علیرضا یاری مدرک لیسانس خود را در زمینه سیستم های کنترل در سال ۱۹۹۳ از دانشکده فنی دانشگاه تهران دریافت کرده است. در ادامه ایشان تحصیلات تکمیلی خود در زمینه مهندسی سیستم در مقاطع فوق لیسانس و دکترا در دانشکده کامپیوتر دانشگاه فنی کیتامی کشور ژاپن ادامه داده که در سال ۲۰۰۰ موفق به دریافت مدرک دکترا از آن دانشگاه شده اند. در حال حاضر زمینه تحقیقاتی ایشان در خصوص مراکزداده، سرویسهای وب و بسترسازی مناسب برای زبان فارسی در محیط رایانه ای می باشد. ایشان در حال حاضر مدیر گروه نرم افزار و سکوهای فناوری اطلاعات در مرکز تحقیقات مخابرات ایران می باشند.

ابوالفضل آل‌احمد در سال ۱۳۸۰ مدرک فوق‌دیپلم خود را در رشته نرم‌افزار کامپیوتر از مرکزآموزش عالی فنی مهندسی شهید باهنر شیراز، در سال ۱۳۸۲ مدرک کارشناسی خود را در رشته مهندسی کامپیوتر از دانشگاه شهید باهنر کرمان و در سال ۱۳۸۷ مدرک کارشناسی ارشد خود را در رشته مهندسی فناوری اطلاعات از دانشگاه تهران دریافت کرده است. ایشان از اعضاء اصلی پروژه‌های تحقیقاتی موفقی در پتروشیمی شیراز و پژوهشگاه صنعت نفت بوده است و از سال ۱۳۸۴ تا کنون عضو گروه تحقیقاتی پایگاه داده‌ها دانشگاه تهران بوده‌اند. زمینه تحقیقاتی ایشان بازیابی اطلاعات، آنالیز وب و داده‌کاوی است.

واکشی شده و کلمات تخصصی موجود در صفحه به عنوان داده‌های بازخوردی جهت توسعه دایره کلمات و یا منابع زبانی استفاده خواهد شد.

مراجع

[1] G. Pant, P. Srinivasan, and F. Menczer, *"Crawling the Web"*, In Web Dynamics: Adapting to Change in Content, Size, Topology and Use. Edited by M. Levene and A. Poulovassilis, Springer Verlag, pp: 153–178, 2004.

[2] M.P.S.Bhatia, Divya Gupta, *"Discussion on Web Crawlers of Search Engine"*, Proceedings of 2nd National Conference on Challenges & Opportunities in Information Technology (COIT-2008), Mandi Gobindgarh. March 29, pp: 227-230, 2008.

[3] George Almpanidis, Constantine Kotropoulos, Ioannis Pitas, *"Combining text and link analysis for focused crawling - An application for vertical search engines"*, Information System 32(6), pp: 886-908, 2007.

[4] A. Badia, T. Muezzinoglu, O. Nas-raoui, *"Focused crawling: experiences in a real world project"*, In Proceedings of the 15International Conference on World Wide Web, Edinburgh S., pp: 1043-1044, 2006.

[5] S. Chakrabarti, M. van den Berg, and B. Dom. "Focused crawling: *a new approach to topic-specific Web resource discovery"*, In Proceedings of the 8th International World Wide Web Conference, Torento, 1999.

[6] F. Menczer, G. Pant, P. Srinivasan and M. Ruiz , *"Evaluating Topic -Driven Web Crawlers"*, In Proceedings of the 24th Annual International ACM/SIGIR Conference, New Orleans, USA, 2001.

[7] J. Cho, H. Garcia-Molina, and L. Page, *"Efficient Crawling Through URL Ordering"*, In Proceedings of the 7th International World-Wide Web Conference, 1998

[8] N. Angkawattanawit, A. Rungsawang, *"Learnable Crawling: An Efficient Approach to Topic-specific Web Resource Discovery"*, In Proceedings of the 2nd International Symposium on Communications and Information Technology (ISCIT), 2002.

[9] P. De Bra, G.-J. Houben, Y. Kornatzky, R. Post, *"Information retrieval in distributed hypertexts"*, In Proceedings of RIAO'94, Intelligent Multimedia, Information Retrieval Systems and Management, New York, NY, 1994.

[10] M. Hersovici, M. Jacovi, Y.S. Maarek, D. Pelleg, M. Shtalheim, and S. Ur, *"The Shark-Search algorithm - an application:tailored Web site mapping"*, In: 7th World-Wide, Web Conference, Brisbane, Australia, online, 1998.

[11] M. Ehrig, A. Maedche, *"Ontology-Focused Crawling of web documents"*, In Proceedings of the ACM Symposium on Applied Computing, 2003.

[12] Yang, K., *"Combining text- and link-based retrieval methods for Web IR"*, In The Ninth Text REtrieval Conf (TREC 9), 2001.

[13] S. Raghavan and H. Garcia-Molina, *"Crawling the hidden web"*, In Proceedings of VLDB '01, pp: 129–138, 2001.

[14] L. Page, S. Brin, R. Motwani, and T.Winograd. *"The pagerank citation ranking: Bringing order to the web"*, 1998.

[15] K. Somboonviwat, M. Kitsuregawa, and T. Tamura. *"Simulation Study of Language Specific Web Crawling"*, icde, 21st International Conference on Data Engineering (ICDE'05), p. 1254, 2005.

[16] B. Novak, *"A Survey of Focused Web Crawling Algorithms"*, SIKDD 2004 Multi-Conference IS 2004, pp: 12–15, 2004.

[17] K. Somboonviwat, T. Tamura, and M. Kitsuregawa, *"Finding thai web pages in foreign web spaces"*, In ICDE Workshops, p. 135, 2006.

[18] G. Botha and E. Barnards, *"Two approaches to gathering text corpora from the World Wide Web"*, In Proceedings of the 16th Annual Symposium of the Pattern Recognition Association of South Africa, Langebaan, South-Africa, p. 194, November, 2005.

[19] C. Castillo, *"Effective Web crawling"*, Ph.D. Thesis, University of Chile, Department of Computer Science, 2004.

[20] J. Kleinberg, *"Authoritative sources in a hyperlinked environment"*, In Proceedings ACM-SIAM Symposium on Discrete Algorithms, 1998, also appears as IBM Research Report RJ 10076(91892) and online at http://www.cs.cornell.edu/home/kleinber/auth.ps.

[21] G. Grefenstette,*"Comparing two language identification schemes"*, In Proceedings of JADT 1995, 3rd International Conference on Statistical Analysis of Textual Data, 1995

[22] W.B. Cavnar, J. M. Trenkle, *"N-gram-based text categorization", In Symposium on Document Analysis and Information Retrieval,* Las Vegas, pp:161-175, 1994.

[23] G. Churcher, *"Distinctive character sequences"*, personal communication by Ted Dunning, 1994

[24] A.H. Keyhanipour, A. Mohammad Zareh Bidoki, M. Mahmoudi, M. Azadnia, *"Evaluation of Iran's Web Content from e-Government perspective"*, 12th International CSI conference, 2007.

[25] E. Darrudi, M. R Hejazi, F. Oroumchian, *"Assessment of a Modern Farsi Corpus"*, In Proceedings of the 2nd Workshop on Information Technology & its Disciplines (WITID) 2004, ITRC, Kish Island, Iran.

[26] Ricardo Baeza-Yates. Carlos Castillo. Vicente Lopez. "*Characteristics of the Web of Spain*".

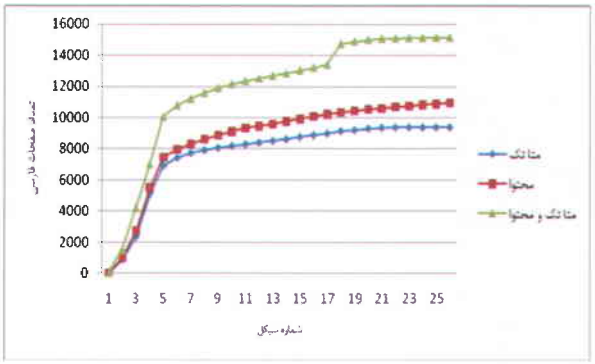
[27] Daniel Gomes. Mario J. Silva. "*A characterization of the portuguese web*". University of Lisbon. Potugal. 2004.

[28] R. Baeza-Yates and C. Castillo. Characterization of national web domains. Technical report, Universitat Pompeu Fabra, 2005.

[۲۹] امیر حسین کیهانی‌پور، علی‌محمد زارع بیدکی، مریم محمودی، محمد آزادنیا، "ارزیابی محتوای وب ایران از منظر دولت الکترونیک"، دوازدهمین کنفرانس انجمن کامپیوتر ایران

[30] http://ece.ut.ac.ir/dbrg/Hamshahri/

معصومه عظیم‌زاده فارغ‌التحصیل رشته مهندسی کامپیوتر گرایش نرم‌افزار از دانشگاه تربیت معلم تهران در سال ۱۳۸۰بوده و در سال ۱۳۸۵ مدرک کارشناسی ارشد خود را در همین رشته- گرایش از دانشگاه آزاد واحد تهران جنوب اخذ نموده است. فعالیتهای پژوهشی ایشان از سال ۱۳۸۰ در مرکز تحقیقات مخابرات آغاز گردیده و هم‌اکنون نیز در زمینه بازیابی اطلاعات از وب و بسترسازی مناسب برای زبان فارسی به همکاری خود با این مرکز ادامه می‌دهد.



نمودار ۶- تعداد صفحات فارسی خزش شده با بکارگیری شاخصهای مختلف خزش زبانی

مطابق با نمودار ۶ نرخ جمع‌آوری صفحات فارسی مبتنی بر اطلاعات فرابرچسب در مقایسه با بکارگیری سایر شاخصها از روند رشد کمتری برخوردار است. نتیجه حاصل از بکارگیری این شاخص مبتنی بر این واقعیت است که درصد قابل توجهی از صفحات فارسی فاقد فرابرچسب زبانی می‌باشند. استفاده از شاخص محتوا در منحنی دیگر این نمودار نشان‌دهنده بهبود نرخ جمع‌آوری صفحات فارسی نسبت به حالت قبل است. این موضوع نشان‌دهنده آن است که استفاده از ایست‌واژه‌ها می‌تواند منجر به شناسائی درصد بیشتری از صفحات فارسی شود.

