نمایه سازی

تهیه کننده:
حیدر قدیم خانی

جلسه اول و دوم مفهوم داده , اطلاعات , دانش

داده data: واقعيات عيني از رويدادها

اطلاعات information: اضافه كردن تفسير به داده ها و ارتباط آنها به هم اطلاعات را شكل مي دهد اطلاعات بايد حاوي عناصر آگاهي دهنده و تغيير دهنده باشد

دانش : مفهومي عميق تر و وسيع تر از اطلاعات دارد . مخلوطي از اطلاعات موجود , تجربه ها , ارزشها و نگرشها.

Knowledge : دانش كمك مي كند كه اطلاعات و داده هاي ناخواسته حذف شده و فهم و درك حاصل شود.

چرخه اطلاعات : داده ها و اطلاعات در پايگاه هاي اطلاعاتي ذخيره مي شوند

مجله , مقاله , كتاب ....)چاپي ديجيتالي

پايگاه اطلاعاتي چيست ؟مجموعه اي از فايلهاي مربوط به هم در يك پايگاه اطلاعاتي كه در مورد يك موجوديت است مجموعه اي از فايلهاي مورد نياز و متناسب با كاركردي كه از پايگاه اطلاعاتي انتظار مي رود ايجاد مي شود . شامل:

فايل ورود اطلاعات: عمل ذخيره سازي داده و تاييد صحت آنها در اين فايل انجام مي شود

فايل اصلي: پس ا ز كسب اطمينان از صحت داده ها , براي نگهداري دائمي وارد فايل اصلي مي شوند . ركوردها پشت سر هم و به ترتيب شماره ركورد وارد مي شود

بايگاني واژه ها و اصطلاحات: مجموعه اي از واژه ها يا عبارات قابل جستجوي در نظر گرفته شده در آن قرار مي گيرد. معكوس شده فايل اصلي است و به همين دليل به ان فايل مقلوب مي گويند.هرواژه يكبار در نظم الفبايي قرار گرفته و در مقابل آن شماره ركوردهاي مربوط قرار دارد . نمايه انتهاي كتاب

فايلهاي جانبي :براي هر پايگاه اطلاعاتي نوع خاصي ساخته مي شود. مانند فايلهاي مستند سازي , فايل هاي تصاوير و

مهمترين عمليات در پايگاه اطلاعاتي

ورود داده ها

اصلاح داده ها

بازيابي داده ها

ضرورت استفاده از فايل مقلوب: جستجو در فايل اصلي مشكل و وقت گير است زيرا ترتيب شماره مدارك يا ركوردها مي باشد و براي جستجو بايد به ترتيب ورود

داده ها اقدام نمود اين جستجو جستجوي ترتيبي نام دارد. مراجعه به فايل مقلوب سرعت بيشتري را در دسترسي به اطلاعات فراهم مي نمايد . (استفاده از اصطلاحات و واژه ها و مراجعه به ركورد مورد نظر. فايل مقلوب با مهارتهاي ويژه نمايه سازي ساخته مي شود كه نقش مهمي در بازيابي اطلاعات صحيح و حذف اطلاعاتناخواسته دارد. در فايل مقلوب پاره هايي از اطلاعات كه قابل جستجو بر اساس نياز واقعي كاربران و قابليتهاي پردازش

در نرم افزار و سخت افزار قرار مي گيرد. پاره هاي اطلاعاتي قرار گرفته در يك فايل مقلوب مي تواند از يك عدد , يك كلمه تا يك عبارت باشد. يك ركورد كه يك بار در رايانه ذخيره مي شود , مي توان مدخلهاي زيادي براي جستجوي آن مشخص كرده و در نمايه مقلوب قرار داد. نمايه سازي از شيوه هاي بازيابي اطلاعات است.

كليد واژه: كلمه , عبارت يا بخشي از محتواي يك پيشينه كه داراي بار اطلاعاتي است و بعنوان شناسه يا مدخل در نمايه مقلوب ذخيره مي شود تا در موقع جستجو بعنوان كليد اصلي دستيابي به آن پيشينه باشد.

افزايش دادن نقاط دسترسي يا بازيابي

ايجاد نمايه ها ي بيشتر

فايلهاي بيشتر

هزينه بيشتر

افزايش محل ذخيره سازي

افزايش هزينه روز آمد سازي و نظم دهي فايلها

بنابراين بازدهي مطلوب ندارد.

راه حل چيست ؟

- تركيب پرسشها و استفاده از نمايه هاي "تركيب كننده فايلها"
- در پايگاه هاي اطلاعاتي از عملگرهاي منطقي جستجو استفاده مي شود:
- در پايگاه هاي اطلاعاتي از عملگرهاي منطقي جستجو براي بازيابي مطلوب استفاده مي شود.

And: تركيب (اشتراك) – خاصتر

Or : اجتماع – عامتر(بيشتر زماني بكار مي رود كه برونداد بسيار كم باشد , همچنين زماني كه جستجو وسيعتر مد نظر قرار گيرد)

Not: جستجو را محدود مي نمايد .

ل جبر بول		AN	ل	OR	ل	NO	ل
And =		T and T =		T or T =		Not T =	
Or =		T and F =		T or F =		Not F =	
Not =		F and T =		F or T =			
		F and F =		F or F =			

در يك پايگاه اطلاعاتي بايد استفاده از اين عملگرها را تعريف نمود .ترتيب و تقدم و تاخر عملگر ها را نيز مي توان تعريف كرد.

- ساير امكانات جستجو
- مجاورت كلماتHistory $ $ iran
- كلمات هم ريشه (محمود ؟ - محموديان , محمود زاده, محمود پور و(...
- عملگرهاي رابطه اي- + > <

نمايه معادل واژه لاتين INDEX به معناي نشان دادن و خاطر نشان كردن ،ريشه واژه لاتين آن INDIC به معني دلالت كردن است كلمه نمايه از مصدر نمودن يا نماييدن فارسي گرفته شده است.

- هر مجموعه اطلاعاتي نيازمند نمايه است.

نمايه سازي(INDEXING SERVICE): يكي از رايج ترين شيوه هاي پذيرفته در سازماندهي اطلاعات در سطح جهان است , به معناي عمل توصيف يا شناسايي محتواي موضوعي يك مدرك به قصد سهولت بازيابي در نمايه سازي سنتي به شكل ساده اساسي ترين كلمات متن يا محتواي اطلاعاتي را جمع آوري كرده , بعد الفبايي نموده وآدرس مكان را مي دهددر بين انواع نمايه ها , نمايه موضوعي به خاطر مفيد بود ن و كارسازتربودن در تنظيم اطلاعات مقبوليت دارد. در اين روش مي توان محتواي اطلاعاتي مثلا" يك كتاب خاص , يك موضوع خاص مانند علم كلام , مجموعه آثار يكانديشمند مانند شهيد مطهري را به گونه اي تنظيم كرد كه بازيابي به

سهولت و بر اساس خواسته كاربرد در سريعترين زمان انجام يافت .Representation هدف از نمايه سازي بازنمود اطلاعات منتشر شده در قالبي مناسب در پايگاه اطلاعاتي مي باشد اين پايگاه بازنمودي مي تواند در قالب چاپي مثل نمايه نامه مهندسي يا در قالب كارتي مثل برگه دان يك كتابخانه سنتي ارائه شود. يا در قالب الكترونيكي مانند اصطلاحنامه هاي آنلاين يا موتورهاي جستجو و راهنماهاي موضوعي

تعريف نمايه:

طبق استاندارد نمايه سازي بريتانيا: نمايه سياهه نظام يافته مدخلهايي است كه به منظور كمك به استفاده كنندگان در جايابي اطلاعات يك مدرك ساخته مي شود .انجمن نمايه سازان امريكا : نمايه عبارت است از يك رشته محلهاي دستيابي سازمان يافته كه استفاده كننده را از اطلاعات معلوم به اطلاعات اضافي ناشناخته ، راهنمايي مي كند .مجموعه شناسه هاي الفبايي يا نظام يافته اي كه كاربر رابه جايگاه اطلاعات درون مدرك هدايت مي كند .دانشنامه كتابداري واطلاع رساني نمايه چنين تعريف شده است: صورتي از موضوعها و واژه هاي مهم وديگر مطالب يك يا چند كتاب با ارجاع به صفحاتي كه اين مطالب در آنها آمده است.

نياز به نمايه

هدف اساسي حفظ و انتقال اطلاعات و دانش از نسلي به نسل ديگر بوده است. اين ساز و كار رويكردي مشترك دارد يعني: جمع آوري ، ذخيره اطلاعات ، بازيابي اطلاعات , استفاده به محض درخواستدهه 1940 عصر افزايش سريع اطلاعاتراهنماهاي كلي ديگر كافي نبود.ما از داشتن ميليون ها سند راضي نبوديم ، بلكه به تحليل موثر و سريع محتواي آن اسناد نياز داشتيم. به موازا ت تغيير نيازهاي اطلاعاتي، تغيير فناوري هم صورت گرفت.يكي از تحولات آرام و بسيار اساسي در نيمه دوم قرن نوزدهم: تغيير بازيابي اطلاعات از اعم به اخصيكي از راههاي سازماندهي اطلاعات نمايهسازي است

اهداف نمايه سازي

- نمايه داراي دو هدف كلي است زمان و تلاش براي يافتن اطلاعات را به حداقل ميزان كاهش دهد و موفقيت جستجوي استفاده كننده را به حداكثر برساند. (انتخاب واژه مناسب در زمان كمتر و با تلاش كمتر)
- نمايه ها و چكيده ها به ارزش سند مي افزايند.
- نمايه يك فرا داده است (داده اي در مورد داده ها)

- فضاي نمايه داراي دو گزاره مربوط به سند نمايه شده است . شمول و عدم شمول يعني چه چيزي در نمايه وجود دارد و چه چيزي در نمايه وجود ندارد.

اهداف نمايه سازي

به طوركلي:

- برقراري ارتباط ميان مفاهيم
- تنظيم شناسه ها به ترتيبي نظام مند و موثر
- شناسايي سريع مدارك در يك مجموعه
- سازماندهي اطلاعات به قصد بازيابي سريع و آسان مدارك

تاريخچه نمايه سازي

تاريخ نمايه سازي و چكيده نويسي با تاريخ نوشتن و مجموعه پيشينه هاي اطلاعات ارتباط نزديكي دارد . نمايه هاي اوليه به اسامي اشخاص و يا ترتيب آمدن كلمات در متن محدود بود. از جمله اولين نمايه ها : كشف اللغات) (concordancesكهن ترين پيشينه اي كه از نمايه سازماندار در تاريخ سراغ داريم، سياهه لوح ههاي گلين كتابخانه هاي اكد و سوم ر است. امانخستين گامهاي جدي در راه تدوين نمايه در خصوص كتب مقدس ديني و در سده چهاردهم ميلادي در چند صومعه صورت گرفت. در مشرق زمين، خاصه در قلمرو تمدن اسلامي كارهايي در زمينه هاي رده بندي، فهرست نويسي و نمايه سازي انجام گرديد. فهرست ابن نديم يكي از اين نمونه هاست . استفاده از نمايه براي طومار پاپيروس مستلزم باز و بسته كردن مكرر طومار بود. و نمايه آن به شكل صفحه كوچكي به طومارمتصل مي شده است و با به وجود آمدن كدكس كه شكل جديد كتاب بود گامي اساسي در نمايه شد. و نمايه كتاب براي اولين بار عملي شد .اختراع چاپ، رشد سريع چاپ و توليد كتاب، نياز به نمايه هاي كتاب را افزايش داد .افزودن آستر بدرقه به كتاب ها باعث خواننده گان مطالب و موضوعات مهم را ياد داشت كنند به عنوان مثال: وكلا از آن براي فهرست الفبايي قوانين مورد علاقه خود استفاده كردند و كاتبان براي نوشتن ارجاعات به كتب مقدس....

- در قرن 16 نمايه هاي با كيفيت براي كتاب ايجاد شد
- در قرن 17 نوع جديد ابزار اطلاعاتي(نشريات ادواري) به وجود آمد و با رشد سريع آن نمايه ها ضرورت يافتند.
- همزمان با رشد مجلات علمي مجلات چكيده نيز توسعه يافت
- قرن 19 روند حركت به سمت انتشارات تخصصي، مجلات چكيده رشد سريع خود را آغاز كردند
- انفجار اطلاعات از دهه 1940

در دهه 1850 دبليو اف. پول نمايه اي منتشر كرد كه مجلات بسياري را شامل مي شد و مفهوم جديد انتشار نمايه مشترك براي شماره هاي متعدد نشريات گوناگون را آغاز كرد. كه اهميت فراواني در رشد نمايه سازي داشته است.

پس از جنگ جهاني دوم، دامنه تحقيقات و پژوهشها روز به روز گسترد ه تر گرديد و اين افزايش تدريجي كنترل اطلاعاتي با استفاده از ابزارها و شيوه هاي قبلي را غيرممكن م يساخت. در سال 1961 اولين كتاب تئوريك در زمينه نظام هاي بازيابي رشد سريع انتشارات و بررسي روشهاي نمايه سازي بود. به « ويكري » اطلاعاتي ارائه شد. انگيزه اصلي تأليف اين كتاب از جانب تدريج بر حجم اين تلاش ها افزوده شد تا اين كه در دهه 1970 تحقيقات گسترده اي به روي بانكهاي اطلاعاتي و سيستم هاي خودكار پيوسته صورت گرفj

- . در طول دهه 1980 استفاده از نظام هاي بازيابي اطلاعات در دو جنبه گسترش پيدا كرد :
- جنبه اول گسترش نظامهاي تمام متن بود (تا قبل از اين دهه در بازيابي اطلاعات فقط از نمايه ها و چكيده ها استفاده مي شد)
- جنبه دوم گسترش نظامهاي پيوسته مورد استفاده غيرمتخصصان بود. در دهه 1990 حجم منابع ذخيره اطلاعات رايانه اي افزايش يافت، تصاوير به همراه اطلاعات ارائه شده و راه نمايش اطلاعات به گونه اي متفاوت گرديد. تحول مهمي كه در اين دهه به وقوع پيوست ظهور اينترنت بود. تا قبل از اين تحول تعداد ناشران اطلاعاتي بسيار محدود بود ولي در حال حاضر افراد زيادي در سرتاسر جهان اطلاعات مربوط به خود را توليد و دسته بندي مي كنند و از طريق صفحات خانگي، آنها را در اختيار ديگران قرار مي دهند، امروزه حجم انبوه اطلاعات به خصوص اطلاعات موجود بر روي شبكه جهاني وب موجب ظهور جنبه هاي جديدي در بازيابي اطلاعات گرديده است.

نمايه سازان چه كاري انجام مي دهند؟

- تحليل سند و تخصيص واژه هاي موضوعي كه فكر مي كند استفاده كننده براساس آن واژه ها جستجو خواهد كرد.

نمايه ساز به چه مواردي توجه مي كند؟

- توجه به خط مشي نمايه سازي
- نياز استفاده كننده
- نقش عامل انساني در نمايه سازي منابع بسيار مهم است.
- متن را جامع و كامل مطالعه تحليل كند
- تعيين موضوع و رشته
- تعين بار اطلاعات و ارزش نمايه سازي آن
- چه ميزان امكان عميق شدن در موضوع وجود دارد
- اصطلاحات را انتخاب كند (گاهي فهم موضوع آسان ولي انتخاب اصطلاح اعم يا اخص يا مترادف مشكل است)
- آشنايي با فن نمايه سازي و اصطلاح نامه نويسي

ويژگي نمايه سازان

- نظم و انضباط و علاقه به جزئيات
- حافظه خوب يك سرمايه است
- سريع خواندن يك ويژگي مثبت است
- نمايه سازان سانسور كننده نيز هستند
- ذهن كنجكاو و خلاق
- نمايه ساز خوب خود را به جاي استفاده كننده از اطلاعات قرار مي دهد
- با وجود تحولات بسيار در حوزه كامپيوتر و نرم افزارها هنوز هم نمايه سازي نيازمند تحليل موضوعي است كه فرآيندهاي ذهني و

انسان مدار است.

مراحل نمايه سازي

با تعارف گفته شده فرآيند نمايه سازي دو مرحله را شامل مي شود:

1 - تحليل مفهومي : يا تشخيص و درك جوهره

2- ترجمه : ارائه جوهره به صورتي كه قابليت پيش بيني داشته و حفظ امانت شود.

1- تحليل مفهومي :

- تصميم گيري درباره محتواست
- نمايه سازي براي برآوردن نياز مخاطبان و بهره گيران از نظام به كار مي رود بنابراين يك نمايه كار آمد نه تنها درباره محتواي مدرك باشد بلكه بايد مشخص شود موضوع يا محتواي آن براي چه گروهي از بهره گيران تهيه مي شود.
- به عبارت ديگر براي بهره گيران مي توان با اهداف مختلف از يك مدرك واحد , نمايه سازي هاي متفاوت انجام داد.

نمايه ساز درباره يك مدرك بايد سوالات زير را مطرح كند:

- مدرك درباره چه چيزي است ؟
- چرا اين مدرك به مجموعه اضافه شده است ؟
- بهره گيران به چه جنبه اي از مدرك علاقمندند؟

مثال : گزارش از يك ماموريت فضايي (اصطلاحاتي كه اين گزارش در هر شركت زير آنها نمايه سازي مي شود)

- شركت لاستيك سازي
- لباسهاي فضايي
- جديد
- مصنوعي
- شركت صنايع فلزي
- ابزارهاي فضايي
- جوشكاري
- فنون

- لاستيك فلزات

تركيبات

2 – ترجمه

- گام دوم در نمايه سازي است
- تبديل تحليل مفهومي يك مدرك به مجموعه اي از اصطلاحات نمايه اي
- در نمايه سازي هر نوع زباني كه براي رسيدن به قابليت پيش بيني و حفظ امانت به كار برده شود , زبان نمايه گفته مي شود.
- هرچه ميزان حفظ امانت و قابليت پيش بيني در يك زبان نمايه بالا باشد ,نظم افزايش مي يابد.
- در نمايه سازي بايد بين اصطلاحات با معاني مختلف تفاوت قائل شد و آن را روشن نمود . (حفظ امانت)

سم گياهي (گياهان سمي يا سمي كه بر عليه گياهان به كار مي رود)

اگر اصطلاحات به كار برده شده از زبان طبيعي (خود مدرك يا ذهن نمايه ساز) مشتق شده باشد زبان نمايه , واژگان طبيعي گفته مي شود . در غير اين صورت از يك اصطلاحنامه استفاده مي شود.

از اين جنبه مي توان دو نوع نمايه سازي متفاوت را در نظر گرفت:

1- نمايه سازي استخراجي
2- نمايه سازي تخصيصي

1- نمايه سازي استخراجي :تعريف : واژگان و عباراتي را كه واقعا" در مدرك وجود دارد براي بيان محتواي موضوعي همان مدرك مورد استفاده قرار داد

زبان هاي نماي هسازي – واژه گرفتي :مثال عنوان (تصوير كلي) : نظر سنجي افكار عمومي درباره نگرش آمريكائيان نسبت به خاورميانه

چكيده (جزئيات بيشتر از عنوان) : يك بررسي تلفني كه در سال 1985 انجام شد مسائل زير را مورد بررسي قرار داده است:

كمك ايالات متحده آمريكا به اسرائيل و مصر ؛ آيا ايالات متحده آمريكا بايد با اسرائيل , مليتهاي عرب يا ساير كشورها طرف شود ؛ آيا سازمان آزادي بخش فلسطين بايد در يك كنفرانس صلح شركت كند ؛ و آيا ايجاد يك كشور مستقل فلسطيني پيش نياز صلح است.

اصطلاحات نمايه سازي استخراجي : افكار عمومي , تحقيقات تلفني , ايالات متحده آمريكا , نگرش , خاورميانه , اسرائيل , مصر ،كمك , صلح

2- نمايه سازي تخصيصي تعريف : اختصاص اصطلاحات به يك مدرك از منبعي غير از خود آن مدرك مثلا“ از ذهن نمايه ساز يا از يك اصطلاحنامه مدون يا واژگان كنترل شده در حوزه موضوعي.

اصطلاحات نمايه سازي تخصيصي سند قبل:

- كمك خارجي
- ارتباطات خارجي آمريكا
- زبان هاي نمايه سازي – واژه گذاري

سوال

به دليل مشكلات و سختي انواع نمايه , چرا بدون دخالت انسان , متون اصلي را ذخيره نكنيم و به اين ترتيب همه لغات متن را

مورد جستجو قرار ندهيم ؟(ذخيره سازي متن كامل)

دلايل طرفداران ذخيره سازي متن كامل

- هيچ چيز از متن حذف نمي شود
- هيچ چيز به آن اضافه نمي شود
- هر لغت يك توصيفگر است
- لازم نيست استفاده كننده قواعد پيچيده زبان مصنوعي را بياموزد

مخالفان

- اين تفكر ساده انديشانه است
- در بازيابي مشكلاتي ايجاد مي شود
- باعث افت اطلاعات مي شود
- ممكن است لغت بكاربرده شده در متن دقيقا" به ذهن جستجو گر خطور نكند . مثلا" اسيد سولفوريك به جاي جوهر نمك
- ممكن است بيش از يك بيان لغوي براي مفهوم در متن نباشد . مثلا" پرخاشگري نوجوانان , خشونت در نوجوانان , بزهكاري

نوجوانان

- ممكن است در متن مفهوم به صورت غير لغوي و تفسيري بيان شده باشد
- در نظام ذخيره سازي متن كامل اطلاعات غير متني مثل جدولها , تصاوير و ... قابل دسترسي نيستند.
- در ذخيره سازي متن كامل تعداد واژه ها زياد و حجم ذخيره سازي افزايش مي يابد
- اتلاف وقت زيادي وجود دارد
- در نظام متن كامل مسئله ابهام در بيان حل نشده است . مثلا" اصطلاح كاج نقره اي (معلوم نيست كاجي از جنس نقره يا نام

يك نوع گياه است)

- در مراحل نمايه سازي , مرحله ترجمه وجود دارد در ذخيره سازي متن كامل اين مرحله حذف مي شود.

- در مرحله انتخاب جوهره متن هم صورت نمي گيرد.

تصاصات روش	خيره سازي متن كامل	ايه سازي كنترل شده
اسايي معنا و رفع ابهام		
ليت پيش بيني جوهره		-)
ليت پيش بيني ارائه جوهره و خاب توصيفگر		-)
كردن خلاها		-)
وي سازي تفسيري		)
ليت ارائه اطلاعات غير متني		-)
ترل مترادفات		
ص بودن كليد واژه ها	+)	+)
يج بودن مفاهيم		
كان نحو در زبان نمايه		

كان جمع آوري توصيفگرها		
سترسي به اصطلاحات جا افتاده در ايه سازي	+	)
سان و حجم لازم براي ذخيره سازي		

راهنماي جدول:

-نقطه ضعف

+نقطه قوت

(+)تا حدودي قوت

(-)تا حدودي ضعف

(-+)دارا بودن حالت بين قوت و ضعف

نمايه به دو دسته كلي نمايه توصيفي و نمايه موضوعي تقسيم بندي مي شود:

نمايه توصيفيDescriptive Index : پرداختن به ويژگيهاي اطلاعاتي غيرموضوعي يك مدرك كه معمولاً به اطلاعات كتابشناختي موسوم مي باشند رامي گويند. « توصيفي نمايه»

نمايه موضوعيSubject Index : ساخت و پرداخت زمينه هاي اطلاعاتي مدرك كه بار موضوعي دارد و دسترسي بر مدارك را از طريق موضوع امكا نپذير مي

خوانده مي شود. « نمايه موضوعي » سازد

زبان هاي نمايه سازي

در نظام هاي اطلاع رساني به زبان ساختگي و قراردادي اطلاق مي شود كه براي مقاصد نمايه سازي به ويژه قابليت بازيابي اطلاعات و مدارك به كار گرفته مي شود. به طور كلي زبان نمايه سازي، استانداردي را مهيا مي كند كه هم نمايه ساز و هم جستجوگر مي توانند از آن استفاده كنند .پس به طور ساده تر: مجموعه اي از روشهاي از پيش تعيين شده براي سازمان دهي، بازيابي و اشاعه اطلاعات

زبان هاي نمايه سازي :

به دو دسته كلي تقسيم مي شوند:

1 - نظام هاي اصطلاح تعيين شدهAssigned term (داراي ابزارهاي كنترل واژگان) نمايه سازي كنترل شده Controled language

2 - نظام هاي اصطلاح مشتق Derived Term توصيفگر ها از متن گرفته مي شود لذا آن را متن آزاد و طبيعي هم مي گويند(نمايه سازي آزاد و طبيعي)

Free language

مزاياي زبان طبيعي

زبان متخصصان هر رشته است و براي ارتباط با ساير متخصصان بسيار مهيا تر مي باشد

- ساختار طبيعي زبان حفظ مي شود
- رو زآمدي در قياس با زبان ساختگي
- ناهمگوني هاي اصطلاحات را تحت سياست واحد قرار مي دهد
- امر نمايه سازي براي نمايه ساز با سهولت و راحتي بيشتري صورت مي پذيرد زيرا نياز چنداني به تحليل موضوعي مدرك نيست
- جامعيت را افزايش مي دهد و از پراكندگي اطلاعات جلوگيري مي كند
- سرعت نمايه سازي افزايش مي يابد

معايب زبان طبيعي

1- از آنجا كه كليدواژه ها مطابق با واژگان مؤلف انتخاب شده كاربر مجبور است جهت بازيابي مدرك تمامي كليدواژه هاي احتمالي را حدس بزند
2- به دليل وجود مترادف ها و واژه هاي مشابه احتمال بازيابي مدارك نامرتبط افزايش مي يابد
3- معمولا وقت و زمان زيادي از كاربر براي جستجو و بازيابي مدرك مورد نظرش گرفته مي شود

مثال : زبان طبيعي

عنوان مقاله: اهميت و نقش كتابخانه هاي آموزشگاهي در آموزش

چكيده

اهميت مطالعه در دوران كودكي و نوجواني و قابليت گسترش آن در سنين بالا بر كسي پوشيده نيست. جامعه دانش آموزي براي به دست آوردن روحيه پژوهش و مطالعه فردي از نخستين سال دبستان تا آخرين سال دبستان به كتابخانه مجهز نياز دارد. در اين مقاله ابتدا كتابخانه هاي آموزشگاهي در مقاطع تحصيلي تعريف و سپس اهداف، وظايف، و تشكيلات آن بيان مي شود. در پايان براي بهبود وضعيت موجود كتابخانه هاي آموزشگاهي پيشنهادهايي ارائه شده است.

كليدواژه ها: كتابخانه هاي آموزشگاهي، مطالعه، پژوهش، آموزش

مثال : زبان كنتر ل شده

عنوان مقاله: اهميت و نقش كتابخانه هاي آموزشگاهي در آموزش

چكيده

اهميت مطالعه در دوران كودكي و نوجواني و قابليت گسترش آن در سنين بالا بر كسي پوشيده نيست. جامعه دانش آموزي براي به دست آوردن روحيه پژوهش و مطالعه فردي از نخستين سال دبستان تا آخرين سال دبستان به كتابخانه مجهز نياز دارد. در اين مقاله ابتدا كتابخانه هاي آموزشگاهي در مقاطع تحصيلي تعريف و سپس

اهداف، وظايف، و تشكيلات آن بيان مي شود. در پايان براي بهبود وضعيت موجود كتابخانه هاي آموزشگاهي پيشنهادهايي ارائه شده است.

كليدواژه ها: كتابخانه هاي آموزشگاهي، مطالعه، پژوهش، آموزش، ارتقاء كيفيت آموزش، پژوهش محوري

مزاياي زبان ساختگي

- بصورت يك بسته اماده به كامپيوتر داده مي شود و بين مدرك و زبان كنترل شده يك پيوند برقرار مي كند.
- به مراتب توانمند تر در دادن اطلاعات است هر چند مانعيت كمتري دارد.
- وقتي ايجاد سياست مي كنيم (مثل قوانين جدا نويسي ...) خود نوعي كنترل است و زبان ساختگي مي شود.
- زبان طبيعي نياز به ارجاع دارد بنابراين از حالت طبيعي خارج مي شود.

