people), öldürülen/yaralanan sayısı (at least 44), kullanılan silah ve cephaneler (rifles and grenades), konum (US) bilgileri

çıkartılabilir [10].

V. YAPILAN HASATLAMA ÇALIŞMALARI

A. *Ağırlıklandırma Değerleri İle Hasatlama*

Amerikan Homeland Security'nin yapmış olduğu bir çalışmada bilgi hasatlama yöntemi ile radikal fikirlere sahip internet sitelerini tespit edilmektedir. Yapılan çalışmada iki çeşit analiz yapılmıştır. Biri link analizi diğer ise içerik analizidir. Radikal grupların amaçlarını anlayabilmek için nitelik tabanlı bir metodoloji geliştirilmiştir. Bunlar; iletişim, para toplama, ideoloji paylaşımı, iç propaganda, dış propaganda, sanal topluluk, komuta kontrol, eleman toplama ve eğitim olarak sıralanır. Seçilen nitelikler 13 yıl tecrübeli bir CIA istihbarat analistinin tecrübesi vasıtasıyla seçilmiştir [14].

Yüksek seviyeli her bir nitelik düşük niteliklerin bileşiminden oluşmaktadır. Örneğin iletişim e-posta bağlantısı, telefon bağlantısı, multimedya dosyaları, online geribildirim formu ve dokümantasyondan oluşmaktadır. Detaylardaki düşük seviyeli nitelikleri tanımlayan kodlama şeması geliştirilmiştir. Kodlama şeması aracı para toplama ya da propaganda gibi belli kaynakları ayırma örüntüsünü bulur. Böylece izlenen grupların internet ağını nasıl kullandığı öğrenilmiş olur. Düşük seviyeli bir niteliğe ağırlık atamak belli amaçlar için kullanım seviyesini ölçmeye yarar.

B. *Kural ve Örüntü Tabanlı Metodlar İle Yapılan Hasatlama*

[15]'te yapılan çalışmada çeşitli bilgi çekme yöntemleri tartışılmıştır. Arama motorlarının işlevselliğinin anlamsal seviyesini yükselten başlıca eğilimler vardır. Amaç otomatik bir şekilde isimlendirilmiş varlıklarla alakalı kapsamlı bir bilgi tabanı, anlamsal sınıflarını ve ortak ilişkilerini yüksek oranda başarı ile oluşturmaktır. Yapılan diğer çalışmada [11] hasatlamanın ilk seviyesinde tüm varlıklar toplanmıştır. Bu varlıklar kişiler, şirketler, şehirler ve ürünler gibi varlıklardır. Varlıklar anlamsal sınıflara ayrılmıştır. Örneğin bu sınıflar sanatçılar, bilim insanları, moleküler biyologlar vs. olabilir. Belli bir varlık birden fazla sınıfa dahil olabilir. Örneğin Angela Merkel politikacı, bilim insanı, başbakan gibi birden fazla sınıf içine dahil olabilir.

WordNet [16] İngilizce kelimelerin sözlüksel anlamlarını barındıran bir çatıdır. YAGO (Yet Another Great Ontology) [17] isimli çalışmada WordNet ile Wikipedia üzerinde çalışılmıştır. YAGO kendi bünyesinde barındırdığı sınıfları WordNet'ten içeri aktarmıştır. YAGO'nun Wikipedia'da bulduğu her varlık YAGO'nun belirlediği sınıflardan birine atanmalıdır. Eğer varlık atama işlemi başarısız olursa varlık bilgi tabanına dahil edilmez. Varlıkların sözlüksel anlamları ve dahil oldukları sınıflar ile ilgili bilgi elde edildikten sonra varlıklar hakkında ilişkisel bilgiler elde edilebilir. İlişkilerden ikili ilişkiler ele alınabilir. Örneğin dogumTarihi ⊆ Kişi x Şehir, evlilik ⊆ Kişi x Kişi, mezuniyet ⊆ Kişi x Üniversite gibi ilişkiler ikili ilişkilere örnek verilebilir [15].

