

Slovenská technická univerzita v Bratislave  
Fakulta informatiky a informačných technológií

FIIT-5220-56312

Bc. Jakub Ševcech

NAVIGÁCIA NA WEBE NA ZÁKLADE POZNÁMOK  
A ZNAČIEK

Diplomová práca

Vedúci práce: prof. Ing. Mária Bieliková, PhD.

máj, 2013

Slovenská technická univerzita v Bratislave  
FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLOGIÍ

FIIT-5220-56312

---

**Bc. Jakub Ševcech**

# **Navigácia na webe na základe poznámok a značiek**

Diplomová práca

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

máj 2013



## Zadanie diplomovej práce

*Meno študenta:* **Bc. Ševcech Jakub**

*Študijný program:* Softvérové inžinierstvo

*Študijný odbor:* Softvérové inžinierstvo

*Názov práce:* **Navigácia na webe na základe poznámok a značiek**

Samostatnou výskumnou a vývojovou činnosťou v rámci predmetov Diplomový projekt I, II, III vypracujte diplomovú prácu na tému, vyjadrenú vyššie uvedeným názvom tak, aby ste dosiahli tieto ciele:

*Všeobecný cieľ:*

Vypracovaním diplomovej práce preukážte, ako ste si osvojili metódy a postupy riešenia relatívne rozsiahlych projektov, schopnosť samostatne a tvorivo riešiť zložité úlohy aj výskumného charakteru v súlade so súčasnými metódami a postupmi študovaného odboru využívanými v príslušnej oblasti a schopnosť samostatne, tvorivo a kriticky pristupovať k analýze možných riešení a k tvorbe modelov.

*Špecifický cieľ:*

Vytvorte riešenie, zodpovedúce návrhu textu zadania, ktorý je prílohou tohto zadania. Návrh bližšie opisuje tému vyjadrenú názvom. Tento opis je záväzný, má však rámcový charakter, aby vznikol dostatočný priestor pre Vašu tvorivosť.

Riadte sa pokynmi Vášho vedúceho.

Pokiaľ v priebehu riešenia, opierajúc sa o hlbšie poznanie súčasného stavu v príslušnej oblasti alebo o priebežné výsledky Vášho riešenia alebo o iné závažné skutočnosti, dospejete spoločne s Vaším vedúcim k presvedčeniu, že niečo v texte zadania a/alebo v názve by sa malo zmeniť, navrhnete zmenu. Zmena je spravidla možná len pri dosiahnutí kontrolného bodu.

*Miesto vypracovania:* Ústav informatiky a softvérového inžinierstva FIIT STU v Bratislave

*Vedúci práce:* **prof. Ing. Mária Bieliková, PhD.**

*Termíny odovzdania:*

podľa harmonogramu štúdia platného pre semester, v ktorom máte príslušný predmet (Diplomový projekt I, II, III) absolvovať podľa Vášho študijného plánu

*Predmety odovzdania:*

V každom predmete dokument podľa pokynov na [www.fiit.stuba.sk](http://www.fiit.stuba.sk) v časti:  
home > Informácie o > štúdiu > organizácia štúdia > diplomový projekt

V Bratislave dňa 13. 2. 2012



prof. Ing. Pavol Návrat, PhD.  
riaditeľ Ústavu informatiky a softvérového  
inžinierstva

## Návrh zadania diplomovej práce

Finálna verzia do diplomovej práce<sup>1</sup>

### Študent:

**Meno, priezvisko, tituly:** Jakub Ševcech, Bc.  
**Študijný program:** Softvérové inžinierstvo  
**Kontakt:** sevo\_jakub@yahoo.fr

### Výskumník:

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

### Projekt:

**Názov:** Navigácia na webe na základe poznámok a značiek  
**Názov v angličtine:** Web navigation based on annotations  
**Miesto vypracovania:** Ústav informatiky a softvérového inžinierstva, FIIT STU, Bratislava  
**Oblasť problematiky:** Navigácia v prostredí Internetu

### Text návrhu zadania<sup>2</sup>

V súčasnosti pozorujeme veľký rozmach rôznych nástrojov na zdieľanie obsahu, či už vo forme odkazov na zaujímavé stránky, obrázkov alebo najrôznejších multimediálnych informácií. Asi najznámejšími nástrojmi podporujúcimi rôzne formy zdieľania sú sociálne siete, ale aj nástroje na vytváranie záložiek a poznámok v prostredí webových stránok. Veľa aplikácií poskytuje funkcie na vytváranie poznámok do elektronických dokumentov, ale len málo z nich využíva poznámky na vytváranie ďalšej pridanej hodnoty pre používateľa.

Zamerajte sa na oblasť podpory navigácie a odporúčania s použitím poznámok priradených k webovým stránkam. Analyzujte existujúce metódy a aplikácie na vytváranie poznámok s dôrazom na kolaboratívne prístupy a metódy na navigáciu pomocou poznámok priradených k dokumentom. Skúmajte možnosti odporúčania dokumentov na základe k nim priradených poznámok.

Navrhňte metódu na podporu navigácie v informačných zdrojoch na webe pomocou záložiek, značiek a iných druhov poznámok. Pri práci so značkami využite vzťahy podobnosti a iné prepojenia medzi používateľmi. Navrhnuté riešenie overte pomocou vytvoreného experimentálneho softvérového prototypu vo vami vybranej doméne.

<sup>1</sup> Vytlačiť obojstranne na jeden list papiera


<sup>2</sup> 150-200 slov (1200-1700 znakov), ktoré opisujú výskumný problém v kontexte súčasného stavu vrátane motivácie a smerov riešenia

### Literatúra<sup>3</sup>

- D. Carmel, H. Roitman, and E. Y. Tov .: Social bookmark weighting for search and recommendation, The VLDB Journal, vol. 19, pp. 761-775, Dec. 2010
- M. Agosti, N. Ferro .: A formal model of annotations of digital content, Transactions on Information Systems (TOIS), ACM, Nov. 2007
- G. Buchanan, J. Pearson .: Improving Placeholders in Digital Documents Research and Advanced Technology for Digital Libraries, Eds. Springer Berlin / Heidelberg, vol. 5173, pp. 1-12, 2008

Vyššie je uvedený návrh diplomového projektu, ktorý vypracoval(a) Bc. Jakub Ševcech, konzultoval(a) a osvojil(a) si ho prof. Ing. Mária Bieliková, PhD. a súhlasí, že bude takýto projekt viesť v prípade, že bude pridelený tomuto študentovi.

V Bratislave dňa 10.1.2012


  
\_\_\_\_\_  
Podpis študenta

  
\_\_\_\_\_  
Podpis výskumníka

### Vyjadrenie garanta predmetov Diplomový projekt I, II, III

Návrh zadania schválený: áno / nie<sup>4</sup>

Dňa: .....7.2.2012.....

  
\_\_\_\_\_  
Podpis garanta predmetov

<sup>3</sup> 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 (uvedte všetky potrebné údaje na identifikáciu zdroja, pričom uprednostnite vedecké príspevky v časopisoch a medzinárodných konferenciách)

<sup>4</sup> Nehodiace sa prečiarknite



# ANOTÁCIA

Slovenská technická univerzita v Bratislave  
FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLOGIÍ

Autor: Jakub Ševcech  
Vedúci diplomového projektu: prof. Ing. Mária Bieliková, PhD.  
Diplomový projekt: Navigácia na webe na základe poznámok a značiek  
Študijný program: Softvérové inžinierstvo

máj 2013

V súčasnosti pozorujeme veľký rozmach rôznych služieb na vytváranie záložiek a pridávanie poznámok k dokumentom. Takto vytvorené poznámky sa dajú použiť na zlepšenie navigácie, na vyhľadávanie, odporúčanie dokumentov a podobne.

V tejto práci sa zaoberáme podporou navigácie medzi dokumentami pomocou vyhľadávania súvisiacich dokumentov k práve študovanému dokumentu na základe jeho obsahu a k nemu pripojených poznámok. Poznámky ako napríklad zvýraznenia v texte a komentáre používame ako indikátory záujmu používateľa o konkrétne časti dokumentu. Na základe obsahu dokumentu a k nemu pripojených poznámok vytvárame dopyt, ktorý používame na vyhľadanie súvisiacich dokumentov. Pri tvorbe dopytu využívame predpoklad, že používateľ pridanými poznámkami zvýrazňuje práve tie časti dokumentu, ktoré ho najviac zaujímajú.

Navrhli sme metódu na používanie rôznych typov poznámok pri tvorbe dopytu na vyhľadanie súvisiacich dokumentov. Navrhnutú metódu sme overili prostredníctvom implementácie nástroja na zbieranie poznámok a vyhľadávanie súvisiacich dokumentov s pomocou týchto poznámok v doméne digitálnych knižníc.





# ANNOTATION

Slovak University of Technology Bratislava  
FACULTY OF INFORMATICS AND INFORMATION TECHNOLOGIES

Autor: Jakub Ševcech  
Supervisor: prof. Ing. Mária Bieliková, PhD.  
Diploma Thesis: Navigation on the web using annotations  
Degree Course: Software engineering

may 2013

Currently we experience a great expansion of various services allowing us to bookmark documents and annotate them. Such annotations can be used to improve navigation, document retrieval, recommendation of documents and so on.

In our work, we aim to improve navigation between documents by searching documents related to studied document using its content and attached annotations. We use annotations such as highlights and comments as indicators of user's interest in particular part of the document. Using document content and attached annotations, we are creating a query that we are using to search for related documents. When creating the query, we use an assumption that by creating annotations, user highlights parts of document, that he is most interested in.

We proposed a method for using different types of annotations in creation of query to search related documents. We evaluated the proposed method by creating a tool for collecting annotations and for searching related documents using these annotations in the domain of digital libraries.



Prehlasujem, že túto prácu som vypracoval samostatne, s použitím uvedených informačných zdrojov.



# Pod'akovanie

Chcem sa pod'akovať vedúcej mojej diplomovej práce, profesorky Márii Bielikovej za jej cenné rady, skúsenosti a vynaložený čas, ktorými mi výrazne pomohla pri písaní tejto práce.

Ďalej sa chcem pod'akovať mojej rodine a priateľom, ktorí ma podporovali počas celého môjho štúdia.

