Uživatelská podpora v prostředí WWW

Jiří Jelínek
Vysoká škola ekonomická, Praha
jelinek@fm.vse.cz

INFORUM 2004: 10. konference o profesionálních informačních zdrojích
Praha, 25. - 27.5. 2004

Abstrakt. *Internet a jeho nejpopulárnější služba WWW daly uživatelům k dispozici obrovské množství dat a informací. Problémem se dnes stává ne technická dostupnost, ale schopnost objevit požadovanou informaci. Pro uživatele je často obtížná i orientace v té části WWW prostoru, kterou již navštívil. Příspěvek se proto zabývá možnými postupy, jak uvedené činnosti technicky podpořit a uživateli WWW jeho každodenní práci s tímto médiem usnadnit. Prezentován bude rovněž podpůrný systém, který je budován na Katedře managementu informací FM VŠE.*

1. Úvod

Při práci se službou WWW je hlavním cílem běžného uživatele nalézt informace, které jsou pro něho v danou chvíli zajímavé. Základním nástrojem je využití některé z dostupných vyhledávacích služeb, která však většinou indexuje pouze část prostoru WWW. Kromě toho bývá určitý problém se začínajícími uživateli, kteří nejsou schopni kvalitně definovat své požadavky. Vyhledávání v plných textech stránek přináší i další problémy vycházející ze současné podoby WWW, kdy stránky jsou strojově obtížně zpracovatelné. Nabízí se tedy otázka, jak běžnému uživateli WWW jeho činnost usnadnit.

Efektivitu procesu vyhledávání je možné podpořit několika způsoby. Jedním z nich je redefinice WWW stránek tak, aby je bylo možné automaticky prohledávat s vysokým stupněm úspěšnosti, tj. shody se zadáním. Tento směr bude jistě upřednostňován při tvorbě nových stránek. Již dnes je možné využít nástrojů jako RDF (Resource Description Framework) či OWL, které umožňují vytvořit popis obsahu stránky (znalostí na ní obsažených) formou orientovaných grafů či ontologií. Kvalita vyhledávacích systémů se tím výrazně zvýší, problémem patrně zůstane jejich omezený záběr (z hlediska počtu stránek na WWW). Druhou cestou je podpořit uživatele a zpříjemnit jejich práci se současnou podobou WWW. To však vyžaduje využití poněkud sofistikovanějších nástrojů a metod.

Jádrem většiny podpůrných systémů je sledování běžného chování uživatele, na jehož základě je definován jeho profil a jsou konstruována doporučení a návrhy na další postup, které jsou mu předkládány. Analýza probíhá po tzv. sezeních, tedy časově omezených blocích, během kterých se předpokládá, že uživatel sleduje jediný cíl (např. nalezení informací o konkrétní problematice). Z jejich průběhu lze zjistit základní vzory chování na WWW a modelovat tu část WWW prostoru, kterou již uživatel objevil. Lze rovněž odkrýt podobnosti mezi chováním a zvyklostmi různých uživatelů. Z jednotlivých sezení lze rovněž vysledovat uživatelův konkrétní zájem a popsat ho ve formě slovních termů. Ty je možno získat z textů odkazů, na které je kliknuto, z popisu navštívených stránek (titulek, klíčová slova) i z plných textů WWW stránek. V kombinaci s údaji o propojení WWW stránek lze termy využít pro zkoumání „okolí“ navštívené stránky ve vztahu k vyhledávané problematice. Termy mohou být užity také při vyhledávání na WWW s pomocí některého z dostupných vyhledávačů. Uživateli jsou pak doporučeny další informační zdroje s obdobným zaměřením. Velmi zajímavou možností je významové zpracování termů a jejich zasazení do pojmových hierarchií.

Pro realizaci uvedeného cíle je možné využít tří oblastí Web Miningu. První je použití metod pro sledování pohybu a chování uživatele (tzv. Web Usage Mining), druhým pak oblast analýzy obsahu WWW stránek (Web Content Mining). Třetí cestou je analýza struktury procházeného WWW prostoru (Web Structure Mining).

