

نوشته: لئارت بچورن بَرْن و پيتر اينگورسن\*  
ترجمه: عليرضا نوروزي و زهرا بيگدلي\*\*

### چکیده

از اواسط دهه ۱۹۹۰ حوزه پژوهشي جديدی به نام «وب سنجي» بر پایه روشهاي اطلاع سنجي جديد به وجود آمد که کار آن، پژوهش درباره ماهيت و خصوصيات وب است. اين مقاله سعی دارد حوزههاي منتخبي از پژوهش وب سنجي را که روند و فضاي مناسبی برای توسعه دارند و همچنين برخی حوزههايی را که در حال حاضر کمتر دارای آیندهاي روشن هستند معرفی کند. پژوهشهاي جديد درباره پوشش و عملکرد موتورهاي کاوش، آنها را به عنوان قالب و چارچوبی برای تحليل گزينشي کیفیت و محتوا، شناخته است. مشکلات مربوط به عوامل تأثیرگذار وب مورد بحث قرار گرفته است و در پایان مقاله راهنماییهاي جديد وب سنجي برای اجرائی کشف دانش و ردیابی موضوع روي وب-که تا حدودی مبتني بر روش کتابسنجي استفاده شده در پایگاههاي اطلاعاتي استنادي و کتابشناختي است- به صورت کلی مطرح شده است. در این چارچوب راهبردهاي نظريه نمودار یکپارچه، شامل: تحليل مسیر، پیوندهاي متقاطع (عرضي)، پیوندهاي ضعیف و پدیده جهان کوچک مورد توجه قرار گرفته است.

**کلیدواژه‌ها:** وب سنجي/ کشف دانش، ردیابی موضوع/ تحليل مسیر/ دانش عمومي کشف نشده/ پیوندهاي متقاطع (عرضي)/ نظريه نمودار/ عوامل تأثیرگذار وب

### مقدمه

از اواسط دهه ۱۹۹۰ تلاشهاي روزافزوني برای بررسی ماهيت و خصوصيات وب جهان گستر که در این مقاله «وب» نامیده می شود، با به کارگیری روش اطلاع سنجي جديد برای فضاي

محتویات آن، ساختار پیوندها<sup>۱</sup> و موتورهاي کاوش، صورت پذیرفت. مطالعه روي وب در سال ۱۹۹۷ را آلمایند و اینگورسن<sup>۲</sup> وب سنجي نامیده و یا در مجلهاي الکترونيکی با عنوان سایبرمتریک<sup>۳</sup> در سال ۱۹۹۷ سایبرمتریک (مجازي سنجي) نامگذاری شد. این مقاله سعی دارد حوزههاي منتخبي از پژوهش وب سنجي را که فضاي مناسبی برای توسعه دارند، معرفی کند. این نوشته مقاله کلامي نیست، اما نسبتاً تخصصي است.

وب سنجي شباهتهاي متعددي با مطالعات علم سنجي، اطلاع سنجي و کاربرد روشهاي کتابسنجي متداول دارد. برای مثال، محاسبات ساده و تحليل محتوای صفحات وب، شبیه به تحليل انتشارات سنتي هستند؛ محاسبه و تجزيه و تحليل پیوندهاي بیرون رونده از صفحات وب که در اینجا پیوندهاي بیروني<sup>۴</sup> و پیوندهايی که به خود صفحات وب داده می شوند، پیوندهاي دروني<sup>۵</sup> نامیده می شوند به ترتیب به عنوان مرجع (مأخذ)<sup>۶</sup> و تحليل استنادي<sup>۷</sup> در نظر گرفته می شوند. بنابراین، در مقالات علمي، پیوندهاي بیروني و دروني به ترتیب شبیه مراجع (مأخذ) و استنادها هستند. هرچند وب به علت ماهيت پويا و توزیعی، غالباً صفحاتش را که به صورت همزمان به یکدیگر پیوند داده شده اند- حالي که در کتاب به شکل سنتي امکان ندارد- در فضایی که بر پایه استناد گذارده شده نمایش می دهد. پوشش موتورهاي کاوش در سراسر وب می تواند به همان شیوه اي که پوشش حوزه<sup>۸</sup> و پایگاههاي اطلاعاتي استنادي

\*Bjorneborn, Lennart & Ingwersen, Peter (2001). Perspectives of webometrics, Scientometrics, 50(1), p. 65-82.

\*\* دانشجوی دکتری اطلاع رسانی کارشناس کتابداری و اطلاع رسانی#

در کل اسناد و مدارک و امکان همپوشانی‌های بین موتورهای شناسایی شده، مورد پژوهش قرار گیرد. از آن زمانی که وب از مشارکت هر کسی که دوست داشت در آن شرکت کند برخوردار شد، در نتیجه فقدان بازبینی‌های دقیق، کیفیت اطلاعات یا ارزش دانش دچار ابهام گردید؛ اما تحلیل پیوندها ممکن است گروهی از سایت‌های تجدیدنظر و بازبینی شده را نشان دهد. الگوهای رفتار کاوش در وب<sup>۹</sup> می‌تواند، مانند مطالعات اطلاع‌یابی سنتی، مورد پژوهش قرار گیرد.

ردیابی موضوع در محیط وب تأمین شده و اقداماتی برای کشف دانش، شبیه به داده‌های رایج (مشترک<sup>۱۰</sup>) یا متن کاوی (استخراج متن<sup>۱۱</sup>) در پایگاه‌های اطلاعاتی متنی (کتابشناختی) یا اداری (اجرایی) انجام گرفته است. از زمانی که وب فضای اطلاعاتی کاملاً متفاوتی از سایر پایگاه‌های اطلاعاتی علمی یا حرفه‌ای متداول، شده است، بعضی وقت‌ها شباهت‌های مذکور ممکن است به صورت کم عمق و سطحی ظاهر گردد. برای مثال، ما با اطمینان نمی‌دانیم که چرا مردم در وب به سایر صفحات پیوند می‌دهند.

هیچ قراردادی برای اسناد در محیط وب همانند آثار علمی چاپی وجود ندارد. علاوه بر این، زمان نقش متفاوتی بر روی وب ایفا می‌نماید. از سوی دیگر، به دلیل این که وب مجموعه‌ای است بسیار پیچیده از کلیه انواع اطلاعاتی که به وسیله انسان‌های متفاوت تولید می‌شود و همچنین توسط کاربران مختلف مورد جستجو قرار می‌گیرد، موضوع جالبی برای پژوهش است، و در واقع، اطلاع‌سنجی روش‌هایی را برای شروع پژوهش در وب ارائه می‌کند. هرچند یک نفر باید با درخواست پیوسته از پایگاه‌های اطلاعاتی استنادی کنترل شده توسط مؤسسه اطلاعات علمی<sup>۱۲</sup> آگاه شود، برای مثال، از طریق زبان دستور<sup>۱۳</sup> پایگاه اطلاعاتی دایالوگ. گردآوری داده‌ها در محیط وب، به ویژگی‌ها

و خصوصیات بازیابی موتورهای کاوش مختلف و روبات‌های وب، بستگی دارد. پیش از ظهور دستور "set posting on" در پایگاه اطلاعاتی دایالوگ<sup>۱۴</sup> در طی دهه ۱۹۹۰، محاسبه اسنادها به صورت پیوسته امکان پذیر نبود. یک نفر باید تمام مدارک اسناد شده را جهت تحلیل به صورت محلی، به منظور شمارش تعداد واقعی اسنادها در فضای اطلاعاتی تعریف شده به وسیله مؤسسه اطلاعات علمی، بارگذاری (فروفرستی) می‌کرد. امروزه این حالتی است که در اکثر موتورهای وب وجود دارد که روسو در سال‌های ۱۹۹۷ و ۱۹۹۹ آن را اثبات کرد. این موتورها کل وب را نمایه‌سازی نمی‌کنند، همپوشانی آنها قابل توجه نیست (لارنس و گیلز، ۱۹۹۹) و ویژگی‌های بازیابی آنها برای تحلیل‌های وب‌سنجی گسترده پیوسته<sup>۱۵</sup> بسیار ساده است. بنابراین، نمونه‌گیری بسیار مهم و حساس است، ولی اجرای آن مشکل است و پالایش آن ضروری است، به همین جهت مهندسی مجدد و پاکسازی اطلاعات در تحلیل‌های وب‌سنجی عنصری مهم است.

این مقاله به برخی از پژوهش‌های اخیر که عملکرد و پوشش موتورهای وب را به عنوان چارچوبی برای تحلیل‌های گزینشی کیفیت و محتوا مورد توجه قرار داده اند، اشاره می‌نماید و سپس ما به دنبال تحلیل سایت‌ها (پیوندها)<sup>۱۶</sup> مانند تحلیل «پیوند-صفحه»<sup>۱۷</sup> در اصطلاحات روسو (۱۹۹۷) و مطالعات عامل تأثیرگذار وب هستیم. تلاش‌هایی به منظور فراهم کردن مسیرهای جدید خروجی (برون رفت) وب‌سنجی از طریق اجرای کشف دانش و ردیابی موضوع، به عنوان مثال به وسیله ساختار پیوندهای متقاطع و پیوندهای ضعیف، در نتیجه‌گیری مقاله مورد بحث قرار گرفته اند.

**پژوهش‌های کیفی و پوششی موتورهای وب**

لورنس و گیلز (۱۹۹۸) مقاله‌ای بنیادی نوشتند که در آن پوشش موتورهای کاوش تجاری در محیط وب را با معرفی محتوای وب قابل نمایه‌سازی<sup>۱۸</sup>، مدنظر قرار دادند. مفهوم وب قابل نمایه‌سازی بر بخشی از وب که می‌تواند با موتورهای کاوش نمایه شود به استثنای اسناد و مدارک پایگاه‌های اطلاعاتی، مانند دایالوگ، دلالت دارد. آزمون گسترده‌ای که در دسامبر ۱۹۹۷ بین ۶ موتور کاوش برتر تجاری اصلی یعنی: آلتاویستا، هات بات، نورثرن لایت، اینفوسیک، لیکوس و اکسایت انجام گرفت، سطح پایینی از قابلیت نمایه‌سازی وب در حدود ۳۲۰ میلیون صفحه را نشان داد. این پژوهش همچنین نشان داد که پوشش هر یک از موتورها به صورت قابل ملاحظه‌ای، با نمایه‌سازی یک سوم «وب قابل نمایه‌سازی» محدود شده‌است.

