Annota: Pieskovisko pre výskum vyhľadávania, organizovania a navigácie s využitím sémantiky v prostredí digitálnych knižníc

Róbert Móro, Michal Holub, Jakub Ševcech, Mária Bieliková

Fakulta informatiky a informačných technológií
Slovenská technická univerzita v Bratislave
Ilkovičova 2, 842 16 Bratislava, Slovensko
meno.priezvisko@stuba.sk

Abstrakt. Digitálne knižnice predstavujú rozsiahly a cenný zdroj informácií pre výskumníkov z rozličných oblastí. Veľkým rozsahom však komplikujú efektívnu správu, hľadanie informácií a prístup k nim. Demonštrujeme Annotu, kolaboratívny nástroj na organizovanie a poznámkovanie výskumných zdrojov. Z hľadiska výskumu je to platforma pre zapojenie a overovanie metód automatizovaného organizovania kolekcií dokumentov, podpory navigácie v neznámej doméne, ako aj inteligentného vyhľadávania v dokumentoch. Prezentujeme aktuálny stav infraštruktúry spolu s výsledkami dosiahnutými pri tvorbe modelu používateľa a modelu domény využívajúcich ľahkú sémantiku.

Kľúčové slová: Annota, digitálna knižnica, prepojené dáta, model domény, model záujmov výskumníka, vyhľadávanie, navigácia, organizácia

1 Motivácia a ciele

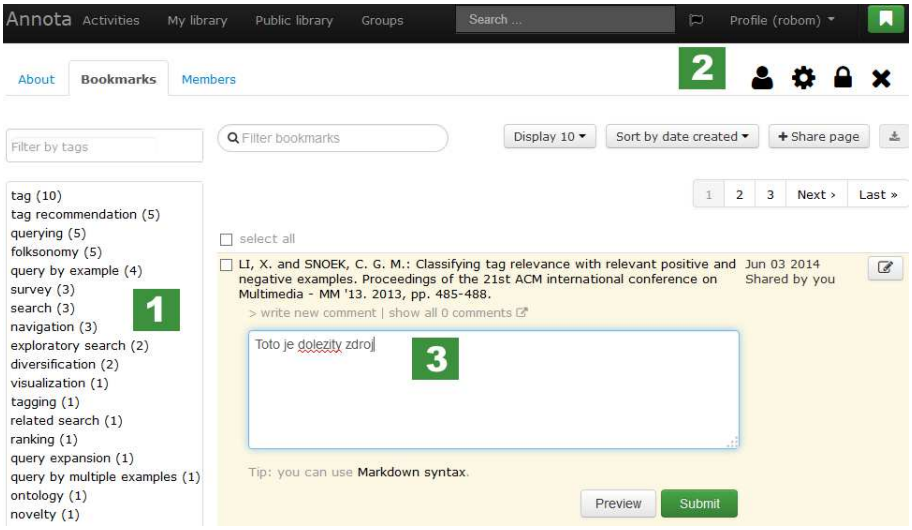
Pri práci s digitálnymi knižnicami musia výskumníci čeliť viacerým problémom vyplývajúcim najmä z rozsiahlosti informačných priestorov, akými sú napr. vyhľadanie a navigácia k relevantným zdrojom v knižnici, pričom existujúce riešenia tieto problémy adresujú len čiastočne. K dispozícii je napr. množstvo metadát o článkoch, ich autoroch a ďalších entitách, tradičné fulltextové vyhľadávanie však tieto súvislosti nevyužíva. Využitím metadát a sémantiky v nich obsiahnutých pritom vieme používateľovi poskytnúť pokročilejšie možnosti vyhľadávania, vrátane rozpoznania významu dopytu v prirodzenom jazyku [2]. Časté sú tiež úlohy vyhľadávania, ktorých charakter je prieskumný, ako je napr. oboznámenie sa s novou doménou [5], čo je typické pre scenár začínajúceho výskumníka, ktorého podpora v súčasných digitálnych knižniciach chýba.

Samostatný problém predstavuje organizácia (osobnej) kolekcie dokumentov. Manuálne zorganizovať väčšiu kolekciu dokumentov je časovo náročné; organizáciu je tiež potrebné priebežne upravovať, aby čo najlepšie vyhovovala aktuálnemu stavu celej kolekcie dokumentov. Výhodu preto majú inovatívne spôsoby (polo)automatickej organizácie pomocou štruktúr, akými sú napr. fazetové stromy [9].

