Vysoká škola ekonomická v Praze

Fakulta managementu v Jindřichově Hradci

Diplomová práce

Jan Dvořák

2007

Vysoká škola ekonomická v Praze

Fakulta managementu v Jindřichově Hradci

Katedra informatiky

Web Usage Mining

Vypracoval:

Jan Dvořák

Vedoucí diplomové práce:

Ing. Jiří Jelínek, CSc.

Velký Rybník, červenec 2007

Prohlášení

Prohlašuji, že diplomovou práci na téma

»Web Usage Mining«

jsem vypracoval samostatně.

Použitou literaturu a podkladové materiály

uvádím v přiloženém seznamu literatury.

Velký Rybník, červenec 2007

podpis studenta

Anotace

Web Usage Mining

- Obecný popis Web Miningu.
- Charakteristika a užití technik web usage miningu.
- Podrobný popis metod a nástrojů zahrnovaných pod pojem "Web Usage Mining" včetně příkladů použití.
- SW nástroje a existující řešení pro web usage mining.
- Praktický návrh konkrétního řešení s využitím výše popsaných metod Web Usage Miningu.

červenec 2007

Poděkování

Za cenné rady, náměty a inspiraci

bych chtěl poděkovat

Ing. Jiřímu Jelínkovi, CSc.,

z Vysoké školy ekonomické v Praze,
Fakulty managementu v Jindřichově Hradci.

Dále bych chtěl za spolupráci poděkovat pánům
Mgr. Davidovi Štroblovi a Michalovi Hajdíkovi,
pracovníkům Centra výpočetní techniky
Fakulty managementu.

Obsah

Úvod

Motto:
"If the point of contact between the product and the people becomes a point of friction, then the industrial designer has failed. If, on the other hand, people are made safer, more comfortable, more eager to purchase, more efficient -- or just plain happier -- the designer has succeeded"

Henry Dreyfuss

Žijeme ve světě informací. Informace dominují světu více než kdy předtím. Evoluce Internetu do globální informační infrastruktury, spojená s obrovskou popularitou World Wide Webu, dovolila obyčejným občanům stát se nejen spotřebiteli informací, ale také jejich rozšiřovateli. Lidé berou výhody Internetu v různých směrech, hledají informace pomocí vyhledávacích strojů, nakupují v e-shopech, inzerují atd. Přirozeně to chtějí dělat rychle, jednoduše a hladce.

Jestliže je zde obrovské, a stále rychleji rostoucí, množství dat a informací, jak může průměrný uživatel nalézt, co hledá? Vyhledávací služby indexují většinou pouze část prostoru WWW. Kromě toho začínající uživatelé nejsou schopni kvalitně definovat své požadavky. Pro mnohé uživatele je obtížná orientace i v té části webu, kterou již dříve navštívili. Nabízí se tedy otázka, jak usnadnit běžnému uživateli jeho činnost. Velkým nástrojem pro tento účel je Web Mining. Web Mining může být široce definován jako objevování a analyzování užitečných informací z WWW, jako aplikace data-miningových technologií k obrovskému skladu webových a jiných dat.

Poskytovatelé obsahu se pokoušejí nabízet uživatelům internetu co nejkvalitnější servis. K tomu potřebují mnoho informací vztahujících se k jejich webovým stránkám a k uživatelům, kteří je navštěvují. Tyto informace jsou obsaženy v logovacích souborech, které zaznamenávají každý pohyb po webových stránkách poskytovatele obsahu. Nástrojem k analýze těchto logů je Web Mining.

Web Usage Mining je součástí Web Miningu. Je to znalostní odhalovací technika zaměřená na analýzu webu. Metody Web Usage Miningu analyzují pohyby na webových stránkách a dávají poskytovatelům obsahu vítanou zpětnou vazbu. Dávají smysluplný náhled na to, jak jsou užívány webové stránky. Uplatnění nalézá v oblastech jako je design webových stránek, podpora obchodního a marketingového rozhodování, personalizace atd.

Objasnění užívaných pojmů

- **Cache** – Rychlá vyrovnávací paměť.
- **Clickstream analýza** – Clickstream analýza představuje širokou škálu analytických prostředků, aplikovaných na informace získané provozem obchodních, či jinak zákaznicky orientovaných řešení v prostředí internetu. K jejich získání dochází, když zákazník přichází na WWW server a prochází se jeho stránkami. Tímto se vytváří spojitý tok kliknutí počítačovou myší, označovaný pojmem clickstream. Těchto dat je následně použito při analýze, která má za cíl pochopit chování zákazníka a těchto poznatků využít k prospěchu online komerčních aktivit (Rehberger 2002a).
- **Cookies** – Informace zasílané webobým serverem prohlížeči k uchování specifických informací (např. o individuálním nastavení, využitelném při dalším požadavku na tento webový server).
- **Hyper-link** – Odkaz v hypertextovém dokumentu. Vazba, definovaná mezi částmi textu (odstavci, stránkami) nebo i mezi textem a objekty (obraz, zvuk, video).
- **IP adresa** – Jedinečná adresa počítače nebo sítě, např. 213.226.245.235
- **Logovací soubor** – Soubor, do kterého se provádí záznamy o stavu nebo o probíhajících procesech.
- **PageRank** – Algoritmus pro ohodnocení důležitosti webových stránek, navržený Larry Pagem a Sergeyem Brinnem a tvořící základ vyhledávače Google.
- **Paket** – Obecně dávka dat, zpracovávaná při komunikaci jako celek. Mívá většinou definovanou hlavičku a strukturu.
- **Proxy** – Zařízení zastupující jiné zařízení, většinou z důvodu efektivity přenosu. Proxy server je server, na kterém běží program pro kešování HTTP dotazů.
- **Rámce** – Frames, umožňují rozdělit stránku na více částí, respektive několik stránek je zobrazeno současně.
- **Server** – Datová stanice sloužící jiným počítačům, nazývaným klient nebo pracovní stanice, jako zdroj dat, programů, služeb. Webový server je server poskytující přístup k webovým zdrojům.

- **Sezení (session)** – Sada uživatelských kliknutí.
- **User session** – Sada uživatelských kliknutí na jednom nebo více webových serverech.
  - **Single user session** – Sada kliknutí jednoho uživatele na jednom nebo více serverech.
  - **Multiple user session** – Sada kliknutí všech uživatelů jednoho nebo více serverů.
- **Server session** – Neboli návštěva, kolekce uživatelských kliknutí na jednom webovém serveru během user session.
- **TCP/IP** – Transmission Control Protocol/Internet Protocol. Sada protokolů vyvinutá v agentuře ARPA. Obsahuje TCP jako primární transportní protokol a IP jako protokol síťové vrstvy.
- **URI** – Uniform Resource Identifer, jedinečná identifikace zdroje, skládá se z adresy objektu a schéma, schématem může být http, ftp, file, atd. Je to celý odkaz i s odkazy na místo v dokumentu. (např. http://www.abc.cz/dokument.htm#uvod)
- **URL** – Uniform Resource Locator, standardizované adresy pro přístup k hypertextovým dokumentům a dalším službám na Internetu pomocí prohlížeče (browser), URL je podmnožinou URI. (např. http://www.abc.cz/dokument.htm)
- **Webový robot** – Část vyhledávače, která prochází webové stránky, ukládá si kopie nebo části stránek, na které se dostane pomocí odkazů.
- **WWW** – World Wide Web, služba Internetu, poskytující hypertextové a další služby klientským aplikacím typu browser.

Web Mining

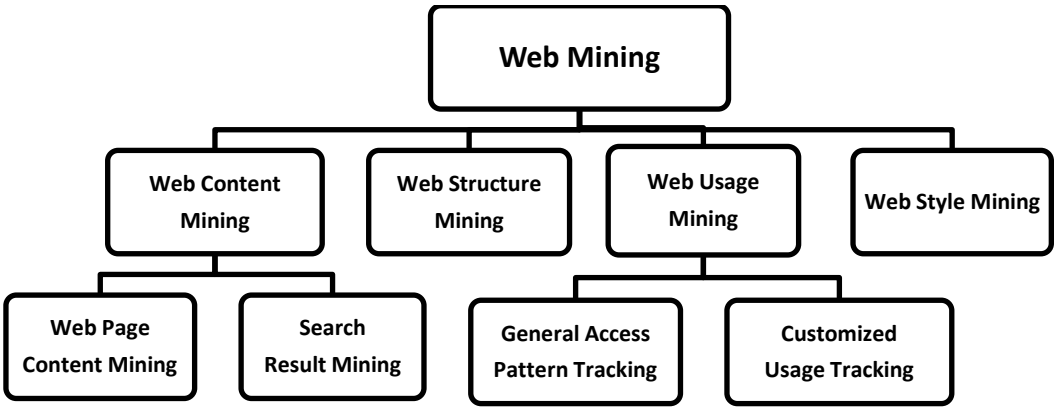
Web Mining je extrakce zajímavých a potenciálně užitečných vzorů, zákonitostí a skrytých informací z artefaktů nebo aktivit provozovaných na World Wide Webu. Web Mining je součástí odvětví nazývaného Data Mining (dolovaní či vytěžování dat), které vzniklo z nutnosti zpracovávat nepřeberné množství dat a v nich objevovat skryté, hlubší vztahy. K tomu využívá nejnovějších poznatků a technologií.

Data Mining znamená mnoho různých postupů a algoritmů, které umožní odhalit užitečné vztahy ukryté v datech. Neexistuje žádná zázračná metoda, která řeší všechny úlohy s libovolným typem dat. Pro různé úlohy a data se hodí různé metody. Velmi častý je případ, kdy nejlepších výsledků dosáhneme vhodnou kombinací různých metod. Typickými úlohami Data Miningu jsou: detekce podvodů, profily zákazníků, udržení zákazníka, stanovení diagnózy, analýza časových řad, analýza prohlížení stránek na Internetu (SPSS 2006).

Jsou přibližně tři až čtyři oblasti odhalování znalostí, které náleží do Web Miningu: Web Content Mining, Web Structure Mining, Web Usage Mining a Web Style Mining. První klasifikace Web Miningu jej rozdělovala na Web Content Mining a Web Usage Mining. Třetí částí Web Miningu se stal Web Structure Mining, v současnosti především v Asii je za další součást považován Web Style Mining. (Galeas 2006), (Ching-Nan 2002).

Metody Web Miningu:

- Web Content Mining je proces extrahování znalostí z obsahu WWW stránek.
- Web Structure Mining je proces odvozování znalostí z uspořádání a propojení WWW.
- Web Usage Mining, také známý jako Web Log Mining, je proces, který sleduje pohyb a analyzuje chování uživatelů WWW stránek.
- Web Style Mining je proces zabývající se designem WWW stránek.



Obr. 1 **Metody Web Miningu**

Metody Web Miningu

Web Content Mining

Web Content Mining je proces dobývání znalostí z obsahu webových stránek. Analyzuje textové složky stránek. Na obrázky, video a zvuk se nezaměřuje, protože dnes ještě není možné z těchto souborů strojově zjistit jejich obsah.

Je to automatický proces, který jde až na extrakci klíčových slov. Protože dokumenty obsahují strojově nečitelnou sémantiku, některé přístupy navrhují restrukturovat obsah dokumentu do podoby, která může být využívána stroji.

Jsou dvě skupiny web content miningových strategií (Galeas 2006):

- ty, které přímo dolují obsah dokumentů – Web Page Content Mining,
- a ty, které vylepšují obsahové pátrání dalších nástrojů, jako jsou vyhledávací stroje – Search Result Mining.

První ukládá polo-uspořádané informace z webu do databází a analyzuje je standardními databázovými dotazy a data-miningovými technikami.

Druhý využívá speciální agenty k hledání, filtrování a třídění dokumentů. Ti využívají různé techniky (doménová charakteristika, uživatelský profil, techniky výběru informací, atd.) k organizování a interpretaci objevených informací.

Většina metod pro analýzu textu používá jako základní jednotku pro popis dokumentu pojem term (Jelínek 2004a). Termy jsou jednotlivá slova nebo víceslovná spojení, která jsou v dokumentu významně zastoupena nebo jsou pro dokument charakteristická. Pro extrakci termů se využívá nejen čistý text na webové stránce, ale i další informace. Například z hlavičky stránky, kde jsou klíčová slova, titulek atd., nebo z jiných URL, kde jsou texty, které odkazují na danou stránku.

Tento postup se snaží nahradit chybějící ontologický popis stránky. Výsledek ontologického zařazení se používá pro objevování podobně zaměřených stránek a pro shromažďování termů, kterých se využívá ve vyhledávání pomocí některé z běžných vyhledávacích služeb. Vyhledávací stroje pak expanzí dotazu přeformulují dotaz do podoby, která zlepší výsledek vyhledávání. Expanze dotazu pomáhá vyhledávat synonyma, různé morfologické formy, nebo ignorovat pravopisné chyby.

Web Structure Mining

World Wide Web může odkrýt více informací, než jsou pouze informace obsažené v dokumentech. Například odkazy směřující k dokumentu signalizují popularitu dokumentu, zatímco odkazy směřující z dokumentu ven signalizují bohatství nebo možná rozmanitost témat pokrytých v dokumentu. To může být přirovnáno k bibliografickým citacím, když je článek citován často, měl by být důležitý (Galeas 2006). PageRank a jiné metody (například algoritmus HITS) berou výhody těchto informací, zprostředkovaných odkazy, k hledání vhodných webových stránek. PageRank pomáhá v detekování významných stránek. Simuluje nahodilé prohlížení stránek na webu, každé stránce jsou přiřazeny body důležitosti. Algoritmus HITS přiřazuje každé stránce dvě hodnoty (autority score a hub score). Jako hub jsou označovány odkazy na kolekce významných stránek o určitém tématu. Jako autorita jsou označovány významné stránky, na které vede mnoho linků (Kryl 2004).

Web Structure Mining analyzuje vzájemné propojení webových stránek, studuje hyper-linkovou strukturu webu. Kategorizuje webové stránky a generuje informace, např. podobnost a vztah mezi webovými stránkami. Ze vzájemných hypertextových vazeb webových stránek je možné usuzovat na jejich tematickou podobnost.

Web Structure Mining například transformuje webový prostor do orientovaného grafu a následně využívá techniky pro práci s grafy. Výsledkem je pak hierarchický model struktury webového prostoru.

Web Usage Mining

Web Usage Mining je proces, který sleduje pohyb a analyzuje chování uživatele WWW stránek.

Této problematice se bude podrobně věnovat následující text.

Web Style Mining

Web Style Mining se zabývá analýzou stylu a prezentace webových stránek. Ze stylu dokumentů, které jsou zobrazeny webovým prohlížečem, mohou být vytěženy cenné informace. Tato metoda je poměrně nová a rozvíjí se především v Asii.

Web Style Mining je možné aplikovat v oblastech: selekce založené na charakteristických stylech, indexování stylů, shlukování stylů, generování stylů a vyhledávání stylů (Ching-Nan 2002).

Zdroje dat pro Web Mining

Data pro Web Mining můžeme shromáždit z mnoha různých zdrojů. Jedním zdrojem dat jsou obsahy webových stránek, odkud můžeme získat zobrazovaný obsah, meta popis stránky, WWW odkazy, URL a jeho struktura, atd. Druhým zdrojem jsou záznamy a data o chování uživatele, která se automaticky ukládají v logovacích souborech, tyto soubory jsou na straně serverů, proxy serverů nebo na straně klienta. Data z těchto zdrojů se liší v jejich původu a klasifikaci. Můžeme je roztřídit do čtyř skupin (Čenovský 2003):

- **Obsah** – Data, která jsou určena k tomu, aby byla prezentována uživatelům. Jsou to data, která se nacházejí na webových stránkách, skládají se z textu a grafiky, přičemž největší význam pro analýzy má textová složka. Zdrojem informací je i obsah hlavičky www stránek, který může obsahovat cenné informace.

- **Struktura** – Uspořádání informací, které charakterizuje strukturu obsahu.
  Mezi-stránková struktura je tvořena prostřednictvím hyper-linků, které spojují stánku s ostatními. Uspořádání HTML a XML tagů tvoří vnitro-stránkovou strukturu.

- **Užívání** – To jsou data, která popisují vzory užívání webových stránek. Jsou to IP adresy, data a časy přístupů, atd. Uživatelská data pochází z rozšíření běžného log formátu (ECLF – Extended Common Log Format).

- **Uživatelský profil** – To jsou data, která poskytují demografické informace o uživatelích webových stránek. Jsou to registrační data a další informace o uživatelích.

Problémy a omezení Web Miningu

- Získávání dat ze serverů je omezeno vlastnostmi logovacích souborů, které byly původně vytvářeny pro účely ladění (Jelínek 2004b). Logovací soubory obsahují velké množství neužitečných informací a na druhou stranu v nich některé užitečné informace mohou chybět. Běžně je užíván Common Log Format, lépe je používat rozšířený formát Extended Log Format.
- Logy neukládají informace o požadavcích, které byly zachyceny při použití webové nebo proxy cache.
- Logovací soubory ukládají pouze URI stránek a nikoliv jejich sémantický popis, také neobsahují data z webových formulářů.
- Problémy přináší i identifikace uživatele, pokud není žádná použita, tak data mohou být zkreslena, protože za jednou IP adresou může být skryto více uživatelů, nebo naopak jeden uživatel může vystupovat pod více IP adresami.
- U identifikace sezení může být problém se zjištěním času, kdy byla stránka opuštěna.
- Problémy přináší stránky, na kterých jsou použity rámce, nebo dynamické stránky.
- Získávání dat na úrovni klienta závisí na jeho spolupráci.
- Určité problémy vyvstávají i v oblasti ochrany soukromí uživatelů.
- Jinými kategoriemi problémů mohou být potíže s náročností prováděných výpočtů.

Web Usage Mining

Web Usage Mining je proces využívající data-miningových technik k objevování vzorů chování uživatelů webových stránek. Sleduje jejich pohyb a analyzuje jejich chování. Pokouší se objevit užitečné informace ze sekundárních dat odvozených z interakcí uživatelů „surfujících po webu“. Analyzování webových záznamů o přístupech (web access logů) z různých webových stránek, může pomoci porozumět uživatelskému chování a struktuře webu, a tím zlepšit tuto ohromnou kolekci zdrojů, například zjednodušit navigaci po stránkách.

Web Usage Mining je možné rozdělit na dva směry sledující odlišné cíle (Galeas 2006):

- General Access Pattern Tracking
- Customized Usage Tracking

General Access Pattern Tracking analyzuje webové logy, abychom porozuměli přístupovým vzorům a trendům. Tyto analýzy mohou pomoci poskytovatelům zdrojů s vylepšením struktury a seskupením dokumentů. Existuje mnoho nástrojů pro analýzy webu. Aplikování data miningových technik na logovací soubory odhaluje zajímavé vzory přístupů, které mohou být použity k restrukturování stránek do vhodnějších skupin, určení efektivnějších reklamních lokalit, a vybrat si za cíl specifické uživatele pro specifický prodej, atd. Tato část Web Usage Miningu nahlíží na uživatele webových stránek jako na celek, zabývá se nejčastějším, nejobvyklejším chováním uživatelů. Sestavuje profil průměrného uživatele.

Customized Usage Tracking analyzuje individuální trendy. Jeho účelem je přizpůsobit webové stránky konkrétním uživatelům. Hloubka stránkové struktury a formát zdrojů mohou být dynamicky přizpůsobovány každému uživateli v průběhu času na základě jejich přístupových vzorů. Customized Usage Tracking se na rozdíl od předchozího proudu zaměřuje na chování konkrétních individuálních uživatelů. Využívá se v oblasti personalizace.

Uplatnění Web Usage miningu

Web Usage mining je nástrojem k analýze velkého objemu nevyužitých dat obsažených v logovacích souborech webových serverů, které obsahují cenná clickstream data. Úspěch všech analýz závisí na tom jaké a kolik znalostí může být objeveno v rozsáhlých surových logových datech, také na jejich validitě a hodnověrnosti.

Cesty po navštívených stránkách mohou být užity k identifikaci typického prohlížecího chování uživatelů, tyto informace jsou užitečné pro personalizaci webu.

Identifikace nejčastějších přístupů na stránky poskytuje podklad pro zlepšení provedení celé kolekce stránek, zvýšení atraktivity obsahu a zefektivnění budoucích přístupů. Informace o nejžádanějších stránkách mohou být využity při kešování.

Identifikace obvyklého prohlížecího chování slouží ke zlepšení provedení navigace po stránkách a je podkladem pro další modifikace na stránkách. Informace jsou užitečné pro předpovídání chování během interakce s webem.

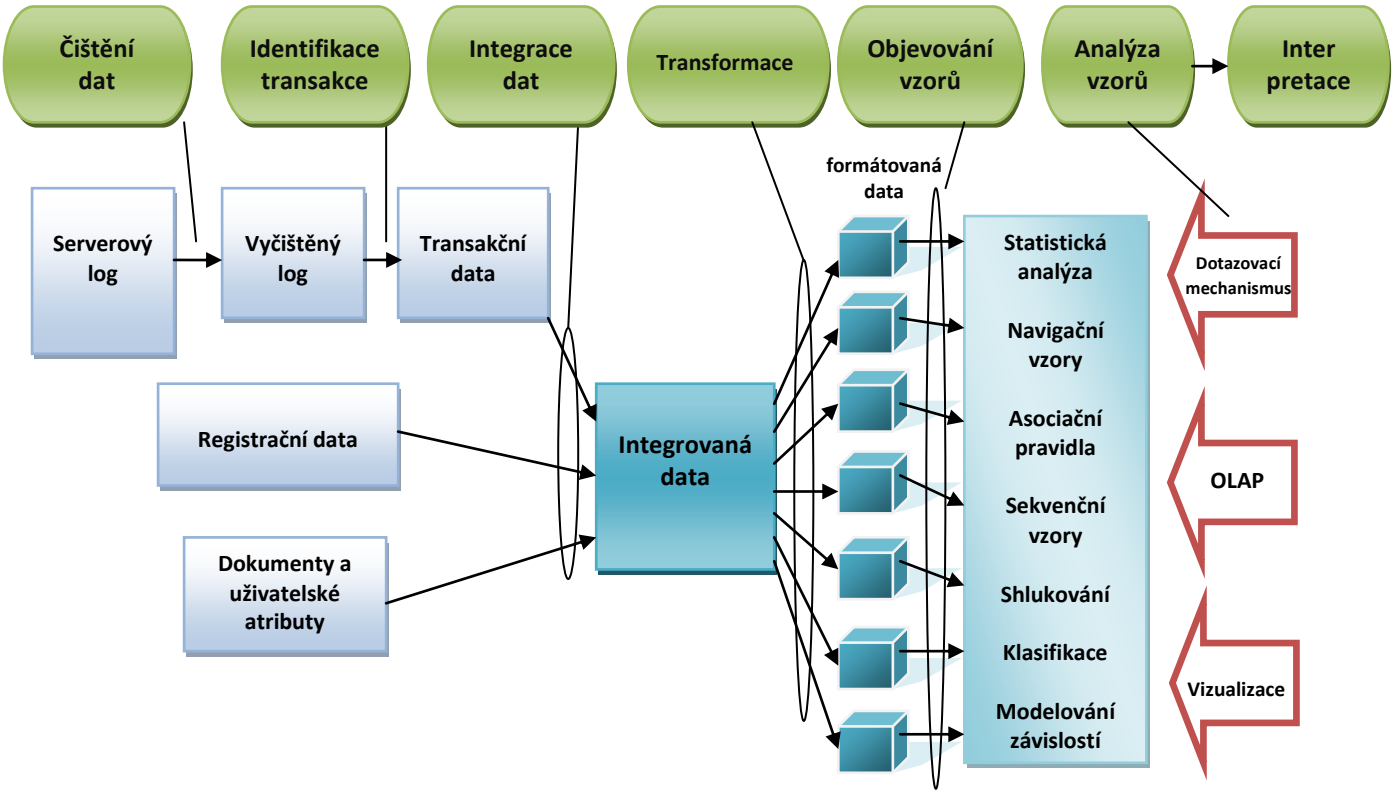
Web Usage Mining přináší informace o tom, odkud uživatelé přicházejí na naše stránky, informace o jejich geografickém rozdělení, ve kterých dnech nebo hodinách nejčastěji přicházejí, jaké používají webové prohlížeče.

Je podkladem pro studie proveditelnosti zabývající se kvalitou provedení uživatelského rozhranní. Umožňuje vytvořit přehled různých navigačních strategií na specifických webových stránkách.

Pomocí Web Usage Miningu je možné detekovat vniknutí, podvody, pokusy o nabourání se do systému.

Web Usage Mining je možné využít v oblastech e-Learningu, e-Business, e-Commerce, e-CRM, e-Services, e-Education, e-Newspapers, e-Government, a Digitálních Knihovnách. Dalšími oblastmi mohou být výroba a plánování, finanční plánování, psychologie, sociologie, biotechnologie atd.

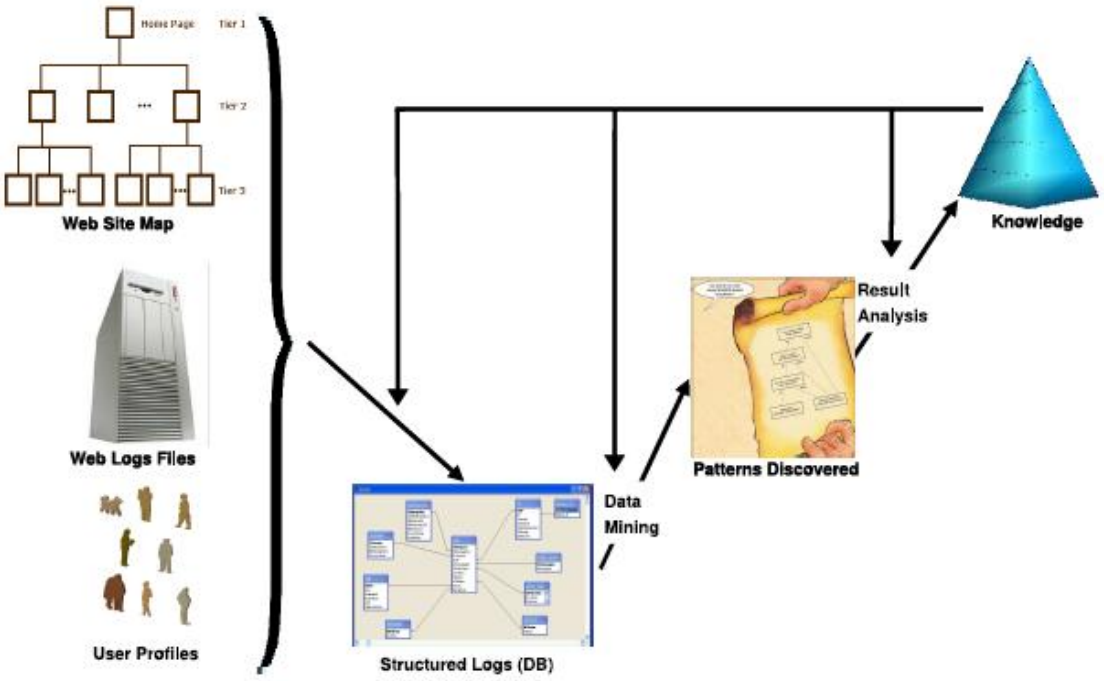
Architektura Web Usage Miningu:



Obr. 2 **Architektura Web Usage Miningu**
Inspirováno (Cooley, Mobasher & Srivastava 1997), (Mobasher 1997b)

- **Čištění dat** - Prvním krokem je čištění dat serverového logu. Analýza bude mít vypovídací schopnost pouze v případě, že data budou obsahovat přesný obraz o uživatelských přístupech. Irelevantní data je možné vypustit například pomocí sledování přípony z URL. Na této úrovni může také dojít ke slučování různých logů.
- **Identifikace transakce** - V tomto kroku jsou položky logu rozděleny do logických shluků. Cílem je vytvořit smysluplné shluky odkazů pro každého uživatele. Velké transakce mohou být rozděleny na menší, nebo naopak malé transakce se slučují do větších shluků. To se odvíjí od daného úkolu těžby dat.
- **Integrace dat** – Logovací soubor není jediným zdrojem dat. Zde se doplňují data získaná z logu o další údaje. Výsledkem jsou integrovaná data.
- **Transformace** – Integrovaná data je nutné přetransformovat, aby byla přizpůsobena příslušnému úkolu těžby dat.

- **Objevování vzorů** - Z transformovaných dat získává analytik pomocí dotazovacích mechanismů vzory chování, pravidla a statistiky.
- **Analýza vzorů** - Objevené vzory následně pomocí vhodných nástrojů analyzuje, aby z nich získal významné a užitečné vzory a pravidla.
- **Interpretace** – Interpretace získaných vzorů a pravidel chování.



Obr. 3 **Hlavní proces Web Usage Miningu**
(Tanasa 2005)

Zjednodušený pohled na proces Webu Usage Miningu. Surová data se integrují do databáze, z které se pomocí Data Miningu dolují vzory. Analýzou vzorů se získávají znalosti. Je zde znázorněna i situace, kdy objevené znalosti ovlivňují samotný proces.

Zdroje dat pro Web Usage Mining

Hlavním zdrojem dat pro Web Usage Mining jsou automaticky generovaná data ukládaná do serverových logů. Dalším zdrojem dat mohou být cookies, logy agentů, uživatelské profily, v oblasti e-commerce změny v nákupních košících, meta-data, atributy stránky, její struktura a obsah. Podle místa sběru dat jsou zdroje rozděleny do třech kategorií (Srivasta, Cooley, Deshpande & Pang-Ning 2000), (Ching-Nan 2002).

Sběr dat na úrovni severu

Logovací soubor webového serveru je nejdůležitějším a nejlepším datovým zdrojem pro Web Usage Mining, protože explicitně zaznamenává chování návštěvníků webových stránek. Logy mohu mít různý formát od základního až po různě rozšířený. Základní formát logu zaznamenává informace o vzdáleném jménu hostitele nebo IP adresy, uživatelské jméno, čas a datum, vyžadované URI, serverový stav a přenesené bajty. Rozšířený formát přidává další položky jako například identifikaci uživatelova agenta (prohlížeče) nebo referrer identifikaci, která umožňuje zjistit, odkud se uživatel dostal na naše stránky.

Logy mají ale i své nevýhody, neobsahují požadavky, které zachytila browserová nebo proxy cache, přinášejí tak neúplný obrázek o uživatelových pohybech, to znesnadňuje identifikaci single-user session, sledování konkrétního návštěvníka. Tato nevýhoda může být redukována například použitím cookies. Logy jsou tak nejlepším nástrojem pro identifikaci multiple-user session, určení chování návštěvníků určitého webového serveru. Další nevýhodou je neznalost času stráveného prohlížením stránek, to je ale možné řešit určitým způsobem, který bude popsán níže.

Dalšími důležitými zdroji dat jsou sledování příchozích a odchozích TCP/IP paketů, nebo logy obsahových a aplikačních serverů, které poskytují obsahy pro dynamické webové stránky.

Sběr dat na úrovni klienta

Tato data mohou být sbírána pomocí vzdálených agentů na základě Java skriptů nebo Java appletů, nebo pomocí upravených webových prohlížečů. Tento sběr dat je vhodný použít pro sledování single-user session na jednom nebo více webových serverech. Tyto metody řeší problémy, které způsobuje browserová nebo proxy cache, a také usnadňují

identifikaci sezení. Zaznamenávají čas, který uživatel strávil na konkrétní stránce, zda použil tlačítka Zpět nebo Obnovit, atd.

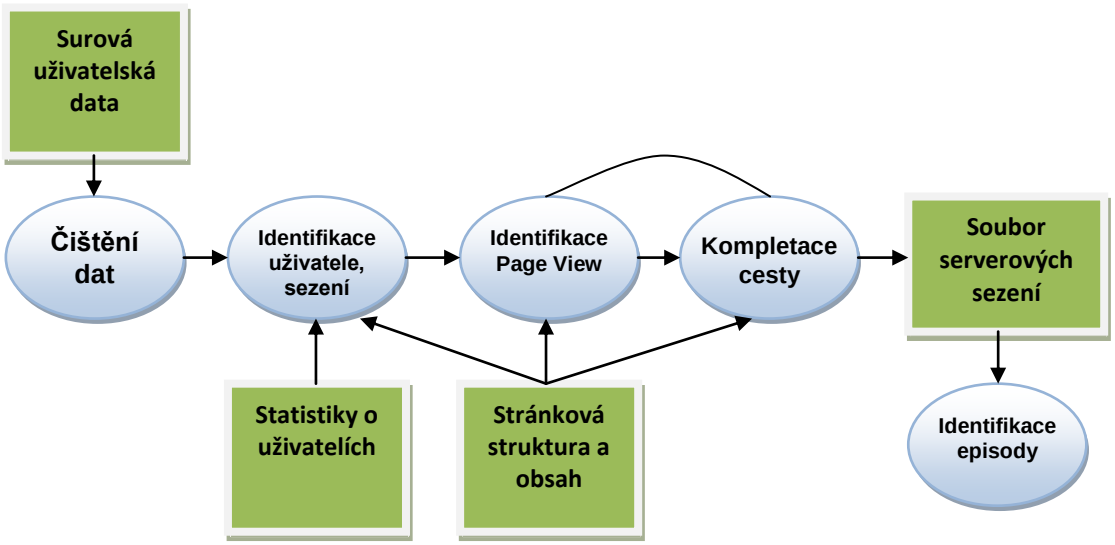
Ale jejich největším nedostatkem je nutnost spolupráce uživatele, kterého musíme nejprve přesvědčit, aby tyto nástroje používal. To podstatně znesnadňuje jejich implementaci. Sběr dat na úrovni klienta je nejčastěji používán v oblasti marketingu a analýzy trhu.

Sběr dat na úrovni proxy

Sběr dat na úrovni proxy je na cestě mezi klientem a serverem. Je to vhodný zdroj dat pro sledování multiple-user session na více webových serverech, protože proxy může zaznamenávat HTTP požadavky klientů na webové servery. Nevýhodou je, že sledujeme chování pouze určité skupiny uživatelů sdílejících stejný proxy server, ne všechny uživatele webového serveru.

Předzpracování dat

Ve fázi předzpracování jsou surová data konvertována do podoby vhodné pro analýzu vzorů. Jak bylo zmíněno výše, zdrojů dat může být velké množství, mají však různou míru dostupnosti. Proto se metody předzpracování dat zaměřují především na informace, které můžeme získat z běžného logu (Common Log Format). Postup při předzpracování dat znázorňuje obrázek.



Obr. 4 **Předzpracování dat**
(Berendt 2003)

Čištění dat

Čištění dat není důležité jen pro Data Mining, ale provádí se při jakékoliv analýze dat. HTTP protokol vyžaduje oddělené spojení pro každý soubor, který je vyžádán z webového serveru. Požadavek uživatele na zobrazení určité stránky je zahrnut v několika položkách logu, protože spolu s HTML souborem se stahují i obrázky, soubory se stylem stránky a skripty. Ve většině případů je relevantní pouze HTML soubor, protože uživatel explicitně nepožaduje grafiku, která se stahuje automaticky. Domnělá irelevantní data je možné vypustit například pomocí sledování přípony z URL. Všechny záznamy o souborech s příponou gif, jpg, jpeg tak mohou být smazány. Seznam je ovšem nutné modifikovat v závislosti na typu webových stránek. Některé stránky obsahují grafické archivy, proto

nemůže analytik smazat všechny požadavky na obrázky. Některé z těchto požadavků tak mohou obsahovat informace o chování uživatele, a jsou důležité pro analýzu. V takovém případě může být vytvořen seznam irelevantních obrázků nepotřebných k dalšímu zpracování a těch, které je nutné pro další zpracování zachovat. Seznam je pak použit při čištění logu.

Stránky také navštěvují weboví roboti, jejichž návštěvy jsou pro analýzu irelevantní. Pokud mají stránky malou návštěvnost, tak podíl těchto návštěv může být značný a výsledky analýzy by tak byly zkresleny. Návštěvy robotů je možné odfiltrovat například pomocí seznamu známých robotů nebo prostřednictvím vypočtené prohlížecí rychlosti. Vychází se z předpokladu, že roboti mají velký podíl zobrazených stránek k časové délce sezení. Další možností by bylo vytvořit odkaz v barvě pozadí, který by byl pro běžné uživatele skrytý. Všechna sezení obsahující tento požadavek, by bylo možné identifikovat jako sezení robotů. Robotovi můžeme procházení odkazů zakázat pomocí elementu meta nebo souboru robots.txt umístěném v kořenové složce webu, to by bylo ale neefektivní z hlediska indexace stránek.

Identifikace uživatele

Tento úkol je velmi obtížný díky existenci lokální cache, společných firewallů a proxy serverů. Několik uživatelů může být schováno za jednou IP adresou, nebo naopak jeden uživatel může na stránky vstoupit pokaždé s jinou IP adresou. Nejlepší, i když ne nejjednodušší, cestou k jeho řešení jsou metody založené na spolupráci uživatelů. Je možné využít uživatelské registrace, cookies atd.

Metody vycházející pouze z logu využívají heuristickou analýzu. Vycházejí z IP adres, záznamech o webových prohlížečích a operačních systémech. Například pokud ve dvou záznamech chybí ID uživatele, IP adresa je stejná, ale pokaždé byl použit jiný software, můžeme tak identifikovat dva různé uživatele.

Další možností je konstrukce navigační cesty pro každého uživatele. Pokud je vyžádána stránka, na kterou nebyl odkaz na stránkách, které již uživatel viděl, jedná se s největší pravděpodobností o různé uživatele se stejnou IP adresou.

Pomocí této metody neodhalíme dva uživatele se stejnou IP adresou, kteří používají stejný webový prohlížeč a operační systém a zajímají se o stejné stránky. Stejně tak

neodhalíme jednoho uživatele, který má spuštěny dva webové prohlížeče, nebo který vypisuje URL přímo bez použití hypertextových odkazů
(Cooley, Mobasher & Srivastava 1999).

Identifikace sezení (návštěvy)

U logů, které obsahují delší časové období, je velmi pravděpodobné, že uživatel navštíví webové stránky v tomto období více než jednou. Cílem identifikace sezení je rozdělit přístupy každého uživatele na stránky na jednotlivá sezení. Nejjednodušší metodou jak toho dosáhnout je využití časového limitu. Z empirických dat bylo zjištěno, že pokud po posledním požadavku nenásleduje další do 25,5 minut, sezení bylo ve většině případů ukončeno (Cooley, Mobasher & Srivastava 1999). Většina produktů proto používá jako přednastavený časový limit 30 minut, jehož délka zaručuje identifikaci maximálního počtu sezení.

Identifikace page view

Když si uživatel vyžádá jednu stránku, je staženo mnoho souborů a o každém je patřičný záznam v logu. Cílem je seskupit tyto záznamy v jeden celek, který nám dá přehled o tom, co uživatel v daný okamžik viděl na monitoru. Je to důležité například, když se používají rámce, které umožňují zobrazit několik stránek najednou (Čenovský 2003).

Page view se obvykle identifikuje pomocí času požadavku. Pro požadavky učiněné ve stejný okamžik (tj. ve stejnou sekundu), ponecháme pouze první požadavek zaznamenaný v logu a vyřadíme následující. Po provedení page view identifikace, obsahuje logovací soubor pouze jeden požadavek pro každou akci uživatele.

Kompletace cesty

Problémem určení unikátního uživatelského sezení je určení, zda jsou všechny přístupy zaznamenány v logovacím souboru. Cílem je určit stránky, které zachytila webová nebo proxy cache. Pokud je učiněn požadavek, který není přímo odkazem z poslední stránky, kterou si uživatel předtím vyžádal, může být zjištěno, z které stránky odkaz pochází. Pokud je stránka s tímto odkazem v nedávné historii uživatelových požadavků, je možné se domnívat,

že se uživatel vrátil nazpět pomocí zpětného tlačítka v prohlížeči. Pokud odkaz obsahuje více stránek v historii, tak se za zdroj nového požadavku považuje stránka, jejíž požadavek je novému nejblíže. Odvozené chybějící odkazy jsou přidány do uživatelského sezení. Doba přístupu na odvozenou pomocnou stránku může být zjištěna například průměrem z předchozího a následujícího požadavku (Cooley, Mobasher & Srivastava 1999).

Identifikace episody

Identifikace episody je krok vytvářející podmnožiny z uživatelských sezení. Cílem je identifikovat dílčí zájmy uživatele. Při identifikaci episody se vychází ze sémantického popisu jednotlivých stránek. Určitý web může mít několik okruhů zájmů, např. novinky, fórum, obchod atd. Podle sémantického popisu jednotlivých stránek se vytvoří sémantický hierarchický strom, na jehož kořeni je domovská stránka a na jednotlivých větvích jsou stánky rozdělené podle jednotlivých témat. Těmto stránkám je přidělena určitá hodnota, čím je význam stránek odlišnější, tím je větší rozdíl mezi hodnotami. Pomocí těchto rozdílů se identifikují episody. Když uživatel cestuje z jednoho konce webu na druhý, zjišťuje se rozdíl mezi hodnotou stránky, na které právě stojí a stránkami, které navštívil předtím. Jakmile je překročena předem stanovená hodnota, začíná nová episoda, příslušející jinému tématu (Tanasa 2005).

K automatickému hledání sémantické hierarchie stránek se využívají techniky Web Content a Text Miningu.

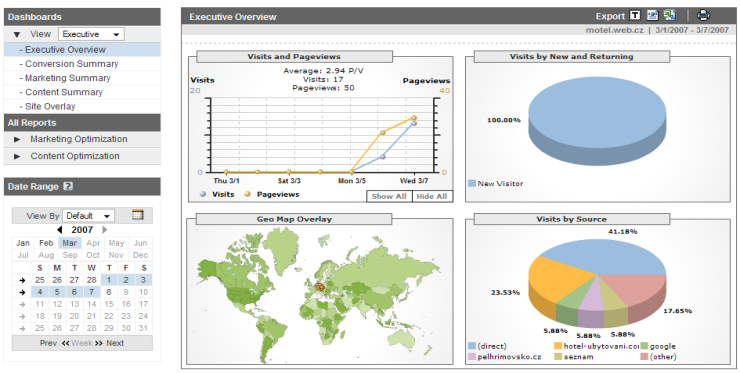
Metody Web Usage Miningu pro objevování vzorů

Statistická analýza

Statistická analýza je nejběžnější metodou používanou k získávání znalostí o návštěvnících webových stránek. Analýza se provádí na předzpracovaném logovacím souboru (soubor se sezeními). Provádějí se různé druhy deskriptivních statistik. Z proměnných page view, času zobrazení a délky navigační cesty se počítají četnosti, průměry, mediány atd. Většina nástrojů pro statistickou analýzu poskytuje pravidelné zprávy, obsahující statistické informace o stránkách s nejčastějším přístupem, průměrné době zobrazení stránky, nebo nejfrekventovanější cestě po kolekci webových stránek. Zpráva také může obsahovat informace o nefunkčních URI či o neautorizovaných přístupech.

Znalosti získané pomocí statistické analýzy mohou být využity pro zlepšení provedení systému, jako podklad při modifikaci stránek, zvýšení bezpečnosti systému a také jako podpora pro marketingová rozhodnutí.

V současné době začala společnost Google nabízet zdarma službu Google Analytics. Služba poskytuje komplexní statistický přehled o uživatelských přístupech na webové stránky. Zjišťuje, jak návštěvníci pracují se stránkami, identifikuje navigační cesty, určuje užitečnost klíčových slov ve vyhledávacích strojích, odkud přichází nejlepší zákazníci, které trhy jsou nejziskovější atd. Vše zobrazuje pomocí grafických výstupů.



Obr. 5 **Google analytics**

Navigační vzory *(Navigation patterns)*

Metoda navigačních vzorů se zaměřuje na techniky studující chování uživatele pohybujícího se po webových stránkách. Porozumění uživatelským navigačním preferencím je základním krokem procesu přizpůsobení a adaptace uživatelského rozhraní individuálnímu uživateli, a zlepšení statické struktury hypertextového systému.

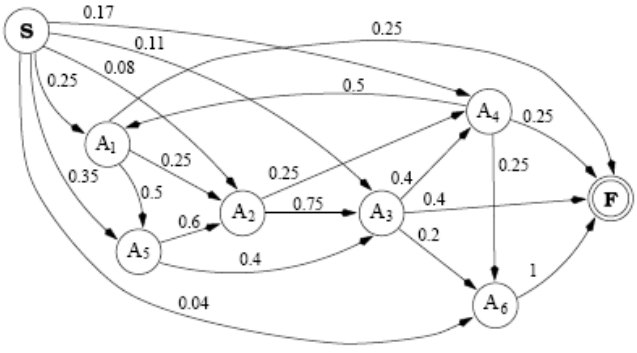
Pro modelování navigace se nejčastěji využívá technika hypertextové pravděpodobnostní gramatiky (Hypertext Probabilistic Grammar – HPG). HPG vytváří ze záznamů o navigaci řetězce, které odpovídají preferovaným cestám uživatelů. Jako vstupní data se využívají uživatelská sezení, která se pomocí HPG převedou do hypertextového pravděpodobnostního jazyka.

Hypertextová Pravděpodobnostní Gramatika

Gramatika využívá prostého zobrazení mezi koncovými a nekoncovými symboly. Každý nekoncový symbol odpovídá jedné webové stránce, produkční pravidlo odpovídá linku mezi stránkami, jako *S* a *F* jsou označeny stavy odpovídající startu a konci navigačního sezení. Z množiny uživatelských sezení je třeba zjistit, kolikrát byla každá stránka požadována, kolikrát byla žádána jako první stránka v sezení a kolikrát jako poslední. Počet sekvencí dvou stránek, které se objevili v sezení, vypovídá o tom, kolikrát byl určitý link použit (Borges, Levene 1999b).

Následující obrázek znázorňuje hypertextovou gramatiku vytvořenou pro množinu šesti cest po šesti stránkách. ID označuje jednotlivé cesty, A_i zastupují jednotlivé stránky.

ID	Trail
1	$A_1 \rightarrow A_2 \rightarrow A_3 \rightarrow A_4$
2	$A_1 \rightarrow A_5 \rightarrow A_3 \rightarrow A_4 \rightarrow A_1$
3	$A_5 \rightarrow A_2 \rightarrow A_4 \rightarrow A_6$
4	$A_5 \rightarrow A_2 \rightarrow A_3$
5	$A_5 \rightarrow A_2 \rightarrow A_3 \rightarrow A_6$
6	$A_4 \rightarrow A_1 \rightarrow A_5 \rightarrow A_3$



Obr. 6 **Tabulka cest a odpovídající Hypertextová gramatika**
(Borges, Levene 1999b)

Jednotlivé webové stránky jsou označeny $A_1 - A_6$. Každé cestě z jedné stránky na druhou je přiřazena pravděpodobnost. Pravděpodobnost cesty z A_i na A_j je podílem počtu cest ze stránky A_i na stránku A_j a počtu všech cest vedených ze stránky A_i. Pokud je v některém sezení stránka A_i stránkou poslední, je to chápáno jako cesta do *F*.

Například: $P(A_1 \rightarrow A_5) = \frac{2}{4} = 0{,}5.$

Pravděpodobnost delší cesty (řetězce) je součinem jednotlivých pravděpodobností.

Například: $P(A_1 \rightarrow A_5 \rightarrow A_2) = 0{,}5 * 0{,}6 = 0{,}3.$

Z počátečního bodu *S* jsou vedeny cesty na všechny stránky. Ohodnocení jednotlivých cest vyjadřuje pravděpodobnost, že konkrétní stránka se stane počáteční stránkou během sezení. Výpočet této pravděpodobnosti je demonstrován na příkladě pro stránku A_1 (Borges, Levene 1999b).

$$\pi(A_1) = \frac{\propto * 4}{24} + \frac{(1-\propto) * 2}{6} = 0{,}25$$

Během šesti sezení bylo provedeno dvacet čtyři požadavků na stránky, z toho stránka A_1 byla shlédnuta čtyřikrát a dvakrát byla počáteční stránkou. Hodnota symbolu α se pohybuje v rozmezí $(0{,}1)$, čím je α nižší, tím více zvýhodňuje stránky, které se objevují spíše na začátku než během cesty. V tomto případě $\propto = 0{,}5$.

Výsledný graf HPG pouze popisuje navigaci po stránkách. K získání navigačních vzorů je nutné graf prořezat. K prořezání slouží hodnoty θ a λ. V prvním kroku prořezání se využívá hodnoty θ vyjadřující limit podpory. Pouze stránky A_i, jejichž pravděpodobnost $S \rightarrow A_i$ překročí tuto hodnotu, se mohou stát výchozími stránkami navigačních vzorů. Tím se ze sledování vyřadí řetězce, které mají sice vysokou pravděpodobnost, ale jsou jen zřídkakdy využívány. Kdyby během všech sezení byl z nějaké stránky učiněn pouze jeden požadavek, měl by odpovídající řetězec přiřazenu pravděpodobnost 1, protože je ale málo častý, bude v tomto kroku odebrán. V druhém kroku prořezání se využívá hodnoty λ vyjadřující limit důvěryhodnosti. Do navigačních vzorů budou zařazeny pouze řetězce, jejichž celková pravděpodobnost překročí hodnotu λ. Hodnoty θ a λ dávají analytikovi možnost kontroly nad kvalitou a kvantitou objevených vzorů.

Příklad objevených navigačních vzorů pro různé hodnoty θ a λ.

Tabulka má dvě části. V obou je použit limit podpory ve výši 0,1. V první části je limit důvěryhodnosti stanoven na hodnotu 0,2 a ve druhé na 0,3.

$\lambda = 0.2$ and $\theta = 0.1$				$\lambda = 0.3$ and $\theta = 0.1$	
String	Confidence	String	Confidence	String	Confidence
$A_1 A_2$	0.25	$A_4 A_1 A_5$	0.25	$A_1 A_5 A_2$	0.3
$A_1 A_5 A_3$	0.2	$A_4 A_6$	0.25	$A_3 A_4$	0.4
$A_1 A_5 A_2 A_3$	0.23	$A_5 A_3$	0.4	$A_4 A_1$	0.5
$A_3 A_4 A_1$	0.2	$A_5 A_2 A_3$	0.45	$A_5 A_3$	0.4
$A_3 A_6$	0.2			$A_5 A_2 A_3$	0.45

Obr. 7 **Objevené navigační vzory**
(Borges, Levene 1999b)

Asociační pravidla *(Association rules)*

Asociační pravidla jsou důležitou součástí Data Miningu a oblastí zaměřujících se na tržní chování a spotřební koše. V oblasti Web Usage Miningu se asociační pravidla využívají především na odhalení stránek (častých položek), které jsou nejčastěji požadovány dohromady během single server session.

Jako zdroj dat je postačující pouze seznam ID sezení a URL.

Asociační pravidlo je sdělení ve formě $A \Rightarrow B$, kde A a B jsou disjunktní podmnožiny množiny položek. Každé pravidlo doprovázejí dvě míry, jistota a podpora (Koutri, Avouris, & Daskalaki 2004).

Jistota vyjadřuje procento z transakcí A, které obsahují také B (pravděpodobnost $P(B|A)$).

Podpora vyjadřuje procento z transakcí obsahujících A nebo B (pravděpodobnost $P(A \cup B)$) (Koutri, Avouris & Daskalaki 2004).

Například: $A.html \Rightarrow B.html, \quad jistota = 80\,\%, \quad podpora = 10\,\%$

Uvedené asociační pravidlo se týká návštěvníků, kteří navštívili stránku A a zároveň také směřovali na stránku B. Pouze 10 % návštěvníků serveru shlédlo jednu z těchto stránek, ale 80 % návštěvníků, které zajímala stránka A, se také zajímalo o stránku B. Pravidlo, ale nepodává žádné další informace o ostatních stránkách shlédnutých během návštěvy.

Podpora (Support) podmnožiny $\{i_1, \dots, i_n\}$ z množiny D je definována jako (Čenovský 2003):

$$S(i_1, \dots, i_n) = \frac{count(\{i_1, \dots, i_n\} \in D)}{count(D)}$$

Jistota (Confidence) je podílem sezení, ve kterém jsou přítomny položky předcházející (i_p) i následující (i_n), se sezením, ve kterém je přítomna pouze položka předcházející. Pro pravidlo $i_p \Rightarrow i_{n1}, \dots, i_{nn}$ platí (Čenovský 2003):

$$C(i_p \Rightarrow i_{n1}, \dots, i_{nn}) = \frac{S(i_p, i_{n1}, \dots, i_{nn})}{S(i_p)}$$

Sekvenční vzory *(Sequential patterns)*

Metoda objevování sekvenčních vzorů se pokouší nalézt vzory uvnitř sezení. Hledá skupiny položek, které jsou následovány jinými položkami v časově řazených sezeních nebo episodách. Pomocí clickstream analýzy může firma získat cenný náhled na chování jejích zákazníků. Marketingoví pracovníci mohou predikovat vzory budoucích návštěv, které jim pomohou s umístěním reklamy cílené na konkrétní skupiny uživatelů. Na sekvenčních vzorech mohou být také prováděny analýzy trendů, detekce bodů změny nebo analýzy podobnosti.

Příklad sekvenčního vzoru:
36 % zákazníků, kteří si objednali knihu na stránce *kniha1.html* si také objednali knihu na stránce *kniha4.html* během *deseti* dní. V takovém případě by bylo vhodné na stránku s první knihou umístit reklamu na druhou knihu a podpořit její prodej. Reklama by tak cíleně upozornila skupinu potenciálních zákazníků mající zájem o obě knihy.

Sekvenční vzory se identifikují na množinách časově seřazených návštěv, takzvaných uživatelských sekvencích. Příklad uživatelské sekvence S (Koutri, Avouris & Daskalaki 2004):

$$S =< (C, D)(A, B, C)(A, B, F)(A, C, D)(E) >$$

Jednotlivá písmena označují konkrétní požadavky na dokumenty. Uvedená sekvence se skládá z pěti uživatelových návštěv. Během první návštěvy byly požadovány dokumenty C, D.

Cílem dolování sekvenčních vzorů je vyjmenovat kompletní množinu skládající se z t frekventovaných sekvencí na dané databázi uživatelských sekvencí, kde t je prahem minimálního výskytu. Možným sekvenčním vzorem z příkladu může být $(* C *)$ s interpretací: uživatel se často během sezení zajímal o dokument C. Nebo $(AB *)$ s interpretací: uživatel se často zajímal o dokumenty A a B a poté se zajímal o jiný dokument.

Většina technik pro dolování sekvenčních vzorů je založena na Apriori-like algoritmu. Další techniky jsou založeny na konceptu maximálně dopředných požadavků (Maximal forward references. Výsledkem je transformovaný logovací soubor na množinu vzorů, založených na statisticky významných cestách a asociačních rolích.

Apriori-like algoritmus je založen na vlastnosti: jestliže vzor s *k* položkami není častý, žádný z jeho super-vzorů s *k+1* nebo více položkami nikdy nebude častý. Algoritmus iterativně generuje množinu kandidátských vzorů o délce *k+1* z množiny častých vzorů (zjištěných pomocí podpory a jistoty) o délce *k*, a kontroluje jejich odpovídající frekvenci výskytu v databázi (Shiwei, Jiawei, Dongqing, Jian, Hongjun & Shojiro 2007).

Technika maximálně dopředných požadavků definuje sekvenci požadavků na dokumenty od první stránky až před stránku, na kterou byl učiněn zpětný požadavek. Zpětný požadavek je stránka, která je už obsažena v současné sekvenci. Dopředný požadavek je stránka, která ještě není obsažena v sekvenci (Čenovský 2003).

Shlukování (Clustering)

Shlukování je významnou metodou Data Miningu. Jsou to techniky využívané pro seskupování množin položek, které mají podobné charakteristiky. Oblast Web Usage Miningu se zaměřuje především na shlukování stránek a uživatelů.

Shlukování stránek

Tato technika objevuje stránky s podobným kontextem. Toho mohou využít vyhledávací stroje nebo poskytovatelé webové asistence. Výsledkem může být například: „Stránky o antivirech a firewallech patří do stejného uživatelského shluku“. Shlukování stránek je obecně založeno na obsahových datech, metadatech a datech o struktuře, ale Web Usage Mining jako vstupních dat využívá především požadavky na dokumenty a uživatelské návštěvy (Koutri, Avouris & Daskalaki 2004).

Například: $P = \{P_1, P_2, \dots, P_n\}$ je množina požadavků na dokumenty korespondující s *n* dokumenty dané kolekce webových stránek. Výsledkem shlukování provedeného na množině *P* může být množina $\{P_1, P_2\}$ sestávající ze dvou objektů určených pro podobné užití.

Shlukování uživatelů

Tato technika vytváří skupiny uživatelů, kteří se projevují stejným nebo podobným chováním. Tato znalost je vhodná pro přizpůsobování obsahu Webu jednotlivým skupinám uživatelů. Uživatelé zařazení do různých shluků se zajímají o různé věci (Koutri, Avouris & Daskalaki 2004).

Například: $V = \{v_1, v_2, \dots, v_m\}$ je množinou uživatelských návštěv, kde každou návštěvu reprezentuje sekvence požadavků na dokumenty, např. $v_1 = (P_1 \rightarrow P_2 \rightarrow P_3)$. Výsledkem shlukování provedeného na množině *V* může být množina $\{v_2, v_3\}$ sestávající ze dvou návštěv s podobnými požadavky.

Tradiční shlukování využívá následující charakterizaci shluků:

- exkluzivní, těžké shluky – objekt náleží pouze do jednoho shluku
- překrývající se shluky – objekt může náležet do několika shluků
- pravděpodobnostní shluky – objekt náleží do všech shluků s určitou pravděpodobností
- fuzzy shluky – objekt náleží do každého shluku s určitým stupněm účasti

Klasifikace

Klasifikace je metodou rozdělující datové položky do několika předdefinovaných tříd. Webmaster nebo marketingový pracovník může tuto metodu použít k vytvoření uživatelských profilů náležejících do specifických tříd nebo kategorií. To vyžaduje extrakci a selekci hlavních rysů, které nejlépe popisují vlastnosti dané třídy nebo kategorie. Klasifikace může být provedena pomocí induktivně se učících algoritmů, jako je klasifikace pomocí rozhodovacích stromů, naivní bayesovský klasifikátor, atd. (Srivasta, Cooley, Deshpande & Pang-Ning 2000).

Rozhodovací stromy identifikují objekty, popsané různými atributy, do tříd. Rozhodovací strom se vytvoří z množiny daných objektů, které musí někdo zařadit do skupin. Jedná se tedy o učení s učitelem. Každý uzel stromu představuje jednu vlastnost objektu, z uzlu vede konečný počet hran (Wikipedia 2007a).

Naivní bayesovský klasifikátor je jednoduchý pravděpodobnostní klasifikátor založený na aplikaci bayesova teorému s naivním nezávislým předpokladem. Klasifikátor může být natrénován velmi efektivně při učení s učitelem. K výpočtu parametrů nezbytných pro klasifikaci vyžaduje pouze malé množství trénovacích dat (Wikipedia 2007b).

Výsledkem klasifikace může být tvrzení:
40 % uživatelů, kteří učinili objednávku v */obchod/knihy/*, mají věk 25-35 let a jsou z Jižních Čech.

Modelování závislostí

Cílem modelování závislostí je vyvinout model, který bude schopný reprezentovat významné závislosti mezi různými proměnnými na webové doméně. Metoda je například schopná vytvořit model popisující uživatelské chování na určité kolekci webových stránek od návštěvníka, který sem přišel poprvé, až po pravidelného uživatele. Získaných znalostí je možné využít k predikci požadavků na web, zlepšení dokumentace a on-line pomoci (Barsagade 2003).

Existuje několik pravděpodobnostních technik užívaných k modelování uživatelského chování, většina je založena na hidden markovově modelu a bayesovských důvěrných sítích.

Hidden markovovův model je matematický aparát podobný konečnému automatu. Model má pevně zadanou množinu stavů, do kterých se může v průběhu výpočtu dostat, přechodová funkce je tvořena maticí přechodu a množina koncových stavů je nahrazena maticí pravděpodobnosti generovaných vzorů (Křivánek).

Bayesovské důvěrné sítě představují grafický model vztahů. Sítě se skládají ze dvou částí. Orientovaného acyklického grafu, kde každý uzel reprezentuje proměnnou a každá hrana pravděpodobnostní závislost, a tabulky podmíněných pravděpodobností, které ke každé proměnné udávají hodnotu pravděpodobnosti vlivu kombinací jejich rodičů (Rychlý 2005).

Modely neposkytují jen teoretický rámec pro analýzu chování uživatelů, ale jsou i potenciálně užitečné pro predikci budoucího užívání webu. To může pomoci při vyvíjení strategií pro zvýšení prodeje nabízených výrobků, nebo napomoci při zlepšování provedení navigace.

Analýza objevených vzorů

Analýza objevených vzorů je posledním krokem celého procesu Web Usage Miningu. Analýza přeměňuje objevené vzory, pravidla a statistiky do znalostí. Analýza se skládá ze dvou částí: validace a interpretace získaných vzorů (Jiang 2003).

Validace odfiltrovává nezajímavé, irelevantní vzory a pravidla z výstupu vytvořeného ve fázi objevování vzorů. Rozlišení toho co je zajímavé nebo nezajímavé je subjektivním problémem. Například marketingového analytika budou zajímat nejčastější přístupové vzory, a naopak bezpečnostního analytika zase zajímají neobvyklé vzory s malou frekvencí.

Interpretace převádí matematický výstup data-miningových algoritmů do podoby, která je srozumitelná lidskému uživateli.

Techniky užívané při analýze vzorů

Analýza neprobíhá automatizovaně, analytik musí zredukovat objevené vzory a určit ty z jeho pohledu nejdůležitější. Různé softwarové nástroje využívají různé techniky k provedení analýzy, nejčastěji nabízejí dotazovací mechanismus, vizualizaci nebo OLAP (Srivasta, Cooley, Deshpande & Pang-Ning 2000).

Dotazovací mechanismus

Dotazovací mechanismus jako SQL dovoluje analytikovi extrahovat pouze relevantní a použitelné vzory chování pomocí specifikace různých omezení.

Vizualizace

Vizualizace pomáhá lidem porozumět abstraktním pojmům. Vizualizační techniky jako sestavování grafů vzorů nebo přidělení barev odlišným hodnotám mohou často zvýraznit celkové, průměrné vzory nebo trendy v datech.

On-Line Analytical Processing (OLAP)

Načtení uživatelských dat do datové kostky umožňuje provádění OLAP operací. Technika umožňuje strukturalizaci a granulaci dat, analýzu souhrnných dat, filtraci dat a jejich uspořádání do smysluplných podmnožin. OLAP poskytuje odpovědi na základní otázky (kdy, kde, co, kdo) i kombinované dotazy. Analytik tak může provádět ad-hoc analýzy na datech ve více dimenzích.

Oblasti aplikace Web Usage Miningu

Oblasti aplikace je možno rozlišit na dvě hlavní oblasti. První je oblast personalizace, která objevuje preference a potřeby individuálních uživatelů k tomu, aby mohla poskytnout přizpůsobené webové stránky určitým typům uživatelů. Druhá oblast aplikace se nezaměřuje na individuální uživatele, ale na hlavní uživatelské navigační vzory, aby porozuměla tomu, jak jsou stránky užívány obvyklými, nejčastějšími uživateli. Informace jsou pak využívány ke zlepšování systému (výkon a bezpečnost), stránkové modifikaci, business inteligenci, webové charakterizaci a uživatelské charakterizaci (Srivasta, Cooley, Deshpande & Pang-Ning 2000).

Personalizace

Cílem personalizace je poskytnout jednotlivým uživatelům dynamický obsah, který je přizpůsoben jejich individuálním zájmům. Zájmy návštěvníka jsou odvozovány z jeho uživatelského profilu, jeho vzorů chování nebo vzorů chování posledního návštěvníka, který má podobný profil. To umožňuje identifikovat potenciálně zajímavé odkazy pro konkrétního uživatele, a následně mu doporučit jednu nebo více položek nebo stránek. V e-shopech je tak možné vytvářet cílenou nabídku konkrétního zboží. Když si zákazník při minulé návštěvě zakoupil fotoaparát, může mu být při další návštěvě doporučena například paměťová karta nebo jiné zboží související s fotografováním.

Výkon a bezpečnost

Web Usage Mining je klíčem k porozumění pohybu po webových stránkách. Vysoký výkon webových aplikací a ostatní atributy kvality služeb jsou rozhodujícími veličinami ovlivňujícími uspokojení uživatelů. Poznatky z analýz jsou podkladem pro vyvinutí politik pro kešování, síťový přenos, rozdělení výkonu a distribuci dat. V poslední době se také neustále zvyšují požadavky na bezpečnost systémů, především v oblasti e-commerce. Web Usage Mining odhaluje vzory chování, které jsou důležité pro detekci vniknutí, podvodů a pokusů o nabourání se do systému.

Stránková modifikace

Struktura webových stránek je dalším rozhodujícím atributem úspěchu. Zatraktivněním webových stránek je možné získat další návštěvníky, kteří byli doposud uživateli jiných serverů s podobným obsahem. Web Usage Mining umožňuje náhled do

navigačního chování uživatelů, což může být nejdůležitějším podkladem pro přepracování stránkové struktury a úpravě obsahu pro zlepšení navigačního pohodlí. Dobře zvolená struktura a obsah jsou důležité ve většině aplikací, ale hlavně v oblasti e-commerce u produktových katalogů.

Business inteligence

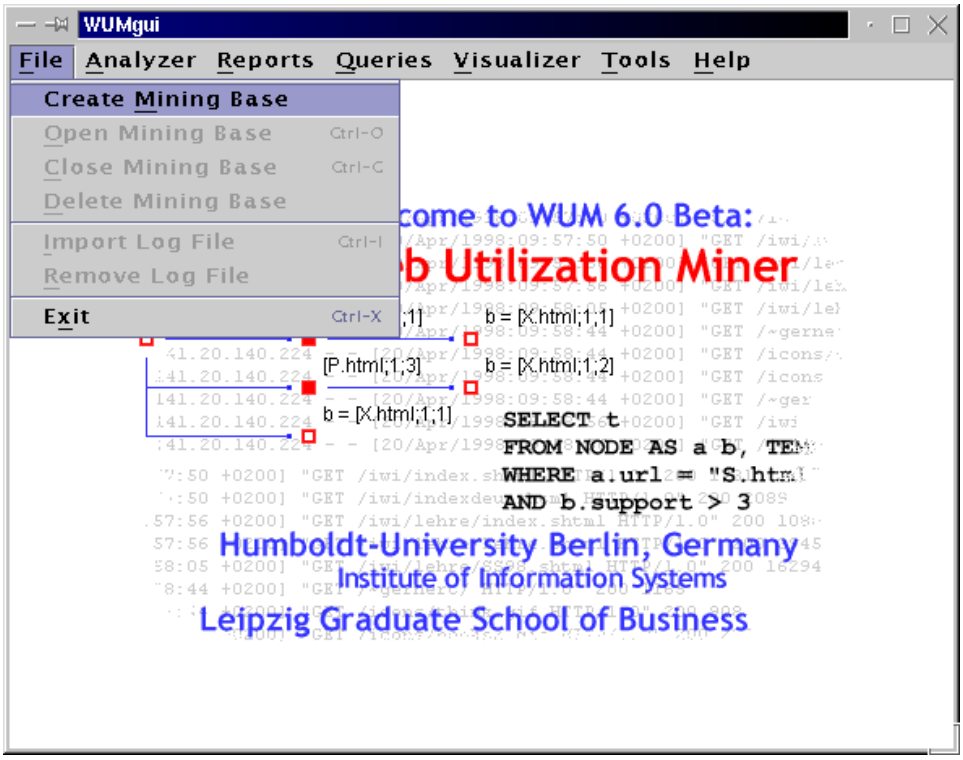
Informace o tom, jak zákazníci užívají webové stránky, jsou rozhodujícími informacemi pro marketingové pracovníky v oblasti e-commerce. Web Usage Mining pomáhá optimalizovat obchodní procesy a provádět marketingová rozhodnutí. Na základě datové kostky, která v sobě spojuje web usage data spolu s marketingovými daty, je možné identifikovat, v které fázi životního cyklu se nacházejí nabízené produkty. Podle jednotlivých fází: zavádění, růst, zralost a pokles, je zvolena vhodná marketingová strategie.

Charakterizace uživatele

Charakterizace uživatele se nezaměřuje na specifické uživatele, ale na charakterizaci obecného uživatele. Metoda pomáhá ve studiu, jak jsou užívány webové prohlížeče a jaká je interakce uživatelů s konkrétními uživatelskými rozhraními. Charakterizace uživatele je prováděna pomocí speciálně upravených prohlížečů, které ukládají logovací soubory na straně klienta, jedním z nich je prohlížeč Xmosaic. Ukládání dat na straně klienta umožňuje vytvářet detailní statistiky o různých proměnných, jako je používání tlačítek Zpět nebo Dopředu, pořizování záložek nebo přímo ukládání stránek.

SW nástroje a existující řešení pro Web Usage Mining

WUM: Web Utilization Miner

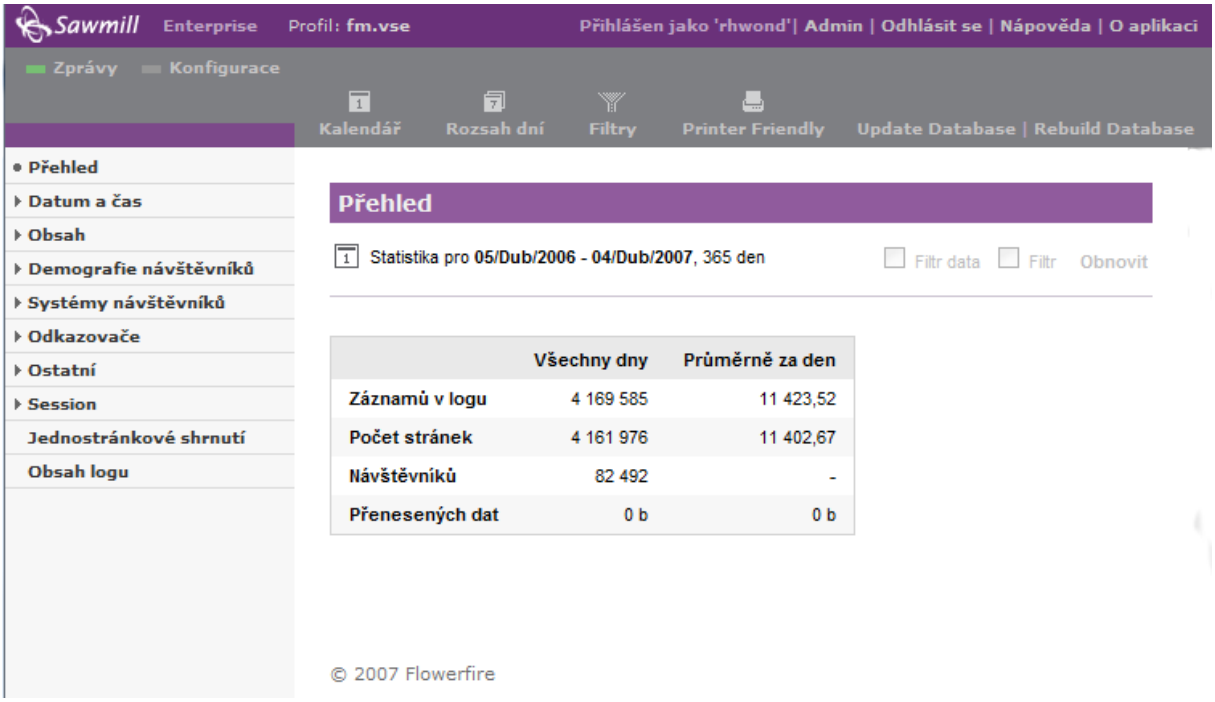


Obr. 8 **Web Utilization Miner**

WUM je nástroj k dolování dat, jehož primárním účelem je analýza navigačního chování uživatelů navštěvujících webový server. Je to integrované prostředí pro preparaci logu, dotazování a vizualizaci. Preparační nástroj předzpracuje data a zorganizuje log na základě sezení podle specifických kritérií. Agregační nástroj transformuje log do stromové struktury, kde jsou stejné sekvence sloučeny. Jeho vizualizační mechanismus zobrazuje uzly zahrnující požadované vzory a odlišuje nefrekventované cesty lokalizované mezi nimi. To je dobré k zjištění, jak jsou webové stránky opravdu navigovány. Program používá vlastní jazyk MINT, pomocí něhož můžeme zadávat dotazy nad daty v databázi. Dotazování je nejdůležitější nabídkou programu. Program je platformově nezávislý. (HypKNOWsys 2005)

Web Utilization Miner je nejčastěji zmiňovaným nástrojem v odborné literatuře.

Sawmill7

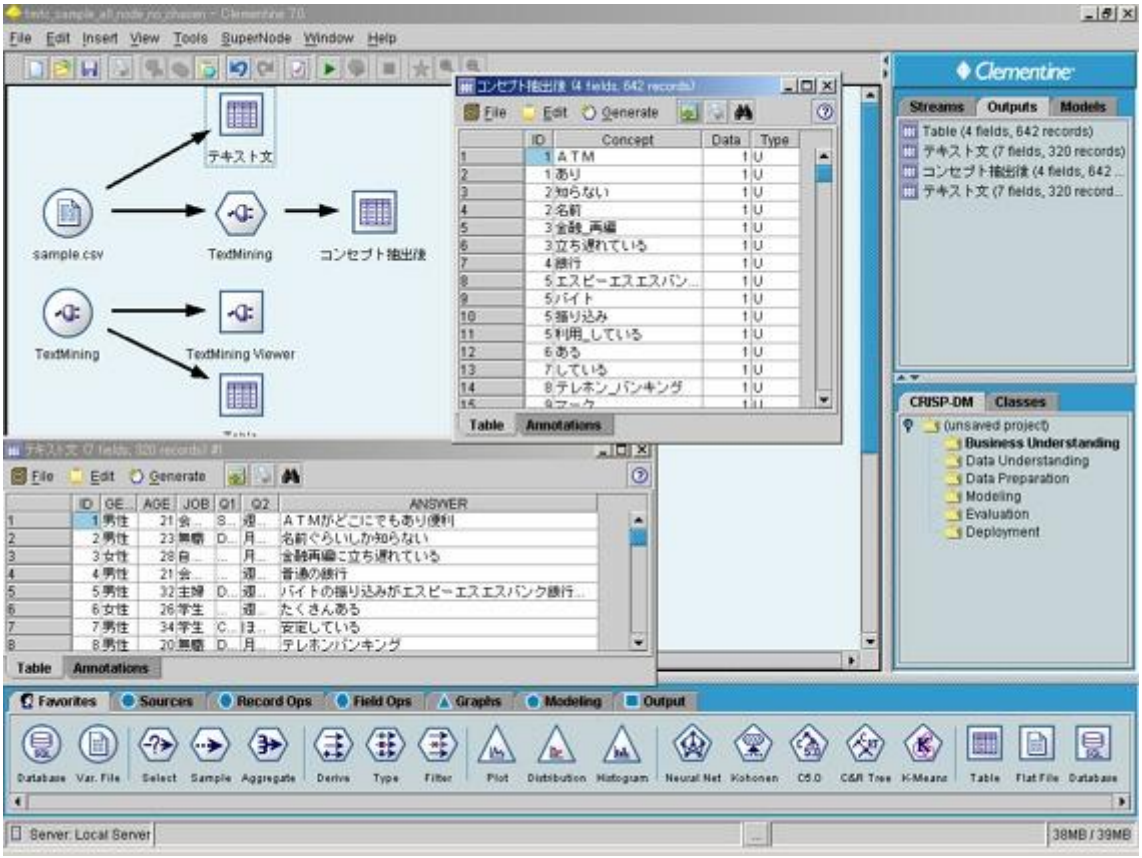


Obr. 9 **Sawmill 7 Enterprise**

Platformově nezávislý nástroj, nabízený ve třech verzích (Lite, Professional, Enterprise), podle množství nabízených funkcí. Podporuje více procesorové systémy. Plně podporuje MySQL databáze. Používá vlastní skriptovací jazyk (Salang), díky kterému je možné kompletně přizpůsobit uživatelský a oznamovací interface. Může být spuštěn lokálně na serveru, který je analyzován, nebo vzdáleně. Je v šesti jazykových verzích včetně češtiny. Podporuje 635 formátů logovacích souborů. Importní funkce umožňuje filtrovat surová logovací data. Podává podrobné zprávy o datu a čase, obsahu, demografii návštěvníků, odkazovačích, sezeních atd. (Sawmill 2007)

Podrobněji v následujícím textu (Praktický návrh konkrétního řešení).

Clementine

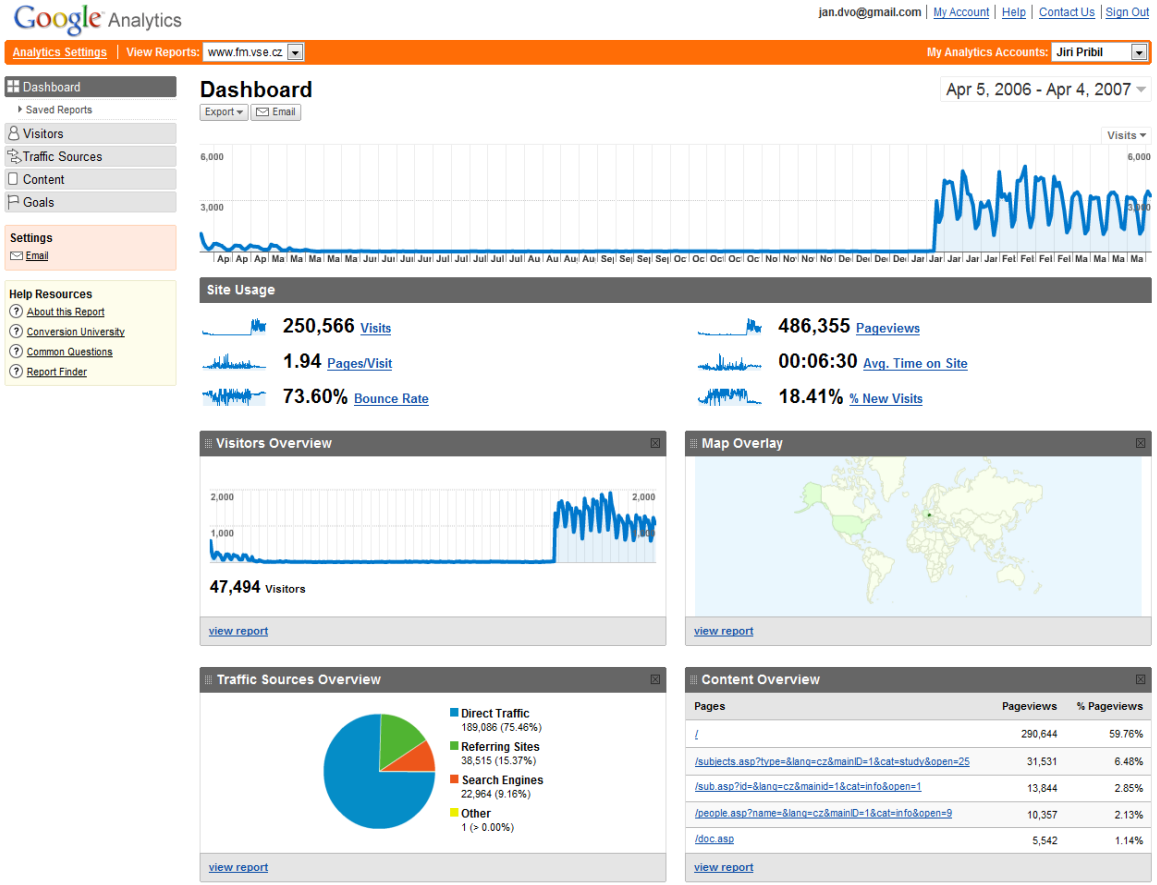


Obr. 10 **Clementine**

Clementine je komplexním nástrojem pro data mining. S Clementine je možné jednoduše zpřístupnit, připravovat či integrovat číselná data, textová data, webová data nebo data z výzkumu. Umožňuje rychle vybudovat a vyhodnotit různé modely s využitím nejpokročilejších statistických a data miningových technik. Efektivně modely aplikuje, a to buď v plánovaných dávkách nebo v reálném čase, a předává výsledky jak uživatelům tak automatickým systémům (SPSS 2007).

Web miningová analýza je zajištěna pomocí doplňkového modulu Web Mining for Clementine. Jednoduše transformuje surová webová data do podoby vhodné k analýze. Rekonstruuje uživatelské navigační cesty a identifikuje uživatelská sezení. Pokročilý dolovací algoritmus odkrývá uživatelské pohyby po webových stránkách. Konečným výsledkem je kolekce cenných navigačních vzorů, které pomáhají web masterům lépe porozumět uživatelskému chování (Galeas 2006).

Google Analytics



Obr. 11 **Google analytics**

Služba je dostupná z adresy: www.google.com/analytics/

Pro využívání služby je nutné mít u společnosti Google zřízen e-mailový účet (Gmail). Po bezplatné registraci je vygenerován skript, který je nutné přidat do HTML kódu každé stránky. Skript se zapisuje až na konec stránky před tag </body>, proto nikterak neprodlužuje dobu načítání stránky. To je vše, co je nutné udělat.

Služba sleduje více než čtyřicet různých cílů, které jsou zaměřeny na sledování návštěvníků (geografický původ, nové versus opakované návštěvy, jazyky, návštěvní trendy, loajalitu, schopnosti prohlížečů, nastavení sítě), odkazující zdroje (přímý přístup, odkazující stránky, vyhledávací stroje, klíčová slova) a obsah (nejžádanější stránky, vstupní a výstupní stránky atd.). Kromě toho je možné si nadefinovat i další cíle. Vše je, jak demonstruje obrázek 11, prezentováno přehlednou grafickou formou.

Další softwarové nástroje

123LogAnalyzer

Nástroj pro rychlou analýzu logu. Analýza návštěvníků, zlepšení PageRank ve vyhledávacích strojích, určení a sledování nechtěných návštěvníků, kteří zneužívají stránky.

AlterWind Log Analyzer Lite

Pomáhá určit základní charakteristiky přístupů na stránky. Zjišťuje jaké vyhledávací stroje a fráze nám přinesly návštěvníky, které odkazující stránky generují největší pohyb, stanovuje chyby objevující se na stránkách.

Analog

Určí nejpopulárnější stránky, z kterých zemí nás lidé navštěvují, na kterých stránkách jsou nefunkční odkazy atd.

ANGOSS KnowledgeWebMiner

Nástroj vhodný pro clickstream analýzy.

Azure Web Log analyzer

Poskytuje informace o nejpopulárnějších stránkách a souborech, počtu návštěvníků a jejich původu, jaké počítače a prohlížeče používají, jaký je čas načítání stránek, jaký je na nich pohyb a zda neobsahují chyby.

ClickTracks

Webový analytický program, zobrazuje vzory chování přímo na souvisejících webových stránkách.

Datanautics

Nástroj pro elektronické obchody, který analyzuje chování zákazníků.

Download Analyzer

Umožňuje sledovat návštěvníky, příchody, downloady, odkazující stránky, hledané výrazy, poskytuje informace pro zlepšení propagace a optimalizaci ve vyhledávacích strojích.

Htminer

Analýza logu objevující unikátní návštěvníky, sezení, transakce. Organizuje data do PostgresSQL datových skladů.

LiveStats from DeepMetrix

Poskytuje sofistikovanou analýzu logu v reálném čase. Zobrazuje navigační cesty, sezónní pohyby, klíčová slova, geografický původ.

Nihuo Web Log Analyzer

Zjišťuje počet návštěvníků, jak se pohybují ze stránky na stránku, z kterých vyhledávací strojů přišli, kolik času strávili na které stránce.

prudsys ECOMMINER

Nástroj pro elektronické obchody. Kombinuje clickstream a databázovou analýzu.

Speed Tracer

Je to analytický nástroj, který sleduje vzory v prohlížení stránek. Vytváří zprávy, které pomáhají webmasterům vylepšit strukturu a navigaci. Využívá odvozovací mechanizmus k rekonstrukci cesty a identifikaci uživatelského sezení. Vytváří statistiky o délce přístupu, nejčastějších cestách a skupinách nejžádanějších stránek.

STstat

Je to sada CGI skriptů, které produkují HTML statistiky založené na logovacích souborech.

Surf Pattern Visual Anylyzer

Identifikuje klíčové navigační vzory. Identifikuje, odkud návštěvníci přišli, které jsou nejčastější příchozí a odchozí stránky, nejfrekventovanější navigační cesty atd.

Visitator

Shlukuje a vizuálně prezentuje skupiny návštěvníků založené na přístupových vzorech.

Web Site Analysis

Grafický nástroj k analýze logu přístupný z webového prohlížeče.

Webalizer

Nástroj pro analýzu logovacích souborů www serverů. Vytváří formátované zprávy, barevné statistické grafy. Umožňuje úpravu řady nastavení, ukládá dlouhodobé statistiky pro použití bez nutnosti opětovného zpracování starých logů.

WebLog Expert 2.0 for Windows

Rychlá a výkonná analýza logu. Poskytuje statistiky o aktivitě, přístupech k souborům, cestách mezi stránkami, požadovaných stránkách, vyhledávacích strojích, prohlížečích operačních systémech atd.

WebLog

Komplexní nástroj pro analýzu logu. Dovoluje sledovat aktivitu na stránkách za měsíc, týden, den, hodinu. Monitoruje chyby, přenesené byty, page views, a sleduje nejpopulárnější stránky.

Weblog_parse

Extrahuje specifická pole z logovacího souboru. Rozebírá je a vypisuje pouze pole, týkající se uživatelského chování, do tabulek pro snadnější manipulaci.

Praktický návrh konkrétního řešení

V následujícím textu provedu analýzu webových logů Fakulty managementu Vysoké školy ekonomické v Praze. Původně jsem zamýšlel provést analýzu nějakého e-shopu. Při analýze elektronického obchodu by bylo možné dosáhnout velice zajímavých výsledků. Porozumět nákupnímu chování uživatelů. Nebo pokusit se zlepšit strukturu navigace, protože elektronické obchody jsou většinou velmi rozsáhlé a mají složitou strukturu odkazů. Takové logovací soubory se mi ale nepodařilo získat. Všichni si jich velice cení a obávají se z úniku informací. V jednom případě mi bylo dokonce sděleno, že mi logy neposkytnou, nicméně analýza uživatelského chování je zajímá natolik, že by si ji na základě mé práce chtěli provést sami.

Protože je získání logů velmi obtížné, tímto bych chtěl poděkovat pracovníkům centra výpočetní techniky naší fakulty za spolupráci a pomoc při získání logovacích souborů.

Tak jak stránky vypadají v současné podobě, byly spuštěny dne 5. dubna 2006. Pátý duben se tak jeví jako vhodný den pro zahájení analýzy. Otázkou bylo jen stanovení koncového termínu, za který ještě budou logy zahrnuty do analýzy. Protože se v průběhu roku mění zájmy uživatelů, v určitých obdobích se přihlašují na zkoušky, v jiných obdobích je zase zvýšený zájem o stránky potenciálními uchazeči o studium, rozhodl jsem se pro časový úsek jednoho roku. V období jednoho roku jsou všechny navigační vzory zastoupeny rovnoměrně. Koncovým termínem je tedy 4. duben 2007.

Logovací soubor nejdříve předzpracuji a vyčistím. Po identifikaci sezení na něm na základě předchozího textu provedu identifikaci navigačních vzorů a statistickou analýzu. Pro tyto dvě metody jsem se rozhodl, protože mě osobně nejvíce zaujaly.

Pracoval jsem na počítači:
s procesorem AMD Athlon™ 64 X2 Dual Core Processor 5200+ 2,6G Hz
a s operační pamětí 2047 MB.

www.fm.vse.cz

Současná podoba webových stránek Fakulty managementu Vysoké školy ekonomické v Praze je od 5. dubna 2006. Protože se v současné době připravuje jejich grafická změna, přikládám obrázek úvodní stránky pro dokumentaci. Současně přikládám mapu webu, která pomůže pochopit strukturu webových stránek.

Obr. 12 **Domovská stránka www.fm.vse.cz**

Mapa webu www.fm.vse.cz

O fakultě managementu
- **Úvodní slovo děkana fakulty**
- **Historie FM**
- **Vedení**
- **Akademický senát FM**
- **Organizační struktura**
- **Hledání osob**
- **Dokumenty**
- **Nabídka zaměstnání na FM**
- **Kontakt**

Věda a výzkum
- **Oddělení pro vědu a výzkum**
- **Publikace**
- **Vědecká rada**
- **Doktorské studium**
- **MATEO**

Katedry
- **Management informací**
  - Personální obsazení
  - Předměty
- **Společenské vědy**
  - Personální obsazení
  - Předměty
  - WWW stránky KSV
- **Management veřejného sektoru**
  - Personální obsazení
  - Předměty
  - WWW stránky katedry
- **Management podnikatelské sféry**
  - Personální obsazení
  - Předměty
  - WWW stránky KMPS
- **Institut managementu zdravotnictví**
  - Personální obsazení
  - Předměty
  - WWW stránky IMZS

Ostatní
- **Program pět P**
- **WWOOF**
- **The sometimes**
- **AFHL**
- **AFbL**
- **Stránky studentů a zaměstnanců**
- **Vstupní vzdělávání úředníků**
  - Obsah vzdělávacího programu
  - Přehled vyučovaných předmětů
  - Termíny kurzů
  - Přihlášky
  - Kontakt
- **Unicef**

Studium
- **Zabezpečené osobní stránky**
- **Předměty**
- **Rozvrhy**
- **Bakalářské a diplomové práce**
- **Studijní řády**
- **Studium na FM**
- **Konzultační hodiny pedagogů**
- **Přijímací řízení**
- **Harmonogram 2007/08**
- **Odevzdávání el. verzí BP a DP**
- **Harmonogram 2006/07**
- **Harmonogram registrací 06/07**
- **Harmonogram registrací 07/08**
- **Studijní plány**
  - Bakalářský stud. program
    - 2006/2007
    - 2007/2008

Oddělení
- **Studijní oddělení**
  - Zpětná vazba
  - Personální obsazení a kontakt
  - Zápisy
- **Centrum výpočetní techniky**
  - Informace CVT
  - Personální obsazení a kontakt
  - eduroam
- **Knihovna**
- **Ediční oddělení**
  - Ediční plán
  - Personální obsazení a kontakt
- **Centrum celoživotního vzdělávání**
  - Mimořádné studium
  - Kurzy pro rok 2006/07
  - Kontakt
  - Rozvoj profesních kompetencí pedag. pracovníků
  - MATEO
  - Workshop Strategické nástroje pro management škol
  - Univerzita třetího věku
- **Koleje a menza**

Služby
- **Webmail**
- **Klokan**
- **Palladium**
- **Skripta On-line**
- **SPEED**
- **KAPR**

Obr. 13 **Mapa webu**

Zdroj dat

Jako zdroj dat pro analýzu jsem použil logovací soubory webového serveru Fakulty managementu Vysoké školy ekonomické v Praze. Jedná se o zdroj dat na úrovni serveru. Druhým zdrojem dat jsou data od vzdáleného agenta (Google Analytics), který sbírá data na základě Java skriptů vložených do každé webové stránky. V tomto případě se jedná o zdroj dat na úrovni klienta. Nejdůležitějším zdrojem dat pro analýzu budou logovací soubory.

Webovým serverem, který používá Fakulta managementu, je Microsoft Internet Information Server. Ten umožňuje sledovat různé atributy, jejich množství je volitelné. Za každý den je vytvořen logovací soubor s názvem exrrmmdd.log (například ex060405.log).

V současné době je server nastaven, aby pro každý požadavek sledoval atributy: date, time, s-sitename, s-ip, cs-method, cs-uri-stem, cs-uri-query, s-port, cs-username, c-ip, cs(User-Agent), cs-status, cs-substatus, cs-win32-substatus. Množství položek je větší než v běžném log formátu (Common Log Format), jedná se tak o rozšířený formát (Extended Log Format). Data získaná z logů, proto budou pro analýzu dostatečná. Bohužel zde ale chybí položka cs(Referrer), do které se zaznamenávají linky na stránku, kterou uživatel navštívil před odesláním požadavku. Položka referrer by tak mohla být cenným zdrojem informací.

