Internetes tartalmak minősítése
a felhasználók modellezésével

SCHLOTTER ILDIKÓ, GÁSPÁR CSABA
Budapesti Műszaki és Gazdaságtudományi Egyetem, Távközlési és Médiainformatikai Tanszék
ildi@cs.bme.hu, gaspar@tmit.bme.hu
LUKÁCS ANDRÁS
Magyar Tudományos Akadémia Számítástechnikai és Automatizálási Kutató Intéze (MTA SZTAKI)
alukacs@sztaki.hu

Kulcsszavak: webtartalmak hitelessége, portálstruktúra, méret- és tartalomfüggetlen minősítés

Az interneten található tartalomszolgáltatók, hírportálok számának növekedésével egyre fontosabb cél e szolgáltató rendszerek megbízható minősítése. A cikkben új, tényalapú megközelítésben vizsgáljuk a minőség meghatározásának kérdését és az ennek kapcsán felmerülő fogalmakat. A minőség mérését a mérhető felhasználói viselkedésre alapozzuk. Megadunk egy a felhasználók böngészését leíró modellcsaládot, amelyet az adott internetes szolgáltató elektronikus forgalmi naplóállományát feldolgozva, paraméterillesztési technikák alkalmazásával optimalizálunk, hangolunk. Az így kapott modell paramétereinek segítségével következtetünk a vizsgált hírportál oldalcsoportjainak minőségére. Bemutatjuk az ehhez szükséges összetett rendszert és eredményeinket egy jelentős hazai tartalomszolgáltató adatain demonstráljuk.

1. Bevezetés

1.1. Motivációk

Az információk szinte végtelennek tűnő tárháza nem csupán előnyöket rejt magában. A weben megtalálható dokumentumok sok esetben hibásak, hiányosak, vagy egyszerűen rossz minőségűek. A felmerülő tartalmi hiányosságok a formai hibáknál nehezebben deríthetők fel, viszont döntően befolyásolják az adott dokumentum hasznosságát és fogyaszthatóságát. Ebben a helyzetben ígéretesnek és fontosnak tűnik egy objektív minőségvizsgálati mérce felállítása.

Ebben a cikkben az internetes tömegkommunikációban kiemelkedő szerepet játszó hírportálokkal foglalkozunk. Ennek egyik oka, hogy egy internetes újság, mint haszon orientált szervezet esetén nem csupán a felhasználók, azaz az olvasók kíváncsiak egy-egy oldal, vagy összetartozó oldalcsoport *(rovat)* minőségére, hanem maga az üzemeltető is. A magasabb színvonal több látogatót vonz a hírportál olvasói táborába, elégedettebb olvasókat eredményez. Ez hosszabb távon lehetőséget nyújt – például a hirdetéseken keresztül – a vállalati profit növelésére. Ezért egy megbízható minősítést segítő módszer nemcsak az olvasók igényeinek kielégítését segíti, hanem egyértelműen az adott portál üzemeltetőjének érdekeit is szolgálná.

A másik okunk, hogy hírportálok minőségét vizsgáljuk, az e portálokat jellemző nagyobb forgalomban és a portál strukturáltságában rejlik. Így lehetővé válik a dokumentumok nagyobb, összetartozó egységeinek, rovatainak vizsgálata és azok tulajdonságainak összehasonlítása.

1.2. Korábbi megközelítések

Az internet elterjedése maga után vonta egy új tudományág, a webes adatbányászat kialakulását. Ennek célja, hogy elemezze, értelmezze, és hasznosíthatóvá tegye a világhálón megjelenő nagymennyiségű adatot és kapcsolataikat. A cél mindig egyfajta tudáskinyerés, azonban a különféle alkalmazásokhoz igazodva egészen eltérő technikák születtek ennek elérésére.

A ma fellelhető publikációk zöme – jó közelítéssel – az alábbi három csoport valamelyikébe sorolható be:

– struktúra analízis,
– tartalom analízis, illetve
– a felhasználói viselkedés elemzése.

A *struktúra analízis* célja a világháló dokumentumai között hiperhivatkozásokkal kialakult struktúrák felismerése és megtalálása [1,2]. A megismert strukturális jellemzőket használják ki például az intelligens keresőrendszerek [3]. Számos kutatás nem pusztán a web szerkezetét igyekszik felderíteni, hanem az elektronikus levelek vagy más kommunikációs forma használatának vizsgálatával az internethasználók közti kapcsolatokat próbálja feltárni [4].

A *tartalom analízis* esetében a cél a webes dokumentumok osztályozása különféle szempontokból. Az eddigi kutatások többsége a dokumentumok tartalom alapján történő klasszifikációjával [5] vagy automatikus feldolgozásával [6] foglalkozik. Ezeken a területeken az adatbányászati technikák mellett sokszor a gépi tanulás, mesterséges intelligencia eredményeit is alkalmazzák, erre adnak példát az információkereső és -osztályozó ágensek [7]. Fontos észrevétel, hogy az osztályozás speciális esetéhez jutunk a dokumentumok minőségének meghatározásával is. Webes dokumentumok minőségének vizsgálata az eddigi irodalomban kizárólagosan csak a dokumentumok keresésével, pontosabban a találatok rangsorolásának keretein belül tárgyalták [3].

A *felhasználók modellezése* a webes adatbányászat legfrissebb területe. Az egyik legtöbbet vizsgált probléma a felhasználók böngészési szokásainak leírása, a felhasználói viselkedés modellezése és elemzése [7].

Statisztikai elemzéseken túl ma már számos módszer ismert gyakori útvonalak és egyéb tipikus viselkedési mintázatok megtalálására [9,10]. Ezeket az eredményeket a felhasználói magatartás előrejelzésében, és az erre épülő adaptív, személyre szabott szolgáltatást kínáló weboldalak fejlesztésében hasznosítják [11]. Ezek mellett a módszerek mellett megjelent néhány modell alapú megközelítés is, ezek közül a legjelentősebbek a rejtett Markov-modelleken *(HMM)* alapuló kutatások, melyeket egyre szélesebb körben alkalmaznak [12,13].

