İNDEKSLEYİCİ İÇİN HTML BELGENİN XML BELGEYE DÖNÜŞTÜRÜLMESİ ÜZERİNE BİR UYGULAMA

Aydın CARUS [1] Eyüp Can DÜNDAR [2] Altan MESUT [3]

[1,2,3] Trakya Üniversitesi, Mühendislik-Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü, Edirne

[1]e-posta: aydinc@trakya.edu.tr [2]e-posta: eyupcan@trakya.edu.tr
[3]e-posta: altanmesut@trakya.edu.tr

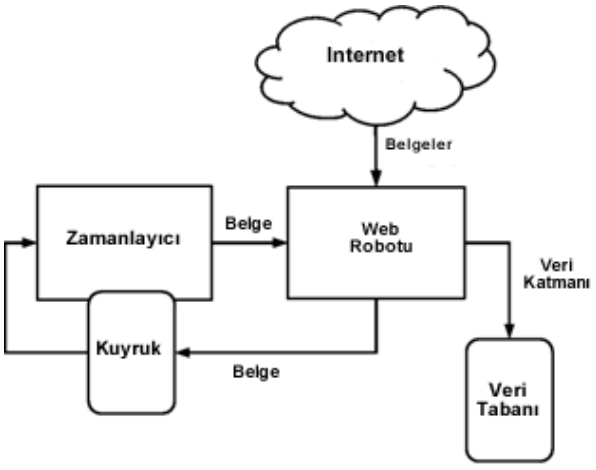
Özetçe

İnternet üzerindeki verilerin büyük bir çoğunluğu HTML yapısındadır. HTML, verilerinin gösterilmesine yönelik geliştirilmiş olan bir dil olduğu için bu biçimdeki bir belgeyi işleyerek içindeki gerekli bilgileri elde edebilmek oldukça çaba gerektirmektedir. Bu çalışmada internet üzerinden Web Robotları tarafından indirilen HTML sayfalarının içindeki verilerin, standart olarak hiyerarşik ve düzenli bir yapıya sahip XML biçimine dönüştürülmesini sağlayan bir uygulama geliştirilmiştir. Geliştirilen bu uygulama ile HTML belge içindeki verilerin indeksleyici gibi internet üzerindeki verileri işleyen programlara daha etkin olarak sunulması sağlanmıştır.

1. Giriş

Günümüzde internet kullanımının yaygınlaşması ile birlikte internet üzerindeki belgelerin sayısı gün geçtikçe artmaktadır. Bu belgelerin büyük bir çoğunluğu HTML (Hypertext Markup Language) [1] standardında hazırlanıp yayınlanmaktadır. HTML, belgelerin birbirine nasıl bağlanacağını ve belge içindeki metin ve resimlerin nasıl yerleşeceklerini belirleyen ve etiket olarak isimlendirilen kod parçalarını içeren bir işaretleme dilidir. Tarayıcılar, HTML kodunu yorumlayarak kullanıcıya bilginin sunulmasını sağlar. HTML standardında, kesin belirleyici bir düzenin olmayışı farklı tarayıcıların aynı kodu farklı yorumlamasına sebep olabilmektedir. Ayrıca bu durum, HTML belgeleri üzerinde gösterilen verinin işlenmesini zorlaştırmaktadır. Verilerin çeşitli çalışma ortamları arasında paylaşımının etkin olarak sağlanabilmesi için XML (Extended Markup Language) standardı [2] geliştirilmiştir. XML'in amacı farklı sistemler arası veri aktarımını sağlamaktır. Verinin tanımlanması ve betimlemesi için kullanılır. HTML'de kullanılacak etiketlerin önceden tanımlı olması gerekirken, XML'de kullanılacak etiketler önceden tanımlı olmak zorunda değildir. XML belgelerinin en önemli özelliği, belgelerin ağaç yapısında olmasıdır.

