دانشگاه علوم پزشکی
و خدمات بهداشتی درمانی لرستان

نمایه سازی(Indexing)

مهوش کلهر
کارشناس ارشد علوم کتابداری و اطلاع رسانی
دانشگاه علوم پزشکی لرستان .دانشکده پرستاری بروجرد .
دی 1397

– اطلاعات وجود دارد، اما ما چیزی پیدا نمی‌کنیم
– اطلاعات پیدا می‌شود، اما آن چیزی نیست که مورد نظر ما بوده
– تنها بخشی از اطلاعات یافت می‌شود
– دقیقاً همان اطلاعاتی را که نیاز داریم پیدا می‌کنیم

نمایه چیست ؟

عمل توصیف یا شناسایی محتوای موضوعی یک مدرک ، در واقع، ثبت و ضبط محتوای اطلاعاتی مدارک با استفاده از روشهای گوناگون(معمولا الفبایی) به منظور سازماندهی اطلاعات به قصد سهولت بازیابی را نمایه‌سازی گویند یعنی تخصیص واژه‌ها یا اصطلاحات به مدارک به منظور توصیف محتوای موضوعی آنها برای بازیابی در مراحل بعد.

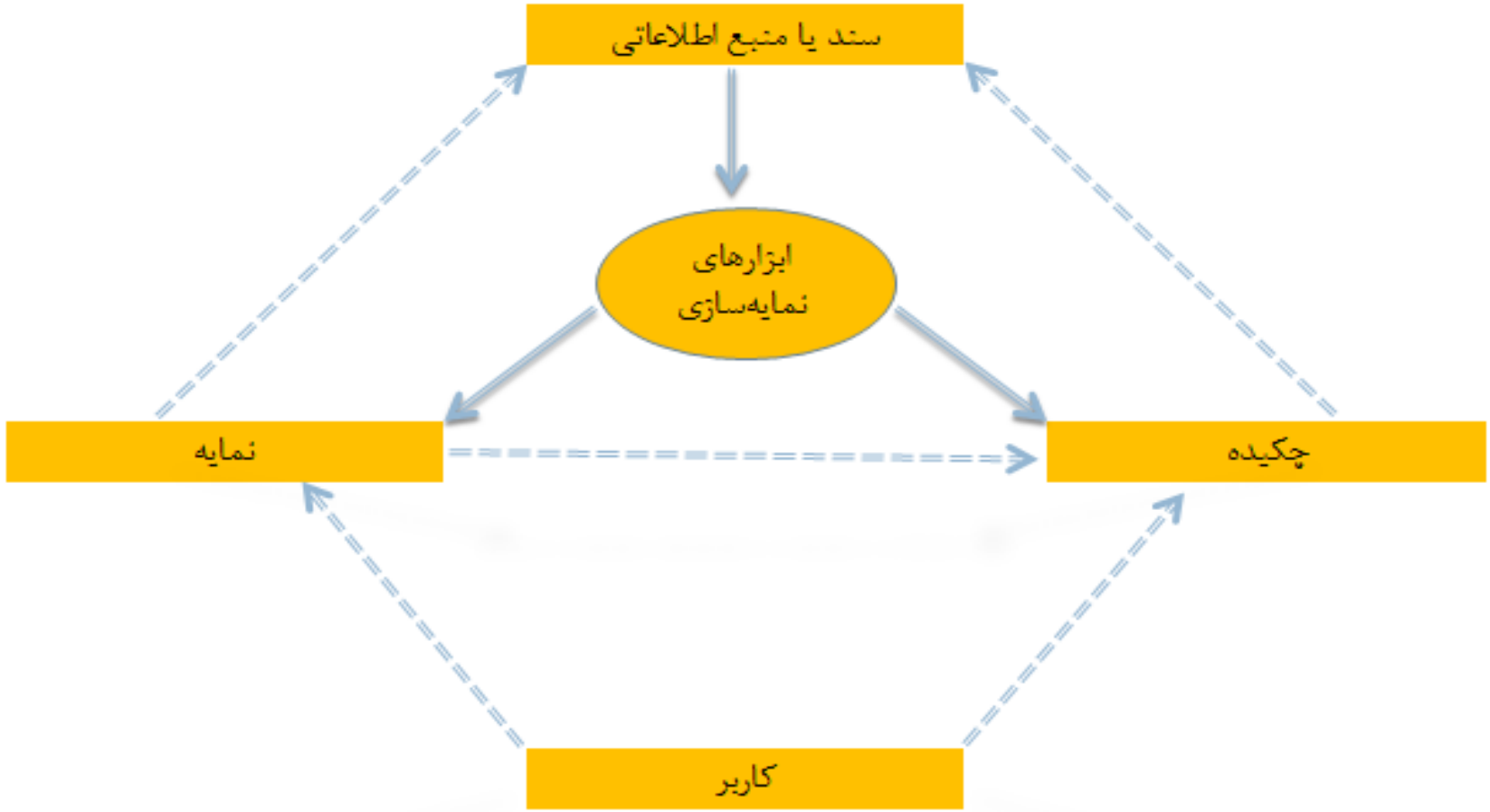
نمایه وسیله ای است برای هدایت منظم به یک متن، محتوا، مجموعه ای از مدارک و یا هر گونه اطلاعات ضبط شده ای که به شکل معمولا الفبایی مرتب شده باشد و با استفاده از شیوه خاص و نظام ارجاعی مشخص ، موقعیت و محل هر مطلب را در بازیابی اطلاعات نشان دهد.

- **هدف نمایه‌سازی اصولاً آماده کردن مدرک برای بازیابی سریع و آسان مدارک است.**

نیاز به نمایه

- نمایه، میانبرهایی سیستماتیک و موثربه اطلاعاتی است که استفاده کنندگان نیاز دارند.
- نمایه ها برای هر مجموعه اطلاعاتی، جزمجموعه های خیلی کوچک، ضروری هستند.
- نمایه سازی عمل ایجاد شاخص ها برای مجموعه ای از رکوردها است. وجود شاخص ها به جستجوگران اجازه می دهد با سرعت بیشتری رکورد ها را برای افراد خاص پیدا کنند که در نبود آن ها جستجوگران ممکن است مجبور باشند به صدها یا هزاران رکورد نگاه کنند تا یک رکورد خاص را بازیابی کنند.
- رکورد نشان دهنده یک عدد است که به فهرستی از اصطلاحات، تعاریف، موضوعات و غیره ارجاع می دهد که الفبایی تنظیم شده اند تا خواننده را به اطلاعات مورد نظر در درون محتوا هدایت کنند.
- نمایه سازی، سازماندهی اطلاعات و محتوا را به نحوی آسان می کند که مدرک مورد نظر برای خواننده به آسانی قابل شناسایی باشد.
- تغییر بازیابی اطلاعات از اعم به اخص

رابطه نمایه‌سازی، چکیده‌نویسی و جستجو



انواع نمایه

- نمایه‌های الفبایی یا واژه‌ای
- نمایه‌های مؤلف
- نمایه‌های کتاب
- نمایه‌های استنادی
- نمایه‌های رده‌ای:
- نمایه‌های همارا
- نمایه‌های درهمکرد
- نمایه‌های چهریزه‌ای
- نمایه‌های سطر اول
- نمایه‌های فرارسانه‌ای
- نمایه‌های اینترنتی
- نمایه‌های چندرسانه‌ای
- نمایه‌های نشریات ادواری
- نمایه‌های گردشی عنوان

انواع نمایه

- **نمایه مولف:نقاط مدخل آنها شامل اسامی افراد، سازمان‌ها، پدیدآورندگان تنالگانی، و... می‌باشند**
- **نمایه استنادی :شامل فهرستی از مقالات و همچنین فهرستی فرعی از مقالات منتشرشده‌ای است که به آن مقالات استناد کرده‌اند**
- **نمایه سطر اول :واژه‌های اولین بیت از یک شعر در موقعیت الفبایی خود در نمایه فهرست می‌گردند**
- **نمایه چند رسانه ای:این نمایه مواد متنی، صوتی و تصویری را با هم ترکیب می‌کند**
- **نمایه رده ای :مدخل‌های آن بر اساس رده‌ها یا سرعنوان‌های موضوعی مرتب می‌شود**
- **نمایه نشریات ادواری:دو نوع می‌باشد: نمایه انفرادی برای مجلات انفرادی و نمایه‌های کلّی برای گروهی از مجلات**

نمایه گردشی : با ورود رایانه به عرصه نمایه سازی دو نوع متداول نمایه گردشی عنوان شکل گرفت:

۱- کوئیک(Keyword In Context)

۲- کواک (Keyword Out of Context)

کوئیک (Keyword In Context)

- *عنوان یک مدرک، محتوای آن را نشان می دهد.*
- *عنوان های مدارک را براساس کلیدواژه های موجود در آن ها مرتب کرده و هر کلیدواژه در نظمی الفبایی، نقش مهمی به عنوان مدخل دارد.*
- نمایه کوئیک نمایه اي گردشی است که اغلب از عنوان هاي انتشارات به وجود می آید بدین مفهوم که هر کلیدواژه موجود در عنوان به منزله نقطه بازیابی (Point Access)در نظر گرفته می شود کلیدواژه هاي موجود در عنوان (نقاط بازیابی) به صورت الفبایی در مرکز واژه هاي غیرکلیدي موجود در عنوان در اطراف کلیدواژه
- سیاهه بازدارنده : (List Stop)شامل واژه هایی است که عملکردي نحوي دارند مثل حروف تعریف، حروف اضافه، حروف ربط و... فاقد هرگونه محتواي موضوعی می باشند