كاركرد نمايه سازي(Indexing Function)

- نمايه سازي عموماً سه كاركرد عمده دارد:
- محتواي اطلاعاتي مدارك را فشرده م يسازد.
- به عنوان واسطه براي تطبيق و يكسان سازي زبان مدرك و زبان كاوش به كار مي رود.
- به عنوان ابزاري كارا، بر شيوه تدوين راهبردي كاوش در جستجوي اطلاعات نظارت دارد.

گام هاي ايجاد يك نمايه

- تعيين خط مشي نمايه سازي
- خواندن سريع متن بدون يادداشت و علامت (عنوان , چكيده , خلاصه , نتيجه گيري , تصاوير , فهرست مندرجات كه باعث
- درك كليت موضوع مي شوند)
-
- مشخص كردن كليد واژه ها: بايد به علايق جامعه و نوع سازمان توجه كرد

- انتقال هركليد واژه (مدخل) وذكر محل بر هر برگه چا پي يا به شكل درون خطي , ثبت روي خود مدرك , ثبت روي نوار صوتي
- الفبايي كردن مدخل ها و ادغام مدخل هاي تكراري
- ايجاد پيوند مفهومي بين مدخل ها
- استفاده از اصطلاح نامه ها براي ايجاد روابط

ويژگي هاي خطي مشي: سياست نمايه سازي

- تعيين تعداد متوسط كليد واژه ها
- تعيين نوع نمايه
- زبان نمايه سازي(آزاد، كنترل شده، تركيبي)
- نوع بازيابي(راهنماي استفاده)
- روشن ساختن بيانگر (جمع يا مفرد،تكليف وضعيت كلمات خارجي)
- سياست مرجح يا نامرجح كردن بيانگرها
- ارجاعات و جاي نماها
- بيان انواع نمايه موجود (موضوع، نويسندگان،).
- استاندارد سازي (ابزارها)
- خودكار سازي نمايه(زبان برنامه نويسي، نرم افزار، چگونگي بازيابي)
- روزآمد سازي

سياست نمايه سازي

سياست سازمان متولي نمايه سازي كه تصميم مي گيرد نمايه جامع باشد يا جزنگر , تفصيلي يا اجماعي , كم عمق يا با عمق بيشتر ؟ نكته موثرديگر در تعيين سياست نمايه سازي , سطح علم و فهم شخص نمايه ساز است , بايد تخصصي علمي او سنخيت داشته و توان تحليل موضوع و استفاده از اصطلاح نامه را داشته باشد.

نمايه سازي پيش همارا: Pre Coordination Indexing

هرگاه بين دو يا چند جزء به طور تصنعي و ساختگي، پيوند برقرار كنيم (ايجاد نحو) به اين عمل و نوع آن، نمايه پيش هماراگويند .به عبارت ديگر در نمايه سازي پيش همارا، تركيب يا همارايي عناصر تشكيل دهنده موضوع مورد جستجو در هنگام نمايه سازي و به عبارتي پيش از بازيابي صورت مي گيرد:مانند : سرعنوان هاي موضوعي

•اصفهان- تاريخ- قبل از اسلام نمايه سازي پيش همارا بيشتر در نمايه هاي چاپي به كار مي رود.

- برگه دان كتابخانه معمولي ترين نظام بازيابي اطلاعات پيش هماراست
- تركيب اصطلاحات در هنگام جستجو ساده نيست
- اولين اصطلاح بايد مهمترين اصطلاح باشد
- نشان دادن روابط در اين نظام مشكل است

نمايه سازي پس همارا: post coordination Index ing

شيوه اي از نمايه سازي كه در آن نمايه ساز، سرشناسه ها را از مفاهيم بسيار ساده انتخاب مي كند و تعدادي شناسه زير هر يك اضافه مي نمايد و نيز تدابيري براي پيوستن آنها به يكديگر به دست مي دهد تا به وسيله آنها جوينده بتواند موضوع مدرك موردنظر خود را بيابد. تركيب يا همارايي عناصر تشكي لدهنده موضوع مورد جستجو در هنگام بازيابي انجام مي شود.

مانند : اصطلاح نامه ها

تاريخ * قبل از اسلام * اصفهان

- اصطلاحات در فايل مقلوب و اطلاعات در فايل اصلي ذخيره مي شوند
- يك نظام بازيابي اطلاعات كه به كاوشگر امكان مي دهد كه به هر طريق اصطلاحات را با هم تركيب كند نظام پس همارا گويند.
- بيشتر در نمايه هاي ماشيني كاربرد دارد

- چند بعدي بودن ارتباط بين اصطلاحات حفظ مي شود
- همه اصطلاحات اختصاص يافته به مدرك وزن يكساني دارند
- مي توان به مدرك مورد نظر با جستجو از طريق يك يا چند اصطلاح به مدرك دست يافت.
- از عملگرهاي منطقي در آن استفاده مي شود

مثلا" تاريخ 10

تاريخ * قبل از اسلام 6

تاريخ * قبل از اسلام * اصفهان 2

نمايه بر اساس حجم و سياست گذاري به دو دسته كلي تقسيم ميشود:

- نمايه تك متني
- نمايه مجموعه اي

نمايه تك متني :نمايه اي كه متن واحدي را در بر مي گيرد و ناظر بر يك متن واحد است. مانند نمايه كتاب Collection Indexنمايه مجموعه اي نمايه مجموعه يا نمايه نامه شامل محتويات مجموعه اي از آثار هم نوع و اغلب هم موضوع است. مانند نمايه نشريات ادواري، نمايه نامه نشريات شيمي و نمايه نامه روزنامه ها كه درواقع ناظر بر مجموعه اي از متون و منابع است، كه بنا به هدف و سياست ويژه اي گردآوري و كنترل واژگان درآن اجتناب ناپذير است. واحد اصلي نمايه مدخل است.از مجموعه مدخلها يك نمايه تشكيل ميشود.

يك مدخل متشكل از عناصرزير مي باشد:

- الف. شناسه

- ب. بیانگرModifier
- ج . جاینما یا ارجاع صفحهLocator/Refrence
- د.ارجاع دو سویهCross-refrence

شناسه :جزء محوري مدخل را شناسه مي گويند كه محل دسترسي به اطلاعات را درنظام تعريف شده ومعمولا الفبايي نشان ميدهد.

- شناسه عبارت است از واژه يا واژه ها، نشانه يا نشانه هايي كه به عنوان اصطلاح هاي نمايه اي از يك متن انتخاب ميشوند و به صورت الفبايي يا نظام انتخابي ديگر در نمايه مرتب ميشوند.

انواع شناسه از نظر شكل:

- شناسه تك واژه اي يا ساده (بسيط) :اين نوع شناسه تنها از يك واژه تشكيل ميشود و اين واژها مي تواند به صورت مفرد يا جمع باشد، مانند موسيقي، هنر، گلها،

قلب

- شناسه عبارتي :اين شناسه از تركيب دو يا چند واژه همراه با حرف ربط يا حرف اضافه ميان آنها تشكيل ميشود . در اين نوع شناسه ها به طور معمول دو واژه كه لازم است رابطه آنها به يكديگر نشان داده شود آورده ميشود مثل :آموزش و پرورش " و " تكنولوژي و تمدن" كه به طور معمول از يك خانواده يا زمينه موضوعي هستند و بيشتر اوقات باهم مطرح ميشوند. مانند: كار و كارگر شناسه هاي متشكل از صفت و موصوف و مضاف و مضاف اليه

اين شناسه ها به طور معمول از اسم و صفت و مضاف و مضاف اليه تشكيل ميشود مانند:

- روان شناسي رشد
- جامعه اطلاعاتي
- شناسه اسم با توضيحگر

گاهي براي تعيين حدود و معني و برطرف كردن ابهام و اخص كردن موضوع و گاه براي تميز قائل شدن ميان دو يا چند شناسه مشابه، واژه توضيحگردر مقابل آن و در مقابل پرانتزافزوده ميشوند، و در واقع واژه توضيحگرمعني از" نظر" مي دهد:

مانند:

حركت از نظر فلسفه

حركت از نظر فيزيك

حركت (فلسفه)

حركت (فيزيك)

شناسه مقلوب يا معكوس

اين شناسه درواقع همان شناسه هاي دو واژه اي صفت وموصوف يا مضاف و مضاف اليه هستند كه به منظور قرارگرفتن واژه

مهمتر در محل الفبايي خود به شكل مقلوب يا معكوس در مي آيند از حالت زبان طبيعي خارج ميشوند وميان دوجزء شناسه

نشا نه ويرگول مي نشيند مانند:

اطلاعات، آلودگي

اطلاعات، اقتصاد

كودك، روانشناسي

كاربرد اين نوع شناسه هاي موضوعي در زبان فارسي زياد رايج نيست.

انتخاب شكل شناسه ها به نوع نمايه وموضوع بستگي دارد.

بیانگرModifier/ Subheading :

در نمايه براي اخص كردن شناسه و نشان دادن جنبه هاي مختلف آن از بيانگر استفاده ميشود. به اين مفهوم كه شناسه تك واژه اي يا ساده (بسيط) را ميتوان را با افزودن واژه هاي مناسب خاص تر كرد . بيانگر اساسا به منظور اخص كردن شناسه و نيز براي حفظ جنبه هاي لاينفك يك شناسه موضوعي با خود شناسه و

محدود كردن دامنه معنايي آن بوجود مي آيد و ممكن است در يك مدخل شناسه، بيانگر نداشته باشد و مدخل تنها متشكل از دو عنصر شناسه و جاينما باشد، مانند:

130 ، ارتباطات، 125

انواع بيانگر

-بيانگر موضوعي يا موضوع زير موضوع

- روان شناسي
- 97 ،76 ، كودك 42
- اطلاعات
- 48 ،17 ، آلودگي 11

-بيانگر تاريخي يا زماني

- اوضاع اجتماعي
- 29 ،14 ،17 قرن 12

- 85 ،73 ،45 قرن 13
- 110 ،104 ،99 قرن 14

-بيانگر مكاني يا جغرافيايي

- آموزش و پرورش
- 55 ،18 ، ايران 15
- 104 ، 75 ، فرانسه 43

جاينما يا ارجاع صفحه:

عدد يا نشانه اي كه محل دستيابي به اطلاعات را مشخص مي سازد به آن جاينما يا ارجاع صفحه گفته ميشود(. (Page refrance

- جاينما ممكن است شماره صفحه، پاراگراف، بخش يا فصل يا آدرس اينترنتي باشد.
- جاينما تا آنجا كه امكان دارد بايد كامل و دقيق باشد.
- ميان شناسه و نخستين جاينما و بيانگر, ويرگول و يك فاصله خالي گذاشته ميشود.

ارجاع

ارجاع: يعني راهنمايي از يك شناسه به شناسه ديگر يا رجوع از بيانگرهاي يك شناسه به خود شناسه و يا شناسه هاي ديگر.

الف - - ارجاع ممكن است يك سويه باشد : شناسه نامرجح به شناسه مرجح

معناي آن اين است كه براي كليه اطلاعات مورد جستجو بايد به اصطلاح مرجح (گزيده نگاه كرد Perferred) . (terms يا

مانند:

بيوشيمي نگاه كنيد به زيست شناسي

ب- ارجاع ممكن است دو سويه باشد:معناي آن اين است كه براي اطلاعات مورد جستجو به جاي ديگر نيز رجوع شود.

مانند:اعتياد نيزنگاه كنيد به مواد مخدر==== كار و كارفرما " نيز نگاه كنيد به" كاروسرمايه مطالب مورد ارجاع به طور معمول اخص تر از مطالب اصلي است.

پرندگان نيز نگاه كنيد به قمري ها، كبوترها، ماكيان

ارجاع هاي دو سويه، جستجوگر را به اطلاعات مرتبط و وابسته راهنمايي مي كند.

ارجاع كورBlind entry :

ارجاع از يك شناسه به شناسه اي كه در نمايه وجود ندارد و يا ارجاع جاينما به صفحه هايي از متن كه فاقد اطلاعات و

مطالب ذكر شده باشد، ارجاع كور يا بي فرجام گفته ميشود، و در برخي از منابع با عنوان مدخل كور از آن ياد شده است.

ارجاع دورCircuit refrence

اشتباه ديگري كه در نمايه سازي رخ مي دهد ارجاع از واژه اي به واژه ديگر است كه به طور معمول به صورت دور تسلسل

انجام مي گيرد كه به آن ارجاع چرخه اي يا دور گفته ميشود.

در نمايه نام ها كه بحث مستند سازي نام اشهر مطرح ميشود گاهي ارجاع دور پيش مي آيد.

مانند:

بوعلي نگاه كنيد به شيخ الرئيس

شيخ الرئيس نگاه كنيد به بوعلي

نمايه كتاب از لحاظ شيوه چاپ و ساختار مدخل

الف – نمايه ساده

ب – نمايه درون بافتي

ج- نمايه برون بافتي

نمايه ساده :به نمايه اي گفته ميشود كه مدخل يعني مجموعه شناسه, بيانگر و جاينما در يك سطرمي آيند بيانگر به صورت طبيعي در ادامه شناسه ذكر مي شود و جاينماها نيز به دنبال آن مي آيند. در نمايه ساده در مواردي كه شناسه به تنهايي خود مدخل قرار گرفته است براي جلوگيري از زياد شدن جاينماها بايد يك توضيح گر در برابر آن بيايد و يا در مقدمه نمايه ذكر شود. د ر نظام نمايه ساده به طور معمول شبكه مفهومي، بين مفاهيم برقرار نمي شود.

نمايه درون بافتي:

به اين نوع نمايه نمايه پاراگرافي، پيوسته يا پيوسته سطري نيز گفته ميشود.زيرا ترتيب نوشتن بيانگر يك شناسه به صورت نحوه ساختاراصطلاحها درهمان بافت سطري و پيوسته است. دراين نوع نمايه به لحاظ ساختار نحوي به زبان طبيعي است.

ترتيب نوشتن بيانگر در نمايه درون بافتي , سطري است يعني تا جايي كه سطر اجازه مي دهد و از سطر دوم به بعد تو رفتگي دارد. توالي بيانگر ها تابع توالي حضور آنها در متن اصلي است. ترتيب پيدايش بيانگرها در متن است. بنابراين الفبايي نيستند . بيشتر مناسب متون علوم انساني , تاريخي و سرگذشتنامه اي مي باشد

مانند:

جاويدان، داريوش

؛20 ،15 ، محل تولد، 12 ؛ دوران كودكي، 13

35 ، ؛تحصيلات و ،33 -24 ، دوران نوجواني، 22

40- آثارعلمي، 35

نمايه برون بافتي:

به اين نوع نمايه، نمايه خطي يا جدا يا سطري يا گسسته نيز گفته ميشود. در اين نمايه بيانگر ها جدا از شناسه و به صورت الفبايي يا تو رفتگي مرتب مي شونديا به عبارت ديگر بيانگرها به صورت منفرد و هر كدام در سطري مستقل همراه با جاينما هاي مربوط است. دراين نمايه توالي بيانگرها تابع نظم الفبايي است.

مانند:

روان شناسي

1950 ،184 ، باليني، 179

232 ،218 ، رشد، 150

215 ،143 ، كودك، 129

تفاوت هاي نمايه برون بافتي و درون بافتي

نمايه درون بافتي صورت نحوي دارد.مثلا ميگوييم رفتار با كودكان ولي دربرون بافتي با را بكار نمي بريم و نمي دانيم آيا رفتار كودكان است يا رفتار با كودكان .

- نمايه درون بافتي بيشتر به حوزه علوم انساني تعلق دارد.

- نمايه درون بافتي كل نگر هستند.
- نمايه برون بافتي بيشتر به حوزه هاي علوم و فنون مرتبط است و بيشتر بن مايه اصطلاحنامه قرار مي گيرد.

نمايه از لحاظ پوشش و محتوي:

الف – نمايه نام ها

- نام پديد آور
- نمايه عناوين
- نمايه نام هاي جغرافيايي
- نمايه نام سازمان و موسسات

ب – نمايه موضوعي

نمايه ها بر حسب روش تنظيم

الفبايي : بيشتر نمايه هاي كنوني به ترتيب الفبايي است و شامل نمايه الفبايي اسامي اشخاص، سازمان ها و نيز موضوعات مي باشد

زماني : به ترتيب زمان از قديم به حال مي باشد و بيشتر براي نمايه هاي تاريخي به كار مي رود

رده اي يا موضوعي : كه براساس رده ها يا سرعنوان موضوعي نظام مند مرتب مي گردند. بيشتر در نمايه هاي علمي به كار ميرود كه البته مي تواند يك نمايه الفبايي رده اي باشد.

تكاملي: براي نمايه هاي زمين شناسي

ملاك انتخاب اصطلاح مرجح

- رواج داشتن . مثال: جامعه شناسي به جاي علم الاجتماع

- بومي بودن . مثال : آرمان گرايي به جاي ايده آليسم
- جديد و امروزي بودن . مثال: هواپيما به جاي طياره
- علمي بودن. مثال: اسيد سولفوريك به جاي جوهر نمك
- سرنام يا اختصار مشهور. مثال: يونسكو
- پرهيز از بكار بردن كلمات تركيبي به جز در مواردي كه اصطلاح قابل جدا سازي نيست مثل : آبله مرغان

جامعيت و مانعيت(Recall, Precision):نمايه سازي مدارك مختلف كه توسط افراد متفاوت، در مكان هاي متفاوت و در زمان هاي متفاوت و احتمالا با واژگان متفاوت است بنابراين نوعي همگوني و هماهنگي بايد پديد آورد كه در عين جامعيت دستيابي ،موارد زائد بازيابي نشود. پس براي ايجاد همگوني 2 نكته مهم است :

جامعيت و مانعيت يعني: از نظر منطق جامع افراد يا افراد همگون زير يك چتر بروند و مانعيت يعني اغيار زير اين چتر نروند. تلاش مي شود سطح هر دو (جامعيت و مانعيت)بالا برود و توازن آن حفظ شود

سایر انواع نمایه

- نمايه هاي الفبايي يا واژه اي

✓ اساسي ترين مانع در استفاده از آن ها مشكل مربوط به مترادف ها و پراكندگي مدخل هاست

✓ نام ها يا مدخل ها در محل الفبايي خود درج شد هاند

✓ شكل كلي و غالب نمايه هاي امروزي است

- نمايه هاي مؤلف

✓ براي ايجاد آ نها مي توان از رويكرد موضوعي غيرمستقيم بهره برد يعني با نويسندگان پيشتاز يا مشهور كار خوشه بندي اسناد و منابع را انجام داد

✓ نقاط مدخل آنها شامل اسامي افراد، سازمان ها، پديدآورندگان تنالگاني، و... م يباشند

- نمايه هاي كتاب

✓ هدف از طراحي آن تسريع در بازيابي اطلاعات موجود در كتاب است و به هيچ عنوان جايگزين اطلاعات موجود در كتاب نيست

✓ در انتهاي اغلب كتاب ها به چشم مي خورد

- نمايه هاي استنادي

- ✓ مزيت اصلي آن اين است كه برخلاف نمايه هاي سنتي حركت رو به جلو نيز دارد يعني مشخص م يكند كه چه مقدار از
- ✓ ايده هاي پيشين در ايجاد ايده هاي جديد نقش داشته اند
- ✓ براي خوشه بندي اسناد و منابع به منظور قراردادن آن ها در گروه هاي موضوعي مشابه بسيار مؤثر است
- ✓ شامل فهرستي از مقالات و همچنين فهرستي فرعي از مقالات منتشرشده اي است كه به آن مقالات استناد كرده اند

- نمايه هاي رد هاي

- ✓ هر چند استفاده از آن ها براي متخصصان ساده و كاربردي تر است اما كاربران عادي در استفاده از اين نمايه ها دچار مشكل م يشوند زيرا اغلب با نمايه هاي الفبايي راحت تر هستند
- ✓ مدخل هاي آن بر اساس رده ها يا سرعنوا نهاي موضوعي مرتب مي شود

- نمايه هاي همارا

- ✓ پايه و اساس بسياري از نظام هاي بازيابي اطلاعات در حال حاضر همين نمايه و نوع "پس همارا" مي باشد
- ✓ دو نوع عمده آن عبارتند از "پيش همارا" و "پس همارا"
- ✓ نوعي نمايه كه به زمان و نحوه تركيب اصطلاحات نمايه اي اشاره دارد

- نمايه هاي درهمكرد

- ✓ بيشتر براي مجلات علمي و كارهاي بزرگ و مهم بكار مي روند
- ✓ تركيبي از چند نمايه كه در طول زمان خاصي منتشر مي شوند

- نمايه هاي فرارسانه اي

- ✓ غالباً در مورد پايگاه هاي اطلاعاتي و منابع الكترونيكي غيرآنلاين كاربرد دارد
- ✓ اين نمايه به كاربر اجازه مي دهد تا از طريق برچسب هاي الكترونيكي مسير خود را به سمت اطلاعات مورد نياز بيابد

- نمايه هاي چندرسانه اي

- ✓ اين نمايه مواد متني، صوتي و تصويري را با هم تركيب مي كند

- نمايه هاي نشريات ادواري

- ✓ اساس و مبناي كار در آن ها مانند نمايه كتاب است با اين تفاوت كه در تدوين آن چندصد مؤلف نقش دارند و اساساً كار پردردسرتري است
- ✓ دو نوع م يباشد: نمايه انفرادي براي مجلات انفرادي و نمايه هاي كلّي براي گروهي از مجلات
- نمايه هاي گردشي عنوان
- ✓ ايرادهايي اساسي دارند مانند: پراكندگي متراد فها و اصطلاحات، عدم شفافيت لازم عنوان براي تبيين محتوا، محدوديت
- ✓ اصطلاحات موجود در عنوان براي بيان صحيح موضوع، و ...
- ✓ ساخت و ايجاد چنين نمايه اي بسيار سريع و فاقد دردسر مي باشد
- ✓ در اين نمايه فرض بر اين است كه عنوان به قدر كافي گوياي محتواي سند مي باشد
- با ورود رايانه به عرصه نمايه سازي دو نوع متداول نمايه گردشي عنوان شكل گرفت
- ✓ (2Keyword Out of Context كواك (- (1Keyword In Context كوئيك

عصر ارتباطات و اطلاعات

عصر انفجار اطلاعات

عصري كه گردآوري، ذخيره ، بازيابي و اشاعه اطلاعات به شدت تغيير يافته است.

از تاثيرات اين تغييرات :

- ✓ كاهش زمان جهت بازيابي اطلاعات مورد نظر
- ✓ اهميت دستيابي به مطالب روزآمدتر
- ✓ نقش مهم تحقيق در پيشرفت و توسعه جوامع بشري
- ✓ رشد پرشتاب دستاوردهاي علمي و فني
- ✓ افزايش متون ومدارك اطلاعاتي

درحوزه اطلاع رساني ، نمايه سازي مورد توجه قرارگرفت .

نمايه سازي حلقه ارتباطي ميان توليد كننده و مصرف كننده اطلاعات ، درجهت دستيابي آسان و سريع به دانش موردنياز كامپيوتر روش هاي نمايه سازي را دچار تحول كرده در حال حاضر نمايه سازي از طريق كامپيوتر و نرم افزارهاي توسعه يافته انجام مي شود.

به طور كلي كاربرد كامپيوتر درنمايه سازي را به دو دسته تقسيم مي كنند:

- ✓ نمايه سازي رايانه اي
- ✓ نمايه سازي با كمك رايانه.

- ✓ نمايه سازي با كمك رايانه :

دراين نظام نخست نمايه ساز با دقت , شناسه هاي موضوعي را انتخاب و علامت گذاري م ي كند و نمايه دستنويس تهيه مي كند. آن گاه اپراتورمدخلها را براساس نمايه دستنويس از طريق صفحه كليد وارد رايانه مي كند بعداز ورود ، رايانه مدخلها را بر اساس نظم خاصي مرتب مي كند، توصيفگرها را زير شنا سه ها مي برد، جاينماهاي هرمدخل را با هم ادغام مي كندو بعد يك نمايه الفبايي ارائه مي دهد. (بر اساس برنامه هاي از پيش طراحي شده)

اين نوع نمايه سازي با دوروش متفاوت انجام مي شود:

1) نمايه سازي در محيط گرافيكي (نمايه سازي درون كاشتي)

Graphic User Interface (GUI) . امروزه بيشتر نرم افزارها با رابط گرافيكي كار مي كنندكار با اين روش به صورت زيراست:

نمايه ساز كليدواژه ها يا كلمات و عبارتهاي مهم را انتخاب مي كند. اين عبارت انتخاب شده از نظر سيستم يك كليدواژه محسوب مي گردد. نمايه ساز همزمان باانتخاب كردن اين كليد واژه ها بايد يك پنجره واژه پرداز مستقل ديگر را باز مي كندو مدخلهاي ارجاعي را به طور دستي

تحرير مي نمايد. نتيجه اين كار يك فايل فرعي نيز به دست مي آيد. نمايه ساز فايل اصلي مدخلها وفايل فرعي را ادغام و توسط يك برنامه ترتيب بندي مناسب الفبايي مي كند.

2) نمايه سازي در محيط دستوري (نمايه سازي فرماني)

Command Line Interface : اين نرم افزار در محيط سي. ال. آي يعني با رابط دستوري (فرماني) كار مي كنند كار با اين روش چنين است:

نمايه ساز صفحه ها را در صفحه نمايش كامپيوتر مطالعه مي كندو كليدواژه ها را بادونماد كاملا استثنايي چپ و راست كه مشابه آن درمتن وجود نداشته باشد، علامت گذاري مي كند(مثلا كليدواژه درون سه كروشه قرار مي گيرديا...) . سپس نمايه ساز در همان فايل ارجاعات مورد نظر تحرير مي كند. پس از اتمام علامت گذاري با كروشه ها وارجاعات و تبديلها يك ابزار نرم افزار مي تواند كل نمايه را تهيه و الفبايي كند. اگر نرم افزارداراي امكانات نمايه سازي نباشد ويراستار مي تواند يك برنامه بسيار ساده (مثلا به زبان بيسيك بنويسد) كه فهرست كليدواژه هاي علامت گذاري شده و شماره صفحه هاي آنها را

استخراج كندويك برنامه يا ابزار ساده ديگر نمايه را ترتيب بندي و مرتب كند. روشهاي نمايه سازي با كمك رايانه داراي مزاياي سرعت ورود اصطلاحات نمايه اي، ظرفيت ذخيره سازي بالا ، سهولت اصلاح ، حروفچيني وتايپ، ويرايش و چاپ نمايه هستندواز لحاظ اقتصادي مقرون به صرفه مي باشند

در نمايه سازي به كمك رايانه ، نمايه ساز از رايانه براي انجام امور نمايه سازي استفاده مي كند.

اولين نمايه هاي رايانه اي مانند كوئيك و كواك از نوع نمايه سازي به كمك رايانه بوده اند.

نمايه سازي كوئيك:kwic

Keyword In Context:

:(پراشر 1989)

اين نظام نمايه سازي بر اين اصل استوار است كه عنوان يك مدرك، محتواي آن را نشان مي دهد. در اين نظام نمايه سازي عنوان و واژه هاي آن نشاندهنده موضوع مدرك مي باشد.

:(نايت 1970)

اين نوع نمايه، عنوان هاي مدارك را براساس كليدواژه هاي موجود در آن ها مرتب كرده و هر كليدواژه در نظمي الفبايي، نقش

مهمي به عنوان مدخل دارد.

:(لنكستر1991)

نمايه كوئيك نمايه اي گردشي است كه اغلب از عنوان هاي انتشارات به وجود مي آيد در نظر گرفته مي شود؛ (Access Point) بدين مفهوم كه هر كليدواژه موجود در عنوان به منزله نقطه بازيابي

مثال :

مباني نظري وتوصيه هاي عملي تحليل موضوعي و نمايه سازي /

توصيه هاي عملي تحليل موضوعي و نمايه سازي / ، مباني نظري و

تحليل موضوعي و نمايه سازي/ ، مباني نظري و توصيه هاي عملي

نمايه سازي/ ، مباني نظري و توصيه هاي عملي تحليل موضوعي و

kwoc نمايه كواك:Keyword Out of Context

نمايه كواك همان نمايه برون بافتي است.

