Univerzita Hradec Králové
Fakulta informatiky a managementu
Katedra informačních technologií

Moderní metody a trendy webové analytiky

Diplomová práce

Autor: Bc. Ondřej Šúkala
Studijní obor: Aplikovaná informatika, prezenční forma

Vedoucí práce: Ing. Karel Mls, Ph.D.

Hradec Králové duben 2016

Prohlášení:

Prohlašuji, že jsem diplomovou práci zpracoval samostatně a s použitím uvedené literatury.

V Hradci Králové dne 25.4.2016 Ondřej Šúkala

Poděkování:

Děkuji vedoucímu diplomové práce Ing. Karlu Mlsovi, Ph.D. za metodické vedení práce, cenné rady, ochotu a čas strávený při konzultacích.

Anotace

Cílem diplomové práce je popsat moderní metody webové analytiky a vytvořit aplikaci pro sledování návštěvníků webu. První kapitola obsahuje definice základních pojmů, které se často vyskytují v celé diplomové práci. Následuje definice webové analytiky, její historický vývoj a popis procesu webové analytiky. Další kapitoly se podrobněji věnují jednotlivým krokům tohoto procesu. Framework webové analytiky pomáhá analytikovi definovat cíle konkrétního webu a měřit jeho úspěšnost. Metody identifikace a sledování návštěvníků popisují způsoby, jak získat potřebná data pro účely webové analytiky. Následující kapitola se věnuje očištění získaných dat od nevyžádaných přístupů. Kapitola analýza se zaměřuje na získávání znalostí ze zpracovaných dat. Hlavním přínosem této práce je nástroj webové analytiky, který slouží jako doplněk ke stávajícím nástrojům nebo jako framework pro vytvoření konkurenčního nástroje. Součástí práce jsou i podpůrné nástroje pro testování této aplikace.

Annotation

Title: Advanced methods and trends of web analytics

The aim of this diploma thesis is to describe advanced methods and trends of web analytics and develop an application for tracking web visitors. The first chapter contains definitions of basic terms, which occur frequently in the entire diploma thesis. This is followed by the definition of web analytics, its historical development and the description of the process of web analytics. The next chapters expand on the various steps of the process. Framework of web analytics helps the analyst to define the objectives of a particular website and measure its success. Methods of identification and tracking of visitors describe ways to obtain the necessary data for the purposes of web analytics. The next chapter deals with cleaning of collected data from spam. The chapter analysis focuses on knowledge extraction from the processed data. The main contribution of the theses is a web analytics tool, which serves as a supplement to existing tools or as a framework for creating a competitive tool. The thesis also contains support tools for testing this application.

Obsah

Seznam obrázků

Seznam tabulek

1 Úvod

Podle letošních výzkumů Asociace pro elektronickou komerci (APEK) se obrat české e-komerce každý rok výrazně zvyšuje a 73 % uživatelů internetu nakoupí alespoň jednou za čtvrt roku [1, 2].

Weboví vývojáři a provozovatelé webů se dříve rozhodovali na základě intuice. Výsledkem byly weby, které odpovídaly jejich představám, ale nedokázaly naplno využít zájmu návštěvníka webu. Bohužel je možné se s tímto přístupem setkat i v dnešní době. Webová analytika umožňuje lépe pochopit zájmy návštěvníka, odhalit nedostatky webu a v některých případech i určit řešení těchto nedostatků. Správné použití webové analytiky a implementace zjištěných poznatků poskytují lepší kontrolu nad provozem webu a zvýšení jeho úspěšnosti. Samotné použití webové analytiky nemusí vždy stačit k dosažení úspěchu. Analytické nástroje nabízejí velké množství údajů a funkcí, které mohou být pro daný web nepotřebné nebo dokonce i škodlivé. Chybná interpretace údajů nebo nesprávné použití funkcí mohou v některých případech vést k větším škodám než samotné nepoužití těchto nástrojů.

Webová analytika je mladý obor, který se stále vyvíjí. Nasvědčuje tomu nedostatek základních definic a velké množství pojmů, které se liší mezi výrobci analytického softwaru. Do českého prostředí se nové poznatky z webové analytiky dostávají převážně ze zahraničí. Z tohoto důvodu jsou v diplomové práci uvedeny některé pojmy v českém i anglickém jazyce.

2 Cíle práce

Cílem teoretické části diplomové práce je popsat současné metody webové analytiky včetně jejich výhod a omezení.

Praktická část práce si klade za cíl vytvoření nástroje webové analytiky pro práci s naměřenými daty a zobrazení webového provozu v podobě mapy.

3 Definice základních pojmů

Tato kapitola obsahuje definice, které se opakovaně vyskytují v následujících kapitolách. Některé tyto pojmy se často nesprávně zaměňují, přestože se jedná o základní a běžně používané pojmy.

3.1 Webový server a prohlížeč

Jako **webový server** se označuje hardware, který obsahuje webové soubory a poskytuje je zařízení návštěvníka. HTTP server je software, který je spuštěn na webovém serveru a komunikuje s klientem. Statický webový server posílá klientovi soubory v podobě, v jaké jsou na něm uloženy. Dynamický webový server musí soubor před jeho odesláním nejdříve zpracovat pomocí serverových skriptů. **Webový prohlížeč** je program, který komunikuje s webovým serverem a zobrazuje jeho obsah návštěvníkovi [3].

3.2 Hypertext transfer protocol

Hypertext Transfer Protocol (**HTTP**) je bezstavový protokol pro přenos dat mezi serverem a klientem. Klientem může být například webový prohlížeč, který zasílá HTTP požadavek webovému serveru. Server zpracuje požadavek a zašle klientovi odpověď. Požadavky a odpovědi jsou popsány pomocí HTTP hlaviček. Hypertext Transfer Protocol Secure (HTTPS) je zabezpečená verze HTTP protokolu [4].

3.3 Web page, website a World Wide Web

Web page (**webová stránka**) je konkrétní dokument na webu. Soubor webových stránek poskytovaných v rámci jedné domény se nazývá **website** (**webové stránky**). **World Wide Web** (**WWW**) je systém vzájemně propojených hypertextových dokumentů přístupných prostřednictvím Internetu. World Wide Web i webové stránky se občas souhrnně označují pojmem web. V rámci této diplomové práce se označení web používá pouze pro webové stránky [5].

3.4 Návštěvník, uživatel a reálný uživatel

Jako **návštěvník** se nejčastěji označuje unikátní návštěvník identifikovaný pomocí unikátních dat z HTTP hlaviček. Občas se pojem návštěvník zaměňuje s pojmem

webový prohlížeč, protože prohlížeč zastupuje návštěvníka při interakci s webem. Po registraci se identifikovaný návštěvník stává **uživatelem**. **Reálný uživatel** je skutečný člověk, který navštívil sledovaný web.

3.5 Identifikátory

Uniform Resource Identifier (**URI**) je identifikátor zdroje. K identifikaci zdroje lze použít jeho umístění, název nebo obě varianty současně. **Uniform Resource Locator** (**URL**) je typ URI, který identifikuje zdroj pomocí jeho umístění. **Uniform Resource Name** (**URN**) na rozdíl od URL identifikuje zdroj pomocí názvu [6].

3.6 Data, informace a znalosti

Data *jsou počítačové reprezentace modelů a vlastností skutečných nebo simulovaných entit* [7].

Informace *jsou data, která představují výsledky výpočetního procesu, jako je statistická analýza, pro přiřazení významu datům, nebo přepisy některých významů přidělených lidmi* [7].

Znalosti *jsou data, která představují výsledky počítačové simulace poznávacího procesu, jako je vnímání, učení, asociace, a uvažování, nebo přepisy některých znalostí získaných lidmi* [7].

3.7 Log a záznam

Log *je soubor záznamů událostí, ke kterým došlo v systémech a sítích organizace. Každý* ***záznam*** *obsahuje informace vztahující se k určité události, k níž došlo v rámci systému nebo sítě* [8].

3.8 Web analytics tool

Web analytics tool (**nástroj webové analytiky**) zpracovává naměřená data a zobrazuje je ve formátu vhodném pro uživatele. V současnosti je Google Analytics nejpoužívanějším nástrojem webové analytiky.

4 Vývoj webové analytiky

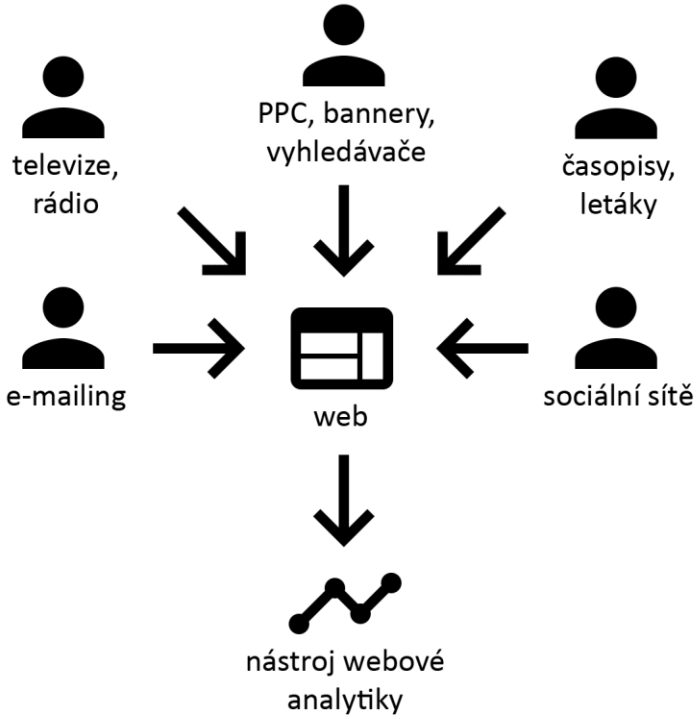
Analytika je proces získávání optimálních a realistických rozhodnutí založených na datech [9].

Definice webové analytiky se vyvíjí současně s vývojem webových technologií a online marketingu. Původně se serverové logy využívaly převážně pro technické účely (záznam chyb atp.). V roce 1995 Dr. Stephen Turner vytvořil Analog, první široce rozšířený program určený k parsování serverových logů pro účely webové analytiky. Ve stejném roce společnost WebTrends vyvinula první komerční nástroj pro webovou analytiku. V roce 1966 vznikla počítadla přístupů (hit counters) a velmi rychle se rozšířila po celém internetu a tím zároveň rozšířila povědomí veřejnosti o webové analytice. O rok později vznikla nová metoda měření webového provozu pomocí značkování stránek. Tato metoda rozšířila webovou analytiku o nové údaje a ještě více posunula zaměření webové analytiky směrem k online marketingu. V roce 2004 byla založena Web Analytics Association, která měla za cíl podporovat rozvoj webové analytiky. V roce 2005 společnost Google koupila Urchin a spustila dnes nejpoužívanější nástroj pro webovou analytiku - Google Analytics [10, 11, 12, 13, 14].

V roce 2006 Web Analytics Association zveřejnila následující definici.
Webová analytika je měření, sbírání, analýza a reportování internetových dat za účelem porozumění a optimalizace využívání webu [15].

Díky rozšíření webové analytiky i mimo web se v roce 2012 Web Analytics Association přejmenovala na Digital Analytics Association a začala používat pojem digitální analytika. Světový odborník na webovou analytiku Avinash Kaushik označuje digitální analytiku jako webovou analytiku 2.0.
Webová analytika 2.0 znamená analýzu kvalitativních a kvantitativních dat z vlastního webu i webů konkurence, a to za účelem kontinuálního vylepšování uživatelského prožitku současných i potenciálních zákazníků, což se promítá do požadovaných výsledků na webu i mimo něj [16].

S využitím digitální analytiky i pro offline prodejní kanály se začíná používat pojem multikanálová analytika (Multi-channel analytics) [17].



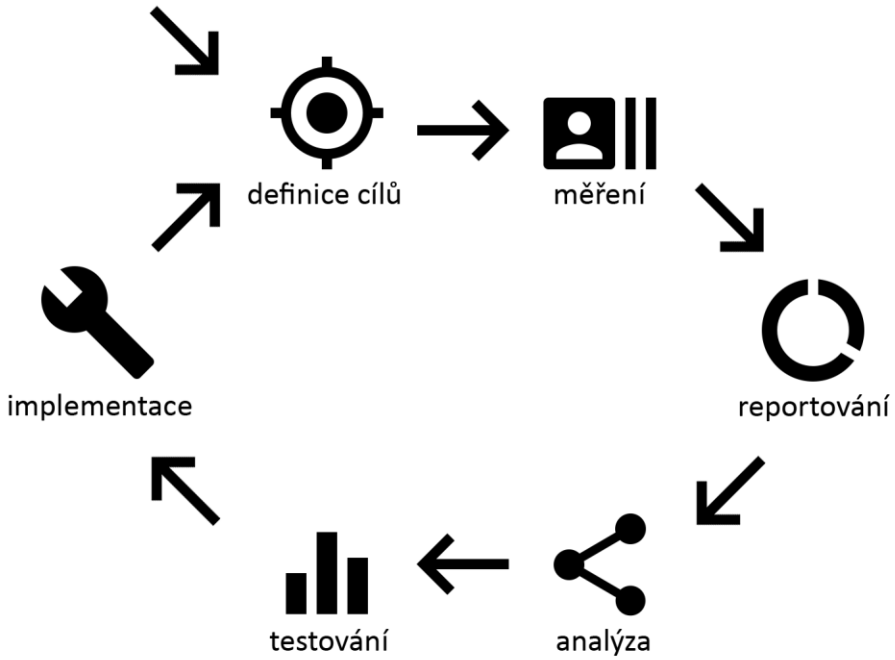
Obrázek 1: Multikanálová analytika [vlastní tvorba]

Obrázek č. 1 znázorňuje různorodost zdrojů webového provozu. Příchod návštěvníků z offline kanálů lze identifikovat například vytvořením speciální stránky, přidáním parametru k existující stránce nebo registrací další domény. Každá stránka, parametr nebo doména slouží pouze pro jednu marketingovou kampaň.

