**نام مقاله: خزنده و ساختاروب**

**نام نشريه: فصلنامه كتابداري و اطلاع رساني (اين نشريه در www.isc.gov.ir نمايه مي شود)**

**شماره نشريه: 34 \_ شماره دوم، جلد 9**

**پديدآور: شعله ارسطوپور**

**مترجم:**

**چكيده**

**وب به عنوان بستر فعاليت موتورهاي جستجو، ساختاري نموداري دارد. اين ساختار حركت خزنده‌ها در موتورهاي جستجو را به روشهايي منطبق بر خود محدود مي‌سازد. مقاله حاضر، به بررسي تأثير ساختار وب بر چگونگي حركت خزنده‌ها و فعاليت نمايه سازها در موتورهاي جستجو مي پردازد. پس از بحثي مقدماتي در باب نمودارهاي جهت‌دار و فرايند كار خزنده، عمده‌ترين روشهاي حركت خزنده در سطح وب شامل حركت «عمق ـ شروع»، «توزيع ـ شروع» و «بهترين ـ شروع» مطرح شده و سپس واحد سازه‌يابي و چگونگي تشكيل درختهاي سازه‌يابي از قالب HTML مورد بررسي قرار خواهد گرفت.**

**كليدواژه‌ها: ساختار نموداري وب، خزنده‌ها، حركت عمق ـ شروع، حركت توزيع ـ شروع، حركت بهترين ـ شروع، نمايه‌سازي وب**

**مقدمه**

**وب، مجموعه‌اي عظيم از مدارك است كه هر يك براي پاسخگويي به يك نياز بالقوه ايجاد و در بستر اينترنت منتشر گرديده است. حجم عظيم صفحات و اطلاعات موجود در وب و لزوم وجود ابزارهايي جهت سازماندهيِ دست كم گوشه‌اي از اين مجموعه تاكنون به حد كفايت مورد بحث قرار گرفته است و اينكه موتورهاي جستجو با تمام كاستيهاي خود از عمده‌ترين ابزارهاي دسترسي به محتواي مدارك پيش گفته هستند، اصلي پذيرفته شده ميان بسياري از كاربران اينترنتي است. وب، ميلياردها صفحه الكترونيكي را از طريق شبكه اي از اتصالها با يكديگر مرتبط مي سازد و از آنجا كه ساختار آن هيچ محدوديتي بر چگونگي نشر صفحات، قالب انتشار، و يا چگونگي برقراري اتصالها و تعداد صفحات وب اعمال نمي‌نمايد، گسترش چشمگيري يافته است. اما اين مسئله هرگز بدين معنا نيست كه وب بدون ساختار است. بسياري از پژوهشها بيانگر وجود روابطي مشخص و منطقي در ساختار وب هستند.**

**(Barabasi & Albert, 1999; Kleinberg etal., 1999; Albert etal., 1999).**

**هر يك از اين تحقيقات به نوبه خود به نتايج متفاوتي رسيده اند. فصل مشترك آنها مسئله ناهماهنگي توزيع صفحات وب است به گونه‌اي كه آلبرت و ديگران[2] (1999) از وب به عنوان يك شبكه جهاني كوچك ياد مي‌كنند در حالي كه برادر و همكاران[3] (2000) توزيع وب را پاپيوني شكل توصيف مي‌كنند. بر اين اساس، وب از يك هسته كه بيشترين اتصالها به آن برقرار شده و دو بال كه با توجه به جهت اتصالها، توزيع متفاوتي از صفحات را در بر مي‌گيرد، تشكيل شده است. كليه پژوهش‌هاي انجام شده و تمام نتايج يك مسئله را اساس كار خود قرار داده اند و آن نموداري بودن ساختار وب[4] است. با توجه به دستاوردهاي پژوهشها، ساختار وب را به يقين مي‌‌توان به صورت يك نمودار عظيم ترسيم نمود؛ نموداري كه در برگيرندة گره‌ها و اتصالهاي متعدد است. در اين نمودار، صفحات وب همان گره‌ها و اتصالها همان لينكهاي برقرار شده ميان صفحات مختلف است**

