دسته بندي معنايي صفحات وب به كمك آنتولوژي

فصل اول

مقدمه

1-1 مقدمه

با افزایش روز افزون منابع متنی در شبکه جهانی وب، هر روز بر گستردگي اطلاعات قابل دسترس براي کاربران افزوده می­شود. اگر چه این رشد سریع مزایاي فراوانی دارد، اما مشکلاتی نیز به همراه داشته است. نخستین چالش اين است که کاربر چگونه می­تواند اطلاعات مورد نیاز خود را بیابد. امروزه، بکارگیري سامانه­هاي بازیابی اطلاعات یکی از عمومی­ترین روش­هاي جستجو و کسب اطلاعات مورد نیاز است. خروجی یک سامانه­هاي بازیابی اطلاعات معمول، به صورت سیاهه­اي از عنوانها است که هر کدام با توضیح کوتاهی همراه شده است. بدبختانه، در بیشتر موارد این سیاهه­ها بسیار بلند هستند و بررسی همه­ي سندهاي بازیابی شده عملی نیست. به طور معمول، کاربران تنها چند ده سند اول را بررسی می­کنند و بقیه را نادیده می­گیرند. بیشتر کاربران گرایش به طرح پرس­وجوهاي کوتاه دارند. ]1[ بنابراين بسیاري از پرس­وجوها نادقیق و مبهم هستند ]2.[ به این ترتیب، کاربران با چالش دیگري مواجه می­شوند: چگونه اطلاعات مفید را گزینش کنند.

1-2 هدف اصلي دسته­بندي صفحات وب

روش­هاي دسته­بندي خودکار صفحات وب، یک راه حل مناسب براي مشكل جستجوي مناسب در صفحات وب است. اگر صفحات وب به شكل مناسب دسته­بندي شوند، موتورهاي جستجو با دقت بيشتري مي­توانند نتايج مورد نظر كاربران را استخراج كنند و اين موتورهاي جستجو يك قدم به سمت موتورهاي جستجوي معنايي نزديك­تر مي­شوند.

در حقیقت، صفحات وب یک ساختار نیمه منظم دارند و بر اساس سليقه طراح و صاحب وب­سايت مي­توانند ساختارهاي متنوعي داشته باشند. که اطلاعات در آن با استفاده از تگ‏هاي اچ تی ام ال نمایش داده می شود و جدا سازی تگها از اطلاعات جز اولین گام­های طبقه­بندی صفحات وب است. همچنین هر تگ مفهوم و جایگاه ویژه‏ای دارد و تفاوت قائل شدن بین این تگ‏ها و محتوی آنها ضروری به نظر می‏رسد ]2[.

1-3 خلاصه­اي از ساختار روش پيشنهادي

چالش اصلی این مسئله مشخص نمودن نوع و دسته یک صفحه وب است. الگوریتم دسته بندی ابتدا لازم است از روی یک پایگاه داده مشخص شروع به یادگیری کند و سپس به طبقه بندی صفحات جدید بپردازد. در اینجا نحوه آموزش الگوریتم و سپس طبقه بندی صفحات جدید با توجه به اطلاعات موجود جز چالش های اصلی این مسئله می باشد ]2[.

در حقیقت الگوریتم برنامه باید به صورتی باشد که با یادگیری ویژگی­های هر صفحه بتواند با ویژگی های خاص هر کلاس را پیدا کرده و سپس با توجه به میزان شباهت هر صفحه وب جدید با صفحات وب قدیمی تر که کلاس آنها مشخص است، کلاس جدید را برای صفحه مشخص نماید. همانطور که مشخص است این مسئله از نوعيادگيري باناظر[[1]](#footnote-1) است که در آن از یک سری داده از قبل مشخص شده برای آموزش الگوریتم استفاده می گردد.

برای انتساب هر صفحه جدید به کلاس های لازم است تا ابتدا الگوریتم یک فاز یادگیری را اجرا نماید و سپس با توجه به آموخته های خود در فاز یادگیری، صفحات جدید را طبقه بندی نماید ]5[.

