**نام مقاله: نمايه سازي سلسله مراتبي مدارك ساخت يافته**

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

**شماره نشريه: 28 \_ شماره چهارم،جلد 7**

**پديدآور: محمد باقر دستغيب**

**چكيده**

**هرروز بر تعداد مدارك ساخت‌يافته (مانند مدارك قابل نشانه‌گذاري) در اينترنت اضافه مي‌شود. اين نوع مدارك ساخت‌يافته، علاوه بر محتواي مدرك، قالب معنايي مدرك را نيز ذخيره مي‌كنند؛ بنابراين مدرك به‌صورت يك درختواره ذخيره مي‌گردد.از طرفي با افزايش اطلاعات موجود در شبكه، تقاضا براي بازيابي اطلاعات، بيشتر و پيچيده‌تر شده است. امروزه كاربران پرسش‌هايي را مطرح مي‌كنند كه داراي ساختار درختي است. براي بازيابي اطلاعات مرتبط، روش‌هاي كلاسيك كه صرفاً از عملگرهاي منطقي براي تطبيق پرسش با مدارك استفاده مي‌كنند، نمي‌توانند چنين پرسش‌هايي را با دقت مناسب بازيابي نمايند. هدف از اين مقاله بررسي نمايه‌سازي سلسله‌مراتبي و تطبيق سلسله‌مراتبي مدارك است.**

**كليدواژه‌ها: نمايه‌سازي سلسله‌مراتبي، تطبيق سلسله‌مراتبي، بازيابي اطلاعات**

**مقدمه**

نظام‌هاي سنتي بازيابي اطلاعات، با مدرك[[1]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn1" \o ") به عنوان كوچكترين واحد اطلاعاتي برخورد مي‌كنند، در صورتي كه غالب كاربران نياز به جستجوي با دقت در اجزاي مدرك دارند. به عنوان مثال درخواست جستجوي «هواپيماي جنگي كه در جنگ جهاني دوم، از آن استفاده شده است» را درنظر بگيريد. در اين مثال هر هواپيماي جنگي مورد قبول نيست، و هواپيمايي منظور است كه در جنگ جهاني دوم از آن استفاده شده. اگر به دو عنصر پرسش[[2]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn2" \o ") به صورت مجزا نگاه كنيم، هريك فقط بخشي از پرسش را مدل مي‌كنند. اين مشكل از ديرباز در برنامه‌هاي جستجوي اطلاعات مانند جستجوگرهاي وب وجود داشته است. به عنوان مثال ديگر، مي‌توان پرسش مدل بْرداري در بازيابي اطلاعات متني را درنظر گرفت. اين پرسش، بخشي از يك كتاب يا مقاله را درخواست مي‌كند، بنابراين كليدواژه‌هاي مورد استفاده بايد سلسله‌مراتبي در نظر گرفته شوند. به عبارت ديگر، بايد بازيابي اطلاعات به عنوان موضوع اصلي، و متني بودن و مدل بْرداري به عنوان موضوع فرعي درنظر گرفته شوند.

امروزه، با افزايش به‌كارگيري «ايكس‌ام‌ال»[[3]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn3" \o ")، تقاضا براي جستجوي اطلاعات ساخت‌يافته بيشتر شده است. «ايكس‌ام‌ال» استانداردهايي را براي مدلسازي ساختار مدرك، در اختيار نويسنده مي‌گذارد. در اين قالب، يك مدرك ساخت‌يافته به صورت مجموعه‌اي از فصل‌ها، بخش‌ها و پاراگراف‌ها ساخته مي‌شود. شكل 1 قالب مدرك ساخت‌يافته را نشان مي‌دهد.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  | **مدرك** |  |   |  |   |
|  |  |  |  |  |  |  |  |  |  |  |
|  | **فصل3** |   |  | **فصل2** |   |  |  | **فصل1** |   | **پاراگراف** |
|  |  |  |  |  |  |  |  |  |  |  |
| **بخش1** | **پاراگراف** |  |  |  | **بخش2** |   | **بخش1** |
|   |  |  |  |  |  |   |  |  |  |   |
| **پاراگراف** |  | **پاراگراف** |  | **پاراگراف** |  | **پاراگراف** |  | **پاراگراف** |  | **پاراگراف** |