Popis jednotlivých položek:

(Microsoft 2007)
date – datum, kdy aktivita proběhla
time – čas, kdy aktivita proběhla
s-sitename – název služby a instanční číslo
s-ip – IP adresa serveru
cs-method – požadovaná akce, například metoda GET
cs-uri-stem – cíl akce, například Default.htm
cs-uri-query – dotaz, který se uživatel pokusil vykonat
s-port – číslo serverového portu, které je konfigurováno pro službu
cs-username – název uživatele, který se přihlásil na server
c-ip – IP adresa klienta, který vykonal požadavek
cs(User-Agent) – typ webového prohlížeče užívaného klientem
cs-status – HTTP stavový kód
cs-substatus – podstav chybového kódu
cs-win32-status – stavový kód Windows

Ukázka webového logu Fakulty managementu VŠE

#Software: Microsoft Internet Information Services 6.0

#Version: 1.0

#Date: 5.4.2006 5:13:11

#Fields:	date	time	s-sitename	s-ip	cs-method	cs-uri-stem	cs-uri-query	s-port	cs-username	c-ip	cs(User-Agent)	sc-status	sc-substatus	sc-win32-status
	5.4.2006	15:29:10	W3SVC1	146.102.248.201	GET	/Default.asp	-	80	-	146.102.249.128	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+(R1+1.1);+.NET+CLR+1.1.4322)	200	0	0
	5.4.2006	15:29:10	W3SVC1	146.102.248.201	GET	/style.css	-	80	-	146.102.249.128	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+(R1+1.1);+.NET+CLR+1.1.4322)	200	0	0
	5.4.2006	15:29:10	W3SVC1	146.102.248.201	GET	/images/fm3.jpg	-	80	-	146.102.249.128	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+(R1+1.1);+.NET+CLR+1.1.4322)	200	0	0
	5.4.2006	15:29:10	W3SVC1	146.102.248.201	GET	/images/novinky.gif	-	80	-	146.102.249.128	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+(R1+1.1);+.NET+CLR+1.1.4322)	200	0	0
	5.4.2006	15:29:10	W3SVC1	146.102.248.201	GET	/images/fmweb.jpg	-	80	-	146.102.249.128	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+(R1+1.1);+.NET+CLR+1.1.4322)	200	0	0
	5.4.2006	15:29:10	W3SVC1	146.102.248.201	GET	/images/banner.jpg	-	80	-	146.102.249.128	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+(R1+1.1);+.NET+CLR+1.1.4322)	200	0	0
	5.4.2006	15:29:10	W3SVC1	146.102.248.201	GET	/hodiny.swf	-	80	-	146.102.249.128	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+(R1+1.1);+.NET+CLR+1.1.4322)	200	0	0
	5.4.2006	15:29:10	W3SVC1	146.102.248.201	GET	/images/footer.jpg	-	80	-	146.102.249.128	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+(R1+1.1);+.NET+CLR+1.1.4322)	200	0	0
	5.4.2006	15:29:16	W3SVC1	146.102.248.201	GET	/english/subjects.asp	Type=Info&ID=SZKVS1&Lang=en	80	-	84.244.122.69	Mozilla/5.0+(Windows;+U;+Windows+NT+5.1;+cs;+rv:1.8.0.1)+Gecko/20060111+Firefox/1.5.0.1	404	0	3
	5.4.2006	15:29:16	W3SVC1	146.102.248.201	GET	/favicon.ico	-	80	-	84.244.122.69	Mozilla/5.0+(Windows;+U;+Windows+NT+5.1;+cs;+rv:1.8.0.1)+Gecko/20060111+Firefox/1.5.0.1	404	0	64
	5.4.2006	15:29:16	W3SVC1	146.102.248.201	GET	/favicon.ico	-	80	-	84.244.122.69	Mozilla/5.0+(Windows;+U;+Windows+NT+5.1;+cs;+rv:1.8.0.1)+Gecko/20060111+Firefox/1.5.0.1	404	0	2
	5.4.2006	15:29:19	W3SVC1	146.102.248.201	GET	/doc.asp	-	80	-	80.78.146.46	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1)	200	0	0
	5.4.2006	15:29:20	W3SVC1	146.102.248.201	GET	/img/folder.gif	-	80	-	80.78.146.46	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1)	200	0	0
	5.4.2006	15:29:20	W3SVC1	146.102.248.201	GET	/images/menubg.gif	-	80	-	80.78.146.46	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1)	200	0	0
	5.4.2006	15:29:23	W3SVC1	146.102.248.201	GET	/img/sipka.gif	-	80	-	212.158.130.217	Mozilla/4.0+(compatible;+MSIE+5.01;+Windows+98;+FunWebProducts)	304	0	0
	5.4.2006	15:29:23	W3SVC1	146.102.248.201	GET	/images/menubg.gif	-	80	-	212.158.130.217	Mozilla/4.0+(compatible;+MSIE+5.01;+Windows+98;+FunWebProducts)	304	0	0
	5.4.2006	15:29:23	W3SVC1	146.102.248.201	GET	/subjects.asp	type=&lang=cz&mainID=1&cat=study&open=251	80	-	146.102.249.128	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+(R1+1.1);+.NET+CLR+1.1.4322)	200	0	0
	5.4.2006	15:29:23	W3SVC1	146.102.248.201	GET	/style.css	-	80	-	146.102.249.128	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+(R1+1.1);+.NET+CLR+1.1.4322)	304	0	0
	5.4.2006	15:29:23	W3SVC1	146.102.248.201	GET	/img/sipka.gif	-	80	-	146.102.249.128	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+(R1+1.1);+.NET+CLR+1.1.4322)	200	0	0
	5.4.2006	15:29:23	W3SVC1	146.102.248.201	GET	/images/fmweb.jpg	-	80	-	146.102.249.128	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+(R1+1.1);+.NET+CLR+1.1.4322)	304	0	0
	5.4.2006	15:29:23	W3SVC1	146.102.248.201	GET	/hodiny.swf	-	80	-	146.102.249.128	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+(R1+1.1);+.NET+CLR+1.1.4322)	304	0	0
	5.4.2006	15:29:23	W3SVC1	146.102.248.201	GET	/images/footer.jpg	-	80	-	146.102.249.128	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+(R1+1.1);+.NET+CLR+1.1.4322)	304	0	0
	5.4.2006	15:29:23	W3SVC1	146.102.248.201	GET	/images/menubg.gif	-	80	-	146.102.249.128	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+(R1+1.1);+.NET+CLR+1.1.4322)	200	0	0
	5.4.2006	15:29:26	W3SVC1	146.102.248.201	GET	/sub.asp	id=&lang=cz&mainid=1&cat=services&open=251	80	-	146.102.249.128	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+(R1+1.1);+.NET+CLR+1.1.4322)	200	0	0
	5.4.2006	15:29:26	W3SVC1	146.102.248.201	GET	/style.css	-	80	-	146.102.249.128	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+(R1+1.1);+.NET+CLR+1.1.4322)	304	0	0
	5.4.2006	15:29:26	W3SVC1	146.102.248.201	GET	/img/sipka.gif	-	80	-	146.102.249.128	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+(R1+1.1);+.NET+CLR+1.1.4322)	304	0	0
	5.4.2006	15:29:26	W3SVC1	146.102.248.201	GET	/images/fmweb.jpg	-	80	-	146.102.249.128	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+(R1+1.1);+.NET+CLR+1.1.4322)	304	0	0
	5.4.2006	15:29:26	W3SVC1	146.102.248.201	GET	/hodiny.swf	-	80	-	146.102.249.128	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+(R1+1.1);+.NET+CLR+1.1.4322)	304	0	0
	5.4.2006	15:29:26	W3SVC1	146.102.248.201	GET	/images/menubg.gif	-	80	-	146.102.249.128	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+(R1+1.1);+.NET+CLR+1.1.4322)	304	0	0
	5.4.2006	15:29:26	W3SVC1	146.102.248.201	GET	/images/footer.jpg	-	80	-	146.102.249.128	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+(R1+1.1);+.NET+CLR+1.1.4322)	304	0	0
	5.4.2006	15:29:26	W3SVC1	146.102.248.201	GET	/english/subjects.asp	Type=Info&ID=SZKVS1&Lang=en	80	-	84.244.122.69	Mozilla/5.0+(Windows;+U;+Windows+NT+5.1;+cs;+rv:1.8.0.1)+Gecko/20060111+Firefox/1.5.0.1	404	0	3

Obr. 14 **Ukázka logovacího souboru**

Předzpracování logu

Pro každý den je serverem vytvořen samostatný logovací soubor. Protože jsem se rozhodl pro analýzu ročních vzorů, datový zdroj obsahuje 365 souborů. To je pro další práci nepraktické, je třeba soubory sloučit do jednoho, nebo alespoň do souborů o takové velikosti, kterou ještě zvládne používaný software. Při své práci jsem nejčastěji používal Microsoft Office Excel 2007, který má technické omezení 1048576 řádků na jeden list. Pro sloučení souborů se osvědčil program „Files to File 2.0“, díky němuž stačí pouze označit soubory, které chceme sloučit a jednoduše je spojit do jednoho.

Výchozí datový zdroj měl velikost 14052385 řádků, které zabíraly 2565976433 bytů. Práce s tak velkým textovým souborem je velice obtížná. Existuje mnoho softwarových nástrojů pro manipulaci s velkými textovými soubory, ale nenabízejí příliš mnoho funkcí. Proto jsem namísto jednoho souboru vytvořil souborů čtrnáct a pracoval s nimi v již výše zmíněném Microsoft Excelu, který umožňuje importovat textový soubor do tabulky. Operace na prvním souboru je pak možné nahrát pomocí makra a na ostatních souborech pouze nahrané makro vykonat.

Dalším krokem v předzpracování bylo odstranění stop ze souboru, které po sobě zanechal Microsoft Internet Information Server. Jak je vidět na předchozím obrázku, začátek souboru obsahuje informace o názvu a verzi software, datum zápisu a hlavičky sloupců logu. Tato informace není jen na začátku souboru, ale několikrát se v souboru opakuje. Odstranění těchto stop je snadné, protože začínají dvojitým křížkem. Stačí tak odstranit všechny řádky mající na začátku #.

Aby bylo možné s logovacím souborem pracovat v softwarových nástrojích určených pro Web Mining, musí mít formát, který jsou jednotlivé programy schopny přečíst. Nejpodporovanějším formátem je Extended Common Log Format (ECLF) definovaný internetovým konsorciem W3C.

ECLF obsahuje následující položky (Rehberger 2002b): remotehost, rfc931, authuser, date, request, status, bytes, referrer, user_agent.

Remotehost je internetová adresa prohlížeče, může mít podobu číselné IP adresy nebo doménového jména. Je důležitá pro identifikaci uživatele, pokud je statická, nebo alespoň

pro identifikaci sezení pokud je dynamická. Nevýhodu dynamických adres eliminuje položka authuser, která obsahuje přihlašovací jméno, pokud se uživatel přihlásil k serveru. Položka rfc931 se v současné době už nepoužívá, dříve sloužila k identifikaci uživatele. Položka date obsahuje datum a čas vzniku požadavku, je důležitou informací pro analýzu průchodu webovou aplikací na serveru. Pod položkou request je zaznamenán uživatelský požadavek na server, URI adresa a popis verze protokolu HTTP, kterou podporuje klient. Položka status obsahuje kód, který server vrací klientovi jako reakci na jeho požadavek. Pro naši potřebu nemá téměř význam. Stejně tak položka bytes, která obsahuje počet bytů vracených serverem. Položka referrer obsahuje URI adresu, ze které byl iniciován požadavek request. User_agent obsahuje informace o klientské aplikaci komunikující se serverem.

Převod původního logu na nový formát jsem provedl načtením dat do tabulky a následným odstraněním a sloučením některých sloupců nebo změnou jejich datového formátu. Data, která původní soubor neobsahoval (rfc931, verze protokolu HTTP, bytes a referrer) jsem nevyplňoval.

Ukázka přeformátovaného logu je na další straně.

Ukázka přeformátovaného logu

remotehost	rfc931	authuser	date	request	status	bytes	referrer	user_agent
146.102.255.68	-	-	[04/Apr/2007:09:06:54 +0000]	"GET /isfmstudent/"	401	0	"-"	"Mozilla/5.0+(Windows;+U;+Windows+NT+5.1;+cs;+rv:1.8.1.3)+Gecko/20070309+Firefox/2.0.0.3"
88.103.178.28	-	-	[04/Apr/2007:09:06:56 +0000]	"GET /koleje/"	404	0	"-"	"Mozilla/5.0+(Windows;+U;+Windows+NT+5.1;+cs;+rv:1.8.1.3)+Gecko/20070309+Firefox/2.0.0.3"
146.102.249.212	-	-	[04/Apr/2007:09:07:00 +0000]	"GET /Default.asp"	200	0	"-"	"Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727;+InfoPath.1;+.NET+CLR+1.1.4322)"
146.102.255.68	-	koman-ma	[04/Apr/2007:09:07:03 +0000]	"GET /isfmstudent/Default.asp"	200	0	"-"	"Mozilla/5.0+(Windows;+U;+Windows+NT+5.1;+cs;+rv:1.8.1.3)+Gecko/20070309+Firefox/2.0.0.3"
146.102.255.68	-	koman-ma	[04/Apr/2007:09:07:05 +0000]	"GET /isfmstudent/termin.asp"	200	0	"-"	"Mozilla/5.0+(Windows;+U;+Windows+NT+5.1;+cs;+rv:1.8.1.3)+Gecko/20070309+Firefox/2.0.0.3"
146.102.254.50	-	-	[04/Apr/2007:09:07:07 +0000]	"GET /Default.asp"	200	0	"-"	"Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+1.1.4322)"
66.249.65.108	-	-	[04/Apr/2007:09:07:07 +0000]	"GET /org_scheme.asp?ID=CVT&lang=cz&mainID=73&cat=study&open=%208"	200	0	"-"	"Mozilla/5.0+(compatible;+Googlebot/2.1;++http://www.google.com/bot.html)"
146.102.255.68	-	koman-ma	[04/Apr/2007:09:07:14 +0000]	"GET /isfmstudent/termin.asp?detail=19758"	200	0	"-"	"Mozilla/5.0+(Windows;+U;+Windows+NT+5.1;+cs;+rv:1.8.1.3)+Gecko/20070309+Firefox/2.0.0.3"
83.208.198.96	-	-	[04/Apr/2007:09:07:16 +0000]	"GET /faq.asp?lang=cz&mainID=55&cat=faq&open=150"	200	0	"-"	"Mozilla/5.0+(Windows;+U;+Windows+NT+5.1;+cs;+rv:1.8.0.11)+Gecko/20070312+Firefox/1.5.0.11"
88.146.207.1	-	-	[04/Apr/2007:09:07:21 +0000]	"GET /Default.asp"	200	0	"-"	"Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727)"
146.102.251.204	-	-	[04/Apr/2007:09:07:27 +0000]	"GET /Default.asp"	200	0	"-"	"Mozilla/4.0+(compatible;+MSIE+7.0;+Windows+NT+5.1;+.NET+CLR+1.1.4322)"
146.102.249.140	-	-	[04/Apr/2007:09:07:38 +0000]	"GET /people.asp?Name=Vop%E1tek&Category=Emploee&MaxRecs=20&Lang=cz"	200	0	"-"	"Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727;+InfoPath.1;+.NET+CLR+1.1.4322)"
146.102.254.50	-	-	[04/Apr/2007:09:07:39 +0000]	"GET /Default.asp"	200	0	"-"	"Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+1.1.4322)"
146.102.249.136	-	-	[04/Apr/2007:09:07:48 +0000]	"GET /Default.asp"	200	0	"-"	"Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727;+InfoPath.1;+.NET+CLR+1.1.4322)"
66.249.65.199	-	-	[04/Apr/2007:09:08:30 +0000]	"GET /konzult.asp?lang=cz&mainID=167&cat=info&open=%208"	200	0	"-"	"Mozilla/5.0+(compatible;+Googlebot/2.1;++http://www.google.com/bot.html)"
146.102.249.183	-	-	[04/Apr/2007:09:08:42 +0000]	"GET /Default.asp"	200	0	"-"	"Mozilla/5.0+(Windows;+U;+Windows+NT+5.1;+cs;+rv:1.8.0.6)+Gecko/20060728+Firefox/1.5.0.6"
146.102.250.143	-	-	[04/Apr/2007:09:08:43 +0000]	"GET /Default.asp"	200	0	"-"	"Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+FunWebProducts;+.NET+CLR+1.1.4322)"
74.6.72.228	-	-	[04/Apr/2007:09:09:00 +0000]	"GET /subjects.asp?Type=Info&ID=HJBIBZ&Lang=cz"	200	0	"-"	"Mozilla/5.0+(compatible;+Yahoo!+Slurp;+http://help.yahoo.com/help/us/ysearch/slurp)"
146.102.255.145	-	-	[04/Apr/2007:09:09:01 +0000]	"GET /Default.asp"	200	0	"-"	"Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1)"
146.102.249.204	-	-	[04/Apr/2007:09:09:02 +0000]	"GET /Default.asp"	200	0	"-"	"Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727;+InfoPath.1;+.NET+CLR+1.1.4322)"
146.102.249.170	-	-	[04/Apr/2007:09:09:09 +0000]	"GET /Default.asp"	200	0	"-"	"Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727;+InfoPath.1;+.NET+CLR+1.1.4322)"
66.249.65.199	-	-	[04/Apr/2007:09:09:12 +0000]	"GET /subjects.asp?type=&lang=cz&mainid=25&cat=others&open=25"	200	0	"-"	"Mozilla/5.0+(compatible;+Googlebot/2.1;++http://www.google.com/bot.html)"
83.208.198.96	-	-	[04/Apr/2007:09:09:17 +0000]	"GET /faq.asp?lang=cz&mainID=55&cat=faq&open=150"	200	0	"-"	"Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1)"
146.102.251.233	-	-	[04/Apr/2007:09:09:18 +0000]	"GET /Default.asp"	200	0	"-"	"Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+1.1.4322)"
193.165.208.98	-	-	[04/Apr/2007:09:09:22 +0000]	"GET /Default.asp"	200	0	"-"	"Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+.NET+CLR+1.1.4322)"
90.176.82.174	-	-	[04/Apr/2007:09:09:22 +0000]	"GET /isfmstudent/"	401	0	"-"	"Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1)"
90.176.82.174	-	pavli-ha	[04/Apr/2007:09:09:24 +0000]	"GET /isfmstudent/Default.asp"	200	0	"-"	"Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1)"
90.176.82.174	-	pavli-ha	[04/Apr/2007:09:09:27 +0000]	"GET /isfmstudent/termin.asp"	200	0	"-"	"Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1)"
194.212.232.6	-	-	[04/Apr/2007:09:09:37 +0000]	"GET /Default.asp"	200	0	"-"	"Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.0;+.NET+CLR+1.1.4322;+.NET+CLR+2.0.50727)"
146.102.249.164	-	-	[04/Apr/2007:09:09:46 +0000]	"GET /Default.asp"	200	0	"-"	"Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727;+InfoPath.1;+.NET+CLR+1.1.4322)"
66.249.65.199	-	-	[04/Apr/2007:09:09:52 +0000]	"GET /subjects.asp?type=&lang=cz&mainid=25&cat=services&open=25"	200	0	"-"	"Mozilla/5.0+(compatible;+Googlebot/2.1;++http://www.google.com/bot.html)"
85.71.167.104	-	-	[04/Apr/2007:09:09:57 +0000]	"GET /Default.asp"	200	0	"-"	"Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1)"

Obr. 15 **Přeformátovaný logovací soubor**

Čištění dat

Předzpracovaný logovací soubor obsahuje velké množství záznamů, které nebyly učiněny na základě vůle uživatelů. Chceme-li sledovat chování uživatelů webových stránek, jsou tyto požadavky pro analýzu irelevantní. Jejich přítomnost v logu by mohla mít dva následky. Protože chceme analyzovat chování a zájmy uživatelů, tak přítomnost požadavků učiněných bez vůle uživatelů by analýzu zkreslila. Čím je jejich podíl větší, tím je zkreslení výraznější. Dalším důvodem pro jejich odstranění je, že jejich přítomnost v souboru zbytečně zvětšuje jeho velikost. S příliš velkými soubory se obtížně pracuje, a zbytečně prodlužují výpočty.

Prvním krokem bylo odstranění všech přímo nevyžádaných souborů. Mezi ně patří všechny řádky, které obsahovaly v položce request přípony: gif, jpg, png, bmp, css, swf, dll, ico, crt. Všechny vyjmenované soubory nejsou přímo žádány uživateli a stahují se spolu s žádanou webovou stránkou. Jsou jimi obrázky, styly webových stránek, flash grafika, knihovny, ikony a certifikáty. Obrázky je možné považovat za nevyžádané, protože fakultní web neobsahuje žádné galerie. V opačném případě by bylo nutné vytvořit seznam obrázků, který by je rozdělil na ty, které musí v logovacím souboru zůstat, a na ostatní, které mohou být odstraněny. Odstranění jsem provedl pomocí dotazu, který odstranil všechny řádky, jež obsahovaly ve sloupci request již zmiňované přípony.

Dále jsem odstranil všechny řádky obsahující na začátku záznamu ve sloupci request požadavek OPTIONS. Metoda OPTIONS slouží ke zjištění informací o daném kontextu, klient může zjistit, které dotazy může na daný kontext zaslat (Zapletal 2001). Požadavků OPTIONS je v logu velké množství a nevycházejí přímo z vůle uživatele. Největší význam pro nás mají požadavky GET sloužící k vyzvednutí objektu (jakéhokoliv souboru) ze serveru.

Tím se velikost souboru rapidně snížila. Z původních 14052385 řádků na 6024920 a z původních 2,5 GB na pouhých 1,1 GB. To podstatně usnadní další práci se souborem.

Pro zjednodušení jsem, po konzultaci s Ing. Jelínkem, odstranil z logu všechny požadavky učiněné ze školních učeben. Důvodem byl, velký provoz na učebnách. Studenti se na počítačích střídají v krátkých intervalech, navíc ve webových prohlížečích jsou stránky fakulty nastaveny jako domovské. To vede k tomu, že na daných IP adresách by nebylo

možné identifikovat sezení. Sezení začíná prvním požadavkem a končí posledním, po kterém nebyl učiněn další během následujících třiceti minut. Časté střídání studentů na počítačích by v některých případech mohlo vést k délce sezení rovnající se době otevření učebny.

Učebny mají přiděleny IP adresy podle vzoru 146.102.249.xxx. V logu tak zůstanou IP adresy vyučujících, studentů na kolejích a všechny ostatní externí adresy. To povede nejen k lepší identifikaci sezení, ale i ke snazší identifikaci uživatelů, protože u ostatních IP adres je velmi pravděpodobné, že se za každou adresou skrývá konkrétní uživatel. Ovšem až na případy, kdy jsou adresy přidělovány dynamicky.

Z logu jsem odstranil všechny řádky obsahující ve sloupci remotehost IP adresy odpovídající výše zmíněnému vzoru. Velikost souboru se zmenšila na 5368221. Redukce nebyla příliš výrazná, v logovacím souboru zůstalo dostatečné množství požadavků, analýza tak nebude výrazně ovlivněna.

Dalším velmi důležitým krokem bylo odstranění stop ze souboru, které po sobě zanechali weboví roboti. Roboti se každý den připojují k serveru a procházejí webovými stránkami kvůli jejich indexaci. Množství požadavků každý den od nich pocházejících je překvapivě vysoké. Protože jejich požadavky nikterak nesouvisí s vůlí uživatelů nebo jejich chováním, jsou pro analýzu irelevantní. Díky jejich vysokému podílu na celkových požadavcích by také mohly vést k zavádějícím výsledkům při analýze. Pokud se struktura odkazů na webu nemění, tak roboti prochází webem každý den po stejných cestách. Velké množství takovýchto opakování by vedlo k vytvoření silných vzorů chování.

Pro jejich odstranění vycházím z faktu, že každý „slušný“ robot si před vstupem na stránky vyžádá soubor /robots.txt, ze kterého se dozví, zda má či nemá zakázanou indexaci. Fakultní web sice nemá soubor robots.txt zřízen, ale požadavky robotů jsou v logovacím souboru zaznamenány. Mají status 404, jenž je chybovou hláškou.

Odstranění sezení robotů je tak možné pomocí odstranění všech požadavků vycházejících z IP adres, ze kterých byl také požadován soubor robots.txt. Pravděpodobnost, že by zmíněný soubor požadoval běžný uživatel, je zanedbatelná. Nicméně ze školních IP adres byl požadavek na soubor několikrát vznesen, dané adresy jsem ovšem v logovacím souboru zanechal.

Čištění proběhlo následujícím způsobem. Logovací soubor jsem načetl do databáze Microsoft Office Access. A na vytvořené tabulce provedl dva SQL dotazy. Logovací soubor je načten v Tabulce1, Pole1 koresponduje s položkou remotehost a Pole5 s položkou request. Dotazem:

```
SELECT DISTINCT Tabulka1.Pole1 FROM Tabulka1
WHERE Tabulka1.Pole5 LIKE "?GET /robots.txt?";
```

jsem zjistil všechny IP adresy související s požadavkem na soubor robots.txt. Výsledek dotazu jsem uložil do Tabulky2 pod Pole1. Otazníky v dotazu jsou nutné, protože položka request je v uvozovkách, otazníky tak nahrazují uvozovky, které není možno do dotazu napsat.
Druhým dotazem:

```
DELETE Tabulka1.* FROM Tabulka1
WHERE Tabulka1.Pole1 IN (SELECT Tabulka2.Pole1 FROM Tabulka2);
```

jsem z Tabulky1 odstranil všechny řádky s IP adresami obsaženými v Tabulce2. Výslednou Tabulku1 jsem následně vyexportoval z databáze do textového souboru. Velikost souboru se rapidně snížila na 4169596 řádků. To bylo velice překvapivé, téměř každého pátého řádku v logu nebyl původcem člověk, ale robot. Odstranění stop po robotech z logu je tak velmi významným krokem.

Identifikace sezení

Na vyčištěném logovacím souboru je nutné provést identifikaci sezení. V určitém okamžiku může být na server připojeno více uživatelů. Jejich požadavky nejsou v logu seskupeny, jsou smíseny dohromady tak, jak chronologicky přicházejí. Identifikace sezení přiřadí k sobě požadavky, které spolu souvisejí. Protože se uživatelé v průběhu času na web opět vracejí, uskutečňují další návštěvy, musí být seskupené požadavky rozděleny na části odpovídající jednotlivým návštěvám.

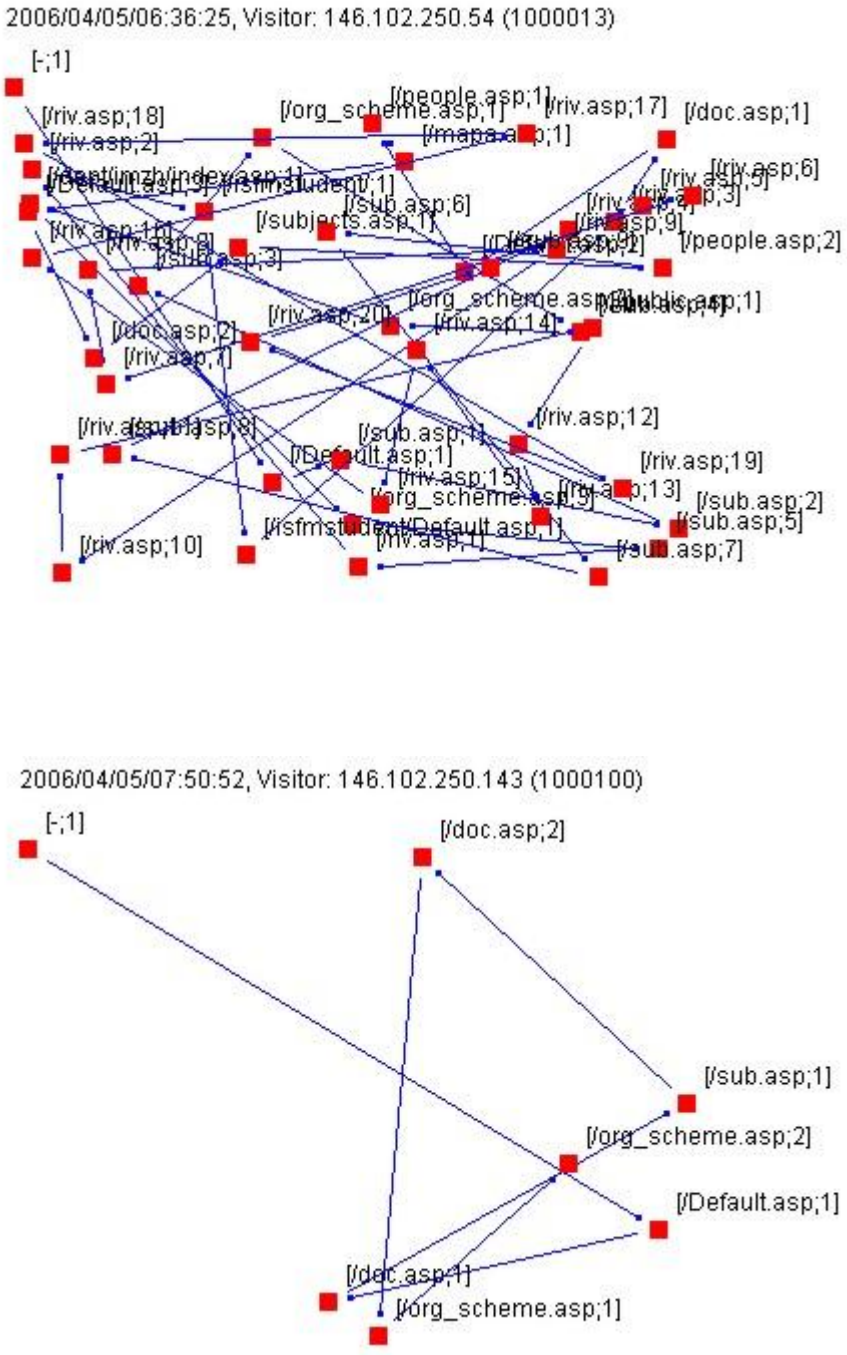
Identifikace sezení sleduje položku remotehost, pomocí níž zjišťuje jednotlivé uživatele. Konkrétním uživatelům přiřazuje požadavky z položky request, které časově řadí pomocí položky date. Pomocí limitu 30 minut jsou požadavky uživatelů rozděleny do jednotlivých sezení. Pokud je mezi po sobě jdoucími požadavky větší časová prodleva než 30 minut, každý patří do jiného sezení. Jeden je konečným požadavkem jednoho sezení a druhý je počátečním požadavkem sezení následujícího.

K identifikaci sezení je nejjednodušší využít nějakého softwarového nástroje. Rozhodl jsem se pro nástroj WUM: A Web Utilization Miner. V oblasti Web Usage Mining je velice propagován. Téměř každý, kdo mluví o nějakém vhodném softwaru, jej neopomene zmínit. Vkládal jsem proto do něj největší naděje.

Pomocí Web Utilization Miner vytvoříme dolovací databázi, do které se načte logovací soubor, načítání je pomalé, téměř jeden den. Poté se log rozdělí na jednotlivá sezení, čekací doba přes jeden den. V tuto chvíli si můžeme pouze zobrazit jednotlivá sezení, kterých jsou ale statisíce. Takový výsledek je pro další práci ještě nepoužitelný. Další volby jako souhrnné zprávy a dotazování jsou dostupné až po vytvoření agregovaného logu. Při agregaci logu program po jednom až dvou dnech dosáhne svých možností a přestane reagovat.

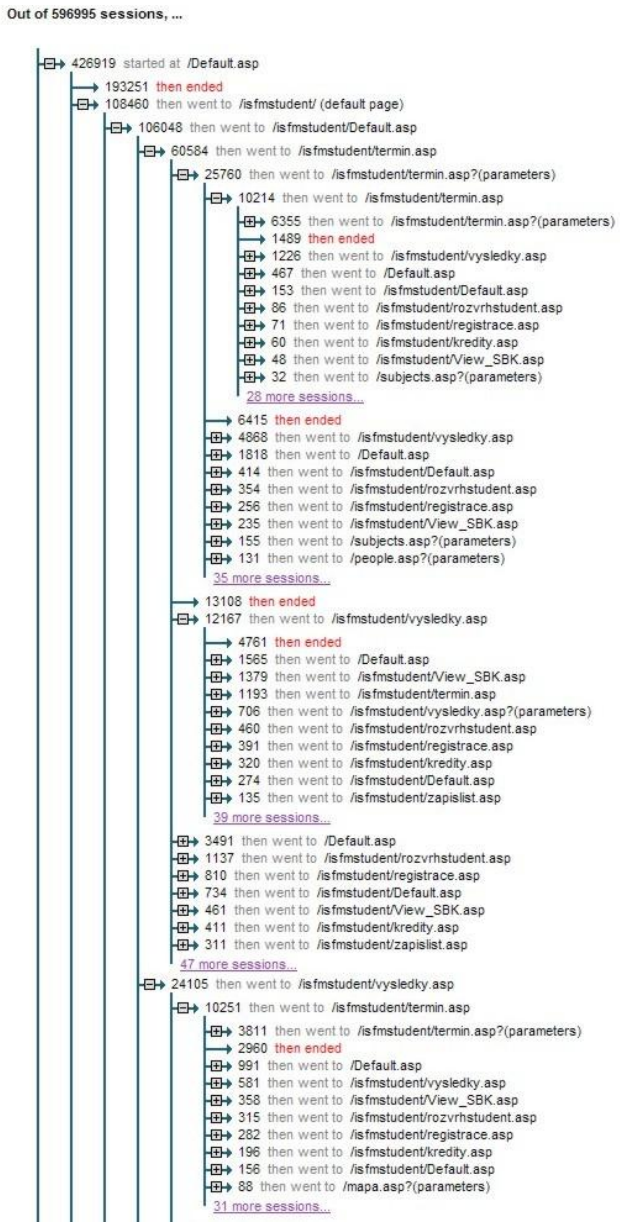
Vzhledem k platformě Java, na které je program vystavěn, je možné jeho použití pouze pro značně zmenšený objem dat. Pro obecné dotazy nad větším počtem údajů (více než dva dny) webového serveru je program (zřejmě špatnou prací z pamětí) nepoužitelný (Křipáč, Novák & Vildová).

Pro ukázku přikládám obrázky dvou sezení. Grafy jsou bohužel vykreslovány velmi chaoticky. Je na nich zajímavé, že vzdálenost bodů (webových stránek) souvisí s délkou page view. Čím více času strávil uživatel na dané stránce, tím je delší spojnice z ní vycházející. To způsobuje chaotické uspořádání. Pro mou další práci jsou bez agregace nepotřebné, protože podobných grafů je přes půl milionu.

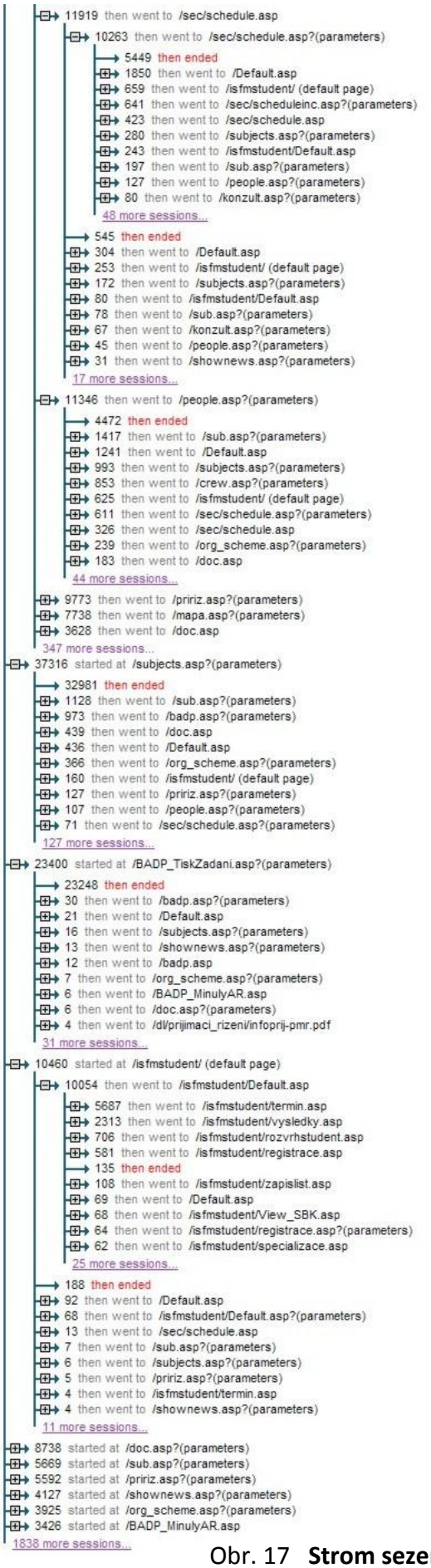


Obr. 16 **Ukázka sezení identifikovaných programem WUM**

Po vyzkoušení několika dalších programů jsem se rozhodl pro Sawmill 7 Enterprise. Program má uživatelské rozhraní řešené pomocí webového prohlížeče. Práce s programem je jednoduchá a velice rychlá. Po krátkém načtení dat do databáze je možné identifikovat sezení. Identifikace sezení trvá pouhých několik minut. Jednotlivá sezení jsou sumarizována a zobrazena pod názvem Session paths. Výsledkem je strom, ve kterém každá cesta od kořene na konec větve odpovídá jedné možné cestě po webu. U každého uzlu je číselná hodnota odpovídající počtu takovýchto cest obsažených v logovacím souboru. Větve je možné podle potřeby dále rozvíjet nebo zavírat. Následující obrázek obsahuje ukázku stromu, kde jsou rozbaleny větve s cestami opakujícími se více než deset tisíckrát.




5629 then ended
2118 then went to /Default.asp
1671 then went to /isfmstudent/View_SBK.asp
1546 then went to /isfmstudent/vysledky.asp?(parameters)
561 then went to /isfmstudent/registrace.asp
466 then went to /isfmstudent/Default.asp
446 then went to /isfmstudent/rozvrhstudent.asp
426 then went to /isfmstudent/kredity.asp
151 then went to /isfmstudent/zapislist.asp
34 more sessions...
8608 then went to /isfmstudent/registrace.asp
5252 then went to /isfmstudent/rozvrhstudent.asp
1448 then went to /isfmstudent/zapislist.asp
775 then went to /isfmstudent/View_SBK.asp
752 then ended
740 then went to /isfmstudent/specializace.asp
592 then went to /isfmstudent/registrace.asp?(parameters)
491 then went to /Default.asp
42 more sessions...
731 then ended
634 then went to /isfmstudent/Default.asp?(parameters)
333 then went to /Default.asp
308 then went to /subjects.asp?(parameters)
107 then went to /sec/schedule.asp
73 then went to /sub.asp?(parameters)
67 then went to /shownews.asp?(parameters)
67 then went to /people.asp?(parameters)
37 then went to /mapa.asp?(parameters)
17 more sessions...
30768 then went to /subjects.asp?(parameters)
12117 then went to /sub.asp?(parameters)
4365 then ended
1842 then went to /pririz.asp?(parameters)
1719 then went to /subjects.asp?(parameters)
1063 then went to /Default.asp
811 then went to /isfmstudent/ (default page)
480 then went to /org_scheme.asp?(parameters)
379 then went to /people.asp?(parameters)
328 then went to /doc.asp
158 then went to /dokumenty/n�st�nka+odd�len�/studijn�+odd�len�/Okruhy_ot�zek_z_teorie_managementu_k_p�ij�mac�m_zkou�k�m.doc
152 then went to /badp.asp?(parameters)
69 more sessions...
4320 then ended
4128 then went to /pririz.asp?(parameters)
3278 then went to /isfmstudent/ (default page)
2106 then went to /badp.asp?(parameters)
1487 then went to /Default.asp
977 then went to /sec/schedule.asp?(parameters)
472 then went to /doc.asp
413 then went to /people.asp?(parameters)
272 then went to /dokumenty/n�st�nka+odd�len�/studijn�+odd�len�/2006+harmonogram+2006+07.doc
55 more sessions...
25496 then went to /sub.asp?(parameters)
9680 then ended
4294 then went to /subjects.asp?(parameters)
3044 then went to /Default.asp
1509 then went to /org_scheme.asp?(parameters)
1401 then went to /isfmstudent/ (default page)
1384 then went to /people.asp?(parameters)
1233 then went to /pririz.asp?(parameters)
407 then went to /doc.asp
224 then went to /sec/schedule.asp
219 then went to /badp.asp?(parameters)
94 more sessions...
17082 then went to /shownews.asp?(parameters)
5645 then ended
3308 then went to /isfmstudent/ (default page)
2256 then went to /Default.asp
1424 then went to /sub.asp?(parameters)
1069 then went to /subjects.asp?(parameters)
799 then went to /pririz.asp?(parameters)
728 then went to /sec/schedule.asp
448 then went to /doc.asp
293 then went to /people.asp?(parameters)
167 then went to /mapa.asp?(parameters)
65 more sessions...



Obr. 17 **Strom sezení identifikovaných programem Sawmill**

Transformace dat

Identifikací sezení bylo objeveno 596995 návštěv. Výsledný strom slučující stejná sezení je vhodným zdrojem informací pro další práci (identifikaci navigačních vzorů). Data je ale nutné transformovat do vhodné podoby.

Protože je strom velice rozsáhlý, nerozbaloval jsem ho celý. Pro zjednodušení jsem stanovil hranici tisíc opakování. Otevíral jsem všechny větve postupně, dokud počet jejich výskytů v sezeních neklesl pod hranici tisíc opakování. Redukce nebude mít na pozdější analýzu vliv, protože budu hledat nejčastější navigační vzory.

Strom jsem přepsal do následující tabulky. Tabulka je řazena abecedně. Pokud se několikrát po sobě opakoval požadavek na stejnou stránku, sloučil jsem opakující se požadavky v jeden. V případech, kde se na konci sezení namísto požadavku objevuje malé x, jednotlivá sezení nekončí, ale rozpadají se na více sezení, jejichž počet výskytů je menší než hranice tisíc opakování. Součet počtů výskytů je tak roven počtu objevených sezení.

ID	Počet výskytů	Cesta						
1	3250	/badp.asp						
2	2927	/BADP_MinulyAR.asp						
3	23248	/BADP_TiskZadani.asp						
4	1618	/Default.asp	/doc.asp	/sub.asp	x			
5	2010	/Default.asp	/doc.asp	x				
6	1014	/Default.asp	/isfmstudent/	/isfmstudent/Default.asp	/isfmstudent/registrace.asp	/isfmstudent/Registrace_Vypis.asp	x	
7	2235	/Default.asp	/isfmstudent/	/isfmstudent/Default.asp	/isfmstudent/registrace.asp	/isfmstudent/rozvrhstudent.asp	x	
8	4025	/Default.asp	/isfmstudent/	/isfmstudent/Default.asp	/isfmstudent/registrace.asp	x		
9	1334	/Default.asp	/isfmstudent/	/isfmstudent/Default.asp	/isfmstudent/registrace.asp			
10	1104	/Default.asp	/isfmstudent/	/isfmstudent/Default.asp	/isfmstudent/rozvrhstudent.asp	/isfmstudent/registrace.asp	x	
11	2940	/Default.asp	/isfmstudent/	/isfmstudent/Default.asp	/isfmstudent/rozvrhstudent.asp	x		
12	1208	/Default.asp	/isfmstudent/	/isfmstudent/Default.asp	/isfmstudent/rozvrhstudent.asp			
13	3499	/Default.asp	/isfmstudent/	/isfmstudent/Default.asp	/isfmstudent/termin.asp	/Default.asp	x	
14	3007	/Default.asp	/isfmstudent/	/isfmstudent/Default.asp	/isfmstudent/termin.asp	/Default.asp		
15	1515	/Default.asp	/isfmstudent/	/isfmstudent/Default.asp	/isfmstudent/termin.asp	/isfmstudent/Default.asp	x	
16	1174	/Default.asp	/isfmstudent/	/isfmstudent/Default.asp	/isfmstudent/termin.asp	/isfmstudent/registrace.asp	x	
17	1672	/Default.asp	/isfmstudent/	/isfmstudent/Default.asp	/isfmstudent/termin.asp	/isfmstudent/rozvrhstudent.asp	x	
18	2352	/Default.asp	/isfmstudent/	/isfmstudent/Default.asp	/isfmstudent/termin.asp	/isfmstudent/vysledky.asp	/Default.asp	x
19	1786	/Default.asp	/isfmstudent/	/isfmstudent/Default.asp	/isfmstudent/termin.asp	/isfmstudent/vysledky.asp	/isfmstudent/termin.asp	x
20	2123	/Default.asp	/isfmstudent/	/isfmstudent/Default.asp	/isfmstudent/termin.asp	/isfmstudent/vysledky.asp	/isfmstudent/View_SBK.asp	x
21	6525	/Default.asp	/isfmstudent/	/isfmstudent/Default.asp	/isfmstudent/termin.asp	/isfmstudent/vysledky.asp	x	
22	7222	/Default.asp	/isfmstudent/	/isfmstudent/Default.asp	/isfmstudent/termin.asp	/isfmstudent/vysledky.asp		
23	6668	/Default.asp	/isfmstudent/	/isfmstudent/Default.asp	/isfmstudent/termin.asp	x		
24	23041	/Default.asp	/isfmstudent/	/isfmstudent/Default.asp	/isfmstudent/termin.asp			
25	2118	/Default.asp	/isfmstudent/	/isfmstudent/Default.asp	/isfmstudent/vysledky.asp	/Default.asp	x	
26	1463	/Default.asp	/isfmstudent/	/isfmstudent/Default.asp	/isfmstudent/vysledky.asp	/isfmstudent/termin.asp	/Default.asp	x
27	4519	/Default.asp	/isfmstudent/	/isfmstudent/Default.asp	/isfmstudent/vysledky.asp	/isfmstudent/termin.asp	x	
28	4269	/Default.asp	/isfmstudent/	/isfmstudent/Default.asp	/isfmstudent/vysledky.asp	/isfmstudent/termin.asp		
29	1671	/Default.asp	/isfmstudent/	/isfmstudent/Default.asp	/isfmstudent/vysledky.asp	/isfmstudent/View_SBK.asp	x	
30	4038	/Default.asp	/isfmstudent/	/isfmstudent/Default.asp	/isfmstudent/vysledky.asp	x		
31	6027	/Default.asp	/isfmstudent/	/isfmstudent/Default.asp	/isfmstudent/vysledky.asp			
32	1448	/Default.asp	/isfmstudent/	/isfmstudent/Default.asp	/isfmstudent/zapislist.asp	x		
33	6051	/Default.asp	/isfmstudent/	/isfmstudent/Default.asp	x			
34	2412	/Default.asp	/isfmstudent/	x				
35	1190	/Default.asp	/mapa.asp	/sub.asp	x			
36	4133	/Default.asp	/mapa.asp	x				
37	2415	/Default.asp	/mapa.asp					

38	1241	/Default.asp	/people.asp	/Default.asp	X			
39	1417	/Default.asp	/people.asp	/sub.asp	X			
40	4216	/Default.asp	/people.asp	X				
41	4472	/Default.asp	/people.asp					
42	1048	/Default.asp	/pririz.asp	/dl/prijimaci_rizeni/infoprij-pmr.pdf	X			
43	1001	/Default.asp	/pririz.asp	/dl/prijimaci_rizeni/zadani.pdf	X			
44	3779	/Default.asp	/pririz.asp	X				
45	3945	/Default.asp	/pririz.asp					
46	1850	/Default.asp	/sec/schedule.asp	/Default.asp	X			
47	4075	/Default.asp	/sec/schedule.asp	X				
48	5994	/Default.asp	/sec/schedule.asp					
49	2256	/Default.asp	/shownews.asp	/Default.asp	X			
50	1306	/Default.asp	/shownews.asp	/isfmstudent/	/isfmstudent/Default.asp	/isfmstudent/termin.asp		
51	1880	/Default.asp	/shownews.asp	/isfmstudent/	/isfmstudent/Default.asp			
52	1424	/Default.asp	/shownews.asp	/sub.asp	X			
53	1069	/Default.asp	/shownews.asp	/subjects.asp	X			
54	3502	/Default.asp	/shownews.asp	X				
55	5645	/Default.asp	/shownews.asp					
56	1812	/Default.asp	/sub.asp	/Default.asp	X			
57	1232	/Default.asp	/sub.asp	/Default.asp				
58	1272	/Default.asp	/sub.asp	/isfmstudent/	/isfmstudent/Default.asp	X		
59	1509	/Default.asp	/sub.asp	/org_scheme.asp	X			
60	1384	/Default.asp	/sub.asp	/people.asp	X			
61	1233	/Default.asp	/sub.asp	/pririz.asp	X			
62	1805	/Default.asp	/sub.asp	/subjects.asp	/sub.asp	X		
63	2489	/Default.asp	/sub.asp	/subjects.asp	X			
64	3080	/Default.asp	/sub.asp	X				
65	9680	/Default.asp	/sub.asp					
66	2106	/Default.asp	/subjects.asp	/badp.asp	X			
67	1487	/Default.asp	/subjects.asp	/Default.asp	X			
68	1905	/Default.asp	/subjects.asp	/isfmstudent/	/isfmstudent/Default.asp	/isfmstudent/termin.asp	X	
69	1098	/Default.asp	/subjects.asp	/isfmstudent/	/isfmstudent/Default.asp	X		
70	1476	/Default.asp	/subjects.asp	/pririz.asp				
71	2652	/Default.asp	/subjects.asp	/pririz.asp				
72	1063	/Default.asp	/subjects.asp	/sub.asp	/Default.asp	X		
73	1842	/Default.asp	/subjects.asp	/sub.asp	/pririz.asp	X		
74	1719	/Default.asp	/subjects.asp	/sub.asp	/subjects.asp	X		
75	3128	/Default.asp	/subjects.asp	/sub.asp	X			
76	4365	/Default.asp	/subjects.asp	/sub.asp				
77	3332	/Default.asp	/subjects.asp	X				
78	4595	/Default.asp	/subjects.asp					
79	7458	/Default.asp	X					
80	193251	/Default.asp						
81	8461	/doc.asp						
82	1730	/english/						
83	1249	/english/subject.asp						
84	1173	/faq.asp	X					
85	1808	/isfmstudent/	/isfmstudent/Default.asp	X				
86	1700	/isfmstudent/	/isfmstudent/Default.asp	/isfmstudent/termin.asp	/isfmstudent/vysledky.asp	X		
87	1805	/isfmstudent/	/isfmstudent/Default.asp	/isfmstudent/termin.asp	X			
88	2182	/isfmstudent/	/isfmstudent/Default.asp	/isfmstudent/termin.asp				
89	2313	/isfmstudent/	/isfmstudent/Default.asp	/isfmstudent/vysledky.asp	X			
90	2054	/isfmstudent/	/isfmstudent/Default.asp	X				
91	1348	/isfmstudent/Default.asp	/isfmstudent/termin.asp	X				
92	1437	/isfmstudent/Default.asp	X					
93	2688	/isfmstudent/termin.asp	X					
94	1545	/isfmstudent/vysledky.asp	X					
95	1034	/mapa.asp	X					
96	3925	/org_scheme.asp						
97	1123	/people.asp						
98	2965	/pririz.asp	X					
99	2627	/pririz.asp						
100	1783	/public.asp						
101	2425	/sec/schedule.asp	X					
102	3631	/shownews.asp						
103	1846	/sub.asp	X					
104	3823	/sub.asp						
105	1128	/subjects.asp	/sub.asp	X				
106	3207	/subjects.asp	X					
107	32981	/subjects.asp						
108	46660	X						

Obr. 18 **Přetransformovaná sezení**

Identifikace navigačních vzorů

Pro identifikaci navigačních vzorů jsem zvolil metodu pravděpodobnostní hypertextové gramatiky. Nejprve jsem pro zjednodušení převedl požadavky do symbolické podoby.

Požadavek	Symbol
/badp.asp	A1
/BADP_MinulyAR.asp	A2
/BADP_TiskZadani.asp	A3
/Default.asp	A4
/dl/prijimaci_rizeni/infoprij-pmr.pdf	A5
/dl/prijimaci_rizeni/zadani.pdf	A6
/doc.asp	A7
/english/	A8
/english/subject.asp	A9
/faq.asp	A10
/isfmstudent/	A11
/isfmstudent/Default.asp	A12
/isfmstudent/registrace.asp	A13
/isfmstudent/Registrace_Vypis.asp	A14
/isfmstudent/rozvrhstudent.asp	A15
/isfmstudent/termin.asp	A16
/isfmstudent/View_SBK.asp	A17
/isfmstudent/vysledky.asp	A18
/isfmstudent/zapislist.asp	A19
/mapa.asp	A20
/org_scheme.asp	A21
/people.asp	A22
/pririz.asp	A23
/public.asp	A24
/sec/schedule.asp	A25
/shownews.asp	A26
/sub.asp	A27
/subjects.asp	A28
x	Ax

Obr. 19 **Převodní tabulka**

Počet výskytů	Cesta
3250	A1
2927	A2
23248	A3
1618	A4→A7→A27→Ax
2010	A4→A7→Ax
1014	A4→A11→A12→A13→A14→Ax
2235	A4→A11→A12→A13→A15→Ax
4025	A4→A11→A12→A13→Ax
1334	A4→A11→A12→A13
1104	A4→A11→A12→A15→A13→Ax
2940	A4→A11→A12→A15→Ax
1208	A4→A11→A12→A15
3499	A4→A11→A12→A16→A4→Ax
3007	A4→A11→A12→A16→A4
1515	A4→A11→A12→A16→A12→Ax
1174	A4→A11→A12→A16→A13→Ax
1672	A4→A11→A12→A16→A15→Ax
2352	A4→A11→A12→A16→A18→A4→Ax
1786	A4→A11→A12→A16→A18→A16→Ax
2123	A4→A11→A12→A16→A18→A17→Ax
6525	A4→A11→A12→A16→A18→Ax
7222	A4→A11→A12→A16→A18
6668	A4→A11→A12→A16→Ax
23041	A4→A11→A12→A16
2118	A4→A11→A12→A18→A4→Ax
1463	A4→A11→A12→A18→A16→A4→Ax
4519	A4→A11→A12→A18→A16→Ax
4269	A4→A11→A12→A18→A16
1671	A4→A11→A12→A18→A17→Ax
4038	A4→A11→A12→A18→Ax
6027	A4→A11→A12→A18
1448	A4→A11→A12→A19→Ax
6051	A4→A11→A12→Ax
2412	A4→A11→Ax
1190	A4→A20→A27→Ax
4133	A4→A20→Ax
2415	A4→A20
1241	A4→A22→A4→Ax
1417	A4→A22→A27→Ax
4216	A4→A22→Ax
4472	A4→A22
1048	A4→A23→A5→Ax
1001	A4→A23→A6→Ax
3779	A4→A23→Ax
3945	A4→A23
1850	A4→A25→A4→Ax
4075	A4→A25→Ax
5994	A4→A25
2256	A4→A26→A4→Ax
1306	A4→A26→A11→A12→A16
1880	A4→A26→A11→A12
1424	A4→A26→A27→Ax
1069	A4→A26→A28→Ax
3502	A4→A26→Ax

Počet výskytů	Cesta
5645	A4→A26
1812	A4→A27→A4→Ax
1232	A4→A27→A4
1272	A4→A27→A11→A12→Ax
1509	A4→A27→A21→Ax
1384	A4→A27→A22→Ax
1233	A4→A27→A23→Ax
1805	A4→A27→A28→A27→Ax
2489	A4→A27→A28→Ax
3080	A4→A27→Ax
9680	A4→A27
2106	A4→A28→A1→Ax
1487	A4→A28→A4→Ax
1905	A4→A28→A11→A12→A16→Ax
1098	A4→A28→A11→A12→Ax
1476	A4→A28→A23
2652	A4→A28→A23
1063	A4→A28→A27→A4→Ax
1842	A4→A28→A27→A23→Ax
1719	A4→A28→A27→A28→Ax
3128	A4→A28→A27→Ax
4365	A4→A28→A27
3332	A4→A28→Ax
4595	A4→A28
7458	A4→Ax
193251	A4
8461	A7
1730	A8
1249	A9
1173	A10→Ax
1808	A11→A12→Ax
1700	A11→A12→A16→A18→Ax
1805	A11→A12→A16→Ax
2182	A11→A12→A16
2313	A11→A12→A18→Ax
2054	A11→A12→Ax
1348	A12→A16→Ax
1437	A12→Ax
2688	A16→Ax
1545	A18→Ax
1034	A20→Ax
3925	A21
1123	A22
2965	A23→Ax
2627	A23
1783	A24
2425	A25→Ax
3631	A26
1846	A27→Ax
3823	A27
1128	A28→A27→Ax
3207	A28→Ax
32981	A28
46660	Ax

Obr. 20 **Sezení převedená do symbolické podoby**

Na vytvořené tabulce je možné v Excelu nebo Accessu provádět jednoduché dotazy a zjistit tak všechny potřebné informace. Z množiny uživatelských sezení je třeba zjistit, kolikrát byla každá stránka požadována, kolikrát byla žádána jako první stránka v sezení a kolikrát jako poslední. Ze získaných informací se vypočte pro každou stránku ohodnocení určující pravděpodobnost, že se konkrétní stránka stane počáteční stránkou během sezení. Výpočet pravděpodobnosti probíhá pomocí vzorce:

$$\pi(A_i) = \frac{\propto * a}{b} + \frac{(1-\propto) * c}{d}$$

Kde: a – počet výskytů stránky A_i během všech sezení
b – počet všech požadavků během všech sezení
c – počet výskytů stránky A_i jako stránky počáteční
d – počet všech sezení
α – hodnota symbolu α se pohybuje v rozmezí (0,1), čím je α nižší, tím více zvýhodňuje stránky, které se objevují spíše na začátku než během cesty. Zvolil jsem hodnotu 0,25

Stránka	Počet výskytů	První	Poslední	π(Ai)
A01	5356	3250	3250	0,005
A02	2927	2927	2927	0,004
A03	23248	23248	23248	0,034
A04	450299	426919	197490	0,620
A05	1048	0	0	0,000
A06	1001	0	0	0,000
A07	12089	8461	8461	0,013
A08	1730	1730	1730	0,002
A09	1249	1249	1249	0,002
A10	1173	1173	0	0,002
A11	127783	11862	0	0,039
A12	129671	2785	1880	0,028
A13	10886	0	1334	0,002
A14	1014	0	0	0,000
A15	9159	0	1208	0,002
A16	85555	2688	30798	0,019
A17	3794	0	0	0,001
A18	49671	1545	13249	0,011
A19	1448	0	0	0,000
A20	8772	1034	2415	0,003
A21	5434	3925	3925	0,006
A22	13853	1123	5595	0,004
A23	22568	5592	10700	0,011
A24	1783	1783	1783	0,003
A25	14344	2425	5994	0,006
A26	20713	3631	9276	0,008
A27	51864	5669	17868	0,017
A28	75166	37316	37576	0,061
Ax	215039	46660	215039	0,098
Suma	1348637	596995	596995	1,000

Obr. 21 **Tabulka pomocných výpočtů**

Dále je nutné zjistit pravděpodobnosti cest ze stránky A_i na stránku A_j. Pravděpodobnost se počítá na základě vzorce:

$$P(A_i \rightarrow A_j) = \frac{a}{b}$$

Kde: a – počet cest vedených ze stránky A_i na stránku A_j
b – počet všech cest vedených ze stránky A_i

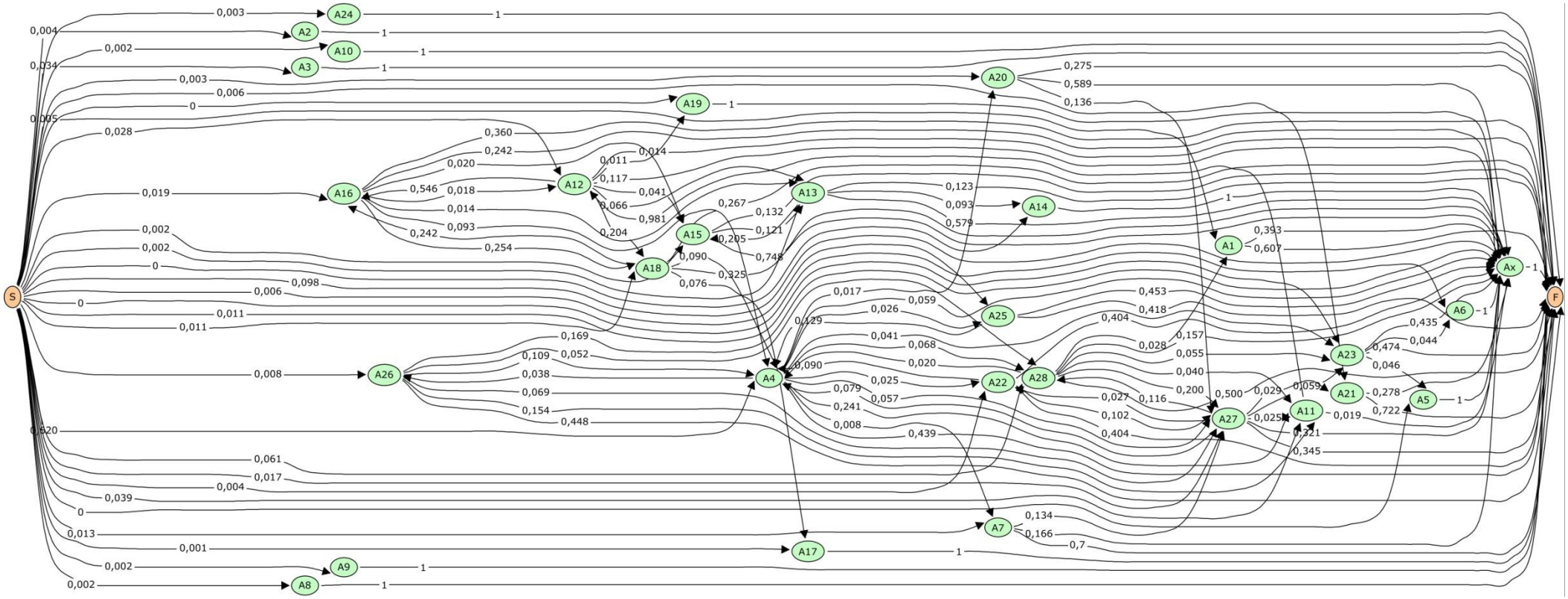
Cesta	Počet	P(Ai→Aj)
A01→Ax	2106	0,393
A01→F	3250	0,607
A02→F	2927	1,000
A03→F	23248	1,000
A04→A07	3628	0,008
A04→A11	108460	0,241
A04→A20	7738	0,017
A04→A22	11346	0,025
A04→A23	9773	0,022
A04→A25	11919	0,026
A04→A26	17082	0,038
A04→A27	25496	0,057
A04→A28	30768	0,068
A04→Ax	26599	0,059
A04→F	197490	0,439
A05→Ax	1048	1,000
A06→Ax	1001	1,000
A07→A27	1618	0,134
A07→Ax	2010	0,166
A07→F	8461	0,700
A08→F	1730	1,000
A09→F	1249	1,000
A10→F	1173	1,000
A11→A12	125371	0,981
A11→Ax	2412	0,019
A12→A13	8608	0,066
A12→A15	5252	0,041
A12→A16	70830	0,546
A12→A18	26418	0,204
A12→A19	1448	0,011
A12→Ax	15235	0,117
A12→F	1880	0,014
A13→A14	1014	0,093
A13→A15	2235	0,205
A13→Ax	6303	0,579
A13→F	1334	0,123
A14→Ax	1014	1,000
A15→A13	1104	0,121
A15→Ax	6847	0,748
A15→F	1208	0,132
A16→A04	7969	0,093
A16→A12	1515	0,018
A16→A13	1174	0,014
A16→A15	1672	0,020
A16→A18	21708	0,254
A16→Ax	20719	0,242
A16→F	30798	0,360

Cesta	Počet	P(Ai→Aj)
A17→Ax	3794	1,000
A18→A04	4470	0,090
A18→A16	12037	0,242
A18→A17	3794	0,076
A18→Ax	16121	0,325
A18→F	13249	0,267
A19→Ax	1448	1,000
A20→A27	1190	0,136
A20→Ax	5167	0,589
A20→F	2415	0,275
A21→Ax	1509	0,278
A21→F	3925	0,722
A22→A04	1241	0,090
A22→A27	1417	0,102
A22→Ax	5600	0,404
A22→F	5595	0,404
A23→F	10700	0,474
A23→A05	1048	0,046
A23→A06	1001	0,044
A23→Ax	9819	0,435
A24→F	1783	1,000
A25→A04	1850	0,129
A25→Ax	6500	0,453
A25→F	5994	0,418
A26→A04	2256	0,109
A26→A11	3186	0,154
A26→A27	1424	0,069
A26→A28	1069	0,052
A26→Ax	3502	0,169
A26→F	9276	0,448
A27→A04	4107	0,079
A27→A11	1272	0,025
A27→A21	1509	0,029
A27→A22	1384	0,027
A27→A23	3075	0,059
A27→A28	6013	0,116
A27→Ax	16636	0,321
A27→F	17868	0,345
A28→A01	2106	0,028
A28→A04	1487	0,020
A28→A11	3003	0,040
A28→A23	4128	0,055
A28→A27	15050	0,200
A28→Ax	11816	0,157
A28→F	37576	0,500
Ax→F	215039	1,000
Suma	1348637	29,000

Obr. 22 **Pravděpodobnosti cest ze stránky A_i na stránku A_j**

Nyní máme dostatek informací k tomu, aby mohl být vytvořen pravděpodobnostní model navigace po stránkách fakulty. Je to orientovaný graf, který má počátek v bodě S. Bod S je propojen na všechny stránky objevující se během sezení. Každá hrana vedoucí z bodu S na konkrétní stránku má přidělenu hodnotu $\pi(A_i)$, která vyjadřuje pravděpodobnost, že se dotyčná stránka stane stránkou výchozí během sezení. Na základě předchozí tabulky jsou vytvořeny hrany mezi všemi stránkami A_1 až A_n a jsou jim přiřazeny příslušné pravděpodobnosti. Ze stránek, na kterých byla některá sezení ukončena, vedou hrany do bodu F. Graf jsem vytvořil za pomoci programu CmapTools.

