Tıp Bilişiminde Veri Ambarı ve Veri Madenciliği Uygulaması

Selda DÜZGÜNOĞLU[a], Adnan YAZICI[b], Ünal YARIMAĞAN[a]
[a]*Hacettepe Üniversitesi,* [b]*ODTÜ, Ankara*

Application of Data Warehouse and Data Mining in Medical Informatics

Abstract

Today the use of data warehouse and OLAP technologies has become widespread. Softwares which are designed for daily operations are designed as operational systems which have low level reporting and high level transactions. Data warehouse applications are designed for analyzing, reporting and supporting analytical operations. Data which is produced from the operational systems is cleaned and necessary transform operations are done before loading aggregated data into data warehouse environment. By defining multidimensional views named 'cube' from the aggregated data by using OLAP technologies, queries are processed for decision support. Using these approaches carries great importance to find decisions for the future. In our study, data which is stored in "Hospital Information System" database and has subjects for decision support was loaded into data warehouse environment. Data was formed by creating multidimensional cubes with OLAP technologies. Also, decision tree data mining model was applied on data which was loaded into data warehouse with clean and free from doubt values, for getting new meanings from the data. It was observed that reports which were generated from data warehouse data produce results more quickly. Reports which are generated from database applications produce results in long periods but reports which are generated from data warehouse cubes produce results in minute levels according to data's size. For reporting data warehouse data a report tool which was designed as a part of Computer Engineering master thesis. As the report tool gets its test data from a "Hospital Information System", this study presents an example of a medical informatics application.

Key Words

Decision Support Systems, Data Warehouse, OLAP (Online Analytical Processing), Data Mining, Decision Tree Algorithm

Özet

Günümüzde veri ambarı ve OLAP teknolojilerinin kullanımı gün geçtikçe yaygınlaşmaktadır. Günlük işlerin takip edilmesi amacıyla hazırlanmış otomasyon yazılımları az seviyede raporlama içeren ve yüksek seviyeli hareket işleme amaçlı tasarlanmış işletimsel sistemlerdir. Veri ambarı uygulamaları, analiz ve raporlama amaçlı tasarlanmış ve analitik işlemleri destekleyen sistemlerdir. İşletimsel sistemlerle üretilmiş verilerin veri ambarı ortamında varsa mevcut hatalarından ayıklanarak ve gerekli veri dönüştürmeleri yapılarak özet bilgilerin saklanması sağlanır. Toparlanmış özet veriler üzerinde OLAP teknolojisi kapsamında tanımlanan küp veri yapıları ile karar destek bağlamında sonuç üretebilecek sorgular hazırlanır. Bu yaklaşımların kullanılması gelecek için alınacak kararların belirlenmesinde büyük önem taşımaktadır. Yapmış olduğumuz çalışmada, "Hastane Bilgi Yönetim Sistemi" veri tabanında bulunan ve karar destek kapsamında bilgi verici konuları içeren veriler veri ambarı ortamına taşınmıştır. OLAP teknolojileri ile hazırlanan çok boyutlu küpler üzerinde verilerin şekillenmesi sağlanmıştır. Ayrıca veri ambarına taşınarak, temizlenmiş ve doğruluğundan emin olunan veriler üzerinde karar ağacı veri madenciliği modeli uygulanarak veriye yeni bir anlam yüklenmiştir. Veri ambarında tutulan verilerle çalışan raporların çok daha hızlı üretildiği gözlenmiştir. Veri tabanı uygulamaları ile veri tabanında tutulan veriler üzerinden alınan raporlar çok uzun sürede sonuç üretirken, veri ambarında yer alan küpler üzerinde hazırlanan raporların çalışma süresi verinin büyüklüğüne göre değişmekle birlikte dakikalar seviyesine inmiştir. Veri ambarındaki verileri raporlamak için Bilgisayar Mühendisliği Yüksek Lisans tezi kapsamında hazırladığım raporlama aracı kullanılmıştır. Geliştirilen raporlama aracı "Hastane Bilgi Yönetim sistemi" verileri ile test edilerek tıp bilişimi kapsamında örnek bir çalışma hazırlanmıştır.

Anahtar Kelimeler:

Karar Destek Sistemleri, Veri Ambarı, OLAP (Online Analytical Processing), Veri Madenciliği, Karar Ağacı Algoritması.

