Web usage mining

Click stream data je analiziranje svih podataka dok on koristi jednu aplikaciju. Koriste se logovi I koriscenje app (tipa ulogujes se koriscenjem gmail naloga I on dobije neke podatke o tebi).

CIlj je znati ko je taj korisnik, ispatrimo put kroz app I da posmatranjem rketanja, za svaki pomeraj ili user event povezemo ponasenje I podatke Kada to sve uradimo mozemo da radimo mining podataka.

Web server logs

Postoji vise formata za logovanje. Ne podrzavaju svi sve. Najcesce je W3C Extended (Microsoft ga podrzava). Sadrzi sirove podatke iz korisnika.

Preces web usage mining-a

Njacesce se poalzi od web logova (najbolje je koriscenje aplikacionih logova jer smanjuju problem oko sesije sto je bitno da za sablon ponasanja korisnika). Kada se dobije log je najteza stavr priprema logova. Treba da se ociste od svih viskova. Definisanje sesije je bitno (osnovni gradivni element). I domen je bitan. Kada se jednom def ponasanje dobiju se transakcije korisnika. Moci ce da se iskoristi klasifikacija, klasterizacija I da se izvrsi karakterizacija ponasanja I njihovo agregiranje. Priprema podataka je zahtevan process. Aplikacioni logovi su razliciti za svaku app, jako informativan, veliki posao da se oni ociste I da dobijemo nesto kako bismo identifikovali akcije korisnika. Cak I ocisceni podaci su sirovi I treba da se primene na nasu konkretu app I da se napravi transakciona baza koja prikazuje ponasanje nashi korisnika.

Pre-processing of web usage data

Data cleaning podrazumeva brisanje nerelevantnih polja, reference (sajt za prodaju preparate za rast kose I jedan page view je situacija kada se korisnik prebaci sa str sa 10 proizvoda predje na str sa jednim proizvodom, sve ucita I css, html, js I ostalo, nama to nije bitno mi hocemo sa cinjenciu da je kosirnik x dosao na str koja opisuje proizvod y). Dodatno, web spideri, pokusaj da se dodaju nedostajuce reference.

Strategije

Heuristika orijentisana ka vremenu (vreme kada je neko pristupio nasoj str). Npr procenjeno je da se ljudi zadrzavaju 30, 10 min, da se krece na 10s.

Heuristika orijentisana ka

Identifikacija korisnika – koriscenje ip adrese I naziva agenta sa koga se pristupio. PP da je taj par jedinstven I dostupno je u sirovim logovima. Moze taj par da ne bude jedinstevn I to se lako radi (koriscenje rotating ip servera, proxy servera).

Path completion

Nesto od alg data mininga I da probamo da dodjemo do nekih zakljucaka (klasifikacija, klasterizacija).

Primena web usage mining

Moze da nam obezbedi podatke sa e commerce sajta (web logovi).