Pravděpodobnostní model pohybu po stránkách www.fm.vse.cz



Obr. 23 **Hypertextová pravděpodobnostní gramatika vytvořená pro cesty po www.fm.vse.cz**

Výsledný graf pouze popisuje navigaci po stránkách. K získání navigačních vzorů je nutné graf prořezat. K prořezání slouží hodnoty θ a λ. Hodnota θ vyjadřuje limit podpory. Pouze stránky A_i, jejichž pravděpodobnost $\pi(A_i)$, překročí tuto hodnotu, se mohou stát výchozími stránkami navigačních vzorů. Hodnota λ vyjadřuje limit důvěryhodnosti. Do navigačních vzorů budou zařazeny pouze řetězce, jejichž celková pravděpodobnost překročí hodnotu λ. Hodnoty θ a λ dávají analytikovi možnost kontroly nad kvalitou a kvantitou objevených navigačních vzorů. Pravděpodobnost delšího řetězce je součinem pravděpodobností $P(A_i \rightarrow A_j)$. Například: $P(A_i \rightarrow A_j \rightarrow A_k) = P(A_i \rightarrow A_j) * P(A_j \rightarrow A_k)$.

Jak je zřejmé z tabulky s pravděpodobnostmi $\pi(A_i)$, nejpravděpodobnější stránkou, kterou začínají sezení je domovská stránka A_{04} (/Default.asp). Pravděpodobnost, že touto stránkou začne sezení je 62 %, o zbylých 38 % se téměř rovnoměrně dělí ostatní stránky. Aby stránka A_{04} nebyla jedinou stránkou v objevených vzorech, kterou začíná sezení, byl jsem nucen stanovit hodnotu θ poměrně nízkou. Při zvolené hodnotě podpory na úrovni 0,05 se do sledovaných výsledků dostala ještě stránka A_{28}. Stránku A_x nemá smyls zvažovat, protože je to fiktivní pomocná stránka. Hodnotu λ jsem přizpůsoboval délce navigační cesty, čím delší cesta tím nižší hodnota důvěryhodnosti.

Objevené navigační vzory pro délku cesty o rozsahu dvou webových stránek:

θ=0,05 λ=0,05	
Cesta	**Důvěra**
A04→A11	0,241
A28→A27	0,200
A28→Ax	0,157
A04→A28	0,068
A04→Ax	0,059
A04→A27	0,057
A28→A23	0,055

Obr. 24

Všech během roku využívaných cest ze stránky na stránku je 93. Jak je zřejmé z tabulky, pouze tři cesty přesáhly hranici pravděpodobnosti 15 %. Pravděpodobnost ostatních cest je rapidně nižší. Z toho třetí cesta nevychází na konkrétní stránku, ale na A_x zastupující jakoukoliv jinou stránku než je stránka A_1, A_{04}, A_{11}, A_{23}, nebo A_{27}. Nejčastějším navigačním vzorem je tedy přechod z domovské stránky do ISFM (Informační systém Fakulty managementu).

Tento vzor není překvapením, bylo možné jej očekávat. Zajímavějším je druhý navigační vzor. Jde o přechod ze stránky A_{28} na stránku A_{27}. Přechod ze stránky Studium na stránku O Fakultě managementu, kde stránka Studium je stránkou výchozí. Proč tomu tak je, se můžeme pouze domnívat, protože v logovacím souboru chybí již zmíněná položka referrer. Položka referrer by na tuto otázku mohla do určité míry poskytnout odpověď. Na stránku

Studium mohou odkazovat webové vyhledávače, nebo odkaz na ni mohou mít uživatelé uložen ve webovém prohlížeči, možnost ručního vypisování URL se mi zdá jako málo pravděpodobná.

Objevené navigační vzory pro délku cesty o rozsahu tří webových stránek:

θ=0,05 λ=0,01	
Cesta	**Důvěra**
A04→A11→A12	0,236
A28→A27→Ax	0,064
A28→A11→A12	0,039
A28→A23→Ax	0,024
A28→A27→A28	0,023
A04→A27→Ax	0,018
A28→A27→A04	0,016
A04→A28→A27	0,014
A04→A25→Ax	0,012
A28→A27→A23	0,012
A28→A01→Ax	0,011
A04→A28→Ax	0,011
A04→A22→Ax	0,010
A04→A20→Ax	0,010

Obr. 25

Protože pravděpodobnost delší cesty je logicky nižší, snížil jsem hodnotu důvěryhodnosti na 0,01. Z výsledků v tabulce můžeme rozdělit chování uživatelů na dvě skupiny. První vzor chování vyjadřuje zájem o ISFM. Z domovské stránky vstoupí do ISFM a zajímají se o úvodní stránku Informačního systému. Jejich navigační chování je předpověditelné s pravděpodobností 23,6 %. Druhým vzorem chování je zájem o něco jiného než ISFM. Ale takových to vzorů je velké množství a mají malou pravděpodobnost. Vzory s pravděpodobností okolo jednoho procenta a nižší jsou vhodné spíše pro zajímavost, než pro vyvozování závěrů. Nízké pravděpodobnosti cest svědčí o jakémsi chaotickém pohybu po webu.

Zjištěné výsledky mohou souviset s tím, že struktura odkazů je plochá. Úvodní stránka fakulty obsahuje velké množství linků, pomocí nichž se přímo dostaneme tam, kam potřebujeme. To přináší uživatelům komfort, že nemusí po webu zbytečně cestovat a hledat co potřebují. Rychlost navigace je ale vyvážena množstvím odkazů, které mohou úvodní stránku tvořit méně přehlednou. To by mohl být problém u rozsáhlejších webů.

Pomocí procenta důvěryhodnosti můžeme usuzovat nejen nad navigací po stránkách, ale i nad zájmy uživatelů nebo jejich vztahu k fakultě. 23,6 % návštěvníků, kteří shlédli alespoň tři stránky, jsou studenty fakulty a zajímají se o ISFM. Dalším zajímavým vzorem je, že 1,2 % návštěv s délkou cesty alespoň tři stránky, uskutečnili pravděpodobně potenciální zájemci o studium. Protože výchozí stránkou nebyla Domovská stránka, ale Stránka studium, na kterou se dostali pravděpodobně z webového vyhledávače. A poté se zajímali o stránku O Fakultě managementu a dále pokračovali na stránku Přijímací řízení.

Objevené navigační vzory pro délku cesty o rozsahu čtyř webových stránek:

θ=0,05 λ=0,01	
Cesta	**Důvěra**
A04→A11→A12→A16	0,129
A04→A11→A12→A18	0,048
A04→A11→A12→Ax	0,028
A28→A11→A12→A16	0,021
A04→A11→A12→A13	0,016
A04→A11→A12→A15	0,010

Obr. 26

V navigaci o délce čtyř stránek již jednoznačně dominuje ISFM. Přestože se délka cesty prodlužuje a její pravděpodobnost klesá, na prvním místě se umístila cesta s pořád ještě silnou pravděpodobností 12,9 %. V tomto případě se studenti po vstupu do Informačního systému zajímají o stránku /isfmstudent/termin.asp, na které jsou vypsány termíny na zkoušky.

Objevené navigační vzory pro délku cesty o rozsahu pěti webových stránek:

θ=0,05 λ=0,01	
Cesta	**Důvěra**
A04→A11→A12→A16→A18	0,033
A04→A11→A12→A16→Ax	0,031
A04→A11→A12→A18→Ax	0,016
A04→A11→A12→A16→A04	0,012
A04→A11→A12→A18→A16	0,012

Obr. 27

V navigaci o délce pěti stránek je již pokles pravděpodobnosti cesty patrný. Druhou nejčastěji navštěvovanou stránkou po stránce s termíny je stránka $A_{18,}$ /isfmstudent/vysledky.asp, která dává studentům přehled o jejich studijních výsledcích. Téměř stejnou pravděpodobnost má také vstup ze stránky termínů na jakoukoliv jinou stránku než je $A_{04,}$ $A_{12,}$ $A_{13,}$ $A_{15,}$ nebo A_{18}.

Objevené navigační vzory pro délku cesty o rozsahu šesti webových stránek:

θ=0,05 λ=0,008	
Cesta	**Důvěra**
A04→A11→A12→A16→A18→Ax	0,011
A04→A11→A12→A16→A18→A16	0,008

Obr. 28

Navigaci o délce šesti cest uvádím spíše jen pro zajímavost, jejich pravděpodobnost je už velice malá. Jak je vidět z tabulky je to nejdelší cesta, kterou je ještě možné ve výchozích datech vystopovat. V prvním případě se cesta rozpadá na velký počet cest s pravděpodobností tak malou, že takovéto cesty nebyly do dat zahrnuty. V druhém případě dochází k zacyklení, kdy se uživatelé vracejí ze stránky s výsledky na stránku s termíny.

Statistická analýza

Statistickou analýzu je nejlepší provést pomocí některého softwarového nástroje, kterých je na trhu dostatek. Pro analýzu jsem použil program Sawmill 7 Enterprise. Do programu jsem nahrál vyčištěný logovací soubor. Čištění logu není v podstatě nutné, ale použití logu, ze kterého jsou odstraněny irelevantní záznamy, zrychlí práci programu.

Přehled

Během sledovaného období jeden rok, bylo uživateli (vyjma webových robotů, které jsem ze souboru odstranil) vytvořeno 4169585 záznamů v logu. 82492 návštěvníků za rok shlédlo na 4161976 stránek.

Statistika pro **05/Dub/2006 - 04/Dub/2007**, 365 den

	Všechny dny	Průměrně za den
Záznamů v logu	4 169 585	11 423,52
Počet stránek	4 161 976	11 402,67
Návštěvníků	82 492	-
Přenesených dat	0 b	0 b

Obr. 29 **Přehled**

Datum a čas

První oblastí, kterou program nabízí sledovat, jsou informace týkající se data a času. Statistiky jsou tvořeny podle roků, dní, dnů v týdnu a hodin. Statistika návštěvnosti v závislosti na čase je nejvděčnějším a nejsledovanějším údajem, který můžeme z logovacích souborů zjistit.

Roky

Rok 2006 je ve sledovaném období zastoupen 271 dny a rok 2007 94 dny. Během roku 2006 bylo denně v logu provedeno 8483 záznamů. V roce 2007 jich bylo již 19900. Důvodem by mohlo být, že zájem o stránky fakulty má rostoucí tendenci. Další příčina by mohla být, že na počátku roku je o stránky zvýšený zájem než po zbytek roku. K určení pravého důvodu by ale byla třeba data za více let.

Statistika pro **05/Dub/2006 - 04/Dub/2007**, 365 den

Záznamů v logu

Roky/měsíce/dny

Řádek 1 - 2 z 2

	▲ Datum/čas	Záznamů v logu	Počet stránek	Návštěvníků	Přenesených dat
1	2006	2 298 916	2 293 054	57 573	0 b
2	2007	1 870 669	1 868 922	36 065	0 b
	Celkem	4 169 585	4 161 976	-	0 b

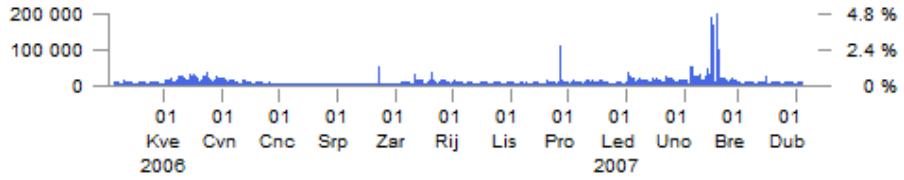
Obr. 30 **Roky**

Rok podle dní

Největší počet záznamů v logu byl vytvořen během několika únorových dní. Největší počet záznamů (195077) byl vytvořen 18. února 2007. Největší počet návštěvníků (1917) byl zaznamenán 5. února 2007. Vysvětlení těchto extrémů je jednoduché. V únoru se konají registrace do následujícího semestru. První kolo registrací 5.-9. února, druhé kolo registrací 14.-.23. února. V prvním kole registrací byl zaznamenán rekordní počet návštěv, všichni studenti se musí registrace zúčastnit. V druhém kole byl zaznamenán rekordní počet záznamů v logu (požadavků). Druhé kolo registrace se už netýká všech studentů, ale ti co se ho účastní si „ladí“ svůj rozvrh, generují velký počet požadavků na server.

Statistika pro **05/Dub/2006 - 04/Dub/2007**, 365 den

Záznamů v logu



Dní

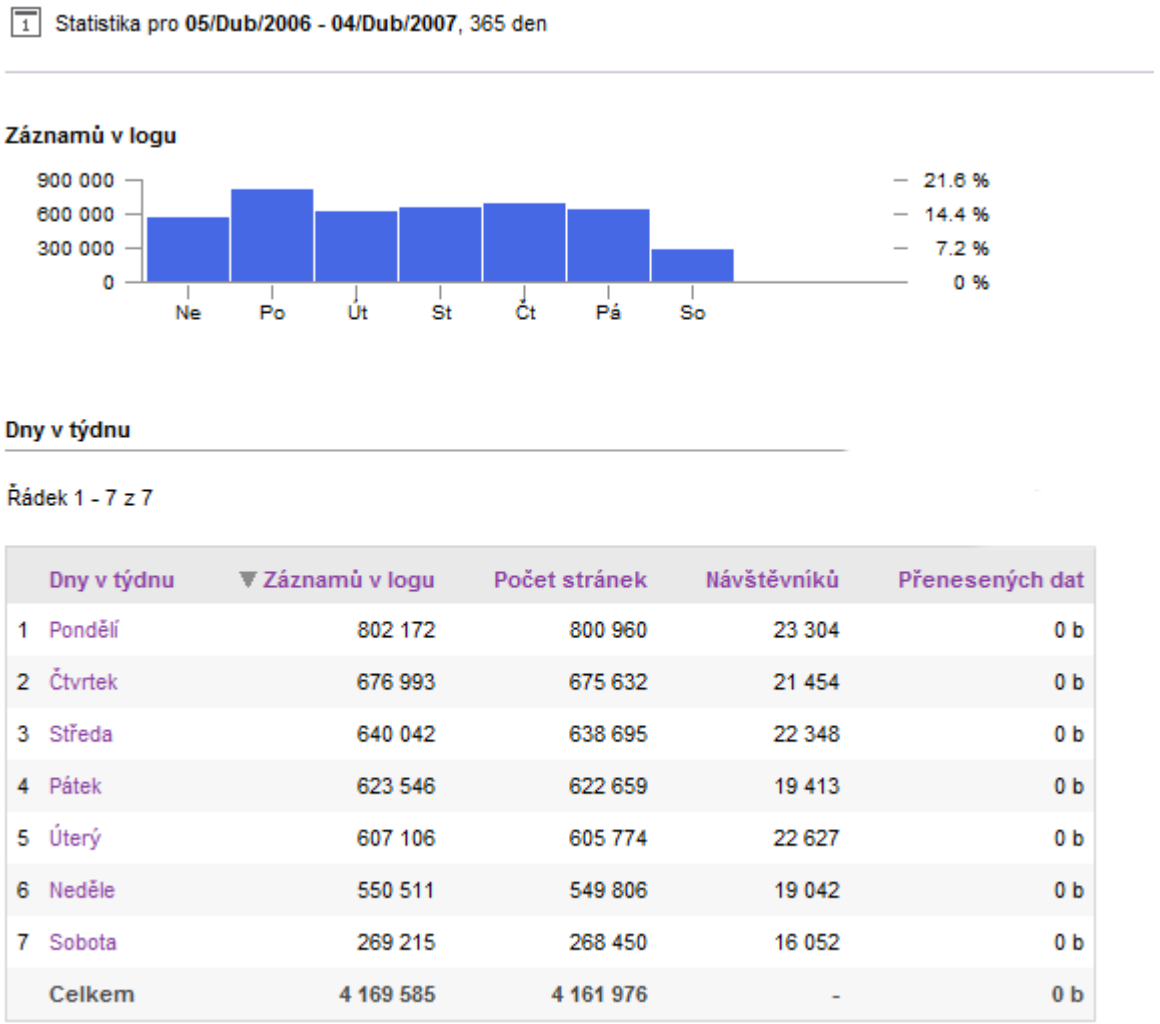
Řádek 1 - 10 z 365 11-20 > >>>

	Datum/čas	▼ Záznamů v logu	Počet stránek	Návštěvníků	Přenesených dat
1	18/Uno/2007	195 077	195 050	1 416	0 b
2	15/Uno/2007	184 468	184 415	1 593	0 b
3	16/Uno/2007	163 259	163 252	1 263	0 b
4	27/Lis/2006	106 089	106 068	1 395	0 b
5	19/Uno/2007	99 912	99 872	1 865	0 b
6	05/Uno/2007	52 342	52 282	1 917	0 b
7	23/Srp/2006	50 864	50 856	721	0 b
8	04/Uno/2007	50 344	50 340	1 363	0 b
9	13/Uno/2007	44 465	44 445	1 649	0 b
10	02/Led/2007	34 166	34 158	1 701	0 b
	355 dalších položek	3 188 599	3 181 238	-	0 b
	Celkem	4 169 585	4 161 976	-	0 b

Obr. 31 **Rok podle dní**

Dny v týdnu

Největší zájem o stránky fakulty je v pondělí a ve čtvrtek. Pokud budeme sledovat počet návštěvníků v jednotlivých dnech, dospějeme k zajímavému úkazu. Největší počet návštěv je učiněn v pondělí a každý následující den je vždy počet návštěv nižší než v předcházející den. Počet návštěvníků klesá s tím, jak týden zraje ke svému konci, s nejmenším počtem návštěv v sobotu.



Statistika pro **05/Dub/2006 - 04/Dub/2007**, 365 den

Záznamů v logu

Dny v týdnu

Řádek 1 - 7 z 7

	Dny v týdnu	▼ Záznamů v logu	Počet stránek	Návštěvníků	Přenesených dat
1	Pondělí	802 172	800 960	23 304	0 b
2	Čtvrtek	676 993	675 632	21 454	0 b
3	Středa	640 042	638 695	22 348	0 b
4	Pátek	623 546	622 659	19 413	0 b
5	Úterý	607 106	605 774	22 627	0 b
6	Neděle	550 511	549 806	19 042	0 b
7	Sobota	269 215	268 450	16 052	0 b
	Celkem	4 169 585	4 161 976	-	0 b

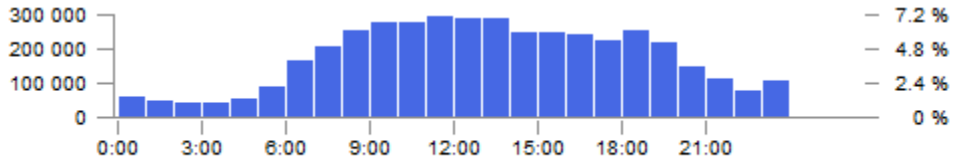
Obr. 32 **Dny v týdnu**

Hodiny

Největší provoz na serveru je mezi šestou a dvacátou hodinou, s maximem kolem poledne. Pokud by z logu nebyly odstraněny požadavky webových robotů, množství záznamů v nočních hodinách by bylo mnohem vyšší.

Statistika pro **05/Dub/2006 - 04/Dub/2007**, 365 den

Záznamů v logu



Hodiny

Řádek 1 - 24 z 24

	Hodiny	▼ Záznamů v logu	Počet stránek	Návštěvníků	Přenesených dat
1	11:00 - 12:00	292 420	291 969	12 548	0 b
2	12:00 - 13:00	284 907	284 190	12 861	0 b
3	13:00 - 14:00	283 319	282 713	12 616	0 b
4	10:00 - 11:00	274 607	274 037	12 168	0 b
5	9:00 - 10:00	273 120	272 538	12 225	0 b
6	8:00 - 9:00	252 700	252 287	11 089	0 b
7	18:00 - 19:00	249 954	249 483	12 936	0 b
8	14:00 - 15:00	244 589	243 963	12 401	0 b
9	15:00 - 16:00	242 618	242 235	12 396	0 b
10	16:00 - 17:00	239 426	239 050	12 815	0 b
11	17:00 - 18:00	218 506	218 007	13 135	0 b
12	19:00 - 20:00	211 884	211 627	12 001	0 b
13	7:00 - 8:00	203 738	203 323	9 032	0 b
14	6:00 - 7:00	161 981	161 687	6 748	0 b
15	20:00 - 21:00	142 945	142 683	10 332	0 b
16	21:00 - 22:00	110 173	109 971	8 339	0 b
17	23:00 - 00:00	100 413	100 377	4 112	0 b
18	5:00 - 6:00	85 387	85 296	4 480	0 b
19	22:00 - 23:00	75 865	75 621	5 944	0 b
20	0:00 - 1:00	58 128	58 092	2 932	0 b
21	4:00 - 5:00	48 506	48 474	2 988	0 b
22	1:00 - 2:00	42 145	42 114	2 451	0 b
23	3:00 - 4:00	36 615	36 609	2 414	0 b
24	2:00 - 3:00	35 639	35 630	2 266	0 b
	Celkem	4 169 585	4 161 976	-	0 b

Obr. 33 **Hodiny**

Obsah

Druhou oblastí, kterou program nabízí k analýze, je obsah. Sledují se zde stránky a typy souborů.

Stránky

Z webového serveru bylo během sledovaného období staženo neuvěřitelných 3901 různých stránek a souborů. Největší podíl (21,1 %) mezi požadavky zaujímá domovská stránka webu (/Default.asp). Mezi další nejžádanější stránky patřila sekce Studium (/subjects.asp), sekce O Fakultě managementu (/sub.asp) a Informační systém Fakulty managementu (/isfmstudent/*), kde byl největší zájem o stránku s termíny (/isfmstudent/termin.asp).

Stránky

Řádek 1 - 20 z 3 901 21-40 > >>>

	Stránka	▼ Záznamů v logu		0 - 100 %	Počet stránek	Návštěvníků	Přenesených dat
1	/Default.asp	879 806	21.1 %		879 806	52 472	0 b
2	/subjects.asp?(parameters)	502 885	12.1 %		502 885	36 072	0 b
3	/sub.asp?(parameters)	266 293	6.4 %		266 293	25 649	0 b
4	/isfmstudent/termin.asp	212 125	5.1 %		212 125	12 196	0 b
5	/isfmstudent/Default.asp	208 744	5.0 %		208 744	14 794	0 b
6	/org_scheme.asp?(parameters)	207 340	5.0 %		207 340	5 602	0 b
7	/sec/schedule.asp?(parameters)	195 005	4.7 %		195 005	6 054	0 b
8	/isfmstudent/ (úvodní stránka)	193 047	4.6 %		193 047	15 834	0 b
9	/isfmstudent/termin.asp?(parameters)	164 189	3.9 %		164 189	6 907	0 b
10	/isfmstudent/vysledky.asp	112 818	2.7 %		112 818	8 815	0 b
11	/pririz.asp?(parameters)	98 561	2.4 %		98 561	15 162	0 b
12	/isfmstudent/registrace.asp?(parameters)	92 446	2.2 %		92 446	2 510	0 b
13	/people.asp?(parameters)	83 953	2.0 %		83 953	9 091	0 b
14	/isfmstudent/registrace.asp	82 344	2.0 %		82 344	4 130	0 b
15	/sec/schedule.asp	66 128	1.6 %		66 128	5 108	0 b
16	/isfmstudent/Default.asp?(parameters)	57 301	1.4 %		57 301	947	0 b
17	/isfmstudent/rozvrhstudent.asp	43 915	1.1 %		43 915	4 927	0 b
18	/BADP_TiskZadani.asp?(parameters)	39 593	0.9 %		39 593	3 326	0 b
19	/shownews.asp?(parameters)	39 329	0.9 %		39 329	7 938	0 b
20	/doc.asp?(parameters)	37 570	0.9 %		37 570	4 517	0 b
	3881 dalších položek	586 138	14.1 %		578 533	-	0 b
	Celkem	4 169 530	100 %		4 161 925	-	0 b

Obr. 34 **Stránky**

Typy souborů

Během sledovaného období bylo na serveru k dispozici 58 různých souborů. V drtivé většině to byly webové stránky, ale i dokumenty, video, zvukové soubory, programy a jiné soubory.

Typy souborů

Řádek 1 - 20 z 58 21-40 > >>>

	Typy souborů	▼ Záznamů v logu		0 - 100 %	Počet stránek	Návštěvníků	Přenesených dat
1	ASP	3 801 981	96.3 %		3 801 981	76 591	0 b
2	PDF	67 859	1.7 %		67 859	8 794	0 b
3	HTML	25 901	0.7 %		25 901	4 350	0 b
4	DOC	22 075	0.6 %		22 075	7 236	0 b
5	HTM	8 519	0.2 %		8 519	2 295	0 b
6	JS	7 609	0.2 %		0	1 475	0 b
7	PPT	5 089	0.1 %		5 089	2 112	0 b
8	PHP	4 898	0.1 %		4 898	151	0 b
9	TXT	4 149	0.1 %		4 149	33	0 b
10	XML	375	0.0 %		375	111	0 b
11	ZIP	242	0.0 %		242	31	0 b
12	CER	230	0.0 %		230	55	0 b
13	AVI	222	0.0 %		222	10	0 b
14	MSO	144	0.0 %		144	37	0 b
15	CZ	77	0.0 %		77	44	0 b
16	XLS	58	0.0 %		58	25	0 b
17	CRL	57	0.0 %		57	8	0 b
18	MP3	49	0.0 %		49	4	0 b
19	PL	43	0.0 %		43	7	0 b
20	EXE	32	0.0 %		32	2	0 b
	38 dalších položek	220	0.0 %		220	-	0 b
	Celkem	3 949 829	100 %		3 942 220	-	0 b

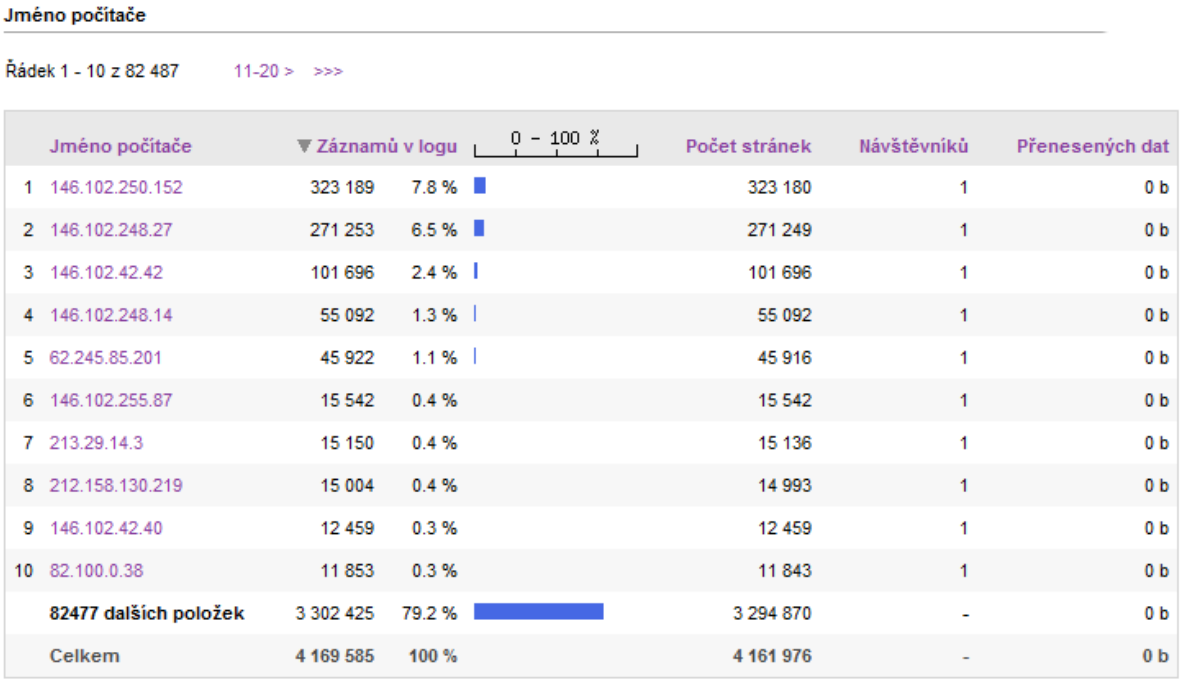
Obr. 35 **Typy souborů**

Demografie návštěvníků

Další oblastí, kterou program nabízí je demografie návštěvníků. Sledují se zde jména počítačů, geografická poloha a ověření uživatelé.

Jméno počítače

Za sledované období se k webovému serveru přihlásili uživatelé prostřednictvím 82487 IP adres. Ke každé IP adrese je možné zjisti počet vytvořených záznamů v logu, počet zobrazených stránek a počet návštěvníků.