برای تمایز صفحات از یکدیگر از تگ ها اچ تی ام ال آن و محتوی درون این تگ ها استفاده می شود. همچنین تمایز قائل شدن بین تگ های هر صفحه جز ویژگی های این روش است و باید الگوریتم پیشنهادی علاوه بر پیماش تگ های صفحات، بین آنها تمایز قائل شود و ترتیب اولیت بین آنها را مشخص نماید. اما يكي از مشكلات اصلي html اين است كه اين زبان يك زبان مفهومي نيست و فقط قابل درك توسط انسان است و در واقع ماشين نمي­تواند مفهوم را از آن درك كند لذا يك دسته­بندي دسته­بندي مفهومي از صفحات وب نمي­توان بر اساس صفحات html انجام داد. البته اگر صفحات وب بر اساس xml يا يك سطح بالاتر، بر اساس RDF يا OWL بودند با همچين مشكلي روبرو نبوديم. به همين دليل در اين پژوهش سعي مي­شود به كمك آنتولوژي مفهوم را به صفحات وب اضافه كنيم و دسته­بندي انجام گرفته مفهومي باشد ]8[.

در اغلب روش­هاي ديگر دسته­بندي صفحات وب در دو مرحله انجام مي­گيرد: مرحله اول انتخاب وبژگي­ها[[2]](#footnote-2) و مرحله دوم دسته­بندي صفحات بر اساس اين ويژگي­ها است. در مرحله اول بايد تعدادي ويژگي مناسب از صفحات وب مانند عنوان صفحه، تگ­هاي مهم و غيره انتخاب مي­شود و بر اساس اينكه محتواي اين ويژگي­ها چه چيزي باشد صفحات را به دسته­هاي مختلف تقسيم­بندي مي­كنند. اما هنگام استفاده از اين روش قطعاً از همه محتواي صفحه استفاده نمي­شود و صرفاً با درنظر گرفتن چند ويژگي دسته­بندي انجام مي­گيرد. در روش پيشنهادي سعي مي­شود تا حد امكان از تمام محتواي صفحه براي دسته­بندي استفاده كرد ]9[.

در روش­هاي انجام گرفته قبلي يك ديتاست با مجموعه­اي از صفحات انتخاب مي­شود كه اين ديتاست شامل چند گروه محدود از صفحات هستند. توسط الگوريتم­هايي مانند svm، دسته­بندي بيزين، knn و ... مرحله يادگيري انجام مي­گيرد ]2[ و موقع آزمايش روش نيز همين چند گروه محدود از صفحات آزمايش مي­شود كه اين كارايي روش را زير سوال مي­برد. اما در روش پيشنهادي سعي بر اين است كه صفحات انتخاب شده شامل دسته­هاي نامحدود باشند. همچنين در روش پيشنهادي يادگيري وجود ندارد.

همان­طور كه اشاره شد در روش پيشنهادي سعي بر اين است كه يك روش دسته­بندي مفهومي به كمك آنتولوژي انجام بگيرد. هدف استفاده از آنتولوژي اين­ است كه ماشين­ها (كامپيوترها و بقيه وسايل) نيز بتوانند مانند انسان صفحات وب را بفهمند و آن­ها را تجزيه و تحليل كنند. از همه مهمتر با استفاه از آنتولوژي مفهوم به صفحات وب اضافه مي­شود. آنتولوژي مي­تواند ابزاري براي طبقه­بندي بهتر صفحات وب باشد. زيرا اطلاعاتي در مورد كلاس­ها، سلسله­مراتب و رابطه بين آن­ها را مشخص مي­كند. با توجه به اين اطلاعات مي­توان دسته­بندي را دقيق­تر انجام داد.

