دانشگاه اصفهان

دانشکده فنی و مهندسی

گروه کامپیوتر

پایان نامه‌ی کارشناسی ارشد رشته‌ی مهندسی کامپیوتر

گرایش نرم‌افزار

بهبود کراولرهای متمرکز مبتنی بر الگوریتم‌های ژنتیک با استفاده از تئوری کولونی مورچه‌ها

استاد راهنما:

دکتر کامران زمانی‌فر

پژوهشگر:

عباس شاهینی شمس آبادی

اردیبهشت ماه ۱۳۸۷

دانشگاه اصفهان

دانشکده فنی و مهندسی

گروه کامپیوتر

پایان نامه‌ی کارشناسی ارشد رشته‌ی مهندسی کامپیوتر

گرایش نرم‌افزار آقای عباس شاهینی شمس آبادی

تحت عنوان

بهبود کراولرهای متمرکز مبتنی بر الگوریتم‌های ژنتیک با استفاده از تئوری کولونی مورچه‌ها

در تاریخ ۸۷/۲/۱۷ توسط هیأت داوران زیر بررسی و با درجه عالی به تصویب نهایی رسید.

۱- استاد راهنمای پایان نامه دکتر کامران زمانی‌فر با مرتبه‌ی علمی استادیار امضا

۲- استاد داور داخل گروه دکتر امیر حسن منجمی با مرتبه‌ی علمی استادیار امضا

۳- استاد داور خارج گروه دکتر پیمان معلم با مرتبه‌ی علمی استادیار امضا

امضای مدیر گروه

چکیده

در موتورهای جستجو برای جمع آوری صفحات وب از کراولرها استفاده می‌شود. امروزه به دلیل تعداد زیاد صفحات وب، رشد سریع آنها و سرعت زیاد تغییرات، کراولرهای معمولی قادر به پوشش دادن آنها نیستند. بنابراین با توجه به منابع و زمان محدود، کراولرها باید وقت خود را برای بازیابی صفحات دارای اهمیت بیشتر صرف کنند. کراولرهای متمرکز برای بازیابی صفحاتی که مربوط به یک موضوع خاص هستند به وجود آمدند.

چون کراولرها از تعدادی صفحه شروع کرده و لینک‌های آنها را دنبال می‌کنند، فقط صفحاتی را بازیابی می‌کنند که به طور مستقیم یا غیر مستقیم لینکی از صفحات اولیه به آنها وجود داشته باشد و سایر صفحات را نمی‌توانند دنبال کنند. به این مساله، مشکل جستجوی محلی گفته می‌شود.برای حل این مشکل روشی مبتنی بر الگوریتم‌های ژنتیک ایجاد شد، کراولر متمرکز مطرح شده در این تحقیق نیز بر مبنای الگوریتم‌های ژنتیک می‌باشد. ولی هنوز راه‌حل‌هایی وجود دارد که بتوان این روش را بهبود بخشید تا صفحاتی که بازیابی می‌شوند به موضوع مورد نظر مربوط‌تر باشند. هدف اصلی در این پایان‌نامه گسترش و بهبود روش کوئین و چن با استفاده از تئوری کولونی مورچه‌ها است. در روش‌هایی که تاکنون مطرح شده‌اند پس از استخراج لینک‌های یک صفحه و امتیازدهی آنها هیچ رابطه‌ای بین آنها در نظر گرفته نمی‌شود. ولی در روش پیشنهادی رابطه خویشاوندی میان لینک‌ها در نظر گرفته شده و با الگو گرفتن از تئوری کولونی مورچه‌ها، پس از دنبال کردن لینکی از یک صفحه وب با توجه به کیفیت صفحه بازیابی شده، امتیاز سایر لینک‌ها اصلاح می‌شود.

علاوه بر این، دو روش برای بهبود مدل فضای برداری که یکی از روش‌های پر طرفدار برای تعیین کیفیت صفحات وب می‌باشد پیشنهاد شده است. در این مدل تنها جهت بردارها در نظر گرفته می‌شود که موجب امتیازدهی ناعادلانه می‌گردد به همین دلیل در این تحقیق استفاده از مقادیر مطلق پیشنهاد شده است. همچنین اگر مهمترین کلمه کلیدی در یک سند بیش از وزن اختصاص داده شده به آن تکرار شود سطح امتیاز کاهش می‌یابد در صورتی که هدف از وزن‌دهی به کلمات کلیدی پررنگ کردن نقش کلمات با اهمیت می‌باشد و هدف محدود کردن آنها نیست. در این تحقیق پیشنهاد شده است که در چنین مواردی بُعد مربوط به مهمترین کلمه کلیدی در محاسبات در نظر گرفته نشود.

