A kereshető webhely

A Webglimpse használata weboldalon történő kereséshez és a keresésen alapuló hirdetések elhelyezéséhez.

Régen, amikor még kíváncsi diák voltam, ellátogattam egy előadásra, amely egy kicsi képzeletbeli lényről, a ***Maxwell*** démonról szót. Ez az okos kis lény ki tudja válogatni a magas hőmérsékletű részecskéket a levegőből, feltéve, ha ismeri azok helyét és sebességét. A démon alapvetően a tudást alakítja energiává (pontosabban fogalmazva az információt és a hőt munkává, illetve hasznosítható energiává). Azt hiszem, hogy azért maradt meg ez annyira a fejemben, mert kézzelfogható módon szemléltette az információ értékét, különösen a rendszerezett információét.
Egy gazdag tartalommal ellátott weboldal vonzza a látogatókat az értékes információn keresztül. Egy keresőmotor hozzáadása ezt az értéket megsokszorozza. És ha nem rendelkezünk ilyen tartalomban gazdag weblappal? Inkább egy bizonyos tárgykörrel kapcsolatos linkgyűjteményünk van? A ***Webglimpse*** segítségével olyan űrlapokat helyezhetünk el az oldalunkon, amellyel a linkgyűjtemény által hivatkozott oldalakon kereshetünk. Így a témakör keresésével, a hasznos információ válogatással, linkgyűjteménnyé szervezésével töltött munkánk segíthet a látogatókat saját oldalunkra csábítani. A továbbiakban ismertetem, hogyan használjuk a ***Webglimpse***-t az oldalunkon más kiválasztott lapokon történő kereséshez, és hogyan jussunk látogatottságunknak köszönhetően gyors hirdetési bevételhez.

A Webglimpse története

A ***Webglimpse*** több összetevő keveredéséből jött létre. Egy ***Perl***-ben íródott letapogató (spider) és kezelőfelületből és a ***C*** nyelven írott ***Glimpse*** a fő kereső és indexelő algoritmusból áll.
A ***Glimpse***-t először ***Udi Manber*** és ***Sun Wu*** számítástudomány-professzorok készítették, akik az ügyes kereső-algoritmus a korábban ***Sun Wu*** által még ***Manbert*** tanítványaként kifejlesztett (és ***agrep*** néven kiadott) úgynevezett fuzzy-minták illesztésére akarták használni.
A ***Glimpse*** eredetileg teljes fájlrendszerek keresésére alkalmas eszközként íródott 1993-ban, amely minden olyan felhasználó számára hasznos, akinek kavarodott már el dokumentuma vagy egy régi e-mail üzenete a merevlemezen.
A ***Webglimpse*** pár évvel később burkolódott a ***Glimpse*** köré, alkalmazva a hatékony kereső és indexelő algoritmust a teljes weben történő keresés helyett a böngészés és az egyéni weboldalakon történő keresés összekapcsolásához. A ***Pavel Klark*** és ***Michael Smith*** végzős hallgatók által ***Webglimpse*** megismertette a weboldalak „szomszédságában" vagyis a hivatkozott oldalakon történő keresés fogalmát. Időközben ***Manber*** és egy másik hallgató ***Burra Gopal*** új szolgáltatásokat írt hozzá, tovább finomítva a ***Glimpse***-t, hogy minél jobban megfeleljen az új környezet elvárásainak.
Ezen a ponton léptem színre én. Épp kiléptem az assembly hálózati kódok javítgatását végző állásomból az ***Artisofttól***, hogy saját céget alapítsak az interneten található információk felfedezésére és rendszerezésére. Az 1996-os korai Web időszakában úgy tűnt, a ***Webglimpse*** a legígéretesebb keresőeszköz. Teljesen új volt, még eléggé sorjás, szóval amikor ***Udi Manber*** elfogadta az ajánlatomat, hogy segítsen a projektben, az első dolgom az volt, hogy újraírtam a telepítést. Egyre inkább beleástam magam a ***Webglimpse***-be, új szolgáltatások hozzáadásával, felhasználók támogatásával, és 2000 januárjában az Arizonai Egyetem