احتمالاً دلایل زیادی برای این نتیجه‌گیری وجود دارد. به عنوان مثال، عمق و جامعیت نمایه‌سازی در خدمت دهنده<sup>۱۹</sup>های محلی بازدید شده به وسیله موتورهای کاوش به چگونگی سازماندهی و ساختار یک سایت، که ممکن است خروجی بازیابی را تحت تأثیر قرار دهد بستگی دارد، همان طور که ممکن است روش نمایه‌سازی، واژه‌های پرسش (سؤال) را تحت تأثیر قرار دهد. برای مثال، از طریق کوتاه‌سازی تحمیلی<sup>۲۰</sup> مدارک وبی طولانی، تلاش‌های دیگری نیز برای ارزشیابی موتورهای وب انجام شده است، مثلاً مشاهده کیفیت فهرست‌های رتبه بندی شده مدارک وب که به وسیله موتورهای اصلی بازیابی شده‌اند (کورتوا و بری، ۱۹۹۹).

در کنار سایر یافته‌ها، این مقاله در مورد دانش کم و بیش در دسترس عموم و درباره ویژگی‌های مختلف نمایه‌سازی و بازیابی که به وسیله هر یک از موتورها استفاده شده است،

بحث می‌کند. روش ارزیابی موتورهای وب به وسیله کلارک و ویلت<sup>۲۱</sup> (۱۹۹۷) که آلتاویستا، لیکوس و اکسایت را با هم مقایسه نمودند، مطرح شده است. علاوه بر این، آن مقاله، ارزیابی انتقادی از پژوهش‌های پیشین را نشان داده و روشی واقع‌بینانه و معقول را که شامل اندازه‌گیری جامعیت نسبی است، فراهم می‌کند. همچنین آن مقاله نتیجه گرفت که آلتاویستا به صورت قابل ملاحظه‌ای بهتر از لیکوس و اکسایت عمل کرده است. اُپنهایم<sup>۲۲</sup> و دیگران (۲۰۰۰) پژوهشی مشروح و روزآمد درباره ارزیابی موتورهای کاوش وب، شامل بحثی در باب روش‌های آزمون، فراهم نمودند. در حالی که بسیاری از مطالعات ارزیابی و پوششی به ربط و تعداد صفحات وب در یک زمان معین توجه می‌کنند، سایر تحلیل‌های انتقادی، بازیابی پیوند- صفحه را در بر می‌گیرند (سیندر و رزن‌باوم، ۱۹۹۹) و یا پژوهش‌های ساختاری یا پوششی وب مبتنی بر سری‌های زمانی<sup>۲۳</sup> را پوشش می‌دهند. اسنایدر و روزنوم نیز مانند اینگورسن (۱۹۹۸) ناهماهنگی‌ها و تفاوت‌های زیادی را به ویژه در رابطه با بازیابی پیوند-صفحه موتور آلتاویستا مشاهده کردند. بی‌نظمی‌های آن موتور را همچنین بار- ایلان (۱۹۹۹) در یک مطالعه طولی<sup>۲۴</sup> و نیز روسو (۱۹۹۹) کسی که دو موتور آلتاویستا و نورثرن لایت را طی ۲۱ هفته به صورت روزانه در طی سال ۱۹۹۹ مورد مقایسه قرار داده بود، گزارش کرده‌اند.

مطالعه بعدی از سه واژه مفرد عام یکسان به عنوان پرسش (واژه‌های جستجو) در طول مدت ارزیابی استفاده کرد. در حالی که نورثرن لایت، همان طور که پیش‌بینی شده بود، یک افزایش ثابت و منظم از موفقیت‌های همزمان با گسترش وب را نشان می‌داد، آلتاویستا در شکل (فرم) بسیار ثابت و جدیدی مجدداً آغاز به کار کرد و گوناگونی و تغییرات زیادی را در طول زمان تا یک

حالت مخفي به سمت نشان دادن طيفي از قابليت‌ها شكل گرفته است (كرونين و مك‌كيم، ۱۹۹۶، ص ۱۷۰). اين عقیده و تصور هم اکنون واقعيت است. بنا براین، تحليل‌هاي وب‌سنجی ماهيت، ساختارها و ویژگی‌هاي محتوایی سایت‌ها و صفحات وب و همچنین ساختار پیوندها به منظور درك بزرگراه‌هاي مجازي و ارتباطات دروني آنها مهم هستند.

لارسون (۱۹۹۶) يکي از اولين متخصصان اطلاع رسانی بود که یک تحليل مقدماتي درباره ساختار فکري و عقلاني فضاي مجازي<sup>۲۶</sup> اجرا نمود. چندي بعد، آلمایند و اینگورسن (۱۹۹۷) روش‌هاي متنوعي شبیه کتابسنجی را براي بخش‌هاي شمالي وب<sup>۲۷</sup> به منظور مشاهده انواع اتصالات<sup>۲۸</sup> (پیوندهاي) صفحه و تعريف نوع شناسي<sup>۲۹</sup> صفحات وب که عملاً در سطح کشورهای شمالي پيدا شده‌اند، به کار بردند. روش پژوهش، شامل نمونه‌گیری طبقه‌اي از صفحات وب و بارگذاري به منظور تحليل محلي بود. در میان یافته‌هاي جالب توجه، این تحليل‌ها نشان داده‌اند که هر صفحه وب قادر به پیوند بیروني، تقریباً به طور متوسط ۹ پیوند بیروني فراهم مي کند. تناسبی که امروزه در رشد تصاعدي فضاي وب حفظ مي‌شود. این مقاله، همچنین تلاش کرده است مقایسه‌اي میان بخش صفحات وب علمي ارزیابی شده و پراکندگی یافته شده در نمایه‌هاي استنادي میان کشورهای شمالي را انجام دهد. مسلماً، دامنه دید روي وب از نمایش در پایگاه‌هاي اطلاعاتي استنادي کاملاً متفاوت است. براي مثال نروژ، در زمان انجام این تحليل در سطح وب نسبت به دنياي چاپ بسیار مشهورتر بود. اهميت ویژه صفحه‌هاي خانگی شخصي به وسیله بتس و لو<sup>۳۰</sup> (۱۹۹۷) و نیز توسط واین و کتر<sup>۳۱</sup> (۱۹۹۷) نشان داده شده است که عمدتاً روي صحنه فضاي مجازي ایالات متحده آمریکا متمرکز شده است که

تاریخ به خصوص (۲۵ اکتبر ۱۹۹۹) نشان مي داد. در آن تاریخ تعداد صفحات وب بازیابی شده به طور چشمگیری همراه با این تأثیر شبه نو اختر<sup>۳۵</sup> مبتني بر پرسش افزایش یافت (روسو، ۱۹۹۹، ص ۵). بعدها تا حدودي احتمالاً به علت حذف پیوند- صفحه‌هاي غیرفعال و خاموش، این تعداد کاهش یافت.

روسو هنوز استفاده از یک صافي میانه را براي کاهش تأثیر گوناگوني نتایج در آن موتور خاص پیشنهاد مي‌کند. نتیجه دیگر آن مطالعه این است که نتایج عوامل تأثیرگذار وب که اینگورسن در سال ۱۹۹۸ منتشر کرد احتمالاً بسیار مشکوک و غیرقابل اطمینان هستند، چون که نتایج مجموعه داده‌هاي او، هم براي صفحات وب و هم براي صفحات پیوند دروني از نسخه بي ثبات و قدیمی آلتاویستا ناشی شده است. دلیل این که چرا کانون تمرکز روي آلتاویستا قرار دارد، این است که موتور کاوش آلتاویستا به طور وسیعی وب را پوشش مي دهد و بدین نحو شرایط جستجوی مناسبی را براي مطالعات اطلاع‌سنجی وب فراهم مي‌کند. سري‌هاي (توالي) زمانی به نظر مي رسد، به عنوان وسیله‌اي که نمایشگر عملکرد موتور وب است بسیار مفید باشد.

### کیفیت و ویژگی‌هاي صفحه وب

مطابق نظر کرونين و مک کیم، وب در حال تغییر شکل روش‌هايي است که از طریق آنها محققان با یکدیگر ارتباط برقرار کنند. انواع جدیدی از انتشارات علمي و آثار دست اول (از قبیل تحقیقات در حال اجرا، آثار آماده چاپ و پیش نویس هاي قدیمی) پدیدار شده‌اند. امروزه آثار در حال اجرا، اعلامیه‌هاي دیواري، طرح هاي مقدماتي پیشین و مقاله‌هاي داوري شده تقریباً بلافاصله قابل اشتراك هستند. پدیدآورندگان مي‌توانند میان پخش محدود و پخش گسترده، انتخاب کنند و بررسی دقیق از

بعدها براي مثال توسط ديلون و گوشروسكي<sup>۳۳</sup> (۲۰۰۰) پيگيري و دنبال شده است.

به طور آشكار اين دستاورد و پيشرفت به هرکسي اجازه مي دهد تا مسائل خود را عملاً و بدون نظارت مقامات صلاحيت دار بيان نمايد و همچنين از طريق ايجاد پيوند به صفحاتي که یک شخص مي خواهد متصل شود، به خاطر وجود پيوند به آن صفحات، اعتبار کسب کند و نیز موجب فراهم کردن امکان دسترسي به داده‌ها، اطلاعات، ارزش‌ها، و دانش در شکل‌ها و مقادير متعددي شود و آزادي اطلاعات را حتي در مناطق و کشورهايي که زيرساخت ضعيفي دارند، ايجاد کند. روي ديگر سکه اين است که وب، پيش از پيش به محيطي نامطمئن و قابل شك براي استفاده کنندگانش تبديل شده است؛ خط قرمز باريک موجود ميان پيچيدگي، حقيقت تاريخ و مبهم، اطلاعات غلط، عقايد، نظرات، تصورات يا تأملات و پايابي، كيفيت، روايي، ربط يا حقيقت، به طور روزافزوني کمتر شده است و اين عين واقعيت است. باستان شناسي وب<sup>۳۳</sup>، در آينده دست در دست روش‌ها و تحليل‌هاي وب‌سنجی پيش خواهد رفت.

