YBS Ansiklopedi

www.YBSAnsiklopedi.com

Cilt 2, Sayı 3, Eylül 2015

Metin Madenciliği (Text Mining)

Sadi Evren SEKER

Istanbul Medeniyet University, Department of Business

Özet

Bu yazı kapsamında, yazılım geliştirme yaşam döngüsü veya sistem geliştirme yaşam döngüsü olarak bilinen SDLC kavramı ve bu kavrama bağlı olarak literatürde sıkça geçen ve temel olarak kabul edilebilecek bazı yazılım geliştirme metotları incelenmiştir. Bu yazılım geliştirme modelleri, şelale modeli (waterfall model), fıskiye modeli (fountain model), v-şekil modeli (v-shaped model), iteratif model (iterative model) ve spiral modeldir. Modellerin temel özelliklerinin tanıtılmasının yanında modellerin üstün ve zayıf oldukları yanların incelemesi ve sistem yaşam döngüsü içerisindeki konumları da incelenmiştir.
***Anahtar Kelimeler**:* Yönetim Bilişim sistemleri, Yazılım Mühendisliği, Proje Yönetimi, Yazılım Proje Yönetimi

Summary

This paper aims to cover the concept of software development life cycle or system development life cycle in the literature. Besides the concept, some well-known and relatively important software development methodologies are also covered in the paper. Some of these methods are waterfall model, fountain model, v-shape model, iterative model and spiral model. Besides the introductory sides of the models, their strengths and weaknesses are also covered together with the software development life cycle.
***Keywords:** Management Information Systems, Software Engineering, Project Management, Software Project Management*

1. Giriş ve Tanımı

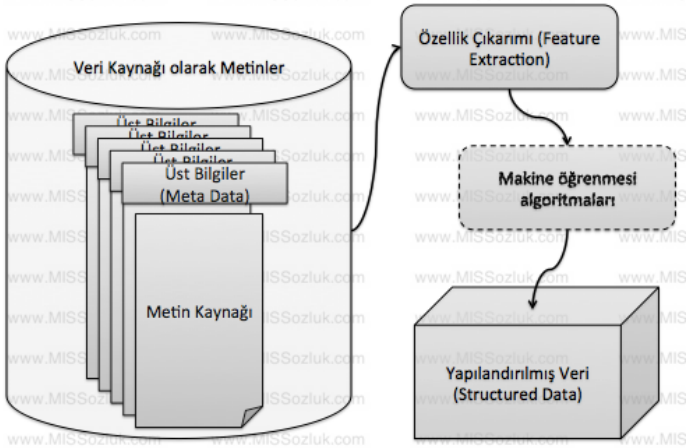
Metin madenciliği çalışmaları metni veri kaynağı olarak kabul eden veri madenciliği (data mining) çalışmasıdır diğer bir tanımla metin üzerinden yapısallaştırılmış (structured) veri elde etmeyi amaçlar. Örneğin metinlerin sınıflandırılması, bölütlenmesi (clustering), metinlerden konu çıkarılması (concept/entity extraction), sınıf taneciklerinin üretilmesi (production of granular taxonomy), duygusal analiz (sentimental analysis), metin özetleme (document summarization), varlık ilişki modellemesi (entity relationship modelling) gibi çalışmaları hedefler.

Yukarıdaki hedeflere ulaşılması için metin madenciliği çalışmaları kapsamında enformasyon getirimi (information retrieval), hece analizi (lexical analysis), kelime frekans dağılımı (Word frequency distribution), örüntü tanıma (pattern recognition), etiketleme (tagging), enformasyon çıkarımı (information extraction), veri madenciliği (data mining) ve hatta görselleştirme (visualization) gibi yöntemleri kullanmaktadır[1].

Peer-review under responsibility of Academic World Education and Research Center.

Metin madenciliği çalışmaları, metin kaynaklı literatürdeki diğer bir çalışma alanı olan doğal dil işleme (natural language processing, NLP) çalışmaları ile çoğu zaman beraber yol yürümektedir. Doğal dil işleme çalışmaları daha çok yapay zeka altındaki dil bilim bilgisine dayalı çalışmalarını kapsamaktadır. Metin madenciliği çalışmaları ise daha çok istatistiksel olarak metin üzerinden sonuçlara ulaşmayı hedefler. Metin madenciliği çalışmaları sırasında çoğu zaman doğal dil işleme kullanılarak özellik çıkarımı da yapılmaktadır.

