



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 (معمولاً 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.

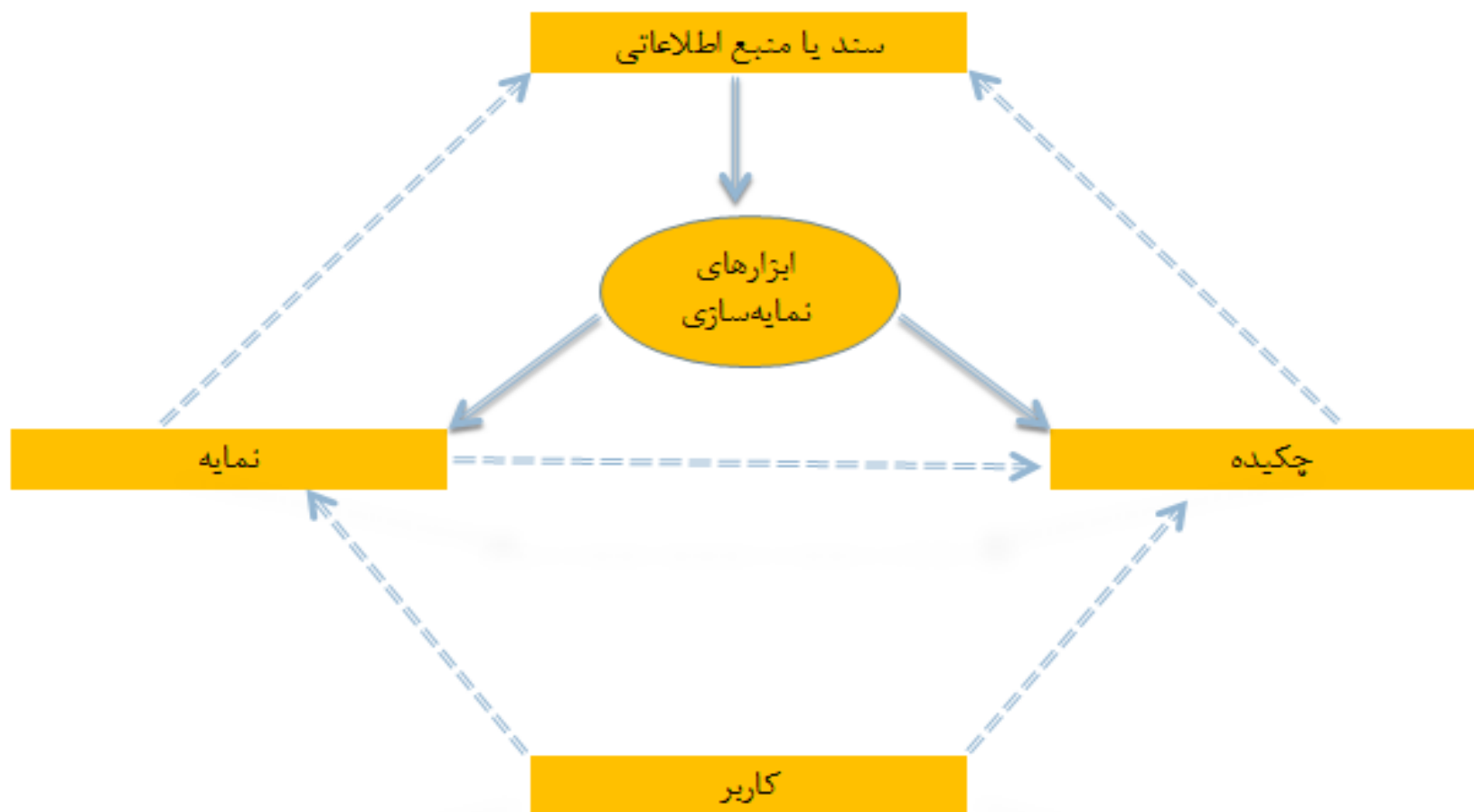
نمایه وسیله ای است برای هدایت منظم به یک متن، محتوا، مجموعه ای از مدارک و یا هر گونه اطلاعات ضبط شده ای که به شکل معمولاً الفبایی مرتب شده باشد و با استفاده از شیوه خاص و نظام ارجاعی مشخص ، موقعیت و محل هر مطلب را در بازیابی اطلاعات نشان دهد.

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

# نیاز به نمایه

- نمایه، میانبرهایی سیستماتیک و موثر به اطلاعاتی است که استفاده کنندگان نیاز دارند.
- نمایه ها برای هر مجموعه اطلاعاتی، جز مجموعه 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 نمایه سازی، چکیده نویسی و جستجو



# انواع نمایه

- ☐ نمایه‌های الفبایی یا واژه‌ای
- ☐ نمایه‌های مؤلف
- ☐ نمایه‌های کتاب
- ☐ نمایه‌های استنادی
- ☐ نمایه‌های رده‌ای:
- ☐ نمایه‌های هم‌ار
- ☐ نمایه‌های درهم‌کرد
- ☐ نمایه‌های چهریزه‌ای
- ☐ نمایه‌های سطر اول
- ☐ نمایه‌های فرارسانه‌ای
- ☐ نمایه‌های اینترنتی
- ☐ نمایه‌های چند رسانه‌ای
- ☐ نمایه‌های نشریات ادواری
- ☐ نمایه‌های گردشی عنوان

# انواع نمایه

- نمایه مولف: نقاط مدخل آنها شامل اسامی افراد، سازمان‌ها، پدیدآوردگان 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 رسانه‌ای: این نمایه مواد متنی، صوتی و تصویری را با هم ترکیب می‌کند
  - نمایه رده‌ای: مدخل‌های آن بر اساس رده‌ها یا سرعنوان‌های موضوعی مرتب می‌شود
  - نمایه نشریات ادواری: دو نوع می‌باشد: نمایه انفرادی برای مجلات انفرادی و نمایه‌های کلی 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 ای گردشی است که اغلب از عنوان های 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 های موجود در عنوان (نقاط بازیابی) به صورت الفبایی در مرکز واژه های غیرکلیدی موجود در عنوان در اطراف کلیدواژه
- **سیاهه بازدارنده : (List Stop)** شامل واژه هایی است که عملکردی نحوی دارند مثل حروف تعریف، حروف اضافه، حروف ربط و... فاقد هرگونه محتوای موضوعی می باشند