İnternet üzerinden bir bilgiye ulaşmak istediğimizde, bu bilgiyi milyonlarca web sitesi içerisinden sayfaları inceleyerek ulaşmak büyük bir çaba gerektirir. Bu sorunu çözmek için arama motorları geliştirilmiştir. Arama motorları, internet üzerindeki web sayfalarını tarayıp, onları kendi içinde indeksleyip, kullanıcıların indekslenmiş veriler üzerinde arama yapmasına olanak sağlayan sistemlerdir. Arama motorları sayesinde kullanıcıların istedikleri bilgiye daha hızlı ulaşması sağlanmaktadır. Arama motorlarında web robotu olarak isimlendirilen programlar kullanılarak, internet üzerindeki sayfalar otomatik olarak kayıt edilir. Web Robotu taradığı tüm sayfaların bir kopyasını alır. İndeksleyici olarak isimlendirilen program, Web Robotunun indirmiş olduğu bu sayfaları çeşitli yöntemler kullanarak [1], arama motorunun veri tabanına kaydeder. Bunun yapılması sonucunda kullanıcıların istedikleri bilgiye, veritabanı üzerinden sorgulamalar yaparak verinin bulunduğu sayfalara çok kısa sürede ulaşabilmesi sağlanmaktadır. Şekil 1'de gösterildiği gibi Web Robotu, web sayfaları üzerindeki bağlantıları belirli bir yöntemle [2] takip ederek sayfalar arasında geçişi sağlamaktadır. Bir Web Robotu için önemli olan veri sayfa içindeki bağlantılardır. Web Robotu sayfanın bir kopyasını oluşturduktan sonra, belge içindeki kelimeleri veritabanına kaydeden indeksleyici için resmin nerede gösterildiği veya yazının kalın mı yazıldığı gibi belgenin gösterimine yönelik ayrıntılar önemli değildir. İndeksleyici, belge içindeki kelimeleri kullanmaktadır. Belge üzerindeki kelimeleri belirli bir metotla indekslemektedir. Web Robotunun belgeyi belirli bir standartta ve HTML'in düzensizliğinden arındırılmış olarak sunması, indeksleyicinin daha etkin olarak indeks verilerine ulaşmasını sağlayacaktır.



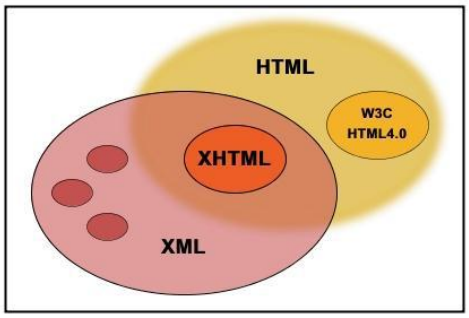
Şekil 1: *Web Robotu yapısı.*

HTML belgenin XML belge biçimine dönüştürülmesi konusunda daha önceden yapılmış çalışmalar mevcuttur. Bu çalışmalar HTML belgedeki verinin okunarak XML standardına çevrilmesinden öte, HTML belgenin XHTML (Extended Hypertext Markup Language) [3] biçimine aktarmasını yapmaktadırlar. [3] [4] Ayrıca HTML belge

[1] http://www.w3.org/TR/1999/PR-html40-19990824/
[2] http://www.w3.org/TR/2006/REC-xml-20060816/
[3] http://www.napersolutions.com/htmltoxml.html

biçiminden XML belge biçimine direk dönüşüm yapan çalışmalar [4] olduğu gibi verilerin yapısı ile tekrar ilgilenmeden başka uygulamalar tarafından yeniden kullanılabilmesine yönelik çalışmalarda mevcuttur [5,6]. Şekil 2'de XML, HTML ve XHTML arasındaki ilişki verilmektedir. Bu çalışmada HTML belge biçimindeki verilerin sadece veri katmanının XML biçimine dönüştürme işlemini gerçekleştiren bir uygulama geliştirilmiştir.

Bu dokümanın 2. bölümünde HTML ve XML standartları karşılaştırılmış, 3. bölümde HTML üzerindeki verilerin alınarak XML biçimine dönüştürülmesi aşamaları anlatılmıştır. Sonuçlar bölümünde uygulamanın kullanılabilirliği değerlendirilmiştir.



Şekil 2: *HTML ve XML ilişkisi.*

2. HTML ve XML Standartları

HTML ve XML standartları W3C (World Wide Web Consortium) olarak isimlendirilen topluluk tarafından belirlenmektedir. Bu topluluk, standartların kullanıcılar tarafından tanınmasını ve kullanıcılar arasında birlikteliğin sağlanmasını amaçlamaktadır. Bunun yanı sıra kullanıcılar tarafından genel olarak kabul görecek çekirdek prensipler ve bileşenleri hazırlamaktadır. W3C; XML, HTML standartlarını desteklemektedir.

