VERİ MADENCİLİĞİ KAVRAMI VE GELİŞİM SÜRECİ

Sertaç ÖĞÜT
Görsel İletişim Tasarımı Bölümü, İletişim Fakültesi
Yeditepe Üniversitesi, İSTANBUL

Giriş

Günümüzde bilişim sistemlerinin hayatın hemen hemen her alanında aktif bir rol oynuyor olması ile birlikte; veri, enformasyon, bilgi vb. bir çok kavrama aşina olmuş durumdayız. Özellikle kişisel bilgisayarların günlük yaşamın herhangi bir kademesinde karşımıza çıkıyor olması, bilgisayar dünyasının jargonunu dilimize güncel olarak sokmaktadır. Çevremizdeki bir çok insan "data"lardan "information"lardan söz etmektedir. Türkçeleştirilmiş haliyle veri ve bilgi kelimeleri içimizde yaşayan birer birey halini almış durumda.

Günlük kullanımda sıkça telaffuz edilen bu kavramlar, salt birer kelime olarak kullanıldıklarında anlam daralmasına hatta anlam kaymasına maruz kalmaktadırlar. Oysa kavram olarak ele alındıklarında oldukça önemli bir yere sahip olan bu hece toplulukları, bilişim dünyasının hiç şüphesiz yapı taşı konumundadırlar. Bu sebepledir ki; bu kavramları doğru adreslemek gerekmektedir.

Veri, Enformasyon, Bilgi ve Bilgelik:

Veri kelimesi Latince'de "gerçek, reel" anlamına gelen "datum" kelimesine denk gelmektedir. "Data" olarak kullanılan kelime ise çoğul "datum" manasına gelmektedir. Her ne kadar kelime anlamı olarak gerçeklik temel alınsa da her veri her daim somut gerçeklik göstermez. Kavramsal anlamda veri, kayıt altına alınmış

her türlü olay, durum, fikirdir. Bu anlamıyla değerlendirildiğinde çevremizdeki her nesne bir veri olarak algılanabilir.

Veri, oldukça esnek bir yapıdadır. Temel olarak varlığı bilinen, işlenmemiş, ham haldeki kayıtlar olarak adlandırılırlar. Bu kayıtlar ilişkilendirilmemiş, düzenlenmemiş yani anlamlandırılmamışlardır. Ancak bu durum her zaman geçerli değildir. İşlenerek farklı bir boyut kazanan bir veri, daha sonra bu haliyle kullanılmak üzere kayıt altına alındığında, farklı bir amaç için veri halini koruyacaktır. Bu konuyu daha iyi açıklayabilmek için enformasyon kavramını incelemek gerekmektedir.

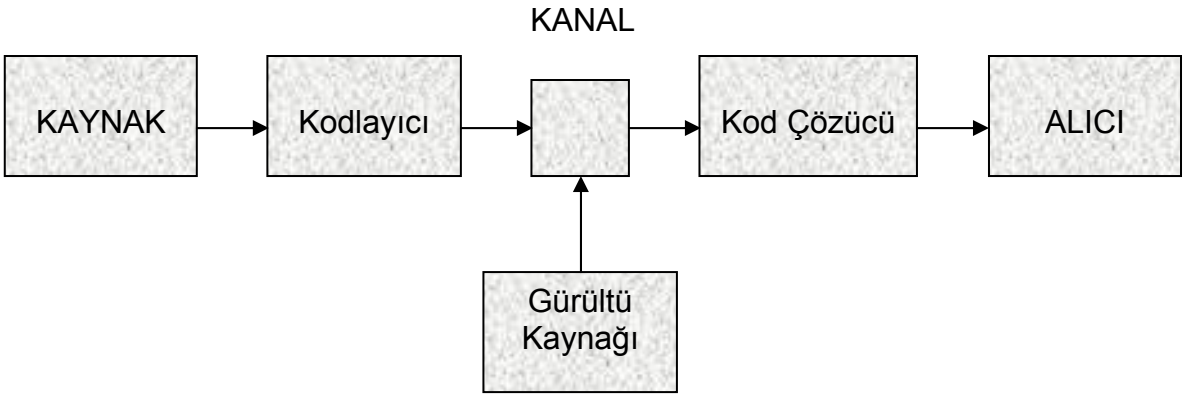
Enformasyon (Information), veri kavramının tanımından yola çıkıldığında, adreslemedeki ikinci safhadır. Yani verilerin ilişkilendirilmiş, düzenlenmiş, anlamlandırılmış, işlenmiş halidir. Bu haliyle enformasyon, potansiyel olarak içinde bilgi barından bir veri halindedir.