Doğal dillerin kısıtlamaları vardır. Bunlardan birisi olarak isim kelimesi sadece bazı fiiller ile birlikte kullanılabilir. Örneğin meyve suyu içilebilir ya da üretilebilir. Fakat yenilemez ya da sürülemez. İsim kelimeleri belli fiil kelimelerine göre kümelenebilir. Heaest örüntüleri POS (Part Of Speech) ile zenginlendirilmiş düzenli ifadelerdir [11]. Hearst örüntüleri [18] serbest formatlı metin ifadelerinden gelen önceden tanımlı ilişkisel modelin örneklerini bulmayı amaçlar. Örneğin instanceOf ilişkisi için isim örnekleri otomatik olarak aşağıdaki örüntüden tespit edilebilir:

$$NP_0 : \{NP_1, NP_2 ... (and \mid or)\}\ NP_n \quad (4)$$

(4) numaralı örüntüde belirtilen NP özel isimler için bir POS etiketidir. Hearst örüntüleri yüksek duyarlılık oranına sahiptir. Fakat düşük duyarlılık değerine sahiptir. Hearst örüntüleri el ile yazılır. Otomatik olarak üretilmez. Bunun için örüntüleri üretmek zordur.

Elle üretilen örüntülerde yüksek duyarlılık değeri mümkün olabilmekte fakat genellikle düşük duyarlılık değerleri çıkmaktadır. Bunun için elle üretilen örüntülerin yanında otomatik örüntü üreten çalışmalar da olmuştur. Örneğin KnowItAll [19] çalışmasından düşük hata positif oranı (false-positive rate) ile birlikte yüksek çağrı değeri elde edilmektedir.

C. *İnternet Ağı Üzerindeki İlişkisel Tablolardan Yararlanarak Yapılan Hasatlama*

İnternet sayfaları üzerinde veri tutan birçok listeler vardır ve bu listelerden bilgi hasatlaması yapılabilir. İnternet sayfası üzerinde bulunan listeler çok kolonlu tablolara dönüştürülebilir. Tablo bilgilerinden öncelikle kolon bilgilerinin çekilmesi gerekmektedir. Bunun için tabloda bulunan alanların kalitesi ölçülmektedir. Kaliteyi ölçen metrik alan kalite skorudur (FQ) [20].

$$FQ(f)=a_{ts}\times S_{ts}(f) + a_{lms}\times S_{lms}(f) + a_{tcs}\times S_{tcs}(f) \quad (5)$$

(5)'te S_{ts} (f) tip desteğidir. S_{lms} dil model desteğidir. S_{tcs} tablo metin desteğidir. Her bilgi kaynağı bir ağırlığa atanır. Bu ağırlıklar sırasıyla a_{ts}, a_{lms} , ve a_{tcs}'dir [20].

Tip Destek Skoru (S_{ts}): Tip destek skoru herhangi bir alanın ayrık tablo kolonlarında sık sık bulunup bulunmadığını anlamaya yarar. Yapılan çalışmada sayısal değerler, tarih değerleri, URL'ler, e-postalar ve telefon numaraları alan olarak belirlenebilir. f'in tipi belirlenirse S_{ts}(f) değeri 1'e eşitlenir aksi takdirde 0'a eşitlenir [20].

Dil Model Destek Skoru (S_{lms}): Dil modeli sözcük dizilerinin oluşma ihtimalini belirler. İki çeşittir. Biri içsel uyum skorudur. Diğeri dışsal uyum skorudur. İçsel uyum skoru S_{ic} ile gösterilir. Dışsal uyum skoru S_{ei} (f) ile gösterilir. Her iki skor (6) ve (7)'da gösterilmiştir [20].

$$S_{ic}(f) = \frac{\sum_{h=1}^{m-1} Pr(w_{i+h}|w_i,\ldots,w_{i+h-1})}{m-1} \quad (6)$$

$$S_{ei}(f) = \frac{2}{Pr(w_i|w_{i-1}) + Pr(w_{i+h+1}|w_{i+h})} \quad (7)$$

(6)'da $Pr(w_i \mid w_1, \ldots, w_{i-1})$, w_i'nin $(w_1, \ldots, w_i)$ sözcük dizisini takip etme ihtimalidir. m ise bir satırdaki sözcük sayısını ifade eder. (7)'de $Pr(w_i \mid w_{i-1})$, f'de bulunan ilk sözcüğün son sırada bulunan sözcüğü takip etmesi ihtimalidir.

