

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

سمینار کارشناسی ارشد

عنوان

مطالعه و ارزیابی تکنیک طبقه بندی در متن کاوی

نگارنده:

استاد راهنما:

دکتر

زمستان ۱۳۹۳

چکیده..... ۶

### فصل اول: کلیات تحقیق

۷

۱-۱- مقدمه ..... ۸

۱-۲- تعریف مسئله و بیان سؤال‌های اصلی تحقیق ..... ۹

۱-۳- سابقه و ضرورت انجام تحقیق ..... ۱۰

۱-۴- هدف‌ها ..... ۱۰

۱-۵- چه کاربردهایی از انجام این تحقیق متصور است؟ ..... ۱۰

۱-۶- استفاده کنندگان از نتیجه تحقیق ..... ۱۰

۱-۷- فرضیه‌ها ..... ۱۰

۱-۸- جنبه جدید بودن و نوآوری طرح در چیست؟ ..... ۱۱

۱-۹- روش انجام تحقیق ..... ۱۱

۱-۱۰- روش و ابزار گردآوری اطلاعات ..... ۱۱

۱-۱۱- جامعه آماری و تعداد نمونه ..... ۱۱

۱-۱۲- روش نمونه‌گیری ..... ۱۱

۱-۱۳- روش تجزیه و تحلیل اطلاعات ..... ۱۱

### فصل دوم: ادبیات تحقیق

۱۲

۲-۱- مقدمه ..... ۱۳

۲-۲- سیستم‌های استخراج اطلاعات ..... ۱۳

۲-۳- داده‌کاوی در مقابل متن‌کاوی ..... ۱۴

۲-۴- متن‌کاوی ..... ۱۶

۲-۴-۱- کشف دانش و ارتباط آن با متن‌کاوی ..... ۱۸

۲-۴-۲- فرآیند متن‌کاوی ..... ۱۹

۲-۴-۳- کاربردهای متن‌کاوی ..... ۲۰

۲-۴-۴- چگونگی انجام متن‌کاوی ..... ۲۱

۲-۴-۵- اجزای متعدد در یک سیستم متن‌کاوی ..... ۲۳

۲۴	..... پرواز از طریق متن..... ۶-۴-۲
۲۴	..... زمینه‌های مرتبط با متن کاوی..... ۷-۴-۲
۲۵	..... روش‌های پیش پردازش کردن متون..... ۵-۲
۲۷	..... مدل فضای برداری..... ۱-۵-۲
۲۸	..... پیش پردازش زبان شناختی..... ۲-۵-۲
۲۹	..... روش‌های استخراج ویژگی..... ۶-۲
۲۹	..... روش فرکانس سند (DF)..... ۱-۶-۲
۲۹	..... روش اطلاعات متقابل (MI)..... ۲-۶-۲
۳۰	..... تکنیک‌های متن کاوی..... ۷-۲
۳۰	..... فازهای اصلی فرآیند متن کاوی..... ۱-۷-۲
۳۲	..... روش‌های طبقه بندی متون..... ۲-۷-۲
۳۲	..... استخراج اطلاعات..... ۳-۷-۲
۳۳	..... رده‌بندی برای استخراج اطلاعات..... ۱-۳-۷-۲
۳۴	..... مدل مارکوف پنهان..... ۲-۳-۷-۲
۳۴	..... فیلدهای رندم شرطی..... ۳-۳-۷-۲
۳۵	..... مقایسه روش‌های استخراج اطلاعات..... ۴-۳-۷-۲
۳۵	..... روش‌های ترکیبی..... ۴-۷-۲
۳۶	..... روش Discotex..... ۱-۴-۷-۲
۳۶	..... مقدمه..... ۱-۱-۴-۷-۲
۳۶	..... یکپارچه کردن داده کاوی و استخراج اطلاعات..... ۲-۱-۴-۷-۲
۳۶	..... سیستم Discotex..... ۳-۱-۴-۷-۲
۳۸	..... روش Textminer..... ۲-۴-۷-۲
۳۸	..... مقدمه..... ۱-۲-۴-۷-۲
۳۹	..... استخراج دانش مفید..... ۲-۲-۴-۷-۲
۴۱	..... الگوریتم خوشه‌بندی..... ۳-۲-۴-۷-۲
۴۴	..... فصل سوم: پیشینه تحقیق
۴۵	..... مقدمه..... ۱-۳
۴۶	..... طبقه بندی..... ۲-۳

۴۷	انتخاب ترم ایندکس
۴۷	Naïve Bayes طبقه بندی
۴۸	طبقه بندی نزدیکترین همسایه
۴۹	درخت تصمیم گیری
۵۰	درخت تصمیم متوالی بر پایه طبقه بندی
۵۰	Hunt روش
۵۰	C4.5 الگوریتم
۵۱	SPRINT الگوریتم
۵۱	فرمول بندی موازی از درخت تصمیم بر پایه طبقه بندی
۵۲	رویکرد ساخت درخت همزمان
۵۲	رویکرد ساخت درخت قسمت بندی شده
۵۴	فرمولاسیون موازی ترکیبی
۵۴	SVM متدهای هسته و
۵۶	شبکه های عصبی
۵۷	ارزیابی الگوریتم های طبقه بندی
۵۸	طبقه بندی اسناد متنی
۵۸	مطالعه مقالات مرتبط با طبقه بندی
۶۰	نتیجه گیری
۶۱	فصل چهارم: نتیجه گیری و پیشنهادات
۶۲	نتیجه گیری و پیشنهادات
۶۳	مراجع

## فهرست اشکال

شکل ۲-۱- داده کاوی	۱۵
شکل ۲-۲- طبقه بندی تکنیک های داده کاوی	۱۶
شکل ۲-۳- فرایند متن کاوی	۲۰
شکل ۲-۴- فازهای متن کاوی	۳۱
شکل ۲-۵- استخراج اطلاعات	۳۳
شکل ۲-۶- دید کلی از چارچوب متن کاوی مبتنی بر IE	۳۷
شکل ۲-۷- دیکشنری واژه های مترادف	۳۸
شکل ۲-۸- مثال یک رویداد خاص	۴۰
شکل ۲-۹- رویدادهای استخراج شده	۴۰
شکل ۲-۱۰- ورودی برای الگوریتم خوشه بندی	۴۰
شکل ۳-۱۱- hyperplane ساخته شدخ توسط SVM با ماگزیمم فاصله	۵۵

## 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 متن‌کاوی، یافتن قطعات کوچک اطلاعات از حجم زیاد داده‌های متنی، بدون نیاز به خواندن تمام آن است. متن‌کاوی اطلاعات متنی غیرساخت‌یافته را استفاده می‌کند و آن را برای کشف ساختار و معنای ضمنی پنهان در متن بررسی می‌کند. در این سمینار با توجه به اهمیت این روش مختصراً به متن کاوی، زمینه‌های مرتبط با آن و برخی روش‌های رایج آن از جمله طبقه‌بندی پرداخته شده است.

**کلمات کلیدی:** متن‌کاوی، طبقه‌بندی، داده کاوی، استخراج اطلاعات

# فصل اول

## کلیات تحقیق

بخش قابل توجهی از اطلاعات قابل دسترس در پایگاه داده‌های متنی (یا پایگاه داده‌های سند) که شامل مجموعه بزرگی از اسناد منابع مختلف (مثلاً 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 کار و 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 غیرساختاریافته هستند و نه به طور کامل ساختاریافته هستند. برای مثال یک سند شامل 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 بطور فزاینده افزایش می‌یابند، ناکارآمد هستند. بدون دانستن محتویات سندها، فرمول‌بندی کردن Query 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 غیرساختاریافته. متن کاوی برای کشف اتوماتیک دانش مورد علاقه یا مفید از متن نیمه ساختاریافته استفاده می‌شود. چندین تکنیک برای متن کاوی پیشنهاد شده است عبارتند از ساختار مفهومی، کاوش association ruleها درخت تصمیم‌گیری، 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 ساختاریافته یا غیرساختاریافته است که توجه زیادی را به خود جلب کرده است. روش‌های داده کاوی سنتی فرض می‌کنند که اطلاعات به فرم پایگاه داده‌های رابطه‌ای هستند بهمین دلیل برای بسیاری از کاربردها مانند اطلاعات الکترونیکی قابل دسترس به فرم نیمه ساختاریافته یا غیرساختاریافته مفید نیستند. بدون عمل متن کاوی پردازش کردن پایگاه داده‌های متنی غیرساختاریافته 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. متن کاوی را می‌توان به عنوان یک روش میان رشته‌ای برای بازیابی اطلاعات، یادگیری ماشین، آماری، زبان دانان محاسباتی و مخصوصاً داده کاوی در نظر گرفت. از آنجا که متن کاوی، 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

متن کاوی 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] متن کاوی 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 متن کاوی با تکنیک‌های مختلف گفته شده بالا جوانب مثبت و منفی و همچنین کاربردهای متن کاوی را مورد بحث قرار می‌دهد. علاوه بر این، بحث مختصری از مزایای متن کاوی و محدودیت‌ها در این مقاله ارائه شده است [2]. متن کاوی عمدتاً غیر ساخت یافته یا نیمه ساخت یافته 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 اینک:

چگونه می‌توان از تکنیک طبقه بندی در متن کاوی استفاده کرد؟ و یا چگونه می‌توان روش‌های طبقه بندی را در طرح پیشنهادی بیان کرد؟