

بِسْمِ اللّٰهِ الرَّحْمٰنِ الرَّحِیْمِ



دانشگاه صنعتی امیرکبیر

(پلی تکنیک تهران)

دانشکده مهندسی کامپیوتر و فناوری اطلاعات

پایان نامه کارشناسی ارشد

گرایش نرم افزار

داده کاوی ساختار وب با استفاده از اتوماتای یادگیر توزیع شده و سلولی و

کاربردهای آن

نگارش: سارا مطیعی

استاد راهنما: دکتر محمدرضا میبیدی

بهار ۱۳۸۷



دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)
معاونت پژوهشی

بسمه تعالی

فرم اطلاعات پایان نامه کارشناسی ارشد و دکترا

تاریخ:

پیوست:

نام و نام خانوادگی:	مطبعی	دانشجوی آزاد	<input checked="" type="checkbox"/>	بورسیه	<input type="checkbox"/>	معادل	<input type="checkbox"/>					
شماره دانشجویی:	۸۴۱۳۱۰۴۵	دانشکده:	مهندسی کامپیوتر	رشته تحصیلی:	مهندسی کامپیوتر							
نام و نام خانوادگی استاد راهنما:	محمد رضا میبیدی											
عنوان پایان نامه به فارسی:	داده کاوی ساختار وب با استفاده از اتوماتای یادگیر توزیع شده و سلولی و کاربردهای آن											
عنوان پایان نامه به انگلیسی:	Web Structure Mining using Distributed and Cellular Learning Automata and its Applications											
نوع پروژه:	کارشناسی ارشد	<input checked="" type="checkbox"/>	دکترا	<input type="checkbox"/>	کاربردی	<input checked="" type="checkbox"/>	بنیادی	<input type="checkbox"/>	توسعه ای	<input type="checkbox"/>	نظری	<input checked="" type="checkbox"/>
تاریخ شروع:	مهر ۸۵	تاریخ خاتمه:	اسفند ۸۶	تعداد واحد:	۶							
سازمان تأمین کننده اعتبار:	مرکز تحقیقات مخابرات ایران											
واژه های کلیدی به فارسی:	داده کاوی ساختار وب، اتوماتای یادگیر توزیع شده، اتوماتای یادگیر سلولی، اتوماتاهای یادگیر، پیمایشگر موضوعی، اجتماعات وب.											
واژه های کلیدی به انگلیسی:	Web Structure Mining ، Distributed Learning Automata ، Cellular Learning Automata ، Focused Crawler ، Web Community											
نظرها و پیشنهادهای به منظور بهبود فعالیت های پژوهشی دانشگاه:	استاد راهنما:											
دانشجو:	تشکیل گروه های تحقیقاتی											
امضاء استاد راهنما:	تاریخ:											
نسخه ۱: معاونت پژوهشی												
نسخه ۲: کتابخانه و به انضمام دو جلد پایان نامه به منظور تسویه حساب با کتابخانه و مرکز اسناد و مدارک علمی												

این پایان نامه توسط مرکز تحقیقات مخابرات ایران حمایت مالی شده است که از این طریق سپاسگزاری می گردد.

تقدیم به مادر عزیزم که از سپاسگزاری زحمات بی دریغش قاصرَم.

و

تقدیم به خواهر عزیزم سیما، با آرزوی موفقیت های هر چه بیشتر.

در ابتدا لازم می‌دانم از استاد ارجمند جناب آقای دکتر محمدرضا میبیدی به پاس راهنمایی‌های ارزنده ایشان در جهت پیشبرد و انجام هرچه بهتر این پایان‌نامه تشکر و قدردانی نمایم.

از آقای علی برادران هاشمی که در انجام این پایان‌نامه من را یاری دادند، سپاسگزارم.

