

حریری (۱۳۹۰)، یوسفی (۱۳۸۸)، بیزا-یتس^۱ و ریبیرو-نتو^۲ (۱۹۹۹) ترجمه قاسمی قاسمی در سال ۱۳۸۵ است که هر کدام ویژگی‌های خاص خود را دارند. کتاب حاضر نیز با توجه به اهمیت موضوع بازایی 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 بازایی 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 بازایی اطلاعات بررسی گردد. در فصل ۱، تاریخچه و تعاریف نظام‌های بازایی ارائه شده است. در فصل ۲، به طراحی نظام‌های بازایی اطلاعات با تأکید بر موتورهای جستجو توجه شده است. این تأکید از آن جهت بود که موتورهای جستجو یکی از پرکاربردترین و مهم‌ترین ابزارهای بازایی اطلاعات هستند. از طرف دیگر، در هر نظام بازایی نیز موتور جستجو مهم‌ترین بخش است. مبحث فصل ۳، مدل‌های بازایی اطلاعات است. در این فصل سعی شده است مدل‌های بازایی اطلاعات بر اساس یک نظام مشخص بررسی شوند. در فصل ۴، موضوع "رابط" از بنیادی‌ترین مفاهیم حوزه بازایی 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 بازایی 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.

فصل ۱

تعاریف و تاریخچه

نظام بازایی اطلاعات شامل دو مفهوم کلیدی ذخیره و بازایی است. نقطه مشترک تمام تعاریف، این دو مفهوم اصلی است. در فرهنگ لغت مریام وبستر^۱ (۲۰۱۴) برای عبارت نظام بازایی اطلاعات تعریفی وجود ندارد؛ اما برای بازایی اطلاعات دو تعریف آمده که در هر دو به ارتباط آن با یک نظام رایانه‌ای اشاره شده است.

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

تعریف دوم جامع‌تر است: "یافتن 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 بازایی 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 بازایی اطلاعات، به‌طور نمونه با تحولات موتورهای جستجوی اینترنتی مدرن، زبان طبیعی، فرامتن و

بازیابی اطلاعات را پژوهشگران زیادی تعریف کرده‌اند. از نظر کرسیانی و پاسی^۱ (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 آن که اغلب مجموعه‌های مدارک در نظام‌های بازیابی اطلاعات باید به محتوای چندین هزار و گاهی میلیون‌ها مدرک رسیدگی کنند.

مطابق تعریف کوالسکی^۲ (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، به جز 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^۳ (۲۰۰۶) در کتاب "مفاهیم هسته در کتابداری و اطلاع‌رسانی" تعاریف بازیابی اطلاعات را جمع‌آوری کرده که به پژوهشگران به نام حوزه بازیابی اطلاعات مربوط می‌شود. به‌طورمثال در این منبع تعاریفی از موئرز (۱۹۵۱)، ون‌ریجسیرگن (۱۹۷۹)، کمپ^۴ (۱۹۸۸)، میکسا^۵ (۱۹۹۹)، وارنر^۶ (۲۰۰۲)، بیزا-یاتز و ریبیرونتو (۱۹۹۹)، راثون^۷ (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

جستجوی کلیدواژه ترکیب شده است. پژوهشگران درگیر با هوش مصنوعی فنون دیگری را که به دنبال دقت بالای بازیابی اطلاعات‌اند مطالعه کرده‌اند.

پاتل^۱ (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^۲ 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" به کار رفته است. تفاوت تعریف فرهنگ لغت تخصصی علوم و فن‌آوری با دیگر فرهنگ‌ها استفاده از عبارت "تفسیر اطلاعات" است.

فرهنگ لغت کالینز^۳ (۲۰۱۴) به عنوان پیشگام در انتشار فرهنگ لغت هم، عبارت "recover" را برای تعریف نظام بازیابی اطلاعات به کار برده است: "نظامی برای یافتن اطلاعات خاص از داده‌های ذخیره شده".

در فرهنگ لغت آکسفورد^۴ (۲۰۱۴) نیز همین اصطلاح برای تعریف بازیابی اطلاعات وجود دارد. اما تفاوت در تعریف آکسفورد استفاده از کلمه ردیابی^۵ است: "ردیابی و یافتن اطلاعات خاص از داده‌های ذخیره شده".

برای یک تعریف استاندارد می‌توان به تعریف ایزو^۶ اشاره کرد. شمس‌اژه‌ای و امیدفر (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 ساخت یافته 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، پیوندهنده یک مفهوم، در مقابل یک رشته کاراکتری (واژه) است."

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

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

در یک یا چند زمینه موضوعی و برای فراهم کردن اطلاعات به وجود آمده است تا به محض درخواست کاربر، اطلاعات را در اختیار او قرار دهد" (جیج و اونوچکوا، ۲۰۱۱). تعریف جدیدتر پاتل در سال ۲۰۱۳ نشان می دهد نوع نگاه به بازیابی اطلاعات تغییر کرده است و پیشرفت های اخیر در حوزه پایگاه های اطلاعاتی بر تعریف آن ها موثر بوده اند:

"بازیابی اطلاعات، علم جستجوی اطلاعات در مدارک، جستجو برای خود مدارک، فراداده ها که توصیف کننده مدارک اند یا جستجو در پایگاه های اطلاعاتی، چه از نوع داده ای رابطه ای مستقل چه از نوع اطلاعاتی شبکه فرامتن مانند اینترنت یا شبکه جهانی وب یا اینترنت، برای متن، صدا، تصاویر یا داده هاست" (پاتل، ۲۰۱۳). همه این تعاریف نشان می دهد "بازیابی اطلاعات" مفهوم گسترده ای است و می توان از دیدگاه های مختلف آن را تعریف کرد و این تعریف ها بسیار تحت تأثیر زمان و پیشرفت فن آوری است. بدین جهت ارائه تعریفی جامع کار دشواری است؛ اما با توجه به مفاهیم کلیدی در تعاریف، می توان نظام بازیابی اطلاعات را این گونه تعریف کرد:

"نظام شناسایی، ذخیره، سازماندهی، بازنمایی، جستجو و ارائه اطلاعات در یک نظام رایانه ای است. اطلاعات به صورت غیرساختاریافته می تواند به هر شکل مانند متن، صدا، تصویر، گفتار، جدول یا نمودار باشد. نظام بر محور پرس و جوی اطلاعات کاربر و ربط اطلاعات با نیاز اطلاعاتی اوست و به همین دلیل فنون بازیابی اطلاعات مانند تحلیل و فرمول بندی پرسش، راهبردهای جستجو، تحلیل ربط و ارزیابی بروندهای اطلاعات در آن اهمیت بسیاری دارد. اطلاعات ضمیمه شده به این اطلاعات مانند فرامتن و فرارسانه یا اطلاعات توصیف کننده آن ها مانند فراداده را نیز می توان 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 فهرست های کتابخانه های که از نظر ماهیت مربوط به حوزه سازماندهی اطلاعات بوده است در واقع برای افزایش نقاط دسترسی به اطلاعات تدوین شده اند. رشد منابع اطلاعاتی سبب شده تا اهمیت بازیابی و دسترسی به اطلاعات بیش تر از قبل شود. ظهور وب و شخصی سازی تولید اطلاعات یعنی فراگیرشدن تولید اطلاعات در میان عموم مردم

پیش از این، شناسایی شده بوده و اکنون دوباره یافت شده است. از نظر او، این دیدگاه شاید در زمانی که اطلاعات در نظام های بسته برای اهداف خاص ارائه می شوند درست باشد اما وقتی که اقلام^۱ اطلاعاتی به صورت غیرمترقبه^۲ یا با جستجوی متنی بازیابی می شوند مشکل ساز است. در نظام بازیابی، اطلاعات یافت یا شناسایی می شوند و لزوماً بازیابی نمی شوند. فرهنگ های لغت مختلف نیز از اصطلاح recover برای تعریف بازیابی استفاده کرده اند که به جز معنی یافتن، معانی بازیافتن و دوباره به دست آوردن را دارد که خود مفهوم بازیابی را به چالش می کشد.

لنکستر^۳ (۱۹۶۸) در تعریف خود اعلام کرد که نظام بازیابی اطلاعات نمی تواند نمی تواند در دانش جستجوگر اطلاعات، تغییری به وجود 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 می کند" (نقل در جیج و اونوچکوا، ۲۰۱۱).

سالتون^۵ (۱۹۸۳) تعریف کاملاً تخصصی از نظام بازیابی اطلاعات ارائه کرده است: "نظام بازیابی اطلاعات، نظامی است که برای ذخیره اقلام اطلاعاتی استفاده می شود. این اقلام باید پردازش و جستجو شوند و در جوامع کاربری مختلف اشاعه یابند (پاتل، ۲۰۱۳). در کنار این تعریف، در کتاب مقدمه ای 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 مجموعه های بزرگ (معمولاً ذخیره در رایانه ها) برآورده می کند (منینگ، رغاوان و شوتز^۶، 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 جمع آوری و سازماندهی اطلاعات

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

بتوان به گونه ای این دسته بندی زمانی را پذیرفت، اما تقسیم تحول نظام های بازیابی اطلاعات بر اساس دوره های زندگی انسان ممکن است وسعت دید را محدود کند. با توجه به تحولات فراوان فن آوری و پیشرفت های عمده بعد از سال ۲۰۱۰ باید پذیرفت نظام های بازیابی اطلاعات همیشه نسبت به دوره بعدی خود، دارای کاستی های فراوانند. پیشرفت فن آوری به شدت بر این نظام ها تأثیر می گذارد و روند توسعه و تکامل آنها به صورت مداوم ادامه دارد.

در ادامه مهم ترین اتفاقات مربوط به نظام های بازیابی اطلاعات آمده است.

از ابتدا تا دهه ۱۹۴۰

اطلاعات چندانی درباره نظام های بازیابی اطلاعات قبل از دهه ۱۹۴۰ وجود ندارد. نظام های بازیابی اطلاعات قبل از این دهه، دستی و بسیار مبتنی بر تکامل فهرست های کتابخانه های بوده اند. در واقع، ریشه های نظام بازیابی اطلاعات به اولین فهرست های کتابخانه های برمی گردد. تفکر اولیه آن به کتابداری و حوزه علم اطلاعات مربوط است (ویر^۲، ۲۰۱۰، ص ۱۲). برخی معتقدند که برای اولین بار بازیابی اطلاعات در چهار هزار سال پیش همزمان با فهرست های اولیه کتابخانه ها 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^۳، ۲۰۰۲). ثبت اطلاعات به صورت مکتوب چالش دسترسی به اطلاعات را برای بشر ایجاد کرد. یکی از دلایل ایجاد اولین فهرست های کتابخانه ای برای افزایش نقاط دسترسی به منابع بوده است.

ایده نخست در حوزه بازیابی اطلاعات در طبیعت خود، اصولاً فلسفی است و به چگونگی رده بندی و سازماندهی اطلاعات مربوط است (ویر^۲، ۲۰۱۰، ص ۱۲) و فهرست های به وجود آمده فهرستی از عناوین کتاب ها یا طبقه بندی گسترده ای از موضوعات بودند. قبل از اختراع کاغذ، رومی ها و یونانیان باستان، اطلاعات خود را روی طومارهای پاپیروس ثبت می کردند. رومیان برای هر پاپیروس، یک برچسب داشتند که شامل خلاصه ای از محتوای پاپیروس بود و در بسیاری از موارد، باعث می شد که نیازی به بازکردن یک پاپیروس بزرگ نباشد. در مدارک یونانی قرن پنجم

باعث شده است تا حوزه بازیابی اطلاعات از یک موضوع مهم به یک چالش بزرگ تبدیل شود. جان نایبست معتقد است "ما در اطلاعات غوطه وریم، اما از کمبود دانش رنج می بریم" (دایتینگر، گات، مائورر و پیوک^۱، ۱۹۹۹ نقل در حسینی، ۱۳۹۰، ص ۶۲۶). شاید به همین دلیل دایتینگر و دیگران، وجود مخازن دانش بسیار بزرگ و بدون ساختار را به همراه افزایش سریع اطلاعات، دو ویژگی جوامع کنونی می دانند (حسینی، ۱۳۹۰). هر چند هدف واقعی "انقلاب ناشی از ظهور وب و تأثیر آن بر ذخیره و دسترسی به اطلاعات، تنها دسترسی پذیری اطلاعات نیست بلکه کارآمدسازی دسترسی به اطلاعات است" (خسروی و فتاحی، ۱۳۸۹، ص ۲۲۰) با این حال، این انقلاب سبب شد تا بازیابی اطلاعات و مسائل مربوط به نظام های بازیابی اهمیت بیشتری یابد. برای درک بهتر این موضوع و چالش های آن، دانستن تاریخچه اجمالی تکامل نظام های بازیابی ضروری است. در این بخش سعی شده است روند تاریخی بازیابی اطلاعات و نظام های بازیابی به صورت کلی و در چند دوره مهم تاریخی بررسی شود.

در بسیاری از متون، تاریخچه نظام های بازیابی اطلاعات از دهه ۱۹۵۰ به بعد بررسی شده است (ساندرسون و کرافت^۲، ۲۰۱۲؛ لسک^۳، ۱۹۹۶؛ دویین و اسمیت^۴، ۱۹۹۹). به دلیل آن که نظام های بازیابی اطلاعات در دوره های قبل، دستی بودند؛ یعنی نمایه ها و فهرست ها به شکل چاپی و کارتی تهیه می شدند (بهمن آبادی، 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 تقسیم بندی آن است که لسک در سال ۱۹۹۶، دوره بعد از سال ۲۰۱۰ را نیز پیش بینی کرده بود. این هفت دوره شامل دوره کودکی^۵ (۱۹۴۵-۱۹۵۵)، دانش آموزی^۶ (دهه ۱۹۶۰)، بزرگسالی^۷ (دهه ۱۹۷۰)، بلوغ^۸ (دهه ۱۹۸۰)، بحران میانسالی^۹ (دهه ۱۹۹۰)، انجام^{۱۰} انجام^{۱۱} (سال ۲۰۰۰ تا ۲۰۱۰) و بازنشستگی^۱ (بعد از سال ۲۰۱۰) است. شاید بتوان

1 Diettinger, 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

1 Retierment

2 Webber

3 Fielden

قبل از میلاد نیز، چکیده‌هایی شبیه به این کشف شده است. با این حال هیچ طرح رده‌بندی واقعی از کتابخانه‌های روم و یونان در آن زمان برجا نمانده است (لانگویل و میر،^۱ ۲۰۰۶). الیوت و روز^۲ (۲۰۰۹) معتقدند اولین فهرست کتابخانه‌ای را شاعر یونان باستان، کالیماکس به وجود آورد (ساندرسون و کرافت، ۲۰۱۲) که هدف آن جستجوی ساده‌تر اطلاعات بود. فهرست مطالب به عنوان یکی از ابزارهای اولیه بازایی اطلاعات در کتیبه‌های یونانی به چشم می‌خورد. کتابخانه اسکندریه، ۲۸۰ سال قبل از میلاد، به عنوان کتابخانه بزرگ عصر خود از پایروس استفاده می‌کرد و حدود هفتصد هزار مدرک را در اختیار داشت. اما پیدایش کتاب، در قرن‌های بعدی و زمان تحریم واردات پایروس 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 پایروس، به صورت طومار درآورند. به همین خاطر آن‌ها را به شکل صفحه درآوردند و به صورت کتاب به هم متصل کردند و کتاب را به وجود 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 به صورت 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 بازایی و دسترسی به اطلاعات پایان یافت. رنسانس تحولات عمده‌ای را به همراه 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 به وجود آورد و جهان انقلابی بزرگ را در انتشار سریع اطلاعات تجربه کرد. این اختراع و افزایش سرعت

انتشار به معنای افزایش انتشارات مکتوب با کیفیت بالاتر بود و این موضوع همراه با اختراعاتی مانند کاغذ و چاپ مطبوعات، نیاز به بازایی 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 بازایی خودکار مقداری زیادی از اطلاعات استفاده کنند (سینال،^۱ ۲۰۰۱). بدین ترتیب، 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)، بازایی به صورت دستی بود و پیشرفت‌ها نیز در حوزه بازایی اطلاعات نسبت به سال‌های بعد از آن کند بود و همان‌طور سنتی باقی ماند.