در حال حاضر ارزيابي و مشاهده كيفيت، بسيار مورد نياز است؛ به ويژه حوزه‌هاي پزشکي و بهداشت، زمينه‌هاي مهمي به منظور پژوهش براي چنين موضوع هايي هستند. لي کوي<sup>۳۴</sup> اخيراً استفاده از روش‌هاي تحليل استنادي روي وب را براي مشخص کردن سايت‌هاي پيوند دروني شده پُرسامد و همپوشاني در زمينه اطلاعات پزشکي (۱۹۹۹) ابداع نموده است. آلن و ديگران نیز به اعتبار و ربط صفحات وب مرتبط به هم توجه کرده‌اند. در مقاله کوي، یک نفر به ساير مطالعات اخير موضوعات هنر و بهداشت روي وب، ارجاع داده است به اين دليل که تحليل استنادي وب به عنوان

يکي از شاخص هاي اساسي كيفيت به کار برده شده است (اشن‌باخ، ۱۹۹۸).

توزيع برادفورد از هزاران پيوند خاطر نشان شده از سوي ۲۵ دانشکده از بهترين دانشکده‌هاي پزشکي آمريکا به عنوان پيوندهاي قوي، توسط کوي، به منظور نشان دادن سايت‌هاي هسته<sup>۳۵</sup> مربوط به موضوع هاي خاص بهداشت به کار برده شده است. مقاله آلن و ديگران، یک پژوهش کارشناسانه درباره پايابي وب‌سايت‌هاي علمي است. همانند موردی که براي مطالعه طولی روسو (۱۹۹۹) که در بالا ذکر شد، اين پژوهش پيمائشي مبتني است بر بازباني سايت‌ها مطابق با سه سؤال نمونه درباره: ۱. ارزيابي ۲. نظام تعريف شده به شکل ژنتيکي ۳. گونه‌هاي در معرض خطر. براي هر پرسش<sup>۳۶</sup>، ۵۰۰ وب‌سايت، نخست به طور متوالي و پي در پي و مستقل توسط دو داور متخصص بررسي شدند. تقريباً ۶۰ سايت حاوي اطلاعات مربوط به موضوع بود. اين روش شبیه روشي است که در آزمايش‌هاي جهاني اخير ارزيابي «بازباني اطلاعات در کنفرانس بازباني متن»<sup>۳۷</sup> استفاده شده است. اين ۶۰ سايت موضوعي براي هر پژوهش پس از بررسي با عناوين ذيل امتياز بندي شدند:

- نادرست<sup>۳۸</sup>: اگر آنها واقعاً حاوي اطلاعات اشتباه بودند.

- گمراه کننده<sup>۳۹</sup>: اگر آنها علم را غلط تعبير نموده بودند يا به طور وقیحانه‌اي حقايقی را که یک موقعيت مخالف را تأييد مي کرد، حذف کرده بودند.

- بي مأخذ<sup>۴۰</sup>: اگر آنها اطلاعات را بدون هيچ گونه مأخذ بررسي شده دقيق ارائه کرده بودند.

امتياز آخر (يعني بي مأخذ) کاملاً عيني و مشخص است. در مجموع، مقدار امتيازهاي مورد توافق داوران به منظور امتيازبندي مقوله‌هاي نادرست و گمراه کننده به اين شرح بودند: ۸۷/۸ درصد براي سايت‌هاي ارزيابي، ۸۲/۸

براي سايت‌هاي نظام تعريف شده به شكل ژنتيكي و ۷۳/۶ درصد براي وب‌سايت‌هاي گونه‌هاي در معرض خطر بازباني و ارزشيابي شدند. سايت‌هاي بي مأخذ براي هر سؤال، بيش از ۴۸ درصد محاسبه شدند.

اين نتايج علت شك و ترديد در مورد قابليت اعتماد اطلاعات را تصديق مي‌کنند. ورودها و كتابخانه‌هاي كاملاً رقومي به راستي ممكن است -مانند پاگاه‌هاي اطلاعاتي علمي متني سنتي- بازبيني اطلاعات دقيقاً بررسي شده معتبر را پيشنهاد نمايند و مانند زمينه‌اي براي تحليل‌هاي وب‌سنجي عمل كنند.

### مطالعات عوامل تأثيرگذار وب

روسو (۱۹۹۷) در مقاله وب‌سنجي معروفش، درباره «پيوند به سايت‌ها»<sup>۴۱</sup>، يعني پيوندهاي دروني، الگوهاي توزيع وب سايت‌ها و پيوندهاي در حال آمدن<sup>۴۲</sup> را مورد تحليل قرار داد. روسو هم مانند اينگورسن (۱۹۹۸) از نسخه قديمي آلتاويستا استفاده کرده‌است، مطالعه‌اش را با ۳۴۳ سايت بارگذاري شده (نقاط داده‌اي<sup>۴۳</sup>) که از طريق یک کاوش درباره «اطلاع‌سنجي يا کتابسنجي يا علم‌سنجي» بازباني شده بودند، انجام داد. بنا بر اين، اين تحليل از ويژگي‌هاي موتور وب مستقل تر و قوي‌تر است. مطالعه نشان مي‌دهد که توزيع حوزه‌هاي سطح بالا براي اين سايت‌ها از توزيع لوتکا پيروي مي‌نمايد. به طور مشابه، روسو نشان داد که توزيع استنادي آن ۳۴۳ سايت نیز از توزيع لوتکا<sup>۴۴</sup> پيروي مي‌کنند و نسبت خود-پيوندها<sup>۴۵</sup> را ۳۰ درصد برآورد کرده است.

تفاوت ميان پيوندها و پيوند- صفحه‌ها به وسيله اينگورسن (۱۹۹۸) در اقدام او براي محاسبه عوامل تأثيرگذار وب براي حوزه‌هاي ملي<sup>۴۶</sup> و سايت‌هاي شخصي نشان داده شده است. پيش از اينگورسن، رودريگوز گابرين (۱۹۹۷) مفهوم «تأثير

اطلاعات روي اينترنت» را در مجله اسپانيايي دکوماناسيون، معرفي کرده است. عقیده و نظر اساسي اين بود که عوامل تأثيرگذار وب ممکن است درباره آگاهی يا شناخت از سايت‌هاي ملي (به طور ميانگين) يا سايت‌هاي شخصي اطلاع دهند. اين مطالعه به سه نتيجه جالب دست يافت:

۱. موتور کاوش آلتاويستا نمي‌تواند تعداد واقعي پيوندهاي دروني به سايت‌هاي خاص را محاسبه نمايد، بلکه فقط تعداد صفحاتي را که حداقل یک پيوند دروني (يا پيوند به سايت) در بر دارند، محاسبه مي‌کند. در مجموع، «خود-پيوندها»<sup>۴۷</sup> عوامل تأثيرگذار وب را تحت تأثير قرار نخواهند داد. بنا بر اين، پيوندهاي دروني خارجي<sup>۴۸</sup> براي مشاهده داراي اهميت خاصي هستند. ميانگين نمره و امتياز «خود- پيوند» در مطالعه اينگورسن تقريباً ۰/۵ و حوزه‌هايي با پسوند «com» ۰/۵۹ بوده است. به طور ميانگين عوامل تأثيرگذار «پيوند- صفحه» خارجي ۰/۳۹ بود.

۲. عوامل تأثيرگذار وب براي وب‌سايت‌هاي شخصي، غيرواقعي‌تر از آن حوزه‌ها بود.

۳. در اين محاسبات، واريانس ابزار ارزشيابي موتور وب بود.

همچنين مسئله اخير، واريانس، استفاده از روش‌هاي كاملاً پيچيده را براي محاسبه کردن عوامل تأثيرگذار و معرفي تنظيم پرسش مشروح پيشنهاد مي‌کند. همان طور که قبلاً روسو (۱۹۹۹) نشان داد موتور کاوش آلتاويستا در زمان تحليل عوامل تأثيرگذار وب در مقايسه با نسخه بعدي اکتبر ۱۹۹۹ واقعاً بي ثبات بود. بنا بر اين، محاسبات عوامل تأثيرگذار وب ممکن است به عنوان شاخص عملکرد موتور کاوش عمل نمايد. در مجموع، دليل استفاده کردن از موتور آلتاويستا، پوشش و توانايي‌هاي دستور بازباني آن، براي جستجوي «صفحات حوزه» در روش‌هاي کنترل شده و نیز پيوند- صفحه‌ها بود.

در رابطه با نتیجه دوم مطالعه اینگورسن درباره بی‌ثباتی و نوسان زیاد عوامل تأثیرگذار وب‌سایت‌های شخصی، اسمیت (۱۹۹۹) و همچنین ثلوال<sup>۴۹</sup> (۲۰۰۰) در مورد این پدیده، بیشتر پژوهش کردند. متأسفانه هنوز (۱۹۹۹) نسخه بی‌ثبات آلتاویستا به کار می‌رود. با این وجود، دقیقاً به علت نوسانات مشاهده شده، آنها هر دو در مورد خصوصیات و ویژگی‌های بازیابی و پوششی موتورهای مشکوک شده‌اند. با توجه به این که نتایج پایدار بوده‌اند و غیره، شاید خیلی ضروری نباشد که روش شناسی فوراً مورد سؤال قرار گیرد. اسمیت (۱۹۹۹) برخی روش‌های گردآوری داده‌های قوی و دوره‌ای را بیان می‌کند و همچنین نشان می‌دهد که چگونه نتایج به دلیل بازیابی صفحات نامربوط تحریف می‌شوند. برای مثال اندونزی (کد حوزه آن id) عوامل تأثیرگذار وب را به علت بازیابی عنصر جایزهای جهانی منع<sup>۵۰</sup> "id" در تعداد زیادی از سایت‌های غیراندونزیایی خیلی بالا نشان می‌دهد. وی همچنین نشان داد که رشته‌های جایزهای جهانی منع طولانی‌تر به منظور نتیجه قابل اعتمادتر جستجو شده‌اند. بافت این رشته‌ها باید منحصر به فرد بودن آنها را تأیید کند. با وجود این، مطالعات منتشر نشده بعدی درباره پوشش واقعی موتورهای -از جمله آلتاویستا- با توجه به صفحات و پیوندهای شناخته شده دانشکده کتابداری کوپنهاگ دانمارک روی خدمت دهنده محلی (<http://ix.db.dk>) نشان می‌دهد که آنها به تمام صفحات و پیوندها نفوذ نمی‌کنند. این نتیجه منفی را ثلوال (۲۰۰۰)-که موتورهای کاوش‌ها، بات، آلتاویستا و اینفوسیک را در تحلیل‌هایش به کار برده است- تأیید می‌کند. در چنین شیوه‌ای که مخرج و صورت کسر عوامل تأثیرگذار وب به همان روش‌ها تحت تأثیر قرار داده شده است، این پوشش تصادفی نیست. به طور خلاصه در وضع حاضر، روش‌های بازیابی و پوشش

موتور کاوش «مفهوم اخیر عوامل تأثیرگذار وب عملاً در ابزار نسبتاً ساده و ابتدایی ظاهر شده‌اند» (ثلوال، ۲۰۰۰، ص ۱۸۸). بدین ترتیب پیامدها و نتایج بسیار مشکل آفرین هستند و همان طور که روسو (۱۹۹۹)، اسمیت (۱۹۹۹) و ثلوال (۲۰۰۰) بیان کرده‌اند، یک نفر باید رویات‌های اختصاصی وب را برای بارگذاری کردن نمونه‌ها به منظور تحلیل‌های محلی به کار برد.

