Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií
FIIT-5220-17475

Bc. Michal Holub

PRISPÔSOBOVANIE NAVIGÁCIE VO WEBOVOM SÍDLE NA ZÁKLADE SPRÁVANIA SA POUŽÍVATEĽOV

Diplomová práca

Študijný program: Softvérové inžinierstvo
Študijný odbor: 9.2.5 Softvérové inžinierstvo
Miesto vypracovania: Ústav informatiky a softvérového inžinierstva, FIIT STU v Bratislave
Vedúca práce: prof. Ing. Mária Bieliková, PhD.

máj 2010

SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE
Fakulta informatiky a informačných technológií

Návrh zadania diplomovej práce

Finálna verzia[1]

Študent

Meno, priezvisko, tituly:	Michal Holub, Bc.
Študijný program:	Softvérové inžinierstvo
Kontakt:	miso.holub@gmail.com

Výskumník:

Meno, priezvisko, tituly:	Mária Bieliková, prof. Ing. PhD.

Projekt:

Názov:	Vylepšovanie štruktúry webových sídiel na základe modelov
Miesto vypracovania:	Ústav informatiky a softvérového inžinierstva
Oblasť problematiky[2]**:**	Webové inžinierstvo

Text zadania

Práca s informáciami na webe je čoraz dôležitejšia pre mnohé ľudské činnosti. V tejto súvislosti vzniká požiadavka na vytváranie efektívnych webových sídiel schopných prispôsobiť sa kontextu, v akom sú použité. Väčšina webových sídiel sa dnes vytvára len s malým ohľadom na informačné potreby a spôsoby práce s informáciami budúcich používateľov. Často tak vznikajú stránky s nevhodnou štruktúrou. Túto štruktúru by bolo možné vylepšiť na základe vzorov použiteľných v určitých kontextoch. Analyzujte problematiku identifikácie vzorov vo webových sídlach. Zamerajte sa pri tom na samotnú stavbu webového sídla, ako aj na analýzu správania sa jeho návštevníkov, ktorých môžeme na základe sledovania ich činnosti na stránke zaraďovať do skupín a podľa charakteristík skupiny následne vylepšiť štruktúru webového sídla tak, aby lepšie zodpovedala potrebám skupiny. Navrhnite metódu na monitorovanie zvoleného webového sídla, pomocou ktorej bude možné vo webovom sídle vyhľadať štruktúrne vzory a vzory správania sa používateľov. Na základe identifikovaných vzorov navrhnite vylepšenia štruktúry webového sídla (zlepšenie navigácie, prispôsobenie sídla určitej skupine používateľov, atď.), Navrhnuté riešenie experimentálne overte prostredníctvom prototypu webového systému, ktorý bude realizovať navrhnutú metódu.

150-200 slov, ktoré opisujú výskumný problém v kontexte súčasného stavu vrátane motivácie a smerov riešenia

[1] Veľkosť jednotlivých polí pre vypĺňanie nemožno meniť. Návrh zadania vytlačiť obojstranne na jeden list papiera.

[2] Identifikácia oblasti v rámci odboru štúdia, na ktorú sa projekt primárne viaže

Literatúra

- *Eirinaki, M., Vazirgiannis M.: Web Mining for Web Personalization. In ACM Transactions on Internet Technology (TOIT), Vol. 3, Issue 1 (February 2003), 1-27.*
- *Melody Y. Ivory, Rodrick Megraw: Evolution of Web Site Design Patterns. In ACM Transactions of Information Systems (TOIS), Vol. 23, Issue 4 (October 2005), 463-497.*
- *Sadagopan, N., Li, J.: Characterizing Typical and Atypical User Sessions in Clickstreams. In Proceeding of the 17th International Conference on World Wide Web, Beijing (2008), 885-894.*

2-3 vedecké zdroje, každý na samostatnom riadku a vo formáte zodpovedajúcom bibliografickým odkazom podľa normy STN ISO 690, ktoré sa viažu k téme zadania a preukazujú výskumnú povahu problému a jeho aktuálnosť (uveďte všetky potrebné údaje na identifikáciu zdroja, pričom uprednostnite vedecké príspevky v časopisoch a medzinárodných konferenciách)

Vyššie je uvedený návrh diplomového projektu, ktorý vypracoval *Michal Holub*, konzultovala a osvojila si ho *prof. Mária Bieliková* a súhlasí, že bude takýto projekt viesť v prípade, že bude pridelený tomuto študentovi.

V Bratislave, dňa 19.1.2009

Podpis študenta

Podpis výskumníka

Vyjadrenie garanta študijného programu

Návrh zadania schválený: áno / nie[3]

Dňa: .....

Podpis garanta

[3] Nehodiace sa prečiarknite

SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE
Fakulta informatiky a informačných technológií

Návrh zadania diplomovej práce

Revízia č.[1]: .1.....

Študent

Meno, priezvisko, tituly:	Michal Holub, Bc.
Študijný program:	Softvérové inžinierstvo
Kontakt:	miso.holub@gmail.com

Vedúci diplomovej práce:

Meno, priezvisko, tituly:	Mária Bieliková, prof. Ing. PhD.

Projekt:

Názov:	Prispôsobovanie navigácie vo webovom sídle na základe správania sa používateľov
Miesto vypracovania:	Ústav informatiky a softvérového inžinierstva
Oblasť problematiky[2]:	Webové inžinierstvo

Text zadania[3]

Práca s informáciami na webe je čoraz dôležitejšia pre mnohé ľudské činnosti. V tejto súvislosti vzniká požiadavka na vytváranie efektívnych webových sídiel schopných prispôsobiť sa kontextu, v akom sú použité. Väčšina webových sídiel sa dnes vytvára len s malým ohľadom na informačné potreby a spôsoby práce s informáciami budúcich používateľov. Často tak vznikajú stránky s nevhodnou štruktúrou. Túto štruktúru by bolo možné vylepšiť na základe vzorov správania sa používateľov počas návštevy webového sídla a jeho jednotlivých stránok. Analyzujte problematiku identifikácie vzorov vo webových sídlach. Zamerajte sa pri tom na samotnú stavbu webového sídla, ako aj na analýzu správania sa jeho návštevníkov, ktorých môžeme na základe sledovania ich činnosti na stránke zaraďovať do skupín a podľa charakteristík skupiny následne vylepšiť štruktúru webového sídla tak, aby lepšie zodpovedala potrebám skupiny. Navrhnite metódu na monitorovanie zvoleného webového sídla, pomocou ktorej bude možné vyhľadať vzory správania sa používateľov pri jeho prehliadaní. Na základe identifikovaných vzorov navrhnite vylepšenia štruktúry navigácie webového sídla prispôsobené konkrétnym používateľom a skupinám používateľov. Navrhnuté riešenie experimentálne overte prostredníctvom prototypu webového systému, ktorý bude realizovať navrhnutú metódu.

150-200 slov, ktoré opisujú výskumný problém v kontexte súčasného stavu vrátane motivácie a smerov riešenia

[1] Uvedie sa poradové číslo revízie

[2] Identifikácia oblasti v rámci odboru štúdia, na ktorú sa projekt primárne viaže

[3] Veľkosť jednotlivých polí pre vypĺňanie nemožno meniť. Formulár vytlačiť obojstranne na jeden list papiera.

Literatúra

- *Eirinaki, M., Vazirgiannis M.: Web Mining for Web Personalization. In ACM Transactions on Internet Technology (TOIT), Vol. 3, Issue 1 (February 2003), 1-27.*
- *Melody Y. Ivory, Rodrick Megraw: Evolution of Web Site Design Patterns. In ACM Transactions of Information Systems (TOIS), Vol. 23, Issue 4 (October 2005), 463-497.*
- *Sadagopan, N., Li, J.: Characterizing Typical and Atypical User Sessions in Clickstreams. In Proceeding of the 17th International Conference on World Wide Web, Beijing (2008), 885-894.*

2-3 vedecké zdroje, každý na samostatnom riadku a s údajmi zodpovedajúcimi bibliografickým odkazom podľa normy STN ISO 690, ktoré sa viažu k téme zadania a preukazujú výskumnú povahu problému a jeho aktuálnosť (uveďte všetky potrebné údaje na identifikáciu zdroja, pričom uprednostnite vedecké príspevky v časopisoch a medzinárodných konferenciách)

Vyššie je uvedená revízia návrhu diplomového projektu, ktorú vypracoval *Bc. Michal Holub*, konzultoval a osvojil si ho *prof. Ing. Mária Bieliková, PhD.*

V Bratislave dňa 5.2.2010

Podpis študenta

Podpis vedúceho diplomovej práce

Vyjadrenie garanta predmetov Diplomový projekt I, II, III

Revízia zadania schválená: áno / nie[4]

Dňa: .....

Podpis garanta predmetov

[4] Nehodiace sa prečiarknite

ANOTÁCIA

Slovenská technická univerzita v Bratislave
FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLÓGIÍ
Študijný program: SOFTVÉROVÉ INŽINIERSTVO

Autor: Bc. Michal Holub
Diplomová práca: Prispôsobovanie navigácie vo webovom sídle na základe správania sa používateľov
Vedúca diplomovej práce: prof. Ing. Mária Bieliková, PhD.
máj, 2010

Neustály nárast množstva informácií dostupných prostredníctvom webu reprezentuje veľkú výzvu. Potrebujeme prostriedky na efektívne prehliadanie webových sídiel bez zbytočnej záplavy nepodstatnými údajmi. Odpoveďou na túto výzvu je personalizácia obsahu webových sídiel. Predkladaná práca sa zaoberá hľadaním vhodnej metódy pre personalizáciu navigácie, ktorá by dávala používateľovi potrebné informácie bez toho, aby o ne musel explicitne žiadať. Analyzujeme existujúce metódy modifikovania štruktúry webových sídiel, ktorými sa zaoberajú adaptívne webové systémy. Predstavujeme návrh novej metódy adaptívnej podpory navigácie. Jej hlavným prínosom je automatické prispôsobovanie zobrazených odkazov podľa správania sa používateľa pri prehliadaní webového sídla. Ďalšou črtou metódy je automatické určovanie záujmu používateľa o prehliadanú stránku. Medzi podobne sa správajúcimi používateľmi odporúčame zaujímavé odkazy, ktoré zobrazujeme v rámci personalizovaného kalendára a personalizovaných noviniek. Implementáciu metódy overujeme v experimentoch s vylepšovaním štruktúry webového sídla našej fakulty.

ANNOTATION

Slovak University of Technology Bratislava
FACULTY OF INFORMATICS AND INFORMATION TECHNOLOGIES
Degree Course: SOFTWARE ENGINEERING

Author: Bc. Michal Holub
Master's Thesis: Adaptation of website's navigation based on behavior of users
Supervisor: prof. Ing. Mária Bieliková, PhD.
2010, May

The constant growth of the amount of information available on the Web represents a big challenge. We need means to effectively browse the websites without being overwhelmed by the irrelevant data. Answer to this challenge lies in personalization of the content of websites. This thesis deals with finding of a suitable method for personalization of website's navigation. With navigation personalized the user should get the right information without the need for explicitly asking for it. We analyse existing methods for modifying the content of websites by adaptive web-based systems. We propose novel method for adaptive navigation support. The main contribution of this method is automatic adaptation of showed links based on the monitoring of user's behaviour during his visit of the website. Another feature of this method is automatic estimation of user's interest in visited web page. We use proposed approach to recommend interesting links among similarly behaving users. The recommended links are presented in a form of personalized calendar and personalized news sections. We evaluate the proposed method in experiments with adaptation of the structure of our faculty's website.

Obsah

Kapitola 1

Úvod

Facebook is the third largest nation on the planet.

Bebo White

Množstvo informácií prístupných prostredníctvom webu sa neustále zvyšuje. Podľa WorldWideWebSize.com[1] obsahovala indexovaná časť webu v máji 2010 približne 20 miliárd stránok. Tento údaj vychádza z informácií o počte indexovaných webových stránok od popredných internetových vyhľadávačov (Google, Bing, Yahoo!, Ask). Spolu s rastúcim počtom dokumentov na webe rastie aj priemerný počet používateľom navštívených stránok za deň. Kým v roku 1994 ich bolo v priemere 14, v roku 2000 ich už bolo 42 a v roku 2005 dokonca 60 [17]. Nové informácie však pribúdajú oveľa rýchlejšie, ako je človek schopný spracovať ich. Z takéhoto prívalu noviniek je pre jednotlivca dôležitý len ich nepatrný zlomok, pričom je ťažké oddeliť užitočné informácie od nepodstatných. Pre návštevníka majú najväčšiu hodnotu "kvalitné" webové stránky, pričom pod kvalitou môžeme rozumieť aj "prispôsobenie sa štruktúry webovej stránky zámeru každej skupiny používateľov, ktorá k nej pristupuje" [8].

Používatelia sú rôznorodí, no webové sídla k nim vo väčšine prípadov pristupujú rovnako [7]. Všetkým ponúkajú rovnaký obsah, pričom neberú do úvahy ich rozdielnosť. Často sa tak stane, že používateľ dostane informácie, o ktoré nemá záujem, sú pre neho príliš špecifické alebo naopak príliš všeobecné.

Snaha o udržanie kroku s narastajúcim množstvom informácií využívaním tradičných techník je vopred odsúdená na neúspech. Človek skončí v stave, kedy nerobí nič iné, len triedi a ukladá nové poznatky, pričom mu neostáva čas na osvojenie si tých podstatných. Akú máme teda možnosť? Ukazuje sa, že webu začína dominovať personalizácia [3]. S jej využitím dokážeme návštevníkovi prezentovať vhodné informácie vďaka poznaniu jeho zámerov a záujmov. Pre každého používateľa budujeme model jeho cieľov, preferencií a znalostí a na základe tohto modelu prispôsobujeme zobrazované informácie [10].

[1] http://www.worldwidewebsize.com

Personalizáciu webu môžeme chápať ako "akúkoľvek akciu, ktorá konkrétnemu používateľovi alebo skupine používateľov sprostredkuje na mieru šitý zážitok z webu" [24]. Cieľom personalizácie webových stránok je dať používateľom informácie, ktoré potrebujú alebo chcú bez toho, aby si ich museli explicitne pýtať [26].

Užitočnosť webovej stránky pre jej návštevníka ovplyvňujú obsah poskytovaný webovým sídlom, dizajn jednotlivých stránok a štruktúra webového sídla. Do posledne menovaného patria hypertextové odkazy medzi jednotlivými stránkami. Pri vytváraní štruktúry webového sídla môže dôjsť k rozporu medzi potrebami návštevníkov, ako ich vidí dizajnér stránok, a skutočnými potrebami návštevníkov. Prvotné pokusy o personalizáciu predstavovali možnosti pre návštevníka zvoliť si obľúbené odkazy na prispôsobenej domovskej stránke internetového portálu. Tento prístup predpokladal, že návštevník pozná obsah celého webového sídla a vie si z neho vybrať pre neho zaujímavé odkazy. Ďalším krokom bolo začlenenie princípov umelej inteligencie. Takým je aj kolaboratívne filtrovanie, ktoré meria podobnosť medzi správaním sa rozličných používateľov. Tí však musia odhaliť časť svojho profilu vrátane záujmov. Ako uvádzame ďalej, toto používatelia nerobia radi. Alternatívu predstavuje pozorovanie používateľovho správania pri minulých návštevách, ktoré sa následne využívajú pri personalizácii. Používateľ tak nemusí explicitne uvádzať údaje o sebe a svojich cieľoch.

Dôležitými aspektmi pri personalizácii sú presnosť analýzy (je nutné správne odhadnúť používateľov aktuálny záujem), flexibilný prístup (potreba kombinovať informácie z rozličných zdrojov, ktoré sa vhodne dopĺňajú) a schopnosť uskutočnenia zmien v reálnom čase bez výrazného oneskorenia. Podstatnou je tiež otázka bezpečnosti, nakoľko používatelia sú citliví na svoje súkromie.

Neľahká úloha je nájsť správnu mieru personalizácie. Používateľom by sme mali posytovať len informácie, ktoré ich práve zaujímajú. Veľké množstvo odporúčaní však môže odvádzať pozornosť používateľa od jeho hlavného cieľa a spôsobovať frustráciu [27].

Návštevníci webového sídla využívajú na prístup k informáciám rozličné techniky. Najčastejšie používanou technikou je nasledovanie hypertextových odkazov. Tento spôsob sa využíva vo viac ako polovici všetkých prípadov [13, 31]. Ďalšou dominantnou technikou je použitie tlačidla *späť* vo webovom prehliadači [23]. Podiel prístupu k stránkam pomocou zoznamu obľúbených položiek, histórie, priameho zadania požadovanej adresy a ostatných techník je zanedbateľný (pohybuje sa v jednotkách percent).

Keďže dominantný podiel pri prístupe k webovým stránkam má používanie hypertextových odkazov, vzniká problém s nevhodnou navigáciou. Pri nevhodnej navigácii prestane mať návštevník po čase prehľad o tom, kde v kontexte webového sídla sa práve nachádza. Priveľké množstvo odkazov tiež spôsobuje, že sa používateľ nevie rozhodnúť, ktorý z nich má použiť. Z tohto dôvodu môže prispôsobenie navigácie na webovej stránke konkrétnemu návštevníkovi zlepšiť jeho prácu s webovým sídlom.

V tejto práci sa zaoberáme návrhom metódy na vylepšovanie štruktúry webových sídiel vo vybranej doméne. Metóda je založená na personalizácii navigácie a odporúčaní vhodných odkazov na základe sledovania správania sa používateľa a hľadania návštevníkov sídla, ktorí sú mu svojím správaním podobní. Opisujeme tiež návrh overenia tejto metódy pomocou implementácie prototypu na báze proxy servera. Pre takéto riešenie sme navrhli experimenty v rámci webového sídla našej fakulty na overenie vhodnosti jeho

použitia a vyhodnotenie jeho prínosu pre používateľov.

Práca sa skladá z 9 kapitol. Analýza problémovej oblasti je uvedená v kapitolách 2, 3 a 4. V nich predstavujeme podklady, z ktorých pri riešení projektu vychádzame. Druhá kapitola obsahuje analýzu metód, ktoré sa používajú na vylepšovanie štruktúry webových sídiel. Opisujeme tu metódy modifikácie webových stránok s dôrazom na prispôsobovanie navigácie. V tretej kapitole uvádzame analýzu získavania informácií o používateľovi a jeho správaní sa. V štvrtej kapitole predstavujeme vybrané existujúce adaptívne webové systémy a ich vlastnosti. Kapitola 5 obsahuje ciele práce. V šiestej kapitole predstavujeme navrhnutú metódu odhadovania záujmu používateľa o zobrazenú stránku zo sledovania jeho aktivít. V kapitole 7 opisujeme metódu vylepšovania webových sídiel využitím prispôsobovania navigácie, ktorá pracuje na základe modelov správania sa návštevníkov. Overenie navrhnutých metód implementovaným prototypom a experimentmi opisujeme v kapitole 8. Deviata kapitola obsahuje záver spolu s možnými využitiami navrhnutých metód a načrtnutie ďalších smerov vývoja.

Kapitola 2

Vylepšovanie štruktúry webových sídiel

Physicists analyze systems. Web scientists, however, can create the systems.

Tim Berners-Lee

Veľké webové sídla obsahujú možstvo rôznodorých informácií, ktorými chcú zasiahnuť širokú skupinu ľudí. Keďže však webové sídlo svojich návštevníkov "nepozná", prezentuje všetkým rovnaké informácie. Každý používateľ je však iný a zaujíma sa len o časť dostupných informácií. Preto má zmysel zaoberať sa prispôsobením webového sídla.

Prispôsobenie webového sídla má za cieľ upraviť štruktúru jeho stránok tak, aby prezentovali používateľovi informácie, o ktoré má záujem, formou, ktorú preferuje. Štruktúrou stránok rozumieme ich stavbu z hľadiska obsahu a použitých prvkov. V tejto kapitole uvádzame rôzne metódy vylepšovania webových sídiel pomocou personalizácie a prezentujeme ich základné vlastnosti.

2.1 Metódy prispôsobovania

Aby sme mohli webové sídlo prispôsobiť návštevníkovi, musíme poznať jeho záujmy a musíme mať model domény, ktorú prispôsobujeme. Na to slúžia [28]:

- analýza obsahu stránky a
- analýza správania návštevníka.

Analýza obsahu pracuje s webovou stránkou ako s dokumentom, z ktorého sa snaží získať dodatočné informácie. Patrí sem lexikálna a sémantická analýza, ktorých výstupom je zoznam jednotiek, z ktorých je dokument zložený, a ich význam. Tieto vstupy môžeme použiť na extrakciu rôznych "vyšších" jednotiek, akými sú kľúčové slová, osoby, geografické

miesta, dátumy, atď. Analýza obsahu nám umožňuje porozumieť téme, ktorej sa webová stránka venuje. Z nej môžeme ďalej odvodiť záujmy používateľa. Pri prispôsobovaní sa hľadajú podobnosti medzi analyzovanou webovou stránkou a modelom používateľa.

Analzýza správania nazerá na webovú stránku ako na súbor prvkov, s ktorými používateľ interaguje. Pri tejto analýze sa sledujú akcie, ktoré používateľ na stránke vykonáva. Podľa akcií môžeme tiež odvodiť záujmy používateľa. Používateľov môžeme zoskupovať podľa spoločných akcií a sledovať správanie celej skupiny. Všetky metódy, ktoré ďalej opisujeme, využívajú analýzu obsahu, správania, prípadne ich kombináciu.

Metódy vylepšujúce štruktúru webových stránok delíme podľa toho, ktoré prvky prispôsobujú. Štandardne sa prispôsobujú tieto časti webových stránok:

- obsah,
- navigácia a
- vzhľad.

Prispôsobovanie obsahu zahŕňa upravovanie hlavného poľa webovej stránky, ktoré obsahuje najdôležitejší text. Ak je text nevhodne štruktúrovaný, môžeme ho rozdeliť na menšie celky. Jednotlivé paragrafy môžeme skrývať a zobrazovať až vo chvíli, keď sú pre používateľa vhodné (napr. pri výučbových textoch môžeme najskôr ukázať teóriu a až po jej prečítaní zobrazíme používateľovi príklad, ktorým si overí, čo sa naučil). Pri prispôsobovaní obsahu sa využíva analýza textu, aby sme porozumeli jeho významu. Často z textu extrahujeme kľúčové slová, ktoré potom porovnávame s kľúčovými slovami v profile používateľa.

Pri prispôsobovaní navigácie upravujeme menu stránky. Odkazy v menu môžeme preusporiadať alebo úplne nahradiť. V druhom prípade sa používateľovi snažíme ponúknuť odkazy na stránky, ktoré by ho mohli zaujať. Nezaujímavé odkazy môžeme z menu čiastočne alebo úplne odstrániť. Pri prispôsobovaní navigácie sledujeme okrem analýzy obsahu odkazovanej stránky aj správanie sa používateľa. Zaujíma nás, ktoré odkazy používa a ktoré odkazy používajú iní.

Prispôsobovanie vzhľadu sa zaoberá rozmiestnením jednotlivých prvkov na stránke. Skúmajú sa pri tom návyky používateľa, sleduje sa jeho interakcia s webovou stránkou (kam sa pozerá, kam ukazuje kurzorom myši, informácie z ktorých častí webovej stránky si najviac zapamätá). Z týchto údajov sa potom tvoria tzv. teplotné mapy (angl. *heat map*). Výsledkom môže byť, že do oblasti, kam smeruje väčšina pozornosti návštevníkov, umiestnime dôležité informácie, zhrnutie článku, najvýznamnejšie odkazy. Prispôsobenie vzhľadu stránky sa často využíva na marketingové účely, kedy sa agentúry snažia do najsledovanejších oblastí stránky umiestňovať reklamu, aby si ju používateľ ľahko všimol.

Do kategórie prispôsobovania vzhľadu patrí aj generovanie personalizovanej webovej stránky. V tomto prípade sa jedná o vytvorenie novej stránky využívajúc obsah z viacerých zdrojov. Základom je pôvodná stránka, ktorej obsah tvorí hlavnú časť tej novovytvorenej. V obsahu sa identifikujú zaujímavé objekty, ku ktorým sa potom dopĺňajú ďalšie informácie. Ako príklad si vezmime osobný blog obsahujúci recenzie filmov [3]. Text recenzie tvorí hlavný obsah novej stránky. K nemu sa v bočnom paneli pridajú základné informácie (režisér, herci, atď.) o danom filme získané z iného zdroja (napr. on-line encyklopédia,

oficiálna stránka filmu). Ďalej je takáto stránka doplnená o informácie o časoch premietania filmu v miestnych kinách. Časy premietania sú získané zo stránok kín, tie sa vyberajú podľa polohy používateľa (ak ju vieme zistiť). Opísaný príklad predstavuje realizáciu jednej z myšlienok webu 2.0, kedy systém komunikuje s rozličnými službami a zo získaného obsahu vytvorí novú stránku, ktorú prezentuje používateľovi.

Druhý pohľad na plnú personalizáciu vzhľadu prinášajú niektoré tradičné internetové portály (napr. Yahoo![1] alebo MSN[2]). Tie najväčšie integrujú množstvo služieb od e-mailovej schránky cez burzové správy až po program TV staníc. Svojim používateľom umožňujú využívanie jedného prihlasovacieho mena a hesla na prístup ku všetkým týmto službám. Čo sa personalizácie týka, umožňujú používateľovi nastaviť si vzhľad a obsah hlavnej stránky portálu. Používateľ si môže zvoliť, ktoré sekcie a v akom rozsahu chce mať zobrazené. Taktiež si môže prispôsobiť dizajn stránky svojmu vkusu. Keďže nastavenia sa ukladajú centralizovane na serveri, personalizovaná verzia portálu je používateľovi dostupná kdekoľvek [21]. Táto metóda však nerieši automatické vylepšovanie štruktúry webu. Používateľ musí stráviť dlhý čas prispôsobovaním stránky svojim potrebám. Nakoľko sa jeho potreby v čase menia, musí tieto nastavenia priebežne upravovať. Práve nutnosť zásahov používateľa eliminujeme návrhom automatickej metódy na modifikáciu webovej stránky.

Z hľadiska vstupných údajov, podľa ktorých štruktúru webového sídla prispôsobujeme, rozlišujeme dva prístupy [22]:

- personalizácia na základe charakteristík individuálnych používateľov a
- personalizácia na základe charakteristík skupiny používateľov.

Pri prispôsobovaní podľa individuálnych charakteristík sa pozeráme na každého používateľa osobitne. Hľadáme zhodu medzi jeho záujmami a doménovým modelom, sledujeme jeho správanie. Následne pre neho vytvoríme stránku šitú na mieru. Každému používateľovi sa prezentuje jedinečná verzia stránky.

Na druhej strane, používateľov môžeme zoskupovať podľa spoločných čŕt. Potom určujeme charakteristiky vytvorených skupín a snažíme sa prispôsobiť webovú stránku konkrétnej skupine. V rámci skupiny by mali mať používatelia podobné záujmy a preferencie.

