Projekt z vyhledávání znalostí v databázích

Web Usage Minig

Miroslav Křipač, Jakub Novák, Helena Vildová

Abstrakt

Cílem projektu bylo prověřit možnosti jednoho z dostupných nástrojů pro web usage mining na reálných datech webového serveru Fakulty informatiky. Projekt se zabývá instalací, zpracováním dat a interpretací výsledků, které nad nimi byly dosaženy.

Úvod

Pro účely projektu byl vybrán nástroj WUM: A Web Utilization Miner (domovská stránka projektu je na http://wum.wiwi.hu−berlin.de/, který umožňuje načíst standardní logy webového (HTTP) server Apache na základě nichž je schopen analyzovat chování uživatelů při přístupu k webovému serveru. Obecně jej lze použít pro hledání častých sekvencí v jakémkoliv zdroji dat.

Samotný program je napsaný v jazyce Java a vyžaduje tedy instalaci javovského interpretu. Vzhledem k tomu, že obsahuje některé platformově závislé prvky, je distribuován ve dvou verzích −− pro Unix/Linux a pro Windows 95/98/NT/2000. Dále se budeme zabývat pouze verzí pro Unix/Linux, která se nám ovšem spolehlivě ve fakultním prostředí osvědčila pouze pod systémem Linux. Instalací ve fakultním prostředí bude věnována další část. Provedli jsme rovněž lokální instalaci pro Windows, která proběhla bez větších problémů.

Druhá část našeho projektu je věnována popisu funkcí, které WUM nabízí. Z nich jsme dále vybrali ty, o kterých se domníváme, že mají význam pro použití při analýze logů fakultního webového serveru. Tato část však může sloužit jako vodítko pro rozhodování o použití tohoto nástroje pro jiné účely.

Zbývající časti se věnují samotné analýze konkrétních dat fakultního webového serveru. Vzhledem k jejich celkovému rozsahu jsme se rozhodli vybrat pouze ta data, která souvisí s oficiální prezentací fakulty. Poměrněvelká část naší práce se tedy zabývala předzpracováním a čištěnímzpracovávaných dat. Na závěr uvádíme konkrétní výsledky analýzy přístupu k fakultnímu webovému serveru na základě nichž se snažíme navrhnout postupy pro zjištěnímožných vylepšení usnadňující uživatelům orientaci a dostupnost potřebných informací.

Instalace programu

Jak bylo naznačeno v úvodu, budeme se dále zabývat pouze verzí pro Linux a to pro prostředí Fakulty informatiky. (Postup instalace na lokální stanici nezávislou na počítačové síti fakulty je však snadno odvoditelný). Vzhledem k tomu, že webový server fakulty je rovněž hlavním studentským serverem, směřovala naše snaha k instalaci programu na tento server pod operační systém IRIX 6.5. Ačkoliv se jedná o program v jazyce Java, nebyli jsme jej schopni korektně interpretovat ani doporučeným interpretem pravděpodobněz důvodů nekompatibility některých použitých tříd. Vzhledem k tomu, že server Aisa nabízí sdílení programových modulů pouze pro stroje s platformou IRIX, pro nainstalování jsme zvolili místo na svazku domovských adresářů (/home), který sdílí jak linuxové stroje, tak stroj Aisa a tedy webový server.

Samotná instalace nejprve vyžaduje extrakci TAR archivu, například příkazem:

```
cd /home/xkripac/WUM
gtar xfvz WUM.v70.tar.gz
```

vzniklý adresář /home/xkripac/WUM/WUM.v70 pak obsahuje všechny potřebné třídy (ve archívu JAR), dokumentaci, příklady použití a příklady spouštěcích skriptu v podadresáři /home/xkripac/WUM/WUM.v70/bin. Zde je rovněž skript pro interpret příkazů (BASH). V něm (na řádcích 28 a 29) jsou umístěny konfigurační proměnnéprostředí, které je třeba nastavit na hodnoty:

```
JAVA_HOME=/home/xkripac/WUM/jre1.2.2
WUM_HOME=/home/xkripac/WUM/WUM.v70
```

Poté je nutné do adresáře /home/xkripac/WUM/jre1.2.2 nainstalovat balík The Java Development Kit verze 1.2.2, který lze obdržet na stránkách společnosti Sun Microsystems. Podle přiložené dokumentace je možné použít i novější verzi javovského interpretu. Podle našich zkušeností je však lépe pro Web Utitlization Miner nainstalovat právě tuto verzi. Instalace je přesně popsaná v dokumentaci a její popis je nad rámec tohoto projektu. Vzhledem k tomu, že jsme při ní nenarazili na žádný problém, dovolíme si její popis vypustit.