چکیده

در سال های اخیر، برای بهره برداری از حجم وسیع داده های وب روش های وب کاوی معرفی شده اند. وب کاوی، به کارگیری روش های داده کاوی برای کشف و استخراج خودکار اطلاعات از اسناد و سرویس های وب می باشد. یکی از انواع داده کاوی، داده کاوی ساختار وب است که با استفاده از پیوندها اطلاعات جدیدی راجع به صفحات به دست می آورد، اما پیوندها اطلاعات کافی راجع به ارتباط بین صفحات به دست نمی دهند. یک راهکار مناسب برای بهبود نتایج روش های داده کاوی ساختار وب، به کارگیری داده های استفاده از وب و نحوه پیمایش کاربران علاوه بر پیوندها در این روش ها می باشد. در این پروژه دو روش برای داده کاوی ساختار وب ارائه می شود که با استفاده از ترکیب پیوندها و داده های استفاده از وب اطلاعات جدید راجع به صفحات و ارتباطشان به دست می آورند. روش اول مبتنی بر اتوماتای یادگیر توزیع شده و روش دوم مبتنی بر اتوماتای یادگیر سلولی است. هر دو روش پیشنهادی از دو مرحله کلی تشکیل شده است. در مرحله اول، با استفاده از اتوماتای یادگیر (توزیع شده یا سلولی)، پیوندهای بین صفحات و رفتار کاربران در مشاهده صفحات وب، ساختار ارتباطی صفحات وب به دست می آید. به آن معنی که صفحات مرتبط با یکدیگر و میزان ارتباط آنها تعیین می شود. در مرحله دوم، ساختار ارتباطی به دست آمده از مرحله قبل، در دو نوع از کاربردهای داده کاوی ساختار وب استفاده خواهد شد. کاربرد اول پیمایش موضوعی صفحات وب و کاربرد دوم تشخیص اجتماعات وب است. همچنین کارایی ساختار به دست آمده، پیمایشگر طراحی شده و اجتماعات وبی که با استفاده از روش های پیشنهادی تشخیص داده می شوند، با روش های مشابه مقایسه شده و رفتار آنها در شرایط گوناگون مورد بررسی قرار می گیرد.

فهرست مطالب

۱	مقدمه	۱
۶	۱-۱- وب کاوی	۶
۸	۲-۱- داده کاوی ساختار وب	۸
۱۰	۱-۲-۱- اصطلاحات داده کاوی ساختار وب	۱۰
۱۱	۲-۲-۱- مدل های بازنمایی ساختار وب	۱۱
۱۱	۱-۲-۲-۱- مدل های مبتنی بر گراف	۱۱
۱۳	۲-۲-۲-۱- مدل های مارکو	۱۳
۱۴	۳-۲-۲-۱- مدل رابطه ای احتمالی	۱۴
۱۴	۳-۲-۱- معیارهای داده کاوی ساختار وب	۱۴
۱۴	۱-۳-۲-۱- معیارهای صفحه منفرد	۱۴
۱۵	۲-۳-۲-۱- معیارهای گروهی از صفحات	۱۵
۱۵	۳-۳-۲-۱- معیارهای کل گراف وب	۱۵
۱۶	۴-۲-۱- الگوریتم های داده کاوی ساختار وب	۱۶
۱۷	HITS -۱-۴-۲-۱	۱۷
۱۸	Page Rank -۲-۴-۲-۱	۱۸
۲۰	الگوریتم جریان بیشینه -۳-۴-۲-۱	۲۰
۲۰	Average Clicks -۴-۴-۲-۱	۲۰
۲۱	۵-۲-۱- کاربردهای داده کاوی ساختار وب	۲۱
۲۳	۳-۱- داده کاوی استفاده از وب	۲۳
۲۳	۱-۳-۱- منابع داده	۲۳
۲۴	۲-۳-۱- الگوریتم های داده کاوی استفاده از وب	۲۴
۲۵	۱-۲-۳-۱- قواعد انجمنی	۲۵
۲۵	۲-۲-۳-۱- الگوهای ترتیبی	۲۵
۲۶	۳-۲-۳-۱- خوشه بندی	۲۶
۲۷	۳-۳-۱- کاربردهای داده کاوی استفاده از وب	۲۷
۲۸	۴-۱- ترکیب داده کاوی ساختار وب و داده کاوی استفاده از وب	۲۸
۲۸	۱-۴-۱- الگوریتم PageRank آگاه از استفاده	۲۸
۲۹	۲-۴-۱- پیمایش در سایت وب	۲۹
۲۹	۳-۴-۱- جستجو در وب محدود	۲۹
۳۰	۴-۴-۱- HITS تغییر داده شده	۳۰
۳۰	۵-۱- اتوماتای یادگیر توزیع شده و سلولی	۳۰

