حریری (۱۳۹۰)، یوسفی (۱۳۸۸)، بیزا-یتس[1] و ریبریو- نتو[2] (۱۹۹۹) ترجمه قاسمی قاسمی در سال ۱۳۸۵ است که هر کدام ویژگی‌های خاص خود را دارند. کتاب حاضر نیز با توجه به اهمیت موضوع بازیابی اطلاعات تهیه و تدوین شده است. در این کتاب سعی شده است مباحث کلی به زبان ساده اما تخصصی بیان شود تا خوانندگان با اصول کلی نظام‌های بازیابی اطلاعات در عصر جدید آشنا شوند.

همچنین تلاش شده است تا حد ممکن از تکرار مباحث پرهیز شود و با زبانی ساده پنج موضوع اصلی حوزه بازیابی اطلاعات بررسی گردد. در فصل ۱، تاریخچه و تعاریف نظام‌های بازیابی ارائه شده است. در فصل ۲، به طراحی نظام‌های بازیابی اطلاعات با تأکید بر موتورهای جستجو توجه شده است. این تأکید از آن جهت بود که موتورهای جستجو یکی از پرکاربردترین و مهم‌ترین ابزارهای بازیابی اطلاعات هستند. از طرف دیگر، در هر نظام بازیابی نیز موتور جستجو مهم‌ترین بخش است. مبحث فصل ۳، مدل‌های بازیابی اطلاعات است. در این فصل سعی شده است مدل‌های بازیابی اطلاعات بر اساس یک نظام مشخص بررسی شوند. در فصل ۴، موضوع "ربط" از بنیادی‌ترین مفاهیم حوزه بازیابی اطلاعات بررسی شده است. در فصل ۵ سعی شده است معیارهای ارزیابی به شکلی اصولی و نظام‌مند به زبانی ساده بیان شوند.

این اثر می‌تواند در درس‌های مرتبط با بازیابی اطلاعات در رشته‌های مختلف بخصوص علم اطلاعات و دانش‌شناسی استفاده شود. امیدواریم مورد توجه صاحب‌نظران، پژوهشگران و دانشجویان این حوزه قرار گیرد و از بازخوردهای مثبت و منفی آن‌ها در ویرایش‌های بعد استفاده کنیم.

1 Baeza-Yatez
2 Riberio- Neto

# فصل ۱

## تعاریف و تاریخچه

نظام بازیابی اطلاعات شامل دو مفهوم کلیدی ذخیره و بازیابی است. نقطه مشترک تمام تعاریف، این دو مفهوم اصلی است. در فرهنگ لغت مریام وبستر[1] (۲۰۱۴) برای عبارت نظام بازیابی اطلاعات تعریفی وجود ندارد؛ اما برای بازیابی اطلاعات دو تعریف آمده که در هر دو به ارتباط آن با یک نظام رایانه‌ای اشاره شده است.

"بازیابی اطلاعات، فنون ذخیره‌سازی، بازیابی و اغلب انتشار اطلاعات ثبت‌شده به‌ویژه با استفاده از یک نظام رایانه‌ای است".

تعریف دوم جامع‌تر است: "یافتن[2] اطلاعات، به‌ویژه در یک پایگاه داده ذخیره شده در یک رایانه. دو رهیافت عمده آن، تطبیق واژگان پرس‌وجو در برابر نمایه پایگاه داده (جستجوی کلمات کلیدی) و پیمایش پایگاه داده با استفاده از پیوندهای فرامتن یا فرارسانه است. جستجوی کلیدواژه‌ای، رویکرد غالب برای بازیابی متنی از اوایل دهه ۱۹۶۰ بوده و فرامتن اکنون تا حد زیادی به برنامه‌های کاربردی ارزیابی اطلاعات شخصی یا شرکت‌های بزرگ محدود شده است. فنون بازیابی اطلاعات، به‌طور نمونه با تحولات موتورهای جستجوی اینترنتی مدرن، زبان طبیعی، فرامتن و

1 Free Merriam-Webster Dictionary
2 Recover

جستجوی کلیدواژه ترکیب شده است. پژوهشگران درگیر با هوش‌مصنوعی فنون دیگری را که به‌دنبال دقت بالای بازیابی اطلاعات‌اند مطالعه کرده‌اند".

پاتل[1] (۲۰۱۳) به نقل از دایرةالمعارف بریتانیکا، بازیابی اطلاعات را دقیقاً مطابق با بخش اول تعریف دوم فرهنگ لغت مریام‌وبستر(۲۰۱۴) تعریف کرده است:

"دریافت اطلاعات به‌ویژه در یک پایگاه اطلاعاتی ذخیره شده در یک رایانه".

همچنین از فرهنگ لغت تخصصی علوم و فن‌آوری[2] تعریف زیر را نقل کرده است:

"بازیابی اطلاعات، فن و فرآیند جستجو، یافتن و تفسیر اطلاعات از حجم زیادی از داده‌های ذخیره شده است."

نکته مهم در تعاریف ذکر شده آن است که برای بازیابی اطلاعات اصطلاح "recover" به‌کار رفته است. تفاوت تعریف فرهنگ لغت تخصصی علوم و فن‌آوری با دیگر فرهنگ‌ها استفاده از عبارت "تفسیر اطلاعات" است.

فرهنگ لغت کالینز[3] (۲۰۱۴) به عنوان پیشگام در انتشار فرهنگ لغت هم، عبارت "recover" را برای تعریف نظام بازیابی اطلاعات به‌کار برده است: " نظامی برای یافتن اطلاعات خاص از داده‌های ذخیره شده".

در فرهنگ لغت آکسفورد[4] (۲۰۱۴) نیز همین اصطلاح برای تعریف بازیابی اطلاعات وجود دارد. اما تفاوت در تعریف آکسفورد استفاده از کلمه ردیابی[5] است: "ردیابی و یافتن اطلاعات خاص از داده‌های ذخیره‌شده".

برای یک تعریف استاندارد می‌توان به تعریف ایزو[6] اشاره کرد. شمس‌اژه‌ای و امیدی‌فر (۱۳۸۶) بازیابی اطلاعات را طبق ایزو ۲۳۸۲/۱ این‌گونه تعریف کرده‌اند:

" اعمال، شیوه‌ها و رویه‌هایی برای بازیابی اطلاعات ذخیره‌شده جهت تهیه اطلاعات حول موضوعی معین. در عمل، تعریف شامل نمایه‌سازی متن، تحلیل پرسش و تحلیل ربط است. استاندارد، متن، جداول، نمودارها، گفتار، تصویر و غیره را به عنوان داده مشخص می‌کند. هم‌چنین ابررسانه را برای تمایز بین متون ساخت‌یافته به صورت غیرخطی و متون (مدارک) خطی تعیین می‌کند و اطلاعات را دانش مرتبطی می‌داند که برای پیشرفت حل مشکل، دانش‌یابی و غیره است. این موضوع، پیونددهنده یک مفهوم، در مقابل یک رشته کاراکتری (واژه) است".

1 Patel
2 Science and Technology Dictionary
3 Colins
4 Oxford dictionary
5 Tracing
6 ISO: International Standard Organization



بازیابی اطلاعات را پژوهشگران زیادی تعریف کرده‌اند. از نظر کرستانی و پاسی[1] (۱۹۹۸) بازیابی اطلاعات، شاخه‌ای از علم رایانه است که به ذخیره و دسترسی سریع به اطلاعات کمک می‌کند. این اطلاعات می‌تواند به هر شکل باشد؛ متنی، دیداری، شنیداری و غیره. وظیفه نظام بازیابی، آسان نیست و مهم‌تر آن‌که اغلب مجموعه‌های مدارک در نظام‌های بازیابی اطلاعات باید به محتوای چندین هزار و گاهی میلیون‌ها مدرک رسیدگی کنند.

مطابق تعریف کوالسکی[2] (۱۹۹۷)، نظام بازیابی اطلاعات، می‌تواند اطلاعات را ذخیره، بازیابی و نگهداری کند. اطلاعات در این زمینه می‌تواند ترکیبی از متن (داده‌های تاریخی و عددی)، تصاویر، صدا و ویدئو و دیگر اشیای چندرسانه‌ای باشد. اگر چه شکل یک شیء در نظام بازیابی مختلف است؛ اما جنبه متنی اطلاعات تنها نوع داده است که به شیوه کاملاً عملی در نظام بازیابی اطلاعات پردازش می‌شود. انواع داده‌های دیگر به عنوان منابع اطلاعاتی بسیار مفید هستند اما در درجه اول برای جستجوی متنی به اطلاعات متنی پیوند خورده‌اند.

در این تعاریف، به‌جز موضوع ذخیره و بازیابی، تأکید بر نوع اطلاعات است که می‌تواند متن، تصویر، صدا و ویدئو باشد. در ادامه به تعاریف حوزه تخصصی کتابداری و اطلاع‌رسانی اشاره می‌شود.

هورلند[3] (۲۰۰۶) در کتاب " مفاهیم هسته در کتابداری و اطلاع رسانی" تعاریف بازیابی اطلاعات را جمع‌آوری کرده که به پژوهشگران به‌نام حوزه بازیابی اطلاعات مربوط می‌شود. به‌طورمثال در این منبع تعاریفی از موئرز (۱۹۵۱)، ون‌ریجسبرگن (۱۹۷۹)، کمپ[4] (۱۹۸۸)، میکسا[5] (۱۹۹۹)، وارنر[6] (۲۰۰۲)، بیزا-یاتز یاتز و ریبریونتو (۱۹۹۹)، راثون[7] (۱۹۹۶) آمده است. از تعریف موئرز (۱۹۵۱) تا وارنر (۲۰۰۲) می‌توان تأثیر فن‌آوری و پیشرفت‌های مختلف را بر حوزه بازیابی اطلاعات در طول زمان درک کرد. هورلند با بررسی تعاریف، خود به نکته مهمی اشاره می‌کند و معتقد است که در واژه retrieve پیشوند "re" به نظر می‌رسد "چیزی" را نشان می‌دهد که دوباره یافت شده است، یعنی "آن چیز" در مراحل

1 Crestani and Pasi
2 Kowalski
3 Hjørland
4 Kemp
5 Miksa
6 Warner
7 Ruthven

پیش از این، شناسایی شده بوده و اکنون دوباره یافت شده است. از نظر او، این دیدگاه شاید در زمانی که اطلاعات در نظام‌های بسته برای اهداف خاص ارائه می‌شوند درست باشد اما وقتی که اقلام[1] اطلاعاتی به‌صورت غیرمترقبه[2] یا با جستجوی متنی بازیابی می‌شوند مشکل‌ساز است. در نظام بازیابی، اطلاعات یافت یا شناسایی می‌شوند و لزوماً بازیابی نمی‌شوند. فرهنگ‌های لغت مختلف نیز از اصطلاح recover برای تعریف بازیابی استفاده کرده‌اند که به‌جز معنی یافتن، معانی بازیافتن و دوباره به‌دست‌آوردن را دارد که خود مفهوم بازیابی را به چالش می‌کشد.

لنکستر[3] (۱۹۶۸) در تعریف خود اعلام کرد که نظام بازیابی اطلاعات نمی‌تواند نمی‌تواند در دانش جستجوگر اطلاعات، تغییری به‌وجود آورد و کار آن جایابی، اعلام وجود یا عدم وجود اطلاعات است:

"[کار] نظام بازیابی اطلاعات، آگاهی دادن، یعنی تغییر دانش کاربر در حوزه موضوعی مورد پژوهش نیست و صرفاً کاربر را از بودن یا نبودن مدارک مربوط به نیاز او و محل نگهداری آن مدارک آگاه می‌کند" (نقل در جیج و اونوچکوا[4]، ۲۰۱۱).

سالتون[5] (۱۹۸۳) تعریف کاملاً تخصصی از نظام بازیابی اطلاعات ارائه کرده است: "نظام بازیابی اطلاعات، نظامی است که برای ذخیره اقلام اطلاعاتی استفاده می‌شود. این اقلام باید پردازش و جستجو شوند و در جوامع کاربری مختلف اشاعه یابند (پاتل، ۲۰۱۳). در کنار این تعریف، در کتاب مقدمه‌ای بر بازیابی اطلاعات، واژه بازیابی اطلاعات از دیدگاه آکادمیک تعریف شده است: بازیابی اطلاعات یافتن مواد (معمولاً مدارک) از یک طبیعت غیرساخت یافته (معمولاً متن) است که نیاز اطلاعاتی را از میان مجموعه‌های بزرگ (معمولاً ذخیره در رایانه‌ها) برآورده می‌کند (منینگ، رغاوان و شوتز[6]، ۲۰۰۸). در تعاریف جدیدتر به سازماندهی اطلاعات در نظام بازیابی اطلاعات نیز توجه شده است:

"نظام بازیابی اطلاعات برای بازیابی مدارک یا اطلاعات مورد نیاز جامعه کاربران طراحی شده است و باید اطلاعات صحیح را در دسترس کاربر مناسب قرار دهد. بنابراین یک نظام بازیابی اطلاعات با هدف جمع‌آوری و سازماندهی اطلاعات

1 items
2 serendipity
3 Lancaster
4 Jegede and Onwuchekwa
5 Salton
6 Manning, Raghavan and Schütze



در یک یا چند زمینه موضوعی و برای فراهم کردن اطلاعات به‌وجود آمده است تا به‌محض درخواست کاربر، اطلاعات را در اختیار او قرار دهد" (جیج و اونوچکوا، ۲۰۱۱). تعریف جدیدتر پاتل در سال ۲۰۱۳ نشان می‌دهد نوع نگاه به بازیابی اطلاعات تغییر کرده است و پیشرفت‌های اخیر در حوزه پایگاه‌های اطلاعاتی بر تعریف آن‌ها موثر بوده‌اند:

"بازیابی اطلاعات، علم جستجوی اطلاعات در مدارک، جستجو برای خود مدارک، فراداده‌ها که توصیف‌کننده مدارک‌اند یا جستجو در پایگاه‌های اطلاعاتی، چه از نوع داده‌ای رابطه‌ای مستقل چه از نوع اطلاعاتیِ شبکه فرامتن مانند اینترنت یا شبکه جهانی وب یا اینترانت، برای متن، صدا، تصاویر یا داده‌هاست" (پاتل، ۲۰۱۳)

همه این تعاریف نشان می‌دهد "بازیابی اطلاعات" مفهوم گسترده‌ای است و می‌توان از دیدگاه‌های مختلف آن را تعریف کرد و این تعریف‌ها بسیار تحت تأثیر زمان و پیشرفت فن‌آوری است. بدین جهت ارائه تعریفی جامع کار دشواری است؛ اما با توجه به مفاهیم کلیدی در تعاریف، می‌توان نظام بازیابی اطلاعات را این‌گونه تعریف کرد:

" نظام شناسایی، ذخیره، سازماندهی، بازنمایی، جستجو و ارائه اطلاعات در یک نظام رایانه‌ای است. اطلاعات به‌صورت غیرساختاریافته می‌تواند به هر شکل مانند متن، صدا، تصویر، گفتار، جدول یا نمودار باشد. نظام بر محور پرس‌وجوی اطلاعات کاربر و ربط اطلاعات با نیاز اطلاعاتی اوست و به همین دلیل فنون بازیابی اطلاعات مانند تحلیل و فرمول‌بندی پرسش، راهبردهای جستجو، تحلیل ربط و ارزیابی بروندادهای اطلاعات در آن اهمیت بسیاری دارد. اطلاعات ضمیمه شده به این اطلاعات مانند فرامتن و فرارسانه یا اطلاعات توصیف کننده آن‌ها مانند فراداده را نیز می‌توان جستجو کرد".

## تاریخچه نظام بازیابی اطلاعات

بازیابی اطلاعات همیشه همزاد و همراه بشر بوده است. یکی از اهداف مدون کردن اطلاعات این بوده که در آینده قابل دسترسی و استفاده باشد. اولین فهرست‌های کتابخانه‌های که از نظر ماهیت مربوط به حوزه سازماندهی اطلاعات بوده است در واقع برای افزایش نقاط دسترسی به اطلاعات تدوین شده‌اند. رشد منابع اطلاعاتی سبب شده تا اهمیت بازیابی و دسترسی به اطلاعات بیش‌تر از قبل شود. ظهور وب و شخصی‌سازی تولید اطلاعات یعنی فراگیرشدن تولید اطلاعات در میان عموم مردم

باعث شده است تا حوزه بازیابی اطلاعات از یک موضوع مهم به یک چالش بزرگ تبدیل شود. جان نابیست معتقد است "ما در اطلاعات غوطه‌وریم، اما از کمبود دانش رنج می‌بریم" (دایتینگر، گات، مائورر و پیوک[1]، ۱۹۹۹ نقل در حسینی، ۱۳۹۰، ص۶۲۶). شاید به همین دلیل دایتینگر و دیگران، وجود مخازن دانش بسیار بزرگ و بدون ساختار را به‌همراه افزایش سریع اطلاعات، دو ویژگی جوامع کنونی می‌دانند (حسینی، ۱۳۹۰). هر چند هدف واقعی "انقلاب ناشی از ظهور وب و تأثیر آن بر ذخیره و دسترسی به اطلاعات، تنها دسترسی‌پذیری اطلاعات نیست بلکه کارآمدسازی دسترسی به اطلاعات است" (خسروی و فتاحی، ۱۳۸۹، ص۲۲۰) با این‌حال، این انقلاب سبب شد تا بازیابی اطلاعات و مسائل مربوط به نظام‌های بازیابی اهمیت بیش‌تری یابد. برای درک بهتر این موضوع و چالش‌های آن، دانستن تاریخچه اجمالی تکامل نظام‌های بازیابی ضروری است. در این بخش سعی شده است روند تاریخی بازیابی اطلاعات و نظام‌های بازیابی به صورت کلی و در چند دوره مهم تاریخی بررسی شود.

در بسیاری از متون، تاریخچه نظام‌های بازیابی اطلاعات از دهه ۱۹۵۰ به بعد بررسی شده است (ساندرسون و کرافت[2]، ۲۰۱۲؛ لسک[3]، ۱۹۹۶؛ دوبین و اسمیت[4]، ۱۹۹۹). به دلیل آن‌که نظام‌های بازیابی اطلاعات در دوره‌های قبل، دستی بودند؛ یعنی نمایه‌ها و فهرست‌ها به شکل چاپی و کارتی تهیه می‌شدند (بهمن‌آبادی، ۱۳۸۶). لسک (۱۹۹۵) تاریخچه نظام بازیابی اطلاعات را به هفت دوره تکامل (بر اساس هفت دوره تکامل انسان از نظر شکسپیر) تقسیم کرده است. ویژگی این تقسیم‌بندی آن است که لسک در سال ۱۹۹۶، دوره بعد از سال ۲۰۱۰ را نیز پیش‌بینی کرده بود. این هفت دوره شامل دوره کودکی[5] (۱۹۴۵–۱۹۵۵)، دانش‌آموزی[6] (دهه ۱۹۶۰)، بزرگسالی[7] (دهه ۱۹۷۰)، بلوغ[8] (دهه ۱۹۸۰)، بحران میانسالی[9] (دهه ۱۹۹۰)، انجام[10] انجام[10] (سال ۲۰۰۰ تا ۲۰۱۰) و بازنشستگی[1] (بعد از سال ۲۰۱۰) است. شاید بتوان

1 Dietinger, Gut, Maurer and pivec
2 Sanderson and Croft
3 Lesk
4 Dubin and Smith
5 Childhood
6 Schoolboy
7 Adulthood
8 Maturity
9 Mid-Life Crisis
10 Fulfillment



بتوان به‌گونه‌ای این دسته‌بندی زمانی را پذیرفت، اما تقسیم تحول نظام‌های بازیابی اطلاعات بر اساس دوره‌های زندگی انسان ممکن است وسعت دید را محدود کند. با توجه به تحولات فراوان فن‌آوری و پیشرفت‌های عمده بعد از سال ۲۰۱۰ باید پذیرفت نظام‌های بازیابی اطلاعات همیشه نسبت به دوره بعدی خود، دارای کاستی‌های فراوانند. پیشرفت فن‌آوری به شدت بر این نظام‌ها تأثیر می‌گذارد و روند توسعه و تکامل آنها به‌صورت مداوم ادامه دارد.

در ادامه مهم‌ترین اتفاقات مربوط به نظام‌های بازیابی اطلاعات آمده است.

## از ابتدا تا دهه ۱۹۴۰

اطلاعات چندانی درباره نظام‌های بازیابی اطلاعات قبل از دهه ۱۹۴۰ وجود ندارد. نظام‌های بازیابی اطلاعات قبل از این دهه، دستی و بسیار مبتنی بر تکامل فهرست‌های کتابخانه‌های بوده‌اند. در واقع، ریشه‌های نظام بازیابی اطلاعات به اولین فهرست‌های کتابخانه‌های برمی‌گردد. تفکر اولیه آن به کتابداری و حوزه علم اطلاعات مربوط است (وبر[2]، ۲۰۱۰، ص۱۲). برخی معتقدند که برای اولین بار بازیابی اطلاعات در چهار هزار سال پیش همزمان با فهرست‌های اولیه کتابخانه‌ها در سومر و به عقیده برخی، دو هزار سال پیش در یونان ظاهر شده است. نکته مسلم آن است که زبان نوشتاری، سه هزار سال قبل از میلاد وجود داشته است (فیلدن[3]، ۲۰۰۲). ثبت اطلاعات به صورت مکتوب چالش دسترسی به اطلاعات را برای بشر ایجاد کرد. یکی از دلایل ایجاد اولین فهرست‌های کتابخانه‌ای برای افزایش نقاط دسترسی به منابع بوده است.

ایده نخست در حوزه بازیابی اطلاعات در طبیعتِ خود، اصولاً فلسفی است و به چگونگی رده‌بندی و سازماندهی اطلاعات مربوط است (وبر، ۲۰۱۰، ص۱۲) و فهرست‌های به‌وجود آمده فهرستی از عناوین کتاب‌ها یا طبقه‌بندی گسترده‌ای از موضوعات بودند. قبل از اختراع کاغذ، رومی‌ها و یونانیان باستان، اطلاعات خود را روی طومارهای پاپیروس ثبت می‌کردند. رومیان برای هر پاپیروس، یک برچسب داشتند که شامل خلاصه‌ای از محتوای پاپیروس بود و در بسیاری از موارد، باعث می‌شد که نیازی به بازکردن یک پاپیروس بزرگ نباشد. در مدارک یونانی قرن پنجم

1 Retierment
2 Webber
3 Fielden

قبل از میلاد نیز، چکیده‌هایی شبیه به این کشف شده است. با این‌حال هیچ طرح رده‌بندی واقعی از کتابخانه‌های روم و یونان در آن زمان برجا نمانده است (لانگویل و میر[1]، ۲۰۰۶). الیوت و روز[2] (۲۰۰۹) معتقدند اولین فهرست کتابخانه‌ای را شاعر یونان باستان، کالیماخوس به‌وجود آورد (ساندرسون و کرافت، ۲۰۱۲) که هدف آن جستجوی ساده‌تر اطلاعات بود. فهرست مطالب به عنوان یکی از ابزارهای اولیه بازیابی اطلاعات در کتیبه‌های یونانی به چشم می‌خورد. کتابخانه اسکندریه، ۲۸۰ سال قبل از میلاد، به عنوان کتابخانه بزرگ عصر خود از پاپیروس استفاده می‌کرد و حدود هفتصد هزار مدرک را در اختیار داشت. اما پیدایش کتاب، در قرن‌های بعدی و زمان تحریم واردات پاپیروس به یونان اتفاق افتاد. کتابخانه پرگاموم یونان به عنوان ماده جانشین، از پوست نازک حیوانات استفاده کرد که به آن پارشمن می‌گفتند (فیلدن، ۲۰۰۲). در پرگاموم نمی‌توانستند پارشمن‌ها را مانند پاپیروس، به‌صورت طومار درآورند. به همین خاطر آن‌ها را به شکل صفحه درآوردند و به صورت کتاب به هم متصل کردند و کتاب را به‌وجود آوردند (لانگویل و میر، ۲۰۱۲). در طی سالیان، تغییر و تحول در شکل کتاب و فهرست‌ها ادامه داشت.

