4th International Conference on Web Research
April 25-26, 2018; Tehran, IRAN
ICWR2018
کنفرانس بین المللی وب پژوهی
چهارمین
۵ و ۶ اردیبهشت ۱۳۹۷؛ تهران، ایران

Clustering Web Documents Using Ontology-Based Fuzzy Method

[1] Najmeh Sakhaee, [2]Fariba Salehi, [3]Majid Khalilian

[1] Ms Student of Software Engineer, Islamic Azad University Karaj Branch, Mechatronics College, Iran
n.sakhaee.star@gmail.com
[2] Faculty member, Islamic Azad University Karaj Branch, Mechatronics College, Iran
Fariba.salehi@kiau.ac.ir
[3]Faculty member Islamic Azad University Karaj Branch, Mechatronics College, Iran
Khalilian@kiau.ac.ir

Abstract

Web documents and web pages are expanding rapidly. Web search engines and web services use different methods to find web pages and documents in the massive amount of documents. However, organizing and analyzing a large amount of data is challenging. The problem with web page retrieval is that the information on the global web is in different formats and from different sources. The accuracy of data selection is essential and their compliance with user requests is a challenge in exploring the web. In order to provide an optimal solution for exploring web documents and organizing and providing quick and accurate access to structured and semi-structured Web documents and web pages, a new approach is proposed. The proposed method is based on the clustering and Web document fuzzation and the semantic and structure of web pages. In the proposed method for the reduction of dimension or features, the mapping of attributes to semantic domains is proposed. The results of the implementation of the proposed method in Python and MATLAB software show that the proposed method in categorizing and organizing web documents is appropriate for the quality of clusters and their density, and in the terms of the davies bouldin and silhouette index, they have suitable values.

Keywords: Clustering, Web Documents, Mining, Semantic Web.

4th International Conference on Web Research
April 25-26, 2018; Tehran, IRAN
ICWR2018
کنفرانس بین المللی وب پژوهی
چهارمین
۵ و ۶ اردیبهشت ۱۳۹۷؛ تهران، ایران

خوشه بندی اسناد وب با استفاده از روش فازی آنتولوژی محور

نجمه سخایی[۱]، فریبا صالحی[۲]، مجید خلیلیان[۳]

[۱] دانشجوی کارشناسی ارشد مهندسی نرم‌افزار، دانشکده مکاترونیک، دانشگاه آزاد اسلامی واحد کرج ، ایران
n.sakhaee.star@gmail.com

[۲] عضو هیات علمی، دانشکده مکاترونیک، دانشگاه آزاد اسلامی واحد کرج ، ایران
Fariba.salehi@kiau.ac.ir

[۳] عضو هیات علمی ، دانشکده مکاترونیک، دانشگاه آزاد اسلامی واحد کرج ، ایران
Khalilian@kiau.ac.ir

چکیده

اسناد و صفحات وب در اینترنت به‌سرعت در حال گسترش هستند. موتورهای جستجو و خدمت رسان‌های وب برای یافتن صفحات وب و اسناد موردنظر در میان حجم انبوهی از اسناد، از روش‌های مختلف استفاده می‌کنند. با این وجود سازمان‌دهی و تحلیل حجم وسیعی از داده‌ها چالش‌برانگیز است. مشکل مطرح درزمینهٔ بازیابی صفحات وب، این است که اطلاعات موجود در وب وسیع جهانی در فرمت‌های مختلف و از منابع مختلف می‌باشند. صحت انتخاب داده‌ها ضروری بوده و تطابق آنها با درخواست کاربران به‌عنوان چالشی در کاوش وب می باشد. به‌منظور ارائه راه‌حلی بهینه برای کاوش در میان اسناد وب و سازمان‌دهی و دسترسی سریع و صحیح به اسناد و صفحات وب ساخت‌یافته و نیمه ساخت‌یافته در این تحقیق روشی جدید پیشنهاد شده است. روش پیشنهادی بر اساس خوشه‌بندی و فازی سازی اسناد وب و با توجه به معنا و ساختار صفحات وب می باشد. در روش پیشنهادی برای کاهش بعد یا ویژگی‌ها، نگاشت ویژگی‌ها به حوزه‌های معنایی پیشنهاد شده است. نتایج حاصل از پیاده سازی روش پیشنهادی در نرم افزار پایتون و متلب نشان می دهد روش پیشنهادی در دسته بندی و سازماندهی اسناد وب، از نظر کیفیت خوشه ها و تراکم آنها مناسب بوده و از نظر شاخص دیویس بولدین و سیلهوئت دارای مقادیر مناسبی می باشد.

کلمات کلیدی

خوشه بندی، اسناد وب، کاوش، وب معنایی.

۱- مقدمه

امروزه اسناد و صفحات وب در اینترنت به‌سرعت در حال گسترش هستند. کاربران با استفاده از اینترنت و بسیاری از موتورهای جستجو می‌توانند اسناد و صفحات وب موردنظرشان را پیدا کنند. سازمان‌دهی و تحلیل حجم وسیعی از داده‌ها چالش‌برانگیز بوده و گاها غیرممکن می‌باشد [1]. موتورهای جستجو و خدمت رسان‌های وب، برای یافتن صفحات وب و اسناد موردنظر در میان حجم انبوهی از اسناد، از روش‌های مختلف استفاده می‌کنند. با این وجود هنوز بخش عظیمی از وب توسط این خدمت رسان‌ها قابل‌کشف نبوده و به‌صورت لایه پنهان وب دور از دسترس بوده و قابل استفاده نمی باشد. مشکل مطرح درزمینهٔ بازیابی صفحات وب، این است که اطلاعات موجود در وب وسیع جهانی در فرمت‌های مختلف و از منابع مختلف می‌باشند [2]. کنسرسیوم W3 اظهار داشته است که HTML، به علت محدودیت‌هایی همچون داده‌های نیمه ساخت‌یافته ، حساسیت به حروف، تگ‌های از پیش تعریف‌شده و مسائلی ازاین‌دست، ساختار معنایی خوبی از محتویات صفحات وب توصیف نمی‌کند. بعدها برای غلبه بر این مشکلات تکنولوژیهایی همچون XML، Flash (با ویژگی‌های خوب طراحی) و مانند آن پدید آمد [3]. امروزه داده‌های نیمه ساخت‌یافته در وب افزایش‌یافته است و به دنبال آن استفاده از

XML به‌عنوان استانداردی برای داده‌های نیمه ساخت‌یافته، گسترش‌یافته است. محققان این حوزه برای بهبود صفحات وب، شروع به مهاجرت به صفحات وب XML ای نمودند و از تکنولوژی ادغام صفحات وب با معانی و محتوای صفحه برای فراهم‌سازی توصیف بهتر ساختار معنایی و توصیف محتوا استفاده کردند.

