پایان نامه دوره کاردانی کامپیوتر

گرایش نرم افزار

موضوع:

پیاده سازي یک موتور جستجوي پیمایشي

استاد راهنما:

مهندس سعید امیریان

نام دانشجو:

محمد صادق شاد

خرداد ماه 1389


بسم الله الرحمن الرحيم

پایان نامه دوره کاردانی کامپیوتر

گرایش نرم افزار

موضوع:

پیاده سازي یک موتور جستجوي پیمایشي

استاد راهنما:

مهندس سعید امیریان

نام دانشجو:

محمد صادق شاد

خرداد ماه 1389

کلام نویسنده

قطعا کم نیستند پروژه هایی از این دست و خصوصا در این زمینه که همانند من با ذوق و شوق فراوان شروع به کار برروی آن کرده اند و به نتایج مطلوبی هم در ابتدای کار دست یافته اند؛ اما پس از مدتی بنا به هزار و یک دلیل دست از کار کشیده و بدنبال موضوعی دیگر رفته اند. از این دست موارد زیاد است و کافی جستجویی در وب بکنید. اما من مصمم تا اینگونه نباشم و تمام تلاشم را خواهم کرد تا این پروژه دچار سرنوشت مذکور نشود. ان شاء الله.

من یاور یقین و عدالتم من زندگی ها خواهم ساخت، من خوشی های بسیار خواهم آورد، من ملتم را سربلند ساحت زمین خواهم کرد، زیرا شادمانی او شادمانی من است.

کوروش کبیر

چکیده

جستجو...، کاریست که اکسپلوریم.نت قصد انجام آن دارد. اما چگونه؟ قطعا باید کار سختی باشد. آن هم جستجوی وب. وبی که نه ابتدایش معلوم است و نه انتهایش. اما کار نشد ندارد. با توجه به وجود ابزارهای قدرتمند و یک برنامه ریزی خوب می توان از پس این کار بر آمد. حال قصد آن را دارم تا به شما نشان دهم این کار چگونه توسط اکسپلوریم.نت انجام می پذیرد و در این راه از چه ابزار هایی استفاده شده است. اصلا جستجو یا بهتر بگویم موتور جستجو چیست؟ اینها مواردی است که با خواندن این سند هر چند سطحی با آن آشنا می شوید. امیدوارم که مورد توجه شما خواننده عزیز واقع شود.

فهرست مطالب

فصل اول

مایکروسافت اس‌کیو ال سرور

مایکروسافت اس‌کیو ال سرور[1] یا مایکروسافت سی‌کو ال سرور یک سیستم مدیریت بانک‌های اطلاعاتی[2] است که توسط شرکت مایکروسافت توسعه داده می‌شود

برخی از ویژگی‌های این سیستم مدیریت پایگاه داده‌ها به این شرح است:

1. بانک اطلاعاتی رابطه‌ای
2. امکان استفاده از روال‌های ذخیره‌شده[3]، نمایه‌ها[4] و تریگر[5]
3. پشتیبانی از ایکس‌ام‌ال
4. بسیار قدرتمند و بدون محدودیت حجم و تعداد رکورد
5. پشتیبانی از جستجوی تمام‌متنی[6] برای سرعت در بازیابی اطلاعات و استفاده از زبان طبیعی[7] در جستجوها

1-1. اس‌کیو ال سرور 2008

نسخه‌ی بعدی اس‌کیو ال سرور، اس‌کیو ال سرور 2008 می‌باشد که در ابتدا با نامگذاری کاتمایی[8]، 27 فوریه سال 2008 برای ارائه به بازار پیشنهاد گردیده بود؛ اما در نهایت نسخه آر تی ام[9] آن در ربع سوم سال 2008 عرضه گردید.

---

[1] Microsoft Sql (Se-Quel) Server
[2] Database Management System
[3] Stored Procedure
[4] View
[5] Trigger
[6] FullText Search
[7] T-Sql
[8] Katmai
[9] RTM

آخرین سی تی پی[1] در 19فوریه سال 2008 عرضه گردید. اهداف این نسخه ایجاد و مدیریت داده‌ها به شکل هماهنگی، سازماندهی و محافظت به شکل اتوماتیک می‌باشد. با توسعه دائمی اس‌کیو ال سرور در عرضه تکنولوژی‌های مختلف، باعث رساندن زمان اتلافی به نزدیکی صفر شده است.

اس‌کیو ال سرور 2008 همیشه در برگیرنده حمایت از داده‌های ساختاری یا نیمه‌ساختاری می‌باشد که این امر شامل قالبهای رسانه‌ای دیجیتال برای عکسها، صوتی، تصویری و دیگر داده‌های چند رسانه‌ای می‌باشد.

در نسخه‌ی جدید، اکثر داده‌های چندرسانه‌ای را می‌توان به عنوان یک مجموعه عظیم بانیری ذخیره‌سازی کرد. آگاهی درونی از داده‌های چندرسانه‌ای به ما این اجازه را خواهد داد که کارکردهای تخصیص یافته را اجرا نماییم. براساس نظر پل فلس‌نر، کاربران اس‌کیو ال سرور 2008 شرکت مایکروسافت می‌تواند به ذخیره‌سازِی داده‌های پشتیبانی شده برای داده‌هایی با تنوع متفاوت بپردازد :ایکس ام ال، پست الکترونیکی، زمان/تقویم، فایل، پرونده و ... از جمله داده‌ها می‌باشند. همین‌طور در این نسخه به خوبی می‌توان به اجرای عملیاتی چون: جستجو، پرس‌وجو، تجزیه‌وتحلیل، تقسیم‌بندی و انطباق همه نوع از داده‌ها پرداخت. از انواع دیگری از داده‌هایی جدید می‌توان از اختصاص داده‌ها و نوع‌های زمانی و نوع‌هایی از داده‌های فضایی نام برد که داده‌های وابسته به مکان می‌باشند.

پشتیبانی بهتر برای داده‌های غیرساختاری یا نیمه‌ساختاری با استفاده از بخش جریان‌فایلی[2] انجام شده‌است. شده‌است. این نوع از داده‌ها می‌توانند اضافه شده، یا اینکه برای بازگرفت به هر فایل ذخیره شده، سیستم فایلها بکار رونده داده‌های منسجم یا فراداده‌ها در هر فایل باید در پایگاه داده‌های اس‌کیو ال سرور ذخیره شوند. و در آنجا اجزا غیرساختاری در سیستم فایل ذخیره می‌شوند. اکثر فایلها می‌توانند هم از طریق کنترل‌کننده فایل وین32[3] و هم از طریق اس‌کیو ال سرور با استفاده از تی اس‌کیو ال در پایگاه داده‌ها قرار گیرند.

انجام و دستیابی به داده‌های فایلی که به عنوان یک حجم عظیم از داده‌های بانیری محسوب می‌شود، پشتیبانی و ذخیره‌سازی در پایگاه داده‌ها پشتیبانی و ذخیره‌سازی فایلهای مرجع انجام می پذیرد.

اس‌کیو ال سرور 2008 همین‌طور از سلسله مراتب داده‌های اصلی پشتیبانی می‌کند و در برگیرنده‌ی مفهوم تی‌اس‌کیو ال می‌باشد که مستقیماً با آنها سروکار دارد بدون اینکه به تحقیق بازگشتی بپردازد.

---

[1] CTP
[2] FileStream
[3] Win32

داده‌های فضایی می‌توانند به دو صورت ذخیره‌سازی شوند. یک زمین صاف (هندسه یا هندسه مسطح) که نوعی از داده‌ها می‌باشند که ارائه‌دهنده‌ی داده‌های هندسی فضایی می‌باشند و به شکلهایی که رد اصل به صورت سیستمهای طراحی کروی و همپایه و... هستند پیش‌بینی شده‌اند. صورت دیگر نوع داده‌های زمین‌کروی (هندسی) هستند که به استفاده از مدلهای بیضی شکل آنچه که در زمین به صورت منفرد و پیوسته تعریف می‌شوند، می‌پردازند.

اس‌کیو ال سرور در برگیرنده ویژگیهای بهتری در زمینه فشردگی و متراکم داده می‌باشد و بنابراین در بهبود یافتن توانایی اسکالر به ما کمک می‌کند. این بخش همین طور دارای اقتدار منابع بوده و به ما این اجازه را می‌دهد که به ذخیره‌سازی منابع برای کاربران بپردازیم.

اس‌کیو ال سرور در بردارنده‌ی قابلیتهایی برای شفاف‌سازی داده‌ها برای فشرده‌سازی و ذخیره آنها می‌باشد اس‌کیو ال سرور کتمایی از موجودیت ساختار پشتیبانی کرده و به ثبت ابزارها، همانندسازی و تعریف داده‌ها می‌پردازد. تعریف داده‌ها به ساختن مدل داده‌های موجود خواهد پرداخت.

سرویس‌های ثبت‌کننده‌ی اس‌کیو ال سرور به ثبت جداول با قابلیتهایی از تطبیق داده‌ها و تجسم محصولات خواهند پرداخت. آنچه که به وسیله‌ی مایکروسافت از مدیریت جانبی حاصل می‌شود اجازه می‌دهد که سیاست پیکربندی و محدودیتها در پایگاه کامل داده‌ها و جداول مورد اطمینان بطور دستوری ایجاد گردد.