رده‌بندی دیویی، رده‌بندی کتابخانه کنگره، و مستندسازی^۲ از پیشرفت‌های مهم قبل از ۱۹۴۰ محسوب می‌شوند. مسأله مهم دیگر در سال‌های ۱۹۲۰ تا ۱۹۴۰، توزیع اطلاعاتی و کتابشناختی شامل ارائه قوانین زیف^۳، پارتو^۴ و لوتکا^۵ بود. هم‌چنین محمل اطلاعاتی به نام میکروفیلم به‌طور گسترده‌ای استفاده شد (داده کاوی: یادگیری ماشینی^۶، ۲۰۰۶) که تأثیرات زیادی بر حوزه بازایی 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 کارایی 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 این که در این دهه، اولین رایانه الکترونیکی به وجود آمد (فیلدن، ۲۰۰۲) و از همان ابتدا برای سازماندهی اطلاعات استفاده گردید.

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

1 Langville and Mayer
2 Eliot and Rose

در سال ۱۹۴۵، وانور بوش^۱ در مقاله پیشگامانه خود با عنوان "همان طور که ممکن است فکر کنیم" ایده دستیابی خودکار به مقدار زیادی از اطلاعات را مطرح کرد که در دهه ۱۹۵۰، با توصیفات واقعی تری از چگونگی جستجوی خودکار اطلاعات در آرشیوهای متنی محقق شد (سینال، ۲۰۰۱). بوش ماشینی به نام "ممکس"^۲ ترسیم کرد که می شد کتاب ها، پیشینه ها و ارتباطات را با آن ذخیره و مکانیزه کرد و از اطلاعات ذخیره شده با سرعت و انعطاف بیش تری 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. مهم ترین آن ها روش تأثیرگذار نمایه سازی لون^۳ بود (سینال، ۲۰۰۱). لون نخستین کسی بود که معتقد بود می توان واژه های معینی را به طور خودکار از متن استخراج کرد و با آن ها محتوای متن را بازنمایی نمود (جلالی منش، علیدوستی، خسرو جردی، ۱۳۹۲). نمایه سازی یکی از ابزارهای مهم در بازنمون و بازایی اطلاعات در نظام های بازایی است. روش نمایه سازی لون، در بعد از دوران خود کاربرد زیادی در نمایه سازی و چکیده نویسی داشت.

تحولات دیگر این دهه، اتفاقات بعد از جنگ جهانی دوم بود. ارتش امریکا با مشکلات نمایه سازی و بازایی مدارک پژوهشی علمی که زمان جنگ از آلمان ها گرفته بود مواجه شد (دوبین واسمیت، ۱۹۹۹) و به دنبال راهی برای رهایی از این مشکلات بود. به طور کلی، بعد از اتمام جنگ افزایش انتشارات مسأله بازایی 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^۴ ماشینی به نام "یونیواک"^۵ را توصیف کرد. یونیواک، می توانست منابع متنی مرتبط را با یک کد موضوعی جستجو کند. کدها و متن ها روی نوارهای مغناطیسی ذخیره شده بودند. ماشین می توانست ۱۲۰ کلمه را در دقیقه پردازش کند. بدین ترتیب اولین منبع برای استفاده از رایانه برای جستجوی

محتوا به وجود آمد (ساندرسون و کرافت، ۲۰۱۲). تحولات بیش تر در دهه بعد پیگیری شد.

دهه ۱۹۵۰

کلون موئرز اصطلاح "بازایی اطلاعات" را خلق کرد (دوبین واسمیت، ۱۹۹۹) و بدین ترتیب با استفاده از نظام های داده پردازی برگه منگنه، نمایه سازی پس همارا به وجود آمد (بهمن آبادی، ۱۳۸۶). اتفاق مهم دیگر آن بود که شوروی ماهواره اسپوتنیک را با موفقیت به فضا فرستاد که سبب ترس و وحشت آمریکا شد. جامعه آمریکا فکر می کرد که در علم عقب مانده است (لسک، ۱۹۹۶). هم چنین آمریکاییان، نگران حمله موشک های دوربرد اتمی شوروی بودند. آن ها می دانستند که شوروی می تواند به شبکه های ارتباطی آن ها حمله کند. افزایش نگرانی ها بخاطر شکاف اطلاعاتی به وجود آمده، آن ها را تشویق کرد تا بودجه هایی برای پژوهش در این زمینه اختصاص دهند و بدین ترتیب پشت صحنه نظام های 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 (اسلاتر^۱، ۲۰۰۲). از طرفی بسیاری از اولین نظام هایی که به گونه ای در بازایی اطلاعات نقش داشتند در دهه ۱۹۵۰ ساخته شدند که شامل نمایه های کوئیک^۲ و کشف اللغات^۳ مورد استفاده در نظام های بازایی اطلاعات بودند (لسک، ۱۹۹۶).

در سال ۱۹۵۶ یوجین گارفیلد کار خود را با پایه گذاری شرکتی کوچک آغاز کرد. در ااتاقک کوچک او تهیه سلف فهرست مندرجات جاری امروزی آغاز شد (بهمن آبادی، ۱۳۸۶). او با این کار یکی از تأثیرگذارترین تحولات حوزه بازایی اطلاعات را یعنی نمایه استنادی علوم پایه گذاری کرد.

دو سال بعد آلن کنت^۴ "انتخاب گر جستجوی دانشگاه وسترن رزرو"^۵ را ساخت (لسک، ۱۹۹۶). از کاربردهای آن، امکان جستجوهای آزمون روی فایل چکیده های کدگذاری شده انجمن فلزات آمریکا^۶ بود (ریز و کنت^۷، 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

1 Vannevar Bush

2 Memex

3 Lohn

4 Holmstrom

5 Univac

در این سالها، میتچل^۱ با استفاده از رایانه یونیواک توانست یک میلیون چکیده را که با شش کد موضوعی نمایه سازی شده بودند جستجو کند. نانوس^۲ نیز، نظامی از جنرال الکتریک را توصیف کرد که می توانست ۳۰۰۰۰۰ چکیده مدرک را جستجو کند (ساندرسون و کرافت، ۲۰۱۲).

دهه ۱۹۶۰

دهه ای تأثیر گذار در بازیابی اطلاعات است و تقریباً در هر سال، اتفاقات مهمی در حوزه بازیابی اطلاعات افتاده است. سال های تجربه های بزرگ بود و در حقیقت دهه جهش و جنبش سریع برای بازیابی اطلاعات شناخته شده است و افراد بسیاری در کنفرانس های بازیابی اطلاعات شرکت کردند (لسک، ۱۹۹۶). دوران بازیابی رایانه ای به شیوه گسسته، پردازش دسته ای و نواری بود و در اوایل همین دهه اولین نظام های بازیابی اطلاعات بزرگ در امریکا در دولت فدرال ایجاد شد. اما فدا کردن یک نظام رایانه ای برای جستجوی گذشته نگر پیشینه های کتاب شناختی توجیه اقتصادی نداشت و هنوز هم ندارد. بسیاری از نظام های رایانه ای این دوره و بعد از آن چند منظوره بودند و طیف وسیعی از محصولات و خدمات را از یک عملیات ورودی منحصربه فرد تولید می کردند. این نظام ها، نتیجه فرآیند خودکارسازی نشر بودند (بلاچ، ۱۹۶۳).

در همین سالها، تحولات بیش تر در حوزه علوم و استاد های علمی به وجود آمد. مؤسسه کوچک یوجین گارفیلد (۱۹۶۰ و ۱۹۶۱) نام کنونی اش یعنی "مؤسسه اطلاعات علمی (آی.اس.آی) را یافت^۳. از طرفی مهم ترین واقعه سال ۱۹۶۱ نشر نخستین ویرایش "نمایه نامه استادی علوم"^۴ برای حوزه ژنتیک بود (نشاط، ۱۳۸۷).

از دیگر تحولات این دهه، روش های متفاوت نمایه سازی بود. از جمله نمایه سازی احتمالی که در این دهه ایجاد شد (بری^۵، ۲۰۰۹ الف). مارون و کان در سال ۱۹۶۰، مقاله ای با عنوان "رابطه، نمایه سازی احتمالاتی و بازیابی اطلاعات" در مجله،

ای سی ام^۱ منتشر کردند.

در این سالها جرالد سالتون یکی از بزرگان حوزه بازیابی اطلاعات و نویسنده کتاب مشهور "درآمدی بر بازیابی اطلاعات"^۲ (سالتون و مک گیل^۳، ۱۹۸۶)، در هاروارد و بعدها در کورنل، کار خود را در این حوزه شروع کرد (دوبین و اسمیت، ۱۹۹۹).

در سال ۱۹۶۲، کلوردون^۴ اولین یافته های کرانفیلد را منتشر کرد که مدل پیشرفته برای ارزیابی نظام های بازیابی بود (دوبین و اسمیت، ۱۹۹۹). بسیاری از افراد در آغاز عصر مدرن بازیابی اطلاعات به تجربه های نمایه سازی کرانفیلد در دهه ۱۹۶۰، استناد کردند (کلوردون، میلز^۵ و کین^۶، ۱۹۶۶). آزمایش های کلوردون در این دهه یکی از تأثیر گذارترین پژوهش ها در حوزه ارزیابی نظام های بازیابی اطلاعات بود. البته آزمایش های اولیه در سال ۱۹۵۸ انجام شد که در آن با استفاده از ۱۸۰۰۰ مدرک و ۱۲۰۰ موضوع، نظام های نمایه سازی همارا با نظام های سنتی مقایسه گردید و نتایج به دست آمده مبنای تنظیم دومین فرضیه های پژوهش قرار گرفت. نتایج این پژوهش، نه تنها الگویی برای بسیاری از پژوهش های ارزشیابی، عملیاتی و تجربی شد؛ بلکه بعدها در متون درسی این رشته نیز اشاعه یافت (زوارقی، ۱۳۹۰). منشا مدل های آزمایشگاهی ارزیابی نظام های بازیابی اطلاعات، به همین مدل ارزیابی کلوردون دوم برمی گردد. این انگاره علوم رایانه در حوزه بازیابی اطلاعات، به دنبال توسعه هر چه بهتر الگوریتم ها و نظام های بازیابی اطلاعات است (ککلاین و جاولین، ۲۰۰۲).

در اواسط این دهه، کتابخانه پزشکی آمریکا، نظام بازیابی و تحلیل متون پزشکی یا همان مدلاز^۷ را توسعه داد. این نظام، اولین نظام بازیابی دسته ای^۸ و اولین پایگاه پایگاه اطلاعاتی بزرگ ماشین خوان بود (دوبین و اسمیت، ۱۹۹۹). مدلاز تحول عمده ای در حوزه دستیابی به اطلاعات پزشکی به وجود آورد که هنوز هم حوزه پزشکی از فواید این تأثیرات بهره می برد. ایجاد نظام مدلاین بر روی دیسک و بعدها به صورت بر خط بسیار مدیون مدلاز 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

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

مطالعات ارزیابی نظام ارزیابی مدلاز را کامل کرد و اولین ویرایش متنی آن را روی نظام بازایی اطلاعات منتشر ساخت (دوبین و اسمیت، ۱۹۹۹). در سال ۱۹۶۶، مطالعه آزمایشی بر روی پروژه مارک^۱ آغاز شد (اورم^۲، ۱۹۷۵) که بر بازایی اطلاعات تأثیر چشمگیری گذاشت. ایجاد فهرستهای ماشین خوان یکی از تحولات مهم بازایی اطلاعات است که مقدمات آن در این دهه پایه ریزی شد و بعد از پیاده سازی به جز لاینفک ساختار سازماندهی کتابخانه ها 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، به وجود آمد در واقع یک فرآیند خودکار و کنترل شده برای فرمول بندی مجدد پرسش بود (سالتون و باکلی^۳، ۱۹۹۷) و به این دلیل مهم است که هنوز هم کاربران با ساختار نظام های بازایی اطلاعات آشنا نیستند و فرمول بندی پرسش به صورت دقیق برای آنها مشکل است. با بازخورد ربط در نظام های بازایی، کاربر می تواند پرسش را دوباره فرمول بندی کند تا اطلاعات بهتری به دست آورد.

یکی از تأثیرگذارترین پیشرفت ها که بر همه حوزه های دانش بشری تأثیر گذاشت و نظام های بازایی اطلاعات سود زیادی از آن برده اند اینترنت است که در این دهه پایه گذاری شد. پروژه آرپانت ابتدا با اهداف سیاسی شکل گرفت ولی مقدمه شکل گرفتن اینترنت جهانی در سال ۱۹۶۸ شد. همان طور که قبلاً نیز گفته شد، اقدام اتحاد جماهیر شوروی در پرتاب موفقیت آمیز ماهواره اسپوتنیک به فضا در دوران جنگ سرد، آمریکائیان را نگران حمله اتمی آنها کرد. آنها هراس داشتند که با پیشرفت علم در شوروی، موشک های دوربرد به شبکه های ارتباطی آنها حمله کند. به همین دلیل در پی آن بودند شبکه ای ایجاد کنند که چند کاربر بتوانند همزمان از آن استفاده کنند و امکان انتقال اطلاعات در آن فراهم باشد. بدین ترتیب آرانس پروژه های تحقیقاتی پیشرفته پنتاگون، تصمیم گرفت بر روی چنین پروژه ای تحقیق کند. در پاییز ۱۹۶۹ اولین نوع از گره های شبکه ای در یوسی ال ای^۴ شکل گرفت. در دسامبر همان سال، چهار رایانه در شبکه اولیه نصب شدند و بعد از مدت کوتاهی توانستند داده ها را در

بین خود انتقال دهند (استرلینگ^۱، ۱۹۹۳). بدین ترتیب اینترنت اولیه شکل گرفت. به طور کلی دهه ۱۹۶۰ را باید دهه نظام های برون خطی^۲ دانست. ویژگی های نظام های بازایی اطلاعات بسیار شبیه به هم بودند. نظام های بازایی پردازش دسته ای^۳ برون خطی که از نوارهای مغناطیسی به عنوان رسانه ذخیره سازی استفاده می کردند (بلاچ، ۱۹۶۳).

دهه ۱۹۷۰

در این دهه، با توسعه سخت افزارها و نرم افزارهای رایانه ای (بهمن آبادی، ۱۳۸۶) و ارتباط از راه دور (بلاچ، ۱۹۶۳)، امکان جستجوهای پیوسته یا تعاملی فراهم آمد (بهمن آبادی، ۱۳۸۶). نظام های پردازش دسته ای برون خطی دهه ۱۹۶۰ با نظام های بازایی پیوسته تعاملی دهه ۷۰ و بعد از آن دنبال شدند (بلاچ، ۱۹۶۳). نظام های پیوسته بازایی، علاوه بر افزایش سرعت، امکان دریافت بازخورد در روند جست و جو و در صورت لزوم، تغییر و اصلاح آن را به استفاده کننده می دادند.

مهم ترین مزیت بازایی پیوسته، امکان دسترسی از راه دور بود (بهمن آبادی، ۱۳۸۶). رسانه ذخیره سازی اطلاعات در این نظام ها، دیسک های مغناطیسی بود (بلاچ، ۱۹۶۳) که امکان انعطاف بیش تری را نسبت به نوارهای مغناطیسی داشت. از اقدامات مهم حاصل از این پیشرفت ها، ایجاد پایگاه های اطلاعاتی درون خطی تجاری مانند مدلاین^۴، ای آی ام - تی دبیو ایکس از کتابخانه پزشکی آمریکا^۵، دایالوگ لا کهید^۶ و اوریت^۷ است. بدین ترتیب نظام های نمایه سازی و چکیده نویسی رشد چشمگیری یافتند (بری، ۲۰۰۹). یکی از مشکلات چکیده نویسی و نمایه سازی، تأخیر زمانی در انجام کارها به صورت دستی بود. فن آوری حروف چینی رایانه ای، بسیاری از کارها را سریع تر کرد (لسک، ۱۹۹۶).

علاوه بر گسترش محتواهای نمایه سازی شده رایانه ای در این سال ها، ایده های جدیدی نیز برای نمایه سازی محتوای مدارک ارائه شد. از این نوع ایده ها می توان به ایده نمایه سازی مارون، کان و ری^۸ 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

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

استفاده کردند. ون ریجسبرگن، اصول رتبه بندی احتمالی را تعریف کرد که نشان می داد چگونه با توجه به معیارهای ارزیابی تعریف شده می توان به صورت بهینه مدارک را رتبه بندی 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. آن ها نشان دادند که این مسأله درست نیست و کلمات مدرک به هم مربوط اند. در نظر گرفتن کلمات به صورت وابسته به یکدیگر در یک مدرک، موضوع پژوهش های سال های بعد شد (ساندرسون و کرافت، ۲۰۱۲).