دراين نمايه هركليدواژه به ترتيب از عنوان خود خارج شده و مقدم برساير اجزاي عنوان قرارمي گيرد.سپس عنوان مدرك به ترتيب طبيعي خود و به طور كامل درزير اين واژه يا به دنبال آن مي آيد. به اين ترتيب براي هر واژه مهم يك مدخل ساخته مي
شود.

:(لنكستر 1991)

همان نمايه كوئيك است با اين تفاوت كه كليدواژه هاي آن در خارج از متن مي آيد.

:(نيزو 1997)

نمايه اي كه در آن هر واژه مهم در زنجيره اي از متن، به عنوان اصطلاح هدايت كننده يا نقطه دسترسي و به دنبال زنجيره
كامل آمده است.

:(راولي نقل در نيزو، 1982)

1- در نمايه كواك همه واژه ها كه به صورت مدخل ظاهر مي شوند، از عناوين مدرك استخراج مي گردند
2- در اين نمايه بعضي مدخل ها معمولاً اصطلاحات تك واژه اي هستند

مثال :

"مباني نظري و توصيه هاي عملي تحليل موضوعي ونمايه سازي "

تحليل موضوعي

مباني نظري و توصيه هاي عملي تحليل موضوعي ونمايه سازي

توصيه هاي عملي

مباني نظري و توصيه هاي عملي تحليل موضوعي ونمايه سازي

مباني نظري

مباني نظري و توصيه هاي عملي تحليل موضوعي ونمايه سازي

نمايه سازي

مباني نظري و توصيه هاي عملي تحليل موضوعي ونمايه سازي

نمايه سازي رايانه اي:

يكي از انواع پركاربرد نمايه سازي كه سرعت ويكدستي را به دنبال دارد، نمايه سازي ماشيني است.

نمايه سازي ماشيني : عبارتست از انتخاب واژه هاي كليدي يك اثر بوسيله روش هاي ماشيني استفاده از كامپيوتر براي بيرون آوردن ونشان دادن واژه هاي نمايه

بدون دخالت انسان در حالي كه يك بار برنامه، و سياست كار آن، به ماشين داده شده است.

نمايه سازي رايانه اي , انجام كليه مراحل نمايه سازي اعم از :

- ✓ -انتخاب واستخراج اصطلاحات نمايه اي ازمتن ،
- ✓ -مدخل آرايي و
- ✓ -ارائه جاينماهاي هر مدخل و
- ✓ -چاپ نمايه توسط رايانه وبدون دخالت انسان را نمايه سازي رايانه اي گويند. نمايه سازي كامپيوتري، رايانه اي ، ماشيني و خودكار همگي با هم مترادفند و بجاي يكديگر به كار مي روند.

نكات :

وقتي شناسه ها توسط نمايه ساز انتخاب مي شوند نمايه سازي دستي گفته مي شود , اگر اين كار را كامپيوتر انجام دهد، نمايه سازي خودكار يا ماشيني ناميده مي شود. جهت استفاده از روش نمايه سازي ماشيني داده ها بايد به صورت ماشين خوان درآيند. در اين نوع نمايه سازي همه امور از انتخاب كليد واژه، شماره گذاري ، ترتيب بندي و غيره توسط كامپيوتر انجام مي گيرد.

<u>نمايه سازاني كه با اصول تحليل و طراحي سيستم و برنامه سازي كاربردي آشنايي دارند، مي توانند در زمينه كامپيوتري سازي نمايه سازي موثر باشند</u>

روال كار :

كامپيوتر مفاهيم مهم را كه بارها تكرار شده اند و جزو كلمات غير موضوعي زبان نيستند، به علاوه اسمهاي اشخاص، مكانها وغيره را به عنوان كليد واژه در نظر مي گيرد.اين كارتوسط يك نرم افزار به نام نرم افزار بسامدي استخراج مي كند. مثلايك مفهوم كه در ذيل يك بخش ازمتن بارها تكرارشده باشند، محتملا كليد واژه است، مگر آن كه جزو واژگان غير موضوعي باشد. واژگان غير موضوعي مانند(است، كه ، را و...) توسط يك فهرست به نام فهرست ايستا مشخص و ناديده گرفته مي شود. قسمت ديگر نرم افزار حاوي اسمهاي خاص است. نرم افزار با استفاده از اين دادگان ، اسمهاي خاص متن را تشخيص مي دهد و به عنوان كليد واژه در نظر مي گيرد.

فرايند انتخاب كليد واژه درنمايه سازي خودكار :

1. .شناسايي واژه هاي انفرادي از متن كه تحليل واژگان ناميده ميشود
2. .برداشتن واژه هاي با بسامد تكرار بالا كه در ارائه محتواي متن بي تأثيرند، با استفاده از فهرست واژه هاي غيرمجاز.
3. .تبديل واژه هاي باقي مانده به شكل ريشه آنها؛ يعني حذف پسوندها يا پيشوندها تا هر كلمه تا حد ريشه اش كوتاه مبدع آن است. porter . شود
4. .محاسبة رايانه اي بسامد ريشه هايي كه در متن تحليل شده اند، به منظور تعيين تابع ارزشگذاري هر ريشه.
5. .ريشه هايي كه ارزشگذاري بزرگتري دارند، براي متني كه در آن ظاهر شده، به عنوان كليدواژه تعيين مي شود.

شركت زافتكس نوعي برنامة كامپيوتري با عنوان IDX طراحي كرده كه به ارائه خدمات نمايه سازي ماشيني مي پردازد. اين نرم افزار از روش استفاده از واژه نامه بهره مي برد. اين نرم افزار امكان تبديل واژگان به ريشة آنها را جهت بازيابي بعدي فراهم مي كند علامتگذاري و محدودكردن واژه هاي ناخواسته را انجام م دهد.شكستن واژ ه هاي مركب و ترجمه و انجام عمل ارجاع و مترادف سازي و ساخت عبارات را نيز انجام مي دهد ماشين امكان تشخيص را تنها از طريق تطبيق واژه هاي استخراج شده از متن يا منتسب شده به متن با فهرستي كه واژه هاي غيرمجاز ناميده مي شود، به دست مي آورد.

در اختيار داشتن فهرستي از اين واژه ها و ارائه آنها به برنامه رايانه اي براي ممانعت از ورود آنها به فهرست واژه هاي مفهومي

مطلوب براي نمايه شدن، يكي از اقدامهاي سودمند در نمايه سازي خودكار مبتني بر كليدواژه هاست.

روش هاي نمايه سازي ماشيني

<u>الف. روش زبان شناختي:</u> اين روشها مي كوشند با كمك تحليل هاي شكل شناسي و ساختار نحوي مدرك توصيفگرها را استخراج نمايند.

- ✓ تجزيه و تحليل ريخت شناسي كه برمبناي ريشه لغات عمل مي كند.
- ✓ كلمات بدون بار معنايي موجود در سياهه بازدارنده را حذف مي كند.
- ✓ شكل هاي دستوري صرف كلمه را به يك شكل مي آورد.
- ✓ ضمايير را براساس اسامي مربوط به آنها مرتب مي كند.
- ✓ تجزيه و تحليل نحوي كه درسطح جملات امكانپذير است و برمبناي علامتهاي نحو لغات انجام مي شود.
- ✓ تجزيه و تحليل معنا شناختي كه درسطح مدارك مشترك در يك پايگاه صورت مي گيرد . ارتباطات معنايي موجود

- ✓ در يك مدرك شناسايي مي شوند تا متون مشترك بتوانند به صورت واحدهاي هم معني تجزيه شوند.

ب. روش هاي آماري: مشخص مي كندكه معني هرمفهوم منفرد در مدرك با حضور آن در جايگاههاي مختلف مدرك ارتباط تنگاتنگ دارد. بنابراين لغات درون متن شمارش مي شوندو ارتباط آنها ارزش گذاري مي شود. هدف آمارگيري از اطلاعات آن است كه لغات داراي بارمعنايي در مدرك به عنوان توصيفگر انتخاب شوند. روش هاي آماري عملا براي بالابردن جامعيت به كار گرفته مي شوند. درحاليكه روشهاي زبان شناختي در جهت بهبود مانعيت كاردارند.

ج. روش هاي مبتني براحتمالات: در اين روشها تئوري احتمالات براي مدل سازي رياضي مراحل بازيابي به كارگرفته مي شود. درحاليكه در توزيع آماري اصطلاحات يك مدرك مورد استفاده قرار مي گيرد.اين روش با عمليات رياضي مفروضات ساده و مطمئني را ارائه مي دهد. فرض برآن است كه مدارك براساس ميزان ربط درهنگام بازيابي مورد ارزيابي قرارمي گيرند.

فهرست ايستا , يا واژگان غير مجاز

تحليل كلمات يك متن نشان مي دهد گروهي از كلمات بي اهميت وجوددارد كه به فراواني در متن ظاهر مي شود (مانند

يك، به، نه، براي، با، چه كسي، چه موقع، است، آن). گروهي نيز وجود دارد كه بندرت در متن مي آيند و ممكن است نشان دهنده محتواي اطلاعاتي متن نباشند

اين دسته از واژه ها به تنهايي بارمعنايي ندارند , بود يا نبود آنها نه تنها در پرسش كاربر تأثيري ندارد بلكه در ميزان ربط يا عدم ربط مدارك بازيابي شده نيز تأثيري ندارد. اين واژه ها باعنوان واژههاي غيرمجاز براي ورود به نمايه معرفي مي شوند.

مزاياي تهيه ليست واژه هاي غير مجاز

درصورتي كه واژه هاي غيرمجاز قبل از فرايند نمايه سازي مدارك مشخص و فهرست آنها براي كنترل به رايانه داده شود، علاوه بر صرفه جويي در زمان و حجم بايگانيهاي نمايه، به ميزان زيادي از بازيابي مدارك نامرتبط و ريزش كاذب در جستجو جلوگيري خواهد شد.

چند نمونه از نرم افزارهاي معروف نمايه سازي كه در سيستمهاي كامپيوتر هاي شخصي كار مي كنند به

شرح زیراست:

INDEXING RESEARCH

محصول شرکتCINDEX

INDEX AID

محصول شرکتSanta Barbara Software Products

INQUIRY

محصول شرکتIndexer assistant

Indexit

محصول شرکتNorman Swartz

Bayside Indexing Service h

MARCREX محصول شرکت

Newberry Library

NLCINDEX محصول شرکت

-Watch City Software

Windex محصول شرکت

طرح هاي موفقي كه دراجراي نمايه سازي ماشيني ارائه گرديده :

1. طرح AIMS از كتابخانه ملي پزشكي آمريكا
2. طرح AIR/X در دانشگاه فني دارمشتات با موضوع فيزيك براي پايگاه اطلاعاتي فيزيك
3. طرح LISAو طرح COPSY توسط زيمنس اجرا شد .
4. 4PASSATتوسط زيمنس، سيستم نمايه سازي براي زبان آلماني
5. saphirدر دانشكده پزشكي هاروارد، سيستمي براي يك سيستم بازيابي هوشمند اطلاعات در مدارك بيولوژي و پزشكي بر پايه روش هاي آماري و زبان شناسي
6. SPECIALISTتوسط كتابخانه ملي پزشكي در آمريكا

عوامل موثر در نمايه سازي خودكار

1 - محدوده ركورد

اولين تصميم گيري مهم براي تهيه هر نوع نمايه ، گزينش حد و حدود ركوردي است كه واحد قابل جستجو را تعريف مي كند. اين تصميم گيري در بازيابي كارآمد نقشي حياتي دارد.

2 – محدوده اصطلاحات

تعيين حد و حدود يك واژه از ديگر مسائلي است كه در نمايه سازي خودكار بايد به آن توجه شود. در نظام هاي نمايه سازي دستي ، گزينش كلمات براي نمايه به سهولت انجام مي شود. اما در نمايه سازي خودكار از آنجا كه ماشين از هوشمندي لازم براي انتخاب كلمات برخوردار نيست بنابراين بايد حدود كلمه را تعريف كرد. معمول حدود كلمات نمايه را با استفاد ه از علائم نقطه گذاري تعريف مي كنند. به طور معمول ، فاصله بين كلمات و علائم دستوري و نقطه گذاري به عنوان مرز كلمات در نظر گرفته مي شود. روش هاي تعيين حد و حدود كلمات در نمايه سازي خودكار، بر اساس نوع برنامه و ميزان پيشرفتگي آنها متفاوت است

كستر نمايه سازي خودكار را به دو دسته استخراجي و تخصيصي تقسيم مي كند:

نمايه سازي استخراجي

ساده ترين روش نمايه سازي در پايگاه هاي اطلاعاتي ، روش نمايه سازي استخراجي است كه در آن واژه ها براي قرار گرفتن در نمايه ، توسط رايانه از متن استخراج مي شوند. در اين روش عموما بسامد تكرار واژه در هر ركورد يا مقاله تعيين شده و كلماتي كه بسامد تكرار آنها زياد است در متن نمايه قرار مي گيرند.

نمايه سازي تخصيصي

انتسابي فرآيند پيچيده اي است كه با استفاده از تحليل هاي آماري ، كلمات و اصطلاحات به مدرك منتسب مي شوند رايانه براي نمايه سازي از اصطلاحنامه يا كنترل واژگان بهره مي گيرد.

انواع نمايه استخراجي

نمايه سازي با استفاده از فهرست كلمات ممنوعه : در اين روش در هنگام نمايه سازي ، رايانه تمام كلمات متن را استخراج مي كند، سپس كلمات ممنوعه را حذف و بقيه كلمات را در يك نظام الفبايي مرتب مي كند.

نمايه سازي بسامدي : در اين روش، بسامد تكرار كلمات در هر ركورد مورد بررسي قرار مي گيرند و براساس بسامد تكرار در فهرست كلمات نمايه قرار مي گيرند.

نمايه سازي ريشه يابي : در بعضي از سيستم هاي نمايه سازي استخراجي ، از پسوند يا ريشه كلمات استفاده مي شود. در اين روش ريشه يا پسوند كلمات جايگزين مجموعه اي از كلمات هم ريشه يا پسوند مشترك مي شود. الگوريتم هاي ريشه يابي مختلفي چون الگوريتم هاي موضوعي خاص مانند الگوريتم هاي پزشكي وجود دارند الگوريتم Lovins. پسوند260

نمايه سازي بر اساس وزن دهي : در اين روش ، كلمات بر اساس محل قرار گرفتن خود در متن (مثل عنوان ، چكيده و)...امتيازدهي مي شوند. حضور كلمات در بخش هاي مختلف ركورد، امتيازات متفاوتي دارد. در نظام هاي رايانه اي بيشتر از روش هاي نمايه سازي استخراجي استفاده مي شود. يكي از عمده ترين مشكلات اين نمايه ها به ويژه هنگام استفاده در پايگاه هاي اطلاعاتي ، عدم بازيابي اطلاعات به دليل نبودن آن كلمه درخواستي در نمايه پايگاه اطلاعاتي است. دليل اين امر آن است كه بهره گيران ,يا همه كلمات مترادف با اصطلاح موجود در درخواست را وارد نكرده اند و يا از مترادفات آن بي خبرند. بنابراين ، بسياري از مدارك مرتبط از دست مي روند. براي رفع اين معضل طراحان پايگاه هاي اطلاعاتي توانايي هاي نمايه اي را با توانايي نرم افزاري درهم مي آميزند: يكي از روش ها، امكان نمايش نمايه و انتخاب واژه درخواستي توسط خود بهره گير است.

روش ديگر، استفاده از نظام بازخورد مرتبط است . اين روش به بهره گيران اجازه مي دهد تا مدارك مرتبط را برگزينند. سپس ازسيستم مي خواهند تا با توجه به اين مدارك ، مدارك مرتبط بيشتري را بازيابي نمايند. امروزه اين روش در اينترنت و پايگاه هاي اطلاعاتي تمام متن كاربرد فراواني دارد

نمايه سازي تخصيصي

درواقع در نمايه سازي تخصيصي براي هر واژه "پرونده اي " از كلمات و عبارات مرتبطي كه به نظر مي رسد تهيه مي شود. بنابراين مي توان از برنامه اي رايانه اي براي انطباق عبارت هاي مهم در مدرك با اين مجموعه پرونده ها استفاده كرد و درصورت انطباق واژه موجود در مدرك با واژه هاي موجود در پرونده هاي كلمات ، اصطلاح نمايه اي را انتخاب نمود موسسهBiosis . (15000 اصطلاح زيست شناسي)

- هر چه نمايه سازي تخصصي تر باشد صرفه اقتصادي آن در يك سيستم خبره بيشتر مي شود .

- سيستمي كه از يك نمايه سازي تخصصي استفاده مي كند مي تواند بعنوان يك سيستم كمي خبره از آن نام برد.

نتيجه تحقيق

پژوهشي را با موضوع نمايه سازي ماشيني متون فارسي براساس قانون زيف انجام دادند. نتايج نشان داد : « داورپناه و بلنديان »

- ✓ توزيع فراواني واژگان در متون فارسي داراي الگوي پيش بيني پذير است.
- ✓ كاربرد واژههاي با بسامد بالا و بسامد پايين درمقاله هاي فارسي، از قانون زيف پيروي ميكند.
- ✓ بسامد واژگاني مي تواند به عنوان معياري براي نمايه سازي ماشيني متون فارسي در نظرگرفته شود.

همخواني كامل بين بسامد واژگاني و كليدواژه هاي موضوعي در شيوة تفكيك صرفا ماشيني بدون دخالت عامل انساني به طور21 % است. در شيوة تفكيك ماشيني با دخالت عامل انساني، ميزان همخواني / متوسط در كل مقاله هاي مورد بررسي به ميزان 50 به 52 % ميرسد.

وضعيت همخواني كامل بسامد واژگاني با كليدواژههاي عنواني در شيوة صرفاً ماشيني بدون دخالت عامل انساني، به طور متوسط9 % است كه در شيوة ماشيني با دخالت عامل انساني اين ميزان بيشتر از 5 برابر شده و به 20 / 9% در كل، مقاله هاي مورد بررسي 14/54 ميرسد .

قانون زيف

جورج كينگزلي زيف استاد زبان شناسي دانشگاه هاروارد، در سال 1949 با آزمايش كلمات كتاب اوليسس جيمز جويس به نتايجي در مورد كلمات و ميزان تكرار آنها در متن رسيد. نتايج او به اين صورت بود كه: هر كلمه با فراواني(بسامد) ٌ اگر تمام كلمات يك كتاب را بشماريم و از زياد به كم مرتب كنبم به اين نتيجه مي رسيم كه رتبه همان كلمه در متن رابطه معكوس ٌ همان كلمه نسبت عكس دارد، يعني تعداد بارهايي كه هر كلمه در متن ظاهر مي شود با رتبه دارد. اين نسبت در كلمات كل متن برقرار است. كه به قانون زيف معروف شده است. قرار دارد و 3 برابر ٌ 1 قرار دارد دوبرابر بيشتر از كلمه اي در متن ظاهر مي شود كه در رتبه ٌ بر طبق زيف كلمه اي كه در رتبه 3 قرار دارد و همينطور تا آخر ٌ بيشتر از كلمه اي ظاهر م يشود كه در رتبه او اين قضيه را با اصل كمترين كوشش توجيه كرد. انسانها بر اساس اين اصل تمايل دارند كارهاي خود را به گونه اي ساده تر انجام دهند و در نوشتن متني سعي دارند بيشتر از كلمات تكراري استفاده كنند. و به همچنين در هنگام صحبت كردن و سخنراني سعي دارند كلمات كمتري را بيشتر تكرار كنند.آن در كل هر متن ٌ حاصل ضرب فراواني(بسامد) واژه در رتبه $r * f = k$: برقرار است كه r و رتبه f اين رابطه بين فراواني عددي (تقريباً) ثابت است.

رابطه لگاريتمي آن شناخته شده تر است و كاربرد بيشتر دارد:log r +log f = log c

اين رابطه به جز كلمات در بسياري از ديگر محيطها از جمله جمعيت شهرها، ميزان بازديد از صفحات اينترنت، شركت ها و كاركنان آن و نيز در نمايه سازي خودكار و... استفاده مي شود. از نظر محققان بسيار عجيب است كه چطور و چرا همچنين ساده اي در بسياري محيطهاي پيچيده اتفاق م ي افتد . ً رابطه اما به هرحال قانون زيف بسيار ساده است چراكه خود زيف زبان شناس بود و به مسائل رياضي چندان علاقه نداشت. و اين فرمولًً نتوانست محيطهاي خيلي پر تكرار را به درستي نشان دهد. بعد از زيف سه عدد ثابت به اين فرمول اضافه شد و كمي محاسبه آن را انعطاف پذيرتر كرد. كه به زيف مندلبرت معروف است .

f=(r+m)^B

نمايه سازي وب

وب به عنوان يكي از جذاب ترين بخش هاي اينترنت كاربردهاي فراواني دارد ومجموعه اي است از صفحات به هم پيوسته كه حاوي اطلاعات مفيدي درزمينه هاي موضوعي متفاوت است. همه دلايل مربوط به چرايي سازماندهي اطلاعات در محيط چاپي با شدت بيشتري درمحيط الكترونيك صادق است.

درحال حاضر موتورهاي كاوش, وب راتحت ضابطه درآورنده اندو با نمايه سازي صفحات ، پاسخي براي پرس وجوي كاربران

فراهم مي آورند.

تحولات اينترنت در حوزه نمايه سازي :

اينترنت بر سرعت ، دقت ، هوشمندي ، قدرت مشاركت ، كاربرپسندي ، جهاني بودن ، جامعيت و چندزباني نمايه ها افزوده است.

- حجم مجموعه

حجم اطلاعات موجود بر روي اينترنت و به طور خاص بر روي وب هر روزه در حال افزايش است . بديهي است هر چه حجم

مجموعه افزايش يابد نمايه سازي آن دشوارتر شده و نيازمند استفاده از روش هاي كامل تر و پيشرفته تري است .

- تفاوت هاي دايره لغات

معمولا" در يك مجموعه يا پايگاه اطلاعاتي يكدستي - چه ازنظر ساختار و چه از نظر موضوعي - وجود دارد. وب برخلاف پايگاههاي اطلاعاتي ، تقريبا از هيچ تجانس موضوعي و زباني برخوردار نيست . علاوه براين ،

بهره گيران از اينترنت نيز بسيار متنوع اند. در يك تحقيق مشخص شد كه احتمال استفاده يكسان دو فرد از اصطلاحي واحد در اينترنت در حدود 20 درصد است .

- راهبردهاي كاوش

كاهش مداخله كاوشگران متخصص در بازيابي اطلاعات، تنوع بهره گيران , فقدان آگاهي درباره شيوه و راهبرد كاوش در اينترنت ، توليد نمايه منسجم را دشوار ساخته است .

- رسانه هاي جديد ارتباطي در اينترنت

امكانات تازه اي كه در وب پديد آمده مانند امكانات فرامتني , لينكها , خدمات چندرسانه اي و چندزبانه بودن اطلاعات اينترنت ، بيش از گذشته ايجاد نمايه را با دشواري مواجه ساخته است . نكته مهم براي نمايه سازان اين است كه نحوه و ساختار ذخيره و بازيابي اطلاعات چگونه است و از چه آلگوريتمي براي ذخيره وبازيابي اطلاعات در آنها استفاده شده است . تفاوتهاي موجود در قابليتهاي جستجو و بازيابي اطلاعات و نحوه نمايش و رتبه بندي اطلاعات در هر پايگاه اطلاعاتي نشان دهنده تفاوت در به كارگيري آلگوريتمهاي مختلف است . سعي شده از روابط آلگوريتمي جهت جايگزين نمودن با پردازش فكري انسان براي نمايه سازي ماشيني استفاده شود . يعني كليد واژه ها بر اساس الگوريتم خاصي در نمايه سازي ماشيني انتخاب شوند . اكثر موتورهاي كاوش و پايگاه هاي تجاري به دليل ماهيت تجاري بودن , الگوريتمهاي نمايه سازي خود را به راحتي در اختياركاربر قرار نمي دهند. هريك با الگوريتمها و سياستهاي متفاوتي به مقوله نمايه سازي مي نگرند.

تفاوت درنتايج بازيابي شده در موتورهاي كاوش مختلف نشان بارزي از وجود تفاوت درالگوريتمهاي نمايه سازي و يا پايگاه هاي داده آنهاست. حجم نمايه در بين موتورهاي كاوش هنوز هم يكي از نكات مهم و اصلي رقابت در بين توليدكنندگان موتورهاي كاوش است گرچه امروزه تهيه بهترين نمايه و نه بزرگترين آن مورد توجه قرار مي گيرد. نمايه هاي وب براي نشان دادن ليستهاي منابع خود ازپيوندهاي فرا متن استفاده مي كنند، آنها اين امكان را دارند تا از طريق لينك ها صدها و بلكه هزاران منبع را در بر بگيرند. در محيط وب به دليل حجم وسيع اطلاعات منتشر شده عملا نمي توان به شيوه هاي دستي نمايه سازي متوسل شد. بنابراين نمايه سازي دروب به صورت خودكار وتوسط موتورهاي كاوش انجام مي شود. موتورهاي كاوش پايگاههاي اطلاعاتي قابل جستجويي هستند كه از طريق برنامه هاي كامپيوتري به شناسايي ونمايه سازي خودكار صفحات وب مي پردازند. موتورهاي كاوش برنامه هاي خودكاري هستند كه هيچ گونه وابستگي به نيروي انساني ندارند.

نوع جمع آوري اطلاعات در اينترنت به دو گروه اصلي تقسيم ميشود:

موتورهاي جستجوگر ياSearch Engine

فهرست هاي وب ياWeb Directory

هر دو مورد اطلاعات را در اختيار كاربران قرار ميدهند اما تفاوت اصلي آنها در رو ش جمع آوري اطلاعات است. در موتورهاي توسط Web Directory جستجوگر اطلاعات توسط نرم افزار جمع آوري و طبقه بندي ميشوند اما در فهرست هاي وب يا عوامل انساني انجام ميشود اما با توجه به رشد بسيار سريع اينترنت عملا وب دايركتوري ها كاربران چنداني ندارند

Yahoo directory

Dmoz.org

دلايل استفاده از راهنماهاي موضوعي

- ✓ رده بنده ي دستي تضميمن كننده ربط مدارك در موضوع مي باشد
- ✓ كانون تمركز ابتدايي فرايند جستجو انجام شده است
- ✓ با مرور حوزه موضوعي به جنبه هاي تخصصي تر به صورت خود كار دست مي يابيم
- ✓ مدارك با موضوعات مشابه با هم گروه بندي شده و ارتباط مدارك انجام مي گيرد

معايب راهنماهاي موضوعي

- ✓ مشكل در روز آمد شدن اطلاعات
- ✓ يك مدرك ممكن است زير يك حوزه موضوعي قرار گيرد اما جستجو گر در حوزه موضوعي ديگر به دنبال آن باشد
- ✓ شناخت طبقه آغازين مرور ممكن است مشكل باشد

فناوري موتورهاي جستجو از دو فرايند مجزا اما يكپارچه تشكيل شده است

1 – ايجاد نمايه اي از مدارك وب : وقتي دكمه جستجو فعال مي شود , اين نمايه جستجو مي شود(فايل مقلوب)

2 – مجموعه اي از مدارك در قالب سياهه اي كه بر مبناي ميزان ربط با پرسش جستجو رتبه بندي شده نشان داده مي شود

روش كار موتورهاي جستجوگر

Spider , crawler, در موتورهاي جستجوگر كار جمع آوري و طبقه بندي اطلاعات بر عهده نرم افزار است. اين نرم افزارها ناميده ميشوند .اسپايدرها هميشه مشغول كار هستند و اطلاعات را از گوشه و كنار جمع worm wanderer,Robot, آوري ميكنند حال تفاوتي ندارد كه اين اطلاعات صفحات وب جديد باشند و يا اطلاعاتي

باشند که قبلا وجود داشتند اما حالا بروزرساني شده اند. هدف تمامي موتورهاي جستجوگر وب در واقع واحد است. همه آنها سعي در جمع آوري اطلاعات مورد نياز براي كاربران خود را با بالاترين ميزان دقت دارند. اين مسئله كه يك وب سايت جستجوگر چگونه اطلاعات را بهتر طبقه بندي كند بستگي مستقيم به نوع موتور جستوي سايت و الگوريتم آن دارد همه جستجوها حتي ساده ترين آن ها , نمايه موتور جستجو يا پايگاه داده در يك كامپيوتر دور را جستجو مي كند

چه اقلامي نمايه سازي مي شوند

- ✓ بزرگترين موتورهاي جستجو هم قادر به نمايه سازي كامل وب نيستند.
- ✓ برخي از اين موتورها نظام نمايه سازي تمام متن دارند و هر واژه موجود درمتن به جز واژه هاي فاقد بار اطلاعاتي مانند حروف
- ✓ اضافه، ربط و تعريف را نمايه مي كند .
- ✓ برخي ديگر، سرعنوانها، عناوين فرعي و فراپيوندها را همراه 20 خط ابتداي متن و 100 كلمه اي كه از بسامد بالايي برخوردار
- ✓ است، نمايه مي كنند.