نسخه‌ی استودیو مدیریت[1] به پشتیبانی از جستجوگر اس‌کیوال می‌پردازد. به وسیله‌ی سی تی پی رایج انتخاب لازم برای تحقیق و بررسی محدود می‌شود. این امر باعث ساختارهای دیگری از تی‌اس‌کیوال در انتشارات بعدی می‌گردد. به ایجاد پایگاه داده‌های موجود از طریق بدنه قدرت ویندوز و کاربرد مدیریت در دسترسی می‌پردازد. بنابراین سرور و همه‌ی نمونه‌های پیوسته می‌توانند به وسیله بدنه‌ی قدرت ویندوز اداره شوند.

شرکت مایکروسافت به ایجاد اس‌کیو ال سرور موجود در نسخه‌های چندگانه کرد که دارای دستگاه‌هایی با ویژگی متفاوت و کاربرانی با اهداف متمایز بود.

2-1. ویرایش هاي مختلف

1-2-1. ویرایش متراکم

[1] Management Studio

این ویراستار فشرده یک موتور با پایگاه داده‌های مستحکم می‌باشد. به جهت اندازه کوچک آن دارای دستگاهی با ویژگیهای کاهش‌دهنده در مقایسه با ویراستارهای دیگر می‌باشد. این وسیله به وسیله‌ی پایگاه داده‌ها با سایز حداکثر 4 گیگابایت محدود شده و نمی‌تواند براساس سرویس ویندوز عمل نماید ویراستار متراکم باید تابع تقاضای کاربرد می‌باشد.

1-2-2. ویرایش پرسرعت

ویرایش پرسرعت یک میزان پایین، ویرایش آزاد از اس‌کیو ال سرور می‌باشد که در برگیرنده موتور مرکزی پایگاه داده‌هاست. در حالیکه هیچ گونه محدودیتی در شماره پایگاه داده‌ها یا کاربران پشتیبانی شده وجود ندارد. پایگاه داده‌های کلی به ذخیره‌سازی در بخشهای مجزا می‌پردازد. هدف از این کار جایگزینی می‌باشد. سرویس جستجوگر متن کامل به عنوان یک بخش ضمیمه در اس‌کیو ال سرور با ویرایش پرسرعت قرار می‌گیرد. بطور کلی نسخه‌ی استودیو مدیریت اس‌کیو ال سرور برای عمل ویراستاری در دسترس می‌باشد.

1-2-3. ویرایش کارگروه

اس‌کیو ال سرور با ویرایش کارگروه در برگیرنده موتور مرکزی پایگاه داده‌ها می‌باشد. این بخش از دیسک ویراژ در شمار نمونه‌هایی با فعالیت کمتر قرار می‌گیرد و در برگیرنده عملکردهایی با دسترسی بالا و شاخصهای برابر نمی‌باشد.

1-2-4. ویرایش احتمالي

اس‌کیو ال سرور با ویرایش احتمالی نسخه‌ای از اس‌کیو ال سرور با ویژگیهای کامل می‌باشد که در برگیرنده‌ی هر دو موتور مرکزی پایگاه داده‌ها و سرویس‌های اضافی می‌باشد و این در حالی است که وجود دامنه‌ی ابزارها برای ایجاد و اداره اس‌کیو ال سرور به صورت خوشه‌ای است.

1-2-5. ویرایش توسعه‌یافته

اس‌کیو ال سرور با ویرایش توسعه یافته دارای همان ویژگیهای اس‌کیو ال سرور با ویرایش احتمالی می‌باشد که به وسیله‌ی مجوز استفاده از سیستم‌های آزمایش و توسعه محدود گردیده و به عنوان یک سرور تولیدی محسوب نمی‌شود. این نسخه برای بازگذاری توسط دانشجویان در شارژ آزاد بخشهایی از برنامه مایکروسافت موجود می‌باشد.

فصل دوم

مایکروسافت ویژوال استدیو[1]

[1] Microsoft Visual Studio

ویژوال استدیو نام مجموعه‌ی برنامه‌نویسی شرکت مایکروسافت است که دارای چند زبان برنامه‌نویسی است. این مجموعه ویژوال سی و ویژوال بیسیک و ویژوال فاکس‌پرو و چند ابزار دیگر را درون خود جای داده‌است.

نرم‌افزار ویژوال استدیو، نرم‌افزاری توسعه یافته برای برنامه نویسان کامپیوتر است که توسط شرکت نرم‌افزاری مایکروسافت تولید شده است. تمرکز اصلی این نرم‌افزار از اولین نسخه‌های آن تا کنون بر روی خصوصیت آی‌دی‌ای بودن آن است که به برنامه نویس اجازه می‌دهد تا برنامه‌های کاربردی مستقل، وب‌گاه، برنامه‌های کاربردی وب و یا سرویس های وب را که بر روی تعدادی از پلتفرم[1]های پشتیبانی شده توسط دات نت فریم‌ورک[2] (البته برای تمام نسخه‌های بعد از ویژوال استودیو 6) همچنین پلتفرم هایی مانند ویندوز سرور[3]، پاکت پی‌سی[4] و مرورگرها[5] که قابلیت اجرا را دارند را براحتی ایجاد نماید .

ویژوال استدیو یک مجموعه از برنامه‌هایی است که ارتباط بسیار نزدیک با هم دارند که مایکروسافت آن را به توسعه دهندگان و برنامه نویسان برنامه‌های کاربردی اهدا نمود تا آنها را وادار نماید در محیطی توسعه یافته بر روی پلت فرم های ویندوز و دات نت به ساخت برنامه‌های خود بپردازند . ویژوال استدیو می‌تواند برای نوشتن برنامه‌های کنسولی ، ویندوزی ، سرویس‌های ویندوز ، برنامه‌های کاربردی موبایل ، برنامه‌های کاربردی ای‌اس‌پی دات نت و سرویسهای وب ای‌اس‌پی دات نت بنا به انتخاب شما همراه با زبانهایی مانند سی شارپ، ویژوال بیسیک دات نت، سی پلاس پلاس و جی شارپ استفاده شود. با ویژوال استدیو واقعا چه کارهایی می‌توان انجام داد؟ در زیر تعدادی از کاربردهایی را که برای تولید آنها می‌توان از ویژوال استدیو استفاده نمود معرفی گردیده اند:

---

[1] Platform
[2] Microsoft .NET Framework
[3] Microsoft Windows servers and workstations
[4] PocketPC Smartphones
[5] World Wide Web browsers

2-1. کاربردهاي ويژوال استديو

2-1-1. برنامه‌هاي تحت کنسول

این کاربرد برای اجرای خطوط دستور البته بدون محیط گرافیکی استفاده می‌شود که از این کاربرد برای برخی از ابزارهای کوچک یا برای اجرا شدن کدها توسط دیگر کاربردها استفاده می‌شود

2-1-2. برنامه‌هاي تحت ويندوز

برای برنامه‌های کاربردی ویندوزی که با استفاده از دات نت فریم‌ورک نوشته می‌شوند.

2-1-3. سرويس‌هاي ويندوز

سرویس‌ها، برنامه‌های کاربردی هستند که در پس زمینه ویندوز اجرا می‌شوند .

2-1-4. اي اس پي دات نت

یک تکنولوژی قدرتمند که برای طراحی و ساخت صفحات وب پویا استفاده می‌شود.

2-1-5. وب سرويس‌ها

تکنولوژی ای اس پی دات نت، مدل سرویسهای وب را بطور کامل فراهم نموده تا شما براحتی و با سرعت سرویسهای وب را تولید نمایید .

2-1-6. برنامه‌هاي ويندوز موبايل

می‌تواند بر روی ابزارهایی که شامل فریم‌ورک هستند مانند پاکت پی‌سی‌ها و همچنین تلفن‌های سلولی که پلت‌فرم مایکروسافت اسمارت‌فون[1] بر روی آنها اجرا می‌شود ، اجرا گردد .

2-1-7. برنامه‌هاي وين32، آي‌تي‌ال و امافسي

شما همچنان می‌توانید برنامه‌های سنتی امافسی، ای‌تی‌ال یا برنامه‌های وین32 را با استفاده از ویژوال سی پلاس پلاس ایجاد نمایید . این برنامه‌ها، برای اجرا به دات‌نت فریم‌ورک نیاز ندارند اما نمی‌توانند از مزایای آن نیز بهره‌ای ببرند.

2-1-8. ويژوال استديو اداينز[1]

[1] Microsoft Smartphone

شما می‌توانید از خود ویژوال استودیو برای ساخت توابعی جدید و قابل اضافه شدن به خود ویژوال استودیو استفاده نمایید .

2-1-9. کاربردهاي ديگر

ویژوال استودیو همچنین شامل پروژه‌هایی برای توسعه برنامه‌های کاربردی شما ، کار با پایگاه‌داده، ساخت گزارشها و ... می‌باشد .

2-2. يرايش هاي ويژوال استاديو

2-2-1. ويرايش ويژه[2]

ویرایش ویژه، گونه‌ی سبک شده ویژوال استدیو است که به طور رایگان عرضه می شود. امکاناتی که در این نسخه ارایه می شود نسبت به سایر ویرایش ها کم‌تر است و نمی توان افزونه ای به آی‌دی‌ای اضافه کرد. از جمله این که امکان برنامه نویسی برای موبایل، کامپایل 64 بیتی، ابزار آفیس، اشکال زدایی ریموت و طراح کلاس[3] وجود ندارد. ویژگی های شی گرا نیز کم‌تر شده است. نسخه های اس‌کیو ال و ام‌اس‌دی‌ان[4] ویرایش ویژه، از نوع کامل نیستند. زبان های تحت ویندوز و وب آن هم از نوع اکسپرس هستند.