Prispôsobovanie jednotlivých častí, príp. celej webovej stránky, podľa uvedených metód sa môže diať manuálne alebo automaticky. Pri manuálnom prispôsobovaní nie je webová stránka priamo upravená. Výstupom takejto metódy je len odporúčanie pre správcu webového sídla, obsahujúce zoznam navrhovaných zmien. Ich realizácia ostáva na zváženie autorovi portálu. Naopak, pri automatickom prispôsobovaní sa webová stránka modifikuje podľa výstupu metódy bez nutnosti zásahu zo strany administrátora.

2.2 Prispôsobovanie navigácie

V tejto práci sa venujeme prispôsobovaniu navigácie vo webovom sídle, preto opíšeme jednotlivé metódy jej vylepšovania podrobnejšie. Všetky majú za cieľ uľahčiť používateľovi

[1] http://www.yahoo.com
[2] http://www.msn.com

pohyb po webovom portáli. Medzi hlavné metódy prispôsobovania navigácie patria [9]:

- *Usporiadanie odkazov* — táto metóda mení poradie odkazov podľa informácií z modelu používateľa. Je pravdepodobné, že používateľ si vyberie jeden z prvých odkazov, ktoré sú mu ponúknuté. Na vyššie miesto tak presúvame najrelevantnejšie odkazy.

- *Skrývanie odkazov* — motivácia tejto metódy je obdobná ako v predošlej, t.j. dostať na popredné miesta relevantné odkazy pre aktuálneho používateľa. Dosiahneme to nezobrazením nerelevantných odkazov. Takéto správanie však nemusia mať používatelia radi, pretože väčšinou si chcú vybrať sami.

- *Anotácia odkazov* — odkaz často tvorí krátke slovné spojenie, z ktorého nie je vždy jasné, čo sa za daným odkazom skrýva. Aj stránka so zaujímavým obsahom môže mať málo návštev v prípade, že sa na ňu používateľ nedostane vinou zle zvoleného textu odkazu. Táto metóda preto dopĺňa odkazy o ďalší popis, ktorý viac ozrejmuje obsah odkazovanej stránky. Zahŕňa tiež farebné zvýrazňovanie odkazov.

- *Generovanie odkazov* — vo webovom sídle môže existovať stránka podobná aktuálne prezeranej. Ak však tieto dve stránky na seba vzájomne neodkazujú, nemusí sa o nej používateľ dozvedieť. Metóda generovania odkazov vytvára nové prepojenia podobných stránok, čím pomáha používateľovi v ďalšej navigácii.

- *Priame smerovanie* — táto metóda na základe modelu používateľa, aktuálne prezeranej stránky a ostatných stránok vo webovom sídle rozhodne o ďalšej stránke, ktorá sa používateľovi zobrazí. Ten sa tak nerozhoduje samostatne, ale iba kliká na možnosť zobraziť nasledujúcu stránku. O tom, ktorá to bude, rozhoduje systém na pozadí.

Ďalší typ vylepšovania štruktúry, ktorý patrí do kategórie prispôsobovania navigácie, opisujú autori [16, 28]. Jedná sa o špeciálny prípad generovania odkazov v podobe indexových stránok. Táto metóda čerpá z údajov o prístupoch k webovým stránkam uložených na serveri. Z nich zisťuje, ktoré stránky sú často navštevované. Následne sa snaží vytvoriť vhodné prepojenia medzi nimi a umiestniť ich na novú hlavnú stránku webového sídla. Ako príklad autori [28] uvádzajú webové sídlo zaoberajúce sa automobilmi. Autor tohto sídla sa rozhodol usporiadať stránky podľa výrobcov automobilov tak, že modely každého výrobcu sú prezentované na samostatnej stránke. Používatelia však môžu mať záujem o porovnanie určitého typu vozidiel, napr. rodinných automobilov, od viacerých výrobcov. V takomto prípade by museli navštíviť stránky všetkých výrobcov a hľadať na nich rodinné automobily. Metóda generovania indexových stránok takého správanie rozpozná a odporučí vytvoriť stránku s odkazmi priamo na jednotlivé rodinné automobily každého výrobcu.

Kapitola 3

Získavanie informácií o správaní sa používateľa

Be a yardstick of quality. Some people aren't used to an environment where excellence is expected.

Steve Jobs

Pre zmysluplné prispôsobovanie štruktúry webovej stránky potrebujeme poznať čo možno najviac informácií o používateľovi, jeho záujmoch a návykoch pri prezeraní stránok. Z nich vytvárame model používateľa, pomocou ktorého následne vykonávame prispôsobovanie štruktúry webu.

3.1 Zber informácií o používateľov

V tejto časti opisujeme niekoľko najvýznamnejších metód zberu týchto informácií.

Jednotlivé prístupy sa odlišujú najmä miestom, kde sa informácie získavajú. Podľa tohto kritéria rozlišujeme:

- zber údajov na strane klienta a
- zber údajov na strane servera.

V [18] definovali podrobnejšie delenie miesta zberu údajov o používaní webu ako:

- *Zber údajov na strane servera* — tieto údaje sa zbierajú najľahšie, pretože ich má pod kontrolou správca webového sídla. Všetky požiadavky na zobrazenie konkrétnej stránky sa uchovávajú v log súboroch spolu s časom požiadavky a zdrojovou stanicou. Veľkou nevýhodou je, že na serveri nemožno rozlíšiť konkrétneho používateľa (za jednou IP adresou môže byť skrytá celá sieť), takisto sa niektoré požiadavky nedajú zachytiť kvôli vyrovnávacej pamäti prehliadača alebo použitiu proxy servera. Nie je tiež možné zistiť čas, ktorý používateľ na stránke strávil.

- *Zachytávanie paketov* — jednotlivé TCP/IP pakety sa zachytávajú ešte pred tým, ako dorazia na webový server. Dá sa tak zistiť viac informácií z hlavičky paketu, na druhej strane je potrebné filtrovať pakety, ktoré neobsahujú HTTP komunikáciu.

- *Zber údajov na strane klienta* — získanie takýchto údajov je podmienené súhlasom používateľa. Dá sa z nich vyčítať kompletná činnosť používateľa na danej stránke (kam klikal, koľko času strávil na stránke, atď.).

- *Zber údajov na strane proxy servera* — tieto údaje sa zbierajú medzi klientskou stranou a serverovou stranou. Sú vhodné na sledovanie skupinového správania používateľov, ktorí používajú rovnaký proxy server. Aj v tomto prípade musí byť ich zber umožnený so súhlasom používateľa (tým, že používateľ používa pripojenie cez proxy server).

- *Zber údajov na aplikačnej vrstve* — prebieha prostredníctvom záznamov (log súborov) vytváraných webovou aplikáciou. Údaje v nich dokážu opísať kompletnú interakciu používateľa so službou.

Najzaujímavejšie sa javí posledný typ, t.j. zber údajov na aplikačnej vrstve. Aplikačná vrstva má úplnú informáciu o prezeranom obsahu, a tak dokáže najlepšie zachytiť interakciu používateľa so stránkou. Údaje z nej získané sú vo forme tzv. postupnosti kliknutí (angl. *clickstream*), čo je postupnosť nasledovaných odkazov usporiadaná podľa poradia ich použitia. Nevýhodou je skutočnosť, že každá webová aplikácia musí implementovať vlastnú metódu zberu údajov. Údaje z rôznych webových služieb môžu byť v rozličných formátoch a nemusia sa dať vzájomne porovnať.

Zaujímavým prístupom k zberu údajov je metóda realizovaná v nástroji SemanticLog z projektu NAZOU [2]. Metóda je založená na použití webovej služby, ktorá na serveri prijíma údaje o udalostiach vykonaných v sledovanom systéme zasielané rozličnými zberačmi údajov pracujúcimi na strane klienta. Nástroje, ktoré lepšie poznajú sledovanú aplikáciu, tak zaznamenávajú udalosti, ktoré v nej nastali, a posielajú ich webovej službe. Udalosti z viacerých aplikácií sa tak zbierajú na jenom mieste. Navyše, k udalostiam sa pridáva aj sémantika, ktorá je následne využiteľná pri analýze charakteristík používateľa. Udalosť je reprezentovaná pomocou ontológie udalostí, ktorá obsahuje atribúty udalosti a je počítačovo spracovateľná.

Podobný prístup zvolili aj autori [20]. Zber údajov o používaní adaptívnych systémov zabezpečuje obaľovač (angl. *wrapper*), ktorý treba implementovať pre každý sledovaný systém. Obaľovač následne poskytuje štandardizované údaje pre metódu prispôsobovania a odporúčania obsahu.

Pri zbere údajov na strane klienta sa ponúka otázka, akým spôsobom tento zber technicky realizovať. Údaje sa najčastejšie zbierajú:

- pomocou tzv. koláčikov (angl. *cookies*), ktoré do počítača používateľa ukladá navštívená webová stránka,

- vložením skriptu monitorujúceho aktivity do zdrojového textu stránky alebo

- vytvorením prídavného modulu do prehliadača, ktorý zber realizuje.

Z uvedených techník sa zdá najviac vyhovujúce vloženie skriptu monitorujúceho aktivity do zdrojového textu stránky. Tento spôsob je nezávislý od použitého prehliadača a neukladá údaje u používateľa, ktorý by ich mohol zmeniť. Navyše využíva výhody zberu údajov na strane klienta, kedy sme schopní zaznamenať interakciu používateľa na stránke (napr. pohyb kurzoru myši), čo na strane servera nedokážeme.

3.2 Model používateľa a spätná väzba

Model používateľa sa vytvára na základe zozbieraných údajov o používateľovi. Problém je, keď k systému pristupuje nový používateľ, o ktorom nič nevieme. V práci [7] navrhujú viacero postupov pre naplnenie modelu používateľa. Pre autora webového sídla je najjednoduchšou možnosťou nechať používateľa vyplniť dotazník. Tento prístup používa mnoho adaptívnych webových systémov. Teoreticky takto môžeme získať ľubovoľné údaje. Ak používateľ odpovie pravdivo, získame o ňom najpresnejšie údaje.

Problém je, že používateľom sa často nechce pred návštevou stránky vypĺňať dotazníky o ich záujmoch a skúsenostiach, takisto môžu niektorí vyplniť dotazník nepravdivo. Spätnú väzbu môžeme tiež skúmať umiestnením stupnice, na ktorej používateľa necháme ohodnotiť zaujímavosť predloženej stránky. Aj tu je však nutná motivácia, aby nám používateľ explicitnú spätnú väzbu poskytol. Je tiež možné, že sa používateľ nebude vedieť správne ohodnotiť. Ak sa ho spýtame na mieru jeho znalosti v určitej oblasti, môže svoje schopnosti podceniť alebo aj preceniť.

Naopak, implicitná spätná väzba sa získava automaticky, bez zásahu návštevníka webového sídla. Daňou za to je jej menšia presnosť vo vyjadrení záujmu používateľa o zobrazené informácie. Záujem sa dá pomocou implicitnej spätnej väzby odhadovať na rôznych úrovniach.

Veľa nám môže napovedať analýza predložených dokumentov. Keďže vieme, aký obsah sa nachádza na stránke, ktorú používateľovi ponúkneme, môžeme tak zistiť jeho oblasti záujmu. Ak na stránke strávi dlhší čas alebo si prezerá podobné stránky, potom ho pravdepodobne daná téma zaujala.

Záujem vieme odhadnúť aj sledovaním akcií, ktoré používateľ na stránke vykonáva. Akcie ako tlač stránky, pridanie stránky do obľúbených položiek, kliknutie na odkaz, skopírovanie textu do schránky predstavujú pozitívny záujem. Naopak, zastavenie načítavania stránky prípadne strávenie podpriemerne krátkeho času jej čítaním vyjadrujú negatívny záujem [34].

Pri webových novinách stačí, že používateľ klikne na nejaký článok [14]. Odkazy na články spravidla tvoria ich nadpisy, pod nimi sa nachádza krátky úvod z článku. To používateľovi stačí na rozhodnutie, či ho daná téma zaujíma. Ak áno, klikne na nadpis a číta ďalej, čím vyjadruje svoj záujem. Ak článok používateľa zaujme, strávi jeho čítaním viac času, ako v prípade, že by ho nezaujal [25]. V prvom prípade je tiež veľká šanca, že si prečíta aj súvisiace články, prípadne hľadá viac informácií o téme. Vo všeobecnosti však nemôžeme odhadovať záujem používateľa len z toho, že na odkaz klikol. Väčšina odkazov totiž nie je tak dobre anotovaná ako v prípade webových novín.

Dá sa tiež sledovať, ako často sa používateľ vracia na dané webové sídlo, či využil možnosť zaregistrovať sa alebo či si prostredníctvom sídla niečo kúpil (v prípade internetových

obchodov). Model používateľa môžeme následne použiť pri filtrovaní a dopĺňaní zobrazených informácií. Ak vieme, do akej miery používateľa predložené informácie zaujali, môžeme mu na ich základe odporučiť podobné stránky, prípadne môžeme ním navštívené stránky odporučiť podobným používateľom.

Tradične sa používal jeden model používateľa. V súčasnosti sa preferujú prístupy využívajúce viacero modelov používateľa. Jednou z možností je rozdeliť model používateľa na dlhodobý a krátkodobý. Zatiaľ čo dlhodobý odráža stabilnejšie preferencie používateľa, jeho vlastnosti a všeobecné zameranie, krátkodobý model odráža aktuálne preferencie, ktoré sa môžu meniť s každou úlohou, ktorú používateľ rieši [1, 30].

Druhý pohľad na model používateľa je podľa kontextu. Používateľ má viacero modelov podľa toho, čo práve vykonáva a akú rolu pri tom zastáva. Príkladom môže byť rozdelenie modelu na pracovný, študijný a voľnočasový. Keď používateľ pracuje, využíva a analyzuje sa časť modelu zaznamenávajúca prácu. Používateľovi odporúčame dokumenty súvisiace s náplňou jeho práce. Naopak, keď sa používateľ vzdeláva, odporúčame mu výučbové materiály.

Svoj vlastný prístup k zberu údajov majú prevádzkovatelia veľkých portálov. Používateľom ponúkajú stiahnutie a inštaláciu svojho vlastného panelu nástrojov (angl. *toolbar*), ktorý sa integruje s ich internetovým prehliadačom. Vďaka nemu má používateľ rýchly prístup k rôznym častiam portálu a nemusí sa neustále prihlasovať do jeho jednotlivých služieb. Navyše, portál takto môže zbierať údaje o jeho návykoch, ktoré používa pri personalizácii [21].

3.3 Analýza postupnosti odkazov

O návštevníkovi stránky a jeho správaní sa nám veľa napovie postupnosť odkazov, ktoré použil pri navigácii. V takýchto postupnostiach sa dajú hľadať opakujúce sa vzory, ktoré odhalia príslušnosť návštevníka k určitej skupine, prípadne aktuálny zámer návštevníka (či hľadá konkrétnu informáciu alebo sa rozhliada po celom webovom sídle). Ako zdroj dát sa používajú údaje v postupnosti kliknutí. Najobľúbenejšou metódou analýzy správania sa používateľa pracujúcou s týmito údajmi je dolovanie používania webu (angl. *Web Usage Mining*) [32]. Výsledky analýzy môžu byť použité na zlepšenie služby, vytvorenie adaptívneho webového sídla, či stránky priamo prispôsobenej konkrétnemu návštevníkovi.

Dolovanie používania webu má tri základné fázy [5]. Prvou fázou je predspracovanie údajov zozbieraných z postupnosti kliknutí. V druhej fáze, rozpoznávanie vzorov, sa v týchto údajoch rozličnými metódami hľadajú vzory navigácie. Medzi tieto metódy patrí napr. štatistická analýza alebo použitie asociačných pravidiel. Poslednou fázou dolovania je analýza vzorov. V tejto fáze sa nájdené vzory vyhodnocujú, napr. sa podľa prevládajúcich vzorov určuje stereotyp používateľa.

Už pri vzniku webu boli opísané a kategorizované typy sledov, ktoré sa používajú v hypertexte. Podľa [11] sú štyrmi základnými sledmi:

- cesta (angl. *path*) — sled, v ktorom sa žiaden uzol neopakuje dvakrát,
- kruh (angl. *ring*) — sled, ktorý začína aj končí v tom istom uzle,

- slučka (angl. *loop*) — sled, ktorý prechádza už navštíveným uzlom a
- hrot (angl. *spike*) — sled, ktorý sa vracia späť po tej istej trase.

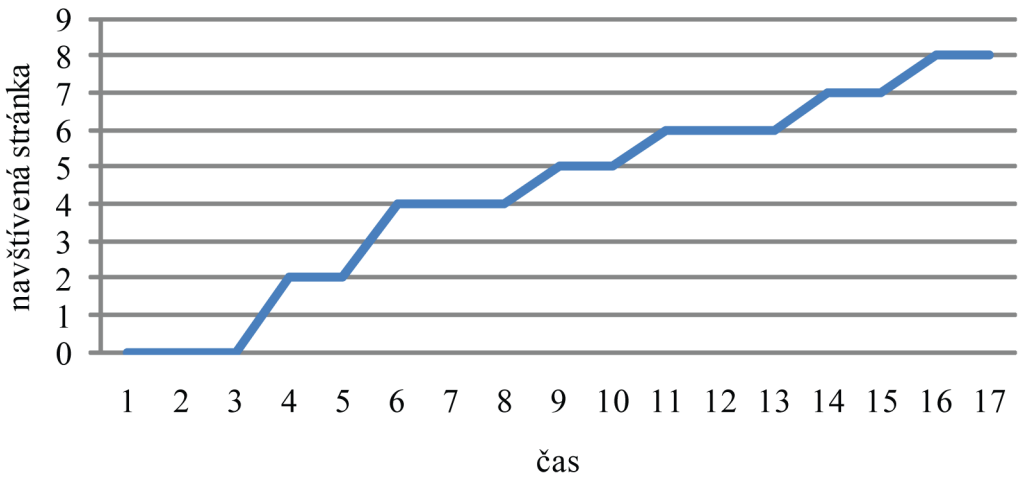
Na základe týchto elementárnych sledov autori následne definovali vzory v navigácii opisujúce rozličné stratégie pri prehliadaní webu. Tieto stratégie sú opísané v tabuľke 3.1.

Tabuľka 3.1: Prejavy rôznych stratégií navigácie [11].

Stratégia	Pravdepodobný cieľ	Použité sledy
skenovanie	pokryť veľkú oblasť, nechodiť do hĺbky	súbor dlhých hrotov a krátkych slučiek
prehliadanie	nasledovať odkazy na stránke až pokým nenarazíme na objekt záujmu	veľa dlhých slučiek, niekoľko rozľahlých kruhov
hľadanie	pátrať po špecifickej informácii na stránke	predlžujúce sa hroty a niekoľko slučiek
objavovanie	prezrieť rozsah a povahu stránky	zmes rôznych sledov
potulovanie sa	pohyb po stránke neorganizovaným spôsobom	veľa stredne veľkých kruhov

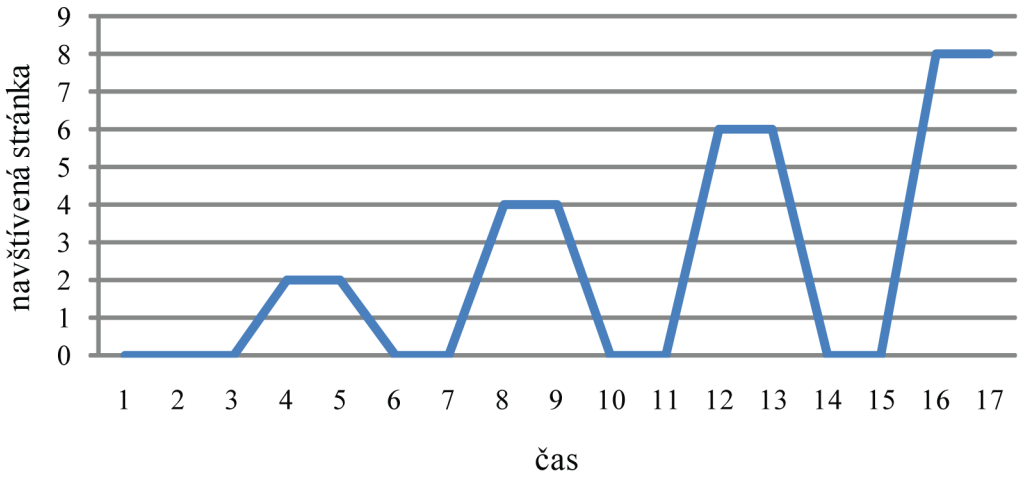
Údaje z usporiadanej postupnosti použitých odkazov môžeme vizualizovať pomocou tzv. grafu krokov (angl. *footstep graph*), čo je dvojrozmerný graf. Os x reprezentuje čas medzi návštevou dvoch po sebe nasledujúcich stránok, os y reprezentuje webovú stránku v slede používateľom navštívených stránok. V takomto grafe sa dajú hľadať vzory v navigácii. Autori [12] ich rozdelili do troch typov: vzor *schody*, vzor *prsty* a vzor *pohorie*.

Vzor schody (angl. *stairs*) opisuje situáciu, kedy sa používateľ vnára hlbšie do webového sídla, pričom sa nevracia. Toto správanie je obvyklé pre nových používateľov, ktorí objavujú, čo všetko dané sídlo ponúka. Graf tohto vzoru je zobrazený na obrázku 3.1.



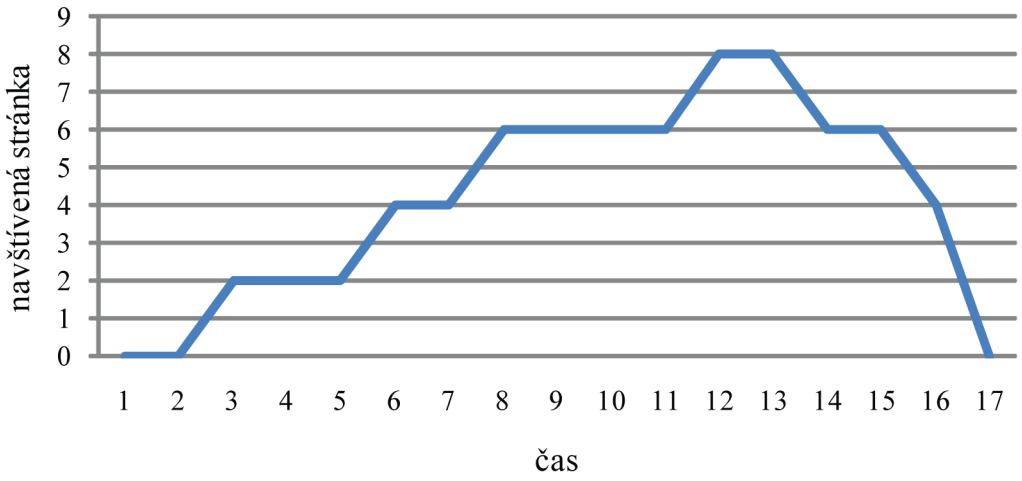
Obr. 3.1: Graf vzoru schody, prevzatý z [12].

Vzor prsty (angl. *fingers*) opisuje situáciu, kedy sa používateľ po návšteve nejakej stránky ihneď vráti na predošlú. Takéto správanie je charakteristické pre pravidelných návštevníkov, ktorí vedia, kde majú hľadať požadovanú informáciu, a idú priamo za ňou. Ak ju nenájdu, ihneď sa vrátia a skúsia inú cestu. Graf tohto vzoru je zobrazený na obrázku 3.2.



Obr. 3.2: Graf vzoru prsty, prevzatý z [12].

Vzor pohorie (angl. *mountain*) opisuje situáciu, kedy používateľ prejde sériou niekoľkých stránok a potom sa postupne vracia. Tento vzor predstavuje kombináciu medzi predchádzajúcimi dvomi. Graf tohto vzoru je zobrazený na obrázku 3.3.



Obr. 3.3: Graf vzoru pohorie, prevzatý z [12].

Vzor schody a vzor pohorie sú podobné v tom, že vzor schody je obsiahnutý v druhom menovanom. Avšak pri vzore schody sa používateľ dostane na nejakú stránku dovnútra webového sídla a na nej ukončí svoju návštevu (zatvorí okno alebo odíde na inú doménu). Na druhej strane, pri vzore pohorie sa používateľ pomocou vzoru schody dostane dovnútra webového sídla, no následne sa vracia späť použitím navigácie na webovom sídle

alebo tlačidla *späť* vo webovom prehliadači. Keď sa dostane na úvod, prípadne na nejakú križovatku, začne sa opäť vnárať, čím sa celý vzor opakuje.

Údaje z postupností kliknutí musíme vhodne vyčistiť, aby sme v nich mohli hľadať navigačné vzory. Pred analýzou je dôležité odfiltrovať prístupy robotov. Sú to systémy, ktoré automaticky prehľadávajú web (napr. kvôli indexácii stránok pre internetový vyhľadávač). Vlastnosti, podľa ktorých možno rozpoznať prístup robota, sú najmä [8]:

- opakovaná požiadavka na tú istú URL adresu z rovnakého zdroja,
- krátky časový interval medzi dvomi bezprostredne nasledujúcimi požiadavkami a
- séria požiadaviek z jedného zdroja, pričom všetky majú prázdnu hodnotu odkazujúcej stránky (tzv. *referer* hlavička v HTTP požiadavke).

3.4 Štandardizované formáty záznamov

Údaje o jednotlivých prístupoch používateľov k webovému sídlu sa na serveri ukladajú vo forme záznamov (tzv. log súborov). Sú to súbory uchovávajúce rozličné informácie o reláciách medzi klientom a serverom. Formáty týchto súborov sa štandardizovali, aby bolo možné súbory použiť v rôznych analytických nástrojoch. Hlavné dva formáty sú Common Log Format a Combined Log Format. Prvý menovaný má nasledovnú štruktúru:

- IP adresa klienta,
- identita klienta (tento atribút sa však príliš nepoužíva),
- ID používateľa (najčastejšie jeho meno),
- čas ukončenia spracovania požiadavky serverom,
- text požiadavky,
- stavový kód odoslaný naspäť klientovi a
- veľkosť odpovede v bajtoch.

Combined Log Format obsahuje všetky atribúty ako predošlý formát a pridáva k nim dva navyše:

- URL odkazujúcej stránky a
- hlavička User Agent protokolu HTTP.

Uvedené dva formáty sa bežne používajú na zaznamenávanie komunikácie na strane servera. Na druhej strane, štandardy pre zaznamenávanie komunikácie na strane klienta prakticky neexistujú, nakoľko takéto sledovanie nie je zatiaľ veľmi rozšírené.

