دانشکده مهندســی

پایان‌نامه کارشناسی ارشد در رشته مهندسی کامپیوتر (نرم افزار)

شخصی سازی خودکار با استفاده از وب کاوی

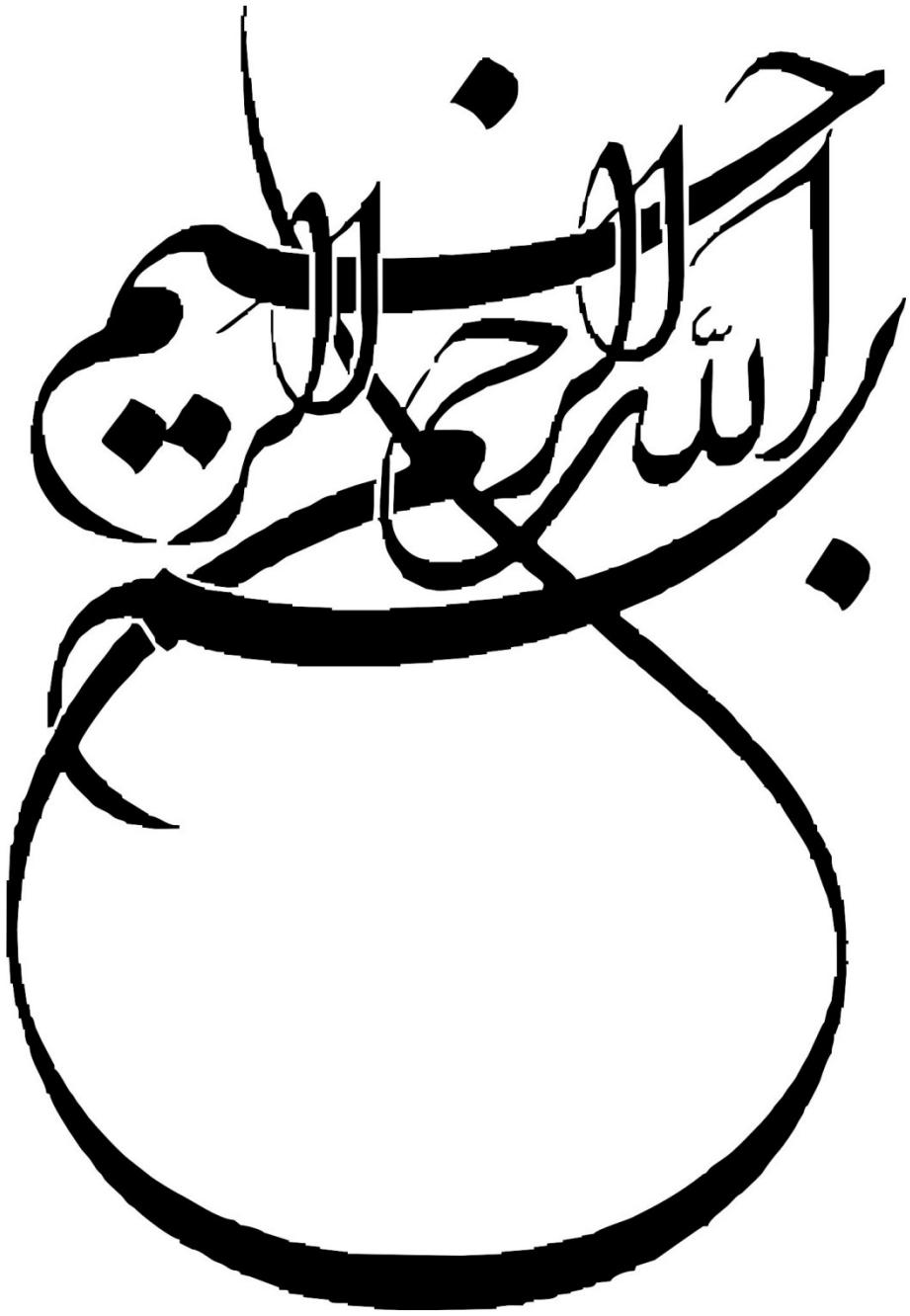
توسط:

سمیرا خنشا

استاد راهنما:

دکترمحمد هادی صدرالدینی

شهریور ماه ۱۳۸۸


بسم الله الرحمن الرحيم

به نام خدا

شخصی سازی خودکار با استفاده از وب کاوی

به وسیله ی:

سمیرا خنشا

پایان نامه

ارائه شده به تحصیلات تکمیلی دانشگاه به عنوان بخشی

از فعالیت‌های تحصیلی لازم برای اخذ درجه کارشناسی ارشد

در رشته:

مهندسی کامپیوتر-نرم افزار

از دانشگاه شیراز

شیراز

جمهوری اسلامی ایران

ارزیابی شده توسط کمیته پایان نامه با درجه: عالی

دکتر محمد هادی صدرالدینی ، دانشیار بخش مهندسی کامپیوتر (رئیس کمیته) .....

