

داده متن کاوی

نویسنده:

Charu C. Aggarwal • Cheng Xiang Zhai

مترجم

مهندس سیده مریم علی کوثری

عنوان و نام پدیدآور	:	داده‌متن‌کاوی/ نویسنده [صحیح: ویراستاران چاروسی. آگروال، چنگ‌شیانگ جای]: مترجم سیده‌مریم علی‌کوثری.
مشخصات نشر	:	تهران: آرنا، ۱۳۹۶.
مشخصات ظاهری	:	۴۹۰ ص.
شابک	:	978-600-356-677-4
وضعیت فهرست نویسی	:	فیا
یادداشت	:	عنوان اصلی: Mining text data, c2012.
یادداشت	:	کتابنامه.
موضوع	:	داده‌کاوی
موضوع	:	Data mining
شناسه افزوده	:	آگروال، چارو سی.، ۱۹۷۰ - م.، ویراستار
شناسه افزوده	:	Aggarwal, Charu C.
شناسه افزوده	:	جای، چنگ‌شیانگ، ویراستار
شناسه افزوده	:	Zhai, ChengXiang
شناسه افزوده	:	علی‌کوثری، سیده‌مریم، ۱۳۶۰ - مترجم
رده بندی کنگره	:	۱۳۹۶ ۲۵۲۳/ ۹/۷۶QA
رده بندی دیویی	:	۳۱۲/۰۰۴
شماره کتابشناسی ملی	:	۵۰۴۰۳۵۹



نشر آرنا

www.arnapub.com

عنوان	:	داده متن کاوی
مترجم	:	سیده مریم علی کوثری
ناشر	:	آرنا
چاپ	:	اول ۱۳۹۶
شمارگان	:	۱۰۰۰ نسخه قیمت: ۳۴۰۰۰ تومان
شابک	:	۹۷۸-۶۰۰-۳۵۶-۶۷۷-۴

www.arnapub.com

فصل ۱	۲۰
معرفی متن کاوی	۲۰
۱ مقدمه	۲۰
۲ الگوریتم های جهت متن کاوی	۲۲
۳ راه کردهای آینده	۲۶
منابع	۲۸
فصل ۲	۲۹
استخراج اطلاعات از متن	۲۹
۱. مقدمه	۲۹
۲ شناسایی موجودیت نام دار	۳۳
۱-۲ رویکرد مبتنی بر قانون	۳۵
۲-۲ رویکرد های یادگیر آماری	۳۶
۱-۲-۲ مدل های مخفی مارکف	۳۷
۴-۲ مدل مارکوف بیشینه آنتروپی	۳۸
۳-۲-۲ حوزه های تصادفی شرطی	۳۹
۳ استخراج رابطه	۴۱
۱-۳ ویژگی مبتنی بر دسته بندی	۴۲
۲-۳ روش های هسته:	۴۶
۳-۳ روش های نظارتی یادگیر ضعیف:	۴۸
۱-۳-۳ بوت استرپینگ:	۴۹

۴۹	۲-۳-۳ نظارت راه دور:
۵۰	۴ استخراج اطلاعات غیر نظارتی:
۵۰	۴-۱ رابطه های اکتشافی و الگو القایی
۵۲	۲-۴ استخراج اطلاعات باز
۵۳	۵ ارزیابی
۵۴	۶ نتیجه گیری و خلاصه
۵۴	منابع
۶۰	فصل ۳
۶۰	بررسی روش های خلاصه سازی متن
۶۰	۱. سیستم خلاصه سازهای اسب اچ چگونه عمل می کند
۶۲	۲. نظریه های ارائه عنوان
۶۳	۲,۱ کلمات عنوانی
۶۴	۲,۲ نظریه های تکرار-گرا
۶۸	۲,۳ تجزیه و تحلیل معنایی نهفته
۶۹	۲,۴ مدل های عنوان ی بیزی
۷۱	۲,۵ دسته بندی جملات و عنوانها وابسته به دامنه
۷۳	۳ تاثیر زمینه
۷۳	۳,۱ خلاصه سازی وب
۷۴	۳,۲ خلاصه سازی مقالات علمی
۷۴	۳,۳ خلاصه سازی مبتنی بر پرس و جو
۷۵	۳,۴ خلاصه سازی ایمیل
۷۶	۴ یادگیری ماشین و ارائه های شاخص برای خلاصه سازی
۷۶	۴,۱ روش های گراف برای اهمیت جمله