Az általunk felhasznált ötlet alapja a fenti megközelítések vegyítése. A felhasználókról megszerezhető tudás segítségével, a böngészési szokásait leíró modellre alapozva próbáljuk meghatározni a portál egy-egy összetartozó oldalcsoportjának, rovatának minőségét.

Ez párhuzamba állítható a legelterjedtebb szabadszavas kereső, a *Google* által alkalmazott minősítési eljárás, a page-rank módszerével. A page-rank a felhasználó böngészését – az egész webre vonatkozó konkrét adatok hiányában – a lehető legegyszerűbben egy a hiperhivatkozásokon történő bolyongással modellezi, majd a modell bizonyos paramétereinek segítségével minősíti a dokumentumokat [14].

Az általunk vizsgált esetben feltételezzük, hogy a minősíteni kívánt oldalakon történő böngészésről naplóállományok, weblogok állnak rendelkezésünkre. Ezek segítségével a felhasználó böngészésének egy részletesebb modelljét tudjuk megadni, kiszámolni. A kapott modell fogja tartalmazni azokat a paramétereket, melyeket a dokumentum csoportok minőségeként lehet értelmezni. Ez a weblogra építő megközelítés megjelenik a page-rank egy változatában is [15].

1.3. A minőség fogalma

Eddig nem terjedt el a minőségnek akár korlátozottan is elfogadott meghatározása. A következőkben végiggondoljuk, hogy mi szükséges egy megfelelő definícióhoz, milyen megfontolásokra támaszkodhatunk.

A minőség definiálása során egy messzemenően szubjektív fogalmat igyekszünk megfelelően absztraktá tenni. Amennyiben egy hírportál rovatai között szeretnénk megtalálni a „legjobbat” vagy éppen a leggyengébb minőségűt, biztosak lehetünk abban, hogy nincsen *tökéletes* választás, ugyanis az általunk hozott döntést nagy valószínűséggel befolyásolja egyéni ízlésünk, értékrendszerünk. Megoldásképpen statisztikai megközelítéssel élhetünk, megpróbálhatjuk kifejezni egy „átlagos felhasználó” nézeteit. Ez általában még mindig nehezen megoldható probléma marad az összes felhasználó viselkedésére vonatkozó adatok hiánya miatt.

Azonban ha az internetes hírportálok felhasználóira korlátozzuk vizsgálatainkat, akkor megfelelő kiindulópontot jelent, hogy ezen hírportálok rögzítik az általuk lebonyolított forgalmat, azaz tárolják a felhasználóktól a portálhoz érkező oldallekérdezéseket. Ezt a folyamatot elektronikus naplózásnak, a kapott adathalmazt – mely többek közt tartalmazza a kliens anonim azonosítóját, a lekért dokumentumok azonosítóját, a kérés időpontját – pedig naplóállománynak, weblognak nevezzük.

Élünk azzal a feltevéssel, hogy egy dokumentum vagy rovat minőségén keresztül befolyással van a böngészés menetére. Tehát a minősítés feladata megfogalmazható úgy, hogy a weblogból, valamint a felhasználókra vonatkozó a priori feltételezéseinkből kiindulva megalkotunk egy böngészési modellt, amely leírja a felhasználó viselkedését a minőség és a hangolandó paraméterek függvényében, majd a rendelkezésre álló weblog alapján ezt a modellt összhangba hozzuk a valósággal. Ily módon lehetőségünk lesz egy implicit módon definiált, reprodukálható minőségi mérce felállítására.

2. Modellezés és szimuláció – egy komplex rendszer

2.1. A megoldás alapgondolata

A minősítési rendszer magvát egy felhasználói modell alkotja Ez a felhasználónak a böngészés során mutatott viselkedését írja le olyan módon, hogy egy adott böngészési helyzetben meghatározza, jellemzi a felhasználó valószínűsíthető következő oldalletöltését. A modell stochasztikus, azaz a felhasználó egyes helyzetekben lehetséges cselekvéseinek valószínűségeit adja meg, és ennek a valószínűségi eloszlásnak megfelelően a felhasználó döntése egy véletlen kísérlettel realizálható.

A modell paraméterein keresztül implicit módon definiáljuk a minőséget. A modellben megjelenik, hogy a böngészés során meglátogatott oldal minősége miként befolyásolja a böngészés további folyamatát. A modell paramétereinek értékei nincsenek előre rögzítve. Célunk azon paraméterértékek meghatározása, amelyeket a modellbe behelyettesítve a modell a legpontosabban írja le egy adott portál valódi böngészéséből származó webes naplóállomány tulajdonságait. Így a minőség meghatározása egy modellillesztési problémára vezethető vissza.

Az ismeretlen értékű paraméterek meghatározására sokféle módszer létezik. A modellben szereplő paramétereket valamilyen optimalizációs eljárás segítségével hangoljuk. Ehhez szükség van az aktuálisan vizsgált paraméterekkel ellátott modell jóságának (vagy hibájának) ismeretére. Ezt az értéket esetünkben a modell segítségével szimulált weblog és az eredeti naplóállomány hasonlósága fogja megadni. Így az iteratívan szimulációkkal, összehasonlításokkal és paraméter-változtatásokkal dolgozó optimalizáció végén megkapjuk azt az – immár paraméterezett – modellt, mely a lehető legközelebb áll a felhasználók mért természetéhez. A kinyert paraméterek között fognak szerepelni a keresett minőséget leíró paraméterek is.

2.2. A felhasználói modell

A felhasználót leíró modell megalkotása során el kell döntenünk, hogy a böngészés folyamatát milyen szempontok szerint vizsgáljuk meg, mik lesznek a modellben szereplő alapfogalmak. Fel kell térképeznünk a modellezendő jelenségeket, és végül a kialakult modellt megfelelő matematikai formába kell öntenünk.