2. Shrnutí současného stavu

Problém uživatelské podpory v prostředí WWW, založené především na technikách Web Miningu (dále jen WM) je dnes velmi aktuální otázkou, kterou se zabývá značný počet jednotlivců i celých pracovišť. Příkladem mohou být např. práce [Cha00, Gau00, Ger03]. Podpůrné systémy vycházejí z údajů o chování a zájmech uživatele shromážděných v minulosti. Ty jsou získávány obvykle z tzv. logovacích souborů, které mohou pocházet z několika zdrojů:

1. Server WWW – soubory zahrnují údaje o přístupech různých uživatelů na stránky právě tohoto serveru.

2. Klient WWW (prohlížeč) – pro tento účel je potřeba obvykle upravit samotného klienta. Tato metoda umožňuje shromáždit údaje o uživatelích jediného počítače. Zahrnuty jsou přístupy do celého prostoru WWW.
3. Proxy server – ukládány jsou údaje o všech uživatelích daného proxy serveru a všech jejich přístupech do WWW.

Používány jsou zejména postupy uvedené pod body 1) a 3). Přístup 1) je typický pro podnikové systémy a pro business inteligence systémy využívané v elektronickém obchodování. Analýza se soustřeďuje na jediný server, na kterém se zkoumá chování uživatelů. Výhodou je obvykle znalost struktury zkoumaného WWW prostoru (jediný server). Přístup 3) lépe vyhovuje tématu tohoto příspěvku, protože poskytuje všechny potřebné údaje pro podporu konkrétních uživatelů a neomezuje se pouze na jeden WWW server.

Problematika Web Miningu se obvykle dělí na následující podúlohy:

1. Web Usage Mining (dále jen WUM)– analýza přístupů, někdy též označována jako click-stream analýza.
2. Web Structure Mining (dále jen WSM) – analýza struktury WWW prostoru často využívaná pro optimalizaci firemních WWW prezentací či zjišťování tématické podobnosti stránek.
3. Web Content Mining (dále jen WCM) – analýza obsahu WWW stránek. Používané postupy mají úzký vztah ke analýze textových dat nástroji umělé inteligence.

Postupy popisované ve všech třech oblastech WM lze velmi efektivně využít při podpoře uživatele a definici jeho profilu a zájmů. Při zpracování dat se uplatní rovněž klasické metody umělé inteligence a operačního výzkumu, především nástroje shlukové analýzy a metody pro práci s reprezentací dat v podobě grafů.

Techniky Web Usage Mining

Mezi technikami WUM použitelnými pro definování profilu uživatele, lze upozornit např. na metodu FGS (Frequent Generalized Sessions). Jak již název vypovídá, jedná se o postup umožňující z údajů o připojení uživatele k síti extrahovat obvyklé vzory jeho chování. Metoda používá na vstupu soubor vektorů jejichž souřadnicemi jsou navštívená URL seřazená podle pořadí, v němž byla navštívena. Výstupem pak jsou n-prvkové posloupnosti, z nichž každá definuje jeden často se opakující vzor chování uživatele. Ty pak mohou tvořit jednu ze součástí uživatelského profilu. Prvky uvedených posloupností jsou buď konkrétní URL nebo zástupný znak „*“ označující jeden nebo více navštívených URL. Tento zástupný znak nesmí být na první a poslední pozici v posloupnosti. Samotný algoritmus spočívá v iterativní tvorbě množiny posloupností od nejkratších (n=1) po nejdelší (maximální hodnotu n zadává uživatel). Každá posloupnost zařazená do výstupní množiny musí splnit podmínku překročení určitého prahu v počtu výskytů ve vstupních datech. Celý algoritmus lze nalézt podrobně popsaný v [Gau00, Ger03]. Výsledné posloupnosti je možno využít při predikci chování uživatele.

