VŠB – Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra informatiky

Nástroj pro monitorování chování uživatelů webových aplikací

A tool for monitoring web application user behaviour

2011 Michal Hantl

Souhlasím se zveřejněním této bakalářské práce dle požadavků čl. 26, odst. 9 *Studijního a zkušebního řádu pro studium v bakalářských programech VŠB-TU Ostrava*.

V Ostravě 20. dubna 2011 .....

Prohlašuji, že jsem tuto bakalářskou práci vypracovala samostatně. Uvedla jsem všechny literární prameny a publikace, ze kterých jsem čerpala.

V Ostravě 20. dubna 2011 .....

Rád bych na tomto místě poděkoval svým rodičům za podporu během celého studia, panu Ing. Michalovi Radeckému, za vedení během vypracování diplomové práce a své přítelkyni Míše.

Abstrakt

Cílem práce je vytvoření nástroje, jakožto podpory pro aplikaci a navržení metody získávání informací o chování uživatele webové aplikace. Nebude se jednat o klasický přístup založený na sběru statistických dat anonymních návštěvníků, ale o využití znalosti o konkrétním uživateli, jeho chování, využívání funkcí a služeb webové aplikace.

Výsledkem praktické části je webová aplikace ...

Klíčová slova: webová analytika, chování uživatelů, webová aplikace, diplomová práce

Abstract

The goal of this thesis is to create a tool to aid web application development and to create a method for gaining knowledge of the web application users' behaviour. It is not the case of usual anonymized data gathering, but make us of the knowledge of concrete user, his behaviour and his usage of web app's functions and services.

Keywords: web analytics, user behaviour, web application, master thesis

Seznam použitých zkratek a symbolů

GUI	–	Grafické uživatelské rozhraní (Graphical User Interface)
HTML	–	Jazyk pro vytváření webových stránek (HyperText Markup Language)

Obsah

1 Úvod (3 stránky) **4**
1.1 Hostorie webové analytiky 4

2 Sběr a interpretace dat (15 stránek) **7**
2.1 Analýza dat z logovacích souborů 7
2.2 Informace získané pomocí HTTP protokolu 8
2.3 Události v prohlížeči 11
2.4 Informace získané pomocí JavaScriptu 11
2.5 Derivované informace 11
2.6 Interpretace a Vizualizace dat 11
2.7 Clickstream analýza 11
2.8 Měření konverze 11

3 Návrh (15 stránek s diagramy) **12**
3.1 Hypotéza 12
3.2 Sběr dat 12
3.3 Interpretace dat 12
3.4 Etická stránka sběru dat 12

4 Implementace (20 stránek i s grafy) **13**
4.1 Popis technického řešení (10 Stran) 13
4.2 Aplikace v reálném provozu (10 stran i s grafy) 13
4.3 Navrhnuté změny na základě získaných dat (2 stránky) 13

5 Zhodnocení (2 stránky) **14**

6 Závěr (1 stránka) **15**

7 Reference **16**

Seznam obrázků

1 Analog - denní návštěvnost 5

Seznam výpisů zdrojového kódu

1 Ukázka záznamu z logovacího souboru ve formátu RFC931 4
2 Formát logovacího souboru dle RFC931 7
3 Ukázka HTTP požadavku 8

1 Úvod (3 stránky)

Tato diplomová práce se zabývá webovou analytikou. Úvodem popisuje jaký problém webová analytika řeší, nastiňuje historii měření na webu a současné nejčastější využití nástrojů pro webovou analytiku.

V první kapitole popisuje nejčastěji používané techniky měření a vizualizace dat, vysvětluje motivace za jednotlivými technikami a případy užití.

Druhá kapitola se zabývá návrhem nové techniky měření, která se zajímá o to, které funkce webové aplikace uživatelé používají. Popisuje jakým způsobem budou data získána a celý systém jako celek.

Třetí kapitola "Implementace"je zaměřena na použité tehcnologie a konkrétní procesy sběru a analýzy dat.

Čtvrtá kapitola následuje aplikací vzniklého nástroje na existující webovou aplikaci. Cílem je ověřit, že vyvinutý nástroj přináší očekávané výsledky.

V poslední kapitole dochází ke zhodnocení zda metoda měření i nástroj plní očekávání, jaký je jejich přínos pro jejich uživatele v praxi a porovnání s podobnými nástroji.