نمونه ای از یک نمایه کوئیک :*واژه نامه ها و دائره المعارف های علوم کتابداری و اطلاع رسانی*

های علوم کتابداری و	اطلاع رسانی#	واژه نامه ها و دائره المعارف
واژه نامه ها و	دائره المعارف های	علوم کتابداری و اطلاع رسانی#
واژه نامه ها و دائره المعارف های	علوم	کتابداری و اطلاع رسانی#
و دائره المعارف های علوم	کتابداری	و اطلاع رسانی# واژه نامه ها
علوم کتابداری و اطلاع رسانی#	واژه نامه ها	و دائره المعارف های

۲- کواک (Keyword Out of Context)

نمایه اي که در آن هر واژه مهم در زنجیره اي از متن، به عنوان اصطالح هدایت کننده یا نقطه دسترسی و به دنبال زنجیره کامل آمده است

1 -در نمایه کواك همه واژه ها که به صورت مدخل ظاهر می شوند، از عناوین مدرك استخراج می گردند

2 -در این نمایه بعضی مدخل ها معمولا اًصطلاحات تک واژه اي هستند

نمونه ای از یک نمایه کواک

”واژه نامه ها و دائره المعارف ها ي علوم كتابداري و اطلاع رساني“

اطلاع رسانی

واژه نامه ها و دائره المعارف هاي علوم كتابداري و اطلاع رسانی

دائره المعارف ها

واژه نامه ها و دائره المعارف هاي علوم كتابداري و اطلاع رسانی

علوم

واژه نامه ها و دائره المعارف هاي علوم كتابداري و اطلاع رسانی

كتابداري

واژه نامه ها و دائره المعارف هاي علوم كتابداري و اطلاع رسانی

واژه نامه ها

واژه نامه ها و دائره المعارف هاي علوم كتابداري و اطلاع رسانی

نمایه سازی کوئیک:

در نمایه سازی کوئیک که همان نمایه درون بافتی است تمام کلمه های موجود درعناوین مدارک راکه توسط برنامه ای به کامپیوتر داده شده ، با سیاهه ای ازواژه های غیرمجاز(واژه های غیرموضوعی که درفهرست ایستی یا بازدارنده قرار می گیردتا درردیف الفبایی خود به عنوان سرشناسه قرار نگیرد) دربرنامه کامپیوتر مطابقت داده می شود.کلمه هایی که درسیاهه واژه های غیرمجاز وجودندارد، بطورخودکارکلیدواژه محسوب می شود. مدخلی که باهریک از این کلیدواژه هادرست می شودبه صورتی است که این واژه را دردرون بافت خود ، یعنی همراه با بقیه کلمه های آن عنوان ، درشکل و نظم طبیعی اش نشان می دهد. این بافت شامل همه واژه ها از جمله واژه های غیر مجاز نیز می شود و به این ترتیب گرچه واژه های غیرمجاز در نمایه نامه مدخل قرار نمیگیرد، اماهمراه با سایر کلمات عنوان، درمدخل می آید تا محیط کلیدواژه و معناومفهوم آن را دقیقتر آشکار سازد و ارتباط میان کلیدواژه های عنوان را بهتر نمایش دهد. به این ترتیب بی درنگ دریافته می شود که عنوان مربوط با نیاز مراجعه کننده مناسبت داردیانه.برون داد رایانه ای می تواند به صورت چاپی دردسترس مراجعان قرارگیرد

نمایه درون بافتی شامل سه بخش است

1. کلیدواژه یا مدخل
2. بافت(متن)
3. نشانه بازیابی(جاینما).

نمایه واژه ها به صورت ستونی درمرکز یا سمت چپ صفحه ویاسمت راست آن(درفارسی) به ترتیب الفبایی می آیدو با همین کلمه هاست که جوینده ، جستجوی خود رادرنمایه آغاز می کند.برای مشخص کردن پایان عنوان، درهرنمایه از علامت قراردادی خاصی نظیرعلامت جمع یا مساوی و یاخط اریب(اسلش/) وغیره استفاده می شود تاجستجوگر بتواند به سهولت ابتدای عنوان را پیدا کرده و عنوان را درشکل طبیعی خود به سرعت مرورکند.

مزایای این نمایه

- سرعت وسهولت
- عدم نیاز به متخصص نمایه سازی زیرا که با وجود فهرست بازدارنده این کار را کامپیوتر انجام می دهد.
- وجود کلمات کلیدی متن که به جستجو گر کمک شایانی می کند.

نمایه کواک

نمایه کواک همان نمایه برون بافتی است. این نمایه برای رهایی از مسائل و مشکلات ناشی از ضرورت کوتاه کردن عنوان و نیز به منظور ساده ساختن خواندن مدخلها طراحی شده است. دراین نمایه هرکلیدواژه به ترتیب از عنوان خود خارج شده و مقدم برسایر اجزای عنوان قرارمی گیرد.سپس عنوان مدرک به ترتیب طبیعی خود و به طور کامل درزیر این واژه یا به دنبال آن می آید. به این ترتیب برای هر واژه مهم یک مدخل ساخته می شود.دراین نمایه نیز کلمات مجاز درمتن عنوان وجود دارد وکلید واژه قرار نمی گیرد

کارکرد نمايه سازي (Indexing Function)

1. محتواي اطلاعاتي مدارک را فشرده مي‌سازد.
2. به عنوان واسطه براي تطبيق و يکسان سازي زبان مدرک و زبان کاوش به کار مي‌رود.
3. به عنوان ابزاري کارا، بر شيوه تدوين راهبردي کاوش در جستجوي اطلاعات نظارت دارد.

مراحل نمایه سازی

1. آشنایي و شناخت
2. تجزیه و تحلیل
3. تشخیص اطاعات قابل نمایه شدن
4. تبدیل مفاهیم به اصطلاحات نمایه یا ترجمه و انتقال مفاهیم قابل نمایه شدن به اصطلاحات پذیرفته شده

زبان نمايه سازي

1. زبان طبيعي National Language

نوعي نمايه سازي است كه از زبان مدارک استفاده مي شود. در اين روش هر اصطلاح ياواژه اي كه در مدرک آمده است، مي تواند براي نمايه در نظر گرفته شود. خصوصيت بارز نمايه سازي به زبان طبيعي، فقدان يک واژگان كنترل شده است. اين خصوصيت باعث تنوع زبان طبيعي در نمايه سازي شده است.

2. زبان آزاد Free Language

نمايه سازي به زبان آزاد عبارت است از نمايه سازي كه در آن هر واژه اي كه بتواند موضوع مدرک را خوب توصيف كند به عنوان اصطلاح نمايه برگزيده مي شود. خواه به وسيله پديدآورنده مدرک به كار رفته باشد و يا به كار نرفته باشد. در اين روش محدوديتي براي واژه‌هايي كه مي تواند در فرايند نمايه سازي به كار رود، وجود ندارد. اين روش مي تواند توسط انسان و يا رايانه انجام گيرد.

3. زبان كنترل شده مقيد (Controlled Language)

در اين گونه نمايه سازي معمولاً فهرستي مستند، مشخص كننده اطلاعاتي است كه ممكن است به موضوعات نسبت داده شود؛ به بيان ديگر در نمايه سازي به زبان كنترل شده، شخص اصطلاحات يا واژه هايي از يک فهرست واژگان را براساس تفسير ذهني كه از مفاهيم مندرج در مدرک دارد به مدرک اختصاص مي دهد

سيستم هاي نمايه سازي (Indexing Systems)

- سيستم هاي نمايه سازي وسايلي هستند كه به وسيله آنها از يک زبان نمايه‌اي براي ايجاد نمايه يا ساير ابزارهاي تجسسي استفاده مي‌شود. بنابراين به مجموعه اي از رويه هاي مقرر كه براي سازماندهي محتواي ركوردها به منظور بازيابي و اشاعه اطلاعات، مورد استفاده قرار مي‌گيرد سيستم نمايه سازي گفته مي شود. پس همارايي و پيش همارايي دو سيستم اساسي نمايه سازي به شمار مي رود.
- نمايه سازي همارا Co – Ordinate Indexing
- نمايه سازي همارا به شيوه اي از نمايه سازي گفته مي شود كه در آن محتواي موضوعي هر مدرک با بيش از يک شناسه يا اصطلاح تعيين و توصيف مي شود و رابطة بين اصطلاح‌ها به وسيله جور كردن و تركيب واژه‌هاي منفرد معين مي‌شود
- . نمايه سازي پس همارا Co – Ordinate Indexing – Post
- 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 يک اضافه مي نمايد و نيز تدابيري براي پيوستن آنها به يكديگر به دست مي دهد تا به وسيله آنها جوينده بتواند موضوع مدرک موردنظر خود را بيابد. به عبارت ديگر پس همارايي اشاره به عملي مي كند كه بعد از روند نمايه سازي انجام مي شود و مربوط به همارايي دو يا چند اصطلاح براي ساختن مفهوم موردنظر در مراحل بازيابي است.
- نمايه سازي پيش همارا Co – Ordinate Indexing – Per
- هرگاه بين دو يا چند جزء به طور تصنعي و ساختگي، پيوند برقرار كنيم به اين عمل و نوع آن، نمايه پيش همارا گويند. به عبارت ديگر در نمايه سازي پيش همارا، تركيب يا همارايي عناصر تشكيل‌دهنده موضوع مورد جستجو در هنگام نمايه سازي و به عبارتي پيش از بازيابي صورت مي‌گيرد

نمایه های قدیمی

- نمایه های گردان: اساس کار آن بر گردش کلیدواژه های عنوان است .
- با استفاده از stop list کلمات Noninformative حذف می شوند.