تنها ویژگی این دوره افزایش خدمات نمایه سازی و چکیده نویسی نبود؛ بلکه توانست اولین نظام های بازایی تمام متن را نیز تجربه کند. یکی از پیشگامان این نظام ها، نظام لکسیس^۱ است. در دوم آوریل ۱۹۷۳، ام دی سی^۲ تنها با چهار شرکت، خدمت مرجع حقوقی لکسیس را راه اندازی کرد. در این نظام، یک وکیل می توانست متن کامل کدهای نیویورک، آهایو و کدهای فدرال هم چنین منابع یک کتابخانه مالیاتی فدرال را جستجو کند (تاریخچه گروه لکسیس - نکسیس^۳، 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 دسترس پذیری در آن ۹۹/۹۹٪ بود (بیزا - بیتس و ریبریوتو، ۲۰۱۰).

از دیگر فن آوری های کلیدی این دوره، دسترس پذیر شدن نظام های اشتراک زمانی بود. در این نظام ها، به جای این که پرسش ها با برخی از عملیات دسته ای پردازش شوند به طور مستقیم به یک پایانه داده می شوند و فوراً پاسخی برای آن دریافت می گردد. این فن آوری جستجو را عملی تر کرد و امکان ارائه خدمات آن به کتابداران به وجود آمد. هم چنین باعث شد قیمت هارد دیسک ها کاهش یابد (لسک، ۱۹۹۶). این موضوع، خود به افزایش انتشارات و پیشرفت های حوزه بازایی اطلاعات کمک شایانی کرد و هارد دیسک ها را به یک رسانه عمومی ذخیره سازی اطلاعات تبدیل نمود.

در سال ۱۹۷۱ یاردین و ون ریجسبرگن^۴ در مورد خوشه بندی سلسله مراتبی^۵ در بازایی اطلاعات صحبت کردند. آن ها راهبردهای بازایی اطلاعات را بر پایه

خوشه بندی سلسله مراتبی عرضه کردند و با استفاده از مجموعه کرافیلد آن را ارزیابی نمودند. آن ها معتقد بودند که این نوع راهبرد می تواند در مجموعه های بزرگ در حال رشد اثربخش باشد. در این دوران، سالتون تأثیرگذارترین مقاله های خود را در حوزه بازایی اطلاعات منتشر کرد. انتشارات وی در ارتباط با نظریه نمایه سازی، مدل فضا برداری در بازایی اطلاعات و اهمیت اصطلاح برای نمایه سازی متنی خودکار بود (دوبین و اسمیت، ۱۹۹۹). او با ارائه رهیافتی در مدل فضا برداری خود، فرآیند بازایی اطلاعات را در بسیاری از نظام های بازایی پژوهشی پایه ریزی کرد. شاید امروزه، فرمول های رتبه بندی سالتون کم تر استفاده شوند؛ اما کاربرد مدارک و پرسش ها به عنوان بردارهایی در فضای بزرگ و دارای بعد، هنوز هم معمول است (ساندرسون و کرافت، ۲۰۱۲).

یکی از پیشرفت های کلیدی حوزه بازایی اطلاعات در این دوره، وزن های بسامد اصطلاح^۱ لون (بر پایه رخداد یک واژه در میان یک مدرک) بود که با کار اسپارک جونز^۲ 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 بازایی، ارتباط معکوسی دارد. یعنی هر چه یک اصطلاح در مدرک کم تر تکرار شود، دارای مفهوم خاص تری است (ساندرسون و کرافت، ۲۰۱۲).

این دوره شاهد توسعه زیاد بخش فهرست های رایانه ای است. اسی ال سی به رکوردهای کتابخانه کنگره آمریکا نیز توسعه یافت (بری، ۲۰۰۹). البته باید گفت که بسیاری از این پژوهش ها در دهه قبل آغاز شده بود اما به ثمرنشتن و توسعه آن ها در دهه ۷۰ میلادی بوده است. به طور مثال پروژه اسی ال سی از سال ۱۹۶۷ کلید خورد اما توسعه آن به فهرست های کتابخانه کنگره در دهه ۷۰ بوده است. این نظام که کیلگور^۳ به وجود آورد از خروجی های مارک کتابخانه کنگره برای ایجاد فهرست نویسی ماشین خوان استفاده کرد. کار دیگر این نظام، ایده کار تعاونی بود و اولین نمونه از سازماندهی هوشمند تعاونی اطلاعات محسوب می شد که بوش در مقاله پیشگامانه خود در سال ۱۹۴۵ پیش بینی کرد (لسک، ۱۹۹۶). تئودور نلسون^۴ مفهوم فرامتن^۱ را در این دهه مطرح کرد (دوبین و اسمیت،

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

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

(۱۹۹۹) و برای متنی به کار برد که به آسانی و به شیوه غیرترتیبی^۲ قابل دسترس است (گال و هانافین^۳، ۱۹۹۴). او پیش بینی کرد که چگونه رایانه ها می توانند فرآیندهای نشر و توزیع را دگرگون کنند و معتقد بود که رایانه به سادگی نمی تواند بهره وری را افزایش دهد و باید چگونگی کار افراد با مدارک بازتعریف شود. در اواخر این دهه، اولین کنفرانس بازایی اطلاعات ای سی ام^۴ برگزار شد که نشان دهنده اهمیت موضوع برای پژوهشگران بود. کنفرانس هرساله به صورت تخصصی برگزار می شود و آخرین یافته ها و پیشرفت های حوزه بازایی اطلاعات در آن عرضه می شود (دوبین و اسمیت، ۱۹۹۹). در سال ۲۰۱۵ سی و هشتمین دوره آن در شیلی برگزار شد و یکی از معتبرترین کنفرانس های حوزه بازایی اطلاعات است.

دهه ۱۹۸۰

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

افزایش مداوم سرعت پردازش کلمات و کاهش پی در پی قیمت فضای دیسک موجب شد که اطلاعات بیش از پیش به شکل ماشین خوان قابل دسترس شود (لسک، ۱۹۹۶). با توجه به افزایش پردازش ماشینی و دیگر فن آوری ها، در این دهه بازایی پیوسته متن کامل رشد قابل توجهی داشت و بسیاری از مجلات و روزنامه ها به صورت پیوسته - البته فقط به صورت متن - قابل دسترسی بودند.

اولین کنفرانس بازایی اطلاعات ای سی ام مشترک با بخش بازایی اطلاعات در انجمن رایانه بریتانیا در سال ۱۹۸۰ برگزار شد (دوبین و اسمیت، ۱۹۹۹). ریاست این کنفرانس را محقق بزرگ حوزه بازایی اطلاعات، ون ریچسبرگن بر عهده داشت و حوزه های جدید اطلاعات در آن دوران بحث شدند. بررسی مقالات ارائه شده در ۱۹۸۰ نشان می دهد که عمده پژوهش ها، در حوزه وزن دهی اصطلاحات، مدل های احتمال و بازایی مفهومی بوده است (خلاصه مقالات سومین کنفرانس سالیانه ای سی ام روی پژوهش و توسعه بازایی اطلاعات^۵، ۱۹۸۰). در سال ۱۹۸۱، حتی مقاله ای در حوزه جستجوی فازی و مقاله دیگری در حوزه بازنامی فضایی دانش ارائه شده است

(کراوچ، کوپر و هر^۱، ۱۹۸۰).

هم چنین بلکین، اودی و بروک^۲ در سال ۱۹۸۲، دیدگاه اسک^۳ را برای بازایی 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 بازایی اطلاعات تعاملی را معرفی کردند که به نیازهای اطلاعاتی توجه داشت (بلکین، اودی و بروک، ۱۹۸۲).

در سال ۱۹۸۳، سالتون کتاب مهم خود "درآمدی بر بازایی اطلاعات" را با کمک انتشارات مک گرو هیل^۴ منتشر کرد. این کتاب بر مدل های بُرداری تأکید فراوانی داشت (دوبین و اسمیت، ۱۹۹۹) و در حوزه بازایی اطلاعات، اثر مهمی محسوب می شد و تا سال ۱۹۸۷ چندین ویرایش از آن منتشر گردید.

اواسط دهه ۱۹۸۰، تلاش ها برای ایجاد نسخه های کاربر نهایی^۵ از نظام های بازایی اطلاعات تجاری آغاز شد (دوبین و اسمیت، ۱۹۹۹). استفاده از نظام های برخط بازایی اطلاعات در این دوره از دو طریق امکان پذیر بود. اول استفاده از نظام های متن کامل و دیگری افزایش استفاده غیرمتخصصان از نظام های بازایی. به طور مثال کتابداران، کارت های فهرست نویسی را با فهرست های عمومی دسترسی پیوسته (اپک ها)^۶ جایگزین کردند. در پایان این دهه نیز فروشندگان تجاری، نرم افزارهای اپک مانند نوتیس^۷ را وارد بازار کردند (لسک، ۱۹۹۶) بدین ترتیب در این دوره استفاده از اپک ها رایج شد و کتابخانه ها توانستند اقتصادی تر و بهینه تر به فهرست های پیوسته دست یابند و نیازهای سازماندهی خود را نیز برطرف کنند. این فهرست ها نقاط جستجوی بیش تر در اختیار کاربران گذاشت. امروزه، اپک ها مهم ترین بخش کتابداری و اطلاع رسانی اند و در طول زمان بسیار پیشرفت کردند. نسل اول اپک ها حاوی اطلاعات اندک شناخته شده برای جستجو بود مثل عنوان، نویسنده و شماره کنترل. نسل دوم اپک ها با پیشرفت فن آوری، امکان جستجو با عملگرهای بولی، سرعنوان های موضوعی و کلیدواژه را نیز مهیا کرد و در نسل بعد بر معماری نظام های باز^۸ تمرکز کرد که در سال ها و دهه های 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

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

رابط کاربری گرافیکی و حمایت استاندارد فراداده هسته دوبلین، فرامتن، برنامه نویسی جاوا^۱ و مجموعه های نتایج رتبه بندی شده پیشرفت قابل ملاحظه ای یافت (بیزایاتار و ریبریو نتو، ۲۰۱۰).

در این دهه، استفاده از سی دی رام ها افزایش قابل ملاحظه ای یافت و توزیع اطلاعات را بیش از پیش آسان کرد. در پایان دهه، هر کتابخانه حداقل یک درایو سی دی رام داشت که با نشر سنتی سازگاری مناسبی یافتند و برای توزیع و استفاده اطلاعات به آسانی استفاده شدند. هم زمان با افزایش استفاده از این محمل جدید، شبکه سازی رایانه ای نیز در این دهه گسترش یافت (لسک، ۱۹۹۶).

تأثیرگذارترین اتفاق در اواخر این دهه آن بود که تیم برنرزی، وب جهان گستر را در آزمایشگاه کرن^۲ پیشنهاد کرد (دوبین و اسمیت، ۱۹۹۹) در سال ۱۹۸۹، این ایده؛ ذخیره، دسترسی و جستجوی مجموعه های مدارک را متحول کرد. مجموعه های مدارک مرتبط^۳ در این زمان، ایده بسیار جدیدی بود. هر چند که چنین طرحی را وانوور بوش (۱۹۴۵) تصویر کرده بود اما تحقق بخشی از طرح او در این سال ها امکان پذیر شد. ایجاد وب جهان گستر 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. کاستی های اینترنت اولیه در مورد رسانه های دیداری و شنیداری و ناتوانی در انتقال و دسترسی به این نوع اطلاعات سبب شد تا پروژه های مختلفی در این حوزه شروع به کار کند. انگلبرت^۴ 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 یو آر ال^۵ بود. این فن آوری به سایت و صفحه یک نشانی منحصر به فرد می داد. زبان نشانه گذاری فرامتن یا اچ تی ام ال^۶ حلقه اتفاقات را کامل کرد تا تیم برنرزی با کنار هم گذاشتن همه این ها بتواند نسخه اولیه وب را ایجاد کند. اولین نسخه در دسامبر ۱۹۹۰ راه اندازی شد و تحول عمده ای را در حوزه بازیابی اطلاعات به وجود آورد (تاریخچه وب جهان گستر^۷، 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

نظام های بازیابی اطلاعات با آزمایش و تجربه بر مجموعه های آزمون، پالایش و ارزیابی می شوند (وورهی و هارمن^۱، ۱۹۹۹). اما در این دهه و اوایل دهه ۱۹۹۰، اندازه مجموعه های مدارکی که برای آزمون و یا ارزیابی ها استفاده می شد نگران کننده بود. جامعه علمی معتقد بود که اندازه های مجموعه های آزمون مورد استفاده در مقایسه با مجموعه هایی که برخی کمپانی های بزرگ استفاده می کنند کوچک است. به همین دلیل دونا هارمن^۲ با همکاران خود، ترک یا کنفرانس بازیابی متن^۳ را به وجود به وجود آوردند. این کنفرانس، در واقع تمرین سالانه برای تعداد زیادی از گروه های پژوهشی بود تا با همکاری یکدیگر، مجموعه های آزمون بزرگ تری را نسبت به دوره قبل طراحی کنند (ساندرسون و کرافت، ۲۰۱۲) و بدین ترتیب توانستند مجموعه های بزرگ را بر اساس مجموعه های آزمون متناسب پالایش و ارزیابی نمایند. این کنفرانس هر سال برگزار و نتایج آن در دیگر کنفرانس های مهم بازیابی اطلاعات مانند ای سی ام استفاده می شود. در واقع، کار این کنفرانس تهیه مجموعه های آزمون استاندارد برای ارزیابی نظام های بازیابی اطلاعات است و در بسیاری از پژوهش های مربوط به ارزیابی این نظام ها استفاده می شود.

به طور کلی در این دهه، استفاده از اطلاعات پیوسته، معمول شد. هر چند که عموم مردم از آنها به صورت دائم استفاده نمی کردند (لسک، ۱۹۹۶). این دهه، شروع انقلابی تازه برای مردم دنیا در انتهای هزاره دوم بود. همه چیز به سرعت به طرف ماشینی شدن پیش رفت. اطلاعات با سرعت و سهولت بیش تر حداقل در اختیار متخصصان قرار می گرفت. کم کم زمینه ورود مردم عادی به عصر اطلاعات بیش از پیش فراهم شد و گام نهادن در مسیر توسعه علمی و آموزشی سریع تر گردید.

دهه ۱۹۹۰

آغاز یک دوره جدید در عصر اطلاعات بود. نسل اول وب به طور رسمی با پیشنهاد تیم برنرزی در سال ۱۹۸۹ آغاز به کار کرد. این پیشنهاد بیست سال بعد از خلق و راه اندازی اولین ارتباط^۴ به شکلی بود که امروزه اینترنت شناخته می شود. پیشنهاد او شامل فن آوری هایی بود که اینترنت را برای مردم دسترس پذیر و قابل استفاده می کرد. قبلاً نیز اشاره شد که سه فن آوری 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 نشانه گذاری فرامتن یا اچ تی ام ال: این زبان فرمت نشر برای وب است و پیوند دادن مدارک در وب را امکان پذیر می کند.
- شناساگر اختصاصی منبع یا یو آر آی^۱: نوعی نشانی که برای هر منبع در وب اختصاصی است.
- پروتکل انتقال فرامتن (اچ تی تی پی): با آن می توان صفحاتی را که به هم پیوند خورده اند بازیابی کرد (تاریخچه وب^۲، ۲۰۰۸).

با وجود پیدایش وب جهان گستر در سال ۱۹۹۰، تعداد وب سایت ها تا سال ۱۹۹۳ بسیار اندک بودند (ساندرسون و کرافت، ۲۰۱۲). در سال ۱۹۹۳ کرن به طور رسمی اعلام کرد که فن آوری وب، رایگان و برای همه افراد در دسترس است. تا سال قبل از این اعلام رسمی، فهرست نویسی 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، به تدریج موتورهای جستجوی پیش تری به وجود آمدند تا پاسخگوی افزایش وب سایت ها و اطلاعات باشند (ساندرسون و کرافت، ۲۰۱۲). بدین ترتیب فهرست نویسی دستی با چالش جدیدی روبه رو شد.