نمودار ۷- نسبت تعداد صفحات فارسی به صفحات پارس‌شده با بکارگیری شاخصهای مختلف خزش زبانی

اگر منحنی مربوط به بکارگیری ترکیبی از شاخصهای فرابرچسب و محتوا را در نمودار ۶ ملاحظه کنید نشان‌دهنده آن است که استفاده از ترکیبی از این دو شاخص تاثیر بسزائی بر افزایش نرخ صفحات جمع‌آوری شده در سیکلهای مختلف خزش دارد. از جمله دلایل این بهبود آن است که در این حالت صفحات فارسی که در روش مبتنی بر محتوا به دلیل غنی نبودن محتوا یا ذخیره‌سازی بر اساس کاراکترستی غیر از Windows-1256 و UTF-8 شناسائی نشده بودند، مورد شناسائی قرار می‌گیرند. همچنین نمودار ۷ نشان‌دهنده تاثیر شاخصهای مختلف خزش بر پوشش بیشتر صفحات فارسی می‌باشد. مطابق این نمودار استفاده از ترکیبی از شاخصهای خزش زبانی منجر به افزایش نسبت صفحات فارسی جمع‌آوری شده به کل صفحات خزش شده می‌گردد.

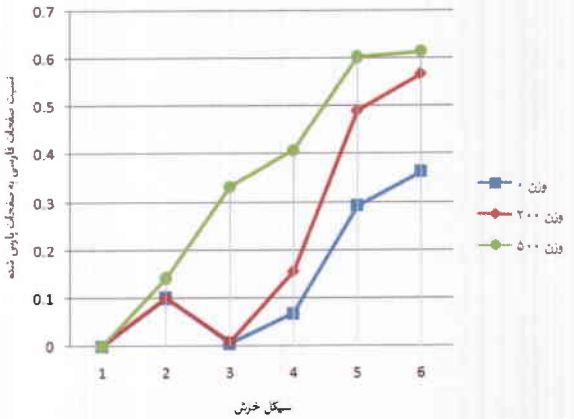
۶- نتیجه‌گیری

در این مقاله خزشگر فارسی با هدف بهبود خزش مستندات فارسی وب ارائه شد. با توجه به ارزیابی‌های صورت گرفته، مشخص شد بخش تشخیص زبان خزشگر فارسی با دقت بالائی عمل می‌کند. همچنین با تاثیر نتایج بخش شناسائی زبان بر اولویت‌دهی به واکشی صفحات مرتبط، ضمن اینکه صفحات فارسی با دقت بالائی جمع‌آوری می‌شوند، امکان پوشش بیشتر صفحات فارسی نیز فراهم می‌شود. در واقع خزشگر فارسی با ایجاد امکان جمع‌آوری صفحات فارسی از سایر دامنه‌های وب، میزان پوشش صفحات فارسی را افزایش می‌دهد. ویژگی دیگر خزشگر ارائه شده سرعت جمع‌آوری صفحات فارسی می‌باشد. در واقع خزشگر فارسی دارای این قابلیت است که در بازه زمانی کوتاهتری تعداد صفحات مرتبط بیشتری جمع‌آوری نماید. بنابراین در مجموع استفاده از خزشگر فارسی منجر به بهبود خزش مستندات فارسی می‌گردد. همچنین در آزمایشات صورت پذیرفته تاثیر انتخاب دسته‌های متفاوت از URLهای اولیه بر سرعت و پوشش خزش در سیکل‌های اولیه بررسی و تحلیل شده است. نتایج حاصله نشان‌دهنده آن است که در صورتیکه URLهای اولیه حاوی پیوندهای خروجی مناسبی به سایر صفحات فارسی باشند، در مراحل اولیه نیز سازوکار وزندهی صفحات فارسی نتیجه مطلوبی برای خزش دربردارد. در مجموع در خزشگر فارسی پیشنهادی انتخاب URLهای اولیه فارسی مناسب به تعداد زیاد، منجر به ایجاد شرایط اولیه مطلوبی جهت پوشش هرچه بیشتر صفحات فارسی وب می‌گردد. همچنین کارائی خزشگر فارسی پیشنهادی با روشهای مبتنی بر محتوا، مبتنی بر فرابرچسب یا ترکیب این دو مورد بررسی قرار گرفته و عملکرد آنها مقایسه شده است. این آزمایش نشان می‌دهد که با ترکیب ویژگیها در مولفه تشخیص زبان عملکرد خزشگر به میزان قابل توجه ای بهبود یافته است. با توجه به بهبود جمع آوری اطلاعات فارسی وب، خزشگر فارسی دارای کاربردهای مختلفی در سامانه‌های جستجو و بازیابی اطلاعات است. به عنوان نمونه می‌توان از آن در موتورهای جستجو و بازیابی اطلاعات فارسی استفاده کرد. در موتورهای جستجوی فارسی، علاوه بر تمرکز بر مستندات فارسی، سرعت بالای این خزشگر در جمع آوری مستندات، بروز رسانی اطلاعات خزش شده را تسریع می‌بخشد. این امر در ارزیابی محتوای فارسی وب نیز صادق بوده و امکان استفاده از این ابزار جهت ارزیابی محتوای فارسی وب بسیار مفید خواهد بود.

این خزشگر هنوز امکان استفاده از هستان‌شناسی و یا گنجیه زبان فارسی را نداشته و امکان تشخیص ابعاد تخصصی و موضوعی صفحات فارسی را ندارد. از جمله فعالیت‌هائی که در کارهای آینده این تحقیق به آن خواهیم پرداخت، توسعه خزشگر موضوعی زبان فارسی می‌باشد. این خزشگر قادر خواهد بود که در ضمن تشخیص زبان صفحات موضوع مرتبط با عنوان پرس و جو را نیز تشخیص داده و جمع آوری کند. به این منظور اولین سازوکار مورد نیاز شناسایی صفحات فارسی وب می‌باشد که در این مقاله ارائه شده است. بعد از شناسائی زبان صفحه، در مرحله بعد می‌بایست کلمات بی‌محتوا و اضافه را دور ریخته و محتوبات باقیمانده صفحه را با منابع زبانی موجود نظیر هستان‌شناس یا گنجینه واژگان مقایسه شود. در صورت مرتبط بودن صفحه با موضوع مورد نظر پیوندهای موجود در آن

۲) نتایج اجرای خزشگر با استفاده از URLهای دسته دوم:

نتایج اجرا با URLهای دسته دوم در نمودار ۴ نشان‌دهنده آن است که نسبت صفحات فارسی خزش شده به کل صفحات پارس شده از روند رشد بسیار بیشتری نسبت به نمودار مشابه (نمودار ۲) مربوط به اجرای با URLهای دسته اول برخوردار است. این مطلب تاثیر انتخاب URLهای اولیه در نحوه خزش وب را نشان می‌دهد. به عنوان مثال در شرایطی که خزشگر با وزن ۵۰۰ برای صفحات فارسی اجرا گردیده است، در سیکل پنجم از نمودار ۲ نسبت صفحات فارسی به صفحات پارس شده حدود ۰,۲ می‌باشد در حالیکه این نسبت برای شرایط مشابه در نمودار ۴ حدود ۰,۶ می‌باشد.

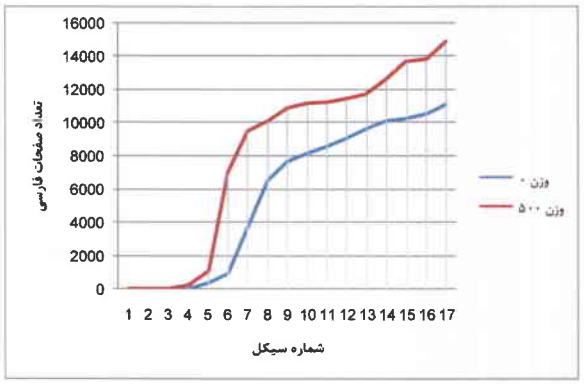


نمودار۴- نسبت تعداد صفحات فارسی به صفحات پارس‌شده براساس افزایش وزن در سیکلهای مختلف

مطابق نمودار ۴ خزشگر با وزن ۵۰۰ برای صفحات فارسی، از همان سیکلهای اولیه در مسیر صحیح خزش و یا در مسیر مناسب جهت خزش صفحات فارسی وب قرار گرفته است و همچنین از روند رشد بهتری نسبت به منحنی مشابه در نمودار ۲ برخوردار است. این امر بیانگر تاثیر انتخاب URLهای اولیه بر روند و نتایج خزش می‌باشد. لازم به یادآوری است که نتایج ارائه شده در نمودار ۴ بیانگر شرایطی است که URLهای اولیه از چندین دایرکتوری فارسی DMOZ انتخاب شده‌اند. همچنین در نمودار ۴ منحنی با وزن ۲۰۰ در سیکلهای اولیه خزش از روند رشد کمتری نسبت به منحنی با وزن ۵۰۰ برخوردار است که از جمله دلایل این امر می‌توان به تاثیر پارامترهای دیگر در اولویت‌دهی صفحات در مراحل اولیه خزش اشاره نمود. پارامتر عمق از جمله پارامترهائی است که در الویت‌دهی به صفحات نقش دارد که با پیمایش عمق بیشتری از صفحات تاثیر کمتری در دنبال نمودن آنها دارد. در واقع می‌توان گفت با توجه به اینکه در مراحل اولیه تعداد لینکهای غیرفارسی بیشتری در دایرکتوری وجود دارد، با تاثیر پارامتر عمق منحنی‌های موجود در نمودار برای وزنهای ۰ و ۲۰۰ در سیکلهای اولیه روند رشد مناسبی ندارند.

۳) نتایج اجرای خزشگر با استفاده از URLهای دسته سوم:

در این اجرا بازهم URLهای انتخاب شده باعث افزایش چشم‌گیر تعداد صفحات فارسی خزش شده نسبت به صفحات فارسی خزش شده در اجرای با URLهای دسته اول می‌باشد. در نمودار ۵ نیز تاثیر انتخاب URLهای اولیه در روند رشد منحنی‌ها دیده می‌شود. با توجه به اینکه URLهای فارسی مورد استفاده در این وضعیت، منجر به دسترسی به سایتهای با درجه پیوند خروجی قابل توجهی می‌شوند، این امر منجر شده که منحنی‌های موجود در این نمودار نسبت به منحنی‌های موجود در نمودار ۱ خصوصا از سیکل ۶ به بعد روند رشد بسیار بیشتری داشته باشند. نکته‌ای که در رابطه با نمودار ۵ وجود دارد آن است که در سیکلهای اولیه روند رشد نمودار کندتر از نمودار ۱ است که دلیل این موضوع را می‌توان در تعداد URLهای اولیه مورد استفاده ذکر نمود.



نمودار ۵- تعداد صفحات فارسی خزش شده در سیکلهای مختلف بر اساس افزایش وزن

نکته مهم دیگری که در رابطه با نمودارهای ارائه شده قابل ذکر است آن است که ماهیت متغییر خزش مانع از رشد خطی نمودارها می‌گردد. در واقع تعداد صفحات فارسی واکشی شده در هر سیکل خزش بسته به عوامل مختلفی از جمله درصد پیوندهای واکشی شده در سیکل ماقبل دارد و همین موضوع موجب می‌شود که شیب نمودار در هر سیکل دچار قدری افزایش یا کاهش شود. از طرفی با توجه به گسترده شدن گراف جستجو همزمان با روند پیشرفت خزش در مجموع تعداد صفحات فارسی یافته شده افزایش می‌یابد.