Bilgi (Knowledge), bu süreçteki üçüncü aşamadır. Enformasyonun, bilgiye dönüşmesi, bireyin onu algılaması, özümsemesi ve sonuç çıkarmasıyla gerçekleşir. Dolayısıyla bireyin algılama yeteneği, yaratıcılık, deneyim gibi kişisel nitelikleri de bu süreci doğrudan etkilemektedir.

Bilgelik (Wisdom) ulaşılmaya çalışılan noktadır ve bu kavramların zirvesinde yer alır. Bilgilerin kişi tarafından toplanıp bir sentez haline getirilmesiyle ortaya çıkan bir olgudur. Yetenek, tecrübe gibi kişisel nitelikler birer bilgelik elemanıdır.

İletişimsel Enformasyon Kuramı

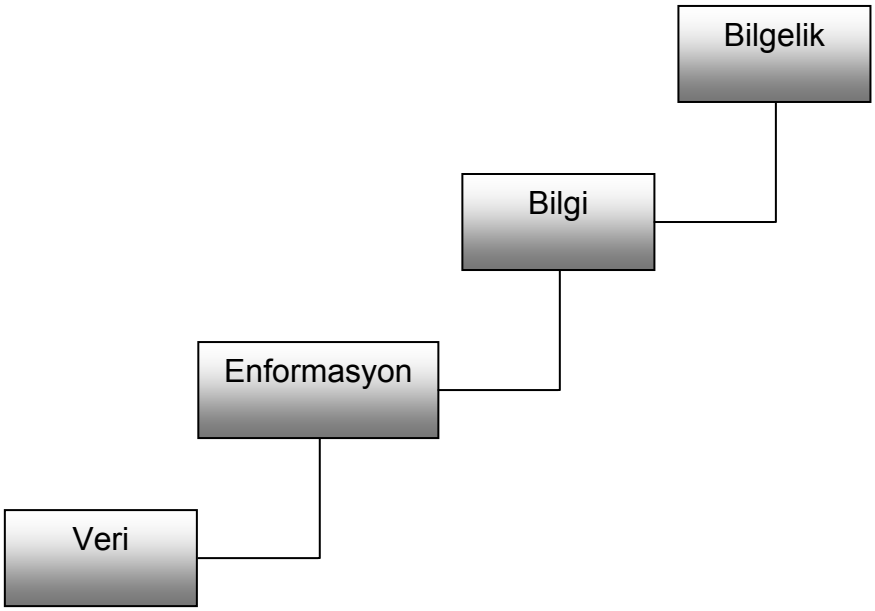
Shannon ve Weaver'in 1949 yılında geliştirmiş oldukları enformasyon kuramı, günümüzde hala bahsi en çok geçen kuramdır. İletişim sürecini bir kaynak ve alıcı arasında açıklayan bu kuramı ayrı ayrı insan ve makine ilişkilerinin her safhasına adapte etmek mümkündür.



Shannon ve Weaver'in iletişim modeli

Bu modelde, ileti kaynakta oluşur ve kodlayıcıda kodlanarak kanal vasıtasıyla alıcının kod çözücüsüne ulaştırılır. Burada çözülen ileti, alıcıya verilir ve iletişim gerçekleşir. Kanal içerisinde yol alan sinyal (ileti) gürültü kaynağı tarafından bir etkiye maruz kalabilir. Bu etkinin boyutları, iletişimin verimlilik boyutları ile ters orantılıdır. Gürültü kaynağının sinyal üzerindeki etkisi arttıkça, iletişim sürecindeki verim azalacaktır.

Bu iletişim sürecinde veri, bir olgu olarak kaynağın bir öğesidir. Yani kaynak olmazsa veri de olmayacaktır. İleti kanalın bir fonksiyonu olarak göze çarparken, enformasyon, alıcıya ulaşan iletinin son halidir ve alıcının bir fonksiyonudur. Alıcıya ulaşan enformasyon, burada anlam kazandırılarak bilgiye dönüştürülecektir. Sürecin sürekli tekrarlanıyor olması sonucunda, topladığı bilgileri sentezleyen alıcı bilgelik sahibi olacaktır.