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

| های علوم کتابداری و              | اطلاع رسانی #     | واژه نامه ها و دایره المعارف  |
|----------------------------------|-------------------|-------------------------------|
| واژه نامه ها و                   | دایره المعارف های | علوم کتابداری و اطلاع رسانی # |
| واژه نامه ها و دایره المعارف های | علوم              | کتابداری و اطلاع رسانی #      |
| و دایره المعارف های علوم         | کتابداری          | و اطلاع رسانی # واژه نامه ها  |
| علوم کتابداری و اطلاع رسانی #    | واژه نامه ها      | و دایره المعارف های           |

## ۲- کواک (Keyword Out of Context)

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

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

2- در این نمایه بعضی مدخل ها معمولاً اصطلاحات تک واژه ای هستند

# نمونه ای از یک نمایه کواک

“واژه نامه ها و دایره المعارف های علوم کتابداری و اطلاع رسانی”

**اطلاع رسانی**

واژه نامه ها و دایره المعارف های علوم کتابداری و اطلاع رسانی

**دایره المعارف ها**

واژه نامه ها و دایره المعارف های علوم کتابداری و اطلاع رسانی

**علوم**

واژه نامه ها و دایره المعارف های علوم کتابداری و اطلاع رسانی

**کتابداری**

واژه نامه ها و دایره المعارف های علوم کتابداری و اطلاع رسانی

**واژه نامه ها**

واژه نامه ها و دایره المعارف های علوم کتابداری و اطلاع رسانی

# نمایه سازی کوئیک:

در نمایه سازی کوئیک که همان نمایه درون بافتی است تمام کلمه های موجود در عناوین مدارک را که توسط برنامه ای به کامپیوتر داده شده ، با سیاهه ای از واژه های غیرمجاز (واژه های غیرموضوعی که در فهرست ایستی یا بازدارنده قرار می گیرند تا در ردیف الفبایی خود به عنوان سرشناسه قرار نگیرد) در برنامه کامپیوتر مطابقت داده می شود. کلمه هایی که در سیاهه واژه های غیرمجاز وجود ندارد، بطور خودکار کلیدواژه محسوب می شود. مدخلی که با هریک از این کلیدواژه هادرست می شود به صورتی است که این واژه را در درون بافت خود ، یعنی همراه با بقیه کلمه های آن عنوان ، در شکل 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، اما همراه با سایر کلمات عنوان، در مدخل 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 دارد یا نه. برون داد رایانه ای می تواند به صورت چاپی در دسترس مراجعان قرار گیرد

# نمایه درون بافتی شامل سه بخش است

1. کلیدواژه یا مدخل

2. بافت (متن)

3. نشانه بازیابی (جایما).

نمایه واژه ها به صورت ستونی در مرکز یا سمت چپ صفحه ویاسمت راست آن (درفارسی) به ترتیب الفبایی می آیدو با همین کلمه هاست که جوینده ، جستجوی خود را درنمایه آغاز می کند. برای مشخص کردن پایان عنوان، درهرنمایه از علامت قراردادی خاصی نظیر علامت جمع یا مساوی و یاخط اریب (اسلش /) و غیره استفاده می شود تا جستجوگر بتواند به سهولت ابتدای عنوان را پیدا کرده و عنوان را در شکل طبیعی خود به سرعت مرور کند.

مزایای این نمایه

➤ سرعت و سهولت

➤ 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. در این نمایه هر کلیدواژه به ترتیب از عنوان خود خارج شده و مقدم بر سایر اجزای عنوان قرار می گیرد. سپس عنوان مدرک به ترتیب طبیعی خود و به طور کامل در زیر این واژه یا به دنبال آن می آید. به این ترتیب برای هر واژه مهم یک مدخل ساخته می شود. در این نمایه نیز کلمات مجاز در متن عنوان وجود دارد و کلید واژه قرار نمی گیرد.

# کارکرد نمایه سازی (Indexing Function)

1. محتوای اطلاعاتی مدارک را فشرده می‌سازد.
2. به عنوان واسطه برای تطبیق و یکسان سازی زبان مدرک و زبان کاوش به کار می‌رود.
3. به عنوان ابزاری کارا، بر شیوه تدوین راهبردی کاوش در جستجوی اطلاعات نظارت دارد.

# مراحل نمایه سازی

1. آشنایی 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 نمایه سازی

## 1. زبان طبیعی National Language

نوعی نمایه سازی است که از زبان مدارک استفاده می شود. در این روش هر اصطلاح یا واژه ای که در مدرک آمده است، می تواند برای نمایه در نظر گرفته شود. خصوصیت بارز نمایه سازی به زبان طبیعی، فقدان یک واژگان کنترل شده است. این خصوصیت باعث تنوع زبان طبیعی در نمایه سازی شده است.

## 2. زبان آزاد Free Language

نمایه سازی به زبان آزاد عبارت است از نمایه سازی که در آن هر واژه ای که بتواند موضوع مدرک را خوب توصیف کند به عنوان اصطلاح نمایه برگزیده می شود. خواه به وسیله پدیدآورنده مدرک به کار رفته باشد و یا به کار نرفته باشد. در این روش محدودیتی برای واژه هایی که می تواند در فرایند نمایه سازی به کار رود، وجود ندارد. این روش می تواند توسط انسان و یا رایانه انجام گیرد.

## 3. زبان کنترل شده مقید (Controlled Language)

در این گونه نمایه سازی معمولاً فهرستی مستند، مشخص کننده اطلاعاتی است که ممکن است به موضوعات نسبت داده شود؛ به بیان دیگر در نمایه سازی به زبان کنترل شده، شخص اصطلاحات یا واژه هایی از یک فهرست واژگان را براساس تفسیر ذهنی که از مفاهیم مندرج در مدرک دارد به مدرک اختصاص می دهد.

# سیستم های نمایه سازی (Indexing Systems)

- سیستم های نمایه سازی وسایلی هستند که به وسیله آنها از یک زبان نمایه ای برای ایجاد نمایه یا سایر ابزارهای تجسسی استفاده می شود. بنابراین به مجموعه ای از رویه های مقرر که برای سازماندهی محتوای رکوردها به منظور بازیابی و اشاعه اطلاعات، مورد استفاده قرار می گیرد سیستم نمایه سازی گفته می شود. پس همارایی و پیش همارایی دو سیستم اساسی نمایه سازی به شمار می رود.
- نمایه سازی همارا Co – Ordinate Indexing
- نمایه سازی همارا به شیوه ای از نمایه سازی گفته می شود که در آن محتوای موضوعی هر مدرک با بیش از یک شناسه یا اصطلاح تعیین و توصیف می شود و رابطه بین اصطلاح ها به وسیله جور کردن و ترکیب واژه های منفرد معین می شود
- . نمایه سازی پس همارا Co – Ordinate Indexing – Post
- 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. به عبارت دیگر پس همارایی اشاره به عملی می کند که بعد از روند نمایه سازی انجام می شود و مربوط به همارایی دو یا چند اصطلاح برای ساختن مفهوم موردنظر در مراحل بازیابی است.
- نمایه سازی پیش همارا Co – Ordinate Indexing – Per
- هرگاه بین دو یا چند جزء به طور تصنعی و ساختگی، پیوند برقرار کنیم به این عمل و نوع آن، نمایه پیش همارا گویند. به عبارت دیگر در نمایه سازی پیش همارا، ترکیب یا همارایی عناصر تشکیل دهنده موضوع مورد جستجو در هنگام نمایه سازی و به عبارتی پیش از بازیابی صورت می گیرد

