Metodika evalvácie prípravy dát v procese objavovania vzorcov správania sa používateľov webu

Pri skúmaní vplyvu prípravy dát na kvantitu a kvalitu získaných znalostí postupujeme nasledovne:

1. Získanie dát z logovacieho súboru webového servera, resp. informačného systému.
2. Príprava dát na rôznych úrovniach.

   Vytvorenie dátových súborov predspracovaných na rôznej úrovni prípravy dát.

   Vo všeobecnosti príprava dát pozostáva z týchto úloh:

   a. Čistenie dát.

      Nevyhnutným krokom je očistenie dát od nepotrebných dát - požiadaviek na obrázky, skripty a štýly. Tento krok je východiskovým - základným bodom pri príprave dát z logovacieho súboru webového servera. Výsledkom je súbor hrubých dát (raw data) obsahujúci prístupy na portál. Čistenie dát z logovacieho súboru webového servera zahrňuje aj očistenie dát od prístupov robotov vyhľadávacích služieb, poprípade očistenie dát od prístupov z NAT/proxy zariadení.

      Čistenie dát z logovacieho súboru informačného systému pozostáva iba z odstránenia prístupov skupín používateľov, ktorých správanie sa nie je predmetom nášho skúmania. Podobne ako v predchádzajúcom prípade tento krok predstavuje východiskový bod pri príprave dát.

   b. Identifikácia používateľov/sedení.
   c. Rekonštrukcia aktivít používateľov webu.
3. Analýza dát.

   Hľadanie vzorcov správania sa používateľov webu v jednotlivých súboroch. Pri hľadaní vzorcov správania sa v skúmaných súboroch je nutné zabezpečiť, aby pravidlá z jednotlivých súborov boli extrahované za rovnakých podmienok.
4. Porozumenie výstupným dátam.

   Vytvorenie dátového súboru z výstupov analýz jednotlivých súborov a výpočet základných charakteristík skúmaných súborov:

   a. počet prístupov,
   b. počet zákazníckych/identifikovaných sekvencií,
   c. počet frekventovaných sekvencií,

d. priemerná veľkosť/dĺžka identifikovaných sekvencií.

Na základe týchto prvotných výsledkov je možné spresniť predpoklady.

5. Porovnanie získaných znalostí zo skúmaných súborov predspracovaných na rôznej úrovni prípravy dát.

   Pri hodnotení získaných znalostí sa zameriavame nielen na kvantitu extrahovaných pravidiel, ale aj na ich kvalitu. Kvalita sekvenčných pravidiel sa posudzuje dvoma ukazovateľmi:

   a. podpora (support),
   b. spoľahlivosť (confidence).

   Ďalej v hodnotení získaných znalostí zohľadňujeme aj ich použiteľnosť v praxi. Od sekvenčných pravidiel, podobne ako od asociačných, požadujeme aby boli nielen zrozumiteľné, ale aj užitočné. Vo všeobecnosti sekvenčná analýza produkuje tri typy pravidiel:

   a. užitočné (useful),
   b. triviálne (trivial),
   c. nevysvetliteľné (inexplicable).

   Užitočné pravidlá obsahujú informáciu vysokej kvality, triviálne pravidlá obsahujú výsledok, ktorý je všeobecne známy pre danú oblasť a nevysvetliteľné sa nedajú objasniť a nevedú k nejakému prospešnému činu. V tejto fáze je veľmi dôležitá spolupráca s expertom na dáta z danej aplikačnej oblasti.

   Získané znalosti hodnotíme z hľadiska kvantity a kvality nájdených sekvenčných pravidiel - vzorcov správania sa používateľov pri prehľadávaní webu v zmysle:

   a. porovnania podielu nájdených pravidiel v skúmaných súboroch,
   b. porovnania podielu užitočných, resp. triviálnych a nevysvetliteľných pravidiel v skúmaných súboroch,
   c. porovnania hodnôt miery podpory (support) a spoľahlivosti (confidence) nájdených pravidiel v skúmaných súboroch.