یازدهمین کنفرانس بین المللی کامپیوتر انجمن کامپیوتر ایران، پژوهشگاه دانشهای بنیادی، پژوهشکده علوم کامپیوتر، ۴ تا ۶ بهمن ۱۳۸۴، تهران، ایران.
11th International CSI Computer Conference (CSICC'2006), School of Computer Science, IPM, Jan. 24-26, 2006, Tehran, Iran.

استخراج اتوماتیک داده از وب برای کاربرد سیستمهای هوشمندی تجاری

دکتر احمد براآنی
استادیار گروه کامپیوتر
دانشگاه اصفهان – گروه کامپیوتر
ahmadb@eng.ui.ac.ir

بهرام امینی
مدرس گروه کامپیوتر
دانشگاه آزاد مبارکه
bahamini@yahoo.com

چکیده: دسترسی به اطلاعات بازار، رقباء و مشتریان از طریق موتورهای جستجوگر و مرور دستی تقریباً ناکارآمد و غیربهینه است. هدف از این تحقیق ارایه یک روش کارآمد و الگوی موفق برای استخراج اتوماتیک داده از وب و مجتمع سازی آنها با اطلاعات انباره داده سازمانی برای کاربرد سیستمهای هوشمندی تجاری است. در این روش با استفاده از یک نرم افزار تجاری و تکنولوژی راپر یک معماری موثر و قوی برای فرآیندهای استخراج، ساختاردهی مجدد و بارگذاری داده ها به سیستم هوشمندی تجاری ایجاد شده است. نتایج بدست آمده نشان میدهد که تکنولوژی راپر برای استخراج داده های خارجی و مجتمع سازی آنها با داده های انباره داده سازمانی بسیار مناسب و کارآمد میباشد.

dataacademy.ir

واژه های کلیدی: استخراج داده، هوشمندی تجاری، انباره داده، وب، راپر

۱- مقدمه

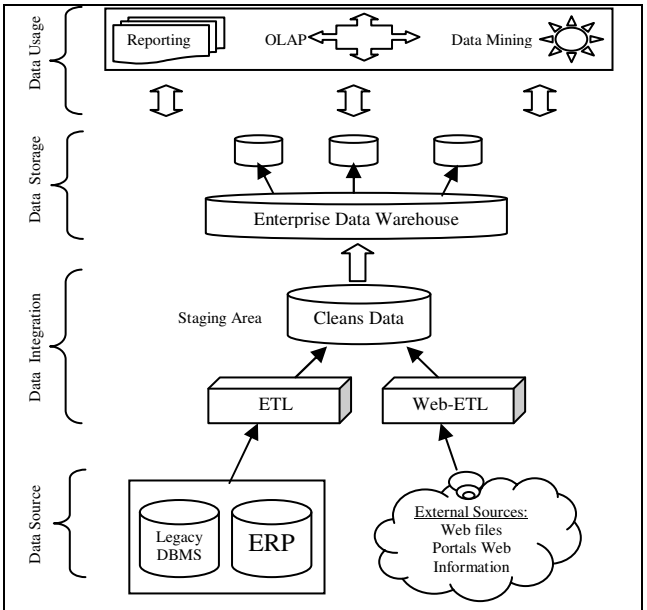
مدلهای بلوغ[1] انباره داده [1] سازمانها را بر اساس میزان استفاده از منابع اطلاعاتی موجود ارزیابی و رتبه بندی میکنند. اغلب سازمانها طبق این مدل در سطوح پائین قرار دارند زیرا اکثر تلاش خود را معطوف مجتمع سازی داده های درون سازمانی نموده اند[2]. برای رسیدن به سطوح عالی مثل مدل تکامل اطلاعات[2] [1] لازم است که یک سازمان محدوده اطلاعات خود را فراتر از مرزهای خودش قرار دهد و از هر فرصت تجاری برای بهره مندی از تجارب و دانش سایرین و در راستای حفظ بقاء خود استفاده کند. رسیدن به این هدف مستلزم چرخش دیدگاه مدیریت اطلاعات در سازمان به سمت دیده بانی بازار و تعدیل سریع خود براساس نیاز بازار میباشد و بنابراین باید سیستم هایی را برای بررسی اطلاعات خارجی مستقر نماید. سازمانهای پیشرفته از انباره داده برای افزایش توان رقابتی و بهبود تصمیم گیری خود استفاده می کنند اما ساختن انباره داده به تنهایی کافی نیست و بایستی وضعیت مشتریان، تامین کنندگان، رقباء و سایر فاکتورهای خارجی را نیز درنظر داشته باشند. برای کسب دید کامل از وضعیت رقباء و پتانسیل های بالقوه سازمان نسبت به سایر سازمانهای مشابه، نیاز به اطلاعات برون سازمانی می باشد. امروزه بهترین منبع برای کسب داده های خارجی وب میباشد زیرا مادر تمام انباره های داده است[1]. وب به عنوان بزرگترین پایگاه داده خارجی شامل حجیم عظیمی از داده های بهم وابسته است که متاسفانه به فرمت قابل فهم انسانی نیست و براحتی توسط برنامه های کامپیوتری پردازش نمی شوند. در این مقاله یک راه حل موفق برای چگونگی اتوماتیک سازی و نرمال سازی داده از منابع اطلاعاتی عمومی مخصوصاً وب سایت ارائه میشود. همچنین شیوه مجتمع سازی[3] آنها با داده های داخلی که در محیط انباره داده قرار دارند بیان میگردد. در بخش دوم درباره هوشمندی تجاری و در بخش سوم، روش Lixto بعنوان ابزار تولید اتوماتیک راپر[4] توضیح داده میشود. در بخش چهارم به شیوه تولید راپر و استخراج اتوماتیک داده از وب بصورت مطالعه موردی پرداخته میشود. در انتها نیز نتیجه گیری و بحث ارایه میگردد.