Yine (7)'de bulunan $Pr(w_{i+h+1}|w_{i+h})$, bir sonraki alanda bulunan ilk sözcüğün f'deki son sözcüğü takip etmesi ihtimalidir. Dil model skoru içsel ve dışsal uyumun ağırlıklı ortalamasıdır [20].

$$S_{lms}(f) = a_{ic} \times S_{ic}(f) + a_{ei} \times S_{ei}(f) \quad (8)$$

a_{ic} ve a_{ei} 0 ve 1 aralığında bulunan değerlerdir ve $a_{ic} + a_{ei} = 1$'dir.

Tablo metin destek skoru (S_{tcs}): Tablo metin destek skoru f'in internet sayfasında bulunan tablolarındaki metinlerde ne kadar desteklendiğini gösterir. tc_support değeri bir tabloda f'in kaç kere hücre değeri olarak bulunduğunu göstersin. tc_support değeri min_tc_support değerinden küçükse S_{tcs} değeri 0, büyükse 1 olarak atanır.

Tablo 1'de yapılan çalışmaya ListExtract adı verilmiştir ve RoadRunner [21] ile karşılaştırma yapılmıştır.

TABLO I
LİSTEXTRACT VE ROADRUNNER KARŞILAŞTIRILMASI [20]

Uygulama	Duyarlılık	Çağrı	Ağırlıklı Ölçüm
ListExtract	0.64	0.63	0.63
RoadRunner	0.39	0.28	0.32

D. *Wikipedia'dan Bilgi Hasatlama Çalışması*

Yapılan bir çalışmada [22] Timely YAGO (T-YAGO) isimli bir bilgi tabanı ortaya konmuştur. Timely YAGO Wikipedia'dan başlıklardan, listelerden ve kategorilerden zamansal olguları çıkarabilmektedir. Aynı zamanda zamansal olgular sorgulanabilmektedir. T-YAGO zamansal bilgi tabanına dayanarak SPARQL isimli bir sorgu dili imkânı sağlar. Olgular özne, nitelik ve nesne üçlüsü halinde gösterilir. Kullanılan zamansal koşulları gösteren sözcükler on, since ve until olarak kullanılmıştır. Zamansal olgular arasındaki ilişkileri gösteren sözcükler şunlardır [2] : *before, after, equal, during, overlaps, sameYear*

Yukarıda kullanılan sözcüklere benzer ilişkisel başka sözcükler de kullanılmaktadır. Fakat temel olarak kullanılanlar yukarıdadır. Örnek olarak David Beckham'la aynı takımda ve aynı zaman aralığında oynayan oyuncuları sorgulayan sorgu aşağıdadır[22].

```
?id1: "David Beckham" playsForClub ?x .
?id2: ?a playsForClub ?x .
?id1 since ?t1 . ?id1 until ?t2 .
?id2 since ?t3 . ?id2 until ?t4 .
[?t1-?t2] overlaps [?t3-?t4] .
?a notEqual "David Beckham"
```

Şekil.2 SPARQL sorguları[22]

[?t1-?t2] zaman aralığını temsil eder. Sorguda overlaps koşulu iki zaman aralığının örtüşüp örtüşmediğini belirler.

VI. KİŞİSEL BİLGİ HASATLAMASI

Bilgi hasatlamasının ilgi alanı içinde kişisel bilgilerin hasatlaması da yer almaktadır. Dolayısıyla herkese açık olan sosyal medya servisi, blog ve diğer internet sitelerinden kişisel bilgilerin hasatlaması yapılabilir.

