

Çok Boyutlu Veri Görselleştirme Teknikleri

T. Tugay BİLGİN¹, A. Yılmaz ÇAMURCU²

¹ Maltepe Üniversitesi, Bilgisayar Mühendisliği Bölümü, İstanbul

² Marmara Üniversitesi, Bilgisayar ve Kontrol Eğitimi Bölümü, İstanbul
ttbilgin@maltepe.edu.tr, camurcu@marmara.edu.tr

Özet: Bu çalışmada, veri madenciliğinde güncel araştırma alanlarından biri olan çok boyutlu veritabanları ve bunların görselleştirilmesinde kullanılan görselleştirme teknikleri incelenmiş ve bu alanda çalışmalar gerçekleştiren araştırma grupları ve bunların geliştirdikleri yeni yöntemler ve teknikler irdelenmiştir.

Anahtar Sözcükler: Veri madenciliği, Veritabanı, Görselleştirme.

High Dimensional Data Visualization Techniques

Abstract: In this paper, high dimensional databases and high dimensional data visualization techniques which are current research areas on data mining are examined. Data visualization research groups and new techniques and methods on high dimensional visualization are briefly explained.

Keywords: Data mining, Databases, Visualization.

1. Giriş

Birçok veri madenciliği uygulamasında verilerin birbiri ile olan ilişkilerinin iyi anlaşılması büyük önem taşır. Bunu gerçekleştirmek için en iyi yol verinin görselleştirilmesidir. Veri görselleştirme teknikleri, bilgisayar grafikleri, görüntü işleme, bilgisayar görüşü (computer vision), kullanıcı arayüzü tasarımı gibi birçok bilim dalının birleşiminden oluşur. Bu teknikler sayesinde bankalar, sayısal kütüphaneler, İnternet siteleri ve metin veritabanları gibi büyük veritabanlarının görselleştirilmesi mümkün olmaktadır.

Veri görselleştirme, insanın algılama yetenekleri ve insanlar arası yorumlama farklarını dikkate alarak analiz gerçekleştirmeye olanak verir. Veri görselleştirme teknikleri ile etkili bir biçimde verinin portresinin çıkarılması sağlanabilir ve veri hakkında genel bir kanıya varılabilir [1, 2, 3].

2. Çok Boyutlu Veritabanları

Çok boyutlu veritabanları bilgi keşfi (information retrieval), görüntü işleme, veri madenciliği, örüntü tanıma ve karar destek sistemleri gibi birçok uygulama alanında önem kazanmaktadır. Günümüzde Veritabanı yönetim sistemleri eski örneklerine göre çok daha karmaşıktır. Modern uygulamalarda veritabanı kavramı yalnızca ilişkisel veya nesne yönelimli olarak iki türe değil, uygulama alanlarına özel birçok farklı türe ayrılmaktadır [15].

2.1. Çokluortam veritabanları (Multimedia Databases)

Çokluortam veritabanları birçok farklı biçimde görüntü, ses ve video verileri içerirler. Fotoğrafik görüntüler, uydu görüntüleri, uzaktan algılama resimleri (remotely sensed images) [16], tıbbi görüntüler (iki boyutlu X ışınları ve üç boyutlu beyin MRI taramaları),

jeolojik görüntüler, biyometrik tanımlama (biometric identification) görüntüleri (parmak izi, retina gibi [17]) gibi farklı çokluortam verileri depolamak üzere özelleştirilmiş birçok uygulama bulunmaktadır. Bu uygulamalarda amaç, hedef olarak seçilmiş bir nesneye en fazla benzeyen nesnelere bulmaktır. Bu sebeple her görüntü renk, şekil, desen gibi özelliklerden oluşan özellik vektörlerine (feature vectors) dönüştürülür. Benzerlik (similarity), özellik vektörleri arasındaki uzaklık hesaplanarak bulunur.