جهت بررسی سرعت جمع‌آوری صفحات مرتبط توسط خزشگر فارسی، تعداد صفحات فارسی شناسائی شده توسط وایر و خزشگر فارسی در دو اجرای مختلف مورد بررسی قرار گرفتند. نتایج اجرا با استفاده از خزشگر وایر نشان داد که در انتهای سیکل ۳۲ از مجموع ۲۸۷۸۷ صفحه پارس شده، تعداد ۴۵۲۸ صفحه فارسی شناسائی شد. در حالیکه با استفاده از خزشگر فارسی در انتهای سیکل ۱۶ تعداد ۶۰۰۰ صفحه فارسی از بین ۲۸۸۳۱ صفحه شناسائی شدند. از نتایج حاصله می‌توان استنباط نمود که خزشگر ارائه شده جهت جمع‌آوری صفحات فارسی دارای سرعتی حداقل ۲ برابر خزشگر وایر است. بنابراین خزشگر فارسی می‌تواند با سرعت بسیار بیشتری نسبت به خزشگر وایر صفحات فارسی را جمع‌آوری نماید.

۴) بررسی تاثیر بکارگیری ویژگیهای مختلف صفحات وب بر کارائی خزش:

جهت نشان دادن تاثیر استفاده از ویژگیهای مختلف صفحات وب بر کارائی خزش، خزشگر با بکارگیری URLهای اولیه دسته دوم با وزن ۵۰۰ و به مدت ۲۵ دور اجرا گردیده است. شاخصهای زبانی مورد استفاده فرابرچسب، محتوا و ترکیبی از این دو شاخص بوده و نتایج در نمودارهای ۶ و ۷ ارائه شده است.

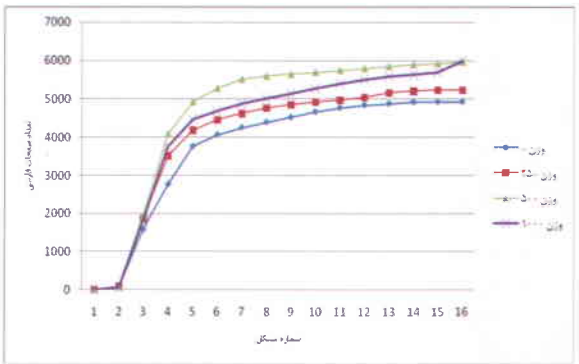
www.sku.ac.ir/fa/academic/members/golestanian/main.htm
www.honar.ac.ir/farhangestan/structure.htm
www.irandoc.ac.ir/library/copy.htm
www.iranculture.org/contactus/index.php
www.bawar.ir/

همانگونه که ملاحظه می‌شود URLهای ارائه شده از دامنه‌های مختلف ir.، com. و org. جمع‌آوری شده‌اند که البته امکان شناسائی دامنه‌های دیگری مانند net. و gov. نیز وجود دارد. در حالیکه در کارهای مرتبطی که از خزشگر وایر استفاده شده مانند [29]، ملاک شناسائی صفحات فارسی تنها تعلق آنها به دامنه ir. بوده است. این عامل موجب می‌شود که بخش قابل توجهی از صفحات فارسی وب نادیده گرفته شود. به عنوان مثال در نمونه بررسی شده تعداد ۴۸ پیوند متعلق به دامنه com. و ۵۰ پیوند متعلق به دامنه ir. بود. همچنین با توجه به اینکه خزشگر در حین عملیات خزش به صفحات فارسی وزن بیشتری می‌دهد، در مراحل بعدی خزش، لینکهای موجود در صفحات فارسی وزن بیشتری دریافت می‌کنند. بنابراین انتظار می‌رود با تاثیر این وزن در سیستم وزندهی خزشگر وایر، نرخ جمع‌آوری صفحات فارسی افزایش یابدکه برای سه دسته URLهای فوق به شکل زیر مورد بررسی قرار گرفته است.

۱) نتایج اجرای خزشگر با استفاده از URLهای دسته اول

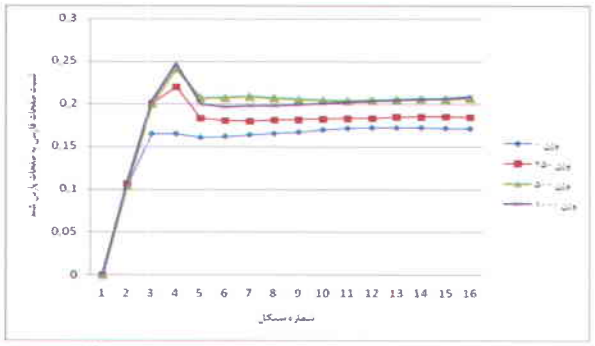
نتایج اجرای خزشگر با دسته اول URLهای اولیه به ازای وزنهای مختلف برای صفحات فارسی در نمودار ۱ نشان داده شده است. ستون افقی نشان دهنده شماره سیکل‌های خزش و ستون عمودی نشان دهنده میزان صفحات فارسی جمع‌آوری شده می‌باشد.

همانگونه که ملاحظه می‌شود در حالتیکه وزن صفحات فارسی با صفر مقداردهی شده است، نمودار رشد منحنی از روند کندتری برخوردار است و با افزایش وزن صفحات فارسی روند رشد نمودار بهبود می‌یابد. نکته قابل توجه در این نمودار این است که افزایش وزن صفحات متعلق به زبان تا حد مشخصی می‌تواند منجر به بهبود نتایج خزش گردد و زمانیکه مقدار آن از حد مشخصی تجاوز کند، با توجه به اینکه تاثیر آن در مقابل تاثیر سایر پارامترها در اولویت‌دهی به صفحات به صورت قابل توجهی بیشتر شده و به حد اشباع رسیده است، این افزایش وزن در نتایج نهائی خزش تغییر قابل توجهی ایجاد نخواهد کرد. به عنوان مثال همانگونه که در نمودار ملاحظه می‌گردد با افزایش وزن صفحات فارسی به مقدار ۱۰۰۰نتایج بدست آمده در سیکلهای مختلف خزش نسبت به وضعیتی است که خزشگر با مقدار ۵۰۰ وزندهی شده است، دستخوش تغییراتی چندانی نگردیده است.



نمودار۱-تعداد صفحات فارسی خزش شده در سیکلهای مختلف بر اساس افزایش وزن

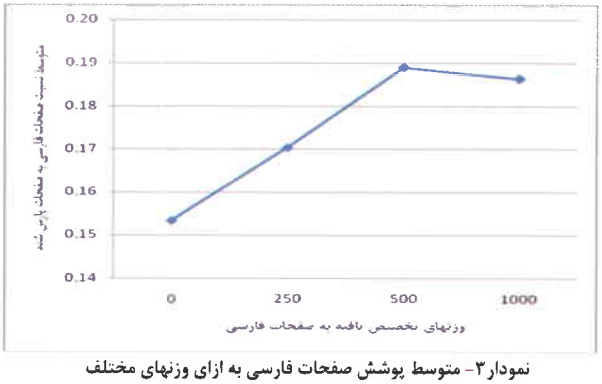
تعیین حداکثر مقدار پارامتر وزندهی به صفحات فارسی وابسته به شرایط اجرا و پارامترهای مختلفی است که در وزن‌دهی به صفحه موثر هستند که از جمله این پارامترها می‌توان به پارامترهای عمق و رندوم اشاره نمود. بنابراین بسته به تنظیمات اولیه خزشگر و شرایط اجرای مختلف، حد به اشباع رسیدن مقدار پارامتر وزن‌دهی به صفحات متعلق به زبان متفاوت می‌باشد که در این مقاله این پارامتر با مقدار ۵۰۰ به حد اشباع رسیده است.



نمودار۲- نسبت تعداد صفحات فارسی به صفحات پارس‌شده براساس افزایش وزن در سیکلهای مختلف

همچنین نمودار ۲ نسبت صفحات فارسی به صفحات پارس شده را نشان می‌دهد. همانطور که از نمودار مشخص است، با افزایش وزن صفحات فارسی، میزان صفحات فارسی خزش شده به کل صفحات پارس شده نیز افزایش یافته است. این موضوع می‌تواند نشاندهنده افزایش پوشش صفحات فارسی وب باشد. همانطور که نمودارهای ۱ و ۲ نشان می‌دهند در اجرای خزشگر با URLهای دسته‌اول، منحنی‌ها در سیکلهای اولیه از روند رشد نسبتا یکسانی برخوردار هستند و پس از آن نحوه تاثیرگذاری وزن صفحات فارسی در بهبود خزش مستندات فارسی دیده می‌شود. دلیل این امر را می‌توان در تاثیر انتخاب URLهای اولیه در نتایج خزش دانست. در واقع با توجه به اینکه در این حالت URLها ترکیبی از URLهای مربوط به دامنه کشورهای مختلف بوده و از نظر تعداد نیز قابل توجه هستند، تاثیر وزندهی به صفحات بعد از پیمایش عمق مشخصی از صفحات وب دیده می‌شود.

نمودار ۳ متوسط نرخ پوشش صفحات فارسی را به ازای وزنهای مختلف نشان می‌دهد. همانگونه که در نمودار ملاحظه می‌شود. از نقطه ۵۰۰ به بعد افزایش وزن صفحات تاثیر چندانی بر متوسط نرخ صفحات جمع‌آوری شده فارسی نسبت به کل صفحات پارس شده نداشته است.



نمودار۳- متوسط پوشش صفحات فارسی به ازای وزنهای مختلف

مقایسه با خزشگر وایر بدون تاثیر زبان، وزن صفحات فارسی را از صفر تا ۵۰۰ تغییر می‌دهیم. در حالتی که خزشگر فارسی با وزن صفر در واقع نشان‌دهنده عملکرد وایر بدون تاثیر اجزاء فارسی اضافه شده در این تحقیق می‌باشد. در ضمن برای پارامتر عمق وزن ۹۵، پارامتر رندوم وزن ۵ و برای سایر پارامترها وزن ۰ در نظر گرفته شده است.