**(Albert et al., 1999; Thelwall, 2002; Evans& Walker, 2004; Yu & Meng, 2004; Coathy, 2004) .**

**بنابراين وب نه تنها بدون ساختار نيست بلكه مي‌‌توان آن را به صورت يك ساختواره نموداري نيز ترسيم نمود. اين ساختواره خواسته يا ناخواسته بر چگونگي گردآوري صفحات و نمايه‌سازي آنها توسط موتورهاي جستجو تأثيرگذار بوده و در نهايت نتايج جستجو و ميزان ربط آنها را رقم مي‌زند.**

**نمودارهاي جهت‌دار[5]**

**نمودار در واقع مجموعه‌اي از گره‌ها و خطوط است كه آن را به صورت رياضيِ G (V, E) نشان مي‌دهند. هر E حتماً دو گره را به يكديگر متصل مي‌كند. نمودار جهت‌دار، نموداري است كه بتوان در آن جهت حركت از هر گره به گره ديگر را به راحتي تشخيص داد.**

**A**

**C**

**B**

**F**

**E**

**D**

**در يك نمودار جهت‌دار، چنانچه تعداد اتصالها را با i نمايش دهيم، براي هر گره كه به عنوان گره اصلي در نظر گرفته شود، رابطه [1,n) I i برقرار خواهد بود. در اين شرايط، هر يك از اتصالها تنها به طرف يك گره حركت خواهد كرد. از ديگر ويژگيهاي عمده يك نمودار جهت‌دار آن است كه هيچ كدام از گره‌ها منفرد و بدون اتصال نخواهد بود.**

**پيشتر بيان شد كه هر صفحه از وب داراي تعدادي اتصال است. اين اتصالها يا از صفحه هسته[6] به ساير صفحات و يا از ساير صفحات به صفحه موردنظر برقرار شده است. بنابراين، تمام ويژگيهاي پيش گفته، در باب صفحات وب و لينكهاي آنها نيز صدق مي‌كند. با توجه به جهت برقراري لينكها، مي‌‌توان وب را نيز به صورت يك نمودار جهت‌دار ترسيم نمود.**

**خزنده**

**تقريباً تمام خزنده‌ها در موتورهاي جستجو داراي چهار بخش گردآورنده[7]، واحد كنترل[8]، واحد سازه‌يابي[9] و واحد كار[10] هستند. گردآورنده تحت نظارت واحد كنترل، به صفحات هسته (گره‌هاي مختلف) رفته و مدارك را به واحد جداسازي لينكها مي فرستد. پس از جداسازي، لينكهاي مناسب به واحد كار ارسال شده و در فهرست دستور كار بعدي گردآورنده قرار مي‌گيرند. واحد سازه‌يابي در واقع از دو بخش جداسازي لينكها[11] و نمايه‌سازي تشكيل شده است.**

**آنچه نهايتاً در پايگاه ذخيره مي‌شود در واقع حاصل فرايند نمايه‌سازي است كه تحت قالب تعريف شده در الگوريتم موتور جستجو به صورت واژگان و عبارات مختل پايگاه نمايه**

**واحد كار**

**واحد جداسازي لينكها**

**نمايه‌ساز**

**واحد كنترل**

**گردآورنده**

**وب**

**ف در آمده است (Cothey, 2004). تصوير 2 ساختار يك خزنده را نشان مي دهد.**

**پيشتر اشاره شد كه ساختار وب همچون نموداري جهت‌دار است، لذا خزنده نيز براي حركت روي اين ساختواره، چاره‌اي جز تبعيت از ويژگيها و جهت‌هاي از پيش تعيين شده ندارد. از نظر جبري، حركت روي نمودارهاي جهت‌دار را مي‌‌توان به صورت ماتريسهاي حركتي[12] ساده كرد. در موتورهاي جستجو نيز، با توجه به شباهت ساختاري، استفاده از قواعد رياضي حاكم بر اين ماتريسها، تحليل و مقايسه را آسان مي‌نمايد. (تصوير3)**