بعد از آن قرون وسطی از راه رسید که به دوران تاریکی شهرت دارد. در این دوره، دانش در انحصار گروه خاصی بود و نسخه‌برداری به کلیسا محدود شد. به همین دلیل اطلاعات به صورت شفاهی (شعر و آهنگ و داستان) ثبت گردید و بیش‌تر در ذهن یک قصه‌گو در یک دهکده کوچک جریان داشت تا به‌صورت مکتوب و در کتاب‌ها (لانگویل و میر، ۲۰۰۶). قوانین محدود کننده کلیسا برای استنساخ نسخ و سانسور باعث شد که دسترسی به اطلاعات بسیار محدود شود. در این دوران منابع از نظر کلیسا به دو شکل بودند. کتب مذهبی که اجازه تکثیر آن‌ها در اختیار کلیسا بود و کتب کفر و الحاد که استفاده و استنساخ آن‌ها ممنوع بود. با آغاز دوره رنسانس و نوزایی دوران سیاه و تاریک بازیابی و دسترسی به اطلاعات پایان یافت.

رنسانس تحولات عمده‌ای را به‌همراه داشت. آزادی‌های بیش‌تر باعث شکوفایی و پیشرفت در همه زمینه‌ها، از جمله حوزه دسترسی به اطلاعات شد. این دوران که از قرن ۱۴ تا ۱۷ میلادی است، شاهد یکی از بزرگ‌ترین اختراعات بشر بعد از اختراع خط بوده است. در سال ۱۴۵۵، گوتنبرگ اولین ماشین چاپ را به‌وجود آورد و جهان انقلابی بزرگ را در انتشار سریع اطلاعات تجربه کرد. این اختراع و افزایش سرعت

1 Langville and Mayer
2 Eliot and Rose



انتشار به معنای افزایش انتشارات مکتوب با کیفیت بالاتر بود و این موضوع همراه با اختراعاتی مانند کاغذ و چاپ مطبوعات، نیاز به بازیابی اطلاعات را در طول سالیان بعد از آن افزایش داد. طولی نکشید که با اختراع رایانه مردم دریافتند که می‌توانند از آن برای ذخیره‌سازی و بازیابی خودکار مقداری زیادی از اطلاعات استفاده کنند (سینال[1]، ۲۰۰۱). بدین‌ترتیب، رایانه بیش از پیش، انتشار اطلاعات مکتوب را تحت تأثیر قرار داد. هر چند تا دهه ۱۹۴۰، که اولین رایانه الکترونیکی جهان اختراع شد (فیلدن، ۲۰۰۲)، بازیابی به‌صورت دستی بود و پیشرفت‌ها نیز در حوزه بازیابی اطلاعات نسبت به سال‌های بعد از آن کند بود و همان‌طور سنتی باقی ماند.

رده‌بندی دیویی، رده‌بندی کتابخانه کنگره، و مستندسازی[2] از پیشرفت‌های مهم قبل از ۱۹۴۰ محسوب می‌شوند. مسأله مهم دیگر در سال‌های ۱۹۲۰ تا ۱۹۴۰، توزیع اطلاعاتی و کتابشناختی شامل ارائه قوانین زیف[3]، پارتو[4] و لوتکا[5] بود. هم‌چنین محمل اطلاعاتی به نام میکروفیلم به‌طورگسترده‌ای استفاده شد (داده کاوی: یادگیری ماشینی[6]، ۲۰۰۶) که تأثیرات زیادی بر حوزه بازیابی اطلاعات داشت. اما شروع دهه جدید و اختراع رایانه الکترونیک مبدأ تحولات بسیار در دهه آینده شد.

## دهه ۱۹۴۰

در قرن نوزدهم کارت‌های فهرست توسعه یافتند. این کارت‌ها، مدخل‌ها و کلیدهای جستجوی بیش‌تری داشتند و روزآمدسازی آن‌ها آسانتر بود؛ اما مشکل آن بود که محلی بودند و تنها یک نسخه از آن وجود داشت و کاربر باید برای یافتن اطلاعات، به محل فیزیکی کتابخانه مراجعه می‌کرد. با رشد کتابخانه‌ها، این کارت‌ها بیش از پیش کارآیی خود را از دست دادند چون نقاط دسترسی محدودی داشتند و به‌خاطر فقر واژگان، کم‌تر می‌توانستند در کتابخانه‌های توسعه‌یافته و مجموعه‌های بزرگ‌تر به بهبود جستجوی اطلاعات کمک کنند (میدو، ۱۹۹۹) تا این‌که در این دهه، اولین رایانه الکترونیکی به‌وجود آمد (فیلدن، ۲۰۰۲) و از همان ابتدا برای سازماندهی اطلاعات استفاده گردید.

1 Singhal
2 dicumenalism
3 Zipf's law
4 Pareto
5 Lotka' law
6 Data Mining: machine learning

در سال ۱۹۴۵، وانور بوش[1] در مقاله پیشگامانه خود با عنوان " همان‌طور که ممکن است فکر کنیم" ایده دستیابی خودکار به مقدار زیادی از اطلاعات را مطرح کرد که در دهه ۱۹۵۰، با توصیفات واقعی‌تری از چگونگی جستجوی خودکار اطلاعات در آرشیوهای متنی محقق شد (سینال، ۲۰۰۱). بوش ماشینی به نام "ممکس[2]" ترسیم کرد که می‌شد کتاب‌ها، پیشینه‌ها و ارتباطات را با آن ذخیره و مکانیزه کرد و از اطلاعات ذخیره شده با سرعت و انعطاف بیش‌تری استفاده نمود (بوش، ۱۹۴۵). این ایده بسیار در زمان خود پیشرو بود و از آن استقبال شد و بسیار زود با کیفیت بهتر به واقعیت تبدیل گردید. اما اتفاقات این دوران فقط به ماشین ممکس بوش ختم نشد.

در این دهه، تلاش شد که ایده اصلی جستجوی متنی با رایانه به دقت تشریح شود. مهم‌ترین آن‌ها روش تأثیرگذار نمایه‌سازی لون[3] بود (سینال، ۲۰۰۱). لون نخستین کسی بود که معتقد بود می‌توان واژه‌های معینی را به‌طور خودکار از متن استخراج کرد و با آن‌ها محتوای متن را بازنمایی نمود (جلالی‌منش، علیدوستی، خسروجردی، ۱۳۹۲). نمایه‌سازی یکی از ابزارهای مهم در بازنمون و بازیابی اطلاعات در نظام‌های بازیابی است. روش نمایه‌سازی لون، در بعد از دوران خود کاربرد زیادی در نمایه‌سازی و چکیده‌نویسی داشت.

تحولات دیگر این دهه، اتفاقات بعد از جنگ جهانی دوم بود. ارتش امریکا با مشکلات نمایه‌سازی و بازیابی مدارک پژوهشی علمی که زمان جنگ از آلمان‌ها گرفته بود مواجه شد (دوبین واسمیت،۱۹۹۹ ) و به دنبال راهی برای رهایی از این مشکلات بود. به‌طورکلی، بعد از اتمام جنگ افزایش انتشارات مسأله بازیابی اطلاعات را بیش از پیش به یک چالش بزرگ تبدیل کرده بود.

در سال ۱۹۴۸ با افزایش انتشارات، در کنفرانسی که در انجمن سلطنتی انگلستان برگزار شد هلمستروم[4] ماشینی به نام "یونیواک[5]" را توصیف کرد. یونیواک، می‌توانست منابع متنی مرتبط را با یک کد موضوعی جستجو کند. کدها و متن‌ها روی نوارهای مغناطیسی ذخیره شده بودند. ماشین می‌توانست ۱۲۰ کلمه را در دقیقه پردازش کند. بدین ترتیب اولین منبع برای استفاده از رایانه برای جستجوی

1 Vannevar Bush
2 Memex
3 Lohn
4 Holmstrom
5 Univac



محتوا به‌وجود آمد (ساندرسون و کرافت، ۲۰۱۲). تحولات بیش‌تر در دهه بعد پیگیری شد.

## دهه ۱۹۵۰

کلوین موئرز اصطلاح " بازیابی اطلاعات" را خلق کرد (دوبین واسمیت، ۱۹۹۹) و بدین‌ترتیب با استفاده از نظام‌های داده‌پردازی برگه منگنه، نمایه‌سازی پس‌همارا به‌وجود آمد (بهمن‌آبادی، ۱۳۸۶). اتفاق مهم دیگر آن بود که شوروی ماهواره اسپوتینگ را با موفقیت به فضا فرستاد که سبب ترس و وحشت آمریکا شد. جامعه آمریکا فکر می‌کرد که در علم عقب مانده است (لسک، ۱۹۹۶). هم‌چنین آمریکاییان، نگران حمله موشک‌های دوربرد اتمی شوروی بودند. آن‌ها می‌دانستند که شوروی می‌تواند به شبکه‌های ارتباطی آن‌ها حمله کند. افزایش نگرانی‌ها بخاطر شکاف اطلاعاتی به‌وجود آمده، آنها را تشویق کرد تا بودجه‌هایی برای پژوهش در این زمینه اختصاص دهند و بدین‌ترتیب پشت صحنه نظام‌های جستجوی متنی ماشینی را فراهم کنند (دوبین و اسمیت، ۱۹۹۹؛ نشاط، ۱۳۸۷). این موضوع در نهایت در سال ۱۹۶۸، به پروژه آرپانت و ایجاد اینترنت امروزی منجر شد (اسلاتر۳[1]، ۲۰۰۲). از طرفی بسیاری از اولین نظام‌هایی که به‌گونه‌ای در بازیابی اطلاعات نقش داشتند در دهه ۱۹۵۰ ساخته شدند که شامل نمایه‌های کوئیک[2] و کشف‌اللغات[3] مورد استفاده در نظام‌های بازیابی اطلاعات بودند (لسک، ۱۹۹۶)

در سال ۱۹۵۶یوجین گارفیلد کار خود را با پایه‌گذاری شرکتی کوچک آغاز کرد. در اتاقک کوچک او تهیه سلف فهرست مندرجات جاری امروزی آغاز شد (بهمن‌آبادی، ۱۳۸۶). او با این کار یکی از تأثیرگذارترین تحولات حوزه بازیابی اطلاعات را یعنی نمایه استنادی علوم پایه‌گذاری کرد.

دو سال بعد آلن کنت[4] "انتخاب‌گر جستجوی دانشگاه وسترن رزرو[5]" را ساخت (لسک، ۱۹۹۶). از کاربردهای آن، امکان جستجوهای آزمون روی فایل چکیده‌های کدگذاری شده انجمن فلزات آمریکا[6] بود (ریز و کنت[7]، ۱۹۵۸).

1 Slater III
2 KWIC
3 concordances
4 Allen Kent
5 Western Reserve University (WRU) Searching Selector
6 American Society for Metals
7 Rees and Kent

در این سال‌ها، میتچل[1] با استفاده از رایانه یونیواک توانست یک میلیون چکیده را که با شش کد موضوعی نمایه‌سازی شده بودند جستجو کند. نانوس[2] نیز، نظامی از جنرال الکتریک را توصیف کرد که می‌توانست ۳۰۰۰۰۰ چکیده مدرک را جستجو کند (ساندرسون و کرافت، ۲۰۱۲).

## دهه ۱۹۶۰

دهه‌ای تأثیرگذار در بازیابی اطلاعات است و تقریباً در هر سال، اتفاقات مهمی در حوزه بازیابی اطلاعات افتاده است. سال‌های تجربه‌های بزرگ بود و در حقیقت دهه جهش و جنبش سریع برای بازیابی اطلاعات شناخته شده است و افراد بسیاری در کنفرانس‌های بازیابی اطلاعات شرکت کردند (لسک، ۱۹۹۶). دوران بازیابی رایانه‌ای به شیوه گسسته، پردازش دسته‌ای و نواری بود و در اوایل همین دهه اولین نظام‌های بازیابی اطلاعات بزرگ در امریکا در دولت فدرال ایجاد شد. اما فدا کردن یک نظام رایانه‌ای برای جستجوی گذشته‌نگر پیشینه‌های کتاب‌شناختی توجیه اقتصادی نداشت و هنوز هم ندارد. بسیاری از نظام‌های رایانه‌ای این دوره و بعد از آن چند منظوره بودند و طیف وسیعی از محصولات و خدمات را از یک عملیات ورودی منحصربه‌فرد تولید می‌کردند. این نظام‌ها، نتیجه فرآیند خودکارسازی نشر بودند (بلاچ، ۱۹۶۳)

در همین سال‌ها، تحولات بیش‌تر در حوزه علوم و استنادهای علمی به‌وجود آمد. مؤسسه کوچک یوجین گارفیلد (۱۹۶۰ و ۱۹۶۱) نام کنونی‌اش یعنی "مؤسسه اطلاعات علمی (آی.اس.آی) را یافت[3]. از طرفی مهم‌ترین واقعه سال ۱۹۶۱ نشر نخستین ویرایش "نمایه‌نامه استنادی علوم[4]" برای حوزه ژنتیک بود (نشاط، ۱۳۸۷). ۱۳۸۷).

از دیگر تحولات این دهه، روش‌های متفاوت نمایه‌سازی بود. از جمله نمایه‌سازی احتمالی که در این دهه ایجاد شد (بری[5]، ۲۰۰۹الف). مارون و کان در سال ۱۹۶۰، مقاله‌ای با عنوان "ربط، نمایه‌سازی احتمالاتی و بازیابی اطلاعات" در مجله،

1 Mitchel
2 Nanus
۳. ISI: قابل ذکر است این مؤسسه را تامسون رویترز خریداری کرد و اکنون به همین نام شهرت دارد.
4 Science citation index
5 Bury



ای‌سی‌ام[1] منتشر کردند.

در این سال‌ها جرالد سالتون یکی از بزرگان حوزه بازیابی اطلاعات و نویسنده کتاب مشهور "درآمدی بر بازیابی اطلاعات[2]" (سالتون و مک‌گیل[3]، ۱۹۸۶)، در هاروارد و بعدها در کورنل، کار خود را در این حوزه شروع کرد (دوبین و اسمیت، ۱۹۹۹).

در سال ۱۹۶۲، کلوردون[4] اولین یافته‌های کرانفیلد را منتشر کرد که مدل پیشرفته برای ارزیابی نظام‌های بازیابی بود (دوبین و اسمیت، ۱۹۹۹). بسیاری از افراد در آغاز عصر مدرن بازیابی اطلاعات به تجربه‌های نمایه‌سازی کرانفیلد در دهه ۱۹۶۰، استناد کردند (کلوردون، میلز[5] و کین[6]، ۱۹۶۶). آزمایش‌های کلوردون در این دهه یکی از تأثیرگذارترین پژوهش‌ها در حوزه ارزیابی نظام‌های بازیابی اطلاعات بود. البته آزمایش‌های اولیه در سال ۱۹۵۸ انجام شد که در آن با استفاده از ۱۸۰۰۰ مدرک و ۱۲۰۰ موضوع، نظام‌های نمایه‌سازی همارا با نظام‌های سنتی مقایسه گردید و نتایج به‌دست آمده مبنای تنظیم دومین فرضیه‌های پژوهش قرار گرفت. نتایج این پژوهش، نه تنها الگویی برای بسیاری از پژوهش‌های ارزشیابی، عملیاتی و تجربی شد؛ بلکه بعدها در متون درسی این رشته نیز اشاعه یافت (زوارقی،۱۳۹۰). منشا مدل‌های آزمایشگاهی ارزیابی نظام‌های بازیابی اطلاعات، به همین مدل ارزیابی کلوردون دوم برمی‌گردد. این انگاره علوم رایانه در حوزه بازیابی اطلاعات، به دنبال توسعه هر چه بهتر الگوریتم‌ها و نظام‌های بازیابی اطلاعات است (ککلاینن و جاولین، ۲۰۰۲).

در اواسط این دهه، کتابخانه پزشکی آمریکا، نظام بازیابی و تحلیل متون پزشکی یا همان مدلارز[7] را توسعه داد. این نظام، اولین نظام بازیابی دسته‌ای[8] و اولین پایگاه پایگاه اطلاعاتی بزرگ ماشین‌خوان بود (دوبین و اسمیت، ۱۹۹۹). مدلارز تحول عمده‌ای در حوزه دستیابی به اطلاعات پزشکی به وجود آورد که هنوز هم حوزه پزشکی از فواید این تأثیرات بهره می‌برد. ایجاد نظام مدلاین بر روی دیسک و بعدها به‌صورت بر خط بسیار مدیون مدلارز است. در همین سال‌ها بود که لنکستر،

1 ACM
2 Introduction to Information Retrieval
3 McGill
4 Cleverdon
5 Mills
6 Keen
7 MEDLARS: Medical Literature Analysis and Retrieval System
8 Batch-information retrieval

مطالعات ارزیابی نظام ارزیابی مدلارز را کامل کرد و اولین ویرایش متنی آن را روی نظام بازیابی اطلاعات منتشر ساخت (دوبین و اسمیت، ۱۹۹۹).

در سال ۱۹۶۶، مطالعه آزمایشی بر روی پروژه مارک[1] آغاز شد (اورم[2]، ۱۹۷۵) که بر بازیابی اطلاعات تأثیر چشمگیری گذاشت. ایجاد فهرست‌های ماشین‌خوان یکی از تحولات مهم بازیابی اطلاعات است که مقدمات آن در این دهه پایه‌ریزی شد و بعد از پیاده‌سازی به جز لاینفک ساختار سازماندهی کتابخانه‌ها تبدیل گردید.

در این دوران مفاهیم دقت و بازیافت تعریف شدند (لسک، ۱۹۹۶) و در پی آن یکی از اتفاقات مهم دیگر در این دوره، یعنی ارزیابی بازخورد ربط اتفاق افتاد (بری، ۱۹۹۹ الف) این فرآیند که در اواسط دهه ۱۹۶۰، به‌وجود آمد در واقع یک فرآیند خودکار و کنترل‌شده برای فرمول‌بندی مجدد پرسش بود (سالتون و باکلی[3]، ۱۹۹۷) و و به این دلیل مهم است که هنوز هم کاربران با ساختار نظام‌های بازیابی اطلاعات آشنا نیستند و فرمول‌بندی پرسش به صورت دقیق برای آن‌ها مشکل است. با بازخورد ربط در نظام‌های بازیابی، کاربر می‌تواند پرسش را دوباره فرمول‌بندی کند تا اطلاعات بهتری به‌دست آورد.

یکی از تأثیرگذارترین پیشرفت‌ها که بر همه حوزه‌های دانش بشری تأثیر گذاشت و نظام‌های بازیابی اطلاعات سود زیادی از آن برده‌اند اینترنت است که در این دهه پایه‌گذاری شد. پروژه آرپانت ابتدا با اهداف سیاسی شکل گرفت ولی مقدمه شکل گرفتن اینترنت جهانی در سال ۱۹۶۸ شد. همان‌طور که قبلاً نیز گفته شد، اقدام اتحاد جماهیر شوروی در پرتاب موفقیت آمیز ماهواره اسپوتینگ به فضا در دوران جنگ سرد، آمریکائیان را نگران حمله اتمی آنها کرد. آن‌ها هراس داشتند که با پیشرفت علم در شوروی، موشک‌های دوربرد به شبکه‌های ارتباطی آن‌ها حمله کند. به همین دلیل در پی آن بودند شبکه‌ای ایجاد کنند که چند کاربر بتوانند همزمان از آن استفاده کنند و امکان انتقال اطلاعات در آن فراهم باشد. بدین‌ترتیب آژانس پروژه‌های تحقیقاتی پیشرفته پنتاگون، تصمیم گرفت بر روی چنین پروژه‌ای تحقیق کند. در پاییز ۱۹۶۹ اولین نوع از گره‌های شبکه‌ای در یوسی‌ال‌ای[4] شکل گرفت. در دسامبر همان سال، چهار رایانه در شبکه اولیه نصب شدند و بعد از مدت کوتاهی توانستند داده‌ها را در

1 MARC (MAchine-Readable Cataloging)
2 Avram
3 Buckley
4 UCLA



بین خود انتقال دهند (استرلینگ[1]، ۱۹۹۳). بدین ترتیب اینترنت اولیه شکل گرفت.

به‌طورکلی دهه ۱۹۶۰ را باید دهه نظام‌های برون‌خطی[2] دانست. ویژگی‌های نظام‌های بازیابی اطلاعات بسیار شبیه به‌هم بودند. نظام‌های بازیابی پردازش دسته‌ای[3] برون‌خطی که از نوارهای مغناطیسی به عنوان رسانه ذخیره‌سازی استفاده می‌کردند (بلاچ، ۱۹۶۳).

## دهه ۱۹۷۰

در این دهه، با توسعه سخت‌افزارها و نرم‌افزارهای رایانه‌ای (بهمن‌آبادی، ۱۳۸۶) و ارتباط از راه دور (بلاچ، ۱۹۶۳)، امکان جستجوهای پیوسته یا تعاملی فراهم آمد (بهمن‌آبادی، ۱۳۸۶). نظام‌های پردازش دسته‌ای برون‌خطی دهه ۱۹۶۰ با نظام‌های بازیابی پیوسته تعاملی دهه ۷۰ و بعد از آن دنبال شدند (بلاچ، ۱۹۶۳). نظام‌های پیوسته بازیابی، علاوه بر افزایش سرعت، امکان دریافت بازخورد در روند جست‌وجو و در صورت لزوم، تغییر و اصلاح آن را به استفاده کننده می‌دادند.

مهم‌ترین مزیت بازیابی پیوسته، امکان دسترسی از راه دور بود (بهمن‌آبادی، ۱۳۸۶). رسانه ذخیره‌سازی اطلاعات در این نظام‌ها، دیسک‌های مغناطیسی بود (بلاچ، ۱۹۶۳) که امکان انعطاف بیش‌تری را نسبت به نوارهای مغناطیسی داشت. از اقدامات مهم حاصل از این پیشرفت‌ها، ایجاد پایگاه‌های اطلاعاتی درون خطی تجاری مانند مدلاین[4]، ای‌آی‌ام-تی‌دبلیوایکس از کتابخانه پزشکی آمریکا[5]، دایالوگ لاکهید[6] لاکهید[6] و اوربیت[7] است. بدین‌ترتیب نظام‌های نمایه‌سازی و چکیده‌نویسی رشد چشمگیری یافتند (بری، ۲۰۰۹). یکی از مشکلات چکیده‌نویسی و نمایه‌سازی، تأخیر زمانی در انجام کارها به‌صورت دستی بود. فن‌آوری حروف‌چینی رایانه‌ای، بسیاری از کارها را سریع‌تر کرد (لسک، ۱۹۹۶).