Jakub Ševcech



# Obsah

|          |  |           |
|----------|--|-----------|
| <b>1</b> | <b>Úvod</b>  | <b>1</b>  |
| <b>2</b> | <b>Poznámky v elektronických dokumentoch</b>                       | <b>5</b>  |
| 2.1      | Rôzne pohľady na poznámky . . . . .                                | 5         |
| 2.2      | Metafora písania poznámok na papier . . . . .                      | 7         |
| <b>3</b> | <b>Podpora orientácie v informačnom priestore pomocou poznámok</b> | <b>9</b>  |
| 3.1      | Navigácia pomocou tagov . . . . .                                  | 10        |
| 3.2      | Poznámky vo vyhľadávaní . . . . .                                  | 11        |
| 3.2.1    | Vyhľadávanie pomocou poznámok . . . . .                            | 12        |
| 3.2.2    | Personalizované vyhľadávanie . . . . .                             | 14        |
| 3.3      | Zhodnotenie využívania poznámok pri navigácii . . . . .            | 18        |
| <b>4</b> | <b>Existujúce nástroje na podporu poznámkovania na webe</b>        | <b>21</b> |
| 4.1      | Delicious . . . . .  | 21        |
| 4.2      | Diigo . . . . .  | 23        |
| 4.3      | Mendeley . . . . .   | 24        |
| 4.4      | Flickr . . . . .   | 25        |
| 4.5      | Bibsonomy . . . . .  | 26        |
| 4.6      | Alef . . . . .   | 27        |
| 4.7      | Diskusia k súčasnému stavu použitia poznámok . . . . .             | 28        |
| <b>5</b> | <b>Pripájanie poznámok k dokumentom</b>                            | <b>33</b> |
| 5.1      | Umiestnenie poznámok do webových stránok . . . . .                 | 33        |
| 5.2      | Realizácia pomocou rozšírenia webového prehliadača . . . . .       | 35        |



|          |   |            |
|----------|---|------------|
| <b>6</b> | <b>Metóda na tvorbu dopytu pomocou poznámok</b>                 | <b>39</b>  |
| 6.1      | Transformácia textu na graf . . . . .                           | 41         |
| 6.2      | Výber slov do dopytu . . . . .                                  | 41         |
| 6.3      | Vlastnosti metódy na tvorbu dopytov . . . . .                   | 44         |
| <b>7</b> | <b>Vyhodnotenie</b>   | <b>47</b>  |
| 7.1      | Stanovenie parametrov navrhnutej metódy pomocou simulácie . . . | 47         |
| 7.1.1    | Dátová sada . . . . .   | 47         |
| 7.1.2    | Generovanie poznámok . . . . .                                  | 49         |
| 7.1.3    | Simulácia . . . . .   | 50         |
| 7.2      | Porovnanie voči metóde založenej na TF-IDF . . . . .            | 51         |
| 7.3      | Vyhodnotenie vyhľadávania súvisiacich dokumentov . . . . .      | 53         |
| 7.3.1    | Používateľská štúdia . . . . .                                  | 54         |
| 7.3.2    | Dlhodobý experiment v nástroji Annota . . . . .                 | 57         |
| <b>8</b> | <b>Zhodnotenie</b>  | <b>59</b>  |
|          | <b>Literatúra</b>   | <b>61</b>  |
| <b>A</b> | <b>Annota - Používateľská príručka</b>                          | <b>65</b>  |
| <b>B</b> | <b>Annota - Technická dokumentácia</b>                          | <b>77</b>  |
| <b>C</b> | <b>Annota - Inštalačná príručka</b>                             | <b>85</b>  |
| <b>D</b> | <b>Analýza údajov z používania nástroja Annota</b>              | <b>87</b>  |
| <b>E</b> | <b>Parametre použité pri generovaní poznámok v simulácii</b>    | <b>93</b>  |
| <b>F</b> | <b>Príspevok publikovaný na konferencii WIKT 2012</b>           | <b>97</b>  |
| <b>G</b> | <b>Príspevok publikovaný na konferencii IIT.SRC 2013</b>        | <b>103</b> |
| <b>H</b> | <b>Príspevok do konferencie ASIR 2013</b>                       | <b>111</b> |
| <b>I</b> | <b>Obsah elektronického média</b>                               | <b>121</b> |

# Úvod

V súčasnosti pozorujeme veľký rozmach rôznych nástrojov na zdieľanie obsahu na webe, či už je to obsah vo forme odkazov na zaujímavé stránky, videá, obrázky alebo iné multimediálne informácie. Dôkazom je rýchly nárast počtu používateľov rôznych sociálnych sietí a služieb na vytváranie poznámok a záložiek do webových stránok. Používatelia sú ochotní vytvárať obsah a zdieľať ho, ale len v tom prípade, ak za to niečo dostanú „na oplátku“.

Zvykli sme si písať poznámky do rôznych tlačенých kníh, časopisov alebo iných dokumentov. Tu používame poznámky na zvýraznenie alebo doplnenie dôležitých informácií. Poznámky, ktoré takto vytvárame, používame na prispôsobenie dokumentov pre vlastné potreby a na zlepšenie navigácie v nich. Rovnako ako pri práci s tlačnými dokumentami aj pri práci s elektronickými dokumentami je možné vytvárať rôzne druhy poznámok. Súčasné nástroje na pridávanie poznámok do elektronických dokumentov nám umožňujú vytvárať väčšinu tých druhov poznámok, ktoré môžeme pridávať aj do tlačných dokumentov. Navyše však tieto poznámky môžeme ďalej spracovávať. Takto vytvorené poznámky môžeme použiť napríklad na zlepšenie navigácie medzi dokumentami alebo na vyhľadávanie dokumentov. V tejto oblasti prebieha živý výskum (Agosti and Ferro 2007; Buchanan and Pearson 2008; Körner et al. 2010; Šimko et al. 2011) a existuje viacero služieb a aplikácií, ktoré umožňujú používateľom vytvárať poznámky a využívať ich napríklad na navigáciu medzi dokumentami.

Ďalej v tomto dokumente budeme pod pomenovaním poznámka uvažovať poznámky, ktoré pridali používatelia do elektronických dokumentov, pokiaľ nebude uvedené inak. V práci sa zaoberáme najmä poznámkami v textových dokumentoch, aj napriek tomu že viaceré typy rozoberaných poznámok je možné pridávať aj do iných, multimediálnych dokumentov. Pod pojmom dokument budeme rozumieť elektronický dokument, pokiaľ nebude uvedené inak.

Oblasť poznámkovania dokumentov sa dá rozdeliť na dva základné smery:

- hľadanie prostriedkov a foriem, pomocou ktorých by používatelia dokázali efektívne pridávať poznámky do dokumentov a
- hľadanie spôsobov ako vytvorené poznámky využiť aj pri iných úlohách, a tak posilniť motiváciu na ich vytváranie.

Ak budeme považovať za dokumenty ľubovoľné webové stránky, narazíme na niekoľko základných problémov spojených s motiváciou používateľov na vytváranie poznámok. Používatelia len veľmi zriedkavo pridávajú poznámky ako komentáre alebo zvýraznenia v texte do tela stránky. Toto je spôsobené tým, že väčšina stránok je príliš krátka na to, aby bolo potrebné v nej zvýrazňovať niektorú jej časť. Za hlavnú príčinu pre malý počet poznámok pridaných do webových stránok považujeme to, že len zriedkakedy sa stáva, že sa používateľ dostane na stránku, o ktorej vie, že ju bude ešte v budúcnosti používať a poznámky, ktoré si do nej pridá mu pri budúcej návšteve môžu poslúžiť.

Napriek tomu veľa používateľov využíva funkcie rôznych nástrojov na vytváranie záložiek (ako špeciálneho typu poznámok) ako je napríklad služba Delicious<sup>1</sup> alebo Diigo<sup>2</sup>. Pomocou týchto služieb si používatelia jednoducho zaradia stránku medzi záložky, prípadne k nej priradia zopár kľúčových slov, aby ju dokázali opätovne rýchlo nájsť v zozname záložiek. Preto oveľa častejšie používatelia pridávajú poznámky k dokumentom ako sú napríklad články v digitálnej knižnici. Teda dokumenty, pri štúdiu ktorých musia používatelia vynaložiť väčšie úsilie na ich pochopenie, ktoré sú dlhšie a je vyššia pravdepodobnosť, že používateľ sa k danému dokumentu opätovne vráti.

Pri používaní nástrojov na pridávanie poznámok do webových stránok, ale aj nástrojov na pridávanie poznámok do iných typov dokumentov, sú používatelia motivovaní najmä tým, že keď budú chcieť niekedy v budúcnosti opätovne navštíviť daný dokument, budú mať k dispozícii navigáciu prostredníctvom tagov, ktoré si k záložke priradili. Prostredníctvom ďalších typov poznámok sa budú môcť rýchlejšie znovu oboznámiť s dokumentom a tak si uľahčiť opätovné štúdium dokumentu.

Vďaka možnosti zdieľania poznámok, ktorú rôzne nástroje na vytváranie poznámok často poskytujú, majú používatelia možnosť navigovať sa nielen medzi dokumentami vo vlastnej zbierke, ale majú taktiež možnosť vyhľadávať dokumenty na základe poznámok ostatných používateľov, prezerat' si zbierky dokumentov iných používateľov a spolupracovať pri vytváraní poznámok vo webových stránkach.

Ďalšou z možných „odmien“ za vytváranie poznámok do dokumentov je vyhľadávanie a odporúčanie ďalších dokumentov na základe doteraz prezretých dokumentov a poznámok priradených k nim. Takéto odporúčanie je možné vykonávať na základe počtu prezretí alebo počtu poznámok priradených

---

<sup>1</sup>Delicious, <http://delicious.com/>

<sup>2</sup>Diigo, <http://www.diigo.com>

---

k dokumentom, prípadne pomocou metód, ktoré berú do úvahy podobnosť jednotlivých dokumentov a používateľov.

Pri vytváraní poznámok sú v súčasnosti používatelia motivovaní najmä tým, že niekedy v budúcnosti sa budú môcť vrátiť k prečítaným dokumentom pomocou záložiek alebo pomocou vyhľadávania na základe značiek (tagov). My sa sústreďujeme na poskytovanie odmeny za vytváranie poznámok v čase ich vytvárania. Navrhujeme metódu na vyhľadávanie dokumentov súvisiacich s práve študovaným dokumentom na základe obsahu dokumentu a poznámok, ktoré do neho používateľ pridáva.

V ďalších častiach dokumentu opisujeme čo to vlastne poznámky v elektronických dokumentoch sú (kapitola 2) a ako ich je možné využiť pri navigácii (kapitola 3). Ďalej v kapitole 4 opisujeme niekoľko aplikácií na pridávanie poznámok do rôznych druhov textových, ale aj multimediálnych dokumentov. V kapitole 5 opisujeme niekoľko spôsobov pre pridávanie poznámok do dokumentov a opisujeme nástroj na pridávanie poznámok do webových stránok a PDF dokumentov, ktorý sme vytvorili. V kapitole 6 navrhujeme metódu na vytváranie dopytu na získanie súvisiacich dokumentov s pomocou obsahu práve študovaného dokumentu a k nemu pripojených poznámok. V kapitole 5 navrhujeme postup pre pripájanie poznámok do webových stránok a opisujeme realizáciu pomocou nástroja Annota. Prácu uzatvára vyhodnotenie navrhnutých metód v doméne digitálnych knižníc (kapitola 7) a zhodnotenie dosiahnutých výsledkov v kapitole 8. Jedným z výsledkov práce je vytvorenie nástroja na pridávanie poznámok do webových stránok Annota, ktorého používateľská príručka je uvedená ako príloha A.



# Poznámky v elektronických dokumentoch

Pri čítaní tlačených kníh a iných dokumentov je prirodzené pridávať si za okraj alebo priamo to textu rôzne druhy poznámok alebo značiek. Pod značkou rozumieme rôzne grafické označenia v texte ako napríklad rôzne šípky, preškrtnutie textu alebo podfarbenie textu. Poznámky vnímame širšie, teda okrem rôznych označení a zvýraznení v texte pod pojmom poznámka rozumieme aj rôzne texty pridané ku konkrétnemu miestu v dokumente alebo k dokumentu ako celku. Značky a poznámky nám slúžia na zvýraznenie alebo doplnenie dôležitých informácií, pomáhajú nám pri opätovnom čítaní dokumentu a pri navigácii v dokumente. Okrem vlastných poznámok, ktoré si pridávame do dokumentu dokážeme pri čítaní dokumentu využiť aj poznámky, ktoré do dokumentu pridali jeho predchádzajúci čitatelia. Pomocou nich priamo vidíme, ktoré časti dokumentu zaujali iných čitateľov, ktoré časti sú najdôležitejšie a aké myšlienky mali čitatelia dokumentu pri čítaní konkrétnych jeho častí.

Rovnako ako pri práci s tlačenými dokumentami aj pri práci s elektronickými dokumentami existujú nástroje a techniky na pridávanie rôznych druhov poznámok k dokumentom.

## 2.1 Rôzne pohľady na poznámky

Jednu z definícií poznámky môžeme nájsť na stránkach World Wide Web konzorcia (W3C) patriacim k Annotea (Kahan 2002) projektu<sup>3</sup>,

„Poznámkami myslíme komentáre, zápisky, vysvetlenia alebo iné typy externých značiek, ktoré môžu byť pripojené k hocakému webovému dokumentu alebo k vybranej časti dokumentu bez toho aby museli dokument meniť.“

---

<sup>3</sup>Annotea, <http://www.w3.org/2001/Annotea/>

Túto definíciu nemusíme používať len pre webové dokumenty, ale všeobecne pre ľubovoľné elektronické dokumenty. Podľa tejto definície môžeme poznámky považovať za hocaké informácie pripojené k dokumentu, ktoré dokument opisujú alebo ho rozširujú.