۲- هوشمندی تجاری

عبارت هوشمندی تجاری[5] یا BI غالباً به عنوان روش جمع آوری ، نمایش و تحلیل داده های سازمان برای پشتیبانی از تصمیم بکار میرود. کلمه هوش رابطه نزدیکی با "دانش" و شفاف سازی دارد بنابراین منظور ما از BI در این مقاله فرآیند تامین بینش بهتر در یک سازمان و زنجیره فعالیتهای آن می باشد. همانگونه که در شکل (۱) نشان داده شده است فرآیند هوشمندی تجاری شامل سه مرحله اصلی میباشد: مجتمع سازی داده، ذخیره سازی داده، کاربرد داده ها.

[3] Integration
[4] Wrapper
[5] Business Intelligence (BI)

آکادمی داده
dataacademy.ir



شکل ۱- فرآیندهای اصلی هوشمندی تجاری

۱- مجتمع سازی داده ها: شامل همه روشهای استخراج داده از منابع داخلی و خارجی مثل سیستم های پایگاههای داده متعارف، سیستم های ERP و وب میباشد . داده ها ابتدا به یک ناحیه پردازش موقت منتقل میشوند تا تبدیلات پاکسازی و نرمال سازی روی آنها انجام شود. سپس فرآیند بارگذاری داده بصورت برنامه ریزی شده(مثلا روزانه، هفتگی، ماهانه) و منظم، داده های پردازش شده و تمیز[۶] را به پایگاه داده مقصد (انباره داده) بارگذاری میکند.

۲- ذخیره سازی در انباره داده : ایده اصلی انباره داده ذخیره کردن داده های وابسته در یک پایگاه داده همگن و انحصاری برای پشتیبانی از تصمیم میباشد. یکی از خصوصیات مهم انباره داده مجتمع سازی داده های ناهمگون توزیع شده داخلی و خارجی می باشد . این کار مستلزم ذخیره سازی داده ها در یک پایگاه داده مرکزی ، و دسته بندی آنها بر حسب نواحی موضوع[۷] شبیه فروش، تولید، مالی و ... برای پردازشهای تجاری میباشد[3].