دکتر احمد توحیدی، استادیار بخش مهندسی کامپیوتر .....

دکتر رضا بوستانی، استادیار بخش مهندسی کامپیوتر .....

شهریور ماه ۱۳۸۸

با سپاس از خداوند متعال،

تقدیم به همه پویندگان این راه

و

خانواده‌ام

سپاسگزاری

اکنون که به مدد و لطف الهی تحصیلات دوره کاشناسی ارشد را به پایان رسانده‌ام، لازم می دانم از همه عزیزانی که در این راه مرا یاری کرده‌اند تشکر و سپاسگزاری کنم.

از جناب آقای دکتر محمد هادی صدرالدینی تشکر و سپاس ویژه را دارم که با صبر و حوصله فراوان زحمت راهنمایی من را در این پروژه به عهده داشتند.

از جناب آقای دکتر احمد توحیدی و جناب آقای دکتر رضا بوستانی برای مشاوره و راهنمایی ارزشمندشان سپاسگزارم.

از استاد گرامی جناب آقای دکتر نایبی که این کار حقیر را قابل ارزیابی دانسته و آن را به داوری نشستند، کمال تشکر را دارم.

از جناب آقای مهندس مرتضی تاجبخش به خاطر فراهم نمودن داده و پیشنهادات ارزشمندشان قدردانی می کنم.

از مرکز تحقیقات مخابرات ایران از بابت حمایت از پایان نامه اینجانب سپاسگزارم.

بر خود واجب می دانم از زحمات پدر و مادرعزیزم قدردانی نمایم که همیشه دعای خیرشان بدرقه راه من بوده است.

و سرانجام از همسر مهربانم کمال سپاس و تشکر را دارم، که در این راه از هیچ کوششی دریغ نکرده و همیشه مشوق من در ادامه این راه بوده است.

چکیده

شخصی سازی خودکار با استفاده از وب‌کاوی

به وسیله‌ی:

سمیرا خنشا

امروزه کاربران وب با مسأله فزونی اطلاعات و سر درگمی به خاطر رشد سریع و تصاعدی حجم اطلاعات و تعداد کاربران، مواجه هستند. در نتیجه، چگونگی ارائه دقیق‌تر اطلاعات مورد نیاز کاربران وب به یک بحث مهم درکاربردهای مبتنی بر وب، تبدیل شده است. هدف ما در این رساله بهبود کارایی سیستم‌های شخصی سازی وب با به کارگیری تکنیک‌های وب کاوی می-باشد.

وب کاوی به پروسه کشف ارتباطات جالب میان داده‌های وب از طریق به کارگیری تکنیک‌های داده کاوی گفته می‌شود، داده‌هایی که به شکل متن، لینک یا اطلاعات استفاده کاربران بیان شده است. به طور دقیق، ما با تغییر در پروسه پیش پردازش و ارائه روشی دقیق‌تر برای تشخیص ربات‌ها، الگوهای استفاده وب را با استفاده از کشف قوانین وابستگی وزن‌دار استخراج می‌کنیم و با ترکیب آنها با داده‌های محتوا و ساختار وب، محتوای شخصی شده‌ای را از طریق سفارشات به کاربران وب ارائه می‌دهیم. نتایج حاصل از آزمایشات بهبود دقت و پوشش سیستم شخصی سازی ارائه شده را نشان می‌دهد.