2.1.**HTML**

HTML, belgelerin birbirlerine nasıl bağlanacaklarını ve belge içindeki metin ve resimlerin nasıl yerleşeceklerini belirleyen ve daha önceden tanımlanmış etiketlerden oluşan bir standarttır. Bir HTML belgeye örnek Şekil 3'te gösterilmektedir.

```
<html>
 <head>
   <title>Bu Bir Başlık</title>
 </head>
 <body>
   <p>
     <b>KALIN</b>
     <i>ITALIK</i>
     <b><i>Kalın ve italik</i></b>
   </p>
 </body>
</html>
```

Şekil 3: HTML belge örneği.

[4] http://www.stylusstudio.com/html_to_xml.html

HTML çok tutarlı bir biçimleme dili değildir. Örneğin, bir belgede başlangıç etiketleri, bitiş etiketleri, diğer etiketler ve metin ile karşılaşılmasına rağmen her başlangıcı olan etiketin bir bitişi olması zorunluluğu yoktur. HTML belgeleri metin, açıklama, basit etiketler ve sonlu etiketler bileşenlerinden oluşur.

HTML üzerindeki etiketler iki gruba ayrılabilir, ilk grup etiketler HTML üzerindeki verileri içeren gruptur. Bu grupta paragraflar, yazılar, resimler ve linkler bulunmaktadır. Diğer grup ise verinin altyapısı olarak adlandırılan, verilerin tarayıcılar tarafından nasıl gösterileceğini belirten etiketlerdir. Arama motorları verinin nasıl gösterileceği ile değil ne olduğu ile ilgilenmektedir. Dolayısıyla aslında Web Robotlarının indeksleyici için hazırlayacağı bilgi ilk grupta belirtilen verilerdir.

2.2.**XML**

XML, HTML ile pek çok açıdan benzerlik gösteren bir işaretleme dilidir. Verinin tanımlanması ve betimlenmesi için kullanılır. HTML'deki yapının aksine XML'de kullanılacak olan etiketler önceden tanımlı değildir. Yani bir XML dokümanının yapısı tamamıyla kullanıcı tarafından oluşturulur. Şekil 4'te örnek bir XML belge verilmiştir. Verinin betimlenmesi için DTD (Document Type Definition) adı verilen yapılar kullanılmaktadır.

XML ve HTML arasındaki en belirgin fark XML'in verinin kendisiyle ilgilenmesi HTML'in ise verinin sunumuyla ilgilenmesidir. HTML dokümanları veriye ilişkin gösterim bilgilerini içerirken XML dokümanları ise verinin tanım bilgilerini içermektedir. XML'in tasarım amaçlarından biri de verinin taşınmasıdır. Bu özellikleri incelendiğinde XML'in birçok önemli işlevi yerine getirdiği görülmektedir.

Günümüz bilişim uygulamalarında XML birçok farklı alanda kullanılmaktadır. Bu nedenle XML'i bir anlamda geleceğin web dili olarak tanımlamak mümkündür.

```
<mesaj>
  <kime>Ali</kime>
  <kimden>Ayşe</kimden>
  <baslik>Nice Yıllara</baslik>
  <yazi>Doğum Günün Kutlu Olsun!</yazi>
</mesaj>
```

Şekil 4: XML belge örneği.

HTML bir sözcüğü etiketler arasına alarak metnin koyu ya da italik yazılmasını sağlar. Oysa XML, yapısal verilerin etiketlenmesi için bir iskelet yapı sağlar. XML, HTML'in yerine geliştirilmemiştir. Farklı amaçlara sahiptir. XML daha çok verinin taşınması, dönüştürülmesi gibi verinin kendisine odaklıdır.