2-2-2. ويرايش استاندارد[5]

ویرایش استاندارد نسبت به ویرایش قبلی قابلیت های بهتری دارد. این نسخه از کامپایلر 64 بیتی، ایکس‌ام‌ال، ام‌اس‌دی‌ان، ابزار خارجی و طراح کلاس به‌طور کامل پشتیبانی می‌کند. اما امکان برنامه نویسی برای موبایل (به جز نسخه‌ی 2005) و آفیس در این ویرایش وجود ندارد. آیتم سرور اکسپلورر[6] در ویرایش استاندارد قرار داده نشده و نسخه‌ی اس‌کیو ال آن، ویژه است.

2-2-4. ويرايش حرفه اي[7]

---

[1] Visual Studio Add-Ins
[2] Express Edition
[3] Class Designer
[4] MSDN
[5] Standard Edition
[6] Server Explorer
[7] Professional Edition

ویرایش حرفه ای علاوه بر این که قابلیت های ویرایش استاندارد را دارد از اس‌کیو ال سرور ویرایش دولوپر[1]، خطایابی راه‌دور[2]، برنامه نویسی موبایل، کریستال ریپورت[3]، سرور اکسپلورر و پروژه‌های نصاب ساز[4] برخوردار برخوردار است. در نسخه‌ی 2008، برنامه نویسی برای آفیس نیز در آی‌دی‌ای گنجانده شده است. به طور کلی این ویرایش جز بهترین ها محسوب می شود.

2-2-5. ویرایش آفیس[5]

این نسخه در حقیقت یک اس‌دی‌کی[6] است که به ویژوال استدیو اضافه می شود تا امکان برنامه نویسی برای برای برنامه های آفیس شامل اکسل، ورد، اینفوپس، اوت‌لوک و اکسس را فراهم آورد. ویژگی آن شبیه ویرایش استاندارد است با این تفاوت که از کامپایلر مخصوص پردازنده های 64 بیتی پشتیبانی نمی کند ولی در عوض از از اس‌کیو ال سرور ویرایش دولوپر بهره می گیرد. تنها زبان هایی که در وی‌اس‌تی‌او[7] کاربرد کاربرد دارند ویژوال بیسیک و ویژوال سی‌شارپ هستند.

2-2-6. ویرایش مخصوص تیم نرم‌افزاري[8]

کامل‌ترین ویرایش ویژوال استدیو می باشد که به طور خلاصه وی‌اس‌تی‌اس[9] نامیده می شود .این نسخه تمام امکانات نسخه‌ی حرفه ای را فراهم می آورد و علاوه بر آن از پردازنده های ایتانیوم[10] هم پشتیبانی می کند. این ویرایش مخصوص گروه های توسعه دهنده نرم‌افزار است و ابزار های ویژه‌ای در این راستا دارد. چهار ویرایش اصلی تیم سیستم عبارت اند از:

1. نسخه معماری[11]
2. نسخه پایگاه‌داده[12]
3. نسخه توسعه[13]

---

[1] Sql Server Developer Edition
[2] Remote Debugging
[3] Crystal Report
[4] Full Setup Project
[5] Tools for Office Edition
[6] SDK
[7] VSTO
[8] Team System Edition
[9] VSTS
[10] Itanium
[11] Architecture Edition
[12] Database Edition
[13] Development Edition

4. نسخه تست[1]

که در یک بسته‌ی کلی با نام مجموعه تیم[2] گرد هم آمده اند. در ویژوال استدیو 2010 نیز این نسخه ها درون نسخه توسعه جای می‌گیرند.

2-3. معرفي ويژگي هاي مهم هر يک از نسخه‌هاي ويژوال استديو

2-3-1. ویژوال استدیو 97

بیش از ده سال از توزیع نسخه اول ویژوال استودیو می‌گذرد. اولین نسخه از این نرم‌افزار سال 1997 به بازار آمد و به نام ویژوال استدیو 97 مشهور شد. برای اولین‌بار برنامه‌ای درست شد که تعداد زیادی ابزار برنامه‌نویسی را در خود جا داده بود و برنامه‌هایی مانند ویژوال بیسیک 5.0[3]، ویژوال سی‌پلاس‌پلاس 5.0[4]، 5.0[4]، ویژوال جی‌پلاس‌پلاس 1.1[5]، ویژوال فاکس‌پرو 5.0[6] و ویژوال اینتردو[7] را شامل می‌شد. کاربرد هر هر یک از زبانهای بالا در زیر آورده شده است.

1. ویژوال بیسیک و ویژوال سی‌پلاس‌پلاس: برای برنامه نویسی تحت ویندوز
2. ویژوال جی‌پلاس‌پلاس: برنامه نویسی با سینتکس‌های جاوا
3. ویژوال فاکس‌پرو: برای برنامه نویسی ایکس‌بیس[8]
4. ویژوال اینتردو: برای تولید صفحات دینامیکی وب‌گاه‌ها با استفاده از ای‌اس‌پی[9]

ویژوال سورس‌سیف[10] بخشی از مجموعه‌ی ویژوال استدیو شرکت مایکروسافت که برای انجام عملیات کنترل سرس طراحی شده‌است. این برنامه‌ی اختیاری در صورت نصب، قابلیت کنترل و پی‌گیری ورژن‌های مختلف کد را به برنامه‌نویس می‌دهد که برای برنامه‌های پیچیده و به خصوص با چند برنامه‌نویس قابلیتی حیاتی است.

---

[1] Test Edition
[2] Team Suite
[3] Visual Basic 5.0
[4] Visual C++ 5.0
[5] Visual J++ 1.1
[6] Visual FoxPro 5.0
[7] Visual InterDev
[8] xBase
[9] Active Server Pages
[10] Visual SourceSafe

ویژوال سی مجموعه‌ی به هم پیوسته‌ای است که تمامی زنجیره‌ی ابزار توسعه‌ی برنامه را یکجا گرد آورده است. زنجیره‌ی ابزار فوق‌الذکر شامل ویرایشگر، کامپایلر، لینکر، ابزار ساخت، دیباگر و اسمبلر مخصوص ویژوال سی می‌باشد که هریک علاوه بر داشتن خواص برنامه‌های قدیمی‌تر، دارای قابلیت‌های منحصر به فردی هم می‌باشند . محیط ویژوال سی‌پلاس‌پلاس بخشی از مجموعه‌ای بزرگ‌تر به نام مایکروسافت ویژوال استدیو است . نسخه 97 همچنین کتابخانه اماس‌دی‌ان[1] که راهنمای کامل برنامه های ویژوال استدیو مایکروسافت است را معرفی نمود . در ویژوال استدیو 97 مذکور همگی از یک محیط استفاده می‌کردند که استدیو توسعه‌دهنده[2] خوانده می‌شد ئر حالی که ویژوال‌بیسیک و ویژوال فاکس‌پرو نیز از محیط های جداگانه‌ای استفاده می‌کردند . ویژوال استودیو 97 در دو نسخه حرفه‌ای و تجاری ارائه گردید . این نسخه از ویژوال استودیو، اولین تلاش شرکت مایکروسافت در تولید محیط تولید نرم‌افزاری برای ساختن برنامه‌هایی با زبان‌های متفاوت بود. (محیط توسعه چند زبانه) این محصول مایکروسافت در آن سال‌ها تقریباً، جوابگوی همه نوع سلیقه‌ای بود و برنامه‌نویسان زیادی را به سوی خود کشید.

2-3-2. ویژوال استدیو 97 یا 6

یک سال پس از ارائه نسخه ویژوال استدیو 97 (یعنی در سال 98)، نسخه 6 ویژوال استدیو بعنوان آخرین نسخه‌ای که می‌توانست در پلتفرم وین9ایکس[3] اجرا شود ، به بازارعرضه شد . از سال 98 تا سال 2002 مایکروسافت نسخه جدیدی از ویژوال استودیو را معرفی نکرد و ورژن تمام فایل های داخلی[4] آن طی این 4 سال به ورژن 6 ارتقاء یافت که بهمین دلیل ویژوال استدیو 98 را با نام ویژوال استدیو 6 نام گذاری نمودند . این آخرین نسخه‌ای بود که شامل ویژوال بیسیک معروف و دوست داشتنی و ویژوال جی‌پلاس‌پلاس بود . نسخه‌های بعدی ویژوال بیسیک کاملا متفاوت از نسخه کلاسیک آن شدند و جزء زبانهای دات نتی قرار گرفتند . اگر چه هدف دراز مدت مایکروسافت متحد کردن ابزارهای برنامه نویسی تحت یک محیط واحد بود و لی در حقیقت این نسخه نسبت به نسخه ویژوال استادیو 97، چند محیط اضافه تر نیز داشت .ویژوال جی‌پلاس‌پلاس و ویژوال اینتردو از محیط ویژوال سی‌پلاس‌پلاس جدا شدند در حالی ویژوال‌بیسیک و ویژوال‌فاکس‌پرو نیز همچنان مانند نسخه قبلی در محیط های جدا بودند .

[1] MicroSoft Developer Network library
[2] Developer Studio
[3] Win9x
[4] File Format Internal

فصل سوم

موتورهاي جستجو

در این فصل قصد ایجاد آشنایی هر چند مختصر با موتورهای جستجو را داریم تا از جنبه های گوناگون آنها را تحلیل و با اصطلاحات موجود در این زمینه اندکی آشنا شویم.

موتور جستجو، در فرهنگ رایانه، به طور عمومی به برنامه‌ای گفته می‌شود که کلمات کلیدی را در یک سند یا بانک اطلاعاتی جستجو می‌کند. در اینترنت به برنامه‌ای گفته می‌شود که کلمات کلیدی[1] موجود در اسناد و صفحات وب[2]، گروه‌های خبری[3]، منوهای گوفر[4] و آرشیوهای اف تی پی[5] را جستجو می‌کند.

