نمایه‌سازی سلسله‌مراتبی مدارک ساخت‌یافته

محمدباقر دستغیب[1]

چکیده

هرروز بر تعداد مدارک ساخت‌یافته (مانند مدارک قابل نشانه‌گذاری) در اینترنت اضافه می‌شود. این نوع مدارک ساخت‌یافته، علاوه بر محتوای مدرک، قالب معنایی مدرک را نیز ذخیره می‌کنند؛ بنابراین مدرک به‌صورت یک درختواره ذخیره می‌گردد.

از طرفی با افزایش اطلاعات موجود در شبکه، تقاضا برای بازیابی اطلاعات، بیشتر و پیچیده‌تر شده است. امروزه کاربران پرسش‌هایی را مطرح می‌کنند که دارای ساختار درختی است. برای بازیابی اطلاعات مرتبط، روش‌های کلاسیک که صرفاً از عملگرهای منطقی برای تطبیق پرسش با مدارک استفاده می‌کنند، نمی‌توانند چنین پرسش‌هایی را با دقت مناسب بازیابی نمایند. هدف از این مقاله بررسی نمایه‌سازی سلسله‌مراتبی و تطبیق سلسله‌مراتبی مدارک است.

کلیدواژه‌ها: نمایه‌سازی سلسله‌مراتبی، تطبیق سلسله‌مراتبی، بازیابی اطلاعات

مقدمه

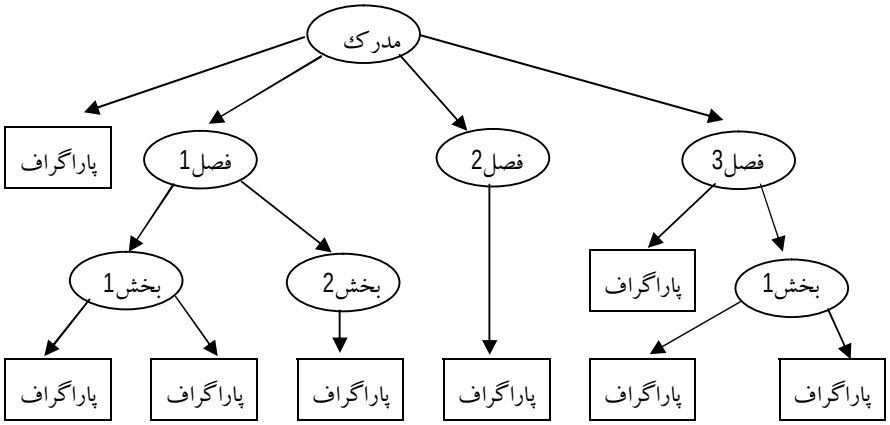
نظام‌های سنتی بازیابی اطلاعات، با مدرک[2] به عنوان کوچکترین واحد اطلاعاتی برخورد می‌کنند، در صورتی که غالب کاربران نیاز به جستجوی با دقت در اجزای مدرک دارند. به عنوان مثال درخواست جستجوی «هواپیمای جنگی که در جنگ جهانی دوم، از آن استفاده شده است» را درنظر بگیرید. در این مثال هر هواپیمای جنگی مورد قبول نیست، و هواپیمایی منظور است که در جنگ جهانی دوم از آن استفاده شده. اگر به دو عنصر پرسش[3] به صورت مجزا نگاه کنیم، هریک فقط بخشی از پرسش را مدل می‌کنند. این مشکل از دیرباز در برنامه‌های جستجوی اطلاعات مانند جستجوگرهای وب وجود داشته است. به عنوان مثال دیگر، می‌توان پرسش مدل بُرداری در بازیابی اطلاعات متنی را درنظر گرفت. این پرسش، بخشی از یک کتاب یا مقاله را درخواست می‌کند، بنابراین کلیدواژه‌های مورد استفاده باید سلسله‌مراتبی در نظر گرفته شوند. به عبارت دیگر، باید بازیابی اطلاعات به عنوان موضوع اصلی، و متنی بودن و مدل بُرداری به عنوان موضوع فرعی درنظر گرفته شوند.