Veri ile başlayan ve Bilgelik ile tamamlanan süreç

Bilginin önemi

Günümüzde bilgelik en değerli varlıktır. Sanayi toplumlarında üretim ve iş gücünün para ile satın alınabilmesi, parayı en değerli varlık haline getirmekteydi. Ancak bugün, bilgelik paranın tahtını almış durumdadır. Bilgi her daim paraya / faydaya dönüştürülebilirken, para her zaman bilgiyi almaya yeterli olmayabilir.

Bilginin bu değerli durumu 1980'lı yılların ortalarında hizmet sektörünün, üretim sektörünün önüne geçmeye başlamasıyla iyice belirginleşmiştir. Daniel Bell'in *The Coming of Post-Industrial Society* isimli çalışmasında ABD'deki aktif nüfüsun 1870 ve 1980 yılları arasındaki sektörel dağılımı incelenmiştir. Buna göre 1900'lü yılların başında Tarım sektörü %37.5, üretim sektörü %31.4 ve hizmet sektörü % 31.1 lik bir paya sahip iken bu oranlar 1960lı yıllarda %5.1, %30.8 ve %64.1 gibi bir dağılıma ulaşmıştır. 1980'li yıllara bakıldığında hizmet sektörü tek başına yaklaşık %70'lik bir paya sahip olarak, bilginin değerini keskin bir biçimde ortaya koyar bir hale gelmiştir. Avrupa'daki durum da ABD'dekinden çok daha farklı bir grafik çizmemektedir.

Bilginin bu denli değerli olması, bilişim teknolojilerinin gelişmesine ön ayak olmuştur. Bilgisayarların bilgi yönetiminde ve üretimde faal olarak yer almaya başlaması kaçınılmaz bir durum haline gelmiştir. Günümüze bakıldığında bir bilgi patlaması söz konusudur. Çevremizin verilerle dolu olması peşi sıra enformasyon ve bilgiyi beraberinde getirmektedir. Internet gibi etkili bir iletişim ortamının varlığı bu durumu körüklemektedir. Makro düzeyde bakıldığında hemen hemen herkes bu veri dağına bir katkıda bulunmakta ve de bundan yararlamaktadır. Ancak bu yanında bazı sorunlar da getirmektedir. Bu kadar çok veri arasından gereken bilgiyi çıkartmabilmek gerekmektedir. Bu aşamada yeni bir kavram karşımıza çıkmaktadır: Veri Madenciliği.

Veri Madenciliği

Günümüzde sadece bilgiye ulaşmak değil, gerekli koşullarda bilgi üretmek de önemli bir konu halini almıştır. Çığ gibi büyüyen sayısal veri ortamları arasından yararlı ve de gerekli olan bilgiye ulaşmayı sağlamak gerçek bir çaba haline gelmiştir. Veri madenciliği bu safhada göze çarpan bir olgudur. *Frawley* veri madenciliğini "*Daha önceden bilinmeyen ve potansiyel olarak yararlı olma durumuna sahip verinin keşfedilmesi*" olarak tanımlamıştır. *Berry ve Linoff* bu kavrama "*Anlamlı kuralların ve örüntülerin bulunması için geniş veri yığınları üzerine yapılan keşif ve analiz işlemleri*" şeklinde bir açıklama getirirken *Sever ve Oğuz* çalışmalarında veri madenciliği hakkında "*Önceden bilinmeyen, veri içinde gizli, anlamlı ve yararlı örüntülerin*[*] *büyük ölçekli veritabanlarından otomatik biçimde elde edilmesini sağlayan veri tabanlarında bilgi keşfi süreci içerisinde bir adımdır.*" tanımını kullanmışlardır.