از جمله مشکلاتی که در زمینه طبقه بندی اسناد وب وجود دارد کشف کردن دانش مفید از متن نیمه ساختیافته یا غیرساختیافته است که توجه زیادی را به خود جلب کرده است. روش های داده کاوی سنتی فرض می کنند که اطلاعات به فرم پایگاه داده های رابطه ای هستند به همین دلیل برای بسیاری از کاربردها مانند اطلاعات الکترونیکی قابل دسترس به فرم نیمه ساختیافته یا غیرساختیافته مفید نیستند. داده های وب، داده های نیمه ساختیافته هستند چون نه به طور کامل غیرساختیافته هستند و نه به طور کامل ساختیافته هستند.

وب معنایی خدمات وبی هوشمندتری را برای هماهنگی و مرتب نمودن داده های روی وب عرضه نموده است. در سال‌های اخیر صحت انتخاب داده‌های ضروری مطابق با درخواست کاربر و صدور آن‌ها در خروجی به‌عنوان چالشی در کاوش وب مطرح شده است. تحلیل و کاوش داده‌های XML از طرف انجمن کشف دانش و داده‌کاوی در سال‌های اخیر به محبوبیت رسیده است. مرور مؤثر صفحات وب و جستجوی مجموعه عظیمی از این اسناد نیازمند سازماندهی و استفاده از روش‌های داده‌کاوی می‌باشد. رشد داده‌های نیمه ساخت‌یافته و گستره استفاده از XML به‌عنوان استانداردی برای داده‌های نیمه ساخت‌یافته دلیل اصلی این نیاز می‌باشد [4]. وب معنایی می‌تواند به‌صورت کارا و مؤثر، منابع مختلف را بر اساس ارتباطات معنایی سازمان‌دهی، اشتراک‌گذاری، دسته‌بندی، ترکیب و مدیریت کند. ارتباطات معنایی، مطالعهٔ مفاهیم در محیط‌های ارتباطی برای پشتیبانی از کاربردهای هوشمند در ارتباطات منابع در محیط‌های معنایی است، به‌طوری‌که ماشین‌ها و انسان‌ها توانایی درک یکدیگر را داشته باشند. زبان‌های معنایی توصیف دقیق نهادهای تعاملی مثل سرویس‌ها، منابع و کاربران در سطوح بالای تجرید را ممکن ساخته‌اند و نتیجه‌گیری خودکار در مورد این نوع ارائه را فراهم می‌سازند، بنابراین به انگیزش و درک دوطرفه میان نهادهای کم اشتراک یا دانش بدون اولویت در مورد همدیگر توجه می‌کنند. در وب معنایی از آنتولوژی استفاده می‌شود. یک آنتولوژی، توصیفی صریح و فرمال از مفاهیم یک دامنه خاص است (کلاس‌ها که گاهی اوقات مفاهیم نیز نامیده می‌شود)، ویژگی‌های هر یک از این مفاهیم و صفات خاص و خصایص مختلف این مفهوم را توصیف می‌کنند. آنتولوژی به‌منزله یک توصیف فرمال و صریح از واژه‌های یک دامنه خاص می‌باشد و ارتباطات میان آنتولوژی ها در وب یک دامنه وسیع را پوشش می‌دهند. هدف این ساختارها تسهیل تعامل عامل‌ها در وب است.

انباره‌های مختلفی وجود دارد که شامل حجم عظیمی از خدمات وب می‌باشد. یافتن روشی که خدمات موردنظر کاربر را از میان این انباره‌ها بیابد و خدمات یافت شده بیشترین ارتباط را به درخواست کاربر داشته باشد به مسئله‌ای چالش‌برانگیز تبدیل‌شده است. خدمت رسان‌های وب اغلب با حجم زیادی از خدمات وب روبرو می‌باشند که مشابه با درخواست کاربر می‌باشد. همان‌طور که اکثراً مشاهده‌شده است برای هر عمل جستجو در موتورهای جستجو معمولاً میلیون‌ها وب‌سایت و سند و تصویر مرتبط با درخواست کاربر در عرض کمتر از یک ثانیه به کاربر ارائه می‌شود درحالی‌که مسلماً کاربران فرصت و حوصله کافی برای مطالعه این چند میلیون پاسخ را نخواهند داشت و احتمالاً فقط پاسخ‌هایی را که در رتبه‌بندی‌های بالا قرار دارد مطالعه نمایند.

در سال‌های اخیر توجه زیادی به سمت خوشه‌بندی معطوف شده است. خوشه‌بندی را می‌توان در رابطه با سازماندهی اسناد استفاده کرد تا بتوان آن‌ها را به‌طور خودکار در دسته‌های معنی‌دار تقسیم کرد. خوشه‌بندی یکی از تکنیک‌هایی است که کاربرد گسترده‌ای در تجزیه‌وتحلیل داده‌های کاوشی دارد چراکه به‌واسطه آن می‌توان ساختار طبیعی داده‌ها را معین نمود. در مقایسه با یادگیری نظارتی، خوشه‌بندی یک روش یادگیری نظارت‌نشده است ازاین‌رو هیچ نیازی به داده‌های برچسب دار در مدت پروسهٔ خوشه‌بندی نیست؛ هدف خوشه بندی، گروه‌بندی اشیاء است، بطوریکه اشیاء در یک خوشه به همدیگر شبیه‌ترند تا با اشیای دیگر خوشه‌ها [5].

برای مرور مؤثر و جستجو در مجموعه عظیمی از اسناد وب و سازمان‌دهی آن‌ها نیازمند استفاده از روش‌های داده‌کاوی هستیم. روش‌هایی که در پاسخ به درخواست‌های کاربران اسناد و صفحاتی را ارائه نمایند که بیشترین ارتباط را با درخواست کاربر داشته باشد. در این تحقیق سعی بر ارائه راه‌حلی بهینه برای کاوش در میان اسناد وب و سازمان‌دهی و دسترسی سریع و صحیح به اسناد و صفحات وب با استفاده از خوشه‌بندی و فازی سازی شده است.

در ادامه پس از مرور کارهای پیشین در بخش ۳ روش پیشنهادی توضیح داده می شود و در بخش ۴ پیاده سازی و ارزیابی نتایج شرح داده می شود و در پایان در بخش ۵ به جمع بندی مطالب بیان شده و در بخش ۶ پیشنهاد کارهای آتی می پردازیم.