علاوه بر گسترش محتواهای نمایه‌سازی شده رایانه‌ای در این سال‌ها، ایده‌های جدیدی نیز برای نمایه‌سازی محتوای مدارک ارائه شد. از این نوع ایده‌ها می‌توان به ایده نمایه‌سازی مارون، کان و ری[8] اشاره کرد. آن‌ها از نظریه احتمال برای این کار

1 Sterling
2 Offline
3 batch-processing system
4 MEDLINE
5 NLM's AIM-TWX
6 Lockheed's Dialog
7 ORBIT
8 Maron, Kuhns and Ray

استفاده کردند. ون‌ریجسبرگن، اصول رتبه‌بندی احتمالی را تعریف کرد که نشان می‌داد چگونه با توجه به معیارهای ارزیابی تعریف شده می‌توان به‌صورت بهینه مدارک را رتبه‌بندی کرد (ساندرسون و کرافت، ۲۰۱۲). بعدها او و اسپارک جونز، مدل را توسعه دادند. در مدل احتمالی پایه، فرض بر این بود که کلمات در یک مدرک ارتباطی به هم ندارند. آن‌ها نشان دادند که این مسأله درست نیست و کلمات مدرک به هم مربوط‌اند. درنظرگرفتن کلمات به صورت وابسته به یکدیگر در یک مدرک، موضوع پژوهش‌های سال‌های بعد شد (ساندرسون و کرافت، ۲۰۱۲).

تنها ویژگی این دوره افزایش خدمات نمایه‌سازی و چکیده‌نویسی نبود؛ بلکه توانست اولین نظام‌های بازیابی تمام‌متن را نیز تجربه کند. یکی از پیشگامان این نظام‌ها، نظام لکسیس[1] است. در دوم آوریل ۱۹۷۳، امدی‌سی[2] تنها با چهار شرکت، خدمت مرجع حقوقی لکسیس را راه‌اندازی کرد. در این نظام، یک وکیل می‌توانست متن کامل کدهای نیویورک، اُهایو و کدهای فدرال هم‌چنین منابع یک کتابخانه مالیاتی فدرال را جستجو کند (تاریخچه گروه لکسیس- نکسیس[3]، ۲۰۰۰). لکسیس- نکسیس در سال ۲۰۱۰ بیش از ۵ میلیون مدرک قابل جستجو و ۴۰ هزار منبع تجاری، حقوقی و خبری را در اختیار داشت و نرخ اطمینان دسترس‌پذیری در آن ٪ ۹۹/۹۹ بود (بیزا-ییتس و ریبریونتو، ۲۰۱۰).

از دیگر فن‌آوری‌های کلیدی این دوره، دسترس‌پذیر شدن نظام‌های اشتراک زمانی بود. در این نظام‌ها، به‌جای این‌که پرسش‌ها با برخی از عملیات دسته‌ای پردازش شوند به‌طور مستقیم به یک پایانه داده می‌شوند و فوراً پاسخی برای آن دریافت می‌گردد. این فن‌آوری جستجو را عملی‌تر کرد و امکان ارائه خدمات آن به کتابداران به‌وجود آمد. هم‌چنین باعث شد قیمت هارد دیسک‌ها کاهش یابد (لسک، ۱۹۹۶). این موضوع، خود به افزایش انتشارات و پیشرفت‌های حوزه بازیابی اطلاعات کمک شایانی کرد و هارددیسک‌ها را به یک رسانه عمومی ذخیره‌سازی اطلاعات تبدیل نمود.

در سال ۱۹۷۱ یاردین و ون‌ریجسبرگن[4] در مورد خوشه‌بندی سلسله‌مراتبی[5] در بازیابی اطلاعات صحبت کردند. آن‌ها راهبردهای بازیابی اطلاعات را بر پایه

1 Lexis
2 MDC: Mead Data Central
3 History of LEXIS-NEXIS Group
4 Jardine and Van Rijsbergen
5 hierarchic clustering



خوشه‌بندی سلسله‌مراتبی عرضه کردند و با استفاده از مجموعه کرانفیلد آن را ارزیابی نمودند. آن‌ها معتقد بودند که این نوع راهبرد می‌تواند در مجموعه‌های بزرگ در حال رشد اثربخش باشد. در این دوران، سالتون تأثیرگذارترین مقاله‌های خود را در حوزه بازیابی اطلاعات منتشر کرد. انتشارات وی در ارتباط با نظریه نمایه‌سازی، مدل فضابرداری در بازیابی اطلاعات و اهمیت اصطلاح برای نمایه‌سازی متنی خودکار بود (دوبین و اسمیت، ۱۹۹۹). او با ارائه رهیافتی در مدل فضابرداری خود، فرآیند بازیابی اطلاعات را در بسیاری از نظام‌های بازیابی پژوهشی پایه‌ریزی کرد. شاید امروزه، فرمول‌های رتبه‌بندی سالتون کم‌تر استفاده شوند؛ اما کاربرد مدارک و پرسش‌ها به عنوان بردارهایی در فضای بزرگ و دارای بعد، هنوز هم معمول است (ساندرسون و کرافت، ۲۰۱۲).

یکی از پیشرفت‌های کلیدی حوزه بازیابی اطلاعات در این دوره، وزن‌های بسامد اصطلاح[1] لون (بر پایه رخداد یک واژه در میان یک مدرک) بود که با کار اسپارک جونز[2] در مورد رخداد کلمه در کل مدارک یک مجموعه تکمیل شد. مقاله او در مورد بسامد مقلوب مدرک به این معنا بود که بسامد رخداد یک کلمه در یک مدرک با میزان اهمیت آن واژه در بازیابی، ارتباط معکوسی دارد. یعنی هر چه یک اصطلاح در مدرک کم‌تر تکرار شود، دارای مفهوم خاص‌تری است (ساندرسون و کرافت، ۲۰۱۲).

این دوره شاهد توسعه زیاد بخش فهرست‌های رایانه‌ای است. اُسی‌ال‌سی به رکوردهای کتابخانه کنگره آمریکا نیز توسعه یافت (بری، ۲۰۰۹). البته باید گفت که بسیاری از این پژوهش‌ها در دهه قبل آغاز شده بود اما به ثمرنشستن و توسعه آن‌ها در دهه ۷۰ میلادی بوده است. به‌طورمثال پروژه اُسی‌ال‌سی از سال ۱۹۶۷ کلید خورد اما توسعه آن به فهرست‌های کتابخانه کنگره در دهه ۷۰ بوده است. این نظام که کیلگور[3] به‌وجود آورد از خروجی‌های مارک کتابخانه کنگره برای ایجاد فهرست‌نویسی ماشین‌خوان استفاده کرد. کار دیگر این نظام، ایده کار تعاونی بود و اولین نمونه از سازماندهی هوشمند تعاونی اطلاعات محسوب می‌شد که بوش در مقاله پیشگامانه خود در سال ۱۹۴۵ پیش‌بینی کرد (لسک، ۱۹۹۶)

تئودور نلسون[4] مفهوم فرامتن[1] را در این دهه مطرح کرد (دوبین و اسمیت،

1 term frequency (tf) weights
2 Spärck Jones
3 Kilgour
4 Theodor Nelson

۱۹۹۹) و برای متنی به‌کار برد که به‌آسانی و به شیوه غیرترتیبی[2] قابل‌دسترس است (گال و هانافین[3]، ۱۹۹۴). او پیش‌بینی کرد که چگونه رایانه‌ها می‌توانند فرآیندهای نشر و توزیع را دگرگون کنند و معتقد بود که رایانه به سادگی نمی‌تواند بهره‌وری را افزایش دهد و باید چگونگی کار افراد با مدارک بازتعریف شود.

در اواخر این دهه، اولین کنفرانس بازیابی اطلاعات ای‌سی‌ام[4] برگزار شد که نشان‌دهنده اهمیت موضوع برای پژوهشگران بود. کنفرانس هرساله به‌صورت تخصصی برگزار می‌شود و آخرین یافته‌ها و پیشرفت‌های حوزه بازیابی اطلاعات در آن عرضه می‌شود (دوبین و اسمیت، ۱۹۹۹). در سال ۲۰۱۵ سی‌وهشتمین دوره آن در شیلی برگزار شد و یکی از معتبرترین کنفرانس‌های حوزه بازیابی اطلاعات است.

## دهه ۱۹۸۰

دهه را شاید بتوان نقطه عطف جدید برای بازیابی اطلاعات دانست که ایده ماشین ممکس وانور بوش (۱۹۴۵) محقق شد.

افزایش مداوم سرعت پردازش کلمات و کاهش پی‌درپی قیمت فضای دیسک موجب شد که اطلاعات بیش از پیش به شکل ماشین‌خوان قابل دسترس شود (لسک، ۱۹۹۶). با توجه به افزایش پردازش ماشینی و دیگر فن‌آوری‌ها، در این دهه بازیابی پیوسته متن کامل رشد قابل توجهی داشت و بسیاری از مجلات و روزنامه‌ها به‌صورت پیوسته -البته فقط به‌صورت متن- قابل دسترسی بودند.

اولین کنفرانس بازیابی اطلاعات ای‌سی‌ام مشترک با بخش بازیابی اطلاعات در انجمن رایانه بریتانیا در سال ۱۹۸۰ برگزار شد (دوبین و اسمیت، ۱۹۹۹). ریاست این کنفرانس را محقق بزرگ حوزه بازیابی اطلاعات، ون‌ریجسبرگن بر عهده داشت و حوزه‌های جدید اطلاعات در آن دوران بحث شدند. بررسی مقالات ارائه شده در ۱۹۸۰ نشان می‌دهد که عمده پژوهش‌ها، در حوزه وزن‌دهی اصطلاحات، مدل‌های احتمال و بازیابی مفهومی بوده است (خلاصه مقالات سومین کنفرانس سالانه ای‌سی‌ام روی پژوهش و توسعه بازیابی اطلاعات[5]، ۱۹۸۰). در سال ۱۹۸۱، حتی مقاله‌ای در حوزه جستجوی فازی و مقاله دیگری در حوزه بازنمای فضایی دانش ارائه شده است

1 Hypertext
2 nonsequential
3 Gall and Hannafin
4 ACM SIGIR conference
5 Proceedings of the 3rd annual ACM conference on Research and development in information retrieval



(کراوچ، کوپر و هر[1]، ۱۹۸۰).

هم‌چنین بلکین، اُدی و بروک[2] در سال ۱۹۸۲، دیدگاه اسک[3] را برای بازیابی اطلاعات پیشنهاد کردند. این مفهوم هر چند که سرانجام به شکست انجامید، اما مفهوم بسیار مهمی بود (دوبین و اسمیت، ۱۹۹۹). بلکین و همکارانش در مقاله خود، نظام بازیابی اطلاعات تعاملی را معرفی کردند که به نیازهای اطلاعاتی توجه داشت (بلکین، اُدی و بروک، ۱۹۸۲).

در سال ۱۹۸۳، سالتون کتاب مهم خود " درآمدی بر بازیابی اطلاعات" را با کمک انتشارات مک‌گروهیل[4] منتشر کرد. این کتاب بر مدل‌های بُرداری تأکید فراوانی داشت (دوبین و اسمیت، ۱۹۹۹) و در حوزه بازیابی اطلاعات، اثر مهمی محسوب می‌شد و تا سال ۱۹۸۷ چندین ویرایش از آن منتشر گردید.

اواسط دهه ۱۹۸۰، تلاش‌ها برای ایجاد نسخه‌های کاربر نهایی[5] از نظام‌های بازیابی اطلاعات تجاری آغاز شد (دوبین و اسمیت، ۱۹۹۹). استفاده از نظام‌های برخط بازیابی اطلاعات در این دوره از دو طریق امکان‌پذیر بود. اول استفاده از نظام‌های متن کامل و دیگری افزایش استفاده غیرمتخصصان از نظام‌های بازیابی. به‌طورمثال کتابداران، کارت‌های فهرستنویسی را با فهرست‌های عمومی دسترسی پیوسته (اپک‌ها)[6] جایگزین کردند. در پایان این دهه نیز فروشندگان تجاری، نرم‌افزارهای اپک مانند نوتیس[7] را وارد بازار کردند (لسک،۱۹۹۶) بدین ترتیب در این دوره استفاده از اپک‌ها رایج شد و کتابخانه‌ها توانستند اقتصادی‌تر و بهینه‌تر به فهرست‌های پیوسته دست یابند و نیازهای سازماندهی خود را نیز برطرف کنند. این فهرست‌ها نقاط جستجوی بیش‌تر در اختیار کاربران گذاشت. امروزه، اپک‌ها مهم‌ترین بخش کتابداری و اطلاع‌رسانی‌اند و در طول زمان بسیار پیشرفت کردند. نسل اول اپک‌ها حاوی اطلاعات اندکِ شناخته‌شده برای جستجو بود مثل عنوان، نویسنده و شماره کنترل. نسل دوم اپک‌ها با پیشرفت فن‌آوری، امکان جستجو با عملگرهای بولی، سرعنوان‌های موضوعی و کلیدواژه را نیز مهیا کرد و در نسل بعد بر معماری نظام‌های باز[8] تمرکز کرد که در سال‌ها و دهه‌های بعد، با استفاده از رابط

1 Crouch, Cooper and Herr
2 Belkin, Oddy and Brooks
3 ASK: Anomalous State of Knowledge
4 McGraw-Hill
5 End user version
6 Online Public Access Catalogs(OPACS)
7 NOTIS
8 open systems architectures

رابط کاربری گرافیکی و حمایت استاندارد فراداده هسته دوبلین، فرامتن، برنامه‌نویسی جاوا[1] و مجموعه‌های نتایج رتبه‌بندی شده پیشرفت قابل ملاحظه‌ای یافت (بیزایاتار و ریبیرو نتو، ۲۰۱۰).

در این دهه، استفاده از سی‌دی‌رام‌ها افزایش قابل ملاحظه‌ای یافت و توزیع اطلاعات را بیش از پیش آسان کرد. در پایان دهه، هر کتابخانه حداقل یک درایو سی‌دی‌رام داشت که با نشر سنتی سازگاری مناسبی یافتند و برای توزیع و استفاده اطلاعات به آسانی استفاده شدند. هم‌زمان با افزایش استفاده از این محمل جدید، شبکه‌سازی رایانه‌ای نیز در این دهه گسترش یافت (لسک، ۱۹۹۶)

تأثیرگذارترین اتفاق در اواخر این دهه آن بود که تیم برنرزلی، وب جهان‌گستر را در آزمایشگاه کرن[2] پیشنهاد کرد (دوبین و اسمیت، ۱۹۹۹) در سال ۱۹۸۹، این ایده؛ ذخیره، دسترسی و جستجوی مجموعه‌های مدارک را متحول کرد. مجموعه‌های مدارک مرتبط[3] در این زمان، ایده بسیار جدیدی بود. هر چند که چنین طرحی را وانور بوش (۱۹۴۵) تصویر کرده بود اما تحقق بخشی از طرح او در این سال‌ها امکان‌پذیر شد. ایجاد وب جهان‌گستر به معنای مرگ عصر صنعتی و تولد عصر اطلاعات بود (لانگویل و میر، ۲۰۰۶).

اینترنت اولیه صفحاتی متنی بود که تنها یک اندازه و فونت داشت. کاستی‌های اینترنت اولیه در مورد رسانه‌های دیداری و شنیداری و ناتوانی در انتقال و دسترسی به این نوع اطلاعات سبب شد تا پروژه‌های مختلفی در این حوزه شروع به‌کار کند. انگلبرت[4] با اختراع ماوس، توانست ایده فرامتن را که تد نلسون ارائه کرده بود به واقعیت تبدیل کند و مقدمات ایجاد وب فراهم شد. پیشرفت دیگر در این حوزه، ایجاد یوآرال[5] بود. این فن‌آوری به سایت و صفحه یک نشانی منحصربه‌فرد می‌داد. زبان نشانه‌گذاری فرامتن یا اچ‌تی‌ام‌ال[6] حلقه اتفاقات را کامل کرد تا تیم برنرزلی با کنار هم گذاشتن همه این‌ها بتواند نسخه اولیه وب را ایجاد کند. اولین نسخه در دسامبر ۱۹۹۰ راه‌اندازی شد و تحول عمده‌ای را در حوزه بازیابی اطلاعات به‌وجود آورد (تاریخچه وب جهان‌گستر[7]، ۲۰۰۴).

1 Java programming
2 CERN
3 Linked
4 Engelbart
5 Uniform Resource Locator(URL)
6 Hypertext Markup Language (HTML)
7 History of the world wide web



نظام‌های بازیابی اطلاعات با آزمایش و تجربه بر مجموعه‌های آزمون، پالایش و ارزیابی می‌شوند (وورهیز و هارمن[1]، ۱۹۹۹). اما در این دهه و اوایل دهه ۱۹۹۰، اندازه مجموعه‌های مدارکی که برای آزمون و یا ارزیابی‌ها استفاده می‌شد نگران کننده بود. جامعه علمی معتقد بود که اندازه‌های مجموعه‌های آزمون مورد استفاده در مقایسه با مجموعه‌هایی که برخی کمپانی‌های بزرگ استفاده می‌کنند کوچک است. به همین دلیل دونا هارمن[2] با همکاران خود، تِرک یا کنفرانس بازیابی متن[3] را به‌وجود به‌وجود آوردند. این کنفرانس، در واقع تمرین سالانه برای تعداد زیادی از گروه‌های پژوهشی بود تا با همکاری یکدیگر، مجموعه‌های آزمون بزرگ‌تری را نسبت به دوره قبل طراحی کنند (ساندرسون و کرافت، ۲۰۱۲) و بدین ترتیب توانستند مجموعه‌های بزرگ را بر اساس مجموعه‌های آزمون متناسب پالایش و ارزیابی نمایند. این کنفرانس هر سال برگزار و نتایج آن در دیگر کنفرانس‌های مهم بازیابی اطلاعات مانند ای‌سی‌ام استفاده می‌شود. در واقع، کار این کنفرانس تهیه مجموعه‌های آزمون استاندارد برای ارزیابی نظام‌های بازیابی اطلاعات است و در بسیاری از پژوهش‌های مربوط به ارزیابی این نظام‌ها استفاده می‌شود.

به‌طور کلی در این دهه، استفاده از اطلاعات پیوسته، معمول شد. هر چند که عموم مردم از آنها به‌صورت دائم استفاده نمی‌کردند (لسک، ۱۹۹۶). این دهه، شروع انقلابی تازه برای مردم دنیا در انتهای هزاره دوم بود. همه چیز به سرعت به طرف ماشینی‌شدن پیش رفت. اطلاعات با سرعت و سهولت بیش‌تر حداقل در اختیار متخصصان قرار می‌گرفت. کم‌کم زمینه ورود مردم عادی به عصر اطلاعات بیش از پیش فراهم شد و گام نهادن در مسیر توسعه علمی و آموزشی سریع‌تر گردید.

## دهه ۱۹۹۰

آغاز یک دوره جدید در عصر اطلاعات بود. نسل اول وب به‌طور رسمی با پیشنهاد تیم برنرزلی در سال ۱۹۸۹ آغاز به‌کار کرد. این پیشنهاد بیست سال بعد از خلق و راه‌اندازی اولین ارتباط[4] به شکلی بود که امروزه اینترنت شناخته می‌شود. پیشنهاد او شامل فن‌آوری‌هایی بود که اینترنت را برای مردم دسترس‌پذیر و قابل استفاده می‌کرد. قبلاً نیز اشاره شد که سه فن‌آوری اصلی، بنیان وب فعلی یا

1 Voorhees and Harman
2 Dona Harman
3 TREC: Text REtrieval Conference
4 Connection

همان وب نسل اول شدند.

- زبان نشانه‌گذاری فرامتن یا اچ‌تی‌ام‌ال: این زبان فرمت نشر برای وب است و پیوند دادن مدارک در وب را امکان‌پذیر می‌کند.
- شناساگر اختصاصی منبع یا یوآرآی[1]: نوعی نشانی که برای هر منبع در وب اختصاصی است.
- پروتکل انتقال فرامتن (اچ‌تی‌تی‌پی): با آن می‌توان صفحاتی را که به هم پیوند خورده‌اند بازیابی کرد (تاریخچه وب[2]، ۲۰۰۸).

با وجود پیدایش وب جهان‌گستر در سال ۱۹۹۰، تعداد وب‌سایت‌ها تا سال ۱۹۹۳ بسیار اندک بودند (ساندرسون و کرافت، ۲۰۱۲). در سال ۱۹۹۳ کِرن به‌طور رسمی اعلام کرد که فن‌آوری وب، رایگان و برای همه افراد در دسترس است. تا سال قبل از این اعلام رسمی، فهرست‌نویسی دستی معمولی برای منابع اطلاعاتی کافی بود. شش ماه بعد، این تعداد صفحات، چهار برابر و شش ماه بعد مجدداً چهار برابر شد. در همان سال، به‌تدریج موتورهای جستجوی بیش‌تری به‌وجود آمدند تا پاسخگوی افزایش وب‌سایت‌ها و اطلاعات باشند (ساندرسون و کرافت، ۲۰۱۲). بدین‌ترتیب فهرست‌نویسی دستی با چالش جدیدی روبه‌رو شد.

اولین موتور جستجو به نام آرچی[3] در سال ۱۹۹۱ ایجاد شد. نظام نمایه‌سازی در موتور جستجوی آرچی فضای محدودی داشت و فقط می‌توانست فهرست‌های قابل‌دانلود را جستجو کند و قابلیت جستجوی محتوا را نداشت. تا پایان این دهه، بیست موتور جستجوی اطلاعات به‌وجود آمد که از مهم‌ترین آن‌ها می‌توان به گوگل[4] و یاهو[5] اشاره کرد (تاریخچه موتورهای جستجو[6]، ۲۰۱۳) که نشان‌دهنده افزایش چشمگیر اطلاعات بود. افزایش تصاعدی اطلاعات در این سال و سال‌های بعد نشان داد که باید برای سازماندهی و البته بازیابی اطلاعات برای کاربران بالقوه و بالفعل وب تدابیر جدیدی اندیشید. افزایش اطلاعات، خود چالش‌ها و مسائل جدیدی به‌همراه داشت. چالش‌های جدید، جوامع پژوهشی و تجاری را در حوزه بازیابی اطلاعات به تعامل واداشت و به آن‌ها فهماند که باید تلاش‌های خود را در حوزه بازیابی اطلاعات بیش‌تر کنند (ساندرسون و کرافت، ۲۰۱۲).