Kwic= نوعی نمایه عنوان است که در آن کلید واژه های عنوان با وسایل ماشینی انتخاب می شوند.

رایانه کار گردش عناوین مقاله های مجلات را با قرار دادن الفبایی عناوینی که واژه های مشابه دارند، انجام می دهد. مانند "عناوین شیمی"

- مثال: چگونگی افزایش تولید سیب زمینی با استفاده از کود حیوانی

عنوان ها بر اساس کلید واژه ها به ترتیب الفبایی مرتب می شوند.

Kwoc

- نوعی نمایه گردان است که کلید واژه ای از عنوان در ابتدا قرار می گیردو بقیه عنوان بعد از این کلمه می آید:
- بررسی رابطه کتابخانه و آموزش در نظام آموزشی ایران
- کتابخانه و آموزش در نظام آموزشی ایران، بررسی رابطه
- آموزش در نظام آموزشی ایران، بررسی رابطه کتابخانه
- نظام آموزشی ایران ، بررسی رابطه کتابخانه و آموزش در
- ایران ، بررسی رابطه کتابخانه و آموزش در نظام آموزشی

Permuterm

- یا جایگشتی اساس کار آن مانند کوئیک و کووک گردش کلید واژه های عنوان است.
- کلید واژه های عنوان به صورت دو تایی انتخاب می شوند و یک اصطلاح را تشکیل می دهند. گردش کار با دو کلید واژه است
- اصطلاح دوم در محدود کردن اصطلاح اول و اخص کردن آن نقش مهمی دارد.
- اصطلاحات جابجا می شوند و شناسه های اصلی الفبایی می شوند

Permuterm

- مثال :
- عنوان:آماده سازی وبرنامه ریزی برای ساختمان کتابخانه 14

آماده سازی/برنامه ریزی 14

آماده سازی/ساختمان کتابخانه 14

برنامه ریزی /آماده سازی 14

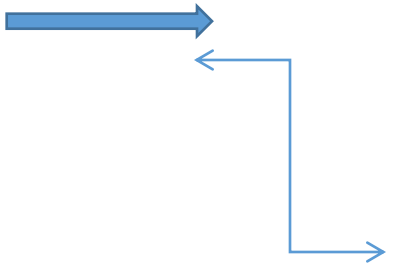
برنامه ریزی /ساختمان کتابخانه 14

ساختمان کتابخانه/ آماده سازی 14

ساختمان کتابخانه / برنامه ریزی 14

انواع نمایه انتهای کتاب

- **نمایه ساده**: مدخل ها به ترتیب الفبایی است و در جلوی آن جاینما وجود دارد. فاقد بیانگر است و هیچگونه تورفتگی وجود ندارد و تعداد جای نما های آن افزایش می یابد.
- **نمایه درون بافتی**: این نمایه به صورت **خطی و نحوی** است و به ترتیب پیدایش بیانگرها در متن است. بنابراین الفبایی نیستند. بیشتر مناسب متون علوم انسانی می باشد.
- مثال: کودکان مدخل
- ~ استثنایی 15- 17؛ ~ عقب مانده ذهنی 21-45؛ ~ تیزهوش 47
- بیانگر



نمایه انتهای کتاب

- **نمایه برون بافتی** : نمایه ای است که بیانگرها در بافت نحوی قرار نمی گیرند، بلکه در ذیل مدخل اصلی به صورت الفبایی می آیند. توالی صفحه ها مطرح نیست. کلماتی مثل حروف اضافه و غیره در آن استفاده نمی شود. برخی از بیانگرها در جای خود به عنوان مدخل قرار می گیرند .
- مثال: فهرست نویسی
- سی دی 36، 34، 32
- کتاب 25
- نرم افزار 12

تفاوت های نمایه برون بافتی و درون بافتی

- نمایه درون بافتی صورت نحوی دارد.مثلا میگوییم رفتار با کودکان ولی دربرون بافتی با را بکار نمی بریم و نمی دانیم آیا رفتار کودکان است یا رفتار با کودکان .
- نمایه درون بافتی بیشتر به حوزه علوم انسانی تعلق دارد.
- نمایه درون بافتی کل نگر هستند.
- نمایه برون بافتی بیشتر به حوزه های علوم و فنون مرتبط است و بیشتر بن مایه اصطلاحنامه قرار مي گيرد.

عناصر نمایه های درون بافتی و برون بافتی

- شناسه(heading):به همه مجموعه سر جمع شناسه گفته می شود.
- هر شناسه از چند عنصر تشکیل شده است:

- مدخل(entry):کلیت ورود یک کلمه یا نماد به شناسه را گویند.عنصر اصلی یک ترکیب است= کانون اصلی واژه ای که به عنوان نمایه به کار می رود.
- بیانگر(modifier):نقش محدود کننده دامنه معنای مدخل را دارد.تعیین کننده نوع یا چگونگی کانون است.
- جاینما (locator):محل منابع را نشان می دهد.
- ارجاع (Reference): برای ارجاع اصطلاح غیر مرجح به مرجح . گاهی به جای نما ارجاع نیز گفته می شود

گام های ایجاد یک نمایه

- تعیین خط مشی نمایه سازی
- خواندن سریع متن بدون یادداشت و علامت = اریب خوانی= درک کلیت موضوع)
- مشخص کردن کلید واژه ها با خط کشیدن زیر واژه ها یا نوشتن در حاشیه هرصفحه.
- انتقال هرکلید واژه (مدخل) بر روی یک برگه 5/7 در 5/12 وذکر شماره صفحه بر هر برگه
- الفبایی کردن مدخل ها و ادغام مدخل های تکراری
- ایجاد پیوند مفهومی بین مدخل ها
- استفاده از اصطلاح نامه ها برای ایجاد روابط

نمایه ها بر حسب روش تنظیم

- الفبایی : بیشتر نمایه های کنونی به ترتیب الفبایی است و شامل نمایه الفبایی اسامی اشخاص، سازمان ها و نیز موضوعات می باشد .
- زمانی : به ترتیب زمان از قدیم به حال می باشد و بیشتر برای نمایه های تاریخی به کار می رود
- رده ای یا موضوعی : که براساس رده ها یا سرعنوان موضوعی نظام مند مرتب می گردند. بیشتر در نمایه های علمی به کار میرود که البته می تواند یک نمایه الفبایی رده ای باشد.
- تکاملی: برای نمایه های زمین شناسی

نمایه استنادی

- نمایه استنادی شامل فهرستی از مقالات و یک فهرست فرعی تحت هر یک از مقالات منتشر شده است که به آن مقالات استناد کرده اند. به عبارت دیگر، در مورد یک مقاله خاص، نمایه استنادی مشخص می کند که این مقاله توسط چه مقالات دیگری که بعد از آن نوشته شده اند ، مورد استناد قرار گرفته اند. که علاوه بر نمایه اصلی، ممکن است نمایه های دیگری همچون مولف و نمایه موضوعی داشته باشد.
- در حقیقت بیان کننده ارتباط موضوعی درونی با مقاله هایی است که به آن استناد کرده اند.
- مزیت اصلی آن ارجاع استفاده کننده به جدیدترین مقالات است

نکات مربوط به ساختار زبانشناسی مداخل

- اسم یا عبارت اسمی فقط می تواند مدخل قرار گیرد(جز اصلی حتما باید اسم باشد و بیانگر صفت یا مضاف الیه مثل: انگشتر طلا،سپید دندان
- صفت و قید مدخل نمایه قرار نمی گیرند مگر جانشین اسم شده باشند. مثل بالا بر می تواند چون یک عنصر عینی است

- وارونه سازی ممنوع (به جز اسامی افراد)
- مثال: استثنايي ، کودکان (غلط است)
- خوارزمي، محمد (صحیح است)

ملاک انتخاب اصطلاح مرجح

- رواج داشتن . مثال: جامعه شناسي به جاي علم الاجتماع
- بومی بودن . مثال : آرمان گرايي به جاي ايده آليسم
- جديد و امروزي بودن . مثال: هواپيما به جاي طياره
- علمی بودن. مثال: اسيد سولفوريك به جاي جوهر نمك
- سرنام يا اختصار مشهور. مثال: يونسكو
- پرهيز از بكار بردن كلمات تركيبی به جز در مواردی كه اصطلاح قابل جدا سازی نيست مثل : آبله مرغان

ویژگی های خطی مشی: سیاست نمایه سازی کتاب)

- تعیین تعداد متوسط کلید واژه ها
- تعیین نوع نمایه (درون بافتی یا برون بافتی)
- زبان نمایه سازی(آزاد، کنترل شده، ترکیبی)
- نوع بازیابی(راهنمای استفاده)
- روشن ساختن بیانگر (جمع یا مفرد،تکلیف وضعیت کلمات خارجی...)
- سیاست مرجح یا نامرجح کردن بیانگرها
- ارجاعات و جای نماها
- بیان انواع نمایه موجود (موضوع، نویسندگان،...)
- استاندارد سازی (ابزارها)
- خودکار سازی نمایه(زبان برنامه نویسی، نرم افزار، چگونگی بازیابی)
- روزآمد سازی

نمایه سازی مجموعه

- نمایه سازی مدارک مختلف که توسط افراد متفاوت، در مکان های متفاوت و در زمان های متفاوت و احتمالا با واژگان متفاوت است و باید در شبکه واحدی مورد جستجو قرار گیرد. مانند: نمایه مقالات کتابداری بنابراین نوعی همگونی و هماهنگی باید پدید آورد که در عین جامعیت دستیابی ،موارد زائد بازیابی نشود. پس برای ایجاد همگونی 2 نکته مهم است : جامعیت و مانعیت

جامعیت و مانعیت
(Recall, Precision)

- شمول و عدم شمول
- یعنی: از نظر منطق جامع افراد یا افراد همگون زیر یک چتر بروند و مانعیت یعنی اغیار زیر این چتر نروند.
- تلاش می شود سطح هر دو (جامعیت و مانعیت)بالا برود و توازن آن حفظ شود
- یکی دیگر از موارد مهم مسئله ربط Relevance است یعنی اطلاعاتی که در نمایه است تا چه حد با نیازهای کاربر همخوانی دارد.
- ربط را کاربر تعیین می کند.

