Vyhledávání častých frází pro generování uživatelských profilů

Petr Grolmus[1], Jiří Hynek[2] a Karel Ježek[1]

[1] Západočeská Univerzita, Univerzitní 8, 306 14 Plzeň, Česká Republika
[2] InSITE s.r.o., Rubešova 29, 326 00 Plzeň, Česká Republika

Abstrakt Systém ProGen (*Profile Generator*) je na Západočeské univerzitě v Plzni navržen a implementován s cílem zdokonalit vyhledávací služby pro vědeckovýzkumné účely. Předpokládá se jeho využití i v dalších oblastech (komerční sféra, vzdělávání, zábava). K tomu se využívá generování uživatelských profilů na základě dokumentů Internetu navštívených uživateli. Získání seznamu těchto dokumentů je podmíněno instalací vytvořeného paketového filtru na klientské stanici – tento způsob stírá morální problém, neboť uživatel není sledován bez jeho vědomí. Navštívené dokumenty jsou staženy off-line a z nich je vytvořen zájmový profil daného uživatele. K nalezení charakteristických frází z dokumentů je použit algoritmus *Suffix Tree Clustering*. Vytvořený uživatelský profil je použit ke splnění primárních cílů systému, kterými je: doporučování dokumentů na Internetu na základě nalezeného uživatelského profilu a vyhledávání doménových expertů.
Klíčová slova: generování profilu, zájmový profil, text mining, STC, suffix tree clustering, expert, vyhledávání, analýza sociálních sítí, doporučovací systém

1 Úvod

Systém ProGen (*profile generator*) je vyvíjen na Západočeské univerzitě v Plzni za účelem automatického generování uživatelského profilu. Nalezené uživatelské profily plánujeme využít pro naplnění dvou primárních cílů: doporučování dokumentů na Internetu a vyhledávání doménových expertů.

Systémy pro automatické generování uživatelských profilů nejsou ve světě elektronických dokumentů žádnou novinkou, nicméně nacházejí praktické uplatnění až v poslední době. Popularitu získávají zejména aplikace pro vyhledávání a doporučování expertů (*recommender systems*, *expert finding systems*). Doporučovací systémy lze obecně dělit na systémy založené na obsahu (*content-based*) a systémy tzv. kolaborativního filtrování. Systémy z první skupiny zjišťují podobnost mezi obsahem položek (např. článků, filmů apod.), které uživatel v minulosti označil za zajímavé a snaží se doporučovat jiné (dosud nenavštívené) položky s podobným obsahem. Využívá se zde algoritmů strojového učení, které klasifikují dostupné materiály jako zajímavé nebo naopak nezajímavé. Systémy z druhé skupiny naproti tomu nepředkládají doporučení na základě profilu uživatelových preferencí, ale podle podobností mezi profilem aktivního uživatele a

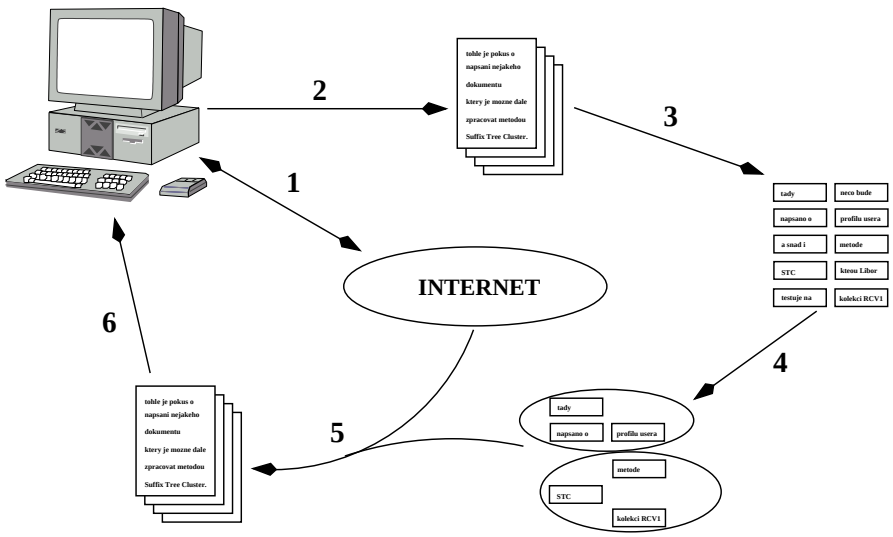
ostatních uživatelů registrovaných v systému. Oba přístupy se pak kombinují v hybridních doporučovacích systémech.