۲- مرور کارهای پیشین

یادگیری نیمه نظارت شده، یادگیری با داده های برچسب دار و بدون برچسب، اخیرا توسط محققان زیادی مورد مطالعه واقع شده است. انواع الگوریتم های نیمه نظارت شده ای پیشنهاد شده اند، شامل آموزش همزمان[6] ، بیز ساده ی نیمه نظارت شده[7] ، ماشین های بردار پشتیبانی[8] TSVM ، مدل خوشه بندی فازی[9] و روش های مبتنی بر گراف می باشند.

لیو و همکاران [10]یک الگوریتم جدید قطعه بندی صفحات وب بر اساس یافته های درخت گوموری هیو در گراف برنامه ریزی ارائه نمودند. الگوریتم ابتدا اطلاعات دیداری و ساختاری را از یک صفحه وب جدا کرده و در یک گراف وزندار بدون جهت قرار می داد به طوریکه رئوس این گراف گره های برگ درخت DOM بوده و لبه ها یک رابطه وضعیت قابل رویت بین رئوس بودند. در نهایت گراف با الگوریتم خوشه بندی بر اساس درخت گرومی هیو پارتیشن بندی می شد.

در سال ۲۰۱۴ وانگ و همکاران [6] از نوعی خوشه بندی آنلاین برای خوشه بندی اطلاعات مربوط به ترافیک شبکه استفاده نمودند. لین و همکاران [11] نوعی خوشه بندی بازگشتی K-mean با استفاده از ماشین شاخص گذاری ناخالصی برای جاگذینی خوشه بندی K-mean معرفی کردند که در واقع ترکیبی از مرکز خوشه و نزدیکترین همسایگی بود و k نزدیکترین همسایگی را به صورت محلی انجام می داد.

در سال ۲۰۱۶ سو و همکاران [12] روش جدیدی از خوشه بندی نظارت نشده با عنوان خوشه بندی حلقه ای، ارائه دادند. در روش پیشنهادی آنها برای هر خوشه یک سرخوشه تعریف می شود که در آن سرخوشه اطلاعات مربوط به اعضای خوشه و فرمت خوشه ذکر می گردد.

در سال ۲۰۱۶ ایوان و همکـاران[13] روش خوشـه بنـدی دسـته ای داده های فازی با استفاده از رابطه غیر قابـل تشـخیص را ارائـه دادنـد. آنهـا یـک الگوریتم جایگزین با تصحیح روش K پارتیشن فازی برای خوشه بندی دسته ای داده های فازی ارائه دادند . الگوریتم پیشنهادی با اسـتفاده از رابطـه غیـر قابل تشخیص میان داده ها خوشه بندی را انجام می داد ، نتـایج نشـان مـی دهد که روش پیشنهادی در مقایسه با الگوریتم های قبلی با پایه K-means خوشه بندی خالص تری انجام می دهد و نیز زمان پاسخ کوتاهتری دارد.

در سال ۲۰۱۶ لیانگ ژانگ و همکاران در [14] از FCM برای دسته بنـدی داده های از دست رفته و نیمه کامل استفاده کردند. در این مقاله نویسـندگان خوشه بندی جدیدی بر اساس FCM برای خوشه بندی داده های نیمه کامل با استفاده از مقادیر احتمالاتی هسته اطلاعات و معیار بیشترین شـباهت ارائـه کردند و در گام بعدی مدل خوشه بندی شان را با استفاده از بهینـه سـاز سـه سطحی جایگزینی و ضریب لاگرانـژ بهبـود دادنـد. صـحت نتـایج بـه وسـیله شاخص میانگین خطای اولیه، با ارجاع به دو شاخص به نام های طبقه بنـدی خطا و نسبت دهی خطا ارزیابی می شود. نتایج همگرایی پایدار، صحت بالا و قدرتمند در روش پیشنهادی را نشان می دهد.

با وجود تحقیق های فراوانی که برای سازمان‌دهی و تحلیل حجم وسـیعی از داده‌ها انجام شده است، اما هنوز بخشهایی از وب به عنوان لایه پنهان دور از دسترس باقی است. مشکل مطرح درزمینهٔ بازیابی صفحات وب، ایـن اسـت که اطلاعات موجود در وب وسیع جهانی در فرمـت‌هـای مختلـف و از منـابع مختلف می‌باشند. صحت انتخـاب داده‌هـا ضـروری بـوده و تطـابق آنهـا بـا درخواست کاربران به‌عنوان چالشی در کاوش وب می باشـد. بـه‌منظـور ارائـه راه‌حلی بهینه برای کاوش در میان اسناد وب و سازمان‌دهی و دسترسی سریع و صحیح به اسناد و صفحات وب سـاخت‌یافتـه و نیمـه سـاخت‌یافتـه در ایـن تحقیق روشی جدید پیشنهاد شده است، که در ادامه به شرح آن می پردازیم.

۳- روش پیشنهادی

به‌منظور ارائه راه‌حلی بهینه برای کاوش در میان اسناد وب و سازمان‌دهی و دسترسی سریع و صـحیح بـه اسـناد و صـفحات وب سـاخت‌یافتـه و نیمـه ساخت‌یافته در این تحقیق روشی جدید با استفاده از خوشه‌بندی و فازی سازی ارائه‌شده است که در آن علاوه بر ویژگی‌های ساختاری، ویژگی‌های محتـوایی اسناد نیز مورد کاوش قرار می‌گیرد. فرایند اصلی روش پیشنهادی مطابق شکل ۱ از پنج گام به شرح زیر تشکیل شده است:

۱. جمع‌آوری اسناد و صفحات وب
۲. استخراج ویژگی‌های ساختاری
۳. پاک‌سازی اسناد
۴. استخراج ویژگی‌های محتوایی و کاهش بعد
۵. خوشه‌بندی و فازی سازی

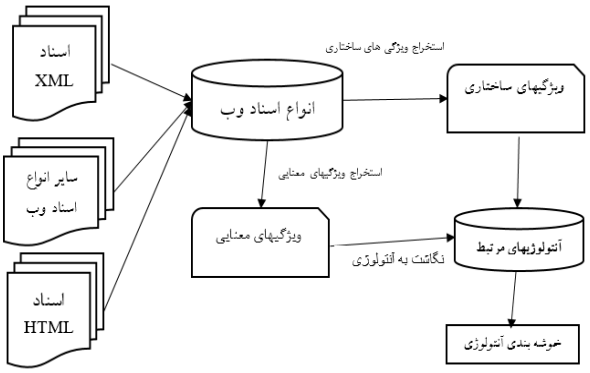
شکل(۱): مراحل روش پیشنهادی

در شکل ۲ معماری روش پیشنهادی نشان داده‌شده است. در ادامه هـر یک از مراحل روش پیشنهادی شرح داده می شود.