### کشف دانش و ردیابی موضوع روی وب

طی دهه اخیر یک حوزه پژوهشی چند رشته‌ای به نام «کشف دانش در پایگاه‌های اطلاعاتی»<sup>۵۱</sup> ظاهر شده است. این حوزه به گسترش روش‌ها برای بهره‌برداری کردن از مخزن در حال رشد به صورت تصاعدي محتواهای ثبت شده در پایگاه‌های اطلاعاتی تجارتي، اداری، علمی و سایر انواع داده‌ها مربوط می‌شود. فراولي و دیگران (۱۹۹۱) کشف دانش در پایگاه‌های اطلاعاتی را به عنوان «استخراج قابل ملاحظه اطلاعات ضمنی (مطلق<sup>۵۲</sup>) که قبلاً ناشناخته و بالقوه مفید از داده‌ها» تعریف نموده‌اند. به منظور شناسایی و استخراج الگوها و روابط جدید که بتواند دانش جدیدی را به ثمر رساند، کشف دانش در پایگاه‌های اطلاعاتی روش‌های متنوع و گسترده‌ای را به کار می‌برد- مانند ترکیب قدرت رایانه و تخصص و مهارت انسان. فنون کشف دانش در پایگاه‌های اطلاعاتی برای مثال، شامل بازیابی اطلاعات، آمار، آموزش ماشینی، شناسایی الگو، مقیاس بندی چندبعدی و تجسم است.

اهداف و روش‌های کشف دانش در پایگاه‌های اطلاعاتی و کتابسنجی، در بسیاری نکات مشترک هستند. برای مثال، روش‌های دسته‌بندی (خوشه‌بندی) کردن کتابسنجی به عنوان یکی از کاربردهای کشف دانش در پایگاه‌های اطلاعاتی استنادی و کتابشناختی توانسته مورد

توجه قرار گیرد. همچنین راهبردهايي در پژوهش کشف دانش در پایگاه‌هاي اطلاعاتي وجود دارد که جزو کتابسنجی است. برای مثال، در حوزه داده کاوي<sup>۳</sup> (استخراج داده‌هاي) متني (لويزويك، اوارد و کوستوف، ۲۰۰۰).

مفهوم «داده‌کاوي» گاهي اوقات در ارتباط با کشف دانش در پایگاه‌هاي اطلاعاتي به طور مترادف به کار برده شده است. کشف دانش در پایگاه‌هاي اطلاعاتي دلالت دارد بر فرایند کلي کشف دانش مفید از داده‌ها، در حالی که داده کاوي، اقدام مهمي است که بر روي شناسايي الگوها متمرکز است (فیاد و دیگران، ۱۹۹۶). حوزه‌هايي که از کشف دانش در پایگاه‌هاي اطلاعاتي استفاده مي کنند، شامل: رفتار مشتري، بررسی‌هاي مربوط به ستاره‌ها، تشخیص سرطان، شناسايي ساختار شيميايي، تحلیل جمعیت، کنترل کیفیت و مدل سازي از تغییر اقليمي جهان است (ویکري، ۱۹۹۷).

وب محیطي است آشکار برای به کارگیری کشف دانش که اتریوني<sup>۴</sup> در سال ۱۹۹۶ مورد بحث قرار داده بود. وب مي‌تواند، به عنوان یک پایگاه اطلاعاتي توزیعي در حال رشد به طور تصاعدي حاوي بیش از يك بيليون صفحات وب در بخش قابل نمایه‌سازي خود (شامل پایگاه‌هاي اطلاعاتي تولید شده درخواستي در قالب هاي سند نیست) و تقریباً ۱۰ بيليون پیوند باشد. علاوه بر این، شامل داده‌هاي توزیعي روي ميليون ها خدمت دهنده است. وب داراي ابعاد دیگری نیز مي باشد که آن را از سایر پایگاه‌هاي اطلاعاتي معمولي متمایز مي‌سازد، مهمترین آنها این است که وب به صورت چند عاملي (چند واسطه‌اي<sup>۵</sup>) بنا گردیده است. ميليون ها حامي و کارگزار مختلف وب، از قبیل مردم عادي، محققان، سازمان ها و غیره، فعالانه صفحات و پیوندهاي وب را خلق مي‌کنند، تغییر مي دهند و حذف مي‌کنند.

ماهیت توزیعي، پویا و متنوع وب- به همراه کمترین استفاده از آبرداده<sup>۶</sup> ترکیب شده است- آن را محیطي مشکل و دشوار برای کشف دانش یا «وب‌کاوي»<sup>۷</sup> مي‌سازد و همان‌طور که در بالا اشاره شد، آلن و همکارانش در مطالعاتشان (۱۹۹۹) خاطر نشان کردند که اطلاعات وب ممکن است نادرست و غیرواقعي باشد. از سوي دیگر، ناهمگني وب مي‌تواند منبع مساعدي برای کشفیات ایجاد کند. همان‌طور که به وسیله دوجانگ و ریپ<sup>۸</sup> (۱۹۹۷) بیان شد کشفیات اغلب از «ساختن ترکیب هاي پیش‌بینی نشده منابع ناهمگن» ناشي مي‌شوند، که به طور ضمنی دلالت دارند بر این که ممکن نیست از قبل بتوان گفت که چه منابعي مورد نیاز است.

سه راهنمای اصلي برای اجرائی «کشف دانش» بر روي وب وجود دارد. آنها مربوط مي‌شوند به بهره برداري از ۱. محتویات صفحه وبی ۲. ساختار پیوندها و ۳. رفتار اطلاع یابی کاربران (مانند جستجو و تورق). در این بخش تمرکز روي بهره‌برداري از ساختار پیوندها برای کشف دانش در وب است. این راهبرد تشابه زیادی با تحلیل‌هاي استنادي کتابسنجی دارد، اما نه صرفاً به وسیله پیوندهاي قوي<sup>۹</sup>.

پیوندها، مدارك وب را در یک پیکره و بستر فرامتن ساختارمند پیچیده، به هم مي‌بافند. ساختار پیوندها، تفسیرهاي تردیدناپذیر و مطلق انسان را نشان مي دهند که مي‌تواند برای کشف دانش به کار گرفته شود، برای مثال شناختن اجتماعات وب<sup>۱۰</sup> (گیسون و کلاین‌برگ و روگوان، ۱۹۹۸)، شناسايي کردن صفحات وب معتبر و موثق (کلاین‌برگ، ۱۹۹۸)، تلخیص‌کردن موضوع<sup>۱۱</sup> (بارات و هینسینگر، ۱۹۹۸)، یا بهبود بخشیدن الگوریتم‌هاي رتبه‌بندی موتور کاوش (برین و پاژه، ۱۹۹۸).

تیم برنرز-لی<sup>۱۲</sup> مخترع وب در سرن<sup>۱۳</sup> (سازمان پژوهش‌هاي هسته‌اي اروپا) در سال



۱۹۸۹-۱۹۹۰ این رشد و توسعه را پیش‌بینی کرده بود. یکی از مهمترین انگیزه‌های او برای توسعه وب، امکان آگاهی از «تارهای پیچیده روابط میان مردم، برنامه‌ها، ماشین‌ها و عقاید بوده است» (برنرس و لی، ۱۹۹۷). روش‌های نظریه نموداری ابزار بسیار مناسبی برای مطالعه ساختار پیوندها روی وب هستند. شکل‌شناسی ساختار پیوندها امکانات عوامل انسانی و رقومی به منظور کاوش و عبور کردن از وب و کشف در آن را تحت تأثیر قرار می‌دهد.