۳- کاربرد داده ها : برای پشتیبانی از تصمیم گیری، داده های انباره داده طوری سازماندهی میشوند که نیازهای کاربران متفاوت را برآورده سازند. بنابراین گزارشهای از پیش تعریف شده برای کاربران معمولی، تحلیل های پیشرفته برای دانشگران[۸] و داده کاوی[۹] برای تحلیل گران داده فراهم شده است.

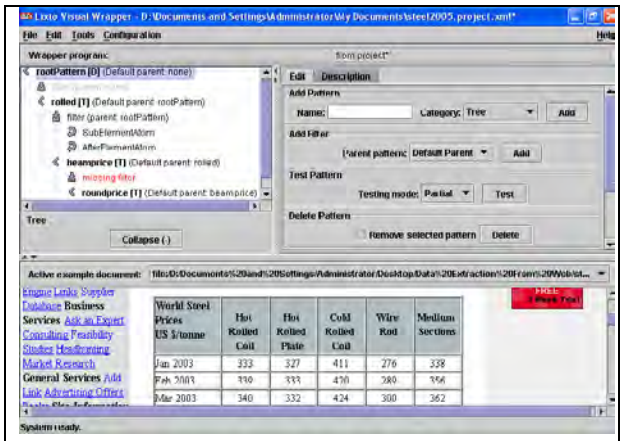
۳- راه حل Lixto

برای استخراج اتوماتیک داده از وب از روش راپر که کارآمد و ساده است استفاده میشود. یک راپر، عامل نرم افزاری هوشمند برای شناسایی و استخراج الگوهای داده مورد نظر از یک وب سایت خاص میباشد. راپر پس از مشاهده و یادگیری الگوهای داده در صفحه وب میتواند انواع داده ها را از محتوای آنها تشخیص داده و استخراج نماید. معمولاً برای تولید اتوماتیک راپر از نرم افزارهای خاص استفاده میشود که علاوه بر کارایی، قابلیت اطمینان و سادگی فرآیند را تامین میکنند[4]. نرم افزار Lixto Suite[۱۰] ابزاری برای تولید اتوماتیک راپر است که دسترسی، استخراج، تبدیل و حمل اطلاعات از منابع شبه ساختیافته مثل صفحات وب را فراهم میکند. این نرم افزار ۱۰۰٪ براساس تکنولوژی جاوا و استانداردهای J2EE , SOAP , XSLT و XML میباشد. این نرم افزار برای تولید راپر از زبان Elog که یک زبان منطق- مبنا[۱۱] است و یک مفسر زبان استفاده میکند. نرم افزار Lixto دارای محیط یادگیری و تولید راپر است و شامل دو ابزار متفاوت میباشد: سرور تبدیل و ویژوال راپر. از تکنولوژی راپر برای استخراج اطلاعات وابسته از اسناد HTML و تبدیل آنها به XML که براحتی توسط ابزارهای نرم افزاری مربوطه مثل XML-GL, XQuery قابل پرس و جو و انجام پس پردازشها[۱۲] است استفاده میگردد. ویژوال راپر از روش جدیدی برای تشخیص و استخراج محتوای بهم وابسته اسناد HTML استفاده میکند که جزییات آن در [2] آمده است. هنگامی که یک راپر ساخته میشود میتوان از آن برای استخراج مستمر و اتوماتیک اطلاعات بهم وابسته از صفحات وب با تغییرات اندک استفاده نمود. البته اگر این تغییرات زیاد باشد مثلاً ساختار جداول و متن بکلی عوض و یا جابجا شود قادر بشناخت اطلاعات نمیباشد زیرا از ساختار درخت HTML بعنوان مبنای شناخت محتوای صفحات وب استفاده میکند. ایجاد یک راپر با LixTo بوسیله انتخاب اطلاعات وابسته در مرورگر وب و کلیک روی آنها صورت میگیرد.