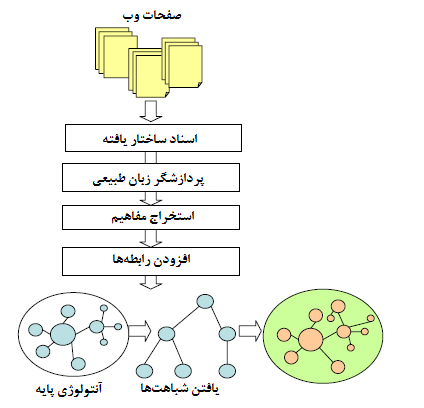
در ابتدا توسط تكنيك­هاي متن­كاوي كلمات مهم صفحه مورد نظر را استخراج مي­كنيم. سپس از يك آنتولوژي دامنه يا آنتولوژي عمومي يا آنتولوژيي كه خودمان ساخته­ايم براي دسته­بندي مفهومي صفحات استفاده مي­كنيم. اما نحوه استفاده از آنتولوژي به اين صورت است كه در ابتدا يك آنتولوژي در مورد دامنه مورد نظر به كمك نرم­افزار protégé طراحي مي­كنيم سپس از روش­هاي انطباق و يافتن ميزان شباهت براي دسته بندي استفاده مي­كنيم.

*برای محاسبه میزان شباهت مفاهیم استخراج شده با مفاهيم موجود در آنتولوژي، نحوه دسترسی به آنها از گره ریشه در نظر گرفته می­شود. محاسبه میزان شباهت مفاهیم، درجه اشتراک آنها را مشخص می­کند. برای محاسبه میزان شباهت از رابطه زير استفاده می­شود.*

**در این رابطه یک مقدار بین صفر و یک دارد که میزان تاثیر عمومی بودن را مشخص می­کند. (x)α مشخص کننده مجموعه گره­هایی است که در ساختار درختی آنتولوژی بالای گره x قرار دارند.(x)∩ α(y)α مشخص کننده عناصر مشترک بالای دو مفهوم است که در واقع نقطه اشتراک دو مفهوم x و y را مشخص می­کند.



ساختار كلي روش پيشنهادي در شكل 1 نشان داده شده است.



شكل 1) ساختار كلي روش پيشنهادي

**1-4 فرضیه ها**

به منظور ايجاد يك دسته­بند مفهومي يك آنتولوژي دامنه ساخته مي­شود و از ساختار سلسله مراتبي آن براي يافتن ارتباطات بين مفاهيم استفاده مي­شود.

در ابتدا صفحات html به صفحات txt تبديل مي­شوند سپس كلمات ايست شناسايي شده و حذف مي­شوند و كلمات با ارزش را شناسايي كرده و ريشه­هاي آنها نيز توسط الگوريتم پورتر استخراج مي­شوند.

به منظور دسته­بندي مفهومي­تر از آنتولوژي wordnet براي ايجاد مفاهيم عمومي استفاده مي­شود.

**1-5 اهداف تحقیق**

* بهبود مرتبه زمانی الگوریتم های طبقه بندی صفحات وب
* افزایش تعداد طبقه بندی های درست
* کاهش تعداد دسته بندی های اشتباه
* افزایش دقت موتور های جستجو با بهبودز crawler هر موتو حستجو
* بهبود پایگاه داده سیستم ها دانش وابسته به پایگاه دانش
* بهینه‏تر نمودن وب دایرکتوری ها
* نیاز به سازماندهی و طبقه بندی صفحات وب
* نیاز به ارائه یک الگوریتم با امکان پشتیبانی از سیستم های طبقه‏بندی یک دسته­ای و چند دسته­ای
* عدم پشتیبانی الگوریتم های ارائه شده برای طبقه‏بندی صفحات وب از مسائل چند هدفه

1-6 نوع روش تحقیق:

در اين تحقيق براي ساختن آنتولوژي از نرم­افزار protégé استفاده مي­كنيم. همچنين نرم­افزار wordnet2.1 براي يافتن مفاهيم عمومي استفاده مي­شود. بقيه­ي كارهاي برنامه­نويسي توسط زبان جاوا و در محيط Eclipse انجام مي­گيرد. همچنين از پكيج jena براي كار با آنتولوژي و از پكيج jwi براي كار با wordnet استفاده مي­شود.