Ohraničenie zberu informácií o používateľovi na strane klienta tkvie v tom, že používateľ musí tento zber povoliť. Skúsenejší používateľ môže ľahko zablokovať odosielanie informácií na server, čím však znemožní prispôsobovanie systému. Na strane klienta získavame viac informácií, ako by sme získali z monitorovania na strane servera. Používateľ

je tiež motivovaný povoliť zbieranie údajov prísľubom prispôsobeného webového portálu, čo je pre neho pozitívne.

Pri získavaní spätnej väzby je problémom automatizované určenie miery záujmu používateľa o zobrazenú stránku. Keď sa používateľa na spätnú väzbu nechceme priamo pýtať, musíme jeho záujem odvodiť z akcií, ktoré na stránke vykonal. Problémom môže byť aj meranie času, ktorý na stránke strávil. Nevieme totiž povedať, či počas tohto času aktívne čítal text stránky alebo bol mimo svojho počítača a stránku jednoducho nechal zobrazenú na monitore. Pri návrhu metódy sledujúcej správanie sa používateľa treba brať všetky tieto ohraničenia a problémy do úvahy.

Kapitola 4

Existujúce adaptívne webové systémy

The Internet is becoming the town square for the global village of tomorrow.

Bill Gates

V tejto kapitole uvádzame prehľad niekoľkých systémov využívajúcich metódy na vylepšovanie štruktúry webových stránok, ktoré sme opísali v kapitole 2.

Pokiaľ sledujeme postupnosť navštívených odkazov, dajú sa v nej nájsť vzory. Na ich objavovanie sa zameriavajú rozličné systémy, ktoré môžeme rozdeliť do dvoch kategórií [29]:

- dolovače postupností navštívených stránok (angl. *sequence miners*) a
- dolovače vo webových záznamoch (angl. *web log miners*).

Prvé menované hľadajú často sa opakujúce postupnosti stránok. Z toho vieme usúdiť, ktoré cesty návštevníci webu najčastejšie volia. Dolovače postupností navštívených stránok však týmto postupnostiam nerozumejú, nevedia odlíšiť triviálnu postupnosť (napr. prechod z hlavnej stránky na stránku s kontaktmi) od užitočnej. Dolovače vo webových záznamoch to síce tiež automaticky nedokážu, ale umožňujú definovať vzory, podľa ktorých následne vyhľadávajú zaujímavé postupnosti použitých odkazov (navštívených stránok).

Systémy a webové sídla, ktoré modifikujú svoj obsah podľa modelu používateľa, môžeme rozdeliť na tri typy [24]:

- systémy s manuálnym definovaním pravidiel (angl. *manual decision rule systems*),
- systémy kolaboratívneho odporúčania (angl. *collaborative filtering systems*) a
- systémy odporúčania založené na analýze obsahu (angl. *content-based filtering systems*).

Systémy s manuálnym definovaním pravidiel nechajú správcu webového sídla definovať pravidlá, na základe ktorých sa mení obsah. Jedná sa napr. o zobrazenie jednej z verzií stránky podľa IP adresy návštevníka. To využívajú spravodajské servery, ktoré návštevníkovi z USA zobrazia stránku s domácimi správami, zatiaľ čo návštevníkovi z Európy zobrazia stránku s medzinárodným spravodajstvom (konkrétne tak robí portál CNN[1]).

Systémy kolaboratívneho odporúčania sledujú správanie používateľa. Obsah navštívenej stránky ho nechajú ohodnotiť a odpoveď uložia. Ak dvaja používatelia hodnotia tú istú stránku rovnako, dá sa predpokladať, že budú mať podobné záujmy a názory [28]. Systém teda odporučí prvému používateľovi stránky, ktoré navštívil druhý, a naopak.

Systémy odporúčania založené na analýze obsahu berú do úvahy obsah prezeranej webovej stránky. Ten následne porovnávajú s modelom používateľa a vyhodnocujú podobnosť, najčastejšie pomocou extrakcie kľúčových slov zo stránky a z modelu používateľa. Ak sa stránka v dostatočnej miere zhoduje, je používateľovi odporučená na prezretie. Na takomto princípe funguje systém WebWatcher [19].

4.1 AHA!

AHA! je všestranná adaptívna webová platforma vyvinutá na technickej univerzite v Eindhoven [15]. Táto platforma v sebe zahŕňa tri modely:

- doménový model,
- model prispôsobovania a
- model používateľa.

Model používateľa je založený na konceptoch a atribútoch. Konceptom môže byť webová stránka, ktorú si používateľ zobrazí, a znalosti, ktoré z nej môže získať. Koncepty reprezentujú témy aplikačnej domény, príkladom môžu byť kapitoly knihy, ktorú používateľ študuje v rámci nejakého predmetu. Použité koncepty sa premietajú do modelu používateľa. Podľa neho sa následne adaptujú ďalšie webové stránky.

Na prispôsobovanie sa využívajú metódy skrývania odkazov, anotovania odkazov a podmieneného zobrazovania fragmentov. Systém tiež umožňuje využiť metódu priameho smerovania. Skrývanie a anotovanie odkazov spočíva v menení ich farieb. Farba odkazov, ktoré používateľ nemá vidieť, sa zmení na farbu okolitého textu, čím sa odkaz akoby skryje (ale jeho text ostane viditeľný). Posledná menovaná metóda spočíva vo vyhodnotení miery vhodnosti jednotlivých častí stránky (fragmentov) pre konkrétneho používateľa. S každým fragmentom je spojená vstupná podmienka na jeho zobrazenie. Pre kapitolu v knihe môže byť podmienkou prečítanie predošlých kapitol. Na stránke sa následne zobrazia iba vhodné fragmenty, t.j. tie, ktorých vstupné podmienky sú splnené.

[1] http://www.cnn.com

4.2 WebWatcher

WebWatcher [19] je virtuálnym sprievodcom po stránke školy. Autori si ako vzor zvolili ľudského sprievodcu po múzeu, ktorý chodí s návštevníkmi, podáva im základné informácie o vystavených exponátoch a v prípade potreby je schopný uviesť podrobnejšie informácie a odpovedať na prípadné otázky. Tento systém sa snaží o niečo podobné vo webovom sídle. Hlavnou náplňou jeho činnosti je odporúčanie odkazov. Systém sa učí na základe reakcií používateľa a následne sa snaží svoje odporúčania vylepšiť.

Hoci sa skôr jedná o odporúčanie, obsahuje aj prvky modifikácie samotnej stránky. Systém sa integruje priamo do prehliadanej stránky pridaním menu so svojimi funkciami. Odporúčania realizuje zvýrazňovaním odkazov na stránke. Na žiadosť používateľa tiež dokáže zobraziť stránky podobné práve prehliadanej (porovnaním kľúčových slov), popularitu odkazu (meranú počtom ľudí, ktorí naň klikli), či poslanie oznamu o zmene obsahu vybranej stránky prostredníctvom e-mailu. Nástroj tiež implementuje metódu priameho smerovania (angl. *direct guidance*).

Systém sa dokáže učiť a vylepšovať svoje odporúčania. Na začiatku sedenia vyjadrí používateľ svoj cieľ zadaním kľúčových slov. Všetky odkazy, ktoré používateľ v rámci sedenia použije, sú týmito kľúčovými slovami anotované. Ďalším návštevníkom portálu systém odporúča tie odkazy, ktorých kľúčové slová sa najviac zhodujú s ich aktuálnym cieľom. Použité odkazy sú opäť anotované ďalšími kľúčovými slovami.

Anotácie jednotlivých odkazov, ako aj používateľov cieľ, sú reprezentované vektorom, ktorého zložky sú tvorené váhami zadaných kľúčových slov použitím metódy TF-IDF. Podobnosť odkazov s cieľom používateľa systém určuje porovnaním príslušných vektorov kosínusovým porovnaním.

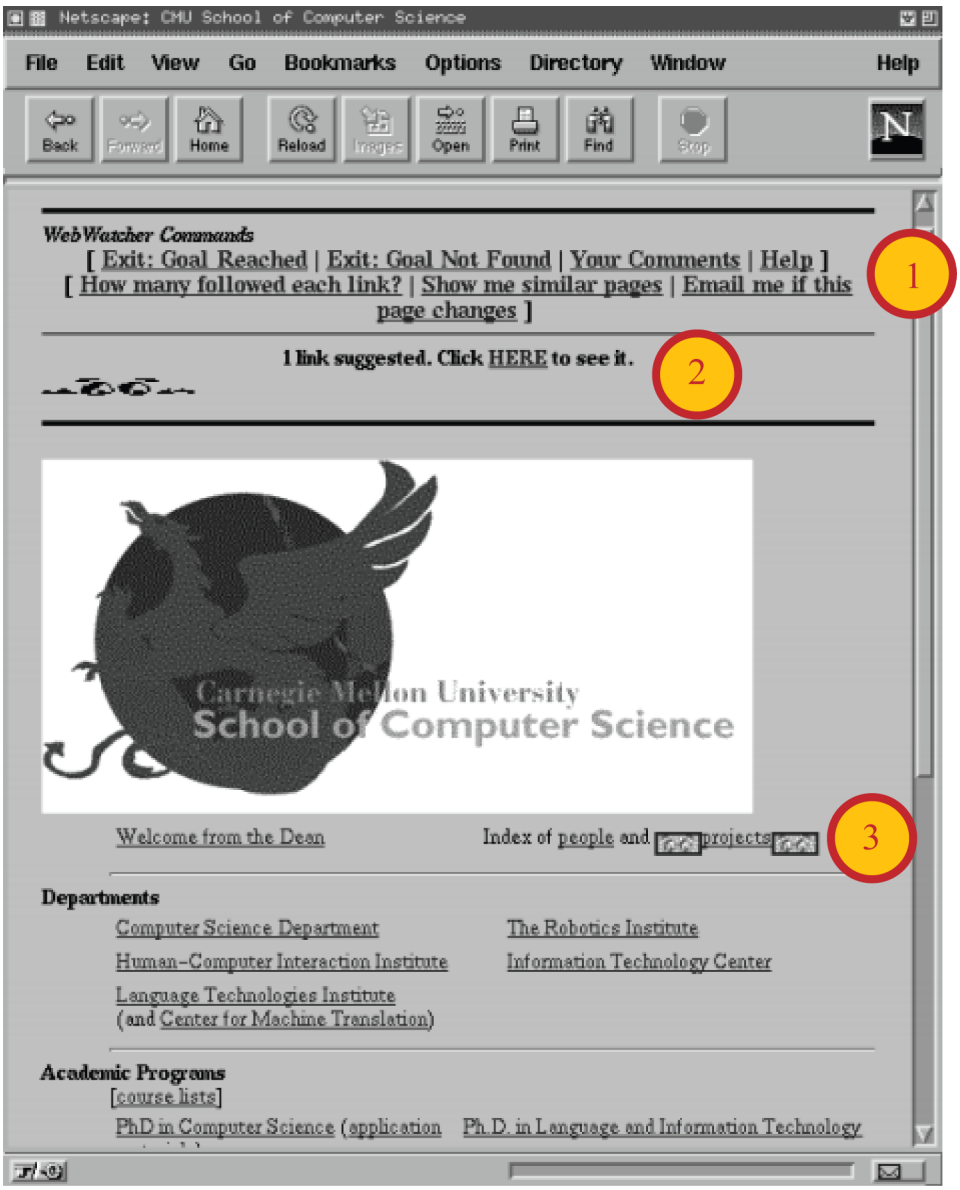
Pri priamom smerovaní systém vedie návštevníka cez webový portál poskytujúc mu vždy jeden odkaz, ktorý má použiť. Cestu vo webovom sídle hľadá s cieľom maximalizovať množstvo používateľom získaných informácií. V rámci predspracovania vypočíta hodnotu TF-IDF pre každé slovo na jednotlivých stránkach portálu. Návštevníka smeruje cestou, na ktorej majú kľúčové slová opisujúce jeho cieľ hodnotu TF-IDF maximálnu. Používateľ môže na konci práce uviesť, či dosiahol alebo nedosiahol svoj cieľ.

Z technologického hľadiska systém funguje prostredníctvom proxy servera. Všetky odkazy presmeruje na seba a do stránok, ktoré si používateľ vyžiadal, pridá menu so svojimi funkciami. Obrazovka s pridaným menu je zobrazená na obrázku 4.1. Používateľ má možnosť prácu so systémom kedykoľvek ukončiť.

Možný problém systému vidíme v inicializácii, pri ktorej musí používateľ vyjadriť svoj cieľ zadaním vhodných kľúčových slov. Vyplnenie dotazníka je síce najspoľahlivejšia metóda získania údajov do modelu používateľa, avšak používatelia nevypĺňajú údaje o sebe radi [7]. Navyše, presnosť odporúčaní závisí od vhodného zadania kľúčových slov.

4.3 IndexFinder

V práci [28] využívajú autori metódu prispôsobovania navigácie generovaním index stránok. Medzi často navštevovanými stránkami hľadajú súvis a odkazy na ne umiestňujú na novo vygenerovanú stránku, prezentovanú návštevníkovi. Generovanie takejto stránky je



Obr. 4.1: Obrazovka systému WebWatcher [19]. 1 – menu s funkciami systému pridané do stránky, 2 – odkaz na priame smerovania, 3 – zvýraznený odporúčaný odkaz.

rozdelené do troch modulov. Modul na spracovanie logov analyzuje log súbory na webovom serveri a počíta počet výskytov danej stránky v jednotlivých návštevách používateľov. Na rozpoznávanie vzorov sa používa metóda štatistickej analýzy. Modul dolovania zhlukov (angl. *cluster mining module*) berie výstupy predošlého modulu spolu s grafom webového sídla a hľadá zhluky často sa opakujúcich výskytov stránok podľa toho, ktoré stránky používateľ v rámci jedného sedenia navštívil. Tieto stránky zoskupí do zhluku. Potom vyhodnocuje, ktoré skupiny stránok sa najčastejšie opakujú v zhlukoch. Modul konceptuálneho zhlukovania (angl. *conceptual clustering module*) používa konceptuálny opis stránok webového sídla a prevádza najčastejšie navštevované zhluky stránok na kandidátske index stránky. Konečné rozhodnutie o použití tej-ktorej index stránky urobí správca sídla.

4.4 Suggest

Systém Suggest sa zameriava na generovanie odkazov, ktoré by mohli návštevníka stránky zaujímať [4]. Systém je realizovaný ako zásuvný modul do webového servera. Pomocou dočasných súborov vytváraných webovou stránkou (tzv. koláčikov, angl. *cookies*) sa identifikujú jednotlivé sedenia. Uchovávajú sa URL adresa aktuálnej stránky a URL adresa stránky, z ktorej používateľ prišiel. Tieto dve adresy sú na strane servera použité na vytvorenie bázy znalostí o spoločne navštevovaných stránkach, ktorá je reprezentovaná neorientovaným grafom. Z bázy znalostí sa extrahujú zhluky stránok, ktoré sú často navštevované spoločne v jednom sedení. Vyberá sa ten zhluk, ktorého množina stránok má najväčší prienik s množinou stránok aktuálnej relácie. Stránky obsiahnuté v rozdieli týchto dvoch množín sa odporučia používateľovi.

Autori dbali pri tvorbe systému na zachovanie súkromia používateľov, stránky nimi navštívené sa neprenášajú na server. Báza znalostí predstavuje agregované informácie o prístupoch všetkých používateľov k stránkam webového sídla, z jednotlivých zhlukov nie je možné zistiť, či používateľ danú stránku už v minulosti navštívil. Toto môže byť nevýhoda, systém môže používateľovi odporučiť tú istú stránku viackrát.

4.5 Kalpana

Kalpana [3] je systém umožňujúci personalizáciu webu na strane klienta. Jeho architektúra sa skladá zo štyroch častí:

- poskytovatelia obsahu, ktorí sú hlavným zdrojom informácií,
- agregátor na strane klienta, ktorý uchováva osobné údaje o používateľovi,
- webové stránky, z ktorých je poskladaný výsledný dokument, a
- platforma prehliadača, ktorá takto poskladaný dokument zobrazí používateľovi.

Najdôležitejším prvkom sú poskytovatelia obsahu. Tí musia mať svoje stránky obohatené o tzv. koncové body pre jazyk SPARQL. SPARQL je dopytovací jazyk pre dáta uložené vo formáte RDF, vhodnom pre reprezentáciu údajov, metaúdajov a vzťahov medzi údajmi pre použitie na webe. Vďaka SAPRQL jazyku je možné automaticky získavať údaje od ich poskytovateľov, ktorými môžu byť rôzne on-line encyklopédie, stránky vládnych agentúr, apod. Agregátor na strane klienta taktiež poskytuje SPARQL rozhranie na prístup k osobným informáciám používateľa. Tie sa získavajú napr. z kalendára používateľa či zo sledovania jeho interakcie s webom.

Webové stránky sú vďaka SPARQL dopytom obohatené o metaúdaje získané od ich poskytovateľov, ako ukazuje obrázok 4.2. Zobrazovacia platforma je implementovaná ako skript pre prehliadač Firefox. Ten funguje takto. Ak si používateľ nechá zobraziť stránku obsahujúcu SPARQL rozhranie, skript tieto značky v zdrojovom texte stránky rozpozná a interpretuje ich. Potom vykoná dopyty na poskytovateľov obsahu (vrátane modelu používateľa). Dopyty sú vykonávané asynchrónne na pozadí, pôvodná stránka sa zobrazí

Obr. 4.2: Obrazovka systému Kalpana [3]. 1 – hlavný obsah pôvodne prezeranej stránky, 2 – doplňujúce informácie získané z externých zdrojov, 3 – informácie získané z bázy znalostí o používateľovi (jeho kalendár aociálna sieť).

bez oneskorenia. Po spracovaní dopytov sa stránka doplní o ich výsledky, ktoré systém Kalpana transformuje do HTML podoby.

Systém Kalpana je veľmi zaujímavý z viacerých hľadísk. Používateľovi poskytuje úplne nový zážitok z prehliadania webu, keďže mu dopĺňa zaujímavé informácie k zobrazovaným stránkam. Ten tak nemusí po týchto informáciách dodatočne pátrať. Takisto jeho implementácia je jednoduchá, systém nevyžaduje inštaláciu nového prehliadača, len rozširuje možnosti populárneho prehliadača Firefox. Takéto rozšírenie by sa dalo implementovať aj pre ďalšie používané prehliadače. Najväčší problém vidíme na strane poskytovateľov obsahu. Aby bol úžitok z používania systému najvyšší, poskytovatelia obsahu by mali údaje uložiť vo forme vhodnej pre web 2.0 napr. prostredníctvom ontológií. Tie zahŕňajú nielen samotné údaje o entitách, ale aj vzťahy medzi nimi.

4.6 Systém na odporúčanie informácií na pozadí

Tento systém (nenašli sme jeho pomenovanie) slúži na automatické personalizované odporúčanie informácií a súvisiaceho obsahu na portáloch [27]. Dokáže dopĺňať obsah stránky poznámkami nad jednotlivými slovami textu, pripájať k objektom informácie od rôznych poskytovateľov obsahu, ako aj informácie z firemného intranetu. Na analýzu obsahu a identifikáciu entít používa webovú službu Calais[2]. Tej pošle ako vstup webovú stránku. Služba systému vráti dokument s poznámkami a identifikovanými entitami. Systém ďalej

[2]http://www.opencalais.com

umožňuje definovať poskytovateľov obsahu, ktorých previaže s konkrétnymi typmi entít (napr. s geografickými názvami môže byť prepojený poskytovateľ máp).

Systém využíva viacero modelov. Doménový model je reprezentovaný ontológiou danej oblasti (definuje napr. oblasť geografie). Naň nadväzuje model úloh, ktorý definuje rôzne činnosti, ktoré môžu používatelia v danej doméne vykonávať (pre geografiu to bude načítanie mapy). Ten je tiež reprezentovaný ontológiou. Model používateľa sa delí na dve časti. Statická časť obsahuje údaje ako dátum narodenia alebo štátna príslušnosť. Dynamická časť modelu používateľa obsahuje jeho skúsenosti a záujmy vo forme referencií na doménový model. Model personalizácie určuje pravidlá prispôsobovania. Tie majú formu trojíc udalosť - podmienka - akcie. Udalosť je výskyt konceptu v obsahu stránky. Podmienka definuje mieru zhody záujmov používateľa s opisom konceptu. Akcie sú jednotlivé entity modelu úloh.

Dôležitou súčasťou je tiež register služieb, ktorý spája akcie (napr. načítanie mapy) s konkrétnym poskytovateľom takejto webovej služby (napr. Google Maps). Na jednu akciu sa môže mapovať viacero poskytovateľov. Pri výbere konkrétnej služby sa sleduje záujem používateľa a jeho odbornosť v danej oblasti. V našom príklade s mapami by sa laikovi načítala základná mapa zo služby Google Maps, odborníkovi by sa načítala podrobná mapa zo špecializovaného geodetického inštitútu.

Na tomto systéme kladne hodnotíme najmä rozdelenie používateľov podľa ich skúseností a záujmu o aktuálny koncept. Každému typu sa tak môžu pridávať informácie presne podľa jeho odbornosti. Tiež sa nám pozdáva definovanie všeobecných akcií a ich následné spojenie s konkrétnymi poskytovateľmi webových služieb. Pri vzniku novej webovej služby sa tak dá toto mapovanie zmeniť bez nutnosti výrazného zásahu do systému.

4.7 Zhodnotenie predstavených adaptívnych systémov

Opísané adaptívne webové systémy realizujú rôzne metódy personalizácie a vylepšovania webových sídiel. Niektoré vylepšenia majú len formu odporúčaní pre správcu (napr. vygenerovanie rôznych index stránok), ktorý sa následne môže rozhodnúť, či zmenu zakomponuje do webového sídla. Iné systémy priamo modifikujú webovú stránku napr. doplnením o informácie získané z iných zdrojov alebo prispôsobením navigácie pre konkrétneho používateľa. Porovnanie systémov podľa spoločných atribútov uvádzame v tabuľke 4.1.

Tabuľka 4.1: Porovnanie adaptívnych webových systémov podľa spoločných atribútov.

Názov	**Explicitné vyjadrenie cieľa**	**Vylepšenie stránky**	**Systém sa učí**
AHA!	nie	automaticky	áno
WebWatcher	áno	automaticky	áno
IndexFinder	N/A	manuálne	nie
Suggest	nie	automaticky	nie
Kalpana	nie	automaticky	áno
WebWatcher	áno	automaticky	áno

Väčšina systémov reprezentuje zámer používateľa a jeho model prostredníctvom kľúčových slov z navštívených stránok. Podľa toho následne vyhodnocujú podobnosť používa-

teľov. Pri prispôsobovaní obsahu webovej stránky sa v menšej miere zaoberajú správaním sa používateľa na webe. Existujúce metódy sledujúce správanie sa používateľa generujú súhrnné správy, na základe ktorých môžu dizajnéri a tvorcovia upraviť svoje sídla. Tu vidíme priestor na využitie správania sa používateľa na personalizáciu webového sídla, konkrétne jeho navigácie.

V našej práci sa zameriavame na sledovanie správania sa používateľa pri navigácii cez celé webové sídlo a pri prehliadaní konkrétnej stránky. Predpokladáme, že aj takýmto spôsobom môžeme zistiť jeho záujem a následne mu odporučiť zaujímavé odkazy alebo prispôsobiť existujúcu navigáciu. Navyše predpokladáme, že na základe vzorov správania sme schopní objaviť podobných používateľov a robiť medzi nimi odporúčania.

Kapitola 5

Ciele práce

Errors using inadequate data are much less than those using no data at all.

Charles Babbage

V tejto práci sa zaoberáme prispôsobovaním navigácie v uzavretom webovom sídle. Našou prácou chceme prispieť do oblasti budovania webových sídiel prispôsobujúcich sa používateľovi, využívajúc sociálny kontext.

Hlavným cieľom našej práce je

- uľahčiť používateľovi využívanie konkrétneho webového portálu prostredníctvom prispôsobenia vybranej časti navigácie a obohatenia jednotlivých stránok.

Toto realizujeme personalizáciou navigáciou na webovej stránke s využitím kolaboratívneho filtrovania a sociálnych aspektov. Zameriavame sa na väčšie webové sídla, ktoré obsahujú informácie pre rôzne skupiny návštevníkov. Príkladom môže byť webové sídlo univerzity alebo intranetový portál pre zamestnancov. Na dosiahnutie hlavného cieľa treba navrhnúť riešenie pre viaceré úlohy, ktoré sme identifikovali ako:

- sledovať pohyb používateľa po webovom sídle,
- získať údaje z používania webového sídla spolu s identifikovaním používateľov,
- získať implicitnú spätnú väzbu o tom, či navštívená stránka používateľa zaujala,
- vytvoriť model používateľa, ktorý odráža jeho záujmy a dá sa flexibilne upravovať,
- identifikovať podobných používateľov na základe správania sa,
- umožniť platformovo nezávislé používanie navrhnutej metódy.

Preferencie používateľa sa v čase menia, preto treba priebežne upravovať model používateľa a určovať podobnosť medzi návštevníkmi. Na portáli denne pribúdajú nové stránky, prípadne sa upravujú už existujúce. Aby model webového sídla odrážal tieto zmeny, je nutné vykonávať analýzu v pravidelných intervaloch. Navrhnuté riešenie má byť dostatočne všeobecné na to, aby sa dalo používať na rôznych sídlach s primeranou mierou počiatočného nastavenia.

Naše predpoklady a návrh metódy overíme v experimentoch v rámci webového sídla fakulty. Ciele experimentov sú:

- zistiť charakteristiky navštevovania webového sídla (ako často používatelia webové sídlo navštevujú, aký čas strávia na stránke, ktoré stránky sú obľúbené),
- overiť predpoklad, že používatelia sa dajú spájať do skupín podľa prevládajúceho vzoru v navigácii po webovom sídle, pričom stránky odporúčané v rámci skupiny budú pre používateľov zaujímavé,
- overiť prínos a mieru využitia prispôsobenia navigácie webového sídla a obohatenia jeho štruktúry.

Kapitola 6

Metóda odhadovania záujmu o stránku

We can't solve problems by using the same kind of thinking we used when we created them.

Albert Einstein

V tejto kapitole uvádzame navrhnutú metódu, ktorou určujeme záujem používateľa o zobrazenú stránku. Mieru záujmu (resp. nezáujmu) o stránku určujeme porovnaním správania sa aktuálneho návštevníka so správaním sa návštevníkov, ktorí videli stránku pred ním.

6.1 Zaznamenanie správania sa používateľa

Správanie sa používateľa je definované činnosťami vykonávanými pri prehliadaní konkrétneho webového sídla. V rámci jednej návštevy webového sídla ho tvoria tieto dve zložky:

- postupnosť navštívených odkazov v celom webovom sídle a
- akcie vykonané používateľom na každej navštívenej stránke.

Pre určenie záujmu o zobrazenú stránku používame akcie, ktoré na nej používateľ vykonal. Postupnosť navštívených odkazov v celom webovom sídle používame na rozdelenie používateľov do skupín podľa podobnosti. Tento proces opisujeme v kapitole **??**.