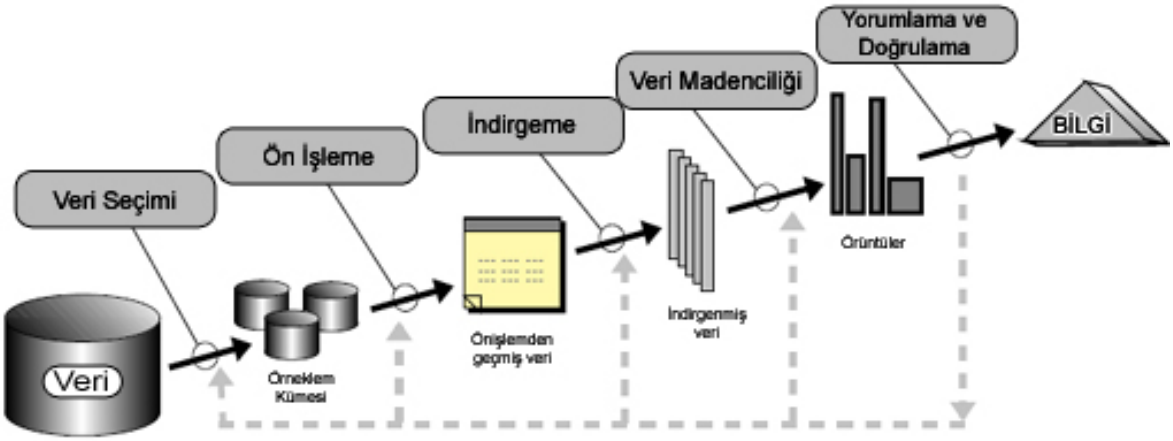
Bu tanımlamaları da göz önünde bulundurarak veri madenciliği kavramına şöyle bir yaklaşım getirmek mümkündür:

[*] İngilizce'de Pattern, Almanca'da Muster ya da Flachenmuster, Fransızca'da Image, Figure, Mosaique kelimelerine eş anlamlı; belirli, ancak bilinmeyen bir sınıfta yer alan nesne ve ya olaylar.

Veri madenciliği, geniş veri yığınları içerisinde, yararlı olma potansiyeline sahip, aralarında beklenmedik / bilinmedik ilişkilerin olduğu verilerin keşfedilerek, veri sahibi için hem anlaşılır hem de kullanılabilir bir biçime getirilmesine yönelik geliştirilmiş yöntemler topluluğudur.

Bahsi geçen bu yöntemler karar verme sürecinde oldukça etkili rol oynamaktadırlar. Nihayetinde amaç bilgiyi keşfederek ona ulaşmak ve bu yolla fayda sağlamaktır.

Veri madenciliği, aynı zamanda bir süreçtir. Veri yığınları arasında, soyut kazılar yaparak veriyi ortaya çıkarmanın yanı sıra, bilgi keşfi sürecinde örüntüleri ayrıştırarak süzmek ve bir sonraki adıma hazır hale getirmek de bu sürecin bir parçasıdır.



Bilgi Keşfi sürecinde veri madenciliğinin yeri

Veri Madenciliğinin Gelişim Süreci

Veri madenciliğinin kökeni hiç şüphesiz ilk sayısal bilgisayar olan ENIAC (**E**lectrical **N**umerical **I**ntegrator **A**nd **C**alculator)'a kadar dayanmaktadır. 1946 yılında geliştirilen ve bugün kullandığımız kişisel bilgisayarların atası olan ENIAC, ABD'li bilimadamları John Mauchly ve J. Presper Eckert tarafından, II. Dünya Savaşı sırasında ABD ordusu için geliştirildi. 30 tonluk ağırlığıyla 170 m2lik bir alanı kaplayan bu "ilk" bilgisayarın 60 sene içerisinde geçirmiş olduğu evrimin nihai boyutlarını şu anda masa üstünüzdeki bilgisayara bakarak anlamanız mümkündür.

Bu evrim tabii ki belli bir süreç ve şartlar altında gerçekleşti. Donanımsal olarak hazırlanan bilgisayarların, yazılımlar vasıtasıyla hayat bulması ve kullanıcılara ulaştırılması, bilgisayar evrim döngüsünün anahtarıdır. Bilgisayar ve yazılım uzmanlarının geliştirdikleri ürünler, kullanıcıların istekleri doğrultusunda zamanla şekillenerek bugünkü halini almış durumdadır. Döngü, donanımın geliştirilmesinin ardından yazılımın bu donanıma entegre edilerek kullanıcıya ulaştırılmasıyla başlar. Kullanıcı ihtiyaçları doğrultusunda yazılımda bulunan eksiklikleri belirler. Yazılım uzmanları bu eksiklikleri göz önünde bulundurarak yeni yazılımlar geliştirirler. Bu yazılımların çalışabilmesi için gerekli donanım güncellenmesinin yapılması için donanım uzmanları uyarılır. Güncellenen bilgisayarlar tekrar kullanıcılara ulaştırılır ve döngü bu şekilde devam eder.