1-3. انواع موتور جستجو

1-1-3. از لحاظ حوزه فعالیت

موتورهای جستجو از لحاظ حوزه فعالیت به دو نوع تقسیم می‌شوند:

1-1-1-3. موتور جستجو وب سایت

برخی از موتور جستجوها برای تنها یک وب سایت به کار برده می‌شوند و در اصل موتور جستجو اختصاصی آن وب‌گاه هستند و تنها محتویات همان وب سایت را جستجو می‌کنند[6].

2-1-1-3. موتور جستجو وب

[1] Keyword's
[2] Web Page's
[3] News Group's
[4] Gopher
[5] File Transfer Protocol
[6] Web Site Search Engine

برخی دیگر نیز ممکن است با استفاده از اسپایدرها[1] محتویات وب‌سایت‌های زیادی را پیمایش کرده و چکیده‌ای از آن را در یک پایگاه اطلاعاتی به شکل شاخص‌گذاری‌شده[2] نگهداری می‌کنند. کاربران سپس می‌توانند با جستجو کردن در این پایگاه داده به پایگاه وبی که اطلاعات موردنظر آن‌ها را در خود دارد پی ببرند.[3]

3-1-2. از لحاظ کارکرد

موتورهای جستجو به دو دسته کلی تقسیم می‌شوند. موتورهای جستجو پیمایشی (خودکار) [4] و فهرست‌های تکمیل‌دستی (غیر خودکار) [5]. هر کدام از آن‌ها برای تکمیل فهرست خود از روش‌های متفاوتی استفاده می‌کنند؛ البته لازم به ذکر است که گونه‌ای جدید از موتورهای جستجو تحت عنوان ابرموتور جستجو[6] نیز وجود دارد که در ادامه به توضیح هر یک از این موارد خواهیم پرداخت. هر چند موتورهای جستجویی نیز وجود دارند که از ترکیبی از موارد فوق برای انجام جستجو بهره می برند.

3-1-2-1. موتورهاي جستجو پيمايشي

موتورهای جستجو پیمایشی مانند گوگل فهرست خود را بصورت خودکار تشکیل می‌دهند. آنها وب را پیمایش کرده، اطلاعاتی را ذخیره می‌کنند؛ سپس کاربران از میان این اطلاعات ذخیره شده، آنچه را که می‌خواهند جستجو می‌کنند. اگر شما در صفحه وب خود تغییراتی را اعمال نمایید، موتورهای جستجو پیمایشی آن‌ها را به طور خودکار می‌یابند و سپس این تغییرات در فهرست‌ها اعمال خواهد شد. عنوان، متن و دیگر عناصر صفحه، همگی در این فهرست قرار خواهند گرفت. وجه مشخصه این گروه از موتور جستجوها وجود نرم‌افزار موسوم به اسپایدر در آن‌هاست. این شبه نرم‌افزار کوچک بصورت خودکار به کاوش در شبکه جهانی پرداخته و از پایگاه‌های وب یادداشت‌برداری و فهرست‌برداری می‌کند سپس این اطلاعات را برای تجزیه و تحلیل و طبقه‌بندی به بانک اطلاعاتی موتور جستجو تحویل می‌دهد.

3-1-2-2. فهرست‌هاي دست‌نويس شده

---

[1] Spider's
[2] Indexed Content's
[3] Web (Internet) Search Engine
[4] Web Crawler (Automated) Search Engine's
[5] Human Powered Directories (Non-Automated) Search Engine's
[6] Meta Based Search Engine's

فهرست‌های دست‌نویس‌شده‌یا فهرست باز[1] مانند پرتال یاهو دایرکتوری[2] وابسته به کاربرانی است که آن را تکمیل می‌کنند. شما صفحه مورد نظر را به همراه توضیحی کوتاه در فهرست ثبت می‌کنید یا این کار توسط ویراستارهایی که برای آن فهرست در نظر گرفته شده، انجام می‌شود. عمل جستجو در این حالت تنها بر روی توضیحات ثبت شده صورت می‌گیرد و در صورت تغییر روی صفحه وب، روی فهرست تغییری به وجود نخواهد آورد. چیزهایی که برای بهبود یک فهرست‌بندی در یک موتور جستجو مفید هستند، تأثیری بر بهبود فهرست‌بندی یک دایرکتوری ندارند. تنها استثناء این است که‌یک سایت خوب با پایگاه داده‌ای با محتوای خوب شانس بیشتری نسبت به‌یک سایت با پایگاه داده ضعیف دارد. البته در مورد موتورهای جستجو مشهور مانند گوگل و یاهو، یک مولفه دیگر هم برای بهبود فهرست‌بندی وجود دارد که کمک مالی[3]است، یعنی وب‌گاه‌هایی که مایل به بهبود مکان وب‌گاه خود در فهرست بندی هستند، می‌توانند با پرداخت پول به این موتور جستجوها به هدف خویش برسند.

3-2-1-3. موتورهاي جستجو تركيبي[4]

به موتورهایی گفته می‌شود که هر دو حالت را در کنار هم نمایش می‌دهند. غالباً، یک موتور جستجو ترکیبی در صورت نمایش نتیجه جستجو از هر یک از دسته‌های فوق، نتایج حاصل از دسته دیگر را هم مورد توجه قرار می‌دهد. مثلاً موتور جستجو ام.اس.ان[5] بیشتر نتایج حاصل از فهرست‌های تکمیل‌دستی را نشان می‌دهد اما در کنار آن نیم نگاهی هم به نتایج حاصل از جستجوی پیمایشی دارد.

4-2-1-3. ابرموتور جستجوها

این گونه جدید از موتور جستجوها که قدمت چندانی نیز ندارند، بصورت هم‌زمان از چندین موتور جستجو برای کاوش در شبکه برای کلید واژه مورد نظر استفاده می‌کنند. بدین معنی که این موتور جستجو عبارت مورد نظر شما را در چندین موتور جستجوِ دیگر جستجو کرده و نتایج آنها را با هم ترکیب کرده و یک نتیجه کلی به شما ارائه می‌دهد. به‌عنوان مثال موتور جستجو داگ پایل[6] از نتایج حاصل از موتورهای اسک[7]، یاهو[8]، بینگ[9] و گوگل[10] استفاده کرده و نتیجه حاصله را به شما ارائه می‌دهد. لازم به ذکر است که

---

[1] Open Directory
[2] Yahoo! Directory
[3] Advertisement
[4] Combined Search Engine's
[5] Microsoft Network (MSN)
[6] Dogpile Web Search at http://www.dogpile.com/
[7] Ask
[8] Yahoo!
[9] Bing
[10] Google

که روش و یا راهکار مشخص و یکسانی برای ترکیب نتایج حاصله از موتورهای پایه (موتورهایی که به عنوان موتور جستجو استفاده می‌شوند مانند یاهو که‌یک موتور پایه برای داگ پایل می‌باشد) وجود ندارد. اما قابلیت جستجو به همه زبانها را ندارد و ظاهرا فقط کلمات انگلیسی را پیدا می‌کند.

3-1-2-5. نوموتور جستجوها

این گونه از موتور جستجوها، نسل جدید و متفاوتی از موتورهای جستجو گذشته هستند. امکان ثبت جستجو و مدل‌سازی فعالیت‌های کاربر و ارائه‌ی نتایج جدید به‌کاربر، به‌صورت متفاوت و تفکیک شده، از امکانات نوموتور جستجوها است.

3-2. بررسي يک موتور جستجو پيمايشي

موتورهای جستجو پیمایشی شامل سه عنصر اصلی هستند. اولی در اصطلاح عنکبوت است که پیمایش‌گر[1] هم نامیده می‌شود. پیمایش‌گر همین که به‌یک صفحه می‌رسد، آن را می‌خواند و سپس پیوندهای آن به صفحات دیگر را دنبال می‌نماید. این چیزیست که برای یک سایت پیمایش‌شده[2] اتفاق افتاده‌است. پیمایش‌گر با یک روال منظم، مثلاً یک یا دو بار در ماه به سایت مراجعه می‌کند تا تغییرات موجود در آن را بیابد. هر چیزی که پیمایش‌گر بیابد به عنصر دوم یک موتور جستجو یعنی فهرست انتقال پیدا می‌کند. فهرست اغلب به کاتالوگی بزرگ اطلاق می‌شود که شامل لیستی از آنچه‌است که پیمایش‌گر یافته‌است. مانند کتاب عظیمی که فهرستی را از آنچه پیمایش‌گرها از صفحات وب یافته‌اند، شامل شده‌است. هرگاه سایتی دچار تغییر شود، این فهرست نیز به روز خواهد شد. از زمانی که تغییری در صفحه‌ای از سایت ایجاد شده تا هنگامی که آن تغییر در فهرست موتور جستجو ثبت شود مدت زمانی طول خواهد کشید. پس ممکن است که‌یک سایت پیمایش‌شده باشد اما فهرست‌شده نباشد. تا زمانی که این فهرست‌بندی برای آن تغییر ثبت نشده باشد، نمی‌توان انتظار داشت که در نتایج جستجو آن تغییر را ببینیم. نرم‌افزار موتور جستجو[3]، سومین عنصر یک موتور جستجو است و به برنامه‌ای اطلاق می‌شود که به صورت هوشمندانه‌ای داده‌های موجود در فهرست را دسته‌بندی کرده و آن‌ها را بر اساس اهمیت طبقه‌بندی می‌کند تا نتیجه جستجو با کلمه‌های درخواست شده هر چه بیشتر منطبق و مربوط باشد.