Nyní by mělo být vše potřebné nainstalováno. Pro spuštění Web Utilization Mineru v grafickém prostředí umožňujícím snadné uživatelské nastavení jednotlivých funkcí včetně startu analýzy slouží příkaz:

```
$ /home/xkripac/WUM/WUM.v70/bin/wumgui
```

Popis funkcí programu

Program po spuštění umožňuje otvírat (zavírat), tvořit a mazat dolovací databáze. To lze pomocí nabídky File a v ní položek Create Mining Base, Open Mining Base, Close Mining Base a Delete Mining Base. Pokud žádnou databázi nemáme, musíme ji vytvořit. Když otevřeme nově vytvořenou databázi, je potřeba do ní importovat data z logu, který chceme analyzovat. To provedeme pomocí File -> Import Log File, při importování lze zadávat různá omezovací kritéria, co při importu vynechat, co naopak zahrnout apod. (např. pokud nemáme předfiltrovaný log, můžeme chtít vynechat ze zpracování obrázky, tedy soubory s příponami .jpg, .gif a další). V databázích s importovanými daty lze tato data smazat a znovu importovat pomocí adekvátních položek v nabídce File.

Aby bylo možné po importu dat s nimi začít pracovat, je nutné vytvořit pomocí nabídky Analyzer tzv. Visitor's Sessions a Aggregated Log (Database). První zpracuje data do podoby popisů jednotlivých návštěv na stránkách (tzn. jak se jeden každý návštěvních pohyboval po stránkách, návštěvníci jsou odlišeni IP adresou nebo jejím doménovým jménem a délka trvání jedné návštěvy je implicitně nastavena na 30 minut − lze změnit), druhá data setřídí do struktury stromu podle toho, jaká byla podle logu hierarchie odkazování na stránkách (nad touto strukturou pak zadáváme MINT dotazy).

Tato dvě předzpracování dat do databáze umožní už opravdovou práci s daty. V nabídce Reports si můžeme nechat vygenerovat souhrnný statistický přehled o datech pomocí položky Comprehensive Summary. V takto vzniklém souboru WumReport.html najdeme statistiky o nejnavštěvovanějších a nejméně navštěvovaných stránkách a adresářích, statistiky používaných webových prohlížečů, statistiky nejčastějších návštěvníků stránek a další informace, které použijeme jako výchozí, až se budeme rozhodovat, na které sekvence dat se zaměříme při zadávání MINT dotazů nad databází. Jak totiž sami autoři v materiálech přiložených k WUMu uvádějí, WUM není expertní systém, rozhodující za uživatele, které sekvence z databáze považovat za zajímavé, ale pouze pomůcka uživatele, aby se tyto sekvence lépe hledaly. Pomocí položky Web Site Pages Report vygenerujeme html stránku WebSitePagesReport.html, která obsahuje statistické informace o počtech přístupů na jednotlivé stránky a také relativní četnosti přístupů, uvedené v procentech.

Nabídka Visualizer dovoluje zobrazit databázové i grafické podoby souborů, vytvořených při předzpracování dat (Visitor’s Sessions a Aggregated Log − lze zobrazit také Observations, což je alternativní zobrazení Visitor’s Sessions se zřetelem na posloupnost návštěvníkových kroků v čase a ne jenom na existenci posloupností těchto kroků) a ostatní soubory databáze. Grafické podoby nejsou moc zdařilé (hlavně u Visitor’s Sessions a Observations není vůbec vyřešeno rozložení grafu po ploše a proto je znázornění každé trochu složitější návštěvy naprosto nepoužitelné k prohlížení).

V nabídce Tools lze exportovat dolovací databázi, případně ji nechat přeorganizovat, aby práce s ní byla co nejefektivnější. Dále zde lze upravovat nastavení samotného programu WUM a jeho GUI.

Nejdůležitější nabídkou je nabídka Queries, pomocí níž zadáváme dotazy v jazyce MINT nad daty v databázi. Můžeme si tvořit dávkové soubory těchto dotazů, ukládat je a nahrávat, nebo je zadávat přímo a ukládat si pouze výsledky až po zpracování dotazu. V nabídce Help a v přiložených materiálech k programu WUM lze najít stručný popis syntaxe MINTu a pár příkladů jeho použití.