---

[8] Knowledge Worker
[9] Data Mining
[10] www.Lixto.com
[11] Logic-base
[12] Post-Processing

dataacademy.ir
آکادمی داده
dataacademy.ir



شکل ۲- انتخاب و استخراج داده از صفحات وب بوسیله LixTo

همچنین میتوان شرایط خاصی را نیز تعریف نمود تا اگر ساختار داده ها در صفحه وب تغییر کرد، داده ها قابل تشخیص باشند. در شکل (۲) یک نمونه از اطلاعات قیمتها در سایت www.steelonthenet.com نشان داده شده است. بعد از اینکه یک راپر تولید گردید اگر ساختار صفحه وب دچار تغییرات جزیی شود مثلا" جدول آن جابجا گردد و یا یک انیمیشن در بالای صفحه اضافه شود، راپر هنوز میتواند اطلاعات وابسته را از وب سایت استخراج کند. از دیدگاه راپر یک صفحه اینترنت یک ساختار درختی HTML است. یک راپر صرفا" متن را از نودهای درخت HTML استخراج نمیکند بلکه از "شرایط هوشمند" بنام الگوهای منطقی استفاده میکند. "شرایط" برای راپر شکل (۲) میتواند چنین باشد: نواحی مرتبط باید در یک جدول باشند، سال و ماه در هرسطر جدول ظاهر شوند، نام محصولات خاصی در سرفصل جدول ظاهر شود. (این نامها در پایگاه داده سیستم ذخیره میگردند). برای اینگونه الگوهای منطقی که شرایط را تشکیل میدهند، نرم افزار با استفاده از روشهای هوشمندانه[۱۳] برای بهترین انطباق در درخت HTML جستجو میکند . با این روش عاملهای راپر در مقابل تغییرات جزئی مستحکم[۱۴] و پایدار خواهند بود. نرم افزار LixTo در طی فرآیند تولید راپر قابلیتهای متفاوتی دارد که برای استخراج اتوماتیک داده کاربرد زیادی دارند. امکان مرور صفحات وابسته بصورت افقی و عمودی با لینکهای مشخص، امکان ضبط پیمایش زیرصفحات[۱۵] بصورت ماکرو، پشتیبانی از SessionID های پویا، logon کردن اتوماتیک به صفحات دارای امنیت، پر کردن صفحات فرم با داده ها و انجام پرس و جو[۱۶] ، پردازش نتایج حاصل از پرس و جوها (از طریق رابط با پایگاه های داده)، پشتیبانی از کوکی ها[۱۷] و SSL ، راحتی در مدیریت نرم افزار و موارد دیگر از آنجمله هستند. این قابلیتها اجازه میدهد که بتوان فرآیند استخراج اتوماتیک داده از وب را بسادگی و بنحو موثر و با قابلیت اطمینان زیاد انجام داد. برای مطالعه بیشتر در مورد قابلیت های راپر میتوان به منبع [2] مراجعه نمود .

۴- مطالعه موردی

این مطالعه در ادامه پروژه طراحی و پیاده سازی انباره داده در ذوب آهن اصفهان [4] انجام شده است. دراین شرکت فرآیندهای متفاوت از قبیل خرید مواد اولیه و قطعات، تامین انرژی، تولید و فروش محصولات، مدیریت منابع مالی، مدیریت منابع انسانی و... وجود دارد . با توجه به رقابت شدید شرکتهای تولید کننده فولاد، برای کاهش هزینه های تولید و کسب سهم بیشتر بازار فروش و تامین ارزان تر مواد اولیه ، نیاز به اطلاعات خارجی توسط تحلیل گران تجاری بخوبی احساس میگردید . از آنجا که حجم قابل توجهی از خرید مواد اولیه اصلی و قطعات از خارج از کشور تامین میشود نیاز به اطلاعات قابل اطمینان قیمت در بازارهای فروش وجود دارد. برای دسترسی به اطلاعات موارد فوق، افراد متفاوتی با صرف وقت زیاد داده های مورد نیاز را بوسیله جستجو در وب سایت های رقباء یا با مرور سایت های اطلاع رسانی قیمتها، اطلاعات مورد نیاز را به طور دستی استخراج و فرمت بندی مجدد مینمودند. معمولا" در تحلیل های تجاری اطلاعات میزان تقاضای بازار و قیمت فروش انواع محصولات بطور روزانه، هفتگی یا ماهیانه مورد استفاده قرار میگیرند. در این تحقیق اطلاعات قیمت خرید مواد اولیه و فروش محصولات فولادی از وب سایت های متفاوت بوسیله LixTo استخراج و در انباره داده منتقل گردیدند. این داده ها بدلیل تجمیع با داده های داخلی شرکت امکان مقایسه سریعتر اطلاعات و کشف حقایق را میسر میسازد.