Először megadjuk a kritériumok azon két csoportját, amelyeket a modellünktől elvárunk. A modellezendő területről, a felhasználók viselkedéséről alkotott előzetes elképzeléseinkből és információinkból nyilvánvalóan kialakulnak azon elvek, amelyeket minden modellalkotási folyamatban érdemes figyelembe venni. Vegyük sorra ezeket.

Általános kritériumok

Ellenőrizhetőség: minden használható modelltől elvárhatjuk, hogy ellenőrizhető legyen, vagyis létezik olyan módszer, amelynek segítségével meg tudunk adni egy hibamértéket a modell és a valóság viszonyának jellemzésére. Látni fogjuk, hogy ez az általunk választott modell esetén többféleképpen is megoldható. A szükséges mértékek megtalálásához a statisztika adja az alapot.

Kiszámíthatóság: a modellezés során a modell helyességének mérésére használt érték gyakran a modell által jósolt események és a valóság összehasonlításán alapul. Ezért szükségszerű, hogy az összehasonlításhoz szükséges jellemzők hatékonyan számíthatók legyenek a modellből. Ez bármely típusú modellillesztés vagy optimalizálás alapja. Előfordulhat, hogy a modell leírásából nem számolható ki közvetlenül az összehasonlítás tárgyát képező érték. Ilyen például a sztochasztikus modellek túlnyomó többsége. Ez ugyan megnehezíti a modellillesztés hatékonyságát, de szimulációk segítségével – sok esetben – kezelhető marad a probléma.

Értelmezhetőség: elvárható még, hogy a modellben használt feltételezések indokolhatóak és a modellben szereplő paraméterek intuitív módon értelmezhetőek legyenek. Az egyes modelljelöltek vizsgálata során az átláthatóság és a kisebb hibázási lehetőség érdekében érdemes az egyszerűbb modelltől a komplexebb felé haladni.

Területspecifikus elvárások

A böngészés, illetve a minőség fogalmának tulajdonságaiból kiindulva az alábbi elvárásaink lehetnek:

Időbeli stabilitás: a modellillesztés eredményeként kapott paraméterértékektől elvárjuk, hogy ne mutassanak erős változásokat rövidtávon. Ennek az a priori feltételezés ad alapot, hogy a vizsgálni kívánt globális jellegű tulajdonságok tekintetében sem a böngészés folyamatának törvényszerűségei, sem a benne résztvevő szereplők (felhasználók és a portál) nem változnak gyorsan.

Térbeli stabilitás: ez alatt azt értjük, hogy a modellnek érzéketlennek kell lennie az aktuálisan vizsgált felhasználók halmazának nagyságára. Azaz ha a felhasználóknak csak egy véletlenszerűen kiválasztott részét tekintjük, akkor azok viselkedését is jellemezze megfelelően a modell, mindaddig, míg számuk elegendő a sztochasztikus megközelítés alkalmazásához. A térbeli stabilitás fogalmát nemcsak a felhasználók oldaláról lehet megközelíteni, hanem a hírportálok rovatainak szemszögéből is. Ekkor azt – az előzővel analóg módon –, csak a portálon szereplő rovatoknak egy véletlenszerűen választott részhalmazánál vizsgáljuk.

A minőség rovatmérettől való függetlensége: a rovatok mérete, azaz a hozzájuk tartozó dokumentumok száma ne befolyásolja nagyobb mértékben a rovat minőségét.

A minőség függetlensége a téma népszerűségétől: a rovathoz kötődő téma popularitásától lehetőleg független legyen a modellillesztésből adódó minőség értéke. Ez jogos kívánalom, hiszen minden témában lehet színvonalas vagy éppen kevésbé jó minőségű rovatot létrehozni. Tipikus példát adnak az utóbbiakra a valószínűleg témájukból kifolyólag magas látogatottságú, ám a többinél gyengébb minőségűnek mért rovatok. Lényeges megszorítás, hogy az eddigi eredmények csak a naplózott portál rovatminőségeinek összehasonlítására alkalmas mutatókat adtak, a portál egészének minősége nem összehasonlítható a mérésen kívüli internetes tartalmakkal.

Sajnos ez utóbbi elvárás megvalósulása nehezen ellenőrizhető, hiszen a népszerűség mérésének nehézsége összemérhető a minőség mérésének problémájával. Első megközelítésként a rovatot összességében meglátogató olvasók száma megfelelő mértéknek tűnik, hiszen egy téma népszerűsége várhatóan megjelenik a témához tartozó rovat látogatottsági számaiban, de a látogatottságot nyilván befolyásolja a rovat minősége is.

Észrevehető, hogy az utolsó két elvárásunk nem közvetlenül a modellre, hanem a kívánt minősítés milyenségére vonatkozik. Mivel a minősítő rendszer magvát a modell képezi, ezért a minőségre vonatkozó elvárásainkat is a modell tulajdonságainak helyes megválasztásával tudjuk elérni.

A modell szereplői és egységei

A modell két legfontosabb elemét egyrészt a böngészést végző felhasználók, másrészt az általuk meglátogatott oldalak, illetve azok csoportjai, a rovatok adják. Ezt a két fogalmat kapcsolja össze a böngészés folyamata, amelynek kapcsán az időbeliségre is ki kell térnünk, hogy definiálhassuk a böngészés egységét.

A felhasználó: az a személy, aki az interneten keresztül meglátogatja az általunk vizsgált hírportál oldalainak valamelyikét. Az így kapott olvasók körét azonban érdemes leszűkíteni azokra a felhasználókra, akik legalább néhány oldalt letöltöttek, hiszen az egy-két oldalkérést tartalmazó böngészések túl rövidek az oldalak hatásainak mérésére.

A felhasználókra vonatkozó legfontosabb feltételezésünk, hogy *homogének*. A valóságban az olvasók nyilvánvalóan nem egyformák, ám a következőkkel indokoljuk feltételezésünket:

– A felhasználók homogenitását valamilyen előfeldolgozás segítségével fokozhatjuk, például osztályozzuk az olvasókat az általuk letöltött oldalak száma alapján.
– A böngészést végző emberek nagy száma miatt a modellben szereplő homogén, de statisztikailag átlagos jellemzőket mutató felhasználók sokasága közelítőleg egyenértékű lesz a valóságban inhomogén felhasználó halmazzal.

Az előző két megfontolást kombinálva egy kevert modellhez jutunk, amelyben az előfeldolgozás osztályozása után minden megkapott felhasználói csoportra külön-külön illesztjük a modellt, majd az utófeldolgozás során a kapott modelleket összevetjük. Az általunk vizsgált modellekben nem használtunk előfeldolgozást, így a későbbiekben a kevert modell megvalósítása egyfajta ellenőrzésként is szolgálhat.

A rovatok: a böngészés tárgyai. Vizsgálatunk tárgyát képező portál közel 40.000 dokumentumot tartalmazott. Mivel az oldalak letöltésszámának eloszlása jó közelítéssel hatványeloszlást mutat, még a portál napi több milliós összletöltésszáma mellett is az oldalak túlnyomó többségét csak néhányszor töltik le. Így ezekről az oldalakról nem lesz elegendő információnk, hogy minőséget mérjünk. Főként hírportálok esetén további probléma, hogy az oldalak időben gyakran változnak. Ezért vizsgálatunk tárgya az oldalak helyett inkább az adott portál rovatai. Rovat alatt *oldalak egy szervesen összetartozó csoportját* értjük. A rovatok és a hozzájuk tartozó dokumentumok pontos kiválasztása a site szerkezete alapján könnyen megoldható volt.

A rovatokon belül *az egyes oldalakat nem különböztetjük meg*. Mégis szükség van néhány, az oldalak szintjét érintő előszűrésre, például:

– nem létező, irreleváns vagy értelmetlen oldalkérések kiszűrése;
– a portál főoldalára vonatkozó kérések kiszűrése, annak túlzott látogatottsága miatt;
– a dokumentumok automatikus frissítéséből adódó ismétlődő letöltések szűrése.

Böngészési sorozat: egy böngészési sorozat, másnéven *session* egy adott felhasználótól egy adott időintervallumban a portálhoz beérkező letöltési kérdések sorozata. A használandó időegység kiválasztásakor a következő szempontokat vehetjük figyelembe:

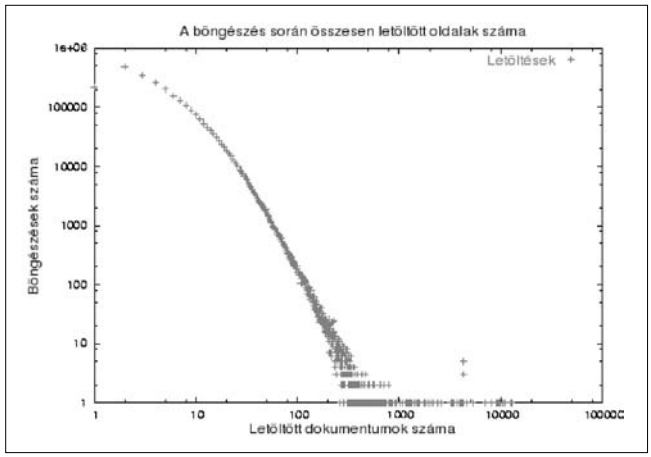
– *A letöltések sűrűsége:* minél sűrűbben követik egymást a felhasználó letöltései, annál valószínűbb, hogy ezek összefüggnek.
– *Periodicitás:* ha valamilyen ismétlődő jelleget fedezünk fel a felhasználók viselkedésében, akkor egy periódus alatt történt letöltések egységnek tekinthetők.

Mivel két egymással összefüggő oldalletöltés között eltelt idő nagyon változatos lehet ezért a gyakorlattól eltérően a nem az oldalletöltések között eltelt idő hosszára alapoztuk a session definícióját. A rendelkezésre álló adatok mennyiségét figyelembe véve megfelelőnek tűnt az egynapos periódus választása. A weblogban egy letöltési kérelemhez, klikkeléshez mint rekordhoz a következő mezők tartoznak: egyedi anonim felhasználó-azonosító *(cookie)*, a session azonosítója, a dokumentum azonosítója, a rovat azonosítója, melyhez a letöltött oldal tartozik, végül a letöltés időbélyege.

A modellezendő jelenségek és események

A böngészést jellemző jelenségek közül a legfontosabb tapasztalat, hogy a felhasználók a böngészés során folyamatosan „fáradnak”.

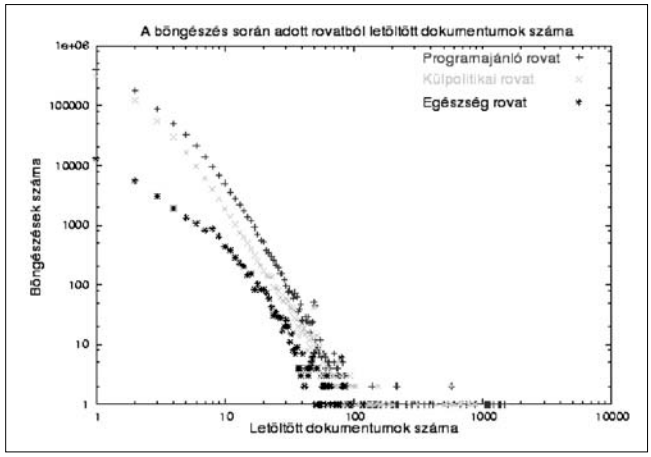
Ha megvizsgáljuk azt a hisztogramot, mely a felhasználók számát mutatja a mérési idő alatt általuk letöltött oldalak számának függvényében *(1. ábra)*, láthatjuk, hogy ez a függvény meredeken csökkenő hatványfüggvény lefutású. Ez azt a feltételezést valószínűsíti, hogy a felhasználót az általa korábban letöltött oldalak száma nagyban befolyásolja annak eldöntésében, hogy letölt-e még egy oldalt, vagy befejezi a böngészést.



1. ábra
A böngészési sorozatok száma a böngészés során letöltött oldalak számának függvényében

Ugyanezt mondhatjuk el, ha csak egy adott rovaton belüli letöltéseket vizsgálunk.

A *2. ábrán* jól látható, hogy különböző rovatok esetén eltérő mértékben jelentkezik az elfáradás jelensége, tehát az adott számú letöltést végző böngészések gyakorisága meredekebben csökken bizonyos rovatok esetén. Ez intuitív módon azt jelenti, hogy bár esetünkben például a külpolitikai rovatot többen nézik meg, mint az egészséggel foglalkozó rovatot, az olvasók mégis jellemzően hosszabb ideig tartozkódnak az utóbbiban. Ez a jelenség nagy valószínűséggel összekapcsolható a két rovat eltérő minőségével.



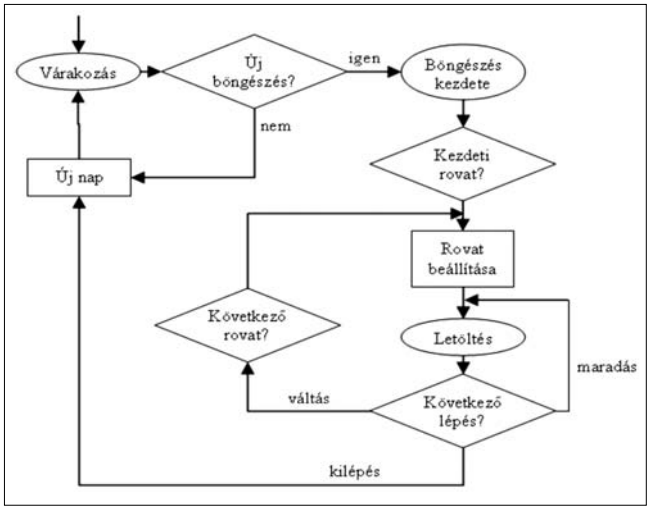
2. ábra
A böngészési sorozatok száma a böngészés során adott rovatból letöltött oldalak számának függvényében

Figyelembe vehetjük még a rovatokban található dokumentumok frissülésének, illetve elévülésének jelenségét is. A rovatban található „friss”, azaz az olvasó számára még új dokumentumok száma érdemben befolyásolja, hogy a böngészést az adott rovatban tovább folytatja-e.

Egy böngészés elemi eseményei a következők:

- *A böngészés (session) kezdete:*
  a felhasználó minden nap dönt arról, hogy böngészik-e aznap, vagy sem.
- *A kezdeti rovatba ugrás:*
  a böngészési sorozat kezdetekor a felhasználó rovatot választ, amelyben megkezdi a böngészést.
- *Rovatban maradás:*
  a böngészés során a felhasználó minden letöltés után dönthet arról, hogy a következő letöltendő oldal szintén az aktuális rovatból kerül-e ki.
- *Rovatváltás:*
  egy letöltés után rovatot vált a felhasználó.
- *A böngészés (session) vége:*
  a felhasználó úgy dönt, befejezi a böngészést.

Ezek alapján a felhasználó viselkedését a *3. ábrán* látható folyamatábrával írhatjuk le.



3. ábra
A böngészést végző felhasználó viselkedésének folyamatábrája

A modell létrehozásánál a böngészés négy elemi valószínűségének definiálására van szükség. Modellcsaládunk egyik legegyszerűbb tagjánál az egyes események valószínűsége a következő módon számolható:

– Az aznapi böngészés elkezdésének valószínűsége konstans.
– A kezdeti rovat kiválasztása során az egyes rovatok közül az alapján választunk, hogy az eredeti weblogban a felhasználók milyen relatív gyakorisággal tették ugyanezt. Hasonlóan viselkedik a modell, ha új rovatra váltásról van szó, azaz elsőrendű Markov-lánccal modellezünk.
– Minden oldalletöltés után a modell eldönti, hogy marad-e az adott rovaton belül. Ennek értékét a következő módon számítjuk:

$$f_{marad}(o_i, f_i, m_i) = f_i \cdot z^{(1-m_i)o_i}$$

ahol

o_i az adott rovatból elolvasott oldalak száma,
f_i az adott rovat oldalainak frissüléséi rátája,
m_i az adott rovat minőségértéke,
z pedig egy 0 és 1 közötti szabad paraméter.

– Ha a felhasználó nem marad a rovaton belül, akkor konstans valószínűséggel (y) befejezi az aznapi böngészését.

2.3. Modellillesztés

A modellillesztés feladata a hírportál által rögzített weblog alapján meghatározni a modellben szereplő ismeretlen paramétereket.

Előfeldolgozás

A modellillesztés a hírportálhoz beérkező kérések sorozatát tároló elektronikus naplóállomány, a *weblog* alapján történik. A weblog „nyers” változatát úgy kell átalakítani, hogy explicit formában is tartalmazza a későbbiekben fontossá váló adatokat, mint amilyen például a rovat azonosítója. Ezen túlmenően a felesleges mezők kiszűrését, és az esetleges egyéb szűréseket – például a főoldalra vonatkozó letöltések kiszűrését – is el kell végezni.

A teljes modell jóságának mérése

A felhasználói modellben szereplő ismeretlen paraméterek értékét - melyek közül számunkra az egyes rovatok minősége a legfontosabb - egy szélsőérték-kereső eljárás segítségével állapítjuk meg. Ehhez becsülni kell annak helyességét. Ezt a kulcsfontosságú problémát szimuláció segítségével oldjuk meg.

A *szimuláció* egy fázisa során egy adott paraméterbeállítást használva a felhasználói modell alapján – sztochasztikus módon – egy böngészési sorozatot állítunk elő.

Ezt megfelelően sokszor megismételve egy *mesterséges weblogot* kapunk, mely megfelel egy olyan weblognak, amelyet az általunk alkalmazott modellel leírható felhasználók oldalkérései generálnak. Mivel a felhasználói modell a rovatokról is tartalmaz információt, a mesterséges weblog egy ezeknek megfelelő tulajdonságú, hasonló minőségű rovatokkal bíró portál működését írja le. Az eredeti és a mesterséges weblog akkor lesz hasonló, ha sikerült jól közelítenünk a valóságos böngészést. Adott modell esetén ez a mérés a paraméterek jóságának meghatározására szolgál, ugyanakkor különböző modelltípusok összehasonlítására is alkalmas.

A feladat tehát két weblog összehasonlítása. Ezt elméletileg megtehetjük, de a közvetlen összehasonlításhoz a weblogok nagy mérete miatt ez jelentős számítási kapacitást tenne szükségessé. Az igen nagyszámú szimuláció szükségessége miatt hatékonyabb megoldásra van szükség. Ezt úgy tudjuk elérni, hogy ha nem direkt módon a weblog adataival, hanem belőlük nyert statisztikákkal mérünk.

Az általunk megvalósított rendszerben tizenhét különböző statisztikát használtunk. Ezek közül néhány:

- *Felhasználó – dokumentum hisztogram:* a felhasználók mekkora hányada tölt le adott számú oldalt.
- *Session – rovatszám hisztogram:* a böngészési sorozatok mekkora hányada tartalmaz adott számú rovatra vonatkozó oldalkéréseket.
- *Rovatváltási mátrix:* adott rovatból mekkora eséllyel lép át a felhasználó egy másik rovatba.
- *Session – dokumentum hisztogram egy rovatra:* a session-ök mekkora hányadában töltöttek le a kérdéses rovatból adott számú dokumentumot.

Hisztogramok összehasonlítását több módszerrel is elvégezhetjük:

- *L_2 norma alapú összehasonlítás:* a hisztogramok azonos oszlophoz tartozó értékeinek különbségét négyzetre emeljük, majd minden oszlopra összegzünk. Minél kisebb az így kapott nemnegatív érték, annál hasonlóbb a két weblog.
- *χ-négyzet próba:* a hisztogramokat gyakoriságokat tartalmazó táblázatként felfogva valójában a feladat megfogalmazható a klasszikus homogenitásvizsgálatként. Ekkor a cél annak a valószínűségnek a megállapítása, hogy az adott gyakoriságértékek mekkora eséllyel származnak azonos eloszlásból – pl. mekkora valószínűsége van annak, hogy a valóságban, illetve a modellben azonos eloszlás szerint változik a letöltések száma egy sessionben. Éppen ezt a feladatot oldja meg a gyakran használt statisztikai χ-négyzet próba. Minél nagyobb a kapott valószínűség, annál inkább hasonlít egymásra a két weblog.

Több statisztika esetén az egyes hisztogramokra kapott hibaértékek (vagy az utóbbi esetben hasonlóságértékeket) súlyozott összegeként kapjuk az adott paraméterekhez tartozó modell hibáját (jóságát).

Optimalizálás

A modellillesztést egy optimalizáló eljárás végzi, mely a modell hibáját minimalizálja (vagy a jóságát maximalizálja). Ennek megoldására sok algoritmus létezik, legtöbbjük a gradiens alapú szélsőérték-kereső eljárások körébe tartozik. Ezek legfontosabb előnye a gyorsaság, azonban működésükhöz szükséges a hibafelület gradiensének kiszámítása, amire esetünkben nincs közvetlen lehetőség. Sok algoritmus létezik, mely nem használja a gradiens fogalmát, azonban ezek jelen esetben nem elégségesek.

Értelmes kompromisszumot kínált a gradiens becslésén alapuló SPSA (Simultaneous Perturbation Stochastic Approximation) algoritmus [16]. A gradiens becsléséhez az SPSA néhány véletlenszerűen kiválasztott irányba lép el a keresési térben (azaz az ismeretlen paraméterek terében), majd az így kapott pontokban végzett szimulációk segítségével számított hibaértékekből approximálható a gradiens. A szimuláció indításakor az optimalizálandó paraméterek kiindulási értékét általában véletlen választással adjuk meg.

Az SPSA algoritmus alkalmazásakor felmerülő főbb problémák a következők lehetnek:

- *Lokális optimumok:* ezek elkerülésére több kezdőpontból is futtathatjuk az optimalizáló eljárást.
- *Lépésköz mérete:* ennek megválasztására széles körben elterjedt heurisztikák léteznek. Gyakran használt módszer például, hogy két jó (az optimumhoz közelebb vivő) lépés után a lépésközt növeljük, egy jó lépés utáni hibás lépés esetén viszont csökkentjük azt. A növelés legtöbbször additív, míg a csökkentés multiplikatív módon történik.
- *Zajosság:* a statisztikák használata miatt természetes módon belép a rendszerbe valamekkora zaj. Ennek csökkenését úgy érhetjük el, hogy a mesterséges weblog létrehozásakor a szimulációk során megfelelően sok sessiont állítunk elő.

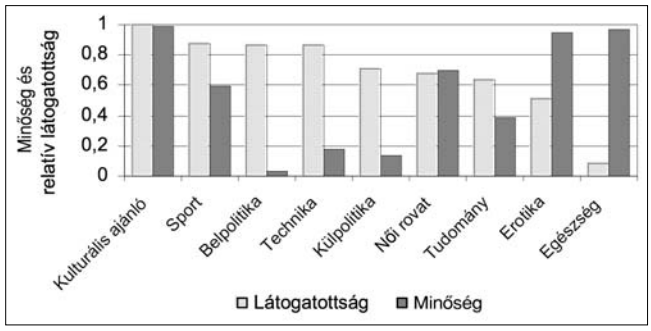
3. Eredmények

Az alkalmazási feladat egy hazai vezető internetes hírportál 9 rovatának minősítése volt. Ehhez rendelkezésünkre állt a hírportál üzemeltetői által rendelkezésünkre bocsátott naplózófájl, melyben 28 egymást követő nap böngészéseinek adatai szerepeltek. A nyers naplóállomány mérete több tíz gigabájtos nagyságrendű volt.

3.1. Az elvégzett szimulációk

A legpontosabb modell kiválasztása érdekében több, szisztematikusan felépített modellcsaládra végeztünk szimulációkat. Ezeket mind a χ-négyzet próba szerint, mind az L_2 norma alapú távolság szerint összehasonlítottuk. A legalkalmasabb modell kiválasztása után a modellillesztés eredményeképpen megkaptuk az optimális paramétereket, ezek között szerepeltek a minőségértékek is.

Az eredményeket a *4. ábra* tartalmazza, a rovatok témája mellett azok felhasználói látogatottságát és az általunk becsült minőségét tüntettük fel.



4. ábra
Egyes rovatok minőségének és látogatottságának értéke

3.2. A szimulációk erőforrás-szükségletei

A szimulációs program végrehajtása – egy P4, 1,4 GHz-es processzorral – letöltésenként (klikkelésenként) mintegy 0,7-0,8 ms nagyságrendű időt vesz igénybe.

Ismerve a modell által generált felhasználói sorozatokban naponta letöltött dokumentumok átlagos számát, kiszámolható, hogy a 28 napos szimulációk során alkalmazott 500-as felhasználói létszám mellett egyetlen szimuláció körülbelül 30 másodpercet vesz igénybe. Mivel egyetlen optimalizációs fázisban néhány száz szimulációt végzünk a modellillesztéshez szükséges futási idő mintegy 130-140 perc. Ha több véletlenszerűen kiválasztott pontból is elindítjuk az optimalizációt indítani, akkor ez 10-12 próbálkozás esetén már kitesz egy teljes napot. A program futásának ez a viszonylagos lassúsága ugyanakkor nem okoz jelentős problémát, hiszen a minősítési feladat nem igényel valóidejű működést.

4. Összefoglalás

Cikkünkben áttekintettük a modell kialakítása során felmerült általános tervezési elveket és felvázoltuk a legfontosabb döntési lehetőségeket. A felhasználói modellek vizsgálatára kidolgoztunk egy komplex rendszert, amely a modellben szereplő paramétereket illeszti a valós adatokhoz, a hírportál weblogjához. A modellillesztés során többféle statisztika felhasználásával, a felhasználói modell segítségével mesterségesen szimulált weblogokat hasonlítunk össze az eredeti naplófájllal. Így megkaphatók a modellbe épített ismeretlen paraméterek legvalószínűbb értékei, azaz a hírportálok rovatainak minősítése.

A rendszert implementáltuk és egy jelentős hazai tartalomszolgáltató weblogján ellenőriztük.

Köszönetnyilvánítás

Köszönetet mondunk Rácz Balázsnak és Szepesvári Csabának hasznos észrevételeikért és tanácsaikért, mellyel munkánkat segítették.

Irodalom

[1] David Gibson, Jon Kleinberg, Prabhakar Raghavan: Inferring web communities from link topology. In Conference on Hypertext and Hypermedia, ACM, 1998 és IEEE Comm. Magazine, July 2001.

[2] E. Spertus: Parasite: Mining structural information on the web. Computer Networks and ISDN Systems: The International Journal of Computer and Telecommunication Networking, Nr.29, 1997, pp.1205–1215.

[3] Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd: The pagerank citation ranking: Bringing order to the web. Technical Report, Stanford Digital Library Technologies Project, 1998.

[4] Wil M. P. van der Aalst, Minseok Song: Mining Social Networks: Uncovering Interaction Patterns in Business Processes. Business Process Management 2004, pp.244–260.

[5] M. Steinbach, G. Karypis, V. Kumar: A comparison of document clustering techniques. In KDD Workshop on Text Mining, 2000.

[6] V. R. Borkar, K. Deshmukh, S. Sarawagi: Automatic Segmentation of Text into Structured Records. In Proc. ACM-SIGMOD International Conference Management of Data (SIGMOD 2001), ACM Press, New York, 2001, pp.175–186.

[7] Eui-Hong (Sam) Han, D. Boley, M. Gini, R. Gross, K. Hastings, G. Karypis: A Web Agent for Document Categorization and Exploration. In Proc. of the 2nd International Conference on Autonomous Agents (Agents'98).

[8] Lara Catledge, James Pitkow: Characterizing browsing strategies in the WWW. Computer Networks and ISDN Systems, Nr.26, Vol.6, 1995, pp.1065–1073.

[9] M.S. Chen, J.S. Park, P.S. Yu: Data mining for path traversal patterns in a web environment. In 16th International Conference on Distributed Computing Systems, 1996, pp.385–392.

[10] J. Pei, J. Han, B. Mortazavi-Asl, H. Zhu: Mining Access Patterns Efficiently from Web Logs. In Proceedings Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2000.

[11] Ralph Kimball, Richard Merz: The Data Webhouse Toolkit: Building the Web-Enabled Data Warehouse. John Wiley & Sons, 2000.

[12] C. Anderson, P. Domingos, D. Weld: Relational Markov Models and their Application to Adaptive Web Navigation. In Proc. 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Canada, 2002, pp.143–152.

[13] A. Ypma, T. Heskes: Clustering web surfers with mixtures of hidden Markov models. In Proc. of the 14th Belgian-Dutch Conference on AI (BNAIC '02), 2002.

[14] Friedman Eszter, Uher Máté, Windhager Eszter: Keresés a Világhálón, Híradástechnika, 2003/3., pp.20–24.

[15] B. Uygar Oztekin, Levent Ertöz, Vipin Kumar, Jaideep Srivastava: Usage Aware PageRank. In Proc. of the 12th International WWW Conference, Budapest, Hungary, 2003.

[16] John L. Maryak, Daniel C. Chin: Global random optimization by simultaneous perturbation stochastic approximation. In Proc. 33nd Conference On Winter simulation, Virginia, 2001, pp.307–312.