Jedním z prvních realizovaných systémů byl *HelpNet* (Maron a kol., 1986). Jeho odpověďí na žádost o informaci je seznam pracovníků setříděný podle jejich schopnosti zodpovědět dotaz. V poslední době se setkáváme s řadou dalších systémů, např. *Expert Finder* (Vivacqua, 1999) – hybridní doporučovací systém hledající odborníky v předdefinovaných komunitách uživatelů, *Expertise Recommender* (McDonald, Ackerman, 2000) – vyvinutý na University of California v Irvine, *RAAP* (Delgado, 2000) – hybridní doporučovací systém filtrující dokumenty a zajišťující on-line adaptabilitu uživatelských profilů nebo *XPERT-FINDER* (Sihn, Heeren, 2001) – analyzuje e-mailovou komunikaci uživatele a na jejím základě připravuje odpovídající znalostní profil.

Publikované postupy systémů pro automatické vytvoření profilu a generování shluků lze nalézt například v [1]–[4].

2 Popis systému

Struktura systému ProGen je znázorněna na obrázku 1.



Obrázek 1. Struktura systému

Prvním krokem při vytváření uživatelského profilu je sběr informací o uživatelem navštívených dokumentech v prostředí Internetu. Shromažďování těchto základních údajů lze realizovat několikerým způsobem, např. procházením logů WWW serveru (možné pouze v rámci intranetu) nebo proxy serveru (ne všichni

jej používají). Obě uvedené možnosti jsou zcela nevhodné pro univerzitní prostředí, neboť např. studenti nejsou v čase vázáni na konkrétní IP adresu. Převážně z tohoto důvodu systém ProGen využívá pro sběr údajů vlastní aplikaci, mající podobu paketového filtru, který filtruje požadavky uživatele na dokumenty v Internetu. Tato metoda navíc stírá morální problém, který by mohl vzejít z nedobrovolného monitorování uživatele na síti, neboť uživatel si je vědom funkce aplikace a má nad ní plnou kontrolu. Aplikaci může kdykoli vypnout, a zamezit tak poškození svého profilu, pokud by nechtěl, aby se některé dokumenty do jeho profilu začlenily. Nevýhody plynoucí z potřeby zapnutí monitorovací aplikace uživatelem považujeme vzhledem k její triviálnosti za nepodstatnou.

Po uplynutí konkrétního časového období, jenž závisí buď na počtu nasbíraných odkazů na dokumenty nebo velikosti prodlevy od poslední regenerace profilu uživatele, jsou dokumenty identifikované nasbíranými URL adresami načteny a upraveny pro další zpracování. V rámci lexikální analýzy jsou dokumenty převedeny do čistého textu (systém akceptuje dokumenty ve formátech PDF, postscript, MS Word a dalších). Z textu jsou následně vypuštěny nežádoucí nebo nadbytečné znaky (např. nevýznamová slova, značky HTML, interpunkční znaménka apod.) tak, aby každý dokument byl reprezentován jen proudem slov a čísel vzájemně oddělených znakem mezery. Vypouštěná nevýznamová slova dále slouží k automatické identifikaci jazyka zpracovávaného dokumentu. Vodítkem je počet nalezených (různých) nevýznamových slov z jednotlivých jazyků a také četnost jejich opakování. Zbylá slova jsou následně podrobena lemmatizaci prostřednictvím kombinace přesnější slovníkové metody (použit i-spell) a Porterova algoritmu (odstraňování koncovek). Porterův algoritmus je aplikován pouze v případě slov, která nebyla nalezena ve slovníku. Určení jazyka v předchozím kroku nám umožňuje volit správný slovník a také množinu odtrhávaných koncovek pro Porterův algoritmus (obě metody jsou již ze své podstaty silně jazykově závislé, nicméně máme k dispozici slovníky pro všechny hlavní světové jazyky).