**شكل 1. نمونه‌اي از ساختار معنايي مدرك در «ايكس‌ام‌ال»**

يك مدرك، غالباً داراي چندين موضوع اصلي است. بنابراين مفاهيم موجود در مدرك به‌صورت يك درخت مفاهيم بازنمايي مي‌گردند كه سطوح بالاي درخت، مفاهيم كلي را بيان مي‌كنند. در درخت مفاهيم، جستجو در زيرموضوعات امكان‌پذير است. سيستم بازيابي اطلاعات كارا، بايد اين قابليت را به‌صورت انتخابي در سطوح مختلف، در اختيار كاربران قرار دهد. اين مقاله دربارة بازيابي مدارك ساخت‌يافته، با روش نمايه‌سازي سلسله‌مراتبي و اطلاعات آماري مدرك مي‌باشد.

در روش‌هاي كلاسيك نمايه‌سازي، فقط پايين‌ترين سطح نمودار درختي نمايه‌ مي‌شود. اين روش نمايه‌سازي با ساختار سلسله‌مراتبي درخت متناسب نيست و رابطة معنايي ميان نهادهاي مدرك از بين مي‌رود (Hang & Wen, 2003). مشكل اصلي در نمايه‌سازي مدارك ساخت‌يافته، وزن‌دهي كلمات به صورتي است كه با پرسش‌هاي گوناگون قابل مقايسه باشد.

بنابر تحقيقات انجام شده، در يك موتور كاوش تنها 16 درصد از اطلاعات موجود در شبكة جهاني كه دسترسي عمومي دارد، نمايه و دسترس‌پذير مي‌شود (Geffet & Feitelson). روش‌هاي مرسوم نمايه‌سازي، موضوعات چندگانه را در يك نمايه تركيب مي‌كنند و بنابراين با توجه به حجم و تعداد مدارك موجود در شبكه، نمي‌توان يك جستجوي كارآ انجام داد. علت بروز چنين مشكلي، تركيب موضوع‌هاي گوناگون در يك نمايه است. در اين حالت چندين مدرك نامرتبط، با يك پرسش تطبيق داده مي‌شوند (Geffet & Feitelson).

راه‌حل مشكل بازيابي اطلاعات نامرتبط، نمايه‌سازي سلسله‌مراتبي است. در اين نمايه‌سازي علاوه بر كلمات سطح آخر نمودار درختي، فصول و بخش‌ها نيز نمايه‌ مي‌شوند. در پيمايش بالا به پايين درخت نمايه، مي‌توان با تطبيق كلمات كليدي[[4]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn4" \o ")، دقت جستجو را اضافه كرد. در سلسله مراتب درخت نمايه، هر سطح اطلاعات، سطوح مافوق را نيز به ارث مي‌برد.

در روش‌هاي مرسوم بازيابي اطلاعات كه از نماية مقلوب استفاده مي‌شود، فايل نمايه فقط حاوي اطلاعات دسترسي به مدارك است و در اين فايل، اطلاعاتي براي توصيف و نماية اجزاي مدارك وجود ندارد (Pottenger & Meling). روش ديگر، ايجاد نمايه‌هاي متمايز براي اجزاي گوناگون است. اين روش با توجه به تعدد فايل‌هاي نمايه، از نظر كارآيي سيستم و هزينة نگهداري، مقرون به صرفه نيست. در ادامه، دو روش نمايه‌سازي سلسله‌مراتبي بررسي مي‌شوند.

**نمايه‌سازي سلسله‌مراتبي**

در اين قسمت روش‌هاي نمايه‌سازي سلسله‌مراتبي بطور خلاصه بررسي مي‌شوند. مدارك مورد استفاده در اين مقاله، مدارك ساخت‌يافته مي‌باشند كه داراي ساختار مشخص هستند. در اين قسمت، روش‌ها و نتايج به‌دست آمده از نمايه‌سازي سلسله‌مراتبي مختصراً مرور مي‌شوند.