زبان های نمایه سازی

نظامهای نمایه سازی(زبان نمایه سازی).

در نظام‌هاي اطلاع رساني به زبان ساختگي و قراردادي اطلاق مي‌شود 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 ای از روشهای از پیش تعیین شده برای سازمان دهی، بازیابی و اشاعه اطلاعات

زبان های نمایه سازی

به دو دسته کلی تقسیم می شوند:

1. نظام های اصطلاح تعیین شده Assigned term دارای ابزارهای کنترل واژگان

2. ۲. اصطلاح مشتق Derived Term توصیفگر ها از متن گرفته می شود لذا آن را متن آزاد و طبیعی هم می گویند

زبان های نمایه سازی

- **نمایه سازی پیش همارا** : Pre Coordination Indexing
- هرگاه بین دو یا چند جزء به طور تصنعی و ساختگی، پیوند برقرار کنیم (ایجاد نحو) به این عمل و نوع آن، نمایه پیش همارا گویند.
- به عبارت دیگر در نمایه سازی پیش همارا، ترکیب یا همارایی عناصر تشکیل‌دهنده موضوع مورد جستجو در هنگام نمایه سازی و به عبارتی پیش از بازیابی صورت می‌گیرد:

مانند : سرعنوان های موضوعی

پرستاری –مراقبت های ویژه

زبان های نمایه سازی

- **نمایه سازی پس همارا** : post coordination Index ing
- شیوه ای از نمایه سازی که در آن نمایه ساز، سرشناسه ها را از مفاهیم بسیار ساده انتخاب می کند و تعدادی شناسه زیر هر یک اضافه می نماید و نیز تدابیری برای پیوستن آنها به یکدیگر به دست می دهد تا به وسیله آنها جوینده بتواند موضوع مدرک موردنظر خود را بیابد. ترکیب یا همارایی عناصر تشکیل‌دهنده موضوع مورد جستجو در هنگام بازیابی انجام می شود.
- مانند : اصطلاح نامه ها
- کتابشناسی تاریخ اصفهان

مزایای زبان طبیعی

- زبان متخصصان هر رشته است و برای ارتباط با سایر متخصصان بسیار مهیا تر است.
- حفظ ساختار طبیعی زبان
- روزآمدی در قیاس با زبان ساختگی.
- ناهمگونی های اصطلاحات را تحت سیاست واحد قرار می دهد. جامعیت را افزایش می دهد و از پراکندگی اطلاعات جلوگیری می کند.
- اشکال:نیاز به ارجاع دارد پس از حالت طبیعی خارج می شود .

مزایای زبان ساختگی

- بصورت یک بسته آماده به کامپیوتر داده می شود و بین مدرک وزبان کنترل شده یک پیوند برقرار میکند. به مراتب توانمند تر در دادن اطلاعات است هر چند مانعیت کمتری دارد.

- وقتی ایجاد سیاست می کنیم (مثل قوانین جدا نویسی و) خود نوعی کنترل است و زبان ساختگی می شود.

اصطلاحنامه

- **گنجواژه يا اصطلاحنامه**، مجموعه اصطلاحات يك رشته است كه ميان آنها روابط معنايي، رده‌اي، و سلسله مراتبي برقرار شده و توانايي آن را دارد كه موضوع آن رشته را با همه جنبه‌هاي اصلي و فرعي و وابسته، به‌گونه‌اي نظام‌يافته و به‌منظور ذخيره و بازيابي اطلاعات ارائه دهد .
- اصطلاحنامه معادل فارسي واژه انگليسي Thesaurus است كه ريشه در زبان لاتيني باستان دارد و به‌معناي گنجينه، ذخيره، و مجموعه به‌كار مي‌رود.
- اصطلاحنامه از نظر وظيفه و كاركرد، ابزار كنترل واژه‌ها به‌منظور برگرداندن زبان طبيعي مدارك به زبان مقيد است و از نظر ساختار، واژگان كنترل‌شده و پوياي زمينه‌اي خاص از دانش بشري است كه براي ذخيره و بازيابي اطلاعات آن حوزه به‌كار مي‌رود

اهداف اصطلاحنامه

اصطلاحنامه داراي هدف‌هاي اساسي زير است:

- 1) نماياندن ساختار زمينه معيني از دانش چنان‌كه هم نمايه‌ساز و هم جست‌وجوگر بتوانند از گستره آن زمينه و ارتباط ميان مفاهيم آن با انديشه‌هاي مرتبط آگاهي يابند
- 2) ارائه اصطلاحات استاندارد در زمينه‌اي معيّن.
- 3) برقراري نظام ارجاعات ميان اصطلاحات و رده‌بندي اصطلاحات به‌صورت سلسله مراتبي
- 4) تأكيد بر توجه به نيازهاي اطلاعاتي استفاده‌كنندگان
- 5) تعيين اصطلاحات مجاز و مشخص كردن حدود معاني اصطلاحات به‌منظور ايجاد هماهنگي در نمايه‌سازي

روابط ميان اصطلاحات

- . بايد توجه داشت كه ويژگي ذاتي اصطلاحنامه، توانايي تعيين و نمايش روابط معنايي ميان واژه‌هاست 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 ارزي** : Equivalance Relation
  ميان اصطلاح پذيرفته‌شده و اصطلاح پذيرفته‌نشده برقرار مي‌شود
- **مثال:** اصطلاح جامعه‌شناسي به جاي علم‌الاجتماع
- ب) **رابطه سلسله مراتبي** Heirachial Relation
- بيان‌كننده رابطه اعم و اخص ميان مفاهيم است كه در واقع، اصطلاحنامه‌ها را از واژه‌نامه‌هاي متداول متمايز مي‌كند .
- **رابطه همبسته يا همايند**: Associative Relation
- رابطه ميان دو اصطلاح كه به دليل وابستگي معنايي ، وجود يكي ديگري را نيز به ذهن متبادر مي‌كند. **مثال :** دو اصطلاح اسب و سوارکاري

ساختار اصطلاحنامه

- **اصطلاح اعم (ا ع):** Brother Term :اصطلاح عام‌تر
- مثال: بیماریهای داخلی
- ا. ع. کلیه ومجاری ادرار
- اصطلاح اخص (ا خ) : Narrow Term : اصطلاح خاص‌تر
- مثال: بیماریهای گوارش
- ا. خ. معده
- مری
- روده
- **اصطلاح وابسته (ا و)** : Related Term : اصطلاحي كه به اصطلاح برگزيده به نوعي ارتباط دارد.
- مثال: انگور
- ا. و. شیره

ساختار اصطلاحنامه

- **اصطلاح به کار ببرید (بك)** : Use : ارجاع از اصطلاح ناگزیده (نامرجح) به برگزیده (مرجح)
- مثال: کانسر
- بك. نئوپلاسم
- **اصطلاح به کار ببرید به جاي (بج):** Used for: ارجاع از اصطلاح گزیده (مرجح) به اصطلاح نا گزیده (نا مرجح) که البته مترادف است.
- مثال: نئوپلاسم
- بج. سرطان

ساختار اصطلاحنامه

اصطلاح راس

مثال: ميوه

راس. خوراكي ها

ياد داشت دامنه :Scope Note: يادداشتي براي توضيح مدخل كه داراي ابهامي و يا داراي چند معنا است و يا معني خيلي كلي دارد استفاده مي شود .

مثال: فرهنگ (آداب و رسوم)

Top Term : اصطلاح سرمقوله مجموعه اي از اصطلاحات

انواع رابطه سلسله مراتبي

- رابطه سلسله مراتبي خود شامل سه نوع است :
- 1. رابطه جنس و نوع: Generic : اين رابطه دقيق ترين و قطعي ترين رابطه سلسله مراتبي است كه با كلمات همه براي جنس و بعضي براي نوع قابل شناسايي است.
- مثال : رابطه ميان اصطلاحات : نشخوار كنندگان (جنس)
- بز(نوع)
- گاو (نوع)
- گوسفند (نوع)
- همه بز ها نشخوار كننده هستند
- بعضي از نشخوار كنندگان بز هستند.

انواع رابطه سلسله مراتبي

- **2. رابطه كل و جزء :** Whole – part : كه نشان دهنده رابطه يك جزء را با كل آن مشخص مي كند و نوعي رابطه اعم و اخص است.
- مثال : پا (جزء)
- بدن (كل)
- **3. رابطه مصداقي** : Istance Relation : اين ارتباط بيشتر در مورد اسامي خاص است و رابطه اي است كه جنس و نوع نيست بلكه بيانگر يك مصداق است.
- مثال: دماوند (مصداق يك قله)

تفاوت واژه نامه و اصطلاح نامه

- واژه نامه ها : نظم الفبايي دارد و به دنبال هم مي آيند. ولي مجرد بوده و حالت خنثي دارند .

- اصطلاحنامه : معمولا نظم الفبايي – رده ا ي دارد و كلمات با هم مرتبط بوده و روابط معنايي دارند.

