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

Seznam obrázků

Seznam výpisů zdrojového kódu

1 Úvod

Tato diplomová práce se zabývá webovou analytikou a novým nástrojem pro analytiku webových aplikací.

Od vzniku webových stránek se web rozvíjel až do dnešní podoby a webová analytika ho provázela. Čím sofistikovanějšími se webové stránky stávaly, tím důmyslnější způsoby byly nacházeny k analýze uživatelského chování.

Z počátku byly používány nástroje na straně serveru, které pasivně analyzovaly data o návštěvách stránek. Až v polovině dvadesátkých let masivní vývoj webových prohlížečů přesunul sběr dat na stranu klienta, kde umožnil aktivní sběr dat o uživatelském chování přímo na stránce.

Sběr dat na straně klienta umožnil rozvoj jednoduše použitelných nástrojů pro webovou analytiku. Ty však svým způsobem stále kopírují konvenční přístup k analýze návštěvníků.

Současné metody měření a interpretace dat dobře vyhovují jednorázovým účelům jako je měření konverze reklamní kampaně, nebo poskytují průměrné statistiky, jako je top 10 nejnavštěvovanějších stránek na serveru v daném období.

Kromě webových stránek jsou na internetu také webové aplikace. Moderní technologie jako HTML5 se snaží smazat rozdíl mezi desktopovou aplikací a tou webovou. Tento trend potvrzuje i to, že Google připravuje svůj vlastní operační systém, který je výhradně založen na webovém prohlížeči a všechny aplikace poběží v něm.

Možnosti webových aplikací jsou již nyní rozšířeny o podporu audia, videa (včetně manipulace) a zobrazení trojrozměrných objektů pomocí technologie WebGl.

Někteří i tak budou tvrdit, že webové aplikace se nikdy nebudou moci měřit s desktopovými. To by možná byla pravda, pokud by jistá posvátná hranice mezi internetovým prohlížečem a nativní aplikací nebyla překročena technologií NaCl[1] - nativním kódem pro webové aplikace. Ano, v budoucnu budeme moci spouštět nativní kód v prohlížeči. Tím končí exkluzivní postavení desktopových aplikací.

S nástupem webových aplikací se mění i požadavky na nástroje webové analytiky. Webová aplikace se od webové stránky liší v několika věcech. Obecně rozdíl mezi stránkou a aplikací je interakce. Do webových aplikací vstupují uživatelská data ať už ve tradiční formě hodnot z formuláře, nebo data, které vznikají samotnou interakcí (například malování).

Když tedy uživatel vkládá, nebo vytváří v aplikaci data, je třeba je nějak uložit k tomu, aby mohly být později vyvolána a dalo se s nimi pracovat. Pro tento účel webové aplikace vyžadují způsob, jak unikátně identifikovat uživatele. V praxi to znamená, že se uživatel musí zaregistrovat a potvrdit svou identitu e-mailem[2].

Právě to, že je uživatel přihlášen dává nové možnosti vývoje nástrojů pro webobou analytiku. Narozdíl od webových stránek, kde nemůžeme s jistotou určit, zda se jedná

[1] NaCl není jediná technologie, která rozšiřuje možnosti prohlížeče o funkce a ryhlost nativního kódu. V současnosti se k těmto účelům používá Adobe Flash, nebo Microsoft Silverlight. Jde však o první implementaci, která si klade za cíl spouštět nativní kód, nepodléhající proprietární technologii.

[2] Stále častější alternativou se stávají identifikační autority jako Facebook, Twitter, nebo OpenId. To umožňuje, že uživatel nemusí vyplňovat email, ani heslo a je přihlášen pomocí jediného kliknutí.

o stejného uživatele můžeme podle přihlášení s jistotou tvrdit, že se o jedná o stejného uživatele a víme kterého.

V prostředí, kde je uživatel přihlášen můžeme na data pohlížet zcela jiným způsobem. Vezměme si statistiku počtu návštěv, kterou zpracovává Google Analytics a podává ji hned na hlavní stránce statistik pro daný web. Denní počet návštěv je vypočítán tak, že pokud přijdu na stránku ráno, v poledne a večer, započítá se to jako tři návštěvy. Pro webovou stránku to dává smysl ze dvou důvodů. Zaprvé ze stejného počítače může přijít ráno v poledne a večer úplně jiný člověk a to nelze nijak rozeznat. Zadruhé webové stránky nemají uživatele, ale návštěvníky tudíž dává smysl sledovat návštěvy.