با توجه به مطالب پیش‌گفته، مهم‌ترین اتفاق در این دهه جستجوی تحت وب

1 URI: Uniform Resource Identifier
2 History Of The Web
3 Archie
4 Google: www.google.com
5 Yahoo: www.yahoo.com
6 History of Search Engines



است. موتورهای جستجوی اولیه از جستجوی فهرست‌ها به موتورهای جستجویی تبدیل شدند که می‌توانستند با امکانات پیشرفته‌تر، متن را نیز جستجو کنند. با وجود چندین برابر شدن وب‌سایت‌ها، باز هم مشکلات فراوانی وجود داشت. در این حوزه وجود یک مرورگر گرافیکی در نظام بازیابی بسیار مهم بود. اما تا سال ۱۹۹۲ هنوز مرور گرافیکی وجود نداشت. چهار دانشجوی یک کالج در فنلاند، مرورگر اروایز[1] را در سال ۱۹۹۱ راه‌اندازی و در ۱۹۹۲ تکمیل کردند. این اولین مرورگر نقطه و کلیکِ وب[2] در دنیا بود. این مرورگر می‌توانست چندین صفحه را بارگذاری کند و سپس با کلیک بر روی یک فرامتن در یک پنجره جدید، صفحه کلیک شده باز شود (هالوردا[3]، ۲۰۰۹). مرورگر مهم دیگر که مادر همه مرورگرهای دیگر شناخته شد موزاییک[4] بود که امکان ذخیره، بازیابی و ارائه اطلاعات را با کاهش زمان بازیابی فراهم می‌کرد و از قابلیت جستجوی مدارک فرامتن با گرافیک بالا برخوردار بود (جادسن[5]، ۱۹۹۶).

در سال ۱۹۹۴ تیم برنرزلی - مبدع وب- و جفری جف[6] برای توسعه استانداردهای وب، کنسرسیوم وب جهان‌گستر را به‌وجود آوردند که هدفش افزایش پتانسیل وب برای ارائه خدمات بیشتر بود. این کار با توسعه استانداردهایی انجام شد که باعث رشد پایدار شدند. هدف آن‌ها "وب برای همه" است و این کار را با دسترس‌پذیر ساختن، بین‌المللی کردن[7] و ایجاد وب برای توسعه اجتماعی[8] انجام می‌دهند (ماموریت W3C، ۲۰۱۲).

در این دهه، مسأله بعدی در بازیابی اطلاعات استفاده از داده‌های استنادی و تحلیل پیوندها[9] بود که مراحل آغازین خود را طی می‌کرد. توسعه دهندگان وب، تشخیص دادند که می‌توانند از پیوند میان صفحات برای ساختن خزنده‌ها و روبوت‌ها استفاده کنند تا صفحات بیش‌تری جمع کنند و بدین‌ترتیب مجموعه‌سازی را به‌صورت خودکار درآورند (ساندرسون وکرافت، ۲۰۱۲). تحلیل پیوندها، فنی است که از اطلاعات ضمیمه‌شده متصل در ساختار فرامتن وب استفاده می‌کند تا کیفیت نتایج

1 Erwise
2 point-and-click web
3 Holwerda
4 Mosaic
5 Judson
6 Jeffry Jeff
7 Internationalization
8 Mobile Web for Social Development
9 Link analysis

جستجو را بهبود ببخشد (لانگویل و مایر، ۲۰۰۶). در این دوره، موتورهای جستجوی وب، بیشتر از ویژگی‌هایی که قبلاً در نظام‌های بازیابی تجربی استفاده شده بود بهره بردند (دوبین و اسمیت، ۱۹۹۹).

یکی از پیشرفت‌های دهه ۱۹۹۰، ایجاد امکان پویش[1] بود. بسیاری از ناشران در پی آن بودند تا تصاویر صفحات را پویش کنند و بفروشند که یا اختراع دیسک‌های ارزان و سی‌دی‌رام‌های ارزان‌تر امکان‌پذیر شد. بدین ترتیب، دو نوع فرمت برای بسیاری از منابع به‌وجود آمد. اول به‌صورت پیوسته و روی وب، و بعد با استفاده از سی‌دی‌رام. بسیاری از این خدمات نیز در این دهه رایگان بود.

## هزاره سوم: سال ۲۰۰۰ تا کنون...

عصر جدید با پیدایش یک پدیده جدید آغاز شد و آنهم وب ۲.۰ بود. خصوصیت وب در مقایسه با مجموعه‌های سنتی، چه در نسل اول وب و چه نسل دوم پویا بودن آن است در مقابل نظام‌های بازیابی سنتی که ایستا و در بسیاری موارد غیرقابل تغییر بودند. اگر هم امکان تغییر وجود داشت، هزینه‌بر بود. ویژگی دیگر وب آن است که خود سازمان‌دهنده محسوب می‌شود؛ یعنی افراد به راحتی و بدون استفاده از استانداردهای پیچیده، می‌توانند محتوا تولید کنند، در حالی که در مجموعه‌های سنتی، تولید محتوا و سازماندهی آن در اختیار متخصصان بود و هزینه داشت. ویژگی دیگر وب، فرامتن بودن و امکان پیوند میان صفحات برای محتواهای تولید شده است (لانگویل و مایر، ۲۰۰۶). این پیشرفت جدید در بدو ظهور، نه تنها صنعت شبکه‌ای را روزآمد کرد؛ بلکه بر روش‌های بازیابی سنتی اطلاعات شبکه‌ای بسیار تأثیر گذاشت و نیازهای جدیدی را مطرح کرد (ژانگ و تانگ، ۲۰۰۷). یافته‌های پژوهشگران در حوزه بازیابی اطلاعات در سه دوره زمانی نشان می‌دهد که حوزه بازیابی در سال‌های۸۴-۱۹۸۰ درگیر پژوهش‌های آموزش، ذخیره و نظام‌ها بوده و در سال‌های ۹۴-۱۹۹۰ در حوزه خدمات اطلاعاتی، رابط‌های کاربری و پایگاه‌های اطلاعاتی و در دوره زمانی ۲۰۰۰ به بعد پژوهش‌های مربوط به وب رواج بیش‌تری یافته است (سوگیموتو و مک‌گین[2]، ۲۰۱۰).

به هر حال با ظهور نسل دو وب، بازیابی اطلاعات تغییرات زیادی کرد. بازیابی اطلاعات سنتی به روش‌هایی وابسته بود که در وب ۲.۰ کارآیی چندانی نداشتند. این

1 Scaninig
2 Sugimoto and McCain



روش‌ها مانند روش‌های بازیابی قبل یا مبتنی بر کلیدواژه‌هایی بودند که به موتورهای جستجو ارائه می‌شدند، یا بر اساس فهرست‌های موضوعی که در نظام استفاده می‌شد یا بر پایه فراداده‌ها، جستجو را انجام می‌دادند که از آن‌جمله می‌توان به شاهراه موضوعی[1] مثل ای‌اچ‌دی‌اس[2] اشاره کرد (ژانگ و تانگ، ۲۰۰۷). اما وب ۲.۰ نیازهای جدیدی را به‌وجودآورد:

۱- تعمیم‌پذیری و گستردگی محصولات اطلاعاتی: در وب ۱.۰ تولید اطلاعات شبکه‌ای به عهده تعداد اندکی از کمپانی‌های حرفه‌ای و ویرایشگران وب‌سایت‌ها بود اما بسیاری از کاربران، محتوای اطلاعاتی زیادی را در شبکه محیط وب ۲.۰ عرضه می‌کردند (ژانگ و تانگ، ۲۰۰۷). در وب ۲.۰ روزنامه‌نگاران، سازندگان وب‌سایت‌ها یا کمپانی‌ها نیستند که اطلاعات را تولید می‌کنند بلکه هر کاربر از طریق رسانه‌های متعدد و خدمات متنوع می‌تواند محتوا تولید کند (پیترز[3]، ۲۰۰۹).

۲- محتواهای کوچک[4] و معنابخشی‌های ساختار اطلاعات[5]: محتواهای کوچک، داده‌های مختلفی‌اند که کابران خلق می‌کنند؛ مانند تصویر، فهرست‌های موسیقی، جاهایی که آرزو دارند بروند، دوستان جدید و مانند آن. در وب ۲.۰ می‌توان حجم زیادی از این محتواهای کوچک را تولید و روزانه استفاده کرد (ژانگ و تانگ، ۲۰۰۷). باید گفت وب ۲.۰ یک وب اجتماعی است که شامل شبکه‌سازی اجتماعی (فیسبوک[6])، برچسب‌گذاری اجتماعی (فلیکر)[7] ، نشانه‌گذاری اجتماعی (دلیشس)[8] ، رده‌بندی‌های مردمی[9]، وبلاگ‌نویسی[10] و ویکی‌ها (ویکی‌پدیا) است (بری، ۲۰۰۹ب). ۲۰۰۹ب). برای استفاده از این محتواهای کوچک، نشانه‌های معنایی[11] لازم است که اچ‌تی‌ام‌ال موجود در وب ۱.۰ این قابلیت را نداشت (ژانگ و تانگ، ۲۰۰۷). وب ۲.۰ تولید این محتواها را با توجه به قابلیت‌های خود امکان‌پذیر کرد. به‌طور کلی یکی از مزیت‌های وب ۲.۰ شبکه نشانه‌گذاری کاربر[12] است که نه تنها می‌تواند برای توصیف دقیق‌تر هر منبع استفاده شود؛ بلکه می‌تواند برای کشف شباهت مشخصه-

1 subject gateway
2 AHDS
3 Peters
4 Microcontent
5 Semantics of information structure
6 Social networking (Facebook)
7 Social tagging (flickr)
8 social bookmarking (delicious)
9 Folksonomies
10 Bloggong
11 Semantic Tags
12 Network of User Tagging

کاربر که برای به‌دست آوردن منابع شخصی دقیق‌تر مفید است به‌کار رود (یانگ و وانگ[1]، ۲۰۰۹).

محتوای اجتماعی وب ۲.۰ بازیابی اطلاعات را واقعاً متحول کرد، زیرا چنان‌که قبلاً اشاره شد، نه فقط متخصصان که افراد عادی توانستند محتوا تولید کنند و این امر چالش‌هایی به‌همراه داشت و آن افزایش نیاز به خلق روش‌ها و امکانات جدید برای ذخیره و بازیابی اطلاعات بود. اشتراک‌گذاری اطلاعات افزایش یافت. همه رسانه‌های اجتماعی مانند فیسبوک و فلیکر، قابلیت‌های فراوانی برای اشتراک اطلاعات و دانش در میان افراد مختلف دارند. ابزارهای اجتماعی توانسته‌اند بسیاری از مشکلات به‌وجود آمده را حل کنند.

یکی از این ابزارهای نسل جدید، رده‌بندی‌های مردمی هستند. رده‌بندی مردمی، ابزار تولید، عرضه، بازیابی و گسترش اطلاعات در وب۲.۰ است که کمپانی‌ها و متخصصان رایانه آن را به‌وجود آوردند تا مردم، محتواهای خود را تولید و ذخیره و با نشانه‌های سفارشی نمایه‌سازی کنند. این خدمت اطلاعاتی تعاونی، شامل اشتراک عکس‌ها، ویدئوها و نشانه‌گذاری‌های اجتماعی است که می‌تواند در بازیابی اطلاعات به افراد کمک کند (پیترز، ۲۰۰۹).

موضوع دیگر در وب۲.۰، ربط بود. هر چند که ربط، چالش همه نظام‌های بازیابی بوده اما به‌طورخاص در بازیابی تحت وب موضوع مهم‌تری محسوب می‌شود. به‌طور مثال، دقت[2]، موضوع مهمی در جستجوی تحت وب است. اگر چه میزان اطلاعات در وب در حال رشد است اما توانایی کاربران برای جستجوی اطلاعات، افزایش نیافته است. کاربران به ندرت بیش از ۱۰ یا ۲۰ یافته اولیه مدارک را مرور می‌کنند و با بی‌حوصلگی، انتظار دارند که دقت جستجو بالا رود. محدودیت دیگر نظام‌های بازیابی در وب، معیارهای اندازه‌گیری است. معیارهای نظام‌های بازیابی سنتی، بر اساس مجموعه‌های کنترل شده بود اما این معیارها برای نظام‌های وبی واقعی به‌نظر نمی‌رسند (لانگویل و مایر، ۲۰۰۶). بدین‌ترتیب رتبه‌بندی ربط در وب نسبت به نظام‌های بازیابی سنتی تغییر کرده است.

رتبه‌بندی ربط در وب ۱.۰ بر توزیع اصطلاح[3] متکی بود که بر اساس بسامد اصطلاح و بسامد مقلوب مدرک و مدل‌های زبانی انجام می‌شد. اما در گذار از

1 Yang and Wang
2 precision
3 Term distribution



وب۱.۰ به وب۲.۰ مهم‌ترین اتفاق تاریخ بازیابی تحت وب، حرکت از تحلیل لینک به رتبه‌بندی صفحات[1] بود. بدین‌ترتیب رتبه‌بندی ربط در وب۲.۰ مجموع امتیاز بازیابی اطلاعات[2] و رتبه صفحات بود (ما، ۲۰۱۰).

با وجود تداوم چالش‌ها نسل بعدی وب از راه رسید. به‌طورکلی در وب۲.۰ کاربران، هم خواننده و هم نویسنده‌اند یعنی محتوا تولید و کنترل می‌کنند. محتوا به‌طورفزاینده‌ای در حال افزایش است و وب به یک محیط پویا تبدیل شده است. کاربران، ارزش‌هایی را با استفاده از برنامه‌های کاربردی به وب اضافه می‌کنند. مشارکت کاربر باعث افزایش هوش جمعی می‌شود. هم‌چنین واسطه‌های کاربرمدار و محتواهای شخصی تجربه کاربر را افزایش می‌دهد. وب به عنوان یک بنیاد برنامه‌نویسی[3]، اجرای برنامه‌های کاربردی را به‌طورکامل با استفاده از یک مرورگر امکان‌پذیر می‌کند. قابلیت انتقال در وب بسیار بالاست. نرم‌افزار در سطحی بالاتر از یک ابزار واحد است و باز بودن[4] نرم‌افزارها ویژگی بسیار مهمی است (گومز–پرز، فرنادز– لوپز و کورکوگارسیا[5]، ۲۰۰۴). همه این‌ها وب۲.۰ را به‌سوی نسل جدیدی از وب پیش برده و ظهور پیشرفت‌های جدیدی را نوید داده است که وب ۳.۰ نام دارد. این نسل وب به نسل وب معنایی مشهور است. نه با آن معنی که در وب ۲.۰ وب معنایی وجود ندارد اما مشخصه ویژه این نوع وب، تمرکز بر معناست (بری، ۲۰۰۹). تیم برنرزلی خالق وب در سال ۱۹۹۸، ایده وب معنایی را مطرح کرد (استایلسویگ[6]، ۲۰۰۶). بعد از سال ۲۰۰۰ این ایده عملی شد. وب ۳.۰ هوشمند است چون فن‌آوری‌های وب‌معنایی، پایگاه‌های جهانی و برنامه‌های کاربردی هوشمند دارد و به‌صورت بی‌سیم در همه‌جا اداره می‌شود. دسترسی به اینترنت سیار[7] و ابزارهای سیار[8] مشخصه وب معنایی است. مهم‌ترین ویژگی این نوع وب، فن‌آوریهای باز[9] است و شامل پروتکل‌ها و برنامه کاربردی باز، شکل‌های داده‌ای باز، باز، بنیادهای نرم‌افزاری منبع باز، و داده‌های باز هستند (گومز–پرز و دیگران، ۲۰۰۴).

1 Page ranks
2 IR Score
3 Programming platform
4 openness
5 Fernández-López, and Corcho-Garcia
6 Styltsvig
7 Mobile Internet access
8 Mobile devices
9 Tecnologías abiertas (Open Technology)

در وب ۳.۰ رتبه‌بندی ربط به سوی آگاهی حرکت کرده و معماری نظام‌های بازیابی اطلاعات از وب‌محور به داده‌محور پیش رفته است. اکتشاف دانش و داده‌کاوی وب از صفحات وب به اشیای وب (موجودیت‌ها)[1]، هم‌چنین از ساختارنیافته ساختارنیافته به ساختاریافته تغییر کرده و ربط به طرف آگاهی و خرد پیش رفته است (ما، ۲۰۱۰). وب ۳.۰ با ویژگی وب‌معنایی بیش‌تر از همه با هستی‌شناسی‌ها[2] بر نظام‌های بازیابی اطلاعات تأثیر گذاشته است (دیویس، هارملن و فنسل[3]، ۲۰۰۲). هستی‌شناسی تعریف دانش در سطح کاربردی است (استایلسویگ، ۲۰۰۶) وب‌معنایی با کاربرد برخی فن‌آوری‌ها به‌ویژه در اطلاعات الکترونیکی کیفیت مدیریت دانش را در سازمان‌های بزرگ بهبود می‌بخشد (دیویس، هارملن و فنسل، ۲۰۰۲). این نوع وب، در حقیقت شکلی از وب فعلی است که اطلاعات در آن به خوبی تعریف شده‌اند و به رایانه‌ها و افراد کمک می‌کند تا همکاری بهتری با هم داشته باشند. اساس آن، داده‌های پیوند خورده است که به گونه مناسبی تعریف شده‌اند تا برای کشف، خودکارسازی و استفاده مجدد در سراسر برنامه‌های کاربردی مختلف موثرتر باشند (گومز–پرز و دیگران، ۲۰۰۴).

نظام‌های بازیابی اطلاعات، دوران پر فراز و نشیبی را پشت سر گذاشته‌اند اما هنوز با چالش‌های زیادی مواجه‌اند. هنوز مسأله ربط، تعامل و ایجاد نظام‌های بازیابی کاربرمدار از مشکلات عمده این نظام‌هاست. تاریخچه این نظام‌ها از دوران اولین فهرست‌ها تا زمان وب ۳.۰ به‌خوبی نشان می‌دهد که بازیابی اطلاعات به مشارکتی و همه‌گیر شدن اطلاعات منجر می‌شود و پیشرفت‌ها و فن‌آوری‌های عمده، نظام‌های بازیابی اطلاعات را مسلماً تعاملی‌تر کرده‌اند.

## خلاصه فصل

در تعاریف مختلف برای بازیابی اطلاعات و نظام‌های بازیابی جنبه‌هایی مانند ذخیره، بازیابی، سازماندهی، جستجو، ردیابی و تفسیر اطلاعات، هم‌چنین وجود یک نظام رایانه‌ای، نوع اطلاعات مانند متن، تصویر، صداو غیره وجود دارند و مسأله مهم در این تعاریف، تحلیل ربط و فنون تحلیل پرسش است. با استفاده از این جنبه‌ها در پایان بخش تعریفی کلی از بازیابی اطلاعات ارائه شد. در بخش دوم تاریخچه

1 Web objects(entities)
2 ontology
3Davis, Harmelen and Fensel



نظام‌های بازیابی بررسی گردید. در ابتدا مشخص شد که از همان ابتدا بازیابی اطلاعات همزمان با تولید اطلاعات بود و نخستین فهرست‌های کتابخانه‌ای، حاصل اولین تلاش انسان برای سازماندهی و قابل دسترس‌کردن اطلاعات بوده است. عصر نظام‌های اطلاعاتی مدرن در دهه ۱۹۴۰، با اختراع اولین رایانه الکترونیکی به‌وجود آمد. ماشین ممکس ایده ایجاد ماشین ذخیره وبازیابی اطلاعات را به واقعیت نزدیک کرد. تحولات بعد در رایانه، ایجاد شبکه‌های رایانه‌ای، فهرست‌های رایانه‌ای عمومی، تولد اینترنت و وب و اجتماعی شدن وب جهان‌گستر بر نظام‌های بازیابی اطلاعات تأثیر بسیاری داشته است.

# فصل دوم

## ساختار نظام‌های بازیابی اطلاعات با تأکید بر موتورهای جستجو

### مقدمه

بحث ساختار نظام‌های بازیابی اطلاعات، بسیار گسترده است. به همین دلیل برای جلوگیری از آشفتگی در ارائه مطالب، در این فصل ساختار و معماری یک موتور جستجو، بررسی شده است. موتورهای جستجو از مهم‌ترین ابزارهای بازیابی و در بسیاری از موارد، نقطه شروع بسیاری از فعالیت‌های جستجوی اطلاعات هستند. این ابزارها ساختار پیچیده‌ای دارند و بحث کامل در مورد آن در یک فصل از کتاب نمی‌گنجد. یکی از کتاب‌های مهم این حوزه را پژوهشگران یاهو، گوگل و دانشگاه ماساچوست در سال ۲۰۰۹ منتشر کرده‌اند که به‌طور کامل به موتورهای جستجو و تمامی ابعاد آن پرداخته‌است (کرافت، متزلر و استرامن[1]، ۲۰۰۸). این فصل کتاب تا حدودی بر اساس ساختار این کتاب نوشته شده است.

در این کتاب سعی شده بخش‌های مختلف یک موتور جستجو به زبان ساده معرفی شوند. قابل ذکر است که نویسندگان تا حد امکان از بحث‌های فنی و نرم‌افزاری که در حوزه کار مهندسین نرم‌افزار است پرهیز کرده‌اند.

---

1 Croft, Metzler and Strohman

## موتور جستجو چیست؟

موتورهای جستجو از اولین ابزارهای کاربردی در وب، برای جستجو و مکان‌یابی اطلاعات‌اند وکاربران وب، ساده‌ترین تا پیچیده‌ترین اطلاعات را از آنها به‌دست می‌آورند. امروزه یک موتور جستجو مانند گوگل به ابزار ضروری و انکارنشدنی اطلاع‌یابی تبدیل شده است. این ابزارها از بدو تاریخ پیدایش، همواره در حال ارتقا و پیشرفت بوده‌اند. به طور مثال اولین موتور جستجو به‌نام آرچی، امکانات بسیار محدودی داشت و تنها برای جستجوی فایل‌های اف‌تی‌پی مناسب بود. ورونیکا اولین موتور جستجویی بود که می‌توانست متن را جستجو کند (موتور جستجو چیست[1]، ۲۰۱۴). موتورهای جستجوی جدید قابلیت‌های زیادی در جستجوی انواع اطلاعات دارند.

این ابزارهای بازیابی اطلاعات نه فقط به صورت مستقل که در نظام‌های بازیابی اطلاعات نیز بخش بسیار مهمی محسوب می‌شوند مانند نظام‌های بازیابی اطلاعات پزشکی یا پایگاه‌های استنادی و هر پایگاه اطلاعاتی که در آن اطلاعات ذخیره شده و قرار است دسترسی به آن وجود داشته باشد. اگر یک نظام اطلاعاتی، موتور جستجو نداشته باشد فقط انبار ذخیره اطلاعات است و کاربرد زیادی ندارد و این خود اهمیت این نرم‌افزار کاربردی سطح بالا را که خود مجموعه‌ای از نرم‌افزارهای دیگر است به‌خوبی نشان می‌دهد.

موتورهای جستجو به زبان ساده، برنامه‌های نرم‌افزاری‌اند که از روبات‌ها یا خزنده‌ها تشکیل شده‌اند. خزنده‌ها، در وب حرکت می‌کنند، صفحات جدید را بازبینی می‌کنند و آن‌ها را برای نمایه‌شدن به درون خود می‌کشند. بعد از ورود نشانی آن‌ها به درون ساختار موتور جستجو، با عملیات مختلف و پیچیده‌ای که روی صفحات مورد نظر انجام می‌شود کل صفحه، نمایه می‌شود و در درون پایگاه اطلاعاتی موتور جستجو ذخیره می‌گردد. هم‌چنین به صفحاتی که قبلاً آن‌ها را نمایه کرده‌اند می‌روند و دوباره آنها را بر اساس آخرین تغییرات نمایه می‌کنند. هنگامی که کاربر اطلاعات را جستجو می‌کند پرس‌وجوی[2] وی در درون نرم‌افزار موتور جستجو ترجمه شده و بر اساس آن، از پایگاه اطلاعاتی نمایه، اطلاعات مرتبط، انتخاب می‌شود و با توجه به الگوریتم‌های رتبه‌بندی، فهرستی از نتایج ظاهر می‌شود.

