
بسم الله الرحمن الرحيم

سمینار

عنوان

مروری جامع بر متن کاوی و تکنیک‌های آن

فهرست مطالب

فصل سوم: مروری بر تکنیک های متن کاوی

فصل چهارم: نتیجه گیری و پیشنهادات

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 ایده‌های 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 متن کاوی، یافتن قطعات کوچک اطلاعات ازحجم زیاد داده‌های متنی، بدون نیاز به خواندن تمام آن است. در این سمینار با توجه به اهمیت این روش مختصراً به متن کاوی، زمینه های مرتبط با آن و برخی روش‌های رایج طبقه‌بندی و خوشه‌بندی پرداخته شده است. اگرچه بیان همه روش‌ها و کاربردها ممکن نیست، اما این سمینار می‌تواند دید کلی از متن کاوی را در ذهن خواننده ایجاد کرده و در صورت علاقه برای مطالعه بیشتر، فرد را به منابع مناسب هدایت کند.

واژه های کلیدی: متن‌کاوی، اطلاعات، دانش، طبقه بندی، خوه بندی.

فصل اول

مقدمه

۱-۱- مقدمه

داده‌ها، نخستین شکل اطلاعات هستند که به منظور ایجاد دانش، مدیریت و کاویده می‌شوند. داده‌ها دارای چندین مشخصه هستند: حجم، سرعت بر حسب زمان، تنوع، صحت، دوام پذیری، اعتبار، ارزش و مدت اعتبار. "حجم"، به مقدار زیاد داده‌ها بر می‌گردد. "سرعت بر حسب زمان"، نرخ تولد داده‌ها در هر واحد زمانی را نشان می‌دهد. "تنوع"، بر شکل‌های مختلف داده، مانند متن (گزارش سلامت)، عدد (آمار و ارقام بازار سهام)، تصاویر (عکسبرداری ماهواره ای)، صوت (تماس‌های تلفنی)، ویدئو و هر فرم و شکل دیگری که بتوان تصور کرد، دلالت دارد. "صحت"، با انحرافات، اختلالات و مویز در داده‌ها سروکار دارد. "دوام پذیری"، به معنای بررسی ارتباط یک متغییر در آرایه ی وسیعی از متغیرهای مربوط به داده‌های چند بعدی است و ارتباطات میان متغیرها است. "اعتبار"، این پرسش را درباره‌ی داده‌ها مطرح می‌کند که آیا آن داده، برای استفاده و کاربرد در نظر گرفته شده، قابل اعتماد و دقیق است؟. "ارزش" حاکی از اهمیت کلیدی داده‌ها است. برخی از داده‌ها میتوانند بسیار مهم باشند؛ در حالی که بعضی دیگر از ارزش کم‌تری برخوردارند. آخرین مورد، یعنی "مدت اعتبار" در این باره است که داده‌ها چه مدت اعبار دارند و می‌بایست ذخیره شوند[1].

هدف از داده کاوی، کشف ضمنی الگوها و روند ناشناخته قبلی از پایگاه داده‌ها است. داده کاوی شامل تکنیکهای بسیاری چون طبقه‌بندی، خوشه‌بندی، شبکه‌های عصبی و درخت‌های تصمیم است. اگر داده‌ها به اندازه تمامی اب سطح زمین باشند، آنگاه داده‌های متنی همانند اقیانوس بیش‌ترین بخش آن را تشکیل میدهند[2]. متن ممکن است در اندازه زیاد و فرم‌های متفاوتی هم چون زبان‌های مختلف، با استفاده از نمادهای مختلف و قالب‌های متفاوت موجود باشد. از این رو، این پرسش ایجاد می‌شود که چگونه اطلاعات را می‌توان از این متن خارج کرد، در این جاست که متن کاوی به ایفای نقش می‌پردازد[1]. متن کاوی، کاربردی از داده کاوی است. تفاوت اصلی این دو، آن است که در متن کاوی، الگوها از متنی با زبان طبیعی استخراج می‌شوند، این در حالی است که داده کاوی بر روی پایگاه داده‌های ساخت یافته عمل می‌کند. بنابراین، داده کاوی، بازیابی اطلاعات، پردازش زبان طبیعی و استخراج اطلاعات از زمینه های مرتبط با متن کاوی هستند.

داده کاوی: روشی بسیار کارا برای کشف اطلاعات از داده های ساخته یافته است. متن کاوی مشابه داده کاوی است، اما ابزارهای داده کاوی طراحی شده اند تا داده های ساخت یافته از پایگاه داده را به کار بیرند. می توان گفت، متن کاوی یک راه حل بهتر برای شرکت ها است.

بازیابی اطلاعات: معمولا در بازیابی اطلاعات با توجه به نیاز مطرح شده از سوی کاربر، مرتبط ترین متون و مستندات و یا در واقع کیسه ی کلمه از میان دیگر مستندات یک مجموعه بیرون کشیده می شود. بازیابی اطلاعات یافتن دانش نیست بلکه تنها آن مستنداتی را که مرتبط تر به نیاز اطلاعاتی جستجوگر تشخیص داده به او تحویل می دهد. این روش در واقع هیچ دانش و حتی هیچ اطلاعاتی را به ارمغان نمی اورد.

پردازش زبان طبیعی: هدف کلی آن رسیدن به یک درک بهتر از زبان طبیعی توسط کامپیوترهاست. تکنیک های مستحکم و ساده ای را برای پردازش سریع متن به کار می برد. همچنین از تکنیک های آنالیز زبان شناسی نیز برای پردازش متن استفاده می کنئ. نقش پردازش زبان طبیعی در متن کاوی فرآهم کردن یک سیستم در مرحله استخراج اطلاعات با داده های زبانی است.

استخراج اطلاعات: هدف استخراج اطلاعات خاص از سندهای متنی است و می تواند به عنوان یک فاز پیش پردازش در متن کاوی به کار رود. استخراج اطلاعات عبارتنداز نگاشت متن های زبان طبیعی به یک نمایش ساخت یافته و از پیش تعریف شده یا قالب هایی که وقتی پر می شوند، منتخبی اط اطلاعات کلیدی از متن اصلی را نشان می دهند. این سیستم های استخراج اطلاعات به شدت بر داده های تولید شده توسط سیستم های پردازش زبان طبیعی تکیه دارند.