Na druhé straně jsou aplikace, které mají uživatele a tak dává smysl sledovat používání. Jestli uživatel přijde třikrát denně, nebo jednou a stráví v aplikaci 3x tolik času je hezké vědět, ale není to tak podstatné. Co je podstatné je kolik uživatelů reálně dnes navštívilo aplikaci a kolik v ní strávili času.

Tento rozdíl se nezdá na první pohled až tak markantní. Je třeba si uvědomit, že se jedná o odlišný úhel pohledu a z toho vycházejí odlišné závěry. Například z Google Analytics se dá zjistit, kolik procent návštěv je z mobilních zařízení.

Řekněme, že mobilní zařízení tvoří 3% všech návštěv. To není jako mnoho, pravděpodobně neznamená, že budeme urpavovat naši aplikaci speciálně pro mobilní zařázení. Na druhou stranu, pokud máme 85% uživatelů, kteří používají free účet a 15% platících[3], mohou tři procenta znamenat až 20% platících uživatelů. Pokud máme více, než jeden placený tarif, například basic ($19), gold ($49) a platinum ($199). V takovém případě bude velmi důležité vědět, kolik z našich 3% je ve kterém tarifu.

Některé věci, které jsem popsal Google Analytics umožňuje. Je pouze na uživateli, aby si svoje Analytics co nejlépe nastavil a svou aplikaci opatřil příslušným nastavením, které umožní sledovat zajímavé údaje o uživateli.

Tento univerzální přístup ovšem znamená, že co si užiavtel sám neudělá, to nemá. Další omezení spočívá v tom, že Analytics zakazuje identifikovat uživatele.

Tato situace mě motivovala k vývoji nástroje, který je výhradně zaměřen na analytiku webových aplikací s důrazem na práci s uživateli. Nejedná se pouze o přesnění výsledků, jde o to zaměřit se na zákazníky a nikoliv na průměrná čísla.

[3] Freemium model, spočívá v účtech zdarma a placených, které poskytují něco navíc.

2 Webová analytika

Tato kapitola popisuje vznik webové analytiky a způsoby sběru a intepretace dat v obecné rovině.

2.1 Historie

Jak již bylo řečeno v úvodu, dnešní nástoje v podstatě kopírují způsob interpretace dat a proto je vhodné nastínit historii webové analytiky. Je proto vhodné se chvíli pozastavit nad tím, jak to všechno začalo.

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.

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.

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.

V roce 1996 byl v obou nejpoužívanějších prohlížečích[4] 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.

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

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

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

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.

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[5]. 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.

Nástroj, kterým se tato práce zabývá je primárně zaměřen na práci s uživateli.

2.2 Sběr dat

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.

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

Analýza dat z logovacích souborů

Webové 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.

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

V query string části adresy se nacházejí údaje, které modifikují obsah stránky. Je to část adresy stránky za otazníkem a není povinná. Například, pokud zadáte "web analytics"do hledání v Google, bude adresa stránky vypadat následovně:

```
www.google.com/search?sourceid=chrome&ie=UTF-8&q=web+analytics
```

Část adresy	Význam
znak "?"	začátek query string
znak "&"	odděluje dvojice klíče a hodnoty
sourceid=chrome	prohlížeč je Google Chrome
ie=UTF-8	požadujeme výslednou stránku v UTF-8
q=web+analytics	vyhledávaný dotaz je "web analytics"

Stejný princip se používá například když uživatel listuje produkty v elektronickém obchodě. V část adresy bude obsahovat informaci o tom, na které stránce se nachází.

Významné využití je například ve zjišťování, co zákazníci v elektronickém obchodě hledají a jaké fráze při tom používají. Zajímavé je také kolik zákazníků vyhledávání používá a jak často.

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.

Podíly jednotlivých prohlížečů se mění podle zaměření stránek a zeměpisných podmínek. Na stránkách o technologiích Microsoftu se dá očekávat vyšší podíl Internet Exploreru[6], na stránkách o vaření se zase mohou objevovat vyšší podíly starých prohlížečů. Každý web by měl fungovat pod většinou prohlížečů, pokud se však jedná o stránku o nových

[6]Internetový prohlížeč od Miscrosoftu

technologiích, kde převažují moderní prohlížeče, může se majitel rozhodnout o využití moderních vlatností těhto prohlížečů.

