ESTUDAM Bilişim Dergisi
Cilt 1, Sayı 2 , 14-16, 2020

Journal of ESTUDAM Information
Volume 1, Issue 2, 14-16, 2020

(Geliş Tarihi / Received Date: 13.01.2020, Kabul Tarihi/ Accepted Date: 19.04.2020)

Endüstri 4.0 İş İlanları Üzerine Veri Madenciliği

Fatih Furkan ARSLAN[*1]

[1]Eskişehir Osmangazi Üniversitesi, Mühendislik-Mimarlık Fakültesi, Bilgisayar Mühendisliği Böl., 26480, Eskişehir

Anahtar Kelimeler:
İş ilanı,
Endüstri 4.0,
Veri Madenciliği

Özet: Endüstri 4.0 ile başlayan temel iş kollarının giderek değerini yitirmesi ve farklı meslek gruplarının ortaya çıkmasıyla birlikte farklı alanlardaki iş ilanlarının analiz edilmesi ihtiyacı da ortaya çıkmaktadır. Günümüzde genel olarak iki tür meslek profili bulunmaktadır. Birisi genel geçer bilgilerle yapılabilecek standart tipteki işlerdir. İkincisi ise uzmanlık ve akademik eğitim gerektiren yazılım mühendisliği, elektrik elektronik mühendisliği, elektronik teknikerliği, endüstri mühendisliği veya tedarik zinciri uzmanlığı gibi işlerdir. Bu nedenle elde bulunan veri setinden faydalanarak iş profilleri hakkında bilgi verici daha fazla detay edinilmesi gerekmektedir. Bu yüzden Indeed iş arama platformundaki ilanlardan oluşan bir veri seti kullanılarak veri madenciliği ile iş ilanları analiz aracı oluşturulmuştur.

Data Mining For Industry 4.0 Jobs

Keywords:
Job Advertisment,
Industry 4.0,
Data Mining

Abstract: There is a need to analyze job advertisements in different fields as the basic business lines starting with Industry 4.0 are gradually losing value and the emergence of different occupational groups. Today, there are generally two types of professional profiles. One is the standard type of work that can be done with general passable information. The second is software engineering, electrical and electronics engineering, electronics technician, industrial engineering or supply chain expertise, which requires expertise and academic training. Therefore, it is necessary to obtain more detailed information about job profiles by making use of the available data set. For this reason, data mining and job search analysis tool was created by using a data set consisting of ads in Indeed job search platform.

1. GİRİŞ

Endüstri 4.0 teknolojilerin ve değer zinciri organizasyonları kavramının kolektif bir bütünüdür. Siber ve fiziksel sistemlerin kavramına, nesnelerin internetine ve hizmetlerin internetine dayalıdır. Bu yapı akıllı fabrikalar vizyonun oluşmasına büyük katkı sağlar. Endüstri 4.0 genel olarak aşağıdaki 3 yapıdan oluşmaktadır.

1. Nesnelerin interneti
2. Hizmetlerin interneti
3. Siber ve fiziksel sistemler

Endüstri 4.0 ile temel iş kollarındaki istihdamın azaldığı görülmektedir.

Buna kıyasen yeni iş kollarının ve yeni iş becerilerinin ortaya çıktığı bir süreçten geçilmektedir. Artık fiziksel güç gerektiren işler yerine daha çok teknolojik becerilerinin ağırlıklı olduğu işler ön planda yer almaktadır.

Bunun doğal sonucu olarak iş ilanlarında da hızla bir değişim söz konusudur. Artık iş ilanları online platformlarda yayınlanmakta ve iş profilleri sürekli olarak güncel teknolojiler ve iş alanlarına göre yenilenmektedir. Bu bağlamda online bir platform olan linkidIn ilanları esas alınarak bir text madenciliği gerçekleştirilmiş ve endüstri 4.0'a bağlı olarak iş ilanları profillerinin nasıl değiştiği gözlemlenmiştir.

Teknolojideki değişiklikler ile birlikte son yıllarda yaklaşık yeni 40 iş grubu ortaya çıkmıştır. Ve bu iş gruplarında geleneksel olan iş profillerinin aksine çalışanların çok daha farklı beceriler edinmesi gerekmektedir. Bu yüzden üniversiteler bu gereksinemlere uygun olarak bölümlerini açmakta ve eğitimlerini vermektedir.

Günümüzde bu alanda olan araştırmalar oldukça azdır ve endüstri 4.0 ile gelen yeniliklerin daha iş profillerini ne derece değiştirdiğine dair daha çok araştırma yapılmalıdır.

*İlgili yazar: Fatih Furkan ARSLAN, arslanfatihfurkan@gmail.com

2. METADOLOJİ

İş ilanları analiz uygulaması yapım aşamasında en önemli bölüm düzgün veri seti ile çalışmış olmaktır. Verilerin yani iş ilanlarının belirli bir formata uygun olması önemli bir noktadır. Verilerin düzgün bir şekilde işlenerek analizin en doğru şekilde gerçekleştirilmesi hedef alınmıştır.
İki çeşit method vardır. İlk aşamada veri çıkarma işlemi daha sonraki aşamada ise machine learning algoritması ile yapılmaktadır. Bu yöntemlerle tanımlayıcı analiz ve veri madenciliğine dayalı analiz yapılmaktadır.