**1**

**2**

**3**

**5**

**4**

**6**

**0 1 1 1 0 0**

**0 0 0 0 0 1**

**0 0 0 0 1 0**

**0 0 1 0 1 0**

**0 1 0 0 0 0**

**0 0 0 1 0 0**

**به طور كلي، سه روش براي حركت خزنده در شبكه لينكهاي وب وجود دارد. اين سه روش عبارتند از حركت عمق ـ شروع[13]، توزيع ـ شروع[14]، و بهترين ـ شروع[15].**

**حركت عمق ـ شروع**

**در اين حركت، واحد كنترل خزنده يك صفحه را به عنوان صفحه هسته براي گردآورنده مشخص مي‌سازد. پس از جداسازي لينكها، واحد كنترل يكي از لينكهاي خارجي صفحه را انتخاب و گره مقصد را به گردآورنده معرفي مي‌كند. اين فرايند تا زماني كه براي واحد كنترل تعريف شده باشد، ادامه پيدا مي‌كند. به عنوان نمونه، ترتيب حركت گردآورنده به صفحه‌هاي مختلف با الگوريتم عمق ـ شروع، مانند تصوير 4 خواهد بود.**

**0**

**7**

**1**

**3**

**2**

**5**

**6**

**4**

**Seed**

**از آنجا كه تقريباً تمام صفحه‌هاي وب لينكهايي به ساير صفحات برقرار مي‌كنند، چنانچه سطح عمق براي واحد كنترل تعريف شده نباشد، حركت به عمق آن‌ قدر ادامه خواهد يافت كه به مرور زمان، عملاً پايگاه نمايه موتور جستجو را از مطالب بي كيفيت خواهد انباشت. به همين دليل، در بيشتر موتورهاي جستجو، سطح عمق براي واحد كنترل تعريف مي‌شود. در تصوير4، چنانچه سطح عمق تعريف شده 2 باشد، ترتيب حركت گردآورنده 1، 2، 5، 6و7 بوده و عملاً صفحه‌هاي 3 و 4 مورد بررسي قرار نخواهند گرفت.**

**حركت توزيع ـ شروع**

**0**

**3**

**1**

**6**

**4**

**2**

**5**

**7**

**Seed**

**در اين حركت، واحد كنترل پس از تعيين صفحه هسته، كليه گره‌هاي هم عمق با يكديگر را تعيين و به ترتيب به گردآورنده معرفي مي‌كند. پس از رجوع به كليه صفحات مشخص شده در آن سطح، واحد كنترل سطح دوم را مورد بررسي قرار مي دهد. به عنوان نمونه، ترتيب حركت گردآورنده تحت نظارت واحد كنترل و با استفاده از الگوريتم توزيع ـ شروع در صفحه هاي مختلف مانند تصوير شماره 5 خواهد بود.**

**روش حركت توزيع ـ شروع مورد علاقه بسياري از طراحان برنامه‌هاي خزنده در موتورهاي جستجوست، زيرا طراحي و اجراي آن به صورت رايانه‌اي بسيار ساده‌تر از روش حركت عمق ـ شروع بوده و در صورت تعيين سياست دقيق، به لحاظ محدود بودن دامنه لينكهاي هر صفحه به عنوان صفحه هسته، حجم پايگاه موتور جستجو بيهوده افزايش نخواهد يافت(Chakrabarti et al., 2002).**