۳۰	۱-۵-۱- اتوماتای سلولی
۳۱	۲-۵-۱- اتوماتای یادگیر
۳۳	۱-۲-۵-۱- الگوریتمهای یادگیری
۳۵	۲-۲-۵-۱- اتوماتای یادگیر با عمل های متغیر
۳۵	۳-۵-۱- اتوماتای یادگیر سلولی (CLA)
۳۷	۱-۳-۵-۱- اتوماتای یادگیر سلولی ناهمگام (ACLA)
۳۹	۴-۵-۱- اتوماتای یادگیر توزیع شده
۴۰	۶-۱- ساختار پایان نامه
۴۱	۲ تشخیص ساختار ارتباطی صفحات وب
۴۱	۱-۲- مقدمه
۴۳	۲-۲- تشخیص ساختار ارتباطی صفحات وب با استفاده از اتوماتای یادگیر توزیع شده
۴۹	۱-۲-۲- ارزیابی ساختار ارتباطی به دست آمده با استفاده از اتوماتای یادگیر توزیع شده
۴۹	۱-۱-۲-۲- مدل شبیه سازی
۵۱	۲-۱-۲-۲- معیار ارزیابی
۵۲	۳-۱-۲-۲- مقایسه با روش های موجود
۵۳	۴-۱-۲-۲- بررسی رفتار روش پیشنهادی
۵۶	۳-۲- تشخیص ساختار ارتباطی صفحات وب با استفاده از اتوماتای یادگیر سلولی ناهمگام
۶۳	۱-۳-۲- ارزیابی ساختار ارتباطی به دست آمده با استفاده از اتوماتای یادگیر سلولی
۶۳	۱-۱-۳-۲- مقایسه با روش های موجود
۶۴	۲-۱-۳-۲- بررسی رفتار روش پیشنهادی
۶۹	۴-۲- نتیجه گیری
۷۰	۳ پیمایش موضوعی صفحات وب
۷۰	۱-۳- مقدمه
۷۲	۲-۳- الگوریتم های پیمایش موضوعی
۷۲	۱-۲-۳- پیمایش موضوعی بدون استفاده از دانش پیش زمینه
۷۲	۱-۱-۲-۳- الگوریتم عرض اول
۷۳	۲-۱-۲-۳- الگوریتم BestFirst
۷۴	۳-۱-۲-۳- الگوریتم SharkSearch
۷۵	۴-۱-۲-۳- الگوریتم PageRank
۷۶	۲-۲-۳- پیمایش موضوعی به کمک دانش پیش زمینه
۷۶	۱-۲-۲-۳- پیمایش مبتنی بر طبقه بندی موضوعی
۷۷	۲-۲-۲-۳- پیمایش با استفاده از گراف زمینه

۷۸	۳-۲-۳-۳- پیمایش مبتنی بر هستان شناسی
۷۹	۳-۲-۳- سایر الگوریتم ها
۷۹	۳-۲-۳-۱- پیمایش هوشمند
۸۰	۳-۲-۳-۲- پیمایشگر یادگیر
۸۱	۳-۳- الگوریتم پیشنهادی برای پیمایش موضوعی
۸۵	۳-۳-۱- ارزیابی پیمایشگر موضوعی وب
۹۱	۳-۴- نتیجه گیری
۹۳	۴ تشخیص اجتماعات وب
۹۳	۴-۱- مقدمه
۹۴	۴-۲- الگوریتم های تشخیص اجتماعات وب
۹۵	۴-۲-۱- روش های مبتنی بر تحلیل پیوندها
۹۵	۴-۲-۱-۱- الگوریتم HITS
۹۶	۴-۲-۱-۲- الگوریتم RPA
۹۷	۴-۲-۲- روش های مبتنی بر تئوری گراف
۹۷	۴-۲-۲-۱- روش های مبتنی بر گراف کامل دوبخشی
۹۸	۴-۲-۲-۲- روش های مبتنی بر الگوریتم جریان بیشینه
۹۹	۴-۳- الگوریتم پیشنهادی برای تشخیص اجتماعات وب
۱۰۳	۴-۳-۱- ارزیابی اجتماعات وب به دست آمده با استفاده از مدل شبیه سازی
۱۰۶	۴-۳-۲- ارزیابی اجتماعات وب به دست آمده با استفاده از داده های واقعی وب
۱۰۶	۴-۲-۳-۱- مجموعه داده های واقعی وب
۱۰۸	۴-۲-۳-۲- پردازش داده های واقعی وب
۱۰۸	۴-۲-۳-۳- معیار ارزیابی
۱۰۹	۴-۲-۳-۴- آزمایش ها
۱۱۱	۴-۴- نتیجه گیری
۱۱۲	۵ نتیجه گیری
۱۱۵	۶ مراجع
۱۲۱	واژه نامه انگلیسی به فارسی
۱۲۶	واژه نامه فارسی به انگلیسی
۱۳۱	ضمائم
۱۳۱	الف- پیاده سازی

۱- مقدمه	۱۳۱
۲- مولفه شبیه ساز وب	۱۳۲
۳- مولفه تعیین ساختار ارتباطی	۱۳۲
۱-۳- ساختار مولفه	۱۳۳
۲-۳- عملکرد مولفه	۱۳۸
۴- مولفه پیمایش موضوعی	۱۳۹
۱-۴- ساختار مولفه	۱۳۹
۲-۴- عملکرد مولفه	۱۴۴
۵- مولفه تشخیص اجتماعات وب	۱۴۵
۱-۵- ساختار مولفه	۱۴۵
۲-۵- عملکرد مولفه	۱۴۸
ب- کد برنامه	۱۵۰