2.1. Veri Seti

Veri seti olarak kariyer.net kullanılmak istendi fakat kariyer.net'in apisi bulunmadığı için daha sonra linkidIn'den veri seti sağlanmaya çalışıldı. Developer partnerlik istenildiği için gerekli veri setine kısa sürede ulaşılamadı. Açık kaynak bir veri sağlayıcısı ile çalışılamadığı için daha sonra indeed uygulaması üzerindeki iş ilanlarının bulunduğu excel formatında bir veri seti bulunmuştur. Gerekli analizler bu veri seti üzerinde yapılmıştır.

2.2. Veri Analizi

Elde edilen veri seti ile yapılacak iki analiz türü vardır. Birisi tanımlayıcı analiz diğeri ise veri madenciliği analiz yöntemleridir.
Tanımlayıcı analizler ile lokasyon, çalışma türü, senyorluk seviyesi gibi çeşitli alanlarda istatistikler çıkartılabilir.
Veri madenciliği kısmı ise bu veri setinden anlamlı bilgi çıkarmaya yönelik yapılan çalışmaları kapsamaktadır. Veri madenciliği analizi ile iş profillerini kümeleyebilir, trendler hakkında bilgi sahibi olunabilir. Önemli nokta kelimeler ve söz öbeklerini kullanarak aralarındaki ilişkiyi yakalayabilmektir. Böylece meslek grubları kategorize edilir.
İlk adım olarak veri setinde sadeleştirme yöntemleri kullanılmalıdır. Bu sadeleştirme yöntemlerinin en önemlisi lemmatization'dır.
İkinci adımda ise en sık geçen kelime ve söz öbeklerini bulmaktır. Genel olarak tekli sözcükler çok kullanışlı olmamaktadır. Bunun yerine ikili veya daha fazla kelimeden oluşan söz öbeklerinin kullanılması daha faydalı olacaktır. Örneğin big ve data ayrı olarak pek bir anlam ifade etmese de big data olarak önemli bir anlam ifade etmektedir.
Daha sonra yapılacak işlem kümeleme yani cluster analiz yöntemidir. Tf-idf'ler çıkartıldıktan sonra kümeleme işlemine başlanılır ve buna göre iş ilanlarına kümeleme işlemi yapılır.

3. BULGULAR

Öncelikle tanımlayıcı analizler üzerinde istatiksel sonuçlara bakılacaktır. Daha sonra çıkartılan söz öbekleri ve iş ilanı metinlerine göre kümeleme işlemeri yapılacaktır.

3.1. Tanımlayıcı Analiz

Bu bölümde çalışma türü ve senyorluk seviyelerini ölçülmektedir. Çalışma türü olarak en çok istenen tam zamanlı olarak belirlenmiştir. Bunun yanında küçük bir miktar part time veya stajyer şeklinde de iş ilanları oluşturulmuştur. Sonuç olarak şirketlerin büyük çoğunluğu full time yani tam zamanlı işçi istihdam etmektedir.
Genel olarak şirketlerin senyorluk seviyesi baz alındığında giriş ve orta seviye çalışan isihdam etme eğilim bulunmaktadır.