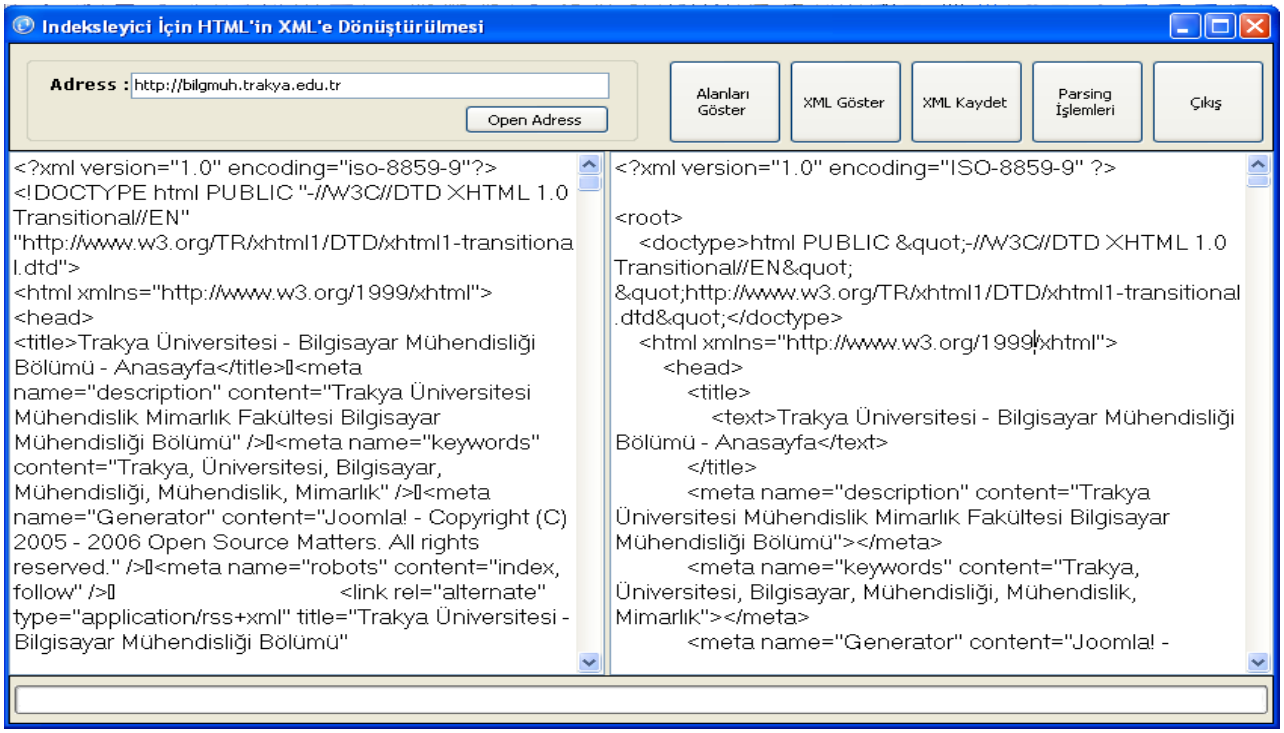
SOAP (Simple Object Access Protocol) internet üzerinde bilginin XML protokolü kullanarak paylaşılmasını sağlayan bir uygulamadır.[5] SOAP uygulamasının birincil fonksiyonu web robotları ile çok benzerdir. Web robotları siteleri tarayarak indeksleyici için gerekli olan bilgiyi siteden alıp getirir, SOAP adresleri de web siteden istenen veriyi getirmek için kullanılır. XML biçimi, HTTP üzerinden istenilen bilgilerin alındıktan sonra, farklı sistemlere aktarmak üzere kullanılması için uygundur.

[5]http://www.w3.org/TR/2000/NOTE-SOAP-20000508/

3. Geliştirilen Uygulama

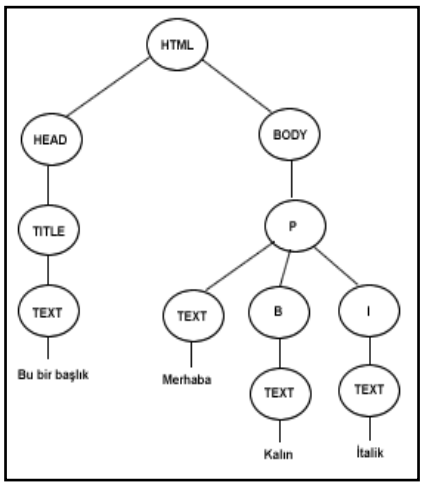
İndeksleyici, web robotlarının getirdiği belgeleri yorumlayarak, arama motorunun veri tabanına kaydeden programdır. İndeksleyici, sadece veri içermeyen karışık HTML kodları içinden, arama motoru için gerekli olan bilgileri alıp belirli bir algoritma dahilinde bu bilgileri ve sayfa hakkında başka gerekli bilgileri arama motorunun veri tabanına kaydeder. Arama motorları, sayfalar üzerindeki bağlantıları, kelimeleri, resimlerin adreslerini ve resimlerin yazıları gibi bilgileri veri tabanına kaydeder.