امروزه، با افزایش به‌کارگیری «ایکس‌ام‌ال»[4]، تقاضا برای جستجوی اطلاعات ساخت‌یافته بیشتر شده است. «ایکس‌ام‌ال» استانداردهایی را برای مدلسازی ساختار مدرک، در اختیار نویسنده می‌گذارد. در این قالب، یک مدرک ساخت‌یافته به صورت مجموعه‌ای از فصل‌ها، بخش‌ها و پاراگراف‌ها ساخته می‌شود. شکل 1 قالب مدرک ساخت‌یافته را نشان می‌دهد.

---

[1]. کتابخانه منطقه‌ای علوم و تکنولوژی شیراز mail: dstghaib@srlst.com

[2]. Document
[3]. Query
[4]. Extensible Markup Language



شکل 1. نمونه‌ای از ساختار معنایی مدرک در «ایکس‌ام‌ال»

یک مدرک، غالباً دارای چندین موضوع اصلی است. بنابراین مفاهیم موجود در مدرک به‌صورت یک درخت مفاهیم بازنمایی می‌گردند که سطوح بالای درخت، مفاهیم کلی را بیان می‌کنند. در درخت مفاهیم، جستجو در زیرموضوعات امکان‌پذیر است. سیستم بازیابی اطلاعات کارا، باید این قابلیت را به‌صورت انتخابی در سطوح مختلف، در اختیار کاربران قرار دهد. این مقاله دربارهٔ بازیابی مدارک ساخت‌یافته، با روش نمایه‌سازی سلسله‌مراتبی و اطلاعات آماری مدرک می‌باشد.

در روش‌های کلاسیک نمایه‌سازی، فقط پایین‌ترین سطح نمودار درختی نمایه می‌شود. این روش نمایه‌سازی با ساختار سلسله‌مراتبی درخت متناسب نیست و رابطهٔ معنایی میان نهادهای مدرک از بین می‌رود (Hang & Wen, 2003). مشکل اصلی در نمایه‌سازی مدارک ساخت‌یافته، وزن‌دهی کلمات به صورتی است که با پرسش‌های گوناگون قابل مقایسه باشد.

بنابر تحقیقات انجام شده، در یک موتور کاوش تنها 16 درصد از اطلاعات موجود در شبکهٔ جهانی که دسترسی عمومی دارد، نمایه و دسترس‌پذیر می‌شود (Geffet & Feitelson). روش‌های مرسوم نمایه‌سازی، موضوعات چندگانه را در یک نمایه ترکیب می‌کنند و بنابراین با توجه به حجم و تعداد مدارک موجود در شبکه، نمی‌توان یک جستجوی کارآ انجام داد. علت بروز چنین مشکلی، ترکیب موضوع‌های گوناگون در یک نمایه است. در این حالت چندین مدرک نامرتبط، با یک پرسش تطبیق داده می‌شوند (Geffet & Feitelson).

راه‌حل مشکل بازیابی اطلاعات نامرتبط، نمایه‌سازی سلسله‌مراتبی است. در این نمایه‌سازی علاوه بر کلمات سطح آخر نمودار درختی، فصول و بخش‌ها نیز نمایه می‌شوند. در پیمایش بالا به پایین درخت نمایه، می‌توان با تطبیق کلمات کلیدی[5]، دقت جستجو را اضافه کرد. در سلسله مراتب درخت نمایه، هر سطح اطلاعات، سطوح مافوق را نیز به ارث می‌برد.