V príspevku demonštrujeme platformu Annota (`annota.fiit.stuba.sk`) [1, 4], ktorá slúži výskumníkom ako: 1) nástroj uľahčujúci prácu s výskumnými zdrojmi a spoluprácu výskumníkov nad nimi, 2) platforma obsahujúca infraštruktúru pre realizáciu a overovanie výskumných metód. Naším hlavným cieľom je podpora efektívnej práce výskumníka s informačnými zdrojmi prístupnými prostredníctvom webových digitálnych knižníc poskytnutím viacerých možností organizovania a poznámkovania kolekcie článkov, inteligentného vyhľadávania v nich, ako aj rôznych foriem podpory navigácie v neznámej doméne. Kladieme pritom veľký dôraz na podporu spolupráce medzi výskumníkmi (či už je to vzťah školiteľ-študent, alebo skupina výskumníkov).

2 Annota – čo poskytuje výskumníkom

Základnou funkcionalitou Annoty je vytváranie a organizácia záložiek na články v digitálnych knižniciach. Tieto je možné pomocou webového rozhrania ďalej zdieľať v skupinách (napr. študent so svojím školiteľom), kde je o nich možné diskutovať (pozri obr. 1), prípadne je možné priamo odporučiť dokument nejakému používateľovi (napr. školiteľ svojmu študentovi). Podporujeme digitálne knižnice z domény informatiky (ACM DL, IEEE Xplore či Springer Link), záložku je však pomocou nami poskytnutého rozšírenia do prehliadača možné vytvoriť na ľubovoľnej webovej stránke.



Obrázok 1. Ukážka webového rozhrania Annoty – používatelia môžu organizovať svoje zdroje pomocou tagov (1) a zdieľať ich v skupinách, ktoré môžu jednoducho manažovať (2). Nad zdieľanými zdrojmi (záložkami) je možné diskutovať (3).

Annota spracúva metadáta navštívených dokumentov, pričom extrahuje napr. názov, autorov, rok vydania, názov publikácie (časopisu alebo zborníka konferencie) či kľúčové slová zadané autormi. Takto spracované dáta transformujeme do podoby RDF s využitím štandardnej ontológie BIBO (`bibliontology.com`) , čo vytvára priestor

pre ich obohatenie pomocou prepojených dát na webe. V súčasnosti naša dátová sada obsahuje vyše 50 000 vedeckých článkov, okolo 230 000 autorov, 8 000 publikácií a 130 000 kľúčových slov zadaných autormi článkov. Dátovú sadu sme zverejnili na výskumné účely (`annota.fiit.stuba.sk/dataset`) a bližšie ju opisujeme v [3]. Metadáta získavame okrem extrakcie aj priamo od používateľov v podobe značiek (tagov) a poznámok (komentárov v texte). Používatelia doteraz pridali k dokumentom v Annote okolo 2 800 unikátnych tagov a okolo 1 700 poznámok.

Zo zaznamenanej aktivity výskumníka v digitálnych knižniciach (čo vyhľadáva, aké dopyty používa a aké dokumenty navštevuje) vytvárame model jeho záujmov v podobe hypergrafu, v ktorom je viacero typov prepojení: medzi používateľom (výskumníkom) a kľúčovými slovami modelujúcimi jeho záujem, medzi kľúčovými slovami navzájom, čo zachytáva ich vzájomný súvis a medzi používateľmi (výskumníkmi), čo vyjadruje vzťahy ich podobnosti. Tento model je ľahko rozšíriteľný o nové typy vzťahov a je možné ho využívať pri prispôsobovaní systému používateľom, ako aj pri odporúčaní.

Annota zahŕňa komplexnú infraštruktúru, ktorá nám umožňuje overovať rôzne metódy a prístupy. Jej hlavnou výhodou je popri konfigurovateľnom a ľahko rozšíriteľnom modeli domény a používateľa podpora rôznych experimentálnych scenárov s využitím A/B testovania. To nám umožňuje nastaviť, čo jednotliví používatelia v rámci testovacích scenárov uvidia, ale aj aké vzťahy (z modelu používateľa a domény), sa v nich zohľadnia. Zameriavame sa pritom najmä na metódy organizácie, navigácie a vizualizácie informačného priestoru digitálnej knižnice a vyhľadávanie v ňom.

Experimentovali sme s organizáciou dokumentov pomocou dôležitých slov vizualizovaných v podobe klasického zoznamu, ako aj v oblaku, ktorý pozostával z používateľských tagov a automaticky extrahovaných kľúčových slov. Okrem frekvencie výskytu slova sme zohľadňovali aj jeho používanie v čase, čo odrážalo jeho aktuálnu dôležitosť [6]. Na hierarchickú organizáciu kolekcie dokumentov podľa ich metadát sme navrhli metódu fazetového stromu, ktorý umožňoval aj tvorbu ad-hoc štruktúr pre podporu aktuálnej úlohy (napr. vyhľadanie konkrétneho článku v kolekcii) [7].

Overili sme tiež metódu na vyhľadávanie súvisiacich dokumentov, v ktorej sa používali poznámky ako indikátory záujmu používateľa o konkrétne časti dokumentu [8].