Tímto způsobem obdržíme „kolekci“ dokumentů příslušející jednomu uživateli. V této množině dokumentů v dalším kroku vytvoříme pomocí metody STC (*suffix tree clustering*) strom frází (viz např. [5]; metoda STC je popsána dále). Z charakteristických frází vybraných ze stromu STC jsou vytvořeny jejich shluky představující jednotlivé zájmy uživatele.

V nalezených shlucích představují charakteristické fráze klíčová slova, pomocí kterých můžeme vyhledat podobné dokumenty, např. pomocí vyhledávacího robota *Google*. Nalezené dokumenty ProGen následně porovná s uživatelským profilem a pokud považuje tyto dokumenty za relevantní a zároveň je uživatel dosud nenavštívil, doporučí je k prohlédnutí.

Po uživateli nežádáme ohodnocení systémem doporučených dokumentů, neboť uživatele tato činnost zpravidla obtěžuje a případný dotazník v praxi vyplňuje náhodným způsobem, což je kontraproduktivní. Je pouze na uživateli, aby svým chováním sám rozhodl o relevantnosti dokumentu. Pokud doporučené dokumenty navštíví, pak se tyto uplatní při příštím regenerování uživatelského profilu.

2.1 Suffix Tree Clustering

Algoritmus „*Suffix Tree*“ vznikl již v sedmdesátých letech, ale až koncem let devadesátých byl modifikován pro vyhledávání frází a nazván „*Suffix Tree Clustering*“. Publikované články uvádějí složitost metody *O(n)*, kde n představuje počet dokumentů v kolekci. Tuto lineární závislost prokazují experimenty na testovacích kolekcích. Výhodou metody je její nezávislost na pořadí zpracování dokumentů v kolekci. Další výhodou je jazyková nezávislost, která navíc umožňuje bezproblémové zpracování kolekcí obsahujících dokumenty v různých jazycích. V ideálním případě pak získáme zájmové shluky v těchto jazycích.

Postup vytváření stromu STC je velmi jednoduchý. Je nutné předem určit maximální délku hledaných frází. Délka hledané fráze mj. určuje hloubku vytvářeného stromu STC, čímž může významnou měrou ovlivnit paměťovou kapacitu nutnou pro výpočet. Délka fráze také představuje velikost „plovoucího okénka“ ve zpracovávaném textu.

Předpokládejme nyní, že maximální velikost hledané fráze je l. Strom STC vytvoříme takto:

1. Založení stromu (obsahuje pouze prázdný kořen);
2. Ze vstupního dokumentu načteme prvních l slov $w_1 \dots w_l$ (nebo méně, pokud jich již více není);
3. Jednotlivá vstupní slova w_i, kde $i = 1, \dots l$, zařadíme do stromu STC tak, že slovo w_i je umístěno do hloubky i tak, aby cesta od kořenu stromu do daného uzlu vedla přes uzly odpovídající slovům $w_1 \dots w_{i-1}$. Pro jednotlivé uzly (představují fráze) je nutné ukládat čísla dokumentů, ve kterých je daná fráze obsažena;
4. Ze vstupu vyřadíme první slovo a pokud je vstup neprázdný, postup opakujeme od kroku 2;
5. Načteme další dokument a pokračujeme bodem 2.

Např. pro následující tři věty (dokumenty) a maximální délku fráze $l = 3$:

1. Dědek tahá řepu.
2. Bába tahá dědka.
3. Bába také tahá řepu.

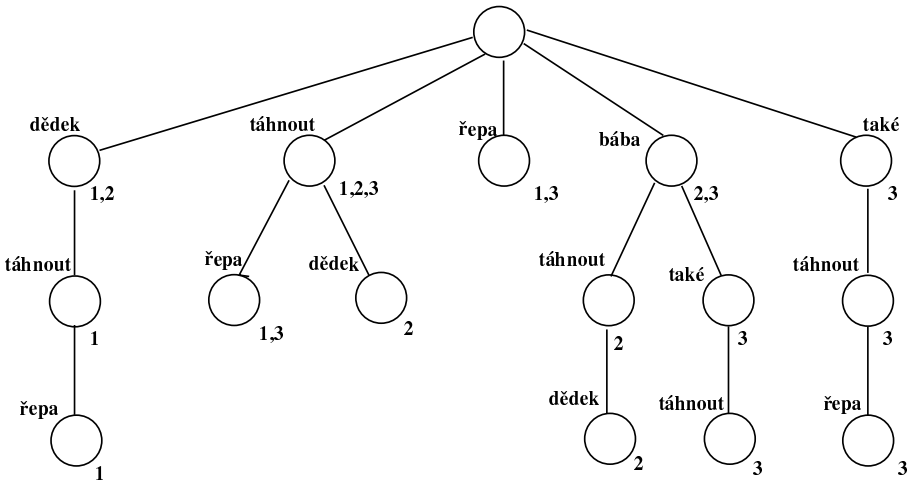
bude strom vygenerovaný po lemmatizaci mít tvar uvedený na obrázku 2.

2.2 Algoritmus generování shluků

Ve vzniklém stromu STC určíme váhu každé fráze takto:

$$w(f_i) = L(f_i) \times N(f_i)^2 \times \sum_{j=1}^{m} (tf_{ij} \times log \frac{m}{df_i}),$$

kde $L(f_i)$ je délka zpracovávané fráze, $N(f_i)$ počet výskytů fráze, tf_{ij} je počet výskytů fráze f_i v dokumentu j, m je celkový počet dokumentů v kolekci a df_i je



Obrázek 2. Příklad stromu STC

celkový počet dokumentů obsahujících frázi f_i. Umocnění $N(f_i)$ má za následek zvýhodnění méně často zastoupených delších frází.

Ze vzniklého stromu STC vybereme p charakteristických frází s největší vahou. Výsledný počet charakteristických frází určíme takto:

$$p = r \times \frac{s}{m} + \frac{m}{k},$$

kde m je počet zpracovávaných dokumentů, s je počet všech výskytů slov ve všech dokumentech, podíl $\frac{s}{m}$ tudíž představuje průměrný počet slov v jednom dokumentu. r určuje číselnou konstantu, která odpovídá procentuálnímu zastoupení průměrné délky dokumentu vůči celkovému počtu všech slov s. K výsledku připočítávaná konstanta $\frac{m}{k}$ zvyšuje vybraný počet charakteristických frází právě o jednu na každých k dokumentů v kolekci.

V takto získané množině charakteristických frází hledáme v dalším kroku podobnosti těchto frází pro vytvoření shluků frází, které představují jednotlivé zájmy uživatele. Nutno předeslat, že počet těchto shluků není předem dán, neboť nelze jakkoli odhadnout počet oblastí zájmu jednotlivých uživatelů.

Podobnost frází je měřena vždy pro dvě různé fráze na základě množiny dokumentů obsahujících obě fráze:

$$(\frac{|D_m \cap D_n|}{|D_m|} \geq \phi) \wedge (\frac{|D_m \cap D_n|}{|D_n|} \geq \phi)$$

kde ϕ určuje stanovenou minimální mez podobnosti dvou frází, D_n množinu dokumentů obsahující frázi f_n a D_m množinu dokumentů obsahující frázi f_m.

Pokud si představíme jednotlivé fráze jako uzly a podobnost dvou frází jako hranu neorientovaného grafu, pak lze hledání shluků převést na klasickou úlohu

z teorie grafů – hledání souvislých komponent grafu. Nalezené shluky představují jednotlivé oblasti zájmu uživatele.

3 Experimenty

Vzhledem ke skutečnosti, že v současné době nemáme statisticky významný počet testovacích uživatelů, prováděli jsme testování navrženého systému na následujících kolekcích – anglické Reuters Corpus Volume One (RCV1) a české kolekci poskytnuté Českou tiskovou kanceláří (ČTK).

	ČTK	RCV1
počet dokumentů	130 955	806 791
počet slov	29×10^6	193×10^6
průměrná délka dokumentu	159,0	88,4
délka nejkratšího dokumentu	10	10
délka nejdelšího dokumentu	5 721	3 996
průměrný počet tříd zařazení jednoho dokumentu	1,7	3,2
počet tříd	42	103

Tabulka 1. Srovnání testovacích kolekcí

Testování jsme prováděli pro obě kolekce shodně. Nejprve jsme z vybraných tříd kolekcí vybrali 2/3 dokumentů, ve kterých jsme za pomoci metody STC nalezli charakteristické fráze a v nich shluky. Tato množina dokumentů simuluje dokumenty navštívené uživatelem. Zbylou 1/3 dokumentů jsme smíchali se shodným počtem dokumentů z jiných tříd – tato množina pak simuluje dokumenty nalezené vyhledávacím robotem a obsahuje dokumenty relevantní i nerelevantní z pohledu uživatelského profilu.

Následně jsme pro každý kandidátní dokument na doporučení vypočítali jeho podobnost s uživatelským profilem. K porovnání byl vytvořen vektor frází jednotlivých shluků a pro každý dokument byl tento vektor naplněn hodnotami odpovídajícími počtu výskytů dané fráze v porovnávaném dokumentu (tedy i 0). Z takto určeného vektoru lze následně určit podobnost dokumentu D se shlukem S_i profilu pomocí kosinové míry:

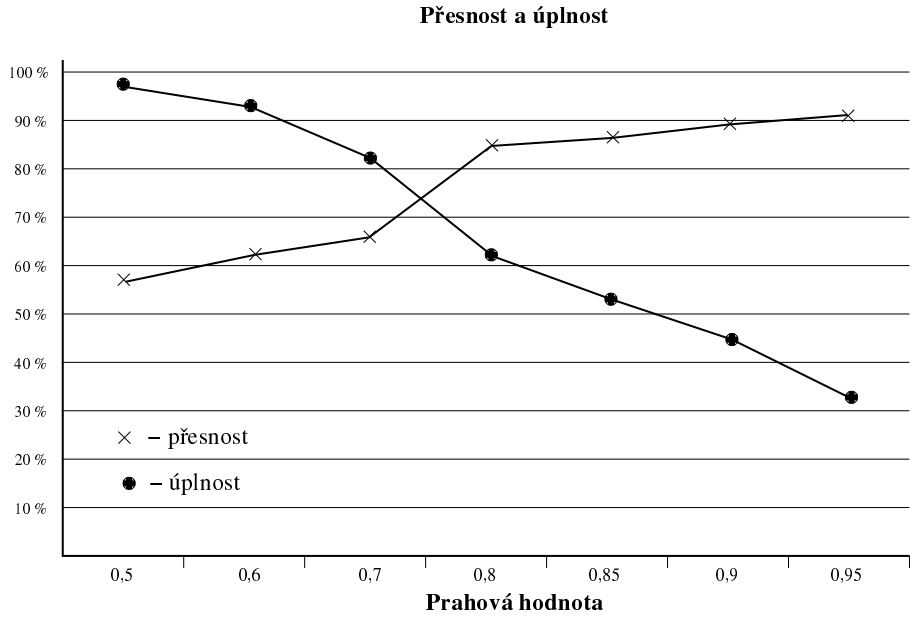
$$Sim(S_i, D) = \frac{\sum_1^H (w_h \times d_h)}{\sqrt{\sum_1^H (w_h)^2 \times \sum_1^H (d_h)^2}},$$

kde H je počet charakteristických frází shluku S_i, w_h udává váhu h-té fráze a d_h počet opakování h-té fráze v dokumentu D. Pokud vypočítaná míra podobnosti překročí stanovenou prahovou hodnotu τ alespoň pro jeden shluk, pak rozhodneme, že jde o dokument relevantní pro daného uživatele.

Podle počtu vybraných relevantních (V_R), nevybraných relevantních (N_R) a vybraných nerelevantních (V_N) dokumentů lze snadno určit přesnost (*precision* – P) a úplnost (*recall* – R) použité metody pomocí následujících vzorců:

$$P = \frac{V_R}{V_R + V_N} \qquad R = \frac{V_R}{V_R + N_R}$$

Výsledky pro různá nastavení prahové hodnoty τ lze vidět na následujících grafech – pro kolekci ČTK a kolekci RCV1 na obrázcích 3 a 4. Podotýkáme, že počítaná úplnost (R) má smysl pouze v případě, že simulujeme chování uživatele kolekcí dokumentů.