6.1.1 Správanie sa používateľa na navštívenej stránke

Na každej zobrazenej stránke sledujeme akcie, ktoré na nej návštevník vykonal. Z nich zisťujeme mieru záujmu používateľa o zobrazenú stránku, aby sme ju mohli ďalej odporučiť podobným používateľom. Vykonané akcie sú teda indikátormi záujmu používateľa o

stránku. Indikátory môžu vyjadrovať kladný alebo záporný záujem o stránku. Indikátory, vyjadrujúce kladný záujem o stránku, sú [34]:

- tlač stránky,
- pridanie stránky medzi obľúbené položky,
- kliknutie na odkaz v rámci stránky a
- skopírovanie textu do schránky.

Negatívny záujem o zobrazenú stránku vyjadrujú tieto indikátory [34]:

- zastavenie načítavania stránky a
- zatvorenie okna s príslušnou stránkou počas jej načítavania.

Ďalej existujú indikátory, ktoré môžu predstavovať ako kladný, tak aj záporný záujem o stránku. To závisí od hodnoty daného indikátora. Takýmito indikátormi sú:

- rolovanie stránky pomocou posuvníkov (angl. *scrolling*) a
- čas strávený na stránke.

Čas strávený na stránke je relatívna veličina. Niekto môže mať celý deň spustený prehliadač, v ktorom má otvorenú webovú stránku, hoci nie je pri počítači. Doba medzi otvorením a zatvorením okna prehliadača teda nie je vhodná na určenie času stráveného na stránke. Tento problém riešime zaznamenávaním času, ktorý používateľ na stránke strávil *aktívne*. Aktivita v našom ponímaní znamená pohyb kurzoru myši a rolovanie stránky. Zaznamenávanie jednotlivých akcií na stránke vykonávame takto:

1. nastav hodnotu *používateľ aktívny* na *false*
2. ak nastal výskyt sledovanej udalosti, uprav príslušný indikátor
3. ak nastala udalosť pohnutia kurzorom myši alebo rolovania stránky kolieskom myši, nastav hodnotu *používateľ aktívny* na *true*
4. vykonávaj v pravidelných intervaloch
   a) ak *používateľ aktívny* = *true* ulož hodnoty indikátorov
   b) *používateľ aktívny* = *false*
   c) vynuluj hodnoty indikátorov

Sledovanou udalosťou môže byť jedna z vyššie spomínaných (napr. skopírovanie textu do schránky). Ak takáto udalosť nastane, upravíme príslušný indikátor (napr. zvýšime počet skopírovaní o 1). Potom sa v pravidelných intervaloch snažíme uložiť hodnoty indikátorov. Robíme tak iba v prípade, že bol používateľ v predposlednom intervale aktívny.

6.2 Analýza záujmu o stránku

Analýzou akcií vykonaných používateľom na konkrétnej stránke určujeme mieru záujmu používateľa o ňu. Čím viac pozitívnych akcií na nej vykonal, tým zaujímavejšia pre neho bude, a naopak. Odkazy na zaujímavé stránky využívame pri vylepšení navigácie pre ostatných používateľov.

V časti 6.1.1 sme opísali rôzne indikátory záujmu o stránku. Kladné indikátory zvyšujú používateľov záujem o stránku, záporné indikátory ho znižujú. Pre indikátory, ktoré môžu znamenať kladný aj záporný záujem (čas strávený na stránke, počet rolovaní stránky) je potrebné určiť, ako budú jednotlivé hodnoty ovplyvňovať celkový odhadovaný záujem o stránku. Na to používame sociálny kontext a porovnávame, ako sa na stránke správajú ostatní používatelia so správaním používateľa, ktorého záujem určujeme. Konkrétne:

- Ak je hodnota indikátora menšia ako je priemer ostatných používateľov, znižujeme odhadovaný záujem o stránku (usudzujeme, že v takom prípade stránka používateľa nezaujala).
- Ak je hodnota indikátora väčšia ako je priemer ostatných používateľov, zvyšujeme odhadovaný záujem (usudzujeme, že v takom prípade stránka používateľa zaujala).

Pre hodnoty indikátorov, ktoré môžu znamenať kladný aj záporný záujem, sme určili tri intervaly:

- nadpriemerná hodnota ($\uparrow$), ak $hodnota > priemer + K\ \%$,
- priemerná hodnota ($-$), ak $hodnota \in < priemer - K\ \%\ ;\ priemer + K\ \% >$,
- podpriemerná hodnota ($\downarrow$), ak $hodnota < priemer - K\%$.

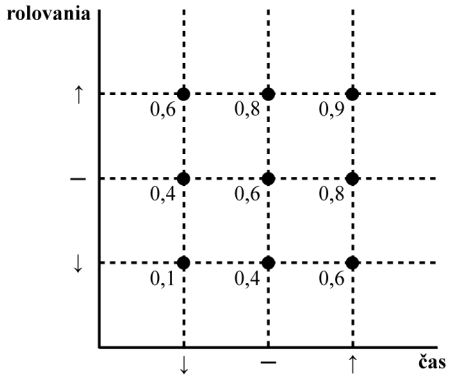
Výber konkrétnych indikátorov záujmu, ktoré sa na odhad použijú, závisí od domény. V našom prípade požadujeme riešenie nezávislé od zvoleného prehliadača. V takom prípade nevieme zistiť výskyt niektorých udalostí, ako je napr. pridanie stránky do zoznamu obľúbených položiek. Sledujeme tieto tri indikátory záujmu, ktoré sa dajú merať nezávisle od zvolenej domény:

- čas strávený na stránke,
- počet rolovaní stránky a
- skopírovanie textu do schránky.

Ako uvádzame vyššie, čas strávený na stránke nemeriame absolútne, zaznamenávame čas strávený aktívne. Tento čas predstavuje počet intervalov, v ktorých bol používateľ na stránke aktívny. Pri ukladaní hodnôt indikátorov zvýšime hodnotu času stráveného na stránke o 1. Čas teda nie je vyjadrený v sekundách, ale v počte periód, v ktorých používateľ hýbal kurzorom myši alebo roloval stránku.

Na obrázku 6.1 uvádzame nami určené hodnoty záujmu pre všetky kombinácie udalostí, ktoré považujeme za prejav záujmu, resp. nezáujmu o stránku. Hodnoty sme najskôr určili podľa nášho predpokladu, ktoré aktivity ako vplývajú na záujem používateľa. Tieto

hodnoty sme následne upravili v experimente, ako opisujeme v kapitole 8. V navrhnutej metóde porovnávame akcie jedného používateľa s akciami, ktoré vykonali na tej istej stránke ostatní. Podľa nich určíme hodnotu záujmu, ako je uvedené na obrázku 6.1. K takto získanej hodnote pripočítame 0,1 v prípade výskytu udalosti skopírovania textu do schránky. V opačnom prípade hodnotu 0,1 odpočítame. Záujem vypočítaný navrhnutou metódou môže byť v intervale <0;1>, pričom hodnota 0 znamená úplný nezáujem a hodnota 1 znamená úplný záujem o zobrazenú stránku.



Obr. 6.1: Určenie záujmu používateľa o zobrazenú stránku.

Po tomto kroku máme vypočítanú mieru záujmu používateľa o ním navštívené stránky. Pre používateľa, ktorý tieto stránky ešte nenavštívil, sa snažíme predpovedať, do akej miery ho budú zaujímať. Na túto predpoveď využívame metódu kolaboratívneho filtrovania, ako ju opísali autori v práci [30]. Metóda sa bežne používa na predpoveď záujmu používateľa o dokumenty na základe kľúčových slov. My ju používame novým spôsobom berúc do úvahy údaje o správaní sa na jednotlivých stránkach. Metódu kolaboratívneho filtrovania sme upravili takto:

- jednotlivé hodnotené prvky tvoria odkazy (a stránky, na ktoré tieto odkazy vedú) a
- ohodnotenie prvkov počítame ako mieru záujmu návštevníka o danú stránku.

Metóda má jedno obmedzenie v tom, že nevieme predpovedať záujem o ľubovoľnú stránku. Predpoveď vieme, prirodzene, urobiť len pre tie stránky, ktoré už navštívili iní používatelia. Na predpoveď záujmu o nenavštívenú stránku využívame mieru podobnosti dvoch používateľov vypočítanú pomocou Pearsonovho korelačného koeficientu. Pearsonov korelačný koeficient je definovaný takto (Sugiyama, 2004):

$$S_{p,u} = \frac{\sum_{i=1}^{I}(r_{p,i} - \bar{r_p}) \times (r_{u,i} - \bar{r_u})}{\sqrt{\sum_{i=1}^{I}(r_{p,i} - \bar{r_p})^2 \times \sum_{i=1}^{I}(r_{u,i} - \bar{r_u})^2}}$$

kde:
$S_{p,u}$ — hodnota Pearsonovho koeficientu medzi používateľmi p a u
$r_{p,i}$ — hodnotenie i-tej položky (odkazu) od používateľa p
$\bar{r_p}$ — priemerné hodnotenie položiek od používateľa p
$r_{u,i}$ — hodnotenie i-tej položky (odkazu) od používateľa u

$\bar{r_u}$ — priemerné hodnotenie položiek od používateľa u
I — celkový počet položiek

Predpoveď záujmu používateľa p o ním nepoužitý odkaz i vypočítame takto:

$$z_{p,i} = \bar{r_p} + \frac{\sum_{u=1}^{N}(r_{u,i} - \bar{r_u}) \times S_{p,u}}{\sum_{u=1}^{N} S_{p,u}}$$

kde:
$z_{p,i}$ — predpoveď záujmu používateľa p pre nenavštívený odkaz i
N — počet najpodobnejších používateľov

Predpoveď záujmu o nenavštívené stránky využívame v metóde prispôsobovania navigácie odporúčaním odkazov (pozri kapitolu 7).

6.3 Diskusia k metóde odhadu záujmu o stránku

V opísanej metóde automatizujeme proces získavania spätnej väzby od používateľa ohľadom jeho záujmu o zobrazenú stránku. Používateľ nemusí vyjadrovať svoj záujem explicitne. Výhoda metódy je v tom, že odstraňuje subjektivitu používateľa pri zadávaní spätnej väzby. Záujem každého používateľa je určený rovnakým postupom.

Hodnota dvoch hlavných indikátorov, ktoré pri určovaní záujmu o stráku používame, závisí od aktivít ostatných používateľov na tejto stránke. Z toho vyplýva aj nevýhoda metódy — nedokážeme určit záujem o stránku pre prvého návštevníka. Jeho záujem vieme určiť len z indikátorov, ktoré nemajú premenlivú hodnotu (ako napr. skopírovanie textu do schránky, čo vždy predstavuje kladný záujem). Odhad záujmu je tým lepší, čím viac používateľov danú stránku navštívi. Extrémne hodnoty jednotlivých indikátorov (príliš nízke alebo príliš vysoké) sú zmiernené vďaka priemerovaniu hodnôt od všetkých návštevníkov.

Kapitola 7

Metóda prispôsobovania navigácie

If you thought that science was certain - well, that is just an error on your part.

Richard Feynman

V tejto kapitole uvádzame navrhnutú metódu, ktorou realizujeme prispôsobovanie navigácie vo webovom sídle na základe modelu správania sa jeho jednotlivých návštevníkov. Model správania sa sme založili na sledovaní správania sa v rámci celého webového sídla, ako aj v rámci konkrétnej navštívenej stránky.

Hlavná myšlienka metódy spočíva v tom, že pre konkrétneho používateľa zlepšujeme štruktúru navigácie vo webovom sídle pridaním odkazov na časti sídla, ktoré by ho mohli zaujímať. Metóda je založená na hypotéze, že podobní používatelia sa pri prezeraní webového sídla správajú podobne. V postupnostiach použitých odkazov identifikujeme vzory navigácie (pozri časť 3.3). Podľa prevažujúcich sledov určíme vzor v navigácii a zaradíme používateľa do skupiny s inými používateľmi navigujúcimi sa podľa tohto vzoru. Do personalizovaného menu vyberáme:

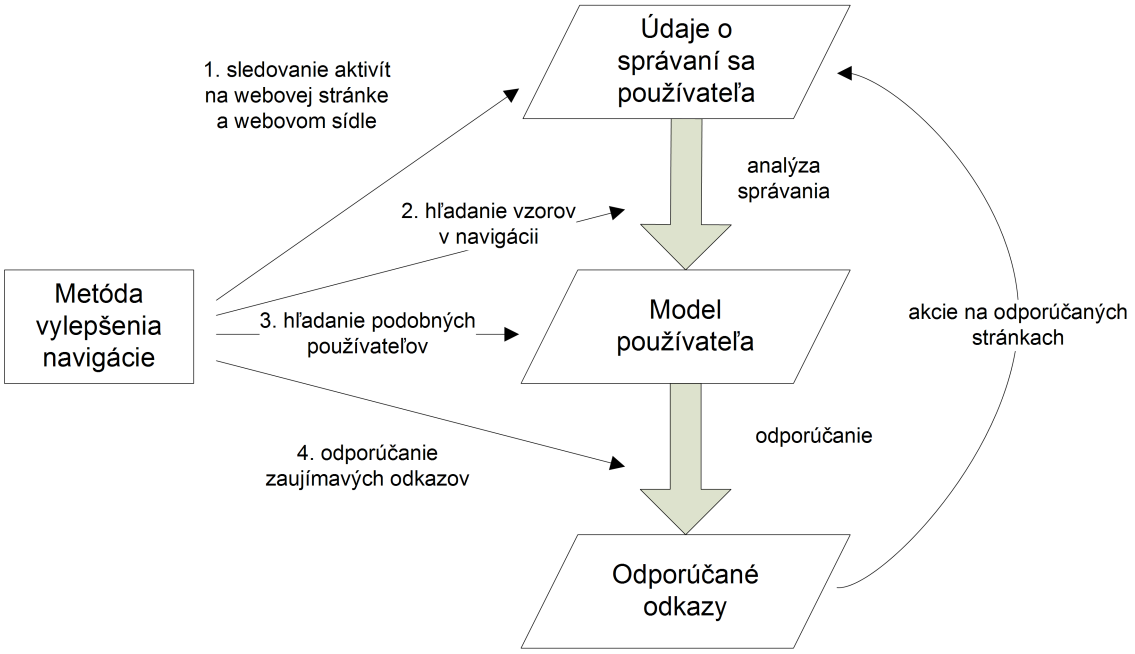
- odkazy na stránky, ktoré používateľ už navštívil a zaujali ho, a
- odkazy na stránky, ktoré ešte nenavštívil, no zaujali jemu podobných používateľov.

Prvým typom odkazov dosiahneme pripomenutie zaujímavej skutočnosti, ktorú mohol používateľ zabudnúť. Príkladom môže byť odkaz na stránku s pozvánkou na podujatie. Ak vyhodnotíme, že táto stránka používateľa zaujala, odkaz na ňu mu odporučíme, aby na podujatie nezabudol.

Druhým typom odkazov odporučíme používateľovi stránky, ktoré zaujali ľudí jemu podobných. Webové sídlo môže byť rozsiahle a zaujímavé stránky môžu byť skryté hlboko v štruktúre jeho navigácie. Ak stránku nejaký používateľ objaví, odporučíme ju ostatným, ktorí by sa k nej inak nemuseli dostať, ale na základe podobnosti ich správania sa predpokladáme, že by mohla byť pre nich zaujímavá.

Metódu vylepšenia navigácie webového sídla (obrázok 7.1) sme navrhli v nasledovných krokoch:

1. zaznamenanie správania sa používateľa na webovej stránke,
2. určenie záujmu o zobrazenú stránku,
3. úprava modelu používateľa, detekcia vzorov v navigácii,
4. nájdenie podobných používateľov na základe správania sa,
5. odporučenie zaujímavých odkazov.



Obr. 7.1: Proces vylepšenia navigácie vo webovom sídle odporúčaním odkazov.

Okrem toho sme navrhli metódu analýzy webového sídla, ktorá sa vykonáva v pravidelných intervaloch nezávisle od vyššie uvedených krokov. Analýza slúži na vytvorenie modelu webového sídla s cieľom získania dodatočných informácií zo stránok, ktoré potom prezentujeme používateľom. Takýmito informáciami sú dátumy a oznamy o nadchádzajúcich udalostiach alebo zmeny na jednotlivých stránkach. Opis metódy je uvedený ďalej v tejto kapitole.

7.1 Zoskupovanie používateľov podľa podobnosti

Zaznamenávame postupnosť všetkých odkazov v rámci webového sídla, ktoré používateľ pri jednej návšteve použil. Každé použitie odkazu predstavuje zobrazenie jednej stránky. Webový prehliadač posiela spolu s požiadavkou na zobrazenie stránky aj adresu referujúcej stránky, z ktorej požiadavka prichádza. Vďaka tomu vieme zrekonštruovať pohyb používateľa po webovom sídle. Ak je adresa referujúcej stránky prázdna, znamená to, že používateľ neklikol na odkaz, ale zadal adresu priamo. Takto vieme identifikovať nové

sedenie (angl. *session*). Pri analýze navštívených odkazov konkrétneho používateľa tieto zoskupujeme do postupností práve podľa sedení, do ktorých patria, podľa algoritmu 7.1.

Algoritmus 7.1 vytvor postupnosti odkazov (používateľ u, doména d)

```
1: zaznamy = načítaj zoznam dvojíc adresa - referujúca adresa pre používateľa p z domény d
2: zaznamy = usporiadaj vzostupne podľa dátumu návštevy
3: sedenia = prázdny zoznam sedení
4: predchadzajuca = null
5: s = nové sedenie
6: for all záznam z in zaznamy do
7:    if predchadzajuca ≠ z.referujuca then
8:       sedenia = sedenia + s
9:       s = nové sedenie
10:   end if
11:   s = pridaj odkaz z.adresa
12:   predchadzajuca = z.adresa
13: end for
14: return sedenia
```

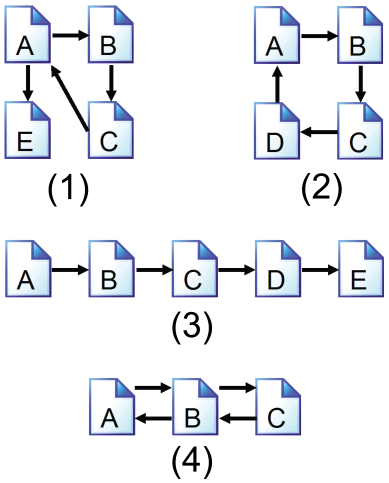
Vytvorenie zoznamu dvojíc adresa – referujúca adresa je krok, ktorý závisí od konkrétneho spôsobu implementácie tohto algoritmu. Iné možnosti zaznamenávania adries máme na strane klienta ako na strane servera.

Vytvorená postupnosť tvorí prúd odkazov (angl. *clickstream*), v ktorom hľadáme vzory. Prúdy odkazov používame na určovanie podobných používateľov, sú teda základným údajovým elementom našej metódy. Spolu s navštíveným odkazom uchovávame jedinečný identifikátor používateľa a čas kliknutia na odkaz.

V každom prúde odkazov (predstavujúcich jedno sedenie) hľadáme opakujúce sa vzory (cesta, kruh, slučka, hrot, pozri časť 3.3). Jednotlivé vzory sú schematicky znázornené na obrázku 7.2. Podľa prevládajúceho vzoru vo všetkých nájdených prúdoch odkazov určujeme, o aký typ používateľa sa jedná a umiestňujeme ho do skupiny zodpovedajúcej tomuto vzoru. Skupín používame 5, jednu pre každý vzor. Piata skupina je pre používateľov, ktorí nemajú ani jeden zo vzorov navigácie dominantný. Rozdelenie používateľov do skupín predstavuje dlhodobý model používateľa. Predpokladáme, že prevažujúci vzor v navigácii konkrétneho používateľa sa pre dané webové sídlo nebude v čase meniť (pokiaľ nedôjde k zmene dizajnu sídla, čo si vyžiada zmenu v návykoch pri navigácii v ňom).

Vzor *cesta* je tvorený postupnosťou odkazov, v ktorej sa ani jeden neopakuje. Toto je elementárny vzor, aj postupnosť dvoch odkazov A a B sa dá považovať za vzor *cesta*. Ak by sme toto pripustili, našli by sme tento vzor aj vo všetkých ostatných vzoroch ako ich súčasť. Tým by sa stal vzor *cesta* zbytočným, pretože by ho obsahovali všetky postupnosti. Z tohto dôvodu považujeme za vzor *cesta* len úplnú postupnosť z jedného sedenia, v ktorej sa ani jeden prvok neopakuje. Hľadáme ho podľa algoritmu 7.2.

Vzor *kruh* je tvorený postupnosťou odkazov, ktorá začína a končí tým istým prvkom. Hľadáme ho podľa algoritmu 7.3. Postupnosť zloženú práve z dvoch prvkov, ktoré majú rovnakú hodnotu, nepovažujeme za vzor *kruh*.



Obr. 7.2: Príklady jednotlivých vzorov. 1 — slučka, 2 — kruh, 3 — cesta, 4 — hrot.

Algoritmus 7.2 obsahuje vzor cesta (prúd odkazov s)

```
1: odkazy = prázdna pomocná množina odkazov
2: for i = 1 to s.length do
3:    if odkazy obsahuje s[i] then
4:       return false
5:    else
6:       odkazy = odkazy + s[i]
7:    end if
8: end for
9: return true
```

Algoritmus 7.3 obsahuje vzor kruh (prúd odkazov s)

```
1: if s[1] == s[s.length] and s.length > 1 then
2:    return true
3: else
4:    return false
5: end if
```

Vzor *slučka* je tvorený postupnosťou odkazov, ktorá obsahuje ten istý odkaz viackrát (pri navigácii po webovom sídle prechádzame už raz navštívenou stránkou). Navyše musí platiť, že postupnosť nie je *kruh* ani *hrot*, a tiež neobsahuje dva zhodné prvky, ktoré po sebe bezprostredne nasledujú. Vzor hľadáme podľa algoritmu 7.4.

Vzor *hrot* je tvorený postupnosťou odkazov, ktorá obsahuje už raz použité odkazy, usporiadené v opačnom poradí. V navigácii po webovom sídle to znamená, že sa vraciame po tej istej ceste späť na niektorú z križovatiek. Vzor hľadáme podľa algoritmu 7.5.

7.2 Odporúčanie odkazov

Pre každého používateľa vytvárame zoznam odporúčaných odkazov. Využívame pri tom jemu podobných používateľov a ich záujem o stránky, ktoré navštívili. Zoznam odporúčaných odkazov ďalej používame pri konkrétnom vylepšovaní navigácie jednotlivých

Algoritmus 7.4 obsahuje vzor slučka (prúd odkazov s)

```
1: odkazy = prázdna pomocná množina odkazov
2: for i = 1 to s.length do
3:   if odkazy obsahuje s[i] and i < s.length then
4:     if i > 1 and s[i] ≠ s[i − 1] then
5:       return true
6:     end if
7:   else
8:     odkazy = odkazy + s[i]
9:   end if
10: end for
11: return false
```

Algoritmus 7.5 obsahuje vzor hrot (prúd odkazov s)

```
1: odkazy = prázdna pomocná množina odkazov
2: for i = 1 to s.length do
3:   if odkazy obsahuje s[i] then
4:     if i > 2 then
5:       if s[i] == s[i − 2] then
6:         return true
7:       end if
8:     end if
9:   else
10:    odkazy = odkazy + s[i]
11:  end if
12: end for
13: return false
```

častí (podľa extrahovaných informácií zo stránky, na ktorú odkaz vedie, ho odporúčame v jednotlivých sekciách webového sídla, pozri časť 7.3). Zoznam odporúčaných odkazov vytvárame podľa algoritmu 7.6.

V prvom kroku algoritmu vyberieme skupinu podobných používateľov k používateľovi, ktorému odporúčame odkazy. Používateľa sme už predtým zaradili do niektorej skupiny podľa objavených vzorov v jeho navigácii. Ako podobných berieme všetkých používateľov z tejto skupiny. Ďalej vypočítame hodnotu Pearsonovho korelačného koeficientu (pozri časť 6.2) medzi aktuálnym používateľom a každým z používateľov z jeho skupiny. Pri tom si zapamätáme, ktoré odkazy používatelia využili. Zo skupiny odkazov odoberieme tie, ktoré už aktuálny používateľ videl. Pre zvyšné odkazy vypočítame jeho predpokladaný záujem (pozri 6.2). Ak je predpokladaný záujem väčší ako konštanta vyjadrujúca hranicu kladného záujmu (v našom prípade 0,5, keďže záujem je číslo z rozsahu <0;1>), odkaz pridáme do zoznamu odporúčaných odkazov.

7.3 Prispôsobovanie navigácie v sekciách sídla

V tejto časti opisujeme konkrétne použitie zoznamu odporúčaných odkazov na prispôsobenie navigácie v rozličných častiach webového sídla. V rámci vylepšenia štruktúry

Algoritmus 7.6 odporuč odkazy (používateľ p)

```
1: skupina = skupina podobných používateľov
2: odkazy = prázdny zoznam odkazov
3: for all používateľ v in skupina do
4:    vypočítaj Pearsonov korelačný koeficient (p, v)
5:    odkazy = pridaj odkazy navštívené používateľom v
6: end for
7: odkazy = odober odkazy navštívené používateľom p
8: for all odkaz l in odkazy do
9:    odhad = vypočítaj predpokladaný záujem p o l
10:   if odhad > LIMIT then
11:      pridaj l do odporúčaných odkazov používateľa p
12:   end if
13: end for
14: z = usporiadaj odporúčané odkazy používateľa p podľa predpovede záujmu
15: return z
```

webového sídla, resp. jeho navigácie, sme identifikovali tri časti, v ktorých uplatňujeme našu metódu:

- personalizovaný kalendár,
- personalizované novinky a
- ďalšie odporúčané stránky

7.3.1 Personalizovaný kalendár

Veľa dokumentov, ktoré na webové sídla umiestňujú univerzity, spoločnosti alebo jednotliví výskumníci, v sebe obsahuje oznam o nadchádzajúcej udalosti. Pod pojmom dokument myslíme článok alebo správu, odkaz na ktorú sa častokrát objavuje na hlavnej stránke, prípadne v sekcii noviniek.

Ak oznam informuje o udalosti, obsahuje aj dátum jej konania. Práve podľa ich prítomnosti identifikujeme udalosti, z ktorých následne zostavujeme používateľov personalizovaný kalendár. V rámci analýzy webového sídla hľadáme na stránkach dátumy, ktoré môžu byť zapísané v rozličných tvaroch (číslom aj slovom, jednoduchý dátum alebo rozpätie dvoch dátumov). Text vyjadrujúci dátum najprv prevádzame na spoločných číselný formát, a potom z neho vytvárame objekt dátum, ktorý ukladáme spolu s odkazom na stránku, na ktorej sa vyskytol. Ak je na stránke dátumov viac, zobrazíme udalosť pri všetkých z nich. Do úvahy neberieme dátumy, ktoré už nastali pred dňom vykonania analýzy. Takto jednoduchým spôsobom vyfiltrujeme napr. dátumy označujúce vytvorenie stránky alebo vykonanie poslednej zmeny na nej, ktoré mnohé redakčné systémy alebo autori do stránok automaticky vkladajú. Na extrakciu dátumov využívame regulárne výrazy.