Na druhou stranu na stránkách s vysokým podílem starých prohlížečů se dá očekávat, že uživatelé stránky zobrazují na starých počítačích a tak je třeba tomu přizpůsobit obsah stránky, aby bylo prohlížení bezproblémové a plynulé.

Většina těhto informací se dá zjistit jak pomocí JavaScriptu na straně klienta, tak pomocí HTTP protokolu na straně serveru.

V sekci 2.2 je tabulka porovnání toho, co je možno měřit oběma metodami. Všechny následující informace vznikají přímo v prohlížeči a jediný způsob, jak se dají měřit je JavaScriptem.

Událostí je myšlen klik na tlačítko, psaní do textového pole, skrolování na stránce, dokonce i pohyb myší a změna velikosti prohlížeče. Jedná se o interakci uživatele se stránkou - ta probíhá v rámci prohlížeče na straně uživatele a tudíž není ji možno měřit jinak, než na straně uživatele.

K sběru dat na straně uživatele se používá JavaScript, který v určitých intervalech posílá tyto informace na server třetí strany, která je vyhodnocuje a smysluplně zobrazuje.

Tato čás popisuje jednotlivé události a jaké mají využití ve webové analytice.

Klikání myší je na stránce způsobuje přechod na jinou stránku, spuštění nějaké funkce, nebo nic, pokud uživatel klikl v místě, kterému není přiřazená žádná funkce.

Interakce s aktiviními prvky na stránce. V dnešní době obsahují stránky aktivní prvky, například fotoalba umožňují větší náhled obrázku bez nutnostni přejít na novou stránku. Protože tento náhled nezpůsobí zobrazení nové stránky, nevznikne ani nový záznam návštěvy stránky a většina nástrojů pro webovou analytiku tuto skutečnost nezaznamenávají.

Odchozí odkazy. Odkaz, který vede na stránku na jiném serveru, říkáme takovému odkazu odchozí. Provozovatele serveru může zajímat, kam z jeho stránky návštěvníci odcházejí. Autora článku zajímá, které odkazy z jeho článku jsou nejzajímavější pro jeho čtenáře.

Sledování **odkazů na soubory** umožňuje získat podobný přehled jako odchozí odkazy. Zatímco analýza logových souborů sleduje stahování souborů na serveru, sběr dat na straně klienta tuto informaci získá, když uživatel klikne na odkaz.

Ze všech kliků na stránce se navíc sestavují takzvané heatmapy[7]

Pohyb myši na stránce není kritický faktor, který je třeba sledovat na všech webech jako například kliky myší. Samy o sobě nemají velkou vypovídají hodnotu, existují však nástroje, které je těhto informací dokážou využít.

Agregace pohybů myší se využívá ke konstrukci heatmap. Účel heatmap je graficky znázornit, kterým částem webové stránky se dostává nejvíce pozornosti, ty se zobrazují teplými barvami (žlutá, červená), místa kterým se nedostává pozornost jsou naopak laděny do studeých barev (zelená, modrá). Více o heatmapách v sekci**??**

Druhým zcela pdlišným způsobem využití znalosti o pohybu myši je mouse-tracking. Tato technika kombinuje sledování kliků a pohybů myši a tak pořizuje záznam o uživatelově činnosti na stránce. Jedná se o velmi specifickou techniku, která se používá k testování

[7]Heatmapy zobrazují TODO, viz sekce TODO

stránek ještě dřív, než se uvolní na veřejnost, nebo k optimalizaci těch částí stránky, kde dochází ke sledované konverzi[8]. Příklad apliakce mouse-trackingu ke sledování konverze je například v proces dokončení objednávky elektronickém obchodě.

Události používání kláves se používá nejčastěji ve spojení s mouse-trackingem (viz Pohyb myši) k testování uživatelské zkušenosti, nebo k optimalizalizaci objednávkových formulářů v podobných situacích.

Některé webové stránky kladou velký důraz na přístupnost a umožňují uživateli se navigovat pomocí klávesových zkratek. V takovém případě by měli měřit i používání těchto zkratek a případně je upravit tak, aby byla navigace na jejich webu co nejjednodušší.

