

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

مروری بر داده کاوی

با ریچارد وب سرویس کاوی

بهروز نیرومندفام

انتشارات ابداع

سرشناسه : نیرومندفام، بهروز، 1360-
عنوان و نام پدیدآور : مروری بر داده‌کاوی با رویکرد وب سرویس‌کاوی/بهروز نیرومندفام.
مشخصات نشر : تهران: نشر ابداع، 1397.
مشخصات ظاهری : 136ص.
شابک : 978-600-93708-7-0
وضعیت فهرست نویسی : فیبا
یادداشت : واژه‌نامه.
موضوع : داده‌کاوی
موضوع : Data mining
رده بندی کنگره : QA76/9 ن9 1397
رده بندی دیویی : 006/312
ماره کتابشناسی ملی : 5506895

عنوان: مروری بر داده‌کاوی با رویکرد وب سرویس‌کاوی
ناشر: ابداع

نوبت چاپ: اول ۹۷

مؤلف: بهروز نیرومندفام

تیراژ: ۱۰۰۰ نسخه

شابک: ۹۷۸-۶۰۰-۹۳۷۰۸-۷-۰

لیتوگرافی: مینیاتور

چاپ: فرشپوه

قیمت : ۲۰۰۰۰ تومان

آدرس: تهران-میدان انقلاب-خیابان کارگر جنوبی بین خ طارک و روانمهر- بن بست

گشتاسب- پلاک ۷- واحد ۱ - تلفن: ۶۶۴۷۲۲۳۰

فهرست مطالب

2	چکیده
4	مقدمه
6	۱- فصل اول: مقدمه ای بر داده کاوی (کلیات و پیشینه و بیان اهداف)
6	۱-۱- داده کاوی چیست؟
15	۱-۲- توصیف داده ها در داده کاوی
18	۱-۳- مدل های پیش بینی: داده ها
18	۱-۳-۱ Classification
18	۱-۳-۲ Regression
19	۱-۳-۳ Time series
20	۲- فصل دوم: مدل ها و الگوریتم ها در داده کاوی
20	۲-۱- تکنیکها داده کاوی
20	۲-۱-۱ شبکه های عصبی
22	۲-۱-۲ Feed-Forward Backpropagation
23	۲-۱-۳ Decision Trees
26	۲-۱-۴ Multivariate Adaptive Regression Splines (MARS)
27	۲-۱-۵ Rule Induction
27	۲-۱-۶ k-nearest neighbour and memory-based reasoning (MBR)
28	۲-۲- مدل های داده کاوی
28	۲-۲-۱ رگرسیون منطقی
29	۲-۲-۲ تحلیل تفکیکی
30	۲-۲-۳ مدل افزودنی کلی (GAM)
30	۲-۲-۴ Boosting
30	۲-۳- سلسله مراتب انتخابها
33	۳- فصل سوم: وب کاوی
37	۳-۱- مقدمه ای برای وب کاوی

38	۳-۲ محتوا کاوی وب
38	۳-۳ ساختار کاوی وب
40	۳-۳-۱ تکنیکهای تحلیل ابرپیوندها
42	۳-۳-۲ الگوریتم رتبه بندی صفحه وب
43	۳-۳-۳ الگوریتم HITS
47	۳-۴ کاوش کاربردی وب
60	۳-۵ سیستم وب کاوی
66	۳-۵-۱ الگوریتم KIM
71	۳-۵-۲ ارزیابی سیستمی
75	۴- فصل چهارم: سرویس‌ها و وب سرویس کاوی
75	۴-۱ سرویس چیست؟
80	۴-۲ وب سرویس استاندارد برای داده کاوی
88	۴-۳ وب سرویس کاوی
99	۵- فصل پنجم: نتیجه گیری و پیشنهادات
99	۵-۱ نتیجه گیری
100	۵-۲ بیان مسئله و پیشنهادات
105	پیوستها
106	پیوست ۱: لغت نامه
114	پیوست ۲: Service Oriented Architecture
120	پیوست ۳: وب سرویس
127	پیوست ۴: XML



چکیده

واژه داده کاوی برای توصیف مجموعه فعالیتهای وسیعی مورد استفاده قرار می گیرد. دنیای مدرون امروز، دنیای اطلاعات است و ما با حجم بسیار وسیعی از داده ها سرکار داریم. برای رسیدن به اطلاعات باید این داده ها تحلیل و بررسی شوند تا اطلاعات از آنها استخراج شود.

توسعه سریع و افزایش قیمت‌های، کاهش هزینه های ذخیره سازی منجر به ذخیره بسیار وسیع داده با هزینه ذخیره سازی پایدار شده است. وجود داده های زیادی بر روی وب سایتها، پایگاه داده ها، مالتی مدیا، وب سرویسها و ... اخیرا استفاده از داده کار را از این منابع مد نظر قرار داده است. کاوش اطلاعات مفید که دانش مناسب را از پایگاه داده های وسیع استخراج می کند، به یکی از زمینه های مهم تحقیقاتی تبدیل شده است.

وب کاوی یکی از زمینه های مهم تحقیقاتی در زمینه داده کاوی برای تعداد وسیعی از سرویسهای وب جهان گستر در سالهای اخیر تبدیل شده است. وب جهان گستر بصورت گسترده ای هم در زمینه ترافیک و هم در زمینه اندازه و پیچیدگی وب سایتها در حال گسترش است. پیچیدگی، نیاز ما را به تحلیل استفاده وب سایت، چگونگی ساختار وب و محتوای منابع وب وادار ساخته است.

تکنولوژی وب سرویس باعث افزایش نقش وب سرویسها در برنامه های اینترنتی بطور کلی و در زمینه تجاری به طور خاص شده است. از زمانی که وب سرویسها نقش مهم و فزاینده ای در تکنولوژی اطلاعات پیدا کرده اند، از سیستمهای سرویسگرا انتظار رشد و پیچیدگی روزافزون می رود. مانند معاضای سیستمهای برای ابزارهای که اجازه تحلیل و نظارت بر سیستمهای سرویسگرا استفاده شده را می دهند.

روشهای مختلفی برای داده کاوی و فرایند کاوی از وب سرویسها و تعاملات بین آنها برای تحلیل تعاملات وب سرویسها بین استفاده کننده گان و مهیا کننده گان و کشف ترکیب سرویسهاست وجود دارد. از زمانیکه استفاده از وب سرویسها افزایش پیدا کرده، تحلیل رفتار کاربرانی که از وب سرویسها استفاده می کنند نیز به

مقدمه

اخیراً توانایی های فنی بشر در تولید و جمع آوری داده ها به سرعت افزایش یافته است عواملی نظیر استفاده گسترده از کامپیوتر در کسب و کار، علوم، خدمات دولتی و پیشرفت در وسائل جمع آوری داده، از اسکن کردن متون و تصاویر تا سیستمهای سنجش از دور ماهواره ای، در این تغییرات نقش مهمی داشته اند.

حجم بسیار بالای سایت های طراحی شده و موجود در بستر وب؛ تنوع و حجم بالای اطلاعات موجود در آنها، تکنیکها و توزیع شدگی منابع به معظلی بزرگ در استفاده از این منابع تبدیل شده است. در واقع مشکل از آنجا ناشی می شود که منابع به اندازه کافی موجود است، ولی بصورت خام و پخش شده در رسانه های مختلف و با حجم زیاد، شکاف اصلی در بدست آوردن دانش موجود در آنها و اینکه دانش مورد نیاز ما در چه منابعی و به چه صورت واقع شده است؛ می باشد.