**حال، چنانچه صفحه هسته به يك مطلب خاص بپردازد، با توجه به آنكه گردآورنده تمام لينكهاي موجود در صفحه و يا صفحات بعد را دنبال نمي‌كند، حركت خزنده تأثير بسيار زيادي بر نمايه‌سازي و در نهايت بازيابي اطلاعات خواهد داشت.**

**در حركت عمق ـ شروع، با انتخاب هر لينك و رفتن به صفحه بعدي و ادامه اين كار، يك مطلب خاص (حوزه موضوعي مربوط به سطح عمق اول حركت) به صورت اختصاصي دنبال شده و از آنجا كه گرايش واحد كنترل نسبت به حركت عمقي گردآورنده بيشتر از حركت در سطح است، در نهايت صفحاتي كه براي نمايه‌ساز فرستاده مي شوند به احتمال، اغلب حول يك مطلب يا موضوع خواهند بود. در حالي كه در حركت توزيع ـ شروع گرايش واحد كنترل به حركت در سطح است و لذا گردآورنده ابتدا به گره‌هاي تعيين شده در سطح سركشي خواهد كرد. در چنين شرايطي صفحاتي براي واحد نمايه‌ساز ارسال مي شوند كه در كل ديدي عامتر دارند (مانند آنچه در صفحه هسته آمده است). اين مسئله، ناشي از آن است كه معمولاً لينكهايي كه از هر صفحه برقرار مي شوند، به بخشي از مطلب مطرح شده در صفحه هسته مربوط مي‌شوند.**

**بنابراين، با حركت از سطح به عمق، ديد نمايه‌ساز خواه نا خواه جزء نگر بوده و با حركت در سطح ديد نمايه‌سازي در حول يك مطلب با گستره‌اي وسيع‌تر و جامع‌تر متمركز خواهد بود (Herrman, 2003).**

**با توجه به مطالعات دهه 1990، استفاده از هر يك از اين روشها براي حركت روي ساختواره وب با توجه به حجم عظيم صفحات و مطالب هر يك، در نهايت تفاوت چنداني در بازيابي بهتر موتورهاي جستجوي مختلف در سطح وب نداشت.**

**تصميم‌گيري در باب انتخاب هر يك از اين دو روش و ترديدهاي موجود، روش سومي را پيش پاي طراحان و برنامه‌نويسان قرار داد و آن، حركت «بهترين ـ شروع» بود.**

**حركت بهترين ـ شروع**

**بهترين در حوزه حركت خزنده روي ساختواره وب، در واقع معاني متفاوتي دارد. الگوريتمهاي مختلفي براي حركت بهترين ـ شروع وجود دارند كه بر اساس فرمول محاسبه بهترين صفحه بعدي، اسامي متفاوت دارند. از اين بين مي‌‌توان به خزنده متمركز[16]، جستجوي كوسه‌اي[17]، عنكبوتهاي اطلاعاتي[18] و... اشاره كرد.**

**در ساده‌ترين حالت، از سياستهاي رتبه‌بندي همچون "رتبه‌بندي صفحات"[19] به عنوان معيار بهترين بودن استفاده مي‌شود. در اين روش واحد كنترل با توجه به رتبه هر صفحه ميان ساير صفحات، گردآورنده را به صفحه بعدي مي فرستد.**

**حركت بهترين ـ شروع، بر اين اصل مبتني است كه پديدآورنده هر صفحه زماني از صفحه خود (A) به صفحه ديگري (B) لينك برقرار مي‌كند كه B از نظر پديدآور A ارزشمند باشد. در بحث رتبه‌بندي صفحات چنين عملي براي B يك امتياز مثبت محسوب مي‌شود. بنابراين، در ساختار خزنده‌ها الگوريتم رتبه‌بندي صفحات در واقع برنامه‌اي است كه اهميت نسبي هر صفحه را بر اساس لينكهاي برقرار شده به آن مشخص مي‌سازد. بر اين پايه در خزنده‌هايي كه براي حركت خود از روش بهترين ـ شروع استفاده مي‌كنند در واقع از سه اصل پيروي مي‌شود:**