# 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 گیرد و 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 و برنامه 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

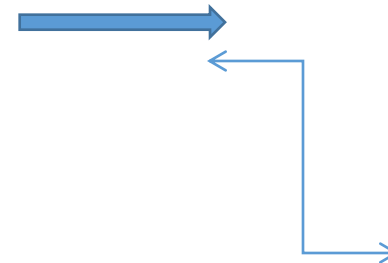
- **نمایه ساده:** مدخل ها به ترتیب **الفبایی** است و در جلوی آن **جاینما** وجود دارد. **فاقد بیانگر** است و هیچگونه تورفتگی وجود **ندارد** و تعداد جای نما های آن افزایش می یابد.
- **نمایه درون بافتی:** این نمایه به صورت **خطی و نحوی** است و به ترتیب **پیدایش بیانگرها** در متن است. بنابراین الفبایی **نیستند**. 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. مثلاً میگوییم رفتار **با** کودکان ولی در برون 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 می گیرد.

# 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 هر صفحه.
- انتقال هر کلید واژه (مدخل) بر روی یک برگه 5/7 در 5/12 **و ذکر 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 **مؤلف 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: استثنایی ، کودکان ( غلط است )
- خوارزمی، محمد ( صحیح است )

## ملاک انتخاب اصطلاح مرجح

- رواج داشتن . مثال: جامعه شناسي به جاي علم الاجتماع
- بومی بودن . مثال : آرمان گرایی به جاي ایده آلیسم
- جدید و امروزي بودن . مثال: هواپیما به جاي طیاره
- علمی بودن. مثال: اسید سولفوریک به جاي جوهر نمک
- سرنام یا اختصار مشهور. مثال: یونسکو
- پرهیز از بکار بردن کلمات ترکیبی به جز در مواردی که اصطلاح قابل جدا سازی نیست 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 احتمالاً 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 نظام‌های اطلاع رسانی **به زبان ساختگی** و قراردادی اطلاق می‌شود که برای مقاصد نمایه سازی به ویژه قابلیت بازیابی اطلاعات و مدارك به کار گرفته می‌شود. به طور کلی زبان نمایه سازی، **استانداردی را مهیا** می‌کند که هم نمایه ساز و هم جستجوگر می‌توانند از آن استفاده کنند.

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 تشکیل دهنده موضوع مورد جستجو در **هنگام نمایه سازی** و به عبارتی پیش از بازیابی صورت می گیرد:

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 تشکیل دهنده 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 و زبان 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 یا اصطلاحنامه**، مجموعه اصطلاحات يك رشته است که میان آنها روابط معنایی، رده‌ای، و سلسله مراتبی برقرار شده و توانایی آن را دارد که موضوع آن رشته را با همه جنبه‌های اصلی و فرعی و وابسته، به‌گونه‌ای نظام‌یافته و به‌منظور ذخیره و بازیابی اطلاعات ارائه دهد.
- اصطلاحنامه معادل فارسی واژه انگلیسی **Thesaurus** است که ریشه در زبان لاتینی باستان دارد و به‌معنای گنجینه، ذخیره، و مجموعه به‌کار می‌رود.
- اصطلاحنامه از نظر **وظیفه و کارکرد**، ابزار کنترل واژه‌ها به‌منظور برگرداندن زبان طبیعی مدارك به زبان مقید است و از نظر ساختار، واژگان کنترل‌شده و پویای زمینه‌ای خاص از دانش بشری است که برای ذخیره و بازیابی اطلاعات آن حوزه به‌کار می‌رود

# اهداف اصطلاحنامه

اصطلاحنامه دارای هدف‌های اساسی زیر است:

- (1) نمایاندن ساختار **زمینه معینی از دانش** چنان‌که هم نمایه‌ساز و هم جست‌وجوگر بتوانند از گستره آن زمینه و ارتباط میان مفاهیم آن با اندیشه‌های مرتبط آگاهی یابند
- (2) ارائه **اصطلاحات استاندارد** در زمینه‌ای معین.
- (3) برقراری **نظام ارجاعات میان اصطلاحات و رده‌بندی اصطلاحات** به‌صورت سلسله‌مراتبی
- (4) تأکید بر توجه به **نیازهای اطلاعاتی** استفاده‌کنندگان
- (5) تعیین **اصطلاحات مجاز و مشخص کردن حدود معانی** اصطلاحات به‌منظور ایجاد هماهنگی در نمایه‌سازی

# روابط میان اصطلاحات

- باید توجه داشت که ویژگی ذاتی اصطلاحنامه، توانایی تعیین و نمایش روابط معنایی میان واژه‌هاست و یک رابطه يك سویه نداریم و رابطه همه جانبه وجود دارد.
- این روابط ممکن است یکی از این سه نوع باشد:
- الف) **رابطه هم ارزی** : Equivalence Relation  
میان اصطلاح پذیرفته‌شده و اصطلاح پذیرفته‌نشده برقرار می‌شود
- مثال: اصطلاح جامعه‌شناسی به جای علم‌الاجتماع
- ب) **رابطه سلسله مراتبی** Heirachial Relation
- بیان‌کننده رابطه اعم و اخص میان مفاهیم است که در واقع، اصطلاحنامه‌ها را از واژه‌نامه‌های متداول متمایز می‌کند.
- **رابطه همبسته یا همایند** : Associative Relation
- رابطه میان دو اصطلاح که به دلیل وابستگی معنایی، وجود یکی دیگری را نیز به ذهن متبادر می‌کند. **مثال :**  
دو اصطلاح اسب و سوارکاری

# ساختار اصطلاحنامه

- اصطلاح اعم ( ا ع ) : Brother Term : اصطلاح عامتر
  - مثال: بیماریهای داخلی
  - ا. ع. کلیه و مجاری ادرار
- اصطلاح اخص ( ا خ ) : Narrow Term : اصطلاح خاصتر
  - مثال: بیماریهای گوارش
  - ا. خ. معده
  - مری
  - روده
- اصطلاح وابسته ( ا و ) : Related Term : اصطلاحی که به اصطلاح برگزیده به نوعی 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 جای ( بج ) : Used for : ارجاع از اصطلاح گزیده ( مرجح ) به اصطلاح ناگزیده ( نا مرجح ) که البته مترادف است.
  - مثال: نئوپلاسم
  - بج. سرطان

# ساختار اصطلاحنامه

## اصطلاح راس

مثال: میوه

راس. خوراکی ها

**یاد داشت دامنه**: Scope Note: یادداشتی برای توضیح مدخل که دارای ابهامی و یا دارای چند معنا است و یا معنی خیلی کلی دارد استفاده می شود .