2.2. Zaman serileri veritabanları

Bu veritabanları finansal, tıbbi ve bilimsel verilerin analizinde, veri madenciliğinde ve karar verme sürecinde kullanılırlar. Zaman serileri veritabanları zaman serisi şeklindeki verileri Ayrık Fourier Dönüşümü (Discrete Fourier Transform) [18] veya Ayrık Dalgacık Dönüşümü (Discrete Wavelet Transform) [19] gibi dönüşüm yöntemleri ile çok boyutlu noktalara dönüştürürler. Benzerlik arama işlemi dönüştürülmüş veriler üzerinde gerçekleştirilir.

2.3. DNA veritabanları

Genetik materyal (DNA) bir canlının tüm hücrel fonksiyonları için gerekli tüm bilgileri depolamaktadır. DNA, dört harfli alfabeti olan bir metin dizisidir. Bu dört harf A,C,G ve T olarak dört farklı çeşit nükleotidi temsil eder. Yeni bir metin dizisi (örneğin bilinmeyen bir hastalığa ait olabilir), eski dizilerin herhangi bir bölümü eşleştirilmeye çalışılır. Eşleştirmenin amacı belirli bir uzaklık fonksiyonu kullanılarak aranan metne en fazla uyan bölümü bulmaktır.

2.4 Doküman veritabanları

Bu veritabanları çoğunlukla belirli bir dile ait kelimeler veya metinlere ait özellik vektörleri içerirler. Çok fazla sayıda boyuta sahip olabilirler. İnternet'in doğuşu ile birlikte gelişme göstermiştir. İnternet arama motorları,

on-line veritabanları, doğal dil işleme, doküman sınıflandırma gibi alanlarda yoğun olarak kullanılmaktadır.

Yukarıda açıklanan veritabanları çok boyutlu veri nesnesi şeklinde temsil edilen ve sayısal verilerden oluşan özellik vektörlerine sahiptir. Bu yüzden bu tür veritabanlarına genel olarak "çok boyutlu veritabanı" adı verilir. Çok boyutlu veri tabanları, anahtar (key) ifade tabanlı geleneksel sorgular yerine "benzerlik tabanlı" (similarity based) veya içerik tabanlı bilgi çekme (content based retrieval) sorgularına gereksinim duyarlar. Bu tür veri tabanlarında benzer örüntüler arama süreci büyük önem taşır. Çünkü bu süreç tahmin etme, karar verme, bilgisayar destekli tıbbi muayene, hipotez doğrulama ve veri madenciliği için kritik öneme sahiptir [20].

3. Veri Görselleştirmede Amaç

İnsanın algılama sistemi yalnızca 3 boyut ile sınırlı olduğu için daha fazla boyut içeren veriler insanın algılama sınırını aşmaktadır. Veri görselleştirme teknikleri çok boyutlu veriyi 2 veya 3 boyuta indirgeyerek görselleştirirken, diğer taraftan da veriler arasındaki ilişkiyi muhafaza edebilmelidir. Bu indirgeme sırasında bir miktar kayıp olması kaçınılmazdır. Görselleştirmede temel hedeflerden biri bu kaybı minimum düzeyde tutmaktır.

Veri görselleştirmenin iki temel amacı bulunmaktadır. Birinci amaç fikirlerin, kuralların ve kavramların daha iyi anlaşılmasıdır. Tüm bunlar bir bilgi olduğu için bu tür görselleştirmelere "bilgi görselleştirmesi" (knowledge visualization) denir. Diğer amaç ise grafiklerin ve resimlerin yeni fikirler oluşturmak, yeni ilişkiler kurmak, bir hipotezin doğruluğunu sınamak, yeni yapılar keşfetmek veya bu yapıları düzenlemektir. Özetle, bu işlemler insanın görsel algılama sistemini mantıksal problemlerin çözümü için kullanmaktır [20]. Bu tür görselleştirmelere "veri görselleştirme" (data

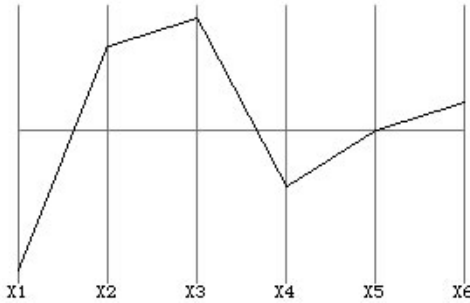
visualization) veya “görsel bilgi keşfi” (visual data exploration) [7] denir.