Účelem webové analytiky je zlepšit uživatelský zážitek (například snadný a rychlý nákup zboží v e-shopu) a tím zvýšit úspěšnost webu (v tomto případě zvýšení počtu objednávek). Správné použití webové analytiky je přínosem pro návštěvníky i majitele webu.

5 Základní kroky webové analytiky

Webová analytika se skládá z několika kroků, které na sebe navazují. Jedná se o nekonečný proces, protože chování návštěvníků i cíle webu se časem mění a je potřeba na tyto změny reagovat.



Obrázek 2: Základní kroky webové analytiky [vlastní tvorba]

Obrázek č. 2 znázorňuje proces webové analytiky. Prvním krokem procesu je vždy definování cílů webu nebo hypotéz, které chce analytik ověřit. Na základě výběru potřebných dat se zvolí vhodná metoda měření. Naměřená data se většinou dále zpracovávají a očišťují od nežádoucích přístupů na web. Ze zpracovaných dat se vypočítávají metriky a sestavují reporty. Pomocí segmentace a dalších nástrojů webové analytiky se získávají znalosti z naměřených dat a informací z reportů. Na základě nových zjištění je možné připravit několik variant a jejich testováním zjistit, která z nich je nejlepší. Po dokončení implementace nové varianty celý proces začíná znovu. Následující kapitoly se podrobně věnují jednotlivým krokům tohoto procesu.

6 Framework

Framework webové analytiky je strategický plán webu. Pomocí jasně definovaných kroků umožňuje identifikovat hlavní i vedlejší cíle a vyhodnotit jejich úspěšnost nebo neúspěšnost.

Framework se skládá z těchto kroků [18]:

1. nalezení obchodních cílů
2. definování cílů pro všechny obchodní cíle
3. nalezení klíčových ukazatelů výkonnosti (KPI)
4. nastavení požadovaných hodnot pro každý KPI
5. definování segmentů

6.1 Business objectives

Business objectives (**obchodní cíle**) popisují účely webu. Příkladem obchodního cíle může být zvýšení počtu objednávek. Je vhodné zvolit zhruba 3-5 obchodních cílů pro každý web. Správně nastavené obchodní cíle jsou proveditelné, srozumitelné, měřitelné a prospěšné [19].

6.2 Goals

Goals (**cíle**) umožňují měřit úspěšnost obchodních cílů. Zvýšení počtů konverzí může být cíl pro měření úspěšnosti obchodního cíle z předchozího příkladu.

Mezi časté typy cílů patří [20]:

- načtení URL (například zobrazení děkovné stránky po dokončení nákupu)
- událost (například spuštění videa)
- délka návštěvy (vhodné pro obsahové weby)
- zobrazení určitého počtu stránek

6.3 Dimenze

Dimenze jsou vlastnosti návštěvníka a vlastnosti jeho aktivit na webu. Komunikaci návštěvníka s webovým severem nejčastěji zprostředkovává webový prohlížeč. Množství a typy naměřených dimenzí závisí na použité metodě měření. Z tohoto důvodu je vhodné vytvořit seznam požadovaných dimenzí ještě před výběrem a implementací měřícího kódu [19].

Obrázek 3: Měření dimenzí [vlastní tvorba]

6.3.1 Nejpoužívanější dimenze

Následující dimenze podporuje většina analytických nástrojů. Analytické nástroje často nabízejí i možnost definovat vlastní dimenze, které jsou specifické pro konkrétní web.

Dimenze **page** (**stránka**) obsahuje identifikátor konkrétní stránky nebo jiného souboru na sledovaném webu. Jako identifikátor se nejčastěji využívá URI. Přesné určení této dimenze záleží na použitém analytickém softwaru a analytikovi, který tento software používá. Velkou výhodou této dimenze je spolehlivost měření. Jedná se o jeden z mála naměřených údajů, který nelze podvrhnout [15].

Event (**událost**) je zaznamenaná akce s přiřazeným konkrétním datem a časem provedení. Události jsou aktivity, které se odehrály na stránce [15].

Entry page (**vstupní stránka**) je první navštívená stránka v rámci návštěvy sledovaného webu. Počet zaznamenaných vstupních stránek odpovídá počtu návštěv, protože každá návštěva obsahuje vstupní stránku [15].

Landing page (**přistávací stránka**) je stránka navržená speciálně pro návštěvníky přicházející z marketingových kampaní. Může se nacházet na odlišném webu a

pouze přesměrovávat návštěvníky na hlavní web. Běžná přistávací stránka neobsahuje navigaci a je navržená pouze pro jeden účel. Přistávací stránka bývá často i vstupní stránkou [21].

Dimenze **exit page** (**výstupní stránka**) je poslední navštívená stránka v rámci návštěvy sledovaného webu. Počet zaznamenaných výstupních stránek odpovídá počtu návštěv, protože každá návštěva obsahuje výstupní stránku [15].

Dimenze **referrer** obsahuje zdroj webového provozu. **Internal referrer** (**interní referrer**) je URL stránky, která se nachází na sledovaném webu. **External referrer** (**externí referrer**) označuje URL stránky nacházející se mimo sledovaný web. **Page referrer** (**referrer stránky**) je zdroj webového provozu na konkrétní stránce. Jako zdroj konkrétní návštěvy se označuje **Session referrer** (**referrer návštěvy**). **Visitor referrer** (**referrer návštěvníka**) je URL první stránky, kterou kdy návštěvník navštívil v rámci všech návštěv sledovaného webu. V případě, že návštěvník zadá URL adresu webu přímo do prohlížeče nebo použije dříve vytvořenou záložku, obsahuje dimenze referrer prázdnou hodnotu [15].

Bounce (**odraz**) je návštěva, která se skládá pouze z jednoho zobrazení stránky. Tato stránka je zároveň vstupní i výstupní stránkou. Velké množství odrazů může naznačovat neúspěch webu přivést návštěvníky, protože plnění cílů webu často vyžaduje zobrazení více stránek [15].

Conversion, goal, **outcome** (**konverze**) je úspěšné splnění cíle webu. **Makro konverze** určuje primární cíl webu (nákup zboží, odeslání poptávky apod.). **Mikro konverze** je dílčí krok makro konverze (process milestone) nebo vedlejší činnost, pomocí které uživatel projevuje zájem o web (stáhnutí propagačních materiálů, přihlášení k odběru novinek apod.) [15, 22].

V této kapitole jsou uvedeny pouze některé dimenze. Za dimenzi lze považovat každý údaj naměřený na webu. Nelze sestavit konečný výčet všech dimenzí, protože průběžně vznikají nové a naopak některé zanikají.

6.4 Metriky

Metrika je číslo, které se používá k měření jedné z vlastností dimenze [23].
Metrika se často používá v kontextu s další dimenzí (například počet návštěv konkrétní stránky). Mohou se lišit mezi dodavateli analytického softwaru. Metriky se nejčastěji rozdělují podle typu (počet nebo poměr) a podle filtrování (agregované, segmentované nebo individuální) [15].

Počet je nejjednodušší jednotka měření. **Poměr** (**míra**) je podíl dvou počtů nebo jiných poměrů. **Agregované metriky** se získávají z celkového provozu webu za určité časové období. **Segmentované metriky** se vypočítávají z celkového provozu webu za určité časové období filtrovaného pomocí dimenzí. **Individuální metriky** se sestavují z aktivit konkrétního návštěvníka za určité časové období [15].

Key Performance Indicator (**klíčové ukazatele výkonnosti**, **KPI**) jsou speciální metriky, které se používají pro měření úspěšnosti webu. Nejčastěji je KPI vyjádřeno jako poměr, ale může to být i počet.

6.4.1 Nejpoužívanější metriky

Následující metriky se vyskytují ve většině analytických nástrojů, ale jejich výpočet se může lišit. Metrika by měla být relevantní ke sledovanému webu, aktuální (web se mění), okamžitě použitelná a neměla by být složitá [16].

Metrika **unique visitors** (**unikátní návštěvníci**) vyjadřuje počet lidí, kteří navštívili sledovaný web. Tato metrika může být z technických důvodů velmi nespolehlivá [15].

Počet hitů (**server request, server call**) je počet požadavků na stažení souboru ze serveru. Například video, obrázek, HTML stránka atp.

Metrika **page views** (**zobrazení stránky**) vyjadřuje počet zobrazení konkrétní stránky. Zaznamenávají se i opakované požadavky na zobrazení stejné stránky. Zobrazení jedné stránky nemusí nutně znamenat pouze jeden požadavek na

stažení souboru z webového serveru, protože webová stránka často vyžaduje načtení další souborů (obrázky, externí soubory apod.) [15, 24].

Metrika **visits, sessions (návštěvy)** vyjadřuje počet návštěv za určité období. **Návštěva** je počet interakcí unikátního návštěvníka, dokud sledovaný web neopustí. Za opuštění webu se považuje neaktivita návštěvníka po určitou dobu. Délka této doby závisí na použitém analytickém softwaru (běžně se používá 30 minut). Návštěva se často skládá z jednoho nebo více zobrazení [15].

Visit duration (**doba trvání návštěvy**) se vypočítá jako rozdíl mezi časem ukončení návštěvy a časem, kdy návštěva začala [15].

New visitors (**noví návštěvníci**) je počet unikátních návštěvníků, kteří poprvé navštívili web během reportovaného období. Metrika **returning visitors** (**vracející se návštěvníci**) vyjadřuje počet unikátních návštěvníků, kteří navštívili web během reportovaného období a alespoň jednou v době před reportovaným obdobím. **Repeat visitors** (**opakovaní návštěvníci**) je počet unikátních návštěvníků, kteří navštívili web během reportovaného období nejméně dvakrát. Návštěvník může být současně nový návštěvník a opakovaný návštěvník nebo vracející se návštěvník a opakovaný návštěvník. Vzhledem k nepřesnosti měření mohou být někteří vracející se návštěvníci nesprávně označeni jako noví návštěvníci [15].

Exit rate, page exit ratio (**míra opuštění**) je procento návštěvníků, kteří opustili konkrétní stránku. **Bounce rate** (**míra okamžitého opuštění**) je procento návštěv ukončených ihned po zobrazení první stránky [15].

Conversion rate (**konverzní poměr**) je podíl počtu návštěv, které vedly ke konverzi a celkového počtu návštěv. Místo počtu návštěv je možné použít i počet unikátních návštěvníků. Pomocí konverzního poměru lze měřit celkovou úspěšnost webu nebo úspěšnost konkrétního segmentu [15].

6.5 Targets

Targets (**cíle KPI**) jsou hodnoty pro jednotlivé klíčové ukazatele výkonnosti, podle kterých se měří úspěšnost nebo neúspěšnost webu. Uvádí se i čas, kdy má ke splnění cíle dojít. Příkladem cíle KPI může být 300 objednávek v e-shopu za příští měsíc. Tento cíl se považuje za splněný v případě, že e-shop měl v daném měsíci minimálně 300 objednávek.

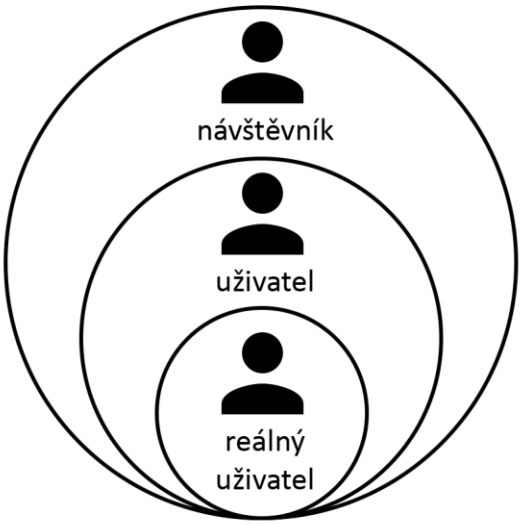
6.6 Segmenty

Segment je část dat z jedné nebo více dimenzí. **Segmentace** je jeden z nejdůležitějších nástrojů webové analytiky. Agregovaná data jsou údaje naměřené v rámci celého webu. Segmentace filtruje agregovaná data podle dimenzí a poskytuje tak konkrétnější informace. Vhodným využitím segmentace může být například segmentace konverzního poměru podle marketingových kampaní. Výpočet konverzní poměru pomocí všech návštěv umožňuje zhodnotit celkovou úspěšnost webu, ale neumožňuje určit příčinu úspěchu. Segmentace podle jednotlivých marketingových kampaní rozlišuje úspěšné a neúspěšné kampaně.

Cross segmentation (**křížová segmentace**) je segmentace, která porovnává jednu množinu dat vzhledem k jiné množině dat.

7 Metody identifikace návštěvníků webu

V současné době neexistuje žádná metoda, která by naprosto spolehlivě a přesně identifikovala **reálného uživatele** webu. Zatím nejspolehlivějším a nejpřesnějším řešením je využití všech možností současně.

