MASARYKOVA UNIVERZITA

Fakulta Informatiky

BibTip – rekomendačný systém

Bc. Róbert Michale

30.11.2009

1 Úvod

Vo všeobecnosti, rekomendačný systém tvoria špecifické typy techník filtrovania informácií, za cieľom podať informáciu (text, hudba, video, kniha...), ktorá by mohla byť pre užívateľa zaujímavá. Takýto systém je veľmi užitočným nástrojom pre vkladanie referenčného komponentu do katalógov a výrazne pomáha vytvárať portály užívateľsky-orientovaných katalógov. Je založený na štatistických modeloch a pomáha nasmerovať užívateľov od jedného záznamu k ďalším podobným záznamom. Z technického hľadiska môže byť braný ako WEB 2.0 aplikácia, pretože užívateľ nepriamo vytvára dáta, ktoré systém používa a ich integrácia do katalógu prebieha formou mashup-u (dva aspekty WEB 2.0 aplikácií).

Jedným z týchto systémov, používaný v prostredí verejne prístupných online knižníc (OPAC), je systém BibTip.

2 História vzniku

V rokoch 2002 až 2007 bežalo na univerzite v Karlsruhe (Nemecko) niekoľko projektov financovaných Nemeckou Nadáciou pre výskum, zameraných na vývoj rekomendačného systému pre knižnice. Funkčná verzia tohto systému dostala názov Karlsruher recommender system, ktorý tvorcovia zmenili na BibTip v roku 2007.

Na projekte spolupracovala knižnica Univerzity Karlsruhe a Inštitút pre Informačné Služby a Elektronický Obchod (Institute for Information Services and Electronic Markets) vedený Prof. Dr. Andreasom Greyer-Shulzom. Inštitút vytvoril algoritmy a systematický základ rekomendačného systému. Univerzitná knižnica mala za úlohu integrovať systém do existujúceho katalógu, získavanie štatistických dát, ale hlavne vývoj a prevádzku BibTip ako služby. V praxi tieto projekty dosiahli úspech a boli predstavené v decembri 2007 na zasadnutí Koalície pre Informačné Siete CNI vo Washington DC.

3 Ako funguje BibTip

BipTip rekomendačný systém je založený na modeloch správania sa užívateľov v interakcii s katalógom informácií knižnice. Táto takzvaná „implicitná“ rekomendačná služba spočíva na pozorovaní užívateľského správania sa a štatistického vyhodnotenia použitých dát. Všetky uložené a spracovávané dáta sú anonymné vo forme identifikačných čísiel a ID sedení.

3.1 Architektúra

Z technického hľadiska môže byt architektúra BibTip popísaná ako agent-architektúra skladajúca sa z troch častí :

1. pozorovací agent integrovaný v OPAC
2. agregačný agent
3. rekomendačný agent

Pozorovací agent sleduje výber titulkov v rámci definovaných sedení v OPAC. Tieto dáta sa prenesú agregačnému agentovi, ktorý následne vykoná výpočty na štatistických materiáloch, za cieľov vytvoriť zoznam rekomendácií. Tento zoznam je potom prezentovaný užívateľovi rekomendačným agentom.

3.2 Algoritmy

Zatiaľ čo v oblasti internetových obchodoch sú dáta generované kúpnymi transakciami alebo klikmi na odkazy stránky, v BibTipe sú dáta generované agregáciou požiadaviek na zobrazenie titulu v plných detailoch počas daného sedenia. Tieto užívateľské zobrazenia sú spočítávané a na základe sedení sú vytvorené páry spoločného výskytu titulov. Podmienkou vytvorenia tohto páru je zobrazenie dvoch titulov naraz v rámci aspoň jedného sedenia.

Tieto páry sú spočítávané a sumarizované v matici spoločného výskytu. V ďalšom kroku je táto matica vyhodnotená za účelom vytvorenia rekomendácií. V tomto kroku hrajú algoritmy kľúčovú rolu. Algoritmické procesy na ktorých je BibTip založený sú odolné voči „rušeniu“ aj vďaka tomu, že boli vytvorené špeciálne pre dáta v katalógoch knižníc.