۴,۲ یادگیری ماشین برای خلاصه سازی	۷۸
۵ انتخاب جملات خلاصه	۸۰
۵,۱ نظریه های حریصانه: ماکزیمم ارتباط حاشیه ای	۸۰
۵,۲ انتخاب خلاصه سراسری	۸۱
۶ نتیجه گیری	۸۱
متابع	۸۲
فصل ۴	۹۰
بررسی الگوریتم های خوشه بندی متن	۹۰
۱. مقدمه	۹۰
۲. روش های انتخاب ویژگی و تغییر شکل مربوط به خوشه بندی متن	۹۳
۱.۲ روش های انتخاب ویژگی	۹۳
۱.۱. ۲ انتخاب سند مبتنی بر فرکانس	۹۳
۲.۱. ۲ قدرت عبارت	۹۴
۲.۱. ۳ رتبه بندی مبتنی بر آنالوژی	۹۵
۲.۱. ۴ عبارت مشارکتی	۹۶
۲.۲ روش های مبتنی بر LSI	۹۶
۲.۱. ۲ تجزیه مفهوم با استفاده از خوشه بندی	۹۸
۲.۳. ۳ فاکتورگیری نامنفی ماتریس	۹۸
۳. الگوریتم های خوشه بندی مبتنی بر مسافت	۱۰۱
۳.۱ الگوریتم های خوشه بندی تجمعی و سلسله مراتبی	۱۰۲
۳.۲ الگوریتم های پارتیشن بندی مبتنی بر مسافت	۱۰۴
۳.۳ روش هیبریدی روش Scatter-Gather	۱۰۶
۳.۳. ۱ برنامه ریزی برای خوشه بندی سند کارآمد	۱۰۹

۴.	خوشه بندی مبتنی بر کلمه و عبارات	۱۱۰
۴.۱	خوشه بندی با الگوی‌های کلمه ی تکراری	۱۱۱
۴.۲	توانمندسازی خوشه‌های کلمه برای خوشه‌های سند	۱۱۲
۴.۳	خوشه بندی همزمان کلمات و اسناد	۱۱۳
۴.۳.۱	خوشه بندی همزمان با پارتیشن بندی گراف	۱۱۴
۴.۳.۱	خوشه بندی همزمان اطلاعات نظری	۱۱۵
۴.۴	خوشه بندی با عبارت‌های تکراری	۱۱۶
۵.	خوشه بندی سند احتمالاتی و عنوان مدلسازی	۱۱۷
۶.	خوشه بندی آنلاین با ریان‌های متنی	۱۲۰
۷.	خوشه بندی متن در شبکه ها	۱۲۶
۸.	خوشه بندی نیم نظارت	۱۲۸
۹.	نتایج و خلاصه	۱۳۰
	منابع	۱۳۱
	فصل ۵	۱۳۷
	کاهش ابعاد و عنوان مدلسازی	۱۳۷
۱.	مقدمه	۱۳۷
۱.۱	رابطه میان خوشه بندی، کاهش ابعاد و عنوان مدلسازی	۱۳۹
۱.۲	نشانه گذاری و مفهومیها	۱۳۹
۲.	شاخص گذاری معنایی پنهان	۱۴۰
۲.۱	فرآیند شاخص گذاری معنایی پنهان	۱۴۱
۲.۲	چالش های پیاده سازی	۱۴۳
۲.۲.۱	نمایش ماتریکس سند-عبارت	۱۴۳
۲.۲.۲	محاسبات	۱۴۳