3-3. رتبه‌بندي صفحات وب توسط موتورهاي جستجو

[1] Crawler
[2] Crawled
[3] Searcher

وقتی شما از موتورهای جستجو پیمایشی چیزی را برای جستجو درخواست می‌نمایید، تقریباً بلافاصله این جستجو از میان میلیون‌ها صفحه صورت گرفته و مرتب می‌شود بطوریکه مربوط‌ترین آنها نسبت به موضوع مورد درخواست شما رتبه بالاتری را احراز نماید. البته باید در نظر داشته باشید که موتور جستجوها همواره نتایج درستی را به شما ارائه نخواهند داد و مسلماً صفحات نامربوطی را هم در نتیجه جستجو دریافت می‌کنید و گاهی اوقات مجبور هستید که جستجوی دقیقتری را برای آنچه می‌خواهید انجام دهید اما موتور جستجوها کار حیرت‌انگیز دیگری نیز انجام می‌دهند. فرض کنید که شما به‌یک کتابدار مراجعه می‌کنید و از وی درباره "سفر" کتابی می‌خواهید. او برای این که جواب درستی به شما بدهد و کتاب مفیدی را به شما ارائه نماید با پرسیدن سؤالاتی از شما و با استفاده از تجارب خود کتاب مورد نظرتان را به شما تحویل خواهد داد. موتور جستجوها همچنین توانایی ندارند اما به نوعی آنها را شبیه‌سازی می‌کنند. پس موتورهای جستجو پیمایشی چگونه به پاسخ مورد نظرتان از میان میلیونها صفحه وب می‌رسند؟ آنها یک مجموعه از قوانین را دارند که الگوریتم نامیده می‌شود. الگوریتم‌های مورد نظر برای هر موتور جستجوی خاص و تقریباً سری هستند اما به هر حال از قوانین زیر پیروی می‌کنند:

1-3-3. مکان و بسامد

یکی از قوانین اصلی در الگوریتم‌های رتبه‌بندی موقعیت، بسامد و تعداد تکرار واژه‌هایی است که در صفحه مورد استفاده قرار گرفته‌اند که بطور خلاصه روش مکان- بسامد[1] نامیده می‌شود. کتابدار مذکور را به خاطر می‌آورِد؟ لازم است که او کتاب‌های در رابطه با واژه "سفر" را طبق درخواست شما بیابد. او در وحله - اول احساس می‌کند که شما به دنبال کتاب‌هایی هستید که در نامشان کلمه "سفر" را شامل شوند. موتور جستجوها هم دقیقاً همان کار را انجام می‌دهند. آنها هم صفحاتی را برایتان فهرست می‌کنند که در برچسب عنوان[2] موجود در کد زبان نشانه‌گذاری اَبَرمتنی[3] حاوی واژه "سفر" باشند. موتور جستجوها همچنین به دنبال واژه مورد نظر در بالای صفحات و یا در آغاز بندها هستند. آنها فرض می‌کنند که صفحاتی که حاوی آن واژه در بالای خود و یا در آغاز بندها و عناوین باشند به نتیجه مورد نظر شما مربوط‌تر هستند. بسامد عامل بزرگ و مهم دیگری است که موتور جستجوها از طریق آن صفحات مربوط را شناسایی می‌نمایند. موتور جستجوها صفحات را تجزیه کرده و با توجه به تکرار واژه‌ای در صفحه متوجه می‌شوند که آن واژه نسبت به دیگر واژه‌ها اهمیت بیش‌تری در آن صفحه دارد و آن صفحه را در درجه بالاتری نسبت به صفحات دیگر قرار می‌دهند.

---

[1] Location/Frequency Method
[2] Title tag
[3] HTML

چگونگی کارکرد دقیق موتور جستجوها درباره روش‌هایی از قبیل مکان- تکرار فاش نمی‌شود و هر موتور جستجوی روش ویژه‌ی خود را دنبال می‌کند. به همین دلیل است که وقتی شما واژه‌های همانندی را در موتورهای متفاوت جستجو می‌کنید، به نتایج متفاوتی می‌رسید. الگوریتم‌های اولیه موتورهای جستجو معتبر و بزرگ همچنان محرمانه نگهداری می‌شوند. برخی موتور جستجوها نسبت به برخی دیگر صفحات بیشتری را فهرست کرده‌اند. نتیجه این خواهد شد که هیچ موتور جستجوی نتیجه جستجوی مشترکی با موتور دیگر نخواهد داشت و شما نتایج متفاوتی را از آن‌ها دریافت می‌کنید. موتور جستجوها همچنین ممکن است که برخی از صفحات را از فهرست خود حذف کنند البته به شرطی که آن صفحات با هرزنامه[1] شدن سعی در گول زدن موتور جستجوها داشته باشند. فرستادن هرزنامه[2] روشی است که برخی از صفحات برای احراز رتبه بالاتر در موتور جستجوها در پیش می‌گیرند و آن به این صورت است که با تکرار بیش از حد واژه‌ها و یا بزرگ نوشتن یا بسیار ریز نوشتن متنها بطور عمدی کوشش در بر هم زدن تعادل و در نتیجه فریب موتور جستجوها دارند. آنها سعی دارند که با افزایش عامل تکرار، در رتبه بالاتری قرار بگیرند. البته آنگونه که گفته شد تعداد تکرارها اگر از حد و اندازه خاصی فراتر رود نتیجه معکوس می‌دهد. موتور جستجوها راه‌های متنوعی برای جلوگیری از فرستادن هرزنامه دارند و در این راه از گزارش‌های کاربران خود نیز بهره می‌برند. امروزه بهینه‌سازی سایت‌های اینترنت برای موتور جستجوها یکی از مهم‌ترین روشهای جلب بازدیدکننده به سایت است.

2-3-3. عوامل خارج از صفحه

موتورهای جستجو گردشی اکنون تجربه فراوانی در رابطه با وب‌دار[3]هایی دارند که صفحات خود را برای کسب رتبه بهتر مرتباً بازنویسی می‌کنند. بعضی از وب‌دارهای خبره حتی ممکن است به سمت روش‌هایی مانند مهندسی معکوس برای کشف چگونگی روش‌های مکان- تکرار بروند. به همین دلیل، تمامی موتورهای جستجو معروف از روش‌های امتیازبندی "خارج از صفحه" استفاده می‌کنند. عوامل خارج از صفحه عواملی هستند که از تیررس وب‌دارها خارجند و آنها نمی‌توانند در آن دخالت کنند و مسأله مهم در آن تحلیل ارتباطات و پیوندهاست. به وسیله تجزیه صفحات، موتور جستجوها پیوندها را بررسی کرده و از محبوبیت آنها می‌فهمند که آن صفحات مهم بوده و شایسته ترفیع رتبه هستند. به علاوه تکنیک‌های پیشرفته به گونه‌ای است که از ایجاد پیوندهای مصنوعی توسط وب‌دارها برای فریب موتور جستجوها جلوگیری می‌نماید. علاوه بر آن موتور جستجوها بررسی می‌کنند که کدام صفحه توسط یک کاربر که واژه‌ای

---

[1] Spam
[2] Spamming
[3] Webmaster

را جستجو کرده انتخاب می‌شود و سپس با توجه به تعداد انتخاب‌ها، رتبه صفحه مورد نظر را تعیین کرده و مقام آن را در نتیجه جستجو جابه‌جا می‌نمایند.

3-4. سرفصل‌هاي بهينه سازي

با توجه به هوشمندتر شدن هر روزه موتورهای جستجو و ارتقا, سطح استاندارد ها در وب، رعایت و بکارگیری موارد زیر می تواند باعث ارتقای سطح وبگاه شما در بین موتورهای جستجو شود. توجه داشته باشید که با رعایت موارد فوق نباید انتظار داشته باشید که طی زمان کوتاهی به تمام خواسته هایتان برسید. مبحث بهینه سازی وب سایت ها برای بدست آوردن رتبه ای بهتر در نتایج موتور های جستجو دارای پروسه ای نسبتا زمانبر و طولانی است. هر چند سعی تمام توسعه دهندگان موتور های جستجو افزایش این سرعت و به حداقل رساندن بازه زمانی این پروسه است. برخی از موارد تاثیر گذار در افزایش رتبه وبگاه شما نزد موتور های جستجو عبارتند از:

1. بازنویسی محتوای سایت با توجه به هدف و با مساعدت شما
2. تحقیق و انتخاب کلمات کلیدی مرتبط با فعالیت و هدف سایت
3. معرفی کامل وب‌گاه به موتورهای جستجو مشهور مانند گوگل، بینگ، **اکسپلوریم**، یاهو و...
4. انتخاب توضیحات متناسب با صفحات سایت
5. بررسی و نحوه تعیین استراتژی ساختار لینک‌ها
6. طراحی مجدد صفحات سایت با توجه به تنوع مطالب
7. قراردادن توضیحات به صورت متنی در قالب جزء و کل
8. ایندکس صفحات سایت
9. افزایش بازدیدکننده هدفمند بر اساس کلمات مرتبط با فعالیت سایت

در پایان این فصل خاطر نشان می کنم که تحقیق درباره موتورهای جستجو باعث شد من نسبت به خیلی موارد دید بهتری پیدا کنم و این باعث شد در ادامه در مواجهه با مشکلات، به قبل بازگشته و با مرور تحقیقات انجام شده و کمی تامل اقدام به کنار زدن مشکلات کنم.

