فصل اول

معرفی نامه

در این فصل به اختصار به معرفی فصول مختلف می پردازیم و با بیان اهداف کلی هر فصل مروری اجمالی بر کل مطالب پایان نامه خواهیم داشت:

معرفی فصل دوم پایان نامه:

۱-۱- روش های مختلف تقسیم بندی صفحات وب

در این بخش همانگونه که از عنوان آن پیداست به معرفی و مقایسه ی روش های متفاوت تقسیم بندی می پردازیم.این بخش شامل سه قسمت مجزاست که در هر قسمت هر یک از این روش ها به تفصیل مورد بحث قرار گرفته اند:

۱-۱-۱- تقسیم بندی صفحات وب بطور مختصر

با رشد سریع جهان پهنه ی گسترده ی وب نیاز فزاینده ای به فعالیت های گسترده در جهت کمک به کاربران صفحات وب برای طبقه بندی و دسته بندی این صفحات وجود دارد . این قبیل کمک ها که در سازمان دهی مقادیر زیاد اطلاعات که با سیستم های جستجو در ارتباط هستند و یا تشکیل کاتالوگ هایی که تشکیلات وب را سامان دهی می کنند ، بسیار مفید هستند.از نمونه های اخیر آن می توان یاهو و فرهنگ لغت looksmart (http://www.looksmart.com) که دارای کاربران زیادی هستند را نام برد.

شواهد نشان می دهد که طبقه بندی نقش مهمی را در آینده سیستم های جستجو بازی می کند. تحقیقات انجام شده نشان می دهد که کاربران ترجیح می دهند برای جستجو ، از کاتولوگ های از پیش دسته بندی شده استفاده کنند . از طرفی رسیدن به این قبیل نیاز های اساسی بدون تکنیک های خودکار دسته بندی صفحات وب و تحت ویرایش دستی و طبیعی توسط انسان بسیار مشکل است . زیرا با افزایش حجم اطلاعات طبقه بندی دستی بسیار وقتگیر و دشوار است.

در نگاه اول ، دسته بندی صفحات وب را می توان از برنامه های اداری دسته بندی متون مقتبس نمود.اگر بخواهیم در یک آزمایش ملموس تر به نتایج دقیق برسیم، راه حل مساله بسیار دشوار خواهد شد.صفحات وب ساختار اصلی و اساسی خود را در قالب زبان HTML قرار می دهند که آنها شامل محتویات پر سروصدا مانند تیترهای تبلیغاتی و یا هدایت گرهای راهنما و غیر متنی هستند.اگز روش های خاص طبقه بندی ویژه متون برای این صفحات بکار گرفته شود چون متمایل به یک الگوریتم خاص دسته بندی متون است ، گمراه شده و باعث از دست رفتن تمرکز بر موضوعات اصلی و محتویات مهم می شود.زیرا این محتویات فقط شامل متن نیست.

پس وظیفه و هدف مهم ما طراحی یک کاوشگر هوشمند برای جستجوی مطالب مهم صفحات وب است که هم شامل اطلاعات متنی و هم سایر اطلاعات مهم باشد.در این مقاله ما نشان می دهیم که تکنیک های خلاصه سازی صفحات وب برای جستجوگر ها دسته بندی صفحات وب ، تکنیکی کاربردی و بسیار مفید است. ما همچنین نشان می دهیم که به جای استفاده از تکنیک های خلاصه سازی در فضای وب که عمدتا برای متون طراحی شده ، می توان از برنامه ها و تکنیک های خاص خلاصه سازی صفحات وب استفاده کرد. به منظور جمع آوری شواهد قابل ملموسی که نشان دهیم تکنیک های خلاصه سازی در صفحات وب سودمند هستند، ما ابتدا یک آزمایش موردی ایده آل را بررسی می کنیم که در آن هر صفحه وب ، با خلا صه این صفحه، که توسط انسان خلاصه شده جابجا شده است.پس از انجام این آزمایش در می یابیم که در مقایسه با استفاه از متن کامل صفحه وب ، حالت خلاصه شده رشد چشمگیر ۱۴،۸ درصدی داشته است که پیشرفت قابل ملاحظه ای شمرده می شود.به علاوه در این مقاله ما یک تکنیک جدید خلاصه سازی صفحات وب را پیشنهاد می کنیم که این روش موضوعات اصلی صفحات وب را با روش آنالیز لایه ای صفحات برای بالا بردن دقت دسته بندی استخراج می کند.

سپس عملیات دسته بندی را به همراه الگوریتم اجرای آن ارزیابی می کنیم و آن را با روش های سنتی دسته بندی خودکار متون که شامل روش های نظارتی و غیر نظارتی می باشد مقایسه می کنیم.در آخر ما نشان می دهیم که یک اسمبل از روش خلاصه سازی حدود ۱۲,۹ درصد پیشرفت را می تواند حاصل کند که این عدد بسیار نزدیک به حدود بالایی است که ما در آزمایش ایده آل خود به آن دست یافتیم.