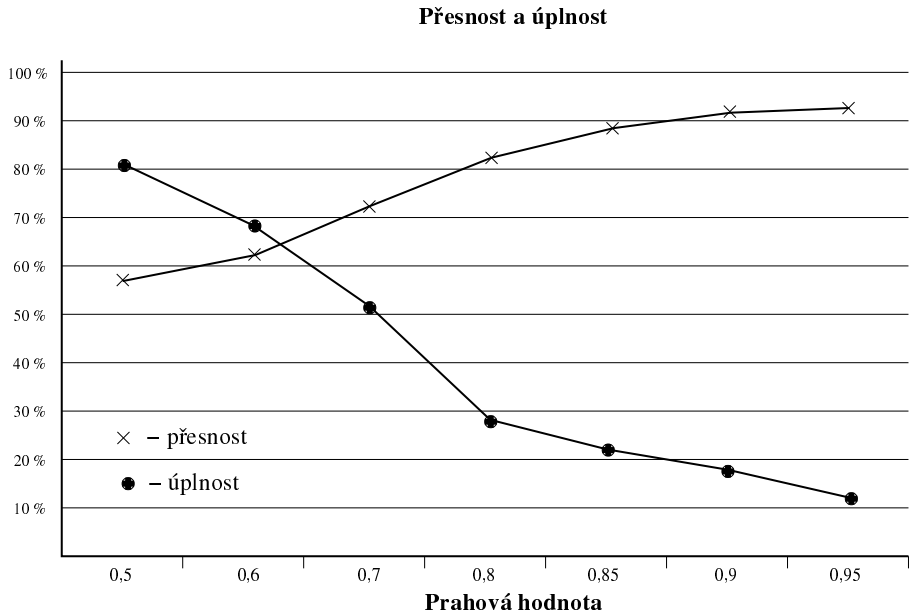
Obrázek 3. Přesnost a úplnost v české kolekci ČTK

Jak je z grafů patrné, dosažené výsledky se pro stejné prahové hodnoty pohybují ve shodných rozmezích pro obě kolekce.

3.1 Nalezené shluky

Příklad nalezených shluků lze demonstrovat na následujících výsledcích získaných z tématu „Politika“ z české kolekce ČTK:

- **C1**: irák, irácký
- **C2**: bělehrad, jugoslávský, kosovský, srbský, albánie, kosovský albánie



Obrázek 4. Přesnost a úplnost v anglické kolekci RCV1

- **C3**: izrael, izraelský
- **C4**: tiskový konference, konference, tiskový
- **C5**: moskva, ruský, rusko
- **C6**: zahraniční, ministr zahraničí

Vzhledem k nalezeným shlukům je nutné upozornit, ze kolekce ČTK obsahuje tiskové zprávy za rok 1999 – což je doba, kdy na území bývalé Jugoslávie probíhaly boje.

Charakteristiky zpracování uvedeného příkladu shluků je možné nalézt v následující tabulce:

počet dokumentů	6 362
slov v dokumentech	1 003 072
uzlů stromu STC	1 397 413
doba zpracování	cca 10 minut

4 Plánovaná rozšíření systému

Důležitým rozšířením popisovaného systému ProGen je zavedení „stárnutí“ profilu. Zájmy každého člověka se s časem vyvíjí a mění, a proto je nutné tomu přizpůsobit i generovaný profil tak, aby systém nedoporučoval uživateli dokumenty, o které již nemá zájem.

Zajímavé by jistě také bylo modifikovat nějaký stávající vyhledávací systém tak, aby nalezené dokumenty řadil s ohledem na relevantnost dokumentu vůči profilu uživatele, který vyhledávací dotaz zadal.

Závěrem bychom rádi poděkovali České tiskové kanceláři za poskytnutí kvalitní české kolekce dokumentů pro testovací účely.

Reference

1. Philip K. Chan – Constructing Web User Profiles: A Non-invasive Learning Approach; příspěvek konference Web Usage Analysis and User Profiling – International WEBKDD'99 Workshop San Diego, USA
2. R.Koval, P.Návrat – Intelligent Support for Information Retrieval in the WWW Environment; příspěvek konference Advances in Databases and Information Systems – 6^{th} East european Conference, ADBIS 2002, Bratislava, Slovensko
3. C.C.Chen, M.C.Chen, Y.Sun: A Web Document Clustering: A Feasible Demonstration; `http://ants.iis.sinica.edu.tw/was/publication.htm`
4. C.C.Chen, M.C.Chen, Y.Sun: PVA: A Self-Adaptive Personal View Agent; `http://ants.iis.sinica.edu.tw/was/publication.htm`
5. Oren Zamir, Oren Etzioni – Web Document Clustering: A Feasible Demonstration; nalezeno na CiteSeer – `http://citeseer.org/` (září 2003)