Geliştirilen uygulama sayesinde indeksleyicinin düzensiz ve karmaşık HTML kodlarından metinleri ve bağlantıları bulmak için zaman harcaması yerine, indeksleyiciye gerekli olan bilginin XML formatında sunulması sağlanmıştır. Bu şekilde, indeksleyicinin HTML üzerindeki düzensiz kodları tasnif etmesine gerek kalmayacaktır.



Şekil 5: Geliştirilen uygulamanın arayüzü.

Geliştirilen uygulamada web belgelerinin incelenmesinde, HTML ve XML belgeleri için yüksek düzeyli, etkin uygulamalar geliştirilmesini sağlayan DOM (Document Object Model) yapısı[6] kullanılmıştır. Geliştirilen uygulamanın arayüzü Şekil 5'te gösterilmektedir. Uygulamada HTML belgenin XML belge olarak dönüştürülmesi için önce HTML belge ağaç yapısında ifade edilmektedir. HTML belgelerde açılan etiketlerin sonlandırılma zorunluluğu yoktur. HTML belgenin, bir ağaç yapısı olarak tanımlanabilmesi için HTML belge içindeki sonlandırılmayan etiketlere, sonlandırma etiketleri eklenir, varsa başlangıç etiketi bulunmayan etiketler de kaldırılır, böylece HTML belgeye XML'deki ağaç yapısına uygun hale getirilir. Bu işlem sonucunda Şekil 6'da gösterildiği gibi HTML belgesini ağaç yapısı şeklinde gösterebiliriz. HTML belgesinin kök düğümü <HTML> </HTML> etiketi olmaktadır.



Şekil 6: HTML belgenin ağaç yapısında ifade edilmesi

XML belgelerin ağaç yapısı şeklinde tanımlandığına göre HTML belgelerinde yukarıda belirtilen düzenlemeleri yaptıktan sonra Şekil 7'de verilen algoritmayı kullanarak belgeler XML formatına çevrilmektedir.

[6] http://www.w3.org/DOM/

```
Verilen web sayfasının HTML kodlarını al
Yap
{
  Etiketi oku ve sonlandırma etiketini oku
  Eğer (etiketin sonlandırma etiketi yoksa)
  {
    Etikete sonlandırma etiketi ekle
  }
  Eğer ( Etiketin başlangıç etiketi yoksa)
  {
    Etiketi sil
  }
} DevamEt (son etikete kadar oku)

XML belgeyi tanımla HTML belge başına gel
Yap
 {
   Etiketi oku ve sonlandırma etiketini oku
   HTML etiketini XML etiketine çevir
 } DevamEt (son etikete kadar oku)
```

Şekil 7: HTML belgeyi XML belgeye dönüştürme algoritması.

Yapılan çalışma ile amacımız, HTML belgeyi XML belge olarak ifade etmekten çok indeksleyici için HTML içindeki gerekli verinin hazırlanmasını sağlamaktır. İndeksleyici "href", "img" etiketleri ve belge üzerindeki metinleri girdi olarak almaktadır. İndeksleyicinin girdi olarak aldığı bu bilgiler Şekil 8'de verilen algoritma kullanılıp XML formatına çevrilerek indeksleyici için hazır hale getirilmektedir. Daha sonra indeksleyici, arama motorunun veritabanına XML olarak ifade edilmiş anlamlı bilgiyi veri tabanına kayıt edecektir.

```
Verilen web sayfasının HTML kodlarını al
Yap
{
  Etiketi oku ve sonlandırma etiketini oku
  Eğer (etiketin bitiş noktası yoksa)
  {
    Etikete sonlandırma etiketi ekle
  }
  Eğer (Etiketin başlangıç etiketi yoksa)
  {
    Etiketi sil
  }
} DevamEt (son etikete kadar oku)

XML belgeyi tanımla HTML belge başına gel
Yap
{
  Etiketi oku
  Eğer (etiket = bağlantı) veya (etiket=yazı) veya
(etiket=resim)
  {
    XML belgede düğümünü oluştur
  }
} DevamEt (son etikete kadar oku)
```

Şekil 8: İndeksleyici için HTML belgedeki bilgilerin XML biçimine dönüştürülme algoritması.