۴-۱- استخراج و تبدیل داده

معماری انباره داده شرکت مطابق با روش Inmon در [4,6] توضیح داده شده است. در این معماری، انباره داده سازمانی با مدل داده 3NF قراردارد. گنجه های داده پس از استخراج داده و انجام پردازشهای تجمعی روی آنها از انباره داده تغذیه میکنند. یک ناحیه میانجی[۱۸] برای ذخیره سازی موقت داده ها و انجام عملیات ETL تکمیلی روی آنها در نظر گرفته شده است که

[13] Heuristic

[17] Cookies

[18] Staging Area

dataacademy.ir
آکادمی داده
dataacademy.ir

میتواند داده های داخلی و خارجی را برای بارگذاری نهایی در انباره داده نگهداری نماید. بستره انباره داده اوراکل انتخاب شده است و دارای رابط با کلیه سیستم های سنتی و ERP میباشد. در این تحقیق برای بارگذاری منظم و با قاعده داده های وب به انباره داده چهار فرآیند متفاوت انجام شد که عبارتند از :

کشف منابع داده: کشف و انتخاب وب سایت های حاوی اطلاعات قیمت، رقباء و مشتریان که در پاسخ به نیازهای تجاری به آنها مراجعه میشود[7]. این سایتها به مرور زمان توسط آن شرکت انتخاب شده است.

استخراج داده: جهت استخراج منظم و اتوماتیک داده از وب، نرم افزار LixTo انتخاب و تنظیمات اولیه روی آن صورت گرفت.

ساختار دهی داده: فیلتر کردن ، معتبرسازی و تبدیل داده های وب به یک فرمت مناسب برای مجتمع سازی در انباره داده توسط سرور تبدیل LixTo صورت گرفت.

مجتمع سازی داده: انباره داده با داده های خارجی (استخراج شده از وب) در دوره های منظم و برنامه ریزی شده بطور اتوماتیک توسط موتور اوراکل تغذیه میگردد.

یکی از اهداف اصلی این تحقیق این است که میزان و نحوه استخراج اتوماتیک دانش از وب سایت های رقباء برای تحلیل های تجاری سازمان مشخص شود. بنابراین نیاز به استقرار و تنظیم یک مکانیزم مناسب برای استخراج داده به روش سیستماتیک میباشد و برای انجام این کار ابزار LixTo انتخاب گردید. این نرم افزار میتواند انواع داده های صفحات وب به فرمت HTML را استخراج نموده و پس از ترکیب و فرمت دهی با سایر داده های خارجی بطور اتوماتیک به ناحیه میانجی منتقل نماید تا به درون انباره داده سازمانی بارگذاری گردند. در شکل (۲) چگونگی انتخاب و استخراج داده از صفحات وب بوسیله Lixto نشان داده شده است. این نرم افزار محتوای هر صفحه وب را جهت انتخاب داده های مرتبط و شناسایی الگوهای یادگیری، به طراح راپر نشان میدهد. در پایین پنجره یک صفحه از وب سایت نوعی www.steelonthenet.com که توسط شرکت مورد استفاده قرار دارد توسط مرورگر وب بارگذاری شده است و اطلاعات وابسته برای استخراج علامت گذاری شده اند. در بالای پنجره و سمت چپ، الگوهای منطقی شبیه عنوان و قیمت به صورت سلسله مراتب تعریف شده اند . این ساختار متناظر با خروجی XML است که بعدا توسط راپر تولید میشوند. کلیه میتوان در منبع [5] مطالعه نمود. بعد از بارگذاری داده های وابسته از یک صفحه وب، یک الگو بصورت "عنوان" تعریف میگردد. این الگو سطرهای اطلاعات عنوان را شناسائی میکند. در این خط سایر الگوها مثل تاریخ ، مقاطع متفاوت فولاد که اطلاعات را مشخص میکنند ایجاد میگردد. بدین منظور ساختارهای اسناد HTML و یا عبارتهای با قاعده که ساختارهای منطقی را نشان می دهند مورد استفاده قرار میگیرند. در هر مرحله میتوان خروجی حاصل از طراحی و اجرای راپر را به فرمت XML مشاهده و کنترل نمود و تغییرات لازم را در ترکیب درختواره الگوها و شرایط بوجود آورد. نتیجه نهایی طراحی، یک برنامه راپر است که به زبان Elog میباشد. متن این برنامه که بصورت مجموعه ایی از قواعد است توسط کاربر قابل مشاهده نیست اما میتواند خروجی XML حاصل از اجرای آنرا بصورت برخط مشاهده کند. با تکرار فرآیند تولید راپر برای هر یک از وب سایتهای مورد استفاده شرکت، مجموعه ایی از راپرهای استخراج داده بدست آمد.

