علوم اطلاع‌رسانی، دوره ۲۰، شماره ۳و۴ (بهار و تابستان ۱۳۸۴) ۹۹-۱۰۷

بازیابی اطلاعات در وب: فراتر از موتورهای جستجوی کنونی[1]

بیزا یاتاز

ترجمه: لیلا دهقانی*

دانشجوی کارشناسی ارشد کتابداری و اطلاع‌رسانی

چکیده

در این مقاله به اختصار چالش‌های مربوط به گسترش بازیابی اطلاعات در وب، بخصوص دیگر انواع داده‌ها ، وب‌کاوی و موضوعات مرتبط با کاوش، همچنین روابط اصلی بازیابی اطلاعات و محاسبات تقریبی، و چگونگی رفع این چالش‌ها با این تکنیک‌ها را بررسی می‌کنیم.

کلیدواژه‌ها: بازیابی اطلاعات/ وب (شبکه سراسری جهانی)/ موتور جستجو/ داده‌ها/ محاسبه تقریبی

۱. مقدمه

وب[2] به صورت بزرگ‌ترین منبع داده‌ها در آمده که به سهولت قابل دسترسی است؛ از این رو طبیعی است که اطلاعات از آن استخراج شود و موتورهای جستجو[3]ی وب به یکی از پراستفاده‌ترین ابزارها در اینترنت تبدیل شده‌اند. اما رشد تصاعدی و آهنگ سریع تغییر و تحول در وب، بازیابی تمام اطلاعات باربط را واقعاً مشکل می‌سازد. در واقع، کاوش[4] در وب شاید تنگنای اصلی موتورهای جستجوی وب است. بعلاوه، این فرض نانوشته وجود دارد که یک فایل فیزیکی یک مدرک منطقی است؛ و این همیشه درست نیست.

تحقیق اخیر دربارهٔ چالش‌های جستجو در وب، شامل مسائل زیر است [۱۰، ۲۰، ۲۵] :

- نمایه (و از جمله، شمول آن بر محتوای پنهان) را به روز و کامل نگه‌داشتن؛
- شناسایی و حذف محتوا و پیوندهای مغرضانه، که اطلاعات ناخواستهٔ[5] موتور جستجو نام دارد . بعضی نویسندگان، آن را «بازیابی اطلاعات متناقض»[6] می‌نامند؛

* نویسنده مکاتبه‌کننده. تلفن: ۰۷۱۱-۶۳۱۰۰۳۷ ؛ نمابر:

پست الکترونیکی:Leiladehghani@yahoo.com

• مشخص‌کردن محتوای با کیفیت خوب. وب پر از محتویات با کیفیت پایین (از نظر نحوی و معنایی) مشتمل بر داده‌های پارازیتی، نامعتبر و متناقض می‌باشد. بنابراین، ما این مشکل را داریم که تا چه حد می‌توان به یک وب‌سایت اعتماد کرد. این، شامل ساختار «اچ‌تی‌ام‌ال»[۷] (که در بیشتر موارد مبهم و نامتجانس است) نیز می‌شود.

• بهره‌برداری از بازخورد کاربر، چه از ارزیابی صریح کاربر یا به طور ضمنی از گزارش‌های وب[۸]. در این جا می‌توانیم اطلاعات ضمنی را که توسط نویسندگان صفحات وب و به شکل چندین قاعدهٔ مورد استفاده در طراحی «اچ‌تی‌ام‌ال» ارائه می‌شود نیز اضافه کنیم؛

• کشف نسخه‌های مشابه از میزبان‌ها و محتویات، به منظور اجتناب از کاوش غیرضروری.

• تشخیص نیاز اطلاعاتی: اطلاعی، راهنمایی[۹] یا اجرایی[۱۰]. برآورد شده است که کمتر از ۵۰ درصد درخواست‌ها از نوع اول می‌باشد.