Vkladanie udalostí do kalendára prispôsobeného pre konkrétneho používateľa prebieha podľa algoritmu 7.7.

Okrem odporúčaných udalostí vkladáme do kalendára aj oznam o udalostiach, o ktorých už používateľ vie. Týmto spôsobom mu pripomíname zaujímavé udalosti, aby na

Algoritmus 7.7 vytvor kalendár (používateľ p)

```
1: odkazy = zoznam odporúčaných odkazov
2: udalosti = zoznam udalostí nájdených pri analýze sídla
3: k = prázdny kalendár
4: for all odkaz l in odkazy do
5:    for all udalosť u in udalosti do
6:       if l == odkaz udalosti (u) then
7:          for all dátum d in dátumy udalosti u do
8:             k = pridaj oznam o u k dátumu d
9:          end for
10:      end if
11:   end for
12: end for
13: return k
```

ne nezabudol. Výber udalostí na pripomenutie prebieha podľa toho istého postupu ako výber odporúčaných udalostí. Rozdiel je v prvom kroku, kedy vyberieme zoznam odkazov na stránky, ktoré používateľa v minulosti zaujali (pozri časť 6.2).

7.3.2 Personalizované novinky

Na webovom sídle často pribúdajú nové oznamy zaujímavé pre návštevníka. Ak však nie sú personalizované, návštevník ich nemusí stíhať sledovať. Môže sa síce prihlásiť na odber najnovších správ, no nie všetky musia byť pre neho zaujímavé. Pri personalizovaných novinkách opäť využívame analýzu správania sa v rámci jednotlivých stránok, čím zisťujeme používateľovu mieru záujmu o ne. Za novinky považujeme dva druhy stránok:

- novo pridané stránky a
- existujúce stránky, ktoré sa zmenili.

Pri analýze webového sídla extrahujeme špeciálne označené sekcie na stránkach, ktoré obsahujú novinky. Každá novinka je spojená s odkazom na stránku, ktorá o nej podrobnejšie informuje. Všetky takto nájdené novinky pridáme do zoznamu noviniek, z ktorých potom odporúčame jednotlivé položky používateľom.

Okrem noviniek zo špeciálne označených sekcií webového sídla považujeme za novinku aj stránku, ktorá sa zmenila. Ak takúto stránku používateľ v minulosti navštívil a zaujala ho, mal by byť informovaní o jej zmene. Keďže sa však nejedná o novo pridanú stránku, štandardnými spôsobmi, ktoré webové sídla využívajú (napr. RSS), sa o nej nedozvie. Preto v rámci analýzy webového sídla porovnávame textový obsah každej stránky s obsahom, ktorý sme si uložili pri poslednej návšteve. Ak sa tento obsah zmení (stačí zmena jedného znaku), pridáme túto stránku do zoznamu noviniek.

V súčasnosti berieme do úvahy všetky zmeny. Takéto riešenie nie je vždy vhodné a jeho výsledkom môže byť, že používateľovi sa zobrazí ako novinka stránka, ktorej podstata sa nezmenila (napr. opravil sa iba preklep v názve). V budúcnosti plánujeme na zisťovanie zmien použiť znalosti o obsahu a type zmeny, ako napríklad [33].

Pri odporúčaní noviniek vychádzame zo zoznamu odporúčaných odkazov (pozri časť 7.2). Ak sa odporúčaný odkaz zhoduje s odkazom niektorej z noviniek, pridáme ju do zoznamu odporúčaných noviniek pre daného používateľa. Postup je podobný ako pri tvorbe personalizovaného kalendára.

7.3.3 Ďalšie odporúčané stránky

Ak stránka zaujme používateľov, je vhodné odporučiť ju ďalším, ktorých má tiež potenciál zaujať (podľa vypočítaného predpokladaného záujmu). Ak takáto stránka neobsahuje informácie o nadchádzajúcej udalosti (nenašli sme na nej dátum alebo je na nej dátum z minulosti) ani sa nejedná o novinku podľa nášho chápania noviniek, nemáme ju používateľovi ako odporučiť. Z tohto dôvodu sme vytvorili ďalšiu sekciu, v ktorej každému používateľovi zobrazujeme zvyšné odkazy zo zoznamu odporúčaných odkazov (pozri časť 7.2).

7.3.4 Identifikovanie častí webového sídla

V súčasnosti nie je definovaný jednotný štandard, ktorý by opisoval stavbu webového sídla z pohľadu významu jednotlivých elementov. Navrhli sme formát, ktorý mapuje jednotlivé elementy z webového sídla na všeobecné elementy, akými sú ľavé menu, pravé menu, hlavný obsah stránky, hlavička, a iné. Jedná sa o obdobu formátu súboru *robots.txt* určeného pre vyhľadávače. Opis formátu tohto súboru, ktorý je zapísaný v jazyku XML, je uvedený v prílohe C. Vytvorenie takéhoto súboru pre konkrétne webové sídlo nie je časovo náročné a mohol by tak spraviť každý správca. Grafické editory webových stránok a systémy na správu obsahu by mohli takýto súbor ľahko generovať, nakoľko tieto informácie už obsahujú, len ich neposkytujú ďalším stranám. Vytvorením takéhoto súboru na opis webového sídla pridávame sémantiku inak nič nehovoriacim prvkom stránky, ktorej znalosť sa dá využiť v rozličných aplikáciách.

Vďaka tomuto súboru rozumieme významu jednotlivých častí stránky. V našej práci ho využívame na to, aby sme analyzovali len hlavný textový obsah stránky. To nám umožňuje hľadať len relevantné dátumy, prípadne rozpoznávať sekciu s novými oznamami. Tiež z neho zistíme, ktorú časť webovej stránky môžeme upraviť (napr. môžeme úplne nahradiť pravé menu). Bez jeho existencie analyzujeme vždy celú webovú stránku a úpravy webovej stránky sú silno zviazané s konkrétnou doménou, v ktorej metódu používame.

Na obrázku 7.3 uvádzame príklad súboru s časťou opisujúcou mapovanie pravého menu, časti s ikonami na tlač, hlavného obsahu stránky a sekcie noviniek na všeobecné elementy. Vzorový súbor pre webové sídlo našej fakulty sa nachádza na CD prílohe.

7.4 Diskusia k prispôsobovaniu navigácie

Pri hľadaní vzorov v navigácii podľa navrhnutej metódy platia isté obmedzenia. Niektoré postupnosti odkazov nemusia byť klasifikované pomocou žiadneho z opísaných vzorov. Naopak, niektoré postupnosti môžu predstavovať viacero vzorov (napr. postupnosť odkazov $A — B — A$ predstavuje vzor *kruh* aj vzor *hrot*).

```
<menuRight>
  <tag>div</tag>
  <type>id</type>
  <value>content_right</value>
</menuRight>
<print>
  <tag>div</tag>
  <type>class</type>
  <value>print_button</value>
</print>
<content>
  <tag>div</tag>
  <type>class</type>
  <value>content_text</value>
</content>
<news>
  <tag>div</tag>
  <type>class</type>
  <value>news_annotation</value>
</news>
</website>
```

Obr. 7.3: Príklad mapovania elementov webového sídla na všeobecné elementy.

V navrhnutej metóde prispôsobovania navigácie úplne nahrádzame niektoré časti navigácie vo webovej stránke a umiestňujeme do nich odporúčané odkazy. Alternatívou k tomuto prístupu je vypočítať záujem používateľa o jednotlivé odkazy v existujúcom menu a upraviť ich pomocou niektorej z metód opísaných v časti 2.2. Odkazy môžeme napr. zoradiť podľa určeného záujmu, prípadne môžeme nezaujímavé odkazy skryť alebo anotovať. Nevýhoda tejto alternatívy je, že používateľ by sa takto nedozvedel o stránkach, na ktoré nevedú odkazy z menu. Zaujímavá stránka môže byť hlbšie v hierarchii webového sídla, čo si vyžaduje použitie viacerých odkazov. Z toho dôvodu sme sa rozhodli niektoré pôvodné časti s navigáciou úplne nahradiť odporúčanými odkazmi.

Kapitola 8

Overenie a experimenty

A man provided with paper, pencil, and rubber, and subject to strict discipline, is in effect a universal machine.

Alan Turing

V rámci overenia sme navrhli softvérové nástroje, ktoré spoločne vykonávajú jednotlivé kroky z metódy pre odhadovanie záujmu používateľa a následné prispôsobovanie navigácie. Vytvorený prototyp umožňuje adaptívnu podporu navigácie odporúčaním zaujímavých odkazov návštevníkom webového portálu.

Overenie metódy sme realizovali na webovom sídle našej fakulty[1]. Softvérový návrh vychádza z architektúry klient — server. Sledovanie správania sa používateľa sa deje na strane klienta. Na serveri sa:

- vykonáva analýza webového sídla,
- uchováva model používateľa,
- hľadajú a analyzujú vzory správania sa v navigácii jednotlivých používateľov,
- vyberajú odkazy určené do jednotlivých personalizovaných sekcií.

Navrhnuté riešenie sme mohli implementovať tromi spôsobmi:

- ako zásuvný modul do webového prehliadača,
- prostredníctvom proxy servera alebo
- priamou úpravou cieľového portálu.

[1] http://www.fiit.stuba.sk

Prvý spôsob vyžaduje implementovať prídavný modul do niektorého z bežne používaných webových prehliadačov (najlepšie všetkých). To by vyžadovalo náročné testovanie, ktoré by sa s každou novou verziou daného prehliadača muselo opakovať. Od používateľa by to vyžadovalo inštaláciu tohto modulu do svojho prehliadača. Inštaláciu by musel opakovať s každou novou verziou prototypu.

Druhé riešenie je založené na predpoklade, že používateľ si nastaví svoj prístup k webu cez nami kontrolovaný proxy server. Nevýhodou je menšia miera kontroly zo strany používateľa a prípadná neochota používateľov pristupovať na web cez prostredníka. Výhodou je naopak fungovanie nezávisle od použitého prehliadača. Použitie jedného proxy servera nám umožňuje ľahšie zdieľať výsledky s ostatnými projektmi. Navyše dochádza k synergickému efektu, kedy nemusíme všetky moduly priamo nesúvisiace s overovanou metódou implementovať odznova (napr. vkladanie skriptov do stránky).

Tretí spôsob vyžaduje plný prístup k zdrojovým textom webového portálu, ktorý chceme vylepšiť. Ak nie sme vlastníkmi vylepšovaného portálu, tento spôsob neprichádza do úvahy. Navyše, riešenie by bolo silno viazané na konkrétne webové sídlo a jeho použiteľnosť na iných webových sídlach by bola nízka.

Po zvážení výhod a nevýhod sme sa rozhodli implementovať navrhnutú metódu ako zásuvný modul do adaptívneho proxy servera, keďže sme ho mali k dispozícii.

8.1 Rozšírenie adaptívneho proxy servera

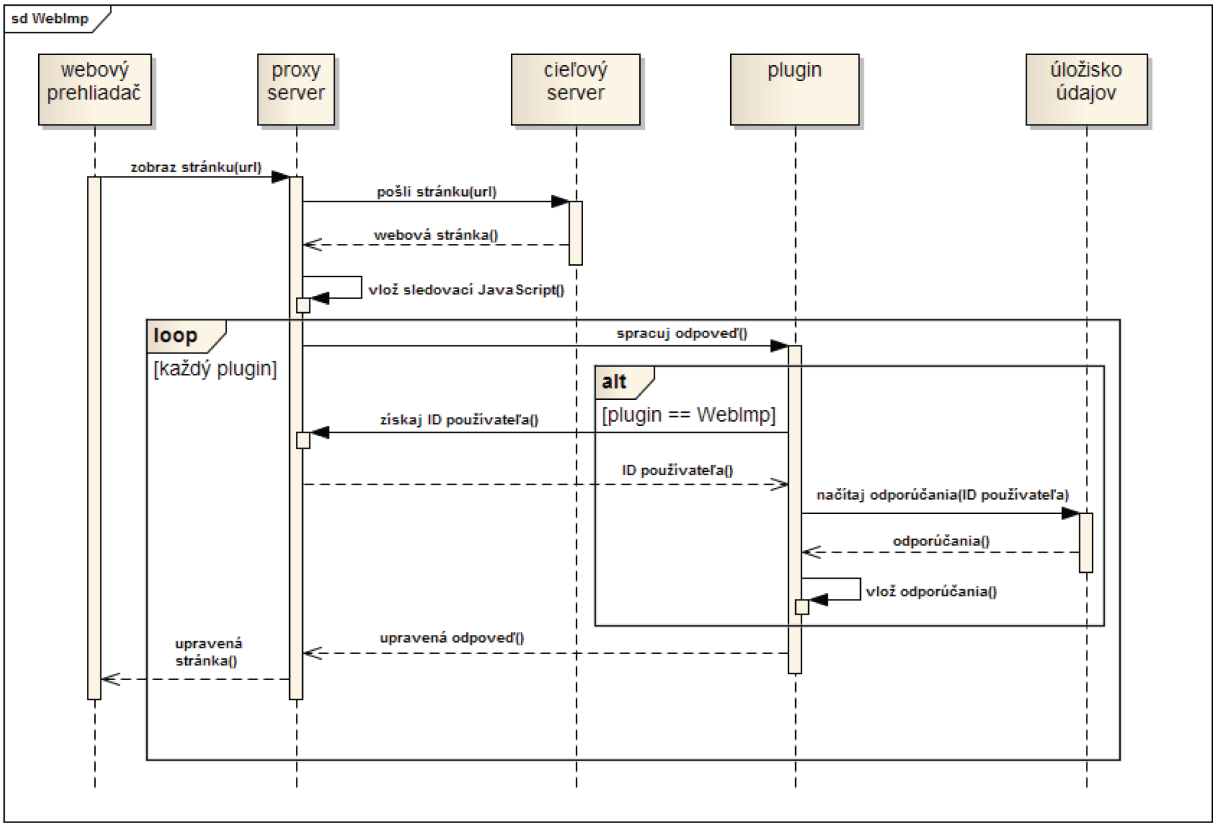
Adaptívny proxy server je projekt vyvíjaný v rámci fakulty [6]. Samotný proxy server len prijíma požiadavky na webové stránky od klienta a preposiela ich cieľovému serveru. Odpoveď od cieľového servera preposiela klientovi. Je však preň možné vytvárať rozšírenia prostredníctvom zásuvných modulov, ktoré môžu túto komunikáciu modifikovať. Adaptívny proxy server tým umožňuje realizáciu rozličných metód personalizácie a vylepšovania štruktúry webových sídiel.

Pri nastavovaní prístupu na web prostredníctvom proxy servera sa nastaví aj jedinečný identifikátor, ktorý sa potom posiela na server pri každej požiadavke o zobrazenie stránky. Vďaka takémuto riešeniu vieme rozlíšiť prístupy na webovú stránku od rôznych používateľov (z pohľadu ich identifikátora). Identifikátor je náhodne vygenerovaná postupnosť znakov, nevieme teda povedať, ktorý človek sa za ním skrýva (ak nám to sám nepovie). Týmto zostáva zachovaná anonymita používateľov.

Obmedzenie spočíva v tom, že nevieme určiť, kto v danej chvíli sedí za počítačom, na ktorom je nastavený konkrétny identifikátor. Ak sa za ním striedajú dvaja ľudia, pre nás to bude vždy rovnaký návštevník. Predpokladáme však, že väčšina používateľov má svoj vlastný počítač, a tak prípadné skreslenie odporúčaní je zanedbateľné. Analýzou charakteristík v správaní sa používateľa sa dá zistiť, či za počítačom sedí ten istý človek. Toto je mimo rozsah tejto práce.

Implementovali sme zásuvný modul, ktorý modifikuje odpoveď od cieľového servera pre klienta. Schéma jeho fungovania je zobrazená na obr. 8.1. Používateľ vyšle požiadavku na zobrazenie stránky. Proxy server prepošle pôvodnú požiadavku na cieľový webový server bezo zmeny. Odpoveď od cieľového servera odovzdá proxy server zásuvnému modulu. Pôvodná odpoveď obsahuje HTML kód webovej stránky. Zásuvný modul vloží do webovej

stránky skript, ktorého účelom je sledovať správanie sa používateľa. Pôvodný kód stránky zmení podľa výsledkov, ktoré mu vrátia nástroje implementujúce metódou vylepšovania štruktúry stránky. Modifikovanú odpoveď následne pošle používateľovi. Zásuvný modul je teda len prostredníkom, odporúčanie odkazov a generovanie personalizovaných sekcií zabezpečujú ostatné nástroje, ktoré sú umiestnené na serveri.



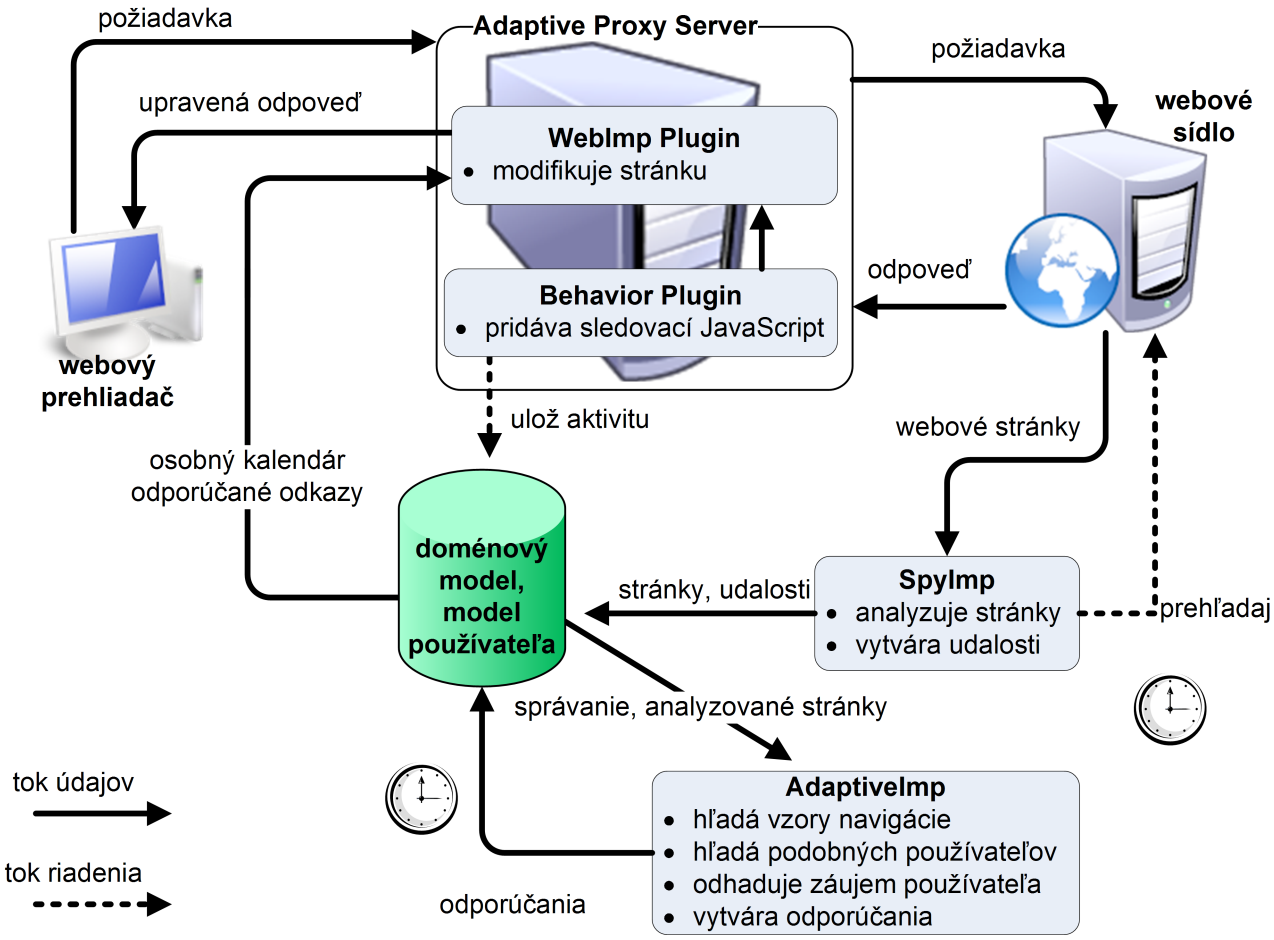
Obr. 8.1: Sekvenčný diagram znázorňujúci postupnosť krokov pri úprave webovej stránky.

Navrhnutú metódu prispôsobovania navigácie sme implementovali v projekte s názvom *WebImp* (angl. *Website Improver*, Vylepšovač stránky). Skladá sa z týchto troch nezávislých komponentov:

- *WebImp* — zásuvný modul (angl. *plugin*) do adaptívneho proxy servera, ktorý prispôsobuje webovú stránku návštevníkovi,
- *SpyImp* — nástroj na analýzu webového sídla, ktorý vytvára doménový model a extrahuje zo sídla údaje používané na odporúčanie odkazov, a
- *AdaptiveImp* — nástroj na určovanie záujmu používateľov o navštívené stránky, hľadanie vzorov v navigácii a odporúčanie odkazov.

Schéma činností jednotlivých komponentov nad spoločným úložiskom dát je zobrazená na obrázku 8.2. Logický model dátového úložiska uvádzame v prílohe C.1. Jednotlivé komponenty pracujú po prvotnej inicializácii nezávisle. Všetky tri komponenty sú implementované v jazyku Java. Inicializácia sa vykoná nasledovnou postupnosťou:

- *SpyImp* vytvorí doménový model tvorený analyzovanými stránkami, z ktorých extrahuje užitočné informácie (dátumy a udalosti),
- *AdaptiveImp* vytvorí model používateľov z údajov o ich aktivite na jednotlivých stránkach, vytvorí zaradenie používateľov do skupín podľa podobnosti a pre každého používateľa vypočíta odporúčané odkazy,
- *WebImp* vloží odporúčania pre konkrétneho návštevníka do webovej stránky.



Obr. 8.2: Schéma činností komponentov navrhnutého riešenia.

Po inicializácii sa akcie jednotlivých komponentov pravidelne opakujú. Analýzu webového sídla vykonávame v závislosti od zvoleného sídla a frekvencie výskytu zmien. Webové sídlo našej fakulty analyzujeme raz za deň. Výpočet podobnosti používateľov a výpočet odporúčaných odkazov vykonávame dvakrát denne. Vhodná periodicita závisí od frekvencie návštev zvoleného webového sídla.

8.2 SpyImp

SpyImp je webový robot, ktorý v pravidelných intervaloch prehľadáva zvolené sídlo. Pracuje podobne ako indexovacie roboty webových vyhľadávačov, pozri algoritmus 8.1.

Algoritmus 8.1 analyzuj sídlo (počiatočná stránka s, doména d)

```
1: odkazy = prázdny zoznam odkazov na stiahnutie
2: odkazy = pridaj s
3: while odkazy obsahuje nejaký element do
4:    l = vyber element z odkazy
5:    stiahni stránku p na ktorú smeruje l
6:    if chyba pri sťahovaní p then
7:       označ l za neplatný
8:       continue
9:    end if
10:   analyzuj p
11:   for all nájdené odkazy o na p do
12:      if o je z domény d then
13:         odkazy = pridaj o
14:      end if
15:   end for
16:   označ l za spracovaný
17: end while
```

Pri prechádzaní webového sídla stiahneme príslušnú stránku a nájdeme na nej všetky hypertextové odkazy. Tie odkazy, ktoré vedú na stránky v rámci domény, pridáme do zoznamu odkazov na stiahnutie. Nesťahujeme stránky mimo zvolenej domény. Odkaz môže byť neplatný a pri sťahovaní nastane chyba. V tomto prípade označíme príslušný odkaz a po analýze ho zaradíme do správy. Správa je užitočná pre správcu webového sídla, ktorý môže chybný odkaz opraviť. Tu vidíme priestor na ďalší výskum a návrh metód na automatické vysporiadanie sa s nefunkčným odkazom. Možno je v adrese len preklep, ak stránka skutočne existuje, mohli by sme ju nájsť a používateľa automaticky presmerovať. Dá sa tiež sledovať kontext, v akom je odkaz použitý, a následne na webovom sídle nájsť stránka, ktorá je svojím obsahom najviac podobná tomuto kontextu.

Po stiahnutí každej stránky vykonávame okrem hľadania odkazov aj jej predspracovanie. Zo stránky odstraňuje komentáre. Hľadáme na nej nadpis (v zdrojovom texte hľadáme elementy *H1* ... *H6* jazyka HTML). Ďalej vyberieme hlavný obsah stránky. Ktorá časť predstavuje hlavný obsah je určené súborom opisujúcim stavbu webového sídla, ktorý vytvorí administrátor. V prípade, že takýto súbor nie je k dispozícii, uchovávame celý textový obsah stránky. Takto extrahované informácie ukladáme do databázy.

Počas analýzy hľadáme na webových stránkach okrem odkazov aj udalosti a novinky. Aby sme označili stránku za oznam o udalosti, musí obsahovať nejaký dátum. Extrakciu dátumov zo stránky realizujeme takto:

1. Pomocou regulárnych výrazov nájdeme všetky výskyty názvov mesiacov a nahradíme ich príslušným číslom.

2. Pomocou ďalšieho regulárneho výrazu nájdeme všetky postupnosti dvoch dvojíc a jednej štvorice čísiel oddelených medzerami alebo bodkami.

3. Z nájdených postupností vytvoríme objekty typu dátum (v tomto kroku sa odstránia neexistujúce dátumy).

Príklad regulárneho výrazu, ktorý rozpozná jednoduchý dátum:

```
\d{1,2}\.\s*\d{1,2}\.\s*\d{4}
```

Dátum musí obsahovať deň, mesiac a rok. Deň a mesiac môžu byť zadané jednou alebo dvomi číslicami, rok musí byť zadaný štyrmi číslicami. Jednotlivé hodnoty musia byť oddelené bodkami a môžu byť oddelené ľubovoľným počtom medzier. Vieme hľadať jednoduché dátumy (zadané dňom, mesiacom a rokom), ako aj rozsahy dátumov (rozsah dní v mesiaci alebo rozsah dní z viacerých mesiacov). Zvyšné regulárne výrazy uvádzame v prílohe C.5.

Po nájdení dátumov vytvoríme udalosť, ktorej názov získame z nadpisu stránky. K udalosti pripojíme všetky dátumy nájdené na stránke, ktoré sú v budúcnosti. Dátumy z minulosti neuvažujeme.