1. Giriş

Günümüzde büyük ölçüde hareket işleme tabanlı sistemler kullanılır ve şirketler bu sistemler tarafından üretilen geniş sayıdaki bilgiye erişirler. Saklanan bilginin boyutu gigabyte hatta terabyte düzeyine erişir. Bu tarz büyük veritabanları ile, bir hastane yönetimi son bir yıl içerisinde en sık karşılaşılan vakaları ve bu vakalar kapsamındaki ortak özellikleri elde edebilir. Bu bilgi ışığında hastane araştırma-geliştirme biriminin hedefi, bu vakalara çözüm bulmak ve oluşumunu sağlayan etmenleri tespit etmek amaçlı strateji geliştirmek olabilir.

Karar destek kapsamında verinin saklanması ve erişimi birçok önemli yaklaşımın oluşmasını sağlar [1].

- Birçok karar destek sorgusu SQL ile yazılabileceği gibi, birçoğunun SQL ile ifade edilmesi imkânsızdır ya da çok zordur. Veri tabanının çok büyük sığalara ulaşmasına rağmen, OLAP veri analizi için kullandığı teknik ve araçlarla, sorgulara anlık cevaplar verebilmektedir.
- Veritabanı sorgu dilleri detaylı istatistiksel analiz için uygun performansta değildir. Veri ambarı ve OLAP uygulamaları için SAS, SAP ve S++ gibi çeşitli yazılım paketleri vardır. Bu paketler büyük ölçekli verilerin veri tabanından alınmasını sağlarlar.
- Bilgi-buluş teknikleri otomatik olarak veriden istatistiksel kuralların ve örüntülerin elde edilmesini sağlar. Veri madenciliği yapay us araştırmacıları ve istatistik çözümleyiciler tarafından bulunan bilgi buluş tekniklerinin birleştirilmesidir.
- Büyük firmaların ticari kararları alırken kullanmak durumunda oldukları çok sayıda veri kaynağı vardır. Veri kaynakları verileri farklı şemalar altında saklarlar. Farklı veri kaynaklarına erişmesi gereken sorguları çalıştırmak için veri ambarları kurulur.

Veri ambarı ve OLAP karar destek sistemlerinin vazgeçilmez elemanları durumuna gelmiştir. Karar destek sistemleri, geleneksel OLTP sistemlerine göre farklı ihtiyaçlar göstermektedir. Ralph Kimball'a göre veri ambarı, özellikle analiz ve sorgulama amaçlı yapılandırılmış hareket verilerinin kopyasıdır [2]. W.H. Inmon'un tanımıyla veri ambarı, temel olarak organizasyonel karar alımında kullanılan konu yönlendirmeli, bütünleşik, zaman değişimli, kararlı/değişime uğramayan veri koleksiyonudur [3].

Karar destek sistemleri işletimsel veritabanlarında unutulmuş veriye ihtiyaç duyar. OLTP işlemleri anlık veriyi içerirken OLAP işlemlerinde tarihsel veriler kullanılır. Karar destek sistemleri çoğunlukla farklı heterojen veri kaynaklarından alınan birleştirilmiş verileri içerir. En genel haliyle veri ambarı, farklı kaynaklarda tutulan verilerin ortak bir çatı altında birleştirilerek, verilerin zaman boyutunda birbiri ile konuşmasını sağlayan, tutarlı ve doğru verilerin yer aldığı sistemdir.

Veri ambarı üzerine ETL (Extract-Transform-Load (Çekme-Dönüştürme-Yükleme)) süreci ile heterojen kaynaklardan alınan veriler üzerinde veri madenciliği algoritmalarının uygulanması sağlanarak, normal sorgular ile elde edilemeyen bilgilerin su yüzüne çıkması sağlanmaktadır. Veri ambarı uygulamalarının veri madenciliği modelleri ile desteklenmesi ile verinin yeniden şekillendirilmesi mümkündür. Yapılan çalışmada 'Karar Ağacı' veri madenciliği modeli kullanılmıştır. Karar ağacı, en yaygın kullanılan sınıflandırma algoritmasıdır. Sınıflandırmayı sağlayacak bir dizi kuralın oluşturulması sağlanır. Ağaçta yer alan her bir düğüm, ilgili düğümün ve o düğümle bağlantılı diğer düğümlerin işlevlerini açıklayan kurallar kümesi olarak ifade edilir. Başlangıç noktasını gösteren kök düğüm üzerinden yaprak düğümlere doğru ilerleyerek, her düğümün temsil ettiği kurallar dizisine göre sonuç kümesi gözlemlenir. Kolay anlaşılır olması tercih edilmesindeki en büyük faktörlerden birisidir.