مثال: فرهنگ ( آداب و رسوم )

Top Term : اصطلاح سرمقوله مجموعه ای از اصطلاحات

# انواع رابطه سلسله مراتبی

- رابطه سلسله مراتبی خود شامل سه نوع است :
- 1. **رابطه جنس و نوع**: Generic : این رابطه دقیق ترین و قطعی ترین رابطه سلسله مراتبی است که با کلمات همه برای جنس و بعضی برای نوع قابل شناسایی است.
- **مثال** : رابطه میان اصطلاحات : نشخوار کنندگان ( جنس )
  - بز ( نوع )
  - گاو ( نوع )
  - گوسفند ( نوع )
- همه بز ها نشخوار کننده هستند
- بعضی از نشخوار کنندگان بز هستند.

# انواع رابطه سلسله مراتبی

- **2. رابطه کل و جزء** : Whole – part : که نشان دهنده رابطه يك جزء را با کل آن مشخص می کند و نوعی رابطه اعم و اخص است.
  - مثال : پا ( جزء )
  - بدن ( کل )
- **3. رابطه مصداقي** : Instance Relation : این ارتباط بیشتر در مورد اسامی خاص است و رابطه ای است که جنس و نوع نیست بلکه بیانگر يك مصداق است.
  - مثال: دماوند ( مصداق يك قله )

# تفاوت واژه نامه و اصطلاح نامه

- واژه نامه ها : **نظم الفبائي** دارد و به دنبال هم مي آیند. ولي **مجرد** بوده و حالت خنثي دارند .

- اصطلاحنامه : معمولاً **نظم الفبائي** – **رده اي** دارد و کلمات با هم مرتبط بوده و **روابط معنائي** دارند.

# واحدهای اندازه‌گیری حجم در کامپیوتر

- کوچک‌ترین واحد حجمی دیجیتال، یک بیت (bit) است
- بایت: برابر با 8 بیت
- کیلوبایت: برابر است با 1024 بایت
- مگابایت: برابر است با 1024 کیلوبایت (1 مگابایت: یک کتاب 400 صفحه‌ای)
- گیگابایت: برابر است با 1024 مگابایت
- ترابایت (TB)
- پتابایت (PB)
- اگزابایت (EB)
- یوتابایت (YB)
- هلابایت (HB)
- Yotta = 24 صفر را در مقابل عدد یک
- Hella = 27 صفر را در مقابل عدد یک در سال 2020

# ISBN( International Standard Book Number )

رده بندی با کمک کد شابک، شماره استاندارد بین المللی کتاب است. این شماره در ابتدا کدی 1 رقمی بود که در سال 1167 توسط آقای 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 رقمی می باشد. 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 بعدی کد خاص آن کتاب در انتشارات است

بخش آخر در اصطلاح Checksum ارقام دیگر است.

# Checksum: روشی برای تشخیص خطا در انتقال داده

## از مبداء به مقصد است

روش به دست آوردن Checksum در شابک 13 رقمی: مثال: در کد شابک زیر، رقم checksum را به دست آورید:

.....-۸۳-۰۵۳۱-۹۶۴-۹۷۸

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.

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

$$115 \bmod 10 = 5$$

$$10 - 5 = 5$$

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 آقای تیم برنرزلی ابداع شد. حاوی صفحات ساده‌ی بدون جذابیت و بدون هوشمندی.

2-وب اشتراکی (Social Web) در این دوران سایتهای اشتراک جمعی یا شبکه‌های اجتماعی (Social Networks) مثل فیسبوک، توئیتر، اینستاگرام و... ظهور کردند و کاربران توانستند داده‌های خود را به راحتی با یکدیگر به اشتراک بگذارند و بنابراین دسترسی به داده‌ها آسانتر شد.

3-وب وب معنایی یا وب هوشمند (Web Semantic) در این وب پیش‌بینی میشود صفحات وب بیش از گذشته برای موتورهای 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 افزارهای 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 از مدرک است که جهت نمایه سازی از آن استفاده می شود و شامل عنوان یا چکیده یا فهرست مندرجات یا حتی کل مدرک می باشد.

**تشخیص اطلاعات:** تشخیص کلمات و جملاتی 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 بر حسب نیاز و با توجه به برنامه ای است که قبلاً به سیستم داده شده است.

**انتخاب:** آن دسته از اصطلاحاتی که در اولویت قرار گرفته اند به عنوان توصیفگر انتخاب 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 بر احتمالات

# نمایه سازی رایانه ای عمدتاً به دوشیوه نجام می شود

الف- حفظ اصطلاح: در سال ۱۹۶۳ توسط دکتر سوزان آرتانندی 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 در متن بود، حفظ کند و آنرا به عنوان اصطلاح نمایه ای انتخاب کند. در این روش استفاده از اصطلاحات از پیش انتخاب شده راهنمای خوب و مفیدی 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 و مهم نمایه سازی شوند.

ب- حذف اصطلاح: در شیوه حذف اصطلاح یک سیاهه بازدارنده به رایانه داده می شود که کلمات فهرست را هرکجا که بود حذف 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 واژه های 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 از متن ،

➤ 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.

# نکات

وقتي شناسه ها توسط نمايه ساز انتخاب مي شوند نمايه سازي دستي گفته مي شود , اگر اين کار را کامپيوتر انجام دهد، نمايه سازي خودکار 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 از انتخاب کلید واژه ،شماره گذاری ،ترتيب بندي و غيره توسط کامپيوتر انجام مي گيرد.

نمايه سازاني که با اصول تحليل و طراحي سيستم و برنامه سازي کاربردی آشنائي دارند، مي توانند در زمينه کامپيوتری سازي نمايه سازي مو ثرباشند

# روش کار

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

- این کار توسط یک نرم افزار به نام نرم افزار بسامدی استخراج می کند.

مثلا یک مفهوم که در ذیل یک بخش از متن بارها تکرار شده باشند، محتملا کلید واژه است، مگر آن که جزو واژگان غیر موضوعی باشد.