اولین موتور جستجو به نام آرچی^۳ در سال ۱۹۹۱ ایجاد شد. نظام نمایه سازی در موتور جستجوی آرچی فضای محدودی داشت و فقط می توانست فهرست های قابل داندلود را جستجو کند و قابلیت جستجوی محتوا را نداشت. تا پایان این دهه، بیست موتور جستجوی اطلاعات به وجود آمد که از مهم ترین آن ها می توان به گوگل^۴ و یاهو^۵ اشاره کرد (تاریخچه موتورهای جستجو^۶، ۲۰۱۳) که نشان دهنده افزایش چشمگیر اطلاعات بود. افزایش تصاعدی اطلاعات در این سال و سال های بعد نشان داد که باید برای سازماندهی و البته بازیابی اطلاعات برای کاربران بالقوه و بالفعل وب تدابیر جدیدی اندیشید. افزایش اطلاعات، خود چالش ها و مسائل جدیدی به همراه داشت. چالش های جدید، جوامع پژوهشی و تجاری را در حوزه بازیابی اطلاعات به تعامل واداشت و به آن ها فهماند که باید تلاش های خود را در حوزه بازیابی اطلاعات بیشتر کنند (ساندرسون و کرافت، ۲۰۱۲).

با توجه به مطالب پیش گفته، مهم ترین اتفاق در این دهه جستجوی تحت وب

است. موتورهای جستجوی اولیه از جستجوی فهرست ها به موتورهای جستجویی تبدیل شدند که می توانستند با امکانات پیشرفته تر، متن را نیز جستجو کنند. با وجود چندین برابر شدن وب سایت ها، باز هم مشکلات فراوانی وجود داشت. در این حوزه وجود یک مرورگر گرافیکی در نظام بازیابی بسیار مهم بود. اما تا سال ۱۹۹۲ هنوز مرورگر گرافیکی وجود نداشت. چهار دانشجوی یک کالج در فنلاند، مرورگر اروايز^۱ را در سال ۱۹۹۱ راه اندازی و در ۱۹۹۲ تکمیل کردند. این اولین مرورگر نقطه و کلیک وب^۲ در دنیا بود. این مرورگر می توانست چندین صفحه را بارگذاری کند و سپس با کلیک بر روی یک فرامتن در یک پنجره جدید، صفحه کلیک شده باز شود (هالورد^۳، ۲۰۰۹). مرورگر مهم دیگر که مادر همه مرورگرهای دیگر شناخته شد موزایک^۴ بود که امکان ذخیره، بازیابی و ارائه اطلاعات را با کاهش زمان بازیابی فراهم می کرد و از قابلیت جستجوی مدارک فرامتن با گرافیک بالا برخوردار بود (جادسن^۵، ۱۹۹۶).

در سال ۱۹۹۴ تیم برنرزی - مبدع وب- و جفری جف^۶ برای توسعه استانداردهای وب، کنسرسیوم وب جهان گستر را به وجود آوردند که هدفش افزایش پتانسیل وب برای ارائه خدمات بیشتر بود. این کار با توسعه استانداردهایی انجام شد که باعث رشد پایدار شدند. هدف آن ها "وب برای همه" است و این کار را با دسترس پذیر ساختن، بین المللی کردن^۷ و ایجاد وب برای توسعه اجتماعی^۸ انجام می دهند (ماموریت W3C، ۲۰۱۲).

در این دهه، مسأله بعدی در بازیابی اطلاعات استفاده از داده های استنادی و تحلیل پیوندها^۹ بود که مراحل آغازین خود را طی می کرد. توسعه دهندگان وب، تشخیص دادند که می توانند از پیوند میان صفحات برای ساختن خزنده ها و روبات ها استفاده کنند تا صفحات بیش تری جمع کنند و بدین ترتیب مجموعه سازی را به صورت خودکار درآورند (ساندرسون و کرافت، ۲۰۱۲). تحلیل پیوندها، فنی است که از اطلاعات ضمیمه شده متصل در ساختار فرامتن وب استفاده می کند تا کیفیت نتایج

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

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

روش ها مانند روش های بازیابی قبل یا مبتنی بر کلیدواژه هایی بودند که به موتورهای جستجو ارائه می شدند، یا بر اساس فهرست های موضوعی که در نظام استفاده می شد یا بر پایه فراداده ها، جستجو را انجام می دادند که از آن جمله می توان به شاهره موضوعی^۱ مثل ای اچ دی اس^۲ اشاره کرد (ژانگ و تانگ، ۲۰۰۷). اما وب ۲.۰ نیازهای جدیدی را به وجود آورد:

۱- تعمیم پذیری و گستردگی محصولات اطلاعاتی: در وب ۱.۰ تولید اطلاعات شبکه ای به عهده تعداد اندکی از کمپانی های حرفه ای و ویرایشگران وب سایت ها بود اما بسیاری از کاربران، محتوای اطلاعاتی زیادی را در شبکه محیط وب ۲.۰ عرضه می کردند (ژانگ و تانگ، ۲۰۰۷). در وب ۲.۰ روزنامه نگاران، سازندگان وب سایت ها یا کمپانی ها نیستند که اطلاعات را تولید می کنند بلکه هر کاربر از طریق رسانه های متعدد و خدمات متنوع می تواند محتوا تولید کند (پیترز^۳، ۲۰۰۹).

۲- محتواهای کوچک^۴ و معنابخشی های ساختار اطلاعات^۵: محتواهای کوچک، داده های مختلفی اند که کاربران خلق می کنند؛ مانند تصویر، فهرست های موسیقی، جاهایی که آرزو دارند بروند، دوستان جدید و مانند آن. در وب ۲.۰ می توان 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 شبکه سازی اجتماعی (فیسبوک^۶)، برچسب گذاری اجتماعی (فلیکر^۷)، نشانه گذاری اجتماعی (دلیشس^۸)، رده بندی های مردمی^۹، وبلاگ نویسی^{۱۰} و ویکی ها (ویکی پدیا) است (بری، ۲۰۰۹). برای استفاده از این محتواهای کوچک، نشانه های معنایی^{۱۱} لازم است که اجتناب از مال موجود در وب ۱.۰ این قابلیت را نداشت (ژانگ و تانگ، ۲۰۰۷). وب ۲.۰ تولید این محتواها را با توجه به قابلیت های خود امکان پذیر کرد. به طور کلی یکی از مزیت های وب ۲.۰ شبکه نشانه گذاری کاربر^{۱۲} است که نه تنها می تواند برای توصیف دقیق تر هر منبع استفاده شود؛ بلکه می تواند برای کشف شباهت مشخصه-

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

یکی از پیشرفت های دهه ۱۹۹۰، ایجاد امکان پوشش^۱ بود. بسیاری از ناشران در پی آن بودند تا تصاویر صفحات را پوشش کنند و بفروشد که با اختراع دیسک های ارزان و سی دی رام های ارزان تر امکان پذیر شد. بدین ترتیب، دو نوع فرمت برای بسیاری از منابع به وجود آمد. اول به صورت پیوسته و روی وب، و بعد با استفاده از سی دی رام. 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 مجموعه های سنتی، چه در نسل اول وب و چه نسل دوم پویا بودن آن است در مقابل نظام های بازیابی سنتی که ایستا و در بسیاری موارد غیرقابل تغییر بودند. اگر هم امکان تغییر وجود داشت، هزینه بر بود. ویژگی دیگر وب آن است که خود سازمان دهنده محسوب می شود؛ یعنی افراد به راحتی و بدون استفاده از استانداردهای پیچیده، می توانند محتوا تولید کنند، در حالی که در مجموعه های 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 شبکه ای را روزآمد کرد؛ بلکه بر روش های بازیابی سنتی اطلاعات شبکه ای بسیار تأثیر گذاشت و نیازهای جدیدی را مطرح کرد (ژانگ و تانگ، ۲۰۰۷). یافته های پژوهشگران در حوزه بازیابی اطلاعات در سه دوره زمانی نشان می دهد که حوزه بازیابی در سال های ۸۴-۱۹۸۰ درگیر پژوهش های آموزش، ذخیره و نظام ها بوده و در سال های ۹۴-۱۹۹۰ در حوزه خدمات اطلاعاتی، رابط های کاربری و پایگاه های اطلاعاتی و در دوره زمانی ۲۰۰۰ به بعد پژوهش های مربوط به وب رواج بیشتری یافته است (سوگیموتو و مک گین^۲، ۲۰۱۰).

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

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

1 Scaninig
2 Sugimoto and McCain

کاربر که برای به دست آوردن منابع شخصی دقیق تر مفید است به کار رود (یانگ و وانگ^۱، ۲۰۰۹).

محتوای اجتماعی وب ۲۰۰ بازیابی اطلاعات را واقعاً متحول کرد، زیرا چنان که قبلاً اشاره شد، نه فقط متخصصان که افراد عادی توانستند محتوا تولید کنند و این امر چالش هایی به همراه داشت و آن افزایش نیاز به خلق روش ها و امکانات جدید برای ذخیره و بازیابی اطلاعات بود. اشتراک گذاری اطلاعات افزایش یافت. همه رسانه های اجتماعی مانند فیسبوک و فلیکر، قابلیت های فراوانی برای اشتراک اطلاعات و دانش در میان افراد مختلف دارند. ابزارهای اجتماعی توانسته اند بسیاری از مشکلات به وجود آمده را حل کنند.

یکی از این ابزارهای نسل جدید، رده بندی های مردمی هستند. رده بندی مردمی، ابزار تولید، عرضه، بازیابی و گسترش اطلاعات در وب ۲۰۰ است که کمپانی ها و متخصصان رایانه آن را به وجود آوردند تا مردم، محتواهای خود را تولید و ذخیره و با نشانه های سفارشی نمایه سازی کنند. این خدمت اطلاعاتی تعاونی، شامل اشتراک عکس ها، ویدئوها و نشانه گذاری های اجتماعی است که می تواند در بازیابی اطلاعات به افراد کمک کند (پیتز، ۲۰۰۹).

موضوع دیگر در وب ۲۰۰، ربط بود. هر چند که ربط، چالش همه نظام های بازیابی بوده اما به طور خاص در بازیابی تحت وب موضوع مهم تری محسوب می شود. به طور مثال، دقت^۲، 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 می کنند و با بی حوصلگی، انتظار دارند که دقت جستجو بالا رود. محدودیت دیگر نظام های بازیابی در وب، معیارهای اندازه گیری است. معیارهای نظام های بازیابی سنتی، بر اساس مجموعه های کنترل شده بود اما این معیارها برای نظام های وبی واقعی به نظر نمی رسند (لانگوئل و مایر، ۲۰۰۶). بدین ترتیب رتبه بندی ربط در وب نسبت به نظام های بازیابی سنتی تغییر کرده است.

رتبه بندی ربط در وب ۱۰۰ بر توزیع اصطلاح^۳ متکی بود که بر اساس بسامد اصطلاح و بسامد مقلوب مدرک و مدل های زبانی انجام می شد. اما در گذار از

وب ۱۰۰ به وب ۲۰۰ مهم ترین اتفاق تاریخ بازیابی تحت وب، حرکت از تحلیل لینک به رتبه بندی صفحات^۱ بود. بدین ترتیب رتبه بندی ربط در وب ۲۰۰ مجموع امتیاز بازیابی اطلاعات^۲ و رتبه صفحات بود (ما، ۲۰۱۰).

با وجود تداوم چالش ها نسل بعدی وب از راه رسید. به طور کلی در وب ۲۰۰ کاربران، هم خواننده و هم نویسنده اند یعنی محتوا تولید و کنترل می کنند. محتوا به طور فزاینده ای در حال افزایش است و وب به یک محیط پویا تبدیل شده است. کاربران، ارزش هایی را با استفاده از برنامه های کاربردی به وب اضافه می کنند. مشارکت کاربر باعث افزایش هوش جمعی می شود. هم چنین واسطه های کاربرمدار و محتواهای شخصی تجربه کاربر را افزایش می دهد. وب به عنوان یک بنیاد برنامه نویسی^۳، اجرای برنامه های کاربردی را به طور کامل با استفاده از یک مرورگر امکان پذیر می کند. قابلیت انتقال در وب بسیار بالاست. نرم افزار در سطحی بالاتر از یک ابزار واحد است و باز بودن^۴ نرم افزارها ویژگی بسیار مهمی است (گومز-پرز، فرناز-لوپز و کورکوگاریا^۵، ۲۰۰۴). همه این ها وب ۲۰۰ را به سوی نسل جدیدی از وب پیش برده و ظهور پیشرفت های جدیدی را نوید داده است که وب ۳۰۰ نام دارد. این نسل وب به نسل وب معنایی مشهور است. نه با آن معنی که در وب ۲۰۰ وب معنایی وجود ندارد اما مشخصه ویژه این نوع وب، تمرکز بر معناست (بری، ۲۰۰۹). تیم برنرزی خالق وب در سال ۱۹۹۸، ایده وب معنایی را مطرح کرد (استایلسویگ^۶، ۲۰۰۶). بعد از سال ۲۰۰۰ این ایده عملی شد. وب ۳۰۰ هوشمند است چون فن آوری های وب معنایی، پایگاه های جهانی و برنامه های کاربردی هوشمند دارد و به صورت بی سیم در همه جا اداره می شود. دسترسی به اینترنت سیار^۷ و ابزارهای سیار^۸ مشخصه وب معنایی است. مهم ترین ویژگی این نوع وب، فن آوری های باز^۹ است و شامل پروتکل ها و برنامه کاربردی باز، شکل های داده ای باز، باز، بنیادهای نرم افزاری منبع باز، و داده های باز هستند (گومز-پرز و دیگران، ۲۰۰۴).

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

1 Yang and Wang
2 precision
3 Term distribution

در وب ۳.۰ رتبه‌بندی ربط به سوی آگاهی حرکت کرده و معماری نظام‌های بازایی اطلاعات از وب‌محور به داده‌محور پیش رفته است. اکتشاف دانش و داده‌کاوی وب از صفحات وب به اشیای وب (موجودیت‌ها)^۱، هم‌چنین از ساختارنیافته ساختارنیافته به ساختارنیافته 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^۲ بر نظام‌های بازایی اطلاعات تأثیر گذاشته است (دیویس، هارملن و فنسل^۳، 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 بازایی اطلاعات، دوران پر فراز و نشیبی را پشت سر گذاشته‌اند اما هنوز با چالش‌های زیادی مواجه‌اند. هنوز مسأله ربط، تعامل و ایجاد نظام‌های بازایی کاربرمدار از مشکلات عمده این نظام‌هاست. تاریخچه این نظام‌ها از دوران اولین فهرست‌ها تا زمان وب ۳.۰ به خوبی نشان می‌دهد که بازایی اطلاعات به مشارکتی و همه‌گیر شدن اطلاعات منجر می‌شود و پیشرفت‌ها و فن‌آوری‌های عمده، نظام‌های بازایی اطلاعات را مسلماً تعاملی‌تر کرده‌اند.

خلاصه فصل

در تعاریف مختلف برای بازایی اطلاعات و نظام‌های بازایی جنبه‌هایی مانند ذخیره، بازایی، سازماندهی، جستجو، ردیابی و تفسیر اطلاعات، هم‌چنین وجود یک نظام رایانه‌ای، نوع اطلاعات مانند متن، تصویر، صدا و 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 بازایی اطلاعات ارائه شد. در بخش دوم تاریخچه

نظام‌های بازایی بررسی گردید. در ابتدا مشخص شد که از همان ابتدا بازایی اطلاعات همزمان با تولید اطلاعات بود و نخستین فهرست‌های کتابخانه‌ای، حاصل اولین تلاش انسان برای سازماندهی و قابل دسترس کردن اطلاعات بوده است. عصر نظام‌های اطلاعاتی مدرن در دهه ۱۹۴۰، با اختراع اولین رایانه الکترونیکی به وجود آمد. ماشین ممکس ایده ایجاد ماشین ذخیره و بازایی 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 بازایی 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 پرداخته است (کرافت، متزler و استرامن^۱، 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 مهندسين نرم‌افزار است پرهیز کرده‌اند.

موتور جستجو چیست؟

موتورهای جستجو از اولین ابزارهای کاربردی در وب، برای جستجو و مکان یابی اطلاعات اند و کاربران وب، ساده ترین تا پیچیده ترین اطلاعات را از آنها به دست می آورند. امروزه یک موتور جستجو مانند گوگل به ابزار ضروری و انکار نشدنی اطلاع یابی تبدیل شده است. این ابزارها از بدو تاریخ پیدایش، همواره در حال ارتقا و پیشرفت بوده اند. به طور مثال اولین موتور جستجو به نام آرچی، امکانات بسیار محدودی داشت و تنها برای جستجوی فایل های اف تی پی مناسب بود. ورونیکا اولین موتور جستجویی بود که می توانست متن را جستجو کند (موتور جستجو چیست^۱، ۲۰۱۴). موتورهای جستجوی جدید قابلیت های زیادی در جستجوی انواع اطلاعات دارند.