Existuje niekoľko rôznych spôsobov ako rozdeľovať poznámky do kategórií (Agosti and Ferro 2007). Jedným zo základných pohľadov na poznámky v elektronických dokumentoch je rozdelenie podľa ich formy na:

- poznámky ako metadáta opisujúce dokument a jeho obsah a
- poznámky ako dodatočné informácie, pridané k informáciám v dokumente.

Pri pohľade na poznámky ako na metadáta, poznámky opisujú dokument a informácie v ňom. Tieto poznámky musia spravidla dodržiavať predpísanú formálnu štruktúru, aby mohli byť spracovateľné strojom. Takéto poznámky sú určené na použitie pri strojovom spracovávaní dokumentov a môžeme ich tak použiť napríklad pri zlepšení vyhľadávania alebo organizovania dokumentov.

Poznámky pridané k dokumentom ako dodatočné informácie, ktoré nemajú formálne špecifikovanú formu, sú ťažšie spracovateľné strojom, keďže stroje nedokážu presne odhaliť ich význam. Takéto poznámky sú určené najmä pre ľudí, pre ktorých predstavujú ďalšie užitočné informácie pripojené k dokumentu. Aj takéto poznámky sa však dajú použiť pri automatickom spracovávaní dokumentov. Jedným z takýchto príkladov je použitie tagov priradených k dokumentom na ich organizovanie a na zlepšenie vyhľadávania.

Ďalším pohľadom na poznámky je ich rozdelenie na základe toho, akým spôsobom poznámky vznikajú, teda na:

- manuálne poznámkovanie a
- (polo)automatické vytváranie poznámok.

Proces manuálneho poznámkovania webových stránok a iných elektronických dokumentov sa snaží napodobniť proces písania poznámok do tlačенých dokumentov (Schilit, Price, and Golovchinsky 1998). Nástroje na manuálne poznámkovanie digitálnych dokumentov v skutočnosti neprinášajú žiadne nové druhy poznámok, len poskytujú nové prostriedky na ich vytváranie, ukladanie a vizualizáciu. Najčastejšie druhy poznámok, ktoré tieto nástroje umožňujú vytvárať, sú rôzne grafické značky, poznámky za okrajom, zvýraznenia v texte a podobne. Na rozdiel od poznámok v tlačенých dokumentoch, poznámky v elektronických dokumentoch sa dajú použiť nie len ako prostriedok na zlepšenie navigácie v rámci jedného dokumentu, ale dajú sa použiť aj na navigáciu medzi dokumentami alebo na vyhľadanie dokumentov.

Pri (polo)automatickom vytváraní poznámok nie je potrebná účasť človeka pri niektorých častiach procesu poznámkovania, alebo celý tento proces prebieha

automaticky. Pri (polo)automatickom vytváraní poznámok je teda snaha o minimalizovanie potreby manuálnych zásahov do procesu poznámkovania na rozdiel od metód na manuálne vytváranie poznámok, ktoré sa zameriavajú hlavne na poskytovanie podporných funkcií pre vytváranie poznámok a na poskytnutie prívetivého používateľského prostredia pre vytváranie poznámok. V tejto práci sa budeme zaoberať práve manuálne vytváranými poznámkami a podporou navigácie s pomocou týchto poznámok.

## 2.2 Metafora písania poznámok na papier

Pri čítaní kníh, novín alebo iných tlačенých dokumentov sme si zvykli písať do textu alebo za okraj krátke poznámky, v ktorých si zachytávame naše myšlienky alebo zvýrazňujeme dôležité časti textu. Často sa stáva, že ak chceme písať poznámky do elektronických dokumentov, tak si tieto dokumenty najskôr vytlačíme a poznámky píšeme do týchto tlačéných dokumentov (Golovchinsky, Price, and Schilit 1999). Takéto poznámky rozširujú alebo opisujú obsah dokumentu a môžu byť teda veľmi užitočné. Tým že si poznámky píšeme do vytlačeného dokumentu však strácame väzby medzi pôvodným dokumentom a poznámkami a nemôžeme ich ďalej používať napríklad na obohacovanie alebo organizovanie dokumentov.

V poslednom období vzniklo množstvo aplikácií, ktoré poskytujú možnosť pridávania poznámok do elektronických dokumentov alebo do webových stránok. Spravidla tieto aplikácie neumožňujú vytvárať nové typy poznámok oproti poznámkam, ktoré sme si už predtým bežne písali do tlačéných dokumentov. Tieto služby však umožňujú zdieľať poznámky medzi rôznymi používateľmi, prípadne umožňujú prepájať a vyhľadávať dokumenty prostredníctvom poznámok. Jedny z najčastejšie podporovaných typov poznámok sú:

- zvýraznenia,
- komentáre,
- záložky a
- značky (tagy).

Za zmienku stoja najmä tagy, ktoré sa používajú s veľkou obľubou v rôznych systémoch ako prostriedok na organizovanie dokumentov. Tagy sú vlastne pojmy, ktorými sa návštevník snaží vystihnúť dokument. Používatelia tagov sa dajú rozdeliť do dvoch skupín podľa spôsobu, akým priradujú tagy k dokumentom na (Körner et al. 2010):

- tých čo používajú značky na zaradenie dokumentov do kategórií a
- tých čo používajú značky na opísanie dokumentov.



Motivácia na tvorbu značiek je rozdielna pre tých, ktorí pomocou nich kategorizujú dokumenty a tých, ktorí pomocou nich opisujú dokumenty. Tí čo pridávajú k dokumentom značky, ktoré predstavujú kategórie tak robia preto, aby si dokumenty zatriedili do vlastných tried. Značky teda umožňujú vytvoriť si vlastnú navigáciu pre každého používateľa a jednotliví používatelia sa nemusia obmedzovať navigáciou, ktorú pre nich vytvoril autor dokumentu.

Tí, čo značkami opisujú dokumenty, tak robia preto, aby si uľahčili neskoršie vyhľadávanie medzi dokumentami (Körner et al. 2010). Priradujú im spravidla slová, podľa ktorých môžu daný dokument neskôr jednoducho nájsť.

Takto vytvorené poznámky nemusia byť užitočné len pre ich autorov, ale môžu byť užitočné pre ďalších čitateľov dokumentu. Veľa služieb umožňuje vytvárať verejné poznámky, ktoré môžu vidieť všetci používatelia, prípadne umožňujú zdieľať poznámky medzi skupinou používateľov ako napríklad služba Diigo alebo Mendeley<sup>4</sup>. Takto dokáže skupina používateľov pridávať a zdieľať poznámky, a tak kolaboratívne obohacovať dokumenty.

V súčasnosti používatelia rôznych nástrojov na vytváranie poznámok do elektronických dokumentov môžu používať poznámky najmä ako prostriedky na zlepšenie organizácie vlastnej sady dokumentov, na zvýraznenie dôležitých častí dokumentov, na zaznamenávanie vlastných myšlienok počas čítania dokumentov a na obohacovanie dokumentov o spoločné poznámky celých skupín používateľov.

---

<sup>4</sup>Mendeley, <http://www.mendeley.com/>

# Podpora orientácie v informačnom priestore pomocou poznámok

Pri súčasnom zahľtení informáciami je jedným z najväčších problémov nájsť spôsob ako sa navigovať medzi dokumentami a ako nájsť najrelevantnejšie dokumenty, ktoré práve potrebujeme. Na vyhľadávanie informácií na webe vzniklo niekoľko rôznych prístupov. Najpopulárnejšie z nich sú:

- katalógy stránok, kde sú ručne vyberané webové stránky zaradené do hierarchie kategórií a
- vyhľadávače, ktoré vyhľadávajú stránky na základe dopytu, najčastejšie v podobe zoznamu kľúčových slov.

Tieto prístupy používajú na podporu navigácie medzi dokumentami informácie z obsahu dokumentov a informácie o prepojení dokumentov medzi sebou. V súčasnosti sledujeme veľký rozmach služieb na zdieľanie obsahu, vytváranie záložiek a pridávanie poznámok do dokumentov. Toto vidíme napríklad na rýchлом nástupe rôznych sociálnych sietí alebo služieb na vytváranie záložiek v prostredí Internetu ako je napríklad aj služba Delicious. Používatelia dosiaľ vytvorili pomocou týchto služieb obrovské množstvo informácií, v ktorých nie je kvôli ich množstvu jednoduché sa navigovať, ale aj naopak, ktoré sa dajú použiť na podporu navigácie. Napríklad záložky a tagy, ktoré k webovým stránkam umožňuje pridávať množstvo služieb a stránok, môžu byť veľmi užitočným prostriedkom na skvalitnenie navigácie a vyhľadávania.

Rôzne služby na vytváranie poznámok najčastejšie podporujú pridávanie poznámok do textových dokumentov (a webových stránok), ale častokrát sú podporované aj ďalšie multimediálne dokumenty ako napríklad videá alebo

hudba. Úlohou týchto služieb je poskytovať možnosti na jednoduché vytváranie záložiek v rámci dokumentov alebo označovať zaujímavé dokumenty ako celok. Možné je teda pridávať záložku na konkrétne miesto v dokumente, napríklad na konkrétnu stranu v textovom dokumente alebo ku konkrétnej scéne vo videu. Označovanie zaujímavých úsekov v dokumentoch môžeme vidieť napríklad v rôznych prehliadačoch textových dokumentov, ktoré umožňujú pridať medzi záložky jednotlivé strany dokumentu. Pohľad na záložku ako na označenie celého dokumentu môžeme vidieť napríklad pri vytváraní záložiek k webovým stránkam, kde sa označujú celé stránky ako zaujímavé a nielen ich časti.

Okrem vytvárania zoznamu záložiek služby na vytváranie záložiek častokrát poskytujú aj prostriedky na pridávanie tagov k dokumentom. Tagy sú veľmi obľúbené najmä vďaka tomu, že sú pre používateľa jednoduché na vytváranie, intuitívne a dajú sa jednoducho použiť na vytvorenie vlastnej organizácie dokumentov.

### 3.1 Navigácia pomocou tagov

Tagy sú jednou z najpoužívanejších foriem poznámok. Nielen služby na pridávanie záložiek do internetových stránok, ale aj množstvo ďalších aplikácií umožňuje pridávať tagy k dokumentom. Príkladom je portál Stackoverflow<sup>5</sup>, ktorý umožňuje pridať poznámky ku každej otázke návštevníkov. Podobne najrôznejšie blogy a ďalšie služby umožňujú pridávať k dokumentom tagy. Tieto tagy slúžia ako dôležitý zdroj informácií pre efektívnu navigáciu medzi dokumentami, ktorú okrem autora dokumentu môžu vytvárať aj jeho návštevníci.

Používatelia vytvárajú tagy dvoma spôsobmi: kategorizovaním dokumentov a opisovaním dokumentov (Körner et al. 2010). Takto vytvorené tagy tvoria akúsi sumarizáciu celého dokumentu a dajú sa pomocou nich dokumenty organizovať a následne filtrovať. Filter dokumentov má najčastejšie formu zoznamu tagov, zoradeného podľa počtu výskytov vo všetkých dokumentoch. Častokrát je pri každom tagu zobrazený aj počet výskytov tagu v dokumentoch, takže tagy tvoria akési fazetové vyhľadávanie (obrázok 3.1). Po kliknutí na konkrétny tag zo zoznamu sa zobrazia len tie dokumenty, ktoré obsahujú daný tag. V zložitejších filtroch sa dajú tieto tagy vo filtri spájať a dajú sa tak vytvárať zložitejšie dopyty na vyhľadanie dokumentov. Takúto formu filtra poskytuje napríklad služba Delicious.