پارامتر دیگر حداکثر تعداد صفحات و سایتهای قابل خزش می‌باشد که جهت اجرای خزش حداکثر ۶۰۰۰۰ صفحه به ازاء هر سایت و حداکثر ۶۰۰۰ سایت برای خزش در نظر گرفته شده است. یکی از پارامترهای مهم دیگری که جهت ارزیابی عملکرد خزشگر فارسی مقداردهی شده است عبارت است از دامنه‌هائی که خزشگر امکان خزش صفحات آنها را دارد. با توجه به اینکه خزشگر وایر جهت خزش دامنه کشورها طراحی شده است در کارهای انجام شده عموما به دامنه اصلی هر کشور محدود می‌شود [26و27]. در برخی تحقیقات صورت پذیرفته نظیر [28] به این موضوع پرداخته شده است که محدود بودن خزشگر به دامنه کشور منجر به از دست دادن بخش زیادی از صفحات مربوط به دامنه یک کشور می‌شود. با توجه به اینکه از اهداف این تحقیق ایجاد امکان پوشش صفحات فارسی موجود در سایر دامنه‌های وب می‌باشد، در پیکربندی خزشگر علاوه بر دامنه وب ایران (ir.) دامنه‌های com. ، org.، net. و gov. نیز در نظر گرفته شده‌اند تا امکان بررسی عملکرد خزشگر در پوشش صفحات فارسی موجود در سایر دامنه‌ها فراهم گردد. همچنین جهت تعیین میزان تمایل خزشگر فارسی به سمت دامنه وب ایران، دامنه‌های کشورهای ژاپن (jp.)، آلمان (de.)، آرژانتین (ar.)، انگلستان (uk.) و آلمان (de.) نیز در فایل پیکربندی مقداردهی شده است.

۵-۲-۳ ارزیابی خزشگر و نتایج به‌دست آمده

جهت ارزیابی عملکرد خزشگر سیستم عامل Linux Fedora Core 8 مورد استفاده قرار گرفت. این سیستم عامل روی سرور Intel با حجم ۳ ترابایت دیسک سخت و حافظه ۴ گیگابایت نصب شد.

با توجه به اینکه عملکرد خزشگر تا حدود زیادی به لیست URLهای اولیه وابسته می‌شود، جهت ارزیابی عملکرد خزشگر فارسی از سه دسته URLهای اولیه استفاده شده است. در ادامه توضیح هریک از لیست URLهای اولیه آمده است.

- *آزمون بر اساس URLهای دسته اول:*

برای هر تست وایر تعداد ۱۶ دور سیکل خزش در نظر گرفته شده است. که هر تست حدود ۳ ساعت زمان به خود اختصاص می‌دهد. مشخصات سه اجرای انجام گرفته به ازاء مقادیر مختلف برای پارامتر وزن صفحات فارسی در جدول ۲ خلاصه بطور خلاصه آمده‌است.

جدول-۲ مشخصات اجراهای انجام گرفته برای ارزیابی خزشگر فارسی به ازاء دسته اول laseed

	اجرای اول	اجرای دوم	اجرای سوم
وزن صفحات فارسی	۰	۲۵۰	۵۰۰
تعداد دور خزش	۱۶	۱۶	۱۶
مدت زمان اجرا	۲:۳۰ ساعت	۲:۳۰ ساعت	۲:۳۰ ساعت
تاریخ	۷ فوریه	۷ فوریه	۷ فوریه
تعداد صفحات پارس شده	۲۸۷۷۷	۲۸۲۹۳	۲۸۸۳۱
تعداد صفحات فارسی	۴۹۲۶	۵۲۳۳	۵۹۶۸

- *آزمون بر اساس URLهای دسته دوم:*

برای هر تست وایر تعداد ۶ دور سیکل خزش در نظر گرفته شده است. که هر تست حدود ۱:۳۰ ساعت زمان به خود اختصاص می‌دهد. مشخصات سه اجرای انجام گرفته به ازاء مقادیر مختلف برای پارامتر وزن صفحات فارسی در جدول ۳ بطور خلاصه آمده است.

جدول-۳ مشخصات اجراهای انجام گرفته برای ارزیابی خزشگر فارسی به ازاء دسته دوم laseed

	اجرای اول	اجرای دوم	اجرای سوم
وزن صفحات فارسی	۰	۲۵۰	۵۰۰
تعداد دور خزش	۶	۶	۶
مدت زمان اجرا	۱:۳۰ساعت	۱:۳۰ساعت	۱:۳۰ساعت
تاریخ	۱۵ فوریه	۱۵ فوریه	۱۵ فوریه
تعداد صفحات پارس شده	۱۲۷۵۰	۱۵۵۰۰	۱۰۰۶۶
تعداد صفحات فارسی	۴۴۸۲	۸۰۰۸	۵۹۶۸

- *آزمون بر اساس URLهای دسته سوم:*

برای هر تست وایر تعداد ۱۶ دور سیکل خزش در نظر گرفته شده است. که هر تست حدود ۳ ساعت زمان به خود اختصاص می‌دهد. مشخصات دو اجرای انجام گرفته به ازاء وزن های ۰ و ۵۰۰ برای پارامتر وزن صفحات در جدول ۴ بطورخلاصه آمده است.

جدول-۴ مشخصات اجراهای انجام گرفته برای ارزیابی خزشگر فارسی به ازاء دسته سوم laseed

	اجرای اول	اجرای دوم
وزن صفحات فارسی	۰	۵۰۰
تعداد دور خزش	۱۶	۱۶
مدت زمان اجرا	۳ساعت	۳ساعت
تاریخ	۱۹ فوریه	۱۹ فوریه
تعداد صفحات پارس	۲۴۸۰۹	۲۲۷۸۳
تعداد صفحات فارسی	۱۱۰۹۵	۱۴۸۸۶

جدول ۵ تعداد صفحات وب جمع‌آوری شده، درصد صفحات تکراری جمع‌آوری شده و همچنین درصد صفحات پویا و ایستای خزش شده را نشان می‌دهد. لازم به ذکر است صفحات پویا به صفحاتی گفته می شود که توسط تکنولوژی‌های برنامه نویسی ایجاد می‌شوند و دارای پسوندهایی مانند asp, .aspx, .php هستند اما صفحات ایستا دارای پسوندهایی مانند html.، .htm هستند. پس از اجرای خزشگر در اجرای دوم از جدول ۲، در کل ۳۶،۵۶۸ صفحه وب توسط خزشگر پردازش شد که از بین ۲۸۲۹۳ صفحه پارس شده تعداد ۵۲۳۳ صفحه فارسی بود.

جدول-۵ آمار صفحات وب خزش شده

تعداد کل صفحات	۳۶،۵۶۸ صفحه	
تعداد صفحات یکتا	۳۵،۳۲۷ صفحه	۹۶،۶۱ درصد
تعداد صفحات تکراری	۱،۲۴۱ صفحه	۳،۳۹ درصد
تعداد صفحات ایستا	۲۲،۴۰۲ صفحه	۶۱،۲۶ درصد
تعداد صفحات پویا	۱۴،۱۶۶ صفحه	۳۸،۷۴ درصد

برای بررسی صحت عملکرد بخش تشخیص زبان فارسی، تعداد ۱۰۰ لینک از بین ۵۲۳۳ لینک فارسی شناسائی شده در اجرای دوم از جدول ۲ بصورت دستی بررسی شدند که همه به درستی شناسائی شده بودند. بخش کوچکی از لینک‌هایی که صفحات آنها به عنوان فارسی شناسائی شدند عبارتند از:

www.farsnews.com/newstext.php?nn=8711290500
www.iust.ac.ir/printme-1.3346.6386.fa.html

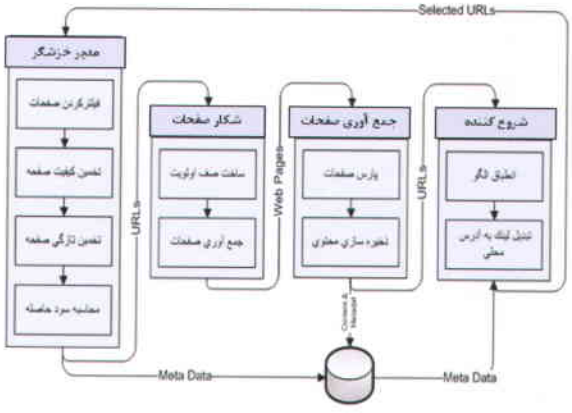
تاثیرگذاری زبان مستندات در وزن صفحات می بایست کلیه وزنها و به ویژه وزن زبان صفحات در فایل پیکربندی خزشگر تنظیم گردد.

۵-پیاده‌سازی و ارزیابی خزشگر زبانی پیشنهادی

۵-۱ پیاده‌سازی

در این بخش شرح پیاده‌سازی چارچوب پیشنهادی در قالب یک خزشگر زبان فارسی آمده است. با توجه به اینکه طراحی یک خزشگر جهت خزش در ابعاد وسیع، دارای پیچیدگیها و چالشهای خاص خود می‌باشد، بنابراین بکارگیری یک خزشگر متن‌باز راه‌حل مناسبتری به نظر می‌رسد. در این راستا جهت پیاده‌سازی خزشگر زبان فارسی، از خزشگر متن‌باز وایر استفاده شد. این خزشگر از بین خزشگرهای متن‌باز موجود به دلیل دارا بودن ویژگیهائی نظیر مبنای علمی قوی، وجود مستندات کافی و بکارگیری آن در تحقیقات مختلف مانند [19]، انتخاب شده است.

معماری این خزشگر شامل چهار بخش اصلی مدیریت[20]، شکار[21]، جمع‌آوری[22] و شروع کننده[23] است که در یک سیکل بصورت گردشی اجرا می‌شوند و در هر سیکل نتایج حاصل از سیکل قبلی استفاده می‌شود. به عبارت دیگر بعد از هر سیکل مجموعه‌ای از پیوندها استخراج شده و به عنوان URLهای قابل واکشی برای سیکل بعدی استفاده می‌شوند. ارتباط بخشهای مختلف خزشگر و مولفه‌های تشکیل‌دهنده آنها در شکل ۳ نمایش داده شده است. با توجه به اینکه در خزشگر وایر ارتباط بین بخشهای مختلف خزشگر توسط منبع ذخیره‌سازی متاداده برقرار می‌شود، بعد از تشخیص زبان صفحات مقداری در بخش منبع ذخیره‌سازی متاداده ذخیره می‌شود تا این موضوع به اطلاع سایر بخشهای خزشگر نیز برسد.



شکل ۳- معماری خزشگر وایر

در طرح پیشنهادی پس از تشخیص زبان صفحات در بخش جمع‌آوری صفحات، با توجه به معماری خزشگر وایر، اولویت دهی به صفحات فارسی از طریق تغییر فرمول وزندهی در بخش مدیریت صورت می‌پذیرد که در این راستا فعالیتهای زیر صورت پذیرفته است:

۱- اضافه کردن یک ساختمان داده به ابرداده‌های مربوط به صفحات در بخش ذخیره سازی ابرداده.

۲- تاثیر وزندهی به صفحات فارسی در فرمول وزن‌دهی جهت اولویت‌دهی به واکشی این صفحات. اضافه نمودن پارامتر وزندهی در فایل پیکربندی خزشگر.

