

## نمایه‌سازی توزیع‌شده وب با استفاده از خزنده مهاجر[۱]

پاپاترو، پاپاستاورو وساماراس

مترجم: رستم مظفري غربا

کارشناسی ارشد علوم کتابداری و اطلاع‌رسانی

پست الکترونیکی: [rostammozaffari@gmail.com](mailto:rostammozaffari@gmail.com)

### چکیده

به علت سرعت بسیار زیاد در افزایش منابع وب و بسامد بالای تغییرات، نگهداری یک نمایه روزآمد برای مقاصد جستجوگری (موتورهای کاوش) به یک چالش تبدیل می‌شود. روش‌های سنتی خزنده‌ها، دیگر این قابلیت را ندارند که با روزآمد شدن و رشد دائمی وب، همگام شوند. با درک این مسئله، در این مقاله ما یک روش جایگزین، یعنی روش خزنده توزیع‌شده با استفاده از عامل‌های سیار [۲] را پیشنهاد می‌کنیم. هدف ما ارائه یک شیوه خزش مقیاس‌پذیر است که میزان بهره‌گیری از شبکه را کم کند، با تغییرات منابع هماهنگ باشد، درک زمانی را به کار گیرد، و به سهولت قابل ارتقا باشد.

**کلیدواژه‌ها:** خزنده مهاجر (وب)، موتور کاوش، عامل متحرک (وب)، خزش وب، نمایه‌سازی توزیع‌شده، روزآمدسازی، یوسیمیکرا (سیستم خزنده وب)

### ۱. مقدمه

در نتیجه ماهیت پویا و رشدیابنده وب، نمایه‌سازی آن به یک چالش تبدیل شده است. منابع وبی که به صورت مستقیم در دسترس هستند (و از آن با عنوان وب سطحی [۳] یاد شده)، بیش‌تر از ۲/۵ میلیارد مدرک است، در حالی که منابع وب غیرمستقیم [۴] (اسنادی که به نحو پویا تولید می‌شوند) در حدود سه برابر این مقدار تخمین زده شده است (Lyman, et al., 2003). بعلاوه، حدود ۴۰ درصد از محتویات وب هر ماه دچار تغییر می‌شود (Kahle, 1996)، در حالی که هیچ موتور کاوشی تاکنون نتوانسته بیش از ۱۶ درصد از این مقدار تخمین زده شده وب را زیرپوشش قرار دهد (Giles, 1999 & Lawrence).

خزیدن در وب (یا خزیدن سنتی) از سال ۱۹۹۳، اقدام غالب در نمایه‌سازی وب توسط موتورهای کاوش معروف و سازمان‌های تحقیقاتی بوده است، اما با وجود منابع وسیع رایانشی و شبکه‌ای که به درون وب ریخته می‌شوند، خزیدن سنتی نمی‌تواند به طور مؤثر با پویایی وب، همگام گردد. به بیان دقیق‌تر، مدل خزیدن سنتی به دلایل زیر موفق نیست:

۱. پردازش داده‌های حاصل از خزیدن، باعث تنگنای شدید پردازش در سایت موتور کاوش می‌شود.
۲. تلاش برای ضبط کردن هزاران سند در ثانیه باعث ایجاد تنگنا در شبکه [۵] می‌شود.
۳. اسناد معمولاً بدون فشرده‌سازی، توسط خزنده ضبط می‌شوند و این امر باعث بروز تنگنا در شبکه می‌شود. عموماً فشرده‌سازی چندان آسان نمی‌باشد، زیرا مستقل از کار خزیدن است و نمی‌توان الزاماً به وسیله خزنده به آن اقدام کرد. به علاوه، خزنده‌ها همه محتویات یک سند- شامل اطلاعات

غیرمفیدی از قبیل توضیحات کد و برنامه- را که به ندرت در نمایه‌سازی اسناد، لازم می‌آیند نیز ضبط می‌کنند.