Nástroj *SpyImp* prechádza webové sídlo našej fakulty pravidelne raz za deň. Ak sa pri ďalšej návšteve obsah niektorej zo stránok zmenil, túto skutočnosť označíme za výskyt novinky, na ktorú môžeme používateľa upozorniť. V súčasnosti považujeme za novinku akúkoľvek zmenu na stránke (t.j. aj zmenu jedného písmena).

8.3 AdaptiveImp

AdaptiveImp je nástroj, ktorý realizuje jadro navrhnutej metódy prispôsobovania. Ako vstupy používa prúdy odkazov od používateľov a údaje o ich aktivitách na jednotlivých stránkach. V prúdoch navštívených odkazov hľadá vzory navigácie.

Prúdy odkazov vytvárame podľa postupu opísaného v časti 7.1. Navštívené odkazy zaznamenávame na strane servera. Tu je obmedzenie nášho prístupu. Postupnosť odkazov je úplná v prípade, ak návštevník využíval iba navigáciu poskytovanú webovým sídlom. V prípade, že využil aj funkcionalitu webového prehliadača (tlačidlá *dopredu* a *späť*), nebude postupnosť úplná. Webový prehliadač pri použití týchto tlačidiel neposiela požiadavku na server. Nevieme zistiť, či používateľ zadal adresu nanovo alebo využil tlačidlá prehliadača. Prejaví sa to tak, že jedno reálne sedenie sa nám javí ako dve nezávislé. Na druhej strane, vieme takto odlíšiť ľudí, ktorí využívajú iba navigáciu poskytovanú webovým sídlom (sedenia sú dlhšie) od ľudí využívajúcich aj tlačidlá webového prehliadača (sedenia sú kratšie a je ich viac).

Aby sme vylúčili elementárne postupnosti (aj prechod zo stránky A na stránku B by sa dal považovať za cestu, no takúto postupnosť budú mať všetci používatelia, takže ich nedokážeme ďalej rozdeliť do skupín) zaviedli sme minimálnu dĺžku pre postupnosť, v ktorej hľadáme vzor. Táto dĺžka je 3, nakoľko pri nej už dokážeme rozlíšiť vzory. V postupnosti odkazov hľadáme všetky vzory, ktoré obsahuje. Keď máme nájdené vzory, rozdelíme používateľov do skupín podľa toho, ktorý zo vzorov u nich prevažuje.

AdaptiveImp určuje záujem používateľa o navštívenú stránku z jeho akcií. Vždy pri tom berie do úvahy akcie ostatných používateľov na tej istej stránke, ktoré sú staršie ako akcia, ktorej význam sa snažíme zistiť. Ak sa správanie používateľov v budúcnosti výrazne zmení, hodnoty ich záujmov z minulosti to neovplyvní (t.j. nepočítame znova hodnotu záujmu o stránku navštívenú v minulosti).

Odporúčame odkazy na udalosti a novinky. Pre každého používateľa vytvoríme jeho personalizovaný kalendár, do ktorého pridáme odporúčané udalosti. Zdrojový text kalendára v jazyku HTML uložíme do databázy. Ukážka tohto textu pre jeden mesiac je v prílohe C.4. Pod kalendárom sa zobrazia odkazy na novinky, t.j. na stránky, ktoré používateľa v minulosti zaujali a vyskytla sa na nich zmena. Návštevníkom webového sídla zobrazujeme aj sekciu s názvom *Pozri aj*, do ktorej umiestňujeme také odporúčané stránky, ktoré neobsahujú informáciu o udalostiach ani novinkách.

Na obrázku 8.3 je zobrazená časť obrazovky s hlavnou stránkou webového sídla našej fakulty a pridanými personalizovanými sekciami. Prvou pridanou sekciou je personalizovaný kalendár. Obsahuje niekoľko odporúčaných udalostí (napr. dňa 10.5.2010). Dátumy pri odporúčaných udalostiach sú farebne zvýraznené. Vieme rozlíšiť viacero typov udalostí (napr. pripomienka už známej udalosti, odporúčanie novej udalosti). Ak používateľ klikne na jeden z vyznačených dátumov, zobrazia sa mu udalosti na daný deň. Po kliknutí na názov udalosti sa dostane na stránku, ktorá o udalosti informuje. Pod kalendárom je sekcia odporúčaných odkazov, ktoré nie sú udalosťou ani novinkou. Sú to odkazy na stránky, pre ktoré sme vypočítali, že by používateľa mohli zaujať. Pod touto sekciou sa nachádzajú osobné novinky, kam umiestňujeme odkazy na nové zaujímavé stránky, ako aj na existujúce zmenené stránky. Tieto dva druhy stránok vizuálne nerozlišujeme.

8.4 WebImp

WebImp je zásuvný modul do adaptívneho proxy servera. Proxy server zachytí každú požiadavku od klienta a následnú odpoveď z cieľového servera. Tieto pred preposlaním posúva zásuvným modulom, ktoré môžu realizovať rozličné metódy prispôsobovania. *WebImp* pre stránky z vybranej domény (portálu) vyberie z databázy pripravené personalizované sekcie podľa identifikátora používateľa. Následne ich vloží do stránky na určené miesto (podľa súboru opisujúceho stavbu webového sídla). Takto modifikovanú odpoveď pošle klientovi.

Na strane klienta zbierame údaje o jeho správaní pomocou skriptu, ktorý adaptívny proxy server vkladá do hlavičky požadovanej stránky. Skript je napísaný v jazyku JavaScript využívajúc rozhranie jQuery[2] a je súčasťou funkcionality proxy servera. V pôvodnej verzii zaznamenával čas aktívne strávený na stránke. My sme doň doplnili zaznamenávanie ďalších akcií, ktorými sú počet rolovaní stránky a počet skopírovaní textu do schránky.

8.5 Návrh experimentov

So systémom sme experimentovali na vzorke študentov našej fakulty. V experimentoch sme sa zamerali na dve oblasti:

- vyhodnotenie metódy určovania záujmu o stránky z akcií používateľa a
- vyhodnotenie metódy odporúčania odkazov.

[2] http://jquery.com

Obr. 8.3: Webová stránka s pridanými personalizovanými sekciami. 1 - osobný kalendár so zvýraznenými dňami s odporúčanými udalosťami, 2 - detail dňa so zobrazenou udalosťou, 3 - sekcia odporúčaných odkazov, 4 - sekcia osobných noviniek.

Prvý experiment sme vykonali na vzorke 5 ľudí, ktorých sme požiadali, aby navštívili vybrané stránky z fakultného sídla a ku každej poskytli spätnú väzbu o tom, ako ich zaujala. Pri tom sme sledovali, aké aktivity používatelia vykonávali. Podľa spätnej väzby sme nastavili parametre metódy určovania záujmu o stránku (pozri časť 6.2). Ukázalo sa, že čas strávený na stránke aj počet rolovaní priamo úmerne ovplyvňujú záujem. Skopírovanie textu do schránky predstavovalo vždy kladný záujem, avšak v experimente bolo použité iba raz (testované osoby nevedeli, ktoré akcie zaznamenávame).

V experimente na overenie správnosti určovania záujmu o stránku sme najprv zbierali údaje o akciách používateľov. Potom sme použitím navrhnutej metódy vypočítali záujem pre 55 stránok. Tieto stránky sme nechali používateľov ohodnotiť v rozmedzí 0 až 10 (0 znamenala úplný nezáujem, 10 úplný záujem) a vypočítali sme, do akej miery sa nami určená hodnota zhoduje s explicitne vyjadreným záujmom od používateľov. V tomto experimente sme dosiahli priemernú presnosť odhadu záujmu 62 %.

Údaje o postupnostiach odkazov v rámci webového sídla fakulty sme zbierali 5 týždňov. Potom sme ich rozdelili na jednotlivé sedenia. Ďalej sme vybrali sedenia, ktorých dĺžka bola minimálne 3. Takto sme našli 52 sedení patriacich 19 rôznym používateľom. V sedeniach sme boli schopní identifikovať všetky typy vzorov (podrobnú tabuľku uvádzame v prílohe B.2). Rozdelenie používateľov do skupín podľa prevažujúceho vzoru je v tabuľke 8.1.

Tabuľka 8.1: Rozdelenie návštevníkov do skupín podľa prevažujúceho vzoru v navigácii.

	cesta	**kruh**	**slučka**	**hrot**	**žiadna skupina**
počet návštevníkov	7	1	3	3	5

Používateľov bolo pomerne málo na to, aby sme im odporúčali odkazy iba v rámci skupín. Preto sme ďalej považovali všetkých používateľov za podobných a odporúčali odkazy medzi všetkými.

Po implementovaní nástroja *AdaptiveImp* sme vykonali off-line experiment na vzorke údajov z proxy servera. Táto vzorka obsahovala aktivity používateľov na jednotlivých stránkach po dobu jedného mesiaca. Rozdelili sme ju na dve dvojtýždňové množiny: trénovaciu a testovaciu. Na trénovacej množine sme vypočítali záujemy používateľov o navštívené stránky a odporúčania nenavštívených odkazov. Následne sme na testovacej množine overovali, či používatelia použili odkazy, ktoré by sme im odporučili. Výsledky tohto experimentu sú uvedené v tabuľke 8.2.

Tabuľka 8.2: Výsledky off-line experimentu.

Navštívené stránky	**Odporučené stránky**	**Navštívené odporučené stránky**	**Odhadnutý záujem**
295	85	21 (25 %)	0,445

Po implementovaní všetkých súčastí prototypu sme vykonali experimenty v reálnej prevádzke. Našou hlavnou hypotézou bol predpoklad, že používatelia s odporúčaniami získajú z webového sídla viac informácií ako tí bez odporúčaní. Sledovali sme pri tom zmeny v prístupe na stránky pred a po spustení experimentu. Predpokladali sme, že používatelia v priemere navštívia viac odkazov ako by navštívili bez odporúčaní.

Pred nasadením odporúčaní sme za dva týždne zaznamenali prístupy na webové sídlo fakulty od 19 používateľov. Tí navštívili celkovo 961 stránok. Po spustení odporúčaní sme za tri týždne zaznamenali prístupy od 24 návštevníkov, ktorí videli celkovo 1352 stránok (nárast o 40 %).

Reálnym používateľom sme podľa ich správania sa odporučili 38 odkazov. Následne sme sa spýtali, koľko z nich by navštívili (nakoľko na stránku fakulty nechodia používatelia príliš často a trvalo by dlho, kým by odporúčané odkazy reálne navštívili). Používatelia sa vyjadrili, že by navštívili 55 % z odporúčaných odkazov.

Prístup na web cez adaptívny proxy server bol nastavený aj na niekoľkých počítačoch voľne prístupných študentom v rámci fakulty. Všetky počítače mali nastavený rovnaký identifikátor. Údaje z nich sme oddelili od údajov ostatných používateľov, keďže sa pri nich návštevníci striedajú. Podľa agregovaných prístupov z týchto počítačov sme vypočítali globálny zoznam zaujímavých stránok. Ten sme použili ako prednastavenú hodnotu

odporúčaní pre používateľov, ktorým sme nevedeli pre nedostatok údajov nič iné odporučiť.

Navštevovanie odporúčaných odkazov sme v menšej miere sledovali aj v reálnej prevádzke. Sledovali sme počet unikátnych používateľov, ktorí navštívili konkrétnu stránku pred tým, než bola odporučená, a potom. Výsledky sú v tabuľke 8.3.

Tabuľka 8.3: Počet unikátnych návštevníkov pred a po odporučení odkazu.

Odkaz	**Počet pred odporučením**	**Počet po odporučení**
A	1	4
B	2	4

Nakoniec sme prácu vyhodnotili použitím dotazníka. Jeho plné znenie a výsledky sú uvedené v prílohe B.1.

8.6 Zhodnotenie experimentov a diskusia

Adaptívny proxy server bol v ostrej prevádzke približne 5 týždňov. Z tohto časového obdobia sme mali údaje o aktivitách používateľov na stránkach, z ktorých sme sledovali tie vykonané na stránkach našej fakulty. Celkovo proxy server používalo 44 ľudí, z ktorých 24 navštívilo aj stránky fakulty.

Cieľom pri zisťovaní záujmu a odporúčaní odkazov je, aby sme používateľovi ponúkli zaujímavé odkazy, ktoré by inak neobjavil, a aby z webového sídla vyťažil viac informácií (čo sa dá dosiahnuť návštevou viacerých stránok).

Pri off-line experimente s trénovacou a testovacou množinou sme nedosiahli príliš dobré výsledky. Priemerná hodnota záujmu bola pod 0,5 čo značí negatívny záujem. Navštívených bolo len 25 % odporučených odkazov. Pri tomto experimente sme však reálne odkazy neodporúčali. Dá sa očakávať, že skutoční používatelia by ich navštívili viac.

Potvrdilo sa to v experimente, pri ktorom sme odporúčané odkazy ukázali používateľom. Tí sa vyjadrili, že by navštívili 55 % z nich. Z odporúčaných odkazov by teda návštevníci klikli na každý druhý, čo potvrdzuje, že odporúčané odkazy ich zaujímajú. Ako je vidieť z tabuľky 8.3, počet unikátnych návštevníkov stránok, odkazy na ktoré sme odporúčali, sa výrazne zvýšila. Absolútne čísla sú síce malé, no treba brať do úvahy pomerne malý celkový počet používateľov.

Používatelia by chceli mať odporúčania hneď pri prvej návšteve stránky. Navrhnuté metódy však počítajú s inicializáciou a fázou učenia. Najprv musí používateľ navštíviť niekoľko stránok, aby sme zaznamenali jeho správanie a mohli ho porovnať s ostatnými. Odporúčania na začiatku je možné vyriešiť globálnou inicializáciou, kedy vezmeme celkovo najnavštevovanejšie alebo najzaujímavejšie stránky. Ukázalo sa tiež, že používatelia majú nedôveru k využívaniu proxy servera na prístup k webu, hoci jeho používanie je bezpečné a plne anonymné. V tomto smere treba lepšie vysvetliť princíp jeho fungovania a vzdelávať používateľov v tejto oblasti.

Z odpovedí používateľov v dotazníku vyplýva, že skoro všetci stránku fakulty navštevujú len niekoľkokrát týždenne (alebo menej). To vidíme aj zo zaznamenaných aktivít, kedy 8 ľudí využívajúcich proxy server zaznamenalo menej ako 20 prístupov na fakultné

stránky. Odporúčania hodnotili vo väčšej miere ako nerelevantné, na druhej strane sa vyjadrili, že o daných stránkach nevedeli. Pozitívne tiež je, že vďaka odporúčaniam sa používatelia rýchlejšie dostanú k cieľovej stránke, ktorá je umiestnenná hlbšie v hierarchii webového sídla. Všetci používatelia sa vyjadrili, že stránka s odporúčaniami im viac vyhovuje ako pôvodný obsah pravého menu fakultného webového sídla.

V tejto práci sme sa zamerali na určovanie záujmu a odporúčanie odkazov čisto na základe sledovania správania, pričom sme dosiahli zaujímavé výsledky. V praxi môžeme skombinovať navrhnuté metódy s metódami pracujúcimi na základe analýzy obsahu stránok na dosiahnutie ešte vyššej presnosti pri odporúčaní. Na odhalenie komplexných návykov pri návšteve webových stránok fakulty navrhujeme vykonať dlhodobý experiment, ktorý bude prebiehať po celý akademický rok. Iba tak sa dá odladiť použitie navrhnutých metód v praxi, keďže veríme, že v rôznych obdobiach (na začiatku semestra, pred skúškami, atď.) sa bude správanie návštevníkov líšiť.

Kapitola 9

Záver

Take risks. Ask big questions. Don't be afraid to make mistakes; if you don't make mistakes, you're not reaching far enough.

David Packard

V tejto práci sme navrhli metódu prispôsobovania navigácie vo webovom sídle na základe sledovania správania sa jeho návštevníkov. V metóde kombinujeme dva rozmery správania: navigáciu po webovom sídle a akcie vykonávané na jednotlivých stránkach. Naša metóda stavia na sociálnych aspektoch, kedy sa snaží odhadovať záujem používateľa porovnávaním jeho správania so správaním ostatných. Predpokladáme, že ľudí s podobným správaním budú zaujímať podobné oblasti, a môžu tak využiť vzájomné odporúčanie odkazov.

Cieľom práce bolo uľahčiť používateľovi využívanie konkrétenho webového sídla prostredníctvom prispôsobenia vybranej časti navigácie a obohatenia jednotlivých stránok. Tento cieľ sme splnili návrhom metódy na odporúčanie zaujímavých odkazov. Overili sme ju implementáciou prototypu pridávajúceho nové sekcie do webových stránok sídla našej fakulty. Konkrétne sa jedná o personalizovaný kalendár a novinky.

Navrhnutá metóda je všeobecná, dá sa použiť na zisťovanie záujmu o ľubovoľné stránky a odkazy. My tieto stránky neodporúčame priamo, ale snažíme sa o extrakciu dodatočných informácií, akými sú napríklad informácie o udalostiach. Až tie následne odporúčame v kalendári. Metóda sa však dá použiť aj na iné typy stránok a odkazov. Môžeme podľa nej napr. zmeniť usporiadanie odkazov v už existujúcom menu podľa predpokladaného záujmu používateľa o ne. Alebo sa môžeme pokúsiť o extrakciu iných údajov (napr. geografických lokalít) a následne odporúčať súvisiace odkazy (vieme zistiť, ktoré lokality používateľa zaujímajú, a potom mu odporučiť vhodný cestopis alebo dovolenku).

Naša metóda pracuje v uzavretých webových portáloch. Nehodí sa na použitie v rámci celého webu. Nemá totiž zmysel porovnávať správanie sa používateľov na dvoch nezávislých portáloch, keďže toto správanie je určené aj štruktúrou daného portálu. Má však

zmysel sledovať toto správanie v rámci konkrétneho portálu, v ktorom majú všetci používatelia rovnaké podmienky (všetci používajú tú istú navigáciu).

Riešenie opísané v tejto práci je doménovo a platformovo nezávislé, na úpravu štruktúry stránok a sledovanie používateľa využívame adaptívny proxy server. Zaznamenávanie údajov o používaní webového sídla je nezávislé od webového prehliadača. V práci sme navrhli formát súboru opisujúceho stavbu webového sídla, vďaka ktorému je riešenie s menším úsilím použiteľné aj na iných webových sídlach.

Webové sídla sú rôznorodé a používajú rozmanité technológie na vytváranie stránok. Nie je možné overiť navrhovanú metódu implementáciou úplne všeobecného systému použiteľného na vylepšovanie akéhokoľvek sídla. Systém musí mať aspoň čiastočnú informáciu o štruktúre tohto sídla, aby vedel sídlo prispôsobiť. Keďže mnohé webové sídla sú tvorené v systémoch správy obsahu (angl. *content management system*, CMS), bolo by možné pomerne jednoducho realizovať rozšírenie pre populárne voľne dostupné CMS systémy, akými sú Drupal[1] alebo Joomla[2]. Zdrojový kód sídiel vytvorených v týchto systémoch má rovnaké prvky, navrhnutá metóda by sa tak mohla používať na väčšom počte webových sídiel.

V práci vidíme viacero častí, do ktorých by sa dali zapojiť ďalšie metódy špecializované na konkrétnu činnosť. Ide napr. o inteligentné zisťovanie zmien na stránke (nie len podľa zmeny jednotlivých písmen), extrakciu udalostí (na jednej stránke môže byť oznam o viacerých udalostiach, pričom je výzvou priradiť dátumy ku konkrétnej udalosti a získať zo stránky jej názov).

S prototypom sme vykonali experimenty na vyhodnotenie navrhnutých metód. Z výsledkov vyplýva, že na základe správania sa na webovej stránke sme schopní určiť záujem používateľa o ňu. Takisto používateľom odporúčame zaujímavé odkazy.

[1] http://drupal.org
[2] http://www.joomla.org

Literatúra

[1] Ahn, J.w., Brusilovsky, P., He, D., Grady, J., Li, Q.: Personalized web exploration with task models. In: *WWW '08: Proceeding of the 17th international conference on World Wide Web*, New York, NY, USA, ACM Press, 2008, pp. 1–10.

[2] Andrejko, A., Barla, M., Bieliková, M., Tvarožek, M.: Tools for User Characteristics Acquisition. In Vojtáš, P., ed.: *Proceedings of Annual Conference Datakon'06*, 2006, pp. 139–148.

[3] Ankolekar, A., Vrandečić, D.: Kalpana - enabling client-side web personalization. In: *HT '08: Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, New York, NY, USA, ACM Press, 2008, pp. 21–26.

[4] Baraglia, R., Lucchese, C., Orlando, S., Serrano', M., Silvestri, F.: A privacy preserving web recommender system. In: *SAC '06: Proceedings of the 2006 ACM symposium on Applied computing*, New York, NY, USA, ACM Press, 2006, pp. 559–563.

[5] Baraglia, R., Silvestri, F.: Dynamic personalization of web sites without user intervention. *Communications of the ACM*, 2007, vol. 50, no. 2, pp. 63–67.

[6] Barla, M., Bieliková, M.: Personalizácia "divokého"webu: adaptívny proxy server. In Babič, F., Paralič, J., eds.: *WIKT '09: Proceedings of the 4th Workshop on Intelligent and Knowledge oriented Technologies*, Equilibria, 2009, pp. 48–51.

[7] Barla, M., Tvarožek, M., Bieliková, M.: Rule-based User Characteristics Acquisition from Logs with Semantics for Personalized Web-Based Systems. *Computing and Informatics*, 2009, vol. 28, no. 4, pp. 399–427.

[8] Berendt, B., Spiliopoulou, M.: Analysis of navigation behaviour in web sites integrating multiple information systems. *The VLDB Journal*, 2000, vol. 9, no. 1, pp. 56–75.

[9] Brusilovsky, P.: Methods and techniques of adaptive hypermedia. *User Modeling and User-Adapted Interaction*, 1996, vol. 6, no. 2-3, pp. 87–129.

[10] Brusilovsky, P.: Adaptive Hypermedia. *User Modeling and User-Adapted Interaction*, 2001, vol. 11, no. 1-2, pp. 87–110.

[11] Canter, D., Rivers, R., Storrs, G.: Characterizing user navigation through complex data structures. *Behaviour & Information Technology*, 1985, vol. 4, no. 2, pp. 93–102.

[12] Clark, L., Ting, I.H., Kimble, C., Wright, P., Kudenko, D.: Combining ethnographic and clickstream data to identify user Web browsing strategies. @online, dostupný

z URL <http://informationr.net/ir/11-2/paper249.html>. *Information Research*, 2006, vol. 11, no. 2, citovaný 1.4.2010.

[13] Cockburn, A., McKenzie, B.: What do web users do? An empirical analysis of web use. *International Journal of Human-Computer Studies*, 2001, vol. 54, no. 6, pp. 903–922.

[14] Das, A.S., Datar, M., Garg, A., Rajaram, S.: Google news personalization: scalable online collaborative filtering. In: *WWW '07: Proceedings of the 16th international conference on World Wide Web*, New York, NY, USA, ACM Press, 2007, pp. 271–280.

[15] De Bra, P., Aerts, A., Berden, B., de Lange, B., Rousseau, B., Santic, T., Smits, D., Stash, N.: AHA! The adaptive hypermedia architecture. In: *HYPERTEXT '03: Proceedings of the fourteenth ACM conference on Hypertext and hypermedia*, New York, NY, USA, ACM Press, 2003, pp. 81–84.

[16] De Bra, P., Brusilovsky, P., Houben, G.J.: Adaptive hypermedia: from systems to framework. *ACM Computing Surveys*, 1999, vol. 31, no. 4es, p. 12.

[17] Hawkey, K., Inkpen, K.: Web browsing today: the impact of changing contexts on user activity. In: *CHI '05: CHI '05 extended abstracts on Human factors in computing systems*, New York, NY, USA, ACM Press, 2005, pp. 1443–1446.

[18] Hu, J., Zhong, N.: Clickstream Log Acquisition with Web Farming. In: *WI '05: Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, Washington, DC, USA, IEEE Computer Society, 2005, pp. 257–263.

[19] Joachims, T., Freitag, D., Mitchell, T.: WebWatcher: A Tour Guide for the World Wide Web. In: *IJCAI '97: Proceedings of the International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, 1997, pp. 770–777.

[20] Krištofič, A., Bieliková, M.: Improving adaptation in web-based educational hypermedia by means of knowledge discovery. In: *HYPERTEXT '05: Proceedings of the sixteenth ACM conference on Hypertext and hypermedia*, New York, NY, USA, ACM Press, 2005, pp. 184–192.

[21] Manber, U., Patel, A., Robison, J.: Experience with personalization of Yahoo! *Communications of the ACM*, 2000, vol. 43, no. 8, pp. 35–39.

[22] Mikroyannidis, A., Theodoulidis, B.: A Theoretical Framework and an Implementation Architecture for Self Adaptive Web Sites. In: *WI '04: Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence*, Washington, DC, USA, IEEE Computer Society, 2004, pp. 558–561.

[23] Milic-Frayling, N., Jones, R., Rodden, K., Smyth, G., Blackwell, A., Sommerer, R.: Smartback: supporting users in back navigation. In: *WWW '04: Proceedings of the 13th international conference on World Wide Web*, New York, NY, USA, ACM Press, 2004, pp. 63–71.

[24] Mobasher, B., Cooley, R., Srivastava, J.: Automatic personalization based on Web usage mining. *Communications of the ACM*, 2000, vol. 43, no. 8, pp. 142–151.

[25] Morita, M., Shinoda, Y.: Information filtering based on user behavior analysis and best match text retrieval. In: *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, Springer-Verlag New York, Inc., 1994, pp. 272–281.

[26] Mulvenna, M.D., Anand, S.S., Büchner, A.G.: Personalization on the Net using Web mining: introduction. *Communications of the ACM*, 2000, vol. 43, no. 8, pp. 122–125.

[27] Nauerz, A., Bakalov, F., König-Ries, B., Welsch, M.: Personalized recommendation of related content based on automatic metadata extraction. In: *CASCON '08: Proceedings of the 2008 conference of the center for advanced studies on collaborative research*, New York, NY, USA, ACM Press, 2008, pp. 57–71.

[28] Perkowitz, M., Etzioni, O.: Adaptive Web sites. *Communications of the ACM*, 2000, vol. 43, no. 8, pp. 152–158.

[29] Spiliopoulou, M.: Web usage mining for Web site evaluation. *Communications of the ACM*, 2000, vol. 43, no. 8, pp. 127–134.

[30] Sugiyama, K., Hatano, K., Yoshikawa, M.: Adaptive web search based on user profile constructed without any effort from users. In: *WWW '04: Proceedings of the 13th international conference on World Wide Web*, New York, NY, USA, ACM Press, 2004, pp. 675–684.

[31] Tauscher, L., Greenberg, S.: How people revisit web pages: empirical findings and implications for the design of history systems. *International Journal of Human-Computer Studies*, 1997, vol. 47, no. 1, pp. 97–137.