فصل چهارم

جستجو در وب با اکسپلوریم.نت[1]

[1] Xplorium.NET Web Search Engine

اکسپلوریم.نت یک پروژه شبیه سازی شده از موتورهای جستجوی امروزی بصورت کلی و سطحی است که با هدف آشنایی با ساختار موتورهای جستجو به عنوان پروژه کاردانی انتخاب شده است. از آنجا که پروژه بصورت پیاده سازی یک موتور جستجو است در نتیجه من مانور زیادی روی بخش مستندات نمی دهم. در پیاده سازی این موتور سعی شده است از مواردی که در لابه لای تحقیق صورت گرفته، یافت شده است استفاده شود تا هم پروژه، پروژه ای خوبی از آب درآید و هم در این بین به اطلاعات من افزوده شود.

1-4. ابزارهاي توسعه دهنده اکسپلوریم.نت

شاید ذکر این نکته خالی از لطف نباشد که در پیاده سازی این پروژه از چه واسطی[1]، چه زبان برنامه نویسی، چه پایگاه داده ای و از چه ابزار های ثالثی[2] استفاده شده است.

آی دی ای یا محیط توسعه مجتمع! مورد استفاده، مایکروسافت ویژوال استودیو 2010[3] می‌باشد. از جمله ویژگی های این نرم افزار می توان به پشتیبانی از 3 زبان قدرتمند، سازگاری کامل با نرم افزار اس کیو ال سرور[4] برای ایجاد برنامه های دی‌دی‌دی[5] ، توانایی ایجاد برنامه های تحت ویندوز و تحت وب، قابلیت بینظیر در اشکال زدایی[6] و تست[7] پروژه ها و همچنین قابلیت ها، توانایی ها و ابزار های فراوانی که مجال پرداختن به آنها نیست اشاره کرد.

---

[1] Integrated Development Environment
[2] Third-Party Tool's
[3] Microsoft Visual Studio 2010
[4] Microsoft SQL Server
[5] Data Driven Development
[6] Debuging
[7] Unit Testing

زبان برنامه نویسی استفاده شده در این پروژه همانند تمام پروژه های ریز و درشت دیگری که تا به حال انجام داده‌ام زبان سی‌شارپ[1] است که نزد تمامی برنامه نویسان کاملا شناخته شده است و نیازی به توضیح درباره‌ی آن نمی بینم.

همان طور که متوجه شده اید پایگاه داده ای پروژه نیر یکی از محصولات مایکروسافت و در واقع تنها محصول این شرکت در این زمینه می‌باشد. همان گونه که در هنگام طراحی و پیاده سازی پروژه (به غیر از موارد مربوط به پایگاه داده) از یک محیط توسعه مجتمع فوق العاده غنی و قدرتمند استفاده کرده‌ام در این مورد نیز یک آی دی ای قدرت مند بنام مایکروسافت اس کیو ال سرور منیجمنت استودیو، در امر طراحی و ساخت پایگاه داده به ما کمک می کند.

از جمله دیگر نرم افزار ها و ابزار های ثالثی که در امر توسعه مورد استفاده قرار گرفته اند می توان به موارد زیر اشاره کرد:

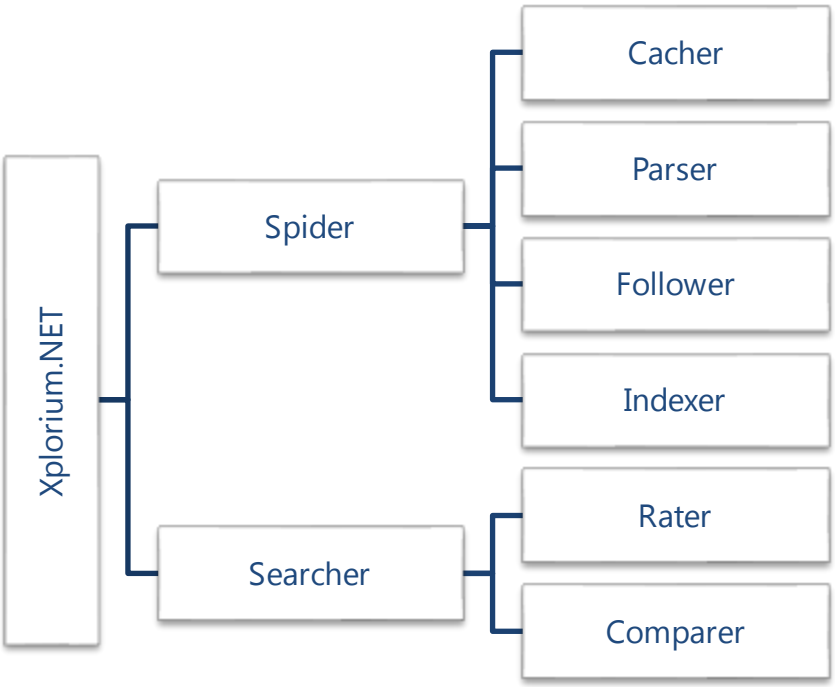
- Visual Studio Add-Ins:
  - JetBrains ReSharper
  - NDepend
- Microsoft IIS 7.5
- Mozilla Firefox + Add-Ons:
  - Web Developer
  - FireBug
  - Yahoo! YSlow

2-4. کمي بيشتر بدانيم

به طور خلاصه اکسپلوریم.نت یک موتور جستجوی پیمایشی است؛ که از چندین و چند بخش کوچک و بزرگ مرتبط با هم تشکیل شده است که در کنار یکدیگر کار می کنند و فعالیت هر یک بر دیگر تاثیر مستقیم دارد. فعالیت این موتور به صورت سلسله مراتبی است که به دنبال هم انجام شده و درصورتی که در فعالیت یکی از بخش ها مشکلی بوجود آید تقریبا بخش بعدی توانایی انجام وظایف خود را بدرستی نخواهد داشت. اکسپلوریوم.نت همانند اکثر موتور های جستجو از 2 بخش اصلی تشکیل شده است. بخش ابتدایی که اسپایدر نامیده می شود وظیفه اسپاید کردن محتوای وب را بر عهده دارد. منظور از اسپایدینگ دریافت محتوای صفحات، تفسیر و ترجمه، حذف موارد ضائد، یافتن متاتگ ها، آدرس ها و کلمات موجود در صفحه و ذخیره آنها و... وظیفه ریز و درشت دیگر است که بر عهده این بخش نهاده شده است. به طور کل

[1] C Sharp (C#)

اسپایدر از 2 زیر بخش کچر[1] و ایندکسر[2] تشکیل شده است که موارد نامبرده شده در بالا از جمله وظایف این دو می باشند. دومین بخش از اکسپلوریم.نت را جستجوگر[3] آن تشکیل می دهد. این بخش عمل جستجو در محتوای آماده شده[4] توسط اسپایدر را انجام داده و خروجی را به کاربر نمایش می دهد. این بخش نیز از زیر بخش هایی از قبیل رتبه دهنده[5] و مقایسه گر[6] متشکل شده است. در زیر ساختار کلی از بخش های مختلف را مشاهده می کنید.



شکل 1-1 نمایی کلی از ساختار اکسپلوریوم.نت

3-4. اسپایدر

تاکنون قطعا ذهنیتی هر چند اندک درباره‌ی ساختار این جستجوگر را پیدا کرده‌اید. این ذهنیت قطعا در ادامه باعث می شود که درک شما از توضیحات بیشتر و بیشتر گردد. اکنون قصد دارم بیشتر بر روی اسپایدر زوم کنم. اینکه این بخش وظیفه‌ی انجام فلان کار را دارد در حد یک تئوری بیان شد. حالا باید ببینیم این وظایف در عمل چگونه پیاده سازی شده‌اند؟ از چه ابزاری برای این کار استفاده شده است؟ آیا

[1] Cacher
[2] Indexer
[3] Searcher
[4] Parsed Content's
[5] Rater
[6] Comparer

روش هایی که من بکار برده‌ام بهترین و بهینه‌ترین اند؟ تمامی موارد در ادامه به تفضیل توضیح داده خواهند شد.