فهرست مطالب

فهرست جداول

فهرست شکل ها

فصل اول

مقدمه

۱- مقدمه

وب طی یک فرایند آشفته و غیرمتمرکز در حال رشد است و به یک منبع فوق العاده قدرتمند برای ذخیره سازی، انتشار و بازیابی اطلاعات و همچنین کاوش دانش مفید تبدیل گردیده است. به خاطر ویژگی‌هایی مانند گستردگی، تنوع، پویایی و طبیعت بدون ساختار داده وب، تحقیقات پیرامون داده وب با چالش‌هایی مانند مقیاس پذیری، چند رسانه‌ای و ... مواجه می-باشد. در نتیجه کاربران وب همیشه غرق در اقیانوسی از اطلاعات هستند و با مسأله سربار اطلاعاتی در هنگام تعامل با وب، مواجه هستند. به عنوان نمونه به چندین مسأله در ارتباط با تحقیقات و کاربردهای وب اشاره می‌شود:

یافتن اطلاعات مرتبط: کاربران برای یافتن اطلاعات خاصی روی وب، یا اسناد وب را به صورت مستقیم مرور می‌کنند یا اینکه از موتورهای جستجوگر بهره می‌گیرند. زمانی که کاربری از یکی از موتورهای جستجوگر برای یافتن اطلاعات استفاده می‌کند، یک یا چند کلمه کلیدی را به عنوان پرس و جو[1] انتخاب می‌کند، سپس موتور جستجو لیستی از صفحات امتیازبندی شده براساس درجه ارتباط با پرس و جو را برمی‌گرداند. با این وجود دو مسأله اساسی در ارتباط با جستجوی وب مبتنی بر پرس و جو وجود دارد [۱]، یکی مشکل پایین بودن دقت است که ناشی از تعداد زیادی صفحات نامرتبط است که توسط موتور جستجوگر برگردانده می‌شود و دومی، مسأله پایین بودن معیار پوشش است که ناشی از فقدان توانایی شاخص گذاری تمام صفحات وبی است که روی اینترنت موجود می‌باشد. که این مسأله منجر به دشواری در یافتن اطلاعات شاخص گذاری نشده، که واقعاً مرتبط هستند، می‌شود.

در دهه اخیر، چگونگی یافتن صفحات مرتبط‌تر به یک پرس و جو، به عنوان موضوع جالبی در مدیریت داده وب مطرح شده است [۲].

یافتن اطلاعات مورد نیاز: اکثر موتورهای جستجو با روش ارائه پرس و جو اجرا می‌شوند، یعنی با وارد کردن یک یا چند کلمه کلیدی نتایج برگردانده می‌شود. در بعضی از مواقع نتایجی که

[1] *query*

توسط جستجوگر برگردانده می‌شود دقیقاً آن چیزی نیست که کاربر به دنبال آن است، که این مسأله ناشی از وجود همسانی کلمات است به عنوان مثال زمانی که کاربری با پیش زمینه تکنولوژی اطلاعاتی می‌خواهد درباره زبان برنامه نویسی *"python"* اطلاعاتی کسب کند، با وارد کردن کلمه *"python"* ممکن است که اطلاعاتی که به او برگردانده می‌شود در باره جانور *"python"* (نوعی مار) باشد، نه زبان برنامه نویسی. به عبارتی معنای[۱] داده وب [۳] در زمینه جستجوی وب، خیلی کم لحاظ می‌شود.

استخراج دانش مفید: در سرویس‌های جستجوی وب سنتی، نتایج مرتبط با پرس و جو به صورت لیستی از صفحات امتیازبندی شده به کاربر برگردانده می‌شود. در بعضی مواقع، کاربران علاوه بر مرور صفحات برگردانده شده، نیازمند کشف دانش مفید از آنها نیز می‌باشند. مطالعه بر روی چگونگی استفاده از وب به عنوان یک پایگاه دانش برای تصمیم و استخراج اطلاعات، توجه محققان زیادی را به خود جلب کرده است [۴-۶].

پیشنهاد یا شخصی‌سازی اطلاعات: زمانی که یک کاربر در حال تعامل با وب است، نیازمندی-های پیمایشی متفاوتی دارد که منجر به ارائه محتوا و اطلاعات متفاوتی خواهد شد. به منظور بهبود کیفیت سرویس‌های اینترنتی و افزایش تعامل کاربران با یک سایت خاص، طراحان و مدیران وب سایت نیازمندند که بدانند کاربران واقعاً به چه اطلاعاتی نیاز دارند و در سایت به دنبال چه محتوایی هستند و صفحاتی که کاربران به دنبال آن هستند را پیش بینی کرده و با یادگیری الگوهای پیمایشی کاربران بتوانند صفحات وب شخصی‌شده را به آنها ارائه دهند [۳-۸].