2. Gereç ve Yöntem

Yapmış olduğum çalışma "Hastane Bilgi Yönetim Sistemi" kapsamında bir veri ambarı oluşturarak OLAP ve veri madenciliği teknolojilerinin kullanılmasını içermektedir. Bu çalışmada 4 farklı katmandan söz etmek mümkündür.

1. katman, veri ambarını besleyen verilerin bulunduğu işletimsel veri tabanı sistemi (Oracle VTYS 9.2.0.1)

2. katman, veri ambarını oluşturan ilişkisel veri tabanı sistemi (Microsoft SQL Server 2000)

3. katman, OLAP sunucusu ve veri madenciliği servisi (Microsoft Analysis Services)

4. katman, çok boyutlu küplerin ve veri madenciliği modellerinin sorgulanabildiği ve raporlanabildiği uygulama katmanı (Microsoft .Net 2003)

Veri ambarının tasarlanması karmaşık bir iştir ve aşağıdaki adımların takip edilmesi gerekir [2];

- Mimariyi belirle, kapasite planlaması yap, sunucu, veritabanı, OLAP sunucu ve araçları seç
- Sunucular, saklama birimleri ve istemci araçları entegre et
- Veri ambarı şema ve görüntülerini tasarla
- Fiziksel ambar organizasyonunu tanımla (veri yerleşimi, bölümleme ve erişim metotları)
- Veri kaynaklarına bağlan
- Veri çekmek, temizlemek, taşımak, yüklemek ve güncellemek için betikler hazırla
- Verileri çek
- Verileri temizle (kaynak veri üzerindeki tutarsızlıkların temizlenmesi)
- Veri dönüştürmelerini yap (farklı veri formatlarına ya da dillere dönüştürüm)
- Yüksek hızlı veri transferi imkânını sağla
- Verileri yükle (Verilerin veri ambarına yüklenmesi)
- Son kullanıcı uygulamalarını tasarla ve gerçekleştir
- Ambar ve uygulamaları bir araya getir.

3. Bulgular

Bildiriniz, "Hastane Bilgi Yönetim Sistemi" verileri için bir veri ambarı ve veri madenciliği uygulaması hazırlanmıştır. Bu uygulamanın gerçekleştirimi sürecinde tamamlanan adımlar ve elde edilen sonuçlar şu şekildedir;

- Veri ambarını besleyecek OLTP sisteminin seçilmesi ve incelenmesi
- Veri ambarı ihtiyaçlarına göre ambar tasarımının yapılması
- OLTP sisteminden veri ambarına verilerin çekilmesi, temizlenmesi ve yüklenmesi (ETL)
- Veri ambarında tutulan tablolardan çok boyutlu küplerin tasarlanması ve üretilmesi
- Üretilen küpler üzerinde veri madenciliği algoritmalarının işletilmesi
- Küplerin sorgulanmasını ve raporlanmasını sağlayan bir aracın gerçekleştirimi
- Sorgular için MDX (MultiDimension Expression) ifadelerinin raporlama aracı kullanılarak kolaylıkla hazırlanması
- Sorguların grid ortamında ve çizge, çubuk ve dairesel grafik olarak gösterilmesi
- Birden çok sorgunun yer aldığı raporların hazırlanması
- Veri madenciliği karar ağacı sonuçlarının ağaç yapısında çizimsel gösteriminin raporlama aracı üzerinde gösterilmesi

Veri madenciliği kapsamında, gerçekleşmiş olan ameliyatların hastanın kişisel nitelikleri ile olan ilişkisi karar ağacı modeli üzerinde üretilmiştir. Bu örnek için kullanılan kişisel nitelikler Tablo-1 ile gösterilmiştir. 'Cinsiyet' ve 'Ameliyat Grupları' nitelikleri için iki ayrı karar ağacı üretilmiştir. Ameliyat grupları niteliği için hazırlanmış karar ağacında ortaya çıkan örnek kurallar Tablo-2 üzerinde yer almaktadır.

Bu kurallar, Tablo-1 de belirtilen nitelikler üzerinden verinin dağılımına göre elde edilmiştir.