سپس با استفاده از سرور تبدیل و مطابق با مدل منطقی انباره داده، مدلسازی جریان داده طراحی گردید. بوسیله سرور تبدیل عملیات تجمیع، فیلتر کردن و تبدیل داده های استخراج شده توسط راپر مدلسازی میشوند. سرور تبدیل برای انتقال داده به محیطهای گوناگون از رابط های مربوطه استفاده میکند که در این مورد از XML و پیامهای email استفاده شد. پیامهای email هنگامی صادر(فعال) میشوند که فرآیند استخراج داده همراه با خطا باشد. بنابراین مسئول سیستم میتواند با آگاهای کامل از نتایج فرآیندها اقدام بموقع بعمل آورد.

در اینجا داده های حاصل از خروجی راپر به سرور تبدیل منتقل شدند تا با ایجاد پیامهای SOAP براساس XML آنها را تبدیل و ساختار دهی مجدد نماید. پیامهای SOAP دارای یک سرفصل با فراداده و اطلاعات مسیردهی و پارامترهای امنیت است. بدنه پیام حاوی داده های تولید شده توسط راپر است که از طریق رابط JDBC به سرور اوراکل منتقل میشوند. اوراکل نیز به عنوان موتور پایگاه داده ERP و BI دارای رابط های متفاوت برای بارگذاری داده از منابع متفاوت میباشد که عبارتند از :

۱- فایلهای مسطح داده
۲- سیستم های پایگاه داده مثل SQL Server , DB2 , Oracle
۳- فایلهای XML از طریق رابط SOAP

dataacademy.ir
آکادمی داده
dataacademy.ir

اما فرمت XML بدلیل تفکیک داده ها از نمایش آنها بیشتر مورد توجه سیستمهای BI قرار داد. بنابراین ارتباط سرور تبدیل با موتور اوراکل برای بارگذاری داده ها در ناحیه میانجی و انتقال برنامه ریزی شده آنها به انباره داده بخوبی مدلسازی گردید.

۲-۴- بارگذاری داده های XML به انباره داده

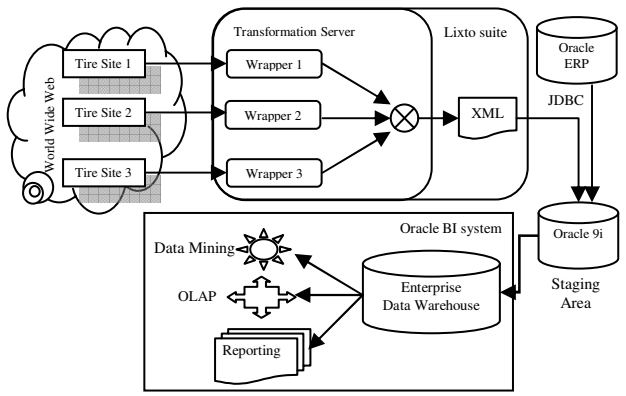
برای انتقال داده های XML به درون ناحیه میانجی انباره داده شرکت از رابط JDBC استفاده شد. این رابط یک API برای تبادل داده در محیطهای ناهمگون و مستقل از پلاتفرم است و براحتی و بدون تبدیل اضافی داده های XML را دریافت میکند. همانگونه که در شکل (۳) نشان داده شده است داده های XML پس از ورود به ناحیه میانجی و ترکیب با سایر داده های داخلی (مثل ERP) مطابق با مدل داده ایی انباره داده وارد آن میشوند. دوره بارگذاری داده ها به انباره داده توسط برنامه ریز[19] موتور هوشمندی تجاری انجام میشود. زمان بندی بطریقی است تا هماهنگ با برنامه زمانبندی سرور تبدیل LixTo اجرا شود. بدین ترتیب همواره داده های جدید وارد انباره داده میشوند و هیچگونه تداخلی بین دو برنامه ریزی صورت نمیگیرد.