Nástroje Web Structure Mining

Použitelným postupem z oblasti WSM může být konstrukce a analýza hierarchických stromových struktur reprezentujících WWW prostor. Tato analýza může být prováděna na úrovni skupiny uživatelů, jednotlivce či např. pouze jednoho sezení. Podle toho je pak možné zvolit vhodnou metodu dalšího zpracování. Vstupem metody je soubor navštívených URL, výstupem pak jejich uspořádání do stromové struktury pomocí rozkladu URL na jednotlivé logické části. Každá cesta stromem od kořene k listům pak vyjadřuje zápis jednoho unikátního URL. Uzly stromu mohou kromě textu příslušné části URL nést rovněž statistické informace o počtu návštěv, případně údaje o celkové době strávené na stránkách obsahujících danou část URL. Tyto doplňkové údaje mohou být vhodným kritériem pro uplatnění prořezávacích technik. Výsledkem pak je generalizovaný hierarchický model struktury WWW prostoru zobrazující pouze uzly splňující určité podmínky [Jel04]. Kritériem pro generalizaci může být kromě počtu návštěv např. maximální počet částí URL. Model navštíveného WWW prostoru lze užít při porovnávání zájmů uživatelů.

Dalšími daty pro následnou WSM analýzu jsou určitě vazby dané stránky na stránky další. Z tohoto úhlu pohledu lze WWW prostor modelovat jako orientovaný graf, jehož uzly zastupují konkrétní URL a hrany jejich vzájemná propojení pomocí hypertextových vazeb. Pro následnou analýzu mohou být využity metody pro práci s grafy. Ze vzájemných hypertextových vazeb WWW stránek je možné usuzovat na jejich tématickou podobnost Základním výchozím předpokladem je, že ve většině případů hypertextové odkazy propojují stránky mající i obsahovou vazbu. To však nemusí být vždy splněno (např. reklamní odkazy).

Postupy Web Content Mining

Pro oblast WCM se velmi často využívají metody určené pro zpracování dokumentů a to buď přímo nebo s určitými obměnami. Většina metod pro analýzu dokumentů pracuje s pojmem „term” jako základní jednotkou pro popis dokumentu. Jako term jsou většinou označována jednotlivá slova či víceslovná spojení, která jsou v daném dokumentu významně zastoupena, případně jsou pro dokument charakteristická. Pro zpracování textového podkladu se nejčastěji používají metody využívající vektorovou reprezentaci dokumentů, kdy sada

dokumentů a termů tvoří matici, ve které jsou uvedeny hodnoty významnosti termů v jednotlivých dokumentech [Kos03]. Uvedený postup pak umožňuje využít další metody směřující ke klasifikaci dokumentů, jichž je možné využít pro získání dalších relevantních termů a tím i pro vyhledávání dalších informačních zdrojů.

Významnou oblastí výzkumu je též výběr relevantních termů a to jak jednoslovných, tak víceslovných [Cha00]. Relevance je často poměřována tzv. *tf-idf* vahou založenou na frekvenci termů v jednotlivých dokumentech a inverzní frekvenci termu v celém souboru dokumentů [Sch03]. Zvláště pro české jazykové prostředí se pak jeví jako zásadní otázka lematizace termů, tj. postup eliminace různých gramatických variant daného slova či slov. V prostředí WWW lze pro extrakci termů využít nejen prostý text obsažený na WWW stránce, ale i další textové informace, např. z hlavičky stránky (např. titulek, klíčová slova, atd) nebo z jiných URL (texty odkazů ukazujících na danou stránku). Zajímavou cestou pro vyhledávání příbuzných termů může být využití pravděpodobnostních přístupů, např. bayesovských pravděpodobnostních sítí.

Jak již bylo uvedeno v úvodu příspěvku, hlavním cílem uživatele WWW je nalézt relevantní informace. Cestou k jejich objevení může být i zasazení dané stránky podle klíčových termů do již existující a dostupné ontologické struktury a následné zkoumání „nejbližších sousedů“ v této struktuře. Tento postup se v podstatě snaží nahradit chybějící ontologický popis stránky, který se v blízké budoucnosti jistě rozšíří, avšak při velikosti WWW prostoru budou zde uváděné postupy jistě aktuální po nezanedbatelnou dobu. Výsledek ontologického zařazení bývá použit jak pro objevení podobně zaměřených stránek, tak pro shromáždění více termů, které lze využít ve vyhledávání pomocí některé z běžných vyhledávacích služeb. Pro výzkumy v popisované oblasti lze využít především klasických ontologických struktur [WordN, Cyc, Suo], ale i nových nástrojů dostupných na Internetu [Google] (např. rozhraní Google s direktivou define:, které WordNet rovněž zahrnuje, ale seskupuje výsledky z více zdrojů). Jako zajímavá možnost se nabízí i využití experimentální služby Google Sets schopné zobrazit termy mající ontologický vztah k zadanému vzoru. Tento vztah však bohužel není blíže definován.