### نظریه نمودار و وب

در نظریه نمودار (گراف)، نمودار عبارت است از بازنمود ریاضی از یک شبکه که شامل رأس‌ها<sup>۶۴</sup> یا گره‌هایی است که یال‌ها (ضلع‌ها<sup>۶۵</sup>) به هم متصل شده‌اند. گره‌ها می‌توانند انسان (در شبکه‌های اجتماعی)، عوامل بوم‌شناختی (در یک شبکه غذایی)، خدمت دهنده‌های اینترنتی، مدارک (در یک شبکه استنادی)، مفاهیم (در یک شبکه معنایی یا یک اصطلاحنامه) و غیره باشند. در یک نمودار جهت دار، یال‌ها روابط جهتی میان گره‌ها را نشان می‌دهند. وب نمونه‌ای است از یک نمودار جهت دار که صفحات وب منطبق با گره‌ها، و فرآیندها<sup>۶۶</sup> منطبق با یال‌ها هستند. روش‌های نظریه نموداری می‌تواند برای تحلیل جنبه‌های ساختاری وب مورد استفاده قرار گیرد. برودر و دیگران (۲۰۰۰) یک مدل کمایی (پایون‌وار<sup>۶۷</sup>) از ساختار نموداری وب ارائه کردند. با استفاده از یک مجموعه وب<sup>۶۸</sup> شامل تقریباً ۲۰۰ میلیون صفحه وب و ۱/۵ بلیون پیوند، می‌توان یک نمونه پایگاه اطلاعاتی از یک نمودار وب تشکیل داد. یک تیم پژوهشی نشان دادند که وب شامل پنج حوزه مجزای مشخص شده به وسیله گره‌هایی است که یا ۱. فقط پیوندهای بیرونی دارند ۲. فقط پیوندهای درونی دارند ۳. هر دو نوع

پیوند، هم بیرونی و هم درونی، دارند ۴. یا هیچ پیوندي ندارند و یا ۵. گره‌هایی که به گره پایون<sup>۶۹</sup> متصل شده‌اند (شامل گره‌هایی با پیوندهای درونی و بیرونی) یا متصل نشده‌اند. یک مدل پایون، مفهومی قابل توجه از ساختار تو در توی وب ارائه می‌نماید. این پیچیدگی، امکانات کشف دانش را وقتی که عوامل انسانی یا ماشینی (رقومی) ساختار پیوند را بازدید و تحلیل می‌نمایند، تحت تأثیر قرار می‌دهد. در چارچوب نظریه نموداری، یک بُعد قابل توجه ساختار پیوندها با پدیده معروف به «جهان کوچک» و «شبکه‌های جهان کوچک» سروکار دارد. این پدیده، امکانات کشف دانش را افزایش می‌دهد.

### شبکه‌های جهان کوچک

در شبکه‌های جهان کوچک، گره‌ها به صورت خوشه‌ای هستند - مانند آنچه در نمودارهای منظم است- اما طول مسیر میان هر جفت گره کوتاه است، مانند آنچه در نمودارهای تصادفی<sup>۷۰</sup> است. در یک شبکه جهان کوچک داشتن درصد بسیار کمی از پیوندها که به عنوان «میان‌برها» عمل می‌کنند و بخش‌های دوردست شبکه را به هم متصل می‌کنند، کافی است.

واتس و استروگاتز<sup>۷۱</sup> (۱۹۹۸) و واتس (۱۹۹۹) نشان دادند که شکل شناسی جهان کوچک در قالب طول مسیرهای کوتاه در شبکه‌های زیست‌شناسی، فناوری و اجتماعی رخ می‌دهد. برای مثال، شبکه عصبی کرم خاکی<sup>۷۲</sup>، شبکه برق فشار قوی ایالات غربی آمریکا و نمودار همکاری بازیگران فیلم. نظریه جهان کوچک از اثری از میلگرام (۱۹۶۷) و کوچن<sup>۷۳</sup> (۱۹۸۹) سرچشمه گرفته است که آنها این نظریه را به واسطه پنداره ۶ جایگاه (نقطه) جداگانه که به «فواصل و مسیرهای کوتاه میان دو شخص دلخواه از طریق حلقه‌های رابط آشنایی» می‌پردازند، عمومیت یافته است.

سایر محققان، سازمان‌ها، طرح‌ها و مقالات در حوزه علمي خودشان می‌سازند.

عوامل رقومي(ماشيني) یا انساني، از طريق پیوندها از يك صفحه وب به صفحه وب دیگر که امکان حرکت از یک خوشه وب به سایر خوشه‌های دوردست را با استفاده از تنها یک پیوند متقاطع به عنوان «میان‌بر» فراهم می‌سازند، به جستجو در وب می‌پردازند. در مثال بالا، يك محقق در حوزه علم اطلاع رسانی توانسته است که پیوندي متقاطع روي یک صفحه وب با پیوندهای مورد علاقه‌اش به سایر حوزه‌های علمي دلخواه خود مثلاً «تحريك خلاقیت»<sup>۷۷</sup> ایجاد کند. بنابراین، ردّپایي (خط سیري)<sup>۷۸</sup> میان دو خوشه ناهمگن وب به وجود می‌آید. چنین خط سیرهای مختلف و دور از هم که به صورت متقاطع در مسیرهای پر رفت و آمد و شلوغ روي وب جریان می‌یابند، می‌توانند امکانات کشف دانش به کمک رایانه و یا بازیابی اتفاقي توسط انسان را تحت تأثیر قرار دهند، در حالی که اطلاعات پیش‌بینی نشده اما احتمالاً مفید، شناسایی و استخراج شده است.

امکان کشف دانش خلاق و سازنده با استفاده از ساختار پیوندها نقش مؤثري را در توسعه نظام‌های اطلاعاتي فرامتن از جمله وب، ایفا نموده است. در این روند تاریخي یک عامل شتاب دهنده مهم، دید سطحی و نوار بوش<sup>۷۹</sup> (۱۹۴۵) از ممکس- یک نظام اطلاعاتي- بوده است که پاراگراف‌های متني متفاوت و مجزا (روي میکروفيلم) می‌توانند به وسیله خط سیرها (ردّپاها) به طور متقاطع به صورت سلسله مراتب رده بندي شوند و به منظور تحريك خلاقیت‌های محققان به هم متصل گردند. [ایده فرامتن از ممکس الهام گرفته است].

جنبه پیوند تقاطعي<sup>۸۰</sup> ممکس و پیوندهای متقاطع می‌تواند به قیاس برای نشان دادن چگونگی انتشار ایده‌ها میان گروه‌های اجتماعي

هنوز فقدان پژوهش در علوم کتابداری و اطلاع رسانی درباره پدیده جهان کوچک در خصوص فواصل کوتاه و نتایج آنها در شبکه‌های اطلاعاتي از قبیل وب جهانگستر، پایگاه‌های اطلاعاتي استنادي، شبکه‌های معنانشناسي و اصطلاحنامه‌های همایند و غیره، وجود دارد. پدیده جهان کوچک امکان پذیر است، وقتی که گره‌ها در یک شبکه اطلاعاتي منطبق با مدارك، اصطلاحات، مؤلفان، مؤلفان استناد شده، مجلات، حوزه‌های علمي<sup>۷۴</sup>، سازمان‌ها، یا کشورها و غیره تعریف شوند، و پیوندهای کاهش فاصله و پیوندهای ایجاد جهان کوچک منطبق با مآخذ، اصطلاحات مرتبط، ظهور هم واژه‌ها (واژه مشترک<sup>۷۵</sup>)، توصیفگرها، مؤلفان همکار، مؤلفان با هم استناد شده، یا مجلات و غیره تعریف شوند.

یکی از پیامدها و نتایج جهان کوچک می‌تواند افزایش و بهبود امکان کشف پیش‌بینی نشده اطلاعات مفید روي اینترنت باشد. در یک طرح جدید یکی از مؤلفان، در حال توسعه این راهبرد است که نظریه نمودار و کتابسنجی را درهم ادغام کند. مفهوم اصلي در این چارچوب به پیوندهای متقاطع معروف است (بجورن‌برن، ۲۰۰۰).

#### پیوندهای متقاطع (عرضي)

پیوندهای متقاطع به عنوان میانبرهای میان خوشه‌های ناهمگن وب<sup>۷۶</sup> عمل می‌کنند. خوشه‌های وب از وب‌سایت‌ها و صفحات وبی بسیار به هم پیوسته تشکیل شده است که گروه‌ها و جوامع مشترک المنافع و حوزه‌های موضوعي هم خانواده و مرتبط را منعکس می‌کنند. صفحات وب درون يك خوشه وب، بسته به ساختار پیوند درون خوشه‌ای می‌توانند کمتر یا بیشتر در مرکز یا در حاشیه مستقر شوند. مثالی برای یک خوشه وب می‌تواند صفحه وب محققانی باشد که پیوندهایی را از صفحات خانگی شخصی شان به

نامتجانس به طور جنبي از طريق افراد وابسته ضعيف كه اصطلاحاً گره‌هاي ضعيف ناميده مي‌شوند، در نظر گرفته شود (گرانووتر، ۱۹۷۳). همچنين پيوندهاي متقاطع ممكن است مورد توجه قرار گيرند براي اين كه چرا نويسندگان آثار علمي اغلب به تعداد بسيار كمی از منابع خارج از حوزه علمي خودشان كه اصطلاحاً تقاطع هاي مرزي<sup>۸۱</sup> ناميده مي‌شود، استناد مي‌کنند (كلاین، ۱۹۹۶ و پیرس، ۱۹۹۹).

پيوندهاي متقاطع، به عنوان گره‌هاي ضعيف و تقاطع هاي مرزي میان خوشه‌هاي ناهمگن وب عمل مي‌کنند. البته، بسياري از پيوندهاي متقاطع داراي خصوصيات منحصر به فردي هستند، مانند سرگرمي‌هاي غيرعلمي شخصي. اما، ساير پيوندهاي متقاطع روي صفحات وب دانشمندان توانسته‌اند پيش پژوهش‌هاي<sup>۸۲</sup> آشكار شده در حوزه‌هاي علمي يا بين رشته‌اي<sup>۸۳</sup> دانشكده‌هاي نامرئي<sup>۸۴</sup> را منعكس سازند. آشكار كردن چنين اتصالات مخفي به صورت شانسي(اتفاقي) توسط انسان يا كشف دانش با كمك رايانه توانسته است اطلاعات مفيدي را درباره ابعاد جديد در تكامل و شكل گيري ارتباط میان علوم و كشف كردن الگوها و روابط ارائه دهد. پيوندهاي متقاطع مرزهاي بين علمي<sup>۸۵</sup> مي‌توانند بينش خلاق و عميقي را به وجود آوردند، بنابراین، مفهوم و درك جديدي براي فهم شدت و ميزان گره‌هاي ضعيف از طريق تحليل شبكه اجتماعي ارائه مي‌دهند (گرانووتر، ۱۹۷۳).

### ملاحظات روش شناسي

تلاش براي محلي كردن پيوندهاي متقاطع روي وب، با كنترل داده‌هاي با بسامد پايين- شبیه توزیع احتمالي برادفورد از صفحات وب مقصد براي پيوندهاي بيروني ايجاد شده در يك حوزه علمي يا مجموعه مورد علاقه - سرو كار دارد. اين هنوز يك

فرضيه است كه در حال آزمون است، اما احتمالاً اغلب پيوندهاي بيروني براي صفحات وب معتبر و محبوب در حوزه‌هاي موضوعي مشابه مورد توجه قرار مي‌گيرند (كويي، ۱۹۹۹).