Jednou z foriem filtra pomocou tagov je oblak tagov (angl. tag-cloud). Oblak tagov je skupina tagov zobrazená používateľovi, pomocou ktorých je možné filtrovať dokumenty. Jednotlivé tagy sú graficky odlišené napríklad veľkosťou alebo farbou písma tak, aby boli významnejšie tagy výraznejšie. Tu odpadá potreba zoradovať

<sup>5</sup>Stackoverflow, <http://stackoverflow.com/>

The screenshot displays the Diigo bookmark management interface. On the left, a sidebar lists various categories with their respective counts: development (19), tutorial (16), reference (9), ruby (9), api (8), tools (7), java (7), learning (6), web (6), javascript (6), rails (5), google (4), code (4), howto (3), linux (3), editor (3), ebook (3), algorithms (3), and rubyonrails (3). The main content area shows a list of bookmarks, each with a date, title, source URL, and a set of tags. The bookmarks include:

- 05 Mar 12: JavaScript Text Highlighting (source: www.nsftools.com, tags: javascript, highlight, search, programming, referrer)
- 06 Feb 12: Sublime Text: The text editor you'll fall in love with (source: www.sublimetext.com, tags: editor, text, programming, software, texteditor, sublime, texmate)
- 04 Feb 12: Udacity - Educating the 21st Century (source: www.udacity.com, tags: education, programming, udacity, university, learning, stanford, science, search)
- 26 Jan 12: 30 books everyone in software business should read (and why) – Micro-ISV (source: www.dextronet.com, tags: books, programming, software, 30, best, engineering)
- 13 Sep 11: Home // RailsTips by John Nunemaker (source: railstips.org, tags: rails, ruby, rubyonrails, development, blog, programming, railstips, tips)
- 11 Sep 11: Learn Code The Hard Way -- Books And Course To Learn To Code (source: ruby.learncodethehardway.org, tags: ruby, tutorial, programming, learning, ebook, reference)

Obr. 3.1: Filtrovanie záložiek pomocou tagov v službe Diigo

zoznam podľa dôležitosti tagov alebo zobrazovať pri jednotlivých tagoch počet výskytov, ale ich dôležitosť sa dá zobrazit' napríklad spomínanou veľkosťou písma.

## 3.2 Poznámky vo vyhľadávaní

Vyhľadávanie je veľmi dôležitý krok v procese hľadania informácií. Na zvýšenie kvality výsledkov vyhľadávania sa úspešne používajú napríklad takzvané anchor texty, teda texty z odkazov na danú stránku. Tieto texty vytvorili autori iných webových stránok a dajú sa považovať za veľmi krátke charakteristiky zdrojov, na ktoré odkazy smerujú. Ďalším zdrojom informácií pri vyhľadávaní dokumentov je explicitná alebo implicitná spätná väzba získaná od návštevníkov dokumentov. Príkladom implicitnej spätnej väzby môže byť čas strávený na stránke, pohyb používateľa v dokumente, kliknutia myši, stlačenia kláves (Claypool et al. 2001; Holub and Bieliková 2011), ale aj dopyty, ktoré zadali používatelia do vyhľadávača na nájdenie nejakého zdroja, prípadne počet zobrazení stránky rôznymi používateľmi pri vyhľadávaní (Joachims 2002) podľa nejakého dopytu. Príkladom explicitnej spätnej väzby sú rôzne formuláre, hodnotenia a podobne.

S nástupom webu 2.0 samotní čitatelia stránok dostávajú možnosť prispievať k obsahu stránok. Jednou z foriem, ktorou môžu prispievať k obsahu stránok sú poznámky a značky, ktoré môžu používatelia vytvárať vďaka rastúcemu počtu služieb umožňujúcim vytváranie záložiek a pridávanie poznámok do webových

stránok (angl. social bookmarking services). Takéto používateľské poznámky opisujú stránky z pohľadu ich čitateľov (Zhang et al. 2009), a teda presnejšie opisujú informácie, ktoré v nich používatelia vidia. Je preto oprávnený predpoklad, že budú užitočné pri vyhľadávaní a navigácii medzi dokumentami.

V poslednej dobe vzniklo niekoľko úspešných služieb, ktoré umožňujú pridávať poznámky k webovým stránkam (Delicious<sup>6</sup>), ale aj k textovým dokumentom (Mendeley, Acrobat Reader<sup>7</sup>, Okular<sup>8</sup>), k fotkám (Flickr<sup>9</sup>), hudbe (Last.fm<sup>10</sup>) alebo videám (YouTube<sup>11</sup>). Sledujeme rastúcu popularitu týchto služieb, a teda aj stále sa zvyšujúce množstvo poznámok pridaných k webovým dokumentom.

Podobne ako informácie o obsahu dokumentov, prepojenia dokumentov a spätná väzba o používaní dokumentov, aj poznámky pripojené k dokumentom sa dajú použiť na vylepšenie výsledkov vyhľadávania. Poznámky sa dajú použiť pri vyhľadávaní dvomi spôsobmi:

- Keďže poznámky pridávajú ďalší obsah do dokumentu, dajú sa považovať sa súčasťou dokumentu (Zhang et al. 2009). Pri najčastejšie používanom vektorovom modeli dokumentu sa dajú priamočiaro pridať do reprezentácie dokumentu a jednoducho pridať do indexu. V takto vytvorenom indexe sa dajú použiť pri vyhľadávaní dokumentov.
- Používateľské poznámky opisujú záujmy používateľa, a preto sa dajú použiť ako spätná väzba o tom, ktorý dokument alebo jeho časť je pre používateľa zaujímavá.

Problémom, ktorý sa postupne objavuje, je vyhľadávanie v stále narastajúcom množstve poznámok (Li et al. 2007) a informácií, ktoré vytvorili používatelia a návštevníci dokumentov a internetových stránok. Množstvo poznámok a zdieľaných informácií, získaných zo služieb na vytváranie poznámok a zo sociálnych sietí, predstavuje ďalší obsah pripojený k webovým stránkam, ktorý treba efektívne prehľadávať a navigovať sa v ňom.

### 3.2.1 Vyhľadávanie pomocou poznámok

Existuje množstvo pokusov o použitie rôznych druhov poznámok vo vyhľadávaní dokumentov. Najrozšírenejším druhom poznámok, ktoré môžu používatelia pridávať do rôznych dokumentov sú tagy. Dá sa predpokladať, že ak je jeden dokument opísaný dvoma tagmi, tak tieto tagy spolu súvisia. Čím častejšie sa objavujú tieto dva tagy spolu, tým silnejší je náš predpoklad. Pomocou takto

---

<sup>6</sup>Delicious, <http://www.delicious.com>

<sup>7</sup>Acrobat Reader, <http://www.adobe.com/products/reader.html>

<sup>8</sup>Okular, <http://okular.kde.org>

<sup>9</sup>Flickr, <http://www.flickr.com>

<sup>10</sup>Last.fm, <http://www.last.fm>

<sup>11</sup>YouTube, <http://www.youtube.com>

vytvorených prepojení sa dajú vytvoriť folksonómie, teda siete konceptov a ich vzájomných vzťahov. Slovo folksonómia je prepojením slov *folks* (ľudia) a *taxonomy* (taxonómia alebo klasifikácia). Samotné slovo folksonómia naznačuje že ide o klasifikáciu, ktorú vytvorili bežní používatelia triedením dokumentov do svojich vlastných kategórií.

Výskumom vlastností folksonómií sa zaoberalo viacero autorov. Ukázali napríklad, že majú vlastnosti sietí malého sveta (Cattuto, La, and Roma 2007), teda že priemerná vzdialenosť medzi uzlami v grafe folksonómie je veľmi malá v porovnaní s počtom uzlov v grafe. Folksonómie vytvorené z používateľských poznámok sa dajú použiť na riešenie najrôznejších problémov, akým je napríklad personalizované vyhľadávanie (Jiao et al. 2008).

Jednou z hlavných prekážok pre širšie rozšírenie sémantického webu je malý počet zdrojov na webe, ktoré by boli vytvorené tak, aby im dokázali stroje „rozumieť“. Na to je potrebné aby boli tieto zdroje prepojené s konceptami a reláciami z rôznych ontológií. Na definovanie takýchto ontológií sú však obvyčajne potrební experti z danej domény a nemalé úsilie. Folksonómie sa dajú použiť ako ľahká forma ontológie (Hotho et al. 2006), a tak sa dajú do istej miery využiť práve pre potreby sémantického webu. Keďže na tvorbu folksonómií niesú potrební experti z danej domény, ale dokážu ich tvoriť bežní používatelia pomocou tagovania, sú veľmi cenným zdrojom údajov. Jeden z prístupov pre použitie tagov na vytvorenie strojovo spracovateľných metadát k webovým dokumentom je opísaný v (Wu, Zhang, and Yu 2006).

Na tagy pridané do dokumentov sa dá pozerat' ako na kategórie, do ktorých používatelia zaradili dokument. Dajú sa preto použiť podobne ako keby sme vyhľadávali nad webovými adresármi ako sú napríklad Open Directory Project<sup>12</sup> alebo Yahoo! Directory<sup>13</sup>.

Pri pohľade na poznámky ako na informácie, ktoré rozširujú obsah dokumentu sa dá použiť pomerne priamočiary spôsob ako využiť poznámky vo vyhľadávaní medzi dokumentami. Podobne ako anchor texty obohacujú informácie o stránke, na ktorú odkazy smerujú, poznámky sa dajú použiť na rozšírenie informácií o dokumente, ku ktorému sú pripojené. Pri reprezentovaní dokumentov ako vektor slov je možné poznámky jednoducho pridať do takto vytvoreného vektoru, prípadne je možné ich použiť na úpravu váh slov vo vektore.

V súčasnosti existuje viacero prístupov pre použitie poznámok pri vyhľadávaní (Zhang et al. 2009; Biancalana 2009). Za zmienku stojí napríklad definícia dokumentu ako kombinácie obsahu dokumentu a k nemu pripojených poznámok (Zhang et al. 2009). Autori tu použili poznámky podobne ako sa používajú anchor texty na doplnenie informácií o webovej stránke pri jej indexovaní. Pri anchor textoch je text odkazu spravidla dôležitejší ako samotný text na stránke, na ktorú odkaz smeruje. Podobne aj text poznámok je v mnohých prípadoch dôležitejší ako ostatný text stránky. Za dokument teda považovali nie len

<sup>12</sup>Open Directory Project, <http://www.dmoz.org>

<sup>13</sup>Yahoo! Directory, <http://dir.yahoo.com>

samotný obsah dokumentu, ale aj poznámky pripojené k dokumentu, čo umožnilo zapojiť do vyhľadávania aj návštevníkov pohľad na dokument.

### 3.2.2 Personalizované vyhľadávanie

Personalizácia vyhľadávania prispôbuje vyhľadávanie pre jednotlivých používateľov na základe ich záujmov alebo preferencií. Rôzne prístupy k personalizácii využívajú informácie o používateľových cieľoch, znalostiach, dokumentoch, ktoré prečítal alebo sa mu páčili a iných používateľských charakteristík. Pomocou týchto informácií je možné vyhľadávať dokumenty, ktoré používateľovi najviac vyhovujú, obohacovať jeho dopyty alebo mu odporúčať dokumenty, ktoré by mohli byť pre neho užitočné.

Pri čítaní tlačeneého ale aj elektronického dokumentu používateľa zvyčajne pridávajú poznámky k tým dokumentom alebo miestam v dokumente, ktoré ich nejakým spôsobom zaujali. Poznámky teda predstavujú významnú spätnú väzbu o tom, ktoré dokumenty používateľov zaujali, ktoré konkrétne časti dokumentov sú zaujímavé pre používateľov, a preto môžu byť cenným zdrojom údajov na personalizáciu vyhľadávania pomocou týchto metód.

