پژوهشگاه اطلاعات و مدارک علمی ایران

فصلنامه علوم و فناوري اطلاعات. دوره ۲۱

شماره۱ (پاييز ۱۳۸۴) صفحه:۷۹-۸۶

نمايه‌سازي توزيع‌شده وب با استفاده از خزنده مهاجر[۱]

پاپاپترو، پاپاستاورو وساماراس

مترجم: رستم مظفري‌غربا

كارشناسي ارشد علوم كتابداري و اطلاع‌رساني

پست الكترونيكي: rostammozaffari@gmail.com

چكيده

به علت سرعت بسيار زياد در افزايش منابع وب و بسامد بالاي تغييرات، نگهداري يک نمايه روزآمد براي مقاصد جستجوگري (موتورهاي كاوش) به يک چالش تبديل مي‌شود. روش‌هاي سنتي خزنده‌ها، ديگر اين قابليت را ندارند كه با روزآمد شدن و رشد دائمي وب، همگام شوند. با درك اين مسئله، در اين مقاله ما يك روش جايگزين، يعني روش خزنده توزيع‌شده با استفاده از عامل‌هاي سيار[۲] را پيشنهاد مي‌كنيم. هدف ما ارائه يك شيوه خزش مقياس‌پذيراست كه ميزان بهره‌گيري از شبكه را كم كند، با تغييرات منابع هماهنگ باشد، درك زماني را به كار گيرد، و به سهولت قابل ارتقا باشد

كليدواژه‌ها: خزنده مهاجر (وب)، موتور كاوش، عامل متحرك (وب)، خزش وب، نمايه‌سازي توزيع‌شده، روزآمدسازي، يوسيميكرا (سيستم خزنده وب)

۱. مقدمه

در نتيجه ماهيت پويا و رشديابنده وب، نمايه‌سازي آن به يك چالش تبديل شده است. منابع وبي كه به صورت مستقيم در دسترس هستند (و از آن با عنوان وب سطحي[۳] ياد شده)، بيش‌تر از ۲/۵ ميليارد مدرک است، در حالي كه منابع وب غيرمستقيم[۴] (اسنادي كه به نحو پويا توليد مي‌شوند) در حدود سه برابر اين مقدار تخمين زده شده است (Lyman, et al., 2003). بعلاوه، حدود ۴۰ درصد از محتويات وب هر ماه دچار تغيير مي‌شود (Kahle, 1996)، در حالي كه هيچ موتور كاوشي تاكنون نتوانسته بيش از ۱۶ درصد از اين مقدار تخمين زده شده وب را زيرپوشش قرار دهد (Giles, 1999 & ,Lawrence).

خزيدن در وب (يا خزيدن سنتي) از سال ۱۹۹۳، اقدام غالب در نمايه‌سازي وب توسط موتورهاي كاوش معروف و سازمان‌هاي تحقيقاتي بوده است، اما با وجود منابع وسيع رايانشي و شبكه‌اي كه به درون وب ريخته مي‌شوند، خزيدن سنتي نمي‌تواند به طور مؤثر با پوياي وب، همگام گردد. به بيان دقيق‌تر، مدل خزيدن سنتي به دلايل زير موفق نيست:

۱. پردازش داده‌هاي حاصل از خزيدن، باعث تنگناي شديد پردازش در سايت موتور كاوش مي‌شود.