Postup algoritmu zjednodušene:

- pre každý titul X, ktorý bol zobrazený v plných detailoch, je vytvorená „história zobrazení“. je to jednoduchý zoznam sedení, v ktorých bol titul X zobrazený
- titul X je porovnaný so všetkými ďalšími Y titulmi, ktoré boli zobrazené v tom istom sedení ako X. Pre každý nájdený pár sa vytvára sekundárna „história zobrazení“
- počet užívateľov, ktorí zobrazili titul X a ďalší titul Y v tom istom sedení je štatisticky analyzovaný. Z tejto analýzy sa vypočíta pravdepodobnosť výskytu spoločného zobrazenia titulov X a Y
- rekomendácia pre titul X je vytvorená vtedy, keď titul Y bol spolu s X v jednom sedení zobrazený častejšie ako sa dá predpokladať z pravdepodobnosti náhodného výberu

Základ algoritmov systému BibTip je Repeat-buying Teória vyvinutá Andrewom Ehrenbergerom. Je to veľmi úspešná a dobre otestovaná štatistická koncepcia, ktorá popisuje pravidelnosti opakovaných nákupov u zákazníkov v rozsiahlych periódach času.

Použitie základných teoretických rozdelení nezávislých procesov v Poissonových sústavách, s použitím logaritmických distribučných algoritmov dáva rekomendačného systému 100% automatizáciu, vysokú presnosť a robustnosť voči odchýlkam.

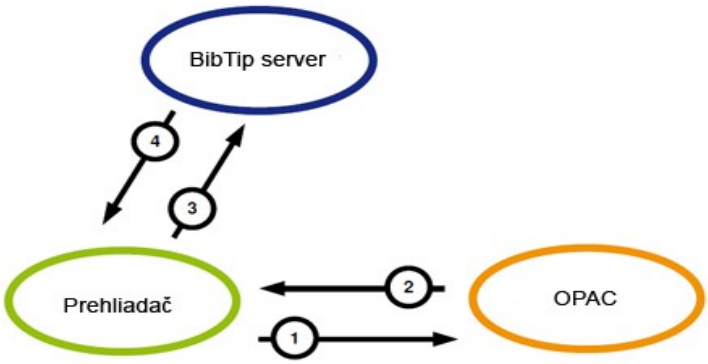
Na obr. 1 je znázornená spolupráca OPAC, klientského prehliadača a serveru BibTip.

1. prehliadač zadá požiadavku na zobrazenie titulu v plnom zobrazení
2. OPAC odošle prehliadaču plné zobrazenie titulu spolu s integračnými kódmi BipTip
3. integračné kódy v klientskom prehliadači iniciujú požiadavku na BipTip server
4. BipTip server prijme požiadavku a po ukončení výpočtov sa generované rekomendácie objavia vo forme hyperlinkov v klientskom prehliadači

Tieto hyperlinky navedú užívateľov na obsahovo-príbuzné tituly zoradené podľa miery relevancie k zobrazenému titulu.

Týmto spôsobom sa tiež získavajú štatistické dáta. Požiadavky putujúce z klientského prehliadača

so sebou nesú všetky ID sedenia, interné ID titulu či ISBN, na základe ktorých je táto interakcia zaradená a následne spracovávaná. Tieto dáta sú v reálnom čase využívané na neustálu aktualizáciu dát pre rozhodovacie procesy pri tvorbe rekomendácií. Je teda možné, že pri danom titule sa po uplynutí istého času objavia úplne iné rekomendácie ako dôsledok zmeny v požiadavkách užívateľov.



Obr.1 – spolupráca OPAC, klientského prehliadača a serveru BibTip

4 Použitie v praxi

BibTip tak ako každý podobný systém má svoje výhody, nevýhody a obmedzenia svojho využitia v praxi. BiBTip funguje ako internetová služba, ktorá je poskytovaná na diaľku za ročný poplatok. Nevyžaduje inštaláciu žiadneho ďalšieho softvéru na strane klienta (OPAC), do ktorého je integrovaná. Analýza štatistických dát, sprostredkovanie a administrácia rekomendačného systému je poskytovaná servermi Univerzity Karlsruhe. Ročný poplatok okrem iného slúži aj na udržiavanie serverov a financovanie ďalšieho vývoju systému.

Systém BibTip je v prevažnej väčšine využívaný knižnicami na území Nemecka. Medzi jeho klientmi sa ale postupne objavujú aj knižnice z iných krajín ako napr. Knižnica Bostonskej Univerzity (USA), či od septembra 2008 aj Vědecká knihovna v Olomouci (ČR). Zatiaľ sa jedná len o akademické a verejné knižnice, no do budúcnosti sa predpokladá využitie BibTip aj v súkromnom sektore.