۳- مقدار دهی به ضرایب وزندهی بر اساس معیارهای مختلف.

۵-۲ ارزیابی

جهت ارزیابی عملکرد خزشگر فارسی معیارهای صحت، پوشش و سرعت خزش مستندات فارسی وب مورد بررسی قرار گرفته‌اند. با توجه به اهمیت بخش شناسائی زبان ابتدا باید صحت عملکرد آن در شناسائی صفحات فارسی تعیین شود که جهت ارزیابی آن صفحات وب شناسائی شده توسط خزشگر فارسی به صورت دستی مورد ارزیابی قرار گرفتند. معیار مهم دیگر معیار پوشش می‌باشد. از جمله شاخصهای ارزیابی معیار پوشش میزان صفحات مرتبط خزش شده و شاخص دیگر امکان شناسائی و جمع‌آوری صفحات فارسی در سایر دامنه‌ها می‌باشد. سرعت خزش مستندات فارسی وب معیار سومی است که جهت ارزیابی عملکرد خزشگر مد نظر قرار گرفته است. در واقع به کمک این فاکتور زمان مورد نیاز جهت جمع‌آوری صفحات فارسی از وب را مبتنی بر سیکلهای خزش مشخص می‌کنیم. در ادامه شرایط اولیه، نحوه آماده‌سازی، شرایط اجرا و ارزیابی نتایج خزشگر فارسی ارائه خواهد شد.

۵-۲-۱ انتخاب URLهای اولیه

جهت ارزیابی عملکرد خزشگر لازم است آن را تحت شرایط اولیه متفاوتی ارزیابی کرد؛ زیرا شرایط اولیه می‌تواند تاثیر زیادی در عملکرد خزشگر داشته باشد. به همین دلیل چندین دسته URL اولیه برای خزشگر در نظر گرفته شده است. منبع انتخاب این URLها دایرکتوری DMOZ[24] می‌باشد که یک دایرکتوری مرجع حاوی دسته‌بندی‌های متنوع از آدرس وب‌سایت-های مربوط به کشورهای مختلف می‌باشد و بصورت دستی بروز رسانی می‌شود.

- URLهای اولیه دسته اول: ارزیابی عملکرد خزشگر مستلزم ایجاد شرایط اولیه‌ای است که تمایل خزشگر به سمت صفحات فارسی را در مقایسه با صفحات زبانهای دیگر نشان دهد. به همین منظور URLهای اولیه از پنج زبان مختلف ژاپنی، آلمانی، اسپانیائی، انگلیسی و فارسی و به ازای هر زبان ۱۵۰ URL انتخاب شد.
- URLهای اولیه دسته دوم: چندین دایرکتوری فارسی از DMOZ
- URLهای اولیه دسته سوم: ترکیبی از URLهای فارسی و انگلیسی

۵-۲-۲ پیکربندی خزشگر

خزشگر وایر دارای یک فایل پیکربندی تحت عنوان WIRE.conf می‌باشد که در این فایل تنظیمات اولیه خزشگر مقداردهی می‌شود.

جهت وزندهی به صفحات فارسی به این فایل پارامتری به نام CONF_MANAGER_SCORE_PERSIAN_WEIGHT اضافه شده است که می‌توان جهت تعیین مقدار وزن مناسب برای صفحات فارسی آن را با مقادیر مختلفی ارزش‌دهی نمود. جهت بررسی عملکرد خزشگر فارسی در

20 Manager
21 Harvester
22 Gatherer
23 Seeder

[24] DMOZ Open Directory: www.dmoz.org

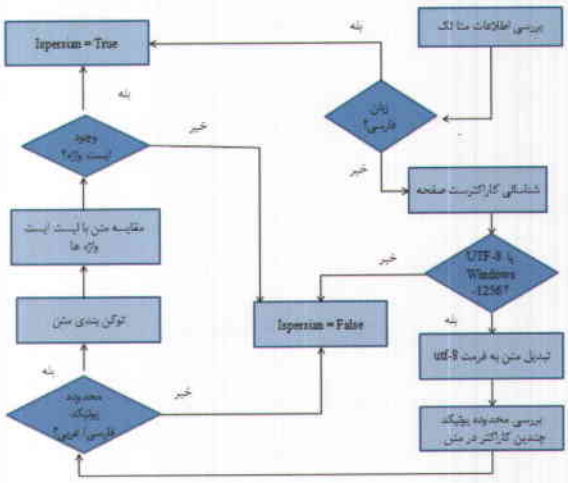
مستندات پس از تجزیه آنها در بخش پارسر انجام می‌شود. ورودی مولفه شناسائی زبان محتوای صفحات وب است. در این بخش پس از بررسی زبان صفحات، آنها به عنوان فارسی یا غیرفارسی علامتگذاری می‌شوند. بخشهای محتوائی مورد استفاده دراین طرح عبارتند از فرابرچسب زبانی، کاراکترست و محتوای صفحات. عملیات تجزیه و پردازش محتوای صفحات در چند مرحله و به کمک شاخصهای زیر صورت می‌پذیرد:

۱. فرابرچسب زبانی صفحه
۲. کاراکترست صفحات
۳. محدوده یونیکد صفحات
۴. ایست واژه‌ها

مولفه تشخیص زبان در مرحله اول به بررسی فرابرچسب زبانی صفحه می‌پردازد. در صورتیکه تگ زبانی صفحه فارسی باشد، صفحه به عنوان فارسی شناسائی می‌شود. اگر زبان صفحه به کمک اطلاعات فرابرچسب شناسائی نشده باشد کاراکترست صفحه مورد بررسی قرار می‌گیرد. با توجه به اینکه کاراکترست صفحات فارسی عموما Windows-1256 یا utf-8 می‌باشد، می‌توان از این مشخصه به عنوان یک شاخص اولیه جهت شناسائی صفحات فارسی استفاده نمود. سپس محتوای صفحات با کاراکترست Windows-1256، به فرمت utf-8 تبدیل می‌شوند.

با توجه به اینکه ممکن است برخی از بخشهای محتوائی صفحه خالی از محتوا بوده و یا محتوای موجود در آنها به زبان فارسی نباشد، محدوده یونیکد چند کاراکتر موجود در متن مورد بررسی قرار می‌گیرد. در صورتیکه کاراکترها در محدوده یونیکد فارسی/عربی باشند، محتوای متنی توکن‌بندی شده و با ایست واژه‌های مورد نظر مقایسه می‌شود. بر اساس وجود ایست واژه‌های فارسی صفحه به عنوان فارسی شناسائی می‌شود. برای توکن‌بندی متن از روش توکن‌بندی مبتنی بر مولفه موجود با معیار کاراکترهای جداکننده نظیر ’ ‘، ’\r‘ ، ’?‘ ،’\n‘ و ’.‘ استفاده شده و کاراکترهای جداکننده فارسی نظیر ’،‘ و ’؟‘ نیز در نظر گرفته شده است.

در نهایت جهت مشخص نمودن صفحات فارسی شناسائی شده یک متغیر بولین به نام ispersian مورد استفاده قرار گرفته است که در متغیر reserved متادیتای صفحه ذخیره می‌گردد تا امکان دسترسی به آن توسط سایر بخشهای خزشگر وجود داشته باشد.



شکل-۲ سازوکار شناسائی زبان صفحات وب

ایست‌واژه‌های[19] مورد استفاده در این مقاله از یک فهرست اولیه شامل ۱۰۰۰ ایست واژه از سایت پیکره همشهری [30] آزمایشگاه DBRG دانشگاه تهران واکشی شده و طی چندین مرحله پردازش نهایی شده است. در مرحله اول از پردازش بخشی از کلمات و واژگانی که تداول کمتری داشتند حذف گردیده‌اند. در مرحله بعد با توجه به اینکه زبان فارسی و عربی دارای اشتراک لغوی زیادی هستند بخشی از ایست‌واژه‌ها که بین دو زبان مشترک بودند حذف شده‌اند. سپس محتوای صفحاتی که به عنوان فارسی شناسائی شده بودند بررسی شده و در صورتیکه صفحه مذکور متعلق به زبان عربی بود و مبتنی بر ایست واژه‌های فراهم شده شناسائی شده باشد، ایست‌واژه‌های مذکور از لیست ایست‌واژه‌های مورد استفاده در این مقاله حذف گردیده است. بطور کلی جهت تسریع در شناسائی زبان صفحات سعی شده است که تعداد مناسبی ایست واژه که دارای فرکانس تکرار زیادی در صفحات فارسی باشند انتخاب شوند. لازم به ذکر است که لیست ایست‌واژه-های مذکور با فرمت UTF-8 ذخیره‌سازی گردیده است. ساز و کار شناسائی صفحات فارسی وب در شکل ۲ ارائه گردیده است.

۴-۲ مولفه وزن دهی به صفحات

در طرح پیشنهادی فرض اولیه مبتنی بر این است که لینکهای موجود در صفحات مرتبط یا فارسی زبان به صفحات فارسی اشاره دارند. بنابراین بعد از شناسائی صفحات مرتبط یا فارسی زبان باید لینکهای موجود در این صفحات به لیست لینکهای قابل دانلود اضافه شوند.