#### Vytváranie a vyhodnocovanie dopytu

Jedným z možných prístupov k personalizácii je vyhľadávanie dokumentov, ktoré by mohli byť zaujímavé pre používateľa. Môžu to byť napríklad dokumenty, ktoré sú podobné tým, ktoré používateľ už čítal a ktoré ho zaujali. Na zistenie či používateľ dokument zaujal môžeme využiť explicitnú alebo implicitnú spätnú väzbu. Poznámky priradené k dokumentu môžeme považovať za istú formu implicitnej spätnej väzby, pomocou ktorej môžeme odvodiť, že tento dokument ho nejakým spôsobom zaujal. Bolo by preto zaujímavé ponúknuť používateľovi ďalšie dokumenty, ktoré sa týkajú podobnej témy ako dokument, ktorý práve čítal a ku ktorému pridával poznámky. Jedným z možných postupov ako používateľovi poskytnúť ďalšie súvisiace dokumenty je vytvorenie dopytu na základe dokumentov, ktoré už prečítal a vyhľadanie ďalších dokumentov.

Existuje množstvo spôsobov na vytváranie dopytov na základe zdrojového dokumentu. Najčastejšie ide o vyberanie významných slov alebo viet z obsahu dokumentu a ich použitie ako dopyt do vyhľadávača. Pomocou týchto metód je možné vytvárať dopyty na získavanie podobných dokumentov (Yang et al. 2009), dajú sa použiť na vyhľadanie citácií na konkrétne miesto v rozpracovanom článku (He et al. 2010), ale aj na odhaľovanie plagiátov (Pereira and Ziviani 2003). Na nájdenie dokumentov na základe zdrojového dokumentu je potrebné dokument najskôr predspracovať a vytvoriť z neho dopyt. Toto môžu byť napríklad dopyty v podobe:

- zoznamu kľúčových slov extrahovaných zo zdrojového dokumentu (Pereira and Ziviani 2003),

- zoznamu významných viet zo zdrojového dokumentu (Yang et al. 2009) alebo
- samotného obsahu dokumentu.

Často používaným postupom pri vytváraní dopytu z obsahu dokumentu je výber malého počtu slov z obsahu dokumentu a použitie tohto zoznamu slov ako dopytu. Na výber slov do tohto zoznamu sa používajú rôzne postupy a algoritmy ako napríklad:

- najmenej časté slová v texte, kde sa predpokladá, že tieto budú špecifické pre dokument (Pereira and Ziviani 2003),
- TF-IDF metrika na ohodnotenie slov v dokumente a na výber slov významných pre dokument a špecifických pre korpus, v ktorom sa vyhľadáva a
- množstvo ATR algoritmov (Korkontzelos, Klapaftis, and Manandhar 2008) a ďalších metód, ktoré vyberajú nejakým spôsobom významné slová z obsahu dokumentu.

Jednou z najčastejšie používaných metód na vytváranie dopytov na získanie súvisiacich článkov používa nástroj ElasticSearch<sup>14</sup>, ktorý umožňuje vyhľadávať dokumenty v indexe na základe dopytu, ktorý má formu obsahu dokumentu. Tento nástroj vnútorne používa nástroj Lucene<sup>15</sup>, a teda používajú spoločný prístup k spracovaniu takéhoto dopytu. Lucene umožňuje použiť takzvaný MoreLikeThis<sup>16</sup> dopyt, pomocou ktorého nástroj z pôvodného textu vyberie niekoľko kľúčových slov a tie použije ako dopyt do klasického vyhľadávania na základe zoznamu kľúčových slov. Pri výbere kľúčových slov sa používa ohodnotenie slov pomocou TF-IDF metriky a následne sa vyberie stanovený počet najvyššie ohodnotených slov.

Metódy, ktoré používajú na vytvorenie dopytu obsah dokumentu je možné obohatiť tak, aby zohľadňovali poznámky pripojené k dokumentu. Z tohto pohľadu je zaujímavý experiment, ktorý vykonali v (Golovchinsky, Price, and Schilit 1999). Požiadali dobrovoľníkov, aby prečítali sadu dokumentov viažúcich sa ku konkrétnej téme a aby v týchto dokumentoch zvýraznili relevantné časti. Na zvýrazňovanie používali tablet, ktorý umožňoval pridávať rôzne druhy poznámok priamo do textu dokumentu. Z takto vytvorených poznámok vytvorili dopyt, ktorý potom použili vo vyhľadávači na získanie podobných dokumentov. Dopyt mal formu zoznamu slov s priradenými váhami, kde váha slova závisela od druhu poznámok, ktoré k nemu boli priradené a od toho, koľkokrát bolo slovo zvýraznené. Vytvorené dopyty porovnávali s dopytmi vytvorenými len na základe

<sup>14</sup>ElasticSearch, <http://www.elasticsearch.org>

<sup>15</sup>Lucene, <http://lucene.apache.org/>

<sup>16</sup>MoreLikeThis, [http://lucene.apache.org/core/3\\_6\\_1/api/contrib-queries/org/apache/lucene/search/similar/MoreLikeThis.html](http://lucene.apache.org/core/3_6_1/api/contrib-queries/org/apache/lucene/search/similar/MoreLikeThis.html)



informácie o relevancii dokumentu ako celku. Poznámky vytvorené na základe poznámok v dokumente dosiahli významne vyššiu presnosť pri vyhľadávaní dokumentov ako dopyty získané na základe relevancie dokumentov.

Zaujímavý pohľad na personalizáciu vyhľadávania je uvedený v (Cai and Li 2010), kde vytvorili model používateľa ako aj model domény pomocou poznámok priradených k dokumentom. Vo väčšine súčasných prístupov založených na analýze obsahu funguje personalizované vyhľadávanie na princípe porovnávania modelu používateľa s modelom dokumentu. V tejto práci zavádzajú nový pohľad na vyhľadávanie dokumentov na základe modelu používateľa. Na problém vyhľadávania sa pozerajú ako na problém splňania ohraničení daných používateľským profilom. Problém pri porovnávaní modelu používateľa a modelu dokumentov spočíva v tom, že váha konceptu v týchto dvoch modeloch má rozdielnu interpretáciu. V modeli dokumentu váha konceptu reprezentuje informáciu o tom, ako relevantný je daný koncept pre dokument. Na rozdiel od toho v modeli používateľa váha konceptu reprezentuje informáciu o tom, ako je koncept pre používateľa zaujímavý, a teda ako veľmi je pre neho dôležitý. Veľká podobnosť medzi modelom používateľa a modelom dokumentu neznamená, že dokument je vhodný pre používateľa, ani naopak nízka podobnosť neznamená, že dokument pre neho nie je vhodný. Riešenie úloh splňania ohraničení sa ukázalo ako vhodné riešenie pre takýto typ úloh.

Zavádzajú teda vyhľadávanie relevantných dokumentov ako úlohu splňania ohraničení.

## Obohacovanie dopytu

Obohacovanie dopytov (angl. query expansion) je technika, pomocou ktorej automaticky pridávame slová do používateľského dopytu alebo meníme ich váhy tak, aby sa upresnil dopyt alebo naopak vytvoril dopyt, ktorý nájde širšie spektrum zaujímavých zdrojov. Používatelia obyčajne používajú dopyty, ktoré obsahujú len malý počet slov. Takéto dopyty sú nepresné a vyhovuje im veľký počet dokumentov, o ktoré používateľ nemá záujem. Existuje viacero metód, ktoré sa používajú na obohatenie takýchto dopytov (Chirita, Firan, and Nejdll 2007). Tieto metódy sa dajú rozdeliť do niekoľkých skupín:

- Metódy založené na úprave výsledkov na základe **spätnej väzby o relevancii nájdených dokumentov** používajú dokumenty získané s pomocou základného dopytu na získavanie slov, ktorými rozšíria dopyt. Tieto metódy pracujú v iteráciách, kde v každej iterácii používateľ vyberie relevantné dokumenty zo zoznamu nájdených dokumentov. Z týchto dokumentov sa vyberú slová, ktoré obohatia dopyt, pomocou ktorého sa vyhľadajú dokumenty v ďalšej iterácii.
- Metódy založené na **výskyte slov v nájdených dokumentoch** používajú dokumenty na prvých priečkach v zozname nájdených dokumentov na

nájdenie slov, ktoré sa v týchto dokumentoch často opakujú. Tieto slová sú následne doplnené do dopytu. Predpokladá sa, že zadaný dopyt vyhľadal na prvých priečkach relevantné dokumenty a tým, že sa pridajú najčastejšie slová z prvých nájdených dokumentov do dopytu sa zvýši vo výsledku počet dokumentov podobných tým, ktoré boli najrelevantnejšie k dokumentom z pôvodného dopytu.

Pri týchto metódach už nie je potrebná interakcia používateľa ako to bolo v predchádzajúcom prípade, kde musel používateľ vyberať relevantné dokumenty.

- Metódy založené na **slovníkoch** dopĺňajú používateľské dopyty o slová zo slovníkov, ktoré sa najčastejšie objavujú v kombinácii so slovami v dopyte. Na použitie týchto metód je potrebné mať pripravený slovník slov spolu s ich vzťahmi k ostatnými slovám. Na zachytenie vzťahov medzi slovami sa najčastejšie používa reprezentácia slov pomocou N-gramov. Medzi tieto metódy možno zaradiť aj metódy, ktoré používajú na nájdenie rozširujúcich slov rôzne externé služby (Yin, Shokouhi, and Craswell 2009) ako sú Wordnet<sup>17</sup> alebo Flickr<sup>18</sup>.
- Ďalšou kategóriou metód na obohacovanie dopytov sú metódy, ktoré používajú **dopyty iných používateľov** (Billerbeck et al. 2003). Do dopytu sa pridávajú tie slová, ktoré použili používatelia na vyhľadanie dokumentov, ktoré sa nachádzajú v zozname získanom pôvodným dopytom. Tento postup sa do istej miery podobá na metódy založené na obohacovaní dopytu o slová z obsahu nájdených dokumentov, len v tomto prípade sa dopyt obohacuje o slová, ktoré použili iní používatelia pri hľadaní týchto dokumentov.

Veľmi dobrý prehľad rôznych metód na obohacovanie dopytu je prezentovaný v (Abbasi 2011), kde autori navrhujú metódu na obohacovanie dopytov použitých na vyhľadávanie vo folksonómiách. Zaujímavé sú najmä metódy, ktoré obohacujú dopyty tagmi z folksonómie na základe podobnosti medzi slovami v dopyte a týmito tagmi.

Zaujímavá obmena metód založených na slovníkovom prístupe je prezentovaná v (Chirita, Firan, and Nejd 2007). Tu je slovník spoločných výskytov slov vytvorený zo všetkých osobných dokumentov konkrétneho používateľa. Pomocou takto rozšírených dopytov dokázali autori významne zvýšiť presnosť vyhľadávania oproti iným bežne používaným metódam.

Použitie poznámok pri obohacovaní dopytov je opísané aj v (Biancalana 2009). Táto práca je postavená na obohacovaní dopytov na základe metód používajúcich na obohacovanie dopytov slová z prvých nájdených dokumentov. V tomto prípade však neboli do dopytu pridávané slová z obsahu nájdených dokumentov, ale tagy,

<sup>17</sup>Wordnet, <http://wordnet.princeton.edu/>

<sup>18</sup>Flickr getRelated, [flickr.com/services/api/flickr.tags.getRelated.ht](http://flickr.com/services/api/flickr.tags.getRelated.ht)

ktoré pridali ostatní používatelia k nájdeným dokumentom. Pri obohacovaní dopytu teda neboli použité poznámky, ktoré vytvoril používateľ, ktorého dopyt rozširovali, ale boli použité poznámky, ktoré k dokumentom priradili iní používatelia.

### Usporiadanie výsledkov vyhľadávania

Najjednoduchším spôsobom ako použiť poznámky vo vyhľadávaní je pridať poznámky priamo k obsahu dokumentu a vytvoriť nad nimi index podobne ako to navrhli v (Zhang et al. 2009). Takto sa však využíva len obsah poznámok a ignorujú sa nielen vzťahy medzi dokumentami, ale aj vzťahy medzi poznámkami.

Poznámky sa dajú pri vyhľadávaní použiť dvoma spôsobmi:

- Dajú sa použiť ako označenia zaujímavých alebo kvalitných dokumentov.
- Obsah poznámok sa dá použiť ako dodatočná informácia k obsahu dokumentu a k informáciám o jeho vzťahoch k ostatným dokumentom.