Dikkat edilirse döngünün anahtar elemanı kullanıcıdır. Kullancıların ihtiyaçları, isteklerini belirler. Dolayısıyla bu istekler mevzu bahis sektörü doğrudan etkiler ve gelişmenin kapıları açılır.

Bilgisayarların efektif kullanımı verilerin depolanması ile başlamaktadır. İlk haliyle karmaşık hesaplamaları yapmaya yönelik geliştirlen bilgisayarlar, kullanıcı ihtiyaçları doğrultusunda veri depolama işlemleri için de kullanılmaya başlandı. Bu sayede veri tabanları ortaya çıktı. Veri tabanlarının genişleme trendi içinde olması donanımsal olarak bu verilerin tutulacakları ortamların da genişlemesini gerektirdi. Veri ambarı kavramının ortaya çıkışı bu dönemlere rastlamaktadır. Kaybedilmek istemeyen veriler, bir ambar misali fiziksel sürücülerde tekrar kullanılmak üzere

saklanmaktaydı. Gittikçe büyüyen veri tabanlarının organizasyonu, düzenlenmesi ve yönetimi de buna paralel olarak güç bir hal almaya başladı. Bu safhada veri modelleme kavramı ortaya çıktı.

İlk olarak basit veri modelleri olan Hiyerarşik ve Şebeke veri modelleri geliştirildi. Hiyerarşik veri modelleri, ağaç yapısına sahip, temelinde bir kök olan ve bu kök vasıtasıyla üstünde her daim bir, altında ise n sayıda düğüm bulunan veri modelleriydi. Şebeke veri modelleri ise kayıt tipi ve bağlantıların olduğu, kayıt tiplerinin varlık, bağlantılarınsa ilişki tiplerini belirlediği bir veri modeliydi. Şebeke veri modelinde herhangi bir eleman bir diğeri ile ilişki içerisine girebiliyordu. Ancak çoklu ilişki kurmak söz konusu değildi. Hiyerarşik veri modellerinde ise bu daha da kısıtlıydı. Dolayısıyla kullanıcıların ihtiyaçlarını tam olarak karşılayamadılar. Bu ihtiyaçlar doğrultusunda Geliştirilmiş Veri Modelleri geliştirildi. Bunlar Varlık – İlişki, İlişkisel ve Nesne – Yönelimli veri modelleri olarak bilinmektedirler. Günümüzde en sık kullanılanı İlişkisel veri modelidir. Nesne – Yönelimli veri modelleri ise hala gelişim süreci içerisindedir.

İhtiyaçlar doğrultusunda şekillenen veri tabanları ve veri modelleme çeşitleri hızla yaygınlaşırken, donanımlar da bu sürece ayak uydurdular. Günümüzde milyarlarca bit veriyi ufacık belleklerde tutmak mümkün hale gelmiştir. İhtiyaçlar her ne kadar teknolojiyi ciddi anlamda şekillendirse de yanında sorunları daim olarak getirmektedir. Verileri saklanması, düzenlenmesi, organize edilmesi her ne kadar bir sorun gibi görünmese de bu kadar çok veri ile istenilen sonuca ulaşmak başlı başına bir sorun halini almıştır.

Veri madenciliği, kavramsal olarak 1960lı yıllarda, bilgisayaların veri analiz problemlerini çözmek için kullanılmaya başlamasıyla ortaya çıktı. O dönemlerde, bilgisayar yardımıyla, yeterince uzun bir tarama yapıldığında, istenilen verilere ulaşmanın mümkün olacağı gerçeği kabullenildi. Bu işleme veri madenciliği yerine önceleri veri taraması (data dredging), veri yakalanması (data fishing) gibi isimler verildi.

1990lı yıllara gelindiğinde veri madenciliği ismi, bilgisayar mühendisleri tarafından ortaya atıldı. Bu camianın amacı, geleneksel istatistiksel yöntemler yerine, veri analizinin algoritmik bilgisayar modülleri tarafından değerlendirmesini vurgulamaktı. Bu noktadan sonra bilimadamları veri madenciliğine çeşitli yaklaşımlar getirmeye başladılar. Bu yaklaşımların kökeninde istatistik, makine öğrenimi (machine learning), veritabanları, otomasyon, pazarlama, araştırma gibi disiplinler ve kavramlar yatmaktaydı.