• بهبود زبان پرس‌وجو، اضافه‌کردن زمینهٔ اطلاعات مورد درخواست، از قبیل نوع یا زمان؛

• بهبود رتبه‌بندی، بخصوص برای مرتبط‌کردن آن با فردی که پرسش را مطرح کرده است. اساس ربط، قضاوت‌های شخصی است؛ بنابراین رتبه‌بندی بر اساس پروفایل‌های کاربر یا اطلاعات زمینه‌ای دیگری که مربوط به کاربر است، می‌تواند مفید باشد. در اینجا می‌توانیم کیفیت، اعتماد و موضوعات بازخورد کاربر را نیز اضافه کنیم.

فهم تمام این مسائل بدون داده‌های واقعی دشوار است، بنابراین نتایج تجربی بیشتری مورد نیاز است. مطالب بیشتر را می‌توان در [۱۱،۲،۱] پیدا کرد.

وب چیزی بیش از«اچ‌تی‌ام‌ال» محض و دیگر ساختارهای متنی متداول است و ما می‌خواهیم دیگر انواع داده‌ها را نیز جستجو کنیم، که در میان آن‌ها صفحات پویا، اشیای چندرسانه‌ای، داده‌های «ایکس‌ام‌ال»[۱۱] و اطلاعات معنایی همراه با آن‌ها را داریم. اگر «وب معنایی»[۱۲] علی‌رغم تمام مسائل اجتماعی که باید حل شوند صورت واقعی به خود گیرد، یک وب بر پایه‌ی «ایکس‌ام‌ال» با طرح کلی و فراداده‌های[۱۳] معنایی استاندارد خواهیم داشت. در چنان محیط محتملی، بازیابی اطلاعات آسان‌تر می‌گردد و حتی جستجوی چندرسانه‌ای ساده می‌شود. در چنین محیطی اطلاعات ناخواسته باید از بین برود و تشخیص محتوای خوب، راحت‌تر است. ازطرف دیگر، مسائل جدید بازیابی مانند پردازش و بازیابی «ایکس‌ام‌ال» و وب‌کاوی[۱۴] بر روی داده‌های ساختاری پیدا می‌شود.

«لطفی‌زاده» [۲۸] مفهوم محاسبات تقریبی[۱۵] را به عنوان همگرایی روش‌ها که در مجموع، پایه‌ای برای مفهوم‌سازی، طرح، ساخت و به‌کارگیری سیستم‌های هوشمند/ اطلاعاتی فراهم می‌آورند ارائه کرد. برخی از روش‌های اصلی محاسبهٔ تقریبی عبارت‌اند از منطق فازی[۱۶]،

الگوریتم‌های تکاملی[۱۷]، شبکه‌های عصبی[۱۸]، مجموعه‌های ناهموار[۱۹]، شبکه‌های بیزی[۲۰] و دیگر شیوه‌های احتمالاتی. ویژگی اصلی محاسبهٔ تقریبی این است که نسبت به عدم دقت، ابهام، حقیقت ناقص، و برآورد، مقاوم است. ذهنیت، ابهام و عدم دقت، ویژگی‌های معمول در هر فرآیند بازیابی اطلاعات هستند. استفاده از فنون محاسبهٔ تقریبی برای بهبود فرآیندهای بازیابی اطلاعات، رضایت‌بخش بوده است. بخصوص، فکر می‌کنیم که کاربرد آن برای حل مسائل مختلف بازیابی اطلاعات که اخیراً در وب پدیدار شده‌اند، مفید است.

ما بحث را از چالش‌های داده‌ای شروع می‌کنیم و به دنبال آن یک مقدمهٔ کوتاه دربارهٔ وب‌کاوی می‌آوریم. سپس به نظراتی دربارهٔ حل نسبی مسئلهٔ کاوش می‌پردازیم و در پایان، یک توضیح کوتاه دربارهٔ کاربرد محاسبهٔ تقریبی در بازیابی اطلاعات می‌آوریم.

۲. چالش‌های داده‌ای

چندین موضوع داده‌ای وجود دارند که لازم است بررسی شوند، که از جمله باید صفحات پویا[۲۱] یا پنهان[۲۲]، داده‌های چندرسانه‌ای[۲۳]، داده‌های ساختاریافته[۲۴] و داده‌های معنایی[۲۵] را ذکر کنیم. سپس هر یک از آن‌ها را توضیح می‌دهیم، بجز داده‌های پنهان که یک مورد خاص از داده‌های عام، با مسئلهٔ دسترسی محدود می‌باشد.