Kişisel bilgi olarak internet sitelerinden örneğin bir kişinin işten atılıp atılmadığı anlaşılabilir. Örneğin Türkçe olarak "işten atıldım", "işten kovuldum" gibi cümle parçacıkları aranarak ilgili cümle içinde kişi ismi de varsa eşleşme yapılabilir [23]. Diğer bir sorun TC Kimlik numarasının internette excel, pdf, word gibi dosyalarda açıkça bulunmasıdır. Kişisel bilgi hasatlamasında kişi ismi ile istenen kişinin TC kimlik numarası bulunabilir.

2009 yılında yapılan bir çalışmada MySpace'de kişilerin benzerlik oranları çıkarılmıştır. Arkadaş olanların birbirine yakın yaşlarda aynı dini görüşlere sahip olduğu ve çocuklara karşı benzer yaklaşımları olduğu çıkarılmıştır. Fakat cinsiyet konusunda benzerliğe dair herhangi bir sonuç bulunamamıştır [24].

Sosyal ağlarda birçok kullanıcının e-posta, okul, iş gibi kişisel bilgileri tüm arkadaşlarına açıktır. Yapılan bir çalışmada kişilerin facebook vb. sosyal ağlarda oturumları çalınarak tüm arkadaş bilgileri çekilmiştir. Arkadaş bilgileri çekildikten sonra bu arkadaşlar taklit edilmiştir. Oturum çalma işleminden sonra kullanıcının doğum günü ya da konum bilgisi ile de saldırı gerçekleştirilebilir [25]. E-mail adresleri, anlık mesajlaşma bilgileri gibi hassas kişisel bilgiler spam mesajları tarafından inandırıcılığı arttırılabilmesi için kullanılabilmektedir. Sosyal ağlardaki kişisel bilgi havuzuna erişim sağlamak ve bir sosyal ağ kullanıcısını taklit etmek çözülmesi kolay olmayan bir zorluktur [26]. İlk yapılan çalışmalar [27,28] sosyal ağlarda bilgi çıkarımına dair bir farkındalık oluşturdu. Çünkü bu çalışmalardan sonra sosyal medya üzerinde yapılan çalışmalar artmıştır. Yapılan diğer bir çalışmada çeşitli sosyal ağlarda bulunan kullanıcıların bilgilerini çekebilen ve bu çekilen bilgileri kullanarak sahte profil üretebilen iCloner isimli bir sistem geliştirilmiştir. iCloner birden fazla sahte profil üretirken CAPTCHA'yı analiz ederek çalışmaktadır.[29].

VII. SONUÇ

İnternet ağı üzerinden bilgi hasatlaması en önemli araştırma konularından birisidir. Bilgi hasatlaması araştırmasında öne çıkan iki temel konu vardır. Birisi bilgi çıkarma (information retrieval) diğeri ise bilgi çekme (information extraction)dir. Bilgi hasatlaması ile ilgili detaylı bir literatür taraması yapılmıştır.

Bilgi çıkarmada en çok kullanılan yöntemlerden biri vektör uzay modeli (VSM)'dir. Vektör uzay modelinde dokümanlar arasındaki benzerlik kosinüs açıları ile bulunur. Aralarındaki açı küçük olan dokümanlar birbirine daha çok benzer olarak kabul edilir. Vektör uzay modeli ise boyutu büyük olan dokümanlarda düşük performans gösterir. Çünkü ilgili dokümanda her bir terim için terim frekansı ve ters doküman frekansı hesaplanır [30]. Vektör uzay modeli incelendikten sonra ikili işlem modeli incelenmiştir. İkili işlem modeli dokümanlar üzerinde çalışırken ilgili terim model için ya vardır ya da yoktur [31]. Bundan dolayı ya çok az ya da çok fazla doküman üzerinde çalışır.

Makalede incelenen bilgi çekme yöntemleri isimlendirilmiş varlık (NER) tanımlaması, çoklu referans çözümlemesi, ilişki çekimi ve olay çekimidir.