**· صفحاتي كه لينكهاي بيشتري به آنها برقرار شده است، اهميت بيشتري دارند. تعداد بيشتر لينكها به نوعي نشانگر شهرت و يا محبوبيت صفحه مذكور در سطح وب است.**

**· چنانچه اين لينكها از صفحات معتبرتري برقرار شده باشند، اعتبار صفحه مورد مطالعه افزايش خواهد يافت.**

**· از طرفي، هرچه تعداد لينكهايي كه از صفحه مورد مطالعه به ساير صفحات برقرار مي‌شود بيشتر باشد، ارزش آن صفحات كمتر خواهد بود.**

**بر اين اساس، در اكثر خزنده‌ها چنانچه U صفحه اصلي، Fu نشانگر صفحاتي كه از U به آنها لينك برقرار شده[20] و Bu نشانگر صفحاتي باشد كه به U لينك برقرار كرده‌اند[21] رتبه صفحه از طريق فرمول حاضر و يا فرمولهايي با عناصر اصلي مشابه، قابل سنجش خواهد بود.**

**در اين فرمول، علامت جمع به معناي وجود رابطه مثبت ميان تعداد لينكهاي برقرار شده به U و رتبه U است. R(V) يا رتبه صفحات برقرار كننده لينك به U نيز چون در صورت كسر قرار گرفته اند، رابطه مثبت با رتبه U دارند، در حالي كه وجود |FU| در مخرج، نشانگر وجود رابطه معكوس رتبه U و تعداد لينكهايي است كه به صفحات ديگر بر قرار كرده است[22].**

**1**

**3**

**2**

**4**

**5**

**6**

**25**

**14**

**18**

**16**

**32**

**11**

**42**

**6**

**14**

**5**

**0 25.0**

**0 10.0 14.0**

**5.0 0 16.0**

**6.0 18.0 0**

**0**

**32.0 42.0 0 14.0**

**11.0 0**

**استفاده از نتيجه اين فرمول در گردآوري صفحات ارزشمندتر، مؤثر خواهد بود. با كمك اين روش و استفاده از رتبه‌بندي محتواي مدارك در واحد نمايه‌سازي به احتمال پاسخهاي بهتري به نياز كاربر موتور جستجو داده خواهد شد. با رتبه‌بندي صفحات، در واقع رتبه لينكهاي منتهي به آنها مشخص شده و خود به خود ترتيب حركت گردآورنده بر روي ساختواره نموداري عظيم وب روشن مي گردد. تصوير 6 نشانگر اين ترتيب بر اساس رتبه‌هاي مشخص شده در ماتريس حركتي خواهد بود.**

**اين مسئله در نظر كاملاً منطقي است، اما ناجورك و وينر[23] (2001) در عمل ثابت مي‌كنند كه با توجه به هزينه بالا و زمان بر بودن فرايند رتبه‌بندي صفحات، استفاده از روش توزيع - شروع به نسبت توجيه پذيرتر مي‌نمايد. آنها برمبناي دستاوردهاي خود بيان مي‌دارند كه با توجه به امكانات فعلي فناوري، روش رتبه‌بندي صفحات بسيار هزينه بر بوده، زمان زيادي را مي‌طلبد و از طرف ديگر با توجه به عدم ثبات رتبه صفحات در طول زمان بايگاني نگهداري رتبه‌هاي صفحات را به سرعت بايد روزآمدسازي نمود. اين در حالي است كه توجه به حركت عمق ـ شروع به عنوان يك گزينه مطرح، كمتر صورت مي‌پذيرد.**