V práci (Bao et al. 2007) autori navrhujú algoritmy, ktoré využívajú oba tieto pohľady. Navrhli dva algoritmy: SocialSimRank pre výpočet podobnosti medzi vytvorenými poznámkami a dopytmi a SocialPageRank, ktorý určuje popularitu dokumentov na základe počtu pripojených poznámok.

Najznámejším algoritmom používaným na usporiadavanie dokumentov na základe k nim priradených poznámok je algoritmus FolkRank (Hotho et al. 2006). FolkRank je algoritmus inšpirovaný algoritmom PageRank (Page et al. 1999). Tento algoritmus zohľadňuje pri usporiadavaní dokumentov štruktúru folksonómie, ktorá vzniká z poznámok pripojených k dokumentom. FolkRank pracuje nad folksonómiou reprezentovanou ako tripartitný graf, kde vrcholy sú tagy, používatelia a dokumenty. Uzly sú prepojené ováňovanými hranami, ktoré reprezentujú počet tagov, ktoré pridali používatelia do dokumentov. Upravený PageRank potom ohodnocuje vrcholy, pričom zohľadňuje ohodnotenie hrán.

Ďalším príkladom algoritmu zohľadňujúceho štruktúru folksonómie pri usporiadavaní opoznámkovaných dokumentov je SocialHITS (Abel et al. 2009) založený na algoritme HITS (Kleinberg 1999), ktorý pracuje s folksonómiou ako s orientovaným grafom a objavuje v nej centrá (*hub*) a autority (*authority*).

## 3.3 Zhodnotenie využívania poznámok pri navigácii

Existuje množstvo nástrojov na pridávanie poznámok do elektronických dokumentov. Vytvorené poznámky sa najčastejšie používajú ako prostriedok na

organizáciu vlastnej zbierky dokumentov, ale aj ako prostriedok na zlepšenie vyhľadávania dokumentov alebo na vytváranie folksonómií.

Poznámky sa používajú ako indikátory záujmu o dokument, ale zatiaľ len vo veľmi obmedzenej miere. Okrem niekoľkých výnimiek, ako napríklad (Carmel et al. 2010), sa pri ohodnocovaní dokumentov na základe priradených poznámok zohľadňuje len počet používateľov, ktorí si pridali záložku k dokumentu. Na poznámky sa spravidla nepozera ako na označenie konkrétnych častí dokumentu, o ktorú má používateľ záujem, ale len ako na indikáciu kvality dokumentu ako celku. Táto informácia by sa však dala použiť napríklad pri indexovaní dokumentov, modelovaní záujmov používateľa, odporúčaní dokumentov alebo pri vyhľadávaní súvisiacich dokumentov.



# Existujúce nástroje na podporu poznámkovania na webe

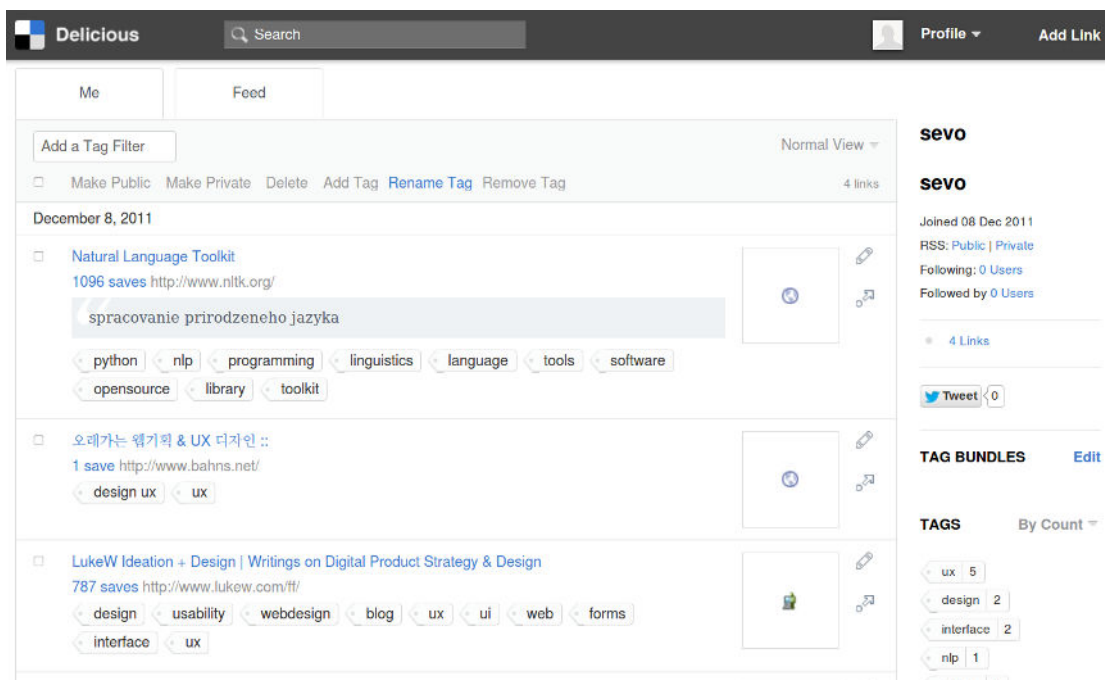
Na podporu manuálneho vytvárania poznámok existuje veľa rôznych nástrojov ako napríklad už spomínané služby Diigo a Delicious, ktoré umožňujú pridávať poznámky do ľubovoľnej webovej stránky. Okrem všeobecne zameraných služieb existuje množstvo služieb, ktoré využívajú poznámky v konkrétnej doméne. Sú to služby ako napríklad služba na správu knižnice zaujímavých odborných článkov CiteULike<sup>19</sup>, výučbový systém ALEF (Šimko et al. 2011) alebo aplikácia na správu zbierky článkov Mendeley. Rôzne služby poskytujú nástroje pre pridávanie poznámok nielen do textových dokumentov, ale aj do videí (YouTube.com), fotiek (flickr.com) alebo hudby (Last.fm). Viaceré z týchto služieb sú zamerané na poznámkovanie (vytváranie záložiek, tagovanie ...) a poskytujú možnosť vytvárať poznámky ako hlavnú funkčnosť. Existuje však veľmi veľké množstvo aplikácií ako napríklad rôzne webové služby ale aj rôzne blogy, ktoré poskytujú možnosť vytvárať poznámky (najčastejšie tagy) ako dodatočnú funkčnosť.

## 4.1 Delicious

Delicious (Obr. 4.1) je asi najznámejšia služba na vytváranie záložiek k webovým stránkam. S použitím záložiek a tagov vytvorených pomocou tejto služby bolo vykonaných množstvo experimentov a výskumov týkajúcich sa správania používateľov pri tagovaní (Golder and Huberman 2006; Körner et al. 2010), použitia značiek vo vyhľadávaní (Bao et al. 2007; Hotho et al. 2006) a výskumov

---

<sup>19</sup>CiteULike, <http://www.citeulike.org>



Obr. 4.1: Zoznam záložiek vytvorený pomocou služby Delicious

týkajúcich sa vlastností folksonómií vytvorených z takto získaných tagov (Cattuto, La, and Roma 2007; Wu, Zhang, and Yu 2006).

Pomocou tejto služby môžeme odkladať webové stránky medzi záložky a k takto vytvoreným záložkám môžeme priradovať tagy, pomocou ktorých je neskôr možné záložky vyhľadávať a filtrovať. Na filtrovanie záložiek pomocou tagov slúži jednoduchý zoznam tagov usporiadaný podľa počtu záložiek, ku ktorým je daný tag priradený. Po kliknutí na tag sa zobrazia len tie záložky, ku ktorým bol tag priradený. Zložitejší filter je možné vytvoriť kombináciou viacerých tagov, kedy sa zobrazia len tie dokumenty, ktoré majú priradené všetky tieto tagy.

Služba poskytuje veľmi jednoduché rozhranie, vo forme rozšírenia do najpoužívanejších prehliadačov. Prezerat' si zbierku záložiek a vyhľadávať vo vlastných a verejných záložkách je možné prostredníctvom webového rozhrania.

Podporované je zaradovanie záložiek do takzvaných vriec (angl. stacks), ktoré sú určené na spájanie záložiek so spoločnou témou. Okrem tagov je k záložkám možné pripojiť krátky komentár. Pridávanie ďalších druhov poznámok ako napríklad zvýraznení v texte nieje podporované. Spolupráca medzi používateľmi je podporovaná prostredníctvom nasledovania rôznych používateľov a nasledovania vriec záložiek. Podobne je možné sledovať zoznam najpopulárnejších záložiek.



Obr. 4.2: Dokument opoznamkovaný pomocou nástroju Diigo

## 4.2 Diigo

Diigo (Obr. 4.2) je aplikácia určená na vytváranie záložiek a pridávanie poznámok a tagov do webových stránok. Tagy je možné použiť na vyhľadávanie a filtrovanie záložiek. Okrem tagov je možné k stránkam pridávať ďalšie typy poznámok ako napríklad zvýraznenia v texte a komentáre. Stránky je možné ukladať ako obrázky, do ktorých je možné pridávať ďalšie grafické značky ako napríklad rôzne orámovania alebo šípky. Aplikácia podporuje zdieľanie vytvorených záložiek a poznámok pre všetkých používateľov alebo v rámci skupiny. Podporovaná je tiež integrácia so sociálnymi sieťami Twitter a Facebook, takže používateľ dokáže zdieľať poznámky a záložky aj prostredníctvom týchto sociálnych sietí.

Aplikácia poskytuje množstvo funkcií na vytváranie a zdieľanie poznámok, sústreďuje sa na to, aby bolo vytváranie poznámok prirodzené a jednoducho použiteľné. Napríklad zvýrazňovanie častí textu funguje tak, že používateľ si zvolí farbu a každý text, ktorý potom označí bude touto farbou zvýraznený. Tento postup sa snaží napodobniť zvýrazňovanie textu v tlačných dokumentoch, keď v ruke držíme zvýrazňovač a zvýrazňujeme zaujímavý text, nemusíme pri každej novej poznámke znova vyberať zvýrazňovač ako to robí napríklad prehliadač dokumentov Okular.

Zvýrazňovanie každého označeného textu je veľmi jednoduché na naučenie, keďže rovnaký proces používajú ľudia pri zvýrazňovaní textu v tlačných dokumentoch. Tento postup je však spojený s rizikom, že používateľ omylom označí veľký úsek textu či už kvôli nepresnosti myši alebo ak zabudne že má aktivovanú túto funkcionality. Preto služba podporuje aj druhý spôsob ako zvýrazniť text vo webovej stránke. Zvýrazniť text v stránke je možné označením ľubovoľného textu myšou, vyvolaním kontextového menu pravým kliknutím na text a zvolením možnosti pre zvýraznenie (voľba „Highlight“).

Spolupráca pri vytváraní poznámok je podporovaná prostredníctvom záujmových



skupín, ktoré vedia zdieľať poznámky a záložky a majú tak vytvorený spoločný priestor na poznámkovanie dokumentov.