جمع‌آوری اسناد و صفحات وب: جمع‌آوری اسناد و صفحات وب اولـین گام برای کاوش صفحات وب می‌باشد. در این گـام فایلهـا و اسـناد وب بـه صـورت اتوماتیـک جمـع‌آوری مـی‌شـوند و در یـک پایگـاه داده‌ای ذخیـره می‌شوند.

استخراج ویژگی‌های ساختاری: استفاده از تگ‌ها متداول‌ترین راه بـرای استخراج ویژگی‌های از اسناد وب می‌باشد. در این روش متون بدون توجه به نام تگها به مفاهیم معنایی نگاشت می‌شوند. سپس مفاهیم استخراج شده از فضای آنتولوژی به نام تگها افزوده می‌شود. ویژگی‌های استخراج شده از نام تگ‌ها می‌توانند با بهره گیری از آنتولوژی لغوی به دانش برسد.

پاک‌سازی اسناد: پاک‌سازی اسناد وب شامل حذف تگها و کلماتی است که در لیست STOPWORD مـی‌باشـد. یـک لیسـت STOPWORD شامل کلماتی است که به صورت متناوب در زبان انگلیسی استفاده می‌شوند ومعنی خاصی ندارند مانند "The".



شکل(۲): معماری روش پیشنهادی

استخراج ویژگی‌های محتوایی و کاهش بعد: داده‌هـای دارای بعـد زیـاد چالش‌های ریاضیاتی زیادی را در وظایف یـادگیری ماشـین پـیش روی مـا می‌گذارند. یکی از مسائل در رابطه با مجموعه داده‌های دارای بعد بالا ایـن است که همهٔ متغیرهای اندازه‌گیری شده بـرای درک پدیـدهٔ مـوردنظر حائز اهمیت هستند. در کاهش ابعاد سعی بر این است که بُعد داده‌هـا طبـق یک تعداد از معیارها پیدا شود. که خود حوزهٔ تحقیقی فعـالی در یـادگیری ماشین است. برای کاهش بعد پیشنهاد ما نگاشت آنتولوژی می‌باشد. از روی ویژگی‌های استخراج شده می‌توان آنتولوژیهای مربوطه را مشـخص نمـود و

در ادامه کار به جای استفاده از ویژگی‌های زیاد از حوزه مربوطه استفاده می‌کنیم. برای نگاشت آنتولوژی از الگوریتم پیشنهادی زیر استفاده می‌شود:

در الگوریتم نگاشت آنتولوژی بر اساس کلمات کلیدی، کلمات کلیدی به عنوان ورودیهای الگوریتم می‌باشد و پس از اجرای الگوریتم، کلمات کلیدی به حوزه‌های معنایی مربوطه نگاشت می‌یابند و خروجی الگوریتم آنتولوژیها می‌باشد. مثلاً اگر در سندی کلمات کلیدی شامل توپ، دروازه، شوط، گل، و غیره باشد، حوزه معنایی ورزش به همراه زیر خوشه‌های فوتبال، والیبال، بسکتبال و موارد مشابه به عنوان آنتولوژی نگاشت خواهد شد.

ورودی: کلمات کلیدی

خروجی: آنتولوژی

۱. شروع
۲. براساس وزن کلمات کلیدی آنها را مرتب کن.
۳. با توجه به وزن کلمات حوزه‌های مربوطه را مشخص کن
۴. برای هر حوزه وزن هر سند را محاسبه کن
۵. برای هر زیر مجموعه از حوزه وزن سند را محاسبه کن
۶. حوزه‌های مربوطه را استخراج کن
۷. پایان

خوشه‌بندی و فازی سازی : برای خوشه‌بندی از الگوریتم نیمه Kmeans-فازی استفاده می‌شود [5]. این الگوریتم نوعی از الگوریتم K-means است ولی در آن از تکنیک EM برای به کارگیری یک تابع عضویت فازی استفاده‌شده و این امکان حاصل می‌شود که اشیاء به بیش از یک خوشه تعلق داشته باشند.

هدف اکثر الگوریتم‌های خوشه‌بندی، حداقل سازی تابع هزینه است که شامل سنجش تفاوتها بین اشیاء و نماینده خوشه می‌باشد. K-means . به‌صورت موضعی فاصله‌ٔ مربع میانگین بین اشیاء و مراکز خوشه‌ای را به حداقل می‌رساند.

EMتکنیک متداول دیگری برای تجزیه و تحلیل الگوریتم‌های خوشه‌بندی است چراکه یک روش فرمول بندی شده‌ٔ آماری می‌باشد و اطلاعات مفصل‌تری راجع به نتیجه‌ٔ خوشه‌بندی می‌دهد. الگوریتم EM یک روش کلی برای یافتن راه حل‌های احتمال ماکزیمم در مدل‌های دارای متغیر نهان است. اگر مجموعه‌ٔ داده‌های مشاهده شده x={x1,….xn} باشد مجموعه پارامترهای مدل با Θ نشان داده می‌شود و مجموعه‌ٔ همه‌ٔ متغیرها نهان Z است، در مرحله‌ٔ E مقادیر پارامتر فعلی Θ^{old} را به کار برده و توزیع پیشین متغیرهای نهان را محاسبه می‌کنیم که با $p(Z|X, \Theta^{old})$ نشان داده می‌شود. سپس توزیع پیشین برای محاسبه‌ٔ مقدار امید ریاضی احتمال استفاده‌شده و مقدار پارامتری جدید Θ^{new} حاصل می‌شود. مقدار امید به ریاضی احتمال داده‌های کامل در توزیع پیشین متغیرهای نهان با $Q(\Theta, \Theta^{old})$ نشان داده می‌شود که در رابطه ۱ می‌بینیم. در مرحله‌ٔ M پارامتر جدید θnew را با ماکزیمم سازی این تابع حاصل می‌کنیم که در رابطه ۲ مشاهده می‌شود.

$$Q(\Theta, \Theta^{old}) \quad (۱)$$

$$= E_{Z|X,\Theta^{old}} [\ln P(X, Z|\Theta)] \quad (۲)$$

$$\Theta^{new} = argmax_{\Theta} Q(\Theta, \Theta^{old})$$

در بسیاری از الگوریتم‌های خوشه‌بندی هدف به حداقل رسانی تابع هزینه است که شامل سنجش اغتشاش بین اشیاء و نماینده خوشه می‌شود-K means. به‌صورت موضعی فاصله‌ٔ مربع میانگین بین اشیاء و مراکز خوشه‌ای را به حداقل می‌رساند .C-Means فازی تابع هدف مشابهی دارد ولی K-means را گسترش داده و درجه‌ٔ اطلاعات عضویت را شامل می‌شود که حاکی از وجود اطمینان در تخصیصشی به خوشه‌ٔ موردنظر است. این دو الگوریتم خوشه‌بندی را می‌توان با بهینه سازی توابع هزینه، حاصل نمود.