همان طور كه پيشتر بيان شد، مسئله روش شناسي در وب‌سنجی با گردآوری داده‌هاي بي غرض از وب سروكار دارد كه آن به عنوان اساس و پايه‌اي براي پژوهش‌هاي تجربي مورد توجه قرار مي‌گيرد. همان گونه كه در بالا بيان شد، در داده‌هاي ثانوي گردآوری شده در موتورهاي كاوش تجاري بزرگ به دليل بي دقتي زياد در مورد پوشش، فاصله روزآمدسازي، اصول نمايه‌سازي، اجرائي محاسبات، الگوريتم هاي رتبه بندي و غيره، اعتبار وجود ندارد. همچنين براي استفاده از داده‌هاي اوليه بارگذاري شده به صورت مستقيم از وب، روش ديگري وجود دارد. روشي كه در مقوله بندي اخير گردش تصادفي ناميده شده كه در نظريه نمودار مورد استفاده قرار گرفته است. به منظور محلي كردن پيوندهاي متقاطع روي نمودار وب، تعداد زيادي از گردش‌هاي تصادفي<sup>۸۶</sup> طولاني، مي‌توانند با پيوندهاي متعاقب (پيرو) در يك روش تصادفي از يك صفحه وب به صفحه وب ديگر مورد استفاده قرار گيرند.

اجرائي تحليل مسير<sup>۸۷</sup> از مسير پيوندهاي طولاني كه بدین نحو ايجاد شده‌اند گاهي اوقات به بن بست ختم مي‌شوند. پيوندهاي متقاطع با استفاده از ملاك‌ها و معيارهاي ناهمگني میان حوزه‌هاي موضوعي منعكس شده در صفحات وب شناسايي مي‌شوند. مفهوم ناهمگني (همانند مفهوم معكوس مرتبط با "تشابه" در نظريه رده‌بندي) پيچيده نيست، و معياري است براي تعريف عملياتي با استفاده از مقياس هاي تشابه واژه‌هاي هم رويداد (هم وقوع<sup>۸۸</sup>).

نقاط شروع تصادفي براي گردش‌هاي تصادفي مسئله ديگري است كه از ماهيت توزيعی

كشوف دانش و توليد فرضيه‌ها مورد استفاده قرار گيرد».

اين شيوه همچنين ممكن است براي محلي كردن پيوندهاي متقاطع روي وب استفاده شود. در يك نمونه نمودار وب توليد شده به وسيله رايانه كه از روش‌هاي الگوريتمي استفاده مي‌نمايد، کوتاه ترين مسير پيوند ميان دو صفحه وب آغازين و پاياني انتخاب شده كه به حوزه‌هاي علمي ناهمگن تعلق دارند، مي‌تواند شناسايي شود. بسته به اين كه همسايگان يك خوشه از صفحات وب انتخابي تا چه حد به هم مرتبط باشند، اين روش طبق مدل «كماني (پايون وار)» پيش گفته به علت بزرگي و گستردگي مناطق صفحات وب كه فقط داراي پيوند دروني يا فقط پيوند بيروني هستند و يا در كل هيچ پيوندي ندارند، شايد تا ۷۵ درصد موارد با شكست مواجه خواهد شد (برودر و ديگران، ۲۰۰۰). اما حداقل تقريباً در ۲۵ درصد از موارد، تعيين کوتاه‌ترين مسير ميان دو حوزه علمي ناهمگن منتخب امكان پذير خواهد بود. بدين ترتيب پيوندهاي متقاطع ميان بر را در طول مسير نشان مي‌دهند.

عقيده اسمال (۱۹۹۹) مبني بر ايجاد مسيرهاي هم استنادي (اشتراك در متن) قوي و شديد چندمرحله‌اي، همچنين مي‌تواند براي تحليل هم پيوند پيوندهاي بيروني<sup>۹۴</sup> هم رويداد، مثلاً روي صفحات وب دانشمندان، به كار برده شود. اين راهبرد مي‌تواند راهنماي مفيد و مؤثري براي اجراي كشف دانش روي وب باشد. روش ديگر كشف دانش استفاده شده در پايدگاه‌هاي اطلاعاتي كتابشناختي، قابل اجرا براي وب، پژوهش سوانسون (۱۹۸۶) روي دانش عمومي كشف نشده است كه در طول سال‌ها توسعه يافته است (وانسون و اسمال‌هايزر، ۱۹۹۹، ۱۹۹۷). سوانسون (۱۹۸۶) بيان نمود كه «دانش مي‌تواند عمومي باشد، ولي هنوز كشف نشده، و اگر اجزاء و قسمت‌هاي آن به طور مستقل خلق شده باشند،

وب كه قبلاً ذكر شد، ناشي مي‌شود. استفاده از شماره‌هاي معروف به شماره IP<sup>۹۵</sup> (تفاهم نامه اينترنت) يا فهرست اسامي خدمت دهنده هاي نام حوزه، مثل حوزه "edu" روش‌هاي احتمالي براي نمونه‌گيري بدون سوگيري و غيرمغرضانه از نقاط شروع هستند. در تفاهم نامه اينترنت به هر ميزان وب يك شماره منحصر به فرد تخصيص يافته است. به عنوان مثال «۱۳۰،۲۳۶،۱۸۶،۶» شماره تفاهم نامه مدرسه سلطنتي علوم كتابداري و اطلاع رساني دانمارك "www.db.dk" است.

### تحليل مسير و دانش عمومي كشف نشده (ناشناخته<sup>۹۰</sup>)

با استفاده از تحليل مسير در پايدگاه‌هاي اطلاعاتي استنادي، اسمال (۱۹۹۹) راه‌ها و مسيرهاي<sup>۹۱</sup> مرزهاي بين رشته‌اي در علوم و خلاقيت بارور شده را كه مي‌توانستند در چنين گذرگاه‌ها و تقاطع هاي مرزي ظاهر شوند مورد پژوهش قرار داد. اسمال به بررسي "گره‌هاي قوي" در مقايسه با "گره‌هاي ضعيف" در حالت هم استنادي (اشتراك در متن<sup>۹۲</sup>) قوي و شديد براي ايجاد مسيرهاي چندمرحله‌اي غيرمستقيم در آثار علمي پرداخته بود. مطابق نظر اسمال، حركت كردن از يك موضوع يا حوزه، به موضوع يا حوزه ديگر در آثار علمي به دليل بافت و ساختار مرتبط و درهم تنيده رشته‌هاي علمي امكان پذير است. اين مسئله با استفاده از مسير خاصي كه از علم اقتصاد شروع و به اختريفيزيك ختم مي‌شود، نشان داده شده است.

كين و نورتون<sup>۹۳</sup> (۱۹۹۹) درباره اسمال اظهار نظر كردند و پيش‌بيني نمودند كه «در نظام هاي آتي بازباني اطلاعات، يك كاربر مي‌تواند دو موضوع يا دو سند را انتخاب نمايد و مسير (خط سير) اسناد يا موضوع هايي را كه به آن دو موضوع مرتبط مي‌شوند پيدا كند كه آن مي‌تواند براي

مفهوم «مدیریت موضوعات»<sup>۱۰۱</sup> که توسط لنکستر (۱۹۸۵) و مفهوم «ردیابی موضوع» که به وسیله ورمل (۲۰۰۰) خلق شدند.

### ردیابی موضوع

لنکستر (۱۹۸۵) در مطالعه موردی خود در مورد این که چگونه موضوع جدید در حال شکل گیری «باران اسیدی»<sup>۱۰۲</sup> در جامعه، توسعه و اشاعه یافته است، مسئله را در چندین پایگاه اطلاعاتی مختلف ردگیری کرده بود که نشان دهد چگونه این مسئله تحقیقی در طول زمان به حوزه علوم کاربردی و بعداً به رسانه‌های گروهی و قانونگذاری کشیده شده است. لنکستر و ورمل (۲۰۰۰) روش‌های اطلاع‌سنجی را برای ردیابی الگوی مباحثه بین المللی درباره «وضعیت رفاهی» جدید در پایگاه‌های اطلاعاتی حوزه‌های مختلف به کار بردند؛ در نتیجه آنان نشان دادند که چگونه یک مفهوم از میان یک مسیر به روش‌های مختلف نشر، حرکت می‌کند. ورمل نتیجه گرفت که فناوری‌های «داده‌کاوی» و «متن‌کاوی»، امکانات زیادی را برای تحلیل اطلاع‌سنجی به منظور استخراج دانش ناشناخته و بالقوه مفید از داده‌های کتابشناختی ارائه می‌کنند. گونه‌ای از ردیابی موضوع روی وب توسط بار-ایلان و پریترز<sup>۱۰۴</sup> (۲۰۰۰) به کار گرفته شده بود. ایشان موضوع انتخاب شده «اطلاع‌سنجی» را برای یک دوره زمانی معین با استفاده از روش‌های کتابسنجی به منظور تحلیل داده‌ها بین شش موتور کاوش اصلی، مورد پژوهش قرار دادند. ماهیت پویای وب، از اسناد وب درباره موضوعاتی که بیشتر اوقات ناپدید هستند، ناشی شده است، در حالی که اسناد جدید اضافه می‌شوند و برخی از آنها تغییر می‌یابند. به دنبال همان روش‌ها بار-ایلان (۲۰۰۰) موضوع مشابهی را به طور مفصل ردیابی کرد و الگوهای

منطقاً به هم مرتبط هستند، اما آنها هرگز با هم بازیابی و تفسیر نمی‌شوند و باهم آورده نمی‌شوند». سوانسون از یک روش آزمون و خطای نظام مند برای یافتن روابط متعددی و ناپایدار<sup>۹۵</sup> میان دو اثر استفاده کرد (دیویس، ۱۹۸۹). با توجه به مثال خود سوانسون (۱۹۸۶)، اگر اثر «الف» مربوط به «روغن ماهی» باشد و اثر «ج» تقریباً در مورد «بیماری ریناد»<sup>۹۶</sup> باشد، آنگاه اثر «ب» روی «پلاکت خون»<sup>۹۷</sup> می‌تواند یک رابطه متعددی کم داشته باشد. اگر «الف» مساوی (اشتراک) «ب» و «ب» مساوی «ج» باشد، در نتیجه، «الف» مساوی «ج» خواهد بود. این روش کشف دانش مبتنی بر اثر<sup>۹۸</sup> برای یافتن «اطلاعات ضمنی (مطلق) جالب اما قبلاً ناشناخته» در میان آثار علمی به کار رفته است (سوانسون و اسمال‌هایزر، ۱۹۹۹) و همچنین برای آشکار کردن روابط میان عقاید و مفاهیمی که قبلاً مورد توجه نبودند (گارفیلد، ۱۹۹۴). با استفاده از این روش روی وب، پیوندهای متقاطع توصیه‌های مفیدی را برای یافتن روابط متعددی و ناپایدار میان رشته‌های علمی ارائه می‌دهند.