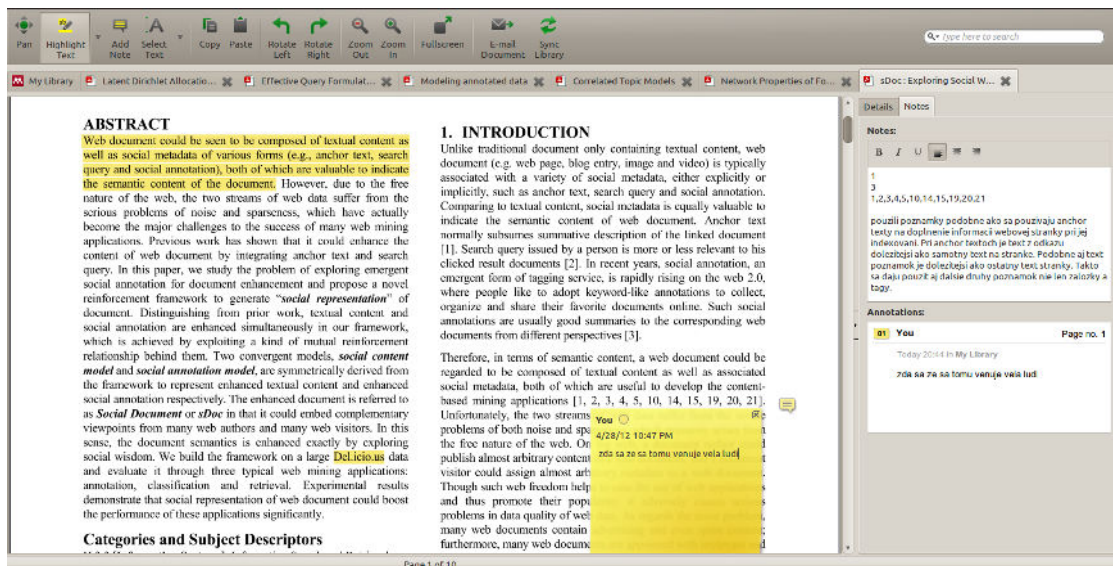
Vyhľadávanie vo vytvorených poznámkach je podporované prostredníctvom filtrovania záložiek podľa tagov. Možné je tiež vyhľadávať vo verejných poznámkach a v obsahu stránok pridaných medzi záložky. Podporované je teda fulltextové vyhľadávanie v metadátach záložiek ako aj v obsahu samotných záložiek.

### 4.3 Mendeley

Mendeley je aplikácia na organizovanie vedeckých publikácií zložená z dvoch častí: webového rozhrania a desktopovej aplikácie. Na obrázku 4.3 je vidieť okno desktopovej aplikácie, v ktorom je zobrazený PDF dokument a do neho pridané poznámky. Možné je pridávať poznámky v podobe zvýraznení v texte, komentárov umiestnených na konkrétne miesto v dokumente a komentárov k celému dokumentu. Okrem toho je možné k dokumentom priradovať tagy.

Na rozdiel od aplikácie Diigo, opísanej v predchádzajúcej časti, Mendeley podporuje len zvýrazňovanie častí textu tým, že si používateľ zvolí funkciu zvýrazňovania a každý text, ktorý následne zvýrazní bude označený. Tento spôsob zvýrazňovania textu spôsobuje problémy ak používateľ omylom označí príliš veľký úsek textu, označí iný text ako chcel alebo zabudne, že má túto funkciu aktivovanú. Nesprávne označenie textu je zvlášť časté pri zvýrazňovaní textu v PDF dokumentoch kvôli spôsobu ich zobrazovania.

Vytvorené poznámky je možné zdieľať medzi ďalšími používateľmi prostredníctvom zdieľania dokumentov v privátnych skupinách. Nie je možné do dokumentov vytvoriť verejné poznámky, čo je funkcia, ktorú podporuje veľká väčšina podobných aplikácií. Rovnako nie je podporované vyhľadávanie medzi dokumentami pomocou poznámok ani vyhľadávanie v samotných poznámkach. Dokumenty je možné vyhľadávať len na základe ich metadát ako je nadpis alebo abstrakt. Poznámky sú použité len pri vizualizácii obsahu dokumentu, ku ktorému sú priradené a nedávajú žiadnu ďalšiu pridanú hodnotu. Vzhľadom na veľký počet používateľov po celom svete má aplikácia k dispozícii veľké množstvo informácií o dokumentoch ako aj o používateľoch, ale v súčasnosti ich nevyužíva na vyhľadávanie alebo na inú formu podpory navigácie.



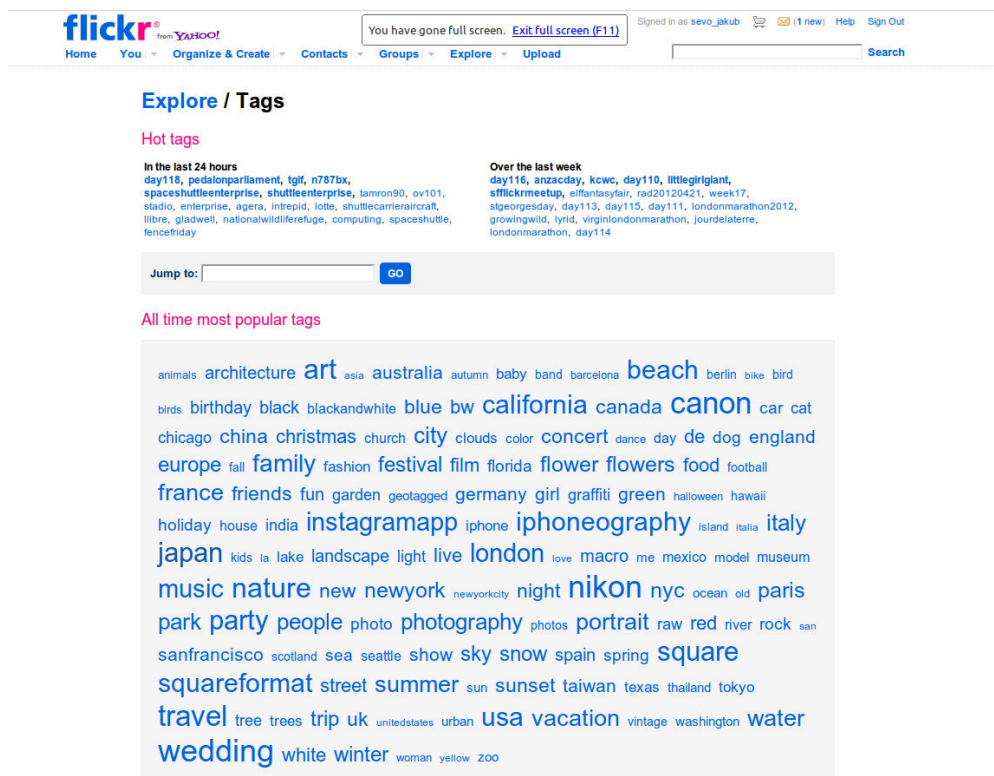
Obr. 4.3: Pridávanie poznámok v nástroji Mendeley

## 4.4 Flickr

Flickr je sociálna sieť na zdieľanie fotografií, ktorá na riešenie problému organizácie a vyhľadávania obrázkov používa poznámky, ktoré k obrázkom pridávajú používatelia. Autor fotografie k nej môže pridávať tagy a tak zlepšiť organizáciu svojich fotiek ako aj zlepšiť možnosti na jej vyhľadanie. Vyhľadávať pomocou tagov je možné vo vlastnej sade fotiek ako aj medzi obrázkami ostatných používateľov. Vyhľadávanie je podporované dvoma spôsobmi, fulltextovým vyhľadávaním medzi tagmi a komentármi pripojenými k obrázkom a pomocou oblaku tagov (obrázok 4.4). Oblak tagov je organizovaný podľa najpopulárnejších tagov za rôzne obdobia a najpopulárnejších tagov vôbec.

Do jednotlivých fotiek je možné umiestňovať komentáre, ktoré sa viažu na konkrétne miesto na fotke. Rovnakým spôsobom je možné na fotkách označovať aj osoby z kontaktov. Komentáre je možné pridávať aj k celej fotke a nielen ku konkrétnemu miestu.

V súčasnosti je pomerne veľký problém spracovávať a vyhľadávať v obrázkoch. Napriek tomu že existuje veľa prístupov na vyhľadávanie medzi obrázkami, najlepšie vedia obrázky spracovávať ľudia. Preto používateľmi vytvorené poznámky pripojené k obrázkom sú veľmi cenou informáciou pre zlepšenie navigácie medzi obrázkami. Tento princíp sa dá použiť nielen pri organizácii fotiek, ale aj pri organizácii ďalších multimediálnych informácií.



Obr. 4.4: Navigácia medzi fotkami pomocou oblaku kľúčových slov v službe Flickr

## 4.5 Bibsonomy

Bibsonomy<sup>20</sup> (obrázok 4.5) je aplikácia, ktorá kombinuje vlastnosti služieb na vytváranie záložiek a systémov na zdieľanie referencií a publikácií. Umožňuje vytvárať záložky na bežných stránkach, ale hlavne pridávanie publikácií a referencií medzi záložky. Tieto záložky je možné organizovať pomocou tagov a pridávať k nim poznámky v podobe textu pripojenému k celému dokumentu. Na rozdiel od iných aplikácií, ktoré používajú tagy na organizovanie zdrojov, Bibsonomy nedovolí používateľom vytvoriť záložku bez toho, aby k nej nepridal žiadny tag. To svedčí o tom, akú významnú rolu hrajú tagy pri organizovaní zdrojov v tejto službe. Navigácia medzi vytvorenými záložkami je podporovaná prostredníctvom filtrovania zoznamu záložiek na základe priradených tagov. Je možné vytvárať takzvané koncepty, ktoré predstavujú skupiny súvisiacich tagov a používať tieto pri filtrovaní.

Aplikácia umožňuje zdieľanie vytvorených záložiek prostredníctvom zdieľania záložiek v skupinách. Takto môžu celé výskumné skupiny spolu zdieľať spoločný zoznam záložiek a referencií. Spolupráca skupiny pri zbieraní poznámok je podporovaná aj prostredníctvom možnosti vytvorenia a zapojenia sa do diskusií.

Záložky je možné vytvárať priamo vložením adresy stránky, alebo aj vložením

<sup>20</sup>Bibsonomy, <http://www.bibsonomy.org/>

Obr. 4.5: Úvodná obrazovka služby na vytváranie záložiek a organizovanie publikácií Bibsonomy

ISBN kódu. Na vytváranie záložiek je možné používať rozšírenia pre najčastejšie používané webové prehliadače. Bibsonomy taktiež podporuje import a export bibliografie medzi rôznymi službami a formátmi vrátane exportu citácií a možnosti definovať si vlastný štýl citácií.

Služba poskytuje veľmi bohaté prostriedky na zbieranie a organizovanie webových stránok a publikácií. Organizácia pomocou tagov je veľmi dobre zvládnutá, ale práca so samotným dokumentom je tu podporovaná len povrhu. Je možné nastaviť rôzne informácie o dokumente, ale napríklad pridanie poznámky k dokumentu je pomerne ukryté a tieto poznámky sa nikde nepoužívajú a nezobrazujú.

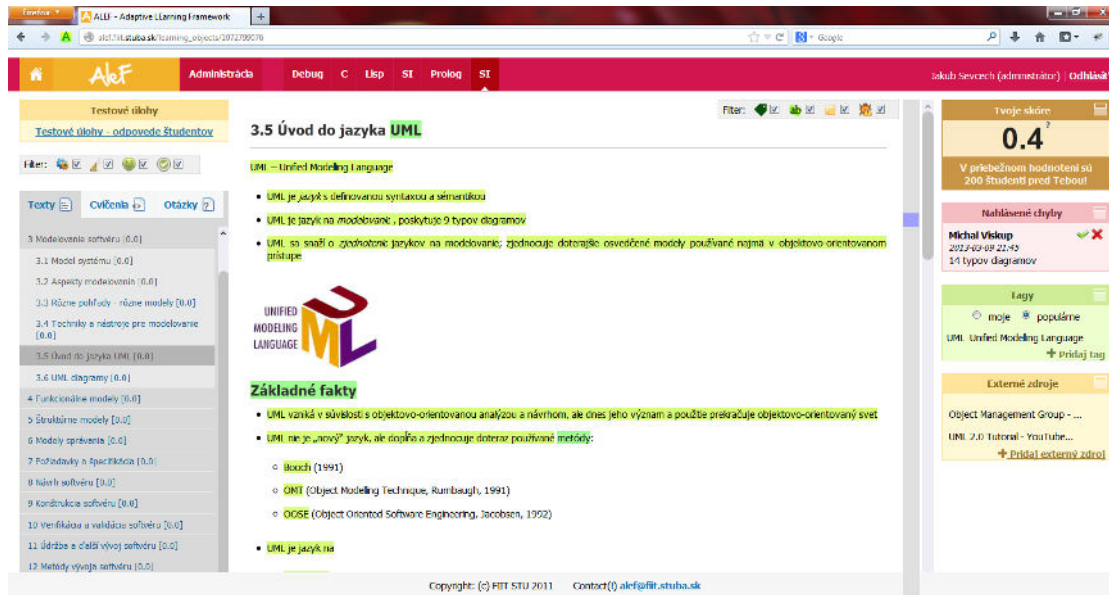
## 4.6 Alef

Alef (Šimko et al. 2011) je výučbový systém vyvíjaný a používaný na Fakulte informatiky a informačných technológií Slovenskej technickej univerzity v Bratislave. Autor tejto práce spolupracoval pri vývoji systému Alef.