- واژگان **غیر موضوعی** مانند (است، که ، را و...) توسط یک فهرست ایستا مشخص و نادیده گرفته می شود.

-قسمت دیگر نرم افزار حاوی **اسمهای خاص** است. نرم افزار با استفاده از این

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

# فرایند انتخاب کلید واژه در نمایه سازی خودکار

1. شناسایی واژه های انفرادی از متن که تحلیل واژگان نامیده میشود
2. برداشتن واژه های با بسامد تکرار بالا که در ارائه محتوای متن بی تأثیرند، با استفاده از فهرست واژه های غیرمجاز.
3. تبدیل واژه های باقی مانده به شکل ریشه آنها یعنی حذف پسوند یا پیشوندها تا هر کلمه تا حد ریشه اش کوتاه مبدع آن است.
4. محاسبه رایانه ای بسامد ریشه هایی که در متن تحلیل شده اند، به منظور تعیین تابع ارزش گذاری بر ریشه.
5. ریشه هایی که ارزش گذاری بزرگتری دارند، برای متنی که در آن ظاهر شده، به عنوان کلید واژه تعیین می شود.

## ادامه

شرکت زافتکس نوعی برنامه کامپیوتری با عنوان IDX طراحی کرده که به ارائه خدمات نمایه سازی ماشینی می پردازد.

این نرم افزار از روش استفاده از واژه نامه بهره می برد.

این نرم افزار امکان تبدیل واژگان به ریشه آنها را جهت بازیابی بعدی فراهم می کند  
علامت گذاری و محدود کردن واژه های ناخواسته را انجام می دهد

شکستن واژه 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 فهرستی که واژه های غیرمجاز نامیده می شود، به دست می آورد.

در اختیار داشتن فهرستی از این واژه ها و ارائه آنها به برنامه رایانه ای برای ممانعت از ورود آنها به فهرست واژه های مفهومی مطلوب برای نمایه شدن، یکی از اقدامهای سودمند در نمایه سازی خودکار مبتنی بر کلید واژه هاست ..

روش های نمایه سازی ماشینی

### الف. روش زبان شناختی:

این روشها می کوشند با کمک تحلیل های شکل شناسی و ساختار نحوی مدرک توصیفگرها را استخراج نمایند.

1. تجزیه و تحلیل ریخت شناسی که بر مبنای ریشه لغات عمل می کند.
2. کلمات بدون بار معنایی موجود در سیاهه بازدارنده را حذف میکند.
3. شکلهای دستوری صرف کلمه را به یک شکل می آورد.
4. ضمائر را بر اساس اسامی مربوط به آنها مرتب می کند.
5. تجزیه و تحلیل نحوی که در سطح جملات امکان پذیر است و بر مبنای علامتهای نحو لغات انجام می شود.
6. تجزیه و تحلیل معنا شناختی که در سطح مدارک مشترک در یک پایگاه صورت می گیرد .  
ارتباطات معنایی موجود در یک مدرک شناسایی می شوند تا متون مشترک بتوانند به صورت واحدهای هم معنی تجزیه شوند.

## • ب. روش های آماری:

1- مشخص می کند که معنی هر مفهوم منفرد در مدرک با حضور آن در جایگاه های مختلف مدرک ارتباط تنگاتنگ دارد. بنابراین لغت درون متن شمارش می شوند و ارتباط آنها ارزش گذاری می شود. هدف آماری از اطلاعات آن است که لغات دارای بارمعنایی در مدرک به عنوان توصیفگر انتخاب شوند.

2- روش های آماری عمل برای بالابردن جامعیت به کار گرفته میشوند. درحالیکه روشهای زبان شناختی در جهت بهبود مانعیت کاردارند.

## ج. روش های مبتنی بر احتمالات:

در این روشها تئوری احتمالات بر مدل سازی ریاضی مراحل بازیابی به کارگرفته می شود. درحالیکه در توزیع آماری اصطلاحات یک مدرک مورد استفاده قرار می گیرد. این روش با عملیات ریاضی مفروضات ساده و مطمئنی را ارائه می دهد. فرض بر آن است که مدارک بر اساس میزان ربط در هنگام بازیابی مورد ارزیابی قرار میگیرند.

# فهرست ایستا , یا واژگان غیر مجاز

- تحلیل کلمات یک متن نشان می دهد گروهی از کلمات بی اهمیت وجود دارد که به فراوانی در متن ظاهر می شود مانند( ی ، به، نه، برای، با، چه کسی، چه موقع، است، آن).

- گروهی نیز وجود دارد که بندرت در متن متی آیند و ممکن است نشان دهنده محتوای اطلاعاتی متن نباشند

- این دسته از واژه ها به تنهایی بار معنایی ندارند , بود یا نبود آنها تنها در پرسش کاربر تأثیری ندارد بلکه در میزان ربط یا عدم ربط مدارك بازیابی شده نیز تأثیری ندارد.

- این واژه ها یا با عنوان واژه های غیرمجاز برای ورود ب نمایه معرفی می شوند.

# مزایای تهیه لیست واژه های غیر مجاز

-در صورتی که واژه های غیر مجاز قبل از فرایند نمایه سازی مدارك مشخص و فهرست آنها برای کنترل به رایانه داده شود، علاوه بر صرفه جویی در زمان و حجم بایگانی های نمایه، به میزان زیادی از بازیابی مدارك نامرتبط و ریزش کاذب در جستجو جلوگیری خواهد شد.

چند نمونه از نرم افزارهای معروف نمایه سازی که در سیستمهای کامپیوتر های شخصی کار می کنند به شرح زیر است:

INDEXING RESEARCH محصول شرکت CINDEX

INDEX AID محصول شرکت Santa Barbara Software Products

INQUIRY محصول شرکت Indexer assistant

# عوامل موثر در نمایه سازی خودکار

## محدوده رکورد

- اولین تصمیم گیری مهم برای تهیه یر نوع نمایه ، گزینش حد و حدود رکوردي است که واحد قابل جستجو را تعریف می کند.
- این تصمیم گیری در بازیابی کارآمد نقشی حیاتی دارد.

## محدوده اصطلاحات

- تعیین حد و حدود یک واژه از دیگر مسائلی است که در نمایه سازی خودکار باید به آن توجه شود. در نظام های نمایه سازی دستتبی ، گزینش کلمات برای نمایه به سهولت انجام می شود. اما در نمایه سازی خودکار از آنجا که ماشین از هوشمندی لازم برای انتخاب کلمات برخوردار نیست بنابراین باید حدود کلمه را تعریف کرد. معمول حدود کلمات نمایه را با استفاد ه از علائم نقطه گذاری تعریف می کنند.
- به طور معمول ، فاصله بین کلمات و علائم دستوري و نقطه گذاری به عنوان مرز کلمات در نظر گرفته می شود. روش های تعیین حد و حدود کلمات در نمایه سازی خودکار، بر اساس نوع برنامه و میزان پیشرفتگی آنها متفاوت است.