این رشد انفجاری در داده ها، ذخیره شده، نیاز مبرم به وجود تکنولوژی های جدید و ابزارهای خودکاری را ایجاد کرده که به صورت هوشمند به انسان های رسانند تا این حجم زیاد داده را به اطلاعات و دانش تبدیل کنند: داده کاوی به عنوان یک راه حل برای این مشکل مطرح می شود.

بطور غیر رسمی داده کاوی فرآیندی است، خوندار برای استخراج الگوهایی که دانش را بازنمایی می کنند، که این دانش به صورت ضمنی در پایگاه داده های عظیم، داده ها، صفحات وب، منابع توزیع شده و دیگر مخازن بزرگ اطلاعات، ذخیره شده است. داده کاوی تکنیکی است که بطور همزمان از چندین رشته علمی نظیر: تکنولوژی پایگاه داده، هوش مصنوعی، یادگیری ماشین، شبکه های عصبی، آمار، شناسایی الگو، سیستم های مبتنی بر دانش، حصول دانش، بازیابی اطلاعات، محاسبات سرعت بالا، رایانه های بصری داده بهره می برد.

- ¹ Data warehouses
- ² Knowledge-based system
- ³ Knowledge-acquisition
- ⁴ Information retrieval
- ⁵ High-performance computing
- ⁶ Data visualization

داده کاوی در اواخر دهه 1980 پدیدار گشته، در دهه 1990 گامهای بلندی در این شاخه از علم برداشته شد و انتظار می رود در این قرن به رشد و پیشرفت خود ادامه دهد.

کشف دانش از منابع داده فرایند شناسایی درست، ساده، مفید، و نهایتاً الگوها و مدلهای قابل فهم در داده ها می باشد. داده کاوی، مرحله ای از فرایند کشف دانش می باشد و شامل الگوریتمهای مخصوص داده کاوی است، بطوریکه، تحت محدودیتهای مؤثر محاسباتی قابل قبول، الگوها و یا مدلهای را در داده کشف می کند. به بیان ساده تر، داده کاوی به فرایند استخراج دانش ناشناخته، درست، و بالقوه مفید از داده اطلاق می شود. تعریف دیگر اینست، داده کاوی گونه ای از تکنیکها برای شناسایی اطلاعات و یا دانش تصمیم گیری از قطعات داده می باشد، به نحوی که با استخراج آنها، در حوزه های تصمیم گیری، پیش بینی، پیشگویی، و تخمین مورد استفاده قرار می گیرند. داده ها اغلب حجیم، اما بدون ارزش می باشند، داده به تنهایی قابل استفاده نیست، بلکه دانش نهفته در داده ها، بر اساس داده می باشد. به این دلیل اغلب به داده کاوی، تحلیل داده ای ثانویه^۷ گفته می شود.

داده کاوی از منابع توزیع شده متفاوت تر از داده های از منابع متمرکز است. مشکل داده کاوی از منابع توزیع شده (مانند وب) در محتویات غیر ساخت یافته و پراکنده است. منابع این یافته ها، منابع توزیع شده، برخلاف منابع متمرکز، دارای یک ساختار استاندارد مناسب نیستند و اسبک و شیوه نگارشی متنوع محتوایی نسبت به آنچه که در مجموعه منابع متمرکز وجود دارد، پیروی می کنند.

فصل اول این سمینار یک مرور سریع بر معرفی داده کاوی؛ مدلهای و روشهای موجود در آن پرداخته است، فصل دوم مدل ها و الگوریتم های داده کاوی به صورت اجمالی بررسی شده؛ در فصل سوم مدل وب کاوی مطرح شده است در این فصل تکنیکها و الگوریتمهای مختلف وب کاوی بررسی شده است، در این بخش الگوریتمهای داده کاوی از وب نیز معرفی شده است در فصل چهارم وب سرویس و معماری وب سرویس و XML و چگونگی داده کاوی از وب سرویسها مطرح شده است. در فصل چهارم یکسری ایده ها و سوالاتی در رابطه با وب سرویس کاوی مطرح شده است.

^۷ Secondary Data Analysis