۱۴۳.....	۲,۲,۳ اداره کردن تغییرات
۱۴۴.....	۲,۲,۴ بروز رسانی فضای مفهومی
۱۴۵.....	۲,۳ بررسی و تحلیل
۱۴۶.....	۳. عنوان مدلسازی و کاهش بعد
۱۴۷.....	۳,۱ شاخص گذاری مفهومی پنهان احتمالی
۱۴۸.....	۳,۱,۱ الگوریتم‌ها
۱۴۹.....	۳,۱,۲ بروز رسانی
۱۴۹.....	۲-۳ تخصیص: بریکائ پنهان
۱۴۹.....	۳-۲-۱-۱
۱۵۲.....	۳-۲-۲-۲- گیبس های دسته شده :
۱۵۳.....	۳-۲-۳- تقریب تغییر
۱۵۵.....	۳-۲-۴- پیاده سازی :
۱۵۵.....	۴- تفسیر و ارزیابی :
۱۵۶.....	۴-۱- تفسیر :
۱۵۶.....	۴-۲- ارزیابی
۱۵۷.....	۴-۳- انتخاب پارامتر
۱۵۷.....	۴-۴- کاهش بُعد
۱۵۸.....	۵. فراتر از تخصیص دیریکله پنهان
۱۵۸.....	۵-۱- مقیاس پذیری
۱۵۹.....	۵-۲- داده‌های دینامیکی
۱۵۹.....	۵-۳- داده‌های شبکه‌ای
۱۶۱.....	۵-۴- مطابقت مدل‌های عنوانی با کاربردها
۱۶۲.....	۶- نتیجه‌گیری

منابع	۱۶۲
فصل ۶	۱۶۷
مطالعه الگوریتم های دسته بندی متن	۱۶۷
۱. مقدمه	۱۶۷
۲. انتخاب ویژگی برای دسته بندی متن	۱۷۰
۱,۲ شاخص جینی	۱۷۱
۲,۲ بهره اطلاعاتی	۱۷۲
۳,۲ اطلاعات مفایا	۱۷۳
۴,۲ آمار X^2	۱۷۴
۵,۲ روش های دیگر گونی ویژگی	۱۷۴
۶,۲ خوشه بندی تحت نظارت برای کاهش ابعاد	۱۷۶
۷,۲ تحلیل متمایز کننده خطی	۱۷۶
۸,۲ تجزیه مقدار منفرد تعمیم یافته	۱۷۸
۹,۲ تعامل انتخاب ویژگی با دسته بندی	۱۷۹
۳. دسته بندی کننده های درخت تصمیم	۱۷۹
۴. دسته بندی کننده های مبتنی بر الگو (قانون):	۱۸۲
۵. دسته بندی کننده های احتمالی و ساده بیزی	۱۸۵
۱,۵ مدل چند متغیره برنولی	۱۸۶
۲,۵ توزیع چند جمله ای	۱۹۱
۳,۵ مدل سازی مخلوط برای دسته بندی متن	۱۹۳
۶. دسته بندی کننده های خطی	۱۹۶
۱,۶ دسته بندی کننده های SVM	۱۹۷
۲,۶ دسته بندی کننده های مبتنی بر رگرسیون	۱۹۸

۳,۶	دسته بندی کننده های شبکه عصبی	۲۰۰
۴,۶	برخی مشاهدات در مورد دسته بندی کننده های خطی	۲۰۲
۷	دسته بندی کننده های مبتنی بر مجاورت	۲۰۳
۸	دسته بندی داده های پیوندی و وب	۲۰۵
۹	فرا الگوریتم های مربوط به دسته بندی متن	۲۱۱
۱,۹	یادگیری گروه دسته بندی	۲۱۱
۲,۹	روش های داده محور: بوستینگ و بگینگ	۲۱۲
۳,۹	بهینه سازی معیار های خاصی دقت	۲۱۳
	منابع	۲۱۵
	فصل هفتم	۲۲۳
	انتقال یادگیری برای متن کلاس	۲۲۳
	۱. مقدمه	۲۲۳
	۲. انتقال یادگیری در دسته بندی متن	۲۲۵
	۲,۱. دسته بندی متون دامنه متقابل	۲۲۵
	۲,۱,۱. ماشین های بردار پشتیبان برای دسته بندی متن	۲۲۵
	۲,۲. انتقال مبتنی بر نمونه	۲۳۱
	۲,۳. یادگیری گروهی دامنه متقابل	۲۳۳
	۲,۴. انتقال معنا بر اساس ویژگی برای دسته بندی سند	۲۳۵
	۳. انتقال یادگیری ناهمگن	۲۴۱
	۳,۱. فضای ویژگی ناهمگن	۲۴۲
	۳,۲. فضای برچسب ناهمگن	۲۴۵
	۳,۳. خلاصه	۲۴۶
	۴. بحث و گفتگو	۲۴۷

۲۴۹	منابع
۲۵۸	فصل هشتم
۲۵۸	مدل های احتمالی برای متن کاوی
۲۵۸	۱-مقدمه
۲۶۰	۲-مدل های ترکیبی
۲۶۰	۱-۲-۱-چارچوب مدل ترکیبی کلی
۲۶۲	۲-۲-تغییرات و کاردها
۲۶۲	۱-۲-۲-مدل های عنوان
۲۶۴	۲-۲-۲-کاربردهای دیگر
۲۶۴	۲-۳-الگوریتم های یادگیری
۲۶۵	۲-۳-۲-الگوریتم EM
۲۶۷	۲-۳-۳-نمونه برداری گیبس
۲۶۸	۳-فرآیندهای اتفاقی در مدل های غیر پارامتری بیزی
۲۶۸	۱-۳-فرآیند رستوران چینی
۲۶۹	۲-۳-فرآیند دیریکله
۲۶۹	۱-۲-۳-مروری بر فرآیند دیریکله
۲۷۱	۲-۲-۳-مدل ترکیبی فرآیند دیریکله
۲۷۲	۳-۲-۳-الگوریتم های یادگیری
۲۷۲	۴-۲-۳-کاربردها در متن کاوی
۲۷۳	۳-۳-فرآیند پیتن-یور
۲۷۴	۴-۳-دیگران
۲۷۵	۴-مدل های گرافیکی
۲۷۵	۱-۴-شبکه های بیزی

۲۷۵.....	۴-۱-۱-مرور کلی.....
۲۷۷.....	۴-۱-۲-الگوریتم های یادگیری.....
۲۷۷.....	۴-۱-۳-کاربردها در متن کاوی.....
۲۷۸.....	۴-۲-مدل های مارکوف پنهانی.....
۲۷۸.....	۴-۲-۱-مرور کلی.....
۲۷۹.....	۴-۲-۲-الگوریتم های یادگیری.....
۲۸۱.....	۴-۲-۳-کاربردها متن کاوی.....
۲۸۲.....	منابع.....
۲۸۷.....	فصل ۹.....
۲۸۷.....	جریان های متن کاوی.....
۲۸۷.....	۱. مقدمه.....
۲۸۸.....	۲. خوشه بندی جریان های متن.....
۲۹۶.....	۲.۱ تشخیص موضوع و ردیابی در جریان های متن.....
۳۰۱.....	۳. دسته بندی جریان های متنی.....
۳۰۴.....	۴. تجزیه و تحلیل تکامل در جریان های متن.....
۳۰۵.....	۵. نتیجه گیری.....
۳۰۶.....	منابع.....
۳۰۹.....	فصل دهم.....
۳۰۹.....	کاوش همه زبانی از داده ی متنی.....
۳۰۹.....	۱. مقدمه.....
۳۱۰.....	۲- متن کاوی سنتی همه زبانه ی - ترجمه ی ماشینی.....
۳۱۱.....	۲-۱-SMT و مدل های ترجمه ی زایشی.....
۳۱۳.....	۲-۲-مدل های مبتنی بر کلمه.....

۳۱۵.....	۲-۳-مدل‌های مبتنی بر عبارت.....
۳۱۹.....	۲-۴-مدل‌های مبتنی بر نحو.....
۳۲۲.....	۳-کاوش خودکار متون موازی.....
۳۲۳.....	۳-۱-استفاده از ساختار وب.....
۳۲۵.....	۳-۲-تطبیق صفحات موازی.....
۳۲۷.....	۴-استفاده از مدل‌های ترجمه در CLIR.....
۳۳۰.....	۵-جمع‌آوری و بهره‌برداری از متون قابل مقایسه.....
۳۳۳.....	۶-انتخاب جملات، عبارات موازی و کلمات ترجمه.....
۳۳۶.....	۷-کاوش روابط همه زبانه از میان تک زبانه.....
۳۳۸.....	۸-کاوش با استفاده از ابرلینک.....
۳۳۹.....	۹-نتایج و مباحث.....
۳۴۱.....	منابع.....
۳۴۶.....	فصل ۱۱.....
۳۴۶.....	متن کاوی در چند رسانه‌ها.....
۳۴۷.....	۱.مقدمه.....
۳۴۸.....	۲. کاوی متن‌های احاطه شده.....
۳۵۰.....	۳. تگ کاوی.....
۳۵۱.....	۳،۱ رتبه‌بندی تگ.....
۳۵۱.....	۳،۲ پالایش تگ.....
۳۵۳.....	۳،۳ غنی‌سازی تگ.....
۳۵۵.....	۴. اتصال متن و کاوی محتوای بصری.....
۳۵۶.....	۴،۱ رتبه‌بندی مجدد بصری.....
۳۵۸.....	۵. متن متقابل و محتوا کاوی بصری.....

۳۶۱	۶. خلاصه و گسترش مسائل
۳۶۱	اتصال متن و رتبه بندی چند رسانه ی محتوای بصری
۳۶۲	متن کاوی مقیاس پذیر برای مدیریت چند رسانه ای در قیاس بزرگ
۳۶۲	شبکه اجتماعی چندرسانه ای کاوی
۳۶۳	منابع
۳۶۸	فصل ۱۲
۳۶۸	آنالیز متن در رسانه های اجتماعی
۳۶۸	۱. مقدمه
۳۷۱	۲. جنبه های متمایز متن در رسانه های اجتماعی
۳۷۱	۲.۱. چارچوب عمومی برای آنالیز متن
۳۷۴	۲.۲. حساسیت زمانی
۳۷۴	۲.۳. طول کوتاه
۳۷۵	۲.۴. عبارت های بدون ساختار
۳۷۶	۲.۵. فراوانی اطلاعات
۳۷۶	۳. به کارگیری آنالیز متن برای رسانه های اجتماعی
۳۷۷	۳.۱. تشخیص رویداد
۳۷۸	۳.۲. پاسخدهی مشترک به سوال
۳۸۰	۳.۳. تگ کردن (برچسب زدن) اجتماعی
۳۸۱	۳.۴. از بین بردن شکاف معنایی
۳۸۲	۳.۵. بهره برداری از قدرت اطلاعات فراوان
۳۸۴	۳.۶. تلاش های مرتبط
۳۸۴	۴. یک مثال گویا
۳۸۵	۴.۱. استخراج عبارت دانه

۳۸۷.....	۴.۲ تولید ویژگی معنایی
۳۸۷.....	۴.۲.۱ پایگاه دانش زمینه
۳۸۸.....	۴.۲.۲ تولید کننده ویژگی
۳۸۸.....	۴.۳ ساخت فضای ویژگی
۳۹۰.....	۵. نتیجه گیری و کار برای آینده
۳۹۱.....	منابع
۳۹۶.....	فصل ۱۳
۳۹۶.....	بررسی نقد کاوش و تحلیل احساسات
۳۹۷.....	۱.۱. تعریف نقد
۴۰۱.....	۱.۲. خلاصه انتقادات مبتنی بر دیدگاه
۴۰۲.....	۲. دسته بندی تمایلی اسناد
۴۰۳.....	۲.۱. دسته بندی مبتنی بر یادگیری نظارتی
۴۰۴.....	۲.۲. دسته بندی مبتنی بر یادگیری نظارت شده
۴۰۷.....	۳. عبارت ذهنیت و دسته بندی تمایلات
۴۰۸.....	۴. توسعه فرهنگ نامه نظرات
۴۰۹.....	۴.۱. روش مبتنی بر فرهنگ لغت
۴۰۹.....	۴.۲. روش مبتنی بر متن و سازگاری تمایلی
۴۱۱.....	۵. تحلیل تمایلات مبتنی بر دیدگاه
۴۱۲.....	۵.۱. دسته بندی تمایلات دیدگاه
۴۱۴.....	۵.۲. قوانین پایه ای انتقادات
۴۱۷.....	۵.۳. استخراج دیدگاه ها
۴۱۹.....	۵.۴. گسترش همزمان فرهنگ لغت انتقادی و استخراج دیدگاه
۴۲۰.....	۶. کاوش انتقادات مقایسه ای