فهرست اشکال

- شکل ۱-۱- طبقه بندی وب کاوی و کاوش ساختار وب ۹
- شکل ۲-۱- مدل های گراف تک گره ای ۱۱
- شکل ۳-۱- مدل های گراف چند گره ای ساده ۱۱
- شکل ۴-۱- مدل های گراف چند گره ای پیچیده ۱۲
- شکل ۵-۱- مدل گراف وب ۱۲
- شکل ۶-۱- الگوریتم جریان بیشینه ۲۰
- شکل ۷-۱- ارتباط بین اتوماتای یادگیر و محیط ۳۲
- شکل ۸-۱- قانون ۵۴ ۳۷
- شکل ۹-۱- اتوماتای یادگیر توزیع شده ۳۹
- شکل ۱-۲- شبه کد الگوریتم به دست آوردن ساختار ارتباطی با استفاده از DLA ۴۸
- شکل ۲-۲- کورولیشن روش ارایه شده در [37] ۵۲
- شکل ۳-۲- کورولیشن روش پیشنهادی ۵۳
- شکل ۴-۲- مقایسه انواع کاربران در ساختار مبتنی بر DLA ۵۴
- شکل ۵-۲- حذف جریمه، پیوند و رابطه تراگذری ۵۶
- شکل ۵-۲- شبه کد الگوریتم به دست آوردن ساختار ارتباطی با استفاده از ACLA ۶۲
- شکل ۶-۲- مقایسه کورولیشن ساختار مبتنی بر DLA و ACLA ۶۴
- شکل ۷-۲- مقایسه انواع کاربران در ساختار مبتنی بر ACLA ۶۵
- شکل ۸-۲- بررسی تاثیر مقدار R بر کورولیشن ساختار مبتنی بر ACLA ۶۶
- شکل ۹-۲- مقایسه اتوماتای یادگیر سلولی همگام و ناهمگام ۶۷
- شکل ۱۰-۲- مقایسه استراتژی های فعال سازی سلول ۶۸
- شکل ۱-۳- الگوریتم عرض اول ۷۳
- شکل ۲-۳- الگوریتم BestFirst ۷۴
- شکل ۳-۳- الگوریتم SharkSearch ۷۵
- شکل ۴-۳- الگوریتم PageRank ۷۶
- شکل ۵-۳- شبه کد الگوریتم پیمایش موضوعی مبتنی بر CLA / DLA ۸۵
- شکل ۶-۳- مقایسه نرخ حاصل برای چهار پیمایشگر ۸۸
- شکل ۷-۳- مقایسه فراخوان هدف برای چهار پیمایشگر ۸۹
- شکل ۸-۳- وابستگی به مجموعه آغازین ۹۰
- شکل ۹-۳- تاثیر به کارگیری امتیاز Hub در پیمایشگر مبتنی بر DLA ۹۱
- شکل ۱۰-۳- تاثیر به کارگیری امتیاز Hub در پیمایشگر مبتنی بر CLA ۹۱

- شکل ۴-۱- الگوریتم تشخیص اجتماعات وب با استفاده از گراف کامل دوبخشی ۹۷
- شکل ۴-۲- مثالی از یک اجتماع وب (مجموعه گره های سمت چپ تصویر) ۹۸
- شکل ۴-۳- الگوریتم تشخیص اجتماعات وب با استفاده از الگوریتم جریان بیشینه ۹۸
- شکل ۴-۴- نحوه عملکرد الگوریتم جریان بیشینه ۹۹
- شکل ۴-۵- شبه کد الگوریتم تشخیص اجتماعات وب مبتنی بر CLA / DLA ۱۰۲
- شکل ۴-۶- مقایسه روش پیشنهادی با الگوریتم HITS ۱۰۴
- شکل ۴-۷- مقایسه روش پیشنهادی با روش مبتنی بر گراف کامل ۱۰۵
- شکل ۴-۸- تاثیر مجموعه ریشه در تشخیص اجتماع وب ۱۰۶
- شکل ۴-۹- مقایسه مدل شبیه سازی با داده های واقعی وب ۱۰۹
- شکل ۴-۱۰- مقایسه با الگوریتم HITS و الگوریتم [42] با استفاده از داده های واقعی وب ۱۱۰
- شکل ۱- مولفه های سیستم ۱۳۱
- شکل ۲- نمودار کلاس مولفه تعیین ساختار ارتباطی ۱۳۳
- شکل ۳- واسط کاربر برای تعیین ساختار ارتباطی ۱۳۹
- شکل ۴- نمودار دنباله برای مولفه تعیین ساختار ارتباطی ۱۴۰
- شکل ۵- نمودار کلاس مولفه پیمایش موضوعی ۱۴۱
- شکل ۶- واسط کاربر برای پیمایش موضوعی ۱۴۴
- شکل ۷- نمودار دنباله برای مولفه پیمایش موضوعی وب ۱۴۵
- شکل ۸- نمودار کلاس مولفه تشخیص اجتماعات وب ۱۴۶
- شکل ۹- واسط کاربر برای تشخیص اجتماعات وب ۱۴۸
- شکل ۱۰- نمودار دنباله برای مولفه تشخیص اجتماعات وب ۱۴۹