1.1 Hostorie webové analytiky

1990 - Zrození WWW stránek

Na začátku devadesátých let došlo ke zrození WWW stránek[reference]. Uživatelé tehdy prohlíželi statické stránky a pokaždé, když si nějakou prohlédli, vznikl záznam v logovacím souboru, takzvaný "hit". Počet hitů se stal ukazatelem úspěšnosti webový stránek.

1994 - Mosaic: První masově úspěšný grafický prohlížeč

V roce devadesát čtyři vznikl grafický webový prohlížeč s názvem Mosaic[reference]. Díky jeho snadné instalaci a srozumitelnému uživatelskému rozhraní se web otevřel široké veřejnosti. Tento prohlížeč byl později přejmenován na Netscape a v roce 1995 ho používalo 80% uživatelů internetu.

1995 - Analog: Analýza logovacích souborů

V roce devadesát pět znikl Analog - nástroj pro analýzu logovacích souborů. Jeho autor Stephen Turner ho poskytoval zdarma jako freeware pro několik platform. Jednalo se o první sofistikovaný nástroj pro analýzu a zobrazení návštěvnosti webových stránek.

```
10.20.30.40 - - [26/Apr/2000:00:23:48 -0400] "GET /index.html HTTP/1.0" 200 6248 "http://www.
    jafsoft.com/asctortf/" "Mozilla/4.05 (Macintosh; I; PPC)"
10.20.30.40 - - [26/Apr/2000:00:23:48 -0400] "GET /background.gif HTTP/1.0" 200 4005 "http://
    www.example.org/" "Mozilla/4.05 (Macintosh; I; PPC)"
```

Výpis 1: Ukázka záznamu z logovacího souboru ve formátu RFC931

Logovací soubor obsahuje jden řádek pro každou návštěvu stránky, nebo souboru na serveru. Nástroj pro analýzu logovacích souborů z toho pak dokáže zjistit kolik měl

server návštěv každý den, z kolika unikátních IP adres, které stránky jsou nejnavštěvovanější a kolik bylo přenesených dat.

The data collection became unusable with the advent of search engines and their robots, proxies servers to surf anonymously, allocation of dynamic IP addresses by ISPs and cached content techniques. All these developments have rendered the use of log files inappropriate to analyze user behavior. The data contained in log files were indeed biased and unique visitor identification almost impossible.

Each unit (▬) represents 30 requests for pages or part thereof.

day	reqs	pages
Sun	3036246	1261
Mon	2012965	626
Tue	2490837	788
Wed	3193776	880
Thu	3188571	941
Fri	3589076	1076
Sat	2749421	1007

Obrázek 1: Analog - denní návštěvnost

1996 - Měření pomocí kódu na stránce

V roce 1996 byl v obou nejpoužívanějších prohlížečích[1] k dispozici JavaScript.

Tvůrce webových stránek na každou stránku umístil kód, který mu vygeneroval poskytovatel měřícího nástroje.

Tento postup se zásadně liší od analýzy logových souborů. Místo statické analýzy zaznamenaných dat na straně serveru tento přístup data dynamicky sbírá na straně klienta.

Tento průlom umožnil použtití třetích stran pro sběr i vyhodnocení dat a tak dal vzniknout prvním online nástrojům pro webovou analytiku. V různých obměnách se tato technika používá dodnes.

2005 - Moderní webová analytika pro každého

V roce 2005 koupila společnost Google analytický systém Urchin on Demand a ještě v témž roce ho poskytla zdarma široké veřejnosti jako webovou aplikaci pod názvem Google Aanlytics.

Pro obrovský zájem byl tento produkt zpřístupněn pouze omezenému počtu uživatelů. Od října roku 2006 byly znovu otevřeny volné registrace a dodnes se jedná o nejpoužívanější systém pro webovou analytiku.

[1] V roce 1996 měl Netscape Navigator zhruba 80% podíl a Internet Explorer 14%

2011 - Sociální média a zařízení

S nástupem sociálních médií a chytrých mobilních zařízení dochází k nárůstu času, který uživatelé tráví na internetu. Poměrně k tomuto nárůstu se zvyšují investice do internetových stránek a webových aplikací.

Na jedné straně je snaha dostat nové uživatele na svou stránku, to se dociluje například pomocí optimalizace pro vyhledávače, reklamních bannerů, nebo PPC reklamy[2]. V tomto případě slouží webová analytika k měření efektivity jednotlivých kampaní a k jejich následné optimalizaci.

Na druhé je snaha pracovat s uživateli, které už webová stránka, nebo aplikace má. Zde je snaha identifikovat problémy a optimalizavat uživatelskou zkušenost. V obou případech je to webová analytika, která nám umožňuje získat přehled, optimalzovat a vyhodnotit návratnost investic.

[2]PPC - Pay per click, platba se za proklik (cena prokliku je určena pomocí aukce).

2 Sběr a interpretace dat (15 stránek)

Podstatou webové analytiky je sběr a interpretace dat. Jak již bylo řečeno, sběr dat se za dobu existence webu vyvinul z pasivní formy analýzy logových souborů na serveru do dnešní - aktivního sběru dat na straně uživatele třetí stranou.

Tato kapitola popisuje popisuje používané způsoby sběru dat s důrazem na to, jaké informace z nich získáváme, jaké další informace z nich můžeme vyčíst a jaké způsoby vizualizace se běžně vyskytují.

2.1 Analýza dat z logovacích souborů

Wbové servery zaznamenávají zobrazení stránek do takzvaných logovacích souborů. Jedná se o soubory, ve kterých každý řádek představuje záznam o jednom zobrazení stránky, nebo například obrázku.

Když na začátku devadesátých let vznikl web, provozovatelé serverů si uvědomili, že tyto soubory umožňují získat informace o popularitě jejich stránek a tak začali měřit počet zobrazení, takzvaných "hitů". V té době se vyskytovaly většinou pouze dlouhé stránky bez obrázků nebo odkazů, takže počet zobrazení postačoval potřebám provozovatelů.

```
127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200 2326
```

Výpis 2: Formát logovacího souboru dle RFC931

`127.0.0.1`	IP adresa návštěvníka. Počet unikátních IP adres byl původně používán jako hrubý odhad počtu uživatelů.
`-`	pomlčka
`frank`	Uživatelské jméno. Tento údaj je přítomen v případě se stránka používá HTTP autentifikaci. To platí pro stránky s omezením přístupu pro uzavřenou skupinu lidí a tudíž u náhodného návštěvníka stránky tento údaj nenajdeme. V takovém případě se místo uživatelského jména zaznamená pomlčka.
`10/Oct/2000`	Datum návštěvy.
`13:55:36`	Čas návštěvy, vztahuje se k časové zóně na serveru.
`-0700`	Časová zóna.
`GET /apache_pb.gif`	Požadavek a verze HTTP protokolu

`HTTP/1.0`	Požadavek HTTP protokolu. GET je základním požadavkem, který signalizuje, že uživatel požaduje nějaký obsah. Další používaný požadavek je například POST, který se používá pro odeslání formuláře na server.
`200`	Kód odpovědi. Kód 200 znamená vše v pořádku. Obvyklé a pro webovou analytiku také zajímavé jsou 404 - stránka nenalezena, 500 - problém na serveru, nebo 303 - přesměrování stránky.
`2326`	Délka odpovědi v bajtech. Sledování této veličiny spolu s počtem návštěv umožňuje identifikovat, který obsah nejvíce zatěžuje internetové připojení daného serveru.

Jak web rostl, tak i jeho složitost a z jednoduchých stránek bez obrázků se staly různě provázené stránky s odkazy a obrázky. S vzůrstající složitostí internetových stránek se i nástroje pro analýzu logovacích souborů stávaly sofistikovanější a začaly na sebe nabalovat další možnosti.

Pole zaznamenávaných dat sbíraných v logovacích souborech byl rozšířen o další informace HTTP protokolu, kterými se zabývá další část kapitoly. Samotný formát je z hlediska anlaytiky vedlejší a není není již nutno ho rozebírat.

2.2 Informace získané pomocí HTTP protokolu

Prohlížení webových stránek na internetu je zajišťěno pomocí HTTP protokolu. HTTP je zkratka pro Hyper Text Tranfer Protokol, neboli protokol pro výměnu hypertextů (hypertextových stránek). Je to způsob jak webový prohlížeč komunikuje s webovým serverem.