4. Çok Boyutlu Veri Görselleştirme Teknikleri

Çok boyutlu veri görselleştirme araçları, bu alandaki birçok çalışmaları ile tanınan Kriegel [7] ve Keim [6] tarafından altı temel sınıfa ayrılmıştır. Bunlar, geometrik izdüşüm teknikleri, ikon tabanlı teknikler, piksel tabanlı teknikler, hiyerarşik teknikler, graf tabanlı teknikler ve karma teknikler olarak literatüre girmiştir.

4.1. Geometrik İzdüşüm Tabanlı Teknikler

Bu tür tekniklerin en bilineni iki boyutlu veri setini x ve y eksenleri boyunca kartezyen koordinat sistemine işaretleyen saçılım grafikleridir (scatterplots).



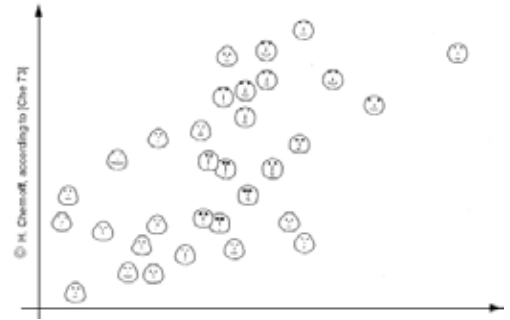
Şekil 1. Altı boyutlu $\{-5,3,4,-2,0,1\}$ bir veri nesnesinin paralel koordinatlar tekniği ile görselleştirilmesi [7].

Paralel Koordinatlar [7], k-boyutlu veri setini 2 boyutlu uzaya haritalayan görselleştirme tekniği Şekil 1’de görüldüğü gibi k adet birbirine paralel konumlandırılmış eksenlerden oluşur. Her eksen veri setine ait bir alan ile ilişkilendirilmiştir. Bir alandaki değer aralığı, o alana ait eksen üzerinde ölçeklenmiştir. Her eksen üzerindeki değer işaretlendikten sonra bu değerler düz çizgiler ile birleştirilir. Bu tek-

niğin en büyük dezavantajı birkaç bin adetten daha fazla nesne içeren veri setleri için uygun olmamasıdır. Nesne sayısı arttıkça üst üste binen çok sayıda çizgi görüntüyü yorumlanabilir olmaktan çıkarmaktadır.

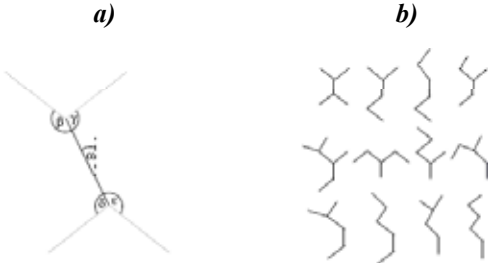
4.2. İkon Tabanlı Teknikler

İkon tabanlı teknikler her birçok boyutlu veri nesnesini bir ikon şeklinde sembolize ederler. İkonun her bir görsel özelliği verinin içerdiği değerlere göre değişir. Bu türün ilk örneklerinden biri Chernoff yüzleri tekniğidir [8]. Her veri nesnesi için bir insan yüzü çizilir. Nesneye ait ilk iki boyut yüz resminin 2 boyutlu düzlemdeki konumu belirtir. Diğer boyutların aldığı değerler ile orantılı olarak insan yüzünün burun, ağız, kulak, göz ve yüz şekli değiştirilir (Şekil 2). Bu tekniğin en büyük dezavantajı insan yüzündeki bazı organların diğerlerine göre daha fazla dikkat çekmesidir. Örneğin gözler kulaklardan daha dikkatli algılandığı için karşılaştırma yanılgıları oluşabilir [6].