Jméno počítače

Řádek 1 - 10 z 82 487 11-20 > >>>

	Jméno počítače	▼ Záznamů v logu		0 - 100 %	Počet stránek	Návštěvníků	Přenesených dat
1	146.102.250.152	323 189	7.8 %		323 180	1	0 b
2	146.102.248.27	271 253	6.5 %		271 249	1	0 b
3	146.102.42.42	101 696	2.4 %		101 696	1	0 b
4	146.102.248.14	55 092	1.3 %		55 092	1	0 b
5	62.245.85.201	45 922	1.1 %		45 916	1	0 b
6	146.102.255.87	15 542	0.4 %		15 542	1	0 b
7	213.29.14.3	15 150	0.4 %		15 136	1	0 b
8	212.158.130.219	15 004	0.4 %		14 993	1	0 b
9	146.102.42.40	12 459	0.3 %		12 459	1	0 b
10	82.100.0.38	11 853	0.3 %		11 843	1	0 b
	82477 dalších položek	3 302 425	79.2 %		3 294 870	-	0 b
	Celkem	4 169 585	100 %		4 161 976	-	0 b

Obr. 36 **Jména počítačů**

Geografická poloha

Stránky Fakulty managementu jsou navštěvovány z celého světa. Téměř 93 % záznamů pochází z České republiky. Zbytek záznamů je směřován z ostatních zemí, celkem ze 124. Mezi nejčastější patří USA, Slovensko, Německo, Rakousko, Velká Británie, Švédsko, Kanada, Belgie, Irsko atd. Z většiny zemí jako je Nepál, Myanmar, Tanzanie, Etiopie… byl učiněn pouze jeden požadavek. Mohlo se jednat například o mylný požadavek, ale všude mohou být lidé se zájmem o naši fakultu.

Geografická poloha

Řádek 1 - 10 z 124 11-20 > >>>

	Geografická poloha	▼ Záznamů v logu		0 - 100 %	Počet stránek	Návštěvníků	Přenesených dat
1	Czech Republic/	3 871 515	92.9 %		3 864 531	66 002	0 b
2	United States/	132 635	3.2 %		132 461	4 736	0 b
3	Slovakia/	59 896	1.4 %		59 823	5 050	0 b
4	Germany/	26 152	0.6 %		26 094	1 218	0 b
5	Austria/	21 027	0.5 %		20 919	202	0 b
6	United Kingdom/	9 357	0.2 %		9 320	795	0 b
7	Sweden/	6 160	0.1 %		6 154	485	0 b
8	Canada/	5 330	0.1 %		5 324	202	0 b
9	Belgium/	5 228	0.1 %		5 227	109	0 b
10	Ireland/	3 303	0.1 %		3 302	151	0 b
	114 dalších položek	27 351	0.7 %		27 190	-	0 b
	Celkem	4 167 954	100 %		4 160 345	-	0 b

Obr. 37 **Geografická poloha**

Ověření uživatelé

Zde jsou sledována uživatelská jména z položky logu authuser a jsou k nim přiřazeny počty záznamů, počty stránek a návštěvníci (za názvem návštěvníci se skrývá množství IP adres, které ověřený uživatel použil k přístupu na webový server).

Celkových záznamů v logu je 4169585, z toho je pouze 1278366 záznamů, kdy byli uživatelé přihlášeni pod svým uživatelským jménem. Počet uživatelů je zde 2925, ale přibližně polovinu z nich tvoří nesprávně napsaná jména.

Jak je vidět z obrázku, největším rekordmanem za sledované období byl uživatel s názvem hurto-ji. Za rok shlédl prostřednictvím čtyř počítačů 13908 stránek, nemluvě o množství shlédnutých stránek, kdy nebyl pod svým jménem přihlášen. To je přes 38 stránek připadajících na každý den. Na druhou stranu jsou zde i studenti, kteří za celý rok shlédli méně než deset stránek. Průměrný počet stránek připadajících na jednoho ověřeného uživatele je 437.

Ověření uživatelé

Řádek 1 - 10 z 2 925 11-20 > >>>

	Ověření uživatelé	▼ Záznamů v logu		0 - 100 %	Počet stránek	Návštěvníků	Přenesených dat
1	hurto-ji	13 908	1.0 %		13 908	4	0 b
2	hurko-iv	7 397	0.5 %		7 397	37	0 b
3	novak-m5	7 119	0.5 %		7 119	3	0 b
4	divis-dr	6 992	0.5 %		6 992	5	0 b
5	kolar-i1	6 878	0.5 %		6 878	21	0 b
6	michladm	6 799	0.5 %		6 799	5	0 b
7	korin-ma	6 687	0.5 %		6 687	64	0 b
8	pruso-ba	6 095	0.5 %		6 095	5	0 b
9	loudi-to	5 616	0.4 %		5 616	28	0 b
10	vlnas-pa	5 147	0.4 %		5 147	3	0 b
	2915 dalších položek	1 278 366	94.6 %		1 278 366	-	0 b
	Celkem	1 351 004	100 %		1 351 004	-	0 b

Obr. 38 **Ověření uživatelé**

Systémy návštěvníků

V části věnované systémům návštěvníků, jsou sledovány rozlišení obrazovky, webové prohlížeče a operační systémy.

Rozlišení obrazovky

Položka rozlišení obrazovky je prázdná, protože logovací soubor takovéto informace neobsahuje.

Www prohlížeče

K zobrazení stránek bylo použito 93 různých webových prohlížečů. Na třech čtvrtinách záznamů se podílel Internet Explorer, který je zde chybně interpretován a označen jako Netscape Navigator. 17,4 % záznamů zaujímá Mozilla Firefox a 1,9 % Opera, podíl ostatních prohlížečů je mizivý. Na devátém místě s 0,1 % se umístil prohlížeč Safari, který se v současné době začíná celosvětově rozšiřovat.

Www prohlížeče

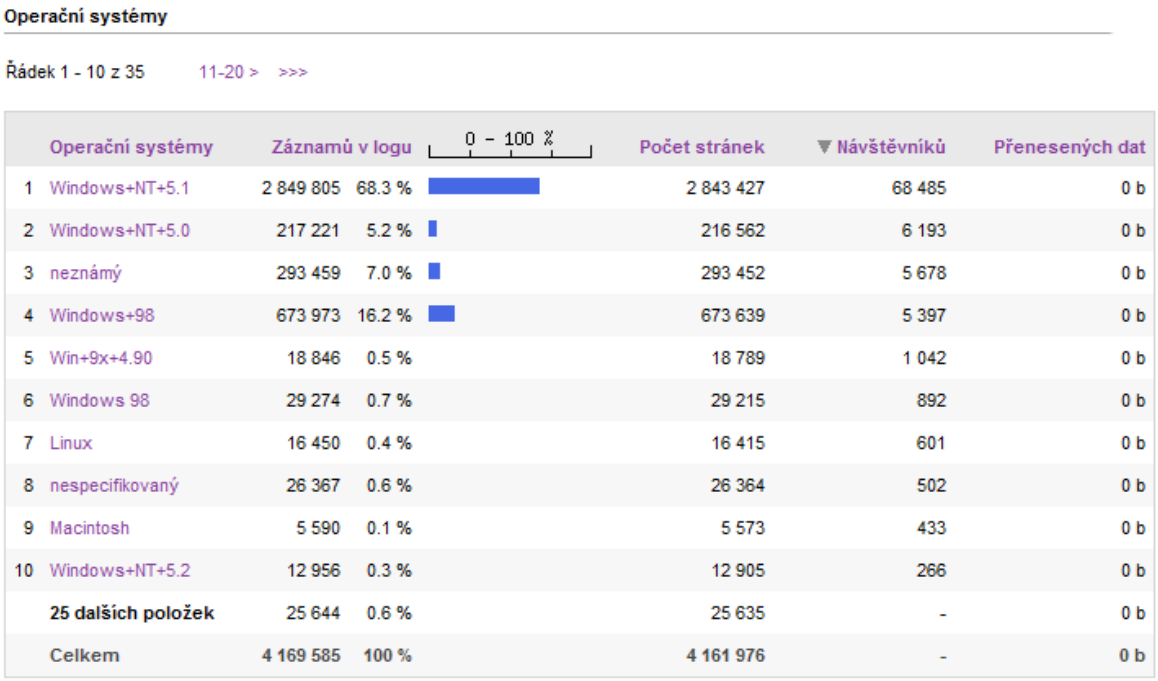
Řádek 1 - 10 z 103 11-20 > >>>

	Www prohlížeče	Záznamů v logu		0 - 100 %	Počet stránek	▼ Návštěvníků	Přenesených dat
1	Netscape Navigator/	3 028 728	74.5 %		3 022 523	61 316	0 b
2	Firefox/	706 537	17.4 %		705 377	19 789	0 b
3	Opera/	77 854	1.9 %		77 726	3 123	0 b
4	Microsoft-WebDAV-MiniRedir/	6 974	0.2 %		6 974	1 395	0 b
5	Mozilla/	35 217	0.9 %		35 124	1 097	0 b
6	MSFrontPage/	1 641	0.0 %		1 641	682	0 b
7	nespecifikovaný/	26 367	0.6 %		26 364	502	0 b
8	neznámý/	2 745	0.1 %		2 745	367	0 b
9	Safari/	3 834	0.1 %		3 823	324	0 b
10	Konqueror/	504	0.0 %		502	69	0 b
	93 dalších položek	175 912	4.3 %		175 907	-	0 b
	Celkem	4 066 313	100 %		4 058 706	-	0 b

Obr. 39 **www prohlížeče**

Operační systémy

Mezi operačními systémy převládá Windows a jeho Windows XP. Překvapivým zjištěním je, že přes 16 % návštěvníků stále pracuje na systému Windows 98. Operační systém Linux využívá pouze 0,4 % návštěvníků.



Operační systémy

Řádek 1 - 10 z 35 11-20 > >>>

	Operační systémy	Záznamů v logu		0 - 100 %	Počet stránek	▼ Návštěvníků	Přenesených dat
1	Windows+NT+5.1	2 849 805	68.3 %		2 843 427	68 485	0 b
2	Windows+NT+5.0	217 221	5.2 %		216 562	6 193	0 b
3	neznámý	293 459	7.0 %		293 452	5 678	0 b
4	Windows+98	673 973	16.2 %		673 639	5 397	0 b
5	Win+9x+4.90	18 846	0.5 %		18 789	1 042	0 b
6	Windows 98	29 274	0.7 %		29 215	892	0 b
7	Linux	16 450	0.4 %		16 415	601	0 b
8	nespecifikovaný	26 367	0.6 %		26 364	502	0 b
9	Macintosh	5 590	0.1 %		5 573	433	0 b
10	Windows+NT+5.2	12 956	0.3 %		12 905	266	0 b
	25 dalších položek	25 644	0.6 %		25 635	-	0 b
	Celkem	4 169 585	100 %		4 161 976	-	0 b

Obr. 40 **Operační systémy**

Referrer

Další část je v programu věnována informacím, které je možné získat při sledování položky referrer. Je zde možné zjistit, ze kterých stránek k nám uživatelé přišli, jaké vyhledávače je k nám přivedli, jaké fráze používali při hledání.

Všechny tyto položky jsou ale bohužel prázdné, protože logovací soubor neobsahuje potřebné informace.

Sezení

Část věnovaná sezením je důležitá nejen pro statistickou analýzu, ale i pro ostatní analýzy. Rozdělení logu na jednotlivá sezení je výchozím krokem všech analýz Web Usage Miningu.

Přehled sezení

V logu bylo identifikováno 596995 sezení. 53930 sezení bylo uskutečněno novými uživateli, 543065 sezení uskutečnili pravidelní návštěvníci, kteří se k serveru připojili prostřednictvím 28544 IP adres. Návštěvníci se připojili nejčastěji dvakrát nebo více než šestkrát. Sezení v průměru trvala 3 minuty a 40 sekund, a na každé sezení připadá 3,6 zobrazených stránek. Sezení celkem trvala 4 roky 60 dnů a 17 hodin. To znamená, že v průměru během sledovaného období byli k serveru v každém okamžiku připojeni 4 uživatelé.

	Všechny dny	Průměrně za den
Celkem přístupů	2 147 619	5 883,89
Celkem session	596 995	1 635,60
Sessions od nový návštěvníků	53 930	-
Session od pravidelných návštěvníků	543 065	-
Celkem uživatelů session	82 475	225,96
Poprvé přišli	53 930	-
Pravidelní návštěvníci	28 544	-
Dvakrát přišli	11 753	-
Třikrát přišli	4 418	-
Čtyřikrát přišli	2 279	-
Pětkrát přišli	1 354	-
Pravidelně chodí (6x +)	8 740	-
Celkové trvání všech session	4r 60d 17:06:24	-
Průměrný počet přístupů během session	3,60	-
Průměr session na uživatele	7,24	-
Medián session na uživatele	1,00	-
Maximum concurrent sessions	182	-
Průměrné trvání session	00:03:40	-

Obr. 41 **Přehled sezení**

Vstupní stránky

Každé sezení bylo zahájeno jednou z 1838 stránek. Většina přístupů logicky připadá na domovskou stránku.

	Stránka	Sessions		0 – 100 %
1	/Default.asp	426 919	71.5 %	
2	/subjects.asp?(parameters)	37 316	6.3 %	
3	/BADP_TiskZadani.asp?(parameters)	23 400	3.9 %	
4	/isfmstudent/ (úvodní stránka)	10 460	1.8 %	
5	/doc.asp?(parameters)	8 738	1.5 %	
6	/sub.asp?(parameters)	5 669	0.9 %	
7	/pririz.asp?(parameters)	5 592	0.9 %	
8	/shownews.asp?(parameters)	4 127	0.7 %	
9	/org_scheme.asp?(parameters)	3 925	0.7 %	
10	/BADP_MinulyAR.asp	3 426	0.6 %	
	1838 dalších položek	67 423	11.3 %	
	Celkem	596 995	100 %	

Obr. 42 **Vstupní stránky**

Výstupní stránky

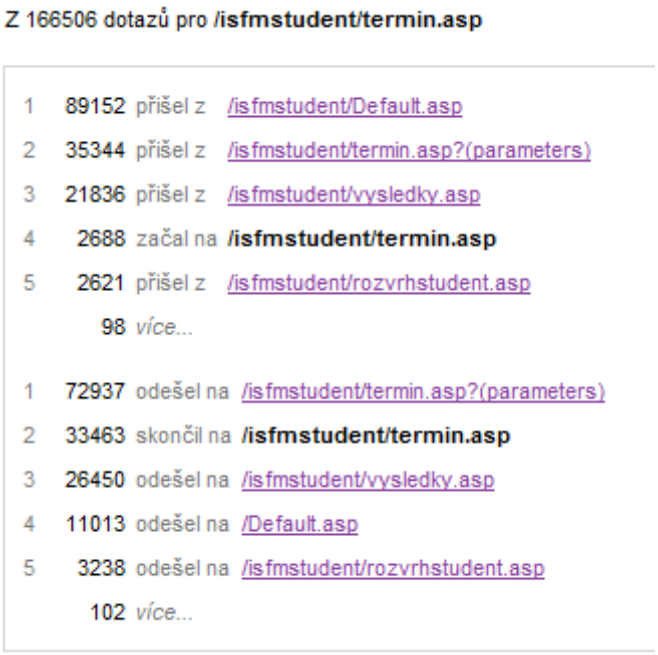
Stránek, kterými byla sezení ukončena, je 1934.

	Stránka	Sessions		0 – 100 %
1	/Default.asp	226 541	37.9 %	
2	/subjects.asp?(parameters)	47 392	7.9 %	
3	/sub.asp?(parameters)	35 983	6.0 %	
4	/isfmstudent/termin.asp	33 463	5.6 %	
5	/isfmstudent/vysledky.asp	25 517	4.3 %	
6	/BADP_TiskZadani.asp?(parameters)	23 986	4.0 %	
7	/sec/schedule.asp?(parameters)	16 244	2.7 %	
8	/isfmstudent/termin.asp?(parameters)	16 059	2.7 %	
9	/pririz.asp?(parameters)	15 678	2.6 %	
10	/shownews.asp?(parameters)	12 830	2.1 %	
	1934 dalších položek	143 302	24.0 %	
	Celkem	596 995	100 %	

Obr. 43 **Výstupní stránky**

Cesty přes stránku

Pro každou stránku je možné zobrazit, odkud na ni uživatelé přišli a kam z ní odešli.



Obr. 44 **Ukázka cesty přes stránku /isfmstudent/termin.asp**

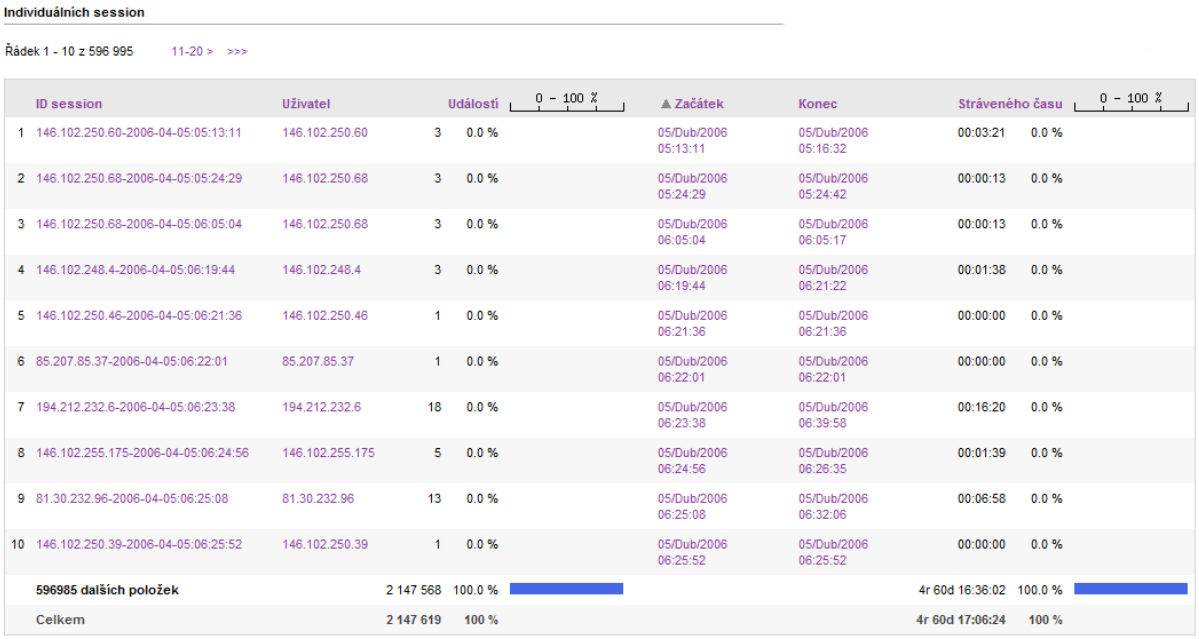
Cesty v sezeních

Zde jsou opakující se sezení agregována do stromové struktury. Podrobný náhled na rozbalený strom je na obrázku 17 v předchozím textu.

Obr. 45 **Cesty v sezení**

Individuální sezení

Stejně jako v programu WUM je zde možné si zobrazit a prohlédnout jednotlivá sezení, kvalita zobrazení v porovnání s obrázkem 16 je ale nesrovnatelně vyšší. U každého sezení můžeme zjistit jméno počítače, počet událostí, čas začátku a konce sezení, strávený čas, nebo zobrazit stromovou strukturu sezení.



Individuálních session

Řádek 1 - 10 z 596 995 11-20 > >>>

	ID session	Uživatel	Událostí	0 - 100 %	▲ Začátek	Konec	Stráveného času	0 - 100 %
1	146.102.250.60-2006-04-05:05:13:11	146.102.250.60	3 0.0 %		05/Dub/2006 05:13:11	05/Dub/2006 05:16:32	00:03:21 0.0 %	
2	146.102.250.68-2006-04-05:05:24:29	146.102.250.68	3 0.0 %		05/Dub/2006 05:24:29	05/Dub/2006 05:24:42	00:00:13 0.0 %	
3	146.102.250.68-2006-04-05:06:05:04	146.102.250.68	3 0.0 %		05/Dub/2006 06:05:04	05/Dub/2006 06:05:17	00:00:13 0.0 %	
4	146.102.248.4-2006-04-05:06:19:44	146.102.248.4	3 0.0 %		05/Dub/2006 06:19:44	05/Dub/2006 06:21:22	00:01:38 0.0 %	
5	146.102.250.46-2006-04-05:06:21:36	146.102.250.46	1 0.0 %		05/Dub/2006 06:21:36	05/Dub/2006 06:21:36	00:00:00 0.0 %	
6	85.207.85.37-2006-04-05:06:22:01	85.207.85.37	1 0.0 %		05/Dub/2006 06:22:01	05/Dub/2006 06:22:01	00:00:00 0.0 %	
7	194.212.232.6-2006-04-05:06:23:38	194.212.232.6	18 0.0 %		05/Dub/2006 06:23:38	05/Dub/2006 06:39:58	00:16:20 0.0 %	
8	146.102.255.175-2006-04-05:06:24:56	146.102.255.175	5 0.0 %		05/Dub/2006 06:24:56	05/Dub/2006 06:26:35	00:01:39 0.0 %	
9	81.30.232.96-2006-04-05:06:25:08	81.30.232.96	13 0.0 %		05/Dub/2006 06:25:08	05/Dub/2006 06:32:06	00:06:58 0.0 %	
10	146.102.250.39-2006-04-05:06:25:52	146.102.250.39	1 0.0 %		05/Dub/2006 06:25:52	05/Dub/2006 06:25:52	00:00:00 0.0 %	
	596985 dalších položek		2 147 568 100.0 %				4r 60d 16:36:02 100.0 %	
	Celkem		2 147 619 100 %				4r 60d 17:06:24 100 %	

Obr. 46 **Individuální sezení**

Ukázka sezení 8

Začátek session: 05/Apr/2006 06:24:56

Jméno počítače: 146.102.255.175

Cekem 1 sezení, ...

1 začátek na /Default.asp
1 pak šel na /doc.asp
1 pak šel na /sub.asp?(parameters)
1 pak šel na /subjects.asp?(parameters)
1 pak šel na /sub.asp?(parameters)
1 pak skončil

Obr. 47 **Ukázka jednoho sezení**

Ostatní

Položka ostatní sleduje údaje, které nejsou v předchozích částech zařazeny.

Viry

Analýza v logu odhalila 51 záznamů nakažených virem Nimda. Červ nazvaný Nimda vznikl v roce 2001, je zaměřený na platformu Win32. Snaží se napadat MS IIS přes několik známých bezpečnostních chyb (Krause 2001). Přestože patche na všechny chyby existují, podařilo se Nimdě napadnout server i v současné době.

	Viry	▼ Záznamů v logu		0 - 100 %	Počet stránek	Návštěvníků	Přenesených dat
1	Nimda	51	100.0 %		51	9	0 b
	Celkem	51	100 %		51	-	0 b

Obr. 48 **Viry**

Roboti

I když jsem ve fázi čištění z logu odstranil 1198625 záznamů, které v něm zanechali weboví roboti, analýza odhalila dalších 87440 záznamů od jedenácti robotů. To je zapříčiněno tím, že níže zobrazení roboti nerespektují soubor robots.txt. Z toho můžeme vyvodit, že pokud nechceme naše stránky z nějakého důvodu nechat indexovat a indexaci zakážeme v souboru robots.txt, většina „slušných“ robotů to bude respektovat, ale existuje i určité procento robotů, kteří naše stránky stejně zaindexují.

	Roboti	▼ Záznamů v logu		0 - 100 %	Počet stránek	Návštěvníků	Přenesených dat
1	Inktomi Slurp	87 440	84.7 %		87 440	3 193	0 b
2	Wget	12 483	12.1 %		12 481	18	0 b
3	Googlebot	2 136	2.1 %		2 136	30	0 b
4	Gigabot	1 042	1.0 %		1 042	19	0 b
5	Internet Explorer Crawler	67	0.1 %		67	31	0 b
6	Snapbot	35	0.0 %		35	24	0 b
7	Collective or e-collector	35	0.0 %		35	5	0 b
8	MSN Robot	19	0.0 %		19	11	0 b
9	larbin	9	0.0 %		9	1	0 b
10	Robozilla	5	0.0 %		5	1	0 b
11	SpiderMan	1	0.0 %		1	1	0 b
	Celkem	103 272	100 %		103 270	-	0 b

Obr. 49 **Roboti**

Odpovědi serveru

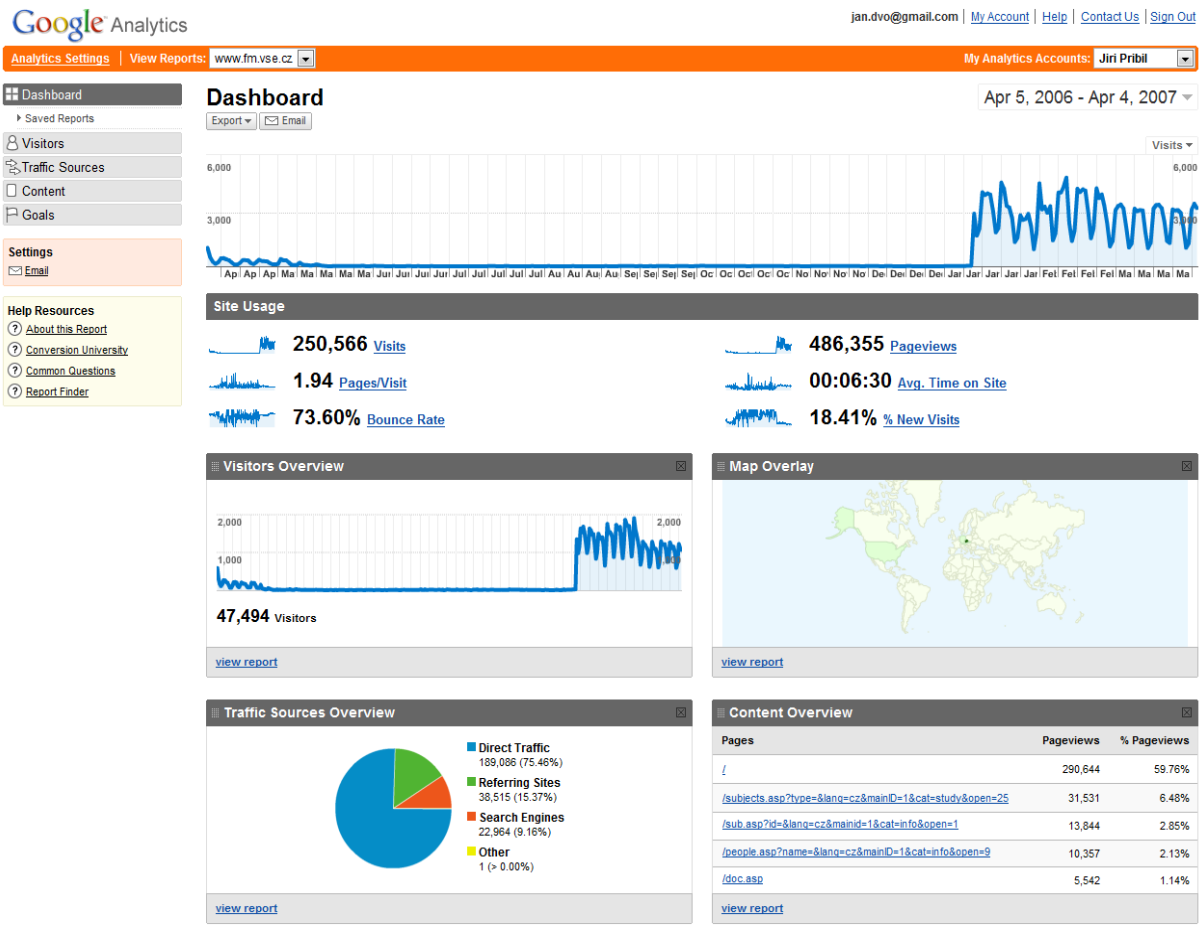
84,7 % požadavků skončilo se stavovým hlášením 200 (OK – vše v pořádku, server zaslal odpověď). Nejčastějším chybovým hlášením bylo 401 (neautorizovaný přístup). Dále 500 (vnitřní chyba serveru – při zpracování dotazu došlo v programu serveru k blíže neurčené chybě), 302 (objekt dočasně přesunut), 206 (neúplný obsah dokumentu), 404 (stránka nenalezena, jedná se o obecně nejrozšířenější chybu, v našem případě tomu ale tak není), 304 (beze změny, odpověď neobsahuje tělo zprávy), 501 (hlášení serveru, pokud je po něm vyžadována metoda, kterou neovládá), 301 (objekt trvale přestěhován na nové URI, klient se musí zeptat na novém umístění), 403 (obecná chyba, server by rád odpověděl, ale nemá to povoleno) (Brbla 2005).

	Odpověďi serveru	▼ Záznamů v logu		0 - 100 %	Počet stránek	Návštěvníků	Přenesených dat
1	200	3 530 851	84.7 %		3 530 777	78 901	0 b
2	401	311 618	7.5 %		311 618	18 652	0 b
3	500	134 028	3.2 %		134 028	2 759	0 b
4	302	76 576	1.8 %		76 576	5 025	0 b
5	206	47 727	1.1 %		47 727	5 954	0 b
6	404	47 025	1.1 %		39 554	10 128	0 b
7	304	7 457	0.2 %		7 393	1 764	0 b
8	501	7 004	0.2 %		7 004	1 414	0 b
9	301	6 093	0.1 %		6 093	1 751	0 b
10	403	1 075	0.0 %		1 075	291	0 b
	2 dalších položek	131	0.0 %		131	-	0 b
	Celkem	4 169 585	100 %		4 161 976	-	0 b

Obr. 50 **Odpovědi serveru**

Google analytics

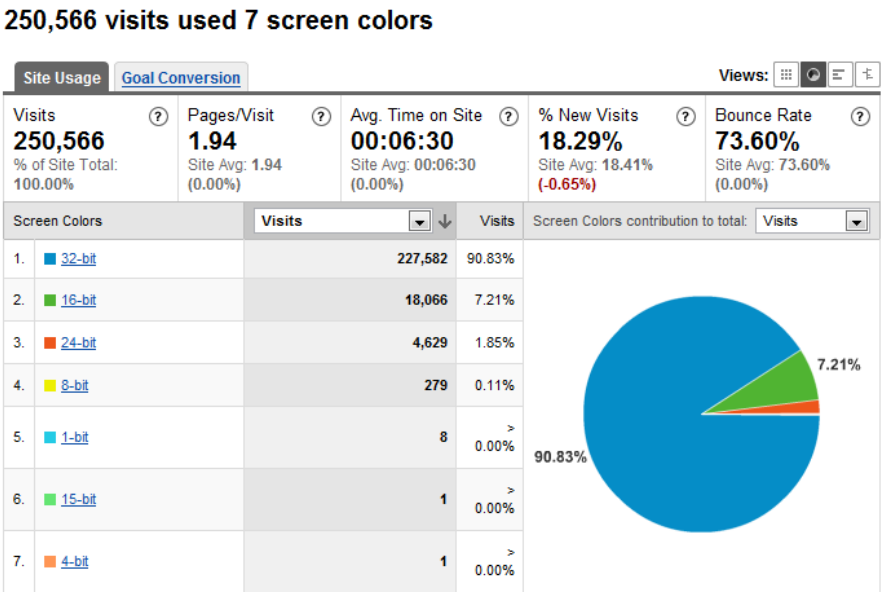
Analýza pomocí Googlu nevychází z logovacího souboru webového serveru. Google sbírá data pomocí skriptů vložených do webových stránek. Jedná se o sběr dat na straně klienta. Při prvním pohledu na úvodní stránku Google analytics jsem zjistil, že analýza webu www.fm.vse.cz pro sledované období nebude možná. Účet u Googlu je sice zřízen už delší dobu, ale skripty byly do stránek přidány až v lednu letošního roku. Není tak možné provést analýzu za sledované období od 5. dubna 2007. Navíc skripty nejsou doposud přidány do všech stránek, například do informačního systému, o kterém jsem v předchozích analýzách zjistil, že je z celého webu nejžádanější. Výsledky analýzy by tak neodpovídaly realitě ani za období od ledna do dubna 2007. Následný text tak bude pouze demonstrativní, ve kterém se zaměřím především na informace, které nejsou z logovacího souboru dostupné.