Rozširovanie systému do ďalších knižníc brzdia vysoké náklady na prevádzku serverov, udržiavanie databáz dát a s nimi spojené ročné poplatky za používanie systému BibTip. Tieto poplatky sa určujú na základe rozsahu (počtu spravovaných titulov) danej knižnice. Najvyššie poplatky majú národné knižnice, ktoré platia ročne približne 4000 €, Vědecká knihovna v Olomouci platí za využívanie týchto služieb približne 2000 € ročne.

4.1 Výhody

Jednou z hlavných výhod použitia BibTip v OPAC je fakt, že rekomendácie sa nikdy nestanú zastaralými. Narozdiel od záznamov v schémach s tradičnou klasifikáciou ako napr. Dewey, rekomendácie v BibTip sú neustále prepočítavané a dynamicky prispôsobované zmenám v použití literatúry užívateľmi knižnice. Pokiaľ začnú užívatelia od istého bodu používať knihy s rozdielnym kontextom ako doteraz, odrazí sa to aj na zozname rekomendácií poskytnutých BibTipom.

Keďže počet rekomendácií daného titulu reprezentuje mieru jeho používania, rekomendačný systém

môže pomôcť pri riešení otázok obsahu knižnice. Pokiaľ nemá daný titul žiadnu rekomendáciu, je zrejmé že pre užívateľov nie je príliš zaujímavý. Na druhú stranu, veľké množstvo rekomendácií daného titulu ukazuje jeho obľúbenosť a časté používanie užívateľmi.

Rekomendácie sú taktiež médium-neutrálne, to znamená, že systém ich vytvára pre položky katalógu bez ohľadu na to, či sú to knihy, videá, zvukové záznamy a pod.

Za zmienku určite stojí aj efektivita nákladov na prevádzku služby BibTip. Po prvotnom zavedení a integrácii totiž nevyžaduje žiadny zásah ani správu od pracovníkov knižnice.

4.2 Nevýhody

BibTip, tak ako ďalšie rekomendačné systémy pracuje tým lepšie, čím rozsiahlejšiu databázu informácií má, pretože štatistické dáta sa vytvárajú až na základe dostatočne veľkého počtu transakcií. Z toho dôvodu je potrebný istý čas pre sledovanie, zbieranie a následnú analýzu dát, pred tým ako sa objavia prvé rekomendácie. Táto časová perióda je tým kratšia, čím viac dát je k dispozícii, čo priamou úmerou súvisí s tým, ako často je katalóg používaný užívateľmi.

Problematické miesto tvorí skupina titulov, ktoré sú veľmi špecifické a ich použitie užívateľmi je veľmi ojedinelé, napriek tomu, že obsahujú relevantný a pre užívateľa atraktívny obsah. Zaradenie týchto titulov do rekomendácií vyžaduje pomerne dlhú periódu času.

Tieto nevýhody však môžu byť čiastočne riešené implementáciou štatistických dát z inej knižnice používajúcej systém BibTip. Podmienkou je ale rovnaký profil užívateľov knižnice a jej obsahové zameranie. Napr. Verejné knižnice, alebo knižnice fakúlt s rovnakým vedeckým zameraním.

4.3 Technické aspekty

Existujú dve formy integrácie BibTip a to štandardná integrácia pomocou ID titulu a zjednodušená integrácia pomocou ISBN.

Spôsob využívajúci ISBN umožňuje integráciu pomocou pridania dvoch riadkov statického HTML kódu do časti s plným zobrazením titulu. V tomto prípade statický znamená, že sa tieto dva riadky kódu opakujú v každom plnom zobrazení titulu. Integrácia sa skladá z vloženia SCRIPT tagu pre načítanie skriptu zo serveru BibTip a z DIV tagu pre zobrazenie rekomendácií

Spôsob využívajúci ID titulu vyžaduje ďalšie tri riadky kódu, ktoré ale obsahujú dynamické časti, ktoré sa menia v závislosti na zobrazenom texte. Integrácia sa skladá z dvoch častí, ktoré používa ISBN integrácia a k nim pribudne DIV tag pre ukladanie hodnôt interného ID titulu, ISBN a skráteného názvu titulu.

Príklady použitia integrácií pomocou:

- ISBN

```
<body>
	Content Full title
	<div style="display:none" id="bibtip_reclist"></div>
	Content Full title
	<script src="http://recommender.ubka.uni-karlsruhe.de/js/bibtip_xxx.js"
		type="text/javascript"></script>
</body>
```

- ID titulu

```
<body>
	Content Full title
	<div id="bibtip_isxn" style="display:none">3-8266-1762-2,978-3-8266-1762-1</div>
	<div id="bibtip_shorttitle" style="display:none">Ajax ge-packt /
		Seeboerger-Weichselbaum, Michael; 2007</div>
	<div id="bibtip_id" style="display:none">26214927</div>
	Content Full title
	<script src="http://recommender.ubka.uni-karlsruhe.de/js/bibtip_xxx.js"
		type="text/javascript"></script>
	Content Full title
	<div style="display:none" id="bibtip_reclist"></div>
	Content Volltitel
</body>
```

Požiadavky pre použitie systému BibTip:

- Webový OPAC s rýchlym a stabilným pripojením k internetu (z dôvodu kooperácie so servermi BipTip)
- konfigurovateľná časť OPAC, ktorá obsahuje plné zobrazenie titulu (operátor OPAC môže meniť HTML kód stránky plného zobrazenia)
- permanentné hypertextové linky na stránku plného zobrazenia daného titulu buď pomocou ISBN, alebo interného ID titulu

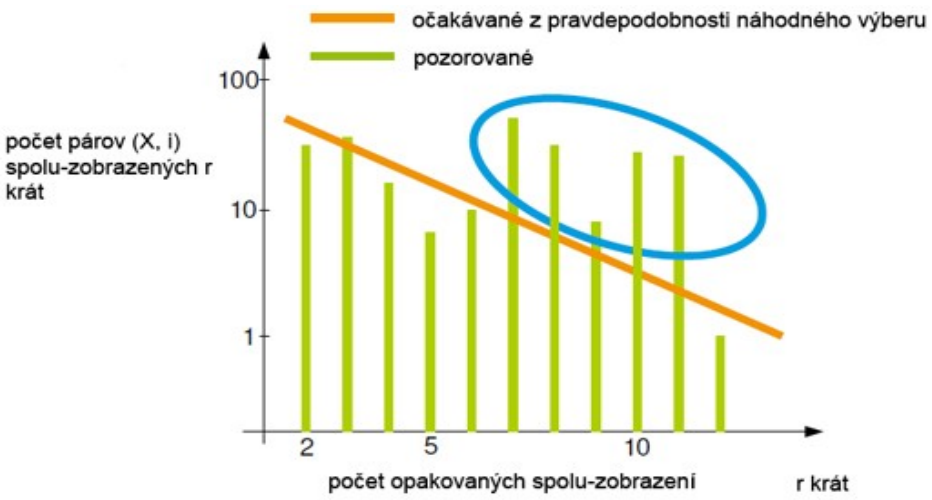
Veľké objemy dát a množstvo súčastne prebiehajúcich výpočtov, ktoré genereuje používanie služby niekoľkými stovkami či tisícami užívateľov sa odrazilo aj na technologickej náročnosti systému.

- použitý hardvér: 4x DELL PowerEdge servery, každý s 4x dual-core procesormi, 16GB operačnej pamäte a 2 TB RAID5 diskovým priestorom
- operačný systém: Ubuntu Linux (oficiálna distribúcia podporovaná Canonical Ltd.)
- virtualizácia: XEN
- databázy: PostgreSQL
- webové servery: Apache, Mongrel
- vývojový framework : Rails
- programovacie jazyky : C, PHP, Ruby, JavaScript

4.4 Postup integrácie do novej knižnice (OPAC)

Vzhľadom na spomínané fakty je integrácia do OPAC veľmi jednoduchá. V prvom rade je nutné kontaktovať správu služby BibTip, vyplnit nimi zaslané registračné dokumenty a zaplatenie poplatku za službu. Pracovníci BibTip skontrolujú či sú splnené požiadavky pre integráciu na strane OPAC. Pokiaľ sú všetky požiadavky v poriadku, vytvoria novú databázu na serveroch BibTip určenú pre ukaladanie dát spojených s novou knižnicou. Medzi tým musí správca, alebo operátor OPAC pridať integračné kódy do HTML zdroju plného zobrazenia titulu. Pracovníci BibTip urobia sériu testov, aby overili správnosť a funkčnosť integrácie a tým prvotná práca končí.