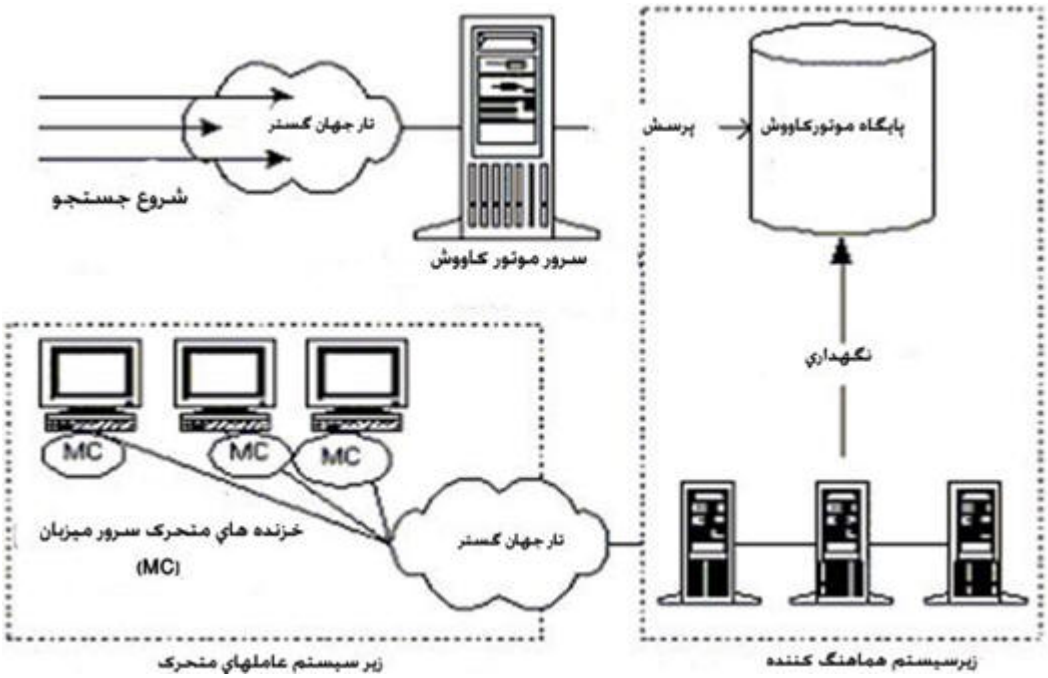
۲. تلاش براي ضبط كردن هزاران سند در ثانيه باعث ايجاد تنگنا در شبكه[۵] مي‌شود.

۳. اسناد معمولاً بدون فشرده‌سازي، توسط خزنده ضبط مي‌شوند و اين امر باعث بروز تنگنا در شبكه مي‌شود. عموماً فشرده‌سازي چندان آسان نمي‌باشد، زيرا مستقل از كار خزيدن است و نمي‌توان الزاماً به وسيله خزنده به آن اقدام كرد. به علاوه، خزنده‌ها همه محتويات يک سند- شامل اطلاعات

غيرمفيدي از قبيل توضيحات كد و برنامه- را كه به‌ندرت در نمايه‌سازي اسناد، لازم مي‌آيند نيز ضبط مي‌كنند.
به دليل نبود روش خزيدن مقياس‌پذير، در چند سال اخير تحقيقات مهمي انجام شده است. خزيدن متمركز[۶] (Chakrabarti et al. ,1999) به عنوان يک روش جايگزين پيشنهاد شد، اما هيچ نوآوري در معماري را موجب نشد، زيرا بر اساس همان كاركردهاي تمركز گرايانه خزيدن سنتي بنا شده بود.
به‌عنوان نخستين تلاش براي ايجاد ماهيت تمركزگرايانه خزيدن سنتي، چند روش توزيع‌شده پيشنهاد شده‌اند (مانند «هاروسِت» (Bowman, et al., 1994) و «گراب» (Ozra, 2001 & ,Kordless, Lajesus).
در اين مقاله، ما «يوسي ميكرا»[۷] را معرفي مي‌كنيم؛ سيستم خزنده‌اي كه از مفاهيمي همانند آنچه در خزيدن موبايل توزيع‌شده (كه در (Hammer, 1999, 2000 & Fiedler) معرفي شده) يافت مي‌شود استفاده مي‌كند. «يوسي ميكرا» اين مفاهيم را توسعه مي‌دهد و مفاهيم جديدي را معرفي مي‌كند تا مدل مؤثرتري براي خزيدن توزيع‌شده در وب بسازد كه قادر باشد بي‌درنگ خود را با تغييرات منابع وب، هماهنگ كند.
«يوسي‌ميكرا» با به‌كارگيري فناوري «عامل‌هاي سيار» يك راهبرد خزيدن كاملاً توزيع‌شده را پيشنهاد مي‌كند. اهداف اين پيشنهاد عبارت‌اند از:
الف) به حداقل رساندن مقدار به كارگيري شبكه؛
ب) هماهنگ‌شدن با تغييرات منابع، با اجراي نظارت درون- سايتي؛
ج) اجتناب از اضافه‌بار غيرضروري سرورهاي وب با به‌كارگيري تحقق همزماني[۸]؛
د) قابليت ارتقا در زمان اجرا.