Bilgi hasatlamasında en büyük iki engel çok anlamlılık (polysemy) ve eş anlamlılıktır (synonymy). Çok anlamlılık bir kelimenin birden fazla anlamı olmasıdır. Eş anlamlılık birden fazla kelimenin bir anlamı olmasıdır [32].

Bilgi hasatlaması bilgi çekme ve bilgi çıkarma olarak kategorik incelendikten sonra bilgi hasatlaması kişisel bilgi hasatlaması özelinde olarak incelenmiştir. Bilgi hasatlamasında kişisel bilgilerin ele geçirilebileceği görülmüş olup araştırmanın sonunda bu konu üzerinde durulmuştur. Özellikle sosyal medyada kişisel bilgilerin internet ağına ne kadar açıkta olduğunun farkında olunması lazımdır. Çünkü bilgileri hasatlanan kişilerin bilgileri kullanılarak sosyal mühendislik saldırıları gerçekleştirilebilmektedir. Bununla birlikte özellikle Facebook başta olmak üzere sosyal ağları tarayan internet robotları bulunmaktadır. Gizlilik özelliklerini yeterince kullanmayan sosyal medya kullanıcıların bilgileri hiç kullanmamış oldukları sahte sosyal ağlarda bulunabilmektedir. Yine arama motorları yardımıyla kişisel bilgileri içeren dosyalar internet üzerinde bulunabiliyorsa bu dosyaların arama motorları tarafından indekslenmesi engellenmelidir.

KAYNAKLAR

[1] A. Sun, E. P. Lim, and W. K. Ng, "Web classification using support vector machine," in *Proceedings of the 4th international workshop on Web information and data management*, 2002, pp. 96–99.

[2] Y. Bassil, "A Survey on Information Retrieval, Text Categorization, and Web Crawling," *J. Comput. Sci. Res.*, vol. 1, no. 6, pp. 1–11, 2012.

[3] E. Greengrass, "Information retrieval: A survey," *Information Retrieval*. p. 224, 2000.

[4] I. Guy, M. Jacovi, E. Shahar, N. Meshulam, V. Soroka, and M. Carmel, "Harvesting with SONAR - The Value of Aggregating Social Network Information," *Soc. Networks*, pp. 1017–1026, 2008.

[5] K. Sharma, G. Shrivastava, and V. Kumar, "Web mining: Today and tomorrow," *2011 3rd Int. Conf. Electron. Comput. Technol.*, vol. 1, pp. 399–403, 2011.

[6] L. Duan, S. Oyama, M. Kurihara and H. Sato, 'Establishing Relationships between Emotion Taxonomies Using the Vector Space Model', *Lecture Notes in Engineering and Computer Science*, vol. 2215, no. 1, pp. 19-24, 2015.

[7] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11. pp. 613–620, 1975.

[8] M. Yamamoto and K. W. Church, "Using Suffix Arrays to Compute Term Frequency and Document Frequency for All Substrings in a Corpus," *Computational Linguistics*, vol. 27, no. 1. pp. 1–30, 2001.

[9] X. Li and W. Cao, "A method for person name disambiguation based on Baidu Encyclopedia," in *Transportation, Mechanical, and Electrical Engineering (TMEE), 2011 International Conference on*, 2011, pp. 423–426.

[10] J. Piskorski and R. Yangarber, "Information extraction: Past, present and future," in *Multi-source, multilingual information extraction and summarization*, Springer, 2013, pp. 23–49.

[11] R. Besancon, G. de Chalendar, O. Ferret, F. Gara, O. Mesnard, M. Laib, and N. Semmar, "LIMA : A Multilingual Framework for Linguistic Analysis and Linguistic Resources Development and Evaluation," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010.

[12] S. Elloumi, A. Jaoua, F. Ferjani, N. Semmar, R. Besançon, J. Al-Jaam, and H. Hammami, "General learning approach for event extraction: Case of management change event," *J. Inf. Sci.*, p. 0165551512464140, 2012.

[13] N. Bach and S. Badaskar, "A review of relation extraction," *Lit. Rev. Lang. Stat. II*, 2007.