Na podporu prieskumného vyhľadávania sme navrhli metódu navigácie pomocou navigačných vodítok, t. j. slov v sumarizácii dokumentu, ktoré slúžia ako prepojenia na relevantné dokumenty. Experimentovali sme s rôznou vizualizáciou vodítok (v súhrne dokumentu, pod súhrnom, v oblaku slov), pričom sme využili aj zariadenie na sledovanie pohľadu, aby sme zistili, ako používatelia s navrhnutým rozhraním interagujú.

Údaje zo sledovania pohľadu pritom možno použiť nielen na overovanie navrhnutých metód a rozhraní, ale aj ako ďalší zo zdrojov implicitnej spätnej väzby a zlepšiť tak modelovanie záujmov výskumníkov. Otvorené výskumné problémy vidíme aj v ďalšom využití doménovo-špecifických metadát pre modelovanie domény, akými sú napr. citácie, či využitie aktuálneho kontextu výskumníkov pre personalizovanú navigáciu a odporúčanie v digitálnej knižnici.

Poďakovanie. Táto publikácia vznikla vďaka čiastočnej podpore projektov VG1/0675/11, APVV-0208-10 a projektu v rámci OP Výskum a vývoj pre projekt ITMS 26240220084, spolufinancovaný zo zdrojov Európskeho fondu regionálneho

rozvoja. Autori by sa radi poďakovali všetkým, ktorí v uplynulých dvoch rokoch používali systém Annota, najmä členom výskumnej skupiny PeWe (pewe.fiit.stuba.sk) za mnohé diskusie a nápady, ako aj študentom priamo sa podieľajúcim na jeho vývoji – Jurajovi Kostolanskému, Martinovi Liptákovi, Samuelovi Molnárovi, Romanovi Burgerovi a Petrovi Mackovi.

English Summary

Annota: Sandbox for Research in the Fields of Semantics-Based Search, Organization and Navigation in the Domain of Digital Libraries.
Digital libraries represent a vast and valuable source of information for researchers in many domains. Nevertheless, their large extent poses complications when trying to effectively organize, access, and search the available information. We demonstrate Annota – a collaborative tool for organizing and annotating research resources on the Web. From the researcher's point of view it is a platform for realization and evaluation of methods for automatic organization of collections of documents, supporting navigation in an unknown domain, as well as intelligent searching in documents. We present the current state of the infrastructure together with the results we achieved while creating the user and domain models utilizing the lightweight semantics.

Literatúra

1. Bieliková, M., Ševcech, J., Holub, M., Móro, R.: Annota – poznámkovanie dokumentov v prostredí digitálnych knižníc. In: DATAKON '13: Proc. of the Annual Database Conference, pp. 143–152. VŠB-Technická univerzita, Ostrava. (2013)
2. Chong, W., Xiong, M., Zhou, Q., Yu, Yong.: PANTO: A portable natural language interface for ontologies. In: LNCS 4519: The Semantic Web: Research and Applications, pp. 473–487. Springer, Berlin, Heidelberg. (2007)
3. Holub, M., Ševcech, J., Móro, R., Bieliková, M.: Annota: Budovanie prepojenej dátovej sady dokumentov a poznámok. In: WIKT '13: Proc. of the 8th Workshop on Intelligent and Knowledge Oriented Technologies, pp. 13–18. Centre for Inf. Tech., Košice, (2013)
4. Holub, M., Móro, R., Ševcech, J., Lipták, M., Bieliková, M.: Annota: Towards enriching scientific publications with semantics and user annotations. D-Lib Magazine (to appear)
5. Marchionini, G.: Exploratory search: From finding to understanding. Communications of the ACM, 49(4), pp. 41–46. (2006)
6. Molnár, S., Móro R., Bieliková, M.: Trending words in digital library for term cloud-based navigation. In: SMAP '13: Proc. of the 8th Int. Workshop on Semantic and Social Media Adaptation and Personalization, pp. 53–58. IEEE CS, Washington, DC. (2013)
7. Móro, R., Bieliková, M., Burger, R.: Facet tree for personalized web documents organization. In: WISE '14: Proc. of 15th Int. Conf. on Web Inf. Systems Engineering, LNCS 8786, pp. 372–387. Springer, Berlin, Heidelberg. (2014)
8. Ševcech, J., Móro, R., Holub, M., Bieliková, M.: User annotations as a context for related document search on the web and digital libraries. Informatica, 38(1), pp. 21–30. (2014)
9. Zhao, J., Drucker, S.M., Fisher, D., Brinkman, D.: TimeSlice: Interactive faceted browsing of timeline data. In: AVI '12: Proc. of the Int. Working Conf. on Advanced Visual Interfaces, pp. 433–436. ACM Press, NY. (2012)