Pomocí HTTP protokolu poskytuje prohlížeč informace o uživateli tak, aby webový server mohl co nejlépe vyhovět požadavku.

Formátem HTTP protokolu se na tomto místě není třeba zabývat. Stačí vědět, že obsahuje podstatné jsou informace, které se používají ve webové analytice. Pro názornost je uveden příklad HTTP požadavku a některých základních informací, které obsahuje.

```
GET /dumprequest HTTP/1.1
Host: djce.org.uk
Connection: keep-alive
Accept: application/xml,application/xhtml+xml,text/html;q=0.9,text/plain;q=0.8,image/png,*/*;q
    =0.5
User-Agent: Mozilla/5.0 (X11; U; Linux i686; en-US) AppleWebKit/534.16 (KHTML, like Gecko)
    Chrome/10.0.634.0 Safari/534.16
Accept-Encoding: gzip,deflate,sdch
Accept-Language: cs-CZ,cs;q=0.8
Accept-Charset: windows-1250,utf-8;q=0.7,*;q=0.3
```

Výpis 3: Ukázka HTTP požadavku

Nyní je třeba si říct, jaké informace můžeme získat a k čemu se dají využít. Jsou to:

Požadovaná stránka

Udává o kterou stránku je zájem. Tento údaj se v nejjednodušším příkladě používá k určení popularity jednotlivých stránek na webovém serveru. Z hlediska toho, jak jsou dnešní webové stránky strukturovány, dá se předpokládat, že čím hlouběji je stránka zanořená ve stromové struktuře webu, tím toto číslo bude menší. Například:

```
--+ Hlavní stránka  (150 návštěv)
  |--+ Podstránka 1 ( 50 návštěv)
  |  \-- Pod-pod stránka ( 10 návštěv)
  \--- Podstránka 2 ( 50 návštěv)
```

Tato informace má různý význam v různých kontextech. Uvažujme informační server na kterém každý den vyjde několik článků v různých kategoriích. Hlavní stránka klade největší důraz na nejnovější články a v postranním panelu obsahuje odkazy na jednotlivé kategorie článků.

```
--+ Hlavní stránka informačního serveru
  |-- 1. Nejnovější článek
  |-- 2. Nejnovější článek
  |-- 3. Nejnovější článek
  ...
  |-- Kategorie článků A
  \-- Kategorie článků B
```

Pro kategorie znamená počet zobrazení popularitu jednotlivých kategorií. Tento údaj by měl být odpovídat počtu zobrazení článků v jednotlivých kategoriích a měl by se porovnávat k počtu zobrazení dalších kategorií. Například v případě že poměr návštěv kategorie vaření je velký ale poměr přečtení článků o vaření je malý, může to například znamenat, že kategorie je zajímavá ale neobsahuje tak zajímavé články.

Pro stránku, která představuje článek je počet zobrazení ukazatelem toho, jak je článek populární. S touto informaci je se dále pracuje a interpretuje se. Je třeba dát do kotextu:

- které webové stránky na daný článek odkázaly
- kolik vzniklo ke článku komentářů
- kolik lidí kliklo na článek z hlavní stránky
- jaký je u článku titulek a obrázek, jestli článek

V kontextu webové aplikace se může jednat o údaj, který udává nejpoužívanější sadu funkcí. Následující příklad toto demonstruje na zjednodušené administraci E-shopu.

```
--+ Administrace E-shopu
  |-+ Objednávky
  | |-- nevyřízené objednávky
  | |-- vyřízené objednávky
  | \-- hledat objednávky
  |-- Zboží
  \-- Nastavení
```

V této fiktivní administraci E-shopu jsou tři hlavní podstránky - Objednávky, Zboží a Nastavení. V tomto případě počet zobrazení jednotlivých kategorií umožňuje zjistit (zhruba), jak často je pracováno s objednávkami v poměru se zbožím a nastavením.

V každém případu počet zobrazených stránek představuje trochu jinou informaci a je třeba je interpretovat v podle typu měřené stránky (případně aplikace) a toho jak ji uživatelé používají.

Referrer

Údaj refferer je stránka, ze které uživatel přišel. Dá se tedy zjistit, odkud uživatelé na stránku přicházejí a spočítat nejčastější zdroje návštěv.