Şekil 2. Chernoff yüzleri [8]

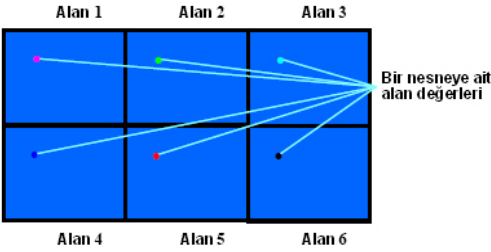
Çubuk şekiller (stick figures) tekniği ise beş kollu çubuk şeklinde ikonlar kullanır [9]. Şekil 3.a’da bir çubuk şekil ve Şekil 3.b’de çubuk şekiller ailesi ile 12 adet veri nesnesi görselleştirilmiştir. Veri nesnesinin ilk iki özelliği çubukların ebatını belirlemede, diğer özellikler ise ikonun kollarının açısını belirlemede kullanılır.



Şekil 3. Çubuk şekiller [9].

4.3. Piksel Tabanlı Teknikler

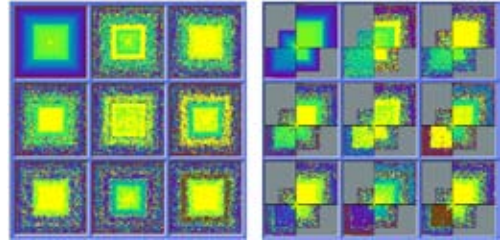
Piksel tabanlı tekniklerde her bir boyuta ait değer renkli bir piksel ile temsil edilir. Şekil 4’de altı boyutlu bir verinin piksel tabanlı görselleştirilmesi görülmektedir.



Şekil 4. Altı boyutlu bir verinin piksel tabanlı görselleştirilmesi [10].

Her boyut ayrı bir dikdörtgen alt pencere içinde konumlandırılarak sahip olduğu değer ile orantılı bir renk ile temsil edilmektedir [10]. Bu teknik çok boyutlu büyük veri setlerinin görselleştirilmesi için elverişlidir.

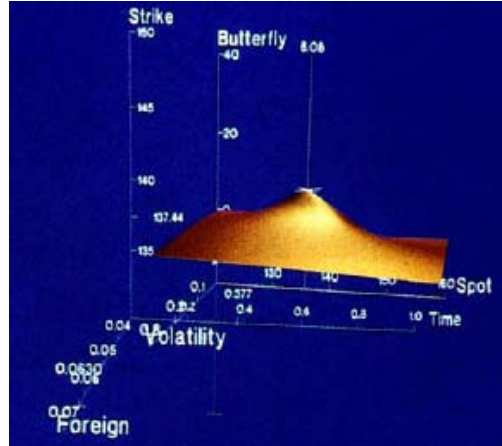
Şekil 5’de 7000 noktadan oluşan 8 boyutlu ve 6 küme içeren sentetik bir veri seti için spiral dizilim ve eksenlere göre dizilim şeklinde iki farklı sorgu bağımlı görselleştirme görülmektedir. Her bir veri noktasının sorgu noktasına uzaklığını parlak sarıdan yeşile, mavi, koyu kırmızı ve siyah renklerle kodlanmıştır. Sorgu noktasına en yakın olan nesnelere parlak sarı, en uzak olanlar ise siyah ile gösterilmiştir.



Şekil 5. 7000 noktadan oluşan 8 boyutlu ve 6 küme içeren sentetik veri seti için spiral dizilim (solda) ve eksenlere göre dizilim (sağda) şeklinde iki farklı sorgu bağımlı görselleştirme [10].