در روش‌های مرسوم بازیابی اطلاعات که از نمایهٔ مقلوب استفاده می‌شود، فایل نمایه فقط حاوی اطلاعات دسترسی به مدارک است و در این فایل، اطلاعاتی برای توصیف و نمایهٔ اجزای مدارک وجود ندارد (Pottenger & Meling). روش دیگر، ایجاد نمایه‌های متمایز برای اجزای گوناگون است. این روش با توجه به تعدد فایل‌های نمایه، از نظر کارآیی سیستم و هزینهٔ نگهداری، مقرون به صرفه نیست. در ادامه، دو روش نمایه‌سازی سلسله‌مراتبی بررسی می‌شوند.

نمایه‌سازی سلسله‌مراتبی

در این قسمت روش‌های نمایه‌سازی سلسله‌مراتبی بطور خلاصه بررسی می‌شوند. مدارک مورد استفاده

[5] . Keywords

در این مقاله، مدارک ساخت‌یافته می‌باشند که دارای ساختار مشخص هستند. در این قسمت، روش‌ها و نتایج به‌دست آمده از نمایه‌سازی سلسله‌مراتبی مختصراً مرور می‌شوند.

1. مراحل ایجاد نمایهٔ سلسله‌مراتبی پویا[6]

نمایه‌سازی سلسله‌مراتبی پویا، روش جدیدی در نمایهٔ اطلاعات انبوه است. روشی که در این نمایه‌سازی استفاده می‌شود مبتنی بر ایجاد زیرگروه، با توجه به معنای محلی مدارک می‌باشد. در روش نمایهٔ سلسله‌مراتبی پویا، نمایه‌های سلسه‌مراتبی مانند فهرست‌های وب موجود در سایت «یاهو!»[7] ایجاد می‌شوند. ایجاد نمایه‌های سلسله‌مراتبی برای بالابردن دقت جستجو و مرور اطلاعات بسیار مفید است، و اگر بتوان آن را به صورت درونخطی[8] تولید کرد، در کاوشگر وب نیز بدون دخالت انسان قابل استفاده است. در این بخش مراحل ساخت نمایهٔ سلسله‌مراتبی مرور می‌گردد (Sykes, 2001).

1ـ1. **استخراج مشخصه**: استخراج مشخصهٔ مدرک در سه مرحله انجام می‌شود: مرور[9] مدرک، علامت‌گذاری، و توصیف مفاهیم.

در مرحلهٔ مرور مدرک، «اچ‌تی‌ام‌ال»[10] یا «ایکس‌ام‌ال»، ورودی سیستم است. ابتدا کلیهٔ مدارک با توجه به ساخت‌یافته بودن آن به مدرک «ایکس‌ام‌ال» تبدیل می‌گردند.

سپس فاز دوم با علامت‌گذاری مدرک «ایکس ام ال» شروع می‌شود. در این مرحله از یک «دستگاه وضعیت نهایی»[11] برای علامت‌گذاری کلمات، با توجه به تلفظ و معنای آن در گنجینهٔ لغت، استفاده می‌شود. در این مرحله، نوع کلمه بطور خودکار توسط دستگاه وضعیت نهایی، شناسایی می‌گردد.

در مرحلهٔ نهایی با توجه به محل و فراوانی کلمه در متن، وزن کلمه[12] محاسبه می‌گردد. از این مجموعه اطلاعات پردازش‌شده، به عنوان ورودی مرحلهٔ بعد استفاده می‌شود (Sykes, 2001).

1ـ2. **تولید ماتریس رخداد**[13] کلمات: ماتریس رخداد، وزن کلمات را در مدرک نشان می‌دهد. یک مدرک از بخش‌های گوناگونی (که با هم در ارتباط می‌باشند) تشکیل می‌شود. به عنوان مثال چکیده، عناوین، گزارش‌ها و نتایج، بخش‌های گوناگون مدرک هستند. رابطهٔ میان اجزای این ماتریس، دوطرفه و متقارن است ولی خاصیت تعدی ندارد. در این مرحله وزن کلمات با توجه به تعداد تکرار و محل ظاهر شدن آن‌ها، محاسبه می‌شود.