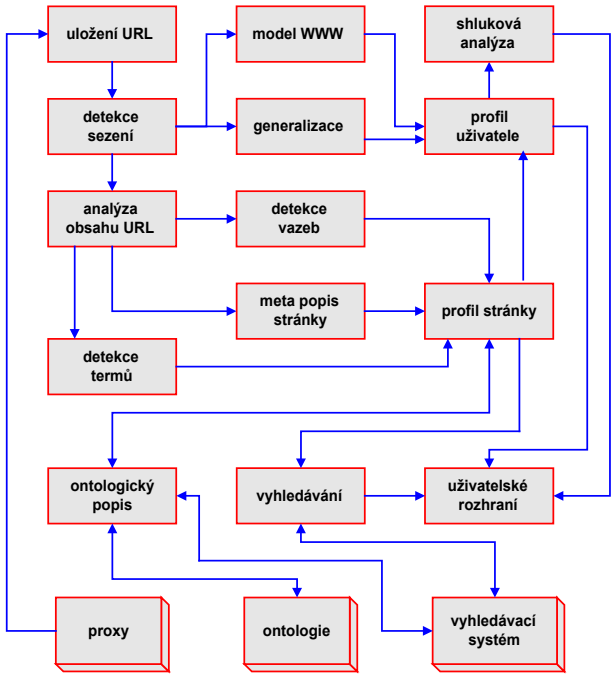
Problémem ontologického přístupu je především velmi obecný záběr stávajících dostupných ontologií, které nejsou prakticky využitelné pro specifické oblasti. Budoucí výzkum bude jistě soustředěn na vývoj tzv. doménových ontologií, které by byly dostupné a přijatelné pro široký okruh uživatelů.

3. Popis systému

Výzkum v oblasti WM na Fakultě managementu VŠE si klade za cíl využít výše obecně popsané metody a to se zaměřením a s ohledem na jednotlivého uživatele. Snahou je navrhnout a prakticky otestovat soubor metod pro zpracování dat z logovacích souborů dostupných pro WWW vycházejících a navazujících na stávající poznatky v oboru a tyto metody postupně implementovat v rámci podpůrného systému pro sledování aktivit uživatele se zaměřením na využití získaných dat pro uživatelovu lepší orientaci v prostředí WWW.

Základními technikami použitými při řešení projektu budou techniky ze všech výše zmíněných oblastí (WUM, WSM, WCM). Pozornost je věnována zejména postupům pro definování profilu uživatelů a navštívených stránek, aby mohly být úspěšně predikovány další kroky uživatele a mohly mu být nabídnuty relevantní doplňkové informace pro další postup. Zamýšlený systém podpory je dále podrobně rozebrán.

Části vytvářeného systému v podstatě kopírují členění uvedené v [Sri00] a postup jejich implementace se shoduje s tokem dat. Prvním úkolem je získání patřičných vstupních údajů, kterými jsou požadavky uživatele na otevření konkrétních URL. Ty jsou získávány z logovacích souborů proxy serveru (viz. [Squid]). Jednotliví uživatelé jsou prozatímně detekováni podle jimi zadaného jména. Vstupní data jsou filtrována tak, aby obsahovala pouze uživatelem skutečně vyžádané odkazy.