کلید واژه: موتور جستجو، کراولر متمرکز، الگوریتم‌های ژنتیک، تئوری کولونی مورچه ها.

فهرست مطالب

عنوان **صفحه**

فصل اول: معرفی

۱-۱ مقدمه ۱

۱-۲ موضوع تحقیق ۱

۱-۳ پیشینه و تاریخچه موضوع تحقیق ۴

۱-۴ اهداف تحقیق ۷

۱-۵ اهمیت و ارزش تحقیق ۸

۱-۶ ساختار پایان نامه ۸

۱-۷ نتیجه‌گیری ۹

فصل دوم: سیاست‌های استفاده شده در کراولرهای معمولی

۲-۱ مقدمه ۱۰

۲-۲ سیاست انتخاب ۱۲

۲-۲-۱ محدود کردن لینکهایی که دنبال میشوند ۱۳

۲-۲-۲ کراول کردن با بالا رفتن از مسیر ۱۴

۲-۲-۳ کراول کردن متمرکز ۱۴

۲-۲-۴ کراول کردن وب عمیق ۱۴

۲-۳ سیاست بازدید مجدد ۱۴

۲-۴ سیاست ادب ۱۷

۲-۵ سیاست موازی سازی ۱۹

۲-۶ نتیجه گیری ۲۰

عنوان | صفحه

فصل سوم: ساختار یک کراولر متمرکز

۳-۱ مقدمه ۲۱

۳-۲ آماده سازی کراولر متمرکز قبل از شروع کار ۲۲

۳-۳ روشهای بررسی کیفیت صفحات ۲۳

۳-۴ روشهای امتیازدهی به لینکها ۲۵

۳-۴-۱ شمارش تعداد صفحات شامل یک لینک ۲۵

۳-۴-۲ تحلیل متن لینک ۲۶

۳-۴-۳ روش pagerank ۲۶

۳-۴-۴ معیار مکان ۲۹

۳-۴-۵ در نظر گرفتن عمق صفحات وب ۲۹

۳-۵ چگونگی انتخاب لینکها ۳۰

۳-۶ روشهای مختلف توقف در کراولرهای متمرکز ۳۰

۳-۷ بررسی تعدادی از کراولرهای متمرکز که تاکنون مطرح شدهاند ۳۱

۳-۷-۱ کراولر متمرکز با قابلیت یادگیری ۳۱

۳-۷-۲ کراول کردن متمرکز وب با استفاده از گرافهای زمینه ۳۳

۳-۸ استفاده از عاملهای وفقی برای جستجوی مقیاس پذیر وب ۳۵

۳-۹ نتیجه گیری ۳۷

فصل چهارم: الگوریتم بهینه‌سازی کولونی مورچه‌ها

۴-۱ مقدمه ۳۸

۴-۲ بهینه سازی کولونی مورچه‌ها ۳۹

عنوان | صفحه

۴-۲-۱ سیستم مورچه ۴۲

۴-۳ الگوریتم مسیریابی AntNet ۴۴

۴-۴ نتیجه گیری ۴۵

فصل پنجم: بهبود کراولرهای متمرکز مبتنی بر الگوریتم‌های ژنتیک با استفاده از تئوری کولونی مورچه‌ها