1ـ3. **ایجاد سلسله‌مراتب**: در این مرحله برای هر ماتریس رخداد، و برای هر مفهوم در ماتریس رخداد، معیار شباهت[14] با مفاهیم دیگر محاسبه می‌شود. رابطهٔ شباهت، یک رابطهٔ یک به چند است، که هر مفهوم را به چند مفهوم مرتبط متصل می‌سازد. این رابطه را می‌توان به‌صورت یک درختواره نمایش داد. در این مرحله، یک دسته‌بندی[15] میان مفاهیم مرتبط انجام می‌شود. در حالت کلی، مجموعهٔ دانش به‌صورت یک گراف[16] جهت‌دار نامتقارن، که گره[17]های آن مفاهیم، و یال[18]های آن وزن‌دار می‌باشند، به دست می‌آید. برای دسته‌بندی اطلاعات، روش‌های زیادی هست، که در این قسمت روش آماری مورد توجه است. برای محاسبه

---

6. Dynamic
7. http://www.yahoo.com
8. online
9. Parse
10. Hyper Text Markup Language) HTML
11. Finite State Machine
12. Term Weight
13. Occurrence (فراوانی کلمات در بخش‌های گوناگون متن)
14. Similarity
15. Clustering
16. Graph
17. Node
18. edge/ arc

وزن دسته‌ها از فرمول 1 استفاده می‌شود.

فرمول 1. محاسبهٔ وزن دسته

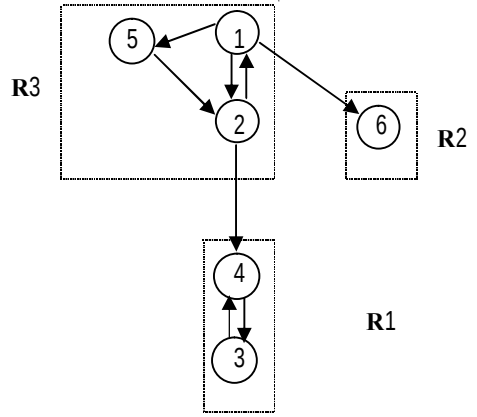
$$\text{ClusterWeight}(C_j,C_k) = \frac{\sum}{\sum} * \text{WeightingFactor}(C_k)$$

فرمول 1 شباهت میان مفهوم C_j به C_k را محاسبه می‌کند. d_{ij} حاصلضرب فراوانی مدرک[19] و فراوانی معکوس[20] است. d_{ijk} حاصلضرب وزن مفاهیم C_j و C_k، در مدرک i است. فاکتور وزن[21] هر مفهوم توسط فرمول 2 مشخص می‌شود. در این فرمول N تعداد کل مدارک و d_{fj}، بیشترین وزن مدرک j است.

فرمول 2. وزن مفاهیم

$$\text{WeightingFactor}(C_j) = \frac{\text{Log } \underline{N}}{\text{Log } N}$$

شکل 2 نمونه‌ای از گراف دسته‌بندی موضوعی را نشان می‌دهد. در این گراف سه دسته موجود است (R3, R2, R1) که در هر دسته، رابطهٔ میان مفاهیم عضو آن با یال‌های وزن‌دار مشخص می‌گردد.



شکل 2. نمایش گراف دسته‌بندی موضوعی

1ـ4. **تطبیق سلسله‌مراتبی**: درنهایت اطلاعات دامنه[22] به‌صورت سلسله‌مراتبی با دسته‌های موضوعی تطبیق داده می‌شود، و ارتباطات لازم میان دسته‌ها ایجاد می‌گردد. ارتباط میان دسته‌ها بر اساس مشابهت اطلاعات هر دسته با دستهٔ دیگر ایجاد می‌شود. هر دسته شامل مدارک شبیه به هم است. در هر دسته نمایندهٔ دسته، اطلاعات آن دسته و گروه را به صورت عام بیان می‌کند. در مرحلهٔ تطبیق سلسله‌مراتبی، رابطهٔ[23] میان دسته‌ها آشکار می‌شود. بنابراین هر دسته، با دسته‌های والد[24] و دسته‌های فرزندانش[25] در ارتباط است. پس از اتمام این مرحله، نمایهٔ سلسله‌مراتبی، اطلاعات کافی برای نمایش رابطهٔ نسبی هر دسته، نسبت به کل اطلاعات مجموعه را دارد (Pottenger & Melling, 2001; Sykes, 2001). در سایت «یاهو» نمونه‌ای از دسته‌های سلسله‌مراتبی وجود دارد.