kizárólagos licencet biztosított a cégem számára az értékesítéshez. Nem éreztem úgy, hogy ha a ***Webglimps*** nyílt forráskódú lenne, akkor tovább élhetne mint életem fő célja, de meghoztam a döntést, hogy mindig adjuk ki a teljes forráskódot, és a legtöbb non-profit szervezet számára tartsuk meg ingyenesnek. Ennek eredményeképp számos felhasználó akart visszajelzést adni és foltokat küldeni, és lehetővé vált hogy ingyenes, sőt továbbadható licenceket biztosítsak mindenkinek, aki segített a projektben.

Fogjuk munkára

A sok évnyi munkának köszönhetően a ***Webglimpse*** szinte minden ***Linux*** változaton és terjesztésen fut. Az egyetlen előfeltétel csupán az 5.004-es vagy újabb változatú ***Perl***-lel ellátott webkiszolgáló, és persze parancssori hozzáférés a kiszolgálóhoz.
A minden részletre kiterjedő útmutató elérhető a ***Webglimpse*** honlapjáról (lásd az on-line erőforrásoknál), így itt csak pár tanácsot említenék.
Ha a rendszerünkön már telepítve van a ***Webglimpse***, ellenőrizzük a változatszámát. A legtöbb előretelepített példány már régi (v 1.6 1998 környékéről), ilyen esetben jó ha van jogosultságunk a frissítéshez.
A ***Webglimpse*** változatszámát a legegyszerűbb módon úgy ellenőrizhetjük, hogy lefuttatunk egy keresést, és megnézzük az eredmény oldal forrását. A változatszám egy megjegyzésben található a keresési eredmények legelején.
A cikk írásának pillanatában a ***Webglimpse 3.0 FTP***-n keresztül telepíthető változatának béta tesztje zajlik. Kipróbálhatjuk ezt, vagy választhatjuk a régebbi 2.0-s változatot, ha van ***SSH*** hozzáférésünk a kiszolgálóhoz. Az ***SSH***-n keresztüli telepítéshez először töltsük le a kipróbálható változat tar fájljait a weboldalról. Kövessük a letöltési oldal tetején hivatkozott telepítési útmutatót, amely elmondja, hogy először le kell fordítani a ***Glimpse***-t, majd telepíteni a szokásos módon:

```
./configure
make
make install
```

Aztán fel kell tenni a ***Webglimse***-t a telepítő parancsfájl segítségével

```
./wginstall
```

A parancsfájl végigmegy a szokásos kérdéseken: hová telepítse a programot, hová tegye a ***CGI*** szkripteket. Megpróbálja még elemezni az Apache fájlt (ha megtalálja), és megerősítést kér az elsődleges tartománynév és a weboldalgyökér beállításokkal kapcsolatosan. Mivel a ***Webglimpse*** beindexeli a helyi merevlemezen található fájlokat ***URL***-ekké alakítva az elérési útvonalat, ennélfogva meg kell adni, hogy hogyan írhatja át az elérési útvonalakat ***URL***-ekké. Ez egy olyan kulcsfontosságú dolog, hogy a webes adminisztrációs felület egy egész képernyőt szentel az ***URL*** >> fájl, fájl >> ***URL*** fordítások ellenőrzésére, hogy megbizonyosodjunk, ezt a részt jól beállítottuk.
Más beállítások a telepítés folyamán a biztonsághoz kapcsolódnak. Annak érdekében, hogy az archívumkezelő webfelületről futhasson, írhatóvá kell tenni az archívumkönyvtárat a webkiszolgáló számára. Ennek legbiztonságosabb módja nem az, hogy mindenki számára írhatóvá tesszük, hanem hogy csak a webszerver felhasználó számára engedjük ezt meg, ami azt a felhasználót takarja, akinek a nevében a webszerver fut. A leggyakrabban ez a ***www***, vagy a ***nobody*** felhasználó. Ezt a folyamatok listájából deríthetjük ki:

```
ps aux | grep httpd
```

