Česká zemědělská univerzita v Praze

Provozně ekonomická fakulta

Katedra informačních technologií

Bakalářská práce

Data Mining

Bohuslav Endt

Souhrn

Bakalářská práce se zabývá Data miningem. V teoretické části práce jsou na základě studia odborné literatury zrekapitulovány základní pojmy z oblasti data miningu a z databázových systémů. V práci jsou přiblíženy postupy užívané pro data mining a objasněny postupy a principy algoritmů využívaných při dolování dat.

Praktická část se zabývá přípravou dat pro dvě analytické metody data miningu a jejich následným zpracováním. V závěru jsou zesumarizovány výsledky celého procesu.

Klíčová slova: Datamining, Dolování dat, Získávání znalostí z databází, Informace, Databázové systémy

Cíl práce

Bakalářská práce má za cíl zrekapitulovat informace z oblasti data miningu z databázových systémů a na základě získaných znalostí provést analýzu dat a sestavit predikční model použitím zvolené metody a softwaru pro data mining. Výsledky analýzy dat a predikce ověřit na testovacích datech, zhodnotit výsledky predikce a popsat nejdůležitější části tohoto modelu.

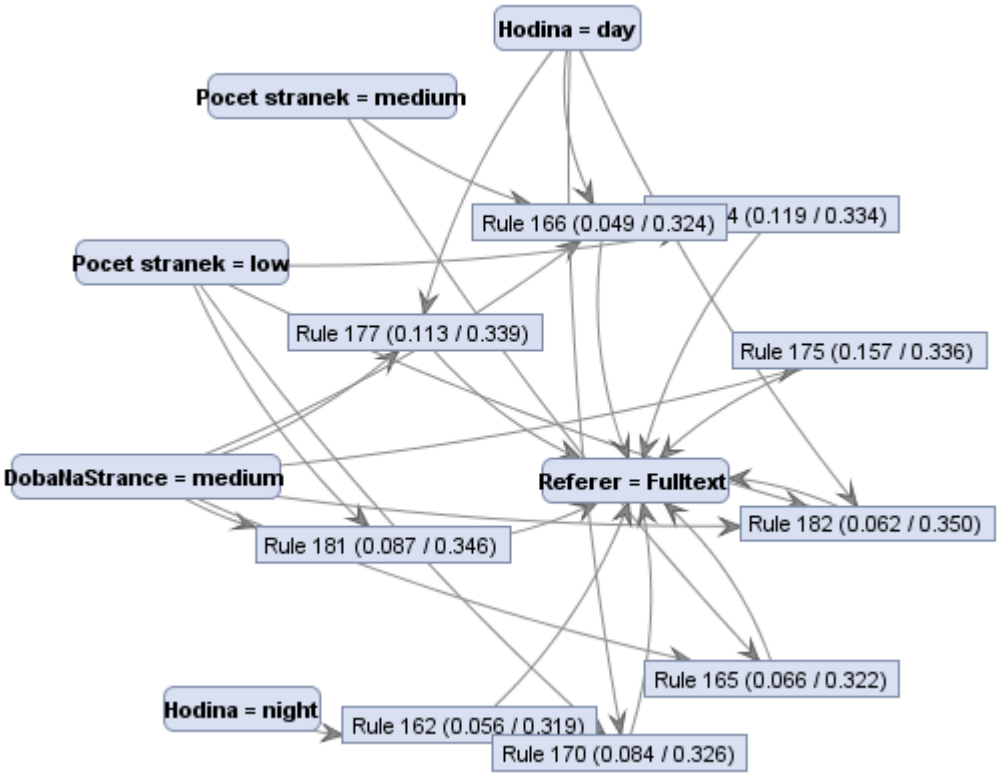
Metodika

Na začátku práce bude provedeno studium literárních zdrojů, následně bude zpracována rešerše těchto zdrojů. Postupy užívané pro DM objasní postup a princip algoritmů využívaných při dolování dat. Fáze úpravy dat se bude zabývat připravením dat pro zpracování zvoleným algoritmem, který bude, společně s vhodným softwarem pro zpracování, zvolen na základě nově získaných znalostí. Výsledky DM budou interpretovány běžným jazykem. Závěrem budou zesumarizovány nejdůležitější informace získané celým procesem DM.

Zhodnocení výsledků

Závěry analýz u shlukování i asociačních pravidel jsem popsal přímo u výsledků. Celkově lze říci, že jsem poměrně úspěšně rozdělil návštěvníky dle jejich zájmů (tedy dle témat stránek, které navštívili) do shluků.

U asociačních pravidel jsem identifikoval pravidla, která naznačují situace, kdy návštěvníci splní definované cíle, ale také jak se chovají, pokud přijdou z vyhledávačů. Tato pravidla dle mého do jisté míry věrně kopírují reálnou situaci.



Obrázek 1: Vizualizace Vztahů mezi asociačními pravidly

Závěr

Z dat jsem ověřil, že výsledky odpovídají očekávaným výstupům, byly zjištěny shluky témat, po kterých se jednotlivé skupiny uživatelů nejvíce pohybují a chování návštěvníků webových stránek, kteří přicházejí z vyhledávačů.

Pro přípravu dat bych do budoucna již nevolil MS Excel, jelikož úprava větších tabulek je v tomto nástroji poněkud těžkopádná. Některé operace s velkým množstvím dat trvají v desítkách vteřin, někdy bylo dokonce třeba data rozdělit a zpracovávat postupně. S pomocí RapidMineru se stejné úpravy prováděli mnohem lépe. Zpracování dat bylo nejnáročnější částí celého procesu. Časová náročnost by se dala zmenšit užitím datového skladu či trhu.

Seznam použitých zdrojů

BERKA, Petr. Dobývání znalostí z databází. 1. vyd. Praha: Academia, 2003. 366s + Computer Press, 2003. 486str + 1CD ROM. ISBN 80-7226-969-0

LACKO, Luboslav. Databáze: datové sklady, OLAP a dolování dat. 1. vyd. Brno: 1CD ROM. 2003. ISBN 80-200-1062-9

C. M. Bishop. Pattern Recognition and Machine Learning. Springer, 2006. ISBN 0-387-31073-8.

LACKO, Luboslav. Business Intelligence v SQL Serveru 2008. 1. Vyd. Brno: Computer Press, 2009. 456str. ISBN 978-80-251-2887-9 978-1-55860-901-3OO

BUDÍKOVÁ, Marie aj. Průvodce základními statistickými metodami. 1. vyd, Praha: Grada, 2010. ISBN 978-80-247-3243-5

COLIN, SHEARER aj,. CRISP-DM 1.0. *CRISP-DM 1.0* [online]. 2000, s. 78 [cit. 2015-02-16]. Dostupné z: http://www.crisp-dm.org/CRISPWP-0800.pdf

BERKA, Petr. Aplikace systémů dobývání znalostí pro analýzu medicínských dat. Aplikace systému KDD. [Online] 2001. [cit. 2015-02-16]. Dostupné z: http://euromise.vse.cz/kdd/index.php?page=proceskdd