فصل دوم

كليات پژوهش

2-1 مقدمه

با افزایش حجم اطلاعات و محتوی در فضای وب، نیاز به طبقه بندی و وجود یک ساختار مشخص که امکان جستجو در این محتوی را فراهم آورد، به شدت حس می شود. از این رو هدف مسئله طبقه بندی صفحات وب طبقه بندی و دسته بندی صفحات وب در گروه‏ها، دسته‏ها و حتی زیر مجموعه های خاصی است. این طبقه بندی امکان جستجو و یا حتی زمینه یافتن میزان ارتباط صفحات وب را با یکدیگر فراهم می آورد. از طرف دیگر طبقه بندی صفحات وب یک تفاوت بزرگ با طبقه بندی اسناد متنی دارد، و آن پویایی این محیط و غیر قابل کنترل بودن محتوی وب است.

2

فصل سوم

معرفي زمينه­هاي تحقيقاتي مرتبط و كارهاي پيشين

3-1 مقدمه

با توجه به اينكه در اين پژوهش از مفهوم آنتولوژي و وب معنايي به عنوان ابزاري جديد براي دسته­بندي معنايي صفحات وب استفاده مي­كنيم در اين فصل ابتدا اين تكنولوژي به صورت كامل توضيح داده مي­شود و در ادامه كارهاي قبلي كه براي دسته­بندي صفحات وب انجام گرفته­اند توضيح داده مي­شوند.

3-2 وب معنايي

آقاي تيم برنرزلي مبتكر صفحات وب با معرفي تكنولوژي وب 3 سعي داشت كه معنا و مفهوم را به صفحات به شكلي اضافه كند كه اين صفحات براي ماشين­ها نيز قابل درك باشند و ماشين­ها نيز براحتي بتوانند محتويات صفحات را درك كنند و در آنها جستجو كنند. اين تكنولوژي تحت عنوان وب معنايي[[3]](#footnote-3) معرفي مي­شود.

واژه وب معنايي براي اولين بار در کتاب تنيدن وب تيم برنرزلي در سال 1999 معرفي شد [16]. بنابراين سال 1999 به­عنوان سال تولد وب معنايي مشخص شد. يک انتشار مهم ديگر براي وب معنايي [17] است. از سال 2001 تاکنون تفکر وب معنايي توسط محققان دانشگاهي، مخترعان صنعتي، و حتي مردم زيادي که زمينه­اي از علم کامپيوتر ندارند، مورد پذيرش واقع شده است. منظور وب معنايي قابل درک کردن وب براي ماشين است. به­طور کلي، سه سطح از کپسوله­سازي دانش براي وب معنايي وجود دارد. سطح اول کپسوله سازي اطلاعات نحوي درباره دانش با استفاده از XML و RDF(S) است. سطح دوم کپسوله سازي اطلاعات معنايي درباره دانش با استفاده از آنتولوژي، از طريق زبان­هاي آنتولوژي از قبيل OWL است. سومين سطح از استنتاج و تکنيک­هاي امنيتي براي فراهم کردن روش دست­کاري و محافظت دانش استفاده مي­کند. در ادامه اين بخش ابتدا تعريفي از وب معنايي ارائه مي شود، سپس ساختار و اجزاي وب معنايي مورد بررسي قرار مي‌گيرند.

وب فعلي، ابزاري براي نمايش و ارائه­ي اطلاعات است و از کامپيوترها فقط به‌عنوان ابزاري براي جمع‌آوري و نمايش اطلاعات استفاده مي‌کند، اما وظيفه­ي درک و تفسير اين اطلاعات را به عهده­ي انسان‌ها قرار مي‌دهد زيرا اين اطلاعات براي ماشين‌ها بي‌معني هستند، چون 1) اطلاعات به­زبان طبيعي ارائه شده‌اند و 2) ماشين‌ها از اطلاعات پيش­زمينه آگاهي ندارند. موارد بسياري وجود دارند که به­وسيله­ي وب نحوي قابل انجام نيستند. به­عنوان مثال يکي از مسائلي که همگي ما آن را در وب تجربه کرده‌ايم، جست­وجوهايي است که توسط جست­وجوگرهاي مبتني بر وب نحوي انجام مي‌گيرد. از آن­جا که در وب نحوي، معنايي وجود ندارد. لذا همواره نتايجي غير از نتايج دلخواه به­عنوان نتايج جست­وجو به ما ارائه مي‌شود.