**1. مراحل ايجاد نماية سلسله‌مراتبي پويا**[[5]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn5" \o ")

نمايه‌سازي سلسله‌مراتبي پويا، روش جديدي در نماية اطلاعات انبوه است. روشي كه در اين نمايه‌سازي استفاده مي‌شود مبتني بر ايجاد زيرگروه، با توجه به معناي محلي مدارك مي‌باشد. در روش نماية سلسله‌مراتبي پويا، نمايه‌هاي سلسه‌مراتبي مانند فهرست‌هاي وب موجود در سايت «ياهو!»[[6]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn6" \o ") ايجاد مي‌شوند. ايجاد نمايه‌هاي سلسله‌مراتبي براي بالابردن دقت جستجو و مرور اطلاعات بسيار مفيد است، و اگر بتوان آن ‌را به صورت درونخطي[[7]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn7" \o ") توليد كرد، در كاوشگر وب نيز بدون دخالت انسان قابل استفاده است. در اين بخش مراحل ساخت نماية سلسله‌مراتبي مرور مي‌گردد (Sykes, 2001).

**1ـ1. استخراج مشخصه:** استخراج مشخصة مدرك در سه مرحله انجام مي‌شود: مرور[[8]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn8" \o ") مدرك، علامت‌گذاري، و توصيف مفاهيم.

در مرحلة مرور مدرك، «اچ‌تي‌ام‌ال»[[9]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn9" \o ") يا «ايكس‌ام‌ال»، ورودي سيستم است. ابتدا كلية مدارك با توجه به ساخت‌يافته بودن آن به مدرك «ايكس‌ام‌ال» تبديل مي‌گردند.

سپس فاز دوم با علامت‌گذاري مدرك «ايكس ام ال» شروع مي‌شود. در اين مرحله از يك «دستگاه وضعيت نهايي»[[10]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn10" \o ") براي علامت‌گذاري كلمات، با توجه به تلفظ و معناي آن در گنجينة لغت، استفاده مي‌شود. در اين مرحله، نوع كلمه بطور خودكار توسط دستگاه وضعيت نهايي، شناسايي مي‌گردد.

در مرحلة نهايي با توجه به محل و فراواني كلمه در متن، وزن كلمه[[11]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn11" \o ") محاسبه مي‌گردد. از اين مجموعه اطلاعات پردازش‌شده، به عنوان ورودي مرحلة بعد استفاده مي‌شود (Sykes, 2001).

**1ـ2. توليد ماتريس رخداد**[[12]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn12" \o ") كلمات: ماتريس رخداد، وزن كلمات را در مدرك نشان مي‌دهد. يك مدرك از بخش‌هاي گوناگوني ( كه با هم در ارتباط مي‌باشند) تشكيل مي‌شود. به عنوان مثال چكيده، عناوين، گزارش‌ها و نتايج، بخش‌هاي گوناگون مدرك هستند. رابطة ميان اجزاي اين ماتريس، دوطرفه و متقارن است ولي خاصيت تعدي ندارد. در اين مرحله وزن كلمات با توجه به تعداد تكرار و محل ظاهر شدن آن‌ها، محاسبه مي‌شود.

**1ـ3. ايجاد سلسله‌مراتب:** در اين مرحله براي هر ماتريس رخداد، و براي هر مفهوم در ماتريس رخداد، معيار شباهت[[13]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn13" \o ") با مفاهيم ديگر محاسبه مي‌شود. رابطة شباهت، يك رابطة يك به چند است، كه هر مفهوم را به چند مفهوم مرتبط متصل مي‌سازد. اين رابطه را مي‌توان به‌صورت يك درختواره نمايش داد. در اين مرحله، يك دسته‌بندي[[14]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn14" \o ") ميان مفاهيم مرتبط انجام مي‌شود. در حالت كلي، مجموعة دانش به‌صورت يك گراف[[15]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn15" \o ") جهت‌دار نامتقارن، كه گره[[16]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn16" \o ")هاي آن مفاهيم، و يال[[17]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn17" \o ")هاي آن وزن‌دار مي‌باشند، به دست مي‌آيد. براي دسته‌بندي اطلاعات، روش‌هاي زيادي هست، كه در اين قسمت روش آماري مورد توجه است. براي محاسبه وزن دسته‌ها از فرمول 1 استفاده مي‌شود.