4.4. Hiyerarşik Teknikler

Hiyerarşik teknikler k-boyutlu uzayı alt uzaylara ayırırlar ve bunları hiyerarşik olarak görüntülemeyi sağlarlar. Bu türün en önemli temsilcilerinden biri n-Vision veya diğer adı ile “dünya içinde dünyalar” (worlds-within-Worlds) [11] adlı sistemdir.



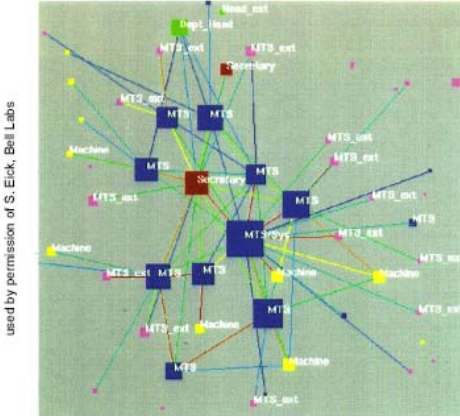
Şekil 6. Altı boyutlu uzayın n-Vision ile görselleştirilmesi [12].

n-Vision aracı k-boyutlu uzayı birçok üç boyutlu alt uzaya ayırarak görselleştirir. Şekil 6’da altı boyutlu uzayın görüntülenmesi görülmektedir. İlk üç boyut dış koordinat sistemi ile sonraki üç boyut ise iç koordinat sistemi ile gösterilmiştir.

4.5. Graf Tabanlı Teknikler

Graf tabanlı teknikler özel yerleşim algoritmaları, sorgulama dilleri ve soyutlama teknikleri kullanarak etkili graflar oluştururlar. Bu alandaki en önemli araçlar Hy+ ve SeeNet araçlarıdır.

Hy+, yapısal veri setlerini görselleştirmek için kullanılan sorgulama ve görselleştirme sistemidir [4]. Bu araç web sörf oturumları e-posta transferleri gibi verilerin görselleştirilmesinde kullanılır.



Şekil 7. Kurum içi e-posta mesajlarının SeeNet ile görselleştirilmesi [12].

SeeNet, hiyerarşik ağların bağlantı ağırlıkları kullanılarak görselleştirilmesini sağlayan bir araçtır [13]. Bu araç anlamsal düğüm yerleşime (semantic node placement), yüksek ağırlıklı bağlantılar arasındaki uzaklıkları en aza indirme gibi özelliklere sahiptir. Şekil 7’de bir işyerinde belirli bir zaman dilimi içerisinde gerçekleşen e-posta bağlantıları görselleştirilmektedir [12]. Şekilde düğümlerin boyu bir kişiye ait olan e-posta sayısını, düğümlerin rengi personelin işyerindeki pozisyonunu, bağlantının kalınlığı ise iki düğüm arasındaki e-posta trafiğinin büyüklüğünü göstermektedir.

4.6. Karma Teknikler

Karma teknikler görselleştirmenin açıklayıcı niteliğini arttırmak için birden fazla görselleştirme tekniğini bir veya daha fazla pencere içerisinde kullanırlar. Görüntüleme farklı pencereler içerisinde yapıldığında pencereler arasında bağlantı kurmak için farklı etkileşimler ve dinamik yöntemler kullanmak gereklidir. Bu konuda kullanılacak yöntemler [14] numaralı referansta incelenebilir.

4. Sonuç

Bu çalışmada çok boyutlu ve çok büyük veritabanlarında etkin olarak çalışabilen görselleştirme araçları incelenmiştir. Bu sistemler, yüksek ölçeklenebilirlik özellikleri ile gelecek yıllarda daha da artacak olan boyut ve büyüklüklere uyum sağlamakta geleneksel yöntemlere göre daha avantajlıdır.