dataacademy.ir

۵- نتیجه گیری

در این مقاله چگونگی استخراج اتوماتیک داده های شبه ساختیافته از وب برای کسب اطلاعات بازار و رقباء و پشتیبانی از تصمیم بیان گردید. همچنین قابلیتهای نرم افزار LixTo در مورد تولید راپر و پردازشهای موثر روی داده های وب توضیح داده شدند. نتیجه این پردازشها یک فایل XML ساختیافته است که براحتی میتواند بوسیله سیستم های هوشمندی تجاری یا هر پایگاه داده با رابط استاندارد مورد استفاده قرار گیرد. همچنین روش ایجاد ناحیه میانجی و بارگذاری داده ها به انباره داده اوراکل با استفاده از رابط JDBC توضیح داده شدند. مجتمع سازی داده های خارجی با سیستمهای هوشمندی تجاری دارای مزیتهای متفاوتی است که در زیر خلاصه میشوند:

۱- مجتمع سازی سریع داده ها برای پشتیبانی از واکنش سریع سازمان به نیازها و تغییرات بازار
۲- فعال سازی مکانیزم های هشدار توسط عامل های گزارش دهی سیستم هوشمندی تجاری
۳- کسب تصویر واقعی تر از بازار
۴- کاهش هزینه های آموزش بعلت داشتن رابط گرافیکی



شکل ۳- فرآیند مجتمع سازی داده های استخراج شده در EDW

۵- کاهش زمان و هزینه تلاش افراد برای بازیابی اطلاعات
۶- کاهش خطاهای جمع آوری و تجمیع داده ها
۷- دسترسی به منابع داده بیشتر با دانه بندی[20] دلخواه
۸- بهبود نمایان سازی[21] و افزایش کیفیت داده

بدین ترتیب دانشگران و تحلیل گران داده قادر به کسب اطلاعات در مورد وضعیت های بازار، رقباء، قیمت محصولات و مواد و ارزیابی رفتار بازار به صورت بلادرنگ خواهند بود. آگاهی سریع در مورد این امر منجر به اخذ تصمیمات درست و به موقع و افزایش توان رقابتی سازمان میگردد. همچنین تکنولوژی راپر برای استخراج داده های خارجی سازمان و مجتمع سازی با داده های انباره داده بسیار مناسب و کارآمد میباشد.

۶- منابع

[1] Hatcher D., Prentice B., 2004, "The Evolution of Information Management", Business Intelligence Journal, TDWI
[2] Baumgartner R., Flesca S., Gottlob G., 2001, "Visual web information extraction with Lixto", ", 27th VLDB Conference, Italy
[3] Alberto H. F., Laender Berthier, A. RibeiroNeto, 2002, "A Brief Survey of Web Data Extraction Tools", www.sigmod.org/sigmod/record/ issues/0206/laender-survey.pdf, June 2005
[4] Amini Bahram, 2005, " MSc thesis: Design and Implementation of Enterprise Data Warehouse in ESCO", Azad Najaf Abad University
[5] Hackathorn Richard, 1999, " Web Farming for the Data Warehouse", ACM Journal
[6] Inmon Bill, 2002, "Building the Data Warehouse", John Wiley
[7] Zhu Y., Buchmann, A., 2002, "Evaluating and Selecting Web Sources as External Information Resources of a Data Warehouse", In Proc. of the 3rd International Conference on Web Information Systems Engineering (WISE02)

[20] Granularity
[21] Transparency

آکادمی داده
dataacademy.ir