**فرمول 1. محاسبة وزن دسته**

|  |  |  |
| --- | --- | --- |
|  \* WeightingFactor(Ck) | ∑ | ClusterWeight (Cj,Ck)= |
| ∑ |

فرمول 1 شباهت ميان مفهوم Cj به Ck را محاسبه مي‌كند. dij حاصلضرب فراواني مدرك[[18]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn18" \o ") و فراواني معكوس[[19]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn19" \o ") است. dijk حاصلضرب وزن مفاهيم Cj  و Ck، در مدرك i است. فاكتور وزن[[20]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn20" \o ") هر مفهوم توسط فرمول 2 مشخص مي‌شود. در اين فرمول N تعداد كل مدارك و dfj، بيشترين وزن مدرك j است.

**فرمول 2. وزن مفاهيم**

|  |  |
| --- | --- |
| Log N | WeightingFactor(Cj)= |
| Log N |

شكل 2 نمونه‌اي از گراف‌ دسته‌بندي موضوعي را نشان مي‌دهد. در اين گراف سه دسته موجود است (R3, R2, R1) كه در هر دسته، رابطة ميان مفاهيم عضو آن با يال‌هاي وزن‌دار مشخص مي‌گردد.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  | **1** | **5** | **3R** |
| **2R** | **6** |  | **2** |
|  |  |  |  |  |  |
|  | **1R** |  | **4** |  |  |
|  |  | **3** |  |  |

**شكل 2. نمايش گراف دسته‌بندي موضوعي**

**1ـ4. تطبيق سلسله‌مراتبي:** درنهايت اطلاعات دامنه[[21]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn21" \o ") به‌صورت سلسله‌مراتبي با دسته‌هاي موضوعي تطبيق داده مي‌شود، و ارتباطات لازم ميان دسته‌ها ايجاد مي‌گردد. ارتباط ميان دسته‌ها بر اساس مشابهت اطلاعات هر دسته با دستة ديگر ايجاد مي‌شود. هر دسته شامل مدارك شبيه به هم است. در هر دسته نمايندة دسته، اطلاعات آن دسته و گروه را به صورت عام بيان مي‌كند. در مرحلة تطبيق سلسله‌مراتبي، رابطة[[22]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn22" \o ") ميان دسته‌ها آشكار مي‌شود. بنابراين هر دسته، با دسته‌هاي والد[[23]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn23" \o ") و دسته‌هاي فرزندانش[[24]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn24" \o ") در ارتباط است. پس از اتمام اين مرحله، نماية سلسله‌مراتبي، اطلاعات كافي براي نمايش رابطة نسبي هر دسته، نسبت به كل اطلاعات مجموعه را دارد (Pottenger & Melling, 2001; Sykes, 2001). در سايت «ياهو» نمونه‌اي از دسته‌هاي سلسله‌مراتبي وجود دارد.

**2. نمايه سلسله‌مراتبي برپاية كتابشناسي بر روي وب**[[25]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn25" \o ")

هدف، به‌كارگيري درونخطي اصول كتابشناسي در محيط شبكة جهاني، براي نمايه‌سازي مدارك است. جستجو در روش سلسله‌مراتبي برپاية كتابشناسي بر روي وب، به صورت سلسله‌مراتبي انجام مي‌گيرد و نتايج براي كاربر فهرست مي‌شود.