4-3-1. کچر

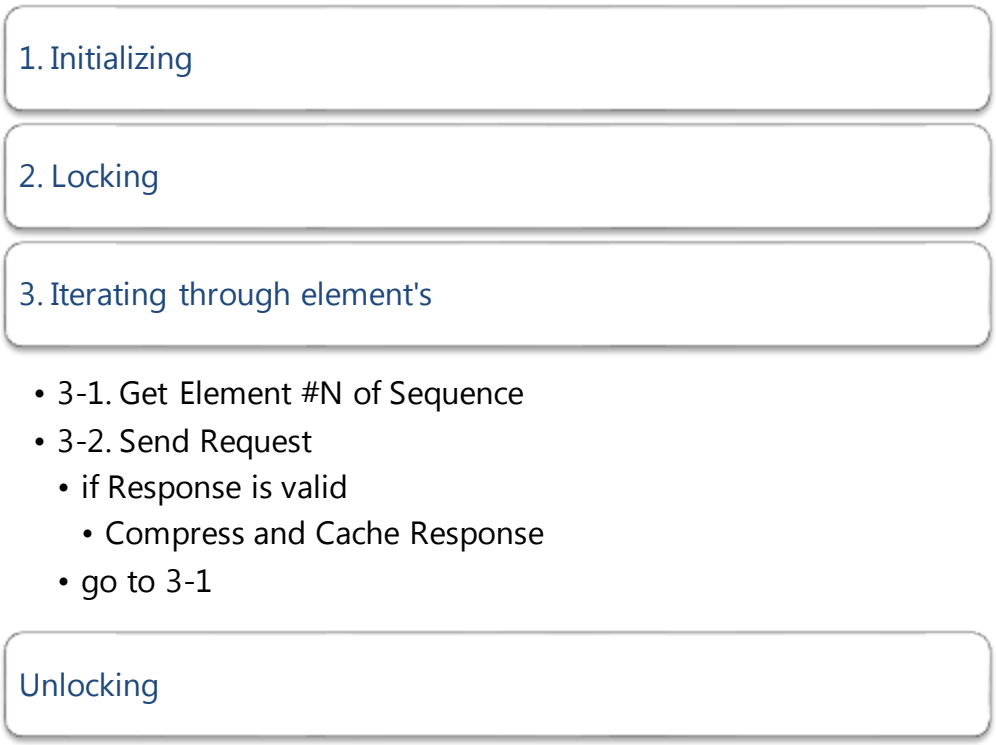
قطعا جایگاه این بخش در اکسپلوریم.نت جایگاه بسیار مهمی می‌باشد. وظیفه این بخش ذخیره محتوای صفحات برای انجام ترجمه، ایندکس و... می‌باشد. در صورتی که این بخش بدرستی عمل نکند، داده های مورد نیاز برای کارکرد صحیح بخش ایندکسر وجود ندارد و در نتیجه بخش سرچر نیز عملا توان انجام وظیفه‌ی خود را ندارد.

از آنجایی که اگر محتوای صفحات را در اختیار نداشته باشیم، امکان انجام هیچگونه کاری را نداریم و در اختیار داشتن محتوای صفحات لزوما نیازمند اتصال دائمی به اینترنت است؛ تصمیم به این شد که‌یک بخش به قسمت های موجود در اسپایدر اضافه شود تا عمل ذخیره سازی محتوای صفحات را سوای انجام هر عمل اضافه ای انجام دهد. انجام عمل کچینگ بسیار کار آسانی است. تنها کافیست N آدرس را از بانک بارگذاری کرده و شروع به کچ کردن صفحات متناظر با آدرس های ذخیره شده کنیم.

تذکر: برنامه ای که وظیفه مدیریت بخش اسپایدر را بر عهده دارد قابلیت انجام بیش از یک عمل کچینگ را در آن واحد دارد. لذا برای جلوگیری از بارگذاری چندباره‌ی یک آدرس توسط چند کچر، هر آدرسی که‌اماده کچینگ می شود، به حالت قفل شده[1] تغییر وضعیت داده تا از این مورد جلوگیری شود. با این مکانیسم ما مطمئن خواهیم بود که که هر آدرس فقط یکبار و توسط یک کچر بارگذاری شده و هیچ بارگذاری بیهوده ای در هنگام کچینگ انجام نمی شود. همین عمل هنگام بارگذاری محتوای ذخیره شده نیز مورد استفاده قرار می گیرد.

به مهز اینکه‌یک آدرس آماده کچ شدن شد، یک درخواست به آدرس تصحیح شده آن ارسال شده و پاسخ برگشتی از وب سرور بعد از فشرده شدن داخل پایگاه داده ذخیره می شود. این روند تا پایان یافتن تمامی آدرس ها ادامه می یابد؛ هر چند کاربر قادر است عمل کچینگ را در هر جایی که نیاز دید لغو کند. بعد از اتمام عمل کچینگ، در هر صورتی، رکورد هایی که فقل شده‌اند از حالت فقل خارج می شوند. در زیر شبه الگوریتمی از اعمال انجام شده توسط کچر را مشاهده می کنید.

[1] Locked

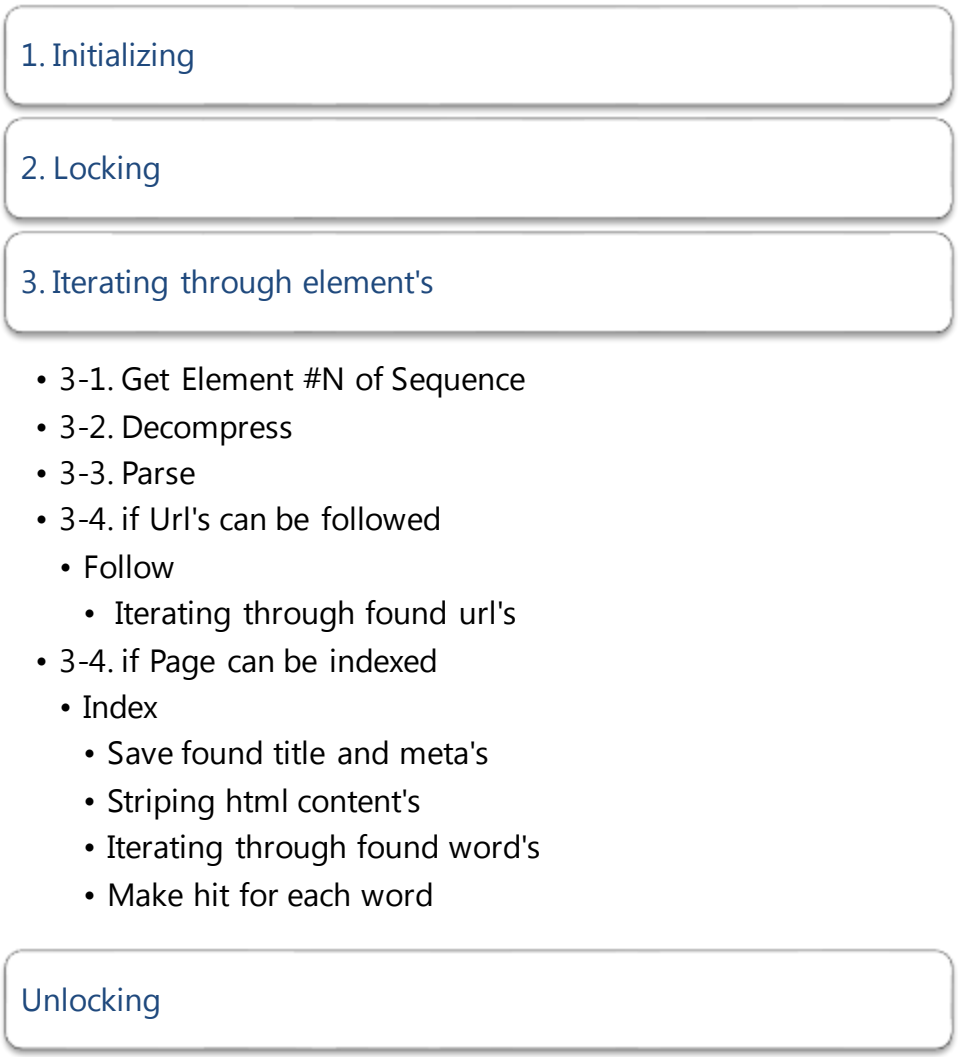


شکل 1-2 شبه الگوریتمی از کچر

تذکر: در این بین موارد زیادی انجام می شود که‌امکان توضیح دادن همه آنها وجود ندارد؛ هر چند شما می توانید برای اطلاع دقیق به سورس پروژه مراجعه فرمایید.

2-3-4. ایندکسر

این بخش بر خلاف کچر دارای ساختار پیچیده تریست، در حالی که دارای قسمت هایی مشابه آنچه در کچر وجود دارد می‌باشد. قبلا به این نکته اشاره کرده بودم که کچر اطلاعات مورد نیاز ایندکسر را آماده می کند. حال ایندکسر با استفاده از این منابع آماده شده، اقدام به انجام عمل ایندکسینگ می کند. در واقع ایندکسر خود به تنهایی قادر به انجام عمل ایندکسینگ نیست و در این بین از بخش های دیگر که در شکل 3-1 آورده شده‌اند کمک می گیرد. بد نیست ابتدا به شبه الگوریتم ایندکسر نگاهی گذرا بیاندازیم.



شکل 2-1 شبه الگوریتمی از ایندکسر

تذکر: در این بین موارد زیادی انجام می شود که‌امکان توضیح دادن همه آنها وجود ندارد؛ هر چند شما می توانید برای اطلاع دقیق به سورس پروژه مراجعه فرمایید.

3-3-4. پارسر

همانطور که از نام این بخش پیداست، وظیفه‌ی ترجمه و استخراج قسمت هایی که لازمه کار ماست را بر عهده دارد. برای ترجمه و استخراج منابع، از رجاکس[1] استفاده شده است. زبان یا بهتر بگویم مکانیزمی که در اکثر زبانهای برنامه نویسی امروزی وجود دارد و با استفاده از قدرت خیرکننده و سینتکس ساده قادریم تا هر نوع رشته ای را با هر پترن خاصی در یک رشته دیگر جستجو کنیم. در صورتی که اگر بخواهیم همین عمل پارسینگ را بوسیله توابع ساده کار با رشته ها پیاده سازی کنیم، نه که نمی شود ولی قطعا

[1] Regular Expression

کاریست بسیار دشوار. لازم به ذکر است عبارات باقاعده مورد استفاده در پروژه درون کلاس PreparedExpressions گردآوری شده‌اند تا خوانایی برنامه بالا رود. ابتدا نگاهی به شبه الگوریتم استفاده شده در پارسر می اندازیم.