با توجه به گسترش روزافزون اطلاعات در وب، نياز به استفاده­ي مؤثرتر از ماشين‌ها براي درک، تفسير و انجام اعمال بر اساس اين اطلاعات، هر روز بيش از پيش احساس مي‌شود. از اين رو، وب معنايي را مي‌توان به‌عنوان بستري براي فراهم آوردن چنين امکاناتي در نظر گرفت.

3-2-1 ساختار وب معنايي

فصل چهارم

معرفي روش پيشنهادي

4-1 مقدمه

در دهه اخیر شاهد توسعه بیش از پیش شبکه جهانی اینترنت و محتوای این شبکه بوده­ایم که این امر منجر به افزایش اهمیت سازماندهی و طبقه­بندی اطلاعات وب گردیده است. دسته­بندی و طبقه­بندی اسناد و صفحات وب جز نیازهای اولیه یک موتور جستجو به حساب می­آید. در حقیقت در هر موتور جستجو Crawler وظیقه جستجو در وب و استخراج صفحات وب و اسناد موجود در شبکه را دارد و در پس زمینه اجرای یک الگوریتم کارآمد موجب دسته­بندی و طبقه­بندی اسناد استخراج شده می­گردد. با طبقه­بندی و دسته­بندی اسناد می­تواند به راحتی در آنها جستجو معنایی نمود و حتی ارتباط اسناد و صفحات را با یکدیگر مشخص نمود. در اين فصل سعي مي­شود مراحل روش پيشنهادي با جزئيات كافي معرفي شوند.

4-2 روش پيشنهادي

در اين پژوهش سعي مي­شود يك روش دسته­بندي معنايي صفحات وب مبتني بر آنتولوژي دامنه معرفي شود. براي بررسي مجموعه آموزشي ابتدا بايد پيش­پردازش­هايي توسط تكنيك­هاي داده­كاوي و متن كاوي انجام بگيرد كه بعداً بتوانيم نتايج دلخواهي را بدست آوريم. توسط تكنيك­هاي موجود لغاتي كه از بهره اطلاعاتي بيشتري برخوردار هستند شناسايي مي­شوند و در نهايت توسط آنتولوژي دامنه لغات مرتبط با آنها نيز شناسايي مي­شوند كه در نتيجه دسته­بندي تاثير مثبتي خواهند داشت. در ادامه مراحل روش پيشنهادي يكي يكي بررسي مي­شوند. در شكل 4-1 ساختار كلي روش پيشنهادي نشان داده شده است.

شكل 4-1) ساختار روش پيشنهادي

تبديل كننده

تجزيه كننده جمله

بسته بند

ريشه يابي

نرمال سازي بردار

حذف كلمات ايست

ايجاد بردار

انتخاب كلمات با حروف بزرگ

انتخاب كلمات مصور

ايجاد ارتباط معنايي بين كلمات

دسته­بندي صفحات توسط كلمات استخراجي

فصل پنجم

پياده­سازي و ارزيابي روش پيشنهادي

5-1 مقدمه

در زندگي امروزي انسان بيشتر نيازهايش را از طريق اينترنت برآورده مي­كند و همچنين حجم اطلاعات موجود در اينترنت روز به روز بيشتر مي­شود. اما نكته مهمي كه وجود دارد اين است كه چگونه بتوانيم درون اين حجم گسترده از اطلاعات، مواردي را كه دلخواه ما است بيابيم. كاملاً مشخص است كه اگر يك دسته­بندي و گروه­بندي مناسبي وجود داشته باشد به مراتب مؤثرتر خواهد بود. در اين پژوهش از يك روش يادگيري بدون ناظر استفاده شده است. برخلاف اغلب روش­هاي ديگر كه روش­هاي يادگيري نظارتي يا نيمه نظارتي را استفاده مي­كنند اما روش پيشنهادي بدون ناظر است و اين خود باعث مي­شود كه محدوديت مجموعه آموزشي را نداشته­ باشيم و همچنين گسترده­گي دسته­ها نسبت به روش­هاي ديگر بسيار بيشتر است. در اين فصل ابتدا مجموعه آزمايشي مورد استفاده معرفي مي­شود و سپس روش پياده­سازي معرفي مي­شود در نهايت نتايج ارزيابي مي­شوند.