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

## نمایه سازی استخراجی

ساده ترین روش نمایه سازی در پایگاه های اطلاعاتی ، روش نمایه سازی استخراجی است که در آن واژه یا برای قرار گرفتن در نمایه ، توسط رایانه از متن استخراج می شوند. در این روش عموماً بسامد تکرار واژه در یک رکورد یا مقاله تعیین شده و کلماتی که بسامد تکرار آنها زیاد است در متن نمایه قرار می گیرند.

## نمایه سازی تخصیصی

انتسابی فرایند پیچیده ای است که با استفاده از تحلیل های آماری ، کلمات و اصطلاحات به مدرک منتسب می شوند رایانه برای نمایه سازی از اصطلاحنامه یا کنترل واژگان بهره می گیرد.

# انواع نمایه استخراجی

**نمایه سازی با استفاده از فهرست کلمات ممنوعه :** در این روش در هنگام نمایه سازی ، رایانه تمام کلمات متن را استخراج میکند، سپس کلمات ممنوعه را حذف و بقیه کلمات را در یک نظام الفبایی مرتب میکند.

**نمایه سازی بسامدی :** در این روش، بسامد تکرار کلمات در یک رکورد مورد بررسی قرار می گیرند و براساس بسامد تکرار در فهرست کلمات نمایه قرار می گیرند.

**نمایه سازی ریشه یابی :** در بعضی از سیستم های نمایه سازی استخراجی ، از پسوند یا ریشه کلمات استفاده می 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 موضوعی خاص مانند الگوریتم های پزشکی 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 متفاوتی دارد.

در نظام های رایانه ای بیشتر از روش های نمایه سازی استخراجی استفاده می شود.

1- یکی از عمده ترین مشکلات این نمایه یا به ویژه هنگام استفاده در پایگاه های اطلاعاتی ، عدم بازیابی اطلاعات به دلیل نبودن آن کلمه درخواستی در نمایه پایگاه اطلاعاتی است.

2- دلیل این امر آن است که بهره گیران ، یا همه کلمات مترادف با اصطلاح موجود در درخواست را وارد نکرده اند و یا از مترادفات آن بی خبرند. بنابراین ، بسیاری از مدارك مرتبط از دست می روند.

برای رفع این معضل طراحان پایگاه های اطلاعاتی توانایی های نمایه ای را با توانایی نرم افزاری درهم می آمیزند به دو روش زیر:

الف: نمایش نمایه و انتخاب واژه درخواستی توسط خود بهره گیر است .

ب: استفاده از نظام بازخورد مرتبط است . این روش به بهره گیران اجازه می دهد تا مدارك مرتبط را برگزینند. سپس از سیستم می خواهند تا با توجه به این مدارك ، مدارك مرتبط بیشتری را بازیابی نمایند. امروزه این روش در اینترنت و پایگاه های اطلاعاتی تمام متن کاربرد فراوانی دارد .

# نمایه سازی تخصیصی

- در واقع در نمایه سازی تخصیصی برای هر واژه "پرونده ای" از کلمات و عبارات مرتبطی که به نظر می رسد تهیه می شود.

بنابراین می توان از برنامه ای رایانه ای برای انطباق عبارت های مهم در مدرک با این مجموعه پرونده یا استفاده کرد و در صورت انطباق واژه موجود در مدرک با واژه های موجود در پرونده های کلمات ، اصطلاح نمایه ای را انتخاب نمود .

- هر چه نمایه سازی تخصیصی تر باشد صرفه اقتصادی آن در یک سیستم خبره بیشتر می شود .

- سیستمی که از یک نمایه سازی تخصیصی استفاده می کند می تواند به عنوان یک سیستم کمی خبره از آن نام برد.

# تعریف وب مرئی و نامرئی

**وب مرئی :** صفحات ثابت با دسترسی آزاد

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

## مهارت های اطلاع یابی

مهارت در بازیابی اطلاعاتی

مهارت در ارزیابی اطلاعاتی

مهارت در سازماندهی اطلاعاتی

مهارت در تبادل ارتباط.

# روشهای جستجوی اطلاعات در محیط وب

ساده

پیشرفته

کنترل واژگان

اصطلاحنامه

**فرایند جستجو**

ارزیابی پایگاه یا

جامعیت

مانعیت

# منابع موجود در وب نامرئی

بخش بزرگی از وب وجود دارد که عنکبوت های موتورهای جستجو آن را نمایه نمی کنند یا نمی توانند نمایه کنند و عبارت اند از سایت های دارای رمز عبور، اسناد موجود در پشت سامانه های حفاظتی، فایل های pdf از متون آرشیو شده، و ابزارهای تعاملی نظیر ماشین حسابها و برخی واژه نامه هاو همچنین محتویات بعضی از پایگاه های اطلاعاتی، منابع محافظت شده از طریق اسم کاربر و گذر واژه، منابع و صفحات وب بدون پیوند، و صفحات افزون بر حداکثر تعداد صفحات قابل مرور در نتایج بازیابی

# اهمیت وب نامرئی

به دو دلیل می توان گفت که وب نامرئی اهمیت دارد . نخست از نظر کمی، باید گفت که حجم اطلاعات موجود در این بخش خیلی بیشتر از سطح آشکار است. موارد زیر، اهمیت وب نامرئی را از نظر کمی نشان می دهند:

1 . بهترین موتورهای کاوش فقط قادر هستند که حدود 16 درصد از اطلاعات موجود در وب را بازیابی کنند و بنابراین 84 درصد آنها جزو وب نامرئی به حساب می آیند

2 . اندازه وب نامرئی تقریباً 500 برابر وب مرئی است: وب نامرئی 550 میلیون سند، و وب مرئی تقریباً 1 میلیون سند را دارا می باشد .