۵-۱ مقدمه ۴۶

۵-۲ استفاده از الگوریتمهای ژنتیک ۴۷

۵-۳ روش پیشنهادی ۴۹

۵-۳-۱ چگونگی محاسبه امتیاز صفحات وب ۵۲

۵-۳-۲ چگونگی محاسبه امتیاز لینکها ۵۳

۵-۳-۳ تعیین ترتیب دنبال کردن لینکها ۵۵

۵-۳-۴ اصلاح امتیاز لینکها با استفاده از الگوریتم مسیریابی AntNet ۵۶

۵-۳-۵ برطرف نمودن مشکل صفحات وب تکراری دارای آدرسهای متفاوت ۵۷

۵-۴ ارزیابی روش پیشنهادی و ارائه نتایج ۵۹

۵-۵ نتیجه گیری ۶۵

فصل ششم: بهبود فضای برداری مبتنی بر TF.IDF

۶-۱ مقدمه ۶۶

۶-۲ استفاده از مقادیر مطلق ۶۸

۶-۳ کاهش ابعاد بردارها ۶۸

۶-۴ نتایج برای استفاده از مقادیر مطلق ۷۱

۶-۵ نتایج برای کاهش ابعاد بردارها ۷۲

عنوان صفحه

۶-۶ نتیجه‌گیری ۷۳

فصل هفتم: نتیجه‌گیری و راه‌کارهای آینده

۷-۱ نتیجه‌گیری ۷۴

۷-۲ راه‌کارهای پیشنهادی برای ادامه کار ۷۲

فهرست منابع و مآخذ ۷۶

فهرست شکل‌ها

عنوان **صفحه**

شکل ۱-۱: کراولر مسئول جمع آوری صفحات از وب جهانی می باشد ۲

شکل ۱-۲:تونل با اندازه حداکثر دو ۶

شکل ۲-۱: ارزیابی سن وتازگی ۱۶

شکل ۳-۱: ساختار کلی یک کراولر متمرکز ۲۱

شکل ۳-۲: نمودار سه فاز مختلف کراولر با قابلیت یادگیری ۳۲

شکل ۳-۳: ساختار کلی کراولر متمرکز با استفاده از گراف های زمینه ۳۴

شکل ۳-۴: معماری کلی عامل عنکبوت اطلاعاتی ۳۶

شکل ۴-۱: تعریفهای کلی یک AS ۴۰

شکل ۴-۲: الگوریتم کلی یک سیستم مورچه ۴۳

شکل ۵-۱: ساختار کلی کراولر مبتنی بر الگوریتمهای ژنتیک ۴۹

شکل۵-۲: روند اصلاح امتیاز صفحات ولینکها پس از بازیابی یک صفحه جدید، که مشابه الگوریتم مسیریابی ANTNET میباشد ۵۰

شکل ۵-۳: ساختار کلی روش پیشنهادی ۵۱

شکل ۵-۴:شبه کد الگوریتم اصلاح امتیاز لینکها بر مبنای الگوریتم مسیریابی ANTNET ۵۷

شکل ۵-۵: قسمتی از نتایج ذخیره شده در حین آزمایشات که نشان دهنده صفحات یکسان با آدرسهای متفاوت است ۵۸

شکل ۵-۶: نمودار مربوط به HARVEST RATE در آزمایش اول ۶۰

شکل۵-۷: نمودار مربوط به میانگین امتیازات صفحات جمعآوری شده در آزمایش اول ۶۰

شکل ۵-۸: نمودار مربوط به HARVEST RATE در آزمایش دوم ۶۱

شکل ۵-۹: نمودار مربوط به میانگین امتیازات صفحات جمعآوری شده در آزمایش دوم ۶۱

شکل ۵-۱۰: نمودار نتایج روشهای مختلف نسبت به زمان ۶۳

شکل ۵-۱۱: نمودار نتایج مربوط به استفاده از مقادیر مختلف α ۶۳

شکل ۵-۱۲: نمودار نتایج مربوط به استفاده از مقادیر مختلف β ۶۴

شکل ۵-۱۳: نمودار نتایج مربوط به استفاده از مقادیر مختلف γ ۶۴

شکل ۵-۱۴: نمودار نتایج مربوط به استفاده از مقادیر مختلف δ ۶۵

عنوان **صفحه**

شکل ۶-۱: مثالی از نقایص روش معمولی فضای برداری ۶۹
شکل ۶-۲: الگوریتم پیشنهادی برای کاهش ابعاد بردارها ۷۰
شکل ۶-۳: استفاده از مقادیر مطلق در مقایسه با روشهای قبلی ۷۱
شکل ۶-۴: نتایج اجرای اول برای کاهش ابعاد بردارها ۷۲
شکل ۶-۵: نتایج اجرای دوم برای کاهش ابعاد بردارها ۷۳

فهرست جدول‌ها

عنوان - **صفحه**

جدول ۲-۱ :سیاست های به کار گرفته شده در یک کراولر کارا ۱۱
جدول ۵-۱: مقایسه HARVEST RATE متوسط در بین سه روش ۶۲
جدول ۵-۲: مقایسه میانگین امتیازات صفحات بازیابی شده در سه روش مختلف ۶۲

فصل اول

معرفی

۱-۱ مقدمه

در این فصل موضوع، تاریخچه، اهداف و اهمیت تحقیق بیان می‌شود. همچنین در اینجا خلاصه‌ای از مسائل و مشکلات مطرح در کراولرهای متمرکز و علت استفاده از الگوریتم‌های ژنتیک و تئوری کولونی مورچه‌ها در این تحقیق، توضیح شده است.