۴۲۲	۷. پاره ای مسایل دیگر.....
۴۲۵	۸. تشخیص اسپم انتقادی.....
۴۲۶	۸,۱. شناسایی اسپم ها براساس یادگیری نظارت شده.....
۴۲۷	۸,۲. شناسایی اسپم براساس رفتارهای غیر معمول.....
۴۲۸	۸,۳. شناسایی اسپم گروهی.....
۴۲۸	۹. سودمندی نظرات.....
۴۳۰	۱۰. نتیجه گیری.....
۴۳۰	منابع.....
۴۴۱	فصل ۱۴.....
۴۴۱	متن کاوی زیست پزشکی: مروری بر پیشرفت های اخیر.....
۴۴۱	۱. مقدمه.....
۴۴۲	۲. منابع برای متن کاوی زیست پزشکی.....
۴۴۳	۱,۲ مجموعه متون.....
۴۴۵	۲,۲ توضیح.....
۴۴۶	۳,۲ منابع دانش.....
۴۴۷	۴,۲ ابزار های پشتیبان.....
۴۴۸	۳. استخراج اطلاعات.....
۴۴۹	۱,۳ شناسایی موجودیت نامدار.....
۴۵۵	۳,۲ استخراج رابطه.....
۴۵۸	۳,۳ استخراج رویداد.....
۴۶۱	۴. خلاصه سازی.....
۴۶۵	۵. پاسخ دهی به پرسش.....
۴۶۶	۵,۱. پاسخ دهی به پرسش پزشکی.....

۴۶۸	۵,۲. پاسخ دهی به پرسش زیستی.....
۴۶۹	۶. کشف مہتہی ہر ادبیات تحقیق.....
۴۷۲	۷. نتیجہ گیری.....
۴۷۳	منابع

www.ketab.ir

مهمترین برنامه های کاربردی متن کاوی در دهه اخیر به دلیل تعداد زیادی از برنامه های کاربردی وب که در حال ایجاد چنین داده های است، افزایش یافته است؛ درحالی که برنامه های کاربردی کلاسیک بر روی پردازش و متن کاوی خام متمرکز شده است، با ظهور برنامه های کاربردی وب یک راه و روش های جدیدی برای پردازشها و داده کاوی مانند استفاده از فرایندها، اطلاعات چند زبانه ای یا ملحق شدن متن کاوی با انواع داده های چند رسانه ای مانند عکس و فیلم نیازمند است. همچنین در خیلی از موارد، آنها را به طرف زمینه ای تحقیقاتی دیگر مانند جریان یادگیرناهمگن توسعه می دهد.

مهمترین ویژگی این زمینه در این است که به وسیله جوامع چندگانه مانند داده کاوی و یادگیری ماشین، بازایی اطلاعات کشف شده است. در خیلی از مواقع، این جوامع تمایل به ترکیب شدن دارند اما تا حد زیادی از هم فاصله دارند و هر کدام نیازمند تحقیقات جداگانه خودش را دارد. یکی از اهداف این کتاب این است که پژوهشگران را به دور هم جمع میکند که به درک انضباط متقابل بالای برساند.

به نظر می رسد که یکی دیگر از عرصه های متن کاوی کارکردن مجموعه ای از محققان به طور مجزا بر روی حالتهای جدید متن کاوی در زمینه سیستم های رم افزاری در حال ظهور مانند جریان داده و شبکه های اجتماعی می باشد. همچنین این کتاب سعی دارد در مورد جنبه های مدرن و کلاسیک متن کاوی بپردازد. فصل های این کتاب در زمینه روشها، کلاسیک همانند خوشه بندی، دسته بندی و عنوان مدلسازی تفسیم بندی شده است. همچنین ما بر روی زمینه های متفاوتی از داده کاوی در زمینه برنامه های کاربردی مدرن و شبکه های اطلاعاتی و رسانه های اجتماعی بپردازیم. و خیلی از برنامه های کاربردی جدید مانند جریان داده در این کتاب برای اولین بار نقد و بررسی شده است.

هر بخش از این کتاب براساس جستجوهای گسترده ای شکل گرفته است که در این کتاب کلیدی و الگوریتم های برای هر ناحیه خاص خود مورد بحث قرار گرفته است. که در آینده و با دستورالعمل های محققان در هر فصل به نمایش گذاشته می شود. امیدوار هستیم که در این کتاب بتوانیم برای دانشجویان، استادان، پرفسورها و محققین زمینه ادراک بزرگی را به وجود بیاوریم.