جایگاه مولفه وزن‌دهی در خزشگر در شکل ۱ نشان داده شده است. همانطور که در شکل مذکور مشاهده می‌گردد، پس از واکشی صفحات می‌بایست تشخیص داده شود که آنها به زبان فارسی هستند یا خیر. سپس لینک‌های موجود در کل صفحات جمع‌آوری شده را استخراج کرده و قبل از اینکه لینک‌های استخراج شده از صفحات را به لیست لینک‌های قابل واکشی اضافه کند، وزن مناسبی به آنها اختصاص داده شود. به طوری که لینک‌های استخراج شده از صفحات فارسی با احتمال بیشتری خزش شوند. بنابراین برای تغییر سازوکار اولویت‌بندی صفحات می‌بایست سازوکار وزندهی به صفحات متعلق به زبان را در فرمول وزن‌دهی در نظر گرفت. فرمول وزن دهی مورد استفاده به شرح زیر است:

URL Weight = *PAGERANK_WEIGHT* pagerank + WLSCORE_WEIGHT * pagerank + HITS_HUB_WEIGHT * hubrank +HITS_AUTHORITY_WEIGHT * authrank + SITERANK_WEIGHT * siterank + QUEUESIZE_WEIGHT * queuesize + DEPTH_WEIGHT* depth_score + STATIC_WEIGHT* static_score + RANDOM_WEIGHT * random_score + PERSIAN_WEIGHT * ispersian;*

در فرمول وزن دهی صفحات، وزن رتبه بندی صفحات نظیر *pagerank*، *hubrank*، *authrank* و *siterank* ، وزن مربوط به صف، عمق، ایستایی و رندم صفحات در نظر گرفته شده است که از فایل تنظیمات خزشگر خوانده می‌شوند. جهت دخیل نمودن زبان صفحات در فرمول وزن دهی صفحات ضریبی تحت عنوان *PERSIAN_WEIGHT* و معیاری تحت عنوان *ispersian* که فارسی بودن یا نبودن زبان صفحات را مشخص می‌کند، در نظر گرفته شده‌اند. بنابراین برای تنظیمات لازم در خصوص میزان

[19] Stop Words

جدول۱- مقایسه روشهای خزش بر اساس ویژگیهای خزش

شاخصهای خزش / روشهای خزش متمرکز	اطلاعات مورد استفاده جهت بررسی مرتبط بودن صفحات مورد اشاره پیوندها			اطلاعات مورد استفاده جهت بررسی مرتبط بودن صفحات واکشی شده						استفاده از پارامتر عمق جهت مواجه با صفحات نامرتبط
	متن پیوند	متن اطراف پیوند	وزنی	هستان شناسی یا کنجیه واژگان	مقایسه صفحه با عبارت پرس و جو	شمارش پیوند	کلمات کلیدی	متا تگ	اطلاعات بازخوردی	
PageRank [14]			●			●				
Fish Search [9]							●			●
Shark Search [10]	●	●	●		●					●
Focused Crawling [4]				●						
Learnable Crawling [8]	●		●				●		●	
Language Specific Web Crawling [15]								●		●

بر اساس ویژگیهای مستخرج از روشهای خزش متمرکز و نیازمندیهای خزش زبانی، شاخصهای مناسب دربردارنده داده‌ها و سازوکار مناسب جهت شناسائی صفحات متعلق به زبان مورد نظر می‌باشد. شاخصهای پایه مستخرج عبارتند از:

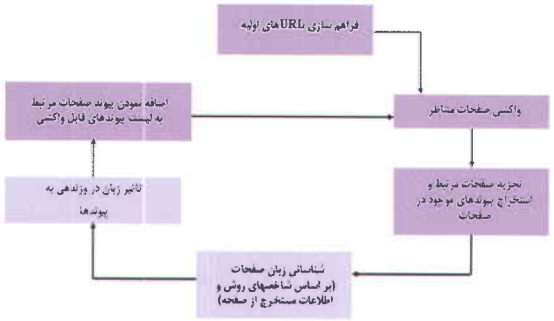
- اطلاعات فرابرچسب
- داده‌های محتوائی یا کلمات زبان

یکی از شاخصهای مناسب جهت شناسائی زبان صفحات اطلاعات مستخرج از فرابرچسب‌ها می‌باشد. با توجه به اینکه فرابرچسب‌ها به سرعت قابل شناسائی هستند و در بسیاری از صفحات وب وجود دارند، به عنوان معیاری مناسب جهت شناسائی زبان محسوب می‌شوند. همچنین چون شناسائی زبان وابستگی زیادی به داده‌های زبانی دارد، استفاده از اطلاعات محتوائی صفحات شامل شاخصهای متن پیوند و محتوای صفحات نیز روش مناسبی جهت شناسائی زبان صفحات می‌باشد. علاوه بر انتخاب شاخصهای پایه جهت خزش زبانی، استفاده از شاخص بازخورد یا یادگیری نیز می‌تواند منجر به افزایش میزان پوشش خزش شود. در واقع می‌توان ترکیبی از شاخصهای ذکر شده را برای خزش زبانی مورد استفاده قرارداد.

۴- چارچوبی برای طراحی یک خزشگر زبانی

در این بخش با توجه به سیاستگذاری واکشی صفحات وب، شاخصها و روش شناسایی زبان صفحات وب که در بخش‌های قبل شرح داده شد، روشی برای خزش مبتنی بر زبان برای زبان فارسی پیشنهاد شده است. همانطوریکه در دیاگرام شکل ۱ آمده است، در خزشگرهای زبانی دو مولفه اصلی وجود دارد که در این مقاله آنها را مولفه تشخیص زبان و مولفه وزن‌دهی به صفحات نامیده‌ایم. در واقع با استفاده از این دو مولفه یک خزشگر امکان اولویت دادن به صفحات با زبان خاص (در این مقاله زبان فارسی) را داشته و می‌تواند لینکهای مرتبط را شناسایی و دنبال کند.

در ادامه این مقاله، تحلیل و طراحی مولفه‌های مذکور برای یک خزشگر زبان فارسی آمده است.



شکل-۱ طرح کلی یک خزشگر زبانی

۴-۱- مولفه تشخیص زبان پیشنهادی

از آنجاییکه این مولفه در عملکرد صحیح خزشگر تاثیر زیادی دارد، از مولفه‌های کلیدی خزشگر زبانی محسوب می‌شود. همانطور که در بخش قبل نیز اشاره شد، روشهای مختلفی جهت شناسائی زبان صفحات وب وجود دارد. در این بخش با بکارگیری ترکیبی از شاخصهای خزش زبانی روشی برای تشخیص زبان فارسی در خزشگرها ارائه گردیده است. با توجه به اینکه شناسائی زبان صفحات به صورت پویا صورت می‌پذیرد باید روشی انتخاب شود که بتوان با سرعت بالایی زبان صفحات را شناسائی نمود. همچنین جهت شناسائی می‌توان از مشخصه‌هائی استفاده نمود که در بیشتر صفحات وب قابل ردیابی باشند. اطلاعات فرابرچسب به دلیل امکان تشخیص سریع زبان صفحه، به عنوان یک شاخص مورد استفاده در خزش زبانی پیشنهادی در نظر گرفته شده است. ولی از آنجاییکه صفحات فاقد فرابرچسب زبانی با این روش قابل تشخیص نمی‌باشند، بنابراین شناسایی زبان صفحات بدون فرابرچسب نیاز به مکانیزمهای مکملی دارد. دراین مقاله روش بکارگیری توکنها با کلمات منتخب زبان به عنوان روش مکمل برای شناسائی زبان صفحه پیشنهاد شده‌است. بخشی که خزشگرها محتوای صفحات را تجزیه می‌کنند پارسر نامیده می‌شود. در طرح پیشنهادی، عملیات شناسائی زبان

➢ معیارهای شناسائی پیوندهای مرتبط

اطلاعاتی که به صورت متداول جهت شناسائی پیوندهای مرتبط مورد استفاده قرار می‌گیرد، شامل اطلاعات برچسب پیوند یا متن اطراف آن می‌باشد. همچنین در برخی از روشهای مبتنی بر وراثت از امتیاز صفحه پدر نیز جهت امتیازدهی به پیوندهای موجود در آن استفاده می‌گردد.

استفاده از متن پیوند جهت تعیین مرتبط بودن صفحاتی که به آن اشاره می‌کنند، منجر به تسریع عملیات شناسائی می‌شود. در مواردی که متن پیوند تهی باشد جهت شناسائی محتوای صفحات ارجاعی از متن اطراف پیوند کمک گرفته می‌شود. محدودیتی که در استفاده از این معیار وجود دارد آن است که گاهی شناسائی مرز اطلاعات مرتبط با پیوند دشوار است.

در روشهای وراثتی اهمیت صفحات در سطوح پائین‌تر با توجه به اهمیت صفحات در سطوح بالاتر افزایش یا کاهش می‌یابد[10,14]. مشکلی که در این‌گونه روشها وجود دارد انتشار خطای شناسائی صفحات مرتبط یا اولویت‌بندی صفحات، در یک مرحله به مراحل دیگر می‌باشد.

➢ معیارهای شناسائی صفحات مرتبط

مرتبط بودن محتوای صفحه مورد جستجو نیز می‌تواند مبتنی بر ویژگیهای متن یا پیوند و یا ترکیبی از هر دو ویژگی بررسی گردد. در روشهای مبتنی بر متن از جستجوی کلمات کلیدی، تطبیق محتوای صفحه با عبارت جستجو و تطبیق متن صفحه با هستان‌شناسی یا گنجینه واژگان استفاده می‌گردد. روشهای مبتنی بر پیوند نظیر [14,20] به شناسائی صفحات محبوب یا پرمراجعه می‌پردازند.

در روشهای مبتنی بر هستان‌شناسی [4,11]، بخش زیادی از صفحات مرتبط با توجه به ارتباط معنایی و مفهومی بازیابی می‌شوند که در روشهای معمول قابل بازیابی نمی‌باشند. روشهای مبتنی بر فضای برداری [6,7] به عنوان دسته دیگری از روشهای خزش متمرکز، به مقایسه عنوان پرس و جو با صفحه می‌پردازند. روش مبتنی بر بازخورد [8] نیز با توجه به اینکه عملیات خزش را طی چندین مرحله اجرا می‌کند و نتایج هر مرحله را در اختیار مراحل بعدی قرار می‌دهد، روش زمانبری می‌باشد. از جمله مزیتهای این روش آن است که با توجه به بهبود حاصله در هر مرحله از خزش نسبت به مراحل قبلی با داشتن دانش از مرحله پیشین نتایج هر مرحله نسبت به مراحل قبلی قابل بهبود می‌باشد. در واقع در این حالت خزشگر از قابلیت یادگیری برخوردار است.

۳-۳ استخراج شاخصهای خزش زبانی

۳-۳-۱ نیازمندیهای خزش زبانی

با توجه به ایده خزش متمرکز که در بخش قبلی مرور شد، خزشگر زبانی نیز دارای دو نیازمندی می باشد:اولا باید دارای روش مناسبی جهت شناسائی صفحات مرتبط یا صفحات متعلق به زبان باشد و ثانیا باید از سیاست هوشمندانه‌ای در جهت دنبال نمودن صفحات مرتبط و برخورد با صفحات نامرتبط برخوردار باشد. علاوه بر این باید از روش اولویت‌بندی قابل قبولی نیز استفاده کند.

➢ شناسایی زبان

روشهای مطرح در شناسائی زبان از داده‌های محتوائی زیر استفاده می‌کنند:

- ایست واژه‌ها
- n -گرم‌های موجود در زبان
- ترکیب حروف انحصاری

استفاده از ایست‌واژه‌ها شامل جداکننده‌ها، حروف عطفی و حروف اضافه روش مناسبی جهت شناسائی زبان می‌باشد، زیرا احتمال وقوع آنها در انواع مختلف متون زیاد بوده و هر زبان تعداد محدود و مشخصی ایست‌واژه دارد. نتایج بکارگیری روش مبتنی بر ایست واژه [21] نشان می‌دهد که این روش جهت تشخیص زبان مستندات با تعداد مشخصی ایست‌واژه، روش مناسبی محسوب می‌شود. روش n-گرم [22] نیز روش متداول و پرکاربردی جهت شناسائی زبان مستندات می‌باشد. در این روش زبان مستنداتی که دارای محتوای متنی کمی باشند، قابل شناسائی نیست.