K-meansرا می‌توان با استفاده از الگوریتم EM روی ترکیبی از گاوسی‌های C تحت فرضیات معینی مدل کرد، که در آن C تعداد خوشه‌های K-means است. از آنجاییکه الگوریتم K-means تخصیص سختی از نقاط داده‌ای را در خوشه‌ها اجرا می‌کند که در آن هر نقطه‌ٔ داده‌ای بطور منحصر بفردی با یک خوشه مرتبط است الگوریتم EM تخصیص نرمی را بر اساس احتمالات پیشین صورت می‌دهد. مدل ترکیبی گاوسی را با اجزای C در نظر بگیرید که در آن میانگین این اجزا $\mu_1 \ldots \mu_C$ هستند و ماتریس‌های کوواریانس معمول اجزای ترکیبی با $\sum=\varepsilon I$داده می‌شوند که در آن I ماتریس یکسانی را نشان می‌دهد و εپارامتر مشترکی با همه‌ٔ اجزا است. با $\varepsilon \to 0$ احتمال داده‌های کامل را می‌توان به شکل رابطه ۳ نوشت.

$$E_{Z|X,\Theta^{old}} [\ln P(X, Z|\Theta)] \quad (۳)$$

$$= \frac{1}{2}\sum_{i=1}^{N}\sum_{c=1}^{C} P(Z_c|X_i;\Theta)||X_i - \mu_c||^2 + const$$

از اینرو با به ماکزیمم رسانی احتمال داده‌های موردانتظار بالا می‌توان به نتیجه رسید. در واقع هر داده به یک خوشه تخصیص داده می‌شود همانطوریه در رابطه ۴ می‌بینیم:

$$P(Z_c|X_i;\Theta) \quad (۴)$$

$$= \begin{cases} 1 \; if \; c = argmin \; ||X_i - \mu_c||^2 \\ 0 \; otherwise \end{cases}$$

اساساً تابع احتمال پیشین به تخصیص سخت محدود نمی‌شود همانطوریکه در رابطه ۴ می‌بینم دیگر تخصیص‌های سخت را نیز می‌توان به خوبی استفاده کرد. ارتباط بین K-means و EM توصیفی در بالا منجر می‌شود که یک الگوریتم K-means نیمه نظارت شده پیشنهاد دهیم. تابع احتمال پیشین با یک تابع عضویت فازی دارای محدودیت جایگزین می‌شود. سپس می‌توان عضویت فازی را داخل مرحله‌ٔ E الگوریتم EM جاگذاری کرد. در این مطالعه از ماتریس U استفاده‌شده و اطلاعات عضویت فازی نگه داشته می‌شود که در آن Uic درجه‌ٔ عضویت سند i در خوشه‌ٔ c را نشان می‌دهد. بعبارت دیگر Uic برای نشان دادن احتمال پیشین P(Zc|Xi) استفاده می‌شود. در مرحله‌ٔ M ، Uic استفاده‌شده و پارامترهای مدل جدید محاسبه می‌شود که مراکز خوشه‌ای μ1…μC از ماکزیمم سازی تابع احتمال لاگ هستند.

علاوه بر این تعداد کمی مثال‌های برچسب دار موجود است. در نیمه-Kmeansفازی از مثال‌های برچسب دار برای دو هدف اصلی استفاده می‌شود. اول اینکه این مثال‌های برچسب دار می‌توانند موجب حدس اولیه‌ٔ مراکز شوند. عملاً الگوریتم‌های خوشه‌بندی نظیر K-means و C-Meansفازی به حدس اولیه حساس هستند. از اینرو بذرها می‌توانند حدس بهتری برای الگوریتم حاصل کرده و اسناد را خوشه‌بندی کنیم. دوم اینکه

مثال‌های برچسب دار می‌توانند خوشه‌بندی را به سمت فضای جستجوی بهتری سوق دهند.

1. Begin
2. Hi ($\frac{Hi}{||Hi||}$ where Hi is the ith row of H and i=1….N
3. Ĥ(PLSA_Clustering(H,K)
4. μC ($\frac{1}{|Sc|}\sum di \in Sc$ Ĥ where μC is the center of cth cluster and c=1…C
5. repeat
6. for i=1 to N do
7. for c=1 to C do
8. If di is a document of Sc then
9. Uic←1
10. Uic´←0. For c´=1…C and c´≠c
11. Break
12. Else
13. Uic($e^{-||\hat{H}i-\mu C||^2/2\sigma^2}$
14. end
15. end
16. Normalize Uic so that the sum of each row of U is 1.
17. end
18. For c=1 to C do
19. μC($\frac{1}{Zc}\sum_{i=1}^{N} Uic \times \hat{H}i$
20. end
21. until convergence
22. return U
23. end

شکل ۳ شبهه کد الگوریتم نیمه K.means فازی

در روش پیشنهادی برای ورودی الگوریتم به جای ماتریس ترم- سند از ماتریس آنتولوژی – سند استفاده می‌شود. ورودی‌های الگوریتم H ماتریس انتولوژی-سند، K تعداد موضوعات،C تعداد خوشه‌ها و S1…..SC هسته‌های خوشه‌های ۱ تا C. و خروجی U ماتریس عضویت سند- خوشه می‌باشد.

شکل ۳ شبهه کد، الگوریتم نیمه Kmeans-فازی را نشان می‌دهد. تابع عضویت فازی مورداستفاده در این الگوریتم، تابع توزیع گاوسی هست که در معادله‌ٔ ۷ می‌بینم که در آن μCموقعیت مرکز و σ برای کنترل درجه‌ٔ عضویت Xi در خوشه‌ٔ c و | Xi- μC | فاصله‌ٔ بین XiوμC است.

$$U_{ic} = e^{-||X_i-\mu_c||^2/2\sigma^2} \quad (۵)$$

نقاط نزدیک به مرکز مهم خواهند بود و نقاط دور از مرکز نسبتاً کم‌اهمیت هستند. فاصله با استفاده از تعداد درجه‌ٔ فازی نشان داده می‌شود چراکه مقدار Uic مقداری است مابین ۰ و ۱. جدای از این مراکز خوشه‌ای μ1…μC پارامتر Θ مدل هستند و به‌صورت مکرر آپدیت می‌شوند

ورودی‌های الگوریتم نیمه Kmeans-فازی شامل ماتریس آنتولوژی – سند $N \times M$ ، تعداد خوشه‌هایC ، تعداد موضوعات K و رده‌بندی سندهای s1,…, Sc می‌شوند. هر ردیف H سندی را نشان می‌دهد و هر ستون خصیصه‌ٔ ترمی سند را. در ابتدا هر ردیف از H باید نرمالیزه شود سپس ماتریس نرمال شده و تعداد موضوعات K به الگوریتم خوشه‌بندی PLSAداده‌شده و ماتریس موضوع–سند $\tilde{H}$حاصل می‌شود. تفاوت اصلی بین H و $\tilde{H}$ این است که هر سند در H با یک بردارترم نشان داده می‌شود؛ درحالی‌که هر ردیف در $\tilde{H}$با بردار موضوعی نشان داده می‌شود. رده‌بندی اولیه اسناد برچسب دار هستند و این امکان را می‌دهند که از رده‌بندی‌های s1,…, Sc برای محاسبه‌ٔ مراکز خوشه‌ای اولیه تحت عناوین μ1…μC استفاده کنیم. پروسه‌های بالا در خطوط ۲ الی ۴ الگوریتم نیمه Kmeans-فازی لیست می‌شوند.