واحدهای اندازه‌گیری حجم در کامپیوتر

- کوچک ترین واحد حجمی دیجیتال، یک بیت (bit)است
- بایت :برابربا 8 بیت
- کیلوبایت :برابر است با 1024 بایت
- مگابایت: برابر است با 1024 کلیوبایت(1 مگابایت: یک کتاب 400 صفحه ای)
- گیگابایت: برابر است با 1024 مگابایت
- ترابایت (TB)
- پتابایت (PB)
- اگزابایت (EB)
- یوتابایت (YB)
- هِلابایت (HB)
- Yotta= 24 صفر را در مقابل عدد یک
- Hella=27 صفر را در مقابل عدد یک درسال 2020

ISBN(International Standard Book Number)

رده بندی با کمک کد شابک، شماره استاندارد بین المللي کتاب است. این شماره در ابتدا کدي 1 رقمي بود که در سال 1167 توسط آقاي گوران فاستر ابداع شد و پس از آن در سال 1170 به ده رقم تغییر یافت و نهایتاً از سال 2007 تاکنون یک کد 13 رقمي میباشد. شابک 13 رقمی دارای 5 بخش است که هر بخش با یک – از هم جدا میشود:

بخش اول از سمت چپ کشور است است.

بخش دوم مربوط به کتاب است.

بخش سوم کد خاص انتشارات کتاب است

رقم بعدي کد خاص آن کتاب در انتشارات است

بخش آخر در اصطلاح Checksumارقام دیگر است.

Checksum:روشي براي تشخيص خطا در انتقال داده از مبداء به مقصد است

روش به دست آوردن Checksumدر شابک 13 رقمی: مثال: در کد شابک زیر، رقم checksumرا به دست آورید:

۹۷۸-۹۶۴ -۵۳۱-۰۸۳ -....

پاسخ:

- از سمت چپ، ارقام را یک در میان در ۱ و ۳ ضرب کنید.
- نتیجه را با هم جمع کرده و بر ۱۰ تقسیم کنید.
- باقیمانده را از ۱۰ کم کنید تا Checksum به دست آید.

9*1 + 7*3 + 8*1 +9*3 +6*1 + 4*3 + 5*1 + 3*3 + 1*1 + 0*3 + 8*1 + 3*3 =115

115 mod 10= 5

10-5=5

نمایه سازی (indexing)موتور های جستجو چیست؟

- جمع آوری، تجزیه (پردازش) و ذخیره سازی داده ها در مورد یک **url** به منظور تسهیل سریع و دقیق بازیابی اطلاعات میباشد.

انواع وب

وب سنتی (Traditional Web) این وب همان وب اولیه است که توسط آقاي تيم برنرزلي ابداع شد. حاوي صفحات سادهي بدون جذابيت و بدون هوشمندي.

2-وب اشتراکی(Social Web) در این دوران سایتهاي اشتراک جمعي یا شبکههاي اجتماعي)(Networks Socialمثل فیسبوک، توئیتر، اینستاگرام و... ظهور کردند و کاربران توانستند داده هاي خود را به راحتي با یکدیگر به اشتراک بگذارند و بنابراین دسترسي به داده ها آسانتر شد.

3-وب وب معنایی یا وب هوشمند(Web Semantic) در این وب پیش بیني میشود صفحات وب بیش از گذشته براي موتورهاي جستجو به طور کامل شناخته شده باشند، بنابراین نسبت به نیاز کاربر بهترین نتایج یافته و نمایش داده خواهد شد.

دو عامل مهم در نمایه سازی وب

الف- زبان نمایه سازی: زبان نمایه سازی آن دسته از واژگان نمایه سازی است که در نظام خاصی از ذخیره و بازیابی مورد استفاده قرار می گیرد. "زبان " می تواند طبیعی، یعنی زبان مدرک نمایه سازی شده باشد، یا ساختگی یا کنترل شده باشد

ب- نرم افزارهای نمایه سازی وب: بستگی به اینکه در وب مورد نظر چه اطلاعاتی را می خواهیم نمایه سازی کنیم (اطلاعات می توانند یک سند، یک Full text، تصاویر و ... باشند

انواع نرم افزارهابرای نمایه سازی وب

نرم افزارهای نمایه سازی وب

Advanced Java Tree Menu
PHP Lightning Portal (PLP)
PHP Portal Builder (PPB)
ActMon Password Recovery XP
Internet Macros Web Test Recorder

نرم افزارهای نمایه سازی لینکها

PHP Lightning Portal (PLP)
PHP Portal Builder (PPB)
Registry First Aid
IEManager
Advanced Java Tree Menu

مراحل نمایه سازی کامپیوتری :

ذخیره اطلاعات: ذخیره اطلاعات بخشی ازمدرک است که جهت نمایه سازی ازآن استفاده می شودو شامل عنوان یا چکیده یا فهرست مندرجات یا حتی کل مدرک می باشد.

تشخیص اطلاعات: تشخیص کلمات وجملاتی است که به صورت الکترونیکی ذخیره شده است و بستگی به توانایی سیستم از لحاظ سخت افزاری و نرم افزاری دارد.

تفکیک اطلاعات به پاره های اطلاعاتی:کل اطلاعات یک مدرک که ذخیره شده است به جملات یا کلمات شکسته شود.

مقایسه پاره های اطلاعاتی: مقایسه جملات یا کلماتی است که در مراحل قبل شکسته شده است.

طبقه بندی: اولویت بندی پاره های اطلاعاتی برحسب نیاز و با توجه به برنامه ای است که قبلا به سیستم داده شده است.

انتخاب :آن دسته از اصطلاحاتی که در اولویت قرارگرفته اند به عنوان توصیفگرانتخاب می شوند و اصطلاحات نامناسب حذف می شوند.

رویکردهای اصلی در نمایه‌سازی

رویکرد سندگرا: ایده اصلی در این رویکرد آن است که نمایه‌ساز، محتوای موضوعی سند را تنها براساس بررسی و تحلیل خود آن سند انجام دهد و البته هدف از این کار، بازنمایی سند به صادقانه‌ترین شکل ممکن و اطمینان از اعتبار بازنمایی موضوعی برای مدت‌زمان طولانی است. در این رویکرد، نمایه‌سازان کاربردهای متفاوت هر فرد از سند مورد بررسی را نادیده می‌انگارند و واژگان نمایه را تنها براساس بررسی و تحلیل سند انتخاب می‌کنند.

رویکرد کاربرگرا: در رویکرد کاربرگرا، واژگان نمایه برحسب نیازها و پرسش‌های کاربران انتخاب می‌شوند، حال آنکه در رویکرد سندگرا در هنگام انتخاب واژگان نمایه‌ای، نیازهای کاربران مورد توجه قرار نمی‌گیرد. ایده اصلی در این رویکرد آن است که نمایه‌ساز می‌بایست در تعیین محتوای موضوعی سند و همچنین انتخاب واژگان نمایه‌ای، نیازهای اطلاعاتی کاربران و واژگان مورد استفاده توسط آنها را در ذهن داشته باشد. در این رویکرد، نمایه‌ساز نیازمند است تا دانش لازم را برای شناخت نیازهای کاربران به‌دست آورد تا بتواند محتوای موضوعی سند را تعیین کند.

روش های نمایه سازی ماشینی

الف. روش های زبان شناختی
ب. روش های آماری
ج. روش های مبتنی براحتمالات

نمایه سازی رایانه ای عمدتا به دوشیوه نجام می شود

الف- حفظ اصطلاح: در سال ۱۹۶۳ توسط دکتر سوزان آرتاندی به کار گرفته شد. سیاهه یا فهرستی به رایانه داده می شود که حاوی کلمات کلیدی است و به نظام گفته می شود، هریک از کلمات این لیست را که درمتن بود ،حفظ کند و آنرا به عنوان اصطلاح نمایه ای انتخاب کند.
دراین روش استفاده از اصطلاحات از پیش انتخاب شده راهنمای خوب ومفیدی است و امکان انتخاب واژه های استفاده شده به وسیله مولف مدرک را فراهم می کند.

ضعف این روش :این است که تضمینی وجود ندارد تا تمامی مفاهیم موضوعات جدید ومهم نمایه سازی شوند.

ب-حذف اصطلاح:درشیوه حذف اصطلاح یک سیاهه بازدارنده به رایانه داده می شود که کلمات فهرست را هرکجا که بودحذف کند و بقیه را به عنوان کلید واژه انتخاب نماید.نمایه سازی کوئیک و کواک از نوع حذف اصطلاح هستند.

(کامپیوتری) نمایه سازی ماشینی

نمایه سازی رایانه ای :یکی از انواع پرکاربرد نمایه سازی که سرعت و یکدستی را به دنبال دارد،

نمایه سازی ماشینی : عبارت ست از انتخاب کلید واژهای یک اثربوسیله روش های ماشینی(استفاده از کامپیوتر) برای بیرون آوردن و نشان دادن واژه های نمایه بدون دخالت انسان در حالی که یک با برنامه و سیاست کار آن، به ماشین داده شده است.

نمایه سازی رایانه ای , انجام کلیه مراحل نمایه سازی اعم از :

- انتخاب و استخراج اصطلاحات نمایه ای ازمتن ،
- مدخل آرایی
- ارائه جاینماهای هر مدخل
- چاپ نمایه توسط رایانه وبدون دخالت انسان را نمایه سازی رایانه ای گویند.
- نمایه سازی کامپیوتری، رایانه ای ، ماشینی و خودکار همگی با هم مترادفند و بجای یکدیگر به کار می 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 ياماشيني ناميده مي شود.

جهت استفاده از روش نمايه سازي ماشيني، داده يا بايد به صورت ماشين خوان درآيند.

در اين نوع نمايه سازي همه امور ازانتخاب كليد واژه ،شماره گذاری ،ترتيب بندي و غيره توسط كامپيوتر انجام مي گيرد.

