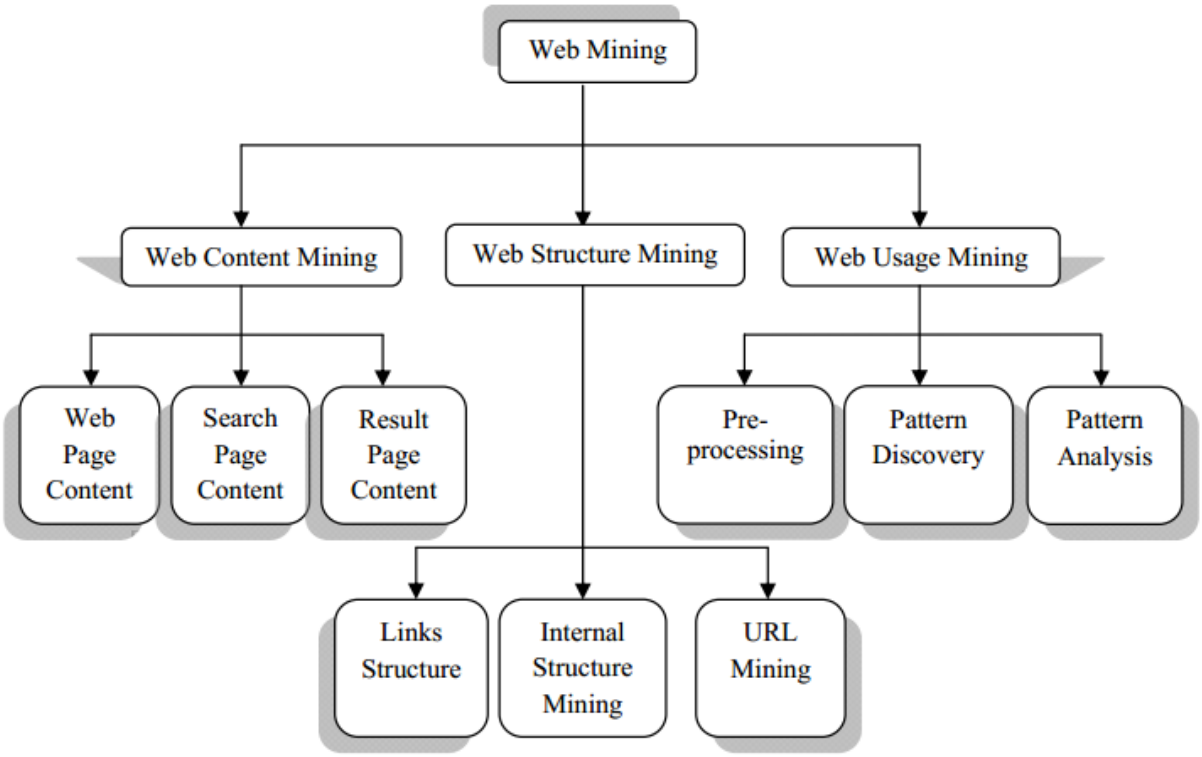
دانشگاه آزاد اسلامی مشهد

مبانی بازیابی اطلاعات و جستجوی وب

کاربرد ها و ابزار های Web Mining



استاد: مهرداد نجاتی

دانشجو: مهرداد محمدیان

Web Mining: استفاده از تکنیک های داده کاوی برای کشف و استفاده از الگو های دنیای وب.

۲ مورد از زمینه های مختلف کاربرد

۱- Robot Detection and Filtering: شناسایی و تفکیک رفتار انسان و ربات در وب.

استفاده غیر مجاز برای جمع آوری اطلاعات برای وب سایت هایی مانند حوزه e-commerce نگران کننده است. ربات ها پهنای باند قابل توجهی مصرف می‌کنند به علاوه ربات ها باعث اختلال در تحلیل جریان کلیک ها در داده های وب می‌شوند.

تکنیک های مختلفی از جمله تشخیص IP و User Agent را برای تشخیص ربات ها می‌توان استفاده کرد، اما این روش ها برای تشخیص ربات های مخفی شده کافی نیستند. برای این منظور میتون از رویکرد های **classification** برای تشخیص ربات بودن یا نبودن کاربران وب سایت، بر اساس داده های رفتاری آن ها استفاده کرد.

۲- User Clustering: جمع آوری اطلاعات و کاوش های کاربران در یک وب سایت و استفاده از این اطلاعات برای دسته بندی افراد با استفاده از الگوریتم های **بدون نظارت** ماشین لرنینگ.

کاربرد در جشنواره ها، اطلاع از تیپ های مختلف کاربران و تصمیم گیری برای استراتژی های بازاریابی، ارسال کد تخفیف های هدفمند و . . . .

۲- Web-Wide Tracking: ردیابی افراد در تمام وب سایت هایی که بازدید می‌کند.

این شیوه می‌تواند درک بسیار خوبی از نحوه زندگی و سلایق افراد را فراهم کند، که برای اهدافی مثل بازاریابی بسیار جذاب است. یکی از نمونه های موفق **DoubleClick** است (توسط گوگل خریداری شده)، تبلیغاتی را به کاربرد ارائه می‌دهد که بر اساس **ویژگی های شناختی و رفتاری** او می‌باشد. اگر هر وب سایت از سرویس DoubleClick استفاده کند، می‌تواند رفتار کاربران را با استفاده از کوکی ها رصد کند و در نتیجه تبلیغات هدفمند به کاربر نشان بدهد.

نمونه های واقعی از کاربرد ها

۱- Google Analytics: تحلیل رفتار کاربران وب سایت و اطلاعات مفید دیگر.

۲- Researchgate: در زمان ثبت نام در این وب سایت، مقالات ایندکس شده ای که نویسنده آن ها مشابه نام شما باشد را از سراسر وب برای اضافه کردن به پروفایل نشان می‌دهد.همچنین در زمان ثبت نام بعد از وارد کردن اسم، تصاویر افرادی که با اسم شما در اینترنت موجود است را برای انتخاب عکس پروفایل پیشنهاد می‌دهد.

۳- Hotjar: بررسی کلیک های کاربر در وب سایت و track کردن نحوه کار کردن کاربر با قسمت های مختلف وب سایت.

ابزار های مختلف برای استخراج و تحلیل داده های وب

Beautiful soup: ابزاری برای Crawl/Scrap کردن وب سایت ها با زبان پایتون. بر اساس تگ های html

Scrapy: ابزاری برای Scrap/Crawl کردن وب سایت ها با زبان پایتون. بر اساس تگ های html

scikit-learn: کتابخانه الگوریتم های کلاسیک یادگیری ماشین. برای اعمال تکنیک های با نظارت و بدون نظارت بر داده ها و در نهایت ساخت یک مدل برای پیش‌بینی، دسته بندی یا طبقه بندی.