**واحد سازه‌يابي**

**واحد نمايه ساز موتورهاي جستجو بايد صفحات حاوي اطلاعات را از گردآورنده دريافت كند و عبارات و واژگان آنها را استخراج و در پايگاه خود ذخيره سازي نمايند. بنابراين، چنانچه هر صفحه در مقام خود يك واحد كلي باشد، نمايه ساز آن را به اجزاي كوچكتر از قبيل واژه و يا عبارت تبديل كرده و در پايگاه خود ذخيره مي‌سازد. نرم‌افزاري كه توانايي انجام اين عمل را داشته باشد، «سازه ياب» نام دارد. در فرايند سازه‌يابي، اولين كار تشخيص زبان رشته نشانه‌هاي ورودي[24] است. پس از آن، بر اساس دستور آن زبان خاص، سازه‌ياب به تعيين ساختار تركيبي آن رشته مي‌پردازد[25].**

**با اين توصيف برنامه سازه‌ياب در ساختار يك موتور جستجو كار جداسازي و يكسان‌سازي آدرسهاي اينترنتي موجود در مدرك، نگهداري فهرست واژگان غير مجاز و تهيه درخت سازه‌يابي را انجام مي دهد. از آنجا كه سازه‌ياب براساس دستور زبان از قبل تعريف شده به هدف دستيابي به محتواي مشخصي عمل مي‌كند، تقسيم‌بندي واژگان استخراج شده و وزن دهي به آنها كار ساده‌‌اي خواهد بود(Fischer, 2005)**

**زبان HTML زبان غالب در سطح وب به شمار مي‌آيد، لذا كليه موتورهاي جستجو داراي نرم افزارهاي سازه‌يابي سازگار با HTML براي زبانهاي مختلف هستند. به واسطه اين نرم افزارها، برچسبهاي HTML و ارزش آنها به سرعت شناسايي مي شوند.**

**براي جداسازي و يكسان نمودن آدرسهاي موجود در يك صفحه، از سازه‌يابها به منظور شناسايي برچسبهاي مختلف و ارزش آنها استفاده مي‌شود. اين كار معمولاً به منظور كمك به واحد كنترل جهت هدايت گردآورنده انجام مي‌شود. اما كار معرفي آدرسها به گردآورنده، اغلب مشكل‌تر از اين است. گاه لازم است بسياري از آدرسهاي ذكر شده در صفحه يكدست و تصحيح شوند. به منظور يكسان سازي آدرسهاي اينترنتي دستورالعملهايي براي تبديل حروف بزرگ آدرسها به حروف كوچك، برداشتن انشعابهاي اضافي از دنباله آدرس، تصحيح و يا تكميل برخي از آدرسها و ...، براي سازه يابهاي مختلف تعريف مي‌شود(Pant, Sirinivan & Meczer, 2004) .**

**سازه‌يابهاي مختلف ممكن است سياهه واژگان غير مجاز متفاوتي داشته باشند و يا اصولاً فاقد ويژگي حذف واژگان بدون بار معنايي در طول فرايند نمايه‌سازي باشند. در سطوح بالاتر، برخي از سازه‌يابها با توجه به دستور زبان از پيش تعريف شده، توانايي تشخيص ريشه كلمات و ذخيره كليه واژگان هم ريشه را در يك محل دارند.**

**در نهايت، وظيفه هر سازه‌ياب ايجاد درخت سازه‌يابي[26] است. در اين مرحله، واحد سازه‌يابي آدرس و يا واژه موجود در صفحه را با كمك محتوا و محل برچسب ارزيابي كرده، درختواره‌اي از ساختار صفحه تشكيل مي‌دهد. نمونه‌اي از اين درخت و قالب HTML متناظر با آن، در تصوير 7 نشان داده شده است.**

**گره اول در اين نمودار نشانگر قالب مدرك، گره‌هاي مياني برچسبهاي مختلف مدرك و آخرين گره‌ها نماينده محتواي ميان دو برچسب ابتدايي و انتهايي است.**