1 What Is A Search Engine
2 Query





هر چند ساختار موتورهای جستجو در کلیات به هم شبیه‌اند اما هر کدام برحسب ویژگی‌ها، راهبردها و الگوریتم‌های مورد استفاده از یکدیگر متفاوت‌اند. همه موتورهای جستجو در بازار رقابتی، تلاش می‌کنند بتوانند هر چه بیش‌تر نظر کاربران را جلب کنند و با برآوردن نیازهای آنان از دیگر رقبای خود پیشی بگیرند. رقابت فن‌آوری و اطلاعاتی در دنیای امروز، بخش مهمی از سهم بازار تجارت را به خود اختصاص داده است و در این میان موتورهای جستجو و شرکت‌های مادر آن‌ها نیز از این قاعده مستثنی نیستند و برای رسیدن به اهداف عمده خود تلاش می‌کنند. برای همه موتورهای جستجو دو هدف عمده تعیین شده است: اثر بخشی و کارآیی[1] (ناومن[2]، ۲۰۱۱) که در حقیقت از مباحث ارزیابی عملکرد موتور جستجو هستند (داسدان، و سیاشلکلس ولاپسگلو[3]، ۲۰۱۰). در ادامه بر اساس این دو هدف می‌توان ساختار موتور جستجو را تشریح کرد.

## اثربخشی در موتورهای جستجو

اثربخشی در کلیه نظام‌ها مسأله‌ای اساسی است. در نظام‌های اطلاعاتی نیز این موضوع اهمیت دارد و بخشی از ارزیابی نظام‌های بازیابی اطلاعات و از جمله موتورهای جستجو است. این‌که چه اندازه از نظر کاربر، نتایج به‌دست آمده در یک جستجوی خاص، دارای کیفیت است و موتور جستجو توانایی عرضه اطلاعات درست و مرتبط را دارد مسأله بسیار مهمی است. اثربخشی موتورهای جستجو و کیفیت آن‌ها به عوامل مختلفی بستگی دارد. کیفیت را با استفاده از معیارهای اندازه‌گیری سنتی نمی‌توان اندازه‌گیری کرد (اسپینک و زیمر[4]، ۲۰۰۸) لواندوسکی[5] (۲۰۱۲)، اثربخشی موتورهای جستجو را در چهار بخش خلاصه می‌کند: کیفیت نمایه[6]، کیفیت نتایج[7]، کاربردپذیری[8] و کیفیت ویژگی‌های جستجو[9].

### کیفیت نمایه

ارائه نتایج مربوط و جامع را کیفیت نمایه گویند. ارزیابان به سه حوزه در

1 Query
2 Naumann
3 Dasdan, Tsioutsiouliklis, and Velipasaoglu
4 Spink and Zimmer
5 Lewandowski
6 Index quality
7 Result quality
8 Usability
9 Quality of search features

ارزیابی کیفیت نمایه، توجه کرده‌اند که شامل پوشش وب[1]، سوگیری کشوری[2] و روزآمدی[3] است (لواندوسکی، ۲۰۱۲). هر کدام از این حوزه‌ها به نوبه خود در کیفیت نتایج به نحو چشمگیری موثرند.

۱. *پوشش وب:* عبارت است از جامعیت نتایج ارائه-شده در موتور جستجو. یک موتور جستجو نمی‌تواند به‌صورت محلی فعالیت کند و در عین حال نیاز کاربران خود را به‌خوبی رفع کند. از طرفی، موتور جستجو هر اندازه قوی باشد تنها برشی از کل وب را می‌تواند در پایگاه داده خود نمایه کند و در پاسخ به پرسش کاربران عرضه نماید. اما کاربران انتظار دارند در میان نتایج، مجموعه‌ای از نتایج جستجوی مورد علاقه خود را ببینند. وقتی موتور جستجو، به کاربرانی از سراسر دنیا و چندین زبان مختلف خدمات می‌دهد و محتوای اطلاعاتی خود را از سراسر وب می‌گیرد، پوشش وب در آن به اندازه، جامعیت و تنوع نمایه برمی‌گردد (داسدان و دیگران، ۲۰۱۰). موتور جستجو هر چه‌قدر بتواند صفحات بیش‌تر، متنوع‌تر و با کیفیت بالاتری نمایه کند، موفق‌تر خواهد بود.

۲. *سوگیری کشوری:* یکی از عوامل موثر در پوشش وب است، یعنی کاربر هنگام جستجو به‌طور محلی به بخشی از موتور جستجو در منطقه خود متصل می‌شود. به‌طورمثال اگر کاربر در خارج از آمریکا در نوار نشانی خود، نشانی موتور جستجوی گوگل را تایپ کند این شانس وجود دارد که کاربر به موتور جستجوی گوگل در کشور خودش تغییر مسیر یابد. گوگل برای این کار دلایلی دارد. اما مهم‌ترین دلیل آن است که جستجوها سریع‌تر انجام شود. به عقیده گوگل کاربران در نتایج جستجوها تا حدی سوگیری کشوری دارند ( ژنالوژی در مجله زمان[4]، ۲۰۱۴) و جستجوهای نزدیک به کشور خود را بیش‌تر بیش‌تر انتخاب می‌کنند. به‌طور مثال، راهنمای یاهو، یاهوی مربوط به کشور چین دارد. وقتی کاربر در چین در نوار نشانی خود، نشانی یاهو را تایپ می‌کند به صورت خودکار به یاهوی حوزه کشور چین تغییر

1 Web coverage
2 Country bias
3 uptodateness
4 Genealogy InTime Magazine



مسیر می‌یابد و با سرعت بیشتری جستجو می‌کند.

۳. *روزآمدی:* گاهی با اصطلاح تازگی[1] بیان می‌شود. برخلاف تصور کاربران، موتور جستجویی مانند گوگل در لحظه پاسخ به پرسش کاربر، وب را جستجو نمی‌کند، بلکه نتایج را از پایگاه نمایه خود گرفته و عرضه می‌نماید. در واقع خزنده‌های موتور جستجو به‌صورت مدام صفحات را بازدید و آن‌ها را در پایگاه اطلاعاتی خود نمایه می‌کنند. به محض ارائه پرس‌وجو، موتور جستجو در پایگاه داده خود به دنبال نتایج احتمالی منطبق با پرسش کاربر می‌گردد و آنها را به کاربر می‌دهد. اما نمایه شدن یک صفحه در موتور جستجو، پایان کار نیست. چون صفحات به شکل ایستا در وب نمی‌مانند و بسیاری از آن‌ها در طول زمان، تغییر می‌کنند. برای کاربر بسیار مهم است که روزآمدترین اطلاعات را ببیند. قدرت یک موتور جستجو یا خزنده‌های آن در روزآمد نگه داشتن پایگاه داده، در کیفیت نمایه بسیار تأثیرگذار است. در واقع یکی از عوامل تبدیل شدن به موتور جستجوی با کیفیت، روزآمد بودن پایگاه داده نمایه است (بارکر[2]، ۲۰۰۳). دلایل زیر به‌طور مستقیم با تازگی و روزآمدی موتور جستجو ارتباط دارد:

- ✓ برخی از موتورهای جستجو چندین پایگاه داده و موتور پرس‌وجو دارند؛
- ✓ نمایه برخی موتورهای جستجو تقسیم‌بندی می‌شود و به‌صورت جزئی درمی‌آید؛
- ✓ وقتی خزنده یک موتور جستجو پایگاه اطلاعاتی خود را تجدید[3] می‌کند شاید برخی صفحات را که قبلاً بازدید کرده به دلیل خرابی خدمت‌دهنده[4] یا قطع ارتباط غیرقابل دسترسی باشند؛
- ✓ ممکن است به علت تغییرات در سیاست نمایه‌سازی موتورهای جستجو یا در اندازه پایگاه داده، نوسان‌هایی وجود داشته باشد (لواندوسکی، ۲۰۰۸).

## کیفیت نتایج

چون با قضاوت ربط ارتباط دارد آن را با آزمون‌های کلاسیک نظام‌های بازیابی اطلاعات می‌سنجند تا کیفیت واقعی یک موتور جستجو مشخص شود. به‌طورکلی

1 freshness
2 Barker
3 refresh
4 server failure

همه قضاوت‌های ربط و قضاوت کاربران در مورد کیفیت یک موتور جستجو، بیش‌تر بر اساس نتایجی است که موتور جستجو به کاربر می‌دهد. در حقیقت کاربران وقتی به نتایج دلخواه خود می‌رسند ترغیب می‌شوند تا دوباره از یک موتور جستجو استفاده کنند. پس برونداد یک موتور-جستجو و میزان ربط آن به پرسش و نیاز کاربر، رابطه بسیار مستقیمی با اثربخشی یک نظام بازیابی و به‌طوراخص یک موتور جستجو دارد.

پژوهشگران سودمندی موتور جستجو را به‌طور مستقیم با کیفیت نتایج یک موتور جستجو و زمان سپری شده برای جستجو مرتبط می‌دانند (مودیب و زیلبرستین[1]، ۱۹۹۹). بسیاری از پژوهش‌ها در نظام‌های بازیابی به افزایش کیفیت نتایج در این نظام‌ها بازمی‌گردد و پژوهشگران به فنون یکپارچگی اطلاعات در این حوزه بسیار علاقمندند (ناومن، ۲۰۰۱)

## کاربردپذیری (قابلیت استفاده)

راهبردهای بسیاری در موتورهای جستجو و بیش‌تر نظام‌های اطلاعاتی به‌کار می‌رود تا میزان کاربردپذیری نظام بازیابی افزایش یابد. تأکید کاربردپذیری بر آن است که فرد در استفاده از خدمت ارائه شده به هدف خود برسد. چون رسیدن به هدف، دلیل همه تلاش‌های کاربران برای استفاده از یک موتور جستجو است. در واقع یک ویژگی چندبعدی از یک نظام رایانه‌ای محسوب می‌شود (تاکسا، اسپینک و گلدبرگ[2]، ۲۰۰۸)

سوال اساسی در کاربردپذیری موتور جستجو این است که چه‌قدر موتور جستجو می‌تواند اطلاعات مورد نیاز کاربر را در اختیارش قرار دهد. تعریف استاندارد ایزو از کاربردپذیری میزانی است که کاربر می‌تواند یک محصول را به‌صورت موثر و کارا و با رضایتمندی در رسیدن به اهداف خاص استفاده کند (تاکسا و دیگران، ۲۰۰۸). مطالعات مختلف، طیف وسیعی از عناصر را به کاربردپذیری مربوط می‌دانند. دسترس‌پذیری[3] و در دسترس بودن[4] و اکتشاف اطلاعات[5] (هنگ[6]، ۲۰۱۴)، بازخورد کاربر، رضایت کاربر، توانایی برای نادیده گرفتن خطاها و فرمول‌بندی پرسش،

1 Mouaddib and Zilberstein
2 Taksa, Spink and Goldberg
3 Accessibility
4 Availability
5 Navigation
6 Heng



کارآیی و اثر بخشی موتور جستجو (تاکسا و دیگران، ۲۰۰۸)، توانایی یادگیری[1] و یادسپاری[2] (مقدمه‌ای بر کاربردپذیری[3]، ۲۰۱۲) برخی از این عناصراند. به‌طورمثال، توانایی یادگیری نشان می‌دهد تا چه اندازه کار با موتور جستجو برای کاربر تازه‌وارد آسان است و یادسپاری به این معنی است که چه‌قدر کارکرد موتور جستجو حتی اگر مدتی از آن استفاده نشود در خاطر می‌ماند.

اصل در کاربردپذیری موتور جستجو آن است که آیا برای کاربر امکان دارد یا آن به‌صورت کارا و موثر تعامل کند یا نه (جویس، بیسکری، گاناسیا، روکس[4]، ۲۰۱۲). لودانسکی (۲۰۰۸) معتقد است کاربردپذیری، بازخوردی از رفتار کاربر می‌دهد و می‌توان آن را با مشاهده پیمایش‌های کاربر، آزمون‌های آزمایشگاهی و تجزیه و تحلیل گزارش‌ها ارزیابی کرد.

## کیفیت ویژگی‌های جستجو

در موتور جستجو باید ویژگی‌هایی چون جستجوی پیشرفته و زبان پرس‌وجوی پیچیده پیش‌بینی شده باشد و با اطمینان و اعتماد کار کند (لواندسکی، ۲۰۰۸). هر موتور جستجو برای جلب رضایت کاربران و افزایش قدرت عرضه منابع مربوط و بالابردن دقت نتایج، راهبردهای مختلفی به‌کار می‌برد. وجود روش‌های مختلف جستجو مانند جستجوی کنترل‌شده با استفاده از اصطلاح‌نامه، جستجوی خوشه‌ای، جستجو بر اساس فیلترهای مختلف مانند جستجوی محدود به یک زبان، زمان و مکان خاص، حتی به یک دامنه یا یک حوزه خاص از ویژگی‌های عمده‌ای است که موتورهای جستجو دارند. بیش‌تر موتورهای جستجو مانند گوگل، جستجوی عکس، ویدئو یا کتاب و مقالات علمی را با امکانات پیشرفته خود انجام می‌دهند.

در کنار بحث اثربخشی هر نظام، بحث اساسی دیگر، کارآیی آن است. در موتورهای جستجو، نیز کارآیی به عملکرد صحیح نظام بازیابی برمی‌گردد.

# کارآیی موتورهای جستجو

این ویژگی به زمان و توان صرف شده برای جستجو در موتور جستجو اشاره می‌کند. زمان صرف شده برای جستجو برای کاربر اطلاعات بسیار مهم است. به‌طور

1 Learnability
2 Memorability
3 Introduction to Usability
4 Jouis, Biskri, Ganascia and Roux

مثال یکی از گروه‌های اصلی کاربری در نظام‌های بازیابی اطلاعات پژوهشگران حوزه پزشکی و بالینی‌اند و به دلیل اشتغال به کارهای درمانی، فرصت کمی برای جستجو دارند و اختصاص زمان زیاد برای جستجو برای آن‌ها مقدور نیست. پژوهشگران معتقدند یکی از عوامل تأثیرگذار بر رفتار اطلاع‌یابی کاربران، میزان زمانی است که آن‌ها برای جستجو در اختیار دارند (جگِد و آنویچکوا[1]، ۲۰۱۱). با این توصیف می‌توان نتیجه گرفت اهمیت بحث کارآیی در موتور جستجو بسیار بالاست. اگر اثربخشی تاکید می‌کند که چه‌قدر یک موتور جستجو در بازیابی اطلاعات درست برای کاربر موفق است کارآیی به این موضوع برمی‌گردد که موتور جستجو چه‌قدر در انجام وظیفه می‌تواند سریع عمل کند (کرافت، متزلر و استرامن، ۲۰۰۸). طراحان موتورهای جستجو تلاش می‌کنند که زمان پردازش پرسش‌ها را کاهش دهند تا بهره‌وری و کارآیی موتور جستجو را افزایش دهند. به‌طورمثال برخی از موتورهای جستجو از الگوریتم‌های فازی تعاملی برای کاهش زمان صرف‌شده در نظام‌های اطلاعاتی مانند پابمد[2] استفاده می‌کنند (وانگ و دیگران، ۲۰۱۰).

به‌طورکلی نوع نگاهی که در مورد بحث کارآیی و اثربخشی موتورهای جستجو در میان طراحان وجود دارد، بر فنون و الگوریتم‌های مورد استفاده در موتور جستجو تأثیر می‌گذارد. موتورهای جستجوی قدرتمند از ویژگی‌های جستجوی متعدد، الگوریتم‌های متنوع و فنون بازیابی پیشرفته استفاده می‌کنند تا بتوانند در صحنه رقابت با دیگر موتورهای جستجو موفق باشند. در ادامه ساختار و معماری موتور جستجو بررسی می‌شود.

## معماری موتور جستجو

یک موتور جستجو از دو بخش اصلی تشکیل شده و معماری یک موتور جستجو شامل دو بخش مهم فرآیند نمایه‌سازی[3] و پرس‌وجو[4] است (لیم، لیو و لی[5]، ۲۰۱۱). در معماری نظام بازیابی یک بخش دیگر هم اضافه می‌شود که بازخورد و بازیابی است (ناسا و ووک[6]، ۲۰۰۲). در ادامه بخش‌های مختلف موتور جستجو تشریح می‌شود.

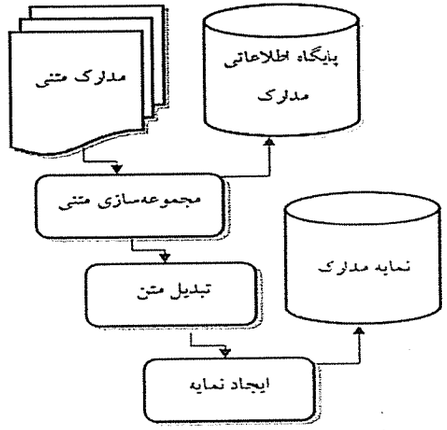
1 Jegede and Onwuchekwa
2 PubMed: www.PubMed.com
3 Indexing Process
4 Query Process
5 Lim, Liu and Lee
6 Gnasa and Woch



## فرآیند نمایه‌سازی در موتورهای جستجو

از مهم‌ترین فرآیندها در موتور جستجو است. قبلاً هم گفته شد که با ارائه پرس‌وجو به موتور جستجو، برنامه نرم‌افزاری آن در پایگاه داده خود جستجو می‌کند و از میان اطلاعات نمایه‌شده در این پایگاه، مدارک مرتبط با پرس‌وجو را به کاربر می‌دهد. در واقع موتور جستجو در هنگام دریافت جستجوی کاربر، اطلاعات را در محیط وب ردیابی نمی‌کند بلکه پایگاه اطلاعاتی خود را به عنوان مخزن اطلاعات جستجو می‌کند.

فرآیند نمایه‌سازی در موتور جستجو اسناد و مدارک را به عنوان درونداد می‌گیرد و از مدارک گرفته‌شده بروندادی به شکل نمایه ایجاد می‌کند. نمایه مدرک، نوعی از اصطلاح‌های نمایه‌ای یا ویژگی‌های مدارک پایگاه داده در موتور جستجو است (لیم و دیگران، ۲۰۱۱). شکل زیر، فرآیند نمایه‌سازی یک موتور جستجو را نشان می‌دهد.




شکل (۲-۱). زیر بخش‌های اصلی در فرآیند نمایه‌سازی در موتورجستجو

همانند شکل فوق، فرآیند نمایه‌سازی موتور جستجو شامل سه بخش مهم است: مجموعه سازی متنی[1]، تبدیل متن و ایجاد نمایه.

1 Text acquisition

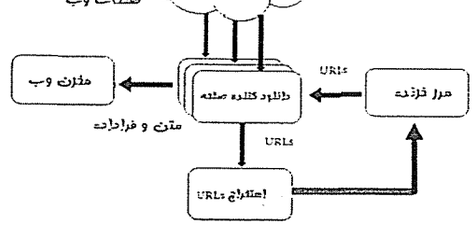
## مجموعه‌سازی متنی

این بخش از موتور جستجو، مدارک را برای نمایه‌شدن، ذخیره و شناسایی می‌کند. خزنده یا عنکبوت موتور جستجو، متن مدرک را از میان بخش عظیمی از صفحات وب انتخاب و برای نمایه‌شدن در پایگاه اطلاعاتی موتور جستجو ذخیره می‌کند. کارهای مختلفی برای رسیدن به این هدف در این بخش انجام می‌شود.

مجموعه سازی متنی خود شامل چهار بخش اصلی است که در مجموع کارشان ذخیره و شناسایی مدارک است. این چهار بخش شامل خزنده، تغذیه یا فید[1]، تبدیل و مخزن داده‌ای مدرک[2] است (ناومن، ۲۰۱۱) که در ادامه به آن‌ها اشاره می‌شود.

### خزنده یا عنکبوت

برنامه‌های نرم‌افزاری‌اند که در وب گردش و از صفحات وب بازدید می‌کنند. سپس آن‌ها را می‌خوانند و صفحات را به‌عنوان ورودی به بخش نمایه موتور جستجو می‌دهند. خزنده بر اساس توان نرم‌افزاری خود در سراسر وب می‌خزد و پیوند به دیگر صفحات را دنبال می‌کند تا مطمئن شود تا حد امکان تمامی صفحات را خوانده و برای نمایه‌شدن آماده کرده است. ساختار وب یک ساختار گرافیکی است و خزنده‌ها از پیوندهای ارائه شده در یک صفحه برای باز کردن صفحات دیگر استفاده می‌کنند. آن‌ها برای بازیابی صفحات و قراردادن آن‌ها، در یک مخزن محلی ساخته شده‌اند (یودپر، کیل و دارمیک[3]، ۲۰۱۴). شکل بعد ساختار یک خزنده را در موتور جستجو نشان می‌دهد.




شکل(۲-۲). معماری یک خزنده وب (یودپر، کیل و دارمیک، ۲۰۱۴)

1 feed
2 Document data store
3 Udapure, Kale and Dharmik



مطابق شکل یک، خزنده دارای سه بخش مجزاست. نخست مرز خزنده[1] است. این بخش فهرستی از نشانی صفحاتی را ذخیره می‌کند که خزنده آن را بازبینی نکرده است. به زبان ساده این بخش مجموعه‌ای از نشانی‌های اینترنتی صفحات است. دوم دانلودکننده صفحه[2] است که مهم‌ترین کار آن دانلود صفحاتی است که نشانی آن‌ها در بخش مرز خزنده ذخیره شده‌اند. مخزن وب[3] نیز بخشی است که دسته بزرگی از اشیای داده‌ای را ذخیره و مدیریت می‌کند. اشیای داده‌ای در این مخزن وب، همان صفحات وب هستند. این مخزن فقط صفحات اچ‌تی‌ام‌ال استاندارد را ذخیره می‌کند (یودپر و دیگران، ۲۰۱۴). به‌طور کلی خزنده وب ابتدا مجموعه‌ای از نشانی‌های اینترنتی را که تاکنون بازدید نشده‌اند یا صفحات آن‌ها روزآمد شده است به بخش مرز خزنده اضافه می‌کند. سپس خزنده از میان مجموعه نشانی‌ها، صفحاتی را انتخاب می‌کند و بخش دانلودکننده صفحات، آن صفحه را می‌یابد و دانلود می‌کند. صفحات دانلودشده به بخش مخزن خزنده برای انجام فرآیندهای بعدی، یعنی تبدیل متن[4] اضافه می‌شوند.