Ami valami hasonló pár sort fog mutatni:

```
nobody      873  0.1  0.5 16492
➥11416 ?
    S    18:03   0:00 /usr/
➥local/apache2/bin/httpd
nobody      874  0.0  0.5 16492
➥11416 ?
    S    18:03   0:00 /usr/
➥local/apache2/bin/httpd
nobody      875  0.0  0.5 16552
➥11620 ?
    S    18:03   0:00 /usr/
➥local/apache2/bin/httpd
```

Az első oszlop a felhasználó neve, akinek a nevében a webkiszolgáló fut, ebben az esetben: ***nobody***. Most már megválaszolhatjuk a ***wginstall*** kérdését, és ha olyan felhasználó nevében futtatjuk a parancsfájlt, aki meg tudja változtatni a fájlok tulajdonosát, akkor az be is állítja azokat. Ha nem, váltsunk rendszergazdai módra a telepítés után, és változtassuk meg kézzel. Feltételezve, hogy az alapértelmezett ***/usr/local/wg2*** területre telepítettük, az alábbi parancsot kell futtatni, hogy az archívumkönyvtárat webről írhatóvá tegyük:

```
chown -R nobody
/usr/local/wg2/archives
```

Ha a telepítés véget ért, itt az ideje, hogy kijelöljük az indexelendő fájlokat és elkészítsük a kereső űrlapot. A ***Webglimpse*** szóhasználatával ez egy archívum beállítása.

Egy archívum beállítása

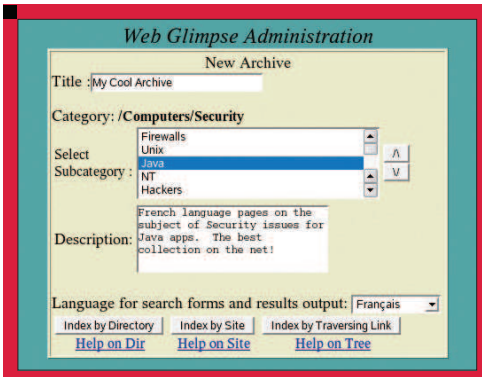
A telepítés végeztével az alábbihoz hasonló sorokat kell látnunk

```
*******************************
Done with install! You may use
 http://mycoolserver.com/
 ➥cgi-bin/wg2/wgarcmin.cgi
```

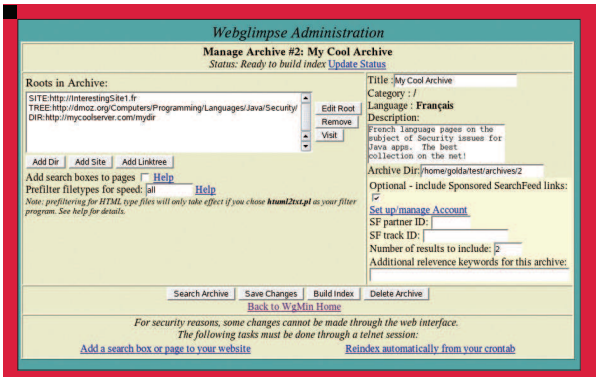
vagy:

```
/usr/local/bin/wgcmd
to configure archives at any
➥time.
(The web version currently has
➥more features)
Run wgcmd to create new archive
➥now? [Y]:
```

Ha már összebarátkoztunk a ***Webglimpse***-szel, a parancssoros eszköz nagyon kézhez álló megoldás a számos archívum kezeléséhez és újak létrehozásához. Első alkalommal, azt javaslom, használjuk a webes felületet. Úgyhogy nyomjunk ***N***-t hogy ne induljon el a ***wgcmd***, ehelyett nyissuk meg a ***wgarcadmin.cgi URL***-t a böngészőnkben, majd írjuk be a felhasználónév-jelszó párost amit a telepítés során választottunk. Ezután betöltődik az archívumkezelő, ami később majd az összes létrehozott archívumot mutatni fogja. Ha ez az első telepítés, akkor a lista üres, tehát nyomjuk meg az ***Új Archívum***



1. ábra Az Új Archívum képernyő lehetővé teszi, hogy megadjunk egy nevet és egy leírást, kijelöljük a kategóriát és kiválasszuk azt nyelvet, amin a keresési eredménynek megjelenik.



2. ábra Az Archívum szerkesztése képernyő lehetővé teszi a különböző helyekről származó weboldalak összekapcsolását egy kereshető indexállományba.

Hozzáadása (*Add New Archive*) gombot.
Most a az *Új Archívum* (*New Archive*) képernyőt kellene látnunk, ahogy az *1. ábra* mutatja.
Itt megadhatunk egy nevet egy leírást és lehetőségünk van kategóriát és nyelvet választani. A nyelv nem korlátozza a keresendő oldalakat, de meghatározza a keresési eredmény oldal sablonjának nyelvét és a karakterkódolási beállításokat. Ezek után kattintsunk az oldal alján található gombok egyikére:

3. ábra A Webglimpse sablonjai módosíthatók, hogy minél jobban illeszkedjenek a weboldalunkba. Mindössze az 'action' címke és a beviteli mezők nevei kötöttek.

- ***Index by Direcotry (Mappákon történő indexelés)***: indexeli a webkiszolgáló meghatározott könyvtárában elhelyezett fájlokat.

- ***Index by Site (Webhelyeken történő indexelés)***: indexeli a megadott weboldalt, akár a mi szerverünkön található, akár valamelyik másikon. A saját weboldalunk dinamikus tartalmainak beindexeléséhez is ezt kell használnunk.

- ***Index by Tree (Webhely bejárásán alapuló indexelés)***: indexeli a megadott kezdőoldalról hivatkozott összes oldalt az olyan beállításoknak megfelelően, mint hogy milyen mélységig és mennyi átugrással kövesse a hivatkozásokat.

Miután megadtuk az indexelendő könyvtárat vagy *URL*-t, és beállítottuk az olyan lehetőségeket, mint például az oldalak legnagyobb megengedett száma, az archívum fő vezérlőoldalára jutunk. Itt újabb oldalforrásokat adhatunk meg az indexelőnek, azaz egy archívumon belül keveredhetnek a helyi fájlok, távoli oldalak, teljes távoli webhelyek, ezekből akár több is, ha a helyzet úgy kívánja. A *2. ábra* az indulásra kész archívumot mutatja.
Amikor rákattintunk az Index felépítése (*Build Index*) gombra, beindul a weboldal-letapogató összegyűjti az oldalakat, kiszűri a *HTML* címkéket, és futtatja a *Glimpse* indexelőjét hogy létrehozza a gyors kereséshez szükséges fordított blokkszintű indexállományt.
Végezetül a a *Keresőlap vagy doboz hozzáadása az oldalhoz* (*Add a search box or page to your website*) linkre kattintva elkészíthetjük az oldalunkba illesztendő kereső űrlapot. Ez egy olyan oldalra visz, ahol háromféle keresőűrlap-forrást találunk ehhez az archívumhoz, az egyszerű keresőmezőtől a szabályos kifejezéseket is támogató, minden lehetőséget felfedő részletes keresőoldalig. A egyszerű keresőűrlap támogatja az *összes*, *bármely*, *pontos egyezés* keresési módokat, ahogyan a *3. ábra* is mutatja.
Ugyanezeket a keresőűrlapokat kapjuk, ha rákattintunk a *Keresés ebben az archívumban* (*Search this Archive*) gombra, vagy beírjuk a közvetlenül a *Webglimpse cgi*-re mutató *URL*-t (*http://mycoolserver.com/cgi-bin/wg2/webglimpse.cgi?ID=2*).
Alaphelyzetben ezek az archívum megadott nyelvén jelennek meg, de itt az angol változatot mutatja.

Tegyük kifizetődővé

Nos, van már egy kereshető archívumunk az adott témában fellelhető legtutibb hivatkozások gyűjteményére vonatkozóan. A felhasználók szerte a világon hasznot húznak az összegyűjtött tartalomból, és az oldalunkra látogatnak, hogy keressenek az igen hatékonyan indexelt tartalomban.
Ha szeretnénk, most már lehetőségünk van hirdetések megjelentetésére, hogy bevételhez jussunk, és működtetni tudjuk az oldalunkat. Ha visszamegyünk a *2. ábrán* látható *Archívum szerkesztése* képernyőre, jelöljük be az opcionális *include Sponsored SearchFeed links* feliratú jelölőnégyzetet. Ezután kattintsunk a *Fiók létrehozása/kezelése* (*Set up/manage Account*) hivatkozásra, amely a *Searchfeed* fiókkezelő oldalára mutat. On-line hirdetéskezelő és tartalomszolgáltató társaságként a Searchfeed.com olyan fizetett keresési eredményeket közvetít, amelyek relevánsnak számítanak

a felhasználók által az oldalunkon megadott keresőszó alapján. Miután elkészült a fiókunk, egyszerűen adjuk meg a partnerazonosítót és a forgalomazonosítót (***track ID***), amit a ***Searchfeed.com***-on kaptunk, és adjuk meg, hogy mennyi hirdetés jelenhet meg a keresési eredményeink elején. Nagyon egyszerű beállítani. Hogy a legtöbbet hozzuk ki a hirdetésekből, használhatjuk a ***Searchfeed*** on-line eszközkészletét, hogy követni tudjuk milyen keresőszavakat használnak a felhasználók, milyen hirdetésekre kattintanak, és mennyi jut nekünk az egyes kattintásokból.

Testreszabás

Függetlenül attól, hogy akarunk-e fizetett hivatkozásokat a keresési eredmény elejére vagy sem, nagyon valószínű, hogy rá akarjuk húzni az oldalunkra jellemző kinézetet, navigációt. Ennek elérése érdekében szerkesszük a ***wgoutput.cfg*** állományt az archívum könyvtárban. (Az archívum könyvtárának helye az Archívumkezelő képernyőn látszik.) A fájl egy ***HTML*** kódrészletet tartalmaz, amely az egyes keresési eredmények elé, közé és mögé kerül. Azt is megtehetjük, hogy a saját fejlécünket és láblécünket behívjuk (***include***), ahelyett, hogy ***HTML***-be irkálnánk bele. Néhány esetben szükségünk lehet az eredményhalmaz rangsorolásának testreszabására is. A ***Webglimpse*** más keresőmotorokkal ellentétben nem szándékozik meghatározni, hogy egy oldal hány százalékban hasznos a felhasználónak. Ehelyett betekintést enged a színfalak mögé, hogy hogyan számítja ki mennyire fontos egy találat, és ha úgy akarjuk, megadhatjuk az általunk készített, személyre szabott rangsoroló képletünket. Egyszerűen csak szerkesszük a ***wgrankhits.cfg*** fájlt, ami szintén az archívum könyvtárban található, és egy ***Perl*** kódrészletet tartalmaz, amely az alábbi változókat használja (itt magyarra fordítva):

```
# A használható változók:
#
# $N            # a kereső-
# szavak előfordulásának száma
# $LineNo       hol fordul elő
# a szüvegben a keresőszó
# $TITLE        # előfordulások
# száma a CÍM címkében
# $FILE         # előfordulások
# száma a fájl elérési útjában
# $Days         dátum (hány
# napja készült a fájl)
# $META         keresőszavak
# összes előfordulása bármilyen
# META címkében
# $LinkPop      hivatkozás
➥népszerűsége az oldalon (hány
➥másik oldal hivatkozik rá)
# %MetaHash     A szavak
# előfordulásának hash értéke
#  az egyes meta címkékben, a
#  NAME= paraméter alapján
#  indexelve
# $LinkString  A hivatkozás
# akutális URL-je
# A következő komment nélküli
# sorok
# az aktuális rangsoroló
# képletet alkotják
# Ez az alapértelmezett
# rangsorolás, magas
# súlyt ad a címben előforduló
# kulcsszavaknak,
# súlyozza az általános
# találatokat, népszerűséget
# és az időbeniséget.
$TITLE * 10 + $N + $LinkPop +
➥5/($Days + 1)
```

