مجله کتابداری

سال سی و نهم، دفتر چهل و سوم

بهار و تابستان ۱۳۸۴، ص۱۱۹-۱۵۱

Journal of Librarianship

Vol 39, No. 43, Spring & Summer 2005

PP.119-151

مروری بر رویکردهای نمایه سازی خودکار وب: محتوامحوری، استناد محوری و معنا محوری

عاطفه شریف

دانشجوی کارشناسی ارشد کتابداری و اطلاع رسانی دانشگاه تهران

چکیده

وب به واسطه ویژگی چند رسانه‌ای، کاربران فراوانی در اینترنت دارد. شمار وب سایت‌ها روز به روز افزایش می‌یابد و انبوهی از اطلاعات در وب منتشر می‌شود. در چنین وضعیتی مهم‌ترین مسئله، سازماندهی و مدیریت اطلاعات منتشر شده است؛ زیرا کیفیت بازیابی در گرو سازماندهی و ذخیره سازی مناسب است. موتورهای کاوش، با نمایه سازی و ذخیره اطلاعات نمایه شده در پایگاه‌های خود، امکان جست و جو، بازیابی، رتبه‌بندی، و نمایش اطلاعات وب را فراهم می‌آورند. در این مقاله سعی بر آن است تاضمن معرفی مختصر برخی ابزارهای کاوش، به تشریح رویکردهای موجود- محتوا محوری، استناد محوری، و معنا محوری - در زمینه نمایه سازی خودکار وب در موتورهای کاوش پرداخته شود. در نهایت حرکت به سمت داده‌های ساختار یافته و وب معنایی[1] با تکنولوزی‌های XML[2] و RDF[3] مورد بررسی قرار می‌گیرد.

کلید واژه‌ها: وب. نمایه سازی خودکار. صفحات متنی

---

1. Semantic Web
2. Extensible Markup Language(زبان نشانه‌گذاری توسعه پذیر)
3. Resource Description Framework(چارچوب توصیف منبع)

مقدمه

وب به عنوان یکی از جذاب‌ترین بخش‌های اینترنت که کاربران فراوانی دارد، مجموعه‌ای است از صفحات به هم پیوسته که حاوی اطلاعات مفیدی در زمینه‌های موضوعی مختلف است. مطالعه‌ای که cyveillance. com انجام داده است نشان می‌دهد که وب بیش‌از ۱/۲ بیلیون سند دارد (اکتبر ۲۰۰۴) و نرخ رشد آن نیز ۷ میلیون صفحه در روز است (رودکی، ۱۳۸۳).

همهٔ دلایل مربوط به چرایی سازماندهی اطلاعات در محیط چاپی با شدت بیشتری در محیط الکترونیکی صادق است. میلیون‌ها نفر در سراسر جهان نیازهای روزمره خود را از طریق وب مرتفع می‌کنند. سرعت رشد وب به حالت انفجارگونه‌ای رسیده است. مسئله اساسی در چنین وضعیتی آن است که چگونه می‌توان بدنه ساختار نیافته[1] و به سرعت رشد کننده وب را تحت کنترل و مدیریت درآورد. اگرچه در سازماندهی وب از روش‌های پایه و اصول سنتی ذخیره و بازیابی اطلاعات استفاده شده است، روشن است که روش‌های سنتی قابلیت لازم را ندارد. در حال حاضر موتورهای کاوش، به‌ظاهر، وب را تحت ضابطه در آورده‌اند(اسدی و جمالی مهموئی، ۲۰۰۴) و با نمایه سازی صفحات پاسخی برای پرس و جوی[2] کاربران فراهم می‌آورند.

نحوه سازماندهی اطلاعات، بازتاب مستقیمی در نحوه بازیابی(رتبه‌بندی و نمایش) نتایج دارد. موتورهای کاوش هریک با الگوریتم‌ها و سیاست‌های متفاوتی به مقولهٔ نمایه سازی می‌نگرند. تفاوت در نتایج بازیابی شده در موتورهای کاوش مختلف نشان بارزی از وجود تفاوت در الگوریتم‌های نمایه سازی آنهاست. سازماندهی و سپس ذخیرهٔ مناسب اطلاعات به بازیابی نتایج مرتبط و مناسب خواهد انجامید. از این جهت است که آشنایی با این مقوله برای کتابداران و اطلاع رسانان که از دیرباز به انتخاب، گردآوری، سازماندهی، ذخیره، بازیابی، و اشاعه اطلاعات پرداخته‌اند خالی از فایده نیست.

---

1. Unstructured
2. Query

مسئله و هدف

ابزارهای کاوش که ظاهراً وب را تحت ضابطه درآورده‌اند از روش‌های گوناگونی جهت نمایه سازی اطلاعات بهره می‌برند. بدون شک هریک از روش‌های اتخاذ شده در نمایه سازی مزایا و معایبی دارد که به واسطه انجام پژوهش و آزمون می‌توان به این نقاط ضعف و قوت پی برد و جهت رفع ضعف‌ها کوشید. تلاش‌ها در این زمینه عمدتاً از سوی متخصصان علوم رایانه و هوش مصنوعی انجام پذیرفته است. این در حالی است که همواره کتابداران و اطلاع رسانان مدعی سازماندهی دانش بشر بوده و هستند. حضور پررنگ‌تر در این حوزه نیازمند آشنایی با اصول پایه است. برداشتن گام اول تنها به معنای برداشتن گام اول است و در این مسیر باید گام‌های فراوانی برداشته شود. این از آن جهت اهمیت دارد که متخصصان سایر علوم اگرچه اطلاعات فنی مناسبی جهت طراحی ابزارهای کاوش دارند؛ متخصصان کتابداری و اطلاع رسانی نیز دارای دانش ارزشمندی در زمینه نیازسنجی، رفتارهای اطلاع‌یابی، و روابط حاکم بر پرسش و پاسخ در مرحله پاسخ‌یابی هستند. موفقیت در این زمینه زمانی حاصل خواهد شد که متخصصان اطلاع رسانی همکاری بیشتری با سایر متخصصان داشته باشند. حضور فعال، انجام پژوهش و ارائه راهکار به سایر متخصصان نیازمند دانشی است که کتابداران و اطلاع رسانان باید از واقعیت وب داشته باشند. لذا مطالعهٔ حاضر بر آن است تا ضمن آشکار ساختن چگونگی نمایه سازی خودکار اطلاعات در محیط وب، رویکردهای موجود را تا حد امکان آشکار سازد.

هدف از انجام این پژوهش رسیدن به درکی واقعی از چگونگی نمایه سازی خودکار در وب و رویکردهای موجود در این زمینه است. انجام این پژوهش می‌تواند فواید زیر را در برداشته باشد:

- رسیدن به درکی از واقعیت نمایه سازی در محیط وب؛
- شناخت رویکردهای موجود در نمایه سازی وب و حرکت‌های آتی؛
- بهره‌گیری از دانش کسب شده جهت پاسخگویی کارآمدتر به کاربران و مراجعان کتابخانه‌ها و مراکز اطلاع رسانی؛

● کسب دانشی که در طراحی و مدیریت وب سایت‌ها و پایگاه‌های اطلاعاتی پیوسته به کتابداران یاری رساند؛

● درک اهمیت مشارکت متخصصان حوزه‌های علوم رایانه، کتابداری و اطلاع رسانی و حوزه هوش مصنوعی و کسب دانشی که این همکاری و مشارکت را تا اندازه‌ای عملی‌تر نماید؛

● کسب آمادگی جهت انجام پژوهش‌های بیشتر در زمینه ارزیابی عملکرد ابزارهای کاوش به خصوص موتورهای کاوش و ارائه راهکارهای عملی.

روش پژوهش حاضر از نوع کتابخانه‌ای است. و از طریق مروری بر نمایه‌سازی در وب و رویکردهای موجود در متون صورت گرفته است.

محدودیت ها

الف. گردآوری اطلاعات

آنچه واقعاً در پشت صحنه جست و جو رخ می‌دهد، به دلایل رقابتی مخفی نگه‌داشته می‌شود و فقط در اختیار تولید کننده خدمات جست و جو است. بسیاری از اطلاعات فنی ارائه نمی‌شود و یا در صورت ارائه شدن از روزآمدی آنها کاسته شده است. لذا در تحقیق حاضر سعی در اتکا بر متونی شده است که تا اندازه‌ای جنبه‌های فنی نمایه سازی را آشکار ساخته‌اند و در ضمن از روزآمدی نسبی برخوردارند. مقالاتی نیز موجود است که تنها امکان دسترسی به چکیده‌های آنها فراهم است و متن کامل آنها در اختیار نیست. به‌این ترتیب ادعایی مبنی بر جامعیت مطالب ارائه شده وجود ندارد.

ب. درک دقیق اطلاعات بازیابی شده

بحث نمایه سازی وب از یک سو به اطلاع رسانی و از سویی به علوم رایانه مربوط می‌شود. این در حالی است که اکثر متون مناسب در حوزه علوم رایانه به نگارش درآمده است، این متون حاوی مفاهیم و لغات تخصصی ویژه‌ای هستند و درک دقیق آنها نیازمند مطالعه و تأمل بسیار است.

پیشینه

الف. خارج از ایران

بسیاری از مقالات موجود در زمینه فن‌آوری موتورهای کاوش در حوزهٔ کامپیوتر به نگارش درآمده‌اند. متأسفانه دسترسی به متن کامل تعداد زیادی از این مقالات تحقیقاتی، که اغلب گزارش طراحی موتورهای کاوش و طرح‌های ذخیره و بازیابی است، امکان‌پذیر نیست. طی جست و جوی انجام شده در پایگاه‌های اطلاعاتی دانشگاه تهران و بررسی نتایج حاصل از کاوش پیشرفته گوگل، تحقیقی با این رویکرد بازیابی نشد.

لنکستر(۲۰۰۳) در مقاله‌ای با عنوان."آیا نمایه سازی و چکیده‌نویسی آینده‌ای را پیش رو خواهد داشت؟" به بحث نمایه سازی و چکیده نویسی گزینشی در محیط وب می‌پردازد. وی ضمن تأیید توانایی نیروی انسانی متخصص در این زمینه وجود نظام‌های خودکار نمایه سازی بر آینده‌ای روشن تأکید می‌کند که: "احتمالاً سرانجام، رده‌بندی خودکار هم قدرت و هم شهرت متخصصان اطلاع رسانی را ارتقاء خواهد داد."

ونگ و برمان[1](۲۰۰۳) ضمن تأکید بر وجود وب نامرئی ساختار موتور کاوش پویایی را شرح می‌دهند که نمایه سازی در آن طبق رویکرد محتوا محور است. در این موتور کاوش پویا از فن‌آوری XML استفاده شده است

برین و پیج[2](۱۹۹۸) با تحلیل دقیق آناتومی گوگل، آن را به کاربران وب معرفی می‌کنند. موتور کاوش گوگل از تحلیل پیوندها بهره می‌گیرد و، علاوه بر نمایه سازی کلید واژه‌ای کلمات متن، بر تحلیل پیوندها نیز تأکید دارد.

