UNIVERSITY OF TARTU
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE
Institute of Computer Science
Computer Science Curriculum

Aleksei Panarin

Logs Mining Based Approach to eCommerce Customer Classification

Master's Thesis (30 ECTS)

Supervisors: Rauno Viin (MSc)
Siim Karus (PhD)

Tartu 2015

Logs Mining Based Approach to eCommerce Customer Classification

Abstract

Fits.me Company has developed a web-tool which helps online shoppers to choose the right size of clothes. The application of Virtual Fitting Room logs users' actions and saves values of entered body measurements into database. Additionally, Google Analytics is used to get data of online shops' website visiting sessions, users' characteristics like location, software and hardware. The main goal of the thesis is to analyse the data, learn to extract useful information. More precisely, we want to develop a method of grouping web-shop customers.

At the first stage we find a way to combine data from different sources. We aggregate the data into user- and session-based profiles. The data is cleaned. It has more informative form, and is ready for further analysis. Data cleaning and pre-processing form a significant part of the thesis.

On the analysis stage we use two methods for the data classification. These are Decision trees and Naïve Bayes. We decide to group customers by one of the important features for eCommerce: we classify user whether he/she makes a purchase or not, whether a user returns purchased item or not. Both, classification tree and Naïve Bayes did not find significant relationship between studied attributes and shopping behaviour. However, regression tree turned to be useful for finding the groups of users with similar behaviour. It shows patterns of behaviour which leads to higher probability of making purchase.

Keywords

Log mining, data analysis, data mining, decision trees, Google Analytics, web usage mining, Naïve Bayes, R

e-Äri klientide klassifitseerimine rakenduse logide põhjal

Magistritöö (30 EAP)

Aleksei Panarin

Lühikokkuvõte

Fits.me ettevõte on arendanud veebipõhise rakenduse, mis aitab veebipoodide külastajatel valida õiget suurust riideid. Virtuaalse Proovikabiini rakendus logib kasutajate tegevusi ja salvestab sisestatud kehamõõdud andmebaasi. Lisaks kasutatakse Google Analytics andmeid, mis annab andmeid veebipoe külastuste sessioonidest ja sellistest kasutajate omadustest, nagu asukoht, kasutatud tarkvara ja riistvara. Käesoleva lõputöö põhiline

ülesanne on analüüsida andmed ja õppida eraldama logidest kasulikku informatsiooni. Täpsemalt, me tahame leida meetodi veebipoe kasutajate grupeerimiseks.

Esimesel etapil me leiame viisi erinevatest allikatest andmete kokkupanemiseks. Me agregeerime andmeid kasutajate- ja sessioonipõhisteks profiilideks. Andmed on puhastatud. Nende vorm on informatiivsem, ning andmed on valmis edaspidiseks analüüsiks. Andmete puhastamine ja eeltöötlus moodustavad lõputöös tähtsa osa.

Analüüsietapil me kasutame kahte andmete klassifitseerimismeetodit. Need on Otsustuspuud ja Naive Bayes. Me otsustame grupeerida kasutajaid e-kaubanduse jaoks ühe tähtsa tunnuse järgi: me klassifitseerime kasutajaid selle järgi, kas nad on teinud ostu või mitte, kas nad on tagastanud ostetud toodet või mitte. Klassifitseerimispuu ega Naive Bayes ei tuvastanud olulisi seoseid uuritud atribuutide ja ostukäitumise vahel. Kuid regressioonipuu osutus kasulikuks sarnase käitumisega kasutajate gruppide leidmises. See näitab, millise käitumismustri korral on ostu tegemise tõenäosus suurem ning millise käitumise korral väiksem.

Võtmesõnad

Logidest kaevamine, andmeanalüüs, andmekaeve, otsustuspuud, Google Analytics, veebikasutuse kaevamine, Naive Bayes, R

II. License

Non-exclusive licence to reproduce thesis

I, **Aleksei Panarin** (date of birth: 06.09.1987),

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

reproduce, for the purpose of preservation, including for the purpose of preservation in the DSpace digital archives until expiry of the term of validity of the copyright

Logs Mining Based Approach to eCommerce Customer Classification,

supervised by Rauno Viin (MSc), Siim Karus (PhD),

2. Making the thesis available to the public is not allowed.

3. I am aware of the fact that the author retains the right refered to in point 1.

4. This is to certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu, 05.08.2015