- ✓ برخي توصيف متني از يك شكل را نمايه مي كنند
- ✓ برخي از موتورهاي جستجو در ازاي هزينه اي , صفحات را بي درنگ نمايه سازي مي كنند
- ✓ برخي از اسپايدر ها در هنگام گردآوري صفحات پيوندهاي فرامتن را دنبال مي كنند كه زمان نمايه سازي را افزايش مي دهد

بازنمود نمايه در وب

نمايه وب به سه صورت نمايش داده مي شود:

1. نمايش تيتر حرف
2. نمايش فرم
3. نمايش سنتي

الگوريتم موتورهاي جستجوگر چيست؟

الگوريتم در واقع مجموعه اي از دستورالعمل هاي گوناگون است كه ترتيب قرارگيري سايت ها را در موتورهاي جستجوگر تعيين مي كند.براي اين كه مشخص شود كه كدام سايت ها بايد در ليست نمايش نتايج جستجو در ابتدا نمايش داده شوند ، موتور جستجو بv اساس پارامترها عمل ميكنند. پس در واقع دو مورد

الگوريتم و پارامترها در نوع نمايش نتايج جستجو موثر هستند و اما معماري هر دو كاملا محرمانه است موتورهاي جستجو پيوسته الگوريتم هاي نمايه سازي خود را بهبود مي بخشند .

الگوريتم وزن دهي:

- ✓ پاره هاي اطلاعاتي را جدا كرده سپس به آن ها بر اساس معيارها مقدار و وزن داده و در نمايه قرار مي دهند.
- ✓ وزن واژگان عامل مهمي در تعيين ربط مدارك يافت شده و نتايج جستجو به شمار مي رود (رتبه بندي مدارك)
- ✓ در بخش FAQ يا راهنماي جستجو موتورهاي جستجو , اطلاعاتي در خصوص معيارهاي رتبه بندي و ربط دارند

معيارهاي ربط يا وزن دهي

ربط آماري خوشه سازي تحليل پيوند clustering

معيارهاي ربط آماري

- ✓ بسامد تكرار واژه ها
- ✓ مكان ظهور واژه مثلا" وجود واژه در برچسبهاي عنوان
- ✓ طول مدرك (واژه در مدرك كوتاه وزن بيشتر در مقايسه با مدرك طولاني دارد)

معيار تحليل پيونديا شهرت(نخستين روش رتبه بندي در موتورهاي جستجو)

pagerank شمارش تعداد پيوند ها به صفحه از نمايه سازي استنادي گرفته شده است(هرچه به اثري بيشتر استناد گردد , اهميت و اعتبار بيشتري دارد)

به كارگيري اين معيار در گوگل مبناي I’m feeling lucky

معيار خوشه سازي

موتور جستجو بر اساس يك الگوريتم , اصطلاحات مشترك موجود در مدارك را در قالب خوشه هاي موضوعي مرتب مي نمايد .

(related terms)

Ask , clusty , google

معيار پيوندهاي مرتبط

مشابه استفاده از " نيز نگاه كنيد " به معني مدارك بيشتري درست همانند اين مدرك بدهيد .

در گوگل Similar pages

قابليتهاي جستجو در مورتورهاي كاوش

جسجوي ساده	جستجوي پيشرفته
عملگرهاي بولي	مجاورت
كوتاه سازي	جستجو در فيلد

انواع موتور كاوش

موتوركاوش عمومي

- ✓ yahoo, google

موتور كاوش تخصصي: الگوريتم آنها مدارك مربوط به حوزه موضوعي را هدف قرار داده و توسعه مي دهند .

- ✓ Search-engine-index.co.uk
- ✓ كتابخانه ديجيتال حاوي مقالات داوري شده)citeseer.ist.psu.edu
- ✓ بخشي براي جستجوي وب پنهان (مقالات مجلات) دارد scirus.com
- ✓ Techxtra.ac.uk

ابر موتور كاوش : پرسش را همزمان به چند موتور كاوش و راهنما مي دهند

- ✓ Dogpile , mamma, metacrawler

اينترنت همانگونه كه ابزاري قدرتمند براي دسترسي به اطلاعات است به همان نسبت نيز ابزاري است كه اطلاعات گمراه كنندهرا اشاعه مي دهد.

تهيه كنندگانhotbot

"ايجاد صفحاتي كه مغرضانه مرورگرهاي كاوش را گول زده و باعث مي شوند تا صفحات نامرتبط با كاوش را بازيابي كنند يا امتياز بالايي را به مدارك آنها اختصاص دهند بسيار رواج يافته است . تكرار هزاران باره كلمات در بخش كليدواژه ها يا اظهارنظرها , يا گنجاندن حجم زيادي از كلمات نامرئي در فونتهاي ريز يا به رنگي مشابه با رنگ پس زمينه صفحه مثال هايي ازحقه زدنهايي است كه عموميت دارند . چنين حقه هايي باعث مي شود تا هات بات به دو روش به اين صفحات امتيازات كمتري را اختصاص دهد : اول حجم اين مدارك طولاني تر است , دوم اگر هات بات يك روش حقه زدن متداول را تشخيص دهد به شدت امتياز آن صفحه را كم خواهد كرد "

وب پنهان : بخشي از وب كه توسط نرم افزار خزنده موتورهاي جستجوي عمومي نمايه نمي شوند يا نمي توانند نمايه شوند , شامل :

- ✓ مبتني بر پايگاه اطلاعاتي html صفحات
- ✓ كتابخانه هاي ديجيتال
- ✓ pdf مانند فايل html قالبهاي غير
- ✓ اطلاعات ناپايدار مانند خبرهاي جاري, آگهي ها
- ✓ آثار غير تجاري مانند گزارشهاي دولتي , خبرنامه ها

موتورهاي جستجو با افزايش قابليتهاي جستجوي خود با چالش وب پنهان روبرو شده اند. از جمله گوگل اسكولار رويكردهاي نمايه سازي خودكار دروب:

- ✓ محتوا محوري
- ✓ معنا محوري

محتوا محوري: اغلب موتورهاي كاوش حاضر از روش نمايه سازي برمبناي كليدواژه هاي متن استفاده مي كنند.دراين شكل فرايند نمايه سازي سه مرحله خواهد داشت : شكستن كلمات، تعديل وحذف كلمات غيرموضوعي، استفاده از الگوريتم ريشه ساز جهت توليد ريشه هاي مفاهيم. درمرحله شكستن كلمات، داده هايي كه به صورت رشته اي از كاراكترها هستندمورد بررسي قرار گرفته وحدود كلمات وفاصله ميان آنها مشخص مي گردد.

درمرحله تعديل كلمات مزاحم، بزرگ نويسي ، نقطه گذاري وملواردي از اين دست مديريت مي شود. كلمات مزاحم درنمايه سازي كلماتي هستند كه بار معنايي خاصي ندارند و تنها براي ايجاد پيوستگي و ارتباط در جمله ها به كار مي روند. بعد از اين مرحله از الگوريتمي جهت توليد ريشه ها و مفاهيم استفاده مي شود.

معنا محوري:

بعضي موتورهاي كاوش رويكردي مكانيكي دارندو به مفاهيم ، الگوها وكليدهايي كه به فهم مفاهيم مي انجامد توجهي ندارند. دراين ابزارها، جست و جوي واقعي صرفابر مبناي كليدواژه هاست. استفاده از فهرست مترادفها وبهره گيري از جست وجوي فازي از راهبردهاي مطرح شده جهت رفع مشكلات جست وجوي كليد واژه اي است. روشي كه دراينجا مطرح است بهره گيري از نمايه سازي معنايي پنهان جهت بهبود مانعيت، جامعيت ورتبه بندي نتايج كاوش

است.نمايه سازي معنايي پنهان به كاربران اين اجازه را ميدهد كه جست وجوي خودرا به مفاهيم و نه فقط كليد واژه ها محدود كنند.درزمينه نمايه سازي خودكار دروب حركتهاي جاري به سمت بهره گيري ازداده هاي ساختار يافته وتحقق وب معنايي است ، اما اين حركتها به طوركامل به انجام نرسيده است و هنوز مشكلات حل نشده فراواني در اين مسير وجود دارد.

وب معنايي(Semantic web)

وب معنايي نسبت به وب يك انقلاب محسوب مي شود كه در آن اطلاعات قابل خواندن و تجزيه و تحليل توسط ماشين است در حالي كه صفحات وب كنوني را فقط انسان مي تواند بخواند وب معنايي اين اجازه را به مرورگرها و ديگر نرم افزارها مي دهد تا اطلاعات را خوانده به راحتي تجزيه و تحليل كنند

فضايي جهاني از جنس محاسبات هوشمند ماشيني را مي توان تصور كرد كه در آن تمامي پايگاه هاي دانش (Knowledge bases) به صورتي معني گرا و با توانايي درك مفهومي همديگر در كنار هم قرار خواهند گرفت. آينده وب نه فقط توسط انسانها قابل فهم است بلكه توسط ماشين ها نيز قابل درك و پردازش است.

در زير سه تعريف مختلف از وب معنايي ارائه شده است :

- ✓ پروژه اي با هدف ايجاد رسانه اي جهاني براي رد و بدل كردن اطلاعات بصورتي كه براي كامپيوتر قابل فهم و پردازش باشد .
- ✓ وب معنايي، شبكه اي از اطلاعات در مقياس جهاني است به نحوي است كه پردازش آنها توسط ماشين ها به سادگي امكان پذيراست .
- ✓ وب معنايي شامل داده هاي هوشمند وب است كه توسط ماشين ها قابل پردازش است

از ديدگاه مديريت نظام اطلاعاتي چهار ويژگي اصلي براي اطلاعات بازيابي شده قابل لحاظ است :

- ✓ دقت، پيوستگي زماني، بهنگام بودن و مرتبط بودن.
- ✓ لذا نمايه سازي بايد بر مبناي چهار ويژگي مذكور انجام پذيرد. بنابراين جهت رسيدن به نمايه سازي جامع و مانع وب به نمايه سازي دقيق موتورهاي جستجو نياز داريم.

درنمايه سازي همواره اين سؤال مطرح است كه يك نماية خوب چگونه نمايه اي است؟

نماية مؤثر كاربر را به اطلاعات دقيق، بدون دشواري، خطا و موارد نامرتبط هدايت مي كند و به ندرت منجر به بازيابي اطلاعات

سطحي مي شود. نمايه از ابزارهاي مهم تحليل مدرك است بنابراين كنترل كيفيت اين ابزار در نظام اطلاع رساني مهم است هدف هر نمايه، بازيابي ركورد ها يا مداركي است كه به وسيلة فرايند نمايه سازي، ذخيره و سازمان دهي شده اند. هدف از ارزيابي نمايه، تعيين ميزان اثربخشي و كارايي آن است

كلولند Cleveland معيارهاي ارزيابي در سه سطح ارزيابي كرده است:

1. فني

2. معنايي

3. ميزان كارآيي

در سطح فني، نمايه بايد داراي زبان مناسب و شكل قابل درك باشد و به سادگي بتوان از آن استفاده كرد

در سطح معنايي،واژه ها بايد معاني را بدون ابهام منتقل كنند

سطح سوم، نمايه بايد اطلاعات مرتبط را به درستي شناسايي كند و در بازيابي اطلاعات مؤثر باشد

معيارهاي ارزيابي نمايه از ديد لنكستر

1. دامنة آن

2. توانايي بازيابي گزينه هاي مورد نياز (جامعيت)

3. توانايي پيشگيري از بازيابي گزينه هاي ناخواسته (مانعيت)

4. مدت زمان پاسخگويي نظام: از يك نظام نمايه سازي خوب، بازيابي سريع مدارك مرتبط انتظار مي رود. اما بايد به خاطر

داشت كه زمان پاسخگويي به نوع پرسش جست وجو بستگي دارد.

5. ميزان تلاشهاي مورد نياز كاربران: اگر دسترسي به مدارك آسان باشد، يعني نظام نمايه سازي اثربخش است.

6. هزينه: در نظام نمايه سازي، هزينه در مقابل سودمندي هاي نظام، هميشه قابل قبول است. اما نظام نمايه سازي مؤثر حداكثرمزايا را با حداقل هزينه ها فراهم مي كند .

كيفيت نظام نمايه سازي به عواملي مانند :

- ✓ دامنة پوشش، جامعيت، مانعيت،تازگي و صحت داده هابستگي دارد.

پراشر عوامل زير را بر كارآيي نمايه سازي مؤثر مي داند:

- ✓ عوامل مخل
- ✓ ضريب ريزش
- ✓ ضريب تازگي و نوظهوري
- ✓ كل نگري و جزءنگري.

ضريب عوامل مخل ، مكمل ضريب مانعيت است كه تعداد مدارك نامرتبط در كل مدارك بازيابي شده را نشان مي دهد و از لحاظ رياضي به صورت زير نشان داده مي شود:

100% ×تعدادكل مدارك نامرتبط بازيابي شده= ضريب عوامل مخل

تعداد كل مدارك بازيابي شده

هر چه ضريب عوامل مخل كم تر باشد، كارآيي نظام نمايه سازي بيشتر است.

ضريب ريزش نشان مي دهد كه چه تعداد مدارك نامرتبط از كل مدارك نامرتبط پايگاه بازيابي شده اند و از لحاظ رياضي فرمول آن به صورت زير است:

100%× تعدادكل مدارك نامرتبط بازيابي شده= ضريب ريزش

تعداد كل مدارك نامرتبط موجود در پايگاه

ضريب تازگي و نوظهوري بخشي از مدارك جديد است كه براي اولين بار مورد توجه كاوشگر اطلاعات قرار گرفته است . در ميان كل مدارك مرتبط، درصد كمي از مدارك جديد وجود دارند كه اطلاعات جديدي ارائه كنند. كارآيي بازيابي در نظام نمايه سازي با اصلاح زبان نمايه سازي و زبان پرس وجو افزايش مي يابد. جزءنگري و كل نگري فنوني هستند كه براي چنين اصلاحي مورد استفاده قرارمي گيرند. سطح بالاتر كل نگري، جامعيت را افزايش و

مانعيت را كاهش مي دهد. اما تعديل جامعيت و مانعيت براي حفظ كارآيي بهينة نمايه سازي الزامي است

ضعف در نمايه سازي

<u>ضعف در تحليل مفهومي</u> :

1 – ضعف در تشخيص موضوعي كه مورد علاقه جامعه بهره گير است

2 – تفسير نادرست از جنبه هايي كه مدرك واقعا" در مورد آن بحث كرده است .

ضعف در ترجمه :

1 – كوتاهي در استفاده ازا اصطلاح اخص تر براي بعضي از موضوعات

2 – استفاده از اصطلاحي نا مناسب با محتواي موضوعي

استانداردهاي حوزة نمايه سازي

ANSI /Z. 39.4معيارهاي اساسي براي نمايه ها

ANSI/NISO Z.39.50 بازيابي اطلاعات: كاربرد تعريف خدمات و پروتكل

BS .3700آماده سازي نمايه براي كتابها، نشريات ادواري و ديگر مدارك

BS 6529 1984 بررسي مدارك، تعيين موضوعهاي آن ها و انتخاب اصطلاحات نمايه اي

ISO 5963 تحليل مدارك، تعيين موضوعها، انتخاب اصطلاحات نمايه اي

ISO 999 دستور عملهايي براي محتوا، ساختار و نمايش نمايه ها

1. دامنة زير پوشش نمايه كامل باشد؛

2. در انتخاب اصطلاح، انسجام داشته باشد؛

3. اصطلاحات انتخاب شده با سطح كاربران متناسب باشد؛

4. ارجاعات به ميزان كافي وجود داشته باشند؛

5. زنجيره هاي بسيار طولاني از سرعنوانهاي فرعي به هم پيوسته، وجود نداشته باشد؛

6. سرعنوانهاي فرعي به درستي سرعنوانهاي اصلي را نشان دهند؛

7. مكان نماي نادرستي وجود نداشته باشد؛

8. زنجيرة بسيار طولاني از مكان نماها وجود نداشته باشد؛

9. نظم الفبايي در متن، يكپارچه و صحيح باشد؛

10 . اشتباه املايي وجود نداشته باشد؛

11 . ارجاع نادرست و دوطرفه وجود نداشته باشد.

اصطلاحنامه

گنجواژه يا اصطلاحنامه، مجموعه اصطلاحات يك رشته است كه ميان آنها روابط معنايي، رده اي، و سلسله مراتبي برقرار شده و توانايي آن را دارد كه موضوع آن رشته را با همه جنب ههاي اصلي و فرعي و وابسته، ب هگونه اي نظام يافته و ب همنظور ذخيره و بازيابي اطلاعات ارائه دهد. اصطلاحنامه معادل فارسي واژه انگليسي Thesaurus است

اصطلاحنامه : از نظر وظيفه و كاركرد، ابزار كنترل واژه ها به منظور برگرداندن زبان طبيعي مدارك به زبان مقيد است . از نظر ساختار، واژگان كنترل شده و پوياي زمينه اي خاص از دانش بشري است كه براي ذخيره و بازيابي اطلاعات آن حوزه به كار مي رود .

اهداف اصطلاحنامه

اصطلاحنامه داراي هدف هاي اساسي زير است:

نماياندن ساختار زمينه معيني از دانش چنان كه هم نمايه ساز و هم جست وجوگر بتوانند از گستره آن زمينه و ارتباط ميان مفاهيم آن با انديشه هاي مرتبط آگاهي يابند

- ✓ ارائه اصطلاحات استاندارد در زمينه اي معين
- ✓ برقراري نظام ارجاعات ميان اصطلاحات و رده بندي اصطلاحات به صورت سلسله مراتبي
- ✓ تأكيد بر توجه به نيازهاي اطلاعاتي استفاده كنندگان
- ✓ تعيين اصطلاحات مجاز و مشخص كردن حدود معاني اصطلاحات به منظور ايجاد هماهنگي در نمايه سازي

روابط ميان اصطلاحات

بايد توجه داشت كه ويژگي ذاتي اصطلاحنامه، توانايي تعيين و نمايش روابط معنايي ميان واژه هاست و يك رابطه يك سويه نداريم و رابطه همه جانبه وجود دارد. اين روابط ممكن است يكي از اين سه نوع باشد:

الف) رابطه هم ارزيEquivalance Relation :

ميان اصطلاح پذيرفته شده و اصطلاح پذيرفت هنشده برقرار مي شود

مثال: اصطلاح گياهان به جاي نباتات

ب) رابطه سلسله مراتبيHeirachial Relation

بيان كننده رابطه اعم و اخص ميان مفاهيم است كه در واقع، اصطلاحنامه ها را از واژه نامه هاي متداول متمايز مي كند .

- گياهان ا.خ درختان
- صنوبر ا.ع درختان

رابطه همبسته یا همایند:Associative Relation

رابطه ميان دو اصطلاح كه به دليل وابستگي معنايي ، وجود يكي ديگري را نيز به ذهن متبادر م يكند. مثال : دو اصطلاح اسب و سوارکاري

- ✓ پشتيباني اصطلاحنامه ها بسيار گران است و به دانش خاص نياز دارد
- ✓ اصطلاحنامه ها نقش مهمي در نظامهاي ذخيره و بازيابي اطلاعات دارند.
- ✓ ظهور وب، همراه با توسعه و پيشرفتهاي اخير در كاربرد اصطلاحنامه ها به عنوان ابزارهاي بازيابي اطلاعات، باعث تولد نسل جديدي از اصطلاحنامه ها شده است.
- ✓ مشكلات روشهاي آماري و زبانشناسي رايانه اي باعث شده است كه اين پروژه ها فكر استفاده از ابزارهاي ديگري را در سر بپرورانند.
- ✓ اصطلاحنامه هاي وب محور، راه خود را به محيطهاي بازيابي و سازماندهي اطلاعات وب محور باز نموده و در تهيه ابر داده ها، نمايه سازي صفحات، سايتهاي وب، پايگاههاي داده و موتورهاي جستجو استفاده مي شوند.
- ✓ ساختار هاي معنايي موجود در اصطلاحنامه ها مي توانند هم در سازماندهي و هم در بازيابي اطلاعات وب و منابع دانش نقش داشته باشند.
- ✓ اصطلاحنامه هاي پيوسته يا به كمك رايانه تدوين و بازنمايي مي شوند يا به صورت رايانه اي توليد مي گردند.

در اين دو رويكرد كلي تدوين، حركت از ذهنيت به عينيت است ،

بدين ترتيب كه روابط اصطلاحنامه اي ممكن است توسط متخصصان تنظيم شوند و از رايانه براي بازنمايي آن ها استفاده گردد كه در اين صورت، رسوخ ذهنيت در تدوين اصطلاحنامه ، بيشتر است. ممكن است روابط به وسيلة استدلال گرهاي خاصي كه مبتني بر قواعد رياضي و جبرهستند توليد شوند كه در اين صورت ، رسوخ ذهنيت در تنظيم روابط بسيار كاهش مي يابد ،ولي انتخاب اصطلاحات همچنان به وسيلة انسان صورت مي گيرد. در روش هاي جديد توليد خودكار اصطلاحنامه از آمارهاي متني و الگوريتم هاي پيچيده براي انتخاب اصطلاحات و تشخيص روابط اصطلاحنامه اي و خوشه بندي لغات استفاده مي شود. رسوخ ذهنيت محدود به تدوين الگوريتم هاي توليد اصطلاحنامه است و پس از اين مرحله، اصطلاحنامه به صورت خودكار و بدون دخالت انسان توليد مي شود . هر دوي اين دورويكرد كلي مي توانند معايب و مزايايي داشته باشند كه با توجه به حوزه ها و موقعيت هاي مختلف، بايد رويكرد مناسب براي تدوين اصطلاحنامه انتخاب شود . اصطلاحنامه هايي كه به صورت خودكار توليد مي شوند ممكن است با نيازهاي كاربران و اصطلاحات مورد نظر آن ها براي انجام جست وجو و بازيابي فاصله داشته باشند و نيازهاي خاص آنان لحاظ نشده باشد. همچنين براي حوزه هاي فعاليت كوچك ، مقرون به

صرفه نيستند . در عوض سرعت توليد، يكدستي روابط اصطلاحنامه اي، قابليت روزآمدسازي، ويرايش سريع و كارآمد، و بازنمايي مناسب اين نوع اصطلاحنامه ها در وب، از نقاط قوت آن ها محسوب مي شوند.

آنتولوژي

- ✓ به معناي هستي شناسي است. (onto+logy) آنتولوژي برگرفته از تركيب يوناني
- ✓ هستي شناسي : فهم چگونگي و علت امور در دنياي بيروني
- ✓ همه علوم بالاخص علوم پايه به دنبال شناخت هستي اند
- ✓ اطلاعات و منابع موجود در وب بصورت فزاينده ايي رو به رشد هستند و استفاده كنندگان وب نيازمند يك درك مشترك از
- ✓ آنها دارند.
- ✓ آنتولوژي نقش اصلي را در مبادله اطلاعات و توسعه وب لغوي بسمت وب معنايي دارد.
- ✓ آنتولوژي يك مدل مفهومي است كه موجوديتهاي واقعي در يك دامنه خاص و روابط بين آنها را به صورت صريح و رسمي
- ✓ مدلسازي مي كند.

در سال 1980 ، مجمع هوش مصنوعي از لغت آنتولوژي براي دو منظور استفاده كرد:

- ✓ نظريه اي در مورد جهان مدل شده
- ✓ مؤلفه اي از سيستم هاي دانش.

آنتولوژي در هوش مصنوعي و همچنين علوم كامپيوتر به مجموعه اي از لغات و فرضيات (عموماً در منطق مرتبه ي اول) گفته مي شود كه با توجه به معني آن لغات ايجاد شده اند و به منظور توصيف يك واقعيت خاص طراحي شده اند. امروزه آنتولوژي درهوش مصنوعي، مهندسي نرم افزار، مهندسي سيستمها و معماري اطلاعات كاربرد دارد.

وب معنايي يا وب 3 كه امروزه از مباحث تازه و جذاب در حوزه كامپيوتر و ارتباطات است،موجب مي شود كه جستجو در اينترنت براي كاربران سريعتر و راحتتر گردد. در اين خصوص عامل اصلي ظهور و موفقيت و بمعنايي و آنچه كه در پس اين فناوري قرار دارد، آنتولوژي مي باشد. دردهه اخير اين واژه وكاربرد آن در حوزه فناوري اطلاعات و ارتباطات وارد شده و نقش مهمي در تبديل وب به وب معنايي دارد .

تفاوت آنتولوژي در فلسفه و كامپيوتر

- ✓ در فلسفه، از نظم و ترتيب ميان مفاهيم به آنتولوژي ميرسيم اما در علوم كامپيوتر، آنتولوژي را از روي ترتيبي كه خود براي

مفاهيم درنظر ميگيريم، استخراج ميكنيم.

- ✓ نگاه آنتولوژي در فسلفه نگاهي جامع و جها نشمول است، درحاليكه آنتولوژي در كامپيوتر داراي دامنه بسيار كوچكتري است.

آنتولوژي در وب

فرض كنيد كه مي خواهيد درمورد موضوعي با كسي صحبت نماييد. براي اينكه طرف مقابل، حرف شما را كامل و درست متوجه شود، احتياج است كه حوزه بحث كاملاً مشخص باشد. واژ ههاي مختلف درحوز ه هاي گوناگون، معاني يا تعبيرهاي متفاوتي دارند وحتي گاهي ممكن است كه با وجود مشخص بودن حوزه بحث، يك كلمه خاص در ذهن افراد مختلف داراي تفاو تهاي اندكي باشد. واضح است كه تنها راه رهايي از چنين وضعيتي، يك مجموعه واژگان مشترك بين افراد است. آنتولوژي در وب معنايي دقيقاًچنين كاربردي را دارد. در هر آنتولوژي، تمام موجوديتهاي يك حوزه به صورت كامل و با ذكر تمام ويژگيها فهرست مي شوند. بعد از اينكه موجوديتهاي حوزه را شناسايي كرديد، بايدارتباطات بين آنها را نيز بيان نماييد.

مرحله بعدي اين است كه تمام اطلاعات فوق را با يك فرمت خاص درون مستندات اينترنتي قرار داده و اطلاعات موجود در آن مستند را به آنتولوژي اتصال دهيد..

يك آنتولوژي، لغات و مفاهيمي را كه در تعريف و نمايش محدود ه اي از دانش به كار ميروند، تعيين كرده و بنابراين معاني را استاندارد ميكند.

آنتولوژي توسط مردم، پايگا ههاي داده و برنامه هاي كاربردي كه نياز به اشترا ك گذاري اطلاعات يك دامنه خاص را دارند، به كار برده مي شود. آنتولوژي در وب معنايي واژه ها و ارتباط بين آنها را در دامنه اي كه استفاده ميگردند، نشان مي دهد.