به دلیل نبود روش خزیدن مقیاس‌پذیر، در چند سال اخیر تحقیقات مهمی انجام شده است. خزیدن متمرکز [۶] (Chakrabarti et al., 1999) به عنوان یک روش جایگزین پیشنهاد شد، اما هیچ نوآوری در معماری را موجب نشد، زیرا بر اساس همان کارکردهای تمرکزگرایانه خزیدن سنتی بنا شده بود. به‌عنوان نخستین تلاش برای ایجاد ماهیت تمرکزگرایانه خزیدن سنتی، چند روش توزیع‌شده پیشنهاد شده‌اند (مانند «هاروسیت» (Bowman, et al., 1994) و «گراب» (Ozra, 2001 & Kordless, Lajesus). در این مقاله، ما «یوسی میکرا» [۷] را معرفی می‌کنیم؛ سیستم خزنده‌ای که از مفاهیمی همانند آنچه در خزیدن موبایل توزیع‌شده (که در (Hammer, 1999, 2000 & Fiedler) معرفی شده) یافت می‌شود استفاده می‌کند. «یوسی میکرا» این مفاهیم را توسعه می‌دهد و مفاهیم جدیدی را معرفی می‌کند تا مدل مؤثرتری برای خزیدن توزیع‌شده در وب بسازد که قادر باشد بی‌درنگ خود را با تغییرات منابع وب، هماهنگ کند.

«یوسی میکرا» با به‌کارگیری فناوری «عامل‌های سیار» یک راهبرد خزیدن کاملاً توزیع‌شده را پیشنهاد می‌کند. اهداف این پیشنهاد عبارت‌اند از:

(الف) به حداقل رساندن مقدار به‌کارگیری شبکه؛

(ب) هماهنگ‌شدن با تغییرات منابع، با اجرای نظارت درون-سایتی؛

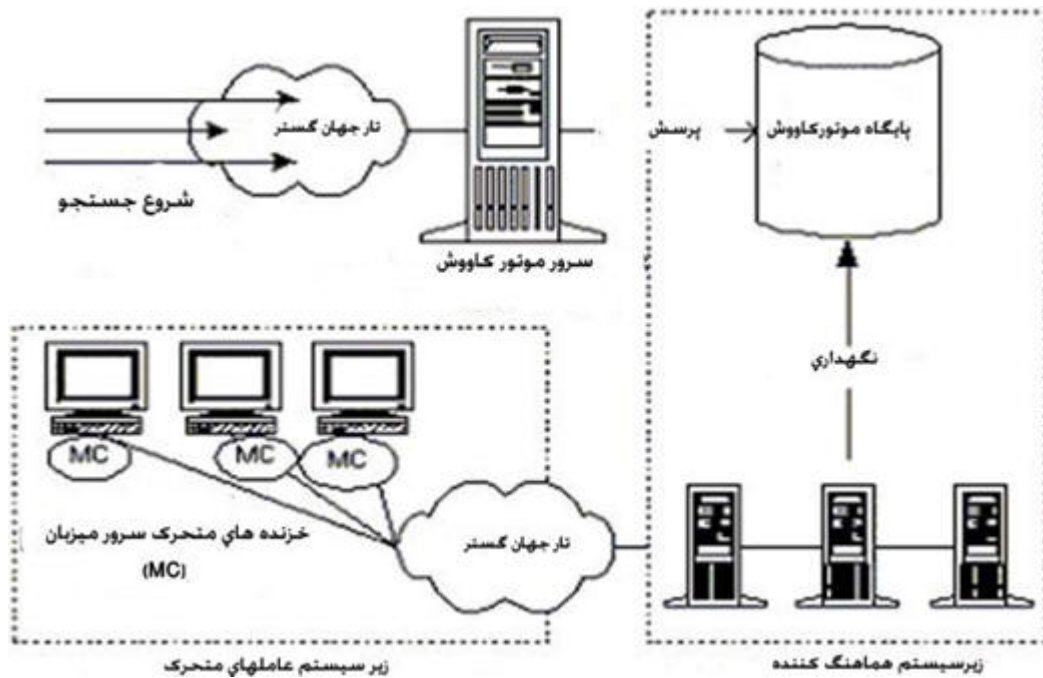
(ج) اجتناب از اضافه‌بار غیرضروری سرورهای وب با به‌کارگیری تحقق همزمانی [۸]؛

(د) قابلیت ارتقا در زمان اجرا.