Výučbový systém umožňuje študentom zobrazovať výučbové texty a pridávať k nim rôzne druhy poznámok. Hlavnou úlohou poznámok v tomto výučbovom systéme je rozširovanie výučbových objektov informáciami od samotných študentov. Podporované je pridávanie:

- komentárov k jednotlivým častiam textu,

- hlásenie chýb v texte,
- pridávanie externých odkazov,
- zvyrazňovanie dôležitých častí textu a
- vytváranie tagov.



Obr. 4.6: Výučbový text v systéme Alef s vytvorenými poznámkami

Poznámky sú zobrazované zvýraznením opoznámkovaného textu, pomocou bočného pásiku, ktorý zobrazuje pozíciu poznámky alebo pomocou rôznych zásuvných modulov (obrázok 4.6). Zásuvné moduly slúžia na vytváranie rôznych druhov poznámok, ich zobrazovanie a ďalšiu prácu s nimi. Alef poskytuje možnosť filtrovať poznámky podľa ich typu, čo zabezpečuje ich prehľadné použitie aj keď ich počet postupne rastie.

Navigácia pomocou poznámok je podporovaná pomocou filtrovania výučbových textov na základe tagov, ktoré k výučbovým objektom priradili samotní študenti. Ďalej sa poznámky využívajú na zlepšovanie obsahu kurzu pomocou oznamovania nájdených chýb a na obohacovanie obsahu pomocou používateľských otázok a komentárov.

## 4.7 Diskusia k súčasnému stavu použitia poznámok

Veľa rôznych nástrojov umožňuje vytvárať poznámky do webových stránok, textových dokumentov, ale aj multimediálnych dokumentov ako sú napríklad

obrázky, videá alebo hudba. Väčšina z týchto nástrojov sa sústreďuje hlavne na vytváranie poznámok a len pomerne jednoduché použitie poznámok na navigáciu. Poznámky však poskytujú viac cenných informácií. Tiež stále pretrváva problém veľmi malého pokrytia dokumentov na webe poznámkami. Treba sa preto sústrediť aj na poskytovanie pridanej hodnoty za to, že používatelia tvoria poznámky. Takouto pridanou hodnotou by mohlo byť napríklad zlepšenie navigácie medzi dokumentami a vyhľadávanie dokumentov.

Pri vytváraní poznámok sa tieto služby často snažia napodobňovať proces pridávania poznámok do tlačенých dokumentov. Vidieť to napríklad pri zvýrazňovaní textu, kde kurzor myši simuluje farebný zvýrazňovač a všetko čo používateľ označí zostane farebne zvýraznené. Takýto proces zvýrazňovania je napríklad v službe Diigo a v aplikácii Mendeley. Dôsledné napodobňovanie procesu vytvárania poznámok do tlačенých dokumentov je intuitívne, má však riziká spojené s rozdielmi v možných spôsoboch interakcie s elektronickými a tlačnými dokumentami.

Najčastejšie poskytovanou pridanou hodnotou za vytváranie poznámok je lepšia navigácia medzi dokumentami v osobnej zbierke dokumentov. Výsledok námahy vlozenej do pridávania poznámok do dokumentov sa však pri tomto druhu motivácie prejavuje až niekedy v budúcnosti, keď je dokumentov dostatočný počet alebo pri znovunavštívení opoznámkovaných dokumentov. Problémom je, že používateľ potrebuje motiváciu pre poznámkovanie skôr, aby vôbec poznámky vytváral.

Najčastejšími formami navigácie medzi dokumentami pomocou poznámok a tagov je filtrovanie dokumentov na základe priradených tagov, oblak tagov a fulltextové vyhľadávanie nových dokumentov pomocou tagov a ďalších priradených poznámok. Najmä pri vyhľadávaní na základe priradených poznámok vidíme veľký priestor na zlepšenie.

V tabuľkách 4.1 a 4.2 opisujeme niekoľko ďalších nástrojov a služieb, ktoré umožňujú používateľom vytvárať rôzne typy poznámok a používajú ich ako prostriedky komunikácie medzi používateľmi, na zaznamenávanie postrehov, na zlepšenie navigácie a podobne. Tabuľka 4.1 poskytuje krátku charakteristiku nástrojov a odkaz na ich domovskú stránku. Tabuľka 4.2 opisuje aký typ poznámok jednotlivé nástroje používajú a k akému obsahu je možné pridávať poznámky.

| Názov         | Účel  | Adresa  |
|---------------|---|---|
| Alef          | Kolaboratívny výučbový systém   | <a href="http://alef.fiit.stuba.sk">http://alef.fiit.stuba.sk</a>   |
| Bibsonomy     | Zdieľanie záložiek a zoznamov literatúry                                | <a href="http://www.bibsonomy.org">http://www.bibsonomy.org</a>     |
| CiteULike     | Zdieľanie vedeckých referencií  | <a href="http://www.citeulike.org">http://www.citeulike.org</a>     |
| Delicious     | Služba na vytváranie záložiek a organizáciu pomocou tagov               | <a href="https://delicious.com">https://delicious.com</a>           |
| Digg          | Zviditeľňovanie webového obsahu pomocou hlasovania                      | <a href="http://www.digg.com">http://www.digg.com</a>               |
| Diigo         | Vytváranie záložiek a pridávanie rôznych druhov poznámok                | <a href="https://www.diigo.com">https://www.diigo.com</a>           |
| Evernote      | Uchovávanie a organizovanie poznámok                                    | <a href="https://www.evernote.com">https://www.evernote.com</a>     |
| Flickr        | Zdieľanie fotiek  | <a href="http://www.flickr.com">http://www.flickr.com</a>           |
| Goodreads     | Používateľmi tvorená databáza kníh, poznámok a hodnotení                | <a href="http://www.goodreads.com">http://www.goodreads.com</a>     |
| Last.fm       | Objavovanie, zdieľanie a organizovanie hudobnej kolekcie                | <a href="http://www.last.fm">http://www.last.fm</a>                 |
| Mendeley      | Organizovanie a zdieľanie výskumných článkov                            | <a href="http://www.mendeley.com">http://www.mendeley.com</a>       |
| Pocket        | Odkladanie webových stránok na neskoršie prečítanie                     | <a href="http://getpocket.com">http://getpocket.com</a>             |
| StackOverflow | Kolaboratívne získavanie odpovedí na otázky týkajúcich sa programovania | <a href="http://stackoverflow.com">http://stackoverflow.com</a>     |
| StumbleUpon   | Vyhľadanie a odporúčanie webových zdrojov pomocou hlasovania            | <a href="http://www.stumbleupon.com">http://www.stumbleupon.com</a> |
| YouTube       | Objavovanie, prezeranie a zdieľanie videí                               | <a href="https://www.youtube.com">https://www.youtube.com</a>       |

Tabuľka 4.1: Služby podporujúce tvorbu poznámok

| Názov         | Podporované typy poznámok   | Typ obsahu                            |
|---------------|---|---------------------------------------|
| Alef          | tagy, zvýraznenia, komentáre, hlásenia o chybách  | výučbové texty                        |
| Bibsonomy     | tagy, poznámky, hodnotenia  | knihy, vedecké články                 |
| CiteULike     | tagy, komentáre   | vedecké referencie                    |
| Delicious     | tagy, zoznamy (stack)   | webové stránky                        |
| Diigo         | zvýraznenia, tagy, poznámky, komentáre, rôzne grafické značky (šípky, čiary, rámy), označenia na neskoršie prečítanie | webové stránky                        |
| Digg          | hlasy   | webové stránky                        |
| Evernote      | tagy  | poznámky                              |
| Flickr        | tagy od autora fotky  | fotky                                 |
| Goodreads     | citáty, hodnotenia, recenzie, označenia na neskoršie prečítanie   | knihy                                 |
| Last.fm       | tagy, playlisty   | piesne, autori                        |
| Mendeley      | zvýraznenia, komentáre k miestu v texte, tagy, poznámky k celému dokumentu  | vedecké články, webové stránky, knihy |
| Pocket        | označenia na neskoršie prečítanie, označenia ako obľúbené   | webové stránky                        |
| StackOverflow | tagy od autora otázky   | otázky                                |
| StumbleUpon   | komentáre, hlasy  | webové stránky                        |
| YouTube       | hlasy, označenia na neskoršie pozretie  | videá                                 |

Tabuľka 4.2: Typy poznámok používané rôznymi službami

Rôzne druhy poznámok pridaných k dokumentom sú veľmi populárnym prostriedkom na zlepšenie organizácie alebo navigácie v obsahu. Veľmi často môžu k obsahu pridávať poznámky všetci používatelia a môžu tak spolupracovať na obohacovaní dokumentov (Diigo, Mendeley) alebo zviditeľňovaní dokumentov (Delicious, Digg, StumbleUpon). V nástrojoch ako napríklad Flickr alebo StackOverflow však poznámky môžu pridávať len autori dokumentov, v prípade týchto dvoch nástrojov autori fotiek alebo otázok. V takýchto nástrojoch nie sú poznámky používané na organizáciu vlastnej sady dokumentov, ale slúžia na zlepšenie nájditel'nosti zdroja a zlepšenie organizácie dokumentov všetkých používateľov.

Častokrát je používanie poznámok spojené s nejakou formou spolupráce viacerých používateľov ako napríklad zdieľanie zdrojov v skupinách, kolaboratívne pridávanie poznámok alebo sledovanie aktivity používateľov. Vtedy sú poznámky používané nie len na organizáciu vlastnej knižnice zdrojov, ale aj ako prostriedok pre spoluprácu a pre zlepšenie navigácie medzi zdrojmi aj pre ostatných používateľov.





# Pripájanie poznámok k dokumentom

Poznámky možno považovať za dodatočné informácie pripojené k dokumentom, ktoré obohacujú obsah dokumentu (Agosti and Ferro 2007). Kľúčovým prvkom pri vytváraní poznámok do dokumentov je výber metódy na prepojenie častí dokumentu, kde má byť poznámka pripojená a samotných poznámok. Pri pripájaní poznámok k dokumentom je potrebné poznať identifikáciu dokumentu a v prípade poznámok, ktoré sa viažu na konkrétne miesto v dokumente aj údaje potrebné na identifikáciu tohto miesta.

## 5.1 Umiestnenie poznámok do webových stránok

Viacere systémy predpokladajú pri pridávaní poznámok že dokumenty sa nebudú meniť (Adobe Acrobat Reader<sup>21</sup>). Toto je veľmi silný predpoklad, ktorý však napríklad v prostredí webových stránok nemôžeme použiť a musíme navrhnúť spôsob na vytváranie a opätovné zobrazovanie poznámok s ohľadom na dokumenty, ktoré sa môžu kedykoľvek zmeniť.

V práci (Phelps and Wilensky 2000) opisujú niekoľko kritérií, ktoré musí spĺňať prístup k umiestňovaniu poznámok do dokumentov na to aby bol robustný:

- Umiestnenie musí byť odolné voči zmenám v dokumente.
- Umiestnenie musí byť odolné voči malým zmenám, ale v prípade veľkých zmien v obsahu dokumentu nesmie pripojiť poznámky k miestam kde nepatria, ale mala by byť zabezpečená možnosť na oznámenie že pripojