WEBES ADATBÁNYÁSZAT

SZOMMER KÁROLY

Összefoglalás

Az internetről összegyűjtött többletinformáció mindig előnyhöz juttathat bennünket egy-egy területen. Munkámban bemutatom, hogy viszonylag egyszerű ilyen információhoz hozzájutni, és ez a folyamat automatizálható is. A Vatera aukciós portálról egy crawler segítségével összegyűjtöttem a mobiltelefonos aukciós oldalon található információkat. Az adatgyűjtés után manuálisan meghatároztam 3500 hirdetés csoportját. Ezt követően adatbányászati algoritmusok segítségével döntési táblát generáltam, amely 95%-os pontossággal sorolta be megfelelően a mobiltelefonokat a megfelelő csoportba. A döntési táblákat az eredmények átvizsgálása után manuálisan tovább pontosítottam, így 99%-os pontosságot sikerült elérnem. Ezt követően a crawler által elmentett adatokat elemeztem. Az elemzés során törekedtem arra, hogy csak olyan eszközöket használjak, amelyek mindenki számára elérhetőek. A megszerzett információkat legális eszközök segítségével, a ráfordított munkaórák bérén kívül mindenféle többletköltség nélkül értem el. Az aukciók elemzése kezd egyre fontosabbá válni, különösen, hogy olyan aukciós házak is megjelentek már, amelyek mind virtuális, mind a valós világbeli fizetőeszközökkel párhuzamosan használhatóak. Ezzel az elemzéssel egy közösség preferenciáit tudhatjuk meg, így gyorsabban, pontosabban reagálhatunk a piacon, mert ismerjük az adott közösség igényeit, reakcióit. De mi lesz akkor, ha már nem csak a közösség, hanem az egyén preferenciáit is pontosan ismerni fogjuk?

Kulcsszavak: *aukció, crawler, adatbányászat, mobiltelefon, elemzés*

Web data mining

Abstract

The information collected from the internet may give us an advantage in certain cases. In my work, I'm going to show that getting such information is relatively easy and this process can be automated. I collected the information about mobile phone auctions with a web crawler from auction site Vatera. I manually assigned the groups of 3500 ads after the data collection, then generated a decision table using data mining algorithms that classified the phones with 95% accuracy. I classified the decision tables further after the examination of the results, and achieved 99% accuracy. Then, I analyzed the data gathered by the crawler. In the analysis I used tools that are available to everyone. I acquired these information with only legal tools without additional costs except the human resource cost. The analysis of the auctions is getting more and more important, particularly because there appeared auction houses where one can pay both by virtual and real money. We get to know the preferences of a community by this analysis, so we can react faster and more precisely to the changes of the market, because we know the need and reactions of the community. But what will happen if we will know not even just the preferences of a society, but the individuals too?

Keywords: *auction, crawler, data mining, mobile phone, analysis*

Bevezetés

Az internetre lépéssel egy időben sokszor mit sem sejtve a digitális lábnyomainkat is a weboldalak üzemeltetőinél hagyjuk. Az interneten található információkat felhasználva csak legális eszközöket használva már többlettudáshoz juthatunk, melyek valamilyen előnyhöz juttathatnak bennünket. Az internetről viszonylag egyszerű olyan információt kinyerni, hogy az számunkra valamilyen tevékenységünket pozitívan befolyásolja. Azt fogom bemutatni, hogy ez az extrakció automatizálható, amit munkámban a Vatera aukciós portál elemzésén keresztül fogok megtenni. (Vatera, 2012) A portálon mi határozzuk meg az egyes termékek árát mind vevői, mind eladói oldalról. Aki ezt az adattömeget feldolgozza, versenyelőnyhöz jut az adott piaci szegmensben: tudni fogja, hogy mely terméket milyen áron érdemes értékesíteni, hogy mekkora a kereslet, milyenek az árváltozási tendenciák, stb. A hivatalos eladók számára is fontos ez az információ, ugyanis manapság már ők is megjelennek az aukciós oldalakon. Munkámban a Vatera mobiltelefon aukcióit elemzem, azon belül is csak bizonyos kategóriájú telefonokat, és csak bizonyos típusokat fogok szerepeltetni.