نتیجه کلی این مقاله این است که جستجو گر هایی که فقط برای متون طراحی شده اند در حالت کلی گزینه مناسبی برای جستجو در فضای وب نیستند و ما نیاز به برنامه هاو جستجو گر هایی داریم که صفحات وب را در لایه های مختلف و همچنین سطوح متفاوت بررسی و جستجو کنند. لذا استفاده از مدل هایی که روش های خلاصه سازی و دسته بندی را بادقت بیشتری انجام می دهند، سرعت و دقت جستجو را افزایش خواهد داد.

۱-۱-۲- تقسیم یندی صفحات وب با استفاده از الگوریتم اجتماع مورچه ها

در این بخش هدف کشف کردن یک مجموعه خوب قوانین تقسیم بندی به منظور رده بندی کردن صفحات وب بر اساس موضوعات آنهاست. الگوریتم استفاده شده در این فصل الگوریتم اجتماع مورچه(اولین الگوریتم بهینه سازی اجتماع مورچه) برای کشف قوانین تقسیم بندی در زمینه ی استخراج مضامین وب می باشد. همچنین مزایا و معایب چندین تکنیک پیش پردازش متنی بر اساس زبان شناسی را به منظور کاهش مقدار زیادی از علائم و نشان های به هم پیو سته با استفاده از استخراج مضامین وب بررسی می کند.

نگهداری صفحات وب بسیار چالش پذیر تر است.زیرا شامل متون غیر سازمان یافته و یا نیمه سازمان یافته بسیاری در صفحات وب یافت می شود. به علاوه تعداد زیادی از لغات و خصوصیات در رابطه با

صفحات وب بالقوه موجود است . و یک تحلیل تئوری از الگوریتم مورچه (تحت یک نگاه بدبینانه) نشان می دهد که زمان محاسباتی شدیدا به مقدار توصیفات و خصوصیات حساس است . پس استنباط اینکه این الگوریتم در رابطه با مجموعه داه هایی که در عمل خصوصیت های زیادی دارند و همچنین در چالش با دنیای وب و نگهداری وب ها چگونه مقیاس بندی می کند ، از اهمیت فراوانی برخوردار است.

در آخر تحقیق در مورد اینکه تکنیک های مختلف جستجوی متون که توصیفات و خصوصیات آنها رو به افزایش است ، چه تاثیری بر عملکرد الگوریتم خواهد گذاشت دارای اهمیت می باشد.

نتیجه کلی این مقاله این است که باافزایش اطلاعات صفحات وب جهت سهولت در برداشت و جستجو نیازمند دسته بندی و طبقه بندی آنها هستیم.برای دسته بندی نیاز به یک الگوی مناسب وجود دارد که این انتخاب الگو نیز به نوبه خود نیازمند قواعد کلی و مناسب ا ست.قواعد شامل مقدمه ها و نتایج هستند که مارا در جهت ایجاد الگوی مناسب برای دسته بندی یاری می دهند.

هدف ما دسته بندی اطلاعات بر حسب موضوع است که نباید به صورت جزئی و خاص این مهم را انجام داد ، بلکه دسته بندی مناسب و معقول باید عمومی ، مفید و جامعه نگر باشد.

۱-۱-۳- تقسیم بندی صفحات وب براساس ساختارپوشه ای

اخیرا در حجم داده های موجود در web یک افزایش نمایی وجود دارد. بر این اساس ، تعداد صفحات موجود در web در حدود ۱ میلیارد است و روزانه تقریبا ۱,۵ میلیون به آن اضافه می شود. این حجم وسیع داده علاوه بر تاثیرات متقابل ،وب رابه شدت مورد توجه عامه مردم قرار داده است.

در هر حال ، در مواردی چون اطلاعات ، محتویات و کیفیت تا حدود زیادی با یکدیگر تفاوت دارند. به علاوه ، سازمان این صفحات اجازه یک تحقیق ساده را نمی دهد. بنابراین ، یک روش دقیق و موثر برای

دسته بندی این حجم از اطلاعات برای بهره برداری از تمام قابلیت های وب بسیار ضروری است. این ضرورت مدت زیادی است که احساس شده است و رویکردهای مختلفی برای حل این مشکل پیشنهاد شده است.

برای شروع ، دسته بندی توسط متخصصین شبکه جهانی به صورت دستی انجام شد. اما خیلی سریع ، دسته بندی به صورت اتوماتیک ونیمه اتوماتیک در آمد. تعدادی از رویکردهای مورد استفاده شامل دسته بندی متن بر اساس الگوریتم های آماری است ، رویکرد K-نزدیکترین همسایه ، یادگیری قوانین القایی ، در خت های تصمیم ، شبکه های عصبی و ماشین های برداری پشتیبان ، از جمله این موارد می باشند. تلاش دیگری که در این زمینه صورت گرفت ، دسته بندی محتویات وب بر اساس ساختمانی وراثتی است.

به هر حال ، علاوه بر محتویات متن در صفحات وب ، تصاویر ، نمایش ها و دیگر موارد رسانه ای در کنار هم و در تعامل با ساختمان متن ، اطلاعات زیادی را برای دسته بندی صفحات می دهند.