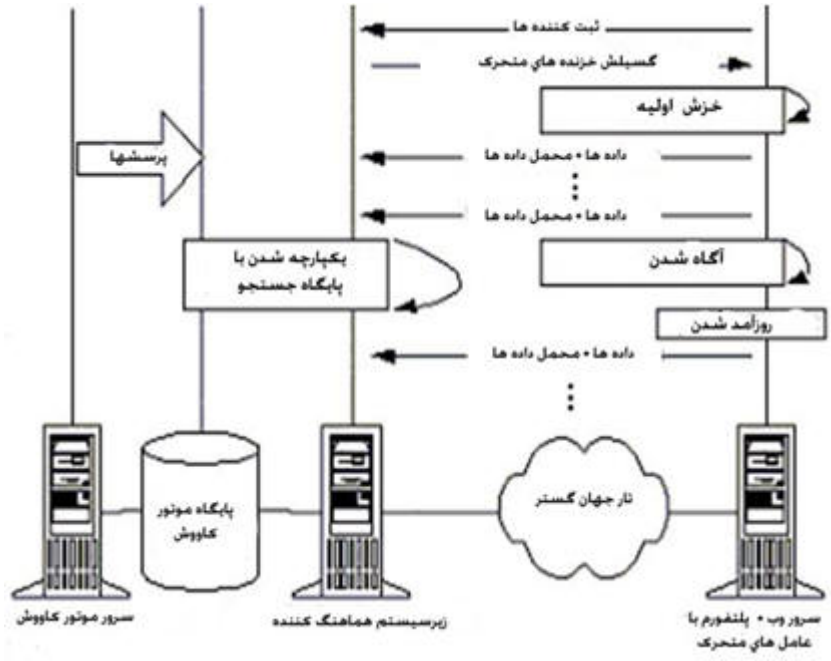
۲. سيستم خزنده «يوسي ميكرا»

نيروي محرک «يوسي‌ميكرا» استفاده از عامل‌هاي سياري است كه از موتور كاوش به سرورهاي وب مهاجرت مي‌كنند و براي خزش، پردازش، و نظارت بر منابع وب براي روزآمدسازي، در آنجا باقي مي‌مانند. از آنجا كه«يوسي‌ميكرا» در سرور وبي كه بايد مورد خزش قرار بگيرد نيازمند به اجرا درآمدن نوع خاصي از كارپايه عامل‌هاي سيار مي‌باشد، در حال حاضر در محيط دانشگاهي داوطلب كه در سطح چند قاره گسترده شده‌اند [به صورت آزمايشي] در حال اجرا مي‌باشد.
«يوسي‌ميكرا» (شكل۱) از سه زيرسيستم تشكيل مي‌شود: الف) زيرسيستم «هماهنگ كننده»[۹]؛ ب) زيرسيستم «عامل‌هاي سيار»[۱۰] ؛ ج) يک «موتور كاوش همگاني» كه پرسش‌هاي كاربر را در پايگاه اطلاعاتي كه توسط زيرسيستم هماهنگ‌كننده نگهداري مي‌شود، جستجو مي‌كند.
زيرسيستم هماهنگ‌كننده در سايت «موتور كاوش» قرار دارد و مسئوليت‌هاي آن عبارت‌اند از: الف) نگهداري پايگاه اطلاعاتي جستجو، ب) ارائه ثبت‌نام برخط[۱۱] براي وب‌سايت‌هاي جديدي كه مي‌خواهند در «يوسي‌ميكرا» مشاركت كنند، ج) اجراي «زيرسيستم عامل‌هاي سيار». «زيرسيستم عامل‌هاي سيار» مسئول خزيدن در وب مي‌باشد و از دو گونه از اين عامل‌ها، يعني «خزنده‌هاي مهاجر» (يا خزنده‌هاي سيار[۱۲]) و «حامل‌هاي داده‌ها» تشكيل مي‌شود. شكل ۲ «يوسي‌ميكرا» را در حين كار نشان مي‌دهد.
همان گونه كه در بالا ذكر شد، هسته سيستم خزنده «يوسي‌ميكرا»، «خزنده‌هاي مهاجر مبتني بر جاوا» هستند. خزنده‌هاي مهاجر بر پايه قابليت ماهوي سيار خود توانايي انجام كارهاي زير را دارند.