مفهوم «کشف شانسی نظام مند»<sup>۹۹</sup> در زمینه کشف علم و کشف دانش با کمک رایانه مفید است. گارفیلد از ۱۹۶۶ چندین مرتبه از این مفهوم برای تشریح روند سازمان یافته کشف روابط علمی قبلاً ناشناخته که در پایگاه‌های اطلاعاتی استنادی مورد استفاده قرار می‌گرفت، استفاده کرد. در حقیقت، «کشف شانسی نظام‌مند» توصیف نسبتاً دقیقی از ضرورت همکاری میان انسان-رایانه در اجرای کشف دانش هم در پایگاه‌های اطلاعاتی سنتی و هم روی وب است. اهمیت همکاری انسان-رایانه برای کشف دانش در حوزه علم توسط ولدس-پرز<sup>۱۰۰</sup> (۱۹۹۹) مورد تأکید قرار گرفته است. آخرین راهبردهایی که در این بخش به استفاده روش‌های کتابسنجی در کشف دانش روی وب مربوط می‌شوند، عبارتند از

روی وب برای پایگاه‌های اطلاعاتی استنادی و کتابشناختی را مقایسه کرد.

### نتیجه‌گیری

این مقاله سعی دارد که به حوزه‌های خاص و منتخبی از پژوهش وب‌سنجی که روند و فضای جالبی برای توسعه دارند، اشاره کند و همچنین برخی حوزه‌های کمتر خوش آتیه را بیان می‌کند. همان طوری که قبلاً ذکر شد، امکان استفاده از روش‌های کتابسنجی روی وب از ماهیت پویا، متنوع و توزیعی وب و نقائص موتورهای وب تأثیر پذیرفته است. تنوع و گوناگونی افرادی که اسناد و مدارک وب و پیوندها را ایجاد می‌کنند البته از کیفیت و اعتبار این عناصر وب، تحت تأثیر قرار می‌گیرند. کمبود آبرده‌های به کارگرفته شده برای مدارک و پیوندهای وب و کمبود موتورهای کاوشی که آبرده‌ها را پشتیبانی کنند بر گزینه‌های پالایش تأثیر می‌گذارد، و در نتیجه گزینه‌های کشف دانش تحت تأثیر قرار می‌گیرند؛ در صورتی که کدهای ناحیه‌ای<sup>۱۰۵</sup> در پایگاه‌های اطلاعاتی سنتی، کشف دانش در پایگاه‌های اطلاعاتی را پشتیبانی می‌کنند. همان طور که در بالا پیشنهاد شد، تنوع در وب می‌تواند امکانات کشف دانش را افزایش دهد. دولت ستیزی و هرج و مرج طلبی، رفتار محلی میلیون‌ها نفر از عاملان وب معمولاً بر روی عملکرد جهانی وب به عنوان یک نظام اطلاعاتی، نتایج و عواقب منفی به همراه داشته است. پیوندهای متقاطع مذکور، توانسته اند از تأثیر مثبت تاکنون فراموش شده این رفتار نامناسب در مسیرهای کوتاه وب نتیجه بگیرند و به این ترتیب موجب امکانات کاوش بهتر برای کارگزاران و واسطه‌های رقومی و انسانی می‌شوند و بدین وسیله موجب فراهم کردن امکانات بهتر برای کشف دانش با کمک رایانه و به صورت شانسی توسط انسان، برای مثال «کشف شانسی نظام‌مند» می‌گردند.

استفاده از این راهبرد برای کشف مواد مفید به منظور اهداف و مقاصد علمی، البته نیازمند کنترل مسئله اعتبار و صحت در وب می‌باشد. یک راه برای انجام این، می‌تواند روش پیش گفته یعنی «انتخاب کیفیت صفحات وب آغازین و پایانی از دو حوزه علمی

ناهمگن» باشد و سپس پیوندهای متقاطع روی یک مسیر پیوند متصل به حوزه‌ها شناسایی شود. این روش می‌تواند احتمال برخورد با محتویات کیفی در صفحات وب میانی در طول مسیر پیوند را افزایش دهد.

وب شامل ساختار پیوندهای واگرا (دور از هم) و همگرا (شبه به هم) است که خوشه‌های وب، با نوع پیشین و پیوندهای متقاطع با نوع بعدی منطبق هستند. این ساختارهای مختلف پیوند می‌تواند کاوش در وب هدایت شده در مسیرهای همگرا (یعنی منطقی، هدف مند) و واگرا (یعنی خلاق، شهودی) را پشتیبانی کند. استفاده از واژه‌های واگرا و همگرا از اثر فورد (۱۹۹۹) و از اثر باودن (۱۹۸۶) در مورد تحریک خلاقیت در نظام‌های اطلاعاتی، الهام گرفته است. درک و استنباط کامل واگرایی و همگرایی -هم در ساختار پیوند کشف شده موجود و هم در رفتار کارگزاران و واسطه‌های رقومی (ماشینی) یا انسانی- به منظور توسعه روش‌های خلاق و سازنده کشف دانش با کمک رایانه روی وب، و نیز در پایگاه‌های اطلاعاتی استنادی، کتابشناختی و سایر پایگاه‌های اطلاعاتی مهم است.

چنین روش‌هایی همچنین می‌توانند استلزام‌هایی برای اصلاح برنامه‌های دروگر<sup>۱۰۶</sup> رویات‌های وب، الگوریتم‌های رتبه بندی موتورهای کاوش و ویژگی‌های بصری/جهت‌یابی<sup>۱۰۷</sup> مرورگرها، داشته باشند. وب‌سنجی حوزه پژوهشی جدیدی است که در حال گذر از یک مرحله آزمایشی و مقدماتی ضروری است. تازگی این حوزه، تعداد قابل توجه مقالات توصیفی وب‌سنجی فراهم شده درباره ابعاد مختلف وب در طی ۴-۵ سال اخیر را توجیه می‌کند. در سال‌های آینده، تجزیه و تحلیل و ترکیب یافته‌های پژوهش‌ها و همچنین پیشبرد و توسعه روش‌ها و نظریه‌ها به منظور فراهم نمودن درک و فهم بهتر درباره ریخت‌شناسی (توپولوژی)، کارکردها و پتانسیل‌های امکانات پیچیده وب، چالشی برای پژوهشگران وب‌سنجی خواهد بود.

### سپاسگزاری

از پروفیسور پیترا اینگورسن و دکتر لئارت بچورن برون که طی نامه مورخه ۲۰ مه ۲۰۰۳ اجازه ترجمه مقاله حاضر را دادند و ما را راهنمایی فرمودند، بسیار سپاسگزار می‌شود.

## بي نوشت‌ها

- 1.Link structure
- 2.Almind and Ingwersen
- 3.Cybermetrics
- 4.Outlinks
- 5.Inlinks
- 6.Reference
- 7.Citation analysis
- 8.Domain
- 9.Web search behavior
- 10.Common data
- 11.Text mining
- 12.ISI: Institute of Scientific Information
- 13.Command language
- 14.Dialog
- 15.Online
- 16.Sitation analysis
- 17.Link-page
- 18.Indexable web
- 19.Servers
- 20.Imposed truncation
- 21.Clarke and Willett
- 22.Oppenheim
- 23.Time series
- 24.Longitudinal study
- 25.Nova-like
- 26.Cyberspace
- 27.Nordic portion of the web
- 28.Connections
- 29.Typology
- 30.Bates and Lu
- 31.Wynn and Katz
- 32.Dillon and Gushrowski
- 33.Web archaeology
- 34.Lei Cui
- 35.Central sites
- 36.Query
- 37.TREC IR: Text Retrieval Conference Information Retrieval
- 38.Inaccurate
- 39.Misleading
- 40.Un-referenced
- 41.Sitations
- 42.Incoming links
- 43.Data points
- 44.Lotka distribution
- 45.Self-sitations
- 46.National domains
- 47.Self-linking
- 48.External inlinking
- 49.TheIWall
- 50.URL element
- 51.Knowledge discovery in databases (KDD)
- 52.Nontrivial extraction of implicit
- 53.Textural data mining
- 54.Etzioni
- 55.Multi-agent
- 56.Metadata
- 57.Web mining
- 58.De Jong and Rip
- 59.Strong ties
- 60.Web communities
- 61.Topic distillation
- 62.Time Berners-Lee
- 63.Conseil Europeen Pour la Recherche Nucleaire (CERN)
- 64.Vertices
- 65.Edges
- 66.Hyperlinks
- 67.Bow tie
- 68.Web crawl
- 69.Bowtie knot
- 70.Random graphs
- 71.Watts and Strogats
- 72.Nematod worm
- 73.Kochen
- 74.Scientific domains
- 75.Co-term
- 76.Heterogeneous web clusters
- 77.Creativity simulation
- 78.Trail
- 79.Vannervar Bush
- 80.Cross-linking
- 81.Boundary crossings
- 82.Research fronts
- 83.Cross-disciplinary
- 84.Invisible colleges
- 85.Crossing scientific boundaries
- 86.Random walk
- 87.Path analysis
- 88.Co-occurrence
- 89.IP-numbers
- 90.Undiscovered public knowledge

91. Pathways  
 92. Co-citation  
 93. Qin and Norton  
 94. Co-link analysis of outlinks co-occurring  
 95. Transitive relations  
 96. Raynaud's disease  
 97. Blood platelets  
 98. literature-based knowledge discovery  
 99. Systematic serendipity  
 100. Valdes-Perez  
 101. Issues management  
 102. Wormell  
 103. Acid rain  
 104. Bar-Ilan and Peritz  
 105. Field codes  
 106. Harvesting  
 107. Visualisation/Navigation features of browsers