دوم اینکه از نظر کیفی، اطلاعات بخش های مختلف این مجموعه بویژه منابع اطلاعاتی موجود در وب عمیق، معمولاً منابع ارزشمند و مفیدی هستند و در بسیاری از موارد پاسخگوی نیاز کاربران می باشند. تقریباً بیش از نیمی از وب نامرئی را پایگاه های اطلاعاتی موضوعی تشکیل می دهند

# بخش های مختلف وب نامرئی

1 . وب مات یا تاریک

2 -وب عمیق

3-وب خصوصی

4-اینترنت واقعاً پنهان

# 1. وب مات یا تاریک

- بخشی از فضای وب نامرئی به وب مات موسوم گردیده که می توانسته مورد استفاده کاربران قرار گیرد، اما به دلیل زیر این اطلاعات در خارج از دسترس کاربران قرار گرفته و موتورهای کاوش نمی توانند آنها را بازیابی کنند:
- از آنجا که اولاً محیط وب دائماً در تغییر است و هر روز منابع و اطلاعات جدید به آن افزوده می گردد
- ثانیاً صفحاتی در وب وجود دارند که هیچ پیوندی بین آن یا با منابع دیگر برقرار نشده، خزنده های موتورهای جستجو قادر به یافتن این صفحات و همگام نمودن خود با این حجم عظیم اطلاعات نیستند.
- به دلیل محدودیت توانایی، نرم افزارهای خزنده فرصت کافی برای روزآمدسازی صفحات جدید وب را ندارند. موتورهای کاوش نیز امکان روزآمدسازی حجم عظیمی از اطلاعات و منابع جدید را ندارند و به همین دلیل بسیاری از این اطلاعات از حوزه موتورهای کاوش دور می مانند .
- محدودیت توان مالی بسیاری از موتورهای کاوش سبب گردیده که موتورهای کاوش نتوانند تمام صفحات وب سایت ها را نمایه سازی کنند، چرا که برای آن یا هزینه های زیادی دارد
- بنابراین موتورهای کاوش بنا بر سیاست های خودشان، تنها بخشی از وب
- سایت ها یا لایه های بیرونی آنها را نمایه سازی می کنند. بنابراین همیشه بخش عظیم لایه های درونی وی سایت ها پنهان می مانند

## 2-وب عمیق

به مجموعه اي از اطلاعات الكترونيكي پيوسته اطلاق مي شود كه بسياري از پايگاه های اطلاع رسانی، آنها را از طریق شبکه جهانگستر وب در دسترس عموم قرار داده اند.

- برخي اين اطلاعات را به رایگان، و برخي دیگر را با دریافت هزینه در دسترس عموم قرار مي دهند.

مندرجات اين پايگاه يا معمولاً خارج از حوزه جستجوی موتورهای كاوش قرار دارند هر يك از اين پايگاه يا صفحه جستجوي مبتني بر وب دارند. كه امكان 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 پزشکی (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 دسترسي به وب عمیق باز نماند

### 3-وب خصوصی و وب ملکی

بخشی دیگر از وب نامرئی وجود دارد که چون اطلاعات موجود در آن جزو دارایی های شخصی یا خصوصی سازمان ها یا افراد می باشد، از حوزه دسترسی موتورهای جستجو پنهان است. مثلاً در برخی از سازمان یا مؤسسات خصوصی یا دولتی، به دلایل امنیتی از اطلاعات مربوط به مسائل کاری و سازمانی و پرسنلی خود حفاظت می کنند اجازه دسترسی به آنها را به دیگران نمی دهند و فقط کسانی که دارای اسم کاربر و گذر واژه هستند می توانند از آنها استفاده کنند این بخش، وب خصوصی محسوب می گردد بخش دیگر، منابع اطلاعاتی از قبیل نشریات الکترونیکی مبتنی بر وب می باشند که دسترسی به آن یا از طریق پرداخت حق اشتراك و خرید محصولات اطلاعاتی شرکت های مختلف صورت می گیرد و وب ملکی نامیده می شود.

## 4 . اینترنت واقعاً پنهان

بخش دیگری از وب پنهان وجود دارد که بنا به مسائل فنی و ناکارآمدی ابزارهای جستجو، از دسترسی کاربران دور مانده است. بسیاری از موتورهای جستجو قادر به بازیابی اطلاعات متنی html هستند، ولی توانایی بازیابی فایل های pdf را ندارند، یا به دلیل کمبود منابع مالی و فنی از جستجوی فایل های غیرمتنی صرف نظر کرده اند. بنابراین منابع اطلاعاتی متنوعی نیز در وب وجود دارند که تنها به دلیل محدودیت های فناوریان یا مالی موتورهای جستجو، از حوزه کاوش آن یا و در نتیجه از دسترس کاربران دور مانده اند.

# دلایل عدم بازیابی و نمایه سازی وب نامرئی توسط موتورهای کاوش

## 1. دلایل فنی:

بسیاری از موتورهای کاوش به دلیل محدودیت های نرم افزاری توانایی روزآمدسازی اطلاعات جدید وب را ندارند. باید یادآور شد که هنوز هیچ موتور کاوشی ادعا نکرده است که قادر به گسترش حوزه کاوش خود به تمام محیط وب می باشد و همیشه این موتورها یک گام از سرعت روزافزون اطلاعات عقب تر هستند.

## 2. دلایل بودجه ای:

همانطور که قبلاً اشاره شد فرآیند نمایه سازی تمام صفحات وب، هزینه بر خواهد بود و موتورهای کاوش نیز بنا به محدودیت بودجه ناگزیرند فقط بخشی از وب سایت ها را نمایه سازی کنند

## 3. دلایل اجتماعی و حقوقی:

از آنجا که اطلاعات موجود در وب در دسترس عموم قرار می گیرد، بسیاری از افراد و سازمان یا به دلیل صرف بودجه های کلان در راه اندازی سایت ها و پایگاه های اطلاعاتی خود، حاضر نیستند این اطلاعات را به صورت رایگان در اختیار همه بگذارند.

البته این از لحاظ اجتماعی و حقوقی جز حق مسلم آنها است.

## شیوه های اطلاع یابی در وب نامرئی

در حال حاضر ابزارهایی به وجود آمده اند که منابع وب نامرئی را شناسایی، و کاربران را به سایت های مناسب راهنمایی می کنند. این رویکرد توسط بزرگراه های اطلاعاتی و کتابخانه های مجازی پذیرفته شده است بطوری که فقط توصیفی از پایگاه های اطلاعاتی و مجلات نامرئی را ارائه می کنند مثلاً سایت Invisibleweb که فهرستی از منابع نامرئی را، و سایت Completeplanet Completeplanet که فهرستی از تقریباً 40000 پایگاه اطلاعاتی وب نامرئی را ارائه می دهند.