Anyag és módszer

Az aukciós oldalról történő információkinyerés során először ki kell nyerni a megfelelő adatokat és el kell azokat tárolni. Az eltárolásra egy MySQL szervert választottam. (MySQL, 2012) A második lépésben az adatok összegyűjtése történt. Erre a feladatra egy crawlert készítettem, PHP nyelven. (Cheng-Hsien, Shi-Jen, 2008). A crawler a Vatera mobiltelefonos oldalait nézte végig, abból reguláris kifejezések segítségével (regular-expressions.info, 2012) összegyűjtötte a termékek részletes adatait, majd azokat eltárolta egy adatbázisban. Ezt követően a crawler futását automatizáltam, amely óránként lefutva átlagosan 9000 rekordot generált. Az automatizáláshoz a linux crontab és php parancsát használtam fel. (Schwarz, 2000)

A második részben az adatok manuális feldolgozása következett. Első lépésben exportáltam az első óra adataiból véletlenszerűen 3500 rekordot, majd manuálisan besoroltam őket a megfelelő kategóriába. Ezután megvizsgáltam, hogy mely kategória fordul elő legalább 20 alkalommal, ezek a következők: Dual SIM Kínai, Egyéb, iPhone 3, iPhone 4, Nokia 5230, Nokia 6500, Nokia c5-03, Nokia N9, Nokia N95, Samsung s5230, Samsung s5620, Sony Ericsson X8, Sony Ericsson Xperia. A kategóriák közül a Dual SIM Kínai-t elemezve később átsoroltam az Egyéb kategóriába a termékek tulajdonságainak nagymértékű inhomogenitása miatt.

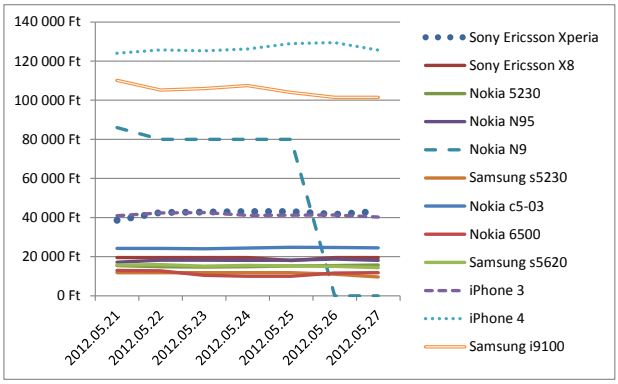
A harmadik részben ki kellett választani egy mindenki számára elérhető, teljesen ingyenes adatbányászati eszközt, amely segítségével a besorolási szabályokat elkészíthettem, továbbá meg kellett határozni, hogy milyen módszerrel készítem el a besorolási szabályokat. A The University of Waikato által fejlesztett WEKA adatbányászati eszköz tökéletesen megfelelt a számomra. (Hall et al., 2009). A szövegek adott kategóriába történő besorolására sokféle módszer áll a rendelkezésünkre, (Sasaki, 2008; Manne, 2011). Azonban minden esetben célszerű a szövegekből szóvektorokat generálni. (Saad & Ashour, 2010). Ezt a WEKA-ban egy felügyelet nélküli szűrő segítségével tehettem meg. A szűrő futtatása előtt beállítottam, hogy a szöveget alakítsa át csupa kisbetűsre, továbbá, hogy csak azokat a szavakat tartsa meg, amelyek legalább 10-szer előfordulnak.

Fontos volt, hogy a besorolási szabály könnyen leprogramozható legyen, így kipróbáltam a döntési fa készítő, (Saad - Ashour, 2010).valamint a döntési táblázat

készítő algoritmusokat. (Kohavi, 1995). A döntési tábla készítésével jobb eredményt tudtam elérni, így ezzel a módszerrel foglalkoztam tovább. Az attribútum keresési módjának több változatát is ki lehet próbálni a WEKA-ban. Több lehetőség közül a legjobb eredményt a scatter search-el kaptam, (Lopez, 2004), ami egy populáció alapú módszer: 95,155%-os pontossággal sorolta be az egyes mobilokat a megfelelő kategóriába, 19 darab szabály segítségével. Megvizsgálva a pontatlanul besorolt rekordokat, néhány további szabály hozzáadásával 99%-os pontosságot értem el.
A feldolgozó modul jelenleg csak a legegyszerűbb statisztikai információkat szolgáltatja az eltárolt adatokból: darabszám, minimum licit, maximum licit, átlag licit, licit szórása, villámár darabszáma, minimum villámár, maximum villámár, átlag villámár, villámár szórása. A licit az, amikor a vásárló miután licitált a termékre a verseny még tovább folytatódik, mások is tovább licitálhatnak rá. A villámár esetében leütéskor már nem folytatódik a verseny, a vevő a terméket megszerzi az adott áron. (vatera, 2012)

Eredmények

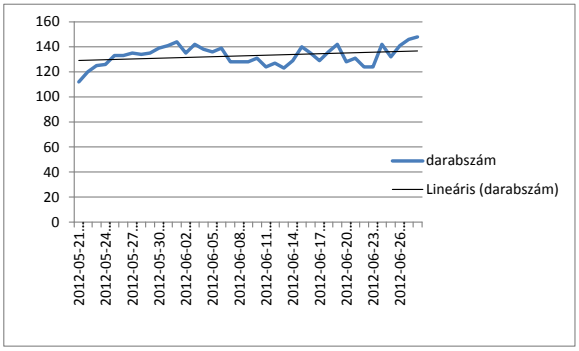
Az elkészített crawlert a Vatera oldalán bármikor le lehet futtatni, akár manuálisan, akár automatizálva, így mindig naprakész információkat tudunk kinyerni a mobiltelefonok aukciós oldaláról. A feldolgozó modul szintén bármikor a rendelkezésre áll. A két modul segítségével tetszőleges percben rendelkezhetünk a legfrissebb információval a piacot illetően.
Az adatgyűjtés 2012. május 20-án, 20:00-kor kezdődött és 2012. június 28-án, 9:03-kor ért véget. Összesen 8 654 698 rekord került felvitelre az adatbázisba, amely 1,6 GB-nyi adatot jelent.
Az adatok feldolgozása után az 1. ábrán jól látszik bizonyos kategóriák közt a különbég. Láthatjuk a felsőkategóriás telefonokat, a közép- és alsókategóriásakat is.



1. ábra: Átlagos villámárak változása és a kategóriák elkülönülése az első hét napon
Forrás: Saját szerkesztés

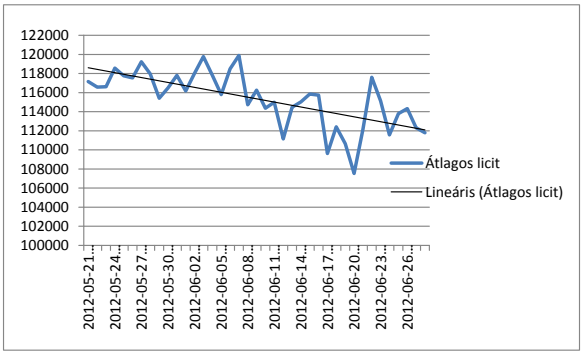
A következőben kiemelek egy típust: az iPhone 4-est, melynek a részletes adatai közül néhányat bemutatok. Azért ezt a típust emelem ki, mert egy szerencsés véletlen folytán az egyik ábrán keresztül lehetőségem van bemutatni a virtuális világok aukciós házaiban oly jellemző árbefolyásolást is.
Fontosnak érzem, hogy kitérjek a virtuális világok aukciós házaira is, ugyanis az interneten már nem az első olyan aukciós ház jön létre, amely mind virtuális, mind valódi fizetőeszközzel párhuzamosan is működik. A legújabb ilyen fejlesztés a Blizzard

egy nemrég kiadott játékához, a Diablo III-hoz készült. (Diablo III. Auction House, 2012)
A virtuális világokban az aukciós házakkal való interakciót legtöbbször egy-egy külön erre a célra készített szoftver segíti. (Auctioneer, 2012)
A játékosok közül az ügyesebbek az árakat is manipulálni tudják, mert a legtöbbjük ezeket a szoftvereket használja a tárgyak árazásához. Erre még a 4. ábránál visszatérek.



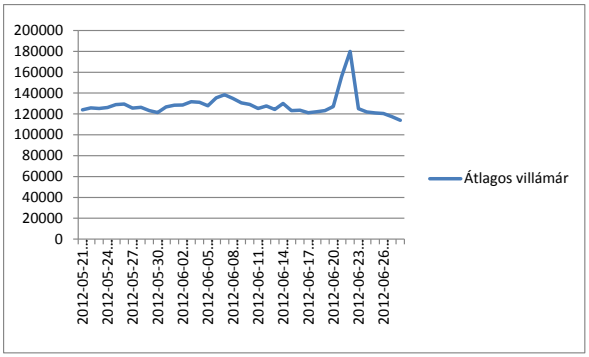
2. ábra: Az iPhone 4 számának változása a vizsgált időszakban
Forrás: Saját szerkesztés

Ahogy azt a 2. ábrán láthatjuk, az iPhone 4 hirdetések száma a vizsgált időszakban megnőtt, azonban ezzel együtt a 3. ábráról jól leolvasható, hogy az átlagos licitáruk csökkent. Az igazán érdekes adatokat majd az iPhone 5 megjelenése és az utáni időszakban fogjuk kapni: vajon hogy alakul majd az eladási mennyiség és az ár?