مسایل مطرح شده در بالا موتورهای جستجوگر موجود و کاربردهای دیگر مبتنی بر وب را با مشکلات جدی مواجه کرده است. تلاش‌های مختلفی برای حل این مشکلات صورت گرفته است، از توسعه تکنیک‌های هوشمند پیشرفته محاسباتی گرفته تا الگوریتم‌های مختلفی از حوزه‌های پایگاه داده، داده کاوی، یادگیری ماشین، بازیابی اطلاعات و مدیریت دانش. بنابراین پیدایش وب باعث شده است که محققان و مهندسان این حوزه برای مدیریت داده‌های مبتنی بر وب و توسعه کاربردهای تحت وب با چالش‌های زیادی مواجه باشند.

---

[۱] *semantic*

۱-۱- ویژگی‌های داده وب

برای داده‌های روی وب در مقایسه با داده‌های موجود در سیستم‌های مدیریت پایگاه داده معمولی ویژگی‌های خاصی وجود دارد. داده وب معمولاً ویژگی‌های زیر را داراست:

داده‌های وب از حجم عظیمی برخوردار هستند. در حال حاضر تخمین داده واقعی روی اینترنت به لحاظ رشد تصاعدی هر روزه آن، امری بسیار مشکل است. برای مثال در سال ۱۹۹۴، یکی از موتورهای جستجوگر اولیه به نام *world wide web worm (wwww)* یک ایندکس که شامل ۱۱۰۰۰۰ صفحه وب و اسنادی که از وب دسترس‌پذیر بودند، می‌شد. در حالی که در نوامبر ۱۹۹۷، موتورهای جستجوگر ادعای ایندکسی شامل ۲ میلیون تا ۱۰۰ میلیون سند وب را داشتند [۹]. امروزه بیش از ۴ بیلیون صفحه ایندکس شده وجود دارد که روزانه در حدود ۱ میلیون صفحه به آن اضافه می‌شود. حجم عظیم داده روی وب باعث می‌شود که تکنیک‌های پایگاه داده سنتی در رسیدگی کردن به داده وب با مشکل مواجه شوند.

داده روی وب توزیع شده و ناهمگون است. به لحاظ ویژگی اساسی وب که اتصال بین گره‌های مختلفی روی اینترنت می‌باشد، داده وب میان کامپیوترها یا سرورهای زیادی در نقاط مختلف جهان توزیع شده است. البته این داده‌ها ذاتاً طبیعت چندرسانه‌ای را نیز دارند، علاوه بر اطلاعات متنی که عموماً برای بیان محتوا استفاده می‌شود، انواع دیگری شامل تصویر، صدا، فیلم و ... نیز در صفحات وب گنجانده می‌شوند. این مسأله نیازمند توسعه تکنیک‌هایی برای پردازش داده وب با توانایی رسیدگی و تحلیل انواع داده چندرسانه‌ای می‌باشد.

داده روی وب ساخت نیافته است. از آنجا که طرح و ساختار واحد و یکپارچه‌ای وجود ندارد که صفحات وب از آن تبعیت کنند، در حالی که این مسأله جزئی از نیازمندی‌های مدیریت پایگاه داده‌های معمولی می‌باشد، طراحان وب قادرند که صفحات را با توجه به سلیقه خود برای نمایش اطلاعات مربوطه سازماندهی کنند. اگرچه وجود قالب‌های *HTML* تعریف شده‌ای مانند تگ‌ها، ابرلینک‌ها و ... اجزای ساخت یافته‌ای هستند، ولی فقط در کیفیت نمایش اسناد وب مؤثرند و قادر نیستند که معنای گنجانده شده در اسناد را آشکار کنند. بنابراین نیاز رو به رشدی در زمینه بهتر مواجه شدن با این طبیعت ساخت نیافته وب احساس می‌شود، تا کاربران بهتر بتوانند اطلاعات مورد نیازشان که در دل این اسناد نهفته است را استخراج کنند.

داده روی وب پویا است. ساختار صریح و ضمنی داده وب دائماً در حال به روز شدن است. به ویژه به خاطر کاربردهای مختلف سیستم‌های مدیریت داده مبتنی بر وب، نمایش‌های متفاوتی از اسناد وب به عنوان محتوا در به روز رسانی‌های پایگاه داده‌ها ایجاد می‌شود. همچنین لینک-های سرگردانی به خاطر مشکلات جابجایی فایل‌ها یا تغییر نام آنها یا تغییر یا حذف دامنه به وجود می‌آید. این ویژگی‌ها منجر به تغییرات مداوم اسناد وب می‌شود که در نتیجه آن، سیستم‌های بازیابی اطلاعات رنج می‌برند.

