thinktech
STM Teknolojik Düşünce Merkezi

Büyük Veri İyi Veri Midir? Efsaneler, Mitler ve Gerçekler

Kendisi de büyük veri ile daha da büyüyen Google'a girip "Büyük Veri (Big Data)" kelimesi üzerinden kitaplar içerisinde arama yaptığınızda ortaya çıkacak kitapların büyük bir kısmı "büyük veri menkıbeleri" olarak nitelendirebileceğimiz kitaplardır. Menkıbeler eski zamanlarda yaşamış önemli şahsiyetlerin olağanüstü hikayelerini anlatan ve çoğu kez motive edici hikâyelerdir. Büyük veri menkıbeleri ise büyük verinin büyük nimetlerini yüzlerce farklı hikâye ile ele alan motive edici hikayelerdir. Bu hikâyelerin ortak formu ise şöyledir: "... vadisindeki ... şirketi, yıllardır .... problemini çözmeye çalışırken büyük veri ile ilgili teknolojileri kullanmaya karar verir ve her şey bir anda değişir, büyük kârlar elde edilir, ilginç sonuçlar alınır."

Kuşkusuz bu hikâyeler kurgusal değildir, içlerinde müşterilerinin hamile olduğunu aylar öncesinden günü gününe tahmin edebilen algoritmalar geliştirmiş bebek mağazalarından, uçak motor seslerinden yapay zekâ yardımıyla olası arızaları tahmin edebilen algoritmalar gibi hatırı sayılır harika hikâyeleri de barındıran bu literatürü okudukça büyük veriye dönük saygı gittikçe artmaktadır. Ancak artan şey sadece saygı değil bir ölçüde agnostizmdir de. Büyük veri üzerine yapılan tanımlar halen çok net olmasa da büyük verinin gerçekten kullanıldığı düzlemde kavramlar bu kadar da bulanık değildir. Talihsiz bir isimlendirme veya "büyük" sıfatına aşırı vurgudan olacaktır ki örneğin ne kadarlık bir veri "büyük" veridir sorusuna farklı cevaplar almanız olasıdır. Gigabaytlar, terrabaytlar, hatta petabaytları içeren cevaplar alabilirsiniz. Peki sorun büyüklük müdür?

Büyüklük Görecelidir

Halihazırda "X bayt veri büyük, geri kalanı küçüktür" gibi sınıflandırma girişimleri olsa da bu büyük resmi (yine göreceli bir sıfat kullandık) görmeyi engelliyor. Aslında büyük veri ile ilgili en büyük şey, verinin büyüklüğünden ziyade icra ettiği fonksiyondur. Kuşkusuz çoğu kez (Örneğin bazı makine öğrenme algoritmalarında) veri büyüdükçe icra edilen fonksiyonun gücü artmaktadır ancak bu her zaman böyle olmayabilir.

Sesler de Veridir!

Diğer bir görecelilik kavramı ise yapılandırılmamış veri (unstructured data) olarak adlandırılan verilerle ilgilidir. Veritabanlarımızda satır ve sütunlar halinde tutulan ve genellikle birbiriyle bağlantılı olarak kayıtlı bulunan ilişkisel veriler (relational data) yapılandırılmış veri olarak adlandırılır ve işlemesi kolaydır. Ancak büyük veri çağı bunun dışındaki verileri de -örneğin yazılı metinler, sesler, resimler, fotoğraflar, videolar, el yazısı metinler, sensörlerden gelen veriler gibi birçok veriyi de- "girdi" kabul etmektedir. Hatta büyük veriyi büyük veri yapan konu çoğu kez bu yapılandırılmamış veridir. Yapılandırılmamış verilerin dünyadaki bilinen verilerin yüzde 80'ini teşkil ettiği bilinmektedir. Büyük veri çağı bu kaynakları da neredeyse yapılandırılmış veri tabanlarıyla aynı saygı içinde işlemektedirler. Örneğin milyonlarca haber metnini okuyarak finansal

tahmin yapmaya çalışan ve belirli ölçüde başarılı olan algoritmalar mevcuttur. Eğer Gmail veya Outlook kullanıyorsanız bir mailin spam olduğunu anlayan algoritmalar da metinsel verileri okuyarak (insanlar değil makineler okuyor) gereksiz olanları sizin yerinize ayırıp zamanınızı boşa harcamayı engellemektedir. Buradaki görecelilik şu şekilde ortaya çıkmaktadır: Aslında birkaç megabayt olan bir kitap üzerinde metin madenciliği (text mining) algoritması kullanarak analiz yapmak istediğinizde kitapta en az bir kez geçen her kelime doküman terim matrisi olarak adlandırılan devasa bir matrisin sütunu haline gelmektedir. Bu matriste milyonlarca hücre (çoğu sıfırlardan oluşur) yer alabilir ve illa baytlarla ölçülecekse bile görece küçük olan bu matrisi işlemek bile artık "büyük" bir iştir. Ses ve resim tanıma algoritmaları ise kabul edilebilir doğrulukta bir öğrenme modeli ortaya çıkana değin devasa miktarda öğrenme verisi (training data) kullanmak zorundadır.

Ekosistem

Başka bir deyimle büyük veri hard disk veya sunucularınızda yer kaplayan verilerin gigabayt cinsinden kapladığı yerden ziyade veriye dayalı yeni dünyamızda büyük veri ekosistemi içerisinde yer alan veri madenciliği, istatistik, bilgisayar bilimleri, matematik, paralel veri işleme, öngörücü modeller ve en önemlisi de makine öğrenmesi ya da yapay zekâyı içeren kavranılması kolay olmayan bir dünyadır.

Büyük veriyle ilgili olarak tanımlayıcı sofistike tartışmalar ya da büyük veri menkıbelerinin etkisinde bir süre kaldıktan sonra yapılacak en doğru hamle büyük veriyle ilgili anılan bu ekosistem içerisindeki unsurları tanıyarak organizasyonlar için en uygun çözümü üretmek ve veri çağına adaptasyondur.