برادشاو و هاموند[3](۲۰۰۲) نمایه سازی خودکار مدارک را با دو روش محتوا محور و براساس استنادها مورد بررسی قرار داده‌اند. در این مقاله تأکید بر نمایه سازی مدارک مطابق با استنادهاست به گونه‌ای که استنادها اساس نمایه سازی مدارک قرار گیرند. این

---

1. Wang and Behrmann
2. Brin and Page
3. Bradshaw and Hammond

دو محقق موتور کاوشی بر این مبنا طراحی کرده‌اند که در کتابخانه دیجیتال رزتا[1] به‌طور آزمایشی مورد استفاده قرار گرفته است.

دمونتیل و ژاکین(۲۰۰۲) استفاده از هستی‌شناسی را به عنوان مبنایی در نمایه سازی وب سایت‌ها معرفی کرده‌اند. در فرایند نمایه سازی بر مبنای هستی‌شناسی، ارزش گذاری بر طبق نوع نشانه‌ها و سایر مفاهیم مرتبط با صفحه انجام می‌پذیرد و ارزش نهایی در نمایه‌ای ساختار یافته به فرمت XML ذخیره می‌شود.

ب. داخل ایران

در زمینه مطالعه روش‌ها یا رویکردهای نمایه سازی وب در منابع فارسی کار چندانی انجام نپذیرفته است. طبق جست و جوی انجام شده چنین به نظر می‌رسد که هیچ پایان‌نامه یا طرح تحقیقاتی به بررسی نمایه سازی در وب نپرداخته است.

کمیجانی(۱۳۸۱) در مقاله‌ای با عنوان"ساختار نمایه سازی در موتورهای کاوش وب" مدل‌های متمرکز[2] و توزیعی[3] نمایه سازی را مورد بررسی و مقایسه قرار داده و برتری‌های مدل توزیعی را برشمرده است.

اسدی و جمالی مهموئی(۲۰۰۴) ضمن مطالعه سیر تحول موتورهای کاوش، گرایش‌های مطالعاتی موجود در این زمینه را بیان کرده‌اند. در این تحقیق موتورهای کاوش هوشمند که توانایی استخراج نتایج از داده‌های ساختار یافته را دارند، به عنوان نسل بعدی موتورهای کاوش معرفی شده‌اند.

صفری(۱۳۸۳)با افزودن چهار ابرداده[4] از مجموعه ابرداده‌های هسته دابلین[5] ـ عنوان، موضوع، پدیدآورنده، و همکار ـ به مطالعهٔ اثربخشی هریک از این ابرداده‌ها پرداخته است. تجزیه و تحلیل آماری در این پایان‌نامه نشان می‌دهد که هیچ‌گونه تفاوت

---

1. Rosetta Digital Library
2. Centralized model
3. Distributed model
4. Metadata
5. Dublin core

معناداری میان دو گروه آزمودنی و گواه در بهبود بازیابی صفحات وب وجود ندارد و فرضیه‌های پژوهش مبنی بر اثربخشی هریک از عناصر ابرداده‌ای بر بازیابی صفحات وب تأیید نگردیده است.

سایر اطلاعات منتشر شده در زمینه نمایه سازی وب بسیار کلی و، بیشتر در قالب کتاب‌ها و مقالات ترجمه شده در زمینهٔ راهبردهای کاوش و چگونگی ارزیابی نتایج کاوش است. در بیشتر این منابع ساختار اصلی موتورهای کاوش به عنوان پر استفاده‌ترین ابزار کاوش به‌طور مختصر مورد بررسی قرار گرفته است. به این ترتیب، در زمینه رویکردهای نمایه سازی خودکار در منابع فارسی نه در حوزه کتابداری و نه در حوزه کامپیوتر مطلبی یافت نشد.

ابزارهای کاوش[1] در وب

رشد سریع وب سایت‌ها، مراجعه فراوان کاربران به وب جهت رفع نیازهای اطلاعاتی و بسیاری دلایل دیگر متخصصان را بر آن داشته است تا به روش‌های گوناگون، دسترس پذیری به اطلاعات وب را فراهم آورند.

ابزارهای کاوش پایگاه‌های اطلاعاتی قابل جست و جو یا مرور هستند که با استفاده از آنها می‌توان به بخشی از اطلاعات موجود در اینترنت دست یافت. این ابزارها به دو شیوه "جست و جوی کلید واژه‌ای" یا "مرور و انتخاب" پیوندهای فرامتنی، کاربران را به سوی اطلاعات مورد نظر هدایت می‌کنند(کوشا، ۱۳۸۰، ص ۲۶).

ابزارهای کاوش را می‌توان به دو نوع اصلی تقسیم بندی کرد:

راهنماهای موضوعی[2] و موتورهای کاوش[3].

راهنماهای موضوعی پایگاه‌هایی هستند که اطلاعات صفحات یا سایت‌های وب

---

1. Search tools
2. Subject directories
3. Search engines

منتخب را توسط نیروی انسانی در پایگاه خود نمایه می‌کنند. این راهنماها می‌توانند عمومی یا خاص باشند(کوشا، ۱۳۸۰، ص ۲۶).

موتورهای کاوش پایگاه‌های اطلاعاتی قابل جست و جویی هستند که از طریق برنامه‌های کامپیوتری به شناسایی و نمایه سازی خودکار صفحات وب می‌پردازند. موتورهای کاوش برخلاف راهنماهای موضوعی، برنامه‌های خودکاری هستند که هیچ گونه وابستگی به نیروی انسانی ندارند(کوشا، ۱۳۸۰، ص ۳۱).

علاوه بر دو نوع اصلی ابزارهای کاوش، ابزارهای دیگری نیز برای دسترسی به اطلاعات در محیط وب وجود دارد که عبارتند از:

- ابزارهای کاوش دو وجهی[1] که ترکیبی از موتور کاوش و راهنمای موضوعی هستند، مانند گوگل و آلتاویستا.
- ابر موتورهای کاوش[2] که ترکیبی از چندین موتور کاوش و راهنمای موضوعی و فاقد پایگاه مستقلی جهت جمع‌آوری و نمایه سازی هستند.
- نرم‌افزارهای کاوش که هیچ گونه وابستگی به محیط وب ندارند و روی کامپیوتر شخصی نصب می‌شوند. این نرم افزارها امکان جست و جوی همزمان در چندین پایگاه اطلاعاتی رادر اختیار قرار می‌دهند و از این لحاظ ابر موتور کاوش به شمار می‌آیند(کوشا، ۱۳۸۰، ص ۴۶).
- حلقه‌ها[3] مجموعه‌ای از سایت‌های هم‌موضوع هستند که توسط نرم‌افزار Web ring به‌هم متصل شده‌اند. این حلقه‌ها گاه مورد ارزیابی قرار می‌گیرند و باتوصیفی همراه می‌شوند.
- درگاه‌ها[4] سایت‌هایی هستند که خود را سرآغاز و نقطه شروع وب می‌شمارند. این درگاه‌ها می‌توانند موضوعی نیز باشند.

---

1. Hybrid search tools
2. Meta search engines
3. Web rings
4. Portals or Gateways

●فن‌آوری فشار[1] صفحات وب را قبل از درخواست به کاربران ارسال می‌کند.
●کارگزاران هوشمند[2] نرم‌افزارهایی هستند که مطابق با نیازهای فردی به جمع‌آوری اطلاعات می‌پردازند، درست همانند موتورکاوشی که مطابق با نیازهای فردی به جست و جو بپردازد(دوشارته،۲۰۰۲، ص ۶۸-۶۹). تامپسون و دیگران[3](۲۰۰۰) کارگزاران هوشمند را نسل چهارم ابزارهای بازیابی می‌نامند که تلاش انسان را در پاسخ‌یابی بسیار کاهش داده است(منصوریان، ۲۰۰۴، ص ۲۱۹).

از میان ابزارهای کاوش معرفی شده، تنها موتورهای کاوش نظام نمایه سازی خودکار خاصی را دنبال می‌کنند؛ لذا در ادامه به بررسی نحوه نمایه سازی در این نوع ابزارهای کاوش پرداخته می‌شود.

۱. فن‌آوری موتورهای کاوش

نمایه سازی با یکی از دو روش دستی یا خودکار انجام می‌پذیرد. در محیط وب به علت حجم وسیع اطلاعات منتشر شده عملاً نمی‌توان به شیوه‌های دستی متوسل شد. به این ترتیب، نمایه سازی در وب به صورت خودکار و توسط موتورهای کاوش انجام می‌شود. موازی با رشد وب، تعداد موتورهای کاوش رو به افزایش است. تفاوت نتایج بازیابی شده در موتورهای کاوش مختلف نشانی از وجود تفاوت در سیاست‌ها و الگوریتم‌های نمایه سازی، ذخیره، رتبه بندی و نمایش آنهاست. با وجود تفاوت‌هایی که موتورهای کاوش با یکدیگر دارند سه جزء اصلی در آنها ثابت است:

●عنکبوت[4] یا روبات خزنده[5] یا روبات اطلاعاتی[6]
●نمایه یا پایگاه اطلاعاتی[7]

---

1. Push technology
2. Intelligent agent
3. Thompson et al
4. Spider
5. Crawler
6. Info-bat
7. Database

●نرم‌افزار کاوش(منتظر، ص ۳۴).

به بیانی ساده، عنکبوت به واسطه پیوندهای فرامتنی، پیوسته در صفحات وب می‌خزد و اطلاعات لازم را جهت نمایه شدن به پایگاه می‌فرستد. فرایند نمایه سازی انجام می‌شود و به وسیله نرم‌افزار کاوش، که واسطه میان کاربر و پایگاه اطلاعاتی به شمار می‌آید، به جویش کاربران پاسخ داده می‌شود.

اولین موتورهای کاوش بر محتوای صفحات وب متمرکز بوده‌اند. آلتاویستا و سایر موتورهای کاوش قدیمی بر مبنای نمایه سازی محتوای صفحات وب طراحی شده‌اند. در واقع، نمایه‌های متمرکز بزرگی می‌ساختند که امروز نیز بخش مهمی از هر موتور کاوش شناخته شده‌ای است. اگرچه مدل‌های سنتی نمایه سازی در پایگاه‌های داده‌ها مناسب بود، این مدل‌ها برای منابع اطلاعاتی ساختار نیافته وسیعی که در وب وجود دارد مناسب به نظر نمی‌رسید. در مورد وب تنها کامل بودن نمایه‌ها عامل کیفیت نتایج جست و جو نیست. در ۱۹۹۶-۱۹۹۷ گوگل با نوآوری جدیدی ظهور کرد. گوگل بر این اندیشهٔ جدید پایه ریزی شد که ساختار پیوندها در وب، منبع مهمی در بهبود نتایج موتورهای کاوش است. بعداز نمایه سازی بر مبنای محتوا[1] و تحلیل پیوندها، فضای جدیدی در مطالعات به ظهور رسید. این فضای جدید بر مطالعه صفحه و ساختارهای ترکیب[2] آن تأکید دارد. در رویکرد جدید، HTML[3] و XML اهمیت می‌یابند. ترکیب صفحه منبع مناسبی جهت بهبود نتایج جست و جو است. به‌طور مثال، ارزش اطلاعاتی که در نشانه[4] <Heading> می‌آید بیشتر از اطلاعاتی است که در نشانه <Paragraph> می‌آید. روند رو به رشد ادامه دارد و چنین انتظار می‌رود که نسل بعدی موتورهای کاوش قادر به استخراج داده‌های ساخت یافته جهت پاسخگویی مناسب و با کیفیت بالا به کاربران باشد(اسدی و جمالی مهموئی،۲۰۰۴).

