Yapılandırılmamış Veriler **Yapılandırılmış Veriler**

Doküman içerikleri

Sosyal Medya Mesajları

e-postalar

Amaç: NLP, istatistiksel modelleme ve makinaca öğrenme teknikleri yoluyla potansiyel değer taşıyan içgörü türetilmesi

Doğal dil metinleri tutarsız sözdizimi ve semantik nedeniyle belirsizlikler içerir. Örn: argo sözcükler, yaş gruplarına özgü dil ve ironi gibi.

MM, genelde girdi metnin yapılandırılması sürecidir: Metnin işlenmesi için birtakım linguistik özellikler eklenir, gramer olarak incelenir(parsing) ve işlevsel sözcükler elenir, ardından elde edilen sözcükler bir veri tabanına girilir; yapılandırılmış veri içerisinden örüntüler türetilir ve son olarak çıktı değerlendirilerek yorumlanır. MM'nde ilgi çekicilik (interestingness) orjinallik (novelty) ve ilgililik (relevance) bileşimleri yüksek kaliteye işaret eder.

metin kümeleme (text clustering),

kavram/varlık çıkarımı (concept/entity extraction),

taksonomi üretimi (production of granular taxonomies),

duygu analizi (sentiment analysis),

doküman özetleme (document summarization) ve

varlık ilişki modeli (entity relation modeling) yani adlandırılmış varlıklar (named entities) arasındaki ilişkileri öğrenmeyi kapsar.

Ronen Feldman MM terimini (2000) 2004 te MA olarak değiştirdi.

Metin madenciliği 6 alanla ilişkilidir;

Veri madenciliği(doküman sınıflama, doküman kümeleme)

Kütüphane ve Enformasyon Bilimi (bilgi erişim)

Veri tabanları (bilgi erişim)

Hesaplamalı Dilbilim (NLP, kavram çıkarımı)

Yapay Zeka, Makinaca Öğrenme, İstatistik, Enformasyon Çıkarımı, NLP

Web Madenciliği

Not: Kırmızılı konular metin madenciliğini oluşturur.

Doküman kümeleme: Veri madenciliğinin kümeleme yöntemlerini kullanarak dokümanların, paragrafların, parçaların/kesitlerin terimlerin gruplanması ve kategorilere ayrılmasıdır;

Doküman sınıflaması: Etiketlenmiş örnekler üzerinde deneyimlenen modellere dayalı sınıflama yöntemleri kullanılarak dokümanların, paragrafların, parçaların/kesitlerin terimlerin gruplanmasıdır;

Enformasyon çıkarımı: Yapılandırılmamış metinlerden, ilgili olgu ve ilişkilerin çıkarımı ve tanımlanmasıdır;

Kavram çıkarımı: Semantik benzerliği olan sözcük ve tamlamaların gruplanmasıdır.

Veri Madenciliği (data mining):

Birbiri yerine kullanılan “veri madenciliği” ve “bilgi keşfi”, 2007 sonrasında öngörü analizi ve 2011 den itibaren de veri bilimi olarak kullanılır olmuştur.

Örüntü Tanıma (Pattern Recognition):

Makinaca öğrenmenin bir dalıdır; veri içindeki örüntü ve düzenlilikleri tanımaya odaklanmıştır. Bazı durumlarda makinaca öğrenme ile eş anlamlı kullanılmaktadır.

Metin Kümeleme (text clustering): Metinsel dokümanlara kümeleme analizlerinin uygulanmasıdır. Otomatik doküman düzenleme, konu çıkarma ve hızlı erişim veya süzme alanında uygulamaları vardır.

MK uygulamaları otomatik olarak bir doküman dermesinin örtük yapısını ortaya çıkarır, derme içindeki sıklıkla geçen konuları tanımlar ve dokümanları çeşitli küme/gruplar halinde düzenler. Bu dağılım, hem aynı grup içindeki dokümanların benzerliğini, hem de farklı gruplar arasındaki farklılığı maksimize eder.

Dokümanlar gruplanırken sadece metinlerin benzerliğine bakılmaz, dokümanların dermedeki konularla olan ilgililiklerine de bakılır ve otomatik olarak her kümeye onun konusunu temsil eden bir başlık, isim atanır. Aynı zamanda içinde, bir terimin bütün varyasyonlarını, stopwordleri ve diğer dilbilimsel unsurları dikkate almayı mümkün kılan, lemmatization teteknolojilerini kullanmaktadır.

Bir sınıflama modelinin yaratılması, makinanın önceden manuel olarak sınıflanmış metinlerle eğitilmesi veya her kategoriye bir dizi kural tanımlayarak (denetimli öğrenme olarak bilinir).

Örnek: Bir araştırma grubu biyomedikal dergi makalelerinden ilgili enformasyonu çıkarmada metin madenciliği yöntemlerinin kullanıldığı bir çalışma yapar. Bu enformasyon daha sonra gene-centric veri tabanlarındaki enformasyon ile entegre edilecek ve belirli bir veri seti ile ilgili yayınlanmış bütün bilginin görsel bir temsilini üretmek için kullanılacaktır. Buradaki hedef, yeni açıklayıcı hipotezlerin tanımlanmasıdır

Araştırma grubu hibe destekli araştırmaları için XML formatında çok büyük miktarda bir tam metin dergi makalesi dermesi oluşturmak, metinler üzerinde madencilik yapma hakkını elde etmek ve elde edilen verileri depolayıp kullanma hakkı almak durumundadır.