4. Sonuçlar

Geliştirilen uygulama kullanılarak Şekil 9'da verilen örnek HTML belgesinin dönüştürülmesi sonucu elde edilen XML belge Şekil 10'da verilmektedir. HTML belge ve XML belge incelendiğinde dönüşüm işleminin indeksleyici için gereksiz olarak nitelendirilen fazla verilerden arındırılmış doğru bir XML belge haline geldiği ve çok karmaşık HTML dokümanlarda bile indeksleyici için uygun olarak dönüşüm yaptığı görülmüştür.

```
<p><b>

<font size="1" face="Verdana">
<img border="0" src="yeni_bullet.gif" width="8" height="6"></font><font size="1" face="Verdana" color="#2388EF">
<a target="_blank" href="http://www.trakya.edu.tr/bidb/">BİDB</a><br>
</font>
<font size="1" face="Verdana">
<img border="0" src="yeni_bullet.gif" width="8" height="6"></font><font size="1" face="Verdana" color="#2388EF">
<a target="_blank" href="http://www.trakya.edu.tr/bidb/internet_hiz_kul_kurallari.htm">
T.Ü. İnternet Hizmetleri Kullanım  Kuralları </a></font></b><br>
<b>
<font size="1" face="Verdana">
<img border="0" src="yeni_bullet.gif" width="8" height="6"></font><font color="#0B4460"><font face="Verdana" size="1">
  <a target="_blank" href="http://www.trakya.edu.tr/yeni/kisisel_web.php">Kişisel Web Siteleri</a></font><a style="text-decoration: none; color
  </a> </font></b><br>
<b>
<font size="1" face="Verdana">
<img border="0" src="yeni_bullet.gif" width="8" height="6">
  <a target="_blank" href="eskiwebler.htm">Eski
  Web Siteleri</a></font></b></td>
</tr>
<tr>
<td class="kaderspecial" bgcolor="#FFFFFF">
  <table border="0" width="100%" id="table17">
<tr>
<td>
<p align="center"><a target="_blank" href="fulbright.htm">
<img border="0" src="fulbright.gif" width="152" height="75"></a></td>
</tr>
</table>
```

Şekil 9: Örnek bir HTML belgesi.

Şekil 10: HTML belgenin geliştirilen uygulama ile XML biçimine dönüştürülmüş hali.

HTML içindeki verilerden, indeksleyici için gerekli olan verileri bulmak için XML formatından ifade edilmesinin oldukça fayda sağladığı görülmektedir. Ayrıca HTML belgelerin XML belgeye çevrilerek işlenmesi arama motorlarının sürekli güncel kalması dışında başka birçok amaçla da kullanılabilir. XML belgenin yeniden kullanılabilirliği XML biçimli belgenin ağaç yapısında olmasından dolayı daha uygundur.

HTML belgeler üzerindeki verilerin XML formatına dönüştürülmesi, internet üzerinden otomatik olarak bilgi alışverişi yapılan tüm uygulamalar için kullanılabilir.
Gelecekte web sayfalarında SOAP desteğinin yaygınlaşması ile birlikte belki de her internet sitesi kendi veri katmanını XML formatında kullanıcılara sunan servisler içerecektir.

Kaynakça

[1] Cho, J., Garcia-Molina, H., "Parallel Crawlers", *In Proceedings of the 11th International Conference on World Wide Web*, USA, 2002.

[2] J., Tsay, "AuToCrawler: An Integrated System for Automatic Topical Crawler", *Proceedings of the Fourth Annual ACIS International Conference on Computer and Information Science* (ICIS'05), 2005.

[3] Raggett, D., "XHTML: The Extensible Hypertext Markup Language", *W3C LA event in Stockholm*,1999.

[4] Sahuguet, F. A., "Web ecology: Recycling html pages as xml documents using w4f", *WebDB*, 1999.

[5] Craig, A. K., Lerman, K., Minton S., Muslea I., "Accurately and reliably extracting data from theweb: A machine learning approach", *Bulletion of the IEEE Computer Society Technical Committee on Data engineering*, 1999.

[6] Bartocci, L. M., Merelli, E., "An xml view of the world", *ICEIS (1)*, 2003.