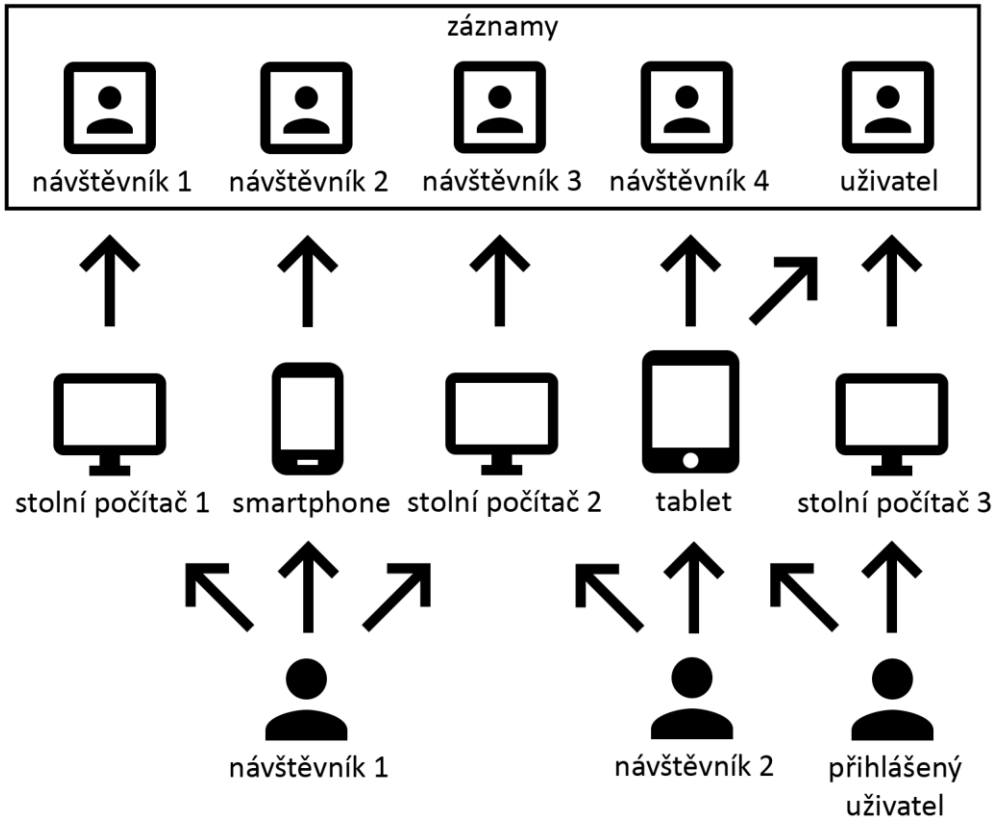


Obrázek 4: Návštěvník, uživatel a reálný uživatel [vlastní tvorba]

Návštěvníka lze do určité míry identifikovat podle jeho chování, unikátního identifikátoru nebo využitím dalších údajů, které odesílá webovému serveru. Některé weby (například sociální sítě) se snaží docílit větší spolehlivosti a přesnosti identifikace zavedením povinné registrace návštěvníka. Pokud návštěvník provede registraci, stává se z něj registrovaný **uživatel**. Uživatele lze rozpoznat podle unikátního identifikátoru uživatele a dalších údajů, které uvedl při registraci. Přestože se uživatel musí vždy identifikovat přihlášením ke svému účtu, není možné ověřit, kdo tento účet fyzicky využívá nebo jestli účet není falešný. Navíc toto částečné řešení není vhodné pro všechny weby (například běžné firemní prezentace).

Provozovatelé českých webů mají povinnost informovat návštěvníky o jejich sledování prostřednictvím registrace na Úřadu pro ochranu osobních údajů a uvedením této skutečnosti v obchodních podmínkách, které se nacházejí přímo na sledovaném webu. Některá pravidla pro sledování návštěvníků definuje Etický kodex webové analytiky, který vytvořili Eric T. Peterson a John Lovett [25].

Obrázek č. 5 znázorňuje, jaké mohou vznikat výrazné rozdíly mezi skutečným provozem sledovaného webu a záznamy vytvořenými měřícím kódem na sledovaném webu. Tyto nesrovnalosti mohou vznikat tím, že více návštěvníků sdílí jedno zařízení, jeden návštěvník používá více zařízení nebo ovlivňuje údaje, které odesílá na web.



Obrázek 5: Provoz sledovaného webu [vlastní tvorba]

7.1 IP adresa

IP adresa je unikátní adresa, která identifikuje zařízení na Internetu nebo na lokální síti. V současnosti existují dva hlavní typy – IPv4 a IPv6 [26].

IP adresa není spolehlivý údaj pro identifikaci návštěvníka webu, protože jednu adresu může sdílet více návštěvníků a lze ji snadno podvrhnout (například pomocí proxy serveru, Virtual Private Network atp.). Z IP adresy je možné za pomocí externích služeb zjistit i další údaje (poskytovatel připojení k internetu, nepřesná poloha návštěvníka atp.). Externí služby využívají databáze obsahující IP adresu, polohu a další údaje. Tyto databáze jsou plněny daty od různých organizací nebo uživatelů, proto jejich údaje mohou být nepřesné a nespolehlivé.

7.2 Unikátní identifikátory

Generování unikátního identifikátoru pro každého návštěvníka je nejpřesnější metoda sledování návštěvníků webu. Rozšířením této metody je generování unikátního identifikátoru pro registrovaného uživatele. Často se používají oba dva identifikátory současně. Nevýhodou tohoto řešení je nespolehlivost při uchovávání identifikátorů. Tato kapitola popisuje nejčastější způsoby uložení unikátních identifikátorů.

HTTP cookie je malé množství dat, které webový server předá prohlížeči pomocí HTTP hlavičky. Prohlížeč potom tato data uloží v zařízení návštěvníka. **Persistent cookies (trvalé cookies)** jsou dočasné soubory, které jsou uloženy v zařízení uživatele, dokud nevyprší jejich platnost nebo je uživatel nevymaže. **Session cookies (dočasné cookies)** jsou dočasné soubory, jejichž platnost vyprší po skončení návštěvy uživatele nebo po uzavření prohlížeče. **First-party cookies (cookies první strany)** jsou dočasné soubory pocházející z navštíveného webového serveru. **Third-party cookies (cookies třetích stran)** jsou dočasné soubory, jejichž obsah pochází z jiného serveru [27, 28].

Session (**Relace**) je malé množství dat, které je uloženo na webovém serveru pomocí textových souborů nebo databáze. Identifikátor relace se přenáší v URL nebo pomocí HTTP cookie [29].

Local shared objects (**flash cookies**) jsou dočasné soubory, které mohou vytvářet webové stránky využívající Adobe Flash. Jejich největší výhodou je spolehlivost. Většina návštěvníků o nich neví, protože se ukládají do jiného adresáře než běžné HTTP cookies [30].

HTML5 WebStorage je lokální úložiště, které se rozděluje podle perzistence na dva druhy. **LocalStorage** ukládá data v prohlížeči, dokud je daný web nebo návštěvník webu neodstraní. **SessionStorage** ukládá data pouze po dobu návštěvy. Výhodou oproti HTTP cookies je snadnější manipulace s daty a větší kapacita dat [31].

Zombie cookie je metoda, která ukládá data do více úložišť najednou. Pokud návštěvník smaže data pouze z některých úložišť, data se obnoví z úložišť, ze kterých je zatím nesmazal. Tento pojem se začal používat v roce 2009, když bylo odhaleno, že společnost Quantcast a několik dalších významných společností zneužívá tuto metodu ke sledování návštěvníků. Nejznámějším příkladem Zombie cookie je Evercookie [32].

ETag je HTTP hlavička, která slouží ke kešování webových souborů. Webový prohlížeč zašle požadavek na stažení souboru z webového serveru. Server zašle prohlížeči soubor společně s etagem obsahujícím kontrolní součet. Pokud prohlížeč znovu požádá o zaslání souboru, zašle s tímto požadavkem i etag. Webový server ověří pomocí kontrolního součtu, zda nedošlo ke změně souboru a zašle odpověď prohlížeči. Tento mechanismus je možné využít ke sledování návštěvníků tím, že se kontrolní součet nahradí unikátním identifikátorem návštěvníka. Obrana proti tomuto způsobu sledování vyžaduje vymazání kešovaných souborů a zakázání kešování v prohlížeči [33].

7.2.1 Browser fingerprint

Browser fingerprint (**otisk prohlížeče**) lze chápat jako alternativu k unikátním identifikátorům z předchozí kapitoly. Otisk prohlížeče je soubor údajů, které webovému serveru poskytuje prohlížeč návštěvníka. Čím více se tyto údaje odlišují od údajů ostatních návštěvníků webu, tím lépe je možné návštěvníka identifikovat. Zajímavou ukázkou možností identifikace návštěvníka pomocí otisku prohlížeče je projekt Panopticlick. Mezi nejčastější používané údaje patří:

- název a verze operačního systému
- název a verze webového prohlížeče
- seznam nainstalovaných doplňků webového prohlížeče
- podpora Javy, Flashe, JavaScriptu a Cookies
- rozlišení obrazovky a barevná hloubka
- seznam nainstalovaných fontů
- preferovaný jazyk návštěvníka

7.2.2 Chování návštěvníka

Návštěvníky webu lze částečně identifikovat podle jejich chování na webu. Tato metoda však není příliš spolehlivá. Návštěvníky, kteří blokují ukládání unikátních identifikátorů a mají shodný otisk prohlížeče, lze v některých případech rozlišit podle navštívených stránek a provedených akcí.

7.3 Hrozby

Všechny výše uvedené údaje je možné podvrhnout. V některých případech je dokonce možné zablokovat jejich záznam. Oklamat nástroj webové analytiky dokáže i běžný uživatel například za pomocí následujících nástrojů.

7.3.1 Doplňky webových prohlížečů

Jedním z nejjednodušších způsobů ochrany soukromí na webu jsou doplňky prohlížečů. Existuje velké množství doplňků jako například Ghostery, NoScript Security Suite, User-Agent Switcher, které umožňují určitou kontrolu soukromí i pro méně zkušené uživatele.

7.3.2 Webové prohlížeče

Tor browser, Epic privacy browser, Comodo dragon jsou webové prohlížeče, které se zaměřují na ochranu soukromí. Většinou se jedná o upravené verze prohlížečů Google Chrome a Mozilla Firefox s předinstalovanými doplňky z předchozí podkapitoly.

7.3.3 Operační systémy

Operační systémy Tails, Whonix a další poskytují lepší úroveň ochrany soukromí než samotné prohlížeče. Kromě již zmíněných prohlížečů nabízejí i další programy pro zajištění soukromí uživatele.

8 Metody sledování návštěvníků webu

Kvalitu sledování návštěvníků ovlivňuje počet záznamů a počet různých údajů v těchto záznamech. S rostoucím počtem sledovaných údajů se nezvyšuje pouze kvalita sledování, ale i náročnost implementace měřícího kódu a nástroje webové analytiky. Nelze jednoznačně určit, která metoda je nejlepší, protože všechny metody uvedené v této kapitole mají svoje výhody a nevýhody.

Pro přehlednější uspořádání údajů v tabulce jsou názvy metod zkráceny (SL – serverové logy, WA – logy webové aplikace, MT - měření tečkou, ZS – značkování stránek, SP - sledování paketů).

Tabulka 1: Přehled metod sledování návštěvníků webu [vlastní tvorba]

	SL	WA	MT	ZS	SP
Lze blokovat záznam	ne	ne	ano	ano	ne
Lze podvrhnout záznam	ano	ano	ano	ano	ano
Viditelné pro návštěvníky	ne	ne	ano	ano	ne
Množství získaných dat	3	3	4	1	2
Náročnost implementace	2	2	1	1	3
Kompatibilita s analytickými nástroji	1	2	3	3	4
Záznam otevření e-mailu	ne	ne	ano	ne	ne
Odolnost proti kešování na straně ISP nebo v prohlížeči	ne	ne	ano	ano	ne
Záznam robotů	ano	ano	omezeně	omezeně	ano

Vlastnosti hybridní metody závisí na kombinaci použitých metod sledování návštěvníků webu. Číselné hodnoty v **tabulce č. 1** vyjadřují hodnocení jednotlivých metod od nejlepších (1) po nejhorší (4). V následujících podkapitolách jsou všechny údaje podrobněji vysvětleny i s uvedením některých výjimek.

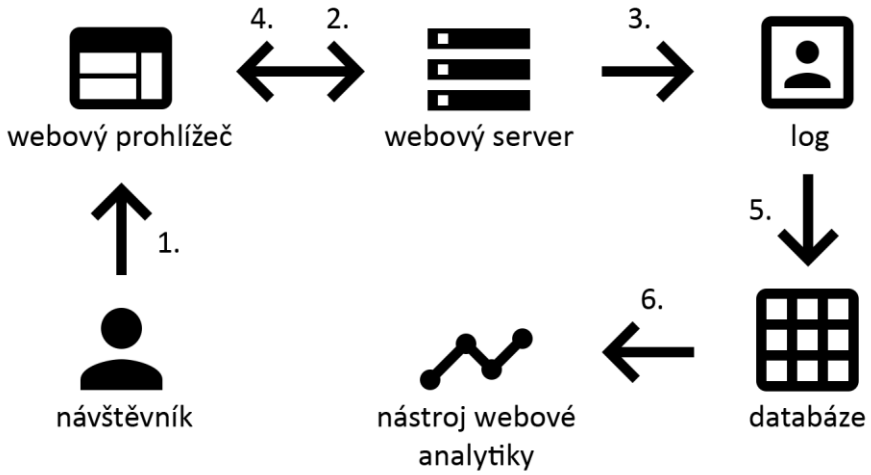
8.1 Web server log file (serverové logy)

Vytváření serverových logů je nejstarší metoda sledování webového provozu, která byla původně vyvinuta pro záznam chyb webových serverů a teprve později se začala používat pro marketingové účely.