شكل١. معماري UCYMicra

١. گسيل‌شدن: به سوي سرورهاي وب جديدي كه مي‌خواهند در «يوسي‌ميكرا» مشاركت كنند؛
٢. خزيدن: يک خزنده مهاجر مي‌تواند (چه از طريق «اچ‌تي‌تي‌پي» يا سيستم فايلي) يك خزش محلي كامل انجام دهد؛
٣. پردازش: مداركي كه مورد خزش قرار گرفته‌اند به كليدواژه‌هايي تقليل داده مي‌شوند و اين كليدواژه‌ها بر اساس ويژگي‌هاي بصري (فونت و رنگ)، موقعيت و فراواني تكرار، و به منظور ايجاد يك نمايه كليدواژه‌اي در محل از محتويات سرور وب، رتبه‌بندي مي‌شوند؛
۴. فشرده‌سازي: نمايه محتويات سرور وب، در محل فشرده مي‌شوند تا زمان انتقال بين خزنده مهاجر و زيرسيستم هماهنگ‌كننده به حداقل برسد؛
۵. انتقال داده‌ها: نمايه فشرده‌شده، توسط حامل‌هاي داده‌ها به زير سيستم هماهنگ‌كننده منتقل مي‌شود. در آنجا، [نمايه فشرده‌شده] مجدداً به حالت اوليه برگردانده مي‌شود و در پايگاه اطلاعاتي جستجو، ادغام مي‌گردد. دليل استفاده از عامل‌هاي سيار براي انتقال داده‌ها بر روي ديگر رابطهای برنامه نويسي كاربردي[١٣] شبكه (مانند RMI, COBRA 14] Sockets]) كاربرد ناهمزماني، انعطاف‌پذيري و هوشمندي آن‌ها به منظور اطمينان از انتقال مستمر داده‌ها مي‌باشد.
۶. نظارت: خزنده مهاجر مي‌تواند تغييرات به وقوع پيوسته در محتويات سرور وب را تشخيص دهد. تغييرات تشخيص داده‌شده بلافاصله پردازش، فشرده‌سازي، و به زيرسيستم هماهنگ‌كننده منتقل مي‌شوند.
٧. ارتقاهاي بي‌درنگ: كدهاي جديد مربوط به انجام هر يك از كارهاي مذكور در بالا به آساني قابل استفاده است، زيرا معماري خزش در «يوسي‌ميكرا» بر پايه جاوا مي‌باشد.



شکل۲. UCYMicra در حین کار

۳. ارزیابي «يوسي ميكرا»

به مقايسه عملكرد سيستم خزنده «يوسي‌ميكرا» با خزيدن سنتي در خصوص الف) حجم داده‌هاي انتقال داده‌شده در اينترنت، ب) كل زمان لازم براي انجام فرآيند خزيدن در يک سري از مدارک مي‌پردازيم. در آغاز كار، فقط اين دو متريک ساده]يعني حجم و زمان[را مورد مطالعه قرار مي‌دهيم و با پارامترهايي مانند بسامد تغييرات مدارک، آزمايش نمي‌كنيم. از آنجا كه اين امكان وجود نداشت كه سرورهاي تجاري وب را در آزمايش‌هاي خود داخل كنيم، يک مجموعه ده‌تايي از سرورهاي دانشگاهي وب را كه در چندين قاره پراكنده بودند در محيط توزيع‌شده داوطلبانه دانشگاهي خود به خدمت گرفته‌ايم. هرسرور وبي، ميزبان حدود ۲۰۰ مدرک با ميانگين حجم ۲۵ كيلو بايت بود. اعداد قبلي بيانگر اين بود كه۴۶/۲ مگابايت از داده‌ها بايد به وسيله خزيدن سنتي وشيوه «يوسي‌ميكرا» پردازش شوند. به دليل محدوديت فضا، يافته‌هاي خود در خصوص حجم داده‌هاي انتقال داده شده را ارائه مي‌كنيم (يافته‌هاي ما در مورد زمان مورد نياز،]به اين يافته‌ها[شبيه هستند).