وقتی‌که پارامترهای μ1…μC حاصل شدند درجه‌ٔ عضویت Uic را می‌توان با استفاده از تابع توزیع گاوسی با اندازه‌گیری فاصله محاسبه کرد. جدای از این هرگونه اطلاعات خوشه‌بندی رده‌ای شناخته می‌شود. سپس ماتریس درجه‌ٔ عضویت U نرمالیزه می‌شود. پروسه‌های بالا در خطوط ۶ الی ۱۷ الگوریتم نیمه Kmeans-فازی لیست می‌شوند.

وقتی‌که ماتریس درجه‌ٔ عضویت U تغییر می‌کند الگوریتم باید از احتمال پیشین جدید متغیر نهان برای محاسبه‌ٔ پارامترهای جدید استفاده کند که مراکز خوشه‌ای با اسناد برچسب دار و بدون برچسب هستند. در این مطالعه از بردار نرمالیزه سازی Z استفاده‌شده و فاکتور نرمالیزاسیون نشان داده می‌شود که در آن هر Zc مجموع درجات عضویت اسناد در خوشه‌ٔ c را نشان می‌دهد. سپس الگوریتم می‌تواند مراکز خوشه‌ای جدا را محاسبه کند. پروسه بالا در خطوط ۱۸ الی ۲۳ الگوریتم نیمه Kmeans-فازی لیست می‌شوند. وقتی‌که الگوریتم همگرا است ماتریس U درجه‌ٔ عضویت را در خروجی می‌دهد.

علاوه بر تابع توزین گاوسی تشابه کسینوسی را می‌توان به‌خوبی استفاده کرد. در بازیابی اطلاعات (IR) یا پردازش زبان طبیعی (NLP)، مدل فضای برداری اغلب برای نشان دادن اسناد استفاده می‌شود که در آن هر سند به‌عنوان برداری نشان داده‌شده و هر بُعد متناظر با یک‌ترم متمایز است. تشابه کسینوسی اغلب برای اندازه‌گیری فاصله در IR یا NLP استفاده می‌شود. نتیجه‌ٔ تشابه کسینوسی عددی است مابین ۰ و ۱. جدای از این تابع عضویتی که در الگوریتم استفاده می‌شود باید نرمالیزه شود چراکه از توزیع احتمال پیشین حاصل می‌شود.

٤- پیاده سازی و ارزیابی نتایج

به‌منظور ارائه راه‌حلی بهینه برای کاوش در میان اسناد وب و سازمان‌دهی و دسترسی سریع و صحیح به اسناد و صفحات وب ساخت‌یافته و نیمه ساخت‌یافته در این تحقیق روشی جدید با استفاده از خوشه‌بندی و فازی سازی ارائه‌شده است که در آن علاوه بر ویژگی‌های ساختاری، ویژگی‌های محتوایی و معنایی اسناد نیز مورد کاوش قرار می‌گیرد. داده های اولیه مورد استفاده در این تحقیق ۴۵ سند خبری از سایت پرس تی وی می باشد. ۱۵ خبر اقتصادی، ۱۵ خبر ورزشی و ۱۵ خبر سیاسی به صورت تصادفی از این سایت به عنوان داده های اولیه انتخاب شده اند. برای پیاده سازی روش پیشنهادی از نرم افزار پایتون و متلب استفاده شده است. برای استخراج ویژگی‌های ساختاری و پاکسازی اسناد از نرم افزار پایتون و از کتابخانه NLTK استفاده شده است. برای استخراج ویژگی‌های مفهومی و کاهش بعد از named entity recognition در کتابخانه NLTK در پایتون استفاده شده است. ابتدا ۶ گروه، کلمه هدف با وزن یکسان تعریف کرده ایم سپس کلمات موجود به گروه مربوطه نگاشت شده است. جدول۱ ویژگیها و خروجی این مرحله را نشان می دهد. در ستون کمترین استفاده، تعداد

کمترین استفاده از این ویژگی در اسناد را نشان می دهد و ستون بیشترین استفاده، تعداد اسنادی که بیشترین استفاده از این گروه کلمات را داشته اند را نشان می دهد. ستون میانگین، میانگین استفاده از گروه کلمه در کل ۴۵ سند را نشان می دهد.

شکل ۴ میزان استفاده از هر ویژگی در خوشه های تعین شده را نشان می دهد. در شکل ۴، تعداد سطرها معرف تعداد خوشه ها می باشد که با توجه به اینکه نوع داده های اولیه، شامل سه نوع خبر ورزشی، اقتصادی و سیاسی می باشد پس ۳ خوشه داریم که در جدول۱ با ۳ سطر نشان داده شده است. هر ستون این جدول بیانگر یک ویژگی می باشد. ستون اول نمایانگر ویژگی اول و میانگین نمونه های موجود در هر خوشه را نشان می دهد. به عنوان مثال ستون اول نمایانگر ویژگی facility بوده و در خوشه اول به طور میانگین ۰.۰۰۵۷ بار از این ویژگی استفاده شده است.

نتایج حاصل از خوشه بندی در شکل ۵نشان داده شده است. در این پیاده سازی پس از مرحله کاهش بعد به ۶ ویژگی رسیدیم و فراوانی این ۶ ویژگی در کل ۴۵ سند طبق نمودار شکل ۵ می باشد.

نمودار شکل ۶ نمایش سه بعدی نتایج خوشه بندی در نرم افزار متلب بر اساس ۳ ویژگی facility, Geopolitical Entity, Geo-Social-Political را نشان می دهد.

نمودار شکل ۷ خوشه بندی اسناد بر اساس ۳ ویژگی Geopolitical Entity, Geo-Social-Political را نشان می دهد.

شکل ۸ خوشه بندی اسناد بر اساس ۳ ویژگی Geo-Social-Political, Location و Organization را نشان می دهد.