**تشكيل اين درخت، كار وزن‌دهي به هر متن و واژه و عبارت استخراج شده از آن را ساده مي‌نمايد. رتبه متون در قسمتهاي مختلف صفحه با توجه به الگوريتم رتبه‌بندي خاص هر موتور متفاوت است. واژگان وزن‌دهي شده را مي‌‌توان به راحتي در قالب يك مقياس عددي ريخته و براي الگوريتم رتبه‌بندي موتور جستجو، امكان سنجش و مقايسه سؤال كاربر و واژگان موجود پايگاه را فراهم آورد.**

**جمع‌بندي**

**با وجود ماهيت متغير وب، باز هم ساختار وب بر نحوة سازماندهي آن تأثيرگذار خواهد بود. با توجه به اينكه هم اكنون موتورهاي جستجو از مهمترين سازمان‌دهندگان به شمار مي‌روند و از طرفي با در نظرگرفتن اينكه ساختار وب روشهاي متفاوت و نه متعدد گردآوري اطلاعات را به طراحان نرم افزارهاي خزنده ديكته مي‌كند، مي‌‌توان بيان داشت كه ساختواره جهت‌دار وب بي‌تأثير بر بازيابي‌هاي مفيد و يا بي‌حاصل تأثير خواهد داشت. آنچه در اين زمان مورد توجه بيشتر محققان حوزه قرار گرفته، چگونگي بهينه‌سازي استفاده از امكاناتي است كه وب در اختيار طراحان قرار مي دهد. تصميم‌گيري در باب انتخاب شيوه حركت خزنده ـ اعم از حركت به عمق يا حركت در سطح و يا انتخاب هر صفحه بسته به كيفيت آن ـ يكي از مباحث مورد توجه علاقه‌مندان به اين حوزه است. مطالعه در باب بازدهي هر روش در طول زمان و يا امكان‌سنجي استفاده از يك روش در حال حاضر از مطالعات مطرح در اين حوزه به شمار مي‌آيد.**

**اين در حالي است كه از زاويه‌اي ديگر، اعمال از پيش تعريف شده براي هر سازه‌ياب ـ چه كوتاه و مختصر (سازه‌ياب‌هاي ساده) و چه پيچيده (سازه‌يابهاي سطح بالا) ـ نيز به بهينه‌سازي استفاده از امكانات وب توجه مي‌كند. با تعريف جزئيات بيشتر، نمايه‌سازي دقيق‌تر شده و در نهايت بازيابي بهتري حاصل خواهد شد.**

**مدارك به واسطة شيوه حركت در سطح وب گردآوري شده اند و نمايه‌سازي روي ساختار ساختمند وب به اجرا در آمده است؛ ساختاري كه حتي زبان نگارش آن را مي‌‌توان در قالب نمودار ترسيم نمود. الگوريتمهاي مختلف رتبه‌بندي بر اساس اين ساختار و اجزاي آن كار خود را انجام مي‌دهند، پس ساختار وب به صورتي غير مستقيم اما با قدرتي بسيار بر آنچه بازيابي مي‌شود تأثير خواهد داشت.**

**منابع**

**- Albert, R., Jeoung, H.& Barabasi, A. (1999). "The Diameter of the Wold Wide Web". Nature . Vol. 401, P. 130 Available online: www10.org/cdrom/papers/208 [Accessed on Oct. 2005]**

**- Barabasi A.L. & Albert, R. (1999). "Emergence of Scaling in random Networks". Science. Vol. 286, P 509 - 512. Available online: http://www.nd.edu/~networks [Accessed on Oct. 2005]**

**- Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, s., Stata, R., (2000). "Graph Structure in the Web". Computer Networks. Vol. 33, P. 309 - 320. Available online: http://www9.org/w9cdrom/160/160.html [Accessed on Oct. 2005]**