Obr. 1: Základní schéma navrhovaného systému

Další poustup zpracování je zaměřen na sestavení profilu uživatele. Především jsou detekována jednotlivá sezení (session). U každého sezení se předpokládá zaměření uživatele na jednu oblast zájmu. Sezení lze detekovat sledováním časových intervalů mezi jednotlivými požadavky, přičemž velikost těchto intervalů nesmí přesáhnout stanovenou hodnotu. Ze získaných dat o sezeních lze sestavit celkový profil uživatele daný jak jeho chováním, tak oblastmi zájmu. Využita může být např. WUM metoda FGS. Limitním kriteriem pro generalizaci může být minimální support daného vzoru. Vzory lze vyhledávat mezi sezeními nebo i v uvnitř jednotlivých sezení. Podstatnou součástí profilu je charakteristika uživatele vytvořená na bázi generalizovaného hierarchického stromu URL odkazů [Bau00, Fu99]. Kritériem pro generalizaci je maximální počet částí URL, případně významová hladina daná procentním podílem na celkovém počtu navštívených URL. Generalizované WWW modely jednotlivých uživatelů lze dále zpracovat metodami shlukové analýzy. Vzdálenost objektů (generalizovaných stromů uživatelů) je odvozena od struktury WWW modelu. Kriteriem kvality rozdělení pro daný počet shluků je poměr průměrné mezishlukové vzdálenosti a průměrné vzdálenosti objektů uvnitř shluku.

Na cíl vyhledávání je možné usuzovat jak ze struktury WWW modelu uživatele, tak analýzou obsahu navštívených stránek. Jejím prvním výstupem může být detekce URL odkazů v příslušné stránce, které lze využít pro tvorbu profilu uživatele. V úvahu budou brány zejména texty URL odkazů.

Dalším výstupem zpracování obsahu stránek jsou jejich plné texty očištěné od tagů jazyka HTML. Text je pak použit pro získání jedno i víceslovných termů, charakterizujících stránku. Kromě toho jsou rovněž zpracovávána metadata stránek, a to především za účelem získání textových informací popisujících stránky. Při zjišťování meta popisu stránky jsou ukládány údaje z hlavičky stránky (titulek, klíčová slova). Významnou roli bude hrát i proces lematizace termů.

Výše uvedené postupy jsou prováděny s cílem definovat profil dané stránky založený především na shromážděných významných termech. Profil stránek navštívených uživatelem bude pak zahrnut do profilu uživatele. Získané významné termy mohou být též použity pro vyhledávání na WWW s pomocí existujících vyhledávacích služeb a také pro zjišťování ontologických vztahů. Výsledkem činnosti by mohlo být zasazení unikátních termů do pojmové hierarchie, uložení výsledku zpět do profilu stránky a přeneseně rovněž do profilu uživatele. Na tomto místě je nutné poznamenat, že tyto pokročilé funkce systému budou v první fázi koncipovány pouze pro stránky v angličtině.

Výsledkem činnosti celého systému pak bude uživateli prezentovaný souhrn dalších informací vztahujících se obsahově k vyhledávanému tématu. Již funkční části systému s uživatelem komunikují prostřednictvím názorného WWW rozhraní jehož obsah je z velké části generován v reálném čase.

4. Další postup

Popisovaný systém je postupně vyvíjen a implementován od podzimu 2003. V současnosti jsou realizovány moduly získávání a filtrování vstupních dat. Dalšími již vytvořenými částmi jsou moduly zaměřené na tvorbu částí uživatelského profilu z modelu WWW a generalizací častých vzorů chování. Rovněž implementovány jsou moduly shlukové analýzy pro zjišťování podobností mezi uživateli a základní koncepce uživatelského rozhraní. V oblasti zpracování obsahu je systém schopen detekovat hypertextové vazby a zjišťovat meta popis stránek. Ve fázi experimentu je forma reprezentace profilu stránek, extrakce termů (jedno i víceslovných) a jejich další ontologické zpracování. Naopak vazba na existující vyhledávací systém pro WWW je již prakticky otestována. Další postup prací bude logicky zaměřen na dosud nerealizované či pouze testované moduly.

Význam navrhovaného a postupně implementovaného systému je možno vidět ve dvou základních směrech. Prvním z nich je detailní seznámení s metodami Web Miningu a jejich tvořivé rozvinutí tak, aby bylo možné je efektivně využít v dalších praktických řešeních. Tato oblast se stále velmi prudce rozvíjí a otevírá se zde tedy značný prostor pro výzkum a vývoj. Druhým přínosem pak bude realizace konkrétního řešení schopného praktického využití běžnými uživateli WWW.