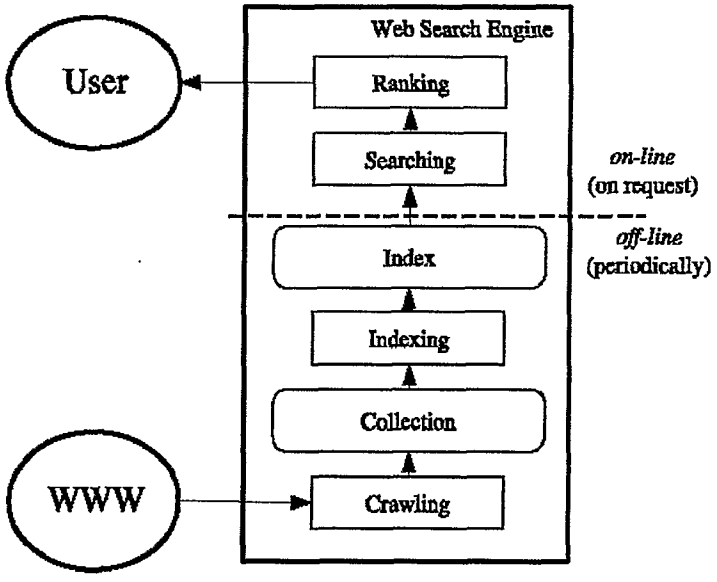
۱-۲ موضوع تحقیق

کراول کردن[1] یکی از مراحلی است که به صورت off to line در موتورهای جستجو انجام می شود. همان-طور که در شکل ۱-۱ نشان داده شده است این مرحله در پایین‌ترین سطح از یک موتور جستجو قرار دارد که صفحات وب جهت ایندکس گذاری جمع آوری می‌شوند. روش کار به این صورت است که در ابتدا آدرس تعدادی از صفحات در اختیار کراولر قرار گرفته، کراولر آن صفحات را بازیابی و ذخیره کرده، سپس لینک های آنها را استخراج می نماید. پس از استخراج کلیه لینک های مربوط به صفحات فعلی، صفحاتی که لینک ها به آنها اشاره می کنند بازیابی و ذخیره شده و کارهای قبلی بر روی صفحات جدید و لینک های آنها برای بدست

---

[1] Crawling

آوردن صفحات بیشتر مجددا تکرار می شوند[۱۰].



شکل ۱-۱: کراولر مسئول جمع آوری صفحات از وب جهانی می باشد [۱۰].

امروزه به علت حجم بسیار بزرگ وب و تغییرات و به روز رسانی‌هایی که همه روزه بر روی صفحات وب انجام می‌شود ، پوشش دادن همه صفحات وب و ارائه دادن اطلاعات به روز ، توسط موتور های جستجوی معمولی امری بسیار دشوار است و در چنین شرایطی برای اینکه یک کراولر بتواند کارایی بالایی داشته باشد اولا باید معماری سخت افزار و نرم افزار پشتیبانی کننده آن بهینه باشد تا بتواند مقادیر زیادی از صفحات وب را در زمان مناسبی جمع‌آوری کند[۵۳،۵۱] که از آن جمله می توان پایداری در مقابل خطا و مدیریت منابع را نام برد. دوم اینکه مجهز به استراتژی‌های هوشمند باشد تا بتواند در انتخاب کارهای مناسب (کدام یک از صفحات بازیابی شوند و غیره) تصمیم گیری کند که از آن جمله می توان استراتژی هایی برای صفحات مهم [۱۴] بازیابی اسناد با موضوع خاص [۴۷،۱۳،۱۱]، کراول کردن دوباره صفحات برای بهینه سازی سرعت بروز رسانی صفحات وب [۱۹,۱۶] و برنامه ریزی فعالیت بازیابی صفحات نسبت به زمان را نام برد . ما در این تحقیق به ویژگی دوم (استفاده از استراتژی‌های هوشمند) می‌پردازیم. آنچه امروزه به آن توجه زیادی شده است استفاده از کراولرهای متمرکز می باشد و از آن جهت به آنها متمرکز گفته می شود که بر روی موضوع خاصی تمرکز یافته-اند و تنها صفحات مربوط به آن موضوع را بازیابی کرده و تنها لینک های آن صفحات را دنبال می‌کنند . معمولا