## 5-2 مجموعه­ي آزمايشي

اين مجموعه شامل 3121 خبر جمع­آوري شده از سايت بي­بي­سي است كه توسط فنگ و لاپاتا در سال 2008 بوجود آمد. هر خبر شامل يك تصوير و يك عنوان نيز است. يعني براي هر خبر سه فايل وجود دارد، يك فايل HTML شامل محتواي خبر، يك فايل txt براي عنوان خبر، و يك فايل JPG كه حاوي تصوير مربوط به خبر است. اندازه تصاوير 152×203 پيكسل است و حجم آنها به طور ميانگين 10 كيلو بايت است. فايل­هاي مربوط به عنوان خبر به طور ميانگين شامل 5.35 كلمه و فايل­هاي مربوط به محتواي خبر شامل 133.85 كلمه هستند. يكي از مزاياي اين مجموعه عمومي بودن آن است، يعني شامل بيشتر زمينه­ها از جمله سياسي، اقتصادي، فرهنگي، هنري، ورزشي و ... است. البته در اين پژوهش از تصاوير استفاده نمي­شود و فقط فايل مربوط به هر خبر كه خود شامل يك صفحه وب است و همچنين عنوان آن مورد استفاده قرار مي­گيرد.

5-3 پياده­سازي روش پيشنهادي

براي پردازش فايل­هاي اوليه و تمامي مراحل مربوط به پيش­پردازش توسط زبان جاوا انجام گرفته است. از آنجا كه براي اين زبان پكيج­هاي آماده زيادي مربوط به پردازش متن وجود دارد انتخاب مناسبي براي اين كار است. براي پياده­سازي كدهاي زبان جاوا محيط Eclipse انتخاب شده است. پكيج­هاي مورد استفاده در ادامه فصل توضيح داده مي­شوند. نمونه­اي از كدهاي مربوط تبديل فايل HTML به فايل متني در شكل 5-1 نشان داده شده است.

*public String HTML2Text(String HTML) {*

*String text = HTML;*

*text = text.replaceAll("\n", "\*\_newline\_\*");*

*text = text.replaceAll("\\<script.\*?\\</script\\>", "");*

*text = text.replaceAll("\\<style.\*?</style.\*\\>", "");*

*text = text.replaceAll("<\\!--.\*?--\\>", "");*

*text = text.replaceAll("\\<.\*?\\>", "");*

*text = text.replaceAll("&nbsp;", " ");*

*text = text.replaceAll("&lt;", "<");*

*text = text.replaceAll("&gt;", ">");*

*text = text.replaceAll("&amp;", "&");*

*text = text.replaceAll("&quot;", "\"");*

*text = text.replaceAll("\\\*\_newline\_\\\*", "\n");*

*return text; }*

شكل 5-1: كد تابع HTML2Text

* کدهاي اسکريپتي موجود در منبع: شامل تمام الگوهايي مي­شود که به صورت نويسه­هايي هستند که در ابتداي آن­ها "<script" و در انتهاي آنها نيز "</script>" باشد.
* تعاریف style ها: شامل تمام الگوهايي مي­شود که به صورت نويسه­هايي هستند که در ابتداي آن­ها "<style" و در انتهاي آن­ها نيز "</style>" است.
* توضیحات: شامل تمام الگوهايي مي­شود که به صورت نويسه­هايي هستند که با "<!--" شروع و با "-->" به پايان مي رسند.
* برچسب­ها: شامل تمام الگوهايي مي­شود که به صورت نويسه­هايي هستند که با "<" شروع و با ">" به پايان مي رسند.
* کاراکتر هاي ويژه که در HTML معني خاص دارند با مقادیر مناسب جایگزین مي­شوند. مثلاً "&amp" با "&" جايگزين مي­شود.