Genel olarak klasik bir metin madenciliği çalışmasını aşağıdaki şekilde özetleyebiliriz.



Şekil 1 Genel olarak Metin Madenciliğinin Adımları

Şekil 1'de de görüldüğü üzere, bir metin veri tabanından alınan veriler öncelikle bir özellik çıkarımına tabi tutulur. Ardından çıkarılan özellikler üzerinde bir makine öğrenmesi algoritması çalışır (sınıflandırma (classification), bölütleme (clustering), tahmin (prediction) v.b.) ve neticede yapılandırılmış veri (structured data) elde edilir.

Buradaki makine öğrenmesi aşaması genelde kullanılmakla birlikte, metin madenciliği için şart olmayan bir aşamadır. Bazı durumlarda, doğrudan çıkarılan özellik aranan yapılandırılmış veri olabilmektedir. Bazı durumlarda ise makine öğrenmesi adımı yerine, istatistiksel bazı farklı yöntemler kullanılabilir.

Metin kaynakları, genelde doğal dilde yazılmış kaynaklardır. Yani bir gazetedeki köşe yazıları, bir kitap, bir makale olabilir. Hatta internet üzerindeki web siteleri bile metin kaynağı olarak görülebilir (bu konu daha özel olarak web madenciliği (web mining) olarak da adlandırılmaktadır). Bu yazıların, metin madenciliği açısından önemli bir de üst bilgileri olması söz konusudur. Örneğin yazının tarihi, yazının yayınlandığı web sitesi, yazar bilgisi gibi, yazının içerisinde yer almayan ancak yazı ile ilgili metin madenciliğinde kullanılabilecek önemli üst bilgiler (meta data) bulunabilir.

Özellik çıkarımı (feature extraction) aşamasında, metinlerin doğrudan içeriğinden veya üst bilgilerinden yararlanılarak istenilen özellikler çıkarılabilir ve çıkarılan özellikler üzerinde işlem yapılabilir.

2. Örnek Metin Madenciliği uygulaması

Örneğin elimizde 100 adet yazı olsun. Bu yazıları yazan yazarları biliyor olalım (diyelim ki 5 farklı yazarın 20'şer adet yazısı olsun). Yeni gelen 101. Yazının bu 5 yazardan hangisine ait olduğunu bulmak, klasik bir metin madenciliği uygulamasıdır ve literatürde yazar tanıma (author recognition) olarak da geçer.

Burada örnek olarak metinlerdeki kelime kullanma sıklıklarını özellik çıkarımı için kullanmak isteyelim. Yani yazarlarımızı kullandıkları kelime sıklıklarından tanıyabileceğimizi düşünüyoruz (author attribution). Her metinde ve dolayısıyla her yazar için hangi kelimeyi ne sıklıkla kullandığı bilgisi bizim özellik çıkarımı aşamamız oluyor.

Ardından kullanılan kelime sıklıklarını örnek olarak makine öğrenme algoritması olan KNN algoritmasına veriyoruz ve diyelim ki yazarını tanımak istediğimiz 101. Yazı için her kelime için en çok kullanan yazarları listeliyoruz. Neticede bize bir olası yazarlar listesi çıkıyor ve biz de en yüksek ihtimalle hangi yazarın bu

yazıyı yazmış olabileceğini söylüyoruz. Bu çıkan sonuç aslında 101. Yazı için anlamlı ve yapılandırılmış bir sonuç olarak kabul edilebilir.

3. Metin madenciliğinin çalışma alanları

Metin madenciliği sırasında genelde aşağıdaki problemlerle ilgilenilir (bunlarla sınırlı değildir).

Enformasyon Getirimi (Information Retrieval): Bu aşama ilgilenilen külliyet (derlem, corpus) hakkında ön bilginin toplandığı aşamadır. Örneğin metin madenciliği web üzerindeki veri kaynakları üzerinde yapılacaksa web sayfaları, adresleri veya dosya sistemi üzerindeyse dosyaların tarihleri, kullanıcı bilgileri, dosya isimleri, dizin bilgileri gibi bilgilerin toplandığı aşamadır.