به این کراولر ها آدرس تعدادی از صفحاتی که به موضوع مورد نظر زیاد مربوط هستند و تعدادی کلمات کلیدی به عنوان ورودی داده می شود، آن صفحات بازیابی شده و لینک های آنها دنبال می‌شود و صفحات جدید بدست می آید. البته تنها صفحاتی ذخیره شده و لینک های آنها دنبال می شود که ارزش کلمات کلیدی در آن صفحات زیاد باشد .

کراولرهای معمولی و متمرکز دارای یک سری از مشکلات مشترک هستند که علت آن محدودیت منابع در دسترس آنها و زمان می‌باشد. به عنوان مثال باید از هدر دادن پهنای باند پرهیز شود، بار خیلی زیاد به سرویس دهندگان وب تحمیل نشود و از روش‌های مناسب برای موازی سازی استفاده گردد. فصل دوم از این پایان‌نامه به بررسی این مشکلات کلی و راه حل‌های مطرح شده برای آنها پرداخته است.

تعدادی از مسائل هستند که در کراولرهای متمرکز از اهمیت بیشتری برخوردار هستند که در فصل سوم، ساختار کلی یک کراولر متمرکز و توضیح در مورد این مسائل مطرح گردیده است. در ادامه نیز به تعدادی از این مسائل اشاره خواهیم کرد.

همان طور که قبلا توضیح داده شد یک کراولر کار خود را با تعدادی از صفحات وب که به آنها صفحات اولیه گفته می‌شود، شروع کرده و با بازیابی آنها و دنبال کردن لینک‌های آنها صفحات جدیدی را بازیابی می‌نماید. در یک کراولر متمرکز صفحات وب اولیه مناسب، به موضوع مورد جستجو بستگی دارد و با توجه به هر موضوعی هر بار کراولر باید صفحات مناسب آن موضوع را در اختیار داشته باشد. معمولا این صفحات اولیه توسط کاربر در اختیار کراولر قرار می‌گیرند البته روش‌های دیگری هم وجود دارد که درباره این موضوع در بخش ۳-۲ بحث شده است.

یک کراولر متمرکز تنها باید صفحات وب مربوط به یک موضوع خاص را جمع‌آوری نموده و لینک‌های آنها را دنبال کند پس باید از روش‌های صحیح و بهینه برای تحلیل محتوای صفحات وب و بررسی مربوط بودن آنها به موضوع مورد نظر استفاده کند. در بخش ۳-۳ در مورد این روش‌ها بحث شده است.

پس از استخراج لینک‌ها از صفحات بازیابی شده یک کراولر متمرکز کارا باید بتواند لینک‌های با ارزش‌تر را تشخیص دهد یعنی بتواند پیش‌بینی کند که چه لینک‌هایی به صفحات با کیفیت‌تری اشاره می‌کند و ابتدا صفحات اشاره شده توسط آنها را بازیابی کند. در این باره روش‌های مختلفی وجود دارد که در بخش‌های ۳-۴ و ۳-۵ مطرح شده است، تمرکز روش پیشنهادی در این پایان نامه نیز در مورد چگونگی اصلاح امتیاز لینک‌ها در

حین کراول کردن می‌باشد.

۱-۳ پیشینه و تاریخچه موضوع تحقیق

وب جهانی شامل میلیاردها صفحه می‌باشد در سال ۱۹۹۹، کراولر مرکاتور[۱] به عنوان سریع‌ترین کراولر ایجاد شده قادر به جمع آوری صفحات با سرعت ۱۱۲ صفحه در ثانیه بود [۳۱] که برای کراول کردن یک بیلیون صفحه از اینترنت سه ماه احتیاج داشت و در این مدت ممکن بود صفحات زیادی اضافه شده یا تغییر داده شوند.