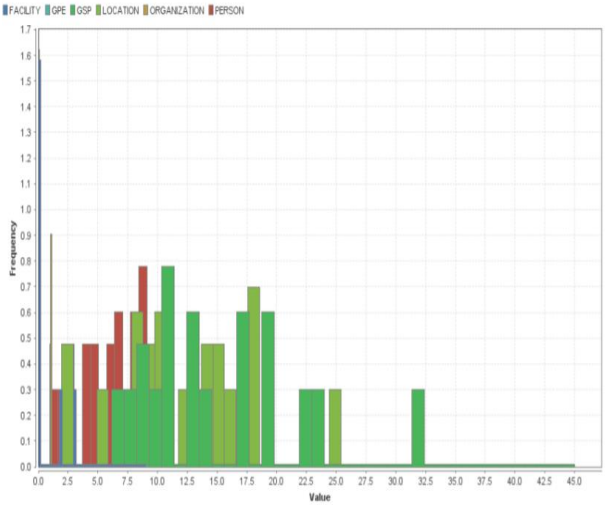
شکل ۹ خوشه بندی اسناد در نرم افزار متلب بر اساس ۳ ویژگی Location ، Person و Organization را نشان می دهد.

جدول۱ استخراج ویژگیهای مفهومی

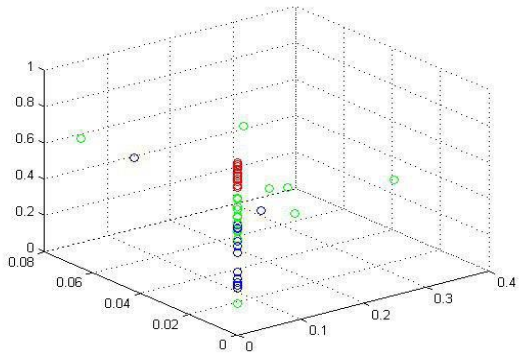
کلمه هدف	کمترین استفاده	بیشترین استفاده	میانگین استفاده
Facility	۰	۹	0.533
Geopolitical Entity	۰	۱	0.067
Geo-Social-Political	۳	۴۵	15.689
Location	۱	۴۰	14
Organization	۰	5	0.400
Person	1	28	9.022

	1	2	3	4	5	6
1	0.0057	0.0012	0.1979	0.1669	0.0044	0.1128
2	0.0056	0.0011	0.1935	0.1632	0.0043	0.1103
3	0.0117	0.0024	0.4043	0.3410	0.0091	0.2304

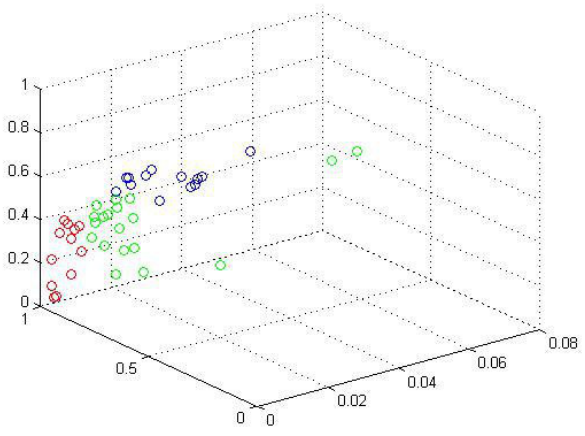
شکل ٤ میزان استفاده از ویژگیها در هر خوشه



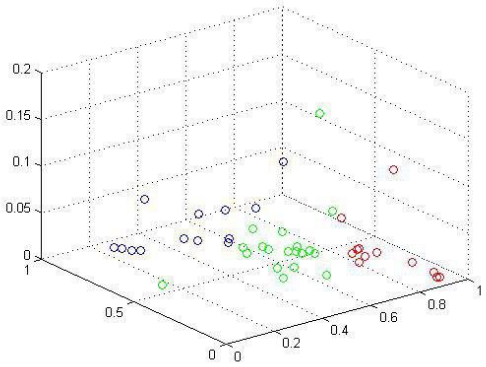
شکل۵ فراوانی ویژگیها در کل اسناد



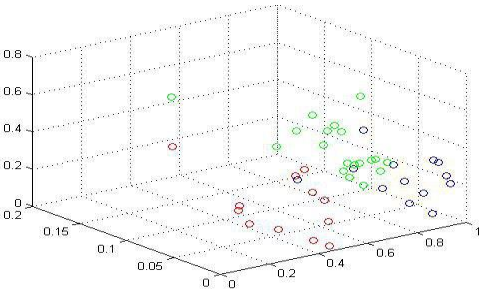
شکل ٦ خوشه بندی بر اساس ۳ ویژگی Facility, Geopolitical Entity, Geo-Social-Political



شکل ۷ خوشه بندی اسناد بر اساس ۳ ویژگی Geopolitical Entity, Locationو Geo-Social-Political



شکل ۸ خوشه بندی اسناد بر اساس ۳ ویژگی Geo-Social-Political, Location و Organization



شکل ۹ خوشه بندی اسناد بر اساس ۳ ویژگی Location و Person و Organization

٤-١- ارزیابی نتایج خوشه بندی با معیار سیلهوئت[1]

کیفیت خوشه‌ها در روشهای یادگیری بدون نظارت از طریق روشهای ارزیابی درونی انجام می‌شود. این روشها ارزیابی می‌کنند که خوشه‌ها تا چه حد از هم جدا هستند و تا چه حد به هم فشرده‌اند. نمونه‌ای از این شاخص‌ها، شاخص سیلوئت می‌باشد که طبق رابطه ۶ تعریف می‌شود.

$$s(i) = \frac{b(i) - a(i)}{max\{b(i), a(i)\}} \quad (۶)$$

a (o)به صورت میانگین فاصله بین o و سایر اشیاء خوشه‌ای که o به آن تعلق دارد، تعریف می‌شود. به روش مشابهی b(o) کمترین میانگین فاصله بین o و همه خوشه‌هایی است که o به آنها تعلق ندارد.

مقدار این شاخص بین ۱- و ۱ است. مقدار a(o) فشردگی خوشه‌ای را که o به آن تعلق دارد نشان می‌دهد. این مقدار هر چه کمتر باشد خوشه فشرده‌تر است. مقدار b(o) نشان می‌دهد که o چه‌قدر از سایر خوشه‌ها جدا است. هر چه b(o) بیشتر باشد o از سایر خوشه‌ها بیشتر جدا شده است. بنابراین در شاخص سیلوئت هر چه مقدار شاخص به یک نزدیک می‌شود خوشه فشرده‌تر است و از سایر خوشه‌ها دورتر است بنابراین حالت مطلوبی است. در شرایطی که شاخص سیلوئت منفی باشد، به این معناست که o به اشیاء خوشه دیگری غیر از خوشه‌ای که به آن تعلق دارد، نزدیکتر است. این حالت نا مطلوب است و باید از بروز آن جلوگیری کرد.