به‌طورکلی خزنده وب عملیات زیر را انجام می‌دهد:

- ✓ مقدار دهی اولیه نشانی‌های اینترنتی (گزینش نشانی‌ها)
- ✓ اضافه کردن آن‌ها به بخش مرزخزنده
- ✓ انتخاب نشانی‌ها از مرز خزنده
- ✓ گرفتن صفحات وب مربوط به این نشانی‌ها
- ✓ تجزیه[5] صفحه بازیابی‌شده برای استخراج[6] نشانی‌ها
- ✓ اضافه کردن تمام پیوندهای بازدید نشده در صفحه به فهرست نشانی‌ها، یعنی به مرز خزنده
- ✓ شروع مرحله دو و تکرار عملیات (یودپر و دیگران، ۲۰۱۴)

### بخش تغذیه یا فید

دومین بخش از مجموعه‌سازی متنی است و به جریان بلادرنگ مدارک[7] در وب اشاره می‌کند (نامن، ۲۰۱۱). این بخش مکانیسمی برای دسترسی به مدارک در وب است. از نمونه‌های مهم آن می‌توان به آراس‌اس[8] اشاره کرد. اما به‌طور کلی، فیدها؛

1 Web frontier
2 Web downloader
3 Web Repository
4 Text transformation
5 parsing
6 extract
7 Real-time stream od document
8 RSS

بلاگ‌ها، اخبار، ویدئو، رادیو و تلویزیون را دربرمی‌گیرند. همان‌طور که در بخش قبلی گفته شد، خزنده وب فقط صفحات اچ‌تی‌ام‌ال را می‌خواند و ذخیره می‌کند؛ اما همه صفحات وب، اچ‌تی‌ام‌ال نیستند. بخش اعظمی از وب صفحات ایکس‌ام‌ال[1] هستند مانند منابع دیداری و شنیداری. بخش فید مربوط به این نوع منابع است. در ادامه یک نمونه از فید یعنی آراس‌اس تشریح شده است.

آراس‌اس به‌طورمعمول استانداردی برای فیدهای وب است. یک آراس‌اس‌خوان[2] می‌تواند برای به اشتراک‌گذاری فیدهای آراس‌اس استفاده شود. شکل فیدهای آراس‌اس، ایکس‌ام‌ال هستند. آراس‌اس‌خوان، فیدها را مانیتور می‌کند و به محض ورود آن‌ها، محتوای جدید را آماده می‌کند (کرافت و دیگران، ۲۰۱۰). منابع تلویزیونی و رادیویی نیز نوعی از فیدها هستند که این بخش از مجموعه‌سازی متنی آن‌ها را می‌خواند. موتورهای جستجوی جدید در حال حاضر می‌توانند داده‌هایی مانند فیلم و یدئو را با تغییر منابع مثل همین دستیابی به فیدهای آراس‌اس و توئیتر ای‌پی‌آی[3] نمایه‌سازی کنند و با استفاده از ابزار تبدیل که در بخش بعدی توضیح داده داده می شود دامنه وسیعی از اطلاعات مانند فایل‌های فلش[4] و پی‌دی‌اف را به اصطلاحات نمایه‌ای تبدیل نمایند (دومان۵، ۲۰۱۱).

**تبدیل**

این بخش مربوط به تغییر قالب مدارک از شکل واقعی خود به شکل استاندارد قابل خواندن در موتور جستجو است. در این‌جا تبدیل حتی برای منابع رمزگذاری شده و منابعی که با زبان‌های برنامه‌نویسی به صورت کد درمی‌آیند هم اتفاق می‌افتد.

منابعی که خزنده یا فید بازبینی می‌کنند همه منابع متنی نیستند؛ بلکه منابع در وب به فرمت‌های مختلف مانند پی‌دی‌اف و ایکس‌ام‌ال، ورد[6] و پاورپوینت[7] و غیره هستند. موتورهای جستجو نیاز دارند که این منابع را به منابع متنی همراه با فراداده تبدیل کند. بخش تبدیل متنی در فرآیند نمایه‌سازی موتور جستجو، صفحات اچ‌تی‌ام‌ال و ایکس‌ام‌ال را به قالب استاندارد درمی‌آورند اما در مورد مدارک

1 XML
2 RSS reader
3 Twitter API
4 Flash
5 Doman
6 word
7 powerpoint



پی‌دی‌اف، پاورپوینت یا متون رمزگذاری شده مانند مدارک رمزگذاری شده با کدهای اسکی[1] اولین گام برای آماده‌سازی مدارک برای پردازش بیش‌تر، این نوع تبدیل متن است (کرافت و دیگران، ۲۰۱۰). این ابزار تبدیل، مدرک را به یک قالب برچسب متنی مانند اچ‌تی‌ام‌ال یا ایکس‌ام‌ال تبدیل می‌کند و برخی از اطلاعات مهم قالب‌بندی را نیز حفظ می‌کند (متزلر و لیوسکی[2]، ۲۰۱۲).

**مخزن داده‌ای مدرک**

این مخزن یا پایگاه داده‌ای، همان‌طور که از نامش مشخص است، محل ذخیره و مدیریت اطلاعاتی است که در بخش‌های قبلی جمع‌آوری شده است و شامل داده‌های ساختاریافته و اطلاعات مرتبط با آن‌هاست. داده‌های ساختاریافته؛ فراداده، پیوندها و متون وابسته به این پیوندها[3] را دربرمی‌گیرند (دومان، ۲۰۱۱). این مخزن می‌تواند یک پایگاه داده‌ای رابطه‌ای[4] باشد و به‌طورخاص از یک نظام ذخیره‌سازی با کارآیی بیش‌تر و ساده‌تر برای ذخیره تعداد بی‌شماری از مدارک استفاده کند (اسمیت۵، ۲۰۱۲). به‌طورکلی دلیل ذخیره‌سازی مدارک آن است که هم موتور جستجو برای پاسخ به پرسش‌های کاربر سریع‌تر به اطلاعات دسترسی داشته باشد و هم این پایگاه بتواند فهرست نتایج را تولید کند و خلاصه‌ای از مدارک را در اختیار کاربر قرار دهد (ناومن، ۲۰۰۱).

**تبدیل متن**

این بخش، اطلاعات شناسایی و ذخیره شده را در بخش مجموعه‌سازی متنی به اصطلاحات نمایه‌ای تبدیل می‌کند (متزلر و لیوسکی، ۲۰۱۲) یعنی متن گرفته شده از اینترنت را از قالب و شکل وبی خود تغییر می‌دهد و به شکل استاندارد قابل خواندن برای برنامه‌های کاربردی موتور جستجو تبدیل می‌کند. مراحل تبدیل متن شامل تجزیه[6]، ایست واژگان[7]، ریشه‌گیری[8]، تحلیل پیوند[9]، استخراج اطلاعات[10] و طبقه‌بندی کننده[11] است.

1 ASCII
2 Leuski
3 Anchor text
4 Relational database
5 Smith
6 Parsing
7 Word Stopping
8 stemming
9 Link Analysis
10 Information Extrction
11 Classifier

**تجزیه**

در این بخش، مدرک انتخاب شده بررسی می‌شود و بخش‌هایی که ارزش اطلاعاتی دارند جدا می‌شوند. این اولین مرحله برای تبدیل متن است. متن به توکن[1] یا کوچک‌ترین واحد اطلاعاتی با ارزش که اغلب هم کلمات‌اند شکسته و تجزیه می‌شود. تبدیل متن یک مدرک به توکن‌ها بسیار اهمیت دارد. برای این‌کار، ابتدا ساختارهای اصلی متن شناسایی می‌شوند. ساختارهای اصلی شامل عنوان، سرعنوان، پیوندها، اشکال و مانند این‌هاست. بعد از شناسایی، مرحله تجزیه به توکن‌ها انجام می‌شود (کرافت و دیگران، ۲۰۱۰) هر تجزیه‌کننده[2]، برای اجرا در تمام وب طراحی شده است و باید بتواند آرایه بزرگی از اشتباهات احتمالی را اداره کند. اشتباهات از غلط‌های املایی در کدهای اچ‌تی‌ام‌ال تا مشکلات بزرگ‌تر را دربرمی‌گیرد (گالی[3] و دیگران، ۲۰۰۵) در تجزیه باید به مسائلی مانند جداکننده‌ها[4]، آپوستروف، حروف بزرگ، کاراکترهای غیرالفبایی توجه کرد (ناومن، ۲۰۰۱). مشکلات احتمالی در زبان‌های غیرلاتین بیش‌تر است. مثلاً در زبان چینی که جداکننده کلمات وجود ندارند (کرافت ودیگران، ۲۰۱۰) یا در زبان فارسی، به‌طور مثال، تشخیص افعال دوکلمه‌ای و کلمات ترکیبی یکی از مشکلات اساسی است که ممکن است تبدیل را با مشکلاتی مواجه کند.

مشکل دیگر در تجزیه متن آن است که جداسازی کلمات خاص از کلمات عام دشوار است. به‌طور مثال در تجزیه متن، برای برنامه جداکننده بسیار مشکل است که میان یک کلمه مانند "میهن" به عنوان یک برند با کلمه میهن به معنای کشور تفاوت قائل شود. البته زبان‌های نشانه‌گذاری اغلب برای تشخیص ساختار از برچسب‌ها و نحو[5] موجود در زبان نشانه‌گذاری استفاده می‌کنند (ناومن، ۲۰۰۱).

**بازدارندگی یا ایست واژگان**

به معنای حذف کلمات عام و آن‌هایی است که در متن ارزش معنایی ندارند و فقط اطلاعات اضافه می‌دهند یا به جملات پیوستگی می‌بخشند. می‌توان به حروف اضافه، افعال کمکی یا معین یا کلماتی اشاره کرد که با فراوانی زیاد در متن ظاهر می‌شوند. به‌طور مثال در جمله زیر :

1 Token
2 Parser
3 Gulli
4 seperators
5 syntax



" برنامه های کاربردی کامپیوتری که به تازگی ارائه شده‌اند بر تلفن‌های همراه تأثیر زیادی گذاشته‌اند."

در این متن که شامل سیزده کلمه است تنها کلمات "برنامه‌های کاربردی، کامپیوتر، تلفن‌های همراه" ارزش نمایه‌شدن دارند و بقیه کلمات حذف می‌شوند. به این نوع کلمات، که بسیار هم در جملات و متون متعارف‌اند سیاهه بازدارنده[1] واژگان گفته می‌شود که بعید به‌نظر می‌رسد به معنای متن کمک کنند. حذف این‌گونه کلمات در کاهش فضای نمایه، زمان جستجو و کارآیی و اثربخشی موتور جستجو تأثیر فراوان دارد و در متن‌کاوی و داده‌کاوی نیز برای بازنمایی موثر متن مهم است.

**ریشه‌گیری**

بخش اصلی این واژه کلمه "stem" به معنای ساقه و ریشه است. می‌توان این اصطلاح را گرفتن ریشه کلمه و واژه معنا کرد. فرهنگ لغت موتورهای جستجو[2] به عنوان راهنمای جامع اصطلاحات موتورهای جستجو این اصطلاح را استفاده از تجزیه و تحلیل زبانی برای رسیدن به فرم ریشه یک کلمه ترجمه کرده است. در این فرهنگ لغت آمده:

"موتورهای جستجویی که از ریشه‌گیری استفاده می‌کنند ریشه واژه‌های مورد جستجو را در مدارک پایگاه داده خود مقایسه می‌کنند. برای مثال، اگر کاربر "مرور" را به عنوان پرسش وارد کند، موتور جستجو کلمه را به ریشه کلمه کاهش می‌دهد و تمام مدارک حاوی آن مانند مرور، مرورگر، مرور و مرورکردن را بازیابی می‌کند (فرهنگ لغت موتور جستجو، ۲۰۰۹)

در برخی موارد این اصطلاح در کنار اصطلاح lemmatization و گاهی به‌جای آن به‌کار می‌رود، هر چند که کاربرد اصطلاح stemming معمول‌تر است؛ باید تأکید کرد که این دو واژه از نظر تخصصی با هم تفاوت دارند. واژه "lemmatization" از "Lemma" به معنای "اصل موضوع" گرفته شده است. در اصل، به معنای "استخراج اصل چیزی" و در این‌جا استخراج اصل یک واژه و اصطلاح است. به همین دلیل شاید گاهی به‌جای واژه stemming یا هم ارز آن به‌کارمی‌رود. با مثال ساده زیر می‌توان تفاوت این دو تکنیک را به‌خوبی درک کرد.

در ریشه‌گیری اگر کاربر کلمه "ماشین" را برای پرس‌وجو به نظام وارد کند

1 stop words
2 Search engine Dictionary

نظام این واژه را با کلمه دیگری مانند "ماشین‌ها" منطبق می‌کند، اما در نظامی که از نوع دوم ریشه‌گیری استفاده می‌کند این انطباق نه فقط میان کلمه ماشین و ماشین‌هاست بلکه انطباق میان واژه "ماشین" و "اتومبیل" و "خودرو" نیز خواهد بود (ایده جدید در مهندسی[1]، ۲۰۱۳)

ریشه‌گیری، بخش‌های زایدی را در کلماتی که قرار است در پایگاه نمایه شود حذف می‌کند. به‌طورمثال پسوندها در کلمات حذف می‌شوند. این مسأله به بهبود بازیابی کمک می‌کند. چون هنگام جستجو، وقتی پرس‌وجو به نظام داده می‌شود همه کلماتی که با کلمه پرس‌وجو هم ریشه‌اند به‌عنوان پاسخ ارائه می‌شود. بدین ترتیب، بازیافت افزایش می‌یابد و کارآیی نظام بالا می‌رود. اما همین کاهش همه اشکال کلمه به ریشه و پایه کلمه، با افزایش بازیافت، باعث کاهش دقت بازیابی می‌شود. به عنوان نمونه فرض کنید وقتی کاربر کلمه "تحلیل" را به نظام می‌دهد تمام مدارک شامل واژه‌های "تحلیل"، "تحلیل‌گر"، "تحلیل‌کننده"، "تحلیل‌ها" برای او بازیابی می‌شود. بدین‌ترتیب، در مواردی که کاربر دقیقاً دنبال یک یا چند واژه مشخص است، ریشه‌گیری، کار را برای وی دشوار می‌کند (لیدی[2]، ۲۰۰۱).

به‌طور کلی در موتورهای جستجو، استفاده از برخی فنون کمک می‌کند که کاربر بتواند دقیقاً عین واژه را جستجو کند. به‌طورمثال در موتورجستجوی گوگل، گذاشتن علامت نقل قول ("") در دو طرف واژه در بخش جستجوی ساده یا استفاده از فیلترهایی مانند "this exact word or phrase" در بخش جستجوی پیشرفته در بخش انگلیسی زبان گوگل یا استفاده از فیلتر معادل آن در بخش فارسی زبان، یعنی، " دقیقاً این کلمه یا عبارت" به جستجوی عین واژه مورد نظر کاربر کمک می‌کند.

**تحلیل پیوند**

این بخش از تبدیل متن، با پیوندها و متن پیوندی یا (انکرتکست[3]) سروکار دارد. بخش زیادی از اطلاعات وب در پیوندها و همین متن‌های پیوندی است که اغلب به‌صورت فرامتن و آبی‌رنگ‌اند و گاهی زیر آن‌ها خط کشیده شده است. به زبان ساده، متن‌های پیوندی متونی‌اند که قابل کلیک کردن هستند. موتورهای جستجو وقتی برای بازدید صفحات به آن‌ها می‌خزند، نه تنها متن صفحه، بلکه کلیه پیوندها و متون پیوندی را نیز ردیابی و به‌صورت جداگانه نمایه‌سازی می‌کنند. حال برای

1 New Idea engineering
2 Liddy
3 Anchor text



تبدیل متن باید این نوع محتوای وب نیز برای موتورهای جستجو تحلیل شود تا وارد مخزن پایگاه داد شوند.

تحلیل پیوند، عمومیت و محبوبیت اطلاعات، یعنی رتبه‌بندی صفحات را شناسایی می‌کند. چون متون پیوندی می‌توانند بازنمایی صفحات نشانه‌گذاری شده را با پیوندها به‌طور معناداری افزایش دهند (ناومن، ۲۰۱۱). تجزیه و تحلیل پیوندها کمک می‌کند تا صفحات با محتوای اطلاعاتی بالاتر شناسایی شوند. برای این کار از الگوریتم‌های تحلیل پیوند استفاده می‌شود. این الگوریتم‌ها از جمله رتبه‌بندی صفحات[1] و جستجوی موضوعی منتج از فرامتن (HITS)[2] از محاسبات بردار ویژه استفاده می‌کنند تا صفحات معتبر را بر اساس ساختار فرامتن شناسایی کنند (زو و دیگران، ۲۰۰۳).

*الگوریتم رتبه‌بندی صفحات:* به‌همراه الگوریتم HITS، الگوریتم محبوب رتبه‌بندی در وب است که از سال ۱۹۹۸ به‌وجود آمد تا نتایج بهتری برای کاربران فراهم کند. این الگوریتم را سرجی برین و لاری پیج[3] دو دانشجوی دانشگاه استنفورد و مؤسسان گوگل درست کردند (دوی، گوپتا و دیکسیت۴، ۲۰۱۴). اما برای این‌که بدانیم این الگوریتم چگونه کار می‌کند باید اصطلاحات پیوند ورودی[5] و پیوند خروجی[6] تعریف شوند.

۱- پیوند ورودی صفحه $A$ : پیوندهایی که از صفحات دیگر به صفحه $A$ داده می‌شود. پیوندهای صفحات درونی یک سایت به خود صفحات آن سایت، به عنوان پیوندهای ورودی محاسبه نمی‌شوند. وجود یک پیوند از یک صفحه به صفحه دیگر، نشان‌دهنده نفوذ[7] صفحه مقصد است. هر چه صفحاتی که به یک صفحه پیوند می‌دهند می‌دهند دارای ارزش بالاتری باشند، از نقطه‌نظر رتبه‌بندی برای صفحه مقصد اهمیت بیش‌تری دارند (طیبی، تشکری هاشمی و محدث خراسانی، ۱۳۸۸).

۲- پیوند خروجی صفحه $A$ : پیوندهایی که از صفحه $A$ به صفحات دیگر داده می‌شود. پیوندهایی که از درون صفحه به صفحات دیگر آن سایت داده شود، به عنوان پیوندهای خروجی محاسبه نمی‌شوند به زبان ساده، امتیاز رتبه‌بندی صفحه $A$ برابر

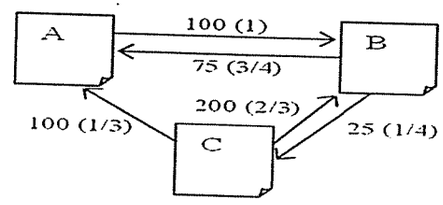
1 PageRank
2 Hyperlink-Induced Topic Search
3 Sergey Brin and Lary Page
4 Devi, Gupta and Dixit
5 In-link
6 Out-link
7 Authority

است با مجموع امتیاز رتبه صفحاتی که به صفحه **A** پیوند داده‌اند (طیبی، تشکری هاشمی و محدث خراسانی، ۱۳۸۸). نسخه ساده‌ای از معادله محاسبه رتبه‌بندی صفحات در ادامه نشان داده می‌شود:

$$PR(A) = (1-\mathrm{d}) + \mathrm{d}\left(\frac{\mathrm{PR(T1)}}{\mathrm{Q(T1)}} \dots\dots\dots + \frac{\mathrm{PR(T}n)}{\mathrm{Q(T}n)}\right)$$

در این معادله:

- ✓ PR (A) رتبه صفحه A است.
- ✓ T1.....Tn همه صفحاتی که به صفحه A پیوند داده‌اند.
- ✓ PR(Tn) رتبه صفحه $T_n$ است.
- ✓ Q (Tn) تعداد صفحاتی است که به $T_n$ پیوند داده‌اند.
- ✓ d ضریب تعدیل[1] که مجموعه‌ای بین صفر و یک است و به صورت اسمی ۸۵/. در نظر گرفته می‌شود (دوی و دیگران، ۲۰۱۴). ضریب تعدیل، احتمالی را نمایش می‌دهد که یک کاربر مدام روی پیوندها کلیک می‌کند
- ✓ (1-d) احتمالی که کاربر به یک صفحه تصادفی پرش می‌کند (ستایش، هارون‌آبادی و رحمانی،۱۳۹۲). حال به شکل ۲ دقت کنید:




شکل (۲-۳): گراف نمونه (اختر و سرور[2]، ۲۰۱۴)

با استفاده از معادله زیر، رتبه صفحات A، B و C به ترتیب با احتساب d=0 برابر است با:

PR (A) =1.2, PR (B) = 1.2, PR (C) = 0.8

در حقیقت رتبه یک صفحه وب به صورت مجموع رتبه تمام صفحاتی است که به صفحه پیوند داده‌اند تقسیم بر تعداد پیوندهایی که از صفحه به صفحات دیگر داده

1 Damping factor
2 Akhtar and Sarwar



شده است. باید گفت هر چه تعداد پیوندهای داده شده به صفحه (پیوندهای ورودی) بیشتر باشند، بهتر است. واضح است که پیوندهای ورودی نمی‌توانند بر رتبه صفحات اثر منفی بگذارند. پیوندهای کم ارزش یا بی‌ارزش هیچ تأثیری بر رتبه صفحات ندارند یعنی اثرشان بر ارزش صفحه و رتبه آن خنثی است. البته در عمل و در بخش موتورهای جستجو این کار بسیار پیچیده و بسیار فنی است که بیش‌تر در حوزه تخصصی مهندسین کامپیوتر و نرم‌افزار است و از حوصله مخاطب این کتاب خارج است. به‌طورمثال یک روش برای محاسبه رتبه یک صفحه، استفاده از power iteration است که قیدهای خاصی را طلب می‌کند و برای شرح آن باید از زنجیره مارکوف[1] استفاده کرد که خود یک مبحث فنی پیچیده است (طیبی و دیگران، ۱۳۸۶).

الگوریتم رتبه‌بندی صفحات، تنها الگوریتم رتبه‌بندی نیست. قبلاً نیز گفته شد که الگوریتم دیگری به نام HITS نیز وجود دارد که کلنبرگ[2] در کنفرانس ای‌سی‌ام[3] در سال ۱۹۹۸ ارائه کرد.

۴. *الگوریتم HITS:* این الگوریتم در حال حاضر در موتور جستجوی اسک[4] استفاده می‌شود (الگوریتم HITS[5]، ۲۰۱۴) و در رتبه‌بندی به پرس‌وجوی کاربر مربوط است و وظیفه آن رتبه دادن هنگام جستجوی کاربر است (دوی و دیگران، ۲۰۱۴) این الگوریتم در دو فاز نمونه‌گیری[6] و تکرار[7] کار می‌کند. در فاز نمونه‌گیری مجموعه‌ای از صفحات مرتبط با پرس‌وجوی مشخص جمع‌آوری می‌شود که شامل صفحات با نفوذ و اعتبار بالاست (زیرگراف S از G). در فاز تکرار، هاب‌ها و نفوذها با استفاده از معادله‌های زیر محاسبه می‌شوند:

$$Hp = \sum_{q \in I(p)} Aq$$

$$Ap = \sum_{q \in B(p)} Hq$$

در این معادله‌ها:

1 markov
2 Kleinberg
3 ACM
4 ASK: www.Ask.com
5 HITS Algorithm
6 sampling
7 Iteration

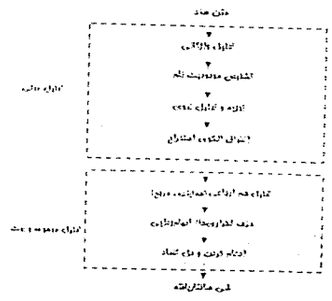
- ✓ $H_p$ وزن هاب
- ✓ $A_p$ وزن نفوذ
- ✓ $I_p$ و $B_p$ در مورد صفحات ارجاع دهنده و مرجع صفحه P بحث می‌کند (اختر و سرور، ۲۰۱۴)

حال برای درک بهتر کارکرد الگوریتم، تصور کنید کاربر پرس‌وجویی به موتور جستجو ارائه می‌کند. برای آن، تعدادی از صفحات، که احتمالاً مرتبط‌ترین نتایج هستند جدا می‌شوند. این مجموعه را S نام‌گذاری می‌کنیم. تمام صفحاتی را که به این مجموعه پیوند داده‌اند و آن‌هایی را که صفحات مجموعه S به آن پیوند داده‌اند G می‌نامیم. تعداد صفحات مجموعه G می‌تواند خیلی زیاد باشد. الگوریتم می‌تواند تعداد صفحات را به عدد خاصی، محدود کند. سپس الگوریتم روی مجموعه G کار می‌کند و برای هر صفحه، وزن هاب ($H_p$) و وزن نفوذ ($A_p$) را با استفاده از معادله‌های بالا برای تمام صفحات محاسبه کرده و از نتیجه به‌دست آمده برای رتبه‌بندی صفحات استفاده می‌کند (طیبی و دیگران، ۱۳۸۶).