Obr. 51 **Úvodní stránka Google analytics**

Hloubka barev

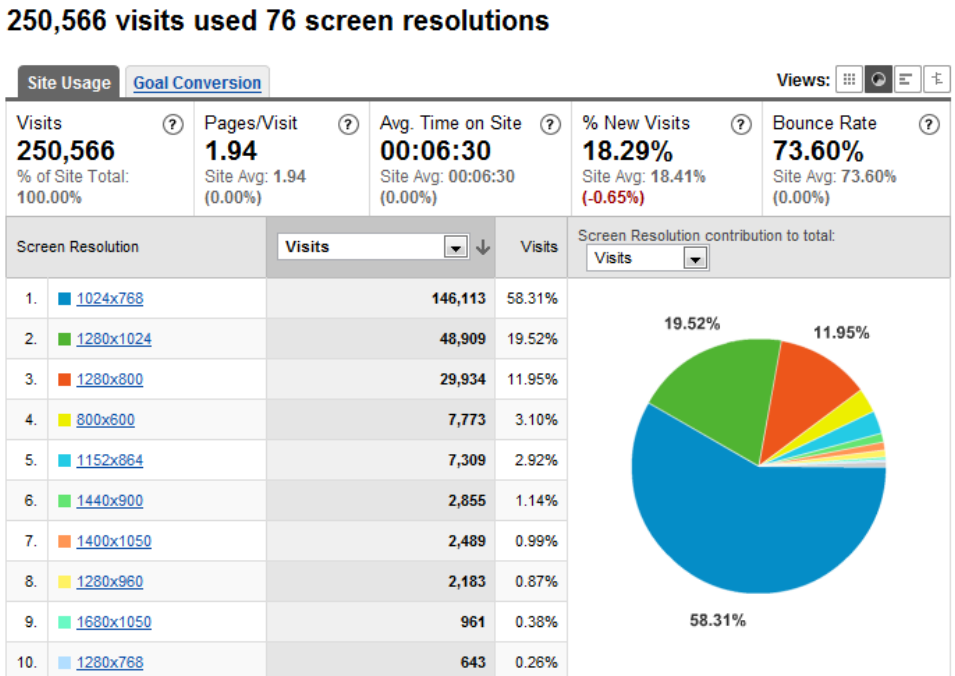
Poskytuje informace o nastavení hloubky barev na straně uživatele.



Obr. 52 **Hloubka barev**

Rozlišení obrazovky

Informace o nejčastěji používaném rozlišení jsou vhodné pro optimalizaci velikosti webové prezentace. Dobré je vyhnout se tomu, aby uživatel na domovské stránce musel stránku svisle posunovat, vodorovnému posunu je vhodné se vyhnout na všech stránkách.



Obr. 53 **Rozlišení obrazovky**

Sítě uživatelů

Dává přehled o tom, z jakých sítí se uživatelé nejčastěji připojují.

250,566 visits came from 2,949 network locations

Site Usage | Goal Conversion | Views:

Visits	Pages/Visit	Avg. Time on Site	% New Visits	Bounce Rate
250,566 % of Site Total: 100.00%	1.94 Site Avg: 1.94 (0.00%)	00:06:30 Site Avg: 00:06:30 (0.00%)	18.29% Site Avg: 18.41% (-0.65%)	73.60% Site Avg: 73.60% (0.00%)

	Network Location	Visits ↓	Pages/Visit	Avg. Time on Site	% New Visits	Bounce Rate
1.	University of Economics	98,198	1.50	00:07:41	9.70%	79.98%
2.	XDSL NETWORK-ADSL	29,379	2.12	00:06:25	19.93%	71.18%
3.	Cesky Telecom, A.S.	8,058	1.98	00:06:11	18.38%	74.55%
4.	UPC Ceska republika, a.s.	7,366	2.28	00:05:40	23.66%	69.06%
5.	JHComp s.r.o.	6,592	1.67	00:06:43	9.74%	77.61%
6.	Eurotel Praha, spol. s r.o.	5,703	2.05	00:07:04	21.55%	70.31%
7.	GTS NOVERA a.s.	3,687	2.50	00:05:55	22.27%	70.19%
8.	GPRS/WBA customer networks.	3,684	1.71	00:06:43	15.34%	76.49%
9.	Sloane Park Property Trust, a.s.	2,919	2.07	00:06:40	18.50%	71.53%
10.	UPC Internet CATV	2,226	2.80	00:05:39	33.56%	59.12%

Obr. 54 **Sítě uživatelů**

Odkazující stránky

Informace o tom, ze kterých stránek uživatelé nejčastěji přicházejí.

Referring sites sent 38,515 visits via 190 sources

Segment: Source

Site Usage | Goal Conversion | Views:

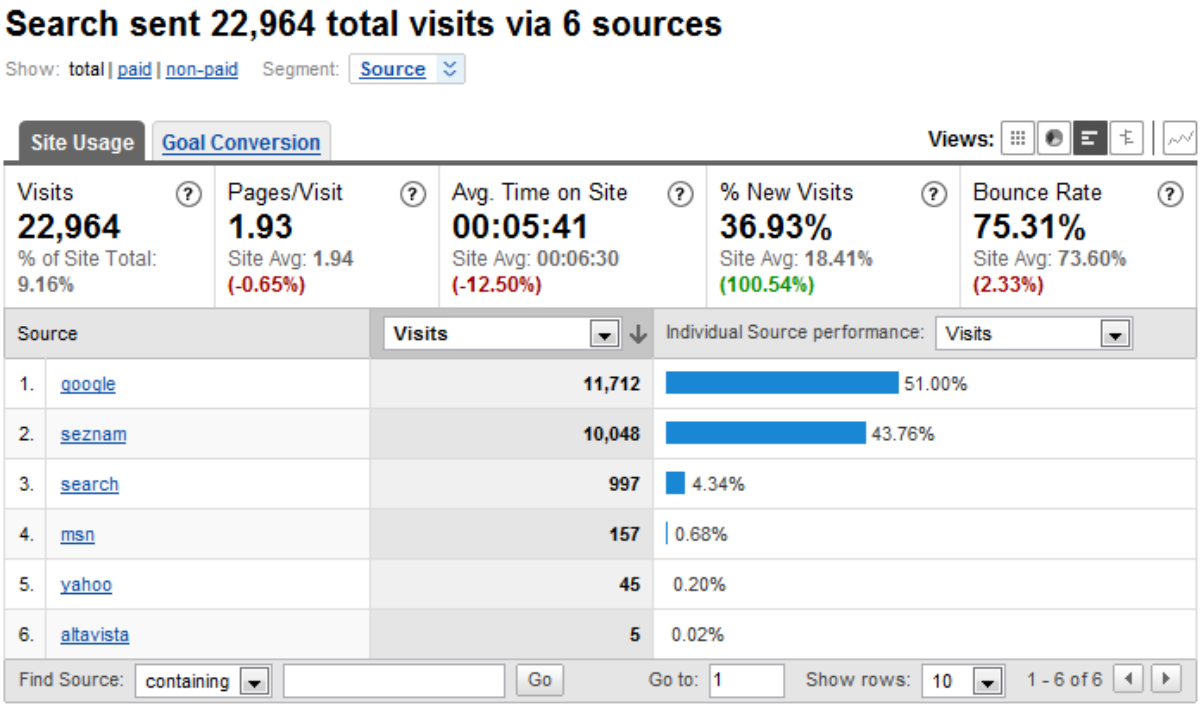
Visits	Pages/Visit	Avg. Time on Site	% New Visits	Bounce Rate
38,515 % of Site Total: 15.37%	3.53 Site Avg: 1.94 (81.63%)	00:04:59 Site Avg: 00:06:30 (-23.25%)	47.41% Site Avg: 18.41% (157.45%)	48.35% Site Avg: 73.60% (-34.30%)

	Source	Visits	Individual Source performance: Visits
1.	vse.cz	21,564	55.99%
2.	fm.vse.cz	7,061	18.33%
3.	firmy.cz	2,941	7.64%
4.	univerzita.net	1,212	3.15%
5.	search.seznam.cz	795	2.06%
6.	icq.com	698	1.81%
7.	skoly.vzdelani.cz	407	1.06%
8.	studentin.cz	270	0.70%
9.	prijimacky.fm.vse.cz	266	0.69%
10.	scio.cz	252	0.65%

Obr. 55 **Odkazující stránky**

Vyhledávací stroje

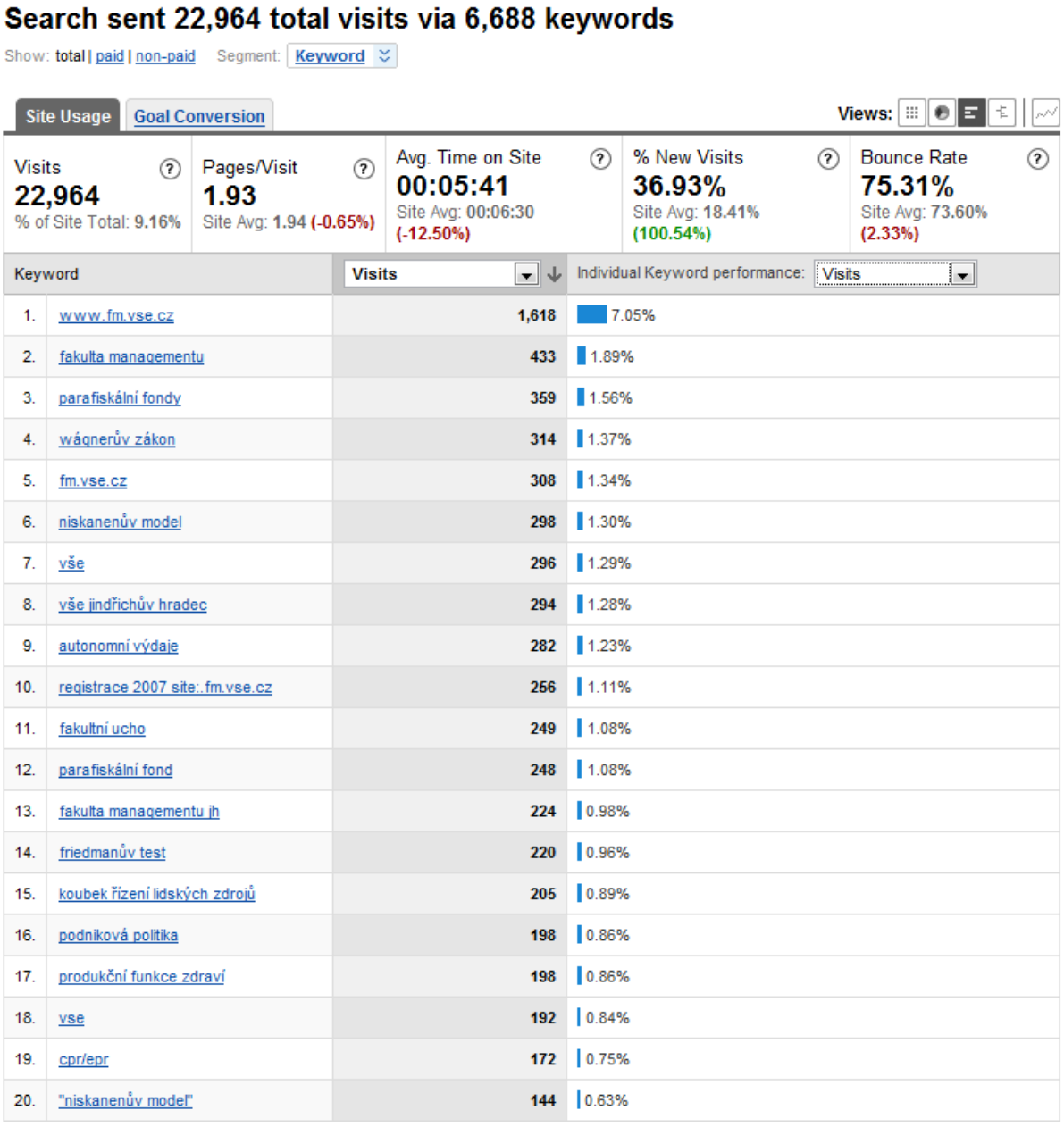
Zde můžeme získat informace o tom, pomocí kterých vyhledávacích strojů se návštěvníci nejčastěji dostanou na naše stránky. Jak je vidět na obrázku, více než polovina uživatelů využívá služeb zahraničních vyhledávačů, především Googlu. Z českých vyhledávačů to byl za sledované období pouze Seznam.



Obr. 56 **Vyhledávací stroje**

Klíčová slova

K vyhledání stránek fakulty použili uživatelé ve vyhledávačích za sledované období 6688 různých klíčových slov. Na základě použitých klíčových slov je možné usuzovat o tom, s jakým zájmem k nám návštěvníci přišli. Ke každému klíčovému slovu je možné zobrazit, v jakém vyhledávači bylo použito a statistiky o uživatelích, kteří jej zadali.



Search sent 22,964 total visits via 6,688 keywords

Show: total | paid | non-paid Segment: Keyword

Site Usage | Goal Conversion Views:

Visits	Pages/Visit	Avg. Time on Site	% New Visits	Bounce Rate
22,964	1.93	00:05:41	36.93%	75.31%
% of Site Total: 9.16%	Site Avg: 1.94 (-0.65%)	Site Avg: 00:06:30 (-12.50%)	Site Avg: 18.41% (100.54%)	Site Avg: 73.60% (2.33%)

	Keyword	Visits	Individual Keyword performance: Visits
1.	www.fm.vse.cz	1,618	7.05%
2.	fakulta managementu	433	1.89%
3.	parafiskální fondy	359	1.56%
4.	wágnerův zákon	314	1.37%
5.	fm.vse.cz	308	1.34%
6.	niskanenův model	298	1.30%
7.	vše	296	1.29%
8.	vše jindřichův hradec	294	1.28%
9.	autonomní výdaje	282	1.23%
10.	registrace 2007 site:.fm.vse.cz	256	1.11%
11.	fakultní ucho	249	1.08%
12.	parafiskální fond	248	1.08%
13.	fakulta managementu jh	224	0.98%
14.	friedmanův test	220	0.96%
15.	koubek řízení lidských zdrojů	205	0.89%
16.	podniková politika	198	0.86%
17.	produkční funkce zdraví	198	0.86%
18.	vse	192	0.84%
19.	cpr/epr	172	0.75%
20.	"niskanenův model"	144	0.63%

Obr. 57 **Klíčová slova**

Nejžádanější stránky

Pomocí této volby můžeme zjistit, které stránky z webové prezentace jsou nejžádanější. Její výhoda spočívá v tom, že na rozdíl například od programu Sawmill, nezobrazuje pouze jména jednotlivých souborů, ale názvy jednotlivých webových stránek. To usnadňuje orientaci ve výsledcích a další práci s nimi.

26 page titles were viewed a total of 486,355 times

Content Performance — Views:

Pageviews	Unique Pageviews	Time on Page	Bounce Rate	% Exit	$ Index
486,355	381,392	00:01:49	73.60%	51.52%	$0.00
% of Site Total: 100.00%	% of Site Total: 100.00%	Site Avg: 00:01:49 (0.00%)	Site Avg: 73.60% (0.00%)	Site Avg: 51.52% (0.00%)	Site Avg: $0.00 (0.00%)

	Page Title	Pageviews	Individual Page Title performance: Pageviews
1.	Fakulta managementu Vysoké školy ekonomické v	459,451	94.47%
2.	.: Fakulta managementu	VŠE	Jindřichův Hradec :	18,643	3.83%
3.	Fakulta managementu Vysoké %u0161koly ekonom	2,612	0.54%
4.	..:: Institut managementu zdravotnických služeb ::..	2,394	0.49%
5.	Přijímací řízení 2006	1,036	0.21%
6.	Fakulta managementu, VŠE	Výběr témat bakalářsl	615	0.13%
7.	.: Fakulta managementu	V%u0160E	Jind%u0159i	535	0.11%
8.	Studijní oddělení Fakulty managementu VŠE	280	0.06%
9.	Frequently asked questions - často kladené dotazy	263	0.05%
10.	Fakulta managementu Vysokй љkoly ekonomickй v	116	0.02%

Find Page Title: containing — Go — Go to: 1 — Show rows: 10 — 1 - 10 of 26

Obr. 58 **Nejžádanější stránky**

Žádanost odkazů na stránce

Unikátní statistika, která umožňuje grafické zobrazení žádanosti jednotlivých odkazů přímo na konkrétní webové stránce. U každého odkazu je malý graf zobrazující relativní podíl kliknutí na daný odkaz vůči všem odkazům na stránce. Výsledků je možné využít například při optimalizaci navigace. Žádanější odkazy v menu stránky by bylo vhodné zvýraznit a seskupit. Usnadnila by se tak orientace uživatelů, kteří by nebyli nuceni pročítat velké množství odkazů, které jsou méně často užívané.

Obr. 59 **Žádanost odkazů na stránce**

Závěr

Webových prezentací na webu neustále přibývá, konkurence se zvyšuje. Pokud chceme se svou webovou prezentací uspět, musíme ji návštěvníkům nabídnout v takové podobě, kterou žádají, a získat tak konkurenční výhodu nad stránkami s podobným obsahem. Nejvhodnějším podkladem pro modifikaci webových prezentací je porozumění uživatelskému chování. K tomu slouží různé metody Web Usage Miningu.

Oblast Web Usage Miningu se dynamicky rozvíjí a v současné době již existuje velké množství více či méně kvalitních softwarových nástrojů. Pokud se ale nespokojíme s výsledky běžných analýz, a ze záznamů o přístupech budeme chtít vydolovat co nejvíce informací, budeme nuceni si analýzu provést sami na základě některé z výše popsaných metod Web Usage Miningu.

Můžeme se vydat různými cestami, podle toho jaká data se pro analýzu rozhodneme použít. Nejčastěji to jsou data na straně serveru (logovací soubory) nebo data na straně klienta. Oba směry mají své výhody a nevýhody.

Využitím logovacích souborů dosáhneme nejlepších výsledků. Tato cesta je ale velice pracná a náročná na výpočty. Logy je nutné složitě upravovat a čistit, teprve pak jsou vhodné k dolování dat, které samo o sobě také není jednoduchou záležitostí. Když se pro tuto metodu rozhodneme dnes, můžeme analýzu provést libovolně do minulosti, protože logovací soubory se vytvářejí a ukládají na každém webovém serveru.

Druhá cesta, o poznání jednodušší, je vhodná pro méně náročné analytiky, kterých je ale většina. Stačí si založit účet například u Google analytics a vložit vygenerovaný skript do všech webových stránek, to je vše co je nutné udělat. Pak už je možné pohodlně procházet výsledky statistik. Nevýhodou je, že data se sbírají od doby začlenění skriptů do stránek. Proto pokud se pro analýzu rozhodneme dnes, nemůžeme ji provést do minulosti. S podobným problémem jsem se potýkal i ve své práci, kdy účet u Googlu byl založen, ale skripty nebyly začleněny do stránek.

Ať se rozhodneme pro jakýkoliv způsob provedení Web Usage Miningu, porozumění uživatelskému chování a jeho reflektování při tvorbě a modifikaci webu, bude vedle obsahu základním kamenem úspěchu většiny webových prezentací.

Literatura:

ADS Lab 2007, *An Introduction to Web Mining*, Advanced Data System Laboratory. Dostupné z: <http://dblab.csie.ncku.edu.tw/~tsengsm/COURSE/WebDB/Web-Mining.ppt> [16. února 2007].

Barsagade, N 2003, *Web Usage Mining and Pattern Discovery: A Survey Paper*, SMU School of Engineering. Dostupné z: <http://engr.smu.edu/~mhd/8331f04/barsagada.doc> [18. března 2007].

Berendt, B 2003, *Web Mining Seminar III: Data mining techniques and resources*, Umboldt-Universität zu Berlin. Dostupné z: <http://vasarely.wiwi.hu-berlin.de/lehre/2002w/wmi/Session3/index.pdf> [25. února 2007].

Borges, J, L, C 2000, *A Data Mining Model to Capture UserWeb Navigation Patterns*, Department of Computer Science University College London. Dostupné z: <http://paginas.fe.up.pt/~jlborges/publications/BorgesPhDthesis.pdf.zip> [16. února 2007].

Borges, J, Levene, M 1999a, *Data Mining of User Navigation Patterns*, Department of Computer Science University College London. Dostupné z: <http://www.informatik.uni-siegen.de/~galeas/papers/web_usage_mining/Data_Mining_of_User_NAvigation_Patterns_(Borges1999a).pdf > [16. února 2007].

Borges, J, Levene, M 1999b*, Mining Navigation Patterns with Hypertext Probabilistic Grammars*, Department of Computer Science University College London. Dostupné z: <http://www.informatik.uni-siegen.de/~galeas/papers/web_usage_mining/Mining_Navigation_Patterns_with_Hypertext_Probabilistic_Grammars(Borges1999c).pdf> [16. února 2007].

Büchner, A, G, Baumgarten, M, Anand, S, S, Mulvenna, M, D, Hughes, J, G 1999*, Navigation Pattern Discovery from Internet Data*, University of Ulster. Dostupné z: <http://www.informatik.uni-siegen.de/~galeas/papers/web_usage_mining/Navigation_Pattern_Discovery_from_Internet_Data_(Buechner 1999).pdf> [16. února 2007].

Brbla 2005, *Stavová hlášení HTTP protokolu, chybové kódy*, Absolut beginer on WWWeb. Dostupné z: <http://www.abowe.brbla.net/2-sitove-protokoly/stavova-hlaseni-http-protokolu.php> [10. července 2007].

Cooley, R, Mobasher, B, Srivastava, J 1997, *Web Mining: Information and Pattern Discovery on the World Wide Web*, University of Minnesota. Dostupné z: <http://maya.cs.depaul.edu/~mobasher/papers/webminer-tai97.pdf> [11. října 2006].

Cooley, R, Mobasher, B, Srivastava, J 1999, *Data Preparation for Mining World Wide Web Browsing Patterns*, University of Minnesota. Dostupné z: <http://www.informatik.uni-siegen.de/~galeas/papers/web_usage_mining/Data_Preparation_for_Mining_WWW_Browsing_Patterns_(Cooley1999a).pdf> [11. října 2007].

Čenovský, L 2003, *Web Usage Mining on is.muni.cz*, Masaryk University. Dostupné z: <http://is.muni.cz/clanky/web_usage_thesis.pl?zpet=../clanky/index.pl> [5. června 2006].

Galeas, P 2006, *Web mining,* Patricio Galeas. Dostupné z: <http://www.galeas.de/webmining.html> [11. října 2006].

Google 2007, *Google Analytics*, Google. Dostupné z: <http://www.google.com/analytics/> [16. února 2007].

HypKNOWsys 2005, *The Web Utilization Miner WUM*, HypKNOWsys. Dostupné z: <http://hypknowsys.sourceforge.net/wiki/The_Web_Utilization_Miner_WUM> [20. února 2007].

Ching-Nan Lin 2002, *Enhancement of Web Sites Security Utilizing Web Logs Mining*, Chung Yuan Christian University. Dostupné z: <http://thesis.lib.cycu.edu.tw/ETD-db/ETD-search/getfile?urn=etd-0807102-102609&&filename=8976022.pdf> [25. února 2007].

Jelínek, J 2004a, *Uživatelská podpora v prostředí WWW*, Inforum. Dostupné z: <http://www.inforum.cz/inforum2004/pdf/Jelinek_Jiri.pdf> [11. října 2006].

Jelínek, J 2004b, *Uživatelská podpora v prostředí WWW*, Inforum. Dostupné z: <http://www.inforum.cz/inforum2004/pdf/Jelinek_Jiri1.pdf> [7. června 2006].

Jelínek, J, Kincl, T 2005, *Techniky a nástroje sémantického vyhledávání*, Inforum. Dostupné z: <http://www.inforum.cz/inforum2005/pdf/Jelinek_Jiri1.pdf> [7. června 2006].

Jiang, Q 2003, *Web Usage Mining: Processes and Applications*, SMU School of Engineering. Dostupné z: <http://engr.smu.edu/~mhd/8331f03/jiang.ppt> [16. února 2007].

Joshi, A 2006, *Web Mining*, UMBC. Dostupné z: <http://www.cs.umbc.edu/~ajoshi/web-mine/> [11. října 2006].

KDnuggets 2007, *Web Mining and Web Usage Mining Software*, KDnuggets. Dostupné z: <http://www.kdnuggets.com/software/web-mining.html> [16. února 2007].

Koster, M, *A standard for Robot Exclucion*, Robots.org. Dostupné z: <http://www.robotstxt.org/wc/norobots.html> [5. března 2007].

Koutri, M, Avouris, N, Daskalaki, S 2004, *A survey on web usage mining techniques for web-based adaptive hypermedia systems*, University of Patras. Dostupné z: <http://hci.ece.upatras.gr/pubs_files/v13_Koutri_Avouris_Daskalaki_2004.pdf> [25. února 2007].

Krause, M 2001, *Červ Nimda jde po MS IIS*, Root.cz. Dostupné z: <http://www.root.cz/zpravicky/cerv-nimda-jde-po-ms-iis/> [10. července 2007].

Kryl, M 2004, *Bloková analýza vylepšuje PageRank*, Lupa server o českém internetu. Dostupné z: <http://www.lupa.cz/clanky/blokova-analyza-vylepsuje-pagerank/> [8. března 2007].

Křipáč, M, Novák, J, Vildová, H, *Projekt z vyhledávání znalostí v databázích, Web Usage Mining*, Masaryk University Brno Faculty of Informatics. Dostupné z: <http://www.fi.muni.cz/~kripac/WUM/> [2. července 2007].

Křivánek, P, *Skryté Markovovy modely (Hidden Markov Models - HMMs)*, Masarykova univerzita, Fakulta informatiky. Dostupné z: <http://nlp.fi.muni.cz/nlp/nlp-prace/referaty/xkrivan/HMM.html> [3. července 2007].

Microsoft 2007, *Microsoft Windows Server 2003 TechCenter*, Microsoft. Dostupné z: <http://www.microsoft.com/technet/prodtechnol/WindowsServer2003/Library/IIS/ffdd7079-47be-4277-921f-7a3a6e610dcb.mspx?mfr=true> [1. července 2007].

Mobasher, B 1997a, *Data cleaning*, University of Minnesota. Dostupné z: <http://maya.cs.depaul.edu/~mobasher/webminer/survey/node11.html#secclean> [8. října 2006].

Mobasher, B 1997b, *Web Usage Mining Architecture*, University of Minnesota. Dostupné z: <http://maya.cs.depaul.edu/~mobasher/webminer/survey/node23.html#SECTION00050000000000000000> [8. října 2006].

Rehberger, I 2002a, *Clickstream analýza: Seznamte se, prosím*, Lupa server o českém internetu. Dostupné z: <http://www.lupa.cz/clanky/clickstream-analyza-seznamte-se-prosim/> [10. března 2007].

Rehberger, I 2002b, *Obsahují webové logy bohatství?,* Lupa server o českém internetu. Dostupné z: <http://www.lupa.cz/clanky/obsahuji-webove-logy-bohatstvi/> [10. března 2007].

Rychlý, M 2005, *Klasifikace a predikce*, Vysoké učení technické v Brně. Dostupné z: <http://www.fit.vutbr.cz/~rychly/public/docs/classification-and-prediction/classification-and-prediction.xml> [5. července 2007].

Sawmill 2007, *Sawmill Enterprise*, Sawmill. Dostupné z: <http://www.sawmill.net/ent.html> [1. června 2007].

Shiwei, T, Jiawei, H, Dongqing, Y, Jian, P, Hongjun, L, Shojiro, N 2007, *H-mine: fast and space-preserving frequent pattern mining in large databases*, Goliath. Dostupné z: < http://goliath.ecnext.com/coms2/gi_0199-6533442/H-mine-fast-and-space.html> [5. července 2007].

Sklenák, V 2005, *Vyhledávání informací v prostředí webu – mírný pokrok v mezích zákona*, Automatizace knihovnických procesů. Dostupné z: <http://www.akvs.cz/akp-2005/10-sklenak.pdf> [7. června 2006].

SPSS 2006a, *Co je to Web mining? Analýza průchodu internetových stránek*, SPSS CR, spol s r.o. Dostupné z: <http://www.spss.cz/sl_webmining.html> [11. října 2006].

SPSS 2006b, *Objevte, jak Data mining změní Vaši organizaci*, SPSS CR, spol s r.o. Dostupné z: <http://www.spss.cz/sl_datamining.html> [11. října 2006]."

SPSS 2007, *Clementine*, SPSS Inc. Dostupné z: <http://www.spss.cz/sw_clementine.htm>, <http://www.spss.com/web_mining_for_clementine/> [1. června 2007].

Srivasta, J, Cooley, R, Deshpande, M, Pang-Ning Tan 2000, *Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data*, University of Minnesota. Dostupné z: <http://www.acm.org/sigs/sigkdd/explorations/issue1-2/srivastava.pdf> [16. února 2007].

Tanasa, D 2005, *Web Usage Mining: Contributions to Intersites Logs Preprocessing and Sequential Pattern Extraction with Low Support*, Université De Nice Sophia Antipolis - UFR Sciences. Dostupné z: <http://www-sop.inria.fr/axis/personnel/Doru.Tanasa/these_TANASA.pdf> [16. února 2007].

Wang, Y 2000, *Web Mining and Knowledge Discovery of Usage Patterns*, University of Waterloo. Dostupné z: <http://se.uwaterloo.ca/~tozsu/courses/cs748t/surveys/wang.pdf> [8. března 2007].

Wikipedia 2006, *Web Mining*, Wikipedia. Dostupné z: <http://en.wikipedia.org/wiki/Web_mining> [11. října 2006].

Wikipedia 2007a, *Rozhodovací stromy*, Wikipedia. Dostupné z: <http://cs.wikipedia.org/wiki/Rozhodovac%C3%AD_stromy> [10. června 2007].

Wikipedia 2007b, *Naive Bayes classifier*, Wikipedia. Dostupné z: <http://en.wikipedia.org/wiki/Naive_Bayes_classifier> [10. června 2007].

Zapletal, L 2001, *Protokol HTTP 1.1 pod lupou*, Root.cz. Dostupné z: <http://www.root.cz/clanky/protokol-http-1-1-pod-lupou/> [1. července 2007].