۱ مقدمه

وب، محیطی وسیع، متنوع و پویا است که کاربران متعدد اسناد خود را در آن منتشر می کنند. وب طی یک فرآیند آشفته و غیر متمرکز رشد می کند و این روند منجر به تولید حجم وسیعی از مستندات متصل به یکدیگر گشته است که از هیچ گونه سازماندهی منطقی برخوردار نیستند. در حال حاضر بیش از ۱۱،۵ بلیون صفحه [34] در وب موجود است و این تعداد با نرخ ۷،۳ میلیون صفحه در روز افزایش می یابد. با توجه به حجم وسیع اطلاعات در وب، مدیریت آن با ابزارهای سنتی تقریباً غیر ممکن است و ابزارها و روش هایی نو برای مدیریت آن مورد نیاز است. به طور کلی کاربران وب در استفاده از آن با مشکلات زیر روبرو هستند:

۱- یافتن اطلاعات مرتبط: یافتن اطلاعات مورد نیاز در وب دشوار می باشد. روش های سنتی بازیابی اطلاعات که برای جستجوی اطلاعات در پایگاه داده ها به کار می روند، قابل استفاده در وب نمی

باشند و کاربران معمولاً از موتورهای جستجو که مهمترین و رایج ترین ابزار برای یافتن اطلاعات در وب می باشند، استفاده می کنند. این موتورها، یک پرس و جوی^۱ مبتنی بر کلمات کلیدی از کاربر دریافت کرده و در پاسخ لیستی از اسناد مرتبط با پرس و جوی وی را که بر اساس میزان ارتباط با این پرس و جوی مرتب شده اند، به وی ارائه می کنند. اما موتورهای جستجو دارای دو مشکل اصلی هستند. اولاً دقت^۲ موتورهای جستجو پایین است، چراکه این موتورها در پاسخ به یک پرس و جوی کاربر صدها یا هزاران سند را بازیابی می کنند، در حالی که بسیاری از اسناد بازیابی شده توسط آنها با نیاز اطلاعاتی کاربر مرتبط نمی باشند. دوماً میزان فراخوان^۳ این موتورها کم می باشد، به آن معنی که قادر به بازیابی کلیه اسناد مرتبط با نیاز اطلاعاتی کاربر نیستند. چرا که حجم اسناد در وب بسیار زیاد است و موتورهای جستجو قادر به نگهداری اطلاعات کلیه اسناد وب، در پایگاه داده های خود نمی باشند.

۲- ایجاد دانش جدید با استفاده از اطلاعات موجود در وب: این مشکل در واقع بخشی از مشکل مطرح شده در قسمت قبل می باشد. در حال حاضر این سوال مطرح است که چگونه می توان داده های فراوان موجود در وب را به دانشی قابل استفاده تبدیل کرد، به طوری که یافتن اطلاعات مورد نیاز در آن به سادگی صورت بگیرد. همچنین چگونه می توان با استفاده از داده های وب به اطلاعات و دانشی جدید دست یافت.

۳- خصوصی سازی^۴ اطلاعات: از آن جا که کاربران متفاوت هر یک درباره نوع و نحوه بازنمایی اطلاعات سلیقه خاصی دارند، این مسئله باید مورد توجه تامین کنندگان اطلاعات در وب قرار بگیرد. برای این منظور با توجه به خواسته ها و تمایلات کاربران متفاوت، نحوه ارائه اطلاعات به آنها باید سفارشی گردد.

برای بهره برداری از حجم وسیع داده و کاهش مشکلات فوق الذکر، در سال های اخیر روش های وب کاوی^۵ معرفی شده اند. وب کاوی به کارگیری روش های داده کاوی^۶ برای کشف و استخراج خودکار اطلاعات از اسناد و سرویس های وب می باشد. البته روش های وب کاوی تنها ابزار موجود برای حل این مشکلات نیستند. بلکه روش های مختلفی از سایر زمینه های تحقیقاتی همچون پایگاه داده

¹ Query

² Precision

³ Recall

⁴ Personalization

⁵ Web Mining

⁶ Data Mining

ها، بازیابی اطلاعات، پردازش زبان طبیعی، ... قابل استفاده در این زمینه می باشند. همچنین وب کاوی با زمینه های مختلف تحقیقاتی علوم کامپیوتر همچون داده کاوی، پایگاه داده، بازیابی اطلاعات، هوش مصنوعی، یادگیری ماشین، پردازش زبان طبیعی، استخراج اطلاعات، انبار داده ها، طراحی واسط کاربر و ... در ارتباط تنگاتنگ است.

