
بسم الله الرحمن الرحيم

سمینار کارشناسی ارشد

عنوان
مطالعه و ارزیابی تکنیک طبقه بندی در متن کاوی

نگارنده:

استاد راهنما:
دکتر

زمستان ۱۳۹۳

فهرست مطالب

عنوان **صفحه**

فهرست اشکال

چکیده:

رشد فزاینده پایگاه داده‌ها در زمینه‌های مختلف از فعالیت انسان باعث شده است که نیاز به ابزارهای قدرتمند جدید، برای تغییر دادن داده به دانش مفید، افزایش یابد. جهت برآوردن این نیاز، محققان به کاوش در زمینه‌های مختلف برای یافتن روش‌ها و ایده های مناسب پرداختند. متن کاوی یکی از زمینه‌های است که به دنبال استخراج اطلاعات مفید، از داده‌های متنی بدون ساختار، به وسیله شناسایی و اکتشاف الگوها می‌باشد. ایده اصلی متن‌کاوی، یافتن قطعات کوچک اطلاعات ازحجم زیاد داده‌های متنی، بدون نیاز به خواندن تمام آن است. متن‌کاوی اطلاعات متنی غیرساخت‌یافته را استفاده می‌کند و آن را برای کشف ساختار و معناهای ضمنی پنهان در متن بررسی می‌کند. در این سمینار با توجه به اهمیت این روش مختصراً به متن کاوی، زمینه های مرتبط با آن و برخی روش‌های رایج آن از جمله طبقه بندی پرداخته شده است.

کلمات کلیدی: متن‌کاوی، طبقه بندی، داده کاوی، استخراج اطلاعات

فصل اول

کلیات تحقیق

۱-۱- مقدمه

بخش قابل توجهی از اطلاعات قابل دسترس در پایگاه داده‌های متنی (یا پایگاه داده‌های سند) که شامل مجموعه بزرگی از اسناد منابع مختلف (مثلا مقالات خبری، مقاله‌ها، کتاب‌ها، ایمیل‌ها و صفحات وب) ذخیره شده‌اند. پایگاه داده‌های متنی به علت افزایش مقدار اطلاعات موجود به فرم الکترونیکی سریع رشد می‌کنند. امروزه بیشتر اطلاعات در صنعت، کسب و کارو سازمان‌های دیگر به صورت الکترونیکی و به فرم پایگاه داده متنی ذخیره شده‌اند. داده‌های ذخیره شده در بیشتر پایگاه داده‌های متنی، داده‌های نیمه ساختاریافته هستند چون نه به طور کامل غیرساختیافته هستند و نه به طور کامل ساختیافته هستند. برای مثال یک سندشامل تعدادی فیلد ساخت یافته مانند عنوان، نویسندگان، تاریخ انتشار، رده و غیره و از طرف دیگر شامل برخی کامپوننت‌های متنی غیرساختاریافته مانند چکیده و محتویات است. تکنیک‌های بازیابی اطلاعات مانند (متدهای ایندکس کردن متن) برای هندل کردن سندهای غیر ساختاریافته ایجاد شده‌اند. تکنیک‌های بازیابی اطلاعات قدیمی برای مقدار زیادی داده متنی که بطورفزاینده افزایش می‌یابند، ناکارآمد هستند. بدون دانستن محتویات سندها، فرمول‌بندی کردنQuery های مناسب برای آنالیز کردن و استخراج کردن اطلاعات مفید از داده، مشکل است. کاربرها نیاز به ابزارهایی برای مقایسه سندهای مختلف، مرتب کردن سندها براساس مربوط بودن آن‌ها و یافتن الگوها دارند. بنابراین یکی از جدیدترین زمینه‌های مورد تحقیق در داده کاوی، متن کاوی برای این منظور گسترش یافت. متن کاوی یعنی جستجوی الگوها در متن غیرساختیافته. متن کاوی برای کشف اتوماتیک دانش مورد علاقه یا مفید از متن نیمه ساختیافته استفاده می‌شود. چندین تکنیک برای متن کاوی پیشنهاد شده است عبارتند از ساختار مفهومی،کاوش association ruleها درخت تصمیم گیری، روش‌های استنتاج قوانین، همچنین تکنیک‌های بازیابی اطلاعات برای کارهایی مانند تطبیق دادن سندها، مرتب کردن کردن، کلاسترینگ و غیره از جمله مشکلاتی که در زمینه متن کاوی وجود دارد کشف کردن دانش مفید از متن نیمه ساختیافته یا غیرساختیافته است که توجه زیادی را به خود جلب کرده است. روش‌های داده کاوی سنتی فرض می‌کنند که اطلاعات به فرم پایگاه داده‌های رابطه‌ایی هستند بهمین دلیل برای بسیاری از کاربردها مانند اطلاعات الکترونیکی قابل دسترس به فرم نیمه ساختیافته یا غیرساختیافته مفید نیستند. بدون عمل متن کاوی پردازش کردن پایگاه داده‌های متنی غیرساختیافته باید به صورت دستی توسط کاربران انجام شود که این امر بسیار طاقت فرساست. بنابراین می‌توان گفت هدف متن کاوی اتوماتیک کردن مقدار زیادی از کار کاربران است. این تحقیق، به بررسی

فیلد متن کاوی می‌پردازد. متن‌کاوی را می‌توان به عنوان یک روش میان رشته‌ای برای بازیابی اطلاعات، یادگیری ماشین، آماری، زبان دانان محاسباتی و مخصوصا داده کاوی در نظر گرفت. از آنجا که متن‌کاوی، در تکنولوژی‌های متفاوتی ریشه دارد، از این رو تعاریف زیادی نیز برای آن وجود دارد. افرادی که دارای پیشینه کار در زمینه‌ی داده کاوی بودند می‌خواستند که همان مفاهیم و روش‌های موجود در داده کاوی را بر متون اعمال کنند و تعاریف‌شان نیز منطبق بر همین زمینه بود. اما کسانی که از جامعه‌ی زبان دانان محاسباتی آمده بودند، قصد داشتند که این توانایی را به کامپیوتر بدهند که بتوانند متن را بفهمند و این غایت چیزی است که از متن کاوی مورد انتظار است.

۱-۲- تعریف مسئله و بیان سؤال‌های اصلی تحقیق

متن‌کاوی می‌تواند به عنوان یک تکنیک برای استخراج اطلاعات جالب و یا دانش از اسناد متنی که معمولاً به صورت بدون ساختار تعریف شده استفاده شود. در این مقاله [1] متن‌کاوی با تکنیک‌های مختلفی مانند خلاصه‌سازی، طبقه‌بندی که تکنیک نظارت شده یعنی دانستن تمام الگوهای ورودی و خروجی برای آموزش مدل استفاده می‌شود، خوشه بندی یعنی برای تقسیم متن به خوشه‌ها با توجه به شباهت اسنادها که این روش آموزش بدون نظارت است که در آن الگوی ورودی و خروجی از پیش تعریف شده وجود ندارد استفاده شده، استخراج اطلاعات یعنی برای استخراج اطلاعات ذخیره شده از متن بدون ساختار که در آن تکنیک‌های داده کاوی می‌توانند برای گرفتن الگوها یا دانش مفید از اسناد استفاده شود و همچنین تجسم سازی به ارائه اطلاعات قابل فهم بهتر برای استخراج اسناد استفاده می‌شود. در این سمینار، یک بحث بر سر چارچوب متن‌کاوی با تکنیک‌های مختلف گفته شده بالا جوانب مثبت و منفی و همچنین کاربردهای متن‌کاوی را مورد بحث قرار می‌دهد. علاوه بر این، بحث مختصری از مزایای متن‌کاوی و محدودیت‌ها در این مقاله ارائه شده است[2]. متن‌کاوی عمدتاً غیر ساخت‌یافته یا نیمه ساخت‌یافته هستند، ابتدا باید توسط روش‌هایی آنها را ساختارمند نمود و سپس از این روش‌ها برای استخراج اطلاعات و دانش استفاده کرد.
سوالاتی که در مورد تکنیک طبقه بندی در متن کاوی مطرح است اینکه:
چگونه می‌توان از تکنیک طبقه بندی در متن کاوی استفاده کرد؟ و یا چگونه می‌توان روش های طبقه بندی را در طرح پیشنهادی بیان کرد ؟