Existuje více druhů serverových logů. Pro účely webové analytiky se používají přístupové logy (access logy) zaznamenávající veškeré HTTP požadavky směřující na webový server. Pro vytváření přístupových logů se často používají standardizované formáty, které zajišťují kompatibilitu s analytickými nástroji. Mezi nejznámější formáty patří NCSA Common log format (CLF), NCSA Combined log format a NCSA Separate log format. CLF je formát, který ukládá některé základní údaje z HTTP hlavičky do jediného souboru. Webové servery Apache a ISS používají CLF jako výchozí formát. Combined log format rozšiřuje CLF o další tři údaje (referrer, user_agent a cookie). NCSA Separate log format ukládá data do tří souborů. První soubor má stejné údaje a strukturu jako CLF. Druhý soubor (Referral log) obsahuje datum, čas a HTTP referrer. Do třetího souboru (Agent logu) se zaznamenává datum, čas a údaje o klientovi, který poslal požadavek na webový server [34, 35, 36, 37].

8.1.1 Postup

1. Návštěvník zadá do webového prohlížeče URL adresu požadovaného webu.
2. Prohlížeč zašle požadavek na stažení souboru z webového serveru.
3. Webový server vytvoří záznam o tomto požadavku.
4. Webový server odešle klientovi požadovaný soubor (HTML stránka, multimediální soubor apod.).
5. Obvykle se data před zpracováním parsují ze serverových logů a ukládají do databáze pro opakované použití.
6. Nástroj webové analytiky pracuje s daty z databáze a zobrazuje je uživateli v čitelnější podobě pomocí metrik, tabulek a grafů (viz **Obrázek č. 6**).



Obrázek 6: Serverové logy [vlastní tvorba]

8.1.2 Výhody

Záznam údajů probíhá na straně serveru. Díky tomu je tato metoda pro návštěvníka neviditelná a není možné ji blokovat. Serverové logy umožňují zaznamenávat i přístupy robotů, stavové kódy HTTP protokolu (404 apod.) a veškeré požadavky na stažení souborů. Pokud webový server již využívá vytváření logů, stačí pouze implementovat nástroj pro jejich zpracování. Použití standardních formátů pro vytváření logů usnadňuje přechod k jinému dodavateli nástroje webové analytiky.

8.1.3 Nevýhody

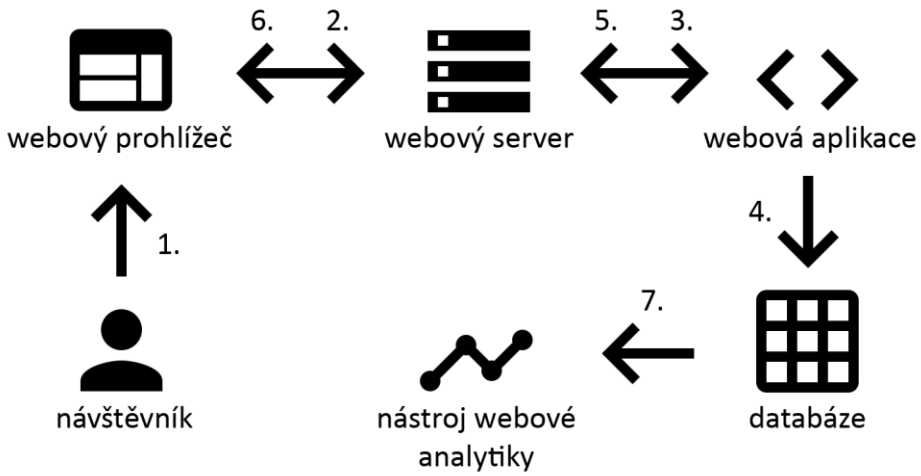
Návštěvník sice nemůže blokovat záznam dat, ale může podvrhnout odeslaná data. Pokud dochází ke kešování na straně ISP nebo webového prohlížeče, nemusí vždy dojít k zaslání požadavku na webový server a vytvoření záznamu na serveru. Tato metoda nepodporuje záznam událostí (přehrávání videa atp.).

8.2 Logy webové aplikace

Vytváření záznamů na úrovni webové aplikace pomocí serverových skriptovacích jazyků (PHP, Java apod.) umožňuje zaznamenávat více údajů o stavech webové aplikace (chybné vyplnění formuláře na webu apod.) než metoda serverových logů. Výhodou je i možnost využití více druhů úložišť. Získaná data se ukládají do logů nebo nejčastěji přímo do databáze.

8.2.1 Postup

1. Návštěvník zadá do webového prohlížeče URL adresu požadovaného webu.
2. Prohlížeč zašle požadavek na stažení webové stránky ze serveru.
3. Webový server spustí webovou aplikaci.
4. Webová aplikace vytvoří záznam o požadavku návštěvníka.
5. Webová aplikace sestaví webovou stránku.
6. Webový server odešle klientovi požadovanou webovou stránku.
7. Nástroj webové analytiky zobrazí data přímo z databáze nebo je parsuje z logů (viz **Obrázek č. 7**).



Obrázek 7: Logy webové aplikace [vlastní tvorba]

8.2.2 Výhody

Logy webové aplikace zaznamenávají i přístupy robotů a chybové stavy aplikace. Metoda je pro návštěvníka neviditelná. V případě použití databáze není nutné parsovat logy a je možné zaznamenané údaje rovnou zobrazit v analytickém softwaru.

8.2.3 Nevýhody

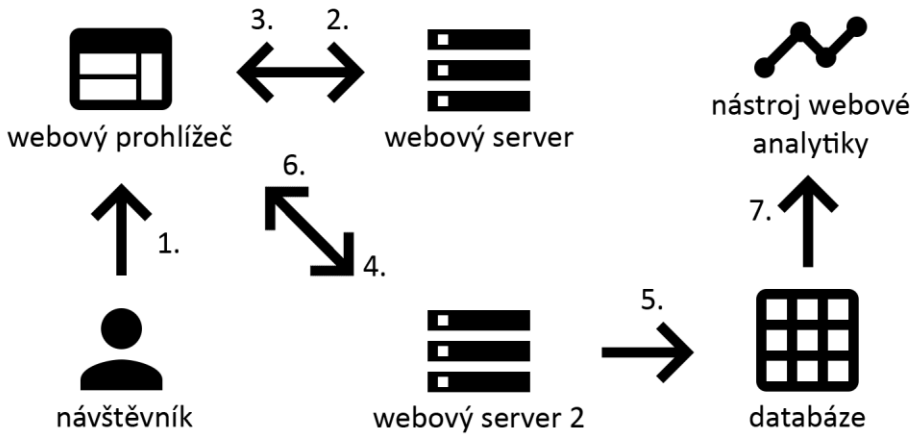
Návštěvník nemůže blokovat záznam dat, ale může podvrhnout odeslaná data. Pokud dochází ke kešování na straně ISP nebo webového prohlížeče, nemusí vždy dojít k zaslání požadavku na webový server a vytvoření záznamu. Tato metoda nepodporuje záznam událostí (přehrávání videa atp.).

8.3 Web beacons (měření tečkou)

Tato metoda využívá zobrazení obrázku (tracking pixel) na sledované webové stránce k zaslání požadavku na webový server. Návštěvník sledovaného webu se může dozvědět o použití této metody zobrazením zdrojového kódu. Nejčastější formou tracking pixelu je transparentní obrázek o velikosti jednoho pixelu na šířku i výšku. Skrytí tracking pixelu lze provést i pomocí kaskádových stylů.

8.3.1 Postup

1. Návštěvník zadá do webového prohlížeče URL adresu požadovaného webu.
2. Prohlížeč zašle požadavek na stažení webové stránky ze serveru.
3. Webový server odešle návštěvníkovi požadovanou webovou stránku spolu s odkazem na obrázek (tracking pixel).
4. Webový prohlížeč zašle požadavek (stažení obrázku) na stejný webový server, jiný server společnosti nebo server třetí strany (záleží na zvolené implementaci).
5. Webový server, který obdržel požadavek, vytvoří záznam.
6. Webový server zašle prohlížeči požadovaný obrázek.
7. Nástroj webové analytiky zobrazí data přímo z databáze nebo je parsuje z logů (viz **Obrázek č. 8**).



Obrázek 8: Měření tečkou a značkování stránek [vlastní tvorba]

8.3.2 Výhody

V případě použití měřícího kódu třetí strany se jedná o nejjednodušší způsob implementace. Metoda umožňuje sledovat návštěvníky napříč různými weby, pokud zasílají údaje na společný webový server. Měření tečkou je možné použít i pro záznam otevření e-mailu.

8.3.3 Nevýhody

Měřící kódy nejsou kompatibilní mezi dodavateli analytického softwaru. Metoda neumožňuje zaznamenat přístup k jiným souborům, než jsou webové stránky. Pokud návštěvník zakáže zobrazení obrázků na webu, není možné ho sledovat. Pomocí metody měření tečkou nelze zaznamenat některé roboty, protože nestahují obrázky. V některých případech může kešování obrázků bránit záznamu údajů. Ke kešování nedojde v případě, že se k požadavku na stažení obrázku přidají i další údaje, které se mění při každém zobrazení stránky.

8.4 Page tagging (značkování stránek)

V dnešní době je značkování stránek nejpoužívanější metoda sledování návštěvníků. Získávání údajů probíhá na straně klienta pomocí JavaScriptu.

8.4.1 Postup

1. Návštěvník zadá do webového prohlížeče URL adresu požadovaného webu.
2. Prohlížeč zašle požadavek na stažení webové stránky ze serveru.
3. Webový server odešle návštěvníkovi požadovanou webovou stránku spolu s měřícím skriptem.
4. Skript získá informace o návštěvníkovi a zašle je formou požadavku na stejný webový server, jiný server společnosti nebo server třetí strany (záleží na zvolené implementaci).
5. Webový server, který obdržel požadavek, vytvoří záznam.
6. Webový server zašle odpověď webovému prohlížeči.
7. Nástroj webové analytiky zobrazí data přímo z databáze nebo je parsuje z logů (viz **Obrázek č. 8**).

8.4.2 Výhody

V případě použití měřícího kódu třetí strany se jedná o nejjednodušší způsob implementace. Metoda umožňuje získávat nejvíce typů údajů. Značkování stránek umožňuje sledovat návštěvníky napříč různými weby, pokud zasílají údaje na společný webový server. Metoda je odolná proti kešování na straně ISP i webového prohlížeče.

8.4.3 Nevýhody

Měření je možné snadno podvrhnout nebo dokonce zablokovat. Návštěvník sledovaného webu se může dozvědět o použití této metody zobrazením zdrojového kódu. Metoda nedokáže zaznamenat většinu robotů, protože jen někteří z nich podporují JavaScript. Je výhodnější zaznamenávat i přístupy robotů, protože mohou poskytovat cenné informace (viz kapitola Filtrování robotů). Měřící kódy nejsou kompatibilní mezi dodavateli nástrojů webové analytiky.
Značkování stránek není příliš spolehlivé, protože [38]:

- uživatelé mohou mít v prohlížeči vypnutou podporu JavaScriptu
- uživatel webu může přejít na jinou stránku dřív, než se spustí měřící kód
- chyba v jiném skriptu na webu znemožní záznam návštěvníka

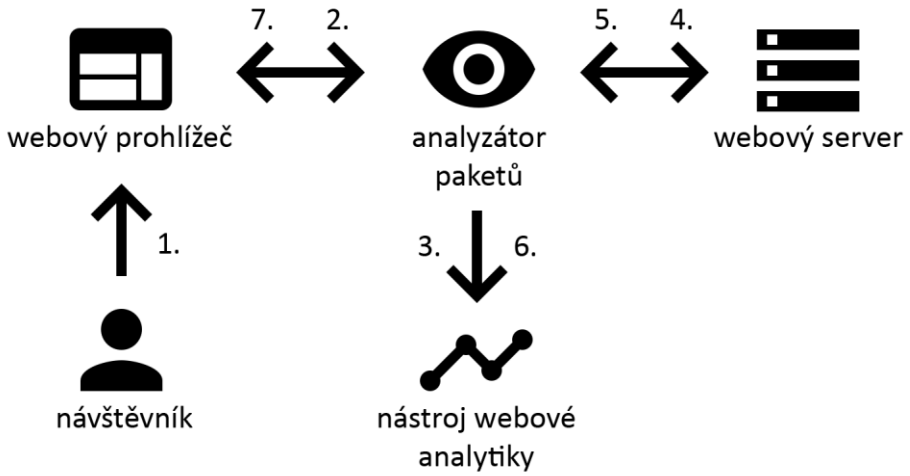
8.5 Packet sniffing (sledování paketů)

Paketové sniffery (**analyzátory paketů**) jsou programy nebo zařízení, které zaznamenávají požadavky směřující na webový server. Tato metoda není příliš oblíbená z důvodu náročnosti implementace.

8.5.1 Postup

1. Návštěvník zadá do webového prohlížeče URL adresu požadovaného webu.
2. Prohlížeč zašle požadavek na stažení souboru z webového serveru.
3. Analyzátor paketů zaznamená požadavek webového prohlížeče.
4. Webový server obdrží požadavek.