روش ترکیب حروف منحصر بفرد [23]، روشی است که در مقابل روش مبتنی بر ایست واژه مطرح شده است. در این روش زیر رشته‌های خاص زبان به عنوان شناسه‌های آن زبان مورد استفاده قرار می‌گیرد. مشکل این روش احتمال کم وقوع رشته منحصربفرد در متنهای کوتاه است. مشکل دیگر در این روش زمانی رخ می‌دهد که که زیررشته‌های منحصربفرد بدرستی انتخاب نشوند. زیرا بسیاری از زیررشته‌ها در زبانهای مختلف تکرار می‌شوند.

➢ سیاست واکشی

خزشگر زبانی در برخورد با صفحات مرتبط دو سیاست را می‌تواند در پیش گیرد. سیاست اول آن است که پیوند موجود در صفحات مرتبط را واکشی نموده و صفحات غیرمرتبط را دور بریزد. سیاست بعدی آن است که مسیر پیوندهای موجود در صفحه غیرمرتبط را تا عمق مشخصی دنبال نماید. زیرا احتمال دارد که بعد از پیمایش عمق مشخصی به یک صفحه مرتبط ختم شود. یکی از روشهای برخورد با این مسئله می‌تواند به این صورت باشد که برای هر صفحه یک پارامتر عمق در نظر گرفت و در صورتیکه پیوند موجود در آن غیرمرتبط باشد، از پارامتر عمق یک واحد کم شده و برای صفحه مورد مراجعه توسط پیوند غیرمرتبط عمق جدید را در نظر گرفت. این روند ادامه می‌یابد تا جائیکه به عمق صفر برسد و خزش در آن مسیر متوقف شود. همچنین در صورتیکه سیاست اولویت‌دهی واکشی پیوندهای موجود در صفحات مرتبط در مقابل صفحات غیرمرتبط وجود داشته باشد، باید با مکانیزمی به پیوندهای موجود در صفحات مرتبط وزن بیشتری اختصاص داد. بنابراین در یک خزشگر زبانی دو مولفه شناسایی زبان و الویت‌دهی واکشی وابسته به زبان بوده و می بایست مد نظر قرار گیرد. این دو مولفه در شکل ۱ نشان داده شده است.

۳-۴ مقایسه روشهای خزش و انتخاب شاخصهای مناسب جهت خزش زبانی

روشهای مختلف خزش با توجه به شاخصها و ویژگیهای آنها که در بخشهای قبل به آنها پرداخته شد، مقایسه و نتایج به طور خلاصه در جدول ۱ آمده است.

قابلیت یادگیری شناخته می‌شوند. در این گونه سیستمها بتدریج مسیرهای غیرمرتبط حذف شده و مسیرهای بهینه‌تری جهت جمع‌آوری صفحات مرتبط پیدا می‌شوند. همچنین روشهای خزش متمرکز وب می‌توانند از سطوح متفاوتی از دانش جهت یافتن صفحات مرتبط استفاده نمایند. به عنوان مثال منبع اطلاعاتی جهت یافتن صفحات مرتبط ممکن است کلمات کلیدی، عبارت پرس و جو [6,7] یا دانش مکمل نظیر هستان‌شناسی[14] یا گنجینه واژگان[15] [4,11] باشد. استفاده از دانش مکمل منجر به پوشش بیشتر و دقیقتر شدن نتایج خزش می‌شود. زیرا تنها به وجود کلمات کلیدی اکتفا نشده و ارتباط معنائی کلمات موجود در متن و موضوع مورد نظر نیز لحاظ می‌گردد. اطلاعات مورد استفاده توسط روشهای خزش متمرکز جهت تعیین سیاست واکشی صفحات می‌تواند مبتنی بر وراثت باشد. این اطلاعات که با توجه به ارتباط پدر و فرزندی بین صفحات موجود در گراف وب حاصل می‌شود، منجر می‌شود که امتیاز فرزندان تحت تاثیر امتیاز صفحات اجداد باشد. برخی روشها نظیر PageRank [14] و الگوریتم کوسه ماهی [10] از نمره وراثتی جهت امتیازدهی به فرزندان استفاده می‌کنند.

تحلیل صفحات وب نیز بسته به نیازها و اهداف خزش، در حین عملیات خزش یا بعد از آن صورت می‌گیرد [13]. در واقع دو امکان تحلیل پویا[16] و ایستا[17] وجود دارد. در تحلیل ایستا ابتدا صفحات با حداقل مشخصات در یک مخزن جمع‌آوری شده و بعدها بسته به کاربرد مورد نظر تحلیل می‌شوند. در این حالت محدودیتهای حافظه‌ای و منابع شبکه باید لحاظ شود. در حالیکه در تحلیل پویا امکان دسته‌بندی اطلاعات متناسب با کاربرد مورد نظر نیز وجود دارد.

مبحث دیگری که در ارتباط با خزش متمرکز وجود دارد، خزش وب عمومی و وب پنهان است. وب عمومی به بخشی از وب اطلاق می‌شود که دسترسی به آن از طریق دنبال نمودن پیوندها امکانپذیر باشد و نیاز به تائید یا ثبت‌نام نداشته باشد. در مقابل وب عمومی، بخش دیگری تحت عنوان وب پنهان وجود دارد که دسترسی به آن نیاز به اخذ برخی مجوزها و پرکردن برخی از فرمها دارد. این بخش از وب حجم وسیعی از اطلاعات مفید و با ارزش وب را دربردارد.

در ارتباط با خزش زبانی فعالیتهای محدودی صورت گرفته است [15,17,18]. روشی تحت عنوان خزش زبانی وب (LSWC)[18] با هدف ایجاد آرشیوی بزرگ از وب کشور تایلند در [17] پیشنهاد شده است. روش LSWC جهت شناسائی زبان از اطلاعات فرابرچسب و روش n-گرم استفاده می‌کند. در واقع شناسائی زبان در دو مرحله صورت می‌پذیرد. ابتدا از اطلاعات فرابرچسب کاراکترست جهت شناسائی زبان استفاده می‌شود و در صورت عدم شناسائی زبان صفحه، در مرحله بعد روش n-گرم مورد استفاده قرار می‌گیرد. در روش معرفی شده در [15] شناسائی زبان صفحه به کمک اطلاعات فرابرچسب کاراکترست یا به کمک ابزار شناسائی کاراکترست تحت عنوان Mozilla Charset Detector انجام می‌شود. به دلیل اینکه برخی از زبانها مانند زبان تایلندی توسط این ابزار شناسائی نمی‌شوند، از اطلاعات فرابرچسب آنها استفاده می‌شود. ایده اصلی روش [18] خزش دامنه وسیعی از اینترنت و یافتن مستندات متعلق به زبان مورد نظر می‌باشد. عملیات شناسائی زبان با جستجوی تعدادی از کلمات زبان توسط یک موتور جستجو صورت می‌گیرد. همچنین جهت محدود نمودن خزشگر، دامنه جستجوی آن به دامنه کشورهای آفریقای جنوبی محدود می‌شود. در مراحل بعدی خزش با توجه به شناسائی کلمات بیشتری از زبان مورد نظر، تعداد مستندات بیشتری یافته می‌شوند. بعد از ایجاد این مجموعه زبانی، روش n-گرم جهت شناسائی تعداد مستندات بیشتر متعلق به زبان هدف استفاده می‌شود.

۳- بررسی نیازمندیهای خزش متمرکز و خزش زبانی از دیدگاه مقایسه‌ای

با توجه به اینکه روشهای خزش زبانی مبتنی بر ایده خزش متمرکز مطرح شده‌اند، در این بخش جهت استخراج نیازمندیهای خزش زبانی ابتدا روشهای خزش متمرکز مورد بررسی قرار گرفته، سپس با تعیین سازوکار و ویژگیهای خزش متمرکز و نیازمندیهای خزش زبانی، با مقایسه ویژگیهای مورد استفاده در روشهای مختلف خزش، شاخصهای مناسب جهت خزش زبانی مشخص می‌شوند.

۳-۱ سازوکار خزش متمرکز

با بررسی روشهای متفاوت خزش [1,19]، می‌توان دریافت که در عملیات خزش چهار موضوع کلیدی زیر در نظر گرفته می‌شود:

- شناسائی صفحات مرتبط
- شناسائی پیوندهای مرتبط در صفحات مرتبط
- سازوکار برخورد با صفحات نامرتبط
- سازوکار دنبال نمودن پیوندهای مرتبط

در واقع یک خزشگر عملیات خزش را از تعدادی URL اولیه آغاز می‌کند، صفحات متناظر با این URLها را واکشی نموده و محتوای این صفحات را تجزیه می‌کند. سپس بر اساس شاخصهای مورد نظر، صفحات مرتبط را شناسائی کرده و اقدام به بررسی اطلاعات مستخرج می‌نمایند. بدیهی است که براساس بررسی‌های صورت گرفته در خصوص دنبال نمودن پیوندهای موجود در این صفحات تصمیم‌گیری می‌نماید. برای دنبال نمودن پیوندها، آنها به لیست URLهای قابل واکشی اضافه می‌شوند. البته در روشهای مبتنی بر اولویت، پیش از اینکه یک آدرس به لیست URLهای قابل واکشی اضافه شود، بر اساس سازوکاری وزندهی می‌شود.

۳-۲ استخراج شاخصهای خزش متمرکز

مهمترین فرآیندی که در خزش صورت می‌گیرد، انتخاب صفحات مرتبط و متعاقبا پیوندهای مرتبط است. به این منظور برخی روشها مانند روش ماهی پیوندهای موجود در صفحات مرتبط را مرتبط فرض نموده و آنها را به لیست URLهای قابل واکشی اضافه می‌کنند. دسته‌ای دیگر از روشها اعتبار خود پیوندها را نیز به کمک اطلاعاتی نظیر متن پیوند و متن اطراف پیوند بررسی می‌کنند. روش کوسه ماهی از جمله این روشها می‌باشد.

[14] Ontology
[15] Thesaurus
[16] Dynamic
[17] Static
[18] Language Specific Web Crawling

یکی از سازوکارهای موثر در بهبود عملکرد موتورهای جستجو در جهت بازیابی اطلاعات مطرح است.