ویژگی‌های بیان شده مشخص می‌کند که داده وب نوع خاصی از داده است که با انواع داده‌ای که در سیستم‌های مدیریت پایگاه داده معمولی قرار دارند متفاوت است. بنابراین درخواست‌های زیادی برای توسعه تکنیک‌های پیشرفته‌ای برای حل مشکلات جستجو یا اطلاعات روی وب و مدیریت داده آن وجود دارد. باتوجه به هدف، این مطالعه و تکنیک‌ها حول دو مقوله از مدیریت داده وب می‌باشد که شامل چگونگی دقیق‌تر پیدا کردن اطلاعات مورد نیاز روی اینترنت و چگونگی بهینه‌تر کردن مدیریت دانش روی اینترنت می‌باشد. با توسعه اینترنت و همه گیر شدن آن در سالهای اخیر، سرویس‌های مبتنی بر وب پیشرفته‌ای پدیدار شده‌اند که به کاربران کمک می‌کند تا به آسانی اطلاعات مورد نیازشان را به دست آورند.

۱-۲- داده کاوی و کاربردهای آن

اخیراً داده کاوی به عنوان یک روش مفید در حوزه مهندسی و کشف دانش مطرح شده است [۱۰]. اساساً داده کاوی به کشف اطلاعات مفید و احیاناً مخفی در حجم عظیمی از داده گفته می‌شود که می‌تواند در انواع مختلفی شامل داده‌های تراکنشی در کاربردهای تجارت الکترونیکی یا اصطلاحات ژنتیکی حوزه اطلاعاتی ژنتیک و ... باشد. بدون توجه به نوع داده، هدف اصلی داده کاوی کشف اطلاعاتی است که تا کنون دیده نشده و مخفی بوده است که در قالب الگوهایی بیان می‌شود.

امروزه داده کاوی توجه زیادی را از حوزه‌های دانشگاهی و صنعتی به خود جلب کرده است و در این بین نیز شاهد پیشرفت‌های وسیعی در عرصه‌های کاربردی گوناگونی بوده است. در دهه اخیر، داده کاوی به طور وسیعی در تحقیقات مدیریت داده وب به کارگرفته شده است. مواردی

شامل اسناد وب، ساختار لینکهای وب، تراکنش‌های کاربردی وب و معانی وب، از جمله اهداف کاوش دراین زمینه بوده‌اند. واضح است که دانش استخراجی از انواع مختلف داده وب کمک بسیار زیادی در ارائه بهتر اطلاعات و بهبود مدیریت داده وب می‌کند.

اگر چه تلاش‌های زیادی در حوزه مدیریت داده مبتنی بر وب صورت گرفته است و نتایج قابل توجهی نیز به دنبال داشته است، با این حال هنوز زمینه‌های تحقیقاتی باز و مشکلات حل نشده بسیاری وجود دارد، که این به خاطر ویژگی‌های خاص داده وب، پیچیدگی مدل‌های داده وب، گوناگونی کاربردهای وب و همچنین افزایش مطالبات کاربران وب در قالب کارایی و موثر بودن این تکنیکها می‌باشد. این مساله که چگونه بطور موثر و کارا با استفاده از تکنیک‌های پیشرفته پردازش داده، داده‌های مبتنی بر وب را مدیریت کنیم، یک زمینه تحقیقاتی فعال است که با چالش‌های زیادی نیز مواجه است. این موضوع انگیزه اصلی انجام این رساله می‌باشد.