اولین بار کراولر متمرکز در سال ۱۹۹۹ توسط چاک رابارتی [۱۱] مطرح شد و اولین مسئله ای که با آن برخورد کرد این بود که بتوان قبل از بازیابی یک صفحه وب پیش بینی کرد که آیا آن صفحه به موضوع مورد نظر مربوط است یا خیر. که البته در سال های قبل از آن روش هایی مطرح شده بود که از آن جمله روش پینکرتون [۴۴] بر اساس متن لینک‌ها[۲] را می‌توان نام برد.

جمع‌آوری صفحات وب در کراولرهای متمرکز به روش جستجوی محلی است، چون یک کراولر متمرکز از یک سری صفحات اولیه شروع می‌کند و تنها می تواند صفحاتی را بازیابی کند که از طریق لینک به طور مستقیم یا غیره مستقیم به صفحات اولیه وصل باشند. روش جستجوی محلی دارای سه مشکل می‌باشد[۴۶]:

۱. بعضی از سایت‌ها که مربوط به یک موضوع مشابه می‌باشند ممکن است به دلایل رقابتی و غیره به همدیگر لینک نداشته باشند که از آن جمله می توان سایت های تجاری را نام برد .

۲. لینک‌ها ممکن است یک طرفه باشند یعنی از یک صفحه وب به دیگری لینک وجود داشته باشد ولی از دومی به اولی لینکی نباشد بنابر این با شروع از صفحه دوم نمی‌توان به اولی رسید .

۳. دسته ای از صفحات وب ممکن است بوسیله تعدادی از صفحات که به موضوع مورد نظر مربوط نیستند از دسته ای از صفحات مربوط به انها جدا شوند ، کراولر متمرکز با رسیدن به صفحات نامربوط کار خود را در آن مسیر متوقف کرده و صفحات بعدی را بررسی نمی‌کند.

رانگساوانگ و همکارانش [۴۸] روش خود را به عنوان کراولر با قابلیت یادگیری مطرح کردند کار این

---

[۱] Mercator.

[۲] Anchor text.

کراولر بر مبنای یک پایگاه دانش بیان شده است که شامل سه قسمت کلمات کلیدی ، URL های[۱] ابتدایی و پیش بینی کننده URL ها می‌باشد. در ابتدا کلمات کلیدی توسط کاربر وارد شده و در پایگاه دانش قرار می-گیرند سپس این کلمات کلیدی توسط یک موتور جستجوی معمولی مانند گوگل جستجو می‌شوند، از میان صفحات پیدا شده تعدادی از صفحات که بیشترین اولویت را دارند انتخاب شده و آدرس آنها به عنوان URLهای ابتدایی در پایگاه دانش قرار می‌گیرد سپس آن صفحات، بازیابی شده و لینکهای آنها دنبال می شوند . در هر مرحله لینکهای صفحات بازیابی شده مرحله قبل استخراج شده و برای هر کدام یک امتیاز محاسبه می شود و در قسمت پیش بینی کننده URLها قرار می‌گیرد و لینکهای با امتیاز بیشتر زودتر دنبال می شوند. اگرچه نام کراولر با قابلیت یادگیری بر روی آن گذاشته شده است ولی طرز کار آن تقریبا شبیه سایر کراولر های متمرکز است با این تفاوت که صفحات ابتدایی به عنوان ورودی به آن داده نمی شود و خودش صفحات ابتدایی برای شروع کار را پیدا می‌کند.

منسر [۴۰] استفاده از عامل های متحرک را در روشی با عنوان عنکبوت های[۲] اطلاعاتی مطرح کرد . در این روش با استفاده از بردار کلمات کلیدی و شبکه های عصبی تصمیم‌گیری می‌شد که کدام لینک ها باید دنبال شوند. در این روش عامل ها به طور مستقل و موازی عمل کراول کردن را انجام می‌دهند. پس از ایجاد هر عامل به آن مقداری انرژی نسبت داده می‌شود که با بازیابی هر صفحه وب از مقدار آن کاسته می‌شود. پس از بدست آوردن مشابهت صفحه با کلمات کلیدی مورد نظر، باتوجه به کیفیت صفحه، مقداری به انرژی نسبت داده شده به عامل اضافه می‌شود. اگر انرژی عامل از حد خاصی بیشتر شود می‌تواند یک عامل دیگر را ایجاد نموده، مقداری از انرژی خود را به او داده و تعدادی از لینکهای استخراج شده از صفحات را برای دنبال کردن در اختیار او بگذارد. همچنین اگر انرژی عامل در اثر بازیابی صفحات نامربوط از حد خاصی کمتر شود عامل می‌میرد.