Tablo-1 Veri Madenciliği Modelinde Kullanılan Nitelikler ve Türleri

Nitelik	**Türü (Girdi, Tahmin Edilebilir)**
Cinsiyet	Girdi, Tahmin Edilebilir
Ameliyat Açıklama	Girdi, Tahmin Edilebilir
Medeni Hali	Girdi
Doğum Yılı	Girdi
Meslek	Girdi
Eğitim Durumu	Girdi
Hasta Grubu	Girdi

Tablo-2 'Ameliyat Grupları' Niteliği İçin Üretilen Karar Ağacı Kuralları Örnekleri

Kurallar	**Örnek Sonuç Kümesi (Ameliyat Grupları)**
Doğum Yılı <= 1958,25	El cerrahisi ve Mikro Cerrahi . % 3.5. Genel İşlemler % 7.49. Fıtıklar % 4.8. Karaciğer ve Safra Yolları % 3.58
Doğum Yılı > 1958,25 ve Cinsiyet = K	El cerrahisi ve Mikro Cerrahi . % 2.70. Genel İşlemler % 5.03. Fıtıklar % 0.74. Karaciğer ve Safra Yolları % 2.21. Gebelik ve Doğum % 29.94
Doğum Yılı > 1958,25 ve Cinsiyet = E	El cerrahisi ve Mikro Cerrahi . % 3.07. Genel İşlemler % 7.55. Fıtıklar % 1.42. Karaciğer ve Safra Yolları % 1.42

Tablo-2 üzerinde hastane uygulamasının bir yıllık işletimsel verilerinden elde edilmiş ameliyat işlemleri üzerinde uygulanan veri madenciliği sonuç kümesine ait örnek kayıtlar bulunmaktadır. Örnek olarak, hastanın doğum yılının 1958 yılı ilk çeyreğinden büyük ve cinsiyetinin 'Kadın' olma durumunu incelersek, yapılacak ameliyat grubunun %2.7 oranında 'El ve Mikro Cerrahi', %29,94 oranında 'Gebelik ve Doğum', %2,21 oranında 'Karaciğer ve Safra Yolları' olabileceği gözlenecektir.
Yapmış olduğum çalışmada OLTP ve OLAP sistemlerini karşılaştırabilmek için, aylık istatistik sonuçlarını üreten, 19 sorgu ve her bir sorgu için tarihsel kümeleme fonksiyonlarını içeren hesaplanmış değerleri gösteren birer sorgu ile toplam da 38 adet sorgu barındıran bir raporu ele aldım. Rapor kapsamında yer alan sorgular Tablo-3 ile gösterilmiştir. Tablo-4 üzerinde OLTP (İşletimsel veri tabanı) ve OLAP (Veri ambarı) sistemlerinin birbirleriyle karşılaştırılması verilmiştir.

4. Tartışma

Veri ambarından üretilen raporların çok karmaşık olmayan sorgu cümleciklerine sahip olmalarına rağmen, normalde işletimsel sistemden üretilebilen raporlara kıyasla çok daha hızlı olduğu ve doğru veriyi içerdiği gözlenmiştir. Ayrıca veri küpleri üzerinde uygulanan veri madenciliği algoritmaları ile karar destek kapsamında bilgi sağlayan modeller geliştirilmiştir.

19 adet ana sorgu, gerçekleştirilen yazılım kapsamında bir rapor bünyesinde toplanmıştır. Toplam çalışma süresi on dk. dır. Aynı raporu OLTP sistemi bünyesinde veri tabanı üzerinde SQL ile hazırlanacak sorgularla elde etmek oldukça maliyetlidir. Bazı sorgular için oluşturulacak sonuç kümesi birden çok rapor ile üretilebilmektedir. Ayrı ayrı elde edilen rapor sonuçlarının tek bir rapora dönüştürülmesi ve bu veriler üzerinden istatistiksel sonuçlar çıkarılabilmesi için de ayrıca bir iş gücü harcanması gerekmektedir. Bu çalışmanın raporların alındığı ortamdan başka bir ortamda (word, excel v.b) yapılması gerekmektedir. Tablo-4 üzerinde belirtildiği gibi, tanımlanmış olan temel 19 sorgunun sonucunu üretebilecek OLTP sisteminde 30 rapor hazırlamak gerekirken, OLAP sisteminde hazırladığım raporlama aracı üzerinde 1 adet raporun tanımlanması yeterlidir. OLTP sisteminde tek bir raporun ortalama çalışma süresi 2 saat'tir. Bazı raporların çalıştırılarak sonuç üretmesi iki günü bulabilmektedir. OLAP sisteminde ise sorgulara çok hızlı cevap verilmekte ve ortalama 2 dk.lık bir sürede bir sorgu çalışmaktadır. Bazı sorgular saniyeler seviyesinde sonuç üretebilmektedir. İşletimsel veri tabanında 60 ilişkisel tablo bulunurken, bu tabloları konu yönlendirmeli olarak özetleyerek veri tabanına alırken gerekli olan