برخی دیگر از ابزارهای اطلاع یابی نیز که تاکنون معرفی شده اند ما می توانیم با استفاده از آن یا به این اطلاعات دسترسی پیدا کنیم به شرح زیر است:

# دروازه های اطلاعاتی

دروازه های اطلاعاتی مجموعه ای از پایگاه یا و سایت یا هستند که به وسیله متخصصان اطلاعاتی و معمولاً کتابداران گردآوری، بررسی و براساس موضوع مرتب شده اند و معمولاً به کاربران نیز

توصیه می شوند. نمونه ای از این 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 مزیت های عمده استفاده از دروازه های اطلاعاتی این است که برای ایجاد آنها افرادی با دانش موضوعی خاص، در اینترنت جستجو کرده اند و به پالایش اطلاعات مفید از غیرمفید پرداخته اند

# پرتال ها يا پایگاه های اطلاعاتی خاص موضوعی

مجموعه اي از پایگاه هاي اطلاعاتي خاص موضوعي هستند که به یک موضوع خاص اختصاص دارند و به وسیله دانشمندان، محققان، متخصصان، مؤسسات دولتي، شرکت هاي بازرگاني و کارشناسان 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 پیوند هاي خبري، فایل هاي چندرسانه اي، آرشیوها، فهرست هاي پستي اشخاص، شغل يابهاو هزاران پایگاه اطلاعاتي که به موضوعات خاص اختصاص دارند استفاده مي شود.

# ویژگی پایگاه های اطلاعاتی

- 1 . جستجوپذیر و مرورپذیر، حاوی توصیف منابع اینترنتی در یک زمینه موضوعی خاصي
- 2 . مشتمل بر معیارهاي شفا و تعریف شده براي ارزیابی کیفیت منابع اطلاعاتي، به جاي انتخاب بدون ارزیابی
- 3 . دخالت کتابداران و متخصصان موضوعي در ایجاد آني
- 4 . تولید دستي رکوردهاي آن، براي توصیف بامعنا و اطلاع بخش از منابع اطلاعاتي
- 5 . فهرستنویسي و رده بندي منابع اطلاعاتي با استفاده از شیوه هاي سنتي کتابخانه اي به منظور بازیابی مو ر

# ابرموتورهای کاوش :

گسترش پذیری حوزه های جستجو نیز یکی از شیوه های دسترسی به وب نامرئی شمرده می شود که نمونه آن، استفاده از ابرموتورهای کاوش است. این ابرموتورها خود، موتورهای

جستجوی واقعی نیستند. بلکه به کاربران این امکان را می دهند که کلیدواژه های خود را همزمان توسط چند موتور، مورد کاوش قرار دهند و نتایج جستجوی همه آنها را با هم ببینند .

## عوامل هوشمند :

این عوامل هوشمند از ابزارهای بازیابی در اینترنت هستند که برای اجرای کارهای بخصوص به کارگرفته می شوند. این عوامل توانایی جستجو و مقایسه و انتخاب منابع اطلاعاتی بر اساس نیاز مطرح شده توسط کاربر را دارند

# توصیه ها و ابزارهایی برای جستجوی وب نامرئی

متخصصان جستجوی وب میگویند موتورهای جستجو مثل گوگل و یاهو فقط 1 درصد (surface Web) از اطلاعات وب برای جستجو اطلاعات، اینترنت را مورد جستجو قرار میدهند مابقی اطلاعات اینترنت را وب نامرئی یا deep Web، invisible Web، dark Web، hidden Web و Web مینامند.

خوشبختانه برای جستجو قسمتهایی از وب پنهان ابزارها و موتور جستجوهای متعددی ایجاد شده و در یافتن اطلاعات دلخواه کمک خوبی هستند. این موتور جستجوها از الگوریتم های پیشرفته رنکینگ و language-analysis استفاده می کنند که اطلاعات مفید و مرتبط را چکیده می کنند

# موتور جستجوهای وب نامرئی

**DeepDyve**: یکی از جدیدترین موتور جستجوها که بر وب نامرئی تمرکز کرده است

**CloserLook Search** جستجوگری برای اطلاعات پزشکی، سلامت، داروها و ...

**CompletePlanet CompletePlanet**: بیش از هفت هزار دیتابیس و موتور جستجو در این پایگاه در دسترس است و یکی از راه های عالی برای جستجوی وب نامرئی می باشد

**Daylife**: جستجوی اخبار به همراه تصاویر، مقالات و ...

**spock**: جستجوی افراد در وب پیدا کردن پروفایلها و تصاویر مخفی دوستانتان

**The WWW Virtual Library**: یکی از قدیمی ترین پایگاه اطلاعاتی وب که با کلمات کلیدی یا دسته بندی ها میتوانید به اطلاعاتش دست یابید.

# موتور جستجوهای وب نامرئی

**pipl:** طراحی شده برای جستجو وب نامرئی ی نتایجی که این وب سایت در دسترس قرار می دهد از نظر بعضی از کاربرانش خطرناک است (حریم شخصی)

**CustomSearchEngine:** لیستی از جستجوگرهای سفارشی سازی شده گوگل

**SurfWax:** موتور جستجویی ساده و پر محتوا

**Freebase:** مجموعه ای از میلیونها بانک اطلاعاتی در موضوعات متنوع

**RefSeek RefSeek:** موتور جستجویی برای دانشجوین و محققان ،بیشتر از یک میلیارد ستند که شامل صفحات وب ، مجلات ،روزنامه ها ، کتابها و دایره المعارف را جستجو می کند

# دسته بندي يا خوشه بندي پايگاه داده

دسته بندي مبتني بر پرس و جو

دسته بندي مبتني بر خزش

يادگيري ماشين (هوش مصنوعي)

# دسته بندی مبتنی 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 جو 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 بندی مبتنی بر خزش

- در واقع یک نرم افزار هستند که می توانند از محتوای صفحات نسخه برداری کنند. از این نسخه ها برای ایندکس کردن صفحات استفاده می شود.
- بدین معنا که علاوه بر آدرس، محتوای صفحات را نیز رتبه بندی کرده، بدین ترتیب کیفیت محتوا به صورت سه گانه در قالب (شی، صفت، مقدار) توصیف میشوند
- خزنده ها در ابتدا لیستی از نشانی های وب را در اختیار دارند که به آنها دانه (Seed) گفته می شود.
- با مرور دانه ها و بررسی کد HTML این صفحات، تمامی پیوندهای صفحه مشخص شده و آنها را به لیست نشانی هایی که باید مرور نماید اضافه می کند

- موتورهای جستجو و برخی از سایت‌ها دارای خزنده‌ها و یا ربات‌هایی 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 معمولاً در بازه‌های زمانی منظمی اطلاعات را بروز کرده و با نسخه‌های قبلی مقایسه می‌کنند.

- به صورت عمومی نحوه کار Web crawler ها به این صورت است که ابتدا لیستی از URL ها (آدرس های وب) که به عنوان seed شناخته 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.