3. ábra: Az iPhone 4 átlagos licitára a vizsgált időszakban
Forrás: Saját szerkesztés

Az átlagos licitáron még nem látszik, hogy június 21-én és 22-én egy extrém magas áru készülék (500 000 Ft kezdő licit, 1 100 000 Ft villámár) is szerepelt az aukciós portálon, azonban a 4. ábra már árulkodik arról, hogy nem volt minden ár rendben.



4. ábra: Az iPhone 4 átlagos villámára és az extrémum hatása
Forrás: Saját szerkesztés

Néhány kilógó érték a szoftverben még korrigálható, azonban a virtuális világokban vannak, akik a szoftverek átveréséből mesterséget űznek.

Következtetések, javaslatok

Érdemes lesz egy új termék megjelenésekor az árakat folyamatosan monitorozni: a megjelenés milyen módon befolyásolja majd azokat, és milyen piaci átrendeződés jön létre?
Érdemes lenne elvégezni a vizsgált kategóriák kibővítését, valamint az egyes készülékek kategórián belüli szeparálását is, természetesen megtartva az aggregált információkat is (kibonthatóság).
A legújabb, Blizzard által elkészített aukciós házon végzett vizsgálat során pedig érdemes lenne megvizsgálni: a virtuális világok termékeinek árbefolyásolása vajon elmozdítja-e a virtuális fizetőeszköz és a valódi pénz közötti konverziós rátát?
Legvégül pedig: mi lenne, ha nem csak az aukcióhoz tartozó adatokat szednénk össze az internetről az eladókról, vevőkről, és azok alapján tovább súlyoznánk őket megbízhatóság szempontjából?

Köszönetnyilvánítás

A kutatás a TÁMOP 4.2.2/B-10/1-2010-0023 projekt keretében készült. Szeretnék köszönetet mondani a lehetőségért és a pénzügyi támogatásért.

Hivatkozott források

Vatera (2012): (http://www.vatera.hu/segitseg/)
Auctioneer (2012): (http://auctioneeraddon.com/)
Cheng-Hsien, Y. - Shi-Jen, L. (2008.). Paralell Crawling and Capturing for On-Line Auction. Lecture Notes In Computer Science, 5075. kötet., 455-466. o.
Diablo III. Auction House (2012): (http://us.battle.net/d3/en/game/guide/items/auction-house)
Hall, M. - Frank, E. - Holmes, G. - Pfahringer, B. - Reutemann, P. - Witten, I. H. (2009): The WEKA Data Mining Software: An Update. SIGKDD Explorations, 11. (1)

Héder, M. - Farkas, T. - Oláh, T. - Illés, S. (2011): Mashing Up Natural Language Processing, Recommender Systems and Search Engines to Support Wiki Article Editing. ESWC 11 AI Mashup Contest. Heraklion.

Kohavi, R. (1995): The Power of Decision Tables. Proceedings of the 8th European Conference on Machine Learning, 174-189. o.

López, F. G. - Torres, M. G. - Batista, B. M. - Pérez, J. A. - Moreno-Vega, J. M. (2004.): Solving feature subset selection problem by a parallel scatter search. European Journal of Operational Research, 169. 477-489. o.

Manne, S. - Fatima, S. S. (2011): A Novel Approach for Text Categorization of Unorganized data based with Information Extraction. International Journal on Computer Science and Engineering, 3. (79, 2846-2854. o.

MySQL (2012): (http://www.mysql.com/)

Regular-Expressions.info, (2012): (http://www.regular-expressions.info/)

Saad, M. K., Ashour, W. (2010): Arabic Text Classification Using Decision Trees. Workshop on computer science and information technologies CSIT'2010, 75-79. o.

Sasaki, Y. (2008): Automatic text classification. – előadás

Schwarz, M. A. (2000):. Linux Job Scheduling. Linux journal, 2000. kötet, 77. sz.

Szerző:

Szommer Károly
Ph D hallgató
Budapesti Corvinus Egyetem
Számítástudományi Tanszék
ifj.szommer.karoly@gmail.com