Hibaelhárítás

Mostanra van némi sejtésünk arról, hogy mi az erőssége, illetve mi a gyengéje a ***Webglimpse***-nek. Egy csomó ügyes sajátosság, amelyeket a felhasználó közvetlenül testre szabhat, és egy csomó ügyes sajátosság valamiféle alkalomszerű módszerrel összekeverve. A ***Webglimpse*** – a nézőponttól függően – meg van áldva, vagy épp átkozva a jó sok trükközéssel, ami lehetővé teszi a sok-sok különálló feladatot. A következő változat, amely a cikk írásának időpontjában javában készül, a tervek szerint egyszerűbben telepíthető és kezelhető lesz, nem beszélve a az ***FTP***-n keresztüli telepítési lehetőségről, azon felhasználók számára, akik nem rendelkeznek héj hozzáféréssel a kiszolgálón. Bárhogy is lesz, a leggyakoribb problémák, amikbe a jelenlegi változat használata során belefuthatunk, a következők.

1. ***Jogosultsági problémák:*** Akkor fordul elő, ha időnként újraindexeljük az archívumot a webfelületről, és néha a héjból vagy ***crontab***-ból történő újraindexelés esetén is. Bármely archívum újraindexelhető, ha a webfelületen az 'Index építése' gombra kattintunk, vagy ha kiadjuk héjból a `./wgreindex` parancsot az archívumkönyvtárban. A legjobban tesszük, ha ragaszkodunk a kiválasztott újraindexelési módhoz, és az archívumot annak a felhasználónak a tulajdonába adjuk, aki a szkriptet futtatni fogja.

2. ***URL/fájl átalakítási problémák:*** Akkor fordulhat elő, ha a `DocumentRoot` változó nincs helyesen beállítva. A webfelület főoldalán található ***Elérési útvonalak átalakításának tesztelése*** gombra kattintva ellenőrizhetjük, hogy az egyes fájlok milyen URL-lé alakulnak át, és fordítva. Minden, a helyi és távoli tartományokra alkalmazható beállítás a ***/usr/local/wg2/archives/wgsites.conf*** fájlban van tárolva. Közvetlenül, és az Archívumkezelő képernyő ***Tartomány szerkesztése*** gombra kattintva egyaránt szerkeszthetjük a ***wgsites.conf*** fájlt.

Még több hibaelhárítási tipp található ***Dokumentációk*** és hogyanok oldalon (lásd az on-line erőforrásoknál).

Köszönetnyilvánítás

A szerző köszönetét fejezi ki ***Udi Manbernek***, amiért rábízta ezeket a nagyszerű kreálmányokat! Még mindig úgy próbálom kezelni, ahogyan megérdemlik. Köszönet továbbá ***Sun Wunak***, ***Burra Gopalnak***, ***Michael Smith***-nek, és ***Pavel Klarknak***, a ***Webglipse*** és ***Glimpse*** társfejlesztőinek, és az összes felhasználónak, aki hibajelentéseket, foltokat, fordításokat és javaslatokat küldött az évek során.

Linux Journal 2006., 147. szám

Golda Valez az Arizonai Tucsonban él, programozó és édesanya. Ő a Webglimpse vezető fejlesztője 1997 óta. Emellett tulajdonosa és alapítója az 1995-ben alapított Internet Workshop nevű tárhelyszolgáltató és tanácsadó cégnek.

KAPCSOLÓDÓ CÍMEK

➲ www.linuxjournal.com/article/9021