Předzpracování

Přestože samotný program WUM umožňuje při importu logů zadat charakteristiku stránek, které se nemají do analýzy zahrnout, při testování jsme se rozhodli, že přístup k webovému serveru fakulty je natolik specifický, že je vhodné z důvodu efektivity zpracování nejprve data předzpracovat podle specifických požadavků fakulty. Toto předzpracování provádíme programem fiwum−filter.pl, který je napsán v jazyce Perl a lze jej tedy interpretovat jak na stroji Aisa (s uloženými webovými logy), tak na strojích s operačním systémem Linux.

Filtr fiwum−filter pracuje ve třech režimech. První režim, spouští se bez parametrů, slouží pro analýzu aktuálního logu webového serveru, umístěného v souboru /usr/local/etc/httpd/logs/access_log. Tento soubor v současné době uchovává logy od noci ze soboty na neděli, kdy jsou vždy všechny starší záznamy přesunuty a zkomprimovány do souboru /usr/local/etc/httpd/logs/access_log.0.gz, z něj přesunuty do /usr/local/etc/httpd/logs/access_log.1.gz a tak dále. Webový server fakulty v současné době uchovává až jeden měsíc záznamů o přístupu. Druhý režim tedy umožňuje jako parametr předat filtru jiný soubor, který bude sloužit jako vstup. Nakonec třetí možnost umožňuje načítat záznamy přímo ze standardního vstupu. Tam lze například zasílat výsledky dekomprimace starších archivů a podobně.

Předzpracovávací filtr provádí několik základních činností:

1. Vzhledem k tomu, že na stránky poměrně v hojné míře přistupují různé automatické roboty −− vyhledávače, ukázalo se jako nezbytné tyto přístupy do analýzy neuvažovat, neboť se jedná o strojový přístup, který nic nevypovídá o logickém postupu uživatele přistupujících na server. Seznam těchto strojů je umístěn v souboru /home/xkripac/WUM/blacklist.hosts. Jedná se o textový soubor obsahující na každém řádku jedno jméno stroje, který se bude ignorovat.

2. Největší zátěží pro webový server fakulty jsou osobní stránky jejich uživatelů. Vzhledem k tomu, že obsah (a rozvržení) těchto stránek je plně na jejich autorech, rozhodli jsme se je do analýzy nezahrnout a ušetřit tak spoustu výsledků, které by pro případnou optimalizaci oficiálních stránek neměly velký význam. Filtr proto nezahrnuje do další analýzy jak stránky studentů (pod adresami http://www.fi.muni.cz/~login/,tak stránky učitelů fyzicky umístěné na jiném serveru a odkazované adresou http://www.fi.muni.cz/usr/login/.

3. Podobně jako v případě strojů, jejichž přístup nechceme analyzovat, existují na serveru také stránky, které nemají vypovídající hodnotu (např. robots.txt což je soubor obsahující právě informace pro různé vyhledávače). Pro základní analýzu je rovněž možné neuvažovat přístup k obrázkům a speciálním skriptům (jako jsou počítadla stránek). Stránky jejichž přístup nechceme analyzovat jsou umístěny analogicky v souboru /home/xkripac/WUM/blacklist.sites.

4. Jako poslední provádí předzpracovávací filtr sjednocování jednotlivých kódování. Webový server fakulty umožňuje přistupovat na jednu stránku pod různými znakovými (a jazykovými) stránkami. Jedna stránka se tak při přístupu z klientů s různým kódováním loguje různě. Předzpracovávací filtr na základě seznamu možných kódování v souboru /home/xkripac/WUM/charsets tyto různé přístupy sjednocuje. Seznam možných kódování je uveden v konfiguraci webového serveru.

Výsledkem předzpracovávacího skriptu je proud relevantních záznamů o přístupu, který filtr posílá na standardní výstup. Lze je jej proto snadno přesměrovat a případně postupně doplňovat o nová data. Tento výstup lze přímo načíst pomocí funkce import log v menu File programu WUM.

Příklad výstupu filtru:

```
Filtr zpracovavajici logy weboveho serveru FI MU

Nacteno 2 stroju pro ignorovani.
Nacteno 8 stranek pro ignorovani.
Nacteno 36 kompatibilnich znakovych sad.
Zpracovavam: /usr/local/etc/httpd/logs/access_log
         . = 1000 zpracovanych radek
         radek obsahuje 50 tecek
..................................................
..................................................
..................................................
..................................................
..................................................
..................................................
..................................................
..................................................
..................................................
..................................................
..................................................
..................................................
..................................................
..................................................
..................................................
..................................................
..................................................
..................................................
..................................................
..................................................
..................................................
......................................

Ignorovano 51372 pristupu ze stroje inktomisearch.com.
Ignorovano 9897 pristupu ze stroje googlebot.com.
Ignorovano 6685 pristupu na stranku .png.
Ignorovano 90096 pristupu na stranku .gif .
Ignorovano 19528 pristupu na stranku cgi-bin/nph-count.
Ignorovano 3530 pristupu na stranku robots.txt.
Ignorovano 59 pristupu na stranku /ftp/.
Ignorovano 8915 pristupu na stranku /pics/.
Ignorovano 2055 pristupu na stranku favicon.ico.
Ignorovano 6735 pristupu na stranku .jpg .
Ignorovano 821433 pristupu k uzivatelskym strankam.
Celkem zpracovano 1088278 stranek.
Cas behu skriptu: 203s.
```

Výsledky analýzy

Analýzu jsme původně prováděli nad všemi dostupnými daty, tedy nad třemi zkomprimovanými soubory a jedním souborem s aktuálními logy pomocí příkazů:

```
        (cd /usr/local/etc/httpd/logs/; \
```

```
              gzcat ‘ls -r access_log*.gz‘) | \
             ./fiwum-filter.pl - > access_log-full
    ./fiwum-filter.pl >> access_log-full
```

Celkový čas potřebný pro předzpracování byl asi 13 minut, přičemž výsledný soubor obsahoval 265 306 řádků a zabíral asi 25MB. Jeho načtení do interních struktur programu WUM (pomocí příkazu Import Log File v menu File) trvalo na stroji nymfe27 s procesorem MD Athlon(tm) XP 1600+ a 512MB paměti přibližně dvě hodiny. Po tomto načtení trvala analýza uživatelských návštěv (pomocí funkce Create visitor’s sessions") přibližně šest minut. Přestože byl původní počet analyzovaných přístupů předzpracovávacím filtrem snížen asi na 10% původní velikosti, nepodařilo se nám zřejmě kvůli jeho velikosti vytvořit agregovaný log, potřebný pro výslednou analýzu. Program spadl s blíže neurčenou chybou javovského interpretu, která byla zřejmě způsobena chynou prácí z pamětí programu.

Na základě tohoto jsme se rozhodli provádět analýzu po menších částech. Vhodný počet dat, které program ještě zvládá zpracovat, odpovídá podle našich testů jednomu týdnu záznamů webového serveru. Další analýzu jsme se proto nakonec rozhodli provádět nad záznamy o přístupu mezi nedělí 8. 12. 2002 a sobotou 14. 12. 2002. Předzpracování takovéhoto množství dat trvá přibližně tři minuty, načtení do interních struktur půl hodiny a jednotlivá analýza okolo pěti minut. Ani toto zmenšené množství dat však nedovoluje spustit funkci zobrazení agregovaného logu (stromu všech průchodů), které padá podobně jako pro větší množství dat samotné vytvoření agregovaného logu.

Základním výstupem, který program WUM generuje (a zobrazuje) jsou standardní statistika přístupů na jednotlivé stránky (dostupná v souboru WebSitePagesReport.html) a statistika celkového přístupu k serveru (dostupná v souboru WumReport.html). Tento druh statistik však je standardní součástí běžných analyzátorů webového přístupu (jako je například Webalizer , které umožňují navíc rozšířené možnosti grafů a lepší strukturu.

Hlavní funkcí sloužící pro analýzu webového přístupu je možnost kladení ad hoc dotazů v jazyce MINT, které umožňují (oprotí zmíněným běžným komerčním analyzátorům webového přístupu) na základě předchozí znalosti struktury serveru zobrazit konkrétní informace o jednotlivých sekvencí uživatelských požadavků. Praktickým zkoumáním této možnosti jsme však zjistili, že interaktivní kladení příkazů (tak, aby uživatel nemusel čekat hodiny na jejich výsledek), je možné pouze u jednoduchých dotazů a to pouze pro velice omezenou množinu dat. Byli jsme proto nuceni opět zmenšit testovací množinu na (předzpracované) záznamy pouze jednoho dne. Aby nedocházelo ke zkreslením, nezahrnuli jsme záznamy z noci. Na takto zmenšené množině dat, se již dají vyhledávat jednoduché sekvence (tedy takové, u kterých jednotlivé uzly grafu navazují na sebe přímo) interaktivně, neboť trvají řádově minuty.

Důsledkem toho je, že program WUM není použitelný pro celkové rozsáhlé analýzy celé prezentace Fakulty informatiky a je potřeba zaměřit se pouze na jednotlivé jednoduché prvky. Jako příklad jsme zvolili umístění odkazů týkajících se přijímacího řízení na hlavní stránku. Tento odkaz byl přidán nedávno a z analýz, které jsme provedli nad různými intervaly vyplývá, že se jedná o v této době jeden z nejnavštěvovanějších zdrojů oficiálních stránek. Na základě celkové statistiky přístupů (WumReport.html) jsme zjistili, že sedmým nejnavštěvovanějším odkazem je dokument prijimacky_kriteria_2003.pdf, který je součástí informací o přijímacím řízení. Pomocí dotazu

```
select t
from node as a b, template a b as t
where a.url = "/"
and b.url = "/studijni/std_dokumenty/prijimacky_kriteria_2003.pdf"
```

jsme zjistili, že tento dokument nebyl ani jednou dotazován z hlavní stránky přímo, naopak dotazem

```
select t
from node as a b c, template a b c as t
where a.url = "/"
and b.url = "/studijni/std_studium/st_prijimaci_rizeni.shtml"
and c.url = "/studijni/std_dokumenty/prijimacky_kriteria_2003.pdf"
```

zjistíme, že 13% všech přístupů k tomuto dokumentu pouze za jeden den bylo přímo přes stránku týkající se přijímacího řízení bez dalších mezikroků. Vzhledem k tomu, že tento dokument byl navštěvován 1,6−krát více než samotná stránka o přijímacím řízení, lze uvažovat, zda dokument neumístit přímo na hlavní stránku. Tam je naopak (zřejmě ne na základě analýzy přístupu) umístěn odkaz na dokument "Zpráva o výsledcích přijímacího řízení v r. 2002", který nebyl dotazován ani jednou a lze jej tedy po této krátké analýze nahradit dokumentem prijimacky_kriteria_2003.pdf.

Podobnými kroky jsme provedli ještě další analýzy, z nichž tato nejlépe vystihuje funkční postup autora serveru při analýze.

Pro rozhodování, jaké odkazy z dané tématiky umístit na nejlépe dostupná místa, jsou tyto dílčí dotazy a analýzy vhodné. Pro zpřesnění výsledků by bylo rovněž vhodné v předzpracovávacím filtru provést výběr pouze těch strojů, které danou zkoumanou stránku navštívili. To by umožnilo (při současném zachování předzpracovávacích pravidel) snížení mohutnosti agregovaného grafu což by vedlo k prohledávání pouze nad relevantními daty a tedy k celkovému zvýšení efektivity.

Naopak, zcela oproti očekávání, nelze nástroj WUM použít pro obecnou analýzu webového serveru fakulty. Ani po radikálním předzpracování se nedaří snížit počet dat tak, aby program vracel výsledky v reálném čase. Dotazy typu "Jaké nejčastější kroky jsou prováděny při přístupu k dané stránce", které by umožnily lépe přizpůsobit prezentaci jako celek uživateli, nejsou v současné verzi možné ať už reálně z důvodu dlouhého zpracování či pádů programu či pouze z důvodu nedostatku této verze (v současné době nelze pro dotazy použít agregované funkce). Případný autor webové prezentace tedy může tento nástroj použít pouze pro analýzu konkrétní stránky (stránek) ve vztahu k ostatním.

Závěr

Projekt Web Usage Mining měl za úkol prozkoumat možnosti programu WUM v konkrétních podmínkách Fakulty informatiky. Jeho součastí je popis instalace samotného programu, výpis možných funkcí, které lze využít pro analýzu webového přístupu, diskuse možných předzpracování dat a výsledky jejich analýzy. Vzhledem k platformě Java, na které je program vystavěn, jej lze podle našich poznatků reálně využít pouze pro analýzu pomocí jednocuchých cílených dotazů na základě přechozí znalosti struktury serveru a to pouze pro značně zmenšený objem dat. Pro obecné dotazy nad větším počtem údajů (více než dva dny) webového serveru je program (zřejmě špatnou prací z pamětí) nepoužitelný.