این ابزارهای بازیابی اطلاعات نه فقط به صورت مستقل که در نظام های بازیابی اطلاعات نیز بخش بسیار مهمی محسوب می شوند مانند نظام های بازیابی اطلاعات پزشکی یا پایگاه های 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 نرم افزار کاربردی سطح بالا را که خود مجموعه ای از نرم افزارهای دیگر است به خوبی نشان می دهد.

موتورهای جستجو به زبان ساده، برنامه های نرم افزاری اند که از روایات یا خزنده ها تشکیل شده اند. خزنده ها، در وب حرکت می کنند، صفحات جدید را بازبینی می کنند و آن ها را برای نمایه شدن به درون خود می کشند. بعد از ورود نشانی آن ها به درون ساختار موتور جستجو، با عملیات مختلف و پیچیده ای که روی صفحات مورد نظر انجام می شود کل صفحه، نمایه می شود و در درون پایگاه اطلاعاتی موتور جستجو ذخیره می گردد. هم چنین به صفحاتی که قبلاً آن ها را نمایه کرده اند می روند و دوباره آنها را بر اساس آخرین تغییرات نمایه می کنند. هنگامی که کاربر اطلاعات را جستجو می کند پرس و جوی^۲ وی در درون نرم افزار موتور جستجو ترجمه شده و بر اساس آن، از پایگاه اطلاعاتی نمایه، اطلاعات مرتبط، انتخاب می شود و با توجه به الگوریتم های رتبه بندی، فهرستی از نتایج ظاهر می شود.

هر چند ساختار موتورهای جستجو در کلیات به هم شبیه اند اما هر کدام برحسب ویژگی ها، راهبردها و الگوریتم های مورد استفاده از یکدیگر متفاوت اند. همه موتورهای جستجو در بازار رقابتی، تلاش می کنند بتوانند هر چه بیشتر نظر کاربران را جلب کنند و با برآوردن نیازهای آنان از دیگر رقبای خود پیشی بگیرند. رقابت فن آوری و اطلاعاتی در دنیای امروز، بخش مهمی از سهم بازار تجارت را به خود اختصاص داده است و در این میان موتورهای جستجو و شرکت های مادر آنها نیز از این قاعده مستثنی نیستند و برای رسیدن به اهداف عمده خود تلاش می کنند. برای همه موتورهای جستجو دو هدف عمده تعیین شده است: اثر بخشی و کارایی^۱ (ناومن^۲، ۲۰۱۱) که در حقیقت از مباحث ارزیابی عملکرد موتور جستجو هستند (داسدان، و سیاشلکلس ولاپسگلو^۳، ۲۰۱۰). در ادامه بر اساس این دو هدف می توان ساختار موتور جستجو را تشریح کرد.

اثربخشی در موتورهای جستجو

اثربخشی در کلیه نظام ها مسأله ای اساسی است. در نظام های اطلاعاتی نیز این موضوع اهمیت دارد و بخشی از ارزیابی نظام های بازیابی اطلاعات و از جمله موتورهای جستجو است. این که چه اندازه از نظر کاربر، نتایج به دست 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 اندازه گیری سنتی نمی توان اندازه گیری کرد (اسپینک و زیمر^۴، ۲۰۰۸) لواندوسکی^۵ (۲۰۱۲)، اثربخشی موتورهای جستجو را در چهار بخش خلاصه می کند: کیفیت نمایه^۶، کیفیت نتایج^۷، کاربردپذیری^۸ و کیفیت ویژگی های جستجو^۹.

کیفیت نمایه

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

1 Query
2 Naumann
3 Dasdan, Tsioutsoulis, 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^۱، سوگیری کشوری^۲ و روزآمدی^۳ 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 به صورت محلی فعالیت کند و در عین حال نیاز کاربران خود را به خوبی 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 چه قدر 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 به طور محلی به بخشی از موتور جستجو در منطقه خود متصل می‌شود. به طور مثال 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^۴، ۲۰۱۴) و جستجوهای نزدیک به کشور خود را بیش‌تر بیش‌تر انتخاب می‌کنند. به طور 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

مسیر می‌یابد و با سرعت بیشتری جستجو می‌کند.

۳. روزآمدی: گاهی با اصطلاح تازگی^۱ 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 به صورت 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^۲، ۲۰۰۳). دلایل زیر به طور 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 به صورت جزئی درمی‌آید؛
- ✓ وقتی خزنده یک موتور جستجو پایگاه اطلاعاتی خود را تجدید^۳ می‌کند شاید برخی صفحات را که قبلاً بازدید کرده به دلیل خرابی خدمت‌دهنده^۴ 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. به طور کلی

1 freshness
2 Barker
3 refresh
4 server failure

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

همه قضاوت های ربط و قضاوت کاربران در مورد کیفیت یک موتور جستجو، بیش تر بر اساس نتایجی است که موتور جستجو به کاربر می دهد. در حقیقت کاربران وقتی به نتایج دلخواه خود می رسند ترغیب می شوند تا دوباره از یک موتور جستجو استفاده کنند. پس برون داد یک موتور جستجو و میزان ربط آن به پرسش و نیاز کاربر، رابطه بسیار مستقیمی با اثربخشی یک نظام بازیابی و به طور اخص یک موتور جستجو دارد.

پژوهشگران سودمندی موتور جستجو را به طور مستقیم با کیفیت نتایج یک موتور جستجو و زمان سپری شده برای جستجو مرتبط می دانند (مودیب و زلیبرستین^۱، ۱۹۹۹). بسیاری از پژوهش ها در نظام های بازیابی به افزایش کیفیت نتایج در این نظام ها بازمی گردد و پژوهشگران به فنون یکپارچگی اطلاعات در این حوزه بسیار علاقمندند (ناومن، ۲۰۰۱)

کاربردپذیری (قابلیت استفاده)

راهبردهای بسیاری در موتورهای جستجو و بیش تر نظام های اطلاعاتی به کار می رود 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 تلاش های کاربران برای استفاده از یک موتور جستجو است. در واقع یک ویژگی چندبعدی از یک نظام رایانه ای محسوب می شود (تاکسا، اسپینک و گلدبرگ^۲، ۲۰۰۸)

سوال اساسی در کاربردپذیری موتور جستجو این است که چه قدر موتور جستجو می تواند اطلاعات مورد نیاز کاربر را در اختیارش قرار دهد. تعریف استاندارد ایزو از کاربردپذیری میزانی است که کاربر می تواند یک محصول را به صورت 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 می دانند. دسترس پذیری^۳ و در دسترس بودن^۴ و اکتشاف اطلاعات^۵ (هنگ^۶، ۲۰۱۴)، بازخورد کاربر، رضایت کاربر، توانایی برای نادیده گرفتن خطاها و فرمول بندی پرسش،

کارایی و اثر بخشی موتور جستجو (تاکسا و دیگران، ۲۰۰۸)، توانایی یادگیری^۱ و یادسپاری^۲ (مقدمه ای بر کاربردپذیری^۳، ۲۰۱۲) برخی از این عناصرند. به طور مثال، توانایی یادگیری نشان می دهد تا چه اندازه کار با موتور جستجو برای کاربر تازه وارد آسان است و یادسپاری به این معنی است که چه قدر کارکرد موتور جستجو حتی اگر مدتی از آن استفاده نشود در خاطر می ماند.

اصل در کاربردپذیری موتور جستجو آن است که آیا برای کاربر امکان دارد یا آن به صورت کارا و موثر تعامل کند یا نه (جویس، بیسکری، گاناسیا، روکس^۴، ۲۰۱۲). لودانسکی (۲۰۰۸) معتقد است کاربردپذیری، بازخوردی از رفتار کاربر می دهد و می توان آن را با مشاهده پیمایش های کاربر، آزمون های آزمایشگاهی و تجزیه و تحلیل گزارش ها ارزیابی کرد.

کیفیت ویژگی های جستجو

در موتور جستجو باید ویژگی هایی چون جستجوی پیشرفته و زبان پرس و جوی پیچیده پیش بینی 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 به کار می برد. وجود روش های مختلف جستجو مانند جستجوی کنترل شده با استفاده از اصطلاح نامه، جستجوی خوشه ای، جستجو بر اساس فیلترهای مختلف مانند جستجوی محدود به یک زبان، زمان و مکان خاص، حتی به یک دامنه یا یک حوزه خاص از ویژگی های عمده ای است که موتورهای جستجو دارند. بیش تر موتورهای جستجو مانند گوگل، جستجوی عکس، ویدئو یا کتاب و مقالات علمی را با امکانات پیشرفته خود انجام می دهند. در کنار بحث اثربخشی هر نظام، بحث اساسی دیگر، کارایی آن است. در موتورهای جستجو، نیز کارایی به عملکرد صحیح نظام بازیابی برمی گردد.

کارایی موتورهای جستجو

این ویژگی به زمان و توان صرف شده برای جستجو در موتور جستجو اشاره می کند. زمان صرف شده برای جستجو برای کاربر اطلاعات بسیار مهم است. به طور

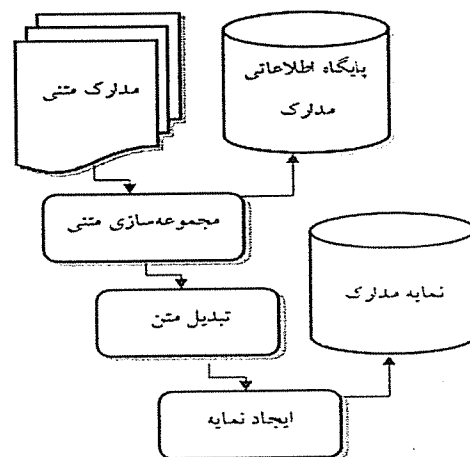
1 Learnability
2 Memorability
3 Introduction to Usability
4 Jouis, Biskri, Ganascia and Roux

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

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 (۱-۲). زیر بخش‌های اصلی در فرآیند نمایه‌سازی در موتور جستجو

همانند شکل فوق، فرآیند نمایه‌سازی موتور جستجو شامل سه بخش مهم است: مجموعه‌سازی متنی^۱، تبدیل متن و ایجاد نمایه.

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 اطلاعاتی کاربران، میزان زمانی است که آن‌ها برای جستجو در اختیار دارند (جگد و آنوچکوا^۱، ۲۰۱۱). با این توصیف می‌توان نتیجه گرفت اهمیت بحث کارایی در موتور جستجو بسیار بالاست. اگر اثربخشی تأکید می‌کند که چه‌قدر یک موتور جستجو در بازیابی اطلاعات درست برای کاربر موفق است کارایی 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 کارایی موتور جستجو را افزایش دهند. به‌طورمثال برخی از موتورهای جستجو از الگوریتم‌های فازی تعاملی برای کاهش زمان صرف‌شده در نظام‌های اطلاعاتی مانند پایمدا^۲ استفاده می‌کنند (وانگ و دیگران، ۲۰۱۰).

به‌طور کلی نوع نگاهی که در مورد بحث کارایی 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^۳ و پرس‌وجو^۴ است (لیم، لیو و لی^۵، ۲۰۱۱). در معماری نظام بازیابی یک بخش دیگر هم اضافه می‌شود که بازخورد و بازیابی است (ناسا و ووک^۶، 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

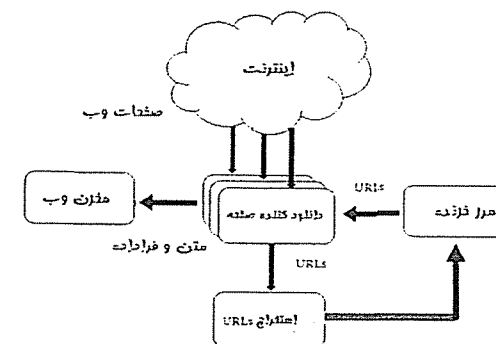
مجموعه سازی متنی

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

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^۱، تبدیل و مخزن داده ای مدرک^۲ است (ناومن، ۲۰۱۱) که در ادامه به آن ها اشاره می شود.

خزنده یا عنکبوت

برنامه های نرم افزاری اند که در وب گردش و از صفحات وب بازدید می کنند. سپس آن ها را می خوانند و صفحات را به عنوان ورودی به بخش نمایه موتور جستجو می دهند. خزنده بر اساس توان نرم افزاری خود در سراسر وب می خزد و پیوند به دیگر صفحات را دنبال می کند تا مطمئن شود تا حد امکان تمامی صفحات را خوانده و برای نمایه شدن آماده کرده است. ساختار وب یک ساختار گرافیکی است و خزنده ها از پیوندهای ارائه شده در یک صفحه برای باز کردن صفحات دیگر استفاده می کنند. آن ها برای بازیابی صفحات و قراردادن آن ها، در یک مخزن محلی ساخته شده اند (یودپر، کیل و دارمیک^۳، ۲۰۱۴). شکل بعد ساختار یک خزنده را در موتور جستجو نشان می دهد.



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

مطابق شکل یک، خزنده دارای سه بخش مجزا است. نخست مرز خزنده^۱ است. این بخش فهرستی از نشانی صفحاتی را ذخیره می کند که خزنده آن را بازیابی نکرده است. به زبان ساده این بخش مجموعه ای از نشانی های اینترنتی صفحات است. دوم دانلود کننده صفحه^۲ است که مهم ترین کار آن دانلود صفحاتی است که نشانی آن ها در بخش مرز خزنده ذخیره شده اند. مخزن وب^۳ نیز بخشی است که دسته بزرگی از اشیای داده ای را ذخیره و مدیریت می کند. اشیای داده ای در این مخزن وب، همان صفحات وب هستند. این مخزن فقط صفحات اچ تی ام ال استاندارد را ذخیره می کند (یودپر و دیگران، ۲۰۱۴). به طور کلی خزنده وب ابتدا مجموعه ای از نشانی های اینترنتی را که تاکنون بازدید نشده اند یا صفحات آن ها روزآمد شده است به بخش مرز خزنده اضافه می کند. سپس خزنده از میان مجموعه نشانی ها، صفحاتی را انتخاب می کند و بخش دانلود کننده صفحات، آن صفحه را می باید و دانلود می کند. صفحات دانلود شده به بخش مخزن خزنده برای انجام فرآیندهای بعدی، یعنی تبدیل متن^۴ اضافه می شوند.

به طور کلی خزنده وب عملیات زیر را انجام می دهد:

- ✓ مقدار دهی اولیه نشانی های اینترنتی (گزینش نشانی ها)
- ✓ اضافه کردن آن ها به بخش مرز خزنده
- ✓ انتخاب نشانی ها از مرز خزنده
- ✓ گرفتن صفحات وب مربوط به این نشانی ها
- ✓ تجزیه^۵ صفحه بازیابی شده برای استخراج^۶ نشانی ها
- ✓ اضافه کردن تمام پیوندهای بازدید نشده در صفحه به فهرست نشانی ها، یعنی به مرز خزنده
- ✓ شروع مرحله دو و تکرار عملیات (یودپر و دیگران، ۲۰۱۴)

بخش تغذیه یا فید

دومین بخش از مجموعه سازی متنی است و به جریان بلادرنگ مدارک^۷ در وب اشاره می کند (نامن، ۲۰۱۱). این بخش مکانیسمی برای دسترسی به مدارک در وب است. از نمونه های مهم آن می توان به آراس اس^۸ اشاره کرد. اما به طور 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

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

بلاگ ها، اخبار، ویدئو، رادیو و تلویزیون را دربرمی گیرند. همان طور که در بخش قبلی گفته شد، خزنده وب فقط صفحات اچ تی ام ال را می خواند و ذخیره می کند؛ اما همه صفحات وب، اچ تی ام ال نیستند. بخش اعظمی از وب صفحات ایکس ام ال^۱ هستند مانند منابع دیداری و شنیداری. بخش فید مربوط به این نوع منابع است. در ادامه یک نمونه از فید یعنی آراس اس تشریح شده است.