الگوریتم های دسته بندی موجود که به تنهایی روی محتویات متن برای دسته بندی ، تکیه دارند ، از این جنبه ها استفاده نمی کنند. به تازگی با رویکردی اتوماتیک بر اساس جنبه ای برای دسته بندی صفحات وب روبرو شده ایم.

ما یک رویکرد برای دسته بندی اتوماتیک صفحات وب توصیف کرده ایم واز تصاویر و ساختمان صفحه برای دسته بندی استفاده می کند.نتایج حاصله کاملا امیدوار کننده است . این رویکرد می تواند در کنار دیگر رویکرد های مبتنی بر متن توسط موتور های جسنجو گر برای دسته بندی صفحات وب ، مورد استفاده قرار گیرد .

عملیات جاری ما روشی را برای دسته بندی استفاده می کند که در آن وزن اختصاص یافته به هر جنبه به طور دستی چند جنبه ابتکاری دیگر (مانند قرار دادن یک صفحه به عنوان صفحه ی خانگی) می تواند دقت دسته بندی را افزایش دهد. در حال حاضر ، ما تنها از تصاویر علاوه بر اطلاعات ساختمان صفحات استفاده کرده ایم و از جنبه هایی چون صوت و نمایش استفاده نکرده ایم.

معرفی فصل سوم پایان نامه:

۱-۲- جستجوی وب با استفاده از طبقه بندی خودکار

پهنه مرزی جستجوی مساعد و مفید کاربر برای صفحات وب هنوز یکی از مهمترین مبارزه طلبی ها درسهل نمودن آن برای عموم می باشد و در حقیقت همه ابزارهای جستجوی اخیر هر یک از ریزه کاری های نا چیز یا فراخوانی ناچیز دستخوش تغییر می شوند.

ما این مسئله را در این فصل با گسترش پهنه مرزی جستجوی که به طبقه بندی خودکار صفحات وب وابسته است مورد توجه قرار داده ایم. تقسیم بندی ما متکی بر علم رده بندی یاهو! می باشد اما از این نظر که آن خودکار می باشد و توانایی در برگرفتن سریع تر همه عظمت وب را در قبال علم رده بندی یاهو! دارد با هم متفاوتند. اعتبار آزمایشات طبقه بندی ما در این فصل اطلاعات جانبی را در زمینه قدرت طبقه بندی خودکار ارائه می کند.

همچنین جستجوی مجدد ما نشان می دهد که تقسیم بندی وب و ابزار های جستجو باید برای مهارت هایی نظیر تشخیص هرزنامه ی وب که از موجودی های چنین ابزار هایی نتیجه شده اند ، پاداشی در نظر بگیرند.

تهیه یک روش تحقیق و جستجو موثر و مطلوب در وب همچنان یکی از چالش های مهم برای در دسترس عموم قرار دادن آن است . تصور کنید که شما به عنوان یک جستجو گر می خواهید وزن متوسط یوز پلنگ رابدانید. اگر شما تصمیم بگیرید که بوسیله ی کلمات کلیدی "یوز پلنگ" و " وزن " جستجو را انجام دهید ، تقریبا ۹۰۰ متن مطابق با کلمات مورد نظر را خواهید یافت . اما متا سفانه شما به سرعت آن جواب مورد نظر را نخواهید یافت . نتایج جستجو با صفحات زیادی که شامل " ماشینهای جاگوار " ، "آتاری جاگوار " به عنوان یک سیستم بازی خانگی ، و احتمالا حتی تیم فوتبال " جاگوار" ، همراه خواهد شد. از این ۹۰۰ صفحه ، یافته ایم که بالاترین متن موجود در لیست که شامل اطلاعات مورد نظر ما است ، متن ۴۱ می باشد . (وزن متوسط جنس نر ، بین ۱۲۵ تا ۲۵۰ پوند است.)

حال سوال این است که آیا ما کم وبیش می توانیم به یک موتور جستجو گر مانند Alta Vista بگوییم که جستجوی ما با این کلمات کلیدی تنها باید محدود به متون مربوط به جانور شناسی و یا موارد دیگری از علوم باشد؟

یک رویکرد برای محدود کردن جستجو استفاده از یک شاخه مانند یاهو! است . متاسفانه این موارد تنها در بخش کوچکی از وب پوشش داده شده اند. در واقع ، تمامی ابزار های جستجو گر موجود در حال حاضر از دو مشکل دقت پایین (به معنای تعداد بیش از اندازه متن های بدون ارتباط) و فراخوانی ضعیف (به این معنی که قسمت کوچکی از وب توسط این ابزار پوشش داده شده است.) رنج می برند.

ما بر این مورد یا توسعه یک جستجو که بر دسته بندی اتوماتیک صفحات وب تکیه دارد ، تاکید می کنیم. دسته بندی ما در یاهو! یک طبقه بندی علمی را می سازد ، اما با این تفاوت که اتوماتیک است و بنابراین این قابلیت را دارد که تمامی وب را تحت پوشش قرار دهد .