با توجه به گسترش روز افزون حجم اطلاعات در وب و ارتباط وب کاوی با تجارت الکترونیکی، وب کاوی به یک زمینه تحقیقاتی وسیع مبدل گشته است. تکنیک ها و روش های وب کاوی از کاربرد وسیعی در حوزه های مختلف همچون موتورهای جستجو، تجارت الکترونیکی، دولت الکترونیکی، آموزش الکترونیکی، آموزش از راه دور، سازمان های مجازی، مدیریت دانش، کتابخانه های دیجیتال، ... برخوردارند. اگرچه وب کاوی با چالش ها و محدودیت های متنوعی رو به رو است که از آن جمله می توان به داده های ناصحیح و نادقیق، حجم وسیع و روبه گسترش داده های وب، تغییر داده های وب در فواصل زمانی کوتاه، کاربران گوناگون با نیازهای مختلف، عدم وجود ابزارها و الگوریتم های مناسب اشاره کرد.

روش های وب کاوی بر اساس آن که چه نوع داده ای را مورد کاوش قرار می دهند، به سه دسته داده کاوی محتوای وب^۲، داده کاوی ساختار وب^۳ و داده کاوی استفاده از وب^۴ تقسیم می شوند. داده کاوی محتوای وب، فرآیند استخراج اطلاعات مفید از محتوای مستندات وب است. داده کاوی ساختار وب به کشف اطلاعات جدید با استفاده از پیوندهای^۵ بین صفحات وب می پردازد. داده کاوی استفاده از وب نیز داده های مربوط به استفاده کاربران از وب را مورد کاوش قرار می دهد و الگوهای استفاده از وب را به منظور درک و برآوردن بهتر نیازهای کاربران استخراج می کند.

بخش عمده ی فعالیت ها و تحقیقات انجام شده در وب کاوی به محتوای صفحات وب می پردازند. اما در سال های اخیر داده کاوی ساختار وب و داده کاوی استفاده از وب نیز مورد توجه قرار گرفته اند.

محور اصلی این پایان نامه داده کاوی ساختار وب می باشد. همان طور که گفته شد، الگوریتم های داده کاوی ساختار وب با استفاده از پیوندها اطلاعات جدیدی راجع به صفحات به دست می آورند. در

¹ Data Warehouse

² Web Content Mining

³ Web Structure Mining

⁴ Web Usage Mining

⁵ Hyperlink

این نوع از وب کاوی، وب به صورت یک گراف مدلسازی می شود که در آن صفحات وب، گره های گراف و پیوندهای بین صفحات، یال های گراف هستند. الگوریتم های داده کاوی ساختار وب در کاربردهای متفاوتی همچون رتبه بندی^۱ صفحات وب، تشخیص اجتماعات وب^۲، پیمایش صفحات وب^۳، تحلیل گراف وب، مدلسازی و شبیه سازی فرآیند تولید گراف وب به کار می روند. به عنوان مثال، الگوریتم HITS، صفحات مرتبط با یک موضوع و یا اجتماعات وب را تشخیص می دهد [47]. الگوریتم Page Rank، رتبه ای برای هر یک از صفحات محاسبه می کند که کیفیت صفحات را مستقل از یک موضوع خاص نشان می دهد [76]. الگوریتم جریان بیشینه^۴، قسمت های متراکم گراف وب را تشخیص می دهد و آن را به عنوان اجتماع وب معرفی می کند [42]. الگوریتم Fish Search برای پیمایش موضوعی صفحات وب^۵ به کار می رود و در حین پیمایش صفحاتی را جمع آوری می کند که مرتبط به یک موضوع خاص هستند [25].

روش هایی که در داده کاوی ساختار وب به کار می رود، تنها از پیوندها استفاده می کنند، در حالی که پیوندها اطلاعات کافی راجع به ارتباط بین صفحات به دست نمی دهند. چرا که بسیاری از پیوندهای بین صفحات، به دلایلی غیر از ارتباط صفحات، بین آنها ایجاد شده اند. از جمله این دلایل می توان به تسهیل پیمایش توسط کاربران و تبلیغات تجاری و ... اشاره کرد. به همین جهت به روش هایی در داده کاوی ساختار وب نیاز است که تاثیر مشکلات فوق الذکر را کاهش دهند.