نمايه سازاني كه با اصول تحليل و طراحي سيستم و برنامه سازی كاربردی آشنايي دارند، مي توانند در زمينه كامپيوتری سازی نمايه سازی مو ثرباشند

روش کار

کامپیوتر مفاهیم مهم را که بارها تکرار شده اند و جزو کلمات غیر موضوعی زبان نیستند، به علاوه اسمهاي اشخاص، مکانها و غیره را به عنوان کلید واژه در نظر مي گیرد.

- این کارتوسط یک نرم افزار به نام نرم افزار بسامدی استخراج مي کند.

مثلا یک مفهوم که در ذیل یک بخش ازمتن بارها تکرارشده باشند، محتملا کلید واژه است، مگر آن که جزو وا ژگان غیر موضوعی باشد.

- وا ژگان غیر موضوعي مانند(است، که ، را و...) توسط یک فهرست ایستا مشخص و نادیده گرفته مي شود.

-قسمت دیگر نرم افزار حاوي اسمهاي خاص است. نرم افزار با استفاده از این

دادگان ، اسمهای خاص متن را تشخیص می دهد و به عنوان کلید واژه درنظر می گیرد.

فرایند انتخاب کلید واژه درنمایه سازی خودکار

1.شناسايي واژ ه های انفرادی از متن که تحلیل واژ گان نامیده میشود

2.برداشتن وا ژه های با بسامد تکرار بالا که در ارائه محتوای متن بی تأثیرند، با استفاده از فهرست واژه های غیرمجاز.

3.تبدیل واژه های باقي مانده به شکل ریشه آنها یعنی حذف پسوند یا پیشوندها تا هر کلمه تا حد ریشه اش کوتاه مبدع آن است.

4.محاسبه رایانه اي بسامد ریشه هایی که در متن تحلیل شده اند، به منظور تعیین تابع ارزش گذاري بر ریشه.

5.ریشه هایی که ارزش گذاری بزرگتری دارند، براي متني که در آن ظاهر شده، به عنوان کلید واژه تعیین مي شود.

ادامه

شرکت زافتکس نوعي برنامه کامپيوتري با عنوان IDXطراحی کرده که به ارائه خدمات نمایه سازی ماشيني مي پردازد.

این نرم افزار از روش استفاده از وا ژه نامه بهره می برد.

این نرم افزار امکان تبدیل وا ژگان به ریشه آنها را جهت بازیابي بعدی فراهم مي کند علامت گذ اری و محدودکردن واژ ه های ناخواسته را انجام مي دهد

.شکستن وا ژه های مرکب و ترجمه و انجام عمل ارجاع و مترادف سازی و ساخت عبارات را نیز انجام میدهد.

ماشین امکان تشخیص را تنها از طریق تطبیق واژه های استخراج شده از متن یا منتسب شده به متن با فهرستي که وا ژه های غیرمجاز نامیده مي شود، به دست مي آورد.

در اختيار داشتن فهرستي از اين واژ ه ها و ارائه آنها به برنامه رايانه
اي براي ممانعت از ورود آنها به فهرست وا ژه هاي مفهومي مطلوب براي نمايه شدن، يكي از اقدامهاي سودمند در نمايه سازي خودكارمبتني بر كليد واژه هاست ..

روش های نمایه سازی ماشینی

الف. روش زبان شناختي:

اين روشها مي كوشند با كمک تحليل های شكل شناسي و ساختار نحوي مدرك توصيفگرها را استخراج نمايند.

1. تجزيه و تحليل ريخت شناسي كه برمبنای ريشه لغات عمل مي كند.
2. كلمات بدون بار معنايی موجود در سياهه بازدارنده را حذف ميكند.
3. شكلهاي دستوری صرف كلمه را به يک شكل می آورد.
4. ضمایر را براساس اسامي مربوط به آنها مرتب می كند.
5. تجزيه و تحليل نحوي كه درسطح جملات امكان پذ ير است و برمبنای علامتهاي نحو لغات انجام مي شود.
6. تجزيه و تحليل معنا شناختي كه درسطح مدارك مشترك در يک پايگاه صورت مي گيرد . ارتباطات معنايي موجود در يک مدرك شناسايي مي شوند تا متون مشترك بتوانند به صورت واحدهاي هم معني تجزيه شوند.

- **ب. روش های آماری:**

1- مشخص مي كندكه معنی هرمفهوم منفرد در مدرك با حضور آن در جايگاه هاي مختلف مدرك ارتباط تنگاتنگ دارد. بنابراين لغت درون متن شمارش مي شوندو ارتباط آنها ارزش گذ اری می شود.هدف آمارگيري از اطلاعات آن است كه لغات داراي بارمعنايی در مدرك به عنوان توصيفگر انتخاب شوند.

2- روش های آماری عمل براي بالابردن جامعيت به كار گرفته ميشوند. درحاليكه روشهاي زبان شناختي در جهت بهبود مانعيت كاردارند.

ج. روش های مبتنی براحتمالات:

در اين روشها تئوري احتمالات برا مدل سازی رياضی مراحل بازيابي به كارگرفته مي شود. درحاليكه در توزيع آماري اصطلاحات يک مدرك مورد استفاده قرار مي گيرد.اين روش با عمليات رياضی مفروضات ساده و مطمئني را ارائه مي دهد. فرض برآن است كه مدارك براساس ميزان ربط درهنگام بازيابي مورد ارزيابي قرارميگيرند.

فهرست ایستا , یا واژگان غیر مجاز

- تحلیل کلمات یک متن نشان مي دهد گروهی از کلمات بی اهمیت وجوددارد که به فراوانی در متن ظاهر مي شود مانند(ی ، به، نه، برای، با، چه کسي، چه موقع، است، آن).
- گروهي نیز وجود دارد که بندرت در متن متي آیند و ممکن است نشان دهنده محتوای اطلاعاتي متن نباشند
- این دسته از واژ ه یا به تنهایي بارمعنایي ندارند , بود یا نبود آنهانه تنها در پرسش کاربر تأ ثیري ندارد بلکه در میزان ربط یاعدم ربط مدارك بازیابي شده نیز تأ ثیری ندارد.
- این وا ژه یا باعنوان وا ژه ها ی غیرمجاز برای ورود ب نمایه معرفی می شوند.

مزایای تهیه لیست واژه های غیر مجاز

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

چند نمونه از نرم افزارهای معروف نمايه سازی كه در سيستمهاي كامپيوتر هاي شخصي كار می كنند به شرح زيراست:

INDEXING RESEARCH محصول شركت CINDEX

INDEX AID محصول شركت Santa Barbara Software Products

INQUIRY محصول شركت Indexer assistant

عوامل موثر در نمايه سازي خودكار

محدوده ركورد

- اولين تصميم گيري مهم براي تهيه ير 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.

محدوده اصطلاحات

- تعيين حد و حدود يک واژ ه از ديگر مسائلی است كه در نمايه سازی خودكار بايد به آن توجه شود. در نظام های نمايه سازی دستتي ، گزينش كلمات براي نمايه به سهولت انجام مي شود. اما در نمايه سازی خودكار از آنجا كه ماشين از هوشمندی لازم براي انتخاب كلمات برخوردار نيست بنابراين بايد حدود كلمه را تعريف كرد. معمول حدود كلمات نمايه را با استفاد ه از عالئم نقطه گذ اري تعريف مي كنند.
- به طور معمول ، فاصله بين كلمات و علائم دستوري و نقطه گذ اري به عنوان مرز كلمات در نظر گرفته مي شود. روش های تعيين حد و حدود كلمات در نمايه سازي خودكار، بر اساس نوع برنامه و ميزان پيشرفتگی آنها متفاوت است.

لنکستر نمایه سازی خودکار را به دو دسته استخراجی و تخصیصی تقسیم می کند

نمایه سازی استخراجی

ساده ترین روش نمایه سازي در پایگاه هاي اطلاعاتي ، روش نمایه سازی استخراجی است که در آن وا ژه یا برای قرار گرفتن در نمایه ، توسط رایانه از متن استخراج مي شوند. در این روش عموما بسامد تکرار وا ژه در یک رکورد یا مقاله تعیین شده و کلماتي که بسامد تکرار آنها زیاد است درمتن نمایه قرار مي گیرند.

نمایه سازی تخصیصی

انتسابي فرآیند پیچیده اي است که با استفاده از تحلیل های آماری ،کلمات و اصطلاحات به مدرك منتسب مي شوند رایانه براي نمایه سازی از اصطلاحنامه یا کنترل واژ گان بهره مي گیرد.

انواع نمایه استخراجی

نمایه سازی با استفاده از فهرست کلمات ممنو عه : در این روش درهنگام نمایه سازی ، رایانه تمام کلمات متن را استخراج میکند، سپس کلمات ممنوعه را حذف و بقیه کلمات را در یک نظام الفبايي مرتب میکند.

نمایه سازی بسامدی : در این روش، بسامد تکرار کلمات در یک رکورد مورد بررسي قرار مي گيرند و براساس بسامد تکرار درفهرست کلمات نمایه قرار مي گيرند.

نمایه سازی ریشه یابی : در بعضی از سیستم های نمایه سازي استخراجی ، از پسوند یا ریشه کلمات استفاده مي شود. در این روش ریشه یا پسوند کلمات جایگزین مجموعه ای از کلمات هم ریشه یا پسوند مشترك میشود. الگوریتم های ریشه یابی مختلفی چون الگوریتم های موضوعي خاص مانند الگوریتم های پزشكي وجود دارند. الگوریتم Lovins 260 پسوند .

- نمایه سازی بر اساس وزن دهی : در این روش ، کلمات بر اساس محل قرار گرفتن خود در متن مثل عنوان ، چکیده و(...امتیازدهی می شوند.