کراولرهای متمرکز مدت زیادی نیست که مطرح شده‌اند و کارهای کمی بر روی آنها انجام شده است همه روش های مطرح شده در بالا با سه مشکل جستجوی محلی مواجه هستند.

در روش ارائه شده توسط دیلیجنتی و همکارانش [۲۲] مشکل دوم مطرح شده در جستجوی محلی (صفحات نامربوط قرار گرفته بین دو دسته از صفحات مربوط به موضوع مورد نظر) تا حدودی برطرف شده

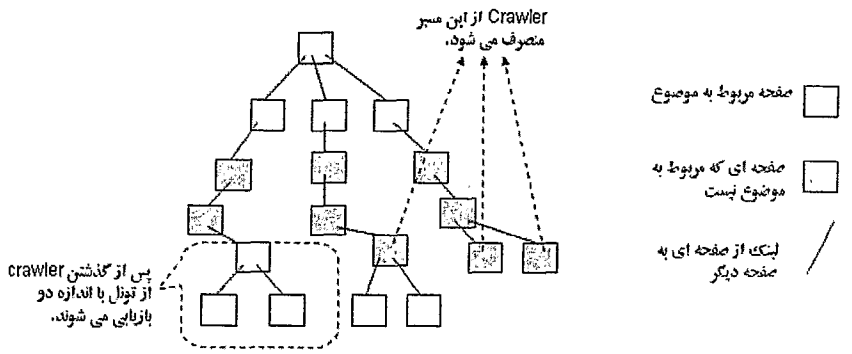
---

[۱] Uniform Resource Locator

[۲] Infospiders

است. در این روش از گراف زمینه[۱] استفاده شده و به ازای هر صفحه اولیه که در اختیار کراولر قرار می‌گیرد یک گراف زمینه ایجاد شده که اولین نود آن خود صفحه می‌باشد که در لایه صفر قرار دارد سپس با استفاده از موتورهای جستجوی معمولی صفحاتی که به آن صفحه لینک دارند پیدا شده و لایه یک را تشکیل می دهند، عمل بازگشت به عقب برای این صفحات نیز تکرار شده و کار تا جایی که تعداد از پیش تعیین شده ای از لایه ها ایجاد شود تکرار می‌شود . با استفاده از گراف زمینه، موضوعاتی که مستقیم یا غیر مستقیم به موضوع مورد نظر مربوط هستند بدست می‌آیند، موضوعات، دسته‌بندی شده و هر صفحه‌ای که بازیابی می‌شود در صف دسته مربوط به خود قرار می‌گیرد. مشکل این روش وابستگی شدید آن به موتورهای جستجوی معمولی می‌باشد.

توسط برگمارک و همکارانش [۶] روشی برای حل مشکلات جستجوی محلی ارائه شده است که تونل زدن[۲] نام دارد . در این روش کراولر متمرکز با برخورد به صفحات نامربوط به تعدادی از پیش تعیین شده (اندازه تونل) از مراحل پیش می رود. البته هیچ تضمینی وجود ندارد که حتما بعد از صفحات نامربوط، صفحاتی مربوط به موضوع وجود داشته باشد، با وادار کردن کراولر به پردازش صفحات نامربوط کارایی آن تاحد زیادی کاهش یافته و هر بار تعدادی صفحات نامربوط به صورت نویز وارد مجموعه صفحات ذخیره شده می‌گردد.



شکل ۱-۲:تونل با اندازه حداکثر دو

معمولا هنگامی که یک فضای جستجوی بزرگ داشته باشیم و اطلاعات ما در مورد آن کم باشد الگوریتم-

---

[۱] context graph

[۲] Tunneling

های ژنتیک برای جستجو در آن فضا کارایی خوبی دارند. کوئین و چن [۴۶] نیز برای حل مشکلات جستجوی محلی روشی مبتنی بر الگوریتم های ژنتیک ارائه دادند که تا حد زیادی می‌تواند این مشکلات را برطرف نماید. این روش در فصل ششم توضیح داده شده است، کراولر متمرکز پیشنهاد شده در این پایان نامه نیز بر مبنای الگوریتم‌های ژنتیک می‌باشد.