---

1. Content - base Indexing
2. Layout Structure
3. Hypertext Markup Language زبان نشانه گذاری فرامتن
4. Tag

رویکردهای نمایه سازی خودکار وب

در بیانی کلی می‌توان گفت که تاکنون در حوزه فن‌آوری ذخیره و بازیابی اطلاعات در وب شاهد رویکردهای نمایه سازی گوناگونی چون محتوا محوری، استناد محوری(تحلیل پیوندها)، معنامحوری، شامل نمایه سازی معنایی پنهان[1](رویکرد ریاضی) و نمایه سازی بر مبنای هستی‌شناسی لغوی[2](رویکرد زبان شناختی) بوده‌ایم که در ادامه، به آنها پرداخته خواهد شد.

۱. محتوا محوری

اغلب موتورهای کاوش حاضر از روش نمایه سازی بر مبنای کلید واژه‌های متن استفاده می‌کنند. در این شکل، فرایند نمایه سازی سه مرحله خواهد داشت:

- شکستن کلمات[3]
- تعدیل[4] و حذف کلمات مزاحم
- استفاده‌از الگوریتم ریشه ساز جهت تولید ریشه‌های مفاهیم

در مرحله شکستن کلمات، داده‌هایی که به صورت رشته‌ای از کاراکترها هستند مورد بررسی قرار گرفته و حدود کلمات و فاصله میان آنها مشخص می‌گردد؛ به نحوی که کلمات از یکدیگر جدا می‌شوند. در مرحله تعدیل طول سند در نظر گرفته می‌شود، کلمات مزاحم، بزرگ نویسی، نقطه‌گذاری، و مواردی از این دست مدیریت شده و حاصل مرحله اول تمیز می‌شود. کلمات مزاحم در نمایه‌سازی کلماتی هستند که بار معنایی خاصی ندارند و تنها برای ایجاد ارتباط و پیوستگی در جمله‌ها به کار می‌روند(ونگ، برمان، ۲۰۰۳) بعداز این مرحله، از الگوریتمی جهت تولید ریشه‌ها و مفاهیم استفاده می‌شود. هریک از اصطلاحات در نمایه دارای وزن یا ارزشی است که از طریق فرمول زیر به دست می‌آید. در این فرمول T_j نشانی برای اصطلاح j با مقدار

---

1. Latent Semantic Indexing
2. Terminology oriented ontology
3. Word breaking
4. Normalization

$j= 1,2,...,t$ و D_i نشانی از سند i با مقادیر $i=1,2,...n$ است. tf_{ij} بسامد اصطلاح T_j در سند D_i و idf_j معکوس بسامد سند است که مطابق با فرمول دوم محاسبه می‌شود. معکوس بسامد سند از طریق محاسبه لگاریتم خارج قسمت تقسیم تعداد سندها بر تعداد اسنادی که اصطلاح T_j در آنها آمده است به دست می‌آید.

$$w_{ij} = tf_{ij} \cdot idf_j$$

$$idf_j = \log\left[\frac{n}{df_j}\right]$$

جهت تعدیل و اعمال طول سند از سه فرمول زیر استفاده می‌شود به گونه‌ای که در فرمول دوم بسامد اصطلاح T_j در سند D_i بر بسامد هر اصطلاحی در سند که دارای بیشترین بسامد باشد تقسیم می‌شود. وزن حاصل، وزن تعدیل شده اصطلاح T_j در سند D_i است(ونگ، برمان، ۲۰۰۳).

$$w_{ij} = ntf_{ij} \cdot nidf_j ,$$

$$ntf_{ij} = \frac{tf_{ij}}{\max\ tf_i},$$

$$idf_j = \log\left[\frac{n}{df_j}\right]$$

در این رویکرد، سند، تنها با توجه به حضور کلید واژه‌ها نمایه سازی می‌شود؛ و وزن‌دهی مطابق با فرمول شناخته شده‌ای که بیان شد محاسبه می‌گردد. در این روش اگر چه واژگان غیرمجاز حذف می‌شود، جامعیت بالا، مانعیت پایین، و معیار رتبه‌بندی مدارک تنها حضور کلیدواژه‌ها در سند است.

۲. استناد محوری

گوگل دو ویژگی مهم دارد که باعث ایجاد مانعیت در بازیابی اطلاعات می‌شود؛ یکی تعیین رتبه صفحه[1] و دیگری استفاده از پیوندها جهت بهبود نتایج جست و جو، تعداد استنادهایی که به صفحه‌ای خاص در وب می‌شود نموداری از اهمیت و کیفیت آن است. رتبه صفحه مدلی از رفتار کاربران است. فرض طراحان گوگل آن است که جست و جوگر به شیوه آزمون و خطا عمل کند. در این‌صورت کاوش را از صفحه‌ای آغاز می‌کند، پیوندها را یکی پس از دیگری دنبال می‌کند و در صورت خسته شدن پیوند دیگری را دنبال خواهد کرد. رتبه صفحه میزان احتمالی است که در حالت جست و جوی تصادفی ممکن است آن صفحه خاص توسط جست و جوگر بازیابی و مشاهده شود. لازم به ذکر است که تنها تعداد استنادها در تعیین رتبه صفحه ملاک نیست و استناد از هر سایتی به یک میزان اهمیت ندارد. اطلاعات مربوط به پیوندها نیز به عنوان توصیفی مناسب در نمایه سازی استفاده می‌شود؛ به‌طور مثال، اگر سند A پیوندی به سند B داشته باشد متن پیوند یا همان متن کد لنگر به سند B مرتبط می‌گردد و جزء کلید واژه‌های سند A قلمداد نمی‌شود. علاوه بر این، جهت تعدیل، تعداد پیوندهای خارج شده از صفحه نیز مورد بررسی قرار می‌گیرد(برین و پیج، ۱۹۹۸).

فرمول محاسبه رتبه صفحه چنین است:

$$PR(A)=(1-d)+d(PR(T1)/C(T1)+...+PR(Tn)/C(Tn))$$

در این حالت، d مقدار متغیری بین ۰ و ۱ است که اغلب ۰/۸۵ در نظر گرفته می‌شود. T1...Tn استنادها به سند A هستند و C(A) تعداد پیوندهایی هستند که از صفحه خارج شده‌اند. این محاسبه توسط الگوریتمی تکراری محاسبه می‌شود(برین و پیج، ۱۹۹۸).

فرایند ذخیره و بازیابی در موتورهای کاوش شامل خزیدن، نمایه سازی، و جست و

---

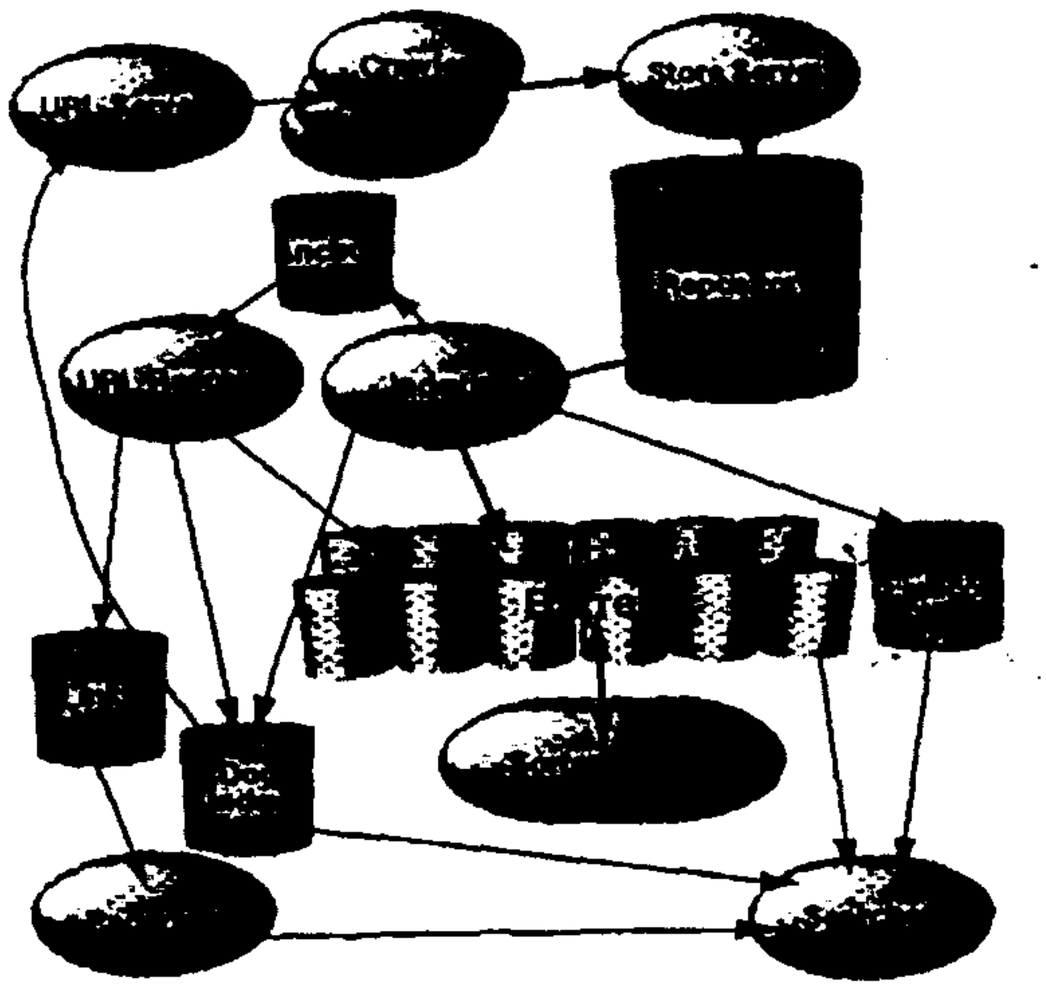
1. Page Rank

جو رتبه بندی و نمایش است. خزیدن در گوگل توسط چند خزنده توزیع شده انجام می‌پذیرد. خدمت دهنده URL[1] سیاهه‌ای از URL ها را جهت واکشی[2] به خزنده‌ها می‌فرستد. URLها بعداز واکشی به خدمت دهنده ذخیره[3] می‌روند. در آن‌جا صفحات وب فشرده و در مخزن[4] ذخیره می‌شوند. در این مرحله به هریک از صفحات شماره شناسایی docID اختصاص می‌یابد. فرایند نمایه سازی توسط نمایه ساز[5] و سورتر[6] انجام می‌پذیرد. نمایه ساز مخزن را می‌خواند، اسناد را از حالت فشرده خارج می‌سازد و تجزیه می‌کند. هر سند به گروهی از رخداد کلمه یا برخورد[7] تبدیل می‌شود. برخوردها، کلمه، موقعیت کلمه در سند، اندازه قلم و بزرگ نویسی را ثبت می‌کنند. نمایه ساز این برخوردها را در گروهی از استوانه‌ها[8] توزیع کرده و نمایه آغازین[9] را که تقریباً مرتب شده است ایجاد می‌کند. نمایه ساز کار مهم دیگری نیز انجام می‌دهد و آن تجزیه همهٔ پیوندها در هر کجای صفحه است، به‌طوری که اطلاعات مهم آنها را در فایل لنگرها[10] ذخیره می‌کند. این فایل حاوی اطلاعاتی است که مسیر پیوندها را مشخص می‌سازد. تطبیق‌گر URL[11] فایل لنگرها را خوانده و URL ها را تثبیت می‌کند. تطبیق‌گر در نمایه آغازین متن لنگر را به همراه docID مربوط به آن صفحه در وب سایت قرار می‌دهد. همچنین پایگاهی از پیوندها به docID خاص ذخیره می‌شود تا با استفاده از اطلاعات آن رتبه صفحه نیز محاسبه شود. سورتر برای ایجاد فایل مقلوب[12] استوانه‌ها را که مطابق

---

1. URLserver
2. Fetch بازیابی دستور یا عنصر داده از حافظه و ذخیره آن در ثبات(فرهنگ تشریحی کامپیوتر میکروسافت. ذیل واژه)
3. Store server
4. Repository
5. Indexer
6. Sorter
7. Hit
8. Barrels
9. Forward Index
10. Anchors file
11. URL resolver
12. Inverted file

باdocID مرتب شده بودند طبق wordID بازآرایی می‌کند. به علاوه، سورتر فهرستی از wordIDها می‌سازد و به فایل مقلوب می‌فرستد. برنامه‌ای به نام Dumplexincon این سیاهه را با سیاهه قبلی که توسط نمایه ساز تولید شده در هم می‌آمیزد و قاموس[1] جدیدی برای استفاده کاربران می‌سازد. کاوشگر، سرویس دهنده وب را اجرا می‌کند و قاموس ایجاد شده راهمراه با فایل مقلوب و رتبه صفحات برای یافتن پاسخی بر جویش خود به کار می‌گیرد(برین و پیج، ۱۹۹۸).



تصویر ۱. معماری موتور کاوش گوگل

در این رویکرد، اگر چه از نمایه سازی کلید واژه‌ای استفاده شده است، به منظور ارتقاء سطح کیفی و افزایش مانعیت، از تحلیل ساختارها به عنوان شاهدی بر اهمیت وب سایت‌ها در رتبه بندی استفاده می‌شود.

---

1. Lexicon

۳. معنامحوری

محققان دریافته‌اند که با استفاده از موتورهای جست‌وجوگر می‌توانند اطلاعات مرتبه‌های بالاتر را نیز به دست آورند. منظور از اطلاعات مرتبه‌های بالاتر، اطلاعات درباره اموری است که مستقیماً در شبکه موجود نیست، بلکه در تراز بالاتر و به صورت انتزاعی در ساختار شبکه مندرج است و باید از کنار هم قرار دادن اطلاعات مرتبه اول یا اطلاعات مرتبه‌های بعدی آشکار شوند. از جمله محققانی که به این نکتهٔ عمیق فلسفی توجه کرده‌اند دو محقق به نام‌های پل ویتانی و رودی سیلیبراستی از مؤسسه ملی ریاضیات و کامپیوتر در آمستردام هلند هستند که در زمینه تولید کامپیوترهای هوشمند فعالیت می‌کنند(موتورهای جست و جو به دنبال معنا، ۱۳۸۳).

این دو محقق برای بهره‌گیری از اطلاعات مرتبه دوم یک مدل آماری طراحی کرده‌اند که نزدیکی و دوری کلمات مختلف را برحسب شمار مواردی که این کلمات به همراه هم در فهرست‌های موتورهای جست و جوگر ظاهر می‌شوند، اندازه‌گیری می‌کند به طور مثال، با کمک این مدل می‌توان دریافت که ارتباط دو واژه "سر" و "کلاه" که در ۹ میلیون مورد به همراه هم در فهرست‌های گوگل ظاهر می‌شوند از رابطه میان دو واژه "سر" و "مو" که تنها در نیم میلیون مورد باهم ظاهر می‌شوند، به مراتب نزدیک‌تر است. به اعتقاد این دو محقق، با کاربرد این مدل آماری در مورد واژگان مختلفی که در فهرست‌های موتورهای جست و جوگر اینترنتی موجود است می‌توان نرم‌افزارهایی تولید کرد که قادر باشند به صورت خودکار و صرفاً با نظر به درجه همبستگی میان واژگان متفاوت معنای آنها را تشخیص دهند(موتورهای جست و جو به دنبال معنا، ۱۳۸۳)

در حال حاضر اغلب موتورهای کاوش رویکردی مکانیکی دارند و به مفاهیم، الگوها، و کلیدهایی که به فهم مفاهیم می‌انجامد توجهی ندارند. در این ابزارها، جست و جوی واقعی صرفاً بر مبنای کلید واژه‌هاست. استفاده از فهرست مترادف‌ها، بهره‌گیری از جست و جوی فازی[1] و تجربیات گیج کننده‌ای در حوزه هوش مصنوعی از جمله

---

1. Fuzzy Search

راهبردهای مطرح شده جهت رفع مشکلات جست و جوی کلید واژه‌ای است. روشی که در اینجا مطرح است بهره‌گیری از نمایه‌سازی معنایی پنهان جهت بهبود مانعیت، جامعیت، و رتبه‌بندی نتایج کاوش است(یو و دیگران،[1] ۲۰۰۲).

۳-۱. LSIنمایه سازی معنایی پنهان

طبق اخبار جدید، گوگل در الگوریتم رتبه‌بندی خود ارزش فراوانی به نمایه سازی معنایی پنهان داده است. نمایه سازی معنایی پنهان به کاربران این اجازه را می‌دهد که جست و جوی خود را به مفاهیم، و نه فقط کلید واژه‌ها، محدود کنند. اگر در ابتدای کلید واژه مورد نظر بدون گذاشتن فاصله، علامت تایپ شود، گوگل به جست و جو در نمایه معنایی خویش پرداخته و نتایج مرتبط با مفهوم را بازیابی می‌کند[2].

به همین ترتیب، از این پس گوگل در بازیابی اطلاعات چندان اعتباری برای رتبه صفحه قائل نمی‌شود. به‌طوری که ممکن است صفحاتی با رتبه صفحه بسیار پایین در ۱۰ نتیجه بازیابی شده در مرتبه بالاتری قرار گیرند[3].

LSI ابتدا در دهه ۱۹۸۰ مطرح شد. در جست و جوی کلید واژه‌ای تنها دو حالت مطرح است یا کلید واژه در سند موجود است و یا موجود نیست، بدون اینکه حد میانی مدنظر باشد. اسنادی که حاویِ کلیدِواژه هستند از سایر اسناد جدا شده و طبق الگوی رتبه‌بندی خاصی به نمایش در می‌آیند. اما در LSI علاوه بر در نظر گرفتن کلید واژه‌های سند، مجموعه اسناد به‌طور کلی مورد بررسی قرار می‌گیرند و به این ترتیب اسناد مرتبط نیز بازیابی می‌شوند.

مشکل اصلی انسان با ماشین آن است که چگونه مفاهیم را به ماشین بیاموزد. درLSI این فرایند تنها با استفاده از محاسبات ریاضی انجام می‌پذیرد، بدون اینکه نیاز به درک

---

1. Yu et at

2. http://www.pacificwebsites. com/news-google-sandbox/latent-semantic-indexing.htm

3. http://www.pacificwebsites. com/news-google-pagerank/google pagerand-news.htm

مفهوم از سوی ماشین باشد. به همین سبب است که نمایه سازی با این روش محدودیت زبانی ندارد و می‌تواند همراه با جست و جوی کلید واژه‌ای مورد استفاده قرار گیرد(یو و دیگران، ۲۰۰۲).

LSI به الگوی توزیع کلمات در اسناد به خصوص همایندی کلمات[1] توجه دارد. آنچه در عمل انجام می‌پذیرد می‌توان چنین برشمرد:

- وجود اسناد الکترونیکی.
- نادیده گرفتن فرمت متن شامل بزرگ نویسی، نظم کلمات، نشانه‌گذاری، و غیره.
- بیرون کشیدن کلمات متن و حذف کلمات زائد از طریق سیاهه بازدارنده[2].

در این مرحله، کلمات متداول که در همهٔ اسناد یافت می‌شوند به علاوه کلماتی که تنها در یک سند از آنها استفاده شده نیز حذف می‌گردد.

- ریشه‌یابی[3] جهت درآوردن ریشه کلمات، ریشه یابی و حذف کلمات زائد، وابسته به زبان است. توسط نرم‌افزارهای خاصی انجام می‌شود؛ به‌طور مثال، می‌توان از Porter Stemmer استفاده کرد.
- تنظیم سیاهه کلمات. در مراحل قبلی نشانه گذاری‌ها، دستور زبان و سبک از میان می‌رفت. در این مرحله نظم کلمات نیز با انتقال کلمات به سیاهه کلمات، از میان می‌رود.
- تولید ماتریس اصطلاح ـ سند[4] به‌طوری که اسناد به صورت افقی و کلمات متن به صورت عمودی ظاهر می‌شوند.
- وزن دهی[5] و تعدیل. در این مرحله دو نوع وزن دهی انجام می‌پذیرد. یکی وزن دهی جایگاهی[6] و دیگری وزن دهی کلّی[7]. در وزن دهی جایگاهی کلماتی که بسامد بالایی

---

1. Word co-occurence
2. Stop list
3. Stemming
4. term-document matrix:TDM
5. Weighing
6. Local weighting
7. Global term weighting

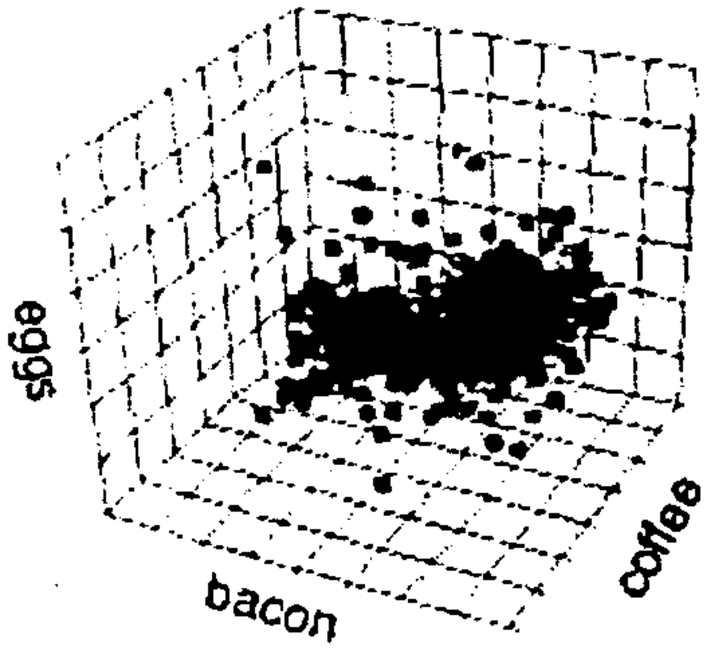
دارند وزن جایگاهی بیشتری نسبت به کلماتی که بسامد کمتری دارند خواهند گرفت. برای محاسبه این وزن از فرمول وزن‌دهی جایگاهی لوگاریتمی[1] استفاده می‌شود. اما وزن‌دهی کلی، سند را در مجموعهٔ اسناد در نظر می‌گیرد. این مقدار به روش‌های مختلفی به دست می‌آید و نمایانگر این مطلب است که در مجموعه اسناد، کلماتی که کمترین میزان حضور را دارند از کلماتی که در اغلب متون توزیع شده و به چشم می‌خورند وزن بیشتری دارند. یکی از طرح‌هایی که وزن کلی را تخمین می‌زند محاسبه معکوس بسامد سند[2] است. در این مرحله تعدیل نیز انجام می‌پذیرد. در نهایت، وزن عددی واقعی مقادیر ماتریس سند ـ اصطلاح تخمین زده می‌شود. طبق محاسباتی که انجام می‌شود مقدار ماتریس تغییر می‌کند. مقادیر مثبت و منفی اعداد، بزرگی و شباهت یا تفاوت اسناد را آشکار می‌سازد. به‌طوری که اعداد منفی در ماتریس نشان دهنده رابطه منفی میان اسناد است.

● اجرای الگوریتم تجزیه مقادیر منفرد(SVD)[3] و شکستن ماتریس اصطلاح ـ سند. این گام کاملاً مکانیکی است و در هر زبانی مستقل از مفاهیم انجام می‌پذیرد.

فرض کنیم می‌خواهیم نوع و پراکندگی سفارش صبحانه در روزی خاص را بررسی کنیم. ترکیبی از سه قلم قابل سفارش را در نظر می‌گیریم، هریک از افراد مجموعه به عنوان سندی در نظر گرفته می‌شود و تقاضای آنها در برداری سه بعدی مشخص می‌شود. آنچه در این فضا مشخص شده است فضای اصطلاح[4] نامیده می‌شود. نمونه‌ای از فضای اصطلاح در تصویر ۲ آمده است(یو و دیگران، ۲۰۰۲).

---

1. Logarithmic Local Weighting
2. Inverse document frequency
3. Singular Value Decomposition
4. Term Space

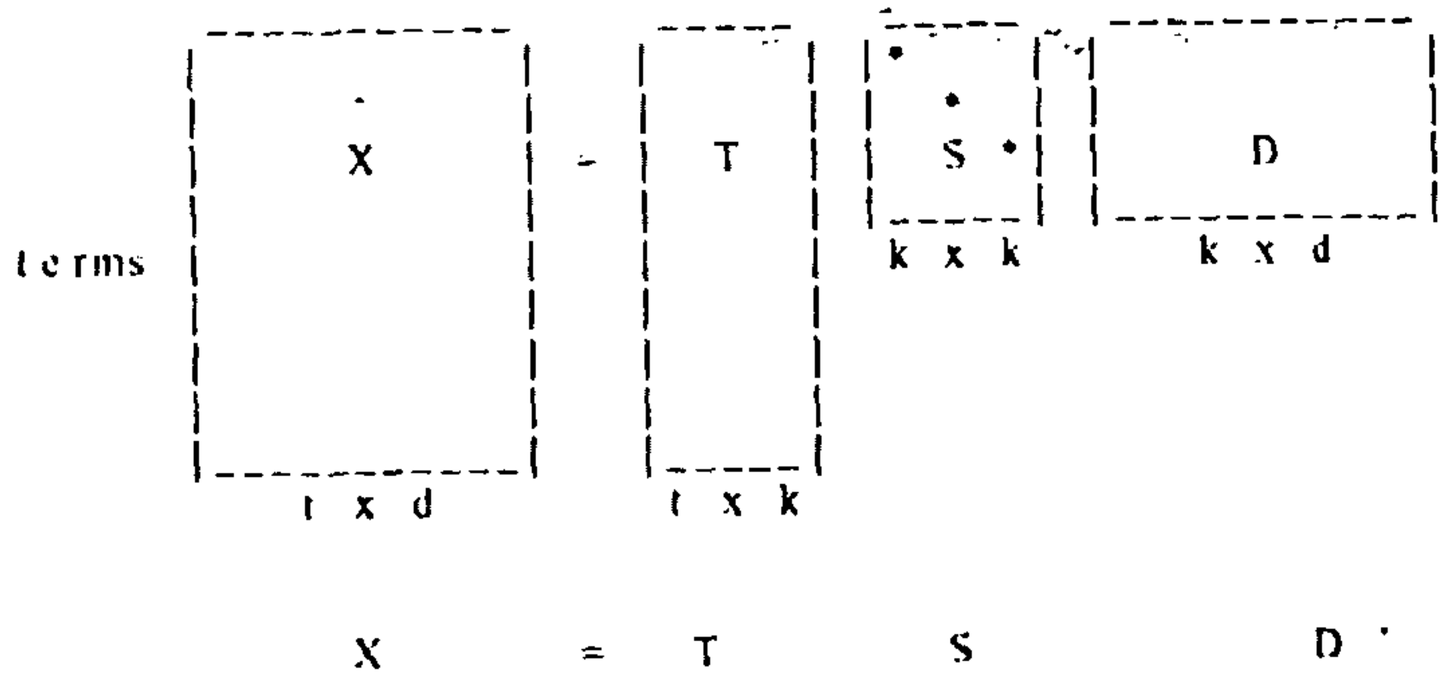


تصویر۲. بردار سه بعدی سفارش صبحانه

در این مثال، چون سه کلید واژه داریم فضای اصطلاح سه بعدی است اما به تبع اضافه شدن کلید واژه‌ها بر ابعاد فضا افزوده می‌شود. هر سند در این مجموعه، برداری تشکیل می‌دهد. در این فضا اسنادی که کلمات مشابه دارند نزدیک هم و اسنادی که کلمات مشترک اندکی دارند از یکدیگر دور هستند. به این ترتیب، فضای اصطلاح ممکن است ده‌ها بُعد داشته باشد. الگوریتم SVD همزمان با کوچک کردن ابعاد تا جایی که امکان دارد اطلاعات مربوط به فواصل بردارها را حفظ می‌کند. در این فرایند، اطلاعاتی از بین خواهد رفت. از دست دادن اطلاعات نکته‌ای منفی به‌نظر می‌رسد، اما در این مورد آنچه حذف می‌شود خشه[1] است. به‌گونه‌ای که شباهت‌های پنهان در اسناد مجموعه آشکار می‌شود. اسناد مشابه به هم نزدیک می‌شوند و اسناد غیر مرتبط از هم فاصله می‌گیرند. SVD ماتریس را شکسته و پس از شکستن، ماتریس‌های دیگری را دوباره ترکیب می‌کند و شبکه‌ای معنایی تشکیل می‌دهد. ماتریس به دست آمده تقریبی از ماتریس اولیه است و جست و جو از طریق آن انجام می‌پذیرد(یو و دیگران،۲۰۰۲). در تصویر ۳ اجرای الگوریتم SVD نمایش داده شده است.

---

1. noise



تصویر ۳. اجرای الگوریتم SVD (دروستر و دیگران[1]، ۱۹۶۰)

از طریق الگوریتم SVD سه نوع مقایسه قابل اجراست:

۱. مقایسه دو اصطلاح i و j با یکدیگر

۲. مقایسه دو سند i و j با یکدیگر

۳. بررسی میزان ارتباط و شباهت اصطلاح i با سند j (دروستر و دیگران، ۱۹۹۰)

هنگامی که پرسشی جهت جست و جو وارد سیستم می‌شود، سیستم به بررسی مقدار وزن‌های هر ترکیب اصطلاح / سند می‌پردازد، میزان شباهت‌ها را محاسبه می‌کند، و اسناد را مطابق با میزان شباهت محاسبه شده، رتبه‌بندی کرده نمایش می‌دهد. در عمل، وزن یا مقداری به عنوان آستانه ارزش[2] در نظر گرفته می‌شود که حد میان اسناد مرتبط و غیرمرتبط است (یو و دیگران، ۲۰۰۲).

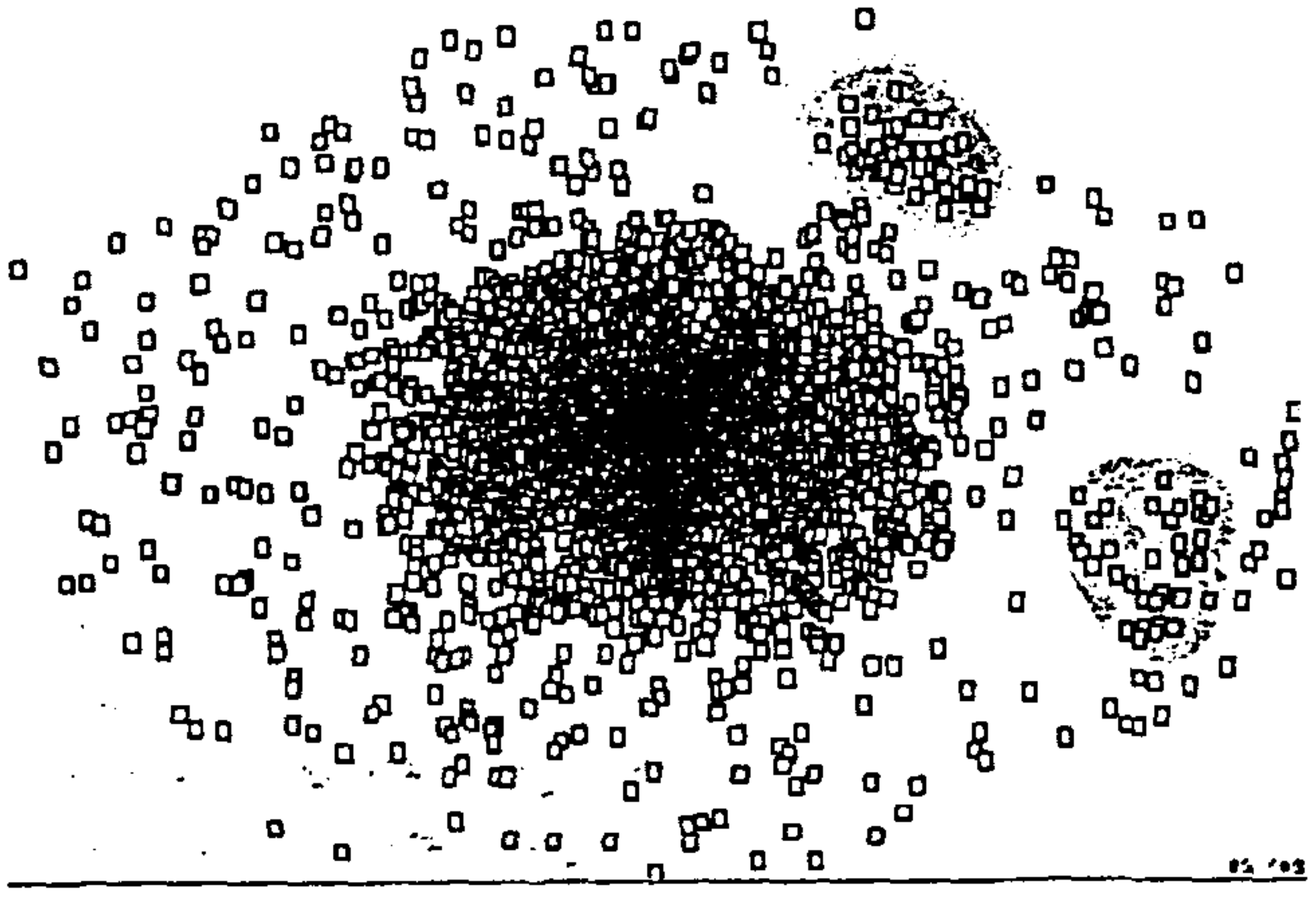
۳-۱-۱. مقیاس گذاری چند بعدی (MDS)[3] تکنولوژی ضمیمه LSI

تلاش الگوریتم SVD کاهش ابعاد ماتریس است، اما در ابعاد بالا نمایش بصری

---

1. Deerwester et al
2. Threshold Value
3. Deerwester et al

بردارها غیرقابل انجام است. لذا از تکنولوژی دیگری به نام MDS استفاده می‌شود. MDS تکنولوژی ضمیمه LSI است که به خلق تصویری دو یا سه بعدی از داده‌های پیچیده مجموعه می‌پردازد. حاصل الگوریتم MDS الگوی پراکندگی داده‌هاست. به طوری که نمایانگر طبقاتی[1] از اسناد است. طبقات به طور خود به خود و به واسطه الگوی همایندی کلمات در مجموعه داده‌ها شکل می‌گیرند، به‌گونه‌ای که مجموعه ساخت نیافته داده‌ها در طبقه‌هایی که نماینده موضوعی خاص هستند جای می‌گیرند. در واقع، الگوی دو بعدی حاصل، نمونه کوچکی است از آنچه LSI در ابعاد بزرگ‌تر ایجاد می‌کند(یو و دیگران،۲۰۰۲). همان‌گونه که در تصویر ۴ مشاهده می‌شود هریک از حوزه‌های تجمع اسناد نشان دهنده طبقه‌های موضوعی خاص است.



تصویر ۴. نمونه‌آی از نمودار MDS(یو و دیگران،۲۰۰۲)

LSI مشکل مربوط به کلمات مترادف را به خوبی رفع می‌کند و به استناد اخبار موجود، گوگل از این روش نمایه سازی نیز بهره می‌گیرد[2].

---

1. Clusters

2. http://www.pacificwebsites.com/news-google-sandbox/latent-semantic-indexing. htm

۳-۲. نمایه سازی بر مبنای هستی‌شناسی لغوی

نمایه سازی بر مبنای کلید واژه‌های صفحات وب بر جامعیت افزوده و مانعیت را کاهش می‌دهد. نمایه‌سازی بر مبنای هستی‌شناسی، فرایندی است نیمه‌خودکار[1] و هدف اصلی آن ساخت نمایه‌ای ساخت یافته از صفحات وب است، به‌گونه‌ای که تنها بر کلید واژه متکی نباشد و بازنمونی از مفاهیم صفحه به شمار آید. در واقع هستی‌شناسی ساختار نمایه را فراهم می‌آورد.

هستی‌شناسی معمولاً متشکل از مجموعه‌ای از رده‌ها(مفاهیم) روابط، عملکردها، اصول بدیهی و نمونه‌ها تصور می‌شود(جمالی مهموئی،۱۳۸۲، ص ۵۹). هستی‌شناسی اصطلاحات را در مجموعه‌های مترادف سازمان می‌دهد و سپس از روابط برای پیوند مجموعه‌های مترادف استفاده می‌کند. از جمله ارتباطات نشان داده شده در هستی‌شناسی لغوی، مترادف‌ها، اصطلاحات همارا، هم‌ارز، اصطلاحات کل به جزء، اصطلاحات مرتبط و متضادهاست(تیلور،۱۳۸۱، ص ۲۱۳).

۳-۲-۱. فرایند نمایه سازی

فرایند نمایه سازی در چندین مرحله انجام می‌پذیرد:

- استخراج اصطلاحات و ساخت نمایهٔ گسترده[2] از صفحات وب. هر اصطلاح در این نمایه با بسامد وزنی[3] مربوط به آن اصطلاح مرتبط می‌شود. استخراج اصطلاحات با حذف نشانه‌های HTML از صفحه انجام می‌پذیرد. متن به جملات مستقل تفکیک شده هر کلمه در صفحه با نقش دستوری آن(اسم،اسم+اسم، صفت+اسم) مشخص و تجزیه و تحلیل می‌شود. سپس برای هر اصطلاح بسامد وزنی محاسبه می‌گردد. این وزن دهی

---

1. Semi-automatic

2. Flat index

فاقد ساختار ، فاقد اطلاعات ساختاری که بازیابی مؤثر اطلاعات با ممکن می‌سازد(فرهنگ تشریحی اصطلاحات کامپیوتر، ذیل واژه flat) کاربران

3. Weighted frequency

از طریق محاسبه بسامد اصطلاح و اعمال نوع نشانه‌های HTML مربوط به آن رخداد انجام می‌پذیرد. دو فرمول برای محاسبه بسامد وزنی مورد استفاده قرار می‌گیرد. طبق جدولی هریک از نشانه‌های HTML وزنی دارد؛ به‌طور مثال، نشانگر عنوان وزن ۱۰ و نشانگر خط کلفت[1] وزن ۲ دارد. در فرمول‌های زیر بسامد وزنی($F(T_i)$) با محاسبه مجموع بسامد اصطلاحات ضرب در وزن نشانه مربوط به آن و سپس تعدیل از طریق تقسیم مقدار ($P(T_i)$) بر مقدار بسامد وزنی اصطلاح K ($P(T_k)$) که بیشترین میزان بسامد وزنی را داراست انجام می‌پذیرد. در فرمول نخست ($M_{i,j}$) مربوط به j بار تکرار اصطلاح T_i با توجه به وزن نشانه‌های HTML آن است(دسمانتیل و جکوئین، ۲۰۰۲).

$$P(T_i) = \sum_{i=1}^{p} (M_{i,j}) \qquad F(T_i) = \frac{P(T_i)}{max_{k=1..n}(P(T_k))}$$

●تعیین مفاهیم صفحه. پس از آنکه اصطلاحات با بسامد وزنی تعیین گردید، به منظور لحاظ کردن مفاهیم صفحه در فرایند نمایه سازی از اصطلاحنامه استفاده می‌شود. از میان مفاهیم استخراج شده در مرحله نخست، مفاهیم کاندید با توجه به اصطلاحنامه مشخص می‌گردد. هر مفهوم با سیاهه‌ای از مترادف‌ها همراه است.

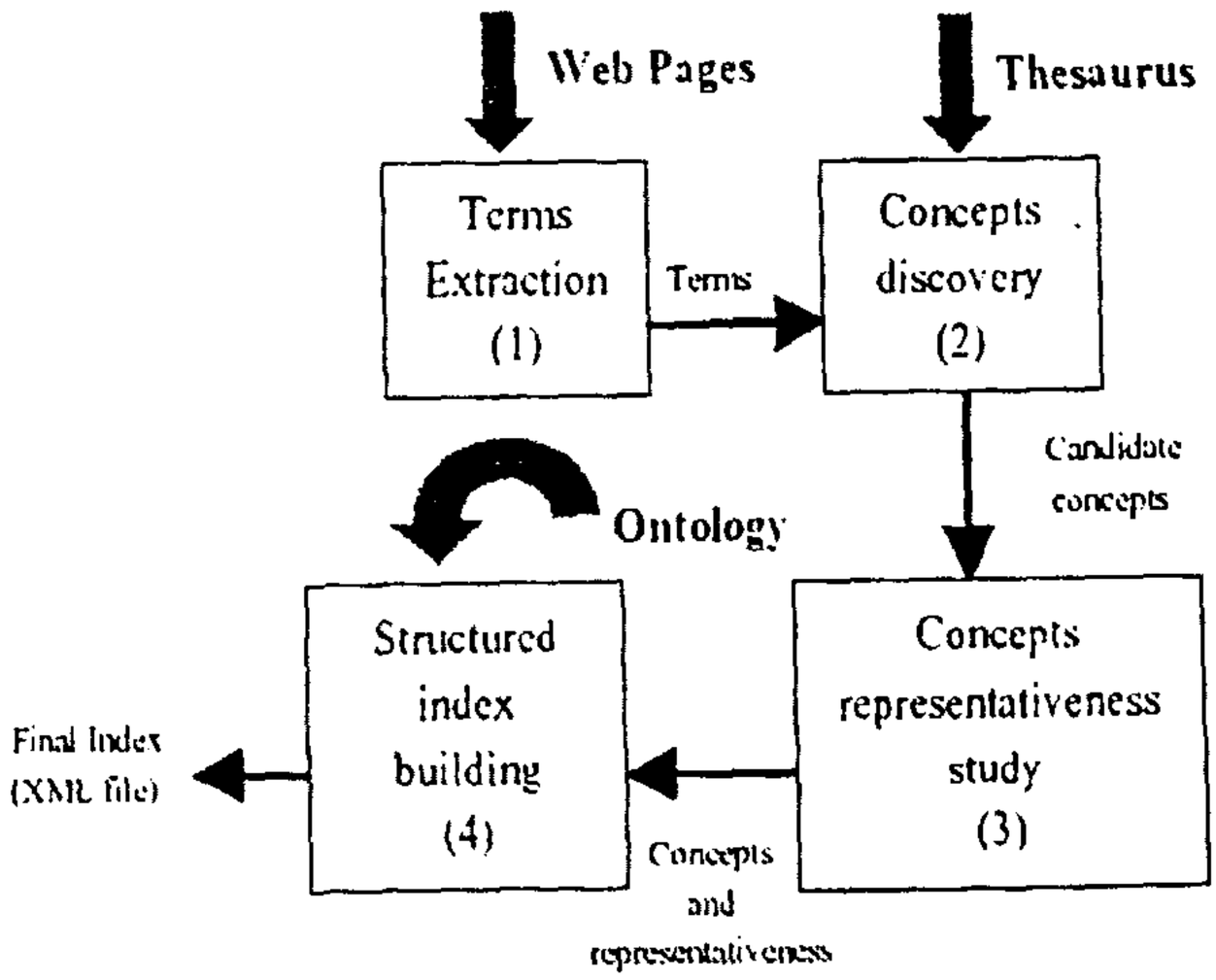
● اندازه‌گیری میزان بازنمونی[2]: برای هریک از مفاهیم کاندید، میزان بازنمونی با توجه به بسامد وزنی و تشابه انباشته[3] میان مفهوم کاندید با سایر مفاهیم مربوط به صفحه محاسبه می‌شود. میزان تشابه مفاهیم به صورت دو به دو محاسبه می‌گردد و سپس مجموع آن در نظر گرفته می‌شود. میزان تشابه میان مفاهیم به اصطلاحنامه مورد استفاده بستگی دارد و، در واقع، در این فرایند فاصله مفهومی میان مفاهیم بررسی می‌شود. در نهایت میزان ضریب بازنمونی، ضریبی از ترکیب خطی بسامد وزنی و ۴ ضریب مربوط به تشابه انباشته مفهوم است. این ضریب عامل مؤثری در بهبود کیفیت پاسخگویی به

---

1. Bold font
2. representative ness
3. cumulative similarity

پرس و جوی کاربران است.

● تا اینجا از طرفی هستی‌شناسی و از طرفی مفاهیم کاندید به همراه میزان بازنمونی را در اختیار داریم. در این مرحله مفاهیم کاندید با مفاهیم موجود در هستی‌شناسی مطابقت داده می‌شود. اگر مفاهیم کاندید در هستی‌شناسی و همچنین در صفحه نمایه شده موجود باشد، URL صفحه به همراه میزان بازنمونی آن به هستی‌شناسی اضافه می‌شود. نمایه در فرمت XML و به صورت مستقل از صفحه ذخیره می‌شود. در تصویر ۵ فرایند نمایه سازی به تصویر کشیده شده است(دسمانتیل و جکوئین، ۲۰۰۲).



تصویر ۵. فرایند نمایه سازی

در بسیاری از موتورهای کاوش از کلید واژه‌ها در نمایه سازی صفحات استفاده می‌شود و بر مبنای فهرستی از واژه‌ها و استفاده از عملگرها فرایند پاسخگویی به پرسش‌ها صورت می‌پذیرد؛ اما در این روش می‌توان از نمایه ساخت یافته‌ای استفاده کرد؛ به این ترتیب پرس و جوها تنها در سطح اصطلاحی[1] پردازش نمی‌شود بلکه در

---

1. terminological level

سطح مفهومی[1] نیز پردازش می‌گردد. بسط پرس و جو از طریق استفاده از هستی‌شناسی امکان‌پذیر است. در این رویکرد اصطلاحاتی که به عنوان پرسش وارد سیستم می‌شوند با مفاهیم مرتبط جا به جا می‌گردند و در ابتدا مفاهیم کاندید در هستی‌شناسی انتخاب می‌شوند. در انتها اگر اصطلاح با مفاهیم کاندید دیگری نیز مرتبط باشد همکاری کاربر ضروری است (دسمانتیل و جکوئین، ۲۰۰۲).

۳-۲-۲. ابرداده

اغلب موتورهای کاوش اهمیت فراوانی به ساختار صفحات می‌دهند. به این معنی که نوع نشانه‌ها گویای ارزش اطلاعاتی است که در آن نشانه نمایان شده است. چنانکه بیان شد، در نمایه سازی با تکیه بر هستی‌شناسی، نشانه‌ها ارزش‌های متفاوتی داشتند که در محاسبه وزن اصطلاح لحاظ می‌شد. از طرفی امکان استفاده از ابرداده‌ها در قالب ابرنشانه‌های رایج در زبان HTML فراهم آمده است. تلاش در زمینهٔ تحقق وب معنایی نیز به استانداردسازی ساختارها و تفسیر و پردازش استاندارد موجودیت‌های توصیف شده در آنها تأکید دارد.

همزمان با گسترش وب بسامد واژه ابر داده در متون مربوط به سازماندهی اطلاعات در این محیط پویا و رو به گسترش افزایش یافت. به‌طوری که می‌توان گفت "در عصر اینترنت، شاید هیچ داده‌ای ارزشمندتر از داده درباره داده نباشد"(السوپ[2]، ۲۰۰۰ نقل در صفری، ۱۳۸۲، ص ۱۷).

در نگاهی تیزبین نمی‌توان ابرداده را تنها به محیط دیجیتالی محدود ساخت. همواره در طول زمان، داده ساختارمند درباره داده از ارزش فراوانی برخوردار بوده است. در مثالی ساده می‌توان فهرستبرگه‌های کتابخانه رامجموعه‌ای ابرداده‌ای دانست که جست و جو، بازیابی و جایابی اقلام اطلاعاتی را تسهیل می‌کند.

برای ابرداده تعاریف متعددی ارائه شده است که ساده‌ترین آنها داده درباره داده

---

1. conceptual level
2. Alsop

است. این داده باید ساختارمند[1] و به‌گونه‌ای ماشین فهم و ماشین خوان باشد(دی،[2] 1999؛ نقل در صفری، ۱۳۸۳، ص ۱۵). می‌توان گفت ابرداده، داده‌ای است که خصوصیات داده‌های اصلی را بیان می‌کند، روابط آن‌ها را توصیف می‌کند، و کشف و استفاده مؤثر از آنها را میسر می‌سازد(برنت[3]؛ نقل در صفری، ۱۳۸۳: ۱۶).

ابرداده در تور جهانگستر به منظور یاری رساندن به کاربران برای یافتن اطلاعات مورد نظر مطرح شده است. وجود توصیفی یک دست، دقیق و خوش ساخت از منابع موجود در تور جهانگستر اجازه حصول دقت بیشتر در مجموعه نتایج به دست آمده از کاوش(چه از طریق راهنما و چه از طریق موتورهای کاوش) را خواهد داد(نقشینه، ۱۳۷۹، ص ۳۵).

طرح دابلین کور یکی از طرح‌های ابرداده‌ای جهت آشکار ساختن مفاهیم و معانی صفحات وب در قالب ۱۵ ابرنشانه است که کاربرد گسترده‌ای پیدا نکرد؛اما استفاده از ابرنشانه‌ها از سوی طراحان وب سایت‌ها مورد توجه قرار گرفته است. برخی از موتورهای کاوش این نشانه‌ها را در فرایند نمایه سازی خویش لحاظ کرده و ارزش خاصی برای آنها قائلند. در زبان HTML می‌توان از ابرنشانه‌ها استفاده کرد که مهم‌ترین آن‌ها ابرنشانه توصیفگر و کلید واژه است.

- ابرنشانه از نوع توصیفگر (ابرنشانه توصیفگر): این نوع ابرنشانه‌ها برای آگاهی از محتوای صفحات وب اهمیت خاصی دارند. زیرا از طریق آنها می‌توان توصیف خلاصه‌ای درباره صفحات وب درج کرد و پس از بازیابی صفحات وب در معرض دید کاربران قرار داد(کوشا، ۱۳۸۰،ص ۱۳۸).
- ابرنشانه از نوع کلید واژه(ابرنشانه کلیدواژه): این نوع ابرنشانه‌ها، در واقع، کلمات و واژه‌هایی هستند که به توصیف محتوای یک صفحه وب می‌پردازند(کوشا، ۱۳۸۰، ص ۱۳۹).

---

1. Structured data
2. Day
3. Burnett

ابرنشانه‌ها در صفحه نمایش داده نمی‌شوند و تنها در فرایند نمایه سازی از سوی برخی موتورهای کاوش مورد استفاده قرار می‌گیرند. جای این ابر نشانه‌ها قبل از نشانه متن <BODY> در واقع در محدوده نشانه سرصفحه <HEAD> و بعداز نشانه عنوان <TTTLE>است(علیمحمدی، ۲۰۰۳). در حالیکه برخی موتورهای کاوش، به علت عدم اعتماد، ابرنشانه‌ها را نادیده می‌گیرند؛ با این همه، به کارگیری مؤثر این ابرنشانه‌ها معمولاً به طراحان وب سایت‌ها توصیه می‌شود.

۳-۲-۳. داده‌های ساختار یافته و وب معنایی

موتورهای کاوش جاری تنها قادرند صفحات محدودی از وب را که مرئی[1] یا نمایه‌پذیر نامیده می‌شوند مرور کنند. بسیاری از اطلاعات علمی مفید در پشت درهای بسته مانده است. اندازه وب مرئی یا وب سطحی[2] از وب نامرئی به مراتب کوچک‌تر است. پایگاه‌های گوناگون فهرست کتابخانه‌ها، کتاب‌ها و مجلات دیجیتال، پروانه‌های ثبت اختراع، گزارش‌های مربوط به تحقیقات و آرشیوهای دولتی از جمله نمونه‌های وب نامرئی است. برای رفع این نقص حرکت به سمت وب معنایی با تکنولوژی‌هایی چون XML و RDF است تا منابع وب ساخت یافته‌تری تولید گردد(اسدی و جمالی مهموئی،۲۰۰۴).

مبتکر وب معنایی آن را چنین تعریف کرده است:" وب متشکل از داده‌ها که به صورت مستقیم یا غیرمستقیم توسط ماشین قابل پردازش هستند"(برنرزلی و فیسکتی، ۱۹۹۹،ص ۱۹۱، نقل در جمالی مهموئی، ۱۳۸۲، ص ۵۱). وب معنایی، داده‌های ساختاربندی شده‌ای به وب کنونی می‌افزاید که پردازش آنها برای رایانه‌ها آسان است(جمالی مهموئی،۱۳۸۲، ص ۵۱). هم اکنون بازیابی اطلاعات وب، عموماً براساس تطابق لغات و عبارات مورد جست و جو با واژه‌ها و عبارات موجود در متن صفحات وب صورت می‌گیرد. وب معنایی از تطبیق صرف واژه‌ها فراتر است و جست و

---

1. invisible
2. Surface web

جو را براساس موضوع، ارتباط میان داده‌ها، نوع داده‌ها، و کیفیت‌های دیگر انجام می‌دهد(جمالی و مهموئی، ۱۳۸۲، ص ۵۲).

وقتی اطلاعات دیجیتال شد، XML وارد عمل می‌شود. جست و جو و نمایش اطلاعات در صورتی که به روشی معنادار ساختاربندی نشده باشد، ناممکن است. به زبان ساده، تولیدکنندگان اطلاعات باید برای استفاده از منابع اطلاعاتی براساس استاندارد خاصی به برنامه نویسی و کدگذاری اسناد الکترونیکی بپردازند(زارع زاده، ۱۳۸۳، ص ۴۱). با استفاده از فرا زبان XML می‌توان عناصر ابرداده‌ای را به اسناد افزود، اما هنوز در این حالت، ابرداده‌ها قابلیت پردازش توسط رایانه را ندارند، لذا برای افزودن قابلیت پردازش از RDF و فرانمای [1] آن استفاده می‌شود.

تا این مرحله رایانه‌ها قادر به تشخیص محتوای صفحات وب شده‌اند و می‌توانند ابرداده‌ها را پردازش کنند؛ اما برای آنکه نظام‌های مختلف رایانه‌ای قادر به تبادل داده‌ها و استفاده از آنها باشند بایستی از سطح قابل قبولی از عمل پذیری درونی [2] برخوردار باشند. زمینه این ویژگی اگرچه با استفاده از XML و RDF تا حدودی فراهم شده اما کامل نیست. لذا برای تکمیل آن از هستی‌شناسی وب استفاده می‌شود. هستی‌شناسی، درک مشترکی میان رایانه‌ها و نیز میان انسان و رایانه ایجاد می‌کند تا بتوان از داده‌های وب در نرم‌افزارهای مختلف استفاده کرد. مرحله نهایی، تدارک موتورهای استنتاج [3] است که از قدرت پردازش دانش موجود در وب برخوردارند(جمالی مهموئی، ۱۳۸۲،ص ۶۰).

پیش نیاز کاربرد هستی‌شناسی‌ها در وب معنایی توسعه استانداردی برای تعریف و

---

1. Resource Description Framework Schema

2. Interoperability اشاره به مؤلفه‌هایی از سیستم‌های کامپیوتری است که می‌توانند در محیط های گوناگون عمل کنند... چنین حالتی در مورد نرم‌افزار زمانی رخ می‌دهد که برنامه‌ها بتواند داده و منابع را به اشتراک گذارند(فرهنگ تشریحی میکروسافت. ۱۳۷۹. ذیل واژه)

3. Inference engine

مبادله هستی‌شناسی یا به عبارتی، زبان‌های بازنمون هستی‌شناسی است(فنسل و دیگران، ۲۰۰۱؛ نقل در صفری، ۱۳۸۳، ۹۷) زبان‌هایی چون RDF را که در توسعه زبان‌های دیگری چون OIL[1] و DAML+OIL[2] نقش داشته است می‌توان به عنوان نمونه‌هایی معروف از این زبان‌ها ذکر کرد(صفری، ۱۳۸۳،ص ۹۷).

در وب معنایی از نمایه سازی معنایی بحث می‌شود. تلاش‌ها جهت تلفیق این نمایه سازی با مدل‌های موجود ـ نمادین و آماری ـ در حال انجام است. این نوع نمایه سازی با توجه به هستی‌شناسی و تحلیل ابرداده‌ها انجام می‌پذیرد(کریاکوف و دیگران[3]، ۲۰۰۴).

رؤیای وب معنایی هنوز تحقق نیافته است، اما تحقیقات و تلاش‌ها در این زمینه در حال پی‌گیری است. موتورهای کاوش حاضر جهت ارتقاء عملکرد به بهبود الگوریتم‌های خویش می‌پردازند و در تلاشند تا راهکاری برای رویارویی با وب نامرئی بیابند.

بحث و نتیجه گیری

در این بررسی، رویکردهای نمایه سازی خودکار در وب نشان داده شد. بدون وجود نظام نمایه سازی مناسب، اطلاعات ارزشمند وب، نامرئی باقی خواهد ماند. هر یک از روش‌های نمایه سازی نقاط ضعف و قوتی دارد که در سایه تحقیقات آشکار خواهد شد. آنچه پشت صحنه موتورهای کاوش به عنوان ابزارهای اصلی کاوش در وب انجام می‌پذیرد بیشتر به نمایه سازی کلید واژه‌ای محدود می‌شود. اگرچه استفاده مناسب از تحلیل پیوندها و به بیانی استناد محوری بر کیفیت نمایه سازی کلید واژه‌ای افزوده است. مدل‌های آماری جدید جهت آشکار ساختن شبکه معنایی مانند آنچه در موتور کاوش گوگل در حال انجام است تا اندازه‌ای امیدوارکننده است. حرکت‌های جاری به سمت بهره‌گیری از داده‌های ساختار یافته و تحقق وب معنایی است، اما این حرکت‌ها تاکنون

---

1. Ontology Inference Layer/Language
2. DARPA Agent Mark up Language
3. Kiryakov et al

به‌طور کامل به سرانجام نرسیده است و هنوز مشکلات حل نشده فراوانی در این مسیر وجود دارد. وب تنها متعلق به متخصصان علوم کامپیوتر نیست. تلاش‌های مشارکتی در دسترس‌پذیر ساختن اطلاعات وب باید از سوی گروهی از متخصصان و به صورت میان رشته‌ای انجام پذیرد. بی گمان کتابداران و اطلاع رسانان به عنوان متولیان سازماندهی دانش بشر، در این مشارکت می‌توانند و باید نقش مؤثری ایفا کنند. تلاش در این زمینه نیازمند مهارت‌ها و دانش پایه‌ای است که لازم است از سوی این متخصصان مورد توجه واقع شود.

به گفتهٔ لینچ[1] (۱۹۹۷) "مهارت‌های کتابدار در طبقه بندی و انتخاب بایستی به وسیله توانایی متخصص رایانه تکمیل شود تا بتوان وظیفه نمایه سازی و ذخیره سازی اطلاعات را خودکار نمود. تنها ترکیبی از دیدگاه‌های مختلف(متخصصان) هر دو حرفه اجازه خواهد داد تا این رسانه جدید پویا باقی بماند(لارج، تد، و هارتلی،۱۳۸۳، ص ۱۵۹).

مآخذ

پفامنرگر، برایان(۱۳۷۷). *فرهنگ تشریحی اصطلاحات کاربران کامپیوتر*، تهران: ناقوس.

تیلور،آرلین جی(۱۳۸۱). *سازماندهی اطلاعات*. ترجمه محمدحسین دیانی. مشهد: کتابخانه رایانه‌ای.

جمالی مهموئی، حمیدرضا(۱۳۸۳). "وب معنایی: شیوه‌ای رو به تکامل برای ذخیره و بازیابی کارآمدتر اطلاعات روی اینترنت". *اطلاع‌شناسی*. (۳):۴۹-۶۶.

رودکی،مهدی(۱ اسفند ماه ۱۳۸۳). "ماگرسنگان اطلاعات". ایتنا ITNA اخبار فناوری اطلاعات. پیوسته: http://www.itna.ir/archives/article/002041.php [۳۰ بهمن‌ماه، ۱۳۸۳]

زارع زاده، فاطمه(۱۳۸۳). "ایکس. ام.ال چیست و چه‌کاربردهایی در کتابخانه دارد؟". *فصلنامه کتاب*. ۱۵(۳): ۳۷-۴۴.

صفری، مهدی(۱۳۸۳). "سنجش میزان اثربخشی عناصر ابرداده‌ای دابلین کور در بازیابی صفحات وب: مطالعه

---

1. Lynch

صفحات وب "Jornal of Science Iranian International". پایان‌نامه کارشناسی ارشد کتابداری و اطلاع رسانی. دانشکده علوم تربیتی. دانشگاه تهران.

صفری، مهدی(۱۳۸۳). "مدل سازی مفهومی در بازنمون رسمی دانش: شناختی از هستی‌شناسی در هوش مصنوعی و نظام‌های اطلاعاتی". *اطلاع‌شناسی*.۱(۴): ۷۵-۱۰۴.

فرهنگ تشریحی میکروسافت(۱۳۷۹). ترجمه رضا حسنودی و داریوش فرسانی. تهران: نشر دانشیار؛ پیک علوم.

کوشا، کیوان(۱۳۸۰). *ابزارهای کاوش اینترنت: اصول، مهارت‌ها و امکانات جست و جو در وب*. تهران: نشر کتابدار.

لارج،آندرو؛ تد، لوسی؛ هارتلی، ریچارد(۱۳۸۳). *جستجوی اطلاعات در عصر اطلاعات: اصول و مهارتها*.ترجمه زاهد بیگدلی؛ ویراسته زهیر حیاتی. تهران: نشر کتابدار.

منتظر، غلامعلی(۱۳۸۱) *موتورهای کاوش اینترنت(درآمدی در بازیابی بهینه اطلاعات)*. تهران: کویر.

"موتورهای جست و جو به دنبال معنا"(۱۰ بهمن ماه ۱۳۸۳). ایتناITNA اخبار فناوری اطلاعات. پیوسته: [۳۰ بهمن‌ماه، ۱۳۸۳]http:// www.itna.ir/archivs/news/001950.php

نقشینه،نادر(۱۳۷۹). "مقدمه‌ای بر فراداد" *علوم اطلاع رسانی*. ۱۵(۳ و ۴): ۳۱-۳۷.

Alimohammadi,Daryoush(2003)."Meta-tag: a means to control the process of Web indexing. Online Information Review. 27(4): 238-242. Available at: http://taddeo.emeraldinsight.com/vl=1364367/cI-158/nw=1/fm=html/rpsv/cw/mcb/14684527/v27n4/s2/p238

Asadi,S., & Jamali M., H.R.(2004). "Shifts in search engine development: A review of past, present and future trends in research on search engines". *Webology,*1(2),Article 6.Available at: http://www.webology.ir/2004/v1n2/a6.html

Bradshaw,Shannon; Hommond, Kristian(2002)."Automatically indexing documents: Content vs. Reference". INternational Conference on Intelligent User Interfaces, Proceedings IUI,2002,p 180-181.{compendex database} Available at: http://www.engineeringvillage2.org/controller/servlet/Controller?CID-quickSearchAbstractFormat&EARCHID=18e261d103560e65f5M5b1bcv2100127DOC

INDEX=17PAGEINDEX=17database=17format=quickSearchAbstractFormat

Brin,Sergey,7 Page,Lawrence(1998)."The Anatomy of a Large-Scale Hypertextual Web Search Engine". Available at: http://www db.stanford.edu/~backrub/google.html.

Deerwester,S.,Dumais,S.T.,Landauer,T.K., Furnas,G.W.and Harshman,R.A.(1990). "Indexing by latent semantic analysis." *Journal of the Society for Information Science,*41(6),391-407.Available at: lsi.research.telcordia.com/lsi/papers/JASIS90.pdf

Desmontils,E;Jacquin,C.(2002)."Indexing a Website with a Terminology Oriented Ontology". Available at: http://www.semanticweb.org/SWWS/program/full/paper5.pdf

Dochartaigh,Niall O.(2002). *The Internet Research Handbook: A Practical for Students and Researchers in the Social Sciences.* London:SAGE Publications.

"Is PageRank Going Away?" (2004). Available at: http://www.pacificwebsites. com/news-google-pagerank/google pagerand-news.htm

Kiryakov et at(2004)."Semantic annotation, indexing, and retrieval". Web Semantics: Science, Science, Services and Agents on the World Wide Web.2(1):49-79.available at: http://www.ontotext.com/publications/SemAIR ISWC169.pdf

Mansourian,Yazdan(2004)"Intelligent Search Agents and Information Seeking Through the Web" *Informology.*1(4):213-231.

Wang,Shuangbao;Behmann,Michael M.(2003)."A Dynamic Visual Search engine for Database-Driven Web Content". Available at: http://www.actapress.com/proceedings/2003proceedings/TOC.

Yu,Clara,Cuadrado,John, Ceglowski,Maciej and Payne,J.Scott.(2002)."Patterns in Unstructured Data: Discovery,Aggregation, and Visualization". Available at: http://javelina,cet.middlebury.edu/lsa/out/cover page.htm

"What is Latent Semantic Indexing Analysis?".(2004). Available at: http://www.pacificwebsites.com/news-google-sandbox/latent-semantic-indexing.htm