---

19. Document Frequency
20. Inverse document frequency
21. Weighting Factor
22. Domain knowledge (اطلاعات جمع‌آوری شده از نمایهٔ مدارک که حاوی ماتریس بردارها نیز می‌باشد)
23. Relationship
24. Parent
25. Child

2. نمایه سلسله‌مراتبی برپایهٔ کتابشناسی بر روی وب[26]

هدف، به‌کارگیری درونخطی اصول کتابشناسی در محیط شبکهٔ جهانی، برای نمایه‌سازی مدارک است. جستجو در روش سلسله‌مراتبی برپایهٔ کتابشناسی بر روی وب، به صورت سلسله‌مراتبی انجام می‌گیرد و نتایج برای کاربر فهرست می‌شود.

نمایهٔ سلسله‌مراتبی از دو قسمت تشکیل شده: اولین قسمت، پیمایش برونخطی[1] مخزن است که به صورت دوره‌های روزانه یا هفتگی تکرار می‌گردد. قسمت دوم، مقایسه‌گر پرسش با بُردار کلمات کلیدی است (Pottenger & Melling,2001; Sykes,2001).

اجزای سیستم نمایه‌سازی سلسله‌مراتبی در این قسمت به اختصار بررسی می‌شود.

2-1. **تجزیه‌کننده**[2]: ابتدا کلیهٔ صفحات مدارک، تجزیه می‌شوند. تجزیه به معنای ایجاد بُرداری از کل اطلاعات موجود در صفحات است. در این مرحله باید معنای کلمه کاملاً مشخص گردد. برای استفاده از کلمات، ابتدا ریشه‌یابی[3] انجام می‌شود. یکی از روش‌های ریشه‌یابی، n- گرم[4] می‌باشد. به عنوان مثال اگر 5 =nباشد، آنگاه زیر رشته‌های کلمهٔ Algorithm عبارت از algor، lgori، gorit، orith و rithm است. این روش ریشه‌یابی، کارآیی قابل قبول دارد. جدول 1 نمونه‌ای از نتایج این روش را نشان می‌دهد (Pottenger & Melling, 2001; Sykes, 2001).

جدول 1. نمایش نتایج روش 5- گرم در مقایسه با کلمهٔ اصلی

موضوع	تعداد دسته‌ها	ضریب توفیق[5]	
		5- گرم	کل کلمه بدون ریشه‌یابی
آشپزی	10	%87	%53
سیستم عامل	16	%85	%47

در حالتی که کلمه کمتر از پنج حرف داشته باشد، خود کلمه به عنوان ریشهٔ کلمه درنظر گرفته می‌شود. محاسبه بر روی ریشهٔ کلمه موجب افزایش دقت مقایسهٔ کلیدواژه‌ها می‌گردد. ضریب توفیق، نسبت جستجوهای موفق به حالت ناموفق را بیان می‌کند. جدول 1 نشان می‌دهد که اگر ریشه‌یابی انجام شود دقت و تعداد پرسش‌های موفق زیاد می‌شود.

2-2. **ترکیب**: دسته‌هایی که در درون یک دسته قرار دارند، زیرشاخه نامیده می‌شوند. هر دسته ممکن است شامل چندین دستهٔ کوچک‌تر به صورت سلسله‌مراتبی باشد. پس از تجزیهٔ تمام زیرشاخه‌ها و دسته‌های موجود، کلیهٔ بُردارهای به‌دست آمده با هم ترکیب می‌شوند. بُردار به‌دست آمده، ترکیبی از کلیه بُردارهای صفحات و مدارک موجود در مخزن است. بنابراین شمارنده‌های کلمات، مجموع تکرار کلمه در متون را نیز محاسبه می‌کنند.