برنامه‌ای که خزش وب توسط آن انجام می‌شود، خزشگر[4] نامیده می‌شود. هدف اصلی از طراحی خزشگرهای وب بازیابی صفحات از وب و ذخیره نمودن آنها در مخازن محلی می‌باشد. چنین مخزنی بعدها برای کاربردهائی مانند موتورهای جستجو مورد استفاده قرار می‌گیرد. خزش به دو صورت همه منظوره[5] یا عمومی[6] و خاص منظوره یا متمرکز[7] [3] امکانپذیر است. در خزش متمرکز موضوع یا حیطه خزش به صورت دقیق مشخص می‌شود در حالیکه یک خزشگر وب همه منظوره با شروع از مجموعه مشخصی URLها[8]، هر تعداد صفحه که بتواند از وب واکشی می‌کند. مهمترین ویژگی خزش متمرکز آن است که نیاز به جمع‌آوری همه صفحات وب ندارد بلکه تنها صفحات مرتبط را انتخاب و جمع‌آوری می‌کند. خزش زبانی به عنوان یکی از روشهای خزش متمرکز به جمع‌آوری صفحات متعلق به یک زبان می‌پردازد. بسیاری از روشهای خزش زبانی با توجه به ویژگی محلی زبانی در وب مطرح شده‌اند. این روشها عموما بر اساس این واقعیت عمل می‌کنند که در گراف وب بین صفحات متعلق به یک زبان معمولا پیوند وجود دارد. از جمله کاربردهای یک خزشگر زبانی استفاده از آن به عنوان یک مولفه مبنائی در جهت توسعه موتور جستجوی وب زبانی می‌باشد. همچنین پیکره ایجاد شده توسط خزشگر زبانی در جهت کاربردهائی نظیر تشخیص الگو و پردازش زبان طبیعی قابل استفاده است. به عنوان مثال در [25] از یک خزشگر غیرفارسی جهت ایجاد یک پیکره متنی فارسی استفاده شده است.

بطور کلی در ارتباط با خزش زبانی فعالیتهای بسیار کمی صورت پذیرفته است که از جمله آنها می‌توان به خزش زبانی وب تایلند اشاره نمود [15]، [17]. در روشهای خزش زبانی مطرح مانند روش مبتنی بر ایستواژه[9] [21]، روش ترکیب حروف انحصاری[10] و روش مبتنی بر n-گرم [22]، عموما تاکید بر استفاده از کاراکترست زبان یا شناسائی زبان به کمک کلمات موجود در زبان است.

در ارتباط با خزش متمرکز مبتنی بر زبان فارسی نیز فعالیتهای زیادی صورت نگرفته است. در [24] و کارهای مشابه در مورد وب سایر زبانها، گستره خزش محدود به دامنه کشورها شده است و همانطور که خود نویسندگان گفته‌اند، این روش کارایی لازم را ندارد. با توجه به پراکندگی وب سایتهای ایرانی روی سرورهای خارج از دامنه ایران (ir.) نیاز به ارائه روشی جهت بهبود خزش مستندات فارسی وب احساس می‌شود. در این مقاله خزشگر زبانی پیشنهاد شده با شناسائی صفحات فارسی وب در هر مرحله امکان پوشش صفحات فارسی بیشتری را برای مراحل بعدی فراهم می‌آورد. سازوکار خزش ارائه شده مبتنی بر ترکیب مناسبی از ویژگیهای فرابرچسب[11] و محتوای صفحات وب امکان شناسائی صفحات فارسی را با دقت و سرعت مناسبی فراهم می‌کند. با توجه به اینکه عموم روشهای خزش زبانی مطرح تنها یکی از این دو ویژگی را بکارمی‌برند، بکارگیری ترکیبی از این ویژگیها در روش پیشنهادی این روش را از سایر روشهای مطرح در حوزه خزش زبانی متمایز نموده است.

همچنین با توجه به اهمیت جمع‌آوری صفحات هدف در روشهای خزش متمرکز بطور کلی و روشهای خزش زبانی به طور خاص، لازم است سیاستهای مورد نیاز جهت انتخاب و اولویت‌بندی واکشی صفحات در حین عملیات خزش در نظر گرفته شود. در واقع با توجه به اینکه در خزشگرها لیست پیوندهای قابل واکشی با توجه به اولویت‌بندی که برای خزشگر تعریف می‌گردد، چیده می‌شوند. لذا در صورتیکه سیاست واکشی در خزشگر مبنی بر اولویت‌دهی به صفحات فارسی تعریف نگردد، علی رغم تشخیص زبان، خزشگر کلیه پیوندهای موجود در صفحات وب را واکشی می‌نماید. بنابراین بخش مهم دیگری که در راستای خزش زبانی در این مقاله به آن پرداخته شده است، سیاست واکشی جهت دنبال نمودن صفحات متعلق به زبان می‌باشد.

در این مقاله جهت ارائه روشی برای خزش بهینه مستندات فارسی وب، ابتدا روشهای مختلف خزش متمرکز مورد بررسی قرار گرفته و سپس دیدگاهی برای خزش زبانی به ویژه زبان فارسی ارائه شده است. براساس دیدگاه پیشنهادی و با استفاده از خزشگر متن باز وایر[12] ، یک خزشگرزبانی برای زبان فارسی طراحی و پیاده سازی شده است. آزمایش‌های عمل‌آمده بر روی این خزشگر نشان می‌دهد که دیدگاه پیشنهادی در خزش موثر صفحات فارسی بسیار کارا عمل کرده است. در ادامه مقاله به شکل زیر سازماندهی شده است. در بخش ۲ کارهای مرتبط در زمینه خزش متمرکز مرور می‌گردد، در بخش ۳ شاخصهای خزش متمرکز استخراج شده و مبتنی بر این شاخصها و نیازمندیهای خزش زبانی، دیدگاه خزش زبانی در بخش ۴ پیشنهاد می‌شود. در بخش ۵ بر اساس دیدگاه پیشنهادی نحوه پیاده‌سازی و ارزیابی خزشگر ارائه می‌گردد.

۲- مروری بر کارهای مرتبط

امروزه روشهای خزش متمرکز به دلیل صرفه‌جوئی در استفاده از منابع مورد توجه محققان قرار گرفته است. این روشها در بسیاری از موارد تحت عنوان خزش موضوعی[13] شناخته می‌شوند. در خزش موضوعی صفحات مرتبط با موضوع مورد نظر جمع‌آوری می‌شوند و بسته به روش آن ممکن است از کلمات کلیدی، عبارت پرس و جو و یا واژگان تخصصی موضوع استفاده شود [1,2,3,4,6,7,9,10,11,14]. در روش خزش متمرکز به جای جمع‌آوری و شاخص‌گذاری تمام صفحات وب (به منظور پاسخگوئی به تمام پرس و جوهای ممکن)، با شناسائی محدوده خزش، مسیرهای مرتبط مشخص شده و خزش در محدوده‌های غیرمرتبط متوقف می‌شود. این امر منجر به صرفه‌جوئی قابل توجه در سخت‌افزار و منابع شبکه شده و بروزرسانی مستندات واکشی شده توسط خزشگر را نیز بهبود می‌بخشد.

عملیات خزش متمرکز می‌تواند به صورت یک مرحله‌ای یا چندمرحله‌ای انجام شود. در روش چندمرحله‌ای که به آنها روشهای مبتنی بر بازخورد نیز گفته می‌شود، در هر مرحله خزش از دانش حاصل از مراحل پیشین استفاده می‌شود [8]. به این ترتیب انتظار می‌رود نتایج هر مرحله نسبت به مراحل قبلی بهبود یابد. روشهای مطرح در این دسته اغلب تحت عنوان روشهای با

[4] Crawler
[5] General Purpose
[6] General
[7] Focused
[8] Uniform Resource Locator
[9] Stop Word
[10] Unique Letter Combination
[11] Meta-tag
[12] WIRE
[13] Topic-Driven

مجله بین المللی علمی- پژوهشی
ارتباطات و فناوری اطلاعات

ITRC

جلد یکم - شماره ۴ - شهریور ماه ۱۳۸۸ (ش ص ۲۷-۱۵)

یادداشت پژوهشی

طراحی و پیاده‌سازی یک خزشگر زبانی جهت بهبود سازوکار خزش در مستندات فارسی وب

معصومه عظیم‌زاده
مرکز تحقیقات مخابرات ایران
پژوهشکده فناوری اطلاعات
azim_ma@itrc.ac.ir

علیرضا یاری
مرکز تحقیقات مخابرات ایران
پژوهشکده فناوری اطلاعات
a_yari@itrc.ac.ir

ابوالفضل آل‌احمد
دانشگاه تهران
گروه تحقیقاتی پایگاه داده‌ها
a.aleahmad@ece.ut.ac.ir

تاریخ دریافت: ۱۳۸۸/۱/۲۹ - تاریخ پذیرش: ۱۳۸۸/۶/۲۴

چکیده- حجم زیاد، ماهیت پویا و غیرقابل کنترل وب چالشهای زیادی را در خصوص خزش وب ایجاد نموده است. روشهای خزش به طور کلی به دو دسته عمومی و متمرکز قابل تقسیم هستند. در روش خزش عمومی همه صفحات وب جمع‌آوری می‌شوند و در روش خزش متمرکز تنها بخشی از صفحات وب که با موضوع خاصی مرتبط هستند،جمع‌آوری می‌گردند. خزش زبانی به نوعی از خزش متمرکز اطلاق می‌شود که صفحات نوشته شده به زبان مورد نظر را جمع-آوری می‌کند. با توجه به اینکه وب حاوی گستره وسیعی از داده‌های بدون ساختار و نوشته شده به زبان‌های مختلف است، نحوه انجام خزش زبانی ازجمله چالشهای بازیابی اطلاعات در محیط وب است. دراین مقاله برای بهبود خزش مستندات فارسی وب، یک خزشگر زبانی پیشنهاد گردیده و تشریح شده است. نتایج حاصل از پیاده سازی و تست این خزشگر نشان می‌دهد خزشگر زبانی در خزش صفحات فارسی وب با کارایی بهتری عمل می‌کند.

کلیدواژه‌ها: **خزشگر فارسی، خزش متمرکز، خزش زبانی، بازیابی اطلاعات.**

۱- مقدمه

وب متشکل از مجموعه صفحاتی است که از طریق فراپیوند[1] به یکدیگر متصل شده‌اند که امکان دسترسی آنها به یکدیگر را تسهیل می‌کند. به طور خلاصه وب در قالب یک گراف قابل تصور است که صفحات آن گره‌های این گراف و فراپیوندها لبه‌های آن را شکل می‌دهند. به کمک این فراپیوندها یا لبه‌ها دسترسی از صفحات مبدا به صفحات مقصد امکانپذیر می‌شود. ماهیت در حال تغییر وب علاوه بر تغییر محتوای صفحات و حذف و اضافه‌شدن آنها، منجر به تغییر ساختار ارتباطی فراپیوندها یا صفحات نیز می‌گردد که این تغییرات وب منجر به عدم کنترل‌پذیری آن و طرح چالشهای فراوانی در ارتباط با خزش و "بازیابی اطلاعات"[2] از وب می‌شود. خزش[3] وب به عنوان

[1] Hyperlink

[2] Information Retrieval
[3] Crawling