MIS – üzleti intelligencia megoldások az LLP-től

A London Logic Budapest Számítástechnikai és Kereskedelmi Kft. (LLP) már hét kelet- és közép-európai országban jelen van szolgáltatásaival. Az angol Management Information System-t magyarul általában Vezetői Információs Rendszernek fordítják. Ez a rövidítés az LLP Budapest esetében nem általánosságban az MIS rendszerekre utal, hanem arra a konkrét megoldásra, melyet az LLP Csoport is képvisel, s melynek szintén ezt a nevet adta a MIS AG, egy német vezetői információs rendszereket és üzleti intelligencia megoldásokat fejlesztő vállalat, melyben 2003 végén az LLP egyik fő partnere, az angliai Systems Union jelentős tulajdont is szerzett.

Az Online Analaitical Processing magyarul talán a következőképpen fordítható: közvetlen elérésű analitikus adatfeldolgozás. Az OLAP egy multidimenziós adatbázis, amelyből a vállalati szükségeletek szerinti üzleti döntések meghozatalához szükséges, különböző mélységű információk és összefüggések nyerhetők ki.

Ma már egyre több vállalatnak van szüksége üzleti intelligencia megoldások alkalmazására, hogy a piaci kihívásokra minél gyorsabban tudjanak reagálni. A MIS Alea üzleti intelligencia megoldás és vezetői információs rendszer a legújabb, amely több mint 900 működő OLAP alkalmazást (vagyis közvetlen elérésű analitikus adatfeldolgozásra képes adatbázist) vizsgált meg.

Az MIS Alea-t az üzleti döntéshozók igényeire támaszkodva fejlesztették ki, amelynek használata nem igényel különösebb IT ismereteket, viszont megkönnyíti többek között a stratégiai vállalatvezetést, a költségtervezést, az anyavállalat és a leányvállalatok közti jelentéskészítést, vagy a hitel- és kockázatkezelést. A nagyvállalatok döntéshozói nap mint nap szembesülnek azzal a problémával, hogy döntéseiket csak számos, különböző forrásból származó információ birtokában hozhatják meg. Ugyanilyen nehézséget jelent az operatív szinten dolgozó kontrollerek és gazdasági elemzők számára, hogy megfelelő információ birtokában készítsék el jelentéseiket. Ezek ugyanis megmutatják, hogy a tervek szerint alakulnak-e az eredmények, és kiváló alapot nyújtanak a döntésekhez, így az eltérés korrigálható lehet.

Egy KPMG tanulmány szerint egy menedzsment idejének 20-30%-át fordítja tervezési feladatok elvégzésére; egy kb. 1 millió dollár forgalmú cég átlagosan 25 munkanapot fordít cége tervezési és elemzési folyamataira. Az éves költségvetés elkészítése mintegy 4-5 hónapot vesz igénybe, az eredmények alakulása alapján viszont a cég menedzsmentjének már csak 20%-a változtat a költségvetési terveken. Felmérések bizonyítják, hogy megfelelő eszközök nélkül a vezetők vagy alulbecsülik a költségvetést, vagy teljesíthetetlennek ítélik meg.

Az MIS csökkenti a tervezésre fordított időt, ezáltal csökkenti a költségeket is. A nagyteljesítményű MIS Alea elemző eszköz segítséget jelent a vállalatok számára, mivel megmutatja, hogy a tervek szerint alakulnak-e az eredmények, és kiváló alapot nyújt azokhoz a döntésekhez, melyekkel az eltérés korrigálható. *Paul Brigitta*

Hírek

A **London Stock Exchange, a Z/Yen Limited és a Sun Microsystems** bejelentette, hogy ágazati összefogással helyre kívánják állítani a világ pénzpiacainak áttekinthető és szabályozott működését. Az együttműködés célja a megbízások legjobb feltételek szerinti teljesítése: a piacfenntartó, folyamatos árjegyzésre vállalkozó piaci szereplők, valamint a brókerek/kereskedők kötelesek ügyfeleik ügyleteit a feladáskor elérhető legjobb áron teljesíteni. A cégek véleménye szerint az ágazati kezdeményezés jelentősen egyszerűsítené a legjobb feltételek szerinti teljesítésre vonatkozó előírás betartását, és segítené a törvényi szabályozás szerinti működést.

A vállalatok egy olyan korszerű rendszert fognak tesztelni, mely meghatározza, hogy az időpont, a volumen, a piaci feltételek és a kötések szokásos jellemzőit figyelembe véve elfogadható-e az ügyletek teljesítése. A rendszerben minden nem elfogadható díjú ügylet átadható vizsgálatra a felügyeleti szerveknek. A projekt a londoni tőzsde, az együttműködésre önként jelentkező vásárló és eladó cégeinek ügyleteire fog kiterjedni.

A vállalatok egyre szélesebb köre ismeri a **grid computing** technológiát és annak előnyeit, egy felmérés eredményei azonban azt jelzik, hogy a legtöbb még nem tett lépéseket annak bevezetése érdekében. Az összesített grid index értékek hasonlóak az egyes régiókban: Észak-Amerika értéke 4.50, Európáé 4.39, míg a délkelet-ázsiai, ausztráliai és óceániai térségé 4.37. Ezek az adatok azt mutatják, hogy az egyes régiók vállalatainak jelentős része vizsgálja, tanulmányozza és értékeli az új technológiát.

Általában elmondható, hogy a grid computing és annak előnyei pozitív visszhangra találnak (az indexek értékei 5.61-4.89 közt vannak), azonban ez egyelőre nem eredményez megfelelő támogatottságot, nem kapcsolódik hozzá sem elért, sem elvárt megtérülési ráta. A támogatottsági index jelenleg 2.45-ös, a megtérülési pedig 1.89-es értéket mutat. Ez a trend jellemzi a hasonló jelentőségű új technológiák bevezetését is.

Az európai vállalatok már elértek bizonyos sikereket a számítóhálós technológia bevezetése terén. Az európai Oracle Grid Index hat hónap alatt 3.1-ről 4.39-re emelkedett, ami jelentős eredmény. A vizsgált európai országok és az összesített európai index alapját képező összes érték növekedő tendenciát mutatott.