5. Webový server odešle klientovi požadovaný soubor (HTML stránka, multimediální soubor apod.).
6. Analyzátor paketů zaznamená odpověď webového serveru.
7. Webový prohlížeč obdrží odpověď (viz **Obrázek č. 9**).



Obrázek 9: Sledování paketů [vlastní tvorba]

8.5.2 Výhody

Sledování paketů umožňuje zaznamenávat i přístupy robotů, dobu mezi odesláním požadavku a odpovědí webového serveru, požadavky na stažení souborů a stavové kódy HTTP protokolu (404 apod.). Metoda je pro návštěvníka neviditelná a nemůže blokovat záznam dat, ale může podvrhnout odeslaná data.

8.5.3 Nevýhody

Nevýhodou sledování paketů je náročnost implementace. Pokud dochází ke kešování na straně ISP nebo webového prohlížeče, nemusí vždy dojít k zaslání požadavku na webový server a vytvoření záznamu.

8.6 Hybridní metody

Spojením několika metod sledování návštěvníků je možné zvýšit množství zaznamenaných typů údajů, zpřesnit jejich měření a zvýšit spolehlivost měření. Nevýhodou této metody je větší náročnost implementace. Kombinací několika metod měření dochází k vícenásobnému záznamu datu. Řešením je kontrola již

existujících záznamů před vytvořením nového záznamu nebo jejich sloučení. Konkrétní vlastnosti této metody závisí na kombinaci použitých metod.

8.7 Přehled dostupných údajů

Tabulka č. 2 porovnává jednotlivé metody sledování návštěvníků webu, vzhledem k jejich možnostem získávání údajů o návštěvnících. Skutečný rozsah získaných údajů závisí na konkrétní implementaci zvolených metod.

Tabulka 2: Porovnání metod sledování návštěvníků webu [vlastní tvorba]

Údaj	SL	WA	MT	ZS	SP
IP adresa návštěvníka	ano	ano	ano	ano	ano
unikátní identifikátor návštěvníka	ano	ano	ano	ano	ano
unikátní identifikátor uživatele	ano	ano	ano	ano	ano
URL adresa požadované stránky	ano	ano	ano	ano	ano
události (přehrávání videa apod.)	ne	ne	ne	ano	ne
HTTP referrer	ano	ano	ano	ano	ano
stavové kódy HTTP protokolu (404 apod.)	ano	ano	ano	ne	ano
chybové stavy aplikace (chybné vyplnění formuláře apod.)	ne	ano	ne	ne	ne
datum a čas zaslání požadavku na server	ano	ano	ne	ne	ano
lokální datum a čas návštěvníka	ne	ne	ne	ano	ne
operační systém (název a verze)	ano	ano	ano	ano	ano
webový prohlížeč (název a verze)	ano	ano	ano	ano	ano
nainstalované doplňky webového prohlížeče	ne	ne	ne	ano	ne
model mobilního zařízení (název a zařízení)	ano	ano	ano	ano	ano
stažení souborů ze serveru (multimédia apod.)	ano	ne	ne	ne	ano
rozlišení obrazovky	ne	ne	ne	ano	ne
maximální využitelná plocha obrazovky	ne	ne	ne	ano	ne
barevná hloubka	ne	ne	ne	ano	ne
HTML5 geolokace	ne	ne	ne	ano	ne
podporované fonty	ne	ne	ne	ano	ne
podpora Javy, Flashe, JavaScriptu apod.	ne	ne	ne	ano	ne
podpora HTTP cookies	ano	ano	ano	ano	ano
podpora HTML5 storage	ne	ne	ne	ano	ne
preferovaný jazyk návštěvníka	ano	ano	ano	ano	ano
pohyb kurzoru	ne	ne	ne	ano	ne
doba odezvy serveru	ne	ne	ne	ne	ano
zapnuté Do Not Track	ne	ne	ne	ano	ne
celkový počet podporovaných údajů	13	13	11	22	14

Údaje získané hybridní metodou závisí na zvolené kombinaci použitých metod sledování návštěvníků webu.

9 Metody filtrování robotů

9.1 Robot

Agent *je výpočetní entita, která autonomně jedná jménem jiných entit, vykonává svou činnost s určitou mírou proaktivity nebo reaktivity, projevuje určitou úroveň klíčových atributů učení, spolupráce a mobility* [39].

Robot je druh internetového agenta. Roboty lze podle účelu rozdělit do dvou hlavních skupin na dobré roboty a škodlivé roboty. Softwarovým agentem je i webový prohlížeč, který návštěvník používá pro přístup k webovému serveru.

Crawler je robot, který prochází internetové dokumenty, aby našel nové informace a odhalil strukturu webu. Tento druh robota často využívají vyhledávače k vytváření indexu. Roboti vyhledávačů jsou na webu žádoucí, ale existují i roboti, kteří kopírují obsah webu bez svolení autora [40].

Webový spambot *je druh internetového robota, který se používá převážně k šíření nevyžádaného obsahu na webu* [41].

Hacking tools jsou nejnebezpečnější formou robotů. Tyto nástroje automaticky provádějí různé typy útoků. Mezi časté útoky patří SQL injekce, Cross-site scripting, útoky hrubou silou na administrace redakčních systémů (WordPress, Joomla apod.).

Malicious pollution je metoda, pomocí které útočník spamuje napadený web prostřednictvím falešných přístupů. Důvodem může být například snaha znehodnotit naměřená data na napadeném webu.

Referral spam je druh malicious pollution. Útočník propaguje svůj web pomocí falešných přístupů na napadeném webu. Při zasílání HTTP požadavku na napadený webový server útočník vloží URL adresu svého webu do hlavičky referrer. Analytik, který pracuje s analytickým nástrojem, uvidí tento odkaz v reportu a přejde na útočníkovu stránku, aby zjistil více informací o tomto zdroji návštěv. V případě, že

napadený web zveřejňuje statistiky přístupů a jsou přístupné pro roboty vyhledávačů, může docházet i ke zlepšení hodnocení útočníkova webu v rámci vyhledávačů. Útočníkův web často nemá dobrou pověst a v těchto případech může docházet i k poškození hodnocení napadeného webu, protože na něj odkazuje. V posledních letech se objevuje varianta, kdy útočník spamuje přímo měřící kód Google analytics nebo měřící kód jiného nástroje webové analytiky bez toho, aby navštívil napadený web [42].

Roboti se nerozlišují na dobré a škodlivé pouze podle jejich činnosti, ale i podle toho, jestli respektují pravidla daného webu (nevstupují na zakázané stránky apod.) a jsou ochotni se identifikovat. Pokud se robot vydává za běžného návštěvníka webu, dochází k zanesení těchto nepravdivých informací do nástroje webové analytiky.

Podle studií společnosti Incapsula za posledních několik let tvoří roboti podstatnou část internetového provozu. Podrobnější údaje jsou uvedeny v **tabulce č. 3**.

Tabulka 3: Vliv robotů na provoz webu [43, 44, 45]

rok	2013	2014	2015
počet testovaných webů	20000	20000	35100
počet bilionů návštěv	1,45	15	19
lidé	38,5%	44%	51,5%
dobří roboti	31%	27%	19,5%
škodliví roboti	30,5%	29%	29%

Existují metody, které umožňují do určité míry zakázat vstup robotům na web nebo zamezit jejich zaznamenání měřícím kódem. Mnohem vhodnějším řešením je záznam veškerého webového provozu a následné odfiltrování nežádoucích přístupů. Záznam robotů poskytuje přehled o indexaci webu a o bezpečnostních hrozbách.

9.2 Filtrování dobrých robotů

Protokol pro zakázání přístupu robotům je textový soubor, který se nachází v hlavním adresáři webu a je pojmenovaný „robots.txt“. Tento soubor umožňuje zakázat robotům přístup na web. Je možné vytvořit pravidla jen pro některé roboty nebo některé stránky webu [46].

Zakázání indexace nebo procházení webové stránky lze robotům sdělit i pomocí meta tagu v hlavičce HTML dokumentu nastavením atributu rel na hodnoty noindex (neindexovat) nebo nofollow (neprocházet). Atribut rel s hodnotou nofollow lze nastavit i pro jednotlivé odkazy.

Zaznamenané dobré roboty je možné identifikovat pomocí údaje user-agent, který se nachází v hlavičce HTTP požadavku. Dobří roboti většinou uvádějí svůj název a v ideálním případě i URL adresu webu s podrobnějšími informacemi a s postupem jak zakázat přístup tohoto robota na web.

9.3 Filtrování škodlivých robotů

Metody pro filtrování dobrých robotů nelze použít pro zakázání přístupu škodlivých robotů, protože dodržování těchto pravidel není povinné.

9.3.1 Captcha

Captcha (Completely Automated Public Turing test to tell Computers and Humans Apart) je kompletně automatický veřejný Turingův test k rozlišení člověka od počítače. Test nejčastěji probíhá tak, že návštěvník opíše kód ze špatně čitelného obrázku. Nevýhodou tohoto řešení je nespolehlivost rozpoznávání robotů a výrazné zhoršení uživatelské přívětivosti webu. Existují firmy, které nabízejí hromadné obcházení této kontroly pomocí lidských pracovníků žijících v zemích s levnou pracovní silou (například Indie) [47, 48].

9.3.2 Rozpoznávání podle jednoduchých filtrů

Filtrování podle jednoduchých filtrů je nejjednodušší metoda rozpoznávání robotů. Na základě údajů získaných z již identifikovaných robotů se vytvoří pravidla, která je rozliší při další návštěvě. Tato metoda je jednoduchá na implementaci, ale

náročná na správu. Množství použitých pravidel je možné výrazně snížit pomocí regulárních výrazů. Metoda je vhodná pro rozpoznávání robotů, které často navštěvují web a jejich údaje se výrazně odlišují od reálných uživatelů.

9.3.3 Strojové učení

Strojové učení je oblast matematiky a informatiky zkoumající metody učení strojů. Oblast využití strojového učení pokrývá v podstatě všechny obory lidské činnosti [49].

Strojové učení je vhodné pro vytvoření nástrojů, které efektivně identifikují i těžko rozpoznatelné roboty. Tyto nástroje kontrolují jednotlivé údaje návštěvníků a na základě jejich vah vyhodnocují, jestli se jedná o robota nebo skutečného uživatele.

Naivní bayesovský klasifikátor je řešení, které na rozdíl od neuronových sítí a genetických algoritmů nevyžaduje převod vstupních údajů na zvláštní formát. Jedinou nutnou úpravou dat je normalizace některých údajů, aby nedocházelo k vícenásobnému ukládání údajů se stejným obsahem, ale odlišnou formou (například označení „firefox“ a „Firefox“ pro konkrétní webový prohlížeč). Výpočet pravděpodobnosti, jestli je návštěvník člověk nebo robot se provádí podle následujícího vzorce.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}; P(B) > 0 \qquad \textbf{(1)}$$

Jev A1 vyjadřuje skutečnost, že návštěvník je robot. Lidský návštěvník je označen jako jev A2. Vlastnosti návštěvníka jsou zde uvedeny jako jev B. P(A) označuje pravděpodobnost jevu A. P(B) je pravděpodobnost jevu B. P(A|B) označuje pravděpodobnost jevu A, jestliže nastal jev B. P(B|A) je pravděpodobnost jevu B, jestliže nastal jev A [50].

9.3.4 Rozpoznávání podle chování

Roboty je možné rozeznat i podle toho, jak se pohybují na webu. Na rozdíl od většiny lidských návštěvníků prohledávají web systematicky a často navštěvují

nesmyslné kombinace webových stránek. Častým znakem jsou i rychlé přechody z jedné webové stránky na druhou.

9.3.5 Bot trap

Bot trap (**past na roboty**) je metoda, která umožňuje identifikaci robotů, kteří nerespektují pravidla webu. Do souboru robots.txt se vloží příkaz, aby roboti nenavštěvovali určitý adresář na webu. Tento adresář nemusí ve skutečnosti na webu existovat a často na něj nevedou žádné odkazy. V případě, že robot tento adresář navštíví, zaznamená se tento přístup měřícím kódem a robot je identifikován jako škodlivý robot.

Honeypot je velmi efektivní metoda pro ochranu webových formulářů před škodlivými roboty. Metoda spočívá ve vytvoření falešného pole ve webovém formuláři. Falešné pole není viditelné pro lidské návštěvníky webu, protože je skryto pomocí kaskádových stylů. Robot, který vyplní toto skryté pole, se identifikuje ihned po odeslání formuláře [51].

10 Analýza dat

Reportování je proces organizace dat do informačních přehledů tak, aby bylo možné sledovat, jak se daří jednotlivým oblastem podnikání [52].

Analýza je proces zkoumání dat a reportů za účelem získání smysluplných znalostí, které mohou být použity k lepšímu porozumění a zlepšení výkonnosti podniku [52].

10.1 Trychtýřová analýza (Funnel analysis)

Trychtýřová analýza je zkoumání přednastavené množiny kroků [53].
Jedná se o kroky (mikro konverze), které vedou k dokončení předem definovaného cíle webu. Každý krok znázorňuje jednu konkrétní stránku nebo událost. Zobrazením všech kroků v jednom diagramu vznikne trychtýř. Celý diagram má tvar trychtýře, protože nejvíce návštěvníků je na začátku procesu. V dalších krocích návštěvníci postupně odcházejí mimo web, a proto se hrdlo trychtýře zužuje. Trychtýřová analýza umožňuje odhalit slabá místa na cestě návštěvníka směrem k cíli. Při trychtýřové analýze se pracuje výhradně s agregovanými daty. Pro přesnější měření je vhodné rozdělit celý proces na co nejmenší kroky. Některé weby například zobrazují obsah nákupního košíku a výběr dopravy na stejné stránce. Pokud tyto dva rozdílné kroky nejsou nějakým způsobem rozlišeny, zobrazí se v rámci trychtýřové analýzy společně a není možné je od sebe rozeznat.

Obrázek 10: Trychtýřová analýza [vlastní tvorba]

10.2 Analýza cesty návštěvníků (Path analýza)

Graf je uspořádaná dvojice G = (V, E), kde V je množina vrcholů a E je množina hran – množina vybraných dvouprvkových podmnožin množiny vrcholů [54].

Orientovaný graf je uspořádaná dvojice D = (V, E), kde $E \subseteq V \times V$ [54].

Vážený graf je graf G spolu s ohodnocením w hran reálnými čísly $w : E(G) \rightarrow R$ [54].

Kladně vážený graf G, w je takový, že $w(e) > 0$ *pro všechny hrany e* [54].

Analýza cesty návštěvníků se často znázorňuje formou seznamu po sobě jdoucích stránek nebo událostí. Pro uživatele je mnohem přehlednější a přívětivější formou znázornění pomocí kladně váženého orientovaného grafu. Webové stránky a události jsou zobrazeny v podobě vrcholů. Přechody mezi nimi tvoří hrany grafu. Každá hrana obsahuje číslo, které vyjadřuje počet přechodů mezi stránkami a událostmi spojenými touto hranou. Na rozdíl od trychtýřové analýzy umožňuje analýza cesty návštěvníků pracovat nejen s agregovanými daty, ale i s daty konkrétních návštěvníků. Při práci s agregovanými daty je vhodné použít segmentaci.

10.2.1 Využití

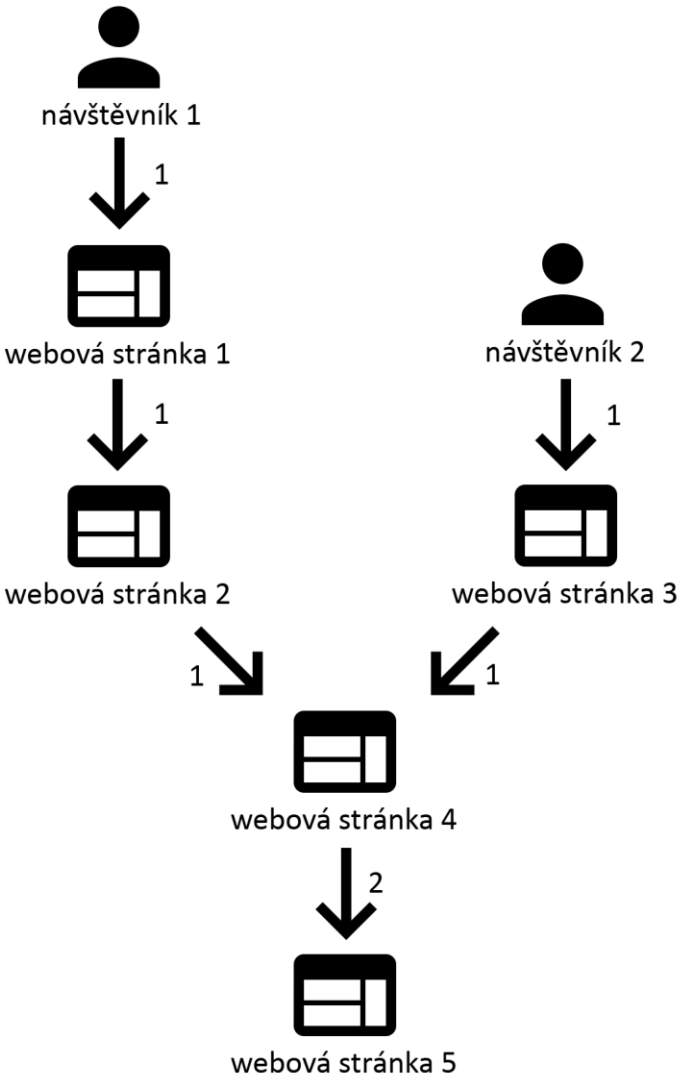
Analýza cesty návštěvníků se využívá například pro zjištění, odkud přichází nejvíce návštěvníků nebo pro identifikaci stránek a událostí, které vedly ke konverzi.

A/B TESTING (**split testing, A/B testování**) je metoda porovnávání dvou variant jedné webové stránky. Cílem testování je zvýšení konverzního poměru testované stránky. Testování spočívá ve střídavém zobrazování obou variant stránky. Při každém zaslání požadavku na stažení stránky z webového serveru se zobrazí jiná varianta. Po získání dostatečného počtu naměřených dat se vyhodnotí, která varianta přivedla více konverzí a implementuje se jako nová verze testované stránky. A/B testování se vždy využívá pro testování jedné vlastnosti jednoho

prvku na webové stránce. Další vlastnosti a prvky je nutné vyčlenit do dalších testů.

Multivariate testing (**multivariantní testování**) na rozdíl od A/B testování umožnuje testovat více prvků současně a najít tak ideální kombinaci prvků. Nevýhodou multivariantního testování je potřeba většího počtu naměřených dat než u A/B testování.

Obrázek č. 11 znázorňuje zobrazení cesty návštěvníka na webu pomocí kladně váženého orientovaného grafu.



Obrázek 11: Analýza cesty návštěvníků [vlastní tvorba]

11 Praktická část

Praktická část se skládá z měřícího kódu, testovacího webu, robota a nástroje webové analytiky. Všechny tyto části jsou spolu propojeny, ale zároveň je možné každou z nich nahradit vlastní aplikací. Robot prochází testovací web a při každém zaslání požadavku na načtení stránky aktivuje měřící kód. Měřící kód zaznamená získané údaje do databáze. Nástroj webové analytiky, který je hlavním přínosem této práce, pomáhá zpracovat naměřená data a zobrazuje je uživateli. Při nasazení v praxi se testovací web nahradí webem, který má být sledován. Robota potom nahradí návštěvníci sledovaného webu.

11.1 Výhody nového řešení

V současnosti existuje mnoho nástrojů webové analytiky, které sice mají některé z následujících výhod, ale nemají všechny výhody současně. Nový nástroj webové analytiky se nesnaží konkurovat stávajícím nástrojům, ale doplnit funkce, které zatím nemají.

11.1.1 Agregovaná data i jednotlivé záznamy

Google Analytics a další nástroje zobrazují pouze agregovaná data a neumožňují přístup k jednotlivým návštěvám a návštěvníkům. Agregovaná data sice mají svoje využití, zejména při použití segmentace, ale neumožňují pozorovat konkrétní zákazníky a jejich vlastnosti. Nástroj, který je cílem této práce, umožňuje přistupovat k údajům všech návštěvníků.

11.1.2 Kontrola nad daty

Velmi častým problémem nástrojů webové analytiky je automatické čištění dat a zamezení přístupu analytika k tomuto procesu. Nový analytický nástroj umožňuje přístup i k datům, které nebyly nijak ošetřeny. Výhodou tohoto řešení je ověření funkčnosti měřícího kódu a možnost sledování i nevyžádaných přístupů na web. Data jsou potom dále upravena pod dohledem analytika. Tento způsob může být mnohem náročnější než plně automatické čištění, ale vede k lepší kvalitě dat a kontrole nad procesem čištění dat.

11.1.3 Okamžitá funkčnost

Smysluplné využití agregovaných dat vyžaduje velké množství naměřených záznamů. Některé weby nemohou mít vysokou návštěvnost, protože jsou úzce zamřené nebo nemají žádnou historii. Nový nástroj webové analytiky poskytuje užitečné informace již po příchodu prvního návštěvníka.

11.1.4 Zabezpečení webu

Záznam veškerého provozu včetně robotů a nevyžádaných přístupů může pomoci odhalit bezpečnostní hrozby (SQL injekce, hledání administrace redakčních systémů atp.). Informace o těchto útocích lze nalézt přímo v naměřených datech. Blokování identifikovaných útoků není součástí této diplomové práce. Další výhodou tohoto nástroje je ukládání dat přímo na sledovaném webu. Naměřená data se neposílají třetím stranám.

11.1.5 Přehledné zobrazení dat

Zobrazení webového provozu formou kladně váženého orientovaného grafu umožňuje lépe pochopit zobrazení celkového provozu na webu nebo chování konkrétního návštěvníka webu.

11.1.6 Testování webu

Zobrazením průchodu konkrétního zákazníka webem je možné odhalit různé druhy chyb včetně nepřehlednosti navigace webu. Výhodou nového nástroje je i možnost A/B testování a multivariantního testování. Úspěšnost jednotlivých variant lze posoudit na základě grafu celkového provozu.

11.1.7 Snadné rozšíření

Google Analytics a podobné nástroje neumožňují přidávání dalších funkcí nebo jakékoliv jiné složitější úpravy. Nový nástroj je možné libovolně upravovat a rozšiřovat.

11.2 Tracker

Pro sledování návštěvníků webu se využívá hybridní metoda. Tracker je PHP třída pro měření základních údajů o návštěvnících sledovaného webu. Měření probíhá

na straně serveru a získané údaje se ukládají přímo do databáze. Po načtení webové stránky se spustí JavaScriptový kód, který zašle asynchronní požadavek na webový server. Následně se znovu spustí třída Tracker, která rozšíří již existující záznam návštěvníka o nově získané údaje nebo vytvoří nový údaj v případě, že nedošlo k první aktivaci Trackeru (například z důvodu kešování). JavaScriptové rozšíření je možné použít i pro záznam událostí. Tracker lze nahradit vlastním měřícím kódem.

11.3 Přehled

Přehled zobrazuje návštěvnost za určité období formou liniového grafu. Každý bod v grafu reprezentuje počet unikátních návštěvníků za jeden den. Pod liniovým grafem se nacházejí výsečové grafy popisující nejčastější preferované jazyky návštěvníků a další údaje.

11.4 Zpracování dat

Data jsou rozdělena do tří skupin. První skupina obsahuje záznamy vytvořené měřícím kódem. Tyto záznamy nejsou žádným způsobem upraveny. Druhá skupina zahrnuje záznamy, které již byly očištěny od dobrých robotů. Třetí skupina slouží k archivaci vyčištěných záznamů. Záznam ve třetí skupině se označuje jako zobrazení, protože odpovídá zobrazení webové stránky.

11.4.1 Záznamy

První fáze čištění dat slouží k odfiltrování dobrých robotů (roboti vyhledávačů apod.). Filtrování je prováděno pomocí regulárních výrazů, které lze nastavit ve správě filtrů. Bez použití funkce pro detekci robotů je možné roboty přesunout do dalších fází čištění i s ostatními záznamy.

11.4.2 Zpracované záznamy

V druhé fázi čištění dat se záznamy automaticky převádějí na zpracované záznamy. Zpracované záznamy obsahují nové dimenze (název prohlížeče, verze prohlížeče a operační systém), které se získávají z pole user_agent. Dochází také k úpravě dimenze preferovaný jazyk.

Během převodu na zpracované záznamy dochází ke spojení záznamů tak, aby tvořily akce jednoho návštěvníka. Spojení záznamů probíhá porovnáním jejich společných údajů a vyhodnocením pravidel, která jsou pro tyto údaje nastavena.

Hlavním účelem druhé fáze je filtrování skrytých robotů. Filtrování robotů funguje na podobném principu jako filtry v první fázi. Na rozdíl od filtrů v první fázi neodhalí pouze jednotlivé záznamy, ale i konkrétního návštěvníka. Druhý způsob využívá naivní bayesovský klasifikátor, který je nutné nejdříve zaučit pomocí ručního roztřídění některých záznamů.

11.4.3 Zobrazení

Poslední fáze čištění dat je určena k archivaci a nepovoluje žádné úpravy. Během převodu zpracovaných záznamů na zobrazení stránek se automaticky provede propojení nově zpracovaných zobrazení s již archivovanými zobrazeními. K propojení se používají pouze unikátní identifikátory.

11.4.4 Testování metod zpracování dat

Pro ověření funkčnosti použitých metod zpracování dat byly provedeny následující testy. K testování metod byl použit soubor dat, který obsahoval 1000 záznamů naměřených na webu během 4 dnů. Sledovaný web patří společnosti, která se zaměřuje na úzkou skupinu zákazníků. Z tohoto důvodu byl soubor tvořen především roboty. Každý test zahrnoval použití metody na naměřených datech a následné vyhodnocení výstupů této metody.

Tabulka 4: Test rozpoznávání dobrých robotů [vlastní tvorba]

Počet vstupních záznamů	1000
Počet záznamů dobrých robotů	629
Počet záznamů ostatních návštěvníků	371
Počet správně zařazených záznamů	1000
Počet chybně rozpoznaných záznamů	0

Na základě naměřených údajů byly sestaveny filtry. Po aplikaci filtrů byly všechny záznamy ručně zkontrolovány a vyhodnoceny. Podle výsledků v **tabulce č. 4** bylo rozhodnuto, že implementace metody rozpoznávání dobrých robotů pomocí jednoduchých filtrů byla úspěšná.

Tabulka 5: Test spojování záznamů [vlastní tvorba]

Počet vstupních záznamů	371
Počet vytvořených skupin	71
Počet správně zařazených záznamů	371
Počet chybně zařazených záznamů	0

Vyčištěný soubor dat byl převeden na zpracované záznamy. Během zpracování záznamů došlo i k jejich spojení do skupin. Následně bylo u každého záznamu vyhodnoceno, zda patří do přiřazené skupiny. Během kontroly nebyly nalezeny žádné chybně zařazené záznamy (viz **tabulka č. 5**).

Tabulka 6: Test rozpoznávání škodlivých robotů [vlastní tvorba]

Počet vstupních záznamů	371
Počet záznamů škodlivých robotů	225
Počet záznamů ostatních návštěvníků	146
Počet správně zařazených záznamů	371
Počet chybně rozpoznaných záznamů	0

Test rozpoznávání škodlivých robotů byl proveden stejným způsobem jako test rozpoznávání dobrých robotů, protože obě dvě metody fungují na stejném principu. Jediným rozdílem bylo použití odlišných filtrů. Výsledek testu prokázal úspěšnost testované metody (viz **tabulka č. 6**).

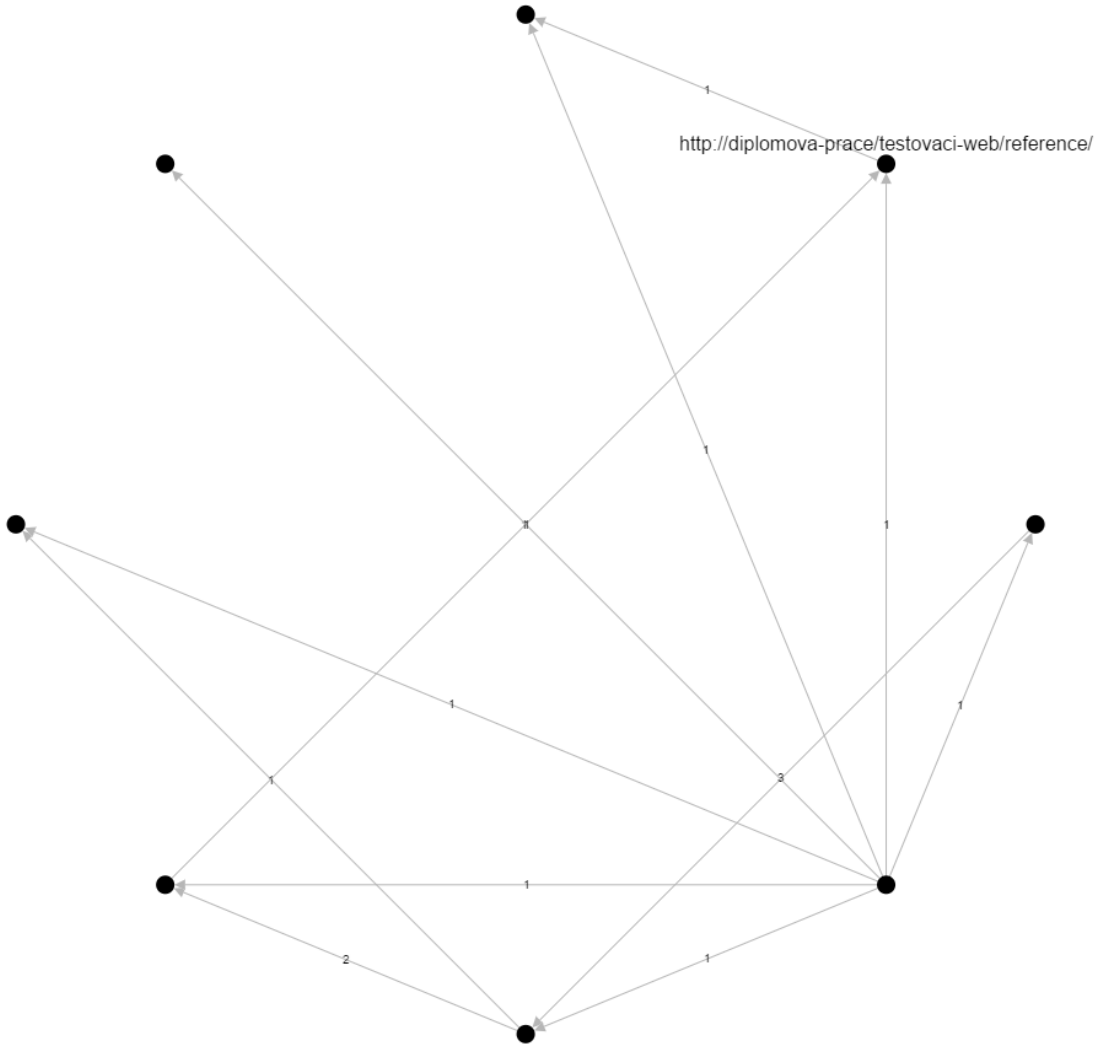
Tabulka 7: Test naivního bayesovského klasifikátoru [vlastní tvorba]

Typ dat	Trénovací data	Testovací data
Počet vstupních záznamů	200	171
Počet záznamů škodlivých robotů	131	94
Počet záznamů ostatních návštěvníků	69	77
Počet správně určených záznamů		161
Počet nesprávně určených záznamů		10

Poslední test ověřil úspěšnost metody rozpoznávání škodlivých robotů pomocí naivního bayesovského klasifikátoru. V tomto případě byly vstupní záznamy rozděleny na dva soubory (testovací a trénovací data). Prvních 200 záznamů bylo určeno ručně. Následně klasifikátor určil 171 zbylých záznamů. Na základě údajů v **tabulce č. 7** bylo rozhodnuto, že metoda poskytuje dostatečné výsledky pro filtrování nevyžádaných záznamů.

11.5 Mapa

Mapa provozu slouží k zobrazení celkového provozu webu nebo ke sledování pohybu konkrétního návštěvníka. Mapa má podobu kladně váženého orientovaného grafu.



Obrázek 12: Mapa celkového provozu webu [vlastní tvorba]

Každý uzel představuje navštívenou stránku nebo událost, která byla provedena na této stránce. Hrany znázorňují přechody mezi stránkami nebo spojení události s příslušnou stránkou. Každá hrana je označena číslem, které vyjadřuje počet přechodů po této hraně. Rozložení uzlů a hran na mapě je možné měnit pomocí přepínače rozložení nebo ručním přetažením uzlů po mapě. Kliknutím na vybraný uzel se zobrazí okno s detailním popisem daného uzlu.

11.6 Filtry

Filtry slouží k rozpoznávání nevyžádaných záznamů. Ve správě filtrů je možné nastavit pravidla pro první i druhou fázi čištění dat. Jednotlivé filtry obsahují následující údaje:

- název dimenze
- regulární výraz
- určení fáze čištění (záznamy nebo zpracované záznamy)
- poznámka se stručným popisem filtru

11.7 Údržba

Sekce údržba slouží k optimalizaci databázových tabulek a k čištění naivního bayesovského klasifikátoru. V sekci je zobrazen výpis všech databázových tabulek a informace o jejich využití.

11.8 Past na roboty

Past na roboty je záznam v souboru robots.txt, který zakazuje vstup robotům do fiktivního adresáře admin. Roboti, kteří ignorují pravidlo v robots.txt se identifikují návštěvou zakázaného adresáře.

11.9 Testovací web

Testovací web se skládá z několika webových stránek propojených pomocí hypertextových odkazů. Strukturou připomíná jednoduchou firemní prezentaci. Na webu je umístěn měřící kód, který zaznamenává přístupy na web. Měřící kód ukládá data přímo do databáze pro další zpracování nástrojem webové analytiky. Testovací web je možné procházet zadáním URL adresy do prohlížeče a následnou aktivací hypertextových odkazů nebo pomocí robota.

11.10 Robot

Robot je jednoduchá webová aplikace, která systematicky prochází testovací web a simuluje aktivaci hypertextových odkazů. Údaje, které robot odesílá testovacímu webu, lze ovlivnit v nastavení robota. Je možné nastavit robota tak, aby jeho aktivita nebyla téměř rozeznatelná od přístupů reálného uživatele webu.

12 Shrnutí výsledků

V teoretické části diplomové práce byly popsány současné metody webové analytiky včetně jejich výhod a omezení. V rámci praktické části byl vytvořen nástroj webové analytiky pro práci s naměřenými daty a zobrazení webového provozu v podobě mapy.

13 Závěry a doporučení

V průběhu implementace praktické části bylo zjištěno, že naměřená data jsou ovlivněna nevyžádanými přístupy na sledovaném webu. Z tohoto důvodu byla diplomová práce rozšířena o metody pro filtrování nevyžádaných přístupů. Metody byly následně implementovány a otestovány. Na základě testů bylo prokázáno, že zvolená implementace metod přináší dostatečné výsledky a filtrovaná data jsou použitelná pro účely webové analytiky.

Výsledná aplikace se zaměřuje pouze na omezenou oblast webové analytiky. Rozšířením této aplikace o další nástroje lze dosáhnout větší kontroly sledovaného webu. Lepších výsledků je možné docílit i rozšířením množství měřených údajů nebo implementací dalších metod filtrování nevyžádaných přístupů na web.

14 Seznam použité literatury

[1] *Česká e-komerce v roce 2015 předčila očekávání, růst se nezastaví ani v roce 2016*. Asociace pro elektronickou komerci [online]. Praha, 2016 [cit. 2016-04-12]. Dostupné z: https://www.apek.cz/clanky/ceska-e-komerce-v-roce-2015-predcila-ocekavani-ru

[2] *Češi nakupují on-line stále častěji, za posledních pět let narostly pravidelné nákupy o polovinu!*. Asociace pro elektronickou komerci [online]. Praha, 2016 [cit. 2016-04-12]. Dostupné z: https://www.apek.cz/clanky/cesi-nakupuji-on-line-stale-casteji-za-poslednich

[3] *What is a web server?* Mozilla Developer Network [online]. Mountain View: Mozilla Corporation, 2016 [cit. 2016-04-12]. Dostupné z: https://developer.mozilla.org/en-US/Learn/Common_questions/What_is_a_web_server

[4] *Hypertext Transfer Protocol -- HTTP/1.1* [online]. FIELDING, R., J. GETTYS, J. MOGUL, H. FRYSTYK, L. MASINTER, P. LEACH a T. BERNERS-LEE. 1999 [cit. 2016-04-12]. Dostupné z: http://www.rfc-editor.org/info/rfc2616

[5] CHOUDHURY, Nupur. *World Wide Web and Its Journey from Web 1.0 to Web 4.0* [online]. 2014 [cit. 2016-04-22]. 0975-9646. Dostupné z: http://www.ijcsit.com/docs/Volume%205/vol5issue06/ijcsit20140506265.pdf. Sikkim Manipal University, Department of Computer Science and Engineering, Sikkim Manipal Institute of Technology.

[6] *Uniform Resource Identifier (URI): Generic Syntax* [online]. BERNERS-LEE, T., R. FIELDING a L. MASINTER. 2005 [cit. 2016-04-12]. Dostupné z: http://www.rfc-editor.org/info/rfc3986

[7] CHEN, Min, David EBERT, Hans HAGEN, et al. *Data, Information and Knowledge in Visualization* [online]. 2009 [cit. 2016-04-22]. Dostupné z: http://homepages.cwi.nl/~robertl/articles/cga2009.pdf

[8] KENT, Karen a Murugiah SOUPPAYA. *Guide to Computer Security Log Management: Recommendations of the National Institute of Standards and Technology* [online]. Gaithersburg, 2006 [cit. 2016-04-12]. Dostupné z: http://csrc.nist.gov/publications/nistpubs/800-92/SP800-92.pdf

[9] HAMEL, Stephane. The Ultimate Definition of Analytics. *Online Behavior* [online]. 2011 [cit. 2016-04-13]. Dostupné z: http://online-behavior.com/analytics/definition

[10] ROY, Jonathan. The History of Web Analytics. *Devrun Web Agency* [online]. 2014 [cit. 2016-04-13]. Dostupné z: http://blog.devrun.com/article/the-history-of-web-analytics

[11] FUSCO, Patricia. WebTrends Log Analyzer: Every Click Matters. *Enterprise Apps Today: CRM, business intelligence and ERP news and research* [online]. 2003 [cit. 2016-04-13]. Dostupné z: http://www.enterpriseappstoday.com/retail/webtrends-log-analyzer-every-click-matters.html

[12] Hit Counters: The Analytics Tool of the Early Web. *Priceonomics: Data Crawling, Visualization & Analysis* [online]. San Francisco, 2015 [cit. 2016-04-13]. Dostupné z: http://priceonomics.com/hit-counters-the-analytics-tool-of-the-early-web/

[13] Mission/Vision. *Digital Analytics Association* [online]. Wakefield (Massachusetts) [cit. 2016-04-13]. Dostupné z: http://www.digitalanalyticsassociation.org/mission-vision

[14] SULLIVAN, Danny. Google To Acquire Urchin Web Analytics Firm. *Search Engine Watch* [online]. New York: ClickZ Group, 2005 [cit. 2016-04-13]. Dostupné z: https://searchenginewatch.com/sew/news/2062461/google-to-acquire-urchin-web-analytics-firm

[15] WEB ANALYTICS ASSOCIATION. *Web Analytics Definitions* [online]. Wakefield (Massachusetts), 2008 [cit. 2016-03-19]. ISBN 1-800-349-1070. Dostupné z: http://www.digitalanalyticsassociation.org/Files/PDF_standards/WebAnalyticsDefinitions.pdf