2-3. **یگانه‌سازی**: برای مقایسهٔ بُردار پرسش با بردارهای موجود در مخزن، بردارهای گنجینهٔ لغت باید در فضایی متناسب با موضوع پرسش، مقایسه شوند. بنابراین یک بردار یگانه از مخزن ایجاد می‌کنیم، که دارای کلیهٔ کلمات به‌کاررفته در مدارک مخزن است. سپس بردار هر موضوع و دسته، نرمالیزه[1] می‌شود

---

[26]. Bibliography on the Web
[1]. Offline
[3]. Stemming
[۴]. ایجاد زیررشته‌هایی به طول n
[5]. Hit Ratio
[1]. .normalize

.(Pottenger & Melling, 2001; Sykes, 2001)

2-4. **نرمال‌سازی شمارنده‌ها**: برای انتخاب کلیدواژه‌های بامعنا، باید تعداد تکرار کلمات در دسته‌های موضوعی در نظر گرفته شوند. ولی مدارک کوچکتر، دارای فراوانی کمتری می‌باشند، بنابراین باید تعداد تکرار کلیدواژه‌ها نسبت به طول مدرک، نرمال گردد.

2-5. **انتخاب کلیدواژه‌ها**: کلیدواژه کلمه‌ای است که خصوصیات مدرک، یعنی مفاهیم و عناوین را توصیف می‌کند. یکی از عامل‌هایی که برای مشخص‌کردن کلیدواژه‌ها به‌کار می‌رود، فراوانی کلمه در متن و دیگر متون مخزن است. اگر فراوانی در کل مدارک پایین ودر یک مدرک، بالا باشد، آنگاه این کلمه، کلیدواژهٔ مدرک است.

2-6. **بهینه‌سازی**: برای بهینه‌سازی مجموعهٔ کلمات کلیدی یک مدرک، کلمه‌هایی که کاربری عمومی دارند حذف می‌گردند. حذف لغات عام، در دو حوزهٔ زبان‌شناسی و حوزهٔ مختص به دامنهٔ مدارک انجام می‌شود. دامنهٔ مدارک در هر حوزه به صورت تخصصی باید تعریف شود. تعریف دامنهٔ تخصصی، نیازمند ایجاد گنجینهٔ لغات تخصصی است. لغات عام که کاربری عمومی دارند و در گنجینهٔ لغت حوزه وجود ندارند، حذف می‌شوند.

2-7. **جستجوی درونخطی و ساماندهی نتایج**: پویشگر اطلاعات باید پرسش را با مدارک مقایسه کند و مدارک مرتبط با پرسش را ارزشگذاری نماید. برای ارزشگذاری مدارک، از گنجینهٔ لغت به‌دست آمده در مراحل قبل استفاده می‌شود.

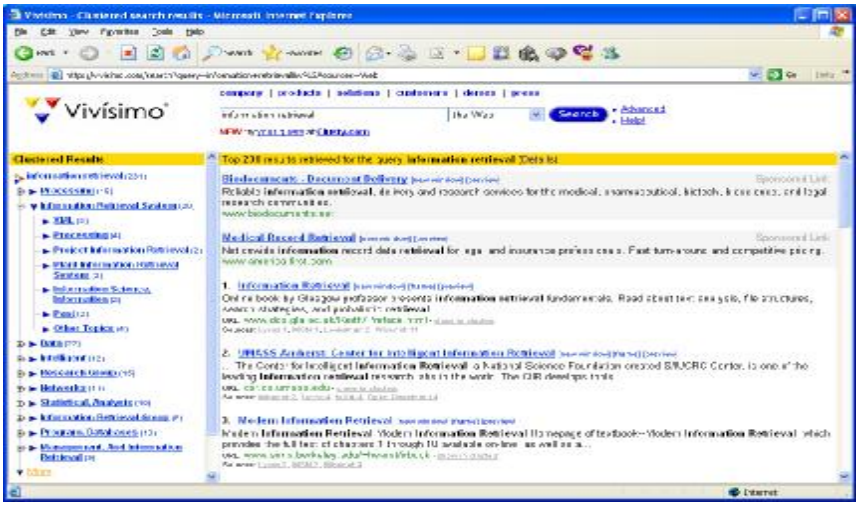
برای مقایسه و ارزشگذاری مدارک، مقایسه از ریشهٔ درخت آغاز می‌شود و زیرشاخه‌هایی که تطبیق بیشتری دارند، مقایسه و فهرست می‌شوند. مقایسهٔ درختی به‌صورت یک تابع بازگشتی[1] پیاده‌سازی می‌شود (Pottenger & Melling, 2001; Sykes, 2001).