## ۲. سیستم خزنده «یوسی میکرا»

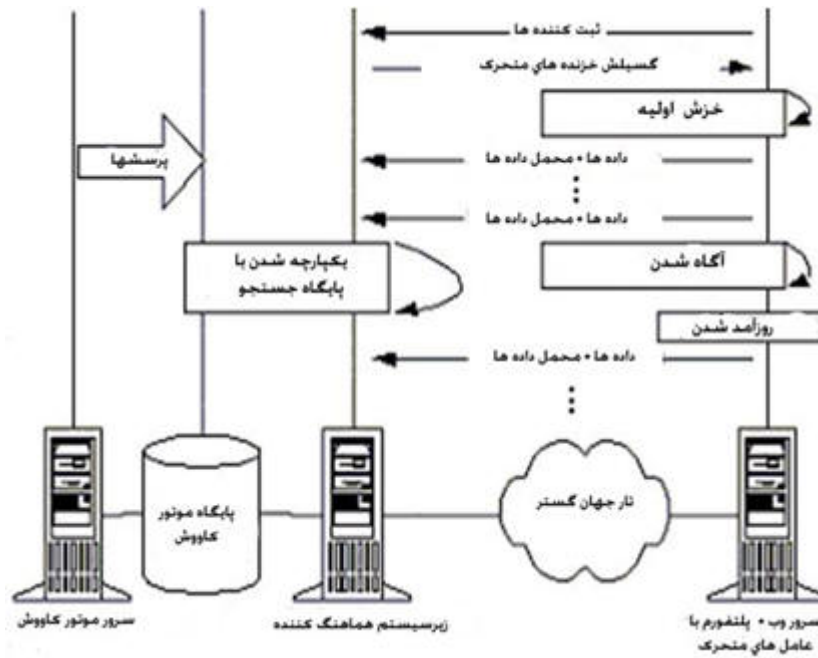
نیروی محرک «یوسی میکرا» استفاده از عامل‌های سیاری است که از موتور کاوش به سرورهای وب مهاجرت می‌کنند و برای خزش، پردازش، و نظارت بر منابع وب برای روزآمدسازی، در آنجا باقی می‌مانند. از آنجا که «یوسی میکرا» در سرور وبی که باید مورد خزش قرار بگیرد نیازمند به اجرا درآمد نوع خاصی از کارپایه عامل‌های سیار می‌باشد، در حال حاضر در محیط دانشگاهی داوطلب که در سطح چند قاره گسترده شده‌اند [به صورت آزمایشی] در حال اجرا می‌باشد. «یوسی میکرا» (شکل ۱) از سه زیرسیستم تشکیل می‌شود: الف) زیرسیستم «هماهنگ کننده» [۹]؛ ب) زیرسیستم «عامل‌های سیار» [۱۰]؛ ج) یک «موتور کاوش همگانی» که پرسش‌های کاربر را در پایگاه اطلاعاتی که توسط زیرسیستم هماهنگ‌کننده نگهداری می‌شود، جستجو می‌کند. زیرسیستم هماهنگ‌کننده در سایت «موتور کاوش» قرار دارد و مسئولیت‌های آن عبارت‌اند از: الف) نگهداری پایگاه اطلاعاتی جستجو، ب) ارائه ثبت‌نام برخط [۱۱] برای وبسایت‌های جدیدی که می‌خواهند در «یوسی میکرا» مشارکت کنند، ج) اجرای «زیرسیستم عامل‌های سیار». «زیرسیستم عامل‌های سیار» مسئول خزیدن در وب می‌باشد و از دو گونه از این عامل‌ها، یعنی «خزنده‌های مهاجر» (یا خزنده‌های سیار [۱۲]) و «حامل‌های داده‌ها» تشکیل می‌شود. شکل ۲ «یوسی میکرا» را در حین کار نشان می‌دهد.

