A felhasználói viselkedés vizsgálata (5. rész)

Cikksorozatom utolsó részében a bemutatott lehetőségek fényében néhány mérést készítek el a saját fejlesztésű programom segítségével. Végül a mérési eredményekből levonható következtetések és a gyakorlati jelentőség elemzésével lezárom a cikksorozatomat.

Eseménynapló források szerkezete, előzetes ismeretek

A mérések elvégzéséhez szükséges eseménynapló fájlok beszerzésénél megpróbáltam a következő szempontokat figyelembe venni:

- Legyen egy olyan eseménynapló, amely tartalmaz munkamenet azonosítókat is. Így lehetőség nyílik az összehasonlításra is: mennyit tévedünk amikor idő alapon különítjük el a látogatóinkat.
- Legyen olyan szerkezetű eseménynapló, amely keletkezéséről (a naplózott portálról) a lehető legtöbb információval rendelkezem. Így bizonyos kiugró adatokra könnyebben lehet magyarázatot találni.
- Szükséges egy nagyméretű, ismeretlen szerverről származó eseménynapló az általános, független következtetések felállításához.

A felhasznált három különböző eseménynapló eltérő portáloktól származik. Mind látogatottságban, mind tartalomban, mind a portálok szerkezetében jelentősek az eltérések. A legkisebb méretű, mindössze 15 MB-os eseménynapló egy apróhirdetéseket kezelő weboldal szerveréről származik. A látogatók legjellemzőbb tevékenysége a keresés az apróhirdetési adatbázisban. A portál összesen 19 jól elkülönülő weboldalból áll (keresés, eredmény lista, felvitel, impresszum, stb.), minden egyes weblap funkciója jól meghatározható. Ez az egyetlen eseménynapló, amely munkamenet azonosító adatokkal rendelkezik. Az eseménynapló összesen 77.562 soros, időben 5 és fél hónap látogatásait tartalmazza. A második szempontnak leginkább megfelelő portál eseménynaplója már lényegesen összetettebb, tekintve, hogy egy eseménynaplóban több különböző portál verzió bejegyzései is benne vannak. Ez egy online folyóirat, rengeteg (több mint 600) letölthető dokumentummal. A legjellemzőbb használat a dokumentumokban történő online keresés (adatbázis segítségével). Ez a portál is jól elkülönülő fájlokból épül fel, de a több, jelentősen eltérő verzió megjelenése egy naplófájlban megnehezíti az analízist. Megfelelő szűrési paraméterekkel a régebbi verzió bejegyzései könnyen kihagyhatóak. Az eseménynapló 93 MB, összesen 430 ezer bejegyzést tartalmaz alig 2 hónapos időtartam alatt.
A legnagyobb és legösszetettebb eseménynapló egy ismeretlen webszerverről származik. Bármiféle vizsgálat előtt az eseménynapló „kézi" átvizsgálásával a következők állapíthatók meg:

- tartalmazza egy galéria lekéréseit rengeteg képpel,
- tartalmazza egy olyan lekéréseit, amelyre jellemző a GET paraméterbe kódolt weboldal megkülönböztetés,
- továbbá tartalmaz sok egyéb letölthető dokumentációt (több szoftver online dokumentációjának tükrözése), amelyre jellemző a sok különálló fájl.

Az eseménynapló egy heti forgalmat tartalmaz, 721 ezer sora 149 MB méretű.
Az analíziseket egységesen egy ***Intel Pentium4 2,60 GHz*** processzorral, 512 MB memóriával felszerelt számítógépen hajtottam végre. Az analízis folyamatok futtatása során törekedtem az egységes szoftverkörnyezet biztosítására.

Különböző paraméterekkel végzett vizsgálatok

Első lépésben minden esetben az összes különböző lekérést kell az eseménynaplókból kiolvasni. Ez azért szükséges, hogy előzetes ismeretek nélkül is tudjuk alapszinten paraméterezni az előszűrési feladatokat. Mivel az eseménynaplókról és a hozzájuk tartozó portálokról nem minden esetben rendelkeztem elégséges információkkal, ezért minden modellkészítés során az 1% alatti elemeket elhagytam. Ez azt jelenti, hogy a ritkán előforduló sorokat nem vettem figyelembe.
Az apróhirdetési portálról származó naplófájl az egyetlen, amely munkamenet azonosítókat is tartalmaz. Először a különböző kéréseket kértem le, amely eredményből jól látszik, hogy ez a portál egyszerű szerkezetű, kevés weboldalból áll, és az egyes aloldalak funkciója is jól elkülönül. Jól látható az is, hogy a portál összes lekérésének (kb. 23 ezer) közel 89%-a 4 aloldal között oszlik el.
A munkamenet azonosítók alapján elvégzett analízis szerint összesen 9672 látogató járt a portálon, vagyis egy átlag látogató 2.1 weboldalt töltött le, ami összhangban van azzal, hogy az összes lekérés 75%-a két aloldalra irányult.
Az *1. táblázatban* látható adatok jelentése:

- Felhasználói útvonalak: a különböző időbeállításokkal vagy munkamenet azonosító felhasználásával lefuttatott analízis során elkülönítésre került látogatók száma.

1. táblázat *Apróhirdetési portál statisztika*

	munkamenet	10 perc	30 perc	60 perc	Átlag (382 sec)	Átlag 2x (764 sec)
Felhasználói útvonalak	9672	9261	8981	8857	9393	9189
Különböző útvonalak	23	18	16	14	18	18
Azonos oldalak egy látogatáson belül	17	12	10	8	12	12
Belépő oldalak	15	15	15	15	15	15
Kilépő oldalak	15	14	14	14	14	14
Átkattintási idők száma	3	3	3	3	3	3
Átkattintási idők átlag (max) sec	382	75	172	292	51	94
Átkattintási idők maximum (max) sec	40743	601	1756	3595	382	760

2. táblázat *Folyóirat portál statisztika*

	10 perc (1%)	10 perc (0,5%)	30 perc	60 perc	Átlag 2x (410 sec)
Felhasználói útvonalak	6785	6785	5857	5515	7265
Különböző útvonalak	4	13	10	9	11
Azonos oldalak egy látogatáson belül	3	9	6	6	7
Belépő oldalak	15	21	21	24	15
Kilépő oldalak	9	16	16	17	14
Átkattintási idők száma	9	15	16	17	3
Átkattintási idők átlag (max)	159	159	191	205	51
Átkattintási idők maximum (max)	600	600	1794	3600	382

- Különböző útvonalak: az összes látogató útvonalában ennyi olyan különböző útvonal van, amelyek gyakorisága a küszöbérték (1%) felett van.
- Azonos oldalak egy látogatáson belül: a különböző útvonalak között előforduló olyan weboldal csoportok száma, amelyek az útvonalon belül ugyanazokat a weboldalakat tartalmazzák, itt a sorrendjük nem számít.
- Belépő és kilépő oldalak: azt mutatja meg, hogy hány olyan weboldal van, amelyre illetve amelyről a küszöbérték feletti számban léptek tovább / érkeztek a látogatók.
- Átkattintási idők száma: azoknak a weboldalaknak a száma, amelyekről a küszöbérték feletti számban történt átkattintás egy másik, portálon belüli oldalra.
- Átkattintási idők átlaga: az összes átkattintási lépésnél átlagolásra kerül a két lekérés között eltelt idő, ez az érték másodpercben mutatja ezen átlagidők maximumát.
- Átkattintási idők maximuma: a legnagyobb olyan idő, amely két lekérés között eltelt.

A fenti táblázatban látható, hogy a különféle idő paraméterek beállítása ellenére sem térnek el több mint 10%-kal a felhasználói útvonal adatok. Ez azt jelenti, hogy a látogatók 90%-a kevesebb, mint 10 percenként (legkisebb időparaméter) lép egyet a portálon belül.
Jól látható, hogy a be- és kilépő oldalak és az átkattintási idők számai nem változnak. Ennek ez azért van, mert a látogatások nagyobbrészt egy jól meghatározott oldalcsoporton belülre irányulnak. Ezek a változatlanságok azt is jelentik, hogy az összes lekérés között ennyi jellemzően használt van. Kirívó adat a munkamenet azonosítóval történő felhasználó elkülönítésnél kapott, az átkattintási idők maximum értékének legnagyobb értéke: 40743 másodperc, közel 11 és fél óra. Ez egy kirívó eset, keletkezése úgy lehetséges, hogy a felhasználó böngészője a reggeli megnyitáskor kapott munkamenet azonosítót az esti visszatérésig nem hatástalanította, és azt újra elküldte a szervernek. Véleményem szerint ez egy böngésző hiba.
Egyenként összehasonlítva a különböző útvonalak és az azokon belüli azonos oldalak analízis eredményeit megállapítható, hogy az eredmény méretében az eltérés a felhasználói útvonalak száma miatt változik, a küszöbérték szerinti elhagyás előbb vagy később dobja el az eredmények egyes részeit.
Az on-line folyóirat portál eseménynaplójának analízise nehezebbnek bizonyult. A bevezetőben említett 1%-os küszöbérték esetén az egyes modellek csak néhány adatot tartottak meg, ezért itt lejjebb kellett venni a vágási értéket 0,5%-ra, vagyis a portál egy jól meghatározható részén van csak érdemi látogatottság. Az analízis eredmények statisztikája olvasható a 2. *táblázatban*.
Az apróhirdetési portálhoz képest itt az elkülönített látogatók száma szélesebb skálán mozog (15% illetve 23%), ami azt jelzi, hogy a látogatók nagyobb hányada maradt hosszabb ideig a portálon.

A különböző útvonalak és az azokon belüli azonos oldalak analízis eredményeit megvizsgálva elmondható, hogy a látogatók jellemzően néhány weboldal köré csoportosulnak: főoldal, keresés, tartalomjegyzék, cikk olvasás. Ez a megállapítás egybevág az összes különböző lekérésben található adatokkal, mert az összes lekérés 75%-a erre a néhány oldalra irányul.

Az ismeretlen webszerver eseménynaplójának vizsgálata az említett több portálos szerkezete miatt nehézkes. Ugyanakkor a fentiekhez hasonló eredményeken kívül megmutatkozott az analízis egy újfajta felhasználási lehetősége: a be- és kilépő oldalak statisztikájának vizsgálatával egyértelműen megállapítható volt, hogy a naplófájl több jól elkülönülő weboldal (portál) lekéréseit tartalmazza, azaz a be- és kilépő oldalak partíciókra estek szét. Jellemzően az egyes partíciókba tartozó oldalak szinte kizárólag az azonos partícióban lévőkre hivatkoztak. Az egyes partíciókból a portálok szerkezetének ismeretében könnyen felállíthatóak azok az előszűrési feltételek, amelyek mentén szétbontva az eseménynaplót, a keletkező résznaplók, mint az egyes portálok önálló eseménynaplói könnyebben és pontosabban analizálhatóak.

Összehasonlítás

A fenti eredményeket összehasonlítva más, hasonló képességű programmal (***Advanced Log Analyzer***, ***http://www.abacre.com/ala/***) az eredményekről elmondható, hogy a szoftverek által előállított modellek hasonlítanak egymásra. Például mindkettőben ugyanazok a weboldalak jelennek meg az eredményekben az első helyeken. Ugyanakkor egyes számszerű értékekben eltérések vannak.

A lekérések számában az eltérés mértéke minimális, a különbség az összevonások és egyszerűsítések, illetve a könyvtár lekérések kezeléséből adódik. A be- és kilépő oldalak statisztikáinál az eltérés valószínűsíthető oka a könyvtár lekérések kezelése, ugyanis az ***Advanced Log Analyzer*** teljesen külön kezeli a könyvtár és weboldal lekéréseket, míg a saját fejlesztésű szoftver képes a megfelelő bejegyzéseket jelentésileg azonosnak tekinteni (tipikus példa: ***http://www.pelda.com/*** azonos a ***http://www.pelda.com/index.php*** lekéréssel).

A felhasználók számában és útvonalaikban az eltérés magyarázata nem triviális, ugyanis az ***Advanced Log Analyzer*** eredményei önmagukban is ellentmondóak: míg a felhasználói profiloknál 5238 látogatót említ, addig a felhasználók által eltöltött idő statisztikáknál már csak 3071 látogatásról van szó (az adatok az első, apróhirdetésekkel foglalkozó weboldalra vonatkoznak).

A munkamenet azonosítóval készült eredményeket etalonnak tekintve a felhasználók számában az ***Advanced Log Analyzer*** eltérése közel 50%-os. Véleményem szerint a különbség egy jelentős részére magyarázat lehet a könyvtár és oldal lekérések összevonásának hiánya. További eltérést okozhat a paraméterezhetőség különbözősége és az ***Advanced Log Analyzer*** idő paraméterének ismeretlensége.

Gyakorlati jelentőség, felhasználhatóság

Az eredmények ismeretében mindössze az eseménynaplók felhasználásával képet lehet alkotni az adott portált felkereső látogatók mozgásáról, így akár régóta üzemelő weboldalak is bevonhatóak az analízisbe.

A felhasználói útvonalak önmagukban már csak méretüknél fogva sem alkalmasak következtetésekre. Ezt az eredményt minden esetben még fel kell dolgozni, hogy az általánosságok is a felszínre kerüljenek.

A különböző útvonalak statisztika elárulja, hogy melyek a legjellemzőbb bejárási útvonalak az adott portálon.

Az adatok segítségével ezeket a gyakori látogatású útvonalakat megfelelően lehet optimalizálni mind marketing szempontokból (például drágábban árult reklámfelület, célzott reklám), mind a szerver terhelésének optimális elosztására (például a frekventált weblapok gyorsítótárazása).

Az adatok ismeretében lehet átalakítani a portál belső szerkezetét is (a látogató előbb jusson hozzá a kívánt információhoz). Hasonló eredményeket és lehetőségeket ad az azonos oldalak egy látogatáson belüli eloszlása is.

Az átkattintási idők vizsgálatával pontosabb kép nyerhető nemcsak a látogatók útvonalairól, hanem az egyes weboldalakon eltöltött idők megoszlásáról is. Egy lehetséges felhasználás: azokat a weboldalakat, ahonnan jellemzően nagyon rövid idő alatt elkattintottak a látogatók, valamilyen módon át kell alakítani, vagy a tartalom módosításával, vagy akár magának a portál szerkezetének módosításával, hiszen az oldal a látogatók számára nem érdekes.

A belépő oldalak statisztika leginkább arra jó, hogy a látogatók portálba lépési pontjait megtaláljuk. A leggyakoribb belépési pontokon megfelelő szolgáltatást kínálva a látogatót sokkal nagyobb valószínűséggel lehet a portálon megtartani, így növelve a látogatottságot és az azzal járó előnyöket (például reklámbevételek, eladások növelése).

A kilépő oldalak statisztika segítségével találhatóak meg azok a weboldalak, amelyek után a felhasználó elnavigál a portálról. Ezen weboldalak tartalmának ismeretében a portál működtetői eldönthetik, hogy normális távozásról van-e szó (például egy címgyűjtemény esetén), vagy a nem megfelelő tartalom miatt hagyják el gyakran az adott oldalon keresztül a portált a látogatók.

A különböző kérések eredménye egyértelműen megmutatja, hogy melyek a portál leggyakrabban látogatott oldalai. Általánosságban elmondható, hogy a különböző analízisek során előálló adatok hatékony felhasználásához szükséges a vizsgált portál beható ismerete úgy tartalmilag, mint szerkezetileg. Fontos továbbá a látogatókról más forrásból is információkat szerezni (például kérdőíves megkérdezés), hogy a többféle eredmény összevetésével az optimális módosításokat lehessen megvalósítani.

Úgy vélem, hogy egy portál hosszú távú és eredményes üzemeltetéséhez elengedhetetlenül szükséges a folyamatos karbantartás és fejlesztés. Ezen folyamat egyik segítője lehet az általam bemutatott vizsgálati módszer: segítségével naprakészen tudjuk követni a látogatóink szokásainak változását.

Beszédes Balázs (beszedes@ei.hu)
24 éves, az e-Média Informatikánál mérnök-informatikus. Hobbija a kerékpározás és a kirándulás.