برای جلوگیری از طولانی‌شدن فهرست جواب‌ها، کلیدواژه‌های سردسته[2] ها را فهرست می‌کنیم. هر دسته دارای یک سردسته به عنوان نمایندهٔ دسته است. سردسته حاوی اطلاعات عمومی دسته است. معمولاً بُردار میانگین مدارک موجود در دسته، به عنوان سردسته انتخاب می‌شود. بنابراین یک کلید عمومی به عنوان نماینده برای کل دسته انتخاب می‌شود. به عبارت دیگر ابتدا عناوین کلی به کاربر داده می‌شوند و کاربر در صورت نیاز می‌تواند زیرشاخه‌ها را مرور کند. شکل 3 نمونه‌ای از جستجوی سلسله‌مراتبی را نشان می‌دهد. در این جستجوگر در سمت چپ فهرستی از دسته‌های موضوعی به صورت درختی نمایش داده شده و در سمت راست، فهرستی از موارد مرتبط فهرست شده است.

در این مرحله، از قوانین کتابشناختی (از قبیل فراوانی عمومی پایین و فراوانی جزئی بالا) یا تطبیق معنایی معادل در گنجینهٔ لغت برای تطبیق و نمایهٔ سلسله‌مراتبی استفاده می‌شود.

---

1. Recursive Function
2. Headings



شکل 3. نمایش دسته‌ای و جستجوی سلسله‌مراتبی

خلاصه

روش نمایه‌سازی و جستجوی سلسله‌مراتبی، نسبت به حالت کلاسیک دارای میانگین دقت بیشتری است. در این روش جستجو به صورت درختی انجام می‌شود، بنابراین نتایج به‌دست آمده قابل انعطاف و گسترده است. از طرفی، دسته‌بندی اطلاعات به کاربران کمک می‌کند اطلاعات مورد نظر را راحت‌تر به‌دست آورند.

در آزمایش‌های انجام شده، سیستم سلسله‌مراتبی دسته‌بندی اطلاعات، دقتی بین 90 تا 95 درصد به‌دست آورده است (Pottenger & Meling, 2001). بطور کلی دسته‌های اطلاعاتی به‌وجود آمده در هر نتیجهٔ جستجو، علاوه بر سهولت کاربرد، کاربر را با موضوعات مرتبط آشنا می‌کنند. همچنین کاربرانی که دارای تجربهٔ کمتری هستند، می‌توانند با مرور دسته‌ها موضوع دلخواه خود را استخراج کنند.

نمایهٔ سلسله‌مراتبی برای مدارک با قالب‌های جدید مانند مدارک «ایکس‌ام‌ال»، به صورت درونخطی تهیه می‌شود. بنابراین با توجه به گسترش قاب‌بندی «ایکس ام ال» در سطح شبکه وب، نمایهٔ سلسله‌مراتبی روش موفقی در نمایه‌سازی و بازیابی اطلاعات مدارک با قالب‌بندی جدید می‌باشد.

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 [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, HDDI^{tm}**. 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]