۱-۴ اهداف تحقیق

روش کوئین و چن تا حد قابل قبولی مشکلات جستجوی محلی را بر طرف می‌کند ولی هنوز راه هایی وجود دارد که بتوان آن را بهبود بخشید، تا صفحاتی که پیدا می شوند به موضوع مورد نظر مربوط‌تر باشند. هدف اصلی در این پایان‌نامه گسترش و بهبود روش کوئین و چن با استفاده از تئوری کولونی مورچه‌ها می‌باشد که تاکنون در هیچ یک از روش های مطرح شده برای کراولرهای متمرکز به کار نرفته‌است. در کراولرهای متمرکزی که تاکنون مطرح شده‌اند پس از اینکه یک صفحه وب بازیابی شد، لینک‌های آن استخراج شده و برای هر لینک یک امتیاز محاسبه می‌شود، پس از آن در ادامه کار هیچ رابطه‌ای بین لینک‌های استخراج شده از یک صفحه در نظر گرفته نمی‌شود. در این پایان نامه با الگو گرفتن از تئوری کولونی مورچه‌ها این رابطه خویشاوندی بین لینک-ها در نظر گرفته شده و از آن در اصلاح امتیاز آنها در حین کراول کردن استفاده می‌شود.

یکی از روش‌هایی که برای بررسی کیفیت صفحات وب استفاده می‌شود مدل فضای برداری[۱] نام دارد در این پایان‌نامه تعدادی از نقاط ضعف و راه‌حل‌هایی برای رفع آنها در مدل فضای برداری مطرح شده‌است.

اهداف کلی که در این پایان نامه دنبال می‌شود به طور خلاصه به صورت زیر می‌باشد:

۱. روش‌های مختلف که برای بهینه سازی همه کراولرها (معمولی و متمرکز) کاربرد دارد بررسی می-شود.

۲. روش‌های مطرح برای قسمت‌های مختلف یک کراولر متمرکز بررسی می‌شود.

۳. به جای استفاده از یک روش در تعیین کیفیت صفحات بازیابی شده یا محاسبه امتیاز لینک‌ها، از

---

[۱] Vector space model

ترکیب مناسبی از روش‌هایی که تاکنون مطرح شده است استفاده می‌گردد.

٤. روشی جهت افزایش دقت روش کوئین و چن با استفادهٔ از تئوری کولونی مورچه‌ها ارائه می‌گردد.

٥. دو تکنیک برای بهبود مدل فضای برداری ارائه می‌گردد.

۱-۵ اهمیت و ارزش تحقیق

یک کاربر به تنهایی نمی‌تواند صفحات مورد نیاز خود را در وب جهانی پیدا کند و معمولا از موتورهای جستجو استفاده می‌نماید. همان طور که قبلا گفته شد کراولرهای معمولی استفاده شده در موتورهای جستجو نمی‌توانند به طور مناسبی کل صفحات وب را پوشش دهند، که این موضوع موجب به وجود آمدن کراولرهای متمرکز شده است.

بحث دیگری که امروزه مطرح است استفاده از کراولرهای شخصی می‌باشد تا هر کاربری صفحات وب مورد نیاز خود را به طور مستقیم و با استفاده از کراولر خود پیدا نموده و جمع آوری کند. یک کراولر متمرکز می‌تواند به عنوان یک کراولر شخصی نیز به کار رود.

یک کراولر متمرکز در هر جایی که استفاده می‌شود در نهایت با حجم عظیمی از صفحات وب مواجه است و ایجاد کوچکترین بهبود در آن می‌تواند از بازیابی و بررسی بی مورد دهها هزار صفحه وب جلوگیری کند. مخصوصا هنگامی که به عنوان کراولر شخصی استفاده شود کاربر به تعداد کمی از صفحات وب نیاز دارد و کراولر متمرکز باید بتواند در کمترین زمان ممکن، باکیفیت‌ترین صفحات وب موجود را در اختیار او بگذارد.

در روش پیشنهاد شده در این پایان‌نامه امتیازدهی به لینک‌ها با دقت بیشتری انجام می‌شود تا از بازیابی و بررسی صفحات نامربوط تا حد ممکن کاسته شود. همچنین در این پایان نامه به منظور تحلیل دقیق‌تر صفحات وب روش‌هایی برای بهبود مدل فضای برداری پیشنهاد شده است.

۱-۶ ساختار پایان نامه

در فصل دوم از این پایان‌نامه روش‌های مختلف استفاده شده برای داشتن یک کراولر (معمولی یا متمرکز)