1. Extracting Title tag
2. Iterating through found meta tag's
   - Iterating through meta tag attribute's
     - Holding meta name and value
     - Processing kept value's
3. Iterating through found A tag's
   - Iterating through A tag attribute's
   - Converting found href value to absolute url
4. Striping page content
   - Removing script's, style's, break line's, white space's, unwanted tag's and ...
5. Iterating through found word's
   - Filtering duplicated word's
   - Removing unwanted word's based on WordFilteringMode

شکل 1-2 شبه الگوریتمی از پارسر

تذکر: در این بین موارد زیادی انجام می شود که‌امکان توضیح دادن همه آنها وجود ندارد؛ هر چند شما می توانید برای اطلاع دقیق به سورس پروژه مراجعه فرمایید.

4-4. سرچر

به طور حتم مهم ترین بخش یک موتور جستجو، سرچر آن است. بخشی که عمل جستجو و واکاوی اطلاعات دخیره شده را بر عهده دارد تا با این کار بتواند فرد جستجوگر را در یافتن اطلاعات مورد یاری کند. تا کنون هر چه گفتیم و انجام داده ایم فقط ذخیره و آماده سازی داده های خام بود و بس. در واقع کار

اصلی ما تماما به این بخش وابسته است. اسپاید کردن محتوای وب هر چند کاریست زمانبر اما انجام آن خارج از دید کاربر است و این روند هر چقدر هم طولانی و طاقت فرسا باشد مهم نیست. چیزی که مهم است این موضوع است که ما تا می توانیم زمان جستجو را بهینه و کم کنیم تا کاربر کمترین معطلی را از زمان فشردن کلید اینتر تا نمایش نتایج جستجو داشته باشد. قطعا اگر بخواهیم در این بین عمل رتبه بندی نتایج و مرتب سازی را نیز لحاظ کنیم فشار زیادی به برنامه خواهد آمد. مثلا شما فرض کنید باید میان 500000 رکورد جستجو کرده، دراین بین 25000 رکورد با عبارت مورد جستجو تطابق دارند. حال در بین این 25000 مورد، ممکن است یکی دارای فقط یک کلمه از عبارت مورد نظر ما باشد (منظور یکبار تکرار شده باشد) و دیگری بیش از یک بار. در ادامه احتمال دارد که در بین این 25000 نتیجه تعدادی مورد تکراری نیز موجود باشد. تشخیس و حذف موارد تکراری را کامپیرر بر عهده دارد. در حالی که رتبه بندی نتایج بر عهده ریتر می‌باشد. در ادامه به شرح هر یک می پردازیم.

4-4-1. ریتر

در ادامه بحث ریتر، حال چگونه باید این عمل مقایسه و رتبه بندی را به نحوی انجام دهیم که نتایج تا جای ممکن به عبارت مورد جستجو توسط کاربر نزدیک تر باشند؟ قطعا امکان این نیست که 25000 رکورد را با یکدیگر مقایسه کنیم؛ زیرا این کار زمان بسیار طولانی را می طلبد؛ و این مهمترین شاخص در انجام جستجوست. پس ما اقدام به محاسبه‌یک ریت برای هر مورد یافت شده می کنیم. هر چند این مکانیزم نیز احتیاج به زمانی نسبتا طولانی دارد. (حدود 250 میلی ثانیه برای 1000 رکورد) تا بحال یکی دو روش دیگر نیر به ذهنم رسیده است که در حال کار برروی آنها نیز هستم و فعلا در حالت آزمایشی قرار دارند! مثلا اینکه از یک مکانیزم جدید برای رتبه بندی استفاده کنم. در این مکانیزم، همانند اسپاید کردن محتوا، قبل از عمل جستجو و کلا در یک پروسه جداگانه عمل ریتینگ انجام می شود. ما سعی می کنیم یه اضای هر کلمه ای که ثبت کرده ایم یک ریت برای آن در صفحات مختلف محاسبه کنیم. سپس در هنگام جستجو براساس امتیازات تعلق گرفته به هر کلمه، مرتب سازی را انجام دهیم. در این صورت دیگر هنگام جستجو نیاز به محاسبه ریت برای هر نتیجه نیست. خوب حال این ریتی که گفتم چگونه محاسبه می شود؟ در اینجا نیز مانند بقیه موارد از خود ابتکار نشان داده‌ام! حال خط زیر را در نظر بگیرید.

```
Rate = new Rater().Rate(brokeQuery, urls.ResolvedPath, parsedContents.Title,
parsedContents.Keywords, parsedContents.Description, words)
```

با ارسال پارامتر های زیر به ریتر، عمل محاسبه‌امتیاز را انجام می شود. اولین پارامتر یک لیست از کوئری وارد شده توسط کاربر است. برای درک بیشتر مثال زیر را در نظر بگیرید.

User raw query: Microsoft –Apple host:microsoft.com

Broke query: Microsoft باشد

-Apple نباشد

host:microsoft.com فقط در این هاست

پارامتر های 2 تا 4 نیازی به توضیح ندارد. و آخرین پارامتر نیز لیستی از کلمات پیدا شده در صفحه مورد نظر است. برای هر عنصری که در پارامتر brokeQuery وجود دارد ما عملیات زیر را انجام می دهیم.

- امتیاز مربوط به آدرس صفحه (بر اساس تعداد تکرار) * 100
- امتیاز مربوط به عنوان صفحه * 50
- امتیاز مربوط به کلمات کلیدی صفحه * 25
- امتیاز مربوط به توضیحات صفحه * 10
- امتیاز مربوط به کلمات موجود در صفحه

منظور از "امتیاز مربوط به x موجود در صفحه" محاسبه تعداد تکرار y در x می‌باشد. فرض کنید x و y مقادیر زیر را دارا می باشند.

y: Microsoft

x: Microsoft, Microsoft Corporation, Microsoft Product, Microsoft Windows, Microsoft Partners etc. (Page keywords)

طبق فرمول بالا، امتیاز محاسبه شده برای این مورد برابر است با:

x.Count(y) * 25 = 125;

حال در صورتی که y با علامت - شروع شده باشد منظور کاربر عدم وجود این کلمه است و بلعکس. در هر حالت ما اقدام به کم یا زیاد کردن امتیاز محاسبه شده از کل امتیازات می کنیم به همین راحتی.

در ضرایبی که برای هر بخش از صفحه در نظر گرفته شده است، سعی بر این است که تا عمل رتبه بندی با دقت بیشتری صورت گیرد. حال شاید بپرسید چرا ضریب آدرس 100 است. اگر تا بحال سورس سایت هایی مثل پی سی دانلود یا هر سایت دیگری که قصد گول زدن موتور های جستجو را دارند را دیده باشید خواهید فهمید که کلمات کیلیدی را علاوه بر اینکه در قسمت Keywords به تعداد زیادی تکرار

کرده اند؛ در خود بدنه صفحه نیز تعداد بسیار بیشتری از همان کلمات کلیدی را تکرار کرده اند. هر چند با مخفی کردن المنت در بر گیرنده این کلمات، عملا کاربر قادر به مشاهده انها نیست. قطعا الگوریتم فوق الذکر هنگام امتیاز دهی به چنین سایت هایی دچار مشکل شده و امتیازی بدور از واقعیت را محاسبه خواهد کرد. حال یکی از راه حل هایی که به ذهنم رسید دادن ضریب به بخش های مختلف صفحه در هنگام امتیاز دهی می‌باشد. مثلا در مثال بالا که کاربر عبارت Microsoft را جستجو می کند قطعا آن صفحه هایی که در آدرس آنها عبارت مذکور وجود دارد در نتایج رتبه بالاتری را از آن خود می کنند و این یعنی سایتی با آدرس http://*microsoft* در جایگاه بالاتری نسبت به دیگران قرار می گیرد. قطعا این راه حل همیشه درست عمل نخواهد کرد. اما حداقل برای شروع کار بد نیست.

احتمالا توضیحات کمی نامفهوم است. در این صورت به سورس کد مراجعه کنید.

سخن آخر

تا بدین جا هر چند بسیار ناچیز اما با این پروژه آشنا شده اید. و من نیز قصد توضیحات بیشتر را ندارم. چون کلا با نوشتن میانه خوبی ندارم و ترجیح می دادم همین چند روزی را که صرف آماده سازی این داکیومنت کردم، صرف کار بر روی پروژه می کردم. در هر صورت اگر کم و کاستی (که قطعا وجود دارد) وجود داشت خواهید گذشت. ان شا, الله برای آینده فکرهایی در سر دارم که اگر تحقق پیدا کرد حتما شما استاد عزیز را نیز با خبر خواهم ساخت تا از راهنمایی های شما همانطور که تاکنون استفاده کرده‌ام در آینده نیز بهره ببرم.

با تشکر - محمد صادق شاد

منابع و ماخذ

برخی از منابعی که در انجام و پیشروی پروژه مورد استفاده قرار گرفته اند در زیر آورده‌ام.

1. **Web search engine** at http://en.wikipedia.org/wiki/Web_search_engine
2. **List of search engines** at http://en.wikipedia.org/wiki/List_of_search_engines
3. **Difference between Spider, Crawler and Robot** at http://forums.seochat.com/search-engine-spiders-27/difference-between-spider-crawler-and-robot-244471.html
4. **The Anatomy of a Large-Scale Hypertextual Web Search Engine** at http://infolab.stanford.edu/~backrub/google.html
5. **How Google Works** at http://www.googleguide.com/google_works.html
6. **Google's New Web Page Spider** at http://www.searchenginepromotionhelp.com/m/articles/search-engine-optimization/googles-new-spider.php
7. **Index (search engine)** at http://en.wikipedia.org/wiki/Index_(search_engine)
8. **Search Engine Glossary** at http://searchenginewatch.com/2156001