[16] KAUSHIK, Avinash. *Web Analytics 2.0: The Art of Online Accountability and Science of Customer Centricity* [online]. Indianapolis: Sybex, 2009 [cit. 2016-03-19]. ISBN 04-705-2939-3. Dostupné z: http://www.webanalytics20.com/

[17] BATRA, Anil. Digital Analytics Association. *Digital Marketing and Analytics by Anil Batra* [online]. 2012 [cit. 2016-04-13]. Dostupné z: http://webanalysis.blogspot.cz/2012/03/digital-analytics-association.html

[18] KAUSHIK, Avinash. Digital Marketing and Measurement Model. *Occam's Razor: Digital Marketing and Analytics Blog* [online]. [cit. 2016-04-14]. Dostupné z: http://www.kaushik.net/avinash/digital-marketing-and-measurement-model/

[19] KAUSHIK, Avinash. Web Analytics 101: Definitions: Goals, Metrics, KPIs, Dimensions, Targets. *Occam's Razor: Digital Marketing and Analytics Blog* [online]. [cit. 2016-04-14]. Dostupné z: http://www.kaushik.net/avinash/web-analytics-101-definitions-goals-metrics-kpis-dimensions-targets/

[20] ALLEN, Rick. Tracking Content Goals With Web Analytics. *Meet Content* [online]. 2012 [cit. 2016-04-14]. Dostupné z: http://meetcontent.com/blog/tracking-content-goals-with-web-analytics/

[21] What is a Landing Page? *Unbounce* [online]. Vancouver [cit. 2016-04-14]. Dostupné z: http://unbounce.com/landing-page-articles/what-is-a-landing-page/

[22] CARDELLO, Jennifer. Define Micro Conversions to Measure Incremental UX Improvements. *Nielsen Norman Group: UX Training, Consulting, & Research* [online]. Fremont, 2014 [cit. 2016-04-14]. Dostupné z: https://www.nngroup.com/articles/micro-conversions/

[23] Complete Guide to Dimensions and Metrics in Google Analytics. SHARMA, Himanshu. *Optimize Smart: Analytics Consulting & Conversion Optimization* [online]. [cit. 2016-04-14]. Dostupné z: https://www.optimizesmart.com/complete-guide-to-dimensions-and-metrics-in-google-analytics/

[24] PETERSON, Eric. *Web Analytics Demystified: A Marketer's Guide to Understanding How Your Web Site Affects Your Business* [online]. Portland: Celilo Group Media, 2004 [cit. 2016-03-19]. ISBN 09-743-5842-8. Dostupné z: http://www.webanalyticsdemystified.com/downloads/Web_Analytics_Demystified_by_Eric_Peterson.pdf

[25] The Web Analyst's Code of Ethics. *Digital Analytics Association* [online]. Wakefield (Massachusetts) [cit. 2016-04-14]. Dostupné z: http://www.digitalanalyticsassociation.org/codeofethics

[26] *Investigations Involving the Internet and Computer Networks* [online]. Washington, D.C.: U.S. Department of Justice, 2007 [cit. 2016-04-23]. Dostupné z: https://www.ncjrs.gov/pdffiles1/nij/210798.pdf

[27] MACVITTIE, Lori. *Cookies, Sessions, and Persistence* [online]. Seattle: F5 Networks, 2016 [cit. 2016-04-23]. Dostupné z: https://f5.com/Portals/1/Cache/Pdfs/2421/cookies-sessions-and-persistence-.pdf

[28] *Cookies* [online]. Brusel: European Commission, 2016 [cit. 2016-04-23]. Dostupné z: http://ec.europa.eu/ipg/basics/legal/cookies/index_en.htm

[29] LHOTÁK, Jaroslav. Pracujeme se session v PHP. *Interval.cz: Svět Internetu, Technologií a Bezpečnosti* [online]. Brno: ZONER software, 2001 [cit. 2016-04-23]. Dostupné z: https://www.interval.cz/clanky/pracujeme-se-session-v-php/

[30] Manage, disable Local Shared Objects. *Adobe Support* [online]. Adobe [cit. 2016-04-23]. Dostupné z: https://helpx.adobe.com/flash-player/kb/disable-local-shared-objects-flash.html

[31] MALÝ, Martin. Webdesignérův průvodce po HTML5: WebStorage. *Zdroják: o tvorbě webových stránek a aplikací* [online]. Praha: Devel.cz Labs, 2010 [cit. 2016-04-23]. Dostupné z: https://www.zdrojak.cz/clanky/webdesigneruv-pruvodce-po-html5-webstorage/

[32] EMERY, Daniel. *Legal action on 'zombie cookies' filed in US court* [online]. Londýn: BBC, 2010 [cit. 2016-04-23]. Dostupné z: http://www.bbc.com/news/technology-10787882

[33] OKRONA, Csaba. *User tracking without cookies* [online]. 2013 [cit. 2016-04-23]. Dostupné z: https://blog.ochronus.com/user-tracking-without-cookies-be9b757f2cea

[34] Logging Control In W3C httpd. *World Wide Web Consortium (W3C)* [online]. 1995 [cit. 2016-04-23]. Dostupné z: https://www.w3.org/Daemon/User/Config/Logging.html

[35] Log Files. *The Apache HTTP Server Project* [online]. Maryland: The Apache Software Foundation [cit. 2016-04-23]. Dostupné z: https://httpd.apache.org/docs/2.4/logs.html

[36] GRACE, L.K. Joshila, V. MAHESWARI a Dhinaharan NAGAMALAI. *ANALYSIS OF WEB LOGS AND WEB USER IN WEB MINING* [online]. 2011 [cit. 2016-04-23]. Dostupné z: https://arxiv.org/ftp/arxiv/papers/1101/1101.5668.pdf

[37] Log File Formats. *IBM Support Portal* [online]. New York: IBM [cit. 2016-04-23]. Dostupné z: http://publib.boulder.ibm.com/tividd/td/ITWSA/ITWSA_info45/en_US/HTML/guide/c-logs.html

[38] Page Tagging (cookies) vs. Log Analysis. *Logaholic Web Analytics* [online]. Amsterdam: Logaholic [cit. 2016-04-24]. Dostupné z: http://www.logaholic.com/page-tagging-vs-log-analysis/

[39] GREEN, Shaw, Leon HURST, Brenda NANGLE, Pádraig CUNNINGHAM, Fergal SOMERS a Richard EVANS. *Software Agents: A review* [online]. Dublin, 1997 [cit. 2016-04-24]. Dostupné z: https://www.scss.tcd.ie/publications/tech-reports/reports.97/TCD-CS-1997-06.pdf

[40] ŠIMKO, Martin. Princip fungování fulltextových vyhledávačů I.: crawler. *Programujte.com: odborný web zaměřený na oblast vývoje, návrhu a designu webových, mobilních a desktopových aplikací* [online]. 2014 [cit. 2016-04-24]. ISSN 1801-1586. Dostupné z: http://programujte.com/clanek/2014010200-princip-fungovani-fulltextovych-vyhledavacu-i-crawler/

[41] HAYATI, Pedram, Kevin CHAI, Vidyasagar POTDAR a Alex TALEVSKI. *HoneySpam 2.0: Profiling Web Spambot Behaviour* [online]. Perth (Austrálie), 2009 [cit. 2016-04-24]. Dostupné z: http://kevinchai.net/wp-content/uploads/2011/06/honeyspam-2.0-profiling-web-spambot-behaviour.pdf. Curtin University, Digital Ecosystem and Business Intelligence Institute.

[42] All Your Google Analytics Are Belong To Us. GEORGIEV, Georgi. *Analytics ToolKit: Google Analytics Tools for SEO, SEM and Digital Agencies* [online]. 2015 [cit. 2016-04-24]. Dostupné z: http://blog.analytics-toolkit.com/2015/google-analytics-data-integrity-attacks/

[43] *Report: Bot traffic is up to 61.5% of all website traffic* [online]. Kalifornie: Imperva, 2013 [cit. 2016-04-24]. Dostupné z: https://www.incapsula.com/blog/bot-traffic-report-2013.html

[44] *2014 Bot Traffic Report: Just the Droids You were Looking for* [online]. Kalifornie: Imperva, 2014 [cit. 2016-04-24]. Dostupné z: https://www.incapsula.com/blog/bot-traffic-report-2014.html

[45] *2015 Bot Traffic Report: Humans Take Back the Web, Bad Bots Not Giving Any Ground* [online]. Kalifornie: Imperva, 2015 [cit. 2016-04-24]. Dostupné z: https://www.incapsula.com/blog/bot-traffic-report-2015.html

[46] A Standard for Robot Exclusion. KOSTER, Martijn. *The Web Robots Pages* [online]. [cit. 2016-04-24]. Dostupné z: http://www.robotstxt.org/orig.html

[47] BAECHER, Paul, Marc FISCHLIN, Lior GORDON, Robert LANGENBERG, Michael LUTZOW a Dominique SCHRODER. *CAPTCHAs: The Good, the Bad, and the Ugly* [online]. 2013 [cit. 2016-04-24]. Dostupné z: http://subs.emis.de/LNI/Proceedings/Proceedings170/353.pdf. Darmstadt University of Technology.

[48] Inside India's CAPTCHA solving economy. DANCHEV, Dancho. *ZDNet: Technology News, Analysis, Comments and Product Reviews for IT Professionals* [online]. 2008 [cit. 2016-04-24]. Dostupné z: http://www.zdnet.com/article/inside-indias-captcha-solving-economy/

[49] BASTL, P., J. KUČERA a D. LÉWOVÁ. *Metody strojového učení* [online]. 2003 [cit. 2016-04-24]. Dostupné z: http://tydenvedy.fjfi.cvut.cz/2003/cd/prispevky/sbpdf/strojuc.pdf

[50] DOWNEY, Allen B. *Think Bayes: Bayesian Statistics Made Simple* [online]. Needham: Green Tea Press, 2012 [cit. 2016-04-25]. Dostupné z: http://www.greenteapress.com/thinkbayes/thinkbayes.pdf

[51] Honeypot Technique: Fast, Easy Spam Prevention. MZARATE. *Solution Factor: Austin Web Development and Database Consulting* [online]. 2014 [cit. 2016-04-25]. Dostupné z: https://solutionfactor.net/blog/2014/02/01/honeypot-technique-fast-easy-spam-prevention/

[52] Reporting vs. Analysis: What's the Difference? DYKES, Brent. *Digital Marketing Blog by Adobe* [online]. Adobe Systems, 2010 [cit. 2016-04-25]. Dostupné z: https://blogs.adobe.com/digitalmarketing/analytics/reporting-vs-analysis-whats-the-difference/

[53] Understanding the differences between path and funnel. LEVY, Tsahi. CoolaData: Big Data Analysis and Visualization [online]. 2015 [cit. 2016-04-25]. Dostupné z: http://www.cooladata.com/path-vs-funnel/

[54] HLINĚNÝ, Petr. *Teorie Grafů* [online]. Brno, 2008 [cit. 2016-04-25]. Dostupné z: http://www.fi.muni.cz/~hlineny/Vyuka/GT/Grafy-text07.pdf. Masarykova univerzita, Fakulta informatiky.

15 Přílohy

- robot
- testovací web a měřící kód
- nástroj webové analytiky
- databázový soubor

Oskenované zadání práce

Univerzita Hradec Králové
Fakulta informatiky a managementu
Akademický rok: 2015/2016

Studijní program: Aplikovaná informatika
Forma: Prezenční
Obor/komb.: Aplikovaná informatika (ai2-p)

Podklad pro zadání DIPLOMOVÉ práce studenta

PŘEDKLÁDÁ:	ADRESA	OSOBNÍ ČÍSLO
Bc. Šúkala Ondřej	Králova Lhota 36, Králova Lhota	I1478

TÉMA ČESKY:

Moderní metody a trendy webové analytiky

TÉMA ANGLICKY:

Advanced methods and trends of web analytics

VEDOUCÍ PRÁCE:

Ing. Karel Mls, Ph.D. - KIT

ZÁSADY PRO VYPRACOVÁNÍ:

Cílem práce je popsat moderní metody webové analytiky a vytvořit aplikaci pro sledování návštěvníků webu.

Osnova: Úvod, metodika zpracování, webová analytika, metriky, metody měření, analytické metody, hrozby, typy nástrojů, shrnutí výsledků, závěry a doporučení.

SEZNAM DOPORUČENÉ LITERATURY:

BURBY, Jason; ATCHISON, Shane. Actionable web analytics: using data to make smart business decisions. John Wiley & Sons, 2007.
CLIFTON, Brian. Advanced web metrics with Google Analytics. Sybex, 2012. ISBN 978-1118168448.
KAUSHIK, Avinash. Webová analytika 2.0: kompletní průvodce analýzami návštěvnosti [online]. Vyd. 1. Brno: Computer Press, 2011, 456 s. [cit. 2015-10-13]. ISBN 978-80-251-2964-7.
KAUSHIK, Avinash. Web Analytics: An Hour a Day. Sybex, 2007. ISBN 978-0470130650.
MORTENSEN, Dennis. Yahoo! Web Analytics: Tracking, Reporting, and Analyzing for Data-Driven Insights. Sybex, 2009. ISBN 0470424249.
PETERSON, Eric. Web Analytics Demystified. Celilo Group Media, 2004. ISBN 978-0974358420.
PETERSON, Eric. Web Site Measurement Hacks. O'Reilly Media, 2005. ISBN 978-0596009885.

Podpis studenta: ..... **Datum:** 13.10.2015

Podpis vedoucího práce: ..... **Datum:** 13.10.2015