[32] Ting, I.H., Kimble, C., Kudenko, D.: UBB Mining: Finding Unexpected Browsing Behaviour in Clickstream Data to Improve a Web Site's Design. In: *WI '05: Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, Washington, DC, USA, IEEE Computer Society, 2005, pp. 179–185.

[33] Tury, M., Bieliková, M.: An approach to detection ontology changes. In: *ICWE '06: Workshop proceedings of the sixth international conference on Web engineering*, New York, NY, USA, ACM Press, 2006, p. 14.

[34] Velayathan, G., Yamada, S.: Behavior-based web page evaluation. In: *WWW '06: Proceedings of the 15th international conference on World Wide Web*, New York, NY, USA, ACM Press, 2006, pp. 841–842.

Príloha A

Príspevky z medzinárodných konferencií

V tejto prílohe uvádzame príspevky z medzinárodných konferencií, v ktorých sme opísali jednotlivé časti nášho riešenia.

A.1 Príspevok prijatý na konferenciu WWW 2010

V tejto prílohe sa nachádza článok prijatý a prezentovaný na konferencii *20th International Conference on World Wide Web – WWW'2010*, ktorá sa konala v apríli 2010 v Raleigh, Severná Karolína, USA. V článku opisujeme automatické odhadovanie záujmu používateľa o navštívenú webové stránku na základe jeho akcií. Článok bol prijatý a prezentovaný v sekcii *Posters*.

Estimation of User Interest in Visited Web Page

Michal Holub
Institute of Informatics and Software Engineering
Faculty of Informatics and Information Technologies, Slovak University of Technology
Ilkovičova 3, 842 16 Bratislava, Slovakia
miso.holub@gmail.com

Maria Bielikova
Institute of Informatics and Software Engineering
Faculty of Informatics and Information Technologies, Slovak University of Technology
Ilkovičova 3, 842 16 Bratislava, Slovakia
bielik@fiit.stuba.sk

ABSTRACT
Nowadays web portals contain large amount of information that is meant for various visitors or groups of visitors. To effectively navigate within the content the website needs to "know" its users in order to provide personalized content to them. We propose a method for automatic estimation of the user's interest in a web page he visits. This estimation is used for the recommendation of web portal pages (through presenting adaptive links) that the user might like. We conducted series of experiments in the domain of our faculty web portal to evaluate proposed approach.

Categories and Subject Descriptors
H.5.4 [Information Interfaces and Presentation (e.g., HCI)]: Hypertext/Hypermedia—*Navigation.* H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—*Relevance Feedback.*

General Terms
Algorithms, Design, Experimentation.

Keywords
Adaptive navigation support, user behavior, user interest estimation, link recommendation

1. INTRODUCTION
Web portals are being visited by various users pursuing different goals. However, most websites offer all visitors the same content. Therefore, the visitors are often presented information in which they have no interest [1]. Another problem is inappropriate navigation that confuses users. They have difficulties to decide which link from the large amount of possibilities they should follow. While navigating, some users can discover interesting information. We believe that the link to the web page with such interesting information can also concern other users with similar goals, so using social recommendation can help in this context.

Often user's interests are determined based on the content of documents the user has read [3]. Interests are then included in the user model which can be expressed by concepts (or just keywords) extracted from these documents. If we know what topics (expressed by keywords) the user prefers, we can recommend him documents (web pages) with similar content.

There are several ways how to get implicit feedback and use it in estimation of user's interests. In areas where links are well annotated (like news portals where links to articles are followed by a short introduction) the event of user clicking on a link can be considered as positive interest [4]. However, in general we cannot predict user's interest in the following web page solely on the fact that the user clicked on the link pointing to this page. Another approach is to track the actions user performs while reading a web page. Actions like printing the page or adding it to bookmarks show positive interest. On the other hand, spending very small amount of time reading it or even closing the browser while the page is being loaded show negative interest [6].

For estimation of user's interest we propose a method of tracking his behavior when visiting a particular web page. With this data we are looking for users who behave similarly and recommend them links based on estimated interest and collaborative filtering.

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

2. DETERMINING USER'S INTEREST
We combine behavioral analysis for deriving user's interest in a web page he currently visits with collaborative filtering technique as described in [5]. We use collaborative filtering for predicting user's interest in a web page he has not yet visited. For our purpose the items are web pages and the rating of an item is an estimate of user's interest based on his behavior.

To determine user's interest we observe actions he makes on a web page. These include *time spent on a web page*, *number of scrolling events that occur* and *number of times he copies text into the clipboard*. Our method is based on comparison of current user's behavior with behavior of others. We compare the values of first two actions with values from other people who visited the same page. If the value for current user is more than X % higher than the average we consider it as a sign of positive interest in the page. On the other hand, when it is more than X % lower than the average we consider it as a sign of negative interest. When the value is around average (± X %) it is a sign of neutral interest. Experiments show that optimal value of X is in range of 20-30.

We estimate the actual value of user's interest in each page he visits according to Figure 1, where the x axis represents time spent on a web page, the y axis represents the number of scrolls done. Symbol ↑ means higher than average value, symbol ↓ means lower than average value, and symbol – means average value. The time spent and the number of scrolls made has the same weight in the final score. We increase this value by 0.1 when the user also copied text into clipboard; otherwise we decrease it by 0.1. Resulting aggregated interest is in the interval <0,1> with 0 meaning no interest and 1 meaning total interest in the visited web page. For the value of spent time we set an inactivity threshold to be 4 times the average time spent on that page by others.

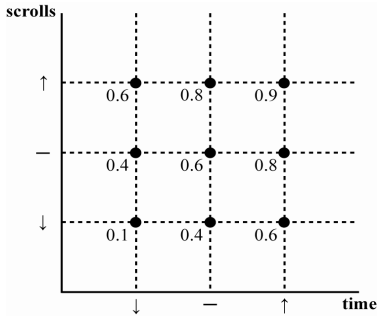


Figure 1. Estimation of user's interest in visited web page.

3. LINK RECOMMENDATION

We recommend links by predicting user's interest in yet unseen pages using collaborative filtering method. We compute the values of Pearson correlation coefficient between the user to whom we want to recommend a web page and all other users [5]:

$$S_{a,u} = \frac{\sum_{i=1}^{I}(r_{a,i} - r_a) \times (r_{u,i} - r_u)}{\sqrt{\sum_{i=1}^{I}(r_{a,i} - r_a)^2 \times \sum_{i=1}^{I}(r_{u,i} - r_u)^2}}$$

In the domain of web pages recommendation $r_{a,i}$ means the interest of user a in page i, r_a means average interest of user a and I is the total number of pages visited by user.

We use this value to predict user's interest in a page which he has not yet visited but which other users have. Pearson correlation coefficient says how similar two users are considering their ratings of items. In our particular case this rating is represented by behavior of users on every page they both visit. We compute the predicted value of interest like this [5]:

$$p_{a,i} = r_a + \frac{\sum_{u=1}^{N}(r_{u,i} - r_u) \times S_{a,u}}{\sum_{u=1}^{N} S_{a,u}}$$

Here $p_{a,i}$ means prediction of interest of user a in page i, $S_{a,u}$ means the similarity of users a and u (value of their Pearson correlation coefficient) and N is the number of similar users.

4. EVALUATION AND CONCLUSIONS

To evaluate proposed method we developed a software tool that supports adaptive navigation for guests of particular web portal. It enhances each original web page by adding links that represent interesting part of the portal. We experimented with web portal of our faculty (www.fiit.stuba.sk) by adding the recommendations to the right navigational menu.

We designed client-server architecture with an adaptive proxy server developed in our group in the middle. Adaptive proxy server is a platform for undisturbed involving of methods and techniques for the adaptation of the content and navigation on the web [2]. It enables developers to control the adaptation process by means of services. On the client side there is a behavior tracking script. It invokes a web service on the server side which collects data about user behavior. Another component on the server is responsible for computing user's interest and making predictions for unseen pages. It then selects the links to be recommended.

We realized a plug-in to the adaptive proxy server, which handles client requests and server responses. It can be extended to conduct various methods of web adaptation. Our plug-in inserts the tracing script to each page together with recommended links. Many web pages on our faculty portal inform about an event. Our tool extracts dates from these pages. We use our interest estimation method to construct a personalized calendar of events. If the interest is positive, we add the event to user's calendar and show it on the page. We can also recommend events between users.

We did a series of experiments on our faculty website. In the experiments the visitors of our website were asked to express their interest in visited web page as an integer ranging from 0 to 10 with 0 meaning no interest and 10 meaning total interest. During the visit of each page interest was estimated by proposed method.

Our results indicate that the most important quantity that determines user's interest is the time spent on a webpage. The more time users had spent on a web page above average the higher they rated it. This is the same result as our method gives.

Scrolling also proved to indicate positive interest in the web page. However, the results showed that when a user does not use scrolling it does not always mean his lack of interest. Thus we should give higher weight to the value of time spent and less weight to number of scrolling events. Copying text into clipboard proved to be a clear sign of interest. Nevertheless, only few people actually used it during our experiments.

We have proposed an approach to automatic estimation of user's interest in a web page. We are able to predict his interest for unvisited pages and recommend him interesting links. In the future work we plan to use this estimation in social adaptive navigation support that employs groups of users determined by observing their paths of navigation through the whole web portal.

Acknowledgements. This work was partially supported by the Scientific Grant Agency of SR, grant VG1/0508/09 and it is partial result of the Research & Development Operational Programme for the project Smart Technologies, Systems and Services, ITMS 26240120005, co-funded by the ERDF.

5. REFERENCES

[1] Barla, M., Tvarožek, M., and Bieliková, M. Rule-based user characteristics acquisition from logs with semantics for personalized web-based systems. Computing and Informatics, Vol. 28, No. 4 (2009), 399-427.

[2] Barla, M., and Bieliková, M. „Wild web" personalization: adaptive proxy server. In Proc. of the 4th Workshop on Intelligent and Knowledge oriented Technologies (Herľany, Slovakia, 2009), 48-51 (in Slovak).

[3] Brusilovsky, P. Methods and techniques of adaptive hypermedia. User Modeling and User-Adapted Interaction, Vol. 6, No. 2-3 (1996), Springer Netherlands, 87-129.

[4] Das, A.S., et al. Google news personalization: scalable online collaborative filtering. In Proc. of the 16th Int. Conf. on WWW (Banff, Canada, 2007), ACM Press, 271-280.

[5] Sugiyama, K., Hatano, K., and Yoshikawa, M. Adaptive web search based on user profile constructed without any effort from users. In Proc. of the 13th Int. Conf. on WWW (New York, NY, USA, 2004), ACM Press, 675-684.

[6] Velayathan, G., and Yamada, S. Behavior-based web page evaluation. In Proc. of the 15th Int. Conf. on WWW (Edinburgh, Scotland, 2006), ACM Press, 841-842

A.2 Príspevok odoslaný na konferenciu RecSys 2010

V tejto prílohe sa nachádza článok odoslaný na konferenciu *ACM Recommender Systems 2010*, ktorá sa bude konať v septembri 2010 v Barcelone, Španielsko. V dobe písania tejto práce bol článok posudzovaný programovým výborom konferencie. V článku opisujeme navrhnutú metódu odhaľovania vzorov v navigácii používateľov a ich využitie pri prispôsobovaní navigácie vo webovom sídle.

Adaptive Link Recommendation Based on User Actions

Michal Holub
Institute of Informatics and Software Engineering
Faculty of Informatics and Information Technologies
Slovak University of Technology
Ilkovičova 3, 842 16 Bratislava, Slovakia
miso.holub@gmail.com

Mária Bieliková
Institute of Informatics and Software Engineering
Faculty of Informatics and Information Technologies
Slovak University of Technology
Ilkovičova 3, 842 16 Bratislava, Slovakia
bielik@fiit.stuba.sk

ABSTRACT
Web portals contain large amount of information. Various groups of users could benefit from it if the information is presented in personalized way. For this to be accomplished the website needs to "know" its users. When surfing the Web users leave digital footprints in the form of navigational paths and actions taken. Users who behave similarly can recommend interesting pages to each other. In this paper we present a method for adaptive navigation support and link recommendation based on an analysis of the user navigational patterns and behavior on the web pages while browsing through a web portal. We also mine the portal to extract interesting information from it. Finally, we evaluate our method by introducing a system which modifies a website and recommends links to the pages which the user should not miss.

Categories and Subject Descriptors
H.5.4 [**Information Interfaces and Presentation (e.g., HCI)**]: Hypertext/Hypermedia—*Navigation*. H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Relevance Feedback.*

General Terms
Algorithms, Design, Experimentation, Verification.

Keywords
Adaptive navigation support, link recommendation, user interest estimation, user behavior, navigational patterns.

1. INTRODUCTION
Web portals are being visited by various users pursuing different goals. However, most websites offer all groups of visitors the same content. Therefore the visitors are often presented information in which they have no interest [4].

While browsing through a web portal some users can discover interesting pages that are hidden deeper in the hierarchy of the portal. If the users with similar goals knew about these pages they could find them interesting, too. Therefore we believe that the navigation should be personalized to include the links to these web pages. In this paper we introduce a method of navigation adaptation based on user's behavior and social recommendation of links among users with similar behavior.

The rest of this paper is organized as follows: In section 2 we present similar solutions to link recommendation and user's interest estimation. In section 3 we propose a method of adaptive link recommendation based on monitoring of users' behavior on a web page as well as on the whole web portal. In section 4 we present implemented software tools and results of their evaluation. Finally, section 5 concludes the paper by presenting plans for future work.

2. RELATED WORK
When accessing the information on a web portal people use different patterns. The most common pattern is to follow hyperlinks, which accounts for more than half of all the possibilities [7, 12]. This introduces a problem with inappropriate navigation containing large number of links. The user has difficulties deciding which link to follow and he eventually gets lost. Therefore adapting the links on a web page could bring significant improvement to user's browsing experience. Other dominant mean is using the browser's back button [10]. Accessing websites through the history, list of bookmarks, typing of exact URL and other means is insignificant.

User's habits can be derived from the navigational patterns found in the sequences of links he uses in a particular web portal. Four basic navigational patterns (path, loop, ring and spike) were described in [6]. From the prevailing patterns different browsing strategies can be identified.

User's interests are often determined based on the content of documents the user has read [5]. The user model can be expressed by concepts or keywords extracted from these documents [2]. If we know what topics (expressed by the keywords) the user prefers, we can recommend him documents (web pages) with similar content. The disadvantage is that documents have to be written in language which we can process.

In [13] authors use rather different approach based on user behavior tracking to estimate his interest. For this to accomplish we need to get feedback from the user. There are several ways how to get implicit feedback and use it in user interest estimation. When links are well annotated (like on news portals where links to articles contain a short introduction) the event clicking on the link is considered as positive interest [8]. However, in general

scenarios we cannot predict interest in the following web page only on the occurrence of click event.

User's actions on a web page can also be used to determine his interest. Actions like printing the page or adding it to bookmarks show positive interest. Spending very small amount of time reading it or even closing the browser while the page is being loaded show negative interest [13].

With user interest determined navigation personalization as well as link recommendation can be done. In [9] authors propose a method of interesting link recommendation by highlighting the links. This method extracts keywords from pages a user visits. Then it recommends links that lead to other pages which contain the same keywords. System Web Watcher, which implements this method can also show similar pages to the page that is currently being viewed based on this principle. The system uses a proxy server to incorporate its toolbar into every web page.

Other method is based on monitoring the context in which the links are being used [1]. This method consists of creating a knowledge base from the links each user has clicked on. Then the clusters of links, which are often used together, are built from the knowledge base. Links from cluster with the largest overlap with the current session are then being recommended to the user.

All mentioned methods share the same feature which is user interest estimation based on his actions. They prefer behavior of users over content of documents which they were shown.

3. ADAPTIVE LINK RECOMMENDATION

We propose a method for adaptive recommendation of interesting links in a particular web portal. For a specific user the method recommends links that similar users found interesting. Moreover, it also recommends links to this user based on his previous surfing sessions. The recommendation is done by modification of the website structure when special sections are added to the web page. When recommending links we do not consider the content of visited sites. We decided to make recommendations solely on the analysis of user's behavior. The recommendation thus does not depend on the language of the site. We are able to analyze interest on different language versions of the same page. Our method of adaptive link recommendation works in two steps:

- Mining web usage history for navigational patterns.
- Recommending of links based on user's behavior.

In the first step we analyze the sequences of followed links from each user's session. We then compare so called *clickstreams* between each pair of users using cosine similarity. As an output we get groups of users with similar navigational patterns.

In the second step we monitor behavior of users on a particular web page of the portal. From their actions we automatically determine their interest in the page. We then recommend links to interesting pages among users of each group from step one.

3.1 Discovering navigational patterns

We find similar users based on comparison of navigational patterns they follow in a closed web portal. We believe that users who follow analogous paths have similar interests. There are four basic navigational patterns as described in [6]:

- *Path* – a sequence in which nodes do not repeat.
- *Ring* – a sequence that starts and ends in the same node.
- *Loop* – a sequence that goes through already visited node.
- *Spike* – a sequence that goes back through the same trail.

In each session a user visits several pages of the web portal. This session is described by a vector whose elements are links to the web pages arranged in order they were visited. During visits to a web portal we create a long-term user model from these vectors. Vectors from older visits have lower weights.

We consider a continuous sequence of links to be a session. For this purpose we use the *referrer* field of HTTP request message. If the URL of previously visited page equals referrer value of currently visited page, we consider the pages to be in the same session. Otherwise we create a new session.

We use the vectors of visited web pages to find similar users. The process of dividing users into groups is presented in Alg. 1.

Algorithm 1 Group users according to their similarity

```
1: for each user u do
2:     find patterns in clickstreams of u
3:     put u to group according to prevailing pattern
4: for each group g do
5:     for each user u in group g do
6:         for each user v in group g do (u ≠ v)
7:             compute cosine similarity of clickstreams (u, v)
8:         sort users in group g according to their similarity to u
```

Alg. 2 presents the process of recommending links among users.

Algorithm 2 Recommend links for user *u* from his similar users

```
1: similar = select top K similar users
2: for each user v in similar do
3:     calculate Pearson coefficient (u, v)
4: for each page p not visited by u do
5:     predict interest of user in page (u, p)
6: recommend top M pages with highest predicted interest
```

Navigational patterns of users have to be of a certain minimal length (so that each sequence of two following pages does not represent a *path* pattern). We have four groups for each pattern and one group for users with no dominant pattern. After finding similar users to user *u* we select top *M* of them to form a recommendation group. The groups change according to new browsing sessions in which the users can behave differently. This reflects the evolution of user's behavior in time.

3.2 Determining user interest

In order to recommend links to a particular user we need to evaluate the interests of the users in his recommendation group. We can recommend pages which other users liked. To determine user's interest in a particular web page we observe actions he makes on this page. These include *time spent on a web page*, *number of scrolling events* that occur and *number of times he copies text into the clipboard*.

Our method is based on comparison of current user's behavior with behavior of other users. We compare the values of first two

actions with values from other people who visited the same page. If the value for the current user is more than X % higher than the average, we consider it as a sign of positive interest in the page. In contrast, when it is more than X % lower than the average we consider it as a sign of negative interest. When the value is around average (± X %) it is a sign of neutral interest. It shows that optimal value of X is in the range of 20-30.

When no behavioral data for a particular web page is available we cannot estimate the user's interest. This is a problem with newly added pages also known as cold start problem. Once new pages have been visited by some users we can estimate their interest.

We estimate the actual value of user's interest in each page he visits according to Figure 1. The time spent and the number of scrolls has the same weight. We increase this value by 0.1 when the user also copied text into clipboard; otherwise we decrease it by 0.1. Resulting interest is in the interval <0,1> with 0 meaning no interest and 1 meaning total interest in the visited web page.

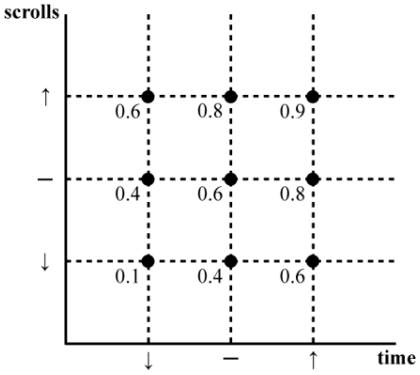


Figure 1. Estimation of user's interest from his behavior.

3.3 Social recommendation of links

We recommend web pages by predicting user's interest in yet unseen pages using collaborative filtering method. We compute the values of Pearson correlation coefficient between the user to whom we want to recommend a web page and all other users from his recommendation group [11]:

$$S_{a,u} = \frac{\sum_{i=1}^{I}(r_{a,i} - r_a) \times (r_{u,i} - r_u)}{\sqrt{\sum_{i=1}^{I}(r_{a,i} - r_a)^2 \times \sum_{i=1}^{I}(r_{u,i} - r_u)^2}}$$

where $r_{a,i}$ means interest of user a in page i, r_a means average interest of user a and I the total number of pages visited by user a.

We use this value to predict user's interest in a page which he has not yet visited but which other users have. Pearson correlation coefficient says how similar two users are considering their ratings of items. In our particular case this rating is represented by behavior of users on every page they both visit. We compute the predicted value of interest [11]:

$$p_{a,i} = r_a + \frac{\sum_{u=1}^{N}(r_{u,i} - r_u) \times S_{a,u}}{\sum_{u=1}^{N} S_{a,u}}$$

where $p_{a,i}$ means prediction of interest of user a in page i, $S_{a,u}$ means the similarity of users a and u (value of their Pearson correlation coefficient) and N is the number of similar users.

4. EVALUATION AND EXPERIMENTS

To evaluate proposed method for user interest estimation we developed software tools which support adaptive navigation by recommending interesting web pages to guests of particular web portal. We experimented with the web portal of our faculty (www.fiit.stuba.sk).

We proposed client-server architecture with an adaptive proxy [3] in the middle as shown in Figure 2. Adaptive proxy can be extended to conduct various methods of web personalization. We use proxy to put behavior tracking script into the web page. It sends logged behavioral data to the server. One component (SpyImp) creates the domain model by analyzing pages of selected web portal. Another server component (AdaptiveImp) is responsible for grouping of users, estimating their interests and making predictions for unseen pages. Then it selects the links to be recommended. The user model consisting of the vectors of clickstreams and his behavior is periodically updated.

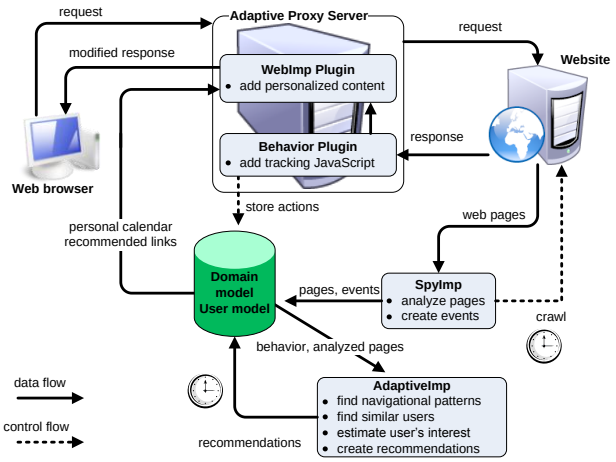


Figure 2. Architecture of proposed link recommender system.

Our plug-in (WebImp) modifies the web page by adding special sections with recommended links. One of those sections is personalized calendar. Many web pages on the web portal of our faculty inform about an upcoming event. We automatically extract dates from these pages and create events. Using proposed method we determine user's interest in such a page. Then, if the interest is positive, we add the event to user's calendar and insert this calendar to the web page. This way we can also recommend events for each user. The calendar is dynamic and personalized to every user. Figure 3 shows part of a web page with added personalized calendar.

We provided a series of experiments on our faculty website. In the first experiment visitors were asked to express their interest in visited page as an integer from 0 to 10 (higher number means higher interest). Results indicate that time is the best interest indicator. This is analogous with result of our method. Scrolling proved to indicate positive interest as well. However, the

Figure 3. Calendar with recommended event on 21/04/2010.

experiment showed that when a user does not use scrolling it does not always mean he is not interested in the page.

We also did an offline experiment with activities collected from 24 users. We divided these data to testing and training sets (2 weeks each). For the purpose of this experiment all users were considered to be in the same recommendation group. We computed users' interests for every page in sessions from the training set. Then we predicted interest for pages each user has not visited. We selected top 10 pages with the highest predicted interest as recommended pages. Then we evaluated if the recommended pages were present in the testing set. We also estimated user's interest in each visited page that was previously recommended. Results are shown in Table 1.

From the Table 1 we can see that people visited 25 % of the pages we would recommend to them. Since we do not perform direct global guidance of user but offer additional links, we consider it as a good result. It is likely that this result would improve during online experiment and recommended links to events would interest more users.

Table 1. Results of the offline experiment.

Visited pages	Recommended pages	Visited recommended pages	Average interest
295	85	21 (25 %)	0.445

5. CONCLUSION AND FUTURE WORK

We have presented a method for adaptive recommendation of interesting links. Our approach is based on collaborative filtering. Instead of content of documents we consider data about user's actions. Our method automatically determines user's interest in a visited web page. We are also able to predict user's interest for yet unvisited pages and use it for link recommendation.

We showed a useful application of our method by creating personalized calendar of events on our faculty's web portal. Using this method we can also personalize other sections of a web page. Users with similar navigational patterns form a social group. In the future work we plan to conduct an online experiment through longer period of time. We plan to study the features of created social groups in more detail and evaluate the relevance of recommended events on larger group of students.

6. REFERENCES

[1] Baraglia, R., et al. 2006. A Privacy Preserving Web Recommender System. In Proc. of the 2006 ACM Symposium on Applied Computing (Dijon, France, 2006). ACM Press, 559-563.

[2] Barla, M. and Bieliková, M. 2009. On Deriving Tagsonomies: Keyword Relations coming from the Crowd. In LNAI 5796, Proc. of Int. Conf. on Computational Collective Intelligence, ICCCI 2009, Springer, 309–320.

[3] Barla, M. and Bieliková, M. 2009. „Wild web" personalization: adaptive proxy server. In Proc. of the 4th Workshop on Intelligent and Knowledge oriented Technologies (Herľany, Slovakia, 2009), F. Babič and J. Paralič (Eds.), 48-51 (in Slovak).

[4] Barla, M., Tvarožek, M. and Bieliková, M. 2009. Rule-based user characteristics acquisition from logs with semantics for personalized web-based systems. Computing and Informatics, Vol. 28, No. 4 (2009), 399-427.

[5] Brusilovsky, P. 1996. Methods and techniques of adaptive hypermedia. User Modeling and User-Adapted Interaction, Vol. 6, No. 2-3, Springer Netherlands, 87-129.

[6] Canter, D., Rivers, R. and Storrs, G. 1985. Characterizing User Navigation through Complex Data Structures. Behaviour and Information Technology, Vol. 4, No. 2, 93-102.

[7] Cockburn, A. and McKenzie, B. 2001. What do web users do? An empirical analysis of web use. International Journal of Human-Computer Studies, Vol. 54, No. 6, 903-922.

[8] Das, A.S., et al. 2007. Google news personalization: scalable online collaborative filtering. In Proc. of the 16th Int. Conf. on World Wide Web (Banff, Alberta, Canada, 2007), ACM Press, 271-280.

[9] Joachims, T., Freitag, D. and Mitchell, T. 1997. WebWatcher: A Tour Guide for the World Wide Web. In Proc. of the 1997 Int. Conf. on Artificial Intelligence (Nagoya, Japan, 1997), Morgan Kaufmann, 770-777.

[10] Milic-Frayling, N., et al. 2004. Smartback: supporting users in back navigation. In Proc. of the 13th Int. Conf. on World Wide Web (NY, USA, 2004), ACM Press, 63-71.

[11] Sugiyama, K., Hatano, K. and Yoshikawa, M. 2004. Adaptive web search based on user profile constructed without any effort from users. In Proc. of the 13th Int. Conf. on World Wide Web (New York, NY, USA, 2004), ACM Press, 675-684.

[12] Tauscher, L. and Greenberg, S. 1997. How people revisit web pages: empirical findings and implications for the design of history systems. International Journal of Human-Computer Studies, Vol. 47, No. 1, 97-137.

[13] Velayathan, G. and Yamada, S. 2006. Behavior-based web page evaluation. In Proc. of the 15th Int. Conf. on WWW (Edinburgh, Scotland, 2006), ACM Press, 841-842.

Príloha B

Výsledky experimentov

V tejto prílohe uvádzame podrobnejšie výsledky niektorých experimentov.

B.1 Vyhodnotenie práce pomocou dotazníka

Používateľov, ktorí vyskúšali stránku fakulty vylepšenú o odporúčané sekcie, sme sa v dotazníku spýtali na ich názor. Otázky mali takéto znenie:

1. Ako často navštevujete stránku FIIT.sk?
   a) niekoľkokrát denne
   b) raz za deň
   c) niekoľkokrát týždenne
   d) raz za týždeň
   e) niekoľkokrát za mesiac
   f) som na upravenej stránke FIIT prvýkrát
2. Všimli ste si pri zapnutom proxy nejaké zmeny na stránke FIIT.sk? Čo pribudlo? Čo ubudlo?
3. Boli pre Vás odporúčané stránky relevantné?
   a) vždy
   b) skoro vždy
   c) častejšie sú ako nie sú
   d) skoro nikdy
   e) nikdy
4. Dozvedeli ste sa vďaka odporúčaniam viac?
   a) áno, o odporúčaných stránkach som nevedel, že na FIIT.sk sú

b) nie, odporúčané stránky som poznal

5. Zefektívnil sa vďaka odporúčaniam prístup k stránkam?

   a) áno, odporúčania mi skrátili cestu (nemusím toľko klikať)

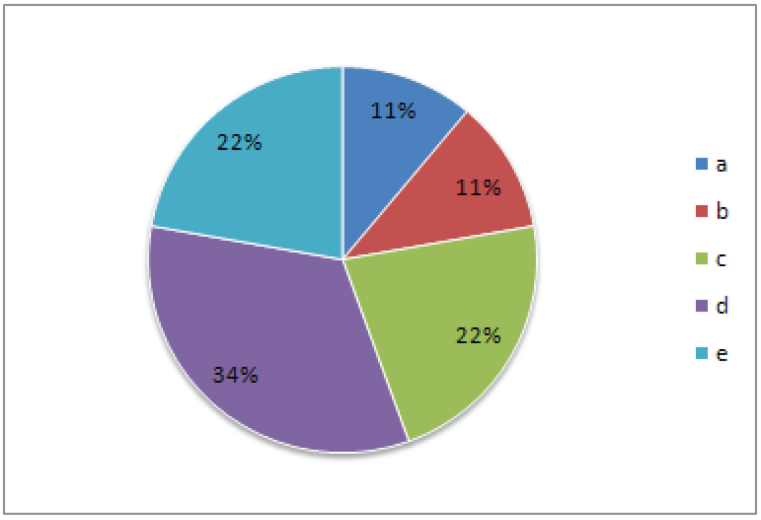
   b) nie, neboli mi odporučené zaujímavé stránky

6. Ako ste spokojný s upravenou stránkou FIIT.sk?

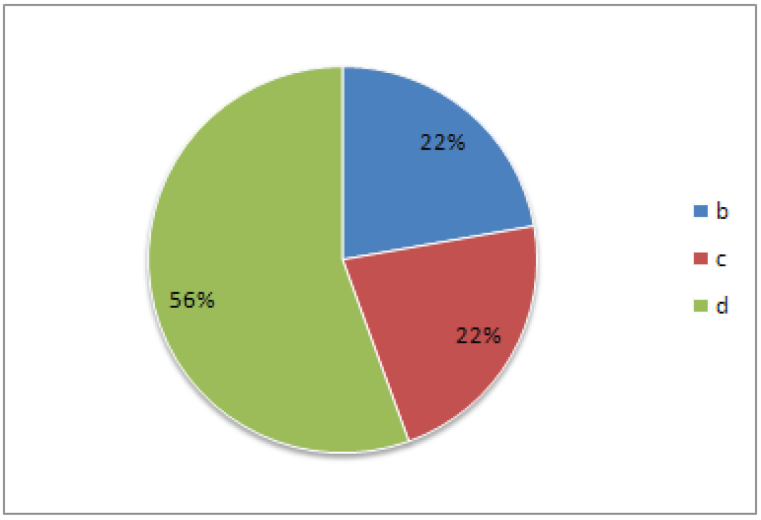
   a) som spokojný a chcem ešte viac odporúčaných sekcií

   b) som spokojný a súčasné odporúčania mi stačia

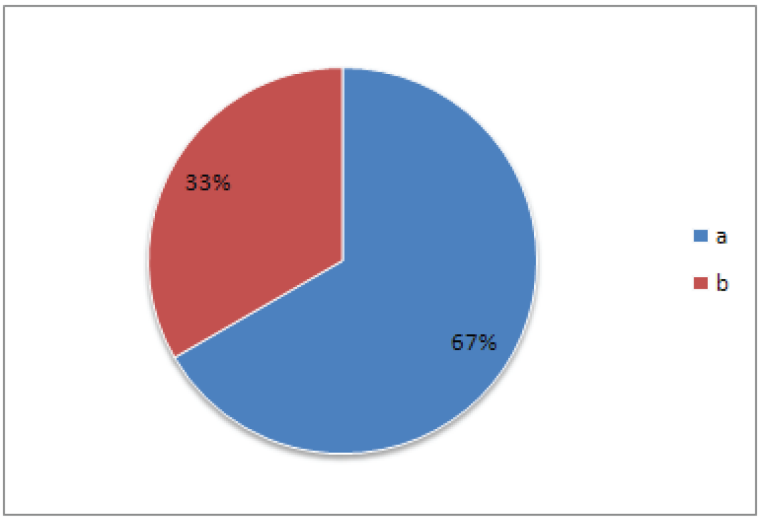
   c) chcem naspäť pôvodnú stránku



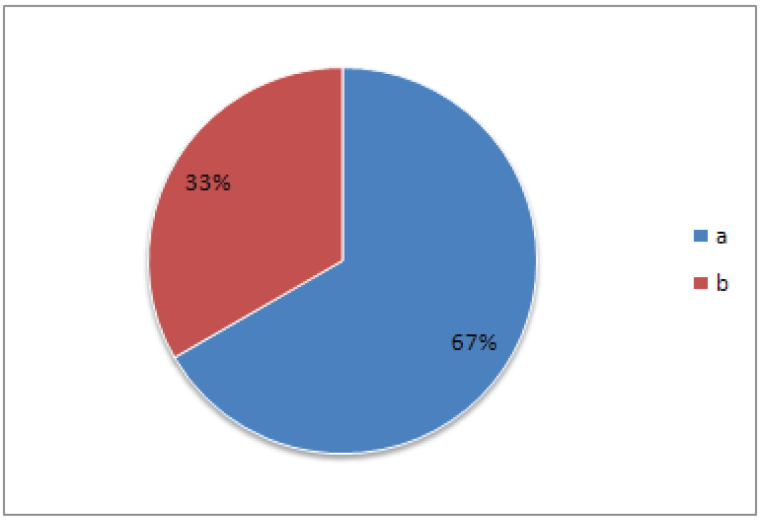
Obr. B.1: Otázka č. 1: Ako často navštevujete stránku FIIT.sk?



Obr. B.2: Otázka č. 3: Boli pre Vás odporúčané stránky relevantné?



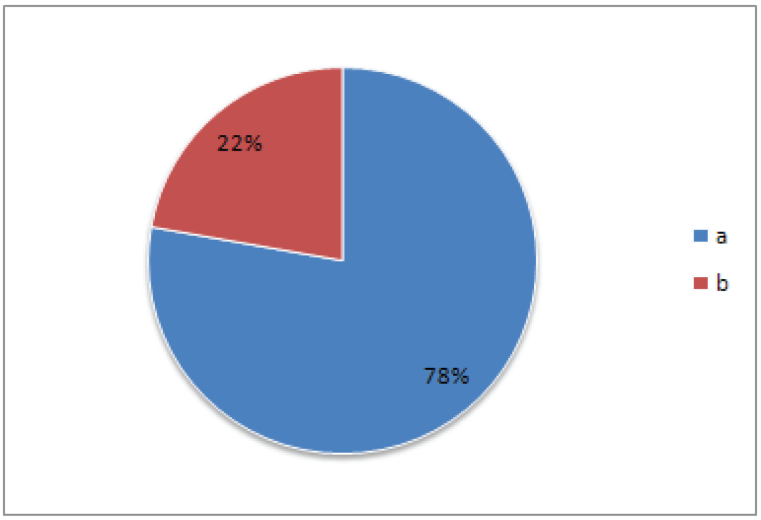
Obr. B.3: Otázka č. 4: Dozvedeli ste sa vďaka odporúčaniam viac?



Obr. B.4: Otázka č. 5: Zefektívnil sa vďaka odporúčaniam prístup k stránkam?

B.2 Experiment s určovaním vzorov v navigácii

V rámci overenia určovania vzorov v navigácii sme vykonali experiment, v ktorom sme identifikovali sedenia všetkých návštevníkov fakultného webového sídla. Z nich sme vybrali tie, ktorých dĺžka bola minimálne 3. Podľa prevažujúceho vzoru sme používateľov rozdelili do piatich kategórí. Hodnoty z experimentu sú uvedené v tabuľke B.1. V niektorých sedeniach sme neidentifikovali žiaden vzor. Napr. postupnosť odkazov používateľa 1 bola A, A, B, C, D, čo nezodpovedá žiadnemu zo vzorov.



Obr. B.5: Otázka č. 6: Ako ste spokojný s upravenou stránkou FIIT.sk?

Tabuľka B.1: Počty identifikovaných vzorov v navigácii používateľov.

návštevník	počet sedení	cesta	kruh	slučka	hrot
1	1				
2	1				
3	9	4		3	2
4	3	1		1	1
5	2	2			
6	1		1		
7	1	1			
8	4	2	1	2	
9	2	1	1		1
10	1	1			
11	1				1
12	12	4	1	1	5
13	1	1			
14	4	1		3	
15	1				1
16	2			2	
17	2			2	
18	1	1			
19	3	2	1		

Príloha C

Technická dokumentácia

V tejto prílohe uvádzame technickú dokumentáciu k implementovanému prototypu.

C.1 Logický dátový model

Väzby medzi jednotlivými entitami implementovaného riešenia sú zobrazené na obrázku C.1. Hlavnými entitami sú *Stránka* a *Návštevník*. Entita *Stránka* vznika pri analýze webového sídla nástrojom *SpyImp*. Webové stránky obsahujú entity *Odkaz*, ktoré sú odkazmi na iné stránky. Pri nich si pamätáme, či je odkaz funkčný alebo nie.

Návštevníka s so stránkou prepájajú *Akcie*. Pre danú stránku môže mať návštevník viac inštancií typu *Akcie*, čo zodpovedá opakovaným návštevám stránky. Pomocou nástroja *AdaptiveImp* vypočítame z akcií návštevníka jeho *Záujem* pre každú navštívenú stránku.

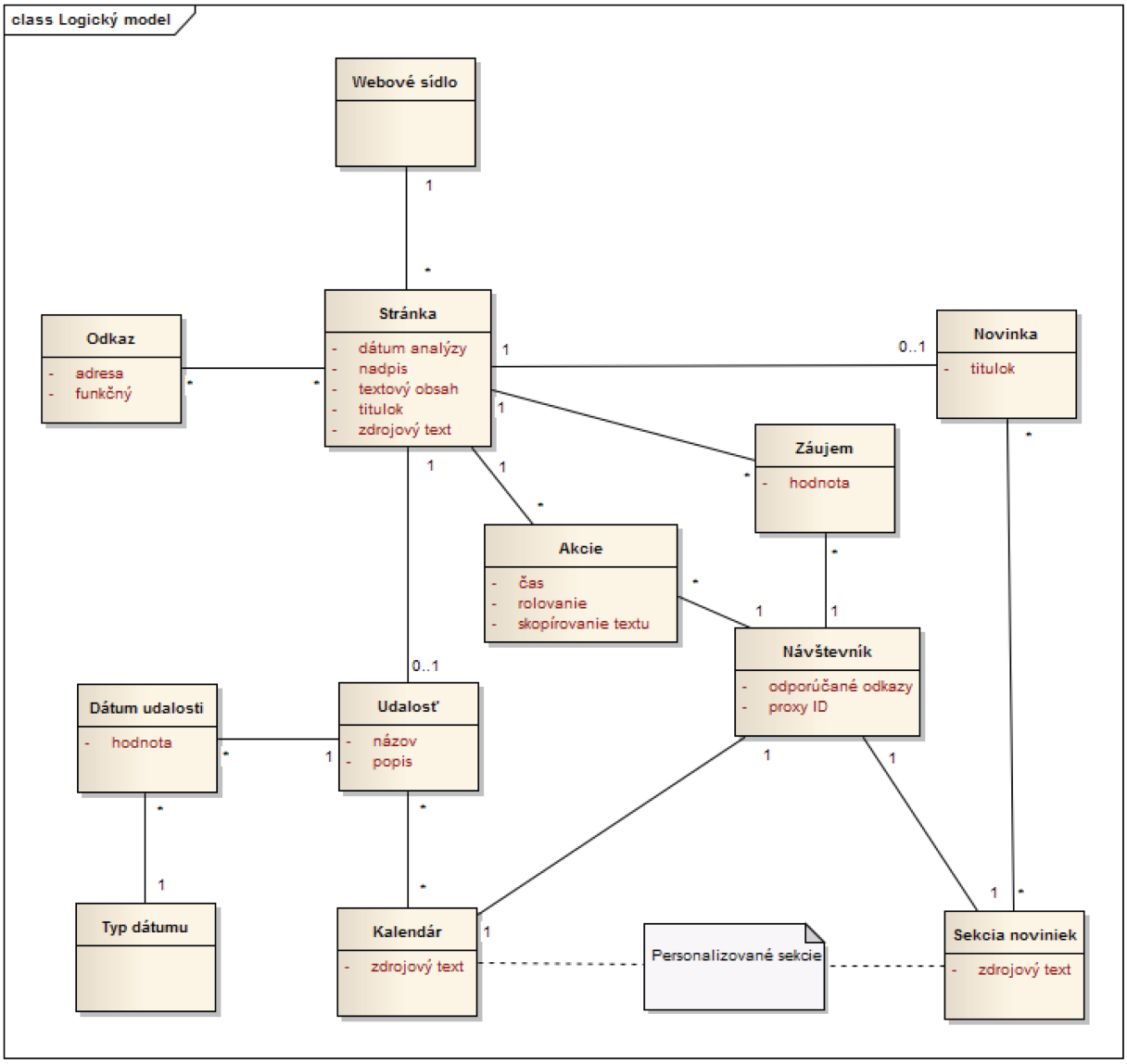
Pri analýze stránky z nej extrahujeme entitiy typu *Dátum udalosti*, ktoré následne prepájame s *Udalosťami*. Každú udalosť prepájame s jednou stránkou, na ktorej bola nájdená. *Stránka* v našom riešení môže informovať najviac o jednej udalosti.

Návštevník má priradený svoj personalizovaný *Kalendár* a *Sekciu noviniek*. V prípade, že nemáme dosť údajov na vytvorenie odporúčaní do týchto sekcií, inicializujeme ich predvolenými hodnotami.

C.2 Dátové úložisko

Ako dátové úložisko používame relačnú databázu MySQL[1]. Tabuľky v nej vytvárame v jazyku Ruby použitím rámca Rails[2]. Ukážka zdrojového textu vytvorenia tabuliek je na obrázku C.2.

[1] http://www.mysql.com
[2] http://rubyonrails.org



Obr. C.1: Logický model znázorňujúci vzťahy medzi jednotlivými entitami.

C.3 Ukážka zdrojového textu algoritmu

V tejto časti uvádzame ukážku implementovaného algoritmu hľadania vzoru *cesta* v zadanej postupnosti odkazov. Algoritmus, ako aj celú aplikačnú logiku prototypu, sme implementovali v jazyku Java. Zdrojový text algoritmu je na obrázku

C.4 Ukážka zdrojového textu kalendára

V tejto časti uvádzame ukážku zdrojového textu personalizovaného kalendára v jazyku HTML. Zdrojový text vychádza z aplikácie Google Calendar[3], ktorej kód sme upravili pre naše potreby. Zdrojový text je generovaný automaticky nástrojom *AdaptiveImp*. Časť zdrojového textu kalendára je uvedená na obrázku C.4.

[3]http://calendar.google.com

```
class CreateWebImpSchema < ActiveRecord::Migration

  def self.up
    create_table :wi_pages do |t|
      t.text :content, :limit => 4.megabytes, :null => false
      t.string :url, :limit => 1024, :null => false
      t.string :heading, :limit => 512
      t.string :title, :limit => 512
      t.timestamp :date_time_accessed, :null => false
    end

    create_table :wi_links do |t|
      t.string :url, :limit => 1024, :null => false
      t.boolean :is_broken, :default => true
    end

    create_table :wi_page_link, :id => false do |t|
      t.integer :page_id, :null => false
      t.integer :link_id, :null => false
    end

    execute "ALTER TABLE wi_page_link ADD PRIMARY KEY (page_id, link_id)"
  end

  def self.down
    drop_table :wi_pages
    drop_table :wi_links
    drop_table :wi_page_link
  end

end
```

Obr. C.2: Zdrojový text vytvorenia databázových tabuliek v rámci Ruby on Rails.

Takúto tabuľku generujeme pre každý mesiac v aktuálnom roku, ktorý nasleduje po aktuálnom mesiaci. Pri dni s odporúčanou udalosťou (9.5.2010) je s udalosťou *onClick* príslušnej bunky spojené volanie funkcie, ktorá zobrazí bublinu s názvom udalosti a odkazom na ňu. Kód tejto bubliny je zobrazený na obrázku C.5. Samotné zobrazenie bubliny realizujeme využitím knižnice *balloon.js*[4].

C.5 Regulárne výrazy na hľadanie dátumov

Dátumy na stránke hľadáme pomocou regulárnych výrazov zapísaných v jazyku Java. Vieme hľadať tri typy dátumov:

1. Jednoduchý dátum zadaný dňom, mesiacom a rokom (napr. 15.4.2010).
2. Rozsah viacerých dní, ktoré zasahujú do viacerých mesiacov (napr. 15.4. – 6.5.2010).
3. Rozsah viacerých dní v rámci jedného mesiaca (napr. 15. - 21.4.2010).

[4] http://gmod.org/wiki/Popup_Balloons

```
/**
 * @param s
 *          list of links from one session ordered by time accessed
 * @return
 *         true when this stream of links represents <i>path<i> pattern,
 *         false otherwise
 */
private boolean isPath(final List<String> s) {
    Set<String> temp = new HashSet<String>();
    for (int i = 0; i < s.size(); i++) {
        if (temp.contains(s.get(i))) {
            return false;
        } else {
            temp.add(s.get(i));
        }
    }
    return true;
}
```

Obr. C.3: Zdrojový text hľadania vzoru *cesta* v zadanej postupnosti odkazov.

4. Rozsah dvoch kompletných dátumov (napr. 15.4.2010 – 6.5.2010).

Regulárny výraz na nájdenie jednoduchého dátumu (typ 1) vyzerá takto (v jazyku Java zadávame regulárny výraz ako textový reťazec, a preto musíme používať jeden znaky \ navyše):

```
\\d{1,2}\\.\\s*\\d{1,2}\\.\\s*\\d{4}
```

Uvedený regulárny výraz využívame aj v ďalších výrazoch, kde ho označujeme ako *DATE PATTERN*. Regulárny výraz na nájdenie dátumu typu 2 vyzerá takto:

```
\\d{1,2}\\.\\s*\\d{1,2}\\.\\s*-\\s*" + DATE_PATTERN
```

Regulárny výraz na nájdenie dátumu typu 3 vyzerá takto:

```
[^.^\\s]\\s*\\d{1,2}\\.\\s*-\\s*" + DATE_PATTERN
```

Regulárny výraz na nájdenie dátumu typu 4 vyzerá takto:

```
DATE_PATTERN + "\\s*-\\s*" + DATE_PATTERN
```

Dátumy môžu byť zadaná buď len pomocou čísiel, alebo kombináciou čísiel pre deň a rok, a textu pre mesiac. V druhom prípade najprv prevedieme slovné vyjadrenie mesiaca na číslo. Rozlišujeme slovenské názvy mesiacov s aj bez diakritiky, rozlišujeme aj mesiac zadaný v inom gramatickom páde. Príklad regulárneho výrazu na nájdenia mesiaca *apríl* vyzerá takto:

```
apr[ií]la?
```

```
<table id="tab_calendar" class="monthtable" cellspacing="0" cellpadding="0">
    <tbody>
        <tr id="tab_header" class="cell tab-heading">
            <td id="month_prev" class="cell">
            </td>
            <td id="dp_0_cur" class="cell month-cur" colspan="5">Máj 2010</td>
            <td id="month_next" class="cell month-next" style="cursor:pointer;"
                onClick="displayNextMonth(4)"> > </td>
        </tr>
        <tr class="dp-days">
            <td class="cell dayh">Po</td>
            <td class="cell dayh">Ut</td>
            <td class="cell dayh">St</td>
            <td class="cell dayh">Št</td>
            <td class="cell dayh">Pi</td>
            <td class="cell dayh">So</td>
            <td class="cell dayh">Ne</td>
        </tr>
        <tr id="day_names_row">
            <td id="day" class="cell day-offmonth"> </td>
            <td id="day" class="cell day-offmonth"> </td>
            <td id="day" class="cell day-offmonth"> </td>
            <td id="day" class="cell day-offmonth"> </td>
            <td id="day" class="cell day-offmonth"> </td>
            <td id="day" class="cell day-onmonth day-weekend">1</td>
            <td id="day" class="cell day-onmonth day-weekend">2</td>
        </tr>
        <tr id="days">
            <td id="day" class="cell day-onmonth day-weekday">3</td>
            <td id="day" class="cell day-onmonth day-weekday">4</td>
            <td id="day" class="cell day-onmonth day-weekday">5</td>
            <td id="day" class="cell day-onmonth day-weekday">6</td>
            <td id="day" class="cell day-onmonth day-weekday">7</td>
            <td id="day" class="cell day-onmonth day-weekend">8</td>
            <td id="day" class="cell day-onmonth day-weekend day-event-important"
                onclick="balloon.showTooltip(event,'load:event09052010', 1);">9</td>
        </tr>
        <tr id="days">
            <td id="day" class="cell day-onmonth day-weekday">10</td>
            <td id="day" class="cell day-onmonth day-weekday">11</td>
            <td id="day" class="cell day-onmonth day-weekday">12</td>
            <td id="day" class="cell day-onmonth day-weekday">13</td>
```

Obr. C.4: Časť zdrojového textu personalizovaného kalendára v jazyku HTML.

```
<div id="event09052010" style="display:none">
    <a href="http://www.fiit.stuba.sk/generate_page.php?page_id=3155">Robo Cup 2010</a>
</div>
```

Obr. C.5: Zdrojový text detailu udalosti v jazyku HTML.

Príloha D

Používateľská príručka

Implementovaný prototyp sa skladá z troch komponentov:

- *AdaptiveImp*,
- *SpyImp* a
- *WebImp*.

Tieto komponenty môžu pracovať nezávisle. K správnemu fungovaniu je potrebný adaptívny proxy server[1], nakoľko prototyp je zásuvný modul do tohto servera. Nastavenie prostredia a spustenie jednotlivých komponentov je možné vykonať nasledovnou postupnosťou krokov:

1. Vytvoriť dátové úložisko a databázu. Tabuľky je možné vytvoriť pomocou zdrojových textov priložených na CD. K tomu je potrebný interpreter jazyka Ruby[2] a rámec Rails[3].
2. Nainštalovať Java[4] prostredie (prototyp bol vyvíjaný pre verziu Java 6).
3. Spustiť analýzu webového sídla nástrojom *SpyImp* príkazom **java -jar spyimp.jar**. V adresári **config** sa nachádza súbor **config.properties**, v ktorom je možné meniť jednotlivé parametre.
4. Spustiť vytvorenie odporúčaní nástrojom *AdaptiveImp* príkazom **java -jar adaptiveimp.jar**. To predpokladá prístup k databáze s údajmi o akciách používateľov na webovom sídle, ktoré zaznamenáva adaptívny proxy server. V adresári **config** sa nachádza súbor **config.properties**, v ktorom je možné meniť jednotlivé parametre.
5. Pridať plugin *WebImp* do adaptívneho proxy servera, ktorý bude upravovať webové stránky vybranej domény pridávaním odporúčaní.

[1] http://peweproxy.fiit.stuba.sk
[2] http://www.ruby-lang.org
[3] http://rubyonrails.org
[4] http://java.sun.com

Príloha E

Obsah elektronického média

Elektronické médium priložené k tejto práci má takúto štruktúru:

/**Guide** – používateľská príručka

/**Papers** – príspevky zaslané na medzinárodné konferencie

/**Prototype** – implementovaný prototyp

/**AdaptiveImp** – nástroj *AdaptiveImp* na výpočet odporúčaných odkazov

/**Database** – zdrojové texty na vytvorenie databázovej schémy

/**Files** – súbory pre webové sídlo

/**Install** – potrebné softvérové vybavenie

/**SpyImp** – nástroj *SpyImp*, ktorý analyzuje vybrané webové sídlo

/**WebImp** – zásuvný modul *WebImp* do adaptívneho proxy servera

/**Thesis** – elektronická verzia diplomovej práce

Každý z nástrojov zahrnutých v prototype obsahuje adresár /**doc** obsahujúci dokumentáciu k tomuto nástroju.