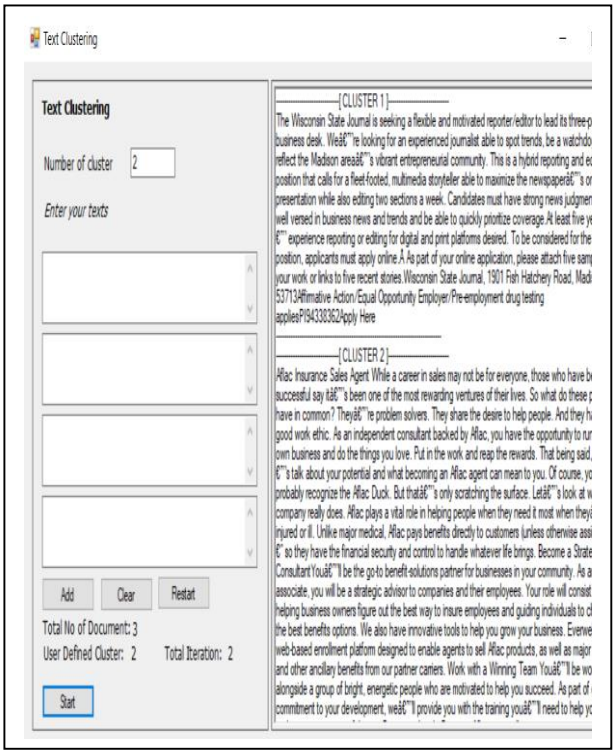
3.2. Veri Madenciliği ile Analiz

Veri madenciliği alanında text veri madenciliği kullanılmıştır. Text madenciliği ile çeşitli sonuçlar elde edilmiştir. Bu sonuçlar içinde ilişkili olan iş ilanlarının tanımlanması, en sık geçen söz ve söz öbekleri bulunmuştur. Bu bölümde terim sıklıkları ve tf-idf üzerinden clustering yani kümeleme işlemleri yapılarak bazı sonuçlara ulaşılmıştır.
İngilizcedeki Term Frequency – Inverse Document Frequency (Terim frekansı – ters metin frekansı) olarak geçen kelimelerin baş harflerinden oluşan terim basitçe bir metinde geçen terimlerin çıkarılması ve bu terimlerin geçtiği miktara göre çeşitli hesapların yapılması üzerine kuruludur.
Klasik olarak TF yani terimlerin kaç kere geçtiğinden daha iyi sonuç verir. Kısaca TF-IDF hesabı sırasında iki kritik sayı bulunmaktadır. Bunlardan birincisi o anda ele alınan dokümandaki terimin sayısı diğeri ise bu terimi külliyatta içeren toplam doküman sayısıdır.[1]
TF-IDF yönteminin diğer yöntemlere göre farkını açıklamaya çalışırsak, TF-IDF ile bir terimin kaç kere geçtiği kadar kaç farklı dokümanda da geçtiği önem kazanır. Örneğin sadece bir dokümanda 100 kere geçen bir terimle 10 farklı dokümanda onar kere geçen terimin ikisi de aslında toplamda 100 kere geçmiştir ancak TF-IDF ikincisine yani daha fazla dokümanda geçene önem verir.[2]
Terim sıklığı baz alındığında en sık geçen söz öbekleri şu şekilde sonuçlanmıştır.
Healthcare Services = 1919
Retail = 1081
Manufacturing - Other = 885
Computer/IT, Services = 822
Legal Services = 466
Business Services = 410

3.3. Kümeleme İşlemi

Her belge, vektör uzay modeli kullanılarak bir vektör olarak temsil edilir. Vektör modeli olarak da adlandırılan vektör uzayı modeli, metin belgesini (veya genel olarak herhangi bir nesneyi) tanımlayıcı vektörleri olarak temsil eden cebirsel bir modeldir. Örneğin, TF-IDF ağırlığı.
Daha sonra K-Means algoritmasını uygulamak için kümeleme işlemi sırasında belgelerin atandığı bir Centroid sınıfı tanımlanır.[3]

Küme merkezi bir sonraki yineleme için başlatılır, burada sayım değişkeni, kullanıcı tanımlı ilk küme merkezinin değerini tutar.
Bu işlev her belge için en yakın küme merkezi indeksini döndürür, belgenin yakınlığını belirlemek için kosinüs benzerliği kullandım. Benzerlik Measure dizisi, her küme merkeziyle ilgili belge için benzerlik puanını korur, en yüksek puana sahip dizin, verilen belgenin en yakın küme merkezi olarak alınır.



Şekil 1. Clustering oluşturma aracı

Sonuç olarak yazılan program ile iş ilanlarının arasındaki yakınlıklar hesaplanır ve kümeleme işlemi gerçekleştirilir.

4. SONUÇ

İş ilanları ile ilgili sağlıklı bir analiz için öncelikle düzgün ve kullanışlı bir veri seti elde etmek gerekmektedir. Burada en önemli kısım ve düzgün bir veri setinin oluşturulmasıdır. Aksi takdirde ne kadar iyi bir analiz programı olursa olsun sağlıklı sonuçlar vermesi beklenmemelidir. Düzgün bir veri seti ile çalışıldığında ise günümüzde en çok aranan iş ilanları içerisinde terim sıklıkları ve bu iş ilanları arasındaki yakınlıklar ve bağlantılar çıkarılmaktadır. Sonuç olarak bu iş ilanları incelendiğinde çoğunluklu olarak bilişim sistemleri ve teknoloji alanında istihdamın ağırlık kazandığı görülmektedir. Bu tür bir analiz sayesinde kesin bir sonuca ulaşmak mümkün olmasa da günümüzdeki iş ilanları sayesinde gelecekte verilecek olan iş ilanların hangi yöne doğru eğilim gösterdiğini dair çıkarım yapılması sağlanabilir. Bu sayede insanların meslek seçimlerinin bu tür analizlere göre şekillenebilmesi toplumun daha sağlıklı bir iş profili oluşmasına katkı sağlayacaktır.

KAYNAKÇA

[1] Inverse Doküman ve Terim Doküman Sıklığı http://www.primaryobjects.com/2013/09/13/tf-idf-in-c-net-for-machine-learning-term-frequency-inverse-document-frequency/ (Erişim Tarihi:29.12.2019)

[2] Tf-idf Kavramı http://bilgisayarkavramlari.sadievrenseker.com/2012/10/22/tf-idf/ (Erişim Tarihi:29.12.2019)

[3] Tf-idf ve Kmeans ile Dökümanların Kümelenmesi https://www.kaggle.com/jbencina/clustering-documents-with-tfidf-and-kmeans (Erişim Tarihi:31.12.2019)