نماية سلسله‌مراتبي از دو قسمت تشكيل شده: اولين قسمت، پيمايش برونخطي1[[26]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn26" \o ") مخزن است كه به صورت دوره‌هاي روزانه يا هفتگي تكرار مي‌گردد. قسمت دوم، مقايسه‌گر پرسش با بْردار كلمات كليدي است (Pottenger & Melling,2001; Sykes,2001).

اجزاي سيستم نمايه‌سازي سلسله‌مراتبي در اين قسمت به اختصار بررسي مي‌شود.

**2-1. تجزيه‌كننده2:** ابتدا كلية صفحات مدارك، تجزيه مي‌شوند. تجزيه به معناي ايجاد بْرداري از كل اطلاعات موجود در صفحات است. در اين مرحله بايد معناي كلمه كاملاً مشخص گردد. براي استفاده از كلمات، ابتدا ريشه‌يابي3[[27]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn27" \o ")انجام مي‌شود. يكي از روش‌هاي ريشه‌يابي، -n گرم4[[28]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn28" \o ")مي‌باشد. به عنوان مثال اگر 5 n= باشد، آنگاه زير رشته‌هاي كلمة Algorithm عبارت از algor، lgori، gorit، orith و rithm است. اين روش ريشه‌يابي، كارآيي قابل قبول دارد. جدول 1 نمونه‌اي از نتايج اين روش را نشان مي‌دهد (Pottenger & Melling, 2001; Sykes, 2001).

**جدول 1. نمايش نتايج روش 5- گرم در مقايسه با كلمة اصلي**

|  |  |  |
| --- | --- | --- |
| **ضريب توفيق5**[[29]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn29" \o ") | **تعداد دسته‌ها** | **موضوع** |
| **كل كلمه بدون ريشه‌يابي** | **5- گرم** |
| **53%** | **87%** | **10** | **آشپزي** |
| **47%** | **85%** | **16** | **سيستم عامل** |

در حالتي كه كلمه كمتر از پنج حرف داشته باشد، خود كلمه به عنوان ريشة كلمه درنظر گرفته مي‌شود. محاسبه بر روي ريشة كلمه موجب افزايش دقت مقايسة كليدواژه‌ها مي‌گردد. ضريب توفيق، نسبت جستجوهاي موفق به حالت ناموفق را بيان مي‌كند. جدول 1 نشان مي‌دهد كه اگر ريشه‌يابي انجام شود دقت و تعداد پرسش‌هاي موفق زياد مي‌شود.

**2-2. تركيب:**دسته‌هايي كه در درون يك دسته قرار دارند، زيرشاخه ناميده مي‌شوند. هر دسته ممكن است شامل چندين دستة كوچك‌تر به صورت سلسله‌مراتبي باشد. پس از تجزية تمام زيرشاخه‌ها و دسته‌هاي موجود، كلية بْردارهاي به‌دست آمده با هم تركيب مي‌شوند. بْردار به‌دست آمده، تركيبي از كليه بْردارهاي صفحات و مدارك موجود در مخزن است. بنابراين شمارنده‌هاي كلمات، مجموع تكرار كلمه در متون را نيز محاسبه مي‌كنند.

**2-3. يگانه‌سازي:**براي مقايسة بْردار پرسش با بردارهاي موجود در مخزن، بردارهاي گنجينة لغت بايد در فضايي متناسب با موضوع پرسش، مقايسه شوند. بنابراين يك بردار يگانه از مخزن ايجاد مي‌كنيم، كه داراي كلية كلمات به‌كاررفته در مدارك مخزن است. سپس بردار هر موضوع و دسته، نرماليزه1[[30]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn30" \o ") مي‌شود (Pottenger & Melling, 2001; Sykes, 2001).

**2-4. نرمال‌سازي شمارنده‌ها:**براي انتخاب كليدواژه‌هاي بامعنا، بايد تعداد تكرار كلمات در دسته‌هاي موضوعي در نظر گرفته شوند. ولي مدارك كوچكتر، داراي فراواني كمتري مي‌باشند، بنابراين بايد تعداد تكرار كليدواژه‌ها نسبت به طول مدرك، نرمال گردد.