5. Odkazy na použitou literaturu

[Bau00] Baumgarten, M., Büchner, A. G., Anand, S.S., Mulvenna, M.D., Hughes, J.G., Navigation Pattern Discovery from Internet Data, In: User-Driven Navigation Pattern Discovery from Intenet Data, Springer-Verlag, Heidelbert, ISBN: 3-540-67818-2, pp74-91, 2000.

[Bor99] Borges J., Levene M., Data Mining of User Navigation Patterns. In: Revised Papers from the International Workshop on Web Usage Analysis and User Profiling, pp. 92-111, ISBN: 3-540-67818-2, Springer-Verlag London, UK, 1999

[Cha00] Chan P. K.: A non-invasive learning approach. In: Web usage Analysis and User Profiling, pp. 39-55, LNAI 1836, Springer-Verlag London, UK, 2000

[Cyc] Cycorp Makers of the Cyc Knowledge Server for artificial intelligence-based Common Sense. http://www.cyc.com/.

[Fu99] Fu Y., Sandhu K., Shih M., Clustering of Web Users Based on Access Patterns. In: Proceedings of International Workshop on Web Usage Analysis and User Profiling (WEBKDD'99), pp. 21-38, San Diego, CA, USA, 1999.

[Gau00] Gaul W., Schmidt-Thieme L.: Mining Web navigation path fragments. In: Proceedings of the Workshop on Web Usage Analysis and User Profiling (WEBKDD '00), pp. 1-5, 2000, Boston, USA

[Ger03] Géry M., Haddad H.: Evaluation of web usage mining approaches for user's next request prediction. In: (Workshop on Web Information and Data Management), Proceedings of the fifth ACM international workshop on Web information and data management, pp. 74-81, ISBN: 1-58113-725-7, ACM Press New York, NY, USA, 2003

[Google] Google, http://www.google.com/

[Jel04] Jelínek J.: Podpora orientace a vyhledávání v prostředí WWW. In: Poster proceedings of the 3rd annual conference Knowledge 04, pp. 13-16, VŠB-TUO Ostrava Czech Rep., 2004

[Kos03] Koštiaĺ I.: Using Latent Semantic Indexing for Intelligent Information Retrieval. In: Proceedings of the 2nd annual conference Knowledge 03. Ostrava 2003. pp. 321-330. ISBN: 80-248-0229-3.

[Kov02] Koval R., Navrat P.: Intelligent support for information retrieval in the WWW environment. In: Advances in Databases and Information Systems, Lecture Notes in Computer Science 2435, Springer Verlag, Berlin 2002, pp. 51-64, ISSN: 302-9743.

[Mysql] MySQL: The World's Most Popular Open Source Database, http://www.mysql.com/

[Nan00] Nanopoulos A., Manolopoulos Y.: Finding Generalized Path Patterns for Web Log Data Mining. In: Current Issues in Databases and Information Systems, Lecture Notes in Computer Science 1884. Springer-Verlag, 2000, 215-228, ISSN: 0302-9743.

[PHP] PHP Shelve, SourceForge.net. http://sourceforge.net/projects/phpshelve .

[Sch03] Schwarz J. Současný stav a trendy automatické indexace dokumentů. In: Proceedings of the 2nd annual conference Knowledge 03. Ostrava 2003. pp. 212-221. ISBN 80-248-0229-3.

[Squid] Squid Web Proxy Cache. http://www.squid-cache.org/.

[Sri00] Srivastava J., Cooley R., Deshpande M., Tan P.: Web usage mining: discovery and applications of usage patterns from Web data, ACM SIGKDD Explorations Newsletter, Volume 1, Issue 2, January 2000, pp. 12 - 23

[SUO] Standard Upper Ontology Study Group, http://suo.ieee.org/.

[WordN] WordNet, http://www.cogsci.princeton.edu/~wn/.

[Xia01] Xiao J., Zhang Y., and Jia X., Li T.: Measuring similarity of interests for clustering web-users, In: (ACM International Conference Proceeding Series), Proceedings of the 12th Australasian conference on Database technologies, pp. 107-114, ISSN: 1530-0919, IEEE Computer Society Washington, DC, USA, 2001