در طول سالیان، خودِ این الگوریتم‌ها توسعه یافته‌اند و هم الگوریتم‌هایی دیگری مانند PCR عرضه شده‌اند که هر کدام مزایا و معایبی دارند. مشخص است هر چه‌قدر وسعت وب افزایش یابد موتورهای جستجو نیز سعی می‌کنند در عرصه رقابت با یکدیگر خود را توسعه دهند.

### استخراج اطلاعات

مرحله بعدی در بخش تبدیل متن، استخراج اطلاعات است که چند مرحله دارد.



شکل (۲-۴): مراحل استخراج اطلاعات در موتور جستجو (آجکتین[1]، ۲۰۰۵)

1 Agichtein



در بخش استخراج متن، ابتدا با تحلیل واژگانی[1] روبه‌رو هستیم. هر موتور جستجو، توانایی جستجو بر اساس تحلیل واژگانی و البته تحلیل گرامری[2] دارد (ما، سانگ، زو و ژانگ[3]، ۲۰۱۰). ایده پایه در تحلیل واژگانی آن است که کلمات، مهم‌ترین توصیف‌گر محتوای یک متن‌اند (آگوستی[4]، ۲۰۰۸). در این بخش متن به تکه‌هایی مثل جمله و پاراگراف تقسیم می‌شود و قوانین و الگوهایی به‌کار می‌روند تا موجودیت‌ها در متن شناسایی شوند (اگیچتین، ۲۰۰۵). بعد از تحلیل واژگانی و شکستن متن به موجودیت‌ها، مرحله تشخیص موجودیتِ نام یا اسامی[5] اجرا می‌شود تشخیص موجودیت نام، ابزاری است برای تشخیص اسامی و نوع آن‌ها اعم از اسامی افراد، اماکن، مقادیر عددی و غیره. برای تشخیص اسم بودن یک کلمه راه‌های مختلفی وجود دارد که از جمله، مراجعه به لغت‌نامه وردنت[6]، توجه به ریشه کلمه، استفاده از قواعد نحوی ساخت واژه و غیره است. در این ابزار، پس از تشخیص اسم‌ها نوع اسم مشخص می‌شود (استیری، ۱۳۹۱). در واقع نظام، معمولاً متن را برای کلیدهای[7] متنی که نشان‌دهنده یک موجودیت مفید[8] هستند پویش می‌کند (آجکتین، ۲۰۰۵). اما مرحله بعد، تحلیل نحوی[9] است. در این مرحله، برنامه؛ موضوع‌ها[10]، فعل‌ها[11]، اشیا[12]، قیدها[13]، صفت‌ها[14] و دیگر متمم‌های مختلف[15] را جدا می‌کند. یعنی بعد از تجزیه و تحلیل واژگانی، تحلیل نحوی در سطح کلمات انجام می‌شود تا رده کلمه را تشخیص دهد (نعیم و آصف[16]، ۲۰۱۱). مرحله بعد، انطباق الگوی استخراج[17] است. برای این بخش، الگوهای استخراج سطح سناریو[18] برای پی بردن به روابط بین موجودیت‌های استخراج‌شده به‌کار می‌روند. برخی از نظام‌ها می‌توانند از آمارهای

1 Lexical analysis
2 Grammar analysis
3 Ma, Song, Xu and Zhang
4 Agosti
5 Named entity recognition(NER)
6 word-net
7 Clue
8 Useful entity
9 syntactic analysis
10 Subject
11 Verb
12 Object
13 Adverb
14 Adjective
15 Complement
16 Naeem and Asif
17 Extraction Pattern Matching
18 scenario-level extraction patterns

جمع‌آوری شده از کل مجموعه استفاده کنند تا نمرات اطمینان را به اشیاء استخراج شده اختصاص دهند. در هر صورت پس از استخراج یا در طول آن، اطلاعات می‌تواند برای رخدادهای همایند چندگانه همان شی[1] استفاده شود (آجکتین، ۲۰۰۵). مراحل تحلیل واژگانی، تشخیص موجودیت نام، تحلیل نحوی، انطباق الگوی استخراج، مراحل تحلیل محلی[2] یک متن مدرک محسوب می‌شوند. مرحله بعد یعنی تحلیل مجموعه[3] شامل چند مرحله دیگر است که مراحل پس پردازش نامیده می‌شوند. این مراحل، اطلاعات را به‌صورت یکپارچه درمی‌آورند (آجکتین، ۲۰۰۵). مرحله اول در پس‌پردازش، تحلیل همایندی مرجع[4] است که در آن مرجع اسمی تعیین می‌شود. همایندی مرجع، ابزاری برای تعیین مرجع اسمی یک اسم یا ضمیر در جملات است. این ابزار در زبان انگلیسی معادل ابزاری است که مرجع ضمیر را که به صورت اسم در جمله‌های قبلی آمده، مشخص می‌کند. استفاده از ضمایر به جای اسامی در زبان انگلیسی بسیار رایج است (استیری، ۱۳۹۱). مرحله بعد، ابهام‌زدایی و حذف موارد تکراری[5] است. چالش وجود نام‌های هم‌معنی در اطلاعات استخراج شده از وب، "ابهام" در اطلاعات استخراج شده نام دارد و تشخیص نام‌های هم‌معنی در موجودیت‌ها و رابطه‌های این اطلاعات را ابهام‌زدایی گویند. به‌طورمعمول، در متن زبان طبیعی، موجودیت‌ها و روابط دنیای واقعی با نام‌های متفاوتی استفاده می‌شوند و سامانه‌های استخراج اطلاعات باید بتوانند این نام‌های متفاوت و متعدد را به یک موجودیت و رابطه دنیای واقعی ملحق کنند (ایمانی، ۱۳۹۲). آخرین مرحله برای یکپارچه‌سازی اطلاعات و رسیدن به یک شی ساختار یافته[6] ادغام و حل تضادهای موجود در اطلاعات استخراج شده است. منابع اطلاعاتی می‌توانند همدیگر را تأیید یا تکمیل کنند یا برخی اطلاعات آن‌ها در تضاد با همدیگر باشد. به‌طور مثال، اگر دو منبع از یک نویسنده با یک شماره شابک نشان داده شوند، این دو شی یک منبع هستند و همدیگر را تأیید می‌کنند. گاهی ممکن است در همین دو منبع با یک نویسنده و شابک یکسان، یک منبع، نام ناشر را داشته باشد و دیگری نام مترجم. این دو منبع همدیگر را تکمیل می‌کنند. برای همین دو منبع، اگر تعداد صفحات برای یکی ۲۰۵ صفحه و برای دیگری ۱۰۵ صفحه توصیف شده باشد، نشان می‌دهد که برخی

1 multiple occurrences of the same object
2 Local Analysis
3 Discourse And Collection Analysis
4 Coreference Resolution
5 Deduplication/Disambiguation
6 Structured object



ویژگی‌های این دو منبع با هم در تضاداند. حل تضادها و ادغام اطلاعات تأیید کننده و تکمیل کننده همدیگر، در مرحله استخراج اطلاعات بسیار چالش برانگیز است که در بخش ادغام و حل تضادها انجام می‌شود (ناومن و هاسلر[1]، ۲۰۰۲). در پایان این مرحله، متن به یک شی ساختار یافته تبدیل می‌شود.

**طبقه‌بندی‌کننده**

آخرین مرحله از بخش تبدیل متن است و فراداده‌های مربوط به رده‌ها[2] را در مدارک و یا بخش‌هایی از مدارک شناسایی می‌کند. کار این بخش در نظر گرفتن برچسب‌ها برای رده‌ها و طبقه‌هایی از اشیاست. این برچسب‌ها، به‌طورمعمول؛ طبقات موضوعی، مانند ورزش، سیاست و تجارت را نشان می‌دهند. این تنها نوع طبقه‌بندی نیست. از انواع دیگر این طبقه‌بندی می‌توان به محتوای هرز[3] یا منابع بی‌محتوا[4] مانند تبلیغات اشاره کرد. از فنون دیگر گروه‌بندی مدارک، فنون خوشه‌بندی مدارک مربوط است. این فنون از طبقه‌بندی از پیش‌تعریف‌شده استفاده نمی‌کند. این گروه‌های مدارک می‌توانند به روش‌های گوناگونی در طول رتبه‌بندی یا تعامل کاربر استفاده شوند (کرافت و دیگران، ۲۰۱۰). چهار مرحله در طبقه‌بندی وجود دارد. اول تولید بردار خصوصیت[5] که بسامد یک کلمه خاص خاص را در متن نشان می‌دهند. دوم کاهش ابعاد[6] که کاهش ابعاد بردار خصوصیت با استفاده از ماتریس کواریانس[7] و فنون پیچیده دیگر است. سوم یادگیری[8] که از بردار وزنی و از بردار میانگین و نمونه یادگیری ایجاد شده در مرحله قبل استفاده می‌کند و چهارم طبقه‌بندی که در آن بردار ویژگی با ابعاد کاهش یافته، در کلاس‌ها (تابع تفکیک[9]) طبقه‌بندی می‌شوند (زو، اوهیاما، واکابایاشی و کیمورا[10]، ۲۰۰۳)

**ایجاد نمایه**

آخرین بخش از فرآیند نمایه‌سازی، بعد از مجموعه‌سازی متنی و تبدیل متن است

1 Naumann and Häussler
2 Class-related Metadata
3 spam
4 Non- content
5 Feature vector generation
6 Dimension reduction
7 covariance matrix
8 Learning
9 discriminant function
10 Zu, Ohyama, Wakabayashi and Kimura

که خود شامل چهار بخش فرعی است. آمارهای مدارک[1]، وزن‌دهی[2]، مقلوب‌سازی[3] و مقلوب‌سازی[3] و توزیع نمایه[4] (ناومن، ۲۰۱۱) در ادامه به کارکرد آن‌ها اشاره شده است.

**آمارهای مدارک**

همان‌طور که از نامش پیداست کارهای آماری و کمی را روی اطلاعات انجام می‌دهد. این آمار می‌تواند در آمار کلمات، خصوصیات و مدارک باشد. شمارش و موقعیت کلمات و دیگر خصوصیات یک مدرک در میان گروهی از مدارک و در بین همه مدارک در این بخش است (ناومن، ۲۰۱۱). شمارش رخداد اصطلاحات نمایه‌ای شامل کلمات و خصوصیات پیچیده‌تر در مدارک و موقعیت آن‌ها در مدرک و در محل رخداد اصطلاح نمایه‌ای، شمارش رخدادها بر روی گروه‌های مدارک (مثلاً در گروهی از مدارک که با " ورزش" برچسب‌گذاری شده‌اند یا مجموعه‌ای از مدارک) و طول مدارک از نظر تعداد توکن از انواع داده‌های مورد نیاز بخش آمارند. آمارها در بخش جداول ارجاعی[5] نگهداری می‌شوند و در بخش رتبه‌بندی به‌کار می‌روند (کرافت و دیگران، ۲۰۱۰)

**وزن‌دهی**

وزن اختصاص داده شده به اصطلاحات در فایل نمایه است. ساده‌ترین موتورهای جستجو از وزن‌دهی دودویی استفاده می‌کنند. در این نظام‌ها عدد یک برای حضور یک اصطلاح و عدد صفر برای عدم حضور اصطلاح به‌کار می‌رود. موتورهای جستجوی پیچیده‌تر از طرح‌های پیچیده‌تر استفاده می‌کنند. اندازه‌گیری بسامد رخداد یک اصطلاح در مدرک با نرمال‌سازی بسامدها هنوز هم طرح پیچیده‌ای برای احتساب وزن است. پژوهش‌های بازیابی اطلاعات به وضوح نشان می‌دهد که وزن مطلوب با استفاده از "بسامد اصطلاح[6]" و "بسامد مقلوب مدرک[7]" به‌دست می‌آید. می‌آید. الگوریتم بسامد، رخداد هر اصطلاح را در درون یک مدرک اندازه می‌گیرد. سپس الگوریتم، بسامد را با بسامد رخداد در کل پایگاه داده مقایسه می‌کند. همه اصطلاحات یک مدرک، به‌خوبی نمی‌توانند میان یک مدرک با مدرک دیگر یا در کل

1 Document Statistics
2 Weighting
3 Inversion
4 Index istribution
5 Lookup Tables
6 Term frequency(tf)
7 inverse document frequency(idf)



پایگاه تمایز قائل شوند. به‌طور مثال، در پایگاه موضوعی ورزش، ممکن است اصطلاح "آنتی‌بیوتیک" شاخص خوبی برای تمایز یک مدرک در کل پایگاه باشد و وزن بسامد خوبی را نیز به خود اختصاص دهد. اما همین اصطلاح ممکن است در یک پایگاه موضوعی بهداشت یا پزشکی ارزش کمی داشته باشد و وزن بسامد پایینی را نیز به خود اختصاص دهد (لیدی، ۲۰۰۱). بسامد مقلوب مدرک در سال ۱۹۷۲ ارائه شد و تاکنون از آن به‌صورت گسترده‌ای در تابع TF*IDF استفاده شده است (رابرتسون[1]، ۲۰۰۴)

بسامد از طریق Log dkN محاسبه می‌شود و در آن N تعداد مدارک مجموعه و dk تعداد مدارکی است که اصطلاح k در آن ظاهر می‌شود. فرمول‌های مختلفی برای محاسبه اوزان اصطلاح وجود دارد که برخی از آن‌ها گونه‌هایی از وزن IDF بوده و از بسامد مدرک (تعداد دفعاتی که اصطلاحی در مدرک ظاهر می‌شود) و نرمال‌سازی بهره می‌گیرند (بهمن‌آبادی، ۱۳۸۹). فنون وزن‌دهی، کاربرد مهمی در رتبه‌بندی صفحات دارند (ناومن، ۲۰۰۱).

**مقلوب‌سازی**

هسته ایجاد نمایه است و کاربرد اصلی آن، تولید نمایه‌هایی است که می‌تواند برای بازیابی استفاده شود. بدون نمایه مقلوب، کار جستجوی اطلاعات بسیار مشکل می‌شود. وقتی اصطلاحات مجموعه مدارک یک پایگاه اطلاعاتی استخراج و به عنوان نماینده‌های مدرک معرفی می‌شوند می‌توانند به چندین صورت فهرست شوند. در یک حالت می‌تواند فقط الفبایی مرتب شوند. در این حالت وقتی کلیدواژه‌ای جستجو و با اصطلاح نمایه‌ای منطبق می‌شود نمی‌توان تعیین کرد مدرک حاوی این کلیدواژه کدام است. در نمایه مقلوب، هر کلیدواژه، شناساگر هر مدرک است. بخش مقلوب‌سازی، اطلاعات را از حالت مدرک – اصطلاح در بخش تبدیل مدرک می‌گیرد و آن را به اطلاعات اصطلاح – مدرک تغییر می‌دهد. در اطلاعات بخش تبدیل، هر مدرک یا اصطلاحاتی نمایه شده است و در بخش نمایه، اصطلاح معرف هر مدرک می‌شود و بدین ترتیب نمایه‌های مقلوب به‌وجود می‌آید. برای ایجاد نمایه‌های مقلوب هم در زمان ایجاد نمایه مقلوب اولیه برای مجموعه بزرگی از مدارک و هم در زمان روزآمدسازی نمایه‌ها با چالش مواجه هستیم. این بخش موتور جستجو، اهمیت زیادی برای سرعت بخشیدن به پردازش پرس‌وجو دارد و تا اندازه‌ای به الگوریتم رتبه‌بندی

1 Robertson

وابسته است (کرافت،۲۰۱۰). در نمایه‌های مقلوب، مجموعه‌ای از مدارک و نماینده‌های آن (کلیدواژه‌ها، ویژگی‌ها و وزن‌ها) وجود دارند. اطلاعات به شکل فهرست مرتب‌شده‌ای از کلیدواژه‌ها منظم شده‌اند. هر کلیدواژه با وزن خود در نمایه پیوند خورده است. با جستجوی کلیدواژه و انطباق با نمایه‌های این بخش، بر اساس وزن هر اصطلاح و بسته به الگوریتم رتبه‌بندی، مدارک مرتبط با اصطلاحات به عنوان نتیجه جستجو؛ به کاربر ارائه می‌شوند.

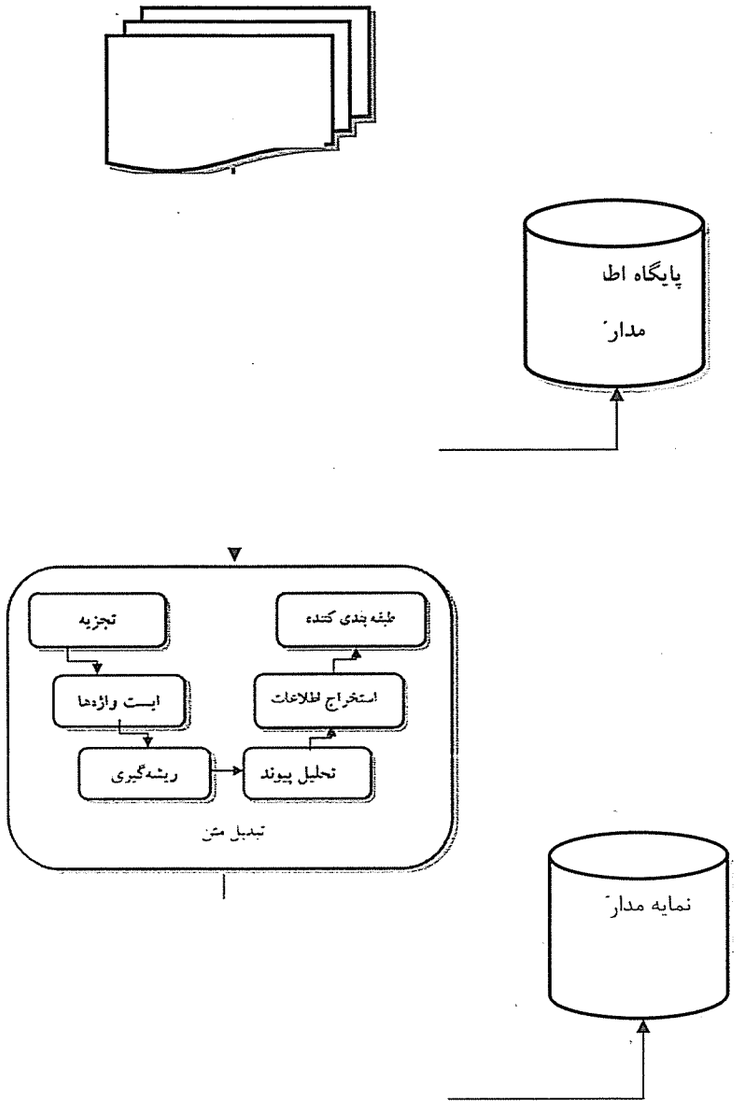
**توزیع نمایه**

در یک موتور جستجو مجموعه‌های بزرگ مدارک وجود دارد. حجم زیاد و اندازه بزرگ آن‌ها تشکیل یک ساختار نمایه‌ای موثر را مشکل می‌کند. برای ایجاد نمایه‌های با کیفیت، باید از الگوریتم‌های نمایه توزیع‌شده استفاده کرد و با آن‌ها نمایه‌ای توزیع‌شده به‌وجود آورد که در بین چند ماشین تقسیم می‌شوند (منینگ، رغاوان و شوتز، ۲۰۰۸ الف). توزیع نه فقط در سطح رایانه‌ها، بلکه در سطح سایت‌ها نیز انجام می‌گیرد. با چنین توزیعی در سطح گروه فرعی از مدارک، نمایه‌سازی و پردازش پرس‌وجو می‌تواند به‌صورت موازی باشد. هم‌چنین توزیع نمایه‌ها در گروه‌های فرعی از اصطلاحات می‌تواند از پردازش مساوی پرس‌وجو حمایت کند. هر جا نسخه‌هایی از یک نمایه یا قسمتی از یک نمایه در سایت‌های مختلف ذخیره شود یک نوع توزیع نمایه صورت می‌گیرد. این تکرار در حقیقت شکلی از نمایه توزیع شده است. یک شکل کم‌تر سازمان‌یافته از نمایه توزیع‌شده، جستجوهای نظیر به نظیر در هر گره[1] در شبکه است. این گره‌ها، نمایه‌ها و مجموعه مدارک خود را نگه می‌دارند (کرافت، ۲۰۱۰).

در ادامه شکل ساده بخش فرآیند نمایه‌سازی در موتور جستجو نمایش داده می‌شود.

1 node





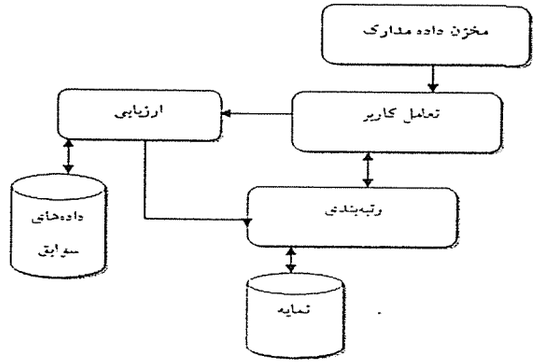

شکل (۲-۵): بخش فرآیند نمایه‌سازی در موتور جستجو

همان‌طور که قبلاً گفته شد موتور جستجو یک بخش اساسی و مهم دیگر نیز دارد و آن فرآیند پرس‌وجو است. در ادامه این بخش مهم از معماری موتورهای جستجو معرفی می‌شود.

## فرآیند پرس‌وجو در موتورهای جستجو

بخش فرآیند نمایه‌سازی یک سوی بازیابی اطلاعات است و با مدارک موجود در وب سروکار دارد. این بخش از موتور جستجو مدرک را آماده می‌کند تا به کاربر داده شود. اما کاربری که در پی اطلاعات است پرس‌وجوی خود را در قالب عبارات و کلیدواژه‌ها به نظام می‌دهد. پرس‌وجوی کاربر در بخش فرآیند پرس‌وجو تجزیه و تحلیل و برای موتور جستجو قابل خواندن می‌شود. موتور جستجو آن‌چه را که ارائه شده با اصطلاحات نمایه‌ای در پایگاه اطلاعاتی خود تطبیق می‌دهد. سپس بر اساس الگوریتم رتبه‌بندی، فهرستی از مدارک مرتبط، به کاربر داده می‌شود. تجزیه و تحلیل و آماده‌سازی پرس‌وجوی کاربر، پیچیده است.