همان گونه که در بالا ذکر شد، هسته سیستم خزنده «یوسی میکرا»، «خزنده‌های مهاجر مبتنی بر جاوا» هستند. خزنده‌های مهاجر بر پایه قابلیت ماهوی سیار خود توانایی انجام کارهای زیر را دارند.



شکل ۱. معماری UCYMicra

۱. گسیل شدن: به سویی سرورهای وب جدیدی که می‌خواهند در «یوسی‌میکرا» مشارکت کنند؛
۲. خزیدن: یک خزنده مهاجر می‌تواند (چه از طریق «اچ‌تی‌تی‌پی» یا سیستم فایل) یک خزش محلی کامل انجام دهد؛
۳. پردازش: مدارکی که مورد خزش قرار گرفته‌اند به کلیدواژه‌هایی تقلیل داده می‌شوند و این کلیدواژه‌ها بر اساس ویژگی‌های بصری (فونت و رنگ)، موقعیت و فراوانی تکرار، و به منظور ایجاد یک نمایه کلیدواژه‌ای در محل از محتویات سرور وب، رتبه‌بندی می‌شوند؛
۴. فشرده‌سازی: نمایه محتویات سرور وب، در محل فشرده می‌شوند تا زمان انتقال بین خزنده مهاجر و زیرسیستم هماهنگ‌کننده به حداقل برسد؛
۵. انتقال داده‌ها: نمایه فشرده‌شده، توسط حامل‌های داده‌ها به زیرسیستم هماهنگ‌کننده منتقل می‌شود. در آنجا، [نمایه فشرده‌شده] مجدداً به حالت اولیه برگردانده می‌شود و در پایگاه اطلاعاتی جستجو، ادغام می‌گردد. دلیل استفاده از عامل‌های سیار برای انتقال داده‌ها بر روی دیگر رابط‌های برنامه نویسی کاربردی [۱۲] شبکه (مانند COBRA [14 RMI Sockets]) کاربرد ناهمزمانی، انعطاف‌پذیری و هوشمندی آنها به منظور اطمینان از انتقال مستمر داده‌ها می‌باشد.
۶. نظارت: خزنده مهاجر می‌تواند تغییرات به وقوع پیوسته در محتویات سرور وب را تشخیص دهد. تغییرات تشخیص داده‌شده بلافاصله پردازش، فشرده‌سازی، و به زیرسیستم هماهنگ‌کننده منتقل می‌شوند.
۷. ارتقاهای بی‌درنگ: کدهای جدید مربوط به انجام هر یک از کارهای مذکور در بالا به آسانی قابل استفاده است، زیرا معماری خزش در «یوسی‌میکرا» بر پایه جاوا می‌باشد.