یکی از راهکار مناسب برای بهبود نتایج روش های داده کاوی ساختار وب، به کارگیری داده های استفاده از وب علاوه بر پیوندها در این روش ها می باشد. در واقع نحوه پیمایش کاربران در وب و صفحاتی که مشاهده می کنند، می تواند اطلاعات مفیدی راجع به این صفحات به دست دهد. چرا که کاربران معمولاً از محتویات صفحه ای که می خواهند آن را در گام بعدی خود انتخاب کنند آگاهی نسبی دارند و بر اساس نیاز اطلاعاتی خود صفحه بعدی را انتخاب می کنند و حرکت کاربران در بین صفحات اتفاقی نیست. در واقع کاربر با استفاده از اطلاعات خود ارتباطی مجازی بین صفحات ایجاد کرده و آنها را مشاهده می کند و انتظار می رود که صفحات مشابه و مرتبط با یک موضوع با یکدیگر مورد استفاده

¹ Ranking

² Web Community

³ Crawling

⁴ Maximum Flow Algorithm

⁵ Focused Crawling

قرار گیرند. به این ترتیب با تحلیل داده‌های استفاده از وب و بدون تلاش مضاعف کاربر یا افراد خبره، اطلاعات با ارزشی راجع به صفحات و ارتباطشان بدست می‌آید.

در این پروژه دو روش برای داده کاوی ساختار وب ارایه می‌شود که با استفاده از ترکیب پیوندها و داده‌های استفاده از وب اطلاعات جدید راجع به صفحات و ارتباطشان به دست می‌آورند. روش اول مبتنی بر اتوماتای یادگیر توزیع شده^۱ و روش دوم مبتنی بر اتوماتای یادگیر سلولی^۲ می‌باشد. هر دو روش پیشنهادی از دو مرحله کلی تشکیل شده است. در مرحله اول، با استفاده از اتوماتای یادگیر (توزیع شده یا سلولی)، پیوندهای بین صفحات و رفتار کاربران در مشاهده صفحات وب، ساختار ارتباطی صفحات وب به دست می‌آید. به آن معنی که صفحات مرتبط با یکدیگر و میزان ارتباط آنها تعیین می‌شود. در مرحله دوم، ساختار ارتباطی به دست آمده از مرحله قبل، در دو نوع از کاربردهای داده کاوی ساختار وب استفاده خواهد شد. کاربرد اول پیمایش موضوعی صفحات وب و کاربرد دوم تشخیص اجتماعات وب است که در ادامه هر یک به اختصار معرفی می‌شود.

پیمایش وب یکی از وظایف موتورهای جستجو می‌باشد. پیمایش وب برای جمع آوری صفحاتی به کار می‌رود که توسط موتور جستجو شاخص گذاری^۳ می‌شوند. موتور جستجو برای پاسخ دادن به پرس و جوی کاربران از این صفحات پیمایش شده و شاخص گذاری شده استفاده می‌کند. پیمایشگرهای وب کار خود را با یک مجموعه اولیه از صفحات آغاز می‌کنند و با استفاده از پیوندهای موجود در این صفحات، صفحات دیگر را پیمایش کرده و این روند را تا رسیدن تعداد صفحات پیمایش شده به حدی مشخص انجام می‌دهند. از آن جا که حجم صفحات وب بسیار بالا و همواره رو به افزایش است، موتورهای جستجو قادر به شاخص گذاری صفحات محدودی هستند. به همین جهت در سال‌های اخیر به جای پیمایش کل وب، الگوریتم‌های پیمایش موضوعی یا متمرکز مورد توجه محققین قرار گرفته است که در حین پیمایش به صورت انتخاب گر عمل نموده و صفحات مرتبط با یک موضوع خاص را جمع آوری می‌کنند.

کاربرد دومی که در این پروژه مورد بررسی قرار خواهد گرفت تشخیص اجتماع وب است. اجتماع وب مجموعه‌ای از صفحات وب است که راجع به یک موضوع مشترک هستند و معمولاً توسط افراد یا سازمان‌های مختلف که علایق مشترک درباره یک موضوع خاص دارند، ایجاد می‌شوند. همچنین تعداد

^۱ Distributed Learning Automata

^۲ Cellular Learning Automata

^۳ Indexing

اتصالات صفحات یک اجتماع وب با یکدیگر بیش از تعداد اتصالاتشان با سایر صفحات وب است. تشخیص اجتماعات وب از این جهت که می تواند به کاربران در بازیابی اطلاعات از وب کمک کند، اهمیت ویژه ای دارد. با تشخیص یک اجتماع وب درباره یک موضوع خاص، کاربران می توانند با استفاده از صفحات اجتماع، اطلاعات مفیدی درباره آن موضوع به دست آورند.