حضور کلمات در بخش های مختلف رکورد، امتیازات متفاوتي دارد.

در نظام های رایانه ای بیشتر از روش های نمایه سازي استخراجی استفاده مي شود.

1- يكي از عمده ترين مشكلات اين نمايه يا به ويژه هنگام استفاده در پايگاه های اطلاعاتي ، عدم بازيابي اطلاعات به دليل نبودن آن كلمه درخواستی در نمايه پايگاه اطلاعاتي است.

2- 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 بی خبرند.بنابراين ، بسياري از مدارك مرتبط از دست مي روند.

براي رفع اين معضل طراحان پايگاه های اطلاعاتي توانايي های نمايه ای را با توانايي نرم افزاري درهم مي آميزندبه دو روش زير:

الف: نمايش نمايه و انتخاب واژه درخواستی توسط خود بهره گير است .

ب:استفاده از نظام بازخورد مرتبط است . اين روش به بهره گيران اجازه می دهد تا مدارك مرتبط را برگزينند. سپس ازسيستم مي خواهند تا با توجه به اين مدارك ، مدارك مرتبط بيشتری را بازيابي نمايند. امروزه اين روش در اينترنت و پايگاه های اطلاعاتي تمام متن كاربرد فراوانی دارد .

نمایه سازی تخصیصی

- درواقع در نمایه سازي تخصيصي براي هر وا ژه "پرونده ای " از کلمات وعبارات مرتبطی که به نظر مي رسد تهيه مي شود.

بنابراين مي توان از برنامه ای رايانه ای براي انطباق عبارت های مهم درمدرك با اين مجموعه پرونده يا استفاده کرد و درصورت انطباق واژ ه موجود در مدرك با وا ژه های موجود در پرونده های کلمات ، اصطلاح نمايه ای راانتخاب نمود .

- هر چه نمايه سازي تخصصي تر باشد صرفه اقتصادی آن در يک سيستم خبره بيشتر می شود .

- سيستمي که از يک نمايه سازي تخصصي استفاده می کند می تواند به عنوان يک سيستم کمي خبره از آن نام برد.

تعریف وب مرئی و نامرئی

وب مرئی : صفحات ثابت با دسترسي آزاد

وب نامرئی:تشکیل شده از صفحات وب متحرك و یا قابل دسترس به صورت محدود مثال با استفاده از رمز ورود ..

مهارت های اطلاع یابی

مهارت در بازيابي اطلاعاتی

مهارت در ارزيابي اطلاعاتی

مهارت در سازماندهی اطلاعاتی

مهارت در تبادل ارتباط.

روشهای جستجوی اطلا عات در محیط وب

ساده

پیشرفته

کنترل واژ گان

اصطلاحنامه

فرایند جستجو

ارزیابي پایگاه یا

جامعیت

مانعیت

منابع موجود در وب نامرئی

بخش بزرگي از وب وجود دارد که عنکبوت های موتورهای جستجو آن یا را نمایه نمی کنند یا نمی توانند نمایه کنند وعبارت اند از سایت های داراي رمز عبور، اسناد موجود در پشت سامانه های حفاظتی، فایل های pdfاز متون آرشیو شده، وابزارهای تعاملی نظیر ماشین حسابها و برخی واژه نامه هاو همچنین محتویات بعضي از پایگاه های اطلاعاتی، منابع محافظت شده از طریق اسم کاربر و گذ رواژه، منابع و صفحات وب بدون پیوند، و صفحات افزون بر حداکثر تعداد صفحات قابل مرور درنتایج بازیابي

اهمیت وب نامرئی

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

1 . بهترين موتورهای كاوش فقط قادر هستند كه حدود 16 درصد از اطلاعات موجود در وب را بازيابي كنند و بنابراين 84 درصد آنها جزو وب نامرئي به حساب مي آيند

2 . اندازه وب نامرئي تقريباً 500 برابر وب مرئي است: وب نامرئي 550 ميليون سند، و وب مرئي تقريباً 1 ميليون سند را دارا مي باشد .

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

بخش های مختلف وب نامرئی

1 . وب مات یا تاریك
2 -وب عمیق
3-وب خصوصی
4-اینترنت واقعاً پنهان

1 . وب مات یا تاریك

- بخشي از فضاي وب نامرئي به وب مات موسوم گرديده كه می توانسته مورد استفاده كاربران قرار گيرد، اما به دليل زير اين اطلاعات در خارج از دسترس كاربران قرار گرفته و موتورهای كاوش نمي توانند انها را بازيابي كنند:
- -از آنجا كه اولاً محيط وب دائماً در تغيير است و هر روز منابع واطلاعات جديد به آن افزوده مي گردد
- -ثانياً صفحاتی در وب وجود دارند كه هيچ پيوندی بين آن يا با منابع ديگربرقرار نشده، خزنده های موتورهای جستجو قادر به يافتن اين صفحات و همگام نمودن خود با اين حجم عظيم اطلاعات نيستند.
- به دليل محدوديت توانايي، نرم افزارهای خزنده فرصت كافي برای روزآمدسازی صفحات جديد وب را ندارند. موتورهای كاوش نيز امكان روزامدسازی حجم عظيمی از اطلاعات و منابع جديد را ندارند و به همين دليل بسياری از اين اطلاعات از حوزه موتورهای كاوش دورمی مانند .

- محدوديت توان مالي بسياري از موتورهای كاوش سبب گرديده كه موتورهای كاوش نتوانند تمام صفحات وب سايت ها را نمايه سازی كنند، چرا كه برای آن يا هزينه های زيادی دارد

- بنابراين موتورهای كاوش بنا بر سياست های خودشان، تنها بخشی از وب

سايت ها يا لايه های بيرونی آنها را نمايه سازی می كنند. بنابراين هميشه بخش عظيم لايه های درونی وی سايت ها پنهان می مانند

2-وب عمیق

به مجموعه اي از اطلاعات الكترونيكي پيوسته اطلاق مي شود كه بسياري از پايگاه های اطلاع رساني، آنها 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 این پایگاه یا معمولاً خارج از حوزه جستجوی موتورهای کاوش قرار دارند هر یک ازاین پایگاه یا صفحه جستجوي مبتني بر وب دارند. که امکان جستجو در آنها برای کاربران رافراهم می کند، اما خزنده های موتورهای جستجو توان ورود به آنها را ندارند و در نتیجه حجم انبوهی از اطلاعات، نمایه نشده باقي می ماند.

مثال

اگر یک متخصص موضوعي مثلاً یک دانشجوی رشته پزشكي(بخواهد خود رابه موتورهای کاوش معمولي محدود کند و نتواند به پایگاه های اطلاعاتي تخصصي مراجعه نماید یا از وجود آن یا آگاه نباشد، از دسترسي به حجم انبوهی از اطلاعات محروم خواهد ماند . بنابراین کاربر باید در این موارد از طریق موتورهای جستجو،پایگاه های مرتبط با موضوع خود را شناسایی کند و سپس،جداگانه به جستجو در آن یا بپردازد تا از دسترسی به وب عمیق باز نماند

3-وب خصوصی و وب ملکی

بخشي ديگر از وب نامرئي وجود دارد كه چون اطلاعات موجود در آن جزو دارایی های شخصي يا خصوصي سازمان ها يا افراد مي باشد، از حوزه دسترسي موتورهای جستجو پنهان است. مثلاً در برخي از سازمان يا مؤسسات خصوصي يا دولتي، به دلايل امنيتي از اطلاعات مربوط به مسائل كاري و سازماني و پرسنلي خود حفاظت مي كنند اجازه دسترسي به آنها را به ديگران نمي دهند و فقط كساني كه داراي اسم كاربر و گذر واژه هستند می توانند از آنها استفاده كنند اين بخش، وب خصوصي محسوب می گردد بخش ديگر، منابع اطلاعاتي از قبيل نشريات الكترونيكی مبتني بر وب می باشند كه دسترسي به آن يا از طريق پرداخت حق اشتراك وخريد محصولات اطلاعاتی شركت های مختلف صورت مي گيرد و وب ملكي ناميده مي شود.

4 . اینترنت واقعاً پنهان

بخش ديگري از وب پنهان وجود دارد كه بنا به مسائل فني وناكارآمدي ابزارهای جستجو، ازدسترسي كاربران دورمانده است.بسياري از موتورهای جستجو قادر به بازيابي اطلاعات متني htmlهستند، ولي توانايي بازيابي فايل های pdfرا ندارند،يا به دليل كمبود منابع مالي وفني از جستجوي فايل های غيرمتني صرف نظر كرده اند. بنابراين منابع اطلاعاتي متنوعي نيز دروب وجود دارند كه تنها به دليل محدوديت های فناوران يا مالي

موتورهای جستجو، از حوزه كاوش آن يا و در نتيجه از دسترس كاربران دور مانده اند.

دلائل عدم بازیابی و نمایه سازی وب نامرئی توسط موتورهای کاوش

1 . دلایل فنی:

بسیاري از موتورهای کاوش به دلیل محدودیت های نرم افزاري توانايي روزآمدسازي اطلاعات جدید وب را ندارند . باید یادآور شد که هنوزهیچ موتور کاوشي ادعا نکرده است که قادر به گسترش حوزه کاوش خود به تمام محیط وب مي باشد و همیشه این موتورها یگ گام از سرعت روزافزون اطلاعات عقب تر هستند .