Od tejto chvíle každý prihlásený užívateľ svojou činnosťou generuje štatistické dáta, ktoré sú okamžite spracovávané. Prvé rekomendácie daného titulu X sa objavia po tom, ako počet spolu-zobrazení páru v ktorom sa nachádza titul X s ďaľšími titulmi Y prevýši počet očakávaný z pravdepodobnosti náhodného výberu, ako je znázornené na obr. 2.



Obr. 2 - hranica od ktorej sú tvorené prvé rekomendácie pre titul X

Každá klientská knižnica má vytvorený vlastný účet na stránkach BibTipu, kde má k dispozícii grafy a podrobné štatistiky o množstve a využívaní štatistických dát, množstve existujúcich párov a rekomendácií pre tituly a mnoho ďalších informačno-štatistických údajov.

4.5 Vývoj do budúcnosti

Postupom času sa do projektu BibTip zapája viac a viac knižníc, čo umožnuje získavanie stále väčšieho množstva dát pre rekomendačné algoritmy a pre poskytovanie údajov knižniciam, ktoré majú menej rozsiahle databázy rekomendačných dát. Vzniká požiadavka vytvorenia rozhrania pre klientské knižnice, pomocou ktorého by mohli určovať externé zdroje štatistických dát bez nutného zásahu pracovníkov BibTip. Takéto rozhranie je predmetom súčastného snaženia vývojového tímu BibTip.

5 Vlastné zhodnotenie systému BibTip

Pre získanie praktických skúseností som si vybral katalóg Vedeckej knižnice v Olomouci (aleph.vkol.cz), ktorá používa systém Aleph. Snažil som sa nájsť titul, ktorý by mohol mať dostatok rekomendácií pre odskúšanie funkčnosti a relevantnosti rekomendácii. Po dlhšom pátraní som sa ale v novinkách knižnice dozvedel, že od 31.11.2009 pozastavili využívanie BibTip vo svojom systéme.

Skúsil som teda knižnicu Univerzity Karlsruhe (ubka.uni-karlsruhe.de). Vyhľadal som si titul o programovacom jazyku používanom aj pri programovaní BibTip s názvom „Design patterns in Ruby“ od Russa Olsena. Po otvorení plného zobrazenia sa v spodnej časti stránky objavil zoznam rekomendácii v prehladnom rámčeku začínajúcom logom BipTip. Relevantnosť odkazov bola pozoruhodná, nachádzali sa medzi nimi ďalšie tituly o Ruby a príbuzných priamo či nepriamo súviciacich témach. Zoznam titulov skôr budil dojem, že bol zostavený človekom, ktorý sa danej téme venuje a vytvoril zoznam doporučenej literatúry, ako že bol vytvorený automaticky na základe štatistických údajov.

6 Záver

Pozorovanie správania sa užívateľov a zbieranie štatistických dát, ich použitie a znovu-použitie pre smerovanie užívateľov k ďalším zdrojom je základným cieľom katalógových systémov súčastnosti. Tieto dáta sú a vždy budú užitočné a vytvoria základ pre budúce systémy inteligentných agentov pre tvorbu rekomendácií. V tejto oblasti sa výskum a vývoj zatiaľ nedostal príliš ďaleko, no systém BibTip je určite krok tým správnym smerom.

Referencie

[1] The Data Driven Library
<*http://datadrivenlibrary.blogspot.com*>

[2] D-lib magazín
<*http://www.dlib.org*>

[3] Domovská stránka University Library Karlsruhe
<*http://www.ubka.uni-karlsruhe.de/index_engl.html*>

[4] Domovská stránka BibTip
<*http://www.bibtip.org*>

Metadata v Dublin Core

```
<link rel="schema.DC" href="http://purl.org/dc/elements/1.1/" />
<link rel="schema.DCTERMS" href="http://purl.org/dc/terms/" />
<meta name="DC.title" lang="SK" content="BibTip - rekomendačný systém" />
<meta name="DC.creator" content="Michale Róbert" />
<meta name="DC.subject" lang="SK" content="BibTip" />
<meta name="DC.description" lang="SK" content="BibTip je rekomendačný systém
jednoducho integrovateľný do existujúcich OPAC, poskytovaný a spravovaný
Univerzitou Karlsruhe (Nemecko)" />
<meta name="DC.date" content="30.11.2009" />
<meta name="DC.type" scheme="DCTERMS.DCMIType" content="Text" />
<meta name="DC.format" scheme="DCTERMS.IMT" content="application/pdf" />
<meta name="DC.language" scheme="DCTERMS.RFC1766" content="SK" />
```