**2-5. انتخاب كليدواژه‌ها:**كليدواژه‌ كلمه‌اي است كه خصوصيات مدرك، يعني مفاهيم و عناوين را توصيف مي‌كند. يكي از عامل‌هايي كه براي مشخص‌كردن كليدواژه‌ها به‌كار مي‌رود، فراواني كلمه در متن و ديگر متون مخزن است. اگر فراواني در كل مدارك پايين ودر يك مدرك، بالا باشد، آنگاه اين كلمه، كليدواژة مدرك است.

**2-6. بهينه‌سازي:**براي بهينه‌سازي مجموعة كلمات كليدي يك مدرك، كلمه‌هايي كه كاربري عمومي دارند حذف مي‌گردند. حذف لغات عام، در دو حوزة زبان‌شناسي و حوزة مختص به دامنة مدارك انجام مي‌شود. دامنة مدارك در هر حوزه به صورت تخصصي بايد تعريف شود. تعريف دامنة تخصصي، نيازمند ايجاد گنجينة لغات تخصصي است. لغات عام كه كاربري عمومي دارند و در گنجينة لغت حوزه وجود ندارند، حذف مي‌شوند.

**2-7. جستجوي درونخطي و ساماندهي نتايج:**پويشگر اطلاعات بايد پرسش را با مدارك مقايسه كند و مدارك مرتبط با پرسش را ارزشگذاري نمايد. براي ارزشگذاري مدارك، از گنجينة لغت به‌دست آمده در مراحل قبل استفاده مي‌شود.

براي مقايسه و ارزشگذاري مدارك، مقايسه از ريشة درخت آغاز مي‌شود و زيرشاخه‌هايي كه تطبيق بيشتري دارند، مقايسه و فهرست مي‌شوند. مقايسة درختي به‌صورت يك تابع بازگشتي1[[31]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn31" \o ") پياده‌سازي مي‌شود (Pottenger & Melling, 2001; Sykes, 2001).

براي جلوگيري از طولاني‌شدن فهرست جواب‌ها، كليدواژه‌هاي سردسته2[[32]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn32" \o ")ها را فهرست مي‌كنيم. هر دسته داراي يك سردسته به عنوان نمايندة دسته است. سردسته حاوي اطلاعات عمومي دسته است. معمولاً بْردار ميانگين مدارك موجود در دسته، به عنوان سردسته انتخاب مي‌شود. بنابراين يك كليد عمومي به عنوان نماينده براي كل دسته انتخاب مي‌شود. به عبارت ديگر ابتدا عناوين كلي به كاربر داده مي‌شوند و كاربر در صورت نياز مي‌تواند زيرشاخه‌ها را مرور كند. شكل 3 نمونه‌اي از جستجوي سلسله‌مراتبي را نشان مي‌دهد. در اين جستجوگر در سمت چپ فهرستي از دسته‌هاي موضوعي به صورت درختي نمايش داده شده و در سمت راست، فهرستي از موارد مرتبط فهرست شده است.

در اين مرحله، از قوانين كتابشناختي (از قبيل فراواني عمومي پايين و فراواني جزئي بالا) يا تطبيق معنايي معادل در گنجينة لغت براي تطبيق و نماية سلسله‌مراتبي استفاده مي‌شود.

**شكل 3. نمايش دسته‌اي و جستجوي سلسله‌مراتبي**

**خلاصه**

روش نمايه‌سازي و جستجوي سلسله‌مراتبي، نسبت به حالت كلاسيك داراي ميانگين دقت بيشتري است. در اين روش جستجو به صورت درختي انجام مي‌شود، بنابراين نتايج به‌دست آمده قابل انعطاف و گسترده است. از طرفي، دسته‌بندي اطلاعات به كاربران كمك مي‌كند اطلاعات مورد نظر را راحت‌تر به‌دست آورند.