**- Chakrabarti, s. , Joshi, M.M., Punera, K. & Pennock, D.M. (2002). "The Structure of Broad Topics On the Web". Proceedings of the 11th World Wide Web Conference (p.508 – 510). Honolulu, Hawaii, May 7- 11 . New York: ACM. Available online: http://http.cs.berkeley.edu/~soumen/doc/www2002t/p338-chakrabarti.pdf [Accessed on Oct. 2005]**

**- Cothey, Viv (2004). "Web Crawling reliability". Journal of the American Society for Information Science and Technology. 55(14). P. 1228 – 1238.**

**- Evans, Michael P. & Walker, Andrew (2004). "Using The Web Graph to Influence Application Behavior". Internet Research. 14(5). P. 372 – 378.**

**- Fischer, Hendrik (2005). Decisions To Go: An Intelligent Mobile Decision Support System[Dissertation]. Georgia: The University of Georgia. Available online: http://graduate.gradsch.uga. edu/etdarchive/summer2005/fischer-hendrik-200508-ms.pdf [Accessed on Oct. 2005]**

**- Herrmann, Frank (2003). Web search engines. Available online: http://graduate.gradsch.uga.edu/etdarchive/summer2005/ fischer-hendrik-200508-ms.pdf [Accessed on Oct. 2005]**

**- Kleinberg, J., Kumar, R., Raghava, P., Rajagopalan, S., & Tomkins, A. (1999). "The Web as a Graph: Measurements, Models, and Methods". Proceedings of the International Conference on Combinatorics and Computing , Tokyo , Japan, July 26 – 28. London: Springer, P. 1 - 17. Available online: http://www.tomkinshome.com/papers/archive/cocoon99.pdf [Accessed on Oct. 2005]**

**- Najork,M. & Wiener, J.L. (2001). "Breadth – First Crawling Yields High Quality Pages". Proceedings of the 10th World Wide Web Conference (p.114 - 118). Hongkong. May 1 - 5 . New York: ACM. Available online: http://www10.org/cdrom/papers/208/ [Accessed on Oct. 2005]**

**- Pant, G., Srinivasan, p. Menczer,F. (2004). "Crawling the Web". Web Dynamics. Springer. Availableonline: http://mia.ece.uic.edu/~papers/MediaBot/pdf00001.pdf [Accessed on Oct. 2005]**

**- Thelwal, Mike (2002). "Methodologies for Crawler Based Web Surveys". Internet Research. 12(2), P. 124 – 138. Available online: www.scms.rgu.ac.uk/staff/fh/CM1008/documents/lecture3.pdf [Accessed on Oct. 2005]**

**- Yu, Clement & Meng, Weiyi (2004). "Web Search Technology". The Internet Encyclopedia. Hoboken, NJ: Wiley. P 738 – 753**

**--------------------------------------------------------------------------------**

**1. دانشجوي كارشناسي ارشد كتابداري و اطلاع‌رساني دانشگاه فردوسي مشهد**

**1. Albert etal.**

**2. Broder etal.**

**3. Web's Graph Structure**

**1. Directed Graph**

**2. Seed Page**

**1. Fetcher**

**2. Controller**

**3. Parsing Unit**

**4. Workload Unit**

**5. Link Extracting**

**6. Traversal Matrix**

**1. Depth - First**

**2. Breadth - First**

**3. Best - First**

**1. Focused Crawler**

**2. Shark Search**

**3. Info Spiders**

**4. Page Rank**

**5. Forward Links**

**6. Back Links**

**1. اين مسئله تابع اين اصل است كه هرچه تعداد صفحات زيرمجموعه صفحه هسته (Child pages) بيشتر باشد، ارزش و رتبه صفحه هسته بين اين صفحات تقسيم مي شود (Yu & Meng, 2004).**

**1. Najork & Wiener**

**2. Input Symbols' String**

**3. پاكروان، اميرحسين (1376)، فرهنگ كامپيوتر يادواره "انگليسي- فارسي"، (تهران: يادواره اسدي؛ فرهنگستان يادواره).**

**1. Parse Tree**