جدول۱. نتايج عملكرد

methodology	Data moved
Traditional crawling	46.9Mb
UCYMicra- no processing, no compression	48.1Mb
UCYMicra- w/processing, no compression	13.3Mb
UCYMicra- no processing, w/compression	8.1Mb
UCYMicra- (w/processing and compression)	2.6Mb

نتايج عملكرد (جدول۱) نشان مي‌دهند كه «يوسي‌ميكرا» (رديف ۵) با توليد تقريباً ۲۰ برابر داده‌هاي كمتر، عملكرد بهتري نسبت به خزيدن سنتي (رديف۱) دارد. دليل اين امر آن است كه خزنده‌هاي مهاجر، منابع وب را به صورت محلي در سرور وب، پردازش و فشرده مي‌كنند. به اين ترتيب، فقط نمايه

رتبه‌بندي‌شده كليدواژه‌هاي فشرده‌شده از محتويات سرور وب، به زيرسيستم هماهنگ‌كننده انتقال داده مي‌شود. در شيوه خزيدن سنتي، بايد كل محتويات يک سرور وب براي پردازش متمركز به وسيله خزنده، ضبط شوند. بعلاوه، خزنده سنتي ممكن است نيازمند فشرده‌سازي محتويات يك سرور وب به منظور ضبط آن‌ها باشد، ولي نتواند آن سرور را وادار به اين كار كند.

براي بدست آوردن تفسير بهتري از نتايج عملكرد خود، سه آزمايش ديگر، اين بار با تغيير شيوه «يوسي‌ميكرا» به منظور انجام دادن (يا انجام ندادن) پردازش و فشرده‌سازي به صورت محلي انجام داديم. نتايج ما نشان داد كه چه با پردازش و چه با فشرده‌سازي، يافته‌هاي عملكردي باز هم تا حدودي مصداق دارند. اما بدون فعال كردن گزينه پردازش يا فشرده‌سازي، نتايج به دست آمده از ميان مي‌روند، زيرا «يوسي‌ميكرا» با شيوه خزيدن سنتي رقابت مي‌كند.

۴. كار جاري

كاري كه ما در حال حاضر روي آن تمركز داريم، توسعه «يوسي ميكرا» براي پشتيباني از يک سازوكار خزيدن دوگانه مي‌باشد كه فناوري‌هايي را، هم از سيستم خزيدن سنتي و هم از سيستم خزيدن كاملاً توزيع‌شده، وام گرفته است. اين سيستم خزيدن دوگانه از يك ساختار مديريت سلسله‌مراتبي پشتيباني خواهد كرد كه شبكه را به صورت محلي در نظر مي‌گيرد. الگوريتم‌هاي كارآمد براي احاله كار، اداره، و تلفيق نتايج كار هم اكنون در حال انجام هستند.

۵. منابع

Bowman, C. M., Danzig, P. B., Hardy, D. R., Manber, U., Schwartz, M. F. (1994). The Harvest information discovery and access System. In Proceedings of the Second International World Wide Web Conference (pp. 763-771). Chicago, Illinois.

Chakrabarti, S., van der Berg, M., & Dom, B. (1999). Focused crawling: a new approach to topic-specific web resource discovery. In A. Mendelzon (Ed.), Proceedings of the 8th International World-Wide Web Conference (pp. 1623-1640). University of Toronto.

Fiedler, J., and Hammer, J. (1999). Using the web efficiently: mobile crawling. In Proceedings of the Seventeenth Annual International Conference of the Association of Management (AoM/IAoM) on Computer Science (pp. 324-329). San Diego, CA.

Fiedler, J., and Hammer, J. (2000). Using mobile crawlers to search the web efficiently. International Journal of Computer and Information Science, 1(1), 36-58.

Kahle, B. (1996). Achieving the internet. Scientific American.

Kordless, Lajesus, Ozra. (2001). Grub: Distributed internet crawler. Available at: http://www.grub.org.

Lawrence, S., & Giles, C. L. (1999). Accessibility of information on the web. Nature, 400(6740), 107-109.

Lyman, P., Varian, H., Dunn, J., Strygin, A. & Swearinfen, L. (2003). How much information? (University of California at Berkeley). Retrieved March 25, 2004, from http://www.sims.berkeley.edu/how-much-info.

پي‌نوشت‌ها

[1]. Odysseas Papapetrou, Stavros Papastavavrou, and George Samaras (2003). Distributed indexing of the web using migrating crawlers. In Proceedings of the Twelfth International World Wide Web Conference (WWW). Retrived at http://softsys.cs.uoi.gr/dbg (obe/publications/ p304-papapetrou.pdf

[2]. mobile agents

[3]. Surface Web

[4]. indirect Web

[5] . Domian Name System (DNS)

[6]. Focused Crawling

[7]. Ucymicra

[8]. employing time realization

[9]. the Coordinator subsystem

[10]. the Mobile agents subsystem

[11]. Online

[12]. Migrating Crawler

[13]. Application Programing Interface (API)

[14]. Remote Method Invocation