Událost skrolování stránky nastává když uživatel posouvá stránku nahoru a dolů, případně doprava nebo doleva. Podle posuvu stránky v určitých intervalech se dá usuzovat, že uživatel stránky čte, nebo pouze prohlíží. Podle délky čtení a toho, jestli se uživatel dostal až na konec článku se dá určit, jestli článek dočetl.

V kombinací s velikostí vnitřní části okna prohlížeče se dá přesně uričit, na kterou část stránky se uživatel dívá. Toho se například využívá k vykreslení heatmapy, která kolik času uživatelé věnují jednotlivým částem stránky.

Odchod ze stránky je událost, která je spuštěna když uživatel přechází na jinou stránku, nebo zavírá okno se stránkou.

Tato událost se používá k vypočtení tzv "bounce rate". Bounce rate je procento uživatelů, kteří na stránku přijdou a hned zase odejdou. Takové chování se dá očekávat například, když uživatel něco hledá ve vyhledávači, prohlédne si a vzápětí odchází, protože na první pohled stránka neobsahuje informace, které hledal.

Bounce rate je třeba chápat v kontextu s tím, odkud uživatel přichází. Měl by se odlišovat podle zdroje, odkud návštěvník přišel. Například návštěvníci z vyhledávače budou mít odlišný bounce rate, než návštěvníci z odkazujících článků, reklamy, nebo ti, kteří přímo napíší adresu stránky do adresního řádku.

Rozlišení obrazovky, stejně jako používaný prohlížeč je informace která vypovídá o návštěvníkovi a zařízení, které používá. Starší monitory mají například poměr stran 4:3, novější zase 16:9 a 16:10. Stejně tak se dá usuzovat, že u nízkých rozlišení moderních úhlopříček se bude jednat o netbooky a notebooky. Nejvyšší rozlišení mají nejmodernější dvaceti sedmi palcové monitory, a ještě větší rozlišení jako je 3200x1200 znamená, že má uživatel dva monitory, konkrétně s rozlišením 1920x1200 a 1280x1024.

Velikost viewportu TODO

(zobrazení stránek, jako analýza logů)

[8] Konverze TODO

Možnosti logfile analýzy a page-taggingu

informace	logfile analýza	page-tagging
objem přenesených dat	●	○
stránka	●	●
referrer	●	●
query string	●	●
jazyk	●	●
prohlížeč	●	●
odchozí odkazy	○	●
doba na stránce	○	●
opuštění stránky	○	●
rozlišení obrazovky	○	●
velikost viewportu	○	●
zobrazená část stránky	○	●

2.3 Interpretace dat

Přidaná hodnota analytických nástrojů spočívá v interpretaci nasbíraných dat. Většina údajů, podle kterých se orientujeme vzniká kombinací několika událostí, nebo informací.

Díky cookies[9] je možné sledovat návštěvnka stránky. Při první návštěvě se návštěvníkovi uloží do cookies jeho unikátní identifikátor a lze tak při příští návštěvě identifikovat, že se jedná o stejného návštěvníka. Označování návštěvníků umožňuje zjistit například počet unikátních návštěvníků. Přesto, že někteří uživatelé si Cookie ručně mažou, většina to nedělá a počet unikátních návštěvníků se dá pokládat za relativně přesný.

Označování uživatelů umožňuje také sledovat počet návštěv. Pokud uživatel stráví pět minut prohlížením stránek, zavře stránku a pak se za dvě hodiny vrátí, vyhodnocuje se to jako dvě návštěvy.

Dlouhodobé opakování návštěv označujeme jako loajalitu návštěvníků. Stránky, které mají tuto hodnotu vysokou mají stabilní bázi návštěvníků, kteří si tam zvykli pravidelně chodit. Tato hodnota je důležitá pro informační servery, elektronické obchody, blogy, fóra a komunitní stránky.

Doba strávená na stránce

Podle IP adresy návštěvníků se zjišťuje jejich poloha a zobrazuje se pomocí jednoduché mapy. Na obrázku 2 je ukázka takové mapy, která znázorňuje počet návštěv pro Ameriku. Jendá se o statiskiku pro rozšíření do prohlížeče Google Chrome. Jelikož se jdná o rozšíření pro technické typy, lze usuzovat, že takoví se vyskytují v Kalifornii (tmavě zelená vlevo) a New Yorku.

Na obrázku 3 je ukázka heatmapy pro stránky výsledků vyhledávačů Google a Bing. Vstupními daty pro vytvoření takových interpretací jsou v ideálním případě data získaná

[9] Do tzv. Cookie je možno uložit uživateli zhruba 4096 znaků a internetový prohlížeč se o tuto hodnotu stará.

pomocí sledování očních pohybů návštěvníků pomocí speciální kamery. Takovéto testování je možnos provádět pouze v laboratoři a s potřebným vybavením.

Mimo laboratoř je možno takové obrázky vygenerovat z pohybů myší a klikání návštěvníků stránky. Cílem heatmap je odhalit, kterým částem stránky se dostává více pozornosti. Na obrázku 3 je vidět, že uživatelé vyhledávače Bing věnují pozornost levému sloupečku.

Testované rozložení stránky vyhledávače Google žádný levý sloupeček nemělo a uživatelé nevěnovali pozornost odkazům pro hledání obrázků a další alternativám. Možná proto má dnes vyhledávač google tyto odkazy nalevo, hned vedle výsledků vyhledávání.

Aplikace v praxi

Teď, když byly nastíněny základy toho, jak webová analytika sbírá a interpretuje data je na místě se zmínit o praktickém využití těchto dat.

Na počátku, kdy se měřil pouze počet zobrazení stránky, byly tyto nástroje měřítkem popularity stránek. Podle denní návštěvnosti se dalo odvodit, kolik by si měla stránka účtovat za reklamu[10], většinou formou bannerů. Návštěvnost stránek i dnes určuje cenu reklamy na stránce, PR článků a obecně hodnotu celé stránky.

Když potenciální zákazník klikne na reklamu a dostane se naši stránku, stojí nás to peníze. Ne každý zákazník objedná, nebo jinak vygeneruje zisk. Takže kolik zákazníků, kteří se na naši stránku dostali přes reklamu u nás vlastně nakoupí? A ještě lépe - kolik nás ve výsledku stojí nová objednávka? Na tyto otázky standardně odpovídá například Google Analytics, který je přímo napojen na AdWords[11].

V okamžiku, kdy se na naši stránku dostane uživatel, vstupuje do hry měření konverze. Ta měří poměr návštěv a uskutečněných cílů - například počet objednávek, nebo počet nových emailů v seznamu odběratelů novinek. Konverzní poměr se zvlášť měří u uživatelů, kteří přišli z vyhledávače, přes reklamu anebo sami napsali adresu webové stránky do prohlížeče.

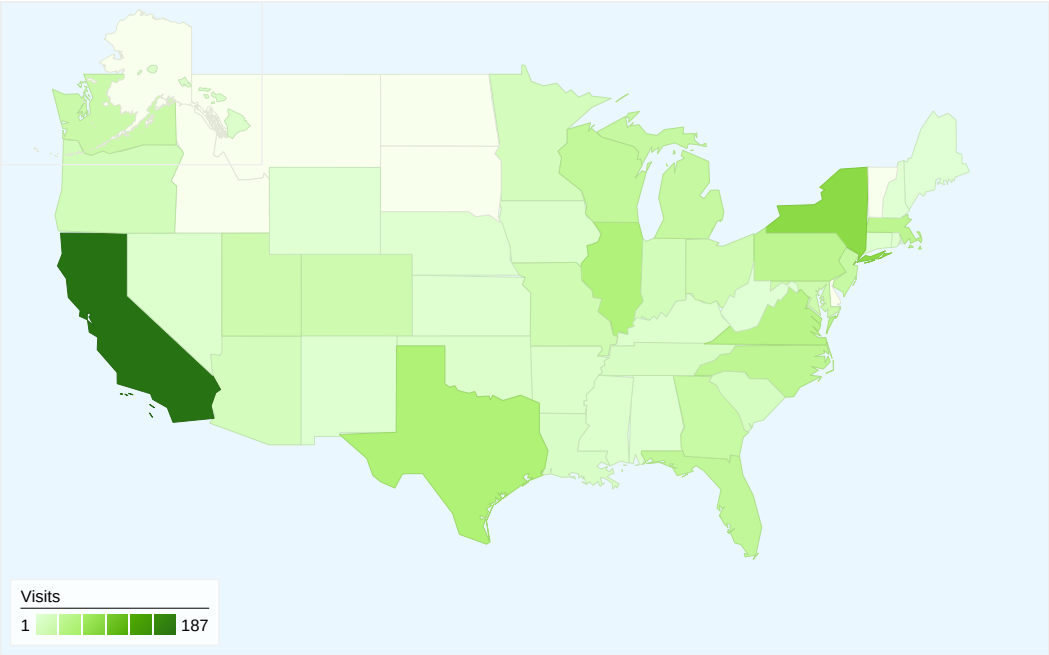
Znalost jednotlivých konverzí a napojení na reklamní systém umžňuje určit cenu za konverzi a tím i efektivitu jednotlivých kampaní.

Typické zdroje návštěvnosti jsou reklama, vyhledávání a dnes nově sociální média. Vyhledávače jsou kapitolou samou pro sebe. Vyhledávače používá velké denně množství lidí a je lákavé dostat svůj web do horních pozic pro vyhledávané fráze související s tématem stránky. Díky tomu, že reklama je relativně drahá je optimalizace pro vyhledávače žádanou službou, která představuje dlouhodobý nárůst návštěvnosti za jednorázovou investici.

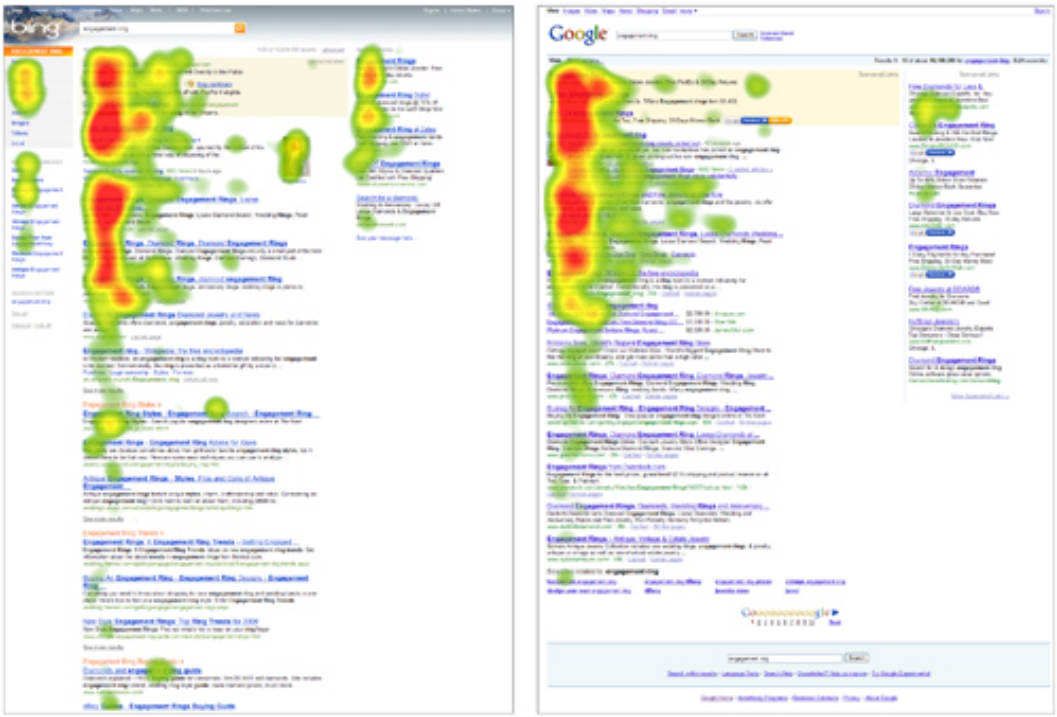
Návštěvnost z vyhledávačů se dá efektivně měřít i s zjištěním návštěvnosti pro jednotlivé klíčové slova a vyhledávací fráze. Analytické nástoje tak umožňují změřit o kolik se zvedl počet objednávek, registrací nebo čehokoli, co majitel stránky sleduje jako cíl konverze.

[10] Dnes jsou více rozšířené sytémy, které použadují platbu za počet prokliků a jejich cena se určuje formou aukce. Například AdWords, u nás Sklik.

[11] Pay Per Click reklama od Googlu



Obrázek 2: Google Analytics - návštěvy podle zemí



Obrázek 3: Heatmapy - vlevo Bing, vpravo Google

Mocným nástrojem pro optimalizaci webových stránek je split-testování, taktéž zvané A/B testování. Tato technika spočívá v tom, že se vytvoří dvě varianty stránky a a některým uživatelům se zobrazí jedna varianta a některým druhá. Na daných stránkách se pak měří konverze a výsledkem testu je zjištění, která stránka měla větší konverzi.

Tímto způsobem se dají optimalizovat stránky, které mají zákazníka přimět k registraci, nákupu, nebo jiné cílové akci. Tato technika umožňuje určit, jaký název produktu bude mít lepší výsledky, alternativní slogany, vzhled produktu, nebo dokonce i cenu produktu.

V případě testování ceny produktu se některým zákazníkům zobrazí například $19 a jiným $49. Je docela možné, že produkt za $49 se bude lépe prodávat, protože zákazníci od něj budou očekávat větší kvalitu. Přesto, že nezní příliš sofistikovaně, umožňuje split-testing každému experimentovat s mnoha aspekty a reálně ověřit, zda navrhované změny opravdu vedou k lepší konverzi.

Druhá strana mince je v případě webové analytiky hlubší porozumění. Do tého kategorie patří informace typu jaké prohlížeče a zařízení používají, jaký jazyk preferují, z jaké části světa jsou, odkud na stránku přicházejí, co na webu dělají a kudy stránku opouštějí.

V závislosti na zaměřené webové stránky je tady větší nebo menší snaha poznat zákazníka. Dnešní analytcké nástroje mají pouze omezené možnosti práce se zákazníkem, především z toho důvodu, že odmítají vědět o tom, že jde o konkrétního zákazníka, jako například Google Analytics, který sledování konkrétních uživatelů přímo zakazuje v podmínkách použití.

Na práci se zákazníky se specializují jiné systémy jako jsou CRM Systémy[12] a jiné úzce specializované produkty, které se již webovou analytikou nezabývají.

Právě znalost konkrétního zákazníka a práce s ním odlišuje nový analytický produkt od ostatních. V dnešní době sodiálních médií, kdy telefonní operátoři mají teamy lidí, kteří se starají o zákaznickou podporu na sociálních sítích, je kritické mít přehled o svých zákaznících a to zvláště u webových aplikací, kde interakce zákazníka s apliakcí je to, za co nás zákazník platí.

Příští kapitola se věnuje vlastnostem nového nástroje a jejich využití.

[12]Customer Relationship Management systémy

3 Nový nástroj

Tato kapitola popisuje problém, který známé analytické nástroje zatím neřeší, a konkrétní způsob jeho řešení. Jak již bylo zmíněno v předchozím textu, nástroje pro webovou analytiku nepracují s konkrétními zákazníky a některé dokonce zakazují jakkoli identifikovat uživatele webů.

Motivace a 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.

Záměrné zahalování identity uživatele je argumentováno zachováním soukromí uživatele. Například Google Analytics obsahuje ve smluvním ujednání služeb Google Analytics následující text:

> 7.1 Nebudete slučovat (nebo nedovolíte jakékoliv třetí osobě slučování) jakákoli data shromážděná z vaší Website (vašich stránek Website) (nebo website takovéto třetí strany) s jakýmikoliv informacemi identifikujícími osoby, které pocházejí z jakéhokoli zdroje jsoucího částí vašeho užívání (nebo užívání takovéto třetí strany) Služby. Budete splňovat všechny zákony na ochranu dat a soukromí ve spojitosti s vašim používáním Služby a shromažďováním informací od návštěvníků na vašich stránkách website. Na své stránce na viditelném místě zpřístupníte (a budete i dodržovat) vhodnou směrnici o ochraně soukromí.
> ...

Na jedné straně je snaha ochránit uživatelovo soukromí, na druhé poskytnout co nejlepší službu a relevantní obsah. Etickou stránku věci se spíše zabývají velké firmy, které mají uloženo mnoho uživatelských dat a nezávislí profesionálové, kteří poukazují na přestupky a nedostatky.

Problém soukromí na webu je možno přirovnat k nakupování v obchodu. Pokud jsem v obchodním řetězci a nakupuji, budou mě sledovat kamery, které mě dokáží identifikovat a sbírají data o tom, kudy chodím. Z hlediska soukromí nechci, aby bez mého souhlasu o mě sbíral obchodní řetězec informace o tom, co nakupuji, jak často do obchodu chodím, nebo co mám na sobě. Toto je filosofie Google Analytics.

Když se svým nákupem přistoupím k pokladně, první na co se mě prodavačka zeptá je, zda mám jejich "kartu". Pokud ano, vezme si ji k zařízení napojeném na pokladnu a v tom okamžiku se můj zatím anonymní nákup stává mým nákupem a je svázán s mou kartou. Tímto způsobem obchodní řetězec dokáže sledovat jak často nakupuji a co nakupuji.

Aby měl zákazník důvod k pořízení karty, která umožňuje jeho sledování je motivován sbíráním bodů, nebo jinými výhodami[13]. Při registraci pro kartu IKEA Family se do formuláře vyplňuje jméno, příjmení, bydliště, email a den narození. O ochraně osobních údajů říká IKEA, že jde o "vztah založený na důvěře". V následujícím výňatku z

[13]Například IKEA svým členům "IKEA Family"věnuje nápoje k jídlu zdarma.

textu "Ochrany osobních údajů"společnost polopatě vysvětluje, že údaje bude poskytovat třetím stranám:

> **Informace, které nám sdělíte, u nás také zůstanou**
> Kdybyste nemohli IKEA věřit, nemohli byste v ní ani nakupovat. Z toho důvodu je pro nás ochrana vašich osobních důvodů navýsost důležitá. IKEA nesděluje vaše osobní údaje jiným firmám mimo IKEA. Informace, které nám poskytnete, využívá jen a jen IKEA. Někdy sice nastanou případy, kdy Vaše osobní údaje poskytneme některé z našich divizí v jiné zemi, ale v takových případech používáme všechny dostupné administrativní, technické i fyzické způsoby ochrany dat, abychom zabránili jejich možnému prozrazení, použití, změně či zničení. Stručně řečeno se všemi dostupnými prostředky snažíme chránit vaše osobní údaje před předvídatelným nebezpečím. V některých případech však opravdu musíme některé vaše údaje sdělit jiným společnostem, které pověříme jejich zpracováním, protože od nich chceme, aby vám poskytly určité služby. Zpracování osobních údajů probíhá v těchto případech podle našich pokynů.

Velké společnosti tedy respektují naše soukromí, ale v rámci vlasntí optimalizace se nás snaží získat do svých klubů, které jim umožňují sbírat libovolná data a vyhodnocovat je.

Pro úplnost metafory o internetovém soukromí je třeba zmínit jak se k soukromí staví ti menší. Zákazník nakupuje v lahůdkářství víno a všelijaké dobroty. Lahůdkář dobře rozumí zboží a dokáže poradit a zákazník si rád nechá poradit, protože sám není odborník a dobrou radu bere jako součást služby lahůdkářství. V této situaci nemá prodejce sepsaný dokument o ochraně osobních dat, nicméně ví, jaké zboží zákazník preferuje, jak často chodí, za kolik průměrně nakoupí a mnohdy i pro jaké příležitosti nakupuje.

To, co má lahůdkář v hlavě se ve větším měřítků nazývá Customer Relationship Management. CRM je nástroj ke shromažďování, zpracování a využití informací o zákaznících. Je to vlastně takový drobnohled, kterým se pohlíží na zákazníka jako jednotlivce. Toto kontrastuje s tím,jak data využívá firma s mnoha zákazníky, která je sohrnně analyzuje.

Nástroj, kterým se tato diplomová zabývá představuje řešení pro webové aplikace, které mají stovky až tisíce zákazníků, které fungují typicky na základě měsíčního předplatného. V tomto případě je potřeba kombinovat způsoby analýzy dat velkých objemů a pohledu na konkrétní zákazníky.

Nástroj si klade za cíl podporovat webovou aplikaci od jejích začátků až k dospělosti. Zabývá se celým životním cyklem zákazníků od jejich prvního vstupu na web, přes registraci trial verze, změnu plánu na placený a používání aplikace.

V metafoře výše se je produkt na půl cesty mezi lahůdkářem, který osobně zná všechy své zákazníky a tato znalost je přidanou hodnotou služby zákazníkovi a obchodním řetězcem, který potřebuje analyzovat chování velkého množsví zákazníků.

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)

3.1 Sběr dat

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

3.2 Zpracování 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?

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 Případová studie

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)

Navrhnuté změny na základě získaných dat (2 stránky)

5 Zhodnocení

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

6 Závěr

Nástroj splnil předpoklady..
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.