عناصر اصلي تشكيل دهنده آنتولوژي عبارتنداز:

1 - مفاهيم

2 - ارتباط بين مفاهيم

3 - خصوصيات آنها

به عبارت ديگر آنتولوژي ارتباط بين مفاهيم در اسناد وب و دنياي واقعي را مشخص ميكند كه با اين كار اسناد مربوطه توسط ماشينها قابل پردازش و فهم مي شوند و اشترا ك گذاري بين عاملها را تسهيل مينمايد. درواقع ميتوان گفت:

- ✓ Vocabulary+Structure=Taxonomy
- ✓ Taxonomy+Relationship و Constraints&Rules=Ontology

- ✓ Ontology+Instance=Knowledge
- ✓ واژگان + ساختار = طبقه بندي
- ✓ طبقه بندي + ارتباط و محدوديت ها و قوانين = هستي شناسي
- ✓ هستي شناسي + نمونه = دانش

5 مرحله در طراحي هستي شناسي ها :

1 - تعيين هدف و دامنه هستي شناسي؛

2 - طراحي هستي شناسي در يك فرآيند سه مرحله اي شامل:

- ✓ گردآوري هستي شناسي (تعيين و تعريف مفاهيم و روابط اصلي)؛
- ✓ كدگذاري هستي شناسي (به كار گرفتن واژ ههاي اصلي براي هستي شناسي (رده، موجوديت، رابطه)؛ انتخاب يك زبان

بازنمون؛ نوشتن كد)؛

را به عنوان زبان نشان هگذاري معنايي به منظور انتشار OWL كنسرسيوم وب جهاني در نوامبر 2002 ، زبان با تعريف كلا سها، نمونه ها و روابط به طور OWL . است RDF هستي شناسي هاي وب پيشنهاد كرد. اين زبان بر مبناي امكانات RDFS و XML،RDF نسبت به OWL . واضح و رسمي در توسعه و ساخت هستي شناسي ها به كار مي رود بيشتري براي بيان مفاهيم و معاني دارد و به دليل قابليت نمايش محتواي ميانكنش پذير رايانه ها در وب، برتر از ساير زبان ها است كه هر يك ويژگ يهاي خاص خود را OWL Full و OWL DL،OWL Lite است. اين زبان داراي سه زبان فرعي دارند و براي گروه خاصي از كاربران طراحي شده اند.

- ✓ يكپارچه سازي هستي شناسي هاي موجود؛

3 - ارزيابي هستي شناسي؛

4 - مستندسازي؛

5 - ارائه راهنماها و دستورالعمل هايي براي هر يك از مراحل قبل

آنتولوژي به عنوان ابزاري قدرتمند براي نمايش و بيان دانش مربوط به يك حوزه، در قالبي رسمي و قابل پردازش توسط ماشين مطرح است.

به كمك آن ميتوان ارتباط بين سيستمهاي ناهمگون را برقرار كرد و تعامل و ارتباط متقابل بين برنامه ها، ماشينها و سيستمهاي ناهمگون را بهبود بخشيد.

موتور جستجوي وب معنايي، براي جستجوي هستي شناسي ها، مدارك، واژه ها و داده هاي منتشر شده در وب .
اين موتور جستجو داراي سيستم جستجوي مدارك RDF, Html ميباشد

نمایه سازی معنایی پنهان

چکیده
اپن مقاله به معرفی و توصیف روش نمایه سازی معنایی پنهان (ال. اس. آی) مي پردازد که پکی از روش هاي نوین نمایه سازی خودکار است. ابتدا در مقدمه ای کوتاه به نمایه سازی و چالش هاي آن اشاره می شود سپس در بخش دودل هاي فضایی برداری که نمایه سازی معنایی پنهان پکی از گسترش هاي آن است، توصیف می شود. در بخش بعدی ضمن تشریح مفهوم نمایه سازی معنایی پنهان، کاربردها و مواردداستفاده از آن بیان می گردد و سپس مبانی ریاضی آن که روش آماری تجزیه مقادیر منفرداست با مثالی تشریح می شود. در بخش بعدی فرآیندکار نمایه سازی معنایی پنهان همراه با مثال توضیح داده می شود و در نهایت طرح ها وبرنامه هایی که هم اکنون در اپن زمینه اجرامی شوند معرفی و به بعضی پیشرفت هاي فناورانه موثر در بهبود عملکرد نمایه سازی معنایی پنهان اشاره می شود.

کلید واژه ها:
نمایه سازی معنایی، نمایه سازی پنهان، بازیابی اطلاعات، تجزیه مقادیر منفرد

امروزه نمایه سازی و به طور کلی حوزه بازیابی اطلاعات، به سبب تغییرات قابل توجه در محیط پیرامون خود متحول شده اند. می توان بعضی از این تغییرات را این گونه برشمرد گسترش قابل توجه دا منه ي بازیابی اطلاعات با ظهور چند رسانه ای ها، اینترنت، و اطلاعات جهانی ظهور پایگاه به داده و بعضی نظریه هاي جدیددر هوش مصنوعی، زبان شناسعي و ریاضیات وویکرد تحلیلی مباحث ریاضی و پیچیده به کار رفته در بازیابی اطلاعات طراحی نظام هاي بازیابی اطلاعات همانند نظام هاي پایگاه
داده ای رابطه ای، و بروز امکانات جدید ناشی از فناوری هاي نوینی چون وب مانند پالایشی اشتراکی (که با عنوان نظام هاي توصیه هاف یا شخصی شناخته می شود).
علاوه بر چالش هاي فوق باید به این نکته اصلي توجه داشت که در بازیابی اطلاعات مفهوم نسبتاً مبهم و گنگی به نام ربط وجود دارد که به روش هاي پیچیده تشخیص نیت کاربر و ماهیت مدرکي بستگی دارد. طبق گفته پاپادیمیتریو و همکارانش نظریه هاي اندکی در این زمینه ارائه شده است برای درك بیشتر این

چالش بررس متون کلاسعیکی چون آثار ریجسبرگن و سا لتون مفید خواهد بود
باید به یاد داشت که چالش هاي جدید، علاوه بر چالش هاي پیشین چون ذهنی بودن فرایند نمایه سازی است که ناشی از نظری بودن این عمل است. اصلي ترین روش تعیین ذهنی بودن نمایه سازی، بررسعي یکدستی آن هنگام تحلیلکار چند نمایه ساز از یک مدرك یا حتی یک نمایه ساز در زمان هاي مختلف است. همه این عوامل نشان دهنده پیچیدگی قابل توجه فرایند نمایه سازی است که در مقاله اندر سون و پرز-کاربالو به طور مفصل تشریح شده است .

وجود این چالش ها موجب شدکه پژوهشگران ازدهه 1950تاکنون به دنبال روش هاي نوین وکارآمد برای رویا رویی با چالش ها باشند و از دهه 1970 تاکنون برای رفع این مسئله اهتمام بسیار داشته اند. همان طور که پالگدرین و کیل ه لیو در پژوهش خود نشان دادند که بیش از 800 اثر پژوهشی طی سال هاي 1956 تا 2000 در مورد نمایه سازی خوکار، نیمه خودکار و رایانه ای نگاشته شده اند

با توجه به نکات فوق، نوعی توافق عمومی وجود دارد مبنی بر اینکه بازیابی سنتی اطلاعات نمی تواند جوابگوی این چالش ها باشد. نتایج بازیابی نظام هاي سنتي بازیابی اطلاعات به دو دلیل عمده ناکارا و غیر دقیق بود. اول اینکه در این نوع نظام ها مفهوم واحدی را می توان به روش هاي مختلف توصیف کرد. عبارت هاي پرسش کاربر ممکن است در مدرك مربوط وجود نداشته باشد.

دوم اینکه بیشتر کلمات بیشر از یک معنا دارنددر نتیجه مطابقت واژگانی عبارت هاي پرسشی کاربر ممکن است به بازیابی مدرکي نامربوط بینجامد
دانشمندان اصلی ترین علل محدودیت مدل ها و نظام هاي بازیابی اطلاعات را ابهام و گویانبودن واژه ها، ناکارآ مدی بازنمون مدارک مجموعه، و در سویی دیگر اغتشاش و عدم صراحت ا` پرسش هاي کاربر مي دانند. راه حل هاي گوناگونی برای این مسائل پیشنهاد شده است که بازیابی اطلاعات براساس هستی شناس یا استفاده از بازنمون معنایی مدارک و پرسش ها از مهم ترین راه حل ها هستند
در این مقاله سعي می شود نمایه سازی معنایی پنهان که نوعی بازنمون معنایی مدارک و نوع گسترش یافته مدل هاي بازیابی برداری است و از مباحث جبر خطی و ماتریسهاي ریاضی و فن تجزیه مقادیر منفرد استفاده می کند، معرفی و ابعاد مختلف استفاده از آن بیان شود

مدل فضا برداري
از آنجا که نمایه سازی معنایی پنهان یکی از راهکارهایی بود که برای رفع مشکلات مدل فضای برداری به وجود آمد، در این بغش سعي می شود خلاصه ای از این مدل و مشکلات آن بیان شود.

مدل فضایی برداری یکی از چند روش تشخیص تشابه میان دو مدرك است که در سال 1975به ومیله سا لتون گسترش یافت و چارچوپی تأثیرگذار و قدر تمند برای ذخیره ، تحلیل و ساختن مدركي است. این مدل ازابتدا به منظور بازیابی اطلاعات گسترش یافت و در نمایه سازی مدركي مبتنی بر فراوانی عبارت ها، به طور گسترده ای به کار رفت. سه مرحله این مدل عبارتند از نمایه سازی مدرك، وزن دهی عبارت، محاسبه ضریب تشابه:

۱. نمایه سازی مدرك: هر مدرك یا (پرسش) به صورت يك بردار در فضایی با ابعاد بزرگ ترترسیم مي شود. تمداد عبارت هاي منحصربه فرد مجمومه مدرك محاسبه می شود. واژگان غیرمهم در بردار مدرك حذف می شوند. در سیاهه واژگان غیر مجار که واژگان عمومی را در خود جای می دهد، برای حذف واژه هاي پرکار برد استفاده می شود (که در کل 40 تا 50درصددر کل واژگان مدرك حذف می شوند).

2. وزن دهی عبارت: وزن دهی برای نشان دادن میزان اهمیت عبارت ها در باز
نمون مدرک انجام مي شود. پیس فرض اغلب روش هاي وزن دهی چون فراوانی معکوس مدرک` این نکته است که اهمیت یک عبارت با افزایش میزان رخداد آن عبارت متناسب است. برای پیشگیری از بازیابی مدارک طولاني تر از شیوه هنجارسازی` استفاده می شود (زیرا با توجه به نکته فوق، احتمال بازیابی مدارک طولاني بیشتر از احتمال بازیابی مدارک کوتاه تر خواهد بود).

3 . محاسبه ضریب تشابه: تشابه میان دو مدرک (یا میان یک پرسش و یک مدرک) با فاصله میان بردارها در فضایی با ابعاد بزرگ تر تعیین می شود. بدین صورت که همپوشانی واژه ، نشان دهنده تشابه خواهد بود. شناخته شده ترین مقیاس تشابه، ضریب کسینوس است و تشابه میان دو مدرک را با کسینوس زاویه میان دو بردار نشان می دهند

در این مدل هنگام ورود یک پرسش توسط کاربر، آن پرسش مانند دیگر مدارک فرض می شود و به صورت یک بردار نشان داده می شود. نظام در فرایند بازیابی، موقعیت پرسش را نسبت به محل هر مدرک در فضای برداری مقایسه می کند و مدارک را با توجه به میزان تشابه با پرسش کاربر رتبه بندی می کند. پس به این ترتیب یا تعداد مدارکی را که بیشترین تشابه را با پرسش دارندبازیابی می کند یا همه مدارکی را که تا حدی از آستانه تشابه (که قبلاً معین شده است) بیشتر باشد، بازیابی می کند. اصلی ترین مزیت مدل فضای برداری، رتبه بندیکارآمد و دقیق مدارک مرتبط با توجه به میزان تشابه آنها با پرسش است، درصورتی که در روش هاي مبتنی بر مطابقت واژگانی یا هیچ طرحی برای رتبه بندی وجود ندارد یا اگر هم وجود داشته باشد فاقد امکان رتبه بندی موارد پیچیده هستند. چنانکه ممکن است به واژه ای که در اول عبارت پرسش ظاهر شده است رتبه ای بالاتر اختصاص داده شود .

کاربران در سیاهه مدارک بازیابی شده ، همیشه مدارک با بیشترین میزان تشابه را مطالعه می کنند به این دلیل که از نظر معنایی بیشتر به پرسش مربوط هستند. با وجود این سنجش میزان تشابه میان بردارها، به دستورالعملی برای چگونگی وزن دهی هر عبارت نمایه نیاز دارد.

روش هاي مختلفی برای این امر وجود دارد که توابع دود ویی ، فراوانی عبارت و لگاریتمی بیشتراز بقیه شناخته شده اند

البته باید توجه داشت که استفاده از این مدل مشکلاتی را نیز دارد. اصلی ترین مشکل مدل هاي فضای برداری، عدم مطابقت واژگان است. مدل هاي فضای برداری، با عبارت هاي نامشابه مانند اقلام نامرتبط برخورد می کنند. برای مثال رایانه و لپ تاپ` اگرچه عبارت هاي مرتبطی هستند اما مدل هاي فضای برداری از کشف چنین رابطه ای عاجز هستند. پس اگر میان پرسش و مدرک در مجموعه متون هیچ کلمه

مشترکی وجود نداشته باشد حتی اگر بعضی مدارک با پرسش مرتبط باشند، مقدار تشابه عملاً منجر خواهد بود و در نتیجه هیچ مدرکی بازیابی نخواهد شد. دومین اشکال این مدل ها در مجموعه مدارک بزرگ` پیس می آید که درصورت تشکیل ماتریس عبارت مدرک، ماتریس حاصل بزرگ و پراکنده خواهد بود که فضای ذخیره زیادی را می طلبدو مدت زمان پردازش و محاسبه آن نیز طولاني می شود. برای رفع این مشکلات از یکی از شاخه هاي این مدل با عنوان نمایه سازی معنایی پنهان استفاده می شود که در سال هاي اخیر رواج یافته است

به طور کلی می توان گفت که بازیابی اطلاعات سنتی با دو مشکل قدیمی متران ها (مانند مدارک حذف شده مربوط به اتومبیل به هنگام پرسش درباره ماشین) و چند معنایی ها (مثل بازیابی مدارک درباره اینترنت به هنگام پرسش درباره موج سواري) روبرواست. در این مدل هیچ تفاوتي میان crane به معنای پرنده ماهی خوار craneبه معنای جرثقیل نیست. برای مقابله با این دو مشکل و سایر مشکلات بازیابی سنتي اطلاعات، سعي می شود مدارک (و پرسش ها) نه با استفاده از خود عبارت ها (همان طور که در روش هاي برداری معمول است)، بلکه با استفاده از مفاهیم ضمنی (پنهان) آن عبارت ها، بازنمون شوند. این ساختار پنهان، نقشه ای ثابت میان عبارت ها و مفاهیم نیست بلکه به کل مدارک مجموعه و رابطه عبارت باکل آن بستگی دارد

نمایه سازي معنایي پنهان

نمایه سازي معنایي پنهان یکی از فنون نمایه سازي مفهومی است که برای غلبه بر مشکلات ناشي ازعدم مطابقت واژگان به وجود آمده است. همان طور که گفته شده در نظام هاي سنتی بازیابی، اطلاعات از مطابقت واژه به واژه عبارت هاي مدارك با پرسش کاربر بازیابی می شود، ولی شواهدناکارآ مدی این روش را نشان داده اند. همچنین از آنجا که معمولاً روش هاي متعددی برای بیان یک مفهوم وجود دارد (ترادف) ا احتمال دارد که میان عبارت ها و واژگان پرسش کاربر با بازنمون واژگانی مدرك هیچ اشتراکی وجود نداشته باشد. علاوه بر آن بیشتر واژه ها چندین معنا دارند(چندمعنایی) ` به طوری که نتیجه بازیابی با عبارت هاي پرسش کاربر، مدرکي نامربوطی را دربرخواهدگرفت. رزاریو معتقد است روشی که کاربر را قادر به بازیابی اطلاعات براساس مفهوم یا معنای یک مدرك خواهد نمود و مشکلات ذکرشده را نیز برطرف می کند، نمایه سازی معنایی پنهان است

تعاریف مختلفی از نمایه سازی معنایی پنهان ارائه شده است. رزاریو آن را این گونه تعریف می کند: "نمایه سازی معنایی پنهان، فنی است که پرسش ها و مدرکي را به فضایی با ابعاد معنایی پنهان وارد می سازد.پاپادیمیتریو و همکارانش آن را فنی برای بازیابی اطلاعات بر اساس تحلیل طیفی ماتریس عبارت مدرك تعریف می کنند که پیشرفت هاي تجربی پیشی از آن فاقد هرگونه پیشی بینی و توصیف محکم بود. به زعم آنها نمایه سازی معنایی پنهان نوعی روش بازیابی اطلاعات است که تلاش دارد تا ساختار معنایی پنهان مجموعه مدرکي را با استفاده از فنون برگرفته از جبر خطی کشف کند

نمایه سازی معنایی پنهان اولین بار توسط گروهی از پژوهشگران (دیر وستر و همکارانش) در بلکور در بازیابی اطلاعات به کار رفت (18) و نمایه سازی معنایی پنهان نامیده شد. پژوهشگران به این دلیل واژه "پنهان" را به آن می افزایند که عبارت هاي جدید که بازنمون اطلاعات معنایی هستند، مستقیماً از مدرکي یافت نمی شوند بلکه حاصل بررسي کل مجموعه مدرکي و استفاده از روش ریاضی خاصی با عنوان تجزیه مقادیر منفرد (اس. وی. دی) است. به عقیده چنین ساختار معنایی پنهان مدرك با توجه به الگوی استفاده از واژگان (با توجه به امکان انتخاب چندین واژه) شکل می گیرد. مدل نمایه سازی معنایی پنهان از فنون آماری خاصی برای نشان دادن ساختار پنهان معنایی و زدودن زواید ناشي از امکان انتخاب چند واژه به جای یک مفهوم که دیر وستر و همکارانش به طور مبسوط آن را توصیف کرده اند،

استفاده می کند و باکشف الگوی استفاده از واژه ها، این ساختار را آشکار می کند و باعث حذف نوفه (پارازیت) می شود (13).

پیشی فرض نمایه سازی معنایی پنهان این است که معمولاً کل محتوای معنایی یک سند چون پاراگراف، چکیده یا کل مدرک با مجموع معانی واژه هاي آن به طور تقریبی برابر است. یعنی بدین صورت که: معنای واژه اول - معنای واژه دوم - معنای واژه سوم-. . . معنای واژه » ام-معنای پاراگراف

همچنین می توان با احتساب هر متن به صورت یک معادله خطی و کل مجموعه مدارک به صورت نظامی از معادلات هم زمان، معنای پایدار بازنمون هاي واژه ها را از کل یک مجموعه مدرک بزرگ به دست آورد ( )

نمایه سازی معنایی پنهان با استفاده از بستر متن، مترادف(یعنی امکان انتخاب چند واژه به جای یک مفهوم) و چند معنایی ها (یعنی وجود چند معنا برای عبارتی واحد) را کنترل و بدین ترتیب از ریزش کاذب پیشگیری می کند. یک عبارت پنهان ممکن است مرتبط به یک مفهوم نمایان (مانند مفهوم تعامل انسان و رایانه) باشد که با چند کلیدواژه توصیف می شود و ترکیبی از چند واژه است

بنا به ادعای پژوهشگرانی چون دیروستر، دو میس، لندئر، فارناس و هارشمن، نمایه سازی معنایی پنهان شناخته شده ترین الگوریتم بازیابی اطلاعات است و برای مقاصد گوناگونی چون جستجو و بازیابی رد ه بندی و پالایش به کار می رود. نمایه سازی معنایی پنهان یکی از روش هاي فضایی برداری برای مدل سازی مدارک است و طبق نظر دیروستر و دیگران دو میس و کنتستاتیس و پتنگر معنای "پنهان" مجموعه مدارک را آشکار می سازد

پژوهش هاي دومیس نشان دادند که نمایه سازی معنایی پنهان با ترجیح مفهوم معنایی مدرک بر واژگان آن، مدل هاي فضایی برداری را بهبود می بخشد

در نمایه سازی معنایی پنهان برخلاف مدل هاي برداری مدارک که عبارت ها و واژه ها را مستقل فرض می کنند، سطوح

ختلفی از همبستگی ، وابستگی یا پیواستگی برای آنها درنظر گرفته می شود و این پیواستگی هاي بین عبارت ها با تشکیل مجموعه جدیدی از عبارت ها با استفاده از روش آماری تجزیه مقادیر منفرد مشخص می شوند همچنین در فضای معنایی پنهان، یک پرسش و یک مدرك، حتی درصورت نداشتن عبارت مشتری، می توانند تشابه کسینوس زیادی داشته باشند زیرا عبارت هاي آنها از نظر معنایی، مشابهت "مفهومی دارنددرصورتی که در مدل هاي برداری چنین چیزی امکان پذیر نیست

مهمترین نکته قوت نمایه سازی معنایی پنهان کار آمدی بازیابی مبتنی بر پرسش کاربر است که از طریق محاسبه ماتریس حاصل می شود. همان طور که با استفاده از این روش مدرکي مربوط، حتی در صورت عدم مطابقت واژگان محتوای دو مدرك با یکدیگر، بازیابی می شوند

به طور کلی می توان گفت که نمایه سازی معنایی پنهان یکی از فنون روبه رشد نمایه سازی مجموعه مدرکي بزرگ است که سعي دارد از یاد گیری آماری ماشین در تحلیل متون استفاده کند (33).

محققان مزیای مختلفی برای نمایه سازی معنایی پنهان برمی شمرند که در ادامه به چند مورد اشاره می شود:

چنگ، یکی از اصلي ترین مزایای نمایه سازی معنایی پنهان را استفاده از مفاهیم معنایی به جای تک تک واژه

ها در نمایه سازی می داند که بدین ترتیب مدرکي مربوط حتی در صورت نبوت واژه مشتری با عبارت پرسش بازیابی می شوند دیروستر نیز مزیت اصلي استفاده از بازنمون نمایه سازی معنایی پنهان راکار آمدی آن در رویا رویی بامدارکی می داندکه حاوی مترادف ها، چند معنایی ها، و عبارت هاي وابسته بهمدیگر است به نظر هازبندز، سیمون و دینگ نمایه سازی معنایی پنهان چون یکی از روش هاي گسترش پرسش است می توانأ جامعیت را بهبود بخشد.

موارد استتفاده کاربرد هاي تمایه سازی معنایي پنهان

پژوهشگران استفاده از نمایه سازی معنایي پنهان را در حوزه هاي مختلف ارزیابی کرده اند و آن را یکی از پرکار بردترین روش هاي نمایه سازی دانسته اند.

موارد استفاده از نمایه سازی معنایی پنهان عبارتنداز:

1. *بازیابی اطلاعات*: همان طور که گفته شد نمایه سازی معنایی پنهان اولین بار توسط گروهی از پژوهشگران در بلکور در بازیابی اطلاعات به کار رفت و به همین نام شناخته شد. این روش بهتر از روش هاي برداری استاندارد عمل می کند و حتی درصورت نبوت اشتراك واژگانی میان سوال و مدرك،کارایی خود را حفظ می کند. استفاده از نمایه سازی پنهان معنایی در بازیابی اطلاعات در مقالات متعددی توسط محققانی چون بری، دومیس و ابریان و دومیس هال و هازبندز، سیمون و دینگ تاکید شده است محققان دیگری چون اندو و لی` بارتل، کترل و بلو دومیس و ژا، مار کوس و سیمون نیز در پژوهش هاي خود نشان دادند که با استفاده از نمایه سازی معنایی پنهان جامعیت و مانعیت نظام بازیابی به طور قابل توجهی افزایش می یابد بشیری نیز از این روش در بازیابی متون فارسي استفاده کرده است
2. *بازخورد ربط:* بیشتر آزمون هایی که از نمایه سازی معنایی پنهان برای بازخورد ربط استفاده می کنند از روشی استفاده می کنند که در آن، حاصلجمع برداری مدارکی که توسط کاربران مربوط شنا سایر شده اند، جایگزین پرسش اولیه می شوند. پژوهش ها نشان داده اند که جایگزین سازی اولین مدرك مربوط با پرسش اولیه، عملکرد را درحد 33درصد و جایگزین سازی سه مدرك مربوط به آن عملکرد را درحد 67 درصد بهبود می بخشد. نمایه سازی معنایی پنهان این قابلیت را دارد که با استفاده از فنون گسترش پرسش حتی بدون استفاده از بازخورد ربط اقدام با شنا سایی مدرکي مربوط کند ولی با استفاده از اطلاعات حاصل از بازخورد ربط، بازدهی عملکرد را به طور قابل توجهی افزایش می دهد.

3. *پالایش اطلاعات:* استفاده از نمایه سازی معنایی پنهان در پالایش اطلاعات بسیار آسان است. ابتدا نمونه ای از مدرك با استفاده از ابزارهاي استاندارد نمایه سازی معنایی پنهان و تجزیه مقادیر منفرد تحلیل می شوند و بدین ترتیب علایق کاربر به صورت برداری با ابعاد کم نشان داده می شود و سپس مدرکي موجود

در مجموعه با آن بردار مطابقت داده شده و در صورت تشابه مدرك با آن برددر، به کاربر توصیه می شود. البته می توان با گذشت زمان با استفاده از روش هاي یاد گیری نظیر بازخورد ربط برای بهبود بازنمون بردارهاي علایق استفاده کرد (زیرا این علایق در طول زمان تغییر می یابند)

4 . *همایش ارزیابی:* بازیابی متني اخیراً از نمایه سازی معنایی پنهان در زمینه هاي پالایش و بازیابی اطلاعات همایش ارزیابی بازیابی متني استفاده می شود. پرسش هاي این برنامه بسیار طولاني و حاوی تومیغات مفصلی هستند که گا ه طول آنها به 50 کلمه نیز می رسد. پرسش هاي این برنامه، دارای ابزارهاي قوی تری نسبت به محاسن نمایه سازی معنایی پنهان یا سایر روش هایی که سعي در غنی سازی پرسش هاي کاربران دارند، است.

اصلي ترین چالش این مجموعه، گسترش ابزارهاي نمایه سازی معنایی پنهان برای کنترل مجموعه بود که این نتایج، کاملاً دلگرم کننده بودند. از آنجا که در زمان همایش هاي ارزیابی بازیابی متني محاسبه کل مدرکي مجموعه منطقی نبود، از نمونه ای در حدود 70هزار مدرك و 90 هزار عبارت استفاده شد. چنین ماتریسهاي عبارت مدرکی بسیار پراکنده هستند و فقط 0.001 تا 0.002 درصد ورودی ها را دربر می گیرند، برای مثال محاسبه دویستمین مقدار فردی بزرگ تر، به حدود 18ساعت زمان کار واحد پردازش مرکزی در یک پایگا ه سانس پارك با ده ایستگا ه کاری نیاز داشت. با وجود دشواری بسیار مقایسه تفصیلی نظام ها (به علت تفاوت هاي قابل توجه پیش پردازش، بازنمون و مطابقت)، عملکرد نمایه سازی معنایی پنهان بسیار خوب گزارش شد. استفاده از اطلاعات درباره مدرکي شناخته شده مربوط برای ایجاد برای هر پرسش، در پالایش بسیار سود مند بود. مزیت بازیافت 31درصدی تا حدی کمتر از میزان مشاهده شده در سایر آزمون هاي پالایش بود که به پرسش هاي اولیه این همایش نسبت داده می شود. در بازیابی نیز با استفاده از نمایه سازی معنایی پنهان در مقایسه با روش هاي برداری کلید واژه ای 16درصد بهبود مشاهده شد

5. *بازیابی در متون چند زبانه*: چون نمایه سازی معنایی پنهان از دستور زبان انگلیسي استفاده نمی کند می توان از آن در هر زبانی استفاده کرد. به علاوه می توان از آن در بازیابی از متونی که چند زبانه هستند استفاده نمود و پرسش هاي کاربران (در هر زبان موجود در متون) با مطابقت زبان پرسش با مدرکي هم زبان صورت می گیرد. لند وئر و لیتمن روشی را برای ایجاد فضایی مشترك که واژه هاي هر زبان در آن ارائه شوند با استفاده از نمایه سازی معنایی پنهان توصیف می کنند .در این نوع موارد ماتریس اولیه عبارت مدرك، با استفاده از مجموعه ای از چکیده هاي چندین (که در پژوهش آنهافرانسویو انگلیسي بود) تشکیل می شود و به هر چکیده به صورت ترکیبی از نسخه هاي انگلیسي وفرانسوي، نگاه می شود. تجزیه مقادیر منفرد کاهش یافته، ماتریس عبارت چکیده ترکیبی محاسبه می شود. فضای حاصل از چکیده هاي ترکیبی انگلیسي وفرانسوي تشکیل می شود. بدین ترتیب در این فضای کاهش یافته، واژه هاي انگلیسي وفرانسویکه در چکیده هاي ترکیبی مشابه ظاهر می شوند، در کنار همدیگر قرار خواهندگرفت. بعد از این تحلیل، چکیده هاي تک زبانی وارد می شوندیک چکیده فرآنسوی به سادگی در بردار حاصل جمع واژه هاي اجزای اصلي که قبلاً در فضای نمایه سازی معنایی پنهان وارد شده بودند قرار می گیرد. پرسش هاي انگلیسي یافرانسوي با چکیده هاي انگلیسي یافرانسوي مطابقت داده می شوند. در فضای نمایه سازی معنایی پنهان نیازی به ترجمه نیست. تجربه ها نشان داده اند که فضای چند زبانی کاملاً خودکار، حتی بهتر از فضایی با زبان واحد عمل می کند و نتایج بهتری عاید کاربران می کند. عملکرد بازیابی مدارك فرانسوي زبان با پرسش هاي به زبان انگلیسي (و برعکس) برابر با عملکردی است که ابتدا پرسش ها به فرانسوي ترجمه وسپس در پایگا ه داد ه منحصرافرانسوي جستجو شوند. این روش نتایج خوبی را در بازیابی چکیده هاي انگلیسي و اندیشه نگاشت هاي ژاپنی کنجي و ترجمه هاي چند زبانی (انگلیسي و یونانی) ارائه داد.