## منابع

- E.S. ALLEN, J.M. BURKE, M.E. WELCH, L.H. RIESEBERG (1999). How reliable is science information on the Web? *Science*, 402: 722.
- T. ALMIND, P. INGWERSEN (1997). Informetric analyses on the World Wide Web: Methodological approaches to "Webometrics", *Journal of Documentation*, 53: 404-426.
- J. BAR-ILAN (1998). The mathematician, Paul Erdos (1913-1996) in the eyes of the Internet, *Scientometrics*, 43: 257-267.
- J. BAR-ILAN (1999). Search engine results over time: A case study on search engine stability, *Cybermetrics*, 2/3, paper 1. ISSN: 1137-5019 (<http://www.cindoc.csic.es/cybermetrics/articles/v2i1p1.html>; visited 08.11.2000).
- J. BAR-ILAN (2000). The Web as an information resource on informetrics? A content analysis, *Journal of the American Society for Information Science*, 51: 432-443.
- J. BAR-ILAN, B.C. PERITZ (1999). The life span of a specific topic on the Web. The case of "informetrics": A quantitative analysis, *Scientometrics*, 46: 371-382.
- M. BATES, S. LU (1997). An exploratory profile of personal home pages: Content, design, metaphors, *Online & CDROM Review*, 21: 331-340
- D. BAWDEN (1986). Information systems and the stimulation of creativity, *Journal of Information Science*, 12: 203-216.
- T. BERNERS-LEE (1997). Realising the full potential of the Web. World Wide Web Consortium. (<http://www.w3.org/1998/02/Potential.html>; visited 08.11.2000).
- K. BHARAT, M. HENZINGER (1998). Improved algorithms for topic distillation in a hyperlinked environment. In: CROFT, W. B. et al. (Eds.). *Proceedings of the 21<sup>st</sup> annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, pp. 104-111.
- L. BJÖRNEBORN (2000). *Verdensvævet som 'small-world'-netværk og mulighedsrum : omridset af en forståelsesmodel for transversale links på World Wide Web*. [Small-World' Web and Possibility Space: outlining a conceptual framework for transversal links on the WWW]. Master's Thesis. Royal School of Library and Information Science, Copenhagen.
- S. BRIN, L. PAGE (1998). The anatomy of a large-scale hypertextual Web search engine, WWW7 Conference. (<http://www-db.stanford.edu/~backrub/google.html>; visited 08.11.2000).
- A. BRODER et al. (2000). Graph structure in the Web, WWW9 Conference. (<http://www.almaden.ibm.com/cs/k53/www9.final>; visited 08.11.2000)
- V. BUSH (1945). As we may think, *The Atlantic Monthly*, 176 (July) 641-649.
- S.J. CLARKE, P. WILLET (1997). Estimating the recall performance of Web search engines, *Aslib Proceedings*, 49: 184-189.
- M.P. COURTOIS, M.W. BERRY (1999). Results ranking in Web search engines, *Online*, (May/June) 39-46.



- B. CRONIN, G. MCKIM (1996). Science and scholarship on the World Wide Web: A North American perspective, **Journal of Documentation**, 52: 163-172.
- L. CUI (1999). Rating health Web sites using the principles of citation analysis: A bibliometric approach. **Journal of Medical Internet Research**, 1(1) e4 (ISSN: 1438-8871) (<http://www.jmir.org/1999/1/e4/index.htm>; visited 08.11.2000).
- R. DAVIES (1989). The creation of new knowledge by information retrieval and classification, **Journal of Documentation**, 45: 273-301.
- H. DE JONG, A. RIP (1997). The computer revolution in science: steps towards the realization of computer-supported discovery environments, **Artificial Intelligence**, 91: 225-256.
- A. DILLON, B.A. GUSHROWSKI (2000). Genres and the Web: Is the personal home page the first uniquely digital genre? **Journal of the American Society for Information Science**, 51: 202-205.
- O. ETZIONI (1996). The World-Wide Web: quagmire or gold mine?, **Communications of the ACM**, 39: 65-68.
- G. EYSENBACH (1998). Towards quality management of medical information on the Internet: Evaluation, labelling, and filtering of information, **British Medical Journal**, 317: 1496-1502.
- U. FAYYAD, G. PIATETSKY-SHAPIRO, P. SMYTH (1996). The KDD process for extracting useful knowledge from volumes of data, **Communications of the ACM**, 39 (Nov.) 27-34.
- N. FORD (1999). Information retrieval and creativity : towards support for the original thinker, **Journal of Documentation**, 55: 528-542.
- W.J. FRAWLEY, G. PIATETSKY-SHAPIRO, C.J. MATHEUS (1991). Knowledge discovery in databases: an overview, In: G. PIATETSKY-SHAPIRO, W.J. FRAWLEY (Eds.). **Knowledge discovery in databases**. Menlo Park, Cal.: AAAI Press.
- E. GARFIELD (1966). The who and why of ISI, *Essays of an Information Scientist*, 1 (1962-73) 33-37. Originally printed in **Karger Gazette**, March 5, 1966.
- E. GARFIELD (1994). Linking literatures: An intriguing use of the citation index, **Current Contents**, 21 (May 23) 3-5.
- D. GIBSON, J. KLEINBERG, P. RAGHAVAN (1998). Inferring web communities from link topology, **Proceedings of the 9<sup>th</sup> ACM Conference on Hypertext and Hypermedia**. (<http://www.cs.cornell.edu/home/kleinber/ht98.pdf>; visited 08.11.2000).
- M.S. GRANOVETTER (1973). The strength of weak ties, **American Journal of Sociology**, 78: 1360-1380.
- P. INGWERSEN (1998). The calculation of Web Impact Factors, **Journal of Documentation**, 54: 236-243.
- J.T. KLEIN (1996). **Crossing boundaries: knowledge, disciplinarity, and interdisciplinarity**, Charlottesville, Virg.: University Press of Virginia.
- J.M. KLEINBERG (1998). Authoritative sources in a hyperlinked environment, **Proceedings of the 9<sup>th</sup> annual ACM-SIAM Symposium on Discrete Algorithms**, pp. 668-677.
- M. KOCHEN (Ed.)(1989). **The small world**. Norwood, N.J.: Ablex Publishing Corporation.
- F.W. LANCASTER, J.-L. LEE (1985). Bibliometric techniques applied to issues management: A case study, **Journal of the American Society for Information Science**, 36: 389-397.
- R. LARSON (1996). Bibliometrics of the World Wide Web: An exploratory analysis of the intellectual structure of cyberspace. In: S. HARDIN (Ed.) **Proceedings of the 59<sup>th</sup> Annual Meeting of the American Society for Information Science**, 33: 71-78.
- S. LAWRENCE, C.L. GILES (1998). Searching the World Wide Web. **Science**, 280: 98-100.
- P. LOSIEWICZ, D.W. OARD, R.N. KOSTOFF (2000). Textual data mining to support science and technology management, **Journal of Intelligent Information Systems**, 15: 99-119.
- S. MILGRAM (1967). The small-world problem, *Psychology Today*, 1: 60-67.

- C. OPPENHEIM, A. MORRIS, C. MCKNIGHT (2000). The evaluation of WWW search engines. **Journal of Documentation**, 56: 190-211.
- S.J. PIERCE (1999). Boundary crossing in research literatures as a means of interdisciplinary information transfer, **Journal of the American Society for Information Science**, 50: 271-279.
- J. QIN, M.J. NORTON (Eds.)(1999). Introduction (In issue: Knowledge Discovery in Bibliographic Databases). **Library Trends**, 48 (Summer) 1-8.
- J.M. RODRIGUEZ I GAIRIN (1997). Volorando el impacto de la informacion en Internet: Altavista, el "Citation Index" de la Red. **Revista Espanola de Documentacion Cientifica** 20 (2): 175-181.
- R. ROUSSEAU (1997). Sitations: An exploratory study. **Cybermetrics**, 1, paper 1. ISSN: 1137-5019. (<http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html>; visited 08.11.2000).
- R. ROUSSEAU (1999). Daily time series of common single word searches in AltaVista and NorthernLight. **Cybermetrics**, 2/3 paper 2. ISSN: 1137-5019. (<http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html>; visited 08.11.2000).
- H. SMALL (1999). A passage through science: Crossing disciplinary boundaries, **Library Trends**, 48 (Summer) 72-108.
- A.G. SMITH (1999). A tale of two web spaces: Comparing sites using web impact factors. **Journal of Documentation**, 55: 577-592.
- H. SNYDER, H. ROSENBAUM (1999). Can search engines be used as tools for web-link analysis? A critical view, **Journal of Documentation**, 55: 375-384.
- D.R. SWANSON (1986). Undiscovered public knowledge, **Library Quarterly**, 56: 103-118.
- D.R. SWANSON, N.R. SMALHEISER (1997). An interactive system for finding complementary literatures: A stimulus to scientific discovery, **Artificial Intelligence**, 91: 183-203.
- D.R. SWANSON, N.R. SMALHEISER (1999). Implicit text linkages between Medline records: using Arrowsmith as an aid to scientific discovery, **Library Trends**, 48 (Summer) 48-59.
- M. THELWALL (2000). Web impact factors and search engine coverage, **Journal of Documentation**, 56: 185-189.
- R.E. VALDÉS-PEREZ (1999). Principles of human-computer collaboration for knowledge discovery in science', **Artificial Intelligence**, 107: 335-346.
- B. VICKERY (1997). Knowledge discovery from databases: an introductory review, **Journal of Documentation**, 53: 107-122.
- D.J. WATTS (1999). **Small worlds: the dynamics of networks between order and randomness**, Princeton University Press, Princeton, N.J.
- D.J. WATTS, S.H. STROGATZ (1998). Collective dynamics of 'small-world' networks, **Nature**, 393 (June 4) 440-442.
- I. WORMELL (2000). Critical aspects of the Danish welfare state - as revealed by issue tracking, **Scientometrics**, 48: 237-250.
- E. WYNN, J.E. KATZ (1997). Hyperbole over cyberspace: Self-presentation and social boundaries in Internet home pages and discourse, **Information Society**, 13: 297-32