علت آنکه روش های پیشنهادی برای داده کاوی ساختار وب به دو مرحله کلی تقسیم شده آن است که با انجام مرحله اول و به دست آوردن ساختار ارتباطی صفحات (صفحات مرتبط و میزان ارتباط آنها با یکدیگر)، می توان از ساختار به دست آمده در کاربردهای گوناگون داده کاوی ساختار وب استفاده کرد. الگوریتم های داده کاوی ساختار وب تنها به پردازش گراف وب می پردازند، اما روش های پیشنهادی در این پروژه در واقع بر یک گراف اصلاح شده عمل می کنند. همچنین این رویکرد امکان استفاده مجدد از اطلاعات به دست آمده راجع به صفحات را در کاربردهای داده کاوی ساختار وب فراهم می کند. به عنوان مثال پس از آن که ساختار ارتباطی تعیین شد، می توان از آن به دفعات متعدد در تشخیص اجتماع وب استفاده کرد. به این ترتیب روش های ارائه شده نسبت به سایر روش هایی که محاسبات و پردازش های خود را در تشخیص اجتماع وب و یا هر کاربرد دیگری تکرار می کنند، به صرفه تر می باشند.

در ادامه این فصل به معرفی وب کاوی، داده کاوی ساختار وب، الگوریتم ها و کاربردهای آن پرداخته می شود. همچنین از آنجا که در روش های پیشنهادی ترکیب پیوندها و داده های استفاده از وب به کار گرفته شده است، به الگوریتم ها و کاربردهای داده کاوی استفاده از وب و الگوریتم هایی که در آنها از ترکیب این دو شاخه از وب کاوی استفاده شده است، اشاره می شود. از آنجا که در الگوریتم های پیشنهادی در این پایان نامه از اتوماتای یادگیر توزیع شده و اتوماتای یادگیر سلولی استفاده شده است، این مفاهیم نیز در انتهای فصل اول معرفی می شود.

۱-۱- وب کاوی

در [48] وب کاوی به صورت زیر تعریف شده است: وب کاوی به کارگیری تکنیک های داده کاوی برای کشف و استخراج خودکار اطلاعات از اسناد و سرویس های وب می باشد. وب کاوی شامل چهار مرحله اصلی می باشد:

۱. پیدا کردن منبع: این مرحله شامل بازیابی صفحات وب مورد نظر می باشد.

۲. انتخاب اطلاعات و پیش پردازش: در این مرحله به صورت خودکار اطلاعات خاصی از صفحات بازیابی شده انتخاب و پیش پردازش می شوند.

۳. تعمیم^۱: در این مرحله به طور خودکار الگوهای عام در صفحات بازیابی شده کشف می شوند.

۴. تحلیل: در این مرحله الگوهای به دست آمده در مرحله قبل اعتبار سنجی^۲ و تفسیر می شوند.

در مرحله اول داده ها از منابع موجود در وب مانند خبرنامه های الکترونیکی، گروه های خبری، اسناد HTML، پایگاه داده های متنی و ... بازیابی می شوند. مرحله انتخاب و پیش پردازش شامل هر گونه فرآیند تبدیل داده های بازیابی شده در مرحله قبل می باشد. این پیش پردازش می تواند کاهش کلمات به ریشه آنها، حذف کلمات زائد، پیدا کردن عبارات موجود در متن و تبدیل بازنمایی داده ها به قالب رابطه ای یا منطق مرتبه اول باشد. در مرحله سوم از تکنیک های داده کاوی و یادگیری ماشین برای تعمیم استفاده می شود. در مرحله آخر، الگوهای به دست آمده ارزیابی می گردند.

به این ترتیب وب کاوی، فرآیند کشف اطلاعات و دانش ناشناخته و مفید از داده های وب می باشد. این فرآیند به طور ضمنی شامل فرآیند کشف دانش در پایگاه داده ها (KDD)^۳ نیز می شود. در واقع وب کاوی گونه توسعه یافته KDD است که بر روی داده های وب عمل می کند.

روش های وب کاوی بر اساس آنکه چه نوع داده ای استفاده می کنند، به ۳ دسته تقسیم می شوند [87]:

۱. کاوش محتوای وب: کاوش محتوای وب فرآیند استخراج اطلاعات مفید از محتوای مستندات وب است. محتوای یک سند وب متناظر با مفاهیمی است که آن سند در صدد انتقال آن به کاربران است. این محتوا می تواند شامل متن، تصویر، ویدئو، صدا و یا رکوردهای ساخت یافته مانند لیست ها و جداول باشد. در این میان کاوش متن بیش از سایر زمینه ها مورد تحقیق قرار گرفته است. از جمله این تحقیقات می توان به تشخیص موضوع^۴، استخراج قواعد انجمنی^۵، خوشه بندی^۶ و طبقه بندی اسناد وب اشاره کرد. روش ها و تکنیک های موجود در این گروه، از تکنیک های بازیابی اطلاعات و پردازش زبان

¹ Generalization

² Validation

³ Stemming

⁴ Stop Words

⁵ Knowledge Discovery in Data Base

⁶ Topic Discovery

⁷ Association Rule

⁸ Clustering