نتایج حاصل از خوشه‌بندی با معیار سیلهوئت در روش پیشنهادی محاسبه شده است و به شرح زیر می باشد:

سیلهوئت = ۰.۵۲

٤-٢- ارزیابی نتایج خوشه بندی با معیار دیویس بولدین[2]

شاخص دیویس بولدین ، معیاری است که از شباهت بین دو خوشه (Rij) استفاده می‌کند که بر اساس پراکندگی یک خوشه (si) و عدم شباهت بین دو خوشه (dij) تعریف می‌شود. معمولا شباهت بین دو خوشه طبق رابطه ۷ تعریف می‌شود.

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \quad (۷)$$

که در آن dij و si با روابط ۸ و ۹ محاسبه می‌شوند.

$$d_{ij} = d(v_i + v_j) \quad (۸)$$

$$s_i = \frac{1}{||c_i||} \sum_x d(x, v_i) \quad (۹)$$

با توجه به مطالب بیان شده و تعریف شباهت بین دو خوشه شاخص دیویس بولدین به صورت رابطه ۱۰ تعریف می‌شود.

$$DB = \frac{1}{|n_c|} \sum_{i=1}^{x_c} R_i \quad (۱۰)$$

این شاخص در واقع میانگین شباهت بین هر خوشه با شبیه‌ترین خوشهٔ به آن را محاسبه می‌کند. می‌توان دریافت که هرچه مقدار این شاخص بیشتر باشد، خوشه‌های بهتری تولید شده است.

نتایج حاصل از خوشه‌بندی با معیار دیویس بولدین برای روش پیشنهادی به قرار زیر می باشد:

دیویس بولدین = ۰.۹۵

۵- جمع بندی

در این تحقیق روشی جدید با استفاده از خوشه‌بندی و فازی سازی، به‌منظور ارائه راه‌حلی بهینه برای کاوش در میان اسناد وب ارائه شده است. روش پیشنهادی علاوه بر کاوش اسناد وب، سازمان‌دهی، دسترسی سریع و صحیح به اسناد و صفحات وب ساخت‌یافته و نیمه ساخت‌یافته را انجام می دهدبه طوری که علاوه بر ویژگی‌های ساختاری، ویژگی‌های محتوایی و معنایی اسناد نیز مورد کاوش قرار می‌گیرد. روش پیشنهادی با ۴۵ سند خبری از سایت پرس تی وی به عنوان داده های اولیه مورد ارزیابی و پیاده سازی می باشد. برای پیاده سازی روش پیشنهادی از نرم افزار های متلب و پایتون استفاده شد و نتایج پیاده سازی با توجه به شاخص دیویس بولدین و سیلهوئت نشان می دهد که روش پیشنهادی از نظر کیفیت خوشه ها و تراکم آنها مناسب می باشد.

[1] silhouette

[2] davies bouldin

٦- پیشنهاد کارهای آتی

برای کارهای آتی پیشنهاد می گردد تا تمرکز بیشتری روی راه کارهای استخراج ویژگی های اسناد و کاهش ابعاد انجام گیرد. همچنین در این تحقیق از روش خوشه بندی و فازی سازی برای دسته بندی اسناد استفاده شد، برای تحقیقات آینده، تمرکز بر روی روش های جدیدی که دقت و سرعت و کارایی بالایی برای دسته بندی اسناد داشته باشند، پیشنهاد می گردد.

مراجع

[1] C. W. Choo, E. Austerm., "Scaning the business environment: acquisition and use of information by managers," *M.E. williams (ED), annual Review of information Science and Technology,Learned Information, Inc. Medford.*, pp. 250-256, 1996.

[2] Maide Abedini Bagha, F. L., " Classifying web pages and documents based on expected cross entropy and weighted vote schema.," in *The international conference on new researches in engineering science*, Tehran, 2016.

[3] Krishna Murthy. A, Suresha, " XML URL classification based on their semantic structure orientation for web mining applications," in *International conference on information and communication technologies (ICICT 2014)*, 2015.

[4] Thasleena N.T., V. S. , "Enhanced Associative Classification of XML Documents Supported by Semantic Concepts," in *International Conference on Information and Communication Technologies (ICICT 2014).* , 2014.

[5] T.-H. H.-H. Chien-Liang Liu, "Clustering documents with labeled and unlabeled documents using fuzzy semi-Kmeans," *Fuzzy SetsandSystems 221* , p. 48–64, 2013.

[6] G. T. Q. R. C. M. M. F. Z. X. Wang W, "Autonomic Intrusion detection : Adaptively detecting anomalies over unlabeled data streams in computer networks," *Expert Systems with Applications* , p. 4062–4080., 2014.

[7] A. M. S. T. a. T. M. K. Nigam, " Text classification from labeled and unlabeled documents using em.," *Machine Learning*, p. 103–134, 2000.

[8] T. Joachims., " Text categorization with support vector machines: Learning with many relevant features," *C. Nedellec and C. Rouveirol, editors, European Conf on Machine Learning (ECML),* , 1998.

[9] Y. Hamasuna, Y. E. , "Semi-supervised fuzzy C-means clustering using clusterwise tolerance based pairwise constraints," in *Proceeding of the 2010 IEEE Initernational Conference on Granular Computing, GRC'10, IEEE Computer Society.*, 2010.

[10] Xinyue Liu, Xianchao Zhang, Ye Tian., " Webpage Segmentation based on Gomory-Hu Tree Clustering," *Undirected Planar Graph.NSFC*, 2010.

[11] K. S. T. C. Lin WC, " CANN: An intrusion detection system based on combining cluster center and nearest neighbors.," *Knowledge-based systems* , pp. 13-21., 2014.

[12] S. P. S.-j. H. Soo-Hoon Moona, "Energy efficient data collection in sink-centric wireless sensor networks: A cluster-ring approach," *Computer Communications*, pp. 1-14, 2016.

[13] Iwan Tri Riyadi Yanto, Younes Saadi, Dedy Hartama, Dewi Pramudi Ismi, Andri Pranolo, "A framework of fuzzy partition based on Artificial Bee Colony for categorical data clustering," *IEEE*, 2016.

[14] Liyong Zhang,Wei Lu, Chongquan Zhong, "Fuzzy C-Means Clustering of Incomplete Data Based on Probabilistic Information Granules of Missing Values," *Knowledge-Based Systems*, 2016.