6. *مطابقت افراد به جای مدارك:* از دیگر کاربردهاي نمایه سازی معنایی پنهان یافتن افراد خبره دریک حوزه

خاص با استفاده از مقالات و آثار آنهاست. در این کاربرد، اشخاص براساس مقالاتی که نوشته اند معرفی می شوند. برای مثال در پژوهش فرناس و دیگران که بلکور ادوایز` نام دارد نظامی برای یافتن خبر گان محلی باتوجه به پرسش هاي کاربران طراحی شد. یک پرسش با جدید ترین مدارك مطابقت داده می شد و توصیف و مشخصات مربوط به پدیدآوران به عنوان مربوط ترین مطلب ارائه می شد .در موردی دیگر، از نمایه سازی معنایی پنهان برای تخصیص مقالات به افراد به منظور ارزیابی داوری آنها استفاده شد. برای انجام اینکار ابتدا صدها منتقد بر اساس متون تألیفیشان توصیف شدند که اینکار بنیان تحلیل نمایه سازی معنایی پنهان شد. صدها مقاله نیز براساس چکیده آنها نشان داده شدنأ و با نزدیک ترین و مرتبط ترین دا وران مطابقت داده شدند. استفاده از امکان تشابه یابی این برنامه برای تخصیص مقالات همایش ´`تعامل انسان رایانه´` به دا وران انجام شد. تحلیل هاي بعدی نشان داد که اینکار کاملأ خودکار، به خوبی کاری بود که توسط افراد خبره انجام شده بود.

ابرداده (فراداده)

Metadata

ابرداده

براي ابرداده تعاريف بسياري آمده كه اكثر آن‌ها بر كاربرد آن تكيه دارند و به عبارت ديگر، تعاريفي عملگرا به شمار مي‌آيند.

به طور خلاصه ابرداده را «داده‌اي براي داده» يا «داده‌اي درباره داده» تعريف مي‌كنند.

فراداده، مكررا "داده درباره داده" يا "اطلاعات درباره اطلاعات" تعريف شده است. به عبارت ديگر، فراداده، داده اي است كه منابع اطلاعاتي را توصيف مي كند.

اين تعريف گسترده، سطوح گوناگوني از توصيف (ساده تا پيچيده) را در بر مي گيرد: يادداشت توصيفي كوتاهي در مورد يک كتاب، توصيف غير رسمي را كه موتورهاي كاوش در مورد پيشينه هاي بازيابي شده ارائه مي دهند، يک پيشينه كتابشناختي چاپي (فهرستبرگه) يا الكترونيكي (پيشينه مارک)، چكيده ها و اصطلاحات نمايه اي، و حتي يک استناد كتابشناختي.

ابرداده

بطور كلي مي توان سه نوع ابرداده براي توصيف يك شيئ ديجيتال ايجاد كرد: ابرداده توصيفي، ساختاري، و اجرايي

- ابرداده توصيفي به صفاتي از عناصر اطلاعاتي مانند عناصر كتابشناختي نظير عنوان، نويسنده، ناشر و غيره، اشاره دارد.
- ابرداده ساختاري، ساختار و رابطه مجموعه اي از عناصر ديجيتال را توصيف مي

كند. به عنوان مثال، براي توليد ابرداده ساختاري يك كتاب، بايد شيوه آرايش صفحات، فهرست مندرجات و ارتباط بين بخش ها و فصل ها را ثبت كرد.

- ابرداده اجرايي، همه اطلاعاتي است كه در تمام دوره حياط يک شيئ ديجيتال، براي مديريت آن مورد نياز است و همه اطلاعات مورد نياز براي حفاظت از آن را نيز در بر خواهد داشت، مانند تاريخ ثبت، ساختار فايل، حق مؤلف و غيره.

ابرداده

- كاربردهاي مهم ابرداده:
  - سازماندهي منابع اطلاعاتي؛
  - توصيف منابع اطلاعاتي اعم از متن، تصاوير، فايلهاي صوتي و نظاير آنها؛
  - تحليل محتوا و نمايه سازي؛
  - تطبيق، اشتراك و يكپارچه سازي منابع اطلاعاتي ناهمگن؛
  - زمينه سازي براي استفاده مجدد از اطلاعات توزيع يافته در ساير كتابخانه هاي ديجيتال و محيط وب؛
  - ايجاد امكان دسترسي به اطلاعات دقيق و مرتبط توسط كاربران؛
  - مديريت دقيق تر بر حجم گسترده اي از اطلاعات در كتابخانه هاي ديجيتال

ابرداده

- يكي از تعاريف جامعي كه از ابرداده ارائه شده، تعريفي است كه «هينز» پس از ارائه تعاريف مختلف از آن، آورده است:
  - «ابرداده داده‌اي است كه محتوا، شكل يا خصوصيات يک ركورد داده‌اي يا يک منبع اطلاعاتي را توصيف مي‌كند. ابرداده را مي‌توان در توصيف منابع كاملاً ساختاريافته، يا اطلاعات ساختارنيافته از قبيل مدارک متني به كار گرفت. همچنين مي‌توان براي توصيف منابع الكترونيكي، داده‌هاي رقمي (شامل تصاوير رقمي) و مدارک چاپي از قبيل كتاب‌ها، مجلات و گزارش‌ها، مورد استفاده قرار داد. مي‌توان آن را در درون يك منبع اطلاعاتي (مانند منابع وب) جاي داد يا به طور جداگانه در

يک پايگاه اطلاعاتي نگهداري کرد» (Haynes 2004).

ابرداده

- مي توان فراداده را، "داده هاي ساختارمند درباره ديگر داده ها" دانست. بنابراين، فراداده عبارت است از مجموعه ساختارمندي از عناصر که منابع اطلاعاتي را به منظور شناسايي، کشف، و استفاده، توصيف مي نمايد.
- اگر چه اين اصطلاح، اصطلاحي نسبتا جديد در حوزه کتابداري و اطلاع رساني است اما، مفهوم توصيف داده ها يا منابع اطلاعاتي از زمان بوجود آمدن طرح هاي اوليه سازماندهي اطلاعات مشاهده مي شود.
- همه ابزارهاي کتابخانه اي که براي توصيف محتواي داده ها استفاده مي شوند را مي توان فراداده ناميد.
  - براي مثال، قواعد فهرستنويسي انگلو – امريکن ابزاري است که براي توصيف کتابشناختي و کمک به دسترسي سازمان يافته به داده ها و منابع الکترونيکي تدوين شده است .

- Data about data
  - Information about information
  - Structured data about data

تاريخچه فراداده

اصطلاح فراداده نخستين بار توسط جک اي. مايرز و اس. کي. کول جيان در سال 1969 به عنوان نام شرکتي تجاري (آمريکايي) براي ايجاد و توسعه محصولات مربوط به "فرا الگوها" به کار رفت و در يک جزوه معرفي محصول در سال 1972 چاپ شد. چند سال بعد، يعني در

سال 1986، علامت تجاري ®MEADATAبراي شركت فراداده ثبت گرديد.

اصطلاح فراداده بعدها بوسيله متخصصان علوم رايانه، آمار، پايگاه هاي اطلاعاتي، و جامعه كتابداري و اطلاع رساني به شكل هاي "فراداده"، "فرا داده"، و "فرا – داده" استفاده شد. اما كاربرد مكرر اين اصطلاح بيشتر به دهه 1990 و شكل گيري شبكه جهاني وب در سال 1993 مربوط مي شود.

انواع قالب هاي ماركي كه در دهه هاي 1960، 1970، و 1980 شكل گرفته را مي توان نخستين قالب هاي فراداده اي دانست. تاريخچه شكل گيري ساير قالب هاي فراداده اي به دهه 1990، به ويژه 1993-1996 باز مي گردد.

از شناخته شده ترين قالب هاي فراداده اي در دهه 1990 مي توان به طرح كدگذاري متن، خدمات مكان يابي اطلاعات دولتي، تبادل رايانه اي اطلاعات موزه ها، چهارچوب توصيف منبع، توصيف كدگذاري شده آرشيوي، و طرح فراداده اي هسته دوبلين اشاره كرد.

- با توجه به ويژگيهاي منابع الكترونيكي و محيط هايي كه اين منابع ايجاد كرده اند و يا در آنها بكار مي روند، نيز تنوع منابع و محصولات اطلاعاتي، ضعف ابزارهاي جستجوي اطلاعاتي (مانند موتورهاي كاوش)، تقاضاي هاي فزاينده استفاده كنندگان، توليد كنندگان و عرضه كنندگان منابع الكترونيكي و شبكه اي ، نياز به ابزارهايي جديد براي سازماندهي و يا مطابقت ابزارهاي سنتي با اين تحولات كاملا بديهي مي نمايد.
- اين همان جايي است كه بحث فراداده مطرح مي شود. بر همين اساس، فراداده ها مي توانند با افزوده شدن (جاسازي) و پيوند يافتن با منابع، باعث توصيف، شناسايي، مكان يابي، و مديريت منابع گردند .

كاركرد فراداده

در بسياري از بحث هاي مربوط به فراداده بويژه در حوزه مديريت اطلاعات (از ديدگاه علم كتابداري و اطلاع رساني)، تمايل به گروه بندي عناصر فراداده اي بر اساس كاركردهايي كه

اين عناصر پشتيباني مي كنند، وجود دارد.

نتايج اين بحث ها شناسايي انواع متفاوت فراداده يا گروه هاي فراداده اي است. در ذيل جدولي از مشهورترين انواع فراداده به همراه كاركردهاي آنها ارائه شده است:

مزايا و كاربردهاي فراداده

- اساس كاربرد فراداده، كاوش، بازيابي، دسترسي، كشف، مستند سازي، ارزيابي و انتخاب منابع الكترونيكي بويژه شبكه اي است كه باعث افزايش دقت بازيابي مي گردد. به تعبير دقيق تر مي توان مهمترين كاربردهاي فراداده را اين چنين توصيف كرد:

دو عبارتي كه معمولا با فراداده همراه هستند

- زبان نشانه گذاري گسترش پذير
- eXtensible Markup Language (XML)
- قالب توصيف منبع
- Resource Description Framework (RDF)
- زمانيكه برخي افراد درباره XMLو RDFصحبت مي كنند گويي كه اين دوآغازه خودشان قالب ابرداده اي هستند، اما اين آشوبي ميان قالب و محتوا است.
- درعمل XMLو RDFقالبهاي داده اي عمومي هستند كه مي توانند در موارد گوناگون استفاده شوند. خصوصا، از XMLاغلب به عنوان قالب مدرک استفاده مي شود و قالب كلي تري است كه از آن HTMLحاصل مي شود.

زبان نشانه گذاري گسترش پذير XML

- هرگاه شما فيلدهاي ركورد مارک را با تگ هايي مانند استفاده از "245" به معني "عنوان" در نظر بگيريد:
- a Hamlet, Prince of Denmark$ 245

- در این صورت XMLتنها روش دیگري براي تگ زدن یک قطعه داده است، اگرچه آن شامل گذاشتن یک تگ شروع و یک تگ پاياني (با یک "/" قبل از نام تگ) در اطراف هر عنصر داده اي است:
  - <title>Hamlet, Prince of Denmark</title>
- تگ ها مي تواند هر چیزي که شما تمایل دارید باشد، به شرط اینکه شما آنها را در یک قالب داده اي تعریف ساختار، از قبل تعیین کرده باشید. بنابراین اگر شما ترجیح مي دهید، تعریفتان مي تواند داراي هر کدام از تگ هاي زیر براي "عنوان" باشد.
  - Hamlet, Prince of Denmark</245><245>
    <ti>Hamlet, Prince of Denmark</ti>
- در اصل XML، مشابه تگ ها و فیلدهاي فرعي مارک، سلسله مراتبي است. مزایاي آن بیش از مارک21 است که مي تواند به اندازه اي که سطوح سلسله مراتبي لازم است را دارا باشد. برخلاف مارک21 که مي تواند داراي دو سطح تگ و فیلد فرعي باشد.

قالب توصیف منبع RDF

- RDF یک مرحله یا دو مرحله بعد از XMLاست.
- بر روابط میان عناصر داده اي تأکید دارد.
- همانطور که از نام RDFبرمي آید ایجاد یک مکانیزمي براي توصیف منابع و اسناد برروي اینترنت است و رابطه کلیدي آن توصیف مي باشد.
- RDFعنصر ضروري و مهمي از وب معنایي است که توسط ائتلاف شبکه جهانگستر وب براي افزودن عنصر معنایي و در چارچوب توصیف منابع براي اشتراک داده ها روي اینترنت ایجاد شده است.
- RDFپیچیده تر و نسبت به XMLکمتر استفاده شده است و هنوز مشخص نیست، آیا آن به عنوان یک زبان عمومي در توصیف جهاني وب موفق است.

برخي از استانداردهاي فراداده مهم

- مارک

- هسته دوبلين
- مودس
- مدس
- متز

مارک MARC

قالب مارک به منظور ماشين خوان و ماشين فهم نمودن پيشينه هاي كتابشناختي طراحي شده است.

قالب مارک در واقع قالبي است براي ثبت داده هاي فهرست نويسي و استفاده بعدي از آنها براي تبادل، جستجو، بازيابي، نمايش، اطلاعات كتابشناختي، و يا چاپ پيشينه هاي كتابشناختي.

مارک داراي قالب هاي مختلف داده اي است كه هر يک به منظور استفاده خاصي ايجاد شده اند

هسته دوبلين (Dublin core)

- استاندارد فراداده اي هسته دوبلين، طرحي بين المللي و ميان رشته اي است كه مجموعه عناصري ساده و كارآمد براي توصيف طيف گسترده اي از منابع اطلاعاتي شبكه اي ارائه مي دهد.
- پديد آمدن اين قالب فراداده اي مانند ديگر قالب هاي فراداده اي برخاسته از اهداف و نياز هايي بوده كه توسعه دهندگان و پديدآورندگان آن در نظر داشته اند.
- اين طرح استاندارد دو سطح دارد: ساده و ويژه (مقيد به توضيحگرها).
- سطح ساده شامل پانزده عنصر، و سطح ويژه شامل هفت عنصر (مخاطب، منشا (ريشه)، نگهدارنده حقوق) بيشتر يعني بيست و دو عنصر مي باشد.
- هسته دوبلين، همچنين داراي گروهي از توضيحگرها (پالايش هاي عنصر و طرح هاي كدگذاري عنصر) است كه عناصر را از لحاظ معناشناختي به منظور فرايند كشف منبع پالايش مي كنند.
- مباحث مربوط به معناشناختي عناصرهسته دوبلين توسط گروهي از متخصصان رشته هاي كتابداري و اطلاع رساني، علم رايانه، كدگذاري متن، جامعه موزه ها، و ديگر حوزه هاي مرتبط پرداخته شده است.

4-5. مجموعه عناصر قراردادهای هسته دوبلین
4-5-1. عناصر سطح ساده

نام عنصر		شرح عنصر
عنوان	Title	نامی که به منبع یا اثر داده می شود و اثر به وسیله آن شناخته می گردد
پدیدآورنده	Creator	شخص یا سازمانی که مسئولیت اصلی خلق محتوای اثر را بر عهده دارد
موضوع	Subject	موضوع محتوای منبع که معمولا از طریق کلیدواژه ها، عبارات کلیدی یا نشانه های رده بندی بیان می شود
توصیف	Description	شرح محتوای منبع که معمولا شامل مواردی چون چکیده، فهرست مطالب و ارجاع به نمایش گرافیکی محتوا می باشد
ناشر	Publisher	شخص حقیقی یا حقوقی (سازمان) که مسئولیت پشتیبانی (امور فنی چاپ) و توزیع منبع یا اثر را بر عهده دارد
همکار	Contributor	شخص حقیقی یا حقوقی که در تهیه محتوای منبع مشارکت دارد و سهم قابل ملاحظه ای نسبت به منبع دارد. نظیر: ویراستار
تاریخ	Date	عنصر تاریخ معمولا در مورد زمان تدوین، ایجاد، انتشار یا توزیع منبع، مورد استفاده قرار می گیرد
نوع منبع	Type	عنصر نوع، معمولا در مورد ماهیت یا گونه محتوای منبع به کار می رود

قالب	Format	شکل فیزیکی یا دیجیتالی منبع که معمولا شامل نوع رسانه یا ابعاد منبع است
شناسه گر	Identifier	یک رشته حرفی [یا دیجیتالی] که منحصرا منبع را مشخص کرده و ارجاع روشن و بدون ابهام به آن دارد. مثل نشانی سایت (URL)
زبان	Language	زبان محتوای فکری منبع، بهترین روش برای این منظور استفاده از عنصر زبانی دو حرفی برای کد زبان است. مثل en برای انگلیسی
ارتباط	Relation	ارجاع به متن مرتبط با اثر یا منبع
منبع اصلی	Source	ارجاع به ماخذی که اثر حاضر از آن اخذ شده است
پوشش	Coverage	پوشش، معمولا شامل موقعیت مکانی و فضایی (نام مکان یا نام جغرافیایی) و دوره زمانی (نشانه دوره یا دامنه زمانی) است
حقوق	Rights	اطلاعات مربوط به حقوق منبع یا اثر، این اطلاعات شامل حقوقی مالکیت معنوی، حق چاپ و نشر و سایر حقوق مالکیت است

عناصر سطح ویژه

مخاطب	Audience	طبقه ای از یک موجودیت که منبع برای او تهیه شده یا مفید است
منشا (ریشه)	Provenance	شرح هر تغییری در مالکیت و حفاظت منبع پس از ایجاد آن که برای صحت و اعتبار، یکپارچگی، و تفسیر منبع مهم قلمداد می شود
مالک حقوق	Rightsholder	شخص یا سازمانی که حقوق منبع را مالکیت یا مدیریت می کند
شیوه آموزشی	Instructional Method	فرایندی که برای تولید دانش صورت می‌گیرد، یا گرایش‌ها و مهارت‌هایی که منبع به منظور پشتیبانی آنها اختصاص یافته است؛ شامل: روش‌های ارائه مطالب آموزشی یا هدایت فعالیت‌های مرتبط (مانند الگو های تعامل یادگیرنده-یادگیرنده؛ یادگیرنده - آموزنده و ...)
شیوه گسترش	Accrual method	روش افزوده شدن آثار به یک مجموعه. بهترین روش توصیه شده، استفاده از واژگان‌های کنترل شده است.
تناوب گسترش	Accrual periodicity	بسامد افزوده شدن آثار به یک مجموعه. بهترین روش توصیه شده، استفاده از واژگان‌های کنترل شده است.
خط مشی گسترش	Accrual Policy	خط مشی افزوده شدن آثار به یک مجموعه (خط مشی مجموعه سازی). بهترین روش توصیه شده، استفاده از واژگان‌های کنترل شده است.

یک پیشینه فراداده اي در قالب استاندارد هسته دوبلین و در بستر نحوي "زبان نشانه گذاري فرامتن (HTML)"

<link rel="schema.DC" href="http://purl.org/dc/elements/1.1/" />

<link rel="schema.DCTERMS"

href="http://purl.org/dc/terms/" />

<meta name="DC.title" lang="فارسي" content="پژوهشگاه علوم و فرهنگ اسلامي </ "

<meta name="DC.creator" content="پژوهشگاه علوم و فرهنگ اسلامي </ "

<meta name="DC.subject" lang="فارسي" content="اخبار علمي، دفتر تبليغات اسلامي حوزه علميه قم - فعاليتهاي پژوهشي،وب سايتها </ "

<meta name="DC.description" lang="فارسي" content="در سال

1380 همزمان با برنامه‌ريزي جامع دفتر تبليغات اسلامي و تدوين بيانيه مأموريت آن، معاونت پژوهشي نيز به بازنگري و اصلاح ساختار، اهداف و خطمشي‌هاي خود پرداخت و سرانجام، چشم‌انداز و سياست‌هاي كلان پژوهشي و همچنين اهداف و برنامه‌هاي پژوهشكده‌ها و گروه‌هاي پژوهشي خود را تنظيم و تدوين كرد. در ساختار جديد دفتر مقرر گرديد كه كليه مراكز پژوهشي در قالب يك پژوهشگاه جامع سازمان‌دهي شود. پس از چند سال پي‌گيري، سرانجام با تصويب چهار پژوهشكده در شوراي عالي گسترش، مجوز قطعي »پژوهشگاه علوم و فرهنگ اسلامي» از وزارت علوم، تحقيقات و فناوري گرفته شد. هم‌اينك پي‌گيري اخذ مجوز رسمي براي «مركز فرهنگ و معارف قرآن» و «پژوهشكده علوم اجتماعي واقتصادي» در دست اقدام است </ "

"=meta name="DC.publisher" contentدفتر تبليغات اسلامي حوزه علميه قم </ "

"=meta name="DC.contributor" contentمركز اطلاعات و مدارك اسلامي </ "

<meta name="DC.date" scheme="DCTERMS.W3CDTF"
content="2007-04-25" />
<meta name="DC.type" scheme="DCTERMS.DCMIType"
content="Text" />
<meta name="DC.format" content="text/html" />
<meta name="DC.format" content="54591 bytes" />
<meta name="DC.identifier" scheme="DCTERMS.URI"
content="http://www.isca.ac.ir" />
<meta name="DC.language" scheme="DCTERMS.URI"
"=contentفارسي </ "
<meta name="DC.relation" scheme="DCTERMS.URI"
"=contentمركز اطلاعات و مدارك اسلامي "http://www.islamicdoc.org
/>

<meta name="DC.rights" scheme="DCTERMS.URI"
content="كليه حقوق وب سايت متعلق به دفتر تبليغات اسلامي حوزه علميه قم
مي باشد </ "

مارک در زبان نشانه گذاري گسترش پذير (MARCXML)
"دفتر استانداردهاي مارک و توسعه شبکه" کتابخانه کنگره، در حال گسترش چارچوبي براي پياده سازي داده-هاي مارک در محيط "زبان نشانه گذاري گسترش-پذيرXML))" است. انعطاف-پذيري و گسترش پذيري اين چارچوب، امکان پاسخگويي به نيازهاي مختلف و خاص را براي کاربران فراهم مي-کند. اين چارچوب خود شامل طرح-ها، نرم-افزارها، و الگوهاي از پيش تعريف شده-اي مي-شود. در ذيل به مهمترين آنها اشاره مي گردد:

توصيف آرشيوي رمزگذاري شده EAD
استاندارد ابرداده اي که توسط "دفتر توسعه شبکه و استانداردهاي مارک کتابخانه کنگره" براي جامعه آرشيويست هاي آمريکا تهيه شده است.
باعث مي شود که اطلاعات آرشيوي به صورت ماشيني ذخيره، پردازش، بازيابي و قابل انتقال شوند.
مبتني بر زبان ايکس ام ال است.
www.lcweb.loc.gov/ead

طرح فراداده اي توصيف شي (MODS)

- اين طرح براي مجموعه عناصر کتابشناختي که با اهداف گوناگون، به خصوص کاربردهاي کتابخانه اي استفاده مي شوند ، تهيه شده است.
- به عنوان يک طرح "زبان نشانه گذاري گسترش-پذير"، امکان انتقال داده هاي کتابشناختي گزيده از پيشينه هاي مارک موجود، نيز توانايي ايجاد پيشينه-هاي توصيف براي منابع جديد را فراهم مي آورد.
- اين طرح بخشي از مناطق (فيلدهاي) مارک 21 را شامل شده و از برچسب هاي زبان-

پايه بيش از برچسب هاي عددي استفاده مي كند.

طرح فراداده اي توصيف شي (MODS)

طرح مذكور با ساير طرح ها و ابزارهاي "مارک در قالب زبان نشانه گذاري گسترش پذير (MARCXML)" كه از طريق وب سايت مارک دسترس پذيرند، سازگار است و به طور ويژه از فهرست نويسي منابع الكترونيكي پشتيباني مي نمايد.

اين طرح بوسيله "دفتر استانداردهاي مارک و توسعه شبكه" و با همكاري گروهي از كاربران علاقه-مند به مارک در سال 2002 گسترش يافت.

قالب مارک استانداردي پيچيده براي رمزگذاري اطلاعات كتابشناختي است.

در محيط شبكه اي در شرايطي كه ابرداده توصيفي مي تواند بر روي نظامها ارسال شود و مي تواند درون يا همراه با انواع ديگر ابرداده باشد، استفاده از ركوردهاي مارک با اين هدف، ايده آل به نظر خواهد رسيد.

با اين وجود، گنجاندن ركوردهاي مارک درون ابرداده، استفاده از ساختار داده اي XMLرا ضروري ساخته است، درحاليكه مارک يک ركورد XMLنيست.

كتابخانه كنگره روشي را براي تبديل ركورد مارک به XMLايجاد كرده است، اما آن با اقبال زيادي روبرو نشد و احتمالا به اين دليل كه ركورد مارک بزرگتر و جزئي تر از اكثر نظامهاي مورد نياز است، و استفاده اش از تگ هاي عددي و كدهاي فيلد فرعي موجب شده درک آن را بدون آموزش قابل توجهي مشكل سازد.

چيزي كه مورد نياز بود نسخه اي خوشايندتر و باكيفيت تر از مارک بود تا بتواند عناصركليدي داده اي از ركورد مارک را قبول نمايد و آنها را به يک قالب آسان و قابل درک XMLارسال نمايد.

بنابراين استاندارد توصيف ابرداده اي شي گرا (مودس) متولد شد.

مودس از تگ هاي قابل درک بشري بجاي تگ هاي سه رقمي و كدهاي فيلدفرعي مارک (يعني "عنوان" بجاي "245") استفاده مي كند.

بيشتر عناصر داده اي فيلد ثابت، به استثناي كدهاي قالب فيزيكي (از007) و بيشتر كدهاي نوع (از 008) را ناديده مي گيرد.

برخي قابليت ها و نوآوري ها را معرفي مي كند.

مودس يک ساختار نام گذاري شده "نام" را تعيين مي کند که فيلدها و فيلدهاي فرعي براي اسامي شخصي، سازماني و براي کنفرانسها را ارائه مي نمايد. اين ساختار مي تواند در هر جايي که نامها ظاهر خواهند شد، يا به شکل سرشناسه، شناسه هاي افزوده يا موضوعات استفاده شوند.

نمونه اي از طرح فراداده اي توصيف شي (MODS)

name type="personal">

<namePart>Shakespeare, William</namePart>

<namePart type="date">1564-1616</namePart>

</name>

مي تواند به عنوان يک فيلد نويسنده استفاده شود يا آن مي تواند قسمتي از يک سرعنوان موضوعي باشد

<subject authority=‘‘lcsh’’>

<name type=‘‘personal’’>

<namePart>Shakespeare, William</namePart>

<namePart type=‘‘date’’>1564–1616</namePart>

</name>

<topic>Bibliography</topic>

<topic>Periodicals</topic>

</subject>

طرح فراداده-اي توصيف مستند (MADS)

طرح فراداده-اي توصيف مستند" براي مجموعه عناصر مستندي که در تهيه پيشينه-هاي فراداده-اي به منظور توصيف افراد حقيقي و حقوقي، رويدادهاي مهم، و شناسه-هاي موضوعي و جغرافيايي، و... بکار مي-روند، و با هدف پياده سازي قالب مستندات مارک 21 در بستر نحوي "زبان نشانه گذاري گسترش-پذير" و استفاده از مزاياي اين محيط توسعه يافته است.