آراس اس به طور معمول استاندارد برای فیدهای وب است. یک آراس اس خوان^۲ می تواند برای به اشتراک گذاری فیدهای آراس اس استفاده شود. شکل فیدهای آراس اس، ایکس ام ال هستند. آراس اس خوان، فیدها را مانیتور می کند و به محض ورود آن ها، محتوای جدید را آماده می کند (کرافت و دیگران، ۲۰۱۰). منابع تلویزیونی و رادیویی نیز نوعی از فیدها هستند که این بخش از مجموعه سازی متنی آن ها را می خواند. موتورهای جستجوی جدید در حال حاضر می توانند داده هایی مانند فیلم و ویدئو را با تغییر منابع مثل همین دستیابی به فیدهای آراس اس و تویتر ای پی آی^۳ نمایه سازی کنند و با استفاده از ابزار تبدیل که در بخش بعدی توضیح داده داده می شود دامنه وسیعی از اطلاعات مانند فایل های فلش^۴ و پی دی اف را به اصطلاحات نمایه ای تبدیل نمایند (دومان، ۲۰۱۱).

تبدیل

این بخش مربوط به تغییر قالب مدارک از شکل واقعی خود به شکل استاندارد قابل خواندن در موتور جستجو است. در این جا تبدیل حتی برای منابع رمزگذاری شده و منابعی که با زبان های برنامه نویسی به صورت کد درمی آیند هم اتفاق می افتد.

منابعی که خزنده یا فید بازیابی می کنند همه منابع متنی نیستند؛ بلکه منابع در وب به فرمت های مختلف مانند پی دی اف و ایکس ام ال، ورد^۵ و پاورپوینت^۶ و غیره هستند. موتورهای جستجو نیاز دارند که این منابع را به منابع متنی همراه با فراداده تبدیل کند. بخش تبدیل متنی در فرآیند نمایه سازی موتور جستجو، صفحات اچ تی ام ال و ایکس ام ال را به قالب استاندارد درمی آورند اما در مورد مدارک

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

پی دی اف، پاورپوینت یا متون رمزگذاری شده مانند مدارک رمزگذاری شده با کدهای اسکی^۱ اولین گام برای آماده سازی مدارک برای پردازش بیشتر، این نوع تبدیل متن است (کرافت و دیگران، ۲۰۱۰). این ابزار تبدیل، مدرک را به یک قالب برچسب متنی مانند اچ تی ام ال یا ایکس ام ال تبدیل می کند و برخی از اطلاعات مهم قالب بندی را نیز حفظ می کند (متزلر و لیوسکی^۲، ۲۰۱۲).

مخزن داده ای مدرک

این مخزن یا پایگاه داده ای، همان طور که از نامش مشخص است، محل ذخیره و مدیریت اطلاعاتی است که در بخش های قبلی جمع آوری شده است و شامل داده های ساختاریافته و اطلاعات مرتبط با آن هاست. داده های ساختاریافته؛ فراداده، پیوندها و متون وابسته به این پیوندها^۳ را دربرمی گیرند (دومان، ۲۰۱۱). این مخزن می تواند یک پایگاه داده ای رابطه ای^۴ باشد و به طور خاص از یک نظام ذخیره سازی با کارایی بیش تر و ساده تر برای ذخیره تعداد بی شماری از مدارک استفاده کند (اسمیت، ۲۰۱۲). به طور کلی دلیل ذخیره سازی مدارک آن است که هم موتور جستجو برای پاسخ به پرسش های کاربر سریع تر به اطلاعات دسترسی داشته باشد و هم این پایگاه بتواند فهرست نتایج را تولید کند و خلاصه ای از مدارک را در اختیار کاربر قرار دهد (ناومن، ۲۰۰۱).

تبدیل متن

این بخش، اطلاعات شناسایی و ذخیره شده را در بخش مجموعه سازی متنی به اصطلاحات نمایه ای تبدیل می کند (متزلر و لیوسکی، ۲۰۱۲) یعنی متن گرفته شده از اینترنت را از قالب و شکل وبی خود تغییر می دهد و به شکل استاندارد قابل خواندن برای برنامه های کاربردی موتور جستجو تبدیل می کند. مراحل تبدیل متن شامل تجزیه^۵، ایست واژگان^۶، ریشه گیری^۷، تحلیل پیوند^۸، استخراج اطلاعات^۹ و طبقه بندی کننده^{۱۱} 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 می شود و بخش هایی که ارزش اطلاعاتی دارند جدا می شوند. این اولین مرحله برای تبدیل متن است. متن به توکن^۱ یا کوچک ترین واحد اطلاعاتی با ارزش که اغلب هم کلمات اند شکسته و تجزیه می شود. تبدیل متن یک مدرک به توکن ها بسیار اهمیت دارد. برای این کار، ابتدا ساختارهای اصلی متن شناسایی می شوند. ساختارهای اصلی شامل عنوان، سرعنوان، پیوندها، اشکال و مانند این هاست. بعد از شناسایی، مرحله تجزیه به توکن ها انجام می شود (کرافت و دیگران، ۲۰۱۰). هر تجزیه کننده^۲، برای اجرا در تمام وب طراحی شده است و باید بتواند آرایه بزرگی از اشتباهات احتمالی را اداره کند. اشتباهات از غلط های املایی در کدهای اچ تی ام ال تا مشکلات بزرگ تر را در بر می گیرد (گالی^۳ و دیگران، ۲۰۰۵). در تجزیه باید به مسائلی مانند جداکننده ها^۴، آپوستروف، حروف بزرگ، کاراکترهای غیرالفبایی توجه کرد (ناومن، ۲۰۰۱). مشکلات احتمالی در زبان های غیر لاتین بیش تر است. مثلاً در زبان چینی که جداکننده کلمات وجود ندارند (کرافت و دیگران، ۲۰۱۰) یا در زبان فارسی، به طور مثال، تشخیص افعال دو کلمه ای 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. به طور مثال در تجزیه متن، برای برنامه جداکننده بسیار مشکل است که میان یک کلمه مانند "میهن" به عنوان یک برند یا کلمه میهن به معنای کشور تفاوت قائل شود. البته زبان های نشانه گذاری اغلب برای تشخیص ساختار از برچسب ها و نحو^۵ موجود در زبان نشانه گذاری استفاده می کنند (ناومن، ۲۰۰۱).

بازدارندگی یا ایست واژگان

به معنای حذف کلمات عام و آن هایی است که در متن ارزش معنایی ندارند و فقط اطلاعات اضافه می دهند یا به جملات پیوستگی می بخشد. می توان به حروف اضافه، افعال کمکی یا معین یا کلماتی اشاره کرد که با فراوانی زیاد در متن ظاهر می شوند. به طور مثال در جمله زیر:

- 1 Token
- 2 Parser
- 3 Gullti
- 4 seperators
- 5 syntax

"برنامه های کاربردی کامپیوتری که به تازگی ارائه شده اند بز تلفن های همراه تأثیر زیادی گذاشته اند."

در این متن که شامل سیزده کلمه است تنها کلمات "برنامه های کاربردی، کامپیوتر، تلفن های همراه" ارزش نمایه شدن دارند و بقیه کلمات حذف می شوند. به این نوع کلمات، که بسیار هم در جملات و متون متعارف اند سیاهه بازدارنده^۱ واژگان گفته می شود که بعید به نظر می رسد به معنای متن کمک کنند. حذف این گونه کلمات در کاهش فضای نمایه، زمان جستجو و کارآیی و اثربخشی موتور جستجو تأثیر فراوان دارد و در متن کاوی و داده کاوی نیز برای بازنمایی موثر متن مهم است.

ریشه گیری

بخش اصلی این واژه کلمه "stem" به معنای ساقه و ریشه است. می توان این اصطلاح را گرفتن ریشه کلمه و واژه معنا کرد. فرهنگ لغت موتورهای جستجو^۲ 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 ریشه گیری استفاده می کنند ریشه واژه های مورد جستجو را در مدارک پایگاه داده خود مقایسه می کنند. برای مثال، اگر کاربر "مرور" را به عنوان پرسش وارد کند، موتور جستجو کلمه را به ریشه کلمه کاهش می دهد و تمام مدارک حاوی آن مانند مرور، مرورگر، مرور و مرور کردن را بازیابی می کند (فرهنگ لغت موتور جستجو، ۲۰۰۹)

در برخی موارد این اصطلاح در کنار اصطلاح lemmatization و گاهی به جای آن به کار می رود، هر چند که کاربرد اصطلاح stemming معمول تر 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 به جای واژه stemming یا هم ارز آن به کار می رود. با مثال ساده زیر می توان تفاوت این دو تکنیک را به خوبی درک کرد.

در ریشه گیری اگر کاربر کلمه "ماشین" را برای پرس و جو 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، 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. بدین ترتیب، در مواردی که کاربر دقیقاً دنبال یک یا چند واژه مشخص است، ریشه‌گیری، کار را برای وی دشوار می‌کند (لیدی، ۲۰۰۱).

به‌طور کلی در موتورهای جستجو، استفاده از برخی فنون کمک می‌کند که کاربر بتواند دقیقاً عین واژه را جستجو کند. به‌طورمثال در موتور جستجوی 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 (انکر تکست^۱) 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 می‌خورند، 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^۱ و جستجوی موضوعی منتج از فرامتن (HITS)^۲ 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^۳ دو دانشجوی دانشگاه استنفورد و مؤسسان گوگل درست کردند (دوی، گوپتا و دیکسیت، ۲۰۱۴). اما برای این که بدانیم این الگوریتم چگونه کار می‌کند باید اصطلاحات پیوند ورودی^۴ و پیوند خروجی^۵ 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^۶ 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

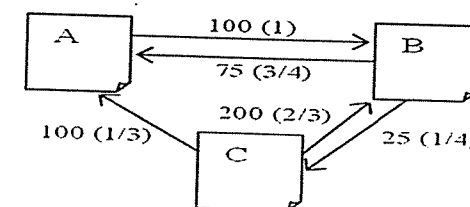
1 PageRank
2 Hyperlink-Induced Topic Search
3 Sergey Brin and Larry Page
4 Devi, Gupta and Dixit
5 In-link
6 Out-link
7 Authority

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

$$PR(A) = (1 - d) + d \left(\frac{PR(T_1)}{Q(T_1)} + \dots + \frac{PR(T_n)}{Q(T_n)} \right)$$

در این معادله:

- ✓ $PR(A)$ رتبه صفحه A است.
- ✓ $T_1, \dots, T_n$ همه صفحاتی که به صفحه A پیوند داده‌اند.
- ✓ $PR(T_n)$ رتبه صفحه T_n است.
- ✓ $Q(T_n)$ تعداد صفحاتی است که به T_n پیوند داده‌اند.
- ✓ d ضریب تعدیل^۱ که مجموعه‌ای بین صفر و یک است و به صورت اسمی ۰/۸۵ 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، هارون آبادی و رحمانی، ۱۳۹۲). حال به شکل ۲ دقت کنید:



شکل (۳-۲): گراف نمونه (اختر و سرور، 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

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 مهندسين 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^۱ استفاده کرد که خود یک مبحث فنی پیچیده است (طبیعی 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^۲ در کنفرانس ای‌سی‌ام^۳ در سال ۱۹۹۸ ارائه کرد.

۴. الگوریتم HITS: این الگوریتم در حال حاضر در موتور جستجوی اسک^۴ استفاده می‌شود (الگوریتم HITS^۵، 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^۶ و تکرار^۷ 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:

$$H_p = \sum_{q \in I(p)} A_q$$

$$A_p = \sum_{q \in B(p)} H_q$$

در این معادله‌ها:

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

1 Damping factor
2 Akhtar and Sarwar

✓ H_p وزن هاب

✓ A_p وزن نفوذ

✓ B_p و I_p در مورد صفحات ارجاع دهنده و مرجع صفحه P بحث می کند (اختر و

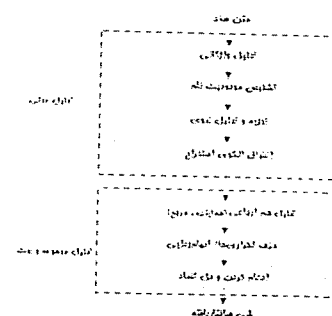
سرور، ۲۰۱۴)

حال برای درک بهتر کارکرد الگوریتم، تصور کنید کاربر پرس و جویی به موتور جستجو ارائه می کند. برای آن، تعدادی از صفحات، که احتمالاً مرتبط ترین نتایج هستند جدا می شوند. این مجموعه را S نام گذاری می کنیم. تمام صفحاتی را که به این مجموعه پیوند داده اند و آنهایی را که صفحات مجموعه S به آن پیوند داده اند G می نامیم. تعداد صفحات مجموعه G می تواند خیلی زیاد باشد. الگوریتم می تواند تعداد صفحات را به عدد خاصی، محدود کند. سپس الگوریتم روی مجموعه G کار می کند و برای هر صفحه، وزن هاب (H_p) و وزن نفوذ (A_p) را با استفاده از معادله های بالا برای تمام صفحات محاسبه کرده و از نتیجه به دست آمده برای رتبه بندی صفحات استفاده می کند (طیبی و دیگران، ۱۳۸۶).

در طول سالیان، خود این الگوریتم ها توسعه یافته اند و هم الگوریتم هایی دیگری مانند PCR عرضه شده اند که هر کدام مزایا و معایبی دارند. مشخص است هر چه قدر وسعت وب افزایش یابد موتورهای جستجو نیز سعی می کنند 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، ۲۰۰۵)

در بخش استخراج متن، ابتدا با تحلیل واژگانی^۱ روبه رو هستیم. هر موتور جستجو، توانایی جستجو بر اساس تحلیل واژگانی و البته تحلیل گرامری^۲ دارد (ما، سانگ، زو و ژانگ^۳، ۲۰۱۰). ایده پایه در تحلیل واژگانی آن است که کلمات، مهم ترین توصیف گر محتوای یک متن اند (آگوستی^۴، ۲۰۰۸). در این بخش متن به تکه هایی مثل جمله و پاراگراف تقسیم می شود و قوانین و الگوهایی به کار می روند تا موجودیت ها در متن شناسایی شوند (اگیچتین^۵، ۲۰۰۵). بعد از تحلیل واژگانی و شکستن متن به موجودیت ها، مرحله تشخیص موجودیت نام یا اسامی^۶ اجرا می شود تشخیص موجودیت نام، ابزاری است برای تشخیص اسامی و نوع آن ها اعم از اسامی افراد، اماکن، مقادیر عددی و غیره. برای تشخیص اسم بودن یک کلمه راه های مختلفی وجود دارد که از جمله، مراجعه به لغت نامه وردنت^۷، توجه به ریشه کلمه، استفاده از قواعد نحوی ساخت واژه و غیره است. در این ابزار، پس از تشخیص اسم ها نوع اسم مشخص می شود (استیری، ۱۳۹۱). در واقع نظام، معمولاً متن را برای کلیدهای^۸ متنی که نشان دهنده یک موجودیت مفید^۹ هستند پوش می کند (آجکتین، ۲۰۰۵). اما مرحله بعد، تحلیل نحوی^{۱۰} است. در این مرحله، برنامه؛ موضوع ها^{۱۱}، فعل ها^{۱۲}، اشیا^{۱۳}، قیدها^{۱۴}، صفت ها^{۱۵} و دیگر متمم های مختلف^{۱۶} را جدا می کند. یعنی بعد از تجزیه و تحلیل واژگانی، تحلیل نحوی در سطح کلمات انجام می شود تا رده کلمه را تشخیص دهد (نعیم و آصف^{۱۷}، ۲۰۱۱). مرحله بعد، انطباق الگوی استخراج^{۱۸} است. برای این بخش، الگوهای استخراج سطح سناریو^{۱۸} برای پی بردن به روابط بین موجودیت های استخراج شده به کار می روند. برخی از نظام ها می توانند 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

جمع آوری 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 می تواند برای رخدادهای همایند چندگانه همان شی^۱ استفاده شود (آجکتین، ۲۰۰۵). مراحل تحلیل واژگانی، تشخیص موجودیت نام، تحلیل نحوی، انطباق الگوی استخراج، مراحل تحلیل محلی^۲ یک متن مدرک محسوب می شوند. مرحله بعد یعنی تحلیل مجموعه^۳ شامل چند مرحله دیگر است که مراحل پس پردازش نامیده می شوند. این مراحل، اطلاعات را به صورت یکپارچه درمی آورند (آجکتین، ۲۰۰۵). مرحله اول در پس پردازش، تحلیل همایندی مرجع^۴ است که در آن مرجع اسمی تعیین می شود. 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 جمله های قبلی آمده، مشخص می کند. استفاده از ضمائر به جای اسامی در زبان انگلیسی بسیار رایج است (استیری، ۱۳۹۱). مرحله بعد، ابهام زدایی و حذف موارد تکراری^۵ است. چالش وجود نام های هم معنی در اطلاعات استخراج شده از وب، "ابهام" در اطلاعات استخراج شده نام دارد و تشخیص نام های هم معنی در موجودیت ها و رابطه های این اطلاعات را ابهام زدایی گویند. به طور معمول، در متن زبان طبیعی، موجودیت ها و روابط دنیای واقعی با نام های متفاوتی استفاده می شوند و سامانه های استخراج اطلاعات باید بتوانند این نام های متفاوت و متعدد را به یک موجودیت و رابطه دنیای واقعی ملحق کنند (ایمانی، ۱۳۹۲). آخرین مرحله برای یکپارچه سازی اطلاعات و رسیدن به یک شی ساختار یافته^۶ ادغام و حل تضادهای موجود در اطلاعات استخراج شده است. منابع اطلاعاتی می توانند همدیگر را تأیید یا تکمیل کنند یا برخی اطلاعات آن ها در تضاد با همدیگر باشد. به طور مثال، اگر دو منبع از یک نویسنده با یک شماره شابک نشان داده شوند، این دو شی یک منبع هستند و همدیگر را تأیید می کنند. 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 می کنند. برای همین دو منبع، اگر تعداد صفحات برای یکی ۲۰۵ صفحه و برای دیگری ۱۰۵ صفحه توصیف شده باشد، نشان می دهد که برخی

ویژگی های این دو منبع با هم در تضادند. 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 می شود (ناومن و هاسلر^۱، ۲۰۰۲). در پایان این مرحله، متن به یک شی ساختار یافته تبدیل می شود.

طبقه بندی کننده

آخرین مرحله از بخش تبدیل متن است و فراداده های مربوط به رده ها^۲ را در مدارک و یا بخش هایی از مدارک شناسایی می کند. کار این بخش در نظر گرفتن برچسب ها برای رده ها و طبقه هایی از اشیاست. این برچسب ها، به طور معمول؛ طبقات موضوعی، مانند ورزش، سیاست و تجارت را نشان می دهند. این تنها نوع طبقه بندی نیست. از انواع دیگر این طبقه بندی می توان به محتوای هرز^۳ یا منابع بی محتوا^۴ مانند تبلیغات اشاره کرد. از فنون دیگر گروه بندی مدارک، فنون خوشه بندی مدارک مربوط است. این فنون از طبقه بندی از پیش تعریف شده استفاده نمی کنند. این گروه های مدارک می توانند به روش های گوناگونی در طول رتبه بندی یا تعامل کاربر استفاده شوند (کرافت و دیگران، ۲۰۱۰). چهار مرحله در طبقه بندی وجود دارد. اول تولید بردار خصوصیت^۵ که بسامد یک کلمه خاص خاص را در متن نشان می دهند. دوم کاهش ابعاد^۶ که کاهش ابعاد بردار خصوصیت با استفاده از ماتریس کواریانس^۷ و فنون پیچیده دیگر است. سوم یادگیری^۸ که از بردار وزنی و از بردار میانگین و نمونه یادگیری ایجاد شده در مرحله قبل استفاده می کند و چهارم طبقه بندی که در آن بردار ویژگی با ابعاد کاهش یافته، در کلاس ها (تابع تفکیک^۹) طبقه بندی می شوند (زو، اوهایما، واکابایاشی و کیمورا^{۱۰}، ۲۰۰۳).

ایجاد نمایه

آخرین بخش از فرآیند نمایه سازی، بعد از مجموعه سازی 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

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

که خود شامل چهار بخش فرعی است. آمارهای مدارک^۱، وزن دهی^۲، مقلوب سازی^۳ و مقلوب سازی^۴ و توزیع نمایه^۵ (ناومن، ۲۰۱۱) در ادامه به کارکرد آنها اشاره شده است.

آمارهای مدارک

همان طور که از نامش پیداست کارهای آماری و کمی را روی اطلاعات انجام می دهد. این آمار می تواند 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 نمایه ای شامل کلمات و خصوصیات پیچیده تر در مدارک و موقعیت آنها در مدرک و در محل رخداد اصطلاح نمایه ای، شمارش رخدادها بر روی گروه های مدارک (مثلاً در گروهی از مدارک که با "ورزش" برچسب گذاری شده اند یا مجموعه ای از مدارک) و طول مدارک از نظر تعداد توکن از انواع داده های مورد نیاز بخش آمارند. آمارها در بخش جداول ارجاعی^۵ نگهداری می شوند و در بخش رتبه بندی به کار می روند (کرافت و دیگران، ۲۰۱۰)

وزن دهی

وزن اختصاص داده شده به اصطلاحات در فایل نمایه است. ساده ترین موتورهای جستجو از وزن دهی دودویی استفاده می کنند. در این نظام ها عدد یک برای حضور یک اصطلاح و عدد صفر برای عدم حضور اصطلاح به کار می رود. موتورهای جستجوی پیچیده تر از طرح های پیچیده تر استفاده می کنند. اندازه گیری بسامد رخداد یک اصطلاح در مدرک با نرمال سازی بسامدها هنوز هم طرح پیچیده ای برای احتساب وزن است. پژوهش های بازیابی اطلاعات به وضوح نشان می دهد که وزن مطلوب با استفاده از "بسامد اصطلاح"^۶ و "بسامد مقلوب مدرک"^۷ به دست می آید. می آید. الگوریتم بسامد، رخداد هر اصطلاح را در درون یک مدرک اندازه می گیرد. سپس الگوریتم، بسامد را با بسامد رخداد در کل پایگاه داده مقایسه می کند. همه اصطلاحات یک مدرک، به خوبی نمی توانند میان یک مدرک با مدرک دیگر یا در کل

پایگاه تمایز قائل شوند. به طور مثال، در پایگاه موضوعی ورزش، ممکن است اصطلاح "آنتی بیوتیک" 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 به صورت گسترده ای در تابع $TF*IDF$ استفاده شده است (رابرتسون^۱، ۲۰۰۴)

بسامد از طریق $\log dkN$ محاسبه می شود و در آن N تعداد مدارک مجموعه و dk تعداد مدارکی است که اصطلاح k در آن ظاهر می شود. فرمول های مختلفی برای محاسبه اوزان اصطلاح وجود دارد که برخی از آنها گونه هایی از وزن IDF بوده و از بسامد مدرک (تعداد دفعاتی که اصطلاحی در مدرک ظاهر می شود) و نرمال سازی بهره می گیرند (بهمن آبادی، ۱۳۸۹). فنون وزن دهی، کاربرد مهمی در رتبه بندی صفحات دارند (ناومن، ۲۰۰۱).

مقلوب سازی

هسته ایجاد نمایه است و کاربرد اصلی آن، تولید نمایه هایی است که می تواند برای بازیابی استفاده شود. بدون نمایه مقلوب، کار جستجوی اطلاعات بسیار مشکل می شود. وقتی اصطلاحات مجموعه مدارک یک پایگاه اطلاعاتی استخراج و به عنوان نماینده های مدرک معرفی می شوند می توانند به چندین صورت فهرست شوند. در یک حالت می تواند فقط الفبایی مرتب شوند. در این حالت وقتی کلیدواژه ای جستجو و با اصطلاح نمایه ای منطبق می شود نمی توان تعیین کرد مدرک حاوی این کلیدواژه کدام است. در نمایه مقلوب، هر کلیدواژه، شناساگر هر مدرک است. بخش مقلوب سازی، اطلاعات را از حالت مدرک - اصطلاح در بخش تبدیل مدرک می گیرد و آن را به اطلاعات اصطلاح - مدرک تغییر می دهد. در اطلاعات بخش تبدیل، هر مدرک با اصطلاحاتی نمایه شده است و در بخش نمایه، اصطلاح معرف هر مدرک می شود و بدین ترتیب نمایه های مقلوب به وجود می آید. برای ایجاد نمایه های مقلوب هم در زمان ایجاد نمایه مقلوب اولیه برای مجموعه بزرگی از مدارک و هم در زمان روزآمدسازی نمایه ها با چالش مواجه هستیم. این بخش موتور جستجو، اهمیت زیادی برای سرعت بخشیدن به پردازش پرس و جو دارد و تا اندازه ای به الگوریتم رتبه بندی

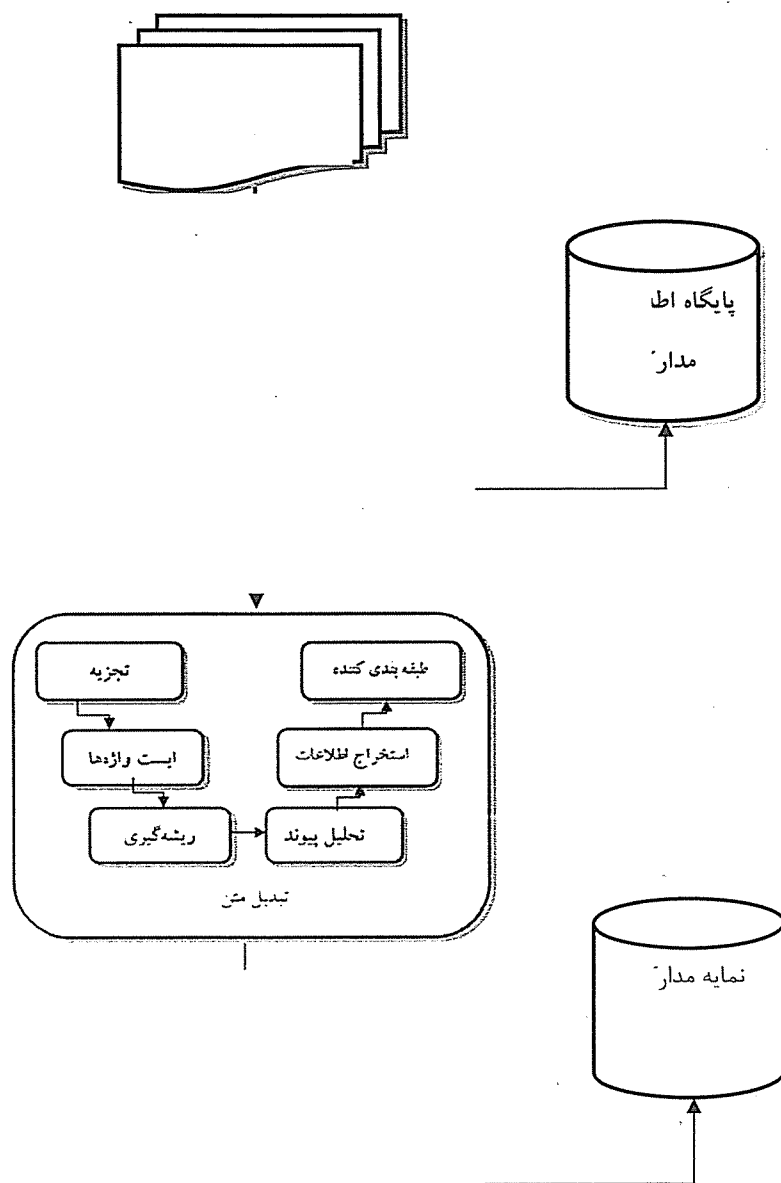
1 Document Statistics
2 Weighting
3 Inversion
4 Index distribution
5 Lookup Tables
6 Term frequency(tf)
7 inverse document frequency(idf)

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 به وجود 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 به صورت 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^۱ 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.

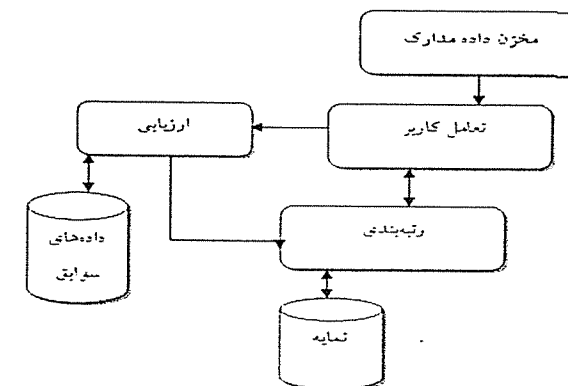


همانطور که قبلاً گفته شد موتور جستجو یک بخش اساسی و مهم دیگر نیز دارد و آن فرآیند پرس و جو است. در ادامه این بخش مهم از معماری موتورهای جستجو معرفی می شود.

فرآیند پرس و جو در موتورهای جستجو

بخش فرآیند نمایه سازی یک سوی بازیابی اطلاعات است و با مدارک موجود در وب سروکار دارد. این بخش از موتور جستجو مدرک را آماده می کند تا به کاربر داده شود. اما کاربری که در پی اطلاعات است پرس و جوی خود را در قالب عبارات و کلیدواژه ها به نظام می دهد. پرس و جوی کاربر در بخش فرآیند پرس و جو تجزیه و تحلیل و برای موتور جستجو قابل خواندن می شود. موتور جستجو آن چه را که ارائه شده با اصطلاحات نمایه ای در پایگاه اطلاعاتی خود تطبیق می دهد. سپس بر اساس الگوریتم رتبه بندی، فهرستی از مدارک مرتبط، به کاربر داده می شود. تجزیه و تحلیل و آماده سازی پرس و جوی کاربر، پیچیده است.

فرآیند پرس و جو شامل سه بخش مهم است. تعامل کاربر^۱، رتبه بندی^۲ و ارزیابی^۳ (کرافت و دیگران، ۲۰۱۰). در شکل زیر فرآیند پرس و جو و بخش های مختلف آن نشان داده می شود.



شکل (۲-۶): بخش های اصلی فرآیند پرس و جو

تعامل کاربر

وظیفه این بخش از لحظه ای شروع می شود که کاربر، درخواست خود را به موتور جستجو می دهد و با مشاهده نتایج تمام می شود. این بخش، خود شامل سه بخش فرعی است. درونداد پرس و جو^۱، تبدیل پرس و جو^۲ و برونداد نتایج^۳.

درونداد پرس و جو

این بخش برنامه رابط کاربری و یک تجزیه کننده است و با زبان پرس و جو سروکار دارد (کرافت، ۲۰۱۰). قبل از بحث زبان پرس و جو باید بدانیم کاربران به طور عمده دنبال چه اطلاعاتی هستند و به خاطر آن از موتورهای جستجو استفاده می کنند. گیبسون^۴ پرسش های کاربران را به سه نوع تقسیم کرده است. انجام دادن^۵، دانستن^۶ و رفتن^۷.

✓ پرسش های "انجام دادن": پرسش های محتوای تراکشی^۸ نیز نامیده می شوند. کاربران در این پرسش ها می خواهند کاری انجام دهند. در واقع تمایل کاربر را برای اجرای یک عمل نشان می دهند؛ مانند خرید ماشین، بلیط سینما یا یک کتاب.

✓ پرسش های دانستن: پرسش های محتوای اطلاعاتی^۹ نامیده می شوند. کاربران می خواهند چیزی را بدانند و از آن آگاهی یابند. این پرسش ها حوزه وسیعی از اطلاعات را پوشش می دهند و هزاران نتیجه مرتبط با درخواست آن ها در وب وجود دارد. در واقع کاربران دنبال پاسخ های سریع و آسانند.

✓ پرسش های رفتن: پرسش های محتوای اکتشافی^{۱۰} نیز نام دارند. کاربران، دنبال یک یا چند صفحه از یک موجودیت وب هستند یا به دنبال یک وب سایت واحد می گردند. به طور مثال یک سایت مانند یوتیوب یا یک مقاله (گیبسون^{۱۱}، ۲۰۱۳). مور^{۱۲} نوع چهارمی از پرسش را نیز معرفی می کند به نام پرسش های اتصال^{۱۳}.

✓ پرسش های اتصال: اتصالی از گراف وب نمایه شده را گزارش می کنند.