9. Kaynaklar

- [1] Carlis, J.V. ; Konstan, J.A.; “Interactive Visualization of Serial Periodic Data.” In UIS-T’98 Conference Proceedings. New York, NY: ACM Press, USA (1998) 29-38.
- [2] Derthick, M.; Kolojechick, J.; Roth, S. F.; “An interactive visualization environment for data exploration”. In Proc. of KDD-97, Kanada (1997) 2-9.
- [3] Keim, D.A; Kriegel, H.P.; “Visualization Techniques for Mining Large Databases: A Comparison”, IEEE Trans. Knowledge and Data Eng., vol. 8, no. 6, USA (1996) 923-938.
- [4] Card, S.K.; Mackinlay, J.D.; Shneiderman, B.; “Readings in Information Visualization: Using Vision to Think”, Morgan Kaufmann Publishers, San Francisco, USA (1999).

- [5] Bertin, J.: “Graphics and Graphic Information Processing”, De Gruyter, Berlin, Germany (1981).
- [6] Keim, D.A.: “Visual Database Exploration Techniques”, Proc. Tutorial KDD '97 Intl. Conf. Knowledge Discovery and Data Mining, California, USA, (1997).
- [7] Inselberg, A.; Dimsdale, B.: “Parallel Coordinates: A Tool for Visualizing Multidimensional Geometry”, Proc. IEEE Visualization'90, USA (1990) 361-375.
- [8] Chernoff, H.: “The Use of Faces to Represent Points in k-Dimensional Space Graphically”, Journal of American Statistical Assoc., vol. 68, USA (1973) 361-368.
- [9] Pickett, R.M.; Grinstein G.G.: “Iconographic Displays for Visualizing Multidimensional Data,” Proc. IEEE Conf. Systems, Man, and Cybernetics, (1988) 514-519.
- [10] Keim, D.A.; Kriegel, H.P.: “VisDB: Database Exploration Using Multidimensional Visualization,” IEEE Computer Graphics and Applications, vol. 14, no. 5, USA (Eylül 1994) 40-49.
- [11] Feiner, S.; Beshers, C.: “Worlds within Worlds: Metaphors for Exploring n-Dimensional Virtual Worlds”, Proc. of User Interface Software and Technology, (1990), 76-83.
- [12] Keim, D.A.: “Visual Data Mining”, Tutorial Notes, Proc of VLDB, Atina, Yunanistan, (1997).
- [13] Becker, R.A.; Eick, S.; Wilks, A.R.: “Visualizing Network Data”, Trans. On Visualization and Computer Graphics 1(1), (1995), 16-28.
- [14] Cristina, M.; Oliveira, F.D.; Levkowitz, H.: “From visual data exploration to visual data mining: a survey.”, IEEE Transactions on Visualization and Computer Graphics, 9(3), (2003), 378-394.
- [15] Li, Y.: “Efficient Similarity Search in High Dimensional Data Spaces”, Doktora Tezi, New Jersey Institute of Technology, Department of Computer Science, (2004).
- [16] Richards, J.: “Remote Sensing Digital Image Analysis, An Introduction”, Wiley and Sons, New York, USA (1993).
- [17] Jain, A.; Lin, H.; Pankanti, S.; Bolle, R.: “An identity-authentication system using fingerprints.”, Proceedings of the IEEE, 85(9), (1997), 1365-1388.
- [18] Agrawal, R., Faloutsos, C.; Swami, A.: “Efficient similarity search in sequence databases.”, In Proc. 4th International Conf. On Foundations of Data Organization and Algorithms (FODO), (1993), 69-84.
- [19] Press, W.; Teukolsky, S.; Vetterling, W.; Flannery, B.: “Numerical Recipes in C, the Art of Scientific Computing”, Cambridge University Press, Cambridge, UK, 2nd Edition, (1992).
- [20] Faloutsos, C.: “Searching Multimedia Databases by Content”, Kluwer Academic Publishers, Boston, MA, (1996).