Doğal dil işleme aşaması (natural language processing): Bu aşama bütün metin madenciliği aşamalarında kullanılmasa bile genelde özellik çıkarımı ve metinden bazı anlamsal bilgilerin elde edilmesinde sıklıkla başvurulan aşamadır. Örneğin, konuşma parçalarının etiketlenmesi (part of speech tagging) veya cümlebilimsel parçalama (syntactic parsing) veya diğer dilbilimsel işlemler doğal dil işleme aşamasında yapılır.

Adlandırılmış varlık tanıma (named entity recognition): Genellikle metin işleme aşamasında istatistiksel bazı özelliklerin çıkarılması için kullanılır. Örneğin, metnin içerisindeki kişi isimleri, yer isimleri, semboller, kısaltmalar v.s. bu yöntemle bulunur. Metin madenciliği çalışmalarının her zaman temiz metinlerde yapılmadığını hatırlatmakta yarar vardır. Örneğin facebook, twitter mesajları, telefonlardan yollanan SMS mesajları gibi mesajların çoğunda yazım hataları hatta kısaltmalar kullanılmaktadır. Metin madenciliği bu ihtimallerin de göz önünde tutulması gereken çalışmalardır. Örneğin ‘’osmanbey’’ kelimesi, istanbulda bir semt ismi olabileceği gibi bir kişi ismi de olabilir. Adlandırılmış varlık tanıma çalışmalarında, hedeflenen kelime gruplarının metin içerisinden çıkarılması, sayılması, yoğunluğunun bulunması, etiketlenmesi gibi işlemler yapılabilir.

Örüntüsü tanımlı varlıkların bulunması (pattern identified entities): Bazı durumlarda, metnin içerisinden özel bazı bilgilerin metin madenciliğine konu olması mümkündür. Örneğin e-posta adresleri, telefon numaraları, adresler, tarihler gibi bazı bilgileri özel olarak almak isteyebiliriz. Genelde bu durumlarda düzenli ifadeler (regular expressions) veya içerik bağımsız gramerler (context free grammers) tanımlanarak metin üzerinde çalıştırılır[2].

Eş Atıf (Coreference): Bir varlığa işaret eden (atıf eden) isim kelime gruplarını ve diğer terimlerin bulunması/ayrılmasını hedefler.

İlişki, kural, olay çıkarımları: Çeşitli amaçlarla metnin içerisinden bazı bilgilerin çıkarılması istenebilir. Örneğin doktora çalışmam sırasında, verilen bir metnin içerisindeki olayları çıkararak sıralamak (event ordering) üzerine çalışmış, Türkçedeki fiil yapılarını, olay belirten kelime gruplarını, zaman kalıplarını ve bütün bu kelime grupları arasındaki olası ilişkileri gösteren özel bir matematik tasarlamıştım[3].

Duygu analizi (sentimental Analysis) : Metinlerde geçen duygusal ifadelerin çıkarılmasını amaçlar. En sık kullanılanı duygusal kutupsallıktır (sentimental polarity). Buna göre bir konu hakkında geçen mesajların veya yazıların olumlu veya olumsuz olmasına göre iki sınıfa ayrılması hedeflenir[4]. Ancak duygu analizi bunun dışında, metinlerdeki ruh hali, kanaat ve daha karmaşık duyguların çıkarılması üzerinde de çalışmaktadır.

Kayanklar

[1]Sadi Evren SEKER, Cihan Mert, Khaled Al-Naami, Nuri Ozalp, Ugur Ayan (2013), Correlation between the Economy News and Stock Market in Turkey., International Journal of Business Intelligence and Review (IJBIR), vol. 4, is. 4, pp. 1-21, 2013

[2] Şadi Evren ŞEKER, “Turkish Query Engine on Library Ontology”, IKE12, Internet Knowledge Engineering, 2012, ISBN:1-60132-222-4, Pages:26-33

[3]Sadi Evren SEKER, Banu DIRI, International Conference on Artificial Intelligence konferansı dahilinde , "Proceedings of International Conference on Artificial Intelligence", bildiri “TimeML and Turkish Temporal Logic”, pp. 881-887, ICAI 2010

[4] Sadi Evren SEKER, Khaled Al-NAAMI “Sentimental Analysis on Turkish Blogs via Ensemble Classifier“, PROCEEDINGS OF THE 2013 INTERNATIONAL CONFERENCE ON DATA MINING, ISBN:1-60132-239-9, DMIN, pp. 10-16, 2013