شکل ۲. UCYMicra در حین کار

### ۳. ارزیابی «یوسی میکرا»

به مقایسه عملکرد سیستم خزنده «یوسی میکرا» با خزیدن سنتی در خصوص الف) حجم داده‌های انتقال داده‌شده در اینترنت، ب) کل زمان لازم برای انجام فرآیند خزیدن در یک سری از مدارک می‌پردازیم. در آغاز کار، فقط این دو متریک ساده [یعنی حجم و زمان] را مورد مطالعه قرار می‌دهیم و با پارامترهایی مانند تغییرات مدارک، آزمایش نمی‌کنیم. از آنجا که این امکان وجود نداشت که سرورهای تجاری وب را در آزمایش‌های خود داخل کنیم، یک مجموعه ده‌تایی از سرورهای دانشگاهی وب را که در چندین قاره پراکنده بودند در محیط توزیع‌شده داوطلبانه دانشگاهی خود به خدمت گرفته‌ایم. هر سرور وبی، میزان حدود ۲۰۰ مدرک با میانگین حجم ۲۵ کیلو بایت بود. اعداد قبلی بیانگر این بود که ۴۶/۲ مگابایت از داده‌ها باید به وسیله خزیدن سنتی و شیوه «یوسی میکرا» پردازش شوند. به دلیل محدودیت فضا، یافته‌های خود در خصوص حجم داده‌های انتقال داده شده را ارائه می‌کنیم (یافته‌های ما در مورد زمان مورد نیاز، [به این یافته‌ها] شبیه هستند).

جدول ۱. نتایج عملکرد

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

نتایج عملکرد (جدول ۱) نشان می‌دهند که «یوسی میکرا» (ردیف ۵) با تولید تقریباً ۲۰ برابر داده‌های کمتر، عملکرد بهتری نسبت به خزیدن سنتی (ردیف ۱) دارد. دلیل این امر آن است که خزنده‌های مهاجر، منابع وب را به صورت محلی در سرور وب، پردازش و فشرده می‌کنند. به این ترتیب، فقط نمایه

رتبه‌بندی‌شده کلیدواژه‌های فشرده‌شده از محتویات سرور وب، به زیرسیستم هماهنگ‌کننده انتقال داده می‌شود. در شیوه خزیدن سنتی، باید کل محتویات یک سرور وب برای پردازش متمرکز به وسیله خزنده، ضبط شوند. بعلاوه، خزنده سنتی ممکن است نیازمند فشرده‌سازی محتویات یک سرور وب به منظور ضبط آن‌ها باشد، ولی نتواند آن سرور را وادار به این کار کند. برای بدست آوردن تفسیر بهتری از نتایج عملکرد خود، سه آزمایش دیگر، این بار با تغییر شیوه «یوسی‌میکرا» به منظور انجام دادن (یا انجام ندادن) پردازش و فشرده‌سازی به صورت محلی انجام دادیم. نتایج ما نشان داد که چه با پردازش و چه با فشرده‌سازی، یافته‌های عملکردی باز هم تا حدودی مصداق دارند. اما بدون فعال کردن گزینه پردازش یا فشرده‌سازی، نتایج به دست آمده از میان می‌روند، زیرا «یوسی‌میکرا» با شیوه خزیدن سنتی رقابت می‌کند.

#### ۴. کار جاری

کاری که ما در حال حاضر روی آن تمرکز داریم، توسعه «یوسی‌میکرا» برای پشتیبانی از یک سازوکار خزیدن دوگانه می‌باشد که فناوری‌هایی را، هم از سیستم خزیدن سنتی و هم از سیستم خزیدن کاملاً توزیع‌شده، وام گرفته است. این سیستم خزیدن دوگانه از یک ساختار مدیریت سلسله‌مراتبی پشتیبانی خواهد کرد که شبکه را به صورت محلی در نظر می‌گیرد. الگوریتم‌های کارآمد برای احاله کار، اداره، و تلفیق نتایج کار هم اکنون 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. & Swearinfe, L. (2003). How much information? (University of California at Berkeley). Retrieved March 25, 2004, from <http://www.sims.berkeley.edu/how-much-info>.

---

### پي نوشت ها

- [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](http://softsys.cs.uoi.gr/dbg(obe/publications/p304-papapetrou.pdf)
- [2]. mobile agents
- [3]. Surface Web
- [4]. indirect Web
- [5]. Domain 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 Programming Interface (API)
- [14]. Remote Method Invocation

پژوهشگاه علوم انسانی و مطالعات فرهنگی  
رتال جامع علوم انسانی