Ještě zajímavější využití této informace je v případě, že uživatel přišel z vyhledávače. Pak se dá z referreru zjistit, jakou frázi uživatel vyhledával a jaké vyhledávače dominují v počtu přivedených zákazníků.

Query string

Akceptovaný jazyk

Pokud má uživatel v operačním systému nastavený jazyk češtinu, HTTP požadavek říká serveru, že preferuje českou verzi obsahu. Tato informace je zajímavá především pro servery s velkou návštěvností, které plánují přidat jazykovou mutaci.

Geolokace

Prohlížeče a jejich verze

—- Porovnání co se dá zjistit z HTTP protokolu a co JavaScriptem. —-

—- JS only informace: —-

2.3 Události v prohlížeči

Pohyb myši

Klikání

Používání kláves

Odchod ze stránky

Skrolování stránky

2.4 Informace získané pomocí JavaScriptu

Rozlišení obrazovky

Velikost viewportu

(zobrazení stránek, jako analýza logů)

2.5 Derivované informace

Doba strávená na stránce

Loajalita = opakování návštěv

Častost návštěv

Počet návštěv

Počet unikátních uživatelů

—- Tabulka jednoduchého srovnání logfile analýzy a page taggingu —-

2.6 Interpretace a Vizualizace dat

Heatmapy

Heatmapy

Světová mapa

Segmentace

Aplikace v praxi - optimalizace kampaní - A/B / split testing

2.7 Clickstream analýza

2.8 Měření konverze

3 Návrh (15 stránek s diagramy)

O čem je tahle kampitola?

3.1 Hypotéza

Popis problému.
Demonstrovat na příkladu.
Popis řešení.
Co se bude měřit a proč.
Co nástroj dělá a proč.
Co nástroj nedělá a proč ne.
Pro koho je nástroj určen.
Existuje nějaký podobný nástroj?
Proč takový nástroj ještě neexistuje? (implementujou si in house řešení tohoto problému)

3.2 Sběr dat

Jaké data nás zajímají a proč
Jaké data nás nezajímají a proč
Jaké jsou omezení sběru dat.

3.3 Interpretace dat

Jaké možnosti interpretace dat jsou?
Jaké možnosti se hodí?
Jaké možnosti intepretace dat jsme zvolili a proč?
Jaké jsou možnosi segmentace. Jaké jsou možnosti do budoucna.
Možnosti aproximace?

3.4 Etická stránka sběru dat

Problém etiky sběru dat. Právní problémy. Anonymní vs. neanonymní data. Webaplikace vs. webová stránka. Zákazník vs. náhodný návštěvník. Analogie z fyzického byznysu.

4 Implementace (20 stránek i s grafy)

O čem je tato kapitola?

4.1 Popis technického řešení (10 Stran)

1) diagram celého systému
2) detail sběru dat
3) detail interpretace dat
4) zobrazení dat
vizualizační technologie, appengine proč jsem vybral tyto technologie detaily funkce měřícího skriptu (AES) detaily sběru dat detaily analýzy dat

4.2 Aplikace v reálném provozu (10 stran i s grafy)

oběcně kolik dat tam lítalo denně, kolik uživatelů analýza dat (na co jsem se zaměřil) zajímavé grafy, co z nich vyplývá závěry (navhrnout změny)

4.3 Navrhnuté změny na základě získaných dat (2 stránky)

5 Zhodnocení (2 stránky)

co splnilo očekávání co nesplnilo / předčilo očekávání přínos pro uživatele v praxi porovnání s podobnými nástroji

6 Závěr (1 stránka)

Byla to bomba :)

Michal Hantl

7 Reference

[1] Pecinovský, Rudolf, *Jak efektivně učit OOP. Tvorba softwaru 2005 – sborník přednášek*, ISBN 80-86840-14-X.

[2] Pecinovský, Rudolf, *Současné trendy v metodice výuky programování*, dostupné z url `http://gynome.nmnm.cz/konference/files/2006/sbornik/pecinovsky.pdf`.

[3] Plamínek, Jiří, *Tajemství motivace – Jak zařídit, aby pro vás lidé rádi pracovali*, ISBN 80-247-1991-6.

[4] Gašparovičová Ľuba, Hvorecký, Josef, *Kamaráti Robota Karla*, ISBN 80-06-00421-8.