برای مثال کدام پیوند به این نشانی اینترنتی اشاره می کند یا چه تعداد صفحات از این نام دامنه^۱ نمایه شده اند (مور، ۲۰۰۸)

پرسش کاربر می تواند از ساده ترین تا پیچیده ترین شکل باشد. رابط کاربری که این پرسش ها را به شکل قابل درک برای موتور جستجو تبدیل می کند، باید قابلیت های فراوانی داشته باشد. زبان پرس و جو در این میان نقش به سزایی دارد. زبان های پرس و جو به کاربران اجازه می دهند تا شکل های خاصی از پرس و جو را مطرح کنند. به طور مثال می توانند از نظام درخواست کنند اصطلاحات نزدیک به هم یا با نظم خاص در یک مدرک بازیابی شود. حتی زبان های جستجو باعث می شوند کاربران بتوانند اهمیت اصطلاحات و عبارات را با وزن آن ها، تعیین کنند یا از نظام بخواهند برخی عبارات در کنار هم و با هم استفاده شوند تا پرس و جو دقیق باشد (استرامن^۲ و دیگران، ۲۰۰۵). استفاده از عملگرها معمول ترین فوننی است که در زبان های پرس و جو استفاده می شود تا کاربران بتوانند پرس و جوی خود را فرمول بندی کنند. عملگرهای بولی یا عملگرهای مجموعه ای شامل اجتماع (و)، اشتراک (یا) و تفاضل (به جز) هستند و با کمک سایر عملگرها مانند علامت نقل قول، علامت * یا # می توانند در جستجوهای خاص تر به کاربران کمک کنند. تمامی این ویژگی ها با استفاده از زبان پرس و جو در نظام بازیابی اطلاعات و به ویژه موتور جستجو میسر می شود.

تبدیل پرس و جو

فوننی مانند بازدارندگی واژگان و ریشه گیری در این بخش نیز کاربرد دارند (کرافت و دیگران، ۲۰۱۰). فتون دیگر مانند کنترل املائی کلمات و فهرست پیشنهاد های پرس و جو نیز از فتون تبدیل متن است. به طور مثال در موتور جستجوی گوگل هر زمان واژه ای با املائی غلط جستجو شود، موتور گوگل به صورت خودکار کلمه را اصلاح و بعد با کلمه اصلاح شده جستجو می کند و نتایج را در صفحه نتایج نشان می دهد یا به محض تایپ یک واژه، فهرستی از واژگان و عبارت ها را پیشنهاد می کند. بازخورد ربط^۳ نیز در این بخش از پرس و جوست که مسأله مهمی است. پورقاسم و قاسمیان (۱۳۸۸) دو روش عمده در بازخورد ربط را حرکت نقطه پرس و جو و اصلاح وزن می دانند.

1 Domain name
2 Turtle
3 Relevance Feedback

۱- حرکت نقطه پرس و جو: در این روش سعی می شود با حرکت نقطه جستجوی جاری به کمک بازخورد کاربر، تخمین نقطه جستجوی ایده آل (به صورت ویژگی های سطح پایین) بهتر شود.

۲- اصلاح وزن ها: در این روش، وزن ها و پارامترهای مورد استفاده در تعیین میزان تشابه بر اساس بازخورد کاربر اصلاح می شود (پورقاسم و قاسمیان، ۱۳۸۸). موضوع مهم دیگر در بخش تبدیل پرس و جو که اغلب در کنار بازخورد ربط مطرح می شود موضوع گسترش پرس و جو^۱ است. برای فهم بهتر این مسأله به یکی از مسائلی که در نظام های بازیابی اطلاعات وجود دارد اشاره می کنیم و آن کلمات مترادف^۲ است که می تواند بر دقت و بازیافت اطلاعات تأثیر بگذارد. برای حل این مشکل دو نوع روش وجود دارد: روش های جهانی^۳ و محلی^۴. در روش های جهانی، گسترش پرس و جو و فرمول بندی مجدد اصطلاحات پرس و جو، مستقل از پرسش و نتایج ارائه شده است. به طوری که تغییرات در جمله بندی پرس و جو باعث خواهد شد پرس و جوی جدید با واژه های دیگر دارای معانی مشابه مطابقت پیدا کند. روش های جهانی عبارتند از گسترش پرس و جو / فرمول بندی مجدد با اصطلاح نامه^۵ یا وردنت^۶، گسترش پرس و جو از طریق تولید اصطلاح نامه خودکار و فوننی مانند تصحیح املا. اما در روش محلی، پرس و جویی در مورد مدارک تنظیم می شود که در ابتدا به نظر می رسد مطابق با پرس و جوی کاربر است. بازخورد ربط^۷، ربط^۸، بازخورد شبه ربط یا بازخورد ربط کور^۹ و بازخورد ربط غیرمستقیم^۹ از انواع انواع روش های محلی است (میننگ، ۲۰۰۸، الف).

برونداد نتایج

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

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 نسخه های مختلف و امکان دسترسی به آنها در هر فهرست نتیجه در گوگل و خوشه بندی نتایج در گروه های موضوعی.

رتبه بندی

این بخش بسیار به مدل های بازیابی وابسته است که در فصل سوم مفصل در مورد آن صحبت می شود. رتبه بندی از سه بخش فرعی تشکیل شده است. نمره دهی^۱، بهینه سازی^۲ و توزیع^۳.

نمره دهی

روشی برای بهبود موتورهای جستجوی وب است و این کار را با اختصاص دادن نمره بر اساس میزان اهمیت صفحات انجام می دهد. الگوریتم های رتبه بندی در این نمره دهی اهمیت زیادی دارند. مثلاً الگوریتم رتبه بندی صفحات گوگل یکی از شناخته ترین الگوریتم های رتبه بندی است (ایکاو و ساداکانی^۴، ۲۰۰۴). وزن دهی بر اساس وزن اصطلاح پرس وجو (q_i) و وزن اصطلاح در مدرک (d_i) محاسبه می شود. معادله محاسبه این امتیاز (امتیاز وضعیت بازیابی^۵) به صورت زیر است (رسلوف و ساوی^۶، ۲۰۰۳):

$$RSV = \sum_i q_i d_i$$

بهینه سازی

موضوع بهینه سازی در موتورهای جستجو مهم و اساسی است. موتور جستجویی مانند گوگل، چهار مکانیسم اساسی دارد: کشف، ذخیره سازی، رتبه بندی و ارائه نتایج؛ که مهم ترین مسائل بهینه سازی اند. بهینه سازی نیازمند فهم عمیق از کشف و رتبه بندی در موتورهای جستجو است (دیویس، ۲۰۰۶). الگوریتم های رتبه بندی صفحات که نقش اساسی در بهینه سازی دارند، در بخش های قبل آمده اند.

توزیع

قبلاً در مورد نمایه توزیع شده در بخش فرآیند نمایه سازی صحبت شد. این بخش در فرآیند پرس وجو نیز هست. پرسش به نظام داده می شود ولی در درون نظام به

همان شکل پذیرفته نمی شود؛ بلکه بعد از تجزیه و تحلیل برای نظام قابل خواندن می شود. تبدیل کردن به توکن، حذف کلمات بی ارزش و ریشه گیری روی پرسش انجام می شود. در نظام، نرم افزار کارگزار پرس وجوی^۱ موجود در بخش توزیع فرآیند، این وظایف را به عهده دارد (بیزا-ییتس، ورین، زاراگوزا، کامبازوگلی و موردادک، ۲۰۱۲).

توزیع به راهبردهای بازیابی مختلف نیازمند است که بر عهده کارگزار پرس وجو است. کارگزار، پرس وجو را به همه خدمت دهنده ها می برد، نتایج بخش ها را در هم ادغام می کند و نتایج نهایی را به کاربر ارائه می دهد (منینگ و دیگران، ۲۰۰۸ ب)

ارزیابی

هیچ نظامی بدون ارزیابی و نظارت نمی تواند به پویایی و تکامل خود ادامه دهد. بازخوردی که از نتایج ارزیابی گرفته می شود می تواند به تغییر راهبردها و افزایش کارایی و اثربخشی نظام ها از بیرون و درون کمک کند. ارزیابی بیرونی نظام از دیدگاه کاربران است و ارزیابی درونی، مکانیسم هایی است که نظام برای سنجش عملکرد خود در ارائه نتایج به کار می گیرد. این نوع ارزیابی در موتورهای جستجو شامل سه بخش فرعی است؛ گزارش سوابق^۲، تحلیل رتبه بندی^۳، و تحلیل عملکرد^۴.

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 نقطه نظر ارزیابی، داده های به دست آمده از این بخش نشان می دهد که کاربران چگونه نتایج ارائه شده در موتور جستجو را مرور می کنند (کرافت و دیگران، ۲۰۰۸). سوابق می توانند برای اصلاح املاهای کلمات، فهرست پیشنهادها، ذخیره پرس وجو^۵ و کمک به به انطباق تبلیغات با جستجو استفاده شوند (کرافت و دیگران، ۲۰۰۸). یکی از فرض ها در این بخش آن است که صفحات کلیک شده به پرس وجوی کاربر مربوط اند (تاومن، ۲۰۱۱). این داده ها برای بهبود کارایی و اثربخشی موتور جستجو

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

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

اهمیت زیادی دارند.

تحلیل رتبه بندی

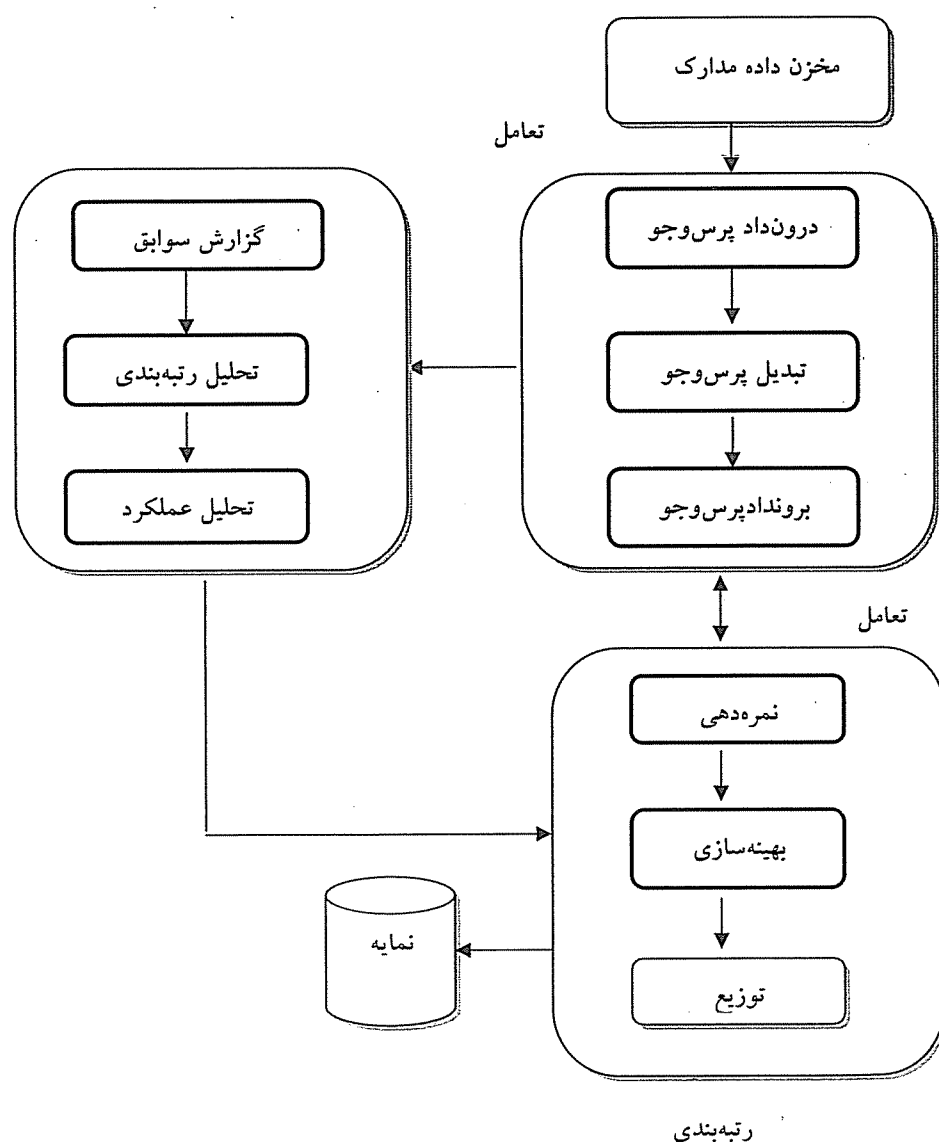
همان طور که قبلاً نیز اشاره شد، در بخش فرآیند نمایه سازی و همچنین پرس و جو از الگوریتم های رتبه بندی برای مدارک و پرس و جو استفاده می شود. در این بخش، رتبه جفت مدارک و پرس و جو تحلیل و ارزیابی می شود (کرافت، ۲۰۱۰) و به ارزیابان فرصت می دهد تا وضعیت موتور جستجوی خود را بررسی کنند. در واقع تحلیل رتبه بندی، اثربخشی موتورهای جستجو را بررسی می کند. معیارهای زیادی برای ارزیابی رتبه بندی و الگوریتم های رتبه بندی موتور جستجو استفاده می شوند. مجموعه های آزمون در این بخش اهمیت زیادی دارند. در بخش معیارهای ارزیابی به طور مفصل به آن اشاره می شود.

تحلیل عملکرد

اگر تحلیل رتبه بندی، اثربخشی موتور جستجو را بررسی می کند، تحلیل عملکرد، ارزیابی کارایی است و در آن عملکرد کلی موتور جستجو به طور کامل تجزیه و تحلیل می شود.

طراحان و مهندسان برای به دست آوردن عملکرد بهتر، طرح های جدیدی ایجاد نمی کنند. بلکه شبیه سازی^۱ می کنند و آزمایش ها را روی همان نمونه های اولیه انجام می دهند (کرفت و دیگران، ۲۰۰۸) موضوع شبیه سازی و مدل سازی از روی نمونه های موجود و ارزیابی عملکرد، مسأله بسیار مهمی در ارزیابی کارکرد نظام هاست.

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



شکل (۷-۲): بخش های اصلی فرآیند نمایه سازی در موتور جستجو

خلاصه فصل

موتورهای جستجو از اساسی ترین ابزارهای جستجو و دسترسی به اطلاعات اند و به بخشی اساسی از زندگی روزمره افراد تبدیل شده اند. در این برنامه های نرم افزاری پیچیده، که گوگل یکی از مهم ترین آن ها است کارایی و اثربخشی موضوع بسیار مهمی است. هر موتور جستجو از دو بخش اصلی تشکیل شده است؛ فرآیند نمایه سازی و پرس و جو. هر بخش، خود شامل بخش های فرعی مختلفی است. فرآیند پرس و جو شامل مجموعه سازی متنی، ایجاد نمایه و تبدیل متن است. نمایه سازی نیز شامل تعامل کاربر، رتبه بندی و ارزیابی است. هر کدام از این ها خود شامل بخش های فرعی متعددی هستند که یک موتور جستجو را به یک برنامه نرم افزاری پیچیده تبدیل می کند. الگوریتم های رتبه بندی هم در بخش نمایه سازی و هم پرس و جو بر کارکرد کلیه بخش ها تأثیر گذارند و می توانند عملکرد آن ها را نیز تحت الشعاع قرار دهند. کارایی و اثربخشی موتورهای جستجو به کارکرد صحیح و درست تک تک این اجزا بستگی دارد. در حال حاضر کمتر طراحی نمونه های کاملاً جدید برای نظام های بازیابی و ارزیابی در دستور کار است و مهندسان بیش تر با استفاده از شبیه سازی و آزمایش روی نمونه های اولیه، موتورهای جستجو را ارزیابی می کنند و آن ها را ارتقا می دهند.

فصل سوم

مدل های بازیابی اطلاعات

مقدمه

نیاز اطلاعاتی کاربران بیش از آن که به صورت 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 می ماند. به دلایل زیادی افراد نیاز اطلاعاتی خود را بیان نمی کنند. شاید یکی از دلایل مهم این است که نیاز اطلاعاتی کاربر در آن لحظه یا بعد از آن اهمیت حیاتی ندارد یا نیاز آن قدر قوی نیست که کاربر زحمت اطلاع جویی 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 محیط های کاری و یا علمی و آکادمیک، سوالی علمی برای فرد ایجاد می شود که دنبال کردن آن، می تواند دانش افزایی 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 می شود فرد از اطلاع جویی پرهیز کند. نبود منابع برای تأمین تقاضای مورد نظر کاربر، پیچیدگی منابع موجود برای اطلاع جویی و مواردی از این قبیل، نیاز اطلاعاتی را در حد یک ابهام یا یک پرسش بی پاسخ در ذهن فرد می گذارد. اما یکی از مهم ترین دلایل آن است که کاربر نمی تواند ابهام یا پرسش خود را به صورت قابل فهم بیان کند. یکی از مسائل مهم اطلاع جویی، مهارت فرد در