در آزمايش‌هاي انجام شده، سيستم سلسله‌مراتبي دسته‌بندي اطلاعات، دقتي بين 90 تا 95 درصد به‌دست آورده است (Pottenger & Meling, 2001). بطور كلي دسته‌هاي اطلاعاتي به‌وجود آمده در هر نتيجة جستجو، علاوه بر سهولت كاربرد، كاربر را با موضوعات مرتبط آشنا مي‌كنند. همچنين كاربراني كه داراي تجربة كمتري هستند، مي‌توانند با مرور دسته‌ها موضوع دلخواه خود را استخراج كنند.

نماية سلسله‌مراتبي براي مدارك با قالب‌هاي جديد مانند مدارك «ايكس‌ام‌ال»، به صورت درونخطي تهيه مي‌شود. بنابراين با توجه به گسترش قاب‌بندي «ايكس ام ال» در سطح شبكه وب، نماية سلسله‌مراتبي روش موفقي در نمايه‌سازي و بازيابي اطلاعات مدارك با قالب‌بندي جديد مي‌باشد.

**منابع**

Hang C., Wen J. (2003). **Hierarchical Indexing and flexible element retrieval for structured documents**. Singapore, Department of computer science. Available Online:

[*http://research.microsoft.com/users/jrwen-files/publications/*
*ScalableRetrieval-ecir2003.pdf*](http://research.microsoft.com/users/jrwen-files/publications/%0BScalableRetrieval-ecir2003.pdf) [Accessed on Feb 2004]

Geffet M and Feitelson D. (2001). **Hierarchical Indexing and Document Matching in BoW**, Israel, School of computer science and engineering. Available Online:
<http://citeseer.ist.psu.edu/397796.html> [Accessed on Jan 2004]

Pottenger W., Kim, Yong-Bin, and Meling D. (2001) **Hierarichical Distributed Dynamic Indexing, HDDItm**. Available Online: <http://citeseer.ist.psu.edu/pottenger01hdditm.html>. [Accessed on Feb 2004].

Sykes J. (2001) **The Value of Indexing**,Dow Jones and Reuters Company Information Management Service. Available Online: <http://www.factiva.com/infopro/indexingwhitepaper.pdf> [Accessed on Feb 2004]

[[1]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftnref1" \o ") . Document

[[2]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftnref2" \o ") . Query

[[3]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftnref3" \o ") . Extensible Markup Language

[[4]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftnref4" \o ") . Keywords

[[5]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftnref5" \o ") . Dynamic

[[6]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftnref6" \o ") . http://www.yahoo.com

[[7]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftnref7" \o ") . online

[[8]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftnref8" \o ") . Parse

[[9]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftnref9" \o ") . Hyper Text Markup Language) HTML

[[10]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftnref10" \o ") . Finite State Machine

[[11]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftnref11" \o ") . Term Weight

[[12]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftnref12" \o ") . Occurrence (فراواني كلمات در بخش‌هاي گوناگون متن)

[[13]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftnref13" \o ") . Similarity

[[14]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftnref14" \o ") . Clustering

[[15]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftnref15" \o ") . Graph

[[16]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftnref16" \o ") . Node

[[17]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftnref17" \o ") . edge/ arc

[[18]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftnref18" \o ") . Document Frequency

[[19]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftnref19" \o ") . Inverse document frequency

[[20]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftnref20" \o ") . Weighting Factor

[[21]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftnref21" \o ") . Domain knowledge (اطلاعات جمع‌آوري شده از نماية مدارك كه حاوي ماتريس‌ بردارها نيز مي‌باشد)

[[22]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftnref22" \o ") . Relationship

[[23]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftnref23" \o ") . Parent

[[24]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftnref24" \o ") . Child

[[25]](http://128.168.0.10/lib/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftnref25" \o ") . Bibliography on the Web

1. Offline

3. Stemming

4. ايجاد زيررشته‌هايي به طول n

5. Hit Ratio

1. .normalize

1. Recursive Function

2. Headings