[14] Y. Zhou, E. Reid, J. Qin, H. Chen, and G. Lai, "US domestic extremist groups on the Web: link and content analysis," *IEEE Intell. Syst.*, vol. 20, no. 5, 2005.

[15] G. Weikum and M. Theobald, "From information to knowledge: harvesting entities and relationships from web sources," *Proc. twenty-ninth ACM SIGMOD-SIGACT-SIGART Symp. Princ. database Syst.*, pp. 65–76, 2010.

[16] P. University, 'About WordNet - WordNet - About WordNet', *Wordnet.princeton.edu*, 2015. [Online]. Available: https://wordnet.princeton.edu. [Accessed: 12- Jul- 2015].

[17] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago," in *Proceedings of the 16th international conference on World Wide Web - WWW '07*, 2007, p. 697.

[18] M. A. Hearst and M. A. Hearst, "Automatic Acquisition of Hyponyms from Large Text Corpora," in *Proceedings of the 14th International Conference on Computational Linguistics*, 1992, pp. 539–545.

[19] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates, "Web-Scale Information Extraction in KnowItAll (Preliminary Results)," in *WWW'04 Proceedings of the 13th international conference on World Wide Web*, 2004, pp. 100–110.

[20] H. Elmeleegy, J. Madhavan, and A. Halevy, "Harvesting relational tables from lists on the web," *VLDB J.*, vol. 20, no. 2, pp. 209–226, 2011.

[21] V. Crescenzi, G. Mecca, and P. Merialdo, "RoadRunner: automatic data extraction from data-intensive web sites," in *Proceedings of the 2002 ACM SIGMOD international conference on Management of data*, 2002, p. 624.

[22] Y. Wang, M. Zhu, L. Qu, M. Spaniol, and G. Weikum, "Timely YAGO : Harvesting , Querying , and Visualizing Temporal Knowledge from Wikipedia," *Proc. 13th Int. Conf. Extending Database Technol. (EDBT), Lausanne, Switzerland, March 22-26*, pp. 697–700, 2010.

[23] D. Wilkinson and M. Thelwall, "Researching Personal Information on the Public Web: Methods and Ethics," *Social Science Computer Review*, vol. 29, no. 4. pp. 387–401, 2011.

[24] M. Thelwall, "Homophily in MySpace," *J. Am. Soc. Inf. Sci. Technol.*, vol. 60, no. 2, pp. 219–231, 2009.

[25] M. Huber, M. Mulazzani, E. Weippl, G. Kitzler, and S. Goluch, "Friend-in-the-middle attacks: Exploiting social networking sites for spam," *IEEE Internet Comput.*, vol. 15, no. 3, pp. 28–34, 2011.

[26] M. Huber, M. Mulazzani, E. Weippl, G. Kitzler, and S. Goluch, "Exploiting social networking sites for spam," *Proc. 17th ACM Conf. Comput. Commun. Secur. - CCS '10*, p. 693, 2010.

[27] T. N. Jagatic, N. A. Johnson, M. Jakobsson, and F. Menczer, "Social phishing," *Communications of the ACM*, vol. 50, no. 10. pp. 94–100, 2007.

[28] H. Jones and H. Soltren, "Facebook : Threats to Privacy," *Soc. Sci. Res.*, vol. December 1, pp. 1–76, 2005.

[29] L. Bilge, T. Strufe, D. Balzarotti, E. Kirda, and S. Antipolis, "All Your Contacts Are Belong to Us : Automated Identity Theft Attacks on Social Networks," *Www 2009*, pp. 551–560, 2009.

[30] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5. pp. 513–523, 1988.

[31] P. Castells, M. Fernández, and D. Vallet, "An adaptation of the vector-space model for ontology-based information retrieval," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 2, pp. 261–272, 2007.

[32] M. W. Berry, Z. Drmac, and E. R. Jessup, "Matrices, Vector Spaces, and Information Retrieval," *SIAM Review*, vol. 41, no. 2. pp. 335–362, 1999.