اين طرح نيز محصول "دفتر استانداردهاي مارک و توسعه شبکه" و برخي متخصصان علاقه-مند به قالب مارک مي-باشد.

استاندارد انتقال و کدگذاري فراداده-ها (METS)

پروژه MOA2، در زمينه تهيه يک قالب کدگذاري براي فراداده-هاي توصيفي، مديريتي، و ساختاري براي منابع متني و تصويري به منظور پاسخگويي به نيازهاي سازماندهي کتابخانه-هاي ديجيتالي فعاليت مي-کند. در همين راستا، "استاندارد انتقال و کدگذاري فراداده-ها" که محصول "فدراسيون کتابخانه-هاي ديجيتالي" است براي تحقق اهداف پروژه MOA2، با ارائه يک قالب مدرک در بستر نحوي "زبان نشانه گذاري گسترش-پذير" به منظور ايجاد فراداده-هاي ضروري براي مديريت اشياي محتوايي کتابخانه ديجيتالي يک سازمان و تبادل اين اشياي محتوايي ميان چند سازمان و کاربران آنها، تلاش مي -کند. بسته به نوع استفاده، اين استاندارد مي-تواند در بسته-هاي گوناگوني بکار گرفته شود.

مديريت کتابخانه هاي متشکل از شيء هاي ديجيتالي نيازمند مديريت ابرداه هاي آن شيء هاست.

ابرداده هاي ضروري براي مديريت موفق و استفاده از شيء هاي ديجيتالي، گسترده تر و متفاوت تر از ابرداده هاي مورد استفاده در مديريت مجموعه آثار چاپي و ساير منابع فيزيکي مي باشد.

در فهرستهاي کتابخانه ها ابرداده توصيفي هر کتاب ثبت مي شود. اگر کتابخانه در ثبت دقيق ابرداده ساختاري سازمان منتشردهنده کتاب ناموفق باشد، به هم پيوستگي صفحه هاي کتاب مذکور از بين نخواهد رفت. همچنين اگر کتابخانه در بيان نحوه چاپ کتاب و استفاده از ابزارهاي ويژه دقت نکرده باشد، در بازيابي کتاب توسط محققان آثار منفي در پي نخواهد داشت

اما اين مسائل درخصوص نسخه ديجيتالي همان کتاب صحيح نيست. بدون ابرداده ساختاري، تصوير صفحه يا فايل متني به وجود آورنده اثر ديجيتالي کمتر مورد استفاده قرار خواهد گرفت.

براي اهداف مديريت داخلي، کتابخانه بايد ابرداده فني مناسب جهت روزآمدسازي و انتقال متناوب داده و اطمينان از ماندگاري منابع باارزش دسترسي داشته باشد.

اين قالب استاندارد با فراهم کردن يک قالب کدگذاري براي ابرداده توصيفي، مديريتي و ساختاري جهت آثار متني و تصويري به حل اين مسئله کمک کند. نتيجه و حاصل از اينکار، مدارک با قالب XMLمي باشد که از اين طريق کدگذاري ابرداده هاي ضروري جهت مديريت شيء هاي ديجيتالي در مخزن و نيز تبادل چنين شيء هايي بين مخزنها (يا بين مخزنها و کاربران آنها) فراهم مي شود(6).

با اين اوصاف، مدرک ابرداده اي وجود دارد که هدف "توصيف" به مفهوم فهرستنويسي نيست. يک مثال از قالب ابرداده اي که در حال استفاده توسط کتابخانه ها و آرشيوهاي ديجيتالي است، استاندارد رمزگذاري و انتقال ابرداده (متس) مي باشد. از متس به عنوان "يک روکش" براي حفظ فايلها از

يكديگر استفاده مي شود تابا همديگر مخلوط نشوندو يک شيء ديجيتالي ايجاد شود. برخلاف جلدكتاب، مدارک ديجيتالي اغلب از تعدادي فايلهاي جداگانه ارائه شده در صفحات يا واحدهاي ديگر ساخته مي شود. و بر خلاف كتاب فيزيكي، جلد يا عنوان صفحه قابل رويت نيست، نه مي توان از انگشت شست براي يافتن صفحه اي خاص در كتاب استفاده نمود. متس را به عنوان صحافي، جلد و رديابي يک گروه از فايلهاي ديجيتالي تصور كنيد (3).

همچنين متس شامل اطلاعات فني است كه نياز به كنترل و شناخت اين فايلها خواهد داشت، مانند قالبهاي فايل، فناوري استفاده شده در پويش، و تغييرشكل ديجيتالي و فشردگي كه برروي فايلها استفاده شده است. چيزي كه متس تعيين نمي كند ابرداده توصيفي است. درعوض امكان ايجاد ركوردهاي متس براي درج در هر ابرداده توصيفي وجود دارد. اين ويژگي مهم، دنياي ابرداده را توصيف مي كند، كه ما آن را در نمونه كار كريپتوكامانز مشاهده نموديم: ابرداده مي تواند دوباره استفاده شود بجاي اينكه دوباره اختراع شود. معمولا ابرداده توصيفي در دابلين كور يا در مودس شامل ركوردهاي متس هستند

فراداده هاي مبتني بر XML براي تصاوير (MIX)

دفتر استانداردهاي مارک و توسعه شبكه، با همكاري كميته استانداردهاي فراداده اي فني براي تصاوير ثابت وابسته به "سازمان استانداردهاي اطلاعات ملي (NISO)"، و ديگر متخصصان علاقه مند، در حال گسترش يكه فرانما XML براي مجموعه اي از عناصر داده اي فني مورد نياز براي مديريت مجموعه تصاوير ديجيتالي هستند. اين فرانما، قالبي براي مبادله و/يا ذخيره سازي داده هاي معين در "فرهنگ داده ها- فراداده هاي فني براي تصاوير ثابت ديجيتالي (ANSI/NISO Z39.87-2006) " فراهم مي نمايند. فرانماي ياد شده، در حال حاضر به "فراداده مبتني بر XML براي تصاوير" مربوط به "سازمان استانداردهاي اطلاعات ملي (NISO)" ارجاع مي دهد. MIX با استفاده از زبان فرانماي XML مربوط به كنسرسيوم وب بيان مي گردد. و بوسيله دفتر استانداردهاي مارک و توسعه شبكه به سفارش "سازمان استانداردهاي اطلاعات ملي (NISO)"، با دروندادي از كاربران نگهداري مي شود.

ابرداده براي آرشيو الكترونيكي

مديريت ركوردها از طريق تنظيم منابع و فراهمكردن بررسي خلاصه مذاكرات، نقش مهمي در مستندسازي تصميمات دولتي ايفا مي‌كند. مديريت ركوردها شامل مديريت فايل‌هاي فيزيكي و

ركوردهاي الكترونيكي مي‌باشد؛ هرچند كه آن‌ها محمل رسانه‌اي متفاوتي داشته باشند، اما اكثر فرايندهاي مشابه بر روي هر نوع ركوردي اعمال مي‌شود. مديريت خوب ركوردها به ايجاد ركوردهاي درست و صحيح بستگي دارد. اين قسمت بر كاربرد ابرداده در مديريت چرخه عمر ركوردها تأكيد مي‌كند.

زماني كه ركورد ايجاد يا دريافت مي‌شود، به ثبت مي‌رسد يا با يك نام معين ذخيره مي‌شود. در طول فرايند فراهم‌آوري، ابرداده‌هاي مربوط به ركورد كه شامل تاريخ ايجاد، مالك، تاريخ بازبيني و رده امنيتي آن است، ايجاد مي‌گردند.

تعدادي از عناصر داده‌اي كه اختصاصاً براي مديريت ركوردها طراحي شده‌اند، در پروفايل‌هاي كاربردي ابرداده براساس «دوبلين كور»[7] تعريف شده‌اند. مثلاً «آرشيو ملي انگلستان» به عنوان يك سازمان موفق در اداره ركوردهاي عمومي و دست‌نوشته‌هاي تاريخي، گروهي از عناصر داده‌اي را اختصاصاً براي مديريت ركوردها در دولت مركزي تعريف كرده است. اما اين عناصر را مي‌توان در يك پروفايل كاربردي عمومي‌تر نيز به كار بست (Public Records Office 2002). اين عناصر داده‌اي عبارت‌اند از: Subject

- Title.
- Identifier.
- Record Type.
- Addressee.
- Date.
- Rights.
- Location.
- Language.
- Description.
- Mandate.
- Preservation.
- Aggregation.
- Relation.
- Creator.

- Digital Signature

موتورهاي جستجو

به طور عمومي به برنامه‌اي گفته مي‌شود كه كلمات كليدي را در يک سند يا بانک اطلاعاتي جستجو مي‌كند.

در اينترنت به برنامه‌اي گفته مي‌شود كه كلمات كليدي موجود در فايل‌ها و سندهاي وب جهاني، گروه‌هاي خبري، منوهاي گوفر و آرشيوهاي FTPرا جستجو مي‌كند.

مهمترين مرجع الكترونيک امروزي هستند

برخي از جويشگرها براي تنها يک وب‌گاه(پايگاه وب) اينترنت به كار برده مي‌شوند و در اصل جويشگري اختصاصي آن وب‌گاه هستند و تنها محتويات همان وب‌گاه را جستجو مي‌كنند.

برخي ديگر نيز ممكن است با استفاده از SPIDERها محتويات وب‌گاه‌هاي زيادي را پيمايش كرده و چكيده‌اي از آن را در يک پايگاه اطلاعاتي به شكل شاخص‌گذاري‌شده نگهداري مي‌كنند.

كاربران سپس مي‌توانند با جستجو كردن در اين پايگاه داده به پايگاه وبي كه اطلاعات موردنظر آن‌ها را در خود دارد پي ببرند.

جويشگرها به دو دسته كلي تقسيم مي‌شوند.

جويشگرهاي پيمايشي (خودكار) و فهرست‌هاي تكميل‌دستي (غير خودكار).

هر كدام از آن‌ها براي تكميل فهرست خود از روش‌هاي متفاوتي استفاده مي‌كنند

البته لازم به ذكر است كه گونه‌اي جديد از جويشگرها تحت عنوان «ابرجويشگر» (Meta Search Engines)نيز وجود دارد كه در ادامه به توضيح هر يک از اين موارد خواهيم

پرداخت :

جويشگرهاي پيمايشي

جويشگرهاي پيمايشي (Crawler-Based Search Enginesمانند گوگل فهرست خود را بصورت خودكار تشكيل مي‌دهند.

آنها وب را پيمايش كرده، اطلاعاتي را ذخيره مي‌كنند، سپس كاربران از ميان اين اطلاعات ذخيره شده، آنچه را كه مي‌خواهند جستجو مي‌كنند.

اگر شما در صفحه وب خود تغييراتي را اعمال نماييد، جويشگرهاي پيمايشي آن‌ها را به طور خودكار مي‌يابند و سپس اين تغييرات در فهرست‌ها اعمال خواهد شد.

عنوان، متن و ديگر عناصر صفحه، همگي در اين فهرست قرار خواهند گرفت.

جويشگرهاي پيمايشي

وجه مشخصه اين گروه از جويشگرها وجود نرم‌افزار موسوم به SPIDERدر آن‌هاست.

اين شبه نرم‌افزار كوچک بصورت خودكار به كاوش در شبكه جهاني پرداخته و از پايگاه‌هاي وب يادداشت‌برداري و فهرست‌برداري مي‌كند سپس اين اطلاعات را براي تجزيه و تحليل و طبقه‌بندي به بانک اطلاعاتي جويشگر تحويل مي‌دهد.

فهرست‌هاي دست‌نويس شده

فهرست‌هاي دست‌نويس‌شده يا Human-Powered (Directories)مانند فهرست بازي Open (Directory) وابسته به كاربراني است كه آن را تكميل مي‌كنند.

شما صفحه مورد نظر را به همراه توضيحي كوتاه در فهرست ثبت مي‌كنيد يا اين كار توسط ويراستارهايي كه براي آن فهرست در نظر گرفته شده، انجام مي‌شود.

عمل جستجو در اين حالت تنها بر روي توضيحات ثبت شده صورت مي‌گيرد و در صورت تغيير روي صفحه وب، روي فهرست تغييري به وجود نخواهد آورد.

فهرست‌هاي دست‌نويس شده

چيزهايي كه براي بهبود يک فهرست‌بندي در يک جويشگر مفيد هستند، تأثيري بر بهبود فهرست‌بندي يک دايركتوري ندارند.

تنها استثناء اين است كه يک سايت خوب با پايگاه داده‌اي با محتواي خوب شانس بيشتري نسبت به يک سايت با پايگاه داده ضعيف دارد.

البته در مورد جويشگرهاي مشهور مانند گوگل و ياهو، يک مولفه ديگر هم براي بهبود فهرست‌بندي وجود دارد كه كمک مالي (يا به اصطلاح اسپانسر) است

يعني وب‌گاه‌هايي كه مايل به بهبود مكان وب‌گاه خود در فهرست بندي هستند، مي‌توانند با پرداخت پول به اين جويشگرها به هدف خويش برسند.

جويشگرهاي تركيبي با نتايج مختلف

به موتورهايي گفته مي‌شود كه هر دو حالت را در كنار هم نمايش مي‌دهند.

غالباً، يک جويشگر تركيبي در صورت نمايش نتيجه جستجو از هر يک از دسته‌هاي فوق، نتايج حاصل از دسته ديگر را هم مورد توجه قرار مي‌دهد.

مثلاً جويشگر ام.اس.ان (MSN)بيشتر نتايج حاصل از فهرست‌هاي تكميل‌دستي را نشان مي‌دهد اما در كنار آن نيم نگاهي هم به نتايج حاصل از جستجوي پيمايشي دارد.

ابرجويشگرها

اين گونه جديد از جويشگرها كه قدمت چنداني نيز ندارند، بصورت همزمان از چندين جويشگر براي كاوش در شبكه براي كليد واژه مورد نظر استفاده مي‌كنند.

بدين معني كه اين جويشگر عبارت مورد نظر شما را در چندين جويشگرِ ديگر جستجو كرده و نتايج آنها را با هم تركيب كرده و يک نتيجه كلي به شما ارائه مي‌دهد.

به‌عنوان مثال جويشگر داگ پايل از نتايج حاصل از موتورهاي Google - Yahoo - MSN - ASKاستفاده كرده و نتيجه حاصله را به شما ارائه مي‌دهد.

نوجويشگرها

اين گونه از جويشگرها، نسل جديد و متفاوتي از جويشگرهاي گذشته هستند.

امكان ثبت جستجو و مدل‌سازي فعاليت‌هاي كاربر و ارائهٔ نتايج جديد به‌كاربر، به‌صورت متفاوت و تفكيک شده، از امكانات نوجويشگرها است

بررسي يک جويشگر پيمايشي

جويشگرهاي پيمايشي شامل سه عنصر اصلي هستند.

اولي در اصطلاح عنکبوت ((Spiderاست که پيمايشگر ((Crawlerهم ناميده مي‌شود.

پيمايشگر همين که به يک صفحه مي‌رسد، آن را مي‌خواند و سپس پيوندهاي آن به صفحات ديگر را دنبال مي‌نمايد.

اين چيزيست که براي يک سايت پيمايش‌شده ((Crawledاتفاق افتاده‌است.

پيمايشگر با يک روال منظم، مثلاً يک يا دو بار در ماه به سايت مراجعه مي‌کند تا تغييرات موجود در آن را بيابد.

هر چيزي که پيمايشگر بيابد به عنصر دوم يک جويشگر يعني فهرست انتقال پيدا مي‌کند.

فهرست اغلب به کاتالوگي بزرگ اطلاق مي‌شود که شامل ليستي از آنچه‌است که پيمايشگر يافته‌است.

مانند کتاب عظيمي که فهرستي را از آنچه پيمايشگرها از صفحات وب يافته‌اند، شامل شده‌است.

هرگاه سايتي دچار تغيير شود، اين فهرست نيز به روز خواهد شد.

از زماني که تغييري در صفحه‌اي از سايت ايجاد شده تا هنگامي که آن تغيير در فهرست جويشگر ثبت شود مدت زماني طول خواهد کشيد. پس ممکن است که يک سايت پيمايش‌شده باشد اما فهرست‌شده نباشد.

تا زماني که اين فهرست‌بندي براي آن تغيير ثبت نشده باشد، نمي‌توان انتظار داشت که در نتايج جستجو آن تغيير را ببينيم.

نرم‌افزار جويشگر، سومين عنصر يک جويشگر است

به برنامه‌اي اطلاق مي‌شود که به صورت هوشمندانه‌اي داده‌هاي موجود در فهرست را دسته‌بندي کرده و آن‌ها را بر اساس اهميت طبقه‌بندي مي‌کند تا نتيجه جستجو با کلمه‌هاي درخواست شده هر چه بيشتر منطبق و مربوط باشد.

رتبه‌بندي صفحات وب توسط جويشگرها

وقتي شما از جويشگرهاي پيمايشي چيزي را براي جستجو درخواست مي‌نماييد، تقريباً بلافاصله اين جستجو از ميان ميليون‌ها صفحه صورت گرفته و مرتب مي‌شود بطوريكه مربوطترين آنها نسبت به موضوع مورد درخواست شما رتبه بالاتري را احراز نمايد.

البته بايد در نظر داشته باشيد كه جويشگرها همواره نتايج درستي را به شما ارائه نخواهند داد و مسلماً صفحات نامربوطي را هم در نتيجه جستجو دريافت مي‌كنيد و گاهي اوقات مجبور هستيد كه جستجوي دقيقتري را براي آنچه مي‌خواهيد انجام دهيد اما جويشگرها كار حيرت‌انگيز ديگري نيز انجام مي‌دهند.

فرض كنيد كه شما به يک كتابدار مراجعه مي‌كنيد و از وي درباره «سفر» كتابي مي‌خواهيد. او براي اين كه جواب درستي به شما بدهد و كتاب مفيدي را به شما ارائه نمايد با پرسيدن سؤالاتي از شما و با استفاده از تجارب خود كتاب مورد نظرتان را به شما تحويل خواهد داد. جويشگرها همچنين توانايي ندارند اما به نوعي آنها را شبيه‌سازي مي‌كنند.

پس جويشگرهاي پيمايشي چگونه به پاسخ مورد نظرتان از ميان ميليونها صفحه وب مي‌رسند؟ آنها يک مجموعه از قوانين را دارند كه الگوريتم ناميده مي‌شود. الگوريتم‌هاي مورد نظر براي هر جويشگري خاص و تقريباً سري هستند اما به هر حال از قوانين زير پيروي مي‌كنند:

مكان و بسامد

عوامل خارج از صفحه

سرفصل‌هاي بهينه سازي

مكان و بسامد

يكي از قوانين اصلي در الگوريتم‌هاي رتبه‌بندي موقعيت و بسامد (تعداد تكرار) واژه‌هايي است كه در صفحه مورد استفاده قرار گرفته‌اند

بطور خلاصه روش مكان-بسامد (Location/Frequency Methode)ناميده مي‌شود.

كتابدار مذكور را به خاطر مي‌آورِد؟ لازم است كه او كتاب‌هاي در رابطه با واژه «سفر» را طبق درخواست شما بيابد. او در وحله اول احساس مي‌كند كه شما به دنبال كتاب‌هايي هستيد كه

در نامشان كلمه «سفر» را شامل شوند. جويشگرها هم دقيقاً همان كار را انجام مي‌دهند.
آنها هم صفحاتي را برايتان فهرست مي‌كنند كه در برچسب عنوان ((Titleموجود در كد زبان نشانه‌گذاري اَبَرمتني (زنگام) ((HTMLحاوي واژه «سفر» باشند.

جويشگرها همچنين به دنبال واژه مورد نظر در بالاي صفحات و يا در آغاز بندها (پاراگراف‌ها) هستند.
آنها فرض مي‌كنند كه صفحاتي كه حاوي آن واژه در بالاي خود و يا در آغاز بندها و عناوين باشند به نتيجه مورد نظر شما مربوطتر هستند.
بسامد عامل بزرگ و مهم ديگري است كه جويشگرها از طريق آن صفحات مربوط را شناسايي مي‌نمايند.
جويشگرها صفحات را تجزيه كرده و با توجه به تكرار واژه‌اي در صفحه متوجه مي‌شوند كه آن واژه نسبت به ديگر واژه‌ها اهميت بيش‌تري در آن صفحه دارد و آن صفحه را در درجه بالاتري نسبت به صفحات ديگر قرار مي‌دهند.

چگونگي كاركرد دقيق جويشگرها درباره روش‌هايي از قبيل مكان-تكرار فاش نمي‌شود و هر جويشگري روش ويژه خود را دنبال مي‌كند.
به همين دليل است كه وقتي شما واژه‌هاي همانندي را در موتورهاي متفاوت جستجو مي‌كنيد، به نتايج متفاوتي مي‌رسيد.
الگوريتم‌هاي اوليه جويشگرهاي معتبر و بزرگ همچنان محرمانه نگهداري مي‌شوند.
برخي جويشگرها نسبت به برخي ديگر صفحات بيشتري را فهرست كرده‌اند. نتيجه اين خواهد شد كه هيچ جويشگري نتيجه جستجوي مشتركي با موتور ديگر نخواهد داشت و شما نتايج متفاوتي را از آن‌ها دريافت مي‌كنيد.

جويشگرها همچنين ممكن است كه برخي از صفحات را از فهرست خود حذف كنند

البته به شرطي كه آن صفحات با هرزنامه (Spam)شدن سعي در گول زدن جويشگرها داشته باشند.

فرستادن هرزنامه (Spamming)روشي است كه برخي از صفحات براي احراز رتبه بالاتر در جويشگرها در پيش مي‌گيرند

به اين صورت است كه با تكرار بيش از حد واژه‌ها و يا بزرگ نوشتن يا بسيار ريز نوشتن متنها بطور عمدي كوشش در بر هم زدن تعادل و در نتيجه فريب جويشگرها دارند.

آنها سعي دارند كه با افزايش عامل تكرار، در رتبه بالاتري قرار بگيرند.

البته آنگونه كه گفته شد تعداد تكرارها اگر از حد و اندازه خاصي فراتر رود نتيجه معكوس مي‌دهد جويشگرها راه‌هاي متنوعي براي جلوگيري از فرستادن هرزنامه دارند و در اين راه از گزارش‌هاي كاربران خود نيز بهره مي‌برند.

امروزه بهينه‌سازي سايت‌هاي اينترنت براي جويشگرها يكي از مهم‌ترين روشهاي جلب بازديدكننده به سايت است.

عوامل خارج از صفحه

جويشگرهاي گردشي اكنون تجربه فراواني در رابطه با وب‌دارهايي دارند كه صفحات خود را براي كسب رتبه بهتر مرتباً بازنويسي مي‌كنند.

بعضي از وب‌دارها (وب‌مسترها)ي خبره حتي ممكن است به سمت روش‌هايي مانند مهندسي معكوس براي كشف چگونگي روش‌هاي مكان-تكرار بروند.

به همين دليل، تمامي جويشگرهاي معروف از روش‌هاي امتيازبندي «خارج از صفحه» استفاده مي‌كنند.

عوامل خارج از صفحه عواملي هستند كه از تيررس وب‌دارها خارجند و آنها نمي‌توانند در آن دخالت كنند و مسأله مهم در آن تحليل ارتباطات و پيوندهاست.

عوامل خارج از صفحه

به وسيله تجزيه صفحات، جويشگرها پيوندها را بررسي كرده و از محبوبيت آنها مي‌فهمند كه آن

صفحات مهم بوده و شايسته ترفيع رتبه هستند.

به علاوه تكنيک‌هاي پيشرفته به گونه‌اي است كه از ايجاد پيوندهاي مصنوعي توسط وب‌دارها براي فريب جويشگرها جلوگيري مي‌نمايد.

علاوه بر آن جويشگرها بررسي مي‌كنند كه كدام صفحه توسط يک كاربر كه واژه‌اي را جستجو كرده انتخاب مي‌شود و سپس با توجه به تعداد انتخاب‌ها، رتبه صفحه مورد نظر را تعيين كرده و مقام آن را در نتيجه جستجو جابه‌جا مي‌نمايند.

تكنيک هاي جستجو براي جستجوي بهينه

يكي از كاراترين و مقتدرترين روش‌هاي جستجوي اطلاعات در دنياي وب استفاده از واژه‌هايي است كه اصطلاحاً كلمات كليدي ناميده مي‌شوند.

اغلب كاربران حرفه‌اي و جستجوگران ورزيده دنياي اينترنت مي‌توانند با طرح بهترين كلمات كليدي و بكار بستن قوانين تركيب آن‌ها با هم براي نيازهاي اطلاعاتي خود پاسخي در خور بيابند.

در اين روش توصيه‌هاي زير براي انتخاب كلمات كليدي و نيز جستجوي دقيق و مفيد پيشنهاد مي‌شود كه بشرح ذيل است :

تكنيک هاي جستجو براي جستجوي بهينه

- حتي‌المقدور سعي شود كلمات كليدي از ميان اصطلاحات منحصر به فرد و اسامي خاص انتخاب شود .
- حتي‌المقدور از آوردن كلمات عمومي كه عناوين بسياري را در زير مجموعه خود شامل مي‌شوند، جداً خودداري كنيد .
- هميشه اسم شخص يا نام شي يا هر چيز ديگري را كه مد نظر داريد به‌طور كامل وارد كنيد .
- دقت كنيد كه اگر موتور جستجو ميان حروف بزرگ و كوچك تفاوتي مي‌گذارد، اين مسأله را در طرح كلمات كليدي خود مدنظر داشته باشيد .
- در نظر داشته باشيد اگر نتيجه جستجو صفر بود به احتمال زياد مي‌تواند از يك اشتباه تايپي باشد .
- اگر املاي صحيح و كامل كلمه‌اي را نمي‌دانيد از كاركتر جانشين كه اغلب * و يا ؟ است

استفاده كنيد .

تكنيک هاي جستجو براي جستجوي بهينه

7. اگر يك كلمه كليدي را براي طرح دقيق و تمام و كمال يك مورد جستجو كفايت نمي‌كند، از تكنيك‌هاي جستجوي عبارتي، استفاده از اپراتورهاي جبر بولين (AND, OR, NOT) استفاده كنيد. جستجوي عبارتي يكي از مهم‌ترين و قدرتمندترين امكانات جستجو در اغلب موتورهاي جستجو مي‌باشد و مي‌توان يك عبارت يا جمله مشخص را به همان ترتيبي كه كلمات وارد شده‌اند مورد جستجو قرار داد. براي اين روش جستجو عبارت مورد نظر را داخل گيومه"" بگذاريد .
8. استفاده از عملگر AND : AND به مفهوم "و" براي محدود كردن دامنه جستجو از طريق تركيب كليد واژه‌هاي مختلف به كار مي‌رود و براي تركيب كليدهاي جستجو زماني كه براي شما مهم است كه دو يا چند كلمه كليدي حتماً وجود داشته باشد و علامت آن در پايگاه‌هاي مختلف به صورت استفاده از عبارتAND ، استفاده از + ، انتخاب عبارت ALL THE WORDاز منو، انتخاب عبارت (MATCH ON ALL WORDS AND)به‌وسيله كليك كردن بر روي دكمه‌هاي راديويي است .

تكنيک هاي جستجو براي جستجوي بهينه

9. استفاده از عملگر :OR اپراتور ORبه مفهوم "يا" و برخلاف عملگر AND باعث گسترش دامنه جستجو و بازيابي اطلاعات بيش‌تر شده براي تركيب كليد واژه‌هاي جستجو زماني كه انتظار داريد تنها يك، دو يا چند كلمه كليدي حضور داشته باشند و علامت آن استفاده از عبارتOR ، نحوه‌ي اجراي ساده و معمولي آن، انتخاب عبارت ANY OF THE WORDSاز منو، انتخاب عبارت (MATCH ON ANY WORDS OR) با كليك بر روي دكمه‌هاي راديويي مي‌باشد. يكي از كاربردهاي مهم اين عملگر پوشش مفاهيم يا اصطلاحات مترادف، مرتبط يا با املاهاي متفاوت است .
10. استفاده از عملگر :NOT اپراتور NOT به مفهوم "نه" و يا به جز كه در اين صورت تمامي جواب‌هاي بازگشتي كه حاوي عبارت يا كلمه كليدي هستند حذف خواهند گرديد و براي اجراي آن تنها كافيست كه NOTرا قبل از عبارت يا كلمه كليدي مورد نظرتان با يك فاصله بياوريد .