2 . دلایل بودجه ای:

همانطور که قبلاً اشاره شد فرآیند نمایه سازي تمام صفحات وب، هزینه بر خواهد بود و موتورهاي کاوش نیز بنا به محدودیت بودجه ناگزیرند فقط بخشي از وب سایت ها رانمایه سازي کنند

3 . دلایل اجتما عی و حقوقی:

از آنجا که اطلاعات موجود در وب در دسترس عموم قرار مي گیرد، بسیاري از افراد و سازمان یا به دلیل صرف بودجه هاي کلان در راه اندازي سایت ها و پایگاه هاي اطلاعاتي خود، حاضرنیستند این اطلاعات را به صورت رایگان در اختیار همه بگذارند.

البتّه این از لحاظ اجتماعی و حقوقی جز حق مسلّم آنها است.

شیوه های اطلاع یابی در وب نامرئی

در حال حاضر ابزارهايي به وجود آمده اند که منابع وب نامرئي را شناسايي، و کاربران را به سايت هاي مناسب راهنمايي مي کنند .اين رويکرد توسط بزرگراه هاي اطلاعاتي و کتابخانه هاي مجازي پذ يرفته شده است بطوري که فقط توصيفي از پايگاه هاي اطلاعاتي و مجلات نامرئي را ارائه مي کنندي مثل سايت Invisibleweb که فهرستي از منابع نامرئي را، و سايت Completeplaset Completeplaset که فهرستي از تقريباً 40000 پايگاه اطلاعاتي وب نامرئي را ارائه مي دهند.

برخي ديگر از ابزارهاي اطلاع يابي نيز که تاکنون معرفي شده اندو ما مي توانيم با استفاده از آن يا به اين اطلاعات دسترسي پيداکنيم به شرح زير است:

دروازه های اطلاعاتی

دروازه هاي اطلاعاتي مجموعه اي از پايگاه يا و سايت يا 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 Academic Information Academic Information

Digital Librarian Digital Librarian

Gaary price direct search price direct search price direct search Infomine

Internet public Internet public Library

از مزيت هاي عمده استفاده از دروازه هاي اطلاعاتي اين است كه براي ايجاد آنها افرادي با دانش موضوعي خاص، در اينترنت جستجو كرده اند و به پالايش اطلاعات مفيد از غيرمفيد پرداخته اند

پرتال ها یا پایگاه های اطلاعاتی خاص موضوعی

مجموعه اي از پايگاه هاي اطلاعاتي خاص موضوعي هستند كه به يک موضوع خاص اختصاص دارند و به وسيله دانشمندان،محققان، متخصصان، مؤسسات دولتي، شركت هاي بازرگاني وكارشناسان موضوعي، افراد بسيار علاقه مند يا داراي دانش حرفه اي و اطلاعات وسيع در حوزه خاص ايجاد مي شوند (Oxford uni. Libraries)از پورتال های در هنگام جستجو براي موضوعات خاص مانند پيوند هاي خبري، فايل هاي چندرسانه اي،آرشيوها، فهرست هاي پستي اشخاص، شغل يابهاو هزاران پايگاه اطلاعاتي كه به موضوعات خاص اختصاص دارند استفاده مي شود.

ویژگی پایگاه های اطلاعاتی

1 . جستجوپذير و مرورپذير، حاوي توصيف منابع اينترنتي در يک زمينه موضوعي خاصي

2 . مشتمل بر معيارهاي شفا و تعريف شده براي ارزيابي کيفيت منابع اطلاعاتي، به جاي انتخاب بدون ارزيابي

3 . دخالت کتابداران و متخصصان موضوعي در ايجاد آني

4 . توليد دستي رکوردهاي آن، براي توصيف بامعنا و اطلاع بخش از منابع اطلاعاتي

5 . فهرستنويسي و رده بندي منابع اطلاعاتي با استفاده از شيوه هاي سنتي کتابخانه اي به منظور بازيابي مو ر

ابرموتورهای کاوش :

گسترش پذ يري حوزه هاي جستجو نيز يکي از شيوه هاي دسترسي به وب نامرئي شمرده مي شود که نمونه آن، استفاده ازابرموتورهاي کاوش است. اين ابرموتورها خود، موتورهاي

جستجوي واقعي نيستند. بلکه به کاربران اين امکان را مي دهند که کليدوا ژه هاي خود را همزمان توسط چند موتور، مورد کاوشقرار دهند و نتايج جستجوي همه آنها را با هم ببينند .

عوامل هوشمند :

اين عوامل يوشمند از ابزارهاي بازيابي در اينترنت هستند که براي اجراي کارهاي بخصوص به کارگرفته مي شوند. اين عوامل توانايي جستجو و مقايسه و انتخاب منابع اطلاعاتي بر اساس نيازمطرح شده توسط کاربر را دارند

توصیه ها و ابزارهایی برای جستجوی وب نامرئی

متخصصان جستجو ي وب ميگويند موتورهاي جستجو مثل گوگل و ياهو فقط 1 درصد (surface Web)از اطلاعات وب برايجستجو اطلاعات، اينترنت را مورد جستجو قرار ميدهند مابقي اطلاعات اينترنت را وب نامرئي يا deep Web deep Web، Deepnet، invisible Web، dark Web dark Web، hidden Web hidden مينامند.

خوشبختانه براي جستجو قسمتهايي از وب پنهان ابزارها و موتورجستجوهاي متعددي ايجاد شده و در يافتن اطلاعات دلخواه كمک خوبي هستند. اين موتور جستجوها از الگوريتم هاي پيشرفته رنكينگ و language -analysis استفاده مي كنند كه اطلاعات مفيد و مرتبط را چكيده مي كنند

موتور جستجوهای وب نامرئی

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

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

CompletePlanet CompletePlanet: بيش از هفت هزار ديتابيس و موتور جستجودر 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 مخفی دوستانتان

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

موتور جستجوهای وب نامرئی

pipl:طراحي شده براي جستجو وب نامرئي ي نتايجي كه اين وب سايت در دسترس قرار مي دهد از نظر بعضی 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 بانک اطلاعاتي درموضوعات متنوع

RefSeek RefSeek :موتور جستجويي براي دانشجوين و محققان ،بيشتر از يک ميليارد ستند كه شامل صفحات وب ، مجلات ،روزنامه ها ، كتابها و دايره المعارف را جستجو مي كند

دسته بندي يا خوشه بندي پايگاه داده

دسته بندي مبتني بر پرس و جو

دسته بندي مبتني بر خزش

یادگیری ماشین (هوش مصنوعی)

دسته بندي مبتني بر پرس و جو

- سیستم های مبتنی بر الگو
  - این سیستم ها سعی دارند تا الگوهای قطعی که در پرس و جو کاربر تکرار می شوند را شناسایی کنند . این الگوها برای تفسیر طبیعت و ماهیت درخواست بازیابی اطلاعات که به صورت ضمنی در پرس و جو وجود دارد به کار می رود.
  - Querix
- سیستم های پرس و جو زبان طبیعی کامل
  - این سیستم ها هیچگونه ساختار گرامری برای زبان که پرسش و جملات درخواست های بازیابی اطلاعات به کار می رود ، تحمیل نمی کنند . در عوض از تکنیکهای پردازش زبان طبیعی پیچیده برای تجزیه ، تفسیر و ترجمه ورودی به زبان پرس و جو سازگار با وب معنایی استفاده می کنند.
  - PANTO
- سیستم های زبان طبیعی کنترل شده
  - این سیستم ها بر زبان طبیعی کنترل شده تکیه دارند، هدف این است که با محدود کردن ورودی های کاربر به یک زیرمجموعه بدون ابهام از درخواست های پرس و جوی امکان پذیر ، یک پردازش گر پرس و جو زبان طبیعی بتواند پرس و جو های کاربر را تفسیر کند و آن را به پرس و جو فرمال معادل معنایی تبدیل کند.
  - CNL خاص (SWAT که از ACE استفاده می کند)
  - هدایتگر (Ginseng)

دسته بندي مبتني بر خزش

- در واقع یک نرم‌افزار هستند که می‌توانند از محتوای صفحات نسخه‌برداری کنند.از این نسخه‌ها برای ایندکس کردن صفحات استفاده می‌شود.
- بدین معنا که علاوه بر آدرس، محتوای صفحات را نیز رتبه‌بندی کرده، بدین ترتیب کیفیت محتوا به صورت سه گانه در قالب (شی، صفت، مقدار) توصیف میشوند
- خزنده‌ها در ابتدا لیستی از نشانی‌های وب را در اختیار دارند که به آنها دانه (Seed) گفته می‌شود.
- با مرور دانه‌ها و بررسی کد HTMLاین صفحات، تمامی پیوندهای صفحه مشخص شده و آنها را به لیست نشانی‌هایی که باید مرور نماید اضافه می‌کند

- موتورهای جستجو و برخی از سایت‌ها دارای خزنده‌ها و یا روبات‌هایی هستند که برای گردآوری اطلاعات وب سایت‌ها و نیز بروز نگه داشتن اطلاعات مورد استفاده قرار می‌گیرند. مهم‌ترین کار بعد از گردآوری اطلاعات، ایندکس کردن آن‌ها برای پردازش سریع هنگام جستجو است. این خزنده‌ها معمولا در بازه‌های زمانی منظمی اطلاعات را بروز کرده و با نسخه‌های قبلی مقایسه می‌کنند.
- به صورت عمومی نحوه کار Web crawlerها به این صورت است که ابتدا لیستی از URLها (آدرس های وب) که به عنوان seedشناخته می‌شوند را برای بازدید پردازش می‌کنند. هنگام پردازش این آدرس‌ها، لیست لینک‌ها و آدرس‌های موجود در صفحات آن‌ها را گردآوری کرده و به لیست ابتدایی اضافه می‌کنند. بقیه اطلاعات را نیز با توجه به نیاز و هدف خود ذخیره و پردازش می‌نمایند.