فرآیند پرس‌وجو شامل سه بخش مهم است. تعامل کاربر[1]، رتبه‌بندی[2] و ارزیابی[3] (کرافت و دیگران، ۲۰۱۰). در شکل زیر فرآیند پرس‌وجو و بخش‌های مختلف آن نشان داده می‌شود.




شکل(۲-۶) : بخش‌های اصلی فرآیند پرس‌وجو

1 User Interaction
2 Ranking
3 Evaluation



### تعامل کاربر

وظیفه این بخش از لحظه‌ای شروع می‌شود که کاربر، درخواست خود را به موتورجستجو می‌دهد و با مشاهده نتایج تمام می‌شود. این بخش، خود شامل سه بخش فرعی است. درونداد پرس‌وجو[1]، تبدیل پرس‌وجو[2] و برونداد نتایج[3].

### درونداد پرس‌وجو

این بخش برنامه رابط کاربری و یک تجزیه‌کننده است و با زبان پرس‌وجو سروکار دارد (کرافت، ۲۰۱۰). قبل از بحث زبان پرس‌وجو باید بدانیم کاربران به‌طور عمده دنبال چه اطلاعاتی هستند و به‌خاطر آن از موتورهای جستجو استفاده می‌کنند. گیبسون[4] پرسش‌های کاربران را به سه نوع تقسیم کرده است. انجام‌دادن[5]، دانستن[6] و رفتن[7].

- ✓ پرسش‌های " انجام دادن": پرسش‌های محتوای تراکنشی[8] نیز نامیده می‌شوند. کاربران در این پرسش‌ها می‌خواهند *کاری* انجام دهند. در واقع تمایل کاربر را برای اجرای یک عمل نشان می‌دهند؛ مانند خرید ماشین، بلیط سینما یا یک کتاب.
- ✓ پرسش‌های دانستن: پرسش‌های محتوای اطلاعاتی[9] نامیده می‌شوند. کاربران می‌خواهند چیزی را بدانند و از آن آگاهی یابند. این پرسش‌ها حوزه وسیعی از اطلاعات را پوشش می‌دهند و هزاران نتیجه مرتبط با درخواست آن‌ها در وب وجود دارد. در واقع کاربران دنبال پاسخ‌های سریع و آسانند.
- ✓ پرسش‌های رفتن: پرسش‌های محتوای اکتشافی[10] نیز نام دارند. کاربران، دنبال یک یا چند صفحه از یک موجودیت وب هستند یا به دنبال یک وب‌سایت واحد می‌گردند. به‌طور مثال یک سایت مانند یوتیوب یا یک مقاله (گیبسون[11]، ۲۰۱۳). مور[12] نوع چهارمی از پرسش را نیز معرفی می‌کند به نام پرسش‌های اتصال[13].
- ✓ پرسش‌های اتصال: اتصالی از گراف وب نمایه شده را گزارش می‌کنند.

1 Transactional Content
2 Query transformation
3 Result output
4 Gibbson
5 Do
6 Know
7 Go
8 Transactional Content
9 Informational Content
10 Navigational Content
11 Gibbson
12 Moor
13 Connectivity

برای مثال کدام پیوند به این نشانی اینترنتی اشاره می‌کند یا چه تعداد صفحات از این نام دامنه[1] نمایه شده‌اند (مور، ۲۰۰۸)

پرسش کاربر می‌تواند از ساده‌ترین تا پیچیده‌ترین شکل باشد. رابط کاربری که این پرسش‌ها را به شکل قابل درک برای موتور جستجو تبدیل می‌کند، باید قابلیت‌های فراوانی داشته باشد. زبان پرس‌وجو در این میان نقش به‌سزایی دارد. زبان‌های پرس‌وجو به کاربران اجازه می‌دهند تا شکل‌های خاصی از پرس‌وجو را مطرح کنند. به‌طور مثال می‌توانند از نظام درخواست کنند اصطلاحات نزدیک به‌هم یا با نظم خاص در یک مدرک بازیابی شود. حتی زبان‌های جستجو باعث می‌شوند کاربران بتوانند اهمیت اصطلاحات و عبارات را با وزن آن‌ها، تعیین کنند یا از نظام بخواهند برخی عبارات در کنار هم و با هم استفاده شوند تا پرس‌وجو دقیق باشد (استرامن[2] و دیگران، ۲۰۰۵). استفاده از عملگرها معمول‌ترین فنونی است که در زبان‌های پرس‌وجو استفاده می‌شود تا کاربران بتوانند پرس‌وجوی خود را فرمول‌بندی کنند. عملگرهای بولی یا عملگرهای مجموعه‌ای شامل اجتماع (و)، اشتراک (یا) و تفاضل (به‌جز) هستند و با کمک سایر عملگرها مانند علامت نقل قول، علامت * یا # می‌توانند در جستجوهای خاص‌تر به کاربران کمک کنند. تمامی این ویژگی‌ها با استفاده از زبان پرس‌وجو در نظام بازیابی اطلاعات و به‌ویژه موتور جستجو میسر می‌شود.

**تبدیل پرس‌وجو**

فنونی مانند بازدارندگی واژگان و ریشه‌گیری در این بخش نیز کاربرد دارند (کرافت و دیگران، ۲۰۱۰). فنون دیگر مانند کنترل املای کلمات و فهرست پیشنهادهای پرس‌وجو نیز از فنون تبدیل متن است. به‌طور مثال در موتور جستجوی گوگل هر زمان واژه‌ای با املای غلط جستجو شود، موتور گوگل به‌صورت خودکار کلمه را اصلاح و بعد با کلمه اصلاح‌شده جستجو می‌کند و نتایج را در صفحه نتایج نشان می‌دهد یا به محض تایپ یک واژه، فهرستی از واژگان و عبارت‌ها را پیشنهاد می‌کند. بازخورد ربط[3] نیز در این بخش از پرس‌وجوست که مسأله مهمی است. پورقاسم و قاسمیان (۱۳۸۸) دو روش عمده در بازخورد ربط را حرکت نقطه پرس‌وجو و اصلاح وزن می‌دانند.

1 Domain name
2 Turtle
3 Relevance Feedback



۱- *حرکت نقطه پرس‌وجو:* در این روش سعی می‌شود با حرکت نقطه جستجوی جاری به کمک بازخورد کاربر، تخمین نقطه جستجوی ایده‌ال (به‌صورت ویژگی‌های سطح پایین) بهتر شود.

۲- *اصلاح وزن‌ها:* در این روش، وزن‌ها و پارامترهای مورد استفاده در تعیین میزان تشابه بر اساس بازخورد کاربر اصلاح می‌شود (پورقاسم و قاسمیان، ۱۳۸۸).

موضوع مهم دیگر در بخش تبدیل پرس‌وجو که اغلب در کنار بازخورد ربط مطرح می‌شود موضوع گسترش پرس‌وجو[1] است. برای فهم بهتر این مسأله به یکی از مسائلی که در نظام‌های بازیابی اطلاعات وجود دارد اشاره می‌کنیم و آن کلمات مترادف[2] است که می‌تواند بر دقت و بازیافت اطلاعات تأثیر بگذارد. برای حل این مشکل دو نوع روش وجود دارد: روش‌های جهانی[3] و محلی[4]

در روش‌های جهانی، گسترش پرس‌وجو و فرمول‌بندی مجدد اصطلاحات پرس‌وجو، مستقل از پرسش و نتایج ارائه شده است. به‌طوری که تغییرات در جمله‌بندی پرس‌وجو باعث خواهد شد پرس‌وجوی جدید با واژه‌های دیگر دارای معانی مشابه مطابقت پیدا کند. روش‌های جهانی عبارتند از گسترش پرس‌وجو/ فرمول‌بندی مجدد با اصطلاح‌نامه[5] یا وردنت[6]، گسترش پرس‌وجو از طریق تولید اصطلاح‌نامه خودکار و فنونی مانند تصحیح املا. اما در روش محلی، پرس‌وجویی در مورد مدارک تنظیم می‌شود که در ابتدا به نظر می‌رسد مطابق با پرس‌و جوی کاربر است. بازخورد ربط[7]، ربط[7]، بازخورد شبه‌ربط یا بازخورد ربط کور[8] و بازخورد ربط غیرمستقیم[9] از انواع انواع روش‌های محلی است (منینگ، ۲۰۰۸، الف).

**برونداد نتایج**

بخش نهایی در تبدیل متن، برونداد نتایج است و بر نوع نمایش نتایج در موتورهای جستجو تمرکز دارد. انواع برونداد نتایج در موتورهای جستجو عبارتند از رتبه‌بندی نتایج بر اساس ربط مدرک به جستجوی کاربر، برجسته کردن اصطلاحات جستجو شده در بخش‌های مختلف مدرک مانند عنوان، چکیده یا بخش‌های

1 Query Expansion
2 Synonyms
3 Global methods
4 Local methods
5 thesaurus
6 WordNet
7 Relevance feedback
8 Pseudo relevance feedback or Blind relevance feedback
9 indirect relevance feedback

قابل نمایش در نتایج، بیان تعداد نسخه‌های مختلف و امکان دسترسی به آن‌ها در هر فهرست نتیجه در گوگل و خوشه‌بندی نتایج در گروه‌های موضوعی.

### رتبه‌بندی

این بخش بسیار به مدل‌های بازیابی وابسته است که در فصل سوم مفصل در مورد آن صحبت می‌شود. رتبه‌بندی از سه بخش فرعی تشکیل شده است. نمره‌دهی[1]، بهینه‌سازی[2] و توزیع[3].

### نمره‌دهی

روشی برای بهبود موتورهای جستجوی وب است و این کار را با اختصاص دادن نمره بر اساس میزان اهمیت صفحات انجام می‌دهد. الگوریتم‌های رتبه‌بندی در این نمره‌دهی اهمیت زیادی دارند. مثلاً الگوریتم رتبه‌بندی صفحات گوگل یکی از شناخته‌ترین الگوریتم‌های رتبه‌بندی است (ایکاوا و ساداکنی[4]، ۲۰۰۴). وزن‌دهی بر اساس وزن اصطلاح پرس‌وجو ($q_i$) و وزن اصطلاح در مدرک ($d_i$) محاسبه می‌شود. معادله محاسبه این امتیاز (امتیاز وضعیت بازیابی[5]) به صورت زیر است (رسلوفو و ساوی[6]، ۲۰۰۳):

$$RSV = \sum_i q_i d_i$$

### بهینه‌سازی

موضوع بهینه‌سازی در موتورهای جستجو مهم و اساسی است. موتور جستجویی مانند گوگل، چهار مکانیسم اساسی دارد: کشف، ذخیره‌سازی، رتبه‌بندی و ارائه نتایج؛ که مهم‌ترین مسائل بهینه‌سازی‌اند. بهینه‌سازی نیازمند فهم عمیق از کشف و رتبه‌بندی در موتورهای جستجو است (دیویس، ۲۰۰۶). الگوریتم‌های رتبه‌بندی صفحات که نقش اساسی در بهینه‌سازی دارند، در بخش‌های قبل آمده‌اند.

### توزیع

قبلاً در مورد نمایه توزیع شده در بخش فرآیند نمایه‌سازی صحبت شد. این بخش در فرآیند پرس‌وجو نیز هست. پرسش به نظام داده می‌شود ولی در درون نظام به

1 Scoring
2 optimization
3 distribution
4 Ikawa an Sadakane
5 Retrieval status scores
6 Rasolofo and Savoy



همان شکل پذیرفته نمی‌شود؛ بلکه بعد از تجزیه و تحلیل برای نظام قابل خواندن می‌شود. تبدیل کردن به توکن، حذف کلمات بی‌ارزش و ریشه‌گیری روی پرسش انجام می‌شود. در نظام، نرم‌افزار کارگزار پرس‌وجوی[1] موجود در بخش توزیع فرآیند، این وظایف را به‌عهده دارد (بیزا-ییتس، وریز، زاراگوزا، کامبازوگلی و مورداک، ۲۰۱۲)

توزیع به راهبردهای بازیابی مختلف نیازمند است که بر عهده کارگزار پرس‌وجو است. کارگزار، پرس‌وجو را به همه خدمت دهنده‌ها می‌برد، نتایج بخش‌ها را در هم ادغام می‌کند و نتایج نهایی را به کاربر ارائه می‌دهد (منینگ و دیگران، ۲۰۰۸ ب)

### ارزیابی

هیچ نظامی بدون ارزیابی و نظارت نمی‌تواند به پویایی و تکامل خود ادامه دهد. بازخوردی که از نتایج ارزیابی گرفته می‌شود می‌تواند به تغییر راهبردها و افزایش کارآیی و اثربخشی نظام‌ها از بیرون و درون کمک کند. ارزیابی بیرونی نظام از دیدگاه کاربران است و ارزیابی درونی، مکانیسم‌هایی است که نظام برای سنجش عملکرد خود در ارائه نتایج به‌کار می‌گیرد. این نوع ارزیابی در موتورهای جستجو شامل سه بخش فرعی است؛ گزارش سوابق[2]، تحلیل رتبه‌بندی[3]، و تحلیل عملکرد[4].

### گزارش سوابق

حاوی اطلاعات بسیار زیاد و ارزشمندی برای بازخورد به نظام است تا بتواند عملکرد خود را بهبود بخشد. گزارش سوابق که شامل تعامل کاربران با موتورهای جستجو است، منبع بسیار مهمی برای توسعه موتورهای جستجوی وب است. از نقطه‌نظر ارزیابی، داده‌های به‌دست آمده از این بخش نشان می‌دهد که کاربران چگونه نتایج ارائه شده در موتور جستجو را مرور می‌کنند (کرافت و دیگران، ۲۰۰۸). سوابق می‌توانند برای اصلاح املای کلمات، فهرست پیشنهادها، ذخیره پرس‌وجو[5] و کمک به به انطباق تبلیغات با جستجو استفاده شوند (کرافت و دیگران، ۲۰۰۸). یکی از فرض‌ها در این بخش آن است که صفحات کلیک شده به پرس‌وجوی کاربر مربوط‌اند (ناومن، ۲۰۱۱). این داده‌ها برای بهبود کارآیی و اثربخشی موتور جستجو

1 Query broker
2 logging
3 Ranking analysis
4 Performance analysis
5 Query caching

اهمیت زیادی دارند.

**تحلیل رتبه‌بندی**

همان‌طور که قبلاً نیز اشاره شد، در بخش فرآیند نمایه‌سازی و هم‌چنین پرس‌وجو از الگوریتم‌های رتبه‌بندی برای مدارک و پرس‌وجو استفاده می‌شود. در این بخش، رتبه جفت مدرک و پرس‌وجو تحلیل و ارزیابی می‌شود (کرافت، ۲۰۱۰) و به ارزیابان فرصت می‌دهد تا وضعیت موتور جستجوی خود را بررسی کنند.

در واقع تحلیل رتبه‌بندی، اثربخشی موتورهای جستجو را بررسی می‌کند. معیارهای زیادی برای ارزیابی رتبه‌بندی و الگوریتم‌های رتبه‌بندی موتور جستجو استفاده می‌شوند. مجموعه‌های آزمون در این بخش اهمیت زیادی دارند. در بخش معیارهای ارزیابی به‌طور مفصل به آن اشاره می‌شود.

**تحلیل عملکرد**

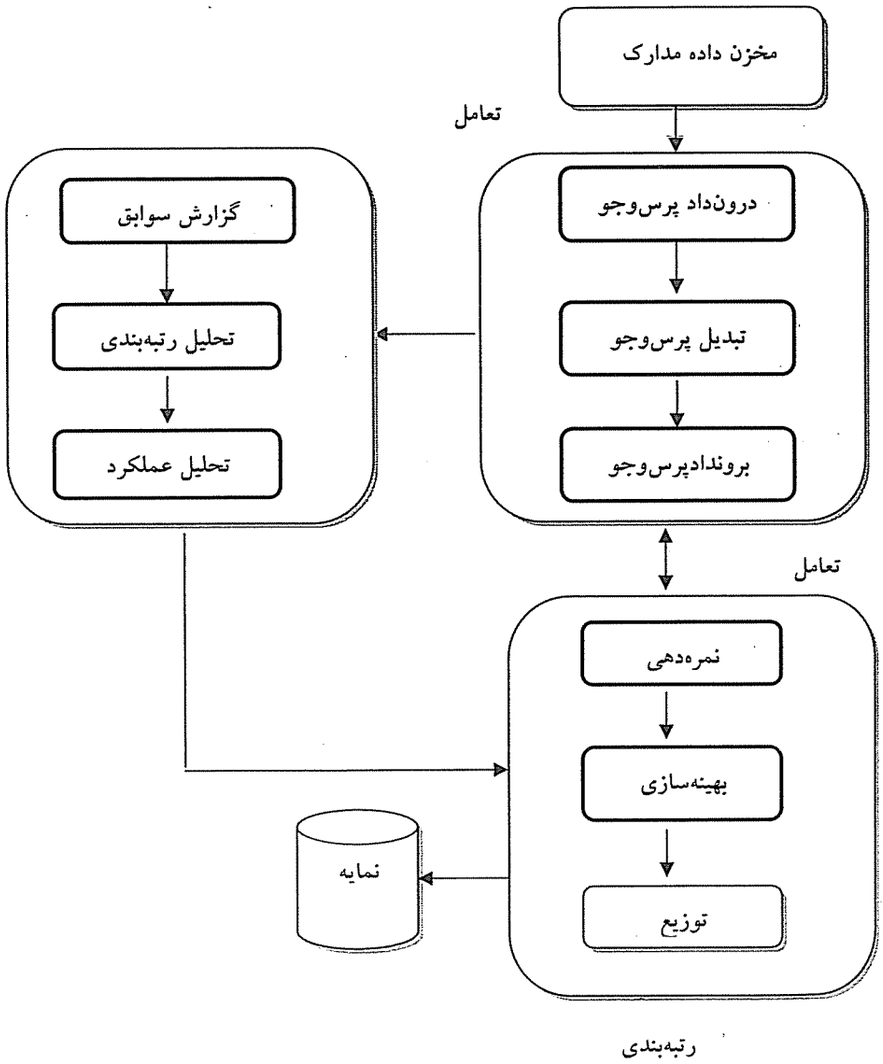
اگر تحلیل رتبه‌بندی، اثر بخشی موتور جستجو را بررسی می‌کند، تحلیل عملکرد، ارزیابی کارآیی است و در آن عملکرد کلی موتور جستجو به‌طور کامل تجزیه و تحلیل می‌شود.

طراحان و مهندسان برای به‌دست آوردن عملکرد بهتر، طرح‌های جدیدی ایجاد نمی‌کنند. بلکه شبیه‌سازی[1] می‌کنند و آزمایش‌ها را روی همان نمونه‌های اولیه انجام می‌دهند (کرفت و دیگران، ۲۰۰۸) موضوع شبیه‌سازی و مدل‌سازی از روی نمونه‌های موجود و ارزیابی عملکرد، مسأله بسیار مهمی در ارزیابی کارکرد نظام‌هاست.

بخش فرایند نمایه‌سازی موتور جستجو در شکل ۶ با جزییات نشان داده شده است.

1 simulation






شکل (۲-۷): بخش‌های اصلی فرآیند نمایه‌سازی در موتور جستجو

## خلاصه فصل

موتورهای جستجو از اساسی‌ترین ابزارهای جستجو و دسترسی به اطلاعات‌اند و به بخشی اساسی از زندگی روزمره افراد تبدیل شده‌اند. در این برنامه‌های نرم‌افزاری پیچیده، که گوگل یکی از مهم‌ترین آن‌هاست کارآیی و اثربخشی موضوع بسیار مهمی است. هر موتور جستجو از دو بخش اصلی تشکیل شده است؛ فرآیند نمایه‌سازی و پرس‌وجو. هر بخش، خود شامل بخش‌های فرعی مختلفی است. فرآیند پرس‌وجو شامل مجموعه‌سازی متنی، ایجاد نمایه و تبدیل متن است. نمایه‌سازی نیز شامل تعامل کاربر، رتبه‌بندی و ارزیابی است. هر کدام از این‌ها خود شامل بخش‌های فرعی متعددی هستند که یک موتور جستجو را به یک برنامه نرم‌افزاری پیچیده تبدیل می‌کند. الگوریتم‌های رتبه‌بندی هم در بخش نمایه‌سازی و هم پرس‌وجو بر کارکرد کلیه بخش‌ها تأثیر گذارند و می‌توانند عملکرد آن‌ها را نیز تحت‌الشعاع قرار دهند. کارآیی و اثربخشی موتورهای جستجو به کارکرد صحیح و درست تک‌تک این اجزا بستگی دارد. در حال حاضر کمتر طراحی نمونه‌های کاملاً جدید برای نظام‌های بازیابی و ارزیابی در دستور کار است و مهندسان بیش‌تر با استفاده از شبیه‌سازی و آزمایش روی نمونه‌های اولیه، موتورهای جستجو را ارزیابی می‌کنند و آن‌ها را ارتقا می‌دهند.

# فصل سوم

# مدل‌های بازیابی اطلاعات

## مقدمه

نیاز اطلاعاتی کابران بیش از آن که به‌صورت یک نیاز واقعی بیان شود یا حتی در قالب کلمات نمود عینی پیدا کند، در بسیاری از مواقع در حد یک ابهام در ذهن باقی می‌ماند. به دلایل زیادی افراد نیاز اطلاعاتی خود را بیان نمی‌کنند. شاید یکی از دلایل مهم این است که نیاز اطلاعاتی کاربر در آن لحظه یا بعد از آن اهمیت حیاتی ندارد یا نیاز آن‌قدر قوی نیست که کاربر زحمت اطلاع‌جویی به خود بدهد. گاهی کاربر از اهمیت رفع نیاز و ارزش افزوده آن آگاه نیست. گاهی حین تعامل با دیگران، خواندن یک منبع علمی یا تجربه در محیط‌های کاری و یا علمی و آکادمیک، سوالی علمی برای فرد ایجاد می‌شود که دنبال کردن آن، می‌تواند دانش‌افزایی خوبی داشته باشد و حتی مقدمه یک پژوهش خوب و ارائه یک مقاله مفید شود. در بسیاری موارد، دلایل مختلفی مانند مشغله کاری، عدم کنجکاوی و نداشتن روحیه جستجوگری، باعث می‌شود فرد از اطلاع‌جویی پرهیز کند. نبود منابع برای تأمین تقاضای مورد نظر کاربر، پیچیدگی منابع موجود برای اطلاع‌جویی و مواردی از این قبیل، نیاز اطلاعاتی را در حد یک ابهام یا یک پرسش بی‌پاسخ در ذهن فرد می‌گذارد. اما یکی از مهم‌ترین دلایل آن است که کاربر نمی‌تواند ابهام یا پرسش خود را به‌صورت قابل فهم بیان کند. یکی از مسائل مهم اطلاع‌جویی، مهارت فرد در