تكنيک هاي جستجو براي جستجوي بهينه

- استفاده از عملگر نزديك‌يابي: در بسياري از موارد استفاده از عملگر AND باعث بازيابي اطلاعاتي مي‌شود كه براي ما مفيد نيست. به اين دليل كه اين عملگر كليد واژه‌ها را در هر كجاي متن كه باشند بازيابي مي‌كند. در اين موارد استفاده از تكنيك نزديك‌يابي مي‌تواند از ريزش كاذب اطلاعات و يا بازيابي اطلاعات غيرمرتبط جلوگيري نمايد. همه موتورهاي جستجو قابليت استفاده از اين تكنيك را ندارند ولي به عنوان مثال در موتور جستجوي آلتاويستا مي‌توان با استفاده از عملگر NEAR از اين تكنيك استفاده نمود .

12. جستجوي تركيبي با استفاده از پرانتز: اين تكنيك يكي از مهم‌ترين تكنيك‌هاي جستجو مي‌باشد كه به وسيله آن مي‌توان تا حدود زيادي از بازيابي موارد غيرمرتبط در محيط وب جلوگيري كرد. در اين روش مي‌توان از همه عملگرهاي جستجو كه در بالا گفته شده يك‌جا استفاده كرد و آن‌ها را با هم‌ديگر تركيب نمود .

تكنيک هاي جستجو براي جستجوي بهينه

13. جستجوي كليدواژه در عنوان صفحات وب: اين تكنيك با اين پيش فرض كه عنوان يك صفحه وب تا حدود زيادي نمايانگر محتواي اطلاعات موجود در آن است به جستجوي واژه‌هاي كليدي در عنوان سايت‌ها مي‌پردازد. علامت آن در موتورهاي جستجو متفاوت است ولي اغلب موتورهاي جستجو از طريق فهرست انتخابي و يا گزينه‌هاي ديگر اين امكان را فراهم مي‌آورند .

14. جستجوي حوزه سايت‌ها: با توجه به اين كه به صورت قراردادي هر كشوري حوزه خاصي در محيط وب دارد، قابليت جستجوي حوزه سايت‌ها به ما اين امكان را مي‌دهد كه فرايند جستجو را به حوزه خاصي نظير سايت‌هاي وب ايران (IR) و يا سايت‌هاي وب سازمان‌هاي غير انتفاعي (ORG) محدود كنيم. دستورات استفاده از اين تكنيك در موتورهاي جستجو مختلف مي‌باشد .

تكنيک هاي جستجو براي جستجوي بهينه

15. محدود كردن جستجو به زبان‌هاي مختلف باعث مي‌شود نتايج جستجو به زبان‌هاي ديگر آورده نشود و انتخاب مطلب مورد نظر آسان‌تر است .

16. محدود كردن جستجو به تاريخ انتشار منابع در وب: تاريخ انتشار يا به اصطلاح روزآمدي مطلب به خصوص در منابع علمي اصل مهمي است و اين‌گونه محدوديت باعث مي‌شود بنا به نياز كاربر جديدترين و يا قديمي‌ترين منبع بازيابي بشود .

17. جستجوي رسانه‌هاي مختلف؛ موسيقي، عكس، ويدئو: زماني كه فقط نوع خاصي از رسانه مورد نياز است به عنوان مثال زماني كه به عكس يك شخصيت نياز داريم، جستجو در ميان عكس‌ها باعث مي‌شود نتيجه جستجو شامل اطلاعات ديگري در مورد آن شخصيت نباشد .

تكنيک هاي جستجو براي جستجوي بهينه

18. جستجوي صفحات با فرمت‌هاي مختلف :PDF, WORD, MP3, MPEG, ….زماني كه فرمت خاصي مورد نظر است مي‌توان از اين تكنيك استفاده كرد . به عنوان مثال اگر مايل باشيم منبع بازيابي شده در فرمت PDF باشد، اين تكنيك مي‌تواند مفيد باشد .

19. آگاهي از پيش‌فرض‌هاي جستجو در موتور جستجو: با توجه به اين كه هر موتور جستجو براي تركيب واژه‌ها يك پيش‌فرض دارد و اگر از هيچ گونه عملگري استفاده نشود، كليد واژه‌ها را به صورت پيش‌فرض با يكي از عملگرهاي جبر بولي تركيب مي‌كند؛ آگاهي از اين پيش‌فرض موتورهاي جستجوي مختلف مهارت ما را در جستجو بالا مي‌برد .

20. وب نامريي: وب نامرئي به دو دليل كمي و كيفي اهميت دارد كمي از اين نظر كه موتورهاي جستجو فقط قادر هستند حدود 16درصد از اطلاعات موجود در اينترنت را بازيابي كنند و اندازه وب نامريي تقريبا 500برابر وب مريي است و كيفي از اين نظر كه منابع اطلاعاتي موجود در وب عميق معمولا ارزشمند و مفيد هستند و در بسياري از موارد پاسخگوي نياز كاربران مي‌باشند .

آشنايي با ابزارهايي كه براي شناسايي منابع وب نامريي به وجود آمده‌اند و كاربران را به سايت‌هاي مناسب راهنمايي مي‌كنند، باعث دسترسي به اين بخش عظيم از اطلاعات مفيد و ارزشمند مي‌شود. مثل سايت INVISIBLEWEB كه فهرستي از منابع نامريي را و سايت COMPLETEPLASETكه فهرستي از تقريبا 40000 پايگاه اطلاعاتي وب نامريي را ارايه مي‌دهد .

جايگاه كتابخانه ها و مراكزاطلاع رساني در جامعه اطلاعاتي

كتابداري و اطلاع رساني جزو مشاغل اطلاعاتي است

كتابخانه ها به عنوان نهادهايي كه وظايف گردآوري ، سازماندهي و اشاعه اطلاعات را بر عهده دارند از عناصر اصلي جامعه اطلاعاتي و زيرساخت هاي اطلاعاتي محسوب مي شوند.

جوامع توسعه يافته، كتابخانه هاي توسعه يافته تري نيز نسبت به جوامع در حال توسعه دارند.

راهنماهاي موضوعي

كتابداران، اطلاع رسانان، آموزشگران و پژوهشگران كتابداري و اطلاع رساني، صرفنظر از اين كه در چه حوزه اي فعاليت مي كنند، مي توانند با بهره گيري از اين رويكردها، خدمات خود را نظام مند كرده و به فعاليتهاي خود جهت دهند.

راهنماهاي موضوعي...

راهنماهاي متون با عناوين مختلفي نظير راهنماهاي منابع اطلاعاتي و راهنماهاي منابع مرجع و دروازه هاي موضوعي شناخته مي شوند و منابع اطلاعاتي را در يک يا چند زمينه موضوعي، معرفي و ساختار آنها را توصيف مي كنند.

منابعي نظير Guide to Reference Book و Walford's Guide to Reference Materials كه كتابداران با آنها آشنا هستند، نمونه هايي از راهنماهاي ارزشمندي هستند كه تمام زمينه هاي موضوعي را پوشش مي دهند.

راهنماهاي موضوعي...

مي توان اين راهنماها را با استفاده از دروازه هاي موضوعي (gateways) مبتني بر وب نيز تهيه كرد چون بيشتر پژوهشگران تمايل دارند منابع مرتبط با موضوعات تخصصي مورد علاقه شان را بدون صرف زمان زيادي، از طريق اينترنت بدست آورند و كتابخانه ها و كتابداران همواره درگير توسعه و طراحي دروازه هاي موضوعي در وب بوده اند.

مطالعات كاربران

دو نوع اصلي از پژوهش هاي مربوط به مطالعات كاربران، نياز اطلاعاتي و رفتار اطلاع يابي است.

جستجوي هدفمند اطلاعات بعنوان پيامد يک نياز اطلاعاتي در راستاي برآوردن يک يا چند هدف.

جستجوي اطلاعات از سوي فرد ممكن است از طريق تعامل وي با نظام هاي اطلاعاتي دستي(نظير يک روزنامه يا يک كتابخانه) يا از طريق تعامل با نظامهاي رايانه محور(نظير وب جهان گستر) صورت پذيرد

مطالعات کاربران

اهميت ديگري كه بررسي نياز اطلاعاتي كاربران دارد در طراحي نظام هاي بازيابي اطلاعاتي است. اين نظام ها را، بدون يافتن دركي روشن از آنچه استفاده كنندگان نياز دارند يا مي خواهند بدانند، چگونگي جستجوي اطلاعات توسط آنها و چگونگي ارزيابي اطلاعاتي كه دريافت مي كنند

مطالعات كتابسنجي

- مطالعات كتابسنجي
- اطلاع سنجي
- علم سنجي
- وب سنجي

افزايش حجم سريع اطلاعات، انفجار اطلاعات، آلودگي اطلاعات، اضافه با اطلاعاتي و ...
سردرگمي متخصصان در تشخيص اطلاعات سره از ناسره

وب سنجي

- تحليل محتواي صفحات وب
- تحليل فناورانه وب(عملكرد موتورهاي كاوش)
- تحليل استنادي وب
- بررسي ضريب تاثيرگذاري وب (web impact factor)

و ...

مهارتهاي اطلاع يابي در محيط وب

با تأكيد بر پايگاه هاي اطلاعاتي

مقدمه

افزايش روزافزون اطلاعات

انتشار منابع اطلاعاتي متعدد

جامعه علمي را با مشكلات و چالش هاي جدي روبرو ساخته است.

در ميان اين حجم انبوه اطلاعات (چاپي، غير چاپي، و الكترونيكي) پيدا كردن اطلاعات كاملا مرتبط و معتبر كار آساني نيست و نيازمند مهارتهاي ويژه اي است.

مهارت هاي اطلاعاتي پيش نياز تحقيق است.

تعريف وب مرئي و نامرئي

تعريف فني

وب مرئي

صفحات ثابت با دسترسي آزاد

وب نامرئي

تشكيل شده از صفحات وب متحرک و يا قابل دسترس به صورت محدود (مثلا با استفاده از رمز ورود ..)

تقسيم بندي وب مرئي و نامرئي

مهارت هاي اطلاع يابي

- مهارت در بازيابي اطلاعات؛
- مهارت در ارزيابي اطلاعات؛
- مهارت در سازماندهي اطلاعات؛
- و مهارت در تبادل ارتباط.

ابزارهاي جستجو

روشهاي جستجوي اطلاعات در محيط وب

- ساده

- پیشرفته
- کنترل واژگان
- اصطلاحنامه

فرایند جستجو

ارزیابي پایگاه ها

- جامعیت
- مانعیت

چگونه مي توان مدارك بیشتر و مرتبط تر را بازیابي کرد؟

بخش هاي مختلف وب نامرئي

وب مات یا تاریك

۱بخشي از فضاي وب نامرئي به وب مات موسوم گردیده که مي توانسته مورد استفاده کاربران قرار گیرد، اما به دلایل زیر این اطلاعات در خارج از دسترس کاربران قرار گرفته و موتورهاي کاوش نمي توانند آن ها را بازیابي کنند :

۱ .از آنجا که اولاً محیط وب دائماً در تغییر است و هر روز منابع و اطلاعات جدید به آن افزوده مي گردد و ثانیاً صفحاتي در وب وجود دارند که هیچ پیوندي بین آن ها با منابع دیگر برقرار نشده، خزنده هاي موتورهاي جستجو قادر به یافتن این صفحات و همگام نمودن خود با این حجم عظیم اطلاعات نیستند .

وب مات یا تاریك

۲ .به دلیل محدودیت توانایي، نرم افزارهاي خزنده فرصت کافي براي روزآمدسازي صفحات جدید وب را ندارند. موتورهاي کاوش نیز امکان روزآمدسازي حجم عظیمي از اطلاعات و منابع جدید را ندارند و به همین دلیل بسیاري از این اطلاعات از حوزه موتورهاي کاوش دور مي مانند .

۳ .محدودیت توان مالي بسیاري از موتورهاي کاوش سبب گردیده که موتورهاي کاوش نتوانند تمام صفحات وب سایت ها را نمایه سازي کنند، چرا که براي آن ها هزینه هاي زیادي دارد و

بنابراين موتورهاي كاوش بنا بر سياست هاي خودشان، تنها بخشي از وب سايت ها يا لايه هاي بيروني آن ها را نمايه سازي مي كنند. بنابراين هميشه بخش عظيم لايه هاي دروني وب سايت ها پنهان مي مانند .

۲ .وب عميق

به مجموعه اي از اطلاعات الكترونيكي پيوسته اطلاق مي شود كه بسياري از پايگاه هاي اطلاع رساني، آن ها را از طريق شبكه جهانگستر وب در دسترس عموم قرار داده اند.
برخي اين اطلاعات را به رايگان، و برخي ديگر را با دريافت هزينه در دسترس عموم قرار مي دهند.
مندرجات اين پايگاه ها معمولاً خارج از حوزه جستجوي موتورهاي كاوش قرار دارند هريك از اين پايگاه ها صفحه جستجوي مبتني بر وب دارند. كه امكان جستجو در آن ها براي كاربران را فراهم مي كند، اما خزنده هاي موتورهاي جستجو توان ورود به آن ها را ندارند و در نتيجه حجم انبوهي از اطلاعات، نمايه نشده باقي مي ماند.

۲ .وب عميق

به عنوان نمونه اگر يك متخصص موضوعي)مثلاً يك دانشجوي رشته پزشكي) بخواهد خود را به موتورهاي كاوش معمولي محدود كند و نتواند به پايگاه هاي اطلاعاتي تخصصي مراجعه نمايد يا از وجود آن ها آگاه نباشد، از دسترسي به حجم انبوهي از اطلاعات محروم خواهد ماند .
بنابراين كاربر بايد در اين موارد از طريق موتورهاي جستجو، پايگاه هاي مرتبط با موضوع خود را شناسايي كند و سپس، جداگانه به جستجو در آن ها بپردازد تا از دسترسي به وب عميق باز نماند .

۳ .وب خصوصي و وب ملكي

بخشي ديگر از وب نامرئي وجود دارد كه چون اطلاعات موجود در آن جزو دارايي هاي شخصي يا خصوصي سازمان ها يا افراد مي باشد، از حوزه دسترسي موتورهاي جستجو پنهان است. مثلاً در برخي از سازمان ها و مؤسسات خصوصي يا دولتي، به دلايل امنيتي از اطلاعات مربوط به مسائل كاري و سازماني و پرسنلي خود حفاظت مي كنند اجازه دسترسي به آن ها را به ديگران نمي دهند و فقط كساني كه داراي اسم كاربر و گذرواژه هستند مي توانند از آن ها استفاده كنند؛ اين

بخش، وب خصوصي محسوب مي گردد (منصوريان، ۱۳۸۲).بخش ديگر، منابع اطلاعاتي از قبيل نشريات الكترونيكي مبتني بر وب مي باشند كه دسترسي به آن ها از طريق پرداخت حق اشتراك و خريد محصولات اطلاعاتي شركت هاي مختلف صورت مي گيرد و «وب ملكي» ناميده مي شود .

۴ .اينترنت واقعاً پنهان

بخش ديگري از وب پنهان وجود دارد كه بنا به مسائل فني و ناكارآمدي ابزارهاي جستجو، از دسترسي كاربران دورمانده است. بسياري از موتورهاي جستجو قادر به بازيابي اطلاعات متني اچ تي ام ال هستند، ولي توانايي بازيابي فايل هاي پي دي اف را ندارند، يا به دليل كمبود منابع مالي و فني از جستجوي فايل هاي غيرمتني صرف نظر كرده اند. بنابراين منابع اطلاعاتي متنوعي نيز در وب وجود دارند كه تنها به دليل محدوديت هاي فناوران يا مالي موتورهاي جست وجو، از حوزه كاوش آن ها و در نتيجه از دسترس كاربران دور مانده اند

منابع موجود در وب نامرئي

بخش بزرگي از وب وجود دارد كه عنكبوت هاي موتورهاي جستجو آن ها را نمايه نمي كنند يا نمي توانند نمايه كنند و عبارت اند از سايت هاي داراي رمز عبور، اسناد موجود در پشت سامانه هاي حفاظتي، فايل هاي پي دي اف از متون آرشيو شده، و ابزارهاي تعاملي نظير ماشين حساب ها و برخي واژه نامه ها و همچنين محتويات بعضي از پايگاه هاي اطلاعاتي، منابع محافظت شده از طريق اسم كاربر و گذرواژه، منابع و صفحات وب بدون پيوند، و صفحات افزون بر حداكثر تعداد صفحات قابل مرور در نتايج بازيابي .

اهميت وب نامرئي

به دو دليل مي توان گفت كه وب نامرئي اهميت دارد .نخست از نظر كمّي، بايد گفت كه حجم اطلاعات موجود در اين بخش خيلي بيشتر از سطح آشكار است[۱۷]. موارد زير، اهميت وب نامرئي را از نظر كمّي نشان مي دهند

۱ .بهترين موتورهاي كاوش فقط قادر هستند كه حدود ۱۶ درصد از اطلاعات موجود در وب را بازيابي كنند و بنابراين ۸۴ درصد آن ها جزو وب نامرئي به حساب مي آيند .

۲ .اندازه وب نامرئي تقريباً ۵۰۰ برابر وب مرئي است: وب نامرئي ۵۵۰ ميليون سند، و وب مرئي تقريباً ۱ ميليون سند را دارا مي باشد .

دوم اينكه از نظر كيفي، اطلاعات بخش هاي مختلف اين مجموعه بويژه منابع اطلاعاتي موجود در وب عميق، معمولاً منابع ارزشمند و مفيد هستند و در بسياري از موارد پاسخگوي نياز كاربران مي باشند. تقريباً بيش از نيمي از وب نامرئي را پايگاه هاي اطلاعاتي موضوعي تشكيل مي دهند .

دلائل عدم بازيابي و نمايه سازي وب نامرئي توسط موتورهاي كاوش

۱ .دلايل فني: بسياري از موتورهاي كاوش به دليل محدوديت هاي نرم افزاري توانايي روزآمدسازي اطلاعات جديد وب را ندارند .بايد يادآور شد كه هنوز هيچ موتور كاوشي ادعا نكرده است كه قادر به گسترش حوزه كاوش خود به تمام محيط وب مي باشد و هميشه اين موتورها يك گام از سرعت روزافزون اطلاعات عقب تر هستند .

۲ .دلايل بودجه اي: همانطور كه قبلاً اشاره شد فرآيند نمايه سازي تمام صفحات وب، هزينه بر خواهد بود و موتورهاي كاوش نيز بنا به محدوديت بودجه ناگزيرند فقط بخشي از وب سايت ها را نمايه سازي كنند .

۳ .دلايل اجتماعي و حقوقي: از آنجا كه اطلاعات موجود در وب در دسترس عموم قرار مي گيرد، بسياري از افراد و سازمان ها به دليل صرف بودجه هاي كلان در راه اندازي سايت ها و پايگاه هاي اطلاعاتي خود، حاضر نيستند اين اطلاعات را به صورت رايگان در اختيار همه بگذارند. البتّه اين از لحاظ اجتماعي و حقوقي حق مسلّم آن ها است .

شيوه هاي اطلاع يابي در وب نامرئي

در حال حاضر ابزارهايي به وجود آمده اند كه منابع وب نامرئي را شناسايي، و كاربران را به سايت هاي مناسب راهنمايي مي كنند .اين رويكرد توسط بزرگراه هاي اطلاعاتي و كتابخانه هاي مجازي پذيرفته شده است؛ بطوري كه فقط توصيفي از پايگاه هاي اطلاعاتي و مجلات نامرئي را ارائه مي كنند؛ مثل سايت Invisibleweb كه فهرستي از منابع نامرئي را، و سايت Completeplasetكه فهرستي از تقريباً ۴۰۰۰۰ پايگاه اطلاعاتي وب نامرئي را ارائه مي دهند.

برخي ديگر از ابزارهاي اطلاع يابي نيز كه تاكنون معرفي شده اند و ما مي توانيم با استفاده از آن ها به اين اطلاعات دسترسي پيدا كنيم به شرح زير است :

دروازه هاي اطلاعاتي

- دروازه هاي اطلاعاتي مجموعه اي از پايگاه ها و سايت ها هستند كه به وسيله متخصصان اطلاعاتي و معمولاً كتابداران گردآوري، بررسي و براساس موضوع مرتب شده اند و معمولاً به كاربران نيز توصيه مي شوند. نمونه اي از اين دروازه ها عبارت اند از :
- Academic Information
- Digital Librarian
- Gaary price direct search
- Infomine
- Internet public Library

پرتال ها يا پايگاه هاي اطلاعاتي خاص موضوعي

مجموعه اي از پايگاه هاي اطلاعاتي خاص موضوعي هستند كه به يك موضوع خاص اختصاص دارند و به وسيله دانشمندان، محققان، متخصصان، مؤسسات دولتي، شركت هاي بازرگاني و كارشناسان موضوعي، افراد بسيار علاقه مند يا داراي دانش حرفه اي و اطلاعات وسيع در حوزه خاص ايجاد مي شوند (Oxford uni. Libraries ٢٠٠٠ .(از ورتال ها در هنگام جستجو براي موضوعات خاص مانند پيوندهاي خبري، فايل هاي چندرسانه اي، آرشيوها، فهرست هاي پستي اشخاص، شغل ياب ها و هزاران پايگاه اطلاعاتي كه به موضوعات خاص اختصاص دارند استفاده مي شود. از مزيت هاي عمده استفاده از دروازه هاي اطلاعاتي اين است كه براي ايجاد آن ها افرادي با دانش موضوعي خاص، در اينترنت جستجو كرده اند و به پالايش اطلاعات مفيد از غيرمفيد پرداخته اند .

ويژگي پايگاه هاي اطلاعاتي :

١ .جستجوپذير و مرورپذير، حاوي توصيف منابع اينترنتي در يك زمينه موضوعي خاص؛

٢ .مشتمل بر معيارهاي شفاف و تعريف شده براي ارزيابي كيفيت منابع اطلاعاتي، به جاي انتخاب بدون ارزيابي؛

۳ .دخالت كتابداران و متخصصان موضوعي در ايجاد آن؛

۴ .توليد دستي ركوردهاي آن، براي توصيف بامعنا و اطلاع بخش از منابع اطلاعاتي؛

۵ .فهرستنويسي و رده بندي منابع اطلاعاتي با استفاده از شيوه هاي سنتي كتابخانه اي به منظور بازيابي موثر .

ابرموتورهاي كاوش :

گسترش پذيري حوزه هاي جستجو نيز يكي از شيوه هاي دسترسي به وب نامرئي شمرده مي شود كه نمونه آن، استفاده از ابرموتورهاي كاوش است. اين ابرموتورها خود، موتورهاي جستجوي واقعي نيستند. بلكه به كاربران اين امكان را مي دهند كه كليدواژه هاي خود را همزمان توسط چند موتور، مورد كاوش قرار دهند و نتايج جستجوي همه آن ها را با هم ببينند .

عوامل هوشمند :

اين عوامل هوشمند از ابزارهاي بازيابي در اينترنت هستند كه براي اجراي كارهاي بخصوص به كارگرفته مي شوند. اين عوامل توانايي جستجو و مقايسه و انتخاب منابع اطلاعاتي بر اساس نياز مطرح شده توسط كاربر را دارند .

توصيه‌ها و ابزارهايي براي جستجوي وب نامرئي

- متخصصان جستجو ي وب ميگويند موتورهاي جستجو مثل گوگل و ياهو فقط 1 درصد (surface Web)از اطلاعات وب را براي جستجو اطلاعات اينترنت را مورد جستجو قرار ميدهند مابقي اطلاعات اينترنت را وب نامرئي يا deep Web، Deepnet، invisible Web، dark Web، hidden Webمي نامند. خوشبختانه براي جستجو قسمتهايي از وب پنهان ابزارها و موتور جستجوهاي متعددي ايجاد شده و در يافتن اطلاعات دلخواه كمك خوبي هستند. اين موتور جستجوها از الگوريتم هاي پيشرفته رنكينگ و language-analysisاستفاده مي كنند كه اطلاعات مفيد و مرتبط را چكيده مي كنند.

موتور جستجوهاي وب نامرئي:

DeepDyve:يكي از جديدترين موتور جستجوها كه بر وب نامرئي تمركز كرده است

CloserLook Search:جستجوگري براي اطلاعات پزشكي ، سلامت ، داروها و...

CompletePlanet:بيش از هفت هزار ديتابيس و موتور جستجو در اين پايگاه در دسترس است و يكي از راه هاي عالي براي جستجوي وب نامرئي مي باشد

Daylife:جستجوي اخبار به همراه تصاوير ، مقالات و...

spock:جستجوي افراد در وب ؛ پيدا كردن پروفايلها و تصاوير مخفي دوستانتان

The WWW Virtual Library:يكي از قديمي ترين پايگاه اطلاعاتي وب كه با كلمات كليدي يا دسته بندي ها مي توانيد به اطلاعاتش دست يابيد

موتور جستجوهاي وب نامرئي:

pipl:طراحي شده براي جستجو وب نامرئي ؛ نتايجي كه اين وب سايت در دسترس قرار مي دهد از نظر بعضي از كاربرانش خطرناك است (حريم شخصي(

CustomSearchEngine:ليستي از جستجوگرهاي سفارشي سازي شده گوگل

SurfWax:موتور جستجويي ساده و پر محتوا

Freebase:مجموعه اي از ميليونها بانك اطلاعاتي در موضوعات متنوع

RefSeek:موتور جستجويي براي دانش جويان و محققان ؛ بيشتر از يك ميليارد سند كه شامل صفحات وب ، مجلات ، روزنامه ها ، كتاب ها و دايره المعارف را جستجو مي كند

چه كنيم كه اطلاعاتمان جزء وب نامرئي محسوب نشود؟

1-بازيابي صفحات وب عميق

يك صفحهي P پويا گفته ميشود اگر بعضي يا تمام محتواي آن در زمان اجرا (زمان بعد از دريافت درخواست صفحه در خدمتگذار) توسط برنامهاي بر روي خدمتگذار يا مشتري توليد شود.

تشخيص فرمها

تشخيص اسكريپتها

2-تشخيص فيلدهاي فرم

كار بسيار دشواري است!

اكثر تكنيكها براساس روشهاي مكاشفهاي است.

خوشبختانه، در اكثر فرمها از عناصر يكساني استفاده شده است.

3-تكميل خودكار فيلدها

انتصاب مقدار مناسب به فيلدهاي استخراج شده براي كشف محتويات داخل دادهپايگاه

تكنيك اول:

استفاده از پرس‌وجوهاي از پيش تعيين شده

تكنيك دوم:

استفاده از مقادير موجود در فيلدها بصورت جايگشتي

تكنيك سوم:

آموزش خزشگر با استفاده از پالايش صفحهي جستجو

3-تكميل خودكار فيلدها - ادامه

تكنيك چهارم:

استفاده از تكنيك‌هاي يادگيري ماشين

تكنيك پنجم:

تبديل مسألهي پيدا كردن بهترين پرس‌وجو براي يك داده پايگاه به مسألهي *پوشش مجموعه* در گراف‌ها

تبديل به مسألهي *مجموعهي غالب وزن‌دار كمينه*

تكنيك ششم:

كار با اسكريپت‌هاي سمت مشتري

4-آناليز نتايج دريافتي از دادهپايگاه‌ها

دلايل:

بدست آوردن كلمات كليدي جديد

تخمين تعداد مستندات دادهپايگاه در يك زمينهي خاص

5-دستهبندي يا خوشهبندي دادهپايگاه‌ها

دستهبندي يا خوشهبندي؟

روش‌هاي مختلف دستهبندي:

دستهبندي مبتني بر پرس‌وجو

دستهبندي مبتني بر خزش

دستهبندي با استفاده از توصيفات كلاس سرويس