همان گونه که بیان کردیم کاربران با حجم زیادی از داده وب و منابع مختلف مواجه هستند که یافتن اطلاعات مناسب و مورد نیاز برای آنها با دشواری‌هایی همراه است. بنابراین محققان به دنبال روش‌هایی هستند که بتواند اطلاعات مورد نیاز کاربران را کشف کند و توانایی پیشگویی آنها را داشته باشد. یکی از زمینه‌های تحقیقاتی فعال در این زمینه شخصی سازی وب می‌باشد. اکثر تحقیقاتی که در زمینه شخصی‌سازی وب انجام شده بر اساس داده کاوی محتوا و یا با استفاده از فایلهای ثبت وقابع سرورهای وب یا برنامه های در سمَت کاربر بوده است یعنی از داده‌های محتوا و داده‌های استفاده کاربران به تنهایی استفاده شده است. اگرچه از خصوصیات ساختار گراف برای شخصی سازی نتایج جستجوی وب به وفور استفاده شده است اما در فرایند شخصی سازی صفحات وب به داده کاوی ساختار کمتر توجه شده است. هدف ما در این پایان‌نامه تلفیق داده محتوا و داده ساختار با داده استفاده کاربران برای بهبود کیفیت شخصی سازی وب می‌باشد. ترتیب نگارش این پایان نامه به این صورت است که در فصل دوم وب‌کاوی و زیر مجموعه های آن یعنی کاوش ساختار، کاوش محتوا و کاوش استفاده توضیح داده خواهد شد و در فصل سوم مروری بر کارهای گذشته صورت خواهد گرفت و در پایان نیز چارچوب پیشنهادی و نتایج حاصل از آزمایشات و نتیجه گیری ارائه خواهد شد.

فصل دوم

کاوش وب

۲- کاوش وب

امروزه اینترنت به عنوان منبع عظیمی از انواع مختلف داده و همچنین حجم زیادی از دانش اطلاعاتی مخفی می‌باشد، که می‌تواند از طریق انواع مختلفی از الگوریتم‌های یادگیری ماشین و داده کاوی کشف شوند.

اگر چه پیشرفت تحقیقات در زمینه مدیریت داده مبتنی بر وب منحصر به توسعه بسیاری از کاربردهای وب و سرویس‌های بر پایه وب شده است، اما کاربران هنوز با مشکلات سربار اطلاعاتی و غرق شدن در اقیانوس عظیم دانش روی وب مواجه هستند، به طور خاص کاربران وب از مشکلات یافتن اطلاعات مطلوب و دقیق روی وب، به خاطر پایین بودن دقت و پوشش، رنج می‌برند. به عنوان مثال اگر کاربری بخواهد با استفاده از کلمات کلیدی به دنبال اطلاعات بگردد با حجم عظیمی از محتوای نامرتبط نیز مواجه خواهد شد، اغلب برای کاربران پیدا کردن دقیق دانش مورد نیاز تنها با استفاده از موتور جستجوگر مشکل است [۱، ۱۱].

بنابراین پیدایش وب چالش‌های زیادی را برای محققان در زمینه مدیریت دانش و بازیابی اطلاعات به دنبال داشته است از این رو محققان در تلاشند که با توسعه تکنیک‌های موثر و کارا به منظور کسب رضایت کاربران برای یافتن اطلاعات مرتبط [۱۲]، استخراج دانش از اطلاعات موجود [۱۳]، کشف الگوهای استفاده کاربران از پیمایش‌های روی وب [۱۴]، پیشنهاد سرویس و اطلاعات به کاربران [۱۵]، به رفع مشکلات موجود بپردازند.

از وب‌کاوی می‌توان تا حدی برای حل مشکلات بیان شده به صورت مستقیم یا غیر مستقیم استفاده کرد.

وب‌کاوی به استفاده از روش‌های داده کاوی برای استخراج دانش از منابع وب گفته می‌شود. تحقیقات وب‌کاوی توجه بسیاری از دانشگاهیان و مهندسان را در زمینه مدیریت پایگاه داده، بازیابی اطلاعات و هوش مصنوعی به خصوص داده کاوی، کشف دانش و یادگیری ماشین به خود جلب کرده است. اساساً وب‌کاوی بر اساس هدف کاوش و اینکه چه قسمتی از داده وب مورد کاوش قرار بگیرد به سه دسته کاوش ساختار وب، کاوش محتوای وب و کاوش استفاده وب تقسیم می‌شود [۱۴، ۱۶]. در بخش‌های بعد این سه دسته توضیح داده خواهد شد.

۲-۱- کاوش ساختار وب

کاوش ساختار وب شامل مدل کردن وب‌سایت از طریق ساختار لینک‌ها می‌باشد که از این اطلاعات استخراج شده از لینک‌ها می‌توان برای یافتن صفحات مرتبط براساس شباهت بین صفحات وب و ارتباط آنها استفاده کرد. با استفاده از تکنیک‌های کاوش ساختار، می‌توان به جمع‌آوری و شاخص گذاری اسناد وب و جستجو و امتیاز بندی صفحات وب با استفاده از محتوای متنی آنها و ساختار ابر پیوندها پرداخت. در ادامه به صورت مختصر مفاهیم پایه در ارتباط با وب توضیح داده می‌شود و بعد از آن روش‌های امتیاز بندی صفحات که اخیراً توسعه داده شده‌اند و بر اساس ساختار ابر پیوندها این کار را انجام می‌دهند ذکر می‌کنیم.