فصل ششم

نتيجه­گيري

6-1 مقدمه

در حقیقت، صفحات وب یک ساختار نیمه منظم دارند و بر اساس سليقه طراح و صاحب وب­سايت مي­توانند ساختارهاي متنوعي داشته باشند. که اطلاعات در آن با استفاده از تگ‏هاي HTML نمایش داده می­شود و جداسازی تگها از اطلاعات جزء اولین گام­های طبقه­بندی صفحات وب است. همچنین هر تگ مفهوم و جایگاه ویژه‏ای دارد و تفاوت قائل شدن بین این تگ‏ها و محتوی آنها ضروری به نظر می‏رسد. چالش اصلی این مسئله مشخص نمودن نوع و دسته یک صفحه وب است. الگوریتم دسته­بندی ابتدا لازم است از روی یک پایگاه داده مشخص شروع به یادگیری کند و سپس به طبقه­بندی صفحات جدید بپردازد. نحوه آموزش الگوریتم و سپس طبقه بندی صفحات جدید با توجه به اطلاعات موجود جز چالش های اصلی این مسئله می باشد. در اين پژوهش يك راه­حل يادگيري بدون ناظر براي دسته­بندي صفحات وب معرفي گرديد و در نهايت با بقيه روش­ها مقايسه شد.

فصل هفتم

پيشنهادات براي كارهاي آينده

7-1 مقدمه

علارغم اينكه كارهاي بسيار زيادي در زمينه دسته­بندي صفحات وب انجام گرفته است اما باز نمي­توان گفت هيچ كاري دقت 100 درصد را دارد و يا اينكه يك روش همواره بهتر از بقيه است. معمولاً برتري­هاي يك روش نسبت به روش ديگر نسبي است و اگر يك روش در يك شرايط بهتر از ديگري است با تغيير شرايط شايد اين برتري برعكس شود. به همين دليل هنوز در اين بحث جاي كار وجود دارد و مي­توان راهكارهاي بهتري را معرفي كرد.

7-2 پيشنهادات

در اين پژوهش دسته­بندي صفحات وب براساس متن موجود در صفحات انجام گرفت اما آنچه كه مسلم است اين است كه ما مي­دانيم كه متن موجود در يك صفحه وب تمام صفحه نيست و يك صفحه وب مي­تواند شامل تصاوير، نمودارها و غيره باشد كه در اين شرايط دسته­بندي صفحات پيچيده­تر نيز مي­شود. مجموعه آزمايشي كه در اين رساله به كار برده شد اتفاقاً داراي تصاوير نيز است. يك مجموعه خبر كه براي هر خبر يك تصوير نيز وجود دارد حال اگر از اين تصاوير نيز براي تقويت دسته­بندي استفاده كنيم به مراتب نتيجه بهتري را بدست مي­آوريم. البته در اين حالت حجم كار بسيار بيشتر مي­شود و نياز به تكنيك­هاي پردازش تصوير خواهد بود.

يكي ديگر از مشكلات عمده­اي كه در موتور­هاي جستجو وجود دارد جستجوي تصاوير است. موتورهاي جستجو به دليل اينكه معمولاً جستجو را كيبورد بيس انجام مي­دهند همواره داراي نتايج نامربوط بسياري هستند لذا به جاي دسته­بندي صفحات وب مي­توان دسته­بندي تصاوير مربوط به اين صفحات را انجام داد تا جستجوي تصاوير نيز با دقت بيشتري انجام بگيرد. يا حتي مي­توان از حاشيه­نويسي تصاوير[[4]](#footnote-4) استفاده كرد. براي اين كار مي­توان از همين مجموعه آموزشي استفاده شده در اين رساله استفاده كرد. براي دسته­بندي تصاوير هم مي­توان از تكنيك­هاي پردازش تصوير استفاده كرد و هم مي­توان از متن مرتبط به اين تصاوير براي دسته­بندي آن­ها استفاده كرد. همچنين مي­توان يك روش تركيبي معرفي كرد كه بهتر نيز خواهد بود.

1. Supervised Learning [↑](#footnote-ref-1)
2. Feature Selection [↑](#footnote-ref-2)
3. Semantic web [↑](#footnote-ref-3)
4. Image Annotation [↑](#footnote-ref-4)