İstatistik, süre gelen zaman içerisinde verilerin değerlendirilmesi ve analizleri konusunda hizmet veren bir yöntemler topluluğuydu. Bilgisayarların veri analizi için kullanılmaya başlamasıyla istatistiksel çalışmalar hız kazandı. Hatta bilgisayarın varlığı daha önce yapılması mümkün olmayan istatistiksel araştırmaları mümkün kıldı. 1990lardan sonra istatistik, veri madenciliği ile ortak bir platforma taşındı. Verinin, yığınlar içerisinden çekip çıkarılması ve analizinin yapılarak kullanıma hazırlanması sürecinde veri madenciliği ve istatistik sıkı bir çalışma birlikteliği içine girmiş bulundular.

Bunun yanısıra veri madenciliği, veri tabanları ve makine öğrenimi disipliniyle birlikte yol aldı. Günümüzdeki Yapay Zeka çalışmalarının temelini oluşturan makine öğrenimi kavramı, bilgisayarların bazı işlemlerden çıkarsamalar yaparak yeni işlemler üretmesidir. Önceleri makineler, insan öğrenimine benzer bir yapıda inşa edilmeye çalışıldı. Ancak 1980lerden sonra bu konuda yaklaşım değişti ve makineler daha spesifik konularda kestirim algoritmaları üretmeye yönelik inşa edildi. Bu durum ister istemez uygulamalı istatistik ile makine öğrenim kavramlarını, veri madenciliği altında bir araya getirdi.

Sonuç

Günümüzde veri madenciliği bir çok alanda kullanılmakta. Operasyonel kararların ötesinde, stratejik ve politik karar verme süreçlerinde önemli bir yere sahip. Gerek özel sektör gerekse kamusal sektör, Müşteri İlişkileri Yönetimi (CRM) Kurumsal Kaynak Planlaması (ERP) gibi çeşitli uygulamalar ve teknikler vasıtasıyla veri madenciliği yapmaktadır. İstatistik ile olan yakın ilişkisi veri madenciliğini tıp ve ekonomi gibi bilimdalları için de önemli kılmaktadır. Bilginin bu kadar değerli olduğu çağımızda, bilgiye ulaşmak için katedilen yolda veri madenciliği oldukça önemli bir safhadır. Zira bu kadar yoğun bir veri akışının altında bilgiyi keşfetmek yalnızca bilimsel bir çaba değil aynı zamanda bir sanat halini almıştır.

Kaynaklar:

- Akpınar, H., Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği, İ.Ü. İşletme Fakültesi Dergisi, C:29, S:1 / Nisan 2000

- Alpaydın, E., Zeki Veri Madenciliği, Bilişim 2000 Eğitim Semineri, 2000

- Chen, M., Han, J., Yu, P., Data Mining: An Overview from Database Perspective

- Collier, K., Carey, B., Grusy, E.; Marjaniemi, C.; Sautter, D., A Perspective on Data Mining, North Arizona University, 1998

- Fayyad, U., Piatetsky-Shapiro, G., Smyht P., From Data Mining to Knowledge Discovery in Databases, AAI, 1996

- Hand, D., Mannila, H., Smyth, P., Principles of Data Mining, The MIT Press, 2001

- Kantardzic, M., Data Mining: Concepts, Models, Methods, and Algorithms, Wiley Pub., 2003

- Mascorola, J., Bolden, R., FROM THE DATA MINE TO THE KNOWLEDGE MILL, 2nd European Symposium on Principles of Data Mining and Knowledge Discovery, 1998

- Nilsson, N. J., Introduction to Machine Learning, 1996

- Rud, O.P., Data Mining Cookbook, Wiley Pub., 2001

- Sever, H., Oğuz, B., Veritabanlarında Bilgi Keşfine Formal Bir Yaklaşım

- Smyth, P., Data Mining Data Analysis on a Grand Scale, UC, 2000

- Sütçü, C, İstatistiksel Veri Sistemleri ve Basın Sektöründe bir Karar Destek Sistemi Uygulaması, Marmara Üniversitesi, 1995

- Sütçü, C., Information Systems Concept – Electronic Communication and Information Systems Approach, 2nd International Symposium of Interactive Media Design, 2004

- Theus, M., What Dataminers Want, 2nd International Workshop on Distributed Statistical Computing, 2001

- Xiao, Z., Statistics and Data Mining, National University of Singapore

- Yao, Y., A Step Towards the Foundations of Data Mining, University of Regina

- Zaiane, O., Fall, A., Rochefort, S., Dahl, V., Tarau, P., Concept Based Retrieval using Controlled Natural Language