۲-۱-۱- مفاهیم پایه وب

وب مجموعه عظیمی از اسناد است که توسط مکانیزم ابر پیوندها که درون *HTML*[۱] قرار دارند با هم لینک دارند. *HTML* در حقیقت زبانی است که توصیف می‌کند چگونه یک سند باید در پنجره مرورگر نمایش داده شود. مرورگرها برنامه‌های کامپیوتری هستند که اسناد *HTML* را می‌خوانند و بر طبق آن اسناد را نمایش می‌دهند، همان گونه که مرورگرهای معروفی مانند *Microsoft Internet Explorer* و *Mozilla* این کار را انجام می‌دهند. این برنامه‌ها مشتریانی[۲] هستند که به سرور وب، که اسناد واقعی را نگهداری می‌کند، متصل می‌شوند و اسناد درخواستی از طریق سرور به آنها فرستاده می‌شود.

هر سند وب دارای یک آدرس وب می‌باشد که به نام *URL*[۳] معروف است که این سند را به طور یکتایی مشخص می‌کند. *URL* توسط مرورگرها برای درخواست اسناد از سرور وب و همچنین در ابر پیوندها برای اتصال صفحات به یکدیگر استفاده می‌شود. اسناد وبی که با یک آدرس به نام *URL* مرتبط شده‌اند معمولاً صفحات وب نامیده می‌شوند.

یک *URL* شامل سه قسمت می‌باشد و قالبی مانند زیر را داراست:

---

[۱] *Hyper Text Markup Language*

[۲] *client*

[۳] *Universal Resource Locator*

< protocol name >: // < machine name > / < file name >،

قسمت اول از سمت چپ پروتکلی (زبانی برای تبادل اطلاعات) است که مرورگر و سرور برای ارتباط از آن استفاده می‌کنند (*HTTP*، *FTP* و ...)، قسمت دوم، نام یا آدرس سروری هست که اسناد روی آن قرار دارند و قسمت آخر مسیری است که نشان دهنده محل ذخیره فایل روی سرور است. به عنوان مثال، *URL* زیر به یک سند *HTML* اشاره می‌کند که در فایلی به نام *index.html* در پوشه *software* روی سرور پاسارگاد ذخیره شده است:

http://pasargad.cse.shirazu.ac.ir/software/index.html

با وارد کردن *URL* در پنجره آدرس، مرورگر به سرور نام برده با استفاده از پروتکل *HTTP*[1] متصل می‌شود. بعد از اینکه این اتصال موفقیت‌آمیز بود سند *HTML* بارگذاری می‌شود و محتوای آن در پنجره مرورگر نمایش داده می‌شود.

همراه با محتوای اطلاعاتی (متن وتصاویر) یک صفحه وب معمولی شامل *URL*هایی نیز می-باشد که به صفحات دیگر اشاره می‌کند. این *URL* ها در ساختارهای خاصی در زبان *HTML* به نام ابرپیوند[2] قرار می‌گیرند. برای مثال سند *index.html* در مثال قبل شامل قطعه زیر می‌باشد:

```
<tr><td valign = right><ul>
<li><a href = " /software/datamining/" ><b>datamining</b></a>
```

ابرپیوند در این قطعه *HTML*، */software/datamining/*، است که درون تگ *<a>* با کلمه *href* مشخص شده است که روی همین سرور قرار دارد. این *URL* در صفحه با متن *datamining* که در ادامه آن آورده شده است قابل بازیابی است البته برای اتصال به صفحات مختلف به جای متن می‌توان از تصاویر و حتی متن‌های رنگی و زیر خط دار نیز استفاده کرد، به طوری که به راحتی برای کاربران قابل دیدن باشد.

نکته‌ای که در ارتباط با متن یا تصاویر نشان دهنده ابرپیوندها باید مد نظر قرار داد این است که دارای معنا و ارتباط با محتوای مورد اشاره باشند.

---

[1] *Hyper TextTransport protocol*

[2] *HyperLink*