Tablo-3 Raporu Oluşturan Sorgu İçerikleri

Sorgu No	İçerik
1	Uzmanlık dallarına göre yatışı yapılan hasta sayıları
2	Yatak türlerine göre toplam yatak sayıları
3	Günlük yatan hasta sayıları
4	Ortalama günlük yatan hasta sayıları
5	Yatak işgal yüzdeleri
6	Ortalama yatış süreleri
7	Doğum türlerine göre doğum sayıları
8	Uzmanlık dallarına göre muayene sayıları
9	Muayenenin yapıldığı bölüme göre muayene sayıları
10	Ameliyat türlerine göre ameliyat sayıları
11	Anestezi yapılarak gerçekleşen ameliyat sayıları
12	El Cerrahisi ameliyat sayıları
13	Patoloji türlerine göre gerçekleşen patoloji işlem sayıları
14	Laboratuar gruplarına göre laboratuar işlemleri sayısı
15	Radyoloji tetkik sayıları
16	İşlem türlerine göre acilde gerçekleşen işlem sayıları
17	İşlem türlerine göre gerçekleşen tıbbi işlem sayıları
18	Hazırlanan reçete sayıları
19	Nedenlere göre gruplanmış ölüm sayıları

Tablo-4 Veri Tabanı ve Veri Ambarı Karşılaştırması

	Veri Tabanı	Veri Ambarı
Toplam tablo sayısı	60	30
Toplam Veri Sığası (1 yıllık)	60 GB	30 MB
Toplam Rapor Sayısı	30	1
Raporların Ortalama Çalışma Süresi	2 saat	2 dk.

tablo sayısı 30 'a düştüğü gözlenmiştir. OLTP sisteminde her türlü detaylı yüksek yoğunlukta veri bulunmaktadır. Bu veriler temizlenerek ve özetlenerek veri ambarına alındığı için veri ambarında aynı veri daha az yer kaplamaktadır. Fakat veri ambarında zaman boyutunda verilerin tutulması gerekmektedir, sadece çalışılan yıl için değil tüm yıllardaki veriler veri ambarına alınmalıdır. Dolayısıyla zaman boyutunda veri ambarı oldukça yoğunluklu veri içerecektir. 1 yıl için veri tabanında 60 GB'lık veri sığası bulunurken veri ambarında 30 MB'a düşebilmektedir.

5. Sonuç

Veri madenciliği ve veri ambarı, veri veritabanı sistemleri, yapay öğrenme, istatistik, algoritma, veri görselliği, yapay sinir ağları gibi konuları kapsayan disiplinler arası alanlardır. Bu nedenle yeni bir disiplin olmasına karşın uygulama alanı oldukça geniştir. Veri ambarı teknolojisi, üretim, satış, finans servisleri (risk ve kredi kartı analizi gibi), taşımacılık, telekomünikasyon, sağlık gibi birçok sektörde başarıyla uygulanmaktadır. Hastane uygulamalarında veri ambarı ve veri madenciliğinin kullanılması, verinin güvenilirliğinin sağlanması, işlemlerde kolaylık ve hızlılık, eldeki veriden karar destek kapsamında yeni bilgilere erişme imkânlarını sağlamaktadır. Veri ambarı ve veri madenciliği sonuçlarının kullanılabilmesi için yeni veri kümelerine erişebilen raporlama araçlarına ihtiyaç duyulmaktadır.

Yapmış olduğum çalışma ile mevcut raporlama araçlarında bulunmayan ek özellikleri de içeren, veri ambarı ve veri madenciliği sonuçlarını kolaylıkla sunabilen bir raporlama aracı geliştirilmiştir. 'Hastane Bilgi Yönetim Sistemi' nde yer alan veriler ile veri ambarı uygulaması gerçekleştirilerek tıp bilişiminde veri ambarı ve veri madenciliğinin yararları ve gerekliliği tespit edilmiştir.

6. Teşekkürler

Hasta Bilgi Yönetim sistemi örnek verilerini bu çalışma kapsamında paylaşan ve tıp bilişimi bilgi birikimini aktaran Tepe Teknolojik Servisler A.ş. yetkililerine teşekkür ederim.

7. Kaynakça

[1] Korth H.F., Silberschatz A., 1991, *Database System Concepts*, McGraw-Hill
[2] Kimball R., 2002*, The Datawarehouse Toolkit : The Complete Guide to Dimensional Modeling*
[3] Chen, Z., 2001, *Data Mining and Uncertain Reasoning : An Integrated Approach*, Wiley Inter-Science Publication.

8. Sorumlu Yazarın Adresi

Hacettepe Üniversitesi, Bilgisayar Mühendisliği Bölümü, Ankara
Selda@cs.hacettepe.edu.tr