İşe dergi makaleleri ile başlarlar ve ilgili literatürden büyük bir derme derlerler. Madenlenecek olan metinler birçok formatta olabilir, XML metin madenciliği için bilgisayara hazır bir formattır; çünkü dokümanın kısımları yapılandırılmıştır. XML 'e 'markup language' deniyor çünkü veri parçalarını betimlemek ve işaretlemek için etiketler (tags) kullanıyor. Açılabilir kısım ise içeriğin tipine göre kullanıcıların onları tanımlayabileceği anlamına gelmektedir.

XML dergi yayıncıları tarafından bir içerik yaratma format standardı olarak uyarlanmıştır; çünkü elektronik ortam için esnek bir formattır. **XML makale kısımlarının başlık, yazar, öz vb etiketlerle kodlanmasını olanaklı kılmaktadır.** Makalenin elektronik olarak editör ve yayıncı arasında iletimi ve kolaylıkla diğer versiyonlarda(basılı, online) formatlanma ve yeniden üretim olanağı sağlamaktadır. XML aynı zamanda metin içindeki belirgin içeriğe işaret edebilir, örn: biyolojik terimler veya kavramlar.

Bir kez makalenin içerik ve kısımları tanımlandıktan sonra metin madenciliği teknikleri makaleye uygulanır. MM metinden kavram biçiminde anlam çıkarır, kavramlar arasındaki ilişkileri veya onlar üzerine gerçekleştirilen eylemleri çıkarır ve bunları olgu(fact) veya değerlendirme olarak sunar.

MM teknikleri makinaca okunabilir formattaki her tip enformasyona uygulanabilir (örn: dergi makalesi, e-kitap). MM ile veri toplandığı zaman bir veri seti yaratılır. Birtakım araçları kullanılarak araştırma grubu, bilgiye

dayalı analiz sistemi ile veri setini analiz eder ve yeni hipotezlere götürme potansiyeli olan bilginin görsel temsilini üretir. MM ve kullanılan tekniklerin bilimsel literatürün içerdiği bilgi parçaları arasında ilişki oluşturma potansiyeli vardır ve bilimde daha hızlı ilerleme ile sonuçlanacak yeni hipotezlere önderlik eder.

Araştırma grubu araştırma makalelerinin bilgiye dayalı analiz sistemini geliştirmek için bu noktada çok önemli olduğunu belirlemiştir. Peki yararlı makale nasıl tanımlanabilir- uzunluğu ile içerik tipini belirleyen XML etiketleri vb ölçütlere göre.

XML etiketleri içine gömülmüş makale özellikleri(attributes) ve karakteristikleri bir makaleyi tanımlamak için kullanılır. Bunlar: Öz, gövde, en az 40 metin satırı, düzeltme, hata, kitap incelemesi, editör, giriş, önsöz, yazışma veya editöre mektup gibi etiketleri dışarıda bırakılır.

Araştırma grubu bundan sonra firmaya ne kadar makale alındığını bildirmelidir. Bu süreç 400 bin makale alınana kadar devam edecektir. MM'nde kullanılacak makale dermesinin oluşturulması yaklaşık bir yıl sürmüştür.

XML formatındaki dokümanlarda MM yapma, kütüphanelere ve kütüphanecilere ve onların içeriğin güvenliği konusundaki rolüne olan talep giderek artmaktadır. Yayıncılar, araştırmacılar ve kütüphaneler dergi makalelerine mm uygulanmasının potansiyel ticari ve araştırma değerini görmektedir. MM kamu fonuyla yapılan araştırmaları tam kullanma potansiyeli sunmaktadır.

Bununla beraber yayıncılar kendi perspektiflerinden MM ile ilgili başlıca 2 engel tanımladılar—içerik formatlarında ve erişim şartlarında (access terms) standartlaşma olmaması ve yayıncılar, araştırma güdümlü madencilik istekleri içi paylaşımlı erişim şartları geliştirmek zorunda olduklarını kabul ettiler.

Araştırmacı ve kütüphaneci perspektifinden birçok engel ve maliyet var. Örn: MM yapılabilir olan materyale erişim hakkı, işlem maliyeti(MM katılımında), giriş (MM hazırlamak), personel ve altyapı. (15) Kütüphaneler üzerine yapılan bir saha araştırmasında bulgular gösterdi ki, **kütüphaneciler** MM yapmak için araştırmacılar ve yayıncılar arasında yardımcı olarak **yeni bir rol umuyorlar**. Kütüphaneciler bu role doğal olarak uyuyor çünkü onların zaten telif hakları izinleri, lisans anlaşmaları konusunda uzmanlığı var. Rehberlik ve danışmalık geliştirilmeli; bu bağlamda ne zaman izin gerekiyor, ne istenecek, düşünülen çalışma en

iyi nasıl açıklanır, araştırma ve telif hakkı sahiplerine yarar nasıl tarif edilir.

Leslie A. Williams, Lynne M. Fox, Christophe Roeder, and Lawrence Hunter(2014). "Negotiating a Text Mining License for Faculty Researchers" INFORMATION TECHNOLOGY AND LIBRARIES SEPTEMBER (5-22)

http://searchbusinessanalytics.techtarget.com/definition/text-mining