۲-۱. داده‌های پویا

وب ایستا در مقایسه با محتوایی که بر اساس درخواست، بخصوص بنابر ارائهٔ درخواست در کسب و کار الکترونیکی یا در سایت‌های خدمات اطلاعاتی ایجاد می‌شود، کوچک شده است. نرم‌افزارهای کنونی کاوش می‌توانند پیوندهای پویا را دنبال کنند، اما این کار باید با دقت انجام شود، زیرا ممکن است هیچ محدودیتی وجود نداشته باشد، یا حتی یک صفحهٔ مشابه، دوباره و دوباره تولید شود. دسترسی به صفحاتی غیر از آنچه در فرم‌های پرسش درخواست می‌شود، از این هم دشوارتر است، زیرا کاوشگر شناختی از پایگاه اطلاعاتی ندارد. از طرف دیگر، حتی اگر پایگاه اطلاعاتی شناخته‌شده باشد، درخواست همهٔ سؤالات ممکن، بسیار وقتگیر خواهد بود (وابسته به اندازهٔ پایگاه اطلاعاتی، به طور تصاعدی افزایش می‌یابد) و حتی اگر فقط به درخواست‌های ساده بسنده کنیم، بعضی از این درخواست‌ها ممکن است هرگز توسط اشخاص حقیقی مطرح نشوند. خدمات وب، اگر امکان یادگیری از پایگاه اطلاعاتی و نحوهٔ پرسش افراد از آن را فراهم کنند، ممکن است راه حل نسبی برای این مسئله باشند. مثلاً به‌دست‌آوردن هزار درخواستی که بیشترین تکرار را دارند، ممکن است کافی باشد. امکان دیگر، تجزیه و تحلیل صفحه است، مانند کاری که در [۲۳] انجام شده.

۲-۲. داده‌های چندرسانه‌ای

داده‌های چندرسانه‌ای شامل تصاویر، تصاویر متحرک، صوت در چندین شکل، و ویدیو است. همهٔ این‌ها قالب استاندارد ندارند. متداول‌ترین آن‌ها JPG و GIF, PNG برای تصاویر، MP3 برای موسیقی، Real Video یا Quicktime برای ویدیو، و ... می‌باشند. راه‌حل ایده‌آل این است که بر روی هر نوع داده (از جمله متن)، با استفاده از الگوی یکسان و با زبان درخواست واحد، جستجو انجام شود. این هدف بلندپروازانه شاید امکانپذیر نباشد.

برای یک نوع دادهٔ خاص می‌توانیم یک مدل تشابه به وجود آوریم، و براساس نوع‌داده‌ها، زبان درخواست تغییر کند. مثلاً درخواست به وسیلهٔ مثال برای تصاویر، یا درخواست به‌وسیلهٔ زمزمه برای صوت. تمام این زمینه‌ها بیشتر متعلق به پردازش تصاویر و علائم[۲۶] است تا به بازیابی اطلاعات به روش کلاسیک.

۲-۳. داده‌های ساخت‌یافته

اغلب داده‌ها تا حدودی دارای ساختار هستند، و نهایتاً داده‌های نیمه ساخت‌یافته نام دارند. نمونه‌های آن پست الکترونیکی، اخبار ارسالی، و... هستند. اگر «ایکس‌ام‌ال» متداول شود، سطح ساختار باز هم بالاتر می‌رود. اولین چالش، طراحی مدل‌های داده‌ای و زبان‌های پرس‌وجوی مربوط به آن‌ها است که امکان می‌دهد محتوا و ساختار با هم درآمیزند. متن‌های ساخت‌یافته را در مرتبهٔ قبل از «ایکس‌ام‌ال» می‌دانستند و چندین واسطهٔ کارآیی/گویایی، طراحی شد [۳]. بعد از «ایکس‌ام‌ال»، «کنسرسیوم وب جهانی»، «ایکس کوئری»[۲۷] را به عنوان استاندارد معرفی کرده است [۲۷].[۲۸]

در هنگام بازیابی داده‌های «ایکس‌ام‌ال»، چندین چالش وجود دارد:

- پاسخ می‌تواند جزئی از «ایکس‌ام‌ال» باشد و حتماً لازم نیست یک شیء کامل باشد. با وجود این، پاسخ‌ها نیز باید داده‌های مبتنی بر«ایکس‌ام‌ال» باشند.
- بسیاری از پاسخ‌ها را می‌توان در یک شیء «ایکس‌ام‌ال» واحد آورد و می‌توانند با یکدیگر، همپوشانی داشته باشند.
- چگونه یک پاسخ را رتبه‌بندی کنیم و اگر لازم باشد پاسخ را در قالب انواع ساختارهای خاصی ارائه دهیم، و چگونه رتبه‌بندی را برای آن‌ها اعمال کنیم؟ گاهی اوقات با ترکیب درختواره‌های فرعی اگر نزدیک به هم باشند، رتبه‌بندی بهتری خواهیم داشت. اما در موارد دیگر اگر کاملاً دور از هم باشند، بهتر است.

تحقیقات اخیر دربارهٔ این موضوعات در [۶ ، ۷ ، ۱۷ ،۱۸] آمده است.

مسئلهٔ دیگر، پردازش جریانات «ایکس‌ام‌ال»، یعنی غربال‌کردن جریانی از اشیای «ایکس‌ام‌ال» به‌وسیله مجموعهٔ گسترده‌ای از پرسش‌ها می‌باشد. در این جا پرسش‌ها را می‌توان نمایه‌سازی کرد، اما داده‌ها را نمی‌توان. برای مطالعهٔ مقدمه‌ای بر این مسئله،[۲۴] را ببینید.

۲-۴. داده‌های معنایی

دو مسئلهٔ اصلی در رابطه با اطلاعات معنایی، استانداردهای مربوط به فراداده‌هایی هستند که معنا، وکیفیت یا درجهٔ اطمینان‌پذیری یک منبع اطلاعاتی را توصیف می‌کنند. در مورد اولین مسئله [یعنی استانداردهای معنا] «کنسرسیوم وب» اقدام می‌کند، اما برای مسئله دوم [یعنی کیفیت یا درجهٔ اطمینان‌پذیری]، نیاز به «طرح‌های تصدیق»[۲۹] می‌باشد که باید در آینده ایجاد شوند.

مسائل دیگر، موضوعات متداول مانند درجه‌بندی، سرعت تغییر، فقدان انسجام ارجاعی (پیوندها فیزیکی می‌باشند نه منطقی)، اختیارات توزیع‌شده، محتوا و کیفیت نامتجانس، منابع چندگانه، و ... می‌باشند. مقدمه‌ای بر این مسائل و دیگر چالش‌های «وب معنایی» در [۸، ۲۱، ۲۶] ارائه شده است.

۳. وب‌کاوی

ما در بازیابی اطلاعات، معمولاً پرسش را می‌دانیم. داده‌کاوی زمانی انجام می‌شود که پرسش را نمی‌دانیم. از این رو، سعی می‌کنیم روابطی در داده‌ها پیدا کنیم که مانند یک پاسخ جالب به نظر برسند، سپس این پاسخ را بررسی می‌کنیم تا پرسش متناظر با آن را پیدا کنیم. در وب، این کار منجر به وب‌کاوی می‌شود، یعنی چالش دیگری فراتر از بازیابی اطلاعات در وب. بعضی نویسندگان بازیابی اطلاعات را نیز جزو وب‌کاوی به شمار می‌آورند، که به عقیدهٔ ما صحیح نیست. وب‌کاوی شامل استخراج اطلاعات، و به دنبال آن تعمیم و تحلیل این اطلاعات است.

سه نوع دادهٔ و بی‌وجود دارد که می‌توان آن‌ها را کاوید: محتوا، کاربرد و ساختار. محتوا، شامل کاویدن متن و چندرسانه‌ای‌ها می‌شود. کاربرد، شامل کاویدن گزارش[۳۰] وب (مشتمل بر گزارش جستجوها و دیگر داده‌های کاربردی) است. ساختار به معنای تحلیل ساختار پیوندهای وب می‌باشد (اما این مطلب، با توجه به امکان کاویدن در ساختار«ایکس‌ام‌ال»، مبهم است). بعلاوه برای هر سه مورد، ما یک بُعد موقتی که مربوط به پویایی چگونگی رشد و تغییرات وب می‌باشد، داریم که دلالت بر داده‌های موقتی دارد. دو نوع اول در [۱۴] بررسی شده‌اند و نوع سوم، موضوع اصلی [۱۲] است. نوع سوم کمتر مورد بررسی قرار گرفته و بعضی از نتایج در ارتباط با آن، در[۴] ارائه شده است.

از وب‌کاوی می‌توان علاوه بر یافتن اطلاعات یا دانش جدید، برای مقاصد گوناگون استفاده کرد: برای طراحی انطباقی وب (مثلاً طراحی وب با انگیزش ناشی از کاربر)، سازماندهی دوبارهٔ وب‌سایت، شخصی‌سازی وب‌سایت، و موارد گوناگون بهبود در اجرا.

۴. به سوی موتور کامل جستجوردر وب

یک موتور کامل جستجو، مسائلی را که قبلاً ذکر شد می‌تواند حل کند، هر نوع داده‌ای را بازیابی نماید و اطلاعات را برای انجام بهتر وب‌کاوی، جمع‌آوری کند. اما مشکل امروزه همچنان باقی خواهد بود: جمع‌آوری داده‌ها. مسئلهٔ کاوش به حجم و رشد داده، همراه با داده‌های متغیر و مشابه، و یک تکنیک بسیار ناکارآمد مربوط می‌شود: بازکشی[۳۱]

موتورهای جستجوی کنونی کارشان را بدون همکاری خدمت‌دهنده‌های وب[۳۲] انجام می‌دهند؛ آن‌ها باید صفحات را با استفاده از پروتکل استاندارد «اچ‌تی‌تی‌پی» از طریق اتصالات «تی‌سی‌پی» اَسکی انتقال دهند، و آن‌ها را بسنجند تا ببینند آیا صفحه‌ای تغییرکرده است یا نه، تا بعد از استخراج صفحات جدید یا روزآمد شده، نمایه‌های خود را به روز کنند.

از همه بهتر این است که یک واسطه[۳۳] برای خدمت‌دهنده بفرستیم، یعنی به جایی که می‌تواند به طور محلی به دنبال صفحات و پیوندهای جدید و صفحات اصلاح‌شده بگردد. همچنین این واسطه می‌تواند همهٔ صفحات روزآمد شده را با هم به صورت یک فایل فشرده جمع‌آوری کند تا به موتور جستجو انتقال یابد. خدمت‌دهندهٔ اصلی جستجو می‌تواند با واسطهٔ دور، در تعامل باشد تا براساس چندین پارامتر مثل تعداد فایل‌ها، اهمیت آن‌ها و ... تصمیم بگیرد آیا ارزش دارد که گروه موجود، انتقال یابد یا نه. سپس می‌توان اطلاعات کاوشگر را بین موتور جستجوی اصلی و واسطهٔ موجود، توزیع کرد. «براندمن» و دیگران [۹] تأثیر باند پهن را، در هنگامی که خدمت‌دهنده‌های شبکه، فراداده‌های صفحات وب خود (مانند تاریخ‌های انجام، اندازه، و ...) را منتشر می‌کنند بررسی نمایند و نشان می‌دهند که ذخیره‌سازی‌هایی وجود دارند و جدید بودن صفحات نیز افزایش می‌یابد. مقالهٔ مشابهی بر جدید بودن تأکید می‌کند [۱۹]. اما می‌توانیم قدمی دیگر به جلو برداریم و به جای بازکشی اطلاعات به تنهایی، اطلاعات را عرضه[۳۴] کنیم.

سپس، تعامل از بازکشی صفحات به عرضه‌کردن تغییرات کشیده می‌شود. طبق معمول، زیاده‌روی مؤثر نیست، و عرضه‌کردن اطلاعات زیادی، بار خدمت‌دهندهٔ مرکزی را زیاد می‌کند. از این رو، بهترین راه حل این است که خدمت‌دهنده، از قبل در این باره که چه موقع و چگونه پیغامی بفرستد تا اطلاع دهد که یک گروه از تغییرات آماده است (یا حتی بهتر، این که تغییرات، نمایه شده است و قسمتی از نمایه در دسترس است) با واسطه مذاکره کند. سپس خدمت‌دهندهٔ

اصلی در موقع مقرر، آن تغییرات را بازکشی خواهد کرد. این به معنای یک برنامه‌ریزی بلندمدت است، که در نتیجه وقتی خدمت‌دهندهٔ وب را، که یک هشدار عرضه می‌کند واقعاً بازبینی می‌نماید، تغییرات بیشتری را پیدا کند. اما این برنامه‌ریزی ساده‌تر از برنامه‌ریزی‌های کنونی است، زیرا که ما اطلاعات بیشتری داریم، و نیازی نیست که نگران رفتار با نزاکت باشیم، چرا که مطمئن هستیم تمام دسترسی‌ها تکراری نیستند و همیشه با موفقیت همراه‌اند.

عموماً خدمت‌دهنده‌های وب می‌خواهند که در این چیدمان همکاری داشته باشند، زیرا امروزه نمایه‌شدن در یک موتور جستجوی مشهور، یک ارزش پذیرفته‌شده می‌باشد. از طرف دیگر، حتی اگر چرخه‌های «سی‌پی‌یو»[۳۵] را به نفع موتور جستجو به کار گیرند، کاوشگر آن‌ها را سنجش نمی‌کند؛ بنابراین بار دسترسی به خدمت‌دهندهٔ وب را به طور مؤثری کاهش می‌دهند. همچنین این چرخه‌ها می‌توانند در دوره‌هایی که بار کمتر است، به کار گرفته شوند.

به عنوان مرحلهٔ اول آزمایش، در هنگامی که یک خط‌مشی[۳۶] ۱ واسطه که به صورت جهانی قابل دسترس باشد وجود ندارد، یک مدول[۳۷] ساده، همراه با خدمت‌دهندهٔ وب، می‌توان برای تأمین کارآیی مشابه و برای اندازه‌گیری میزان بهبود عملکرد، ایجاد کرد. همان طور که قبلاً ذکر کردیم، تغییرات کوچک در خدمت‌دهندهٔ وب، برای ایجاد امکان همکاری با موتورهای جستجو، پیشنهاد شده‌اند [۹ و ۱۹]. اما این [موتورها] فاقد انعطاف‌پذیری هستند و در خط‌مشی‌های کاوشگر، اختلال ایجاد می‌کنند. واسطه‌ها این رفتار را بسیار بهبود می‌بخشند و به الگوریتم‌های خود این امکان را می‌دهند که صفحات را برای این که در کد واسطه‌ها گنجانده شوند، اولویت‌بندی کنند. از این نظر، واسطه جز مهمی از الگوریتم کاوشگر است و منطق آن، از خط مشی‌های یک موتور جستجوی خاص پیروی می‌کند [۵].

محاسبات تقریبی و بازیابی اطلاعات

همان طور که در مقدمه ذکر کردیم، واژهٔ «محاسبات تقریبی» بوسیلهٔ «لطفی‌زاده» ارائه شد و در همگرایی روش‌هایی که برای حل مسائلی که نیاز به نوعی هوش (که از محاسبات کلاسیک ناشی می‌شود) دارند، مفید است. «محاسبات تقریبی» یک مجموعه از فنون مناسب برای رفع ابهام، ذهن‌گرایی، و کلیت موجود در برخی مسائل می‌باشد.

هدف بازیابی اطلاعات مدلسازی، طراحی، و اجرای سیستم‌هایی است که قادر باشند دسترسی سریع و کارآمد بر پایهٔ محتوا را به مقادیر عظیم اطلاعات، تأمین کنند. هدف یک سیستم بازیابی اطلاعات، برآورد ربط اقلام اطلاعاتی با نیازهای اطلاعاتی یک کاربر (که در قالب

یک سؤال بیان شده) می‌باشد. این، کار مشکل و پیچیده‌ای است؛ زیرا با ذهنیت، ابهام و عدم دقت آکنده است.

محاسبات تقریبی روش‌های متفاوتی از قبیل منطق فازی، الگوریتم‌های ژنتیکی، شبکه‌های عصبی، مجموعه‌های نادقیق، و شبکه‌های بیزی را شامل می‌شود. مسئله بازیابی اطلاعات، یک حوزهٔ کاربردی معمول برای محاسبات تقریبی است. بعضی از رویکردهای اصلی محاسبات تقریبی در بازیابی اطلاعات از این قرارند:

منطق و مجموعه‌های فازی: ترکیب اطلاعات، استخراج متن، مدل‌های زبان پرس‌وجو، و خوشه‌بندی مدارک؛

شبکه‌های عصبی: رده‌بندی و خوشه‌بندی اسناد و اصطلاحات، و بازیابی چندرسانه‌ای‌ها؛

الگوریتم‌های ژنتیکی: رده‌بندی مدارک، بازیابی تصویر، بازخورد ربط، و یادگیری پرس‌وجو؛

تکنیک‌های احتمالاتی: رتبه‌بندی، وب‌کاوی.

مجموعه‌های نادقیق و منطق‌های چند ارزشی: خوشه‌بندی مدارک؛

شبکه‌های بیزی: مدل‌های بازیابی، رتبه‌بندی، ساخت اصطلاحنامه، و بازخورد ربط.

حداقل صد مقاله به این مسائلی که ذکر شد، اختصاص یافته و برشمردن همهٔ آن‌ها نیاز به یک بررسی کامل دارد. ولی ما خواننده را به کتاب «میاموتو» [۲۲]، و نیز به یک کتاب عالی با ویراستاری «کرستانی» و «پاسی»[۱۶]، شمارهٔ ویژه‌ای از IP&M [۱۵]، و یک مقالهٔ پیمایشی از «چن»، و شمارهٔ حاضر از این مجله ارجاع می‌دهیم.

حداقل نیمی از مسائلی را که در مقدمه و بخش‌های بعدی ذکر کردیم، می‌توان با شیوه‌های بالا از میان برداشت. از این رو، تحقیقات بیشتری در پیش روی ما قرار می‌گیرد. اصلی‌ترین مسائل شاید موضوعات عملکردی (مثلاً این که آیا می‌توان با زمان پاسخ محدود، در موقعیت‌های عملی استفاده کرد؟) و توضیح پاسخ (مثلاً این که چرا یک مدرک در یک طبقهٔ معین رده‌بندی می‌شود؟) باشند. جدیدترین کاربردهای محاسبات تقریبی در بازیابی اطلاعات در وب، شامل واسطه‌های سازگار، پروفایل‌های کاربر، طبقه‌بندی صفحات وب، سنجش کیفیت، و ... می‌باشند. بنابراین، این نشان می‌دهد که پیشرفت در زمینهٔ بازیابی اطلاعات در وب، با استفاده از شیوه‌های محاسبات تقریبی، امکانپذیر است.

پی‌نوشت

[1]. Baeza-Yates, Ricardo (2003). "Information retrieval in the web: Beyond current search engines" International Journal of Appoximate Reasoning. 34: 97-104.
[2]. web
[3]. search engines
[4]. Crawling
[5]. spam
[6]. adversarial IR
[7]. HTML (HyperText Markup Language) زبان نشانه‌گذاری فرامتن
[8]. web logs
[9]. navigational
[10]. transactional
[11]. XML (Xtended Markup Language) زبان نشانه‌گذاری توسعه‌پذیر
[12]. semantic web
[13]. metadata
[14]. web mining
[15]. Soft Computing (SC)
[16]. fuzzy logic
[17]. genetic algoritm
[18]. neural network
[19]. rough sets
[20]. Bayesian network
[21]. dynamic pages
[22]. hidden
[23]. Multimedia data
[24]. Structures data
[25]. semantic data
[26]. signal
[27]. X Query
[۲۸]. اگرچه Xpath و XSLT را نیز می‌توان زبان پرس‌وجو دانست، اما برای مقاصد دیگری طراحی می‌شوند.
[29]. certification schemes
[30]. log
[31]. pulling
[32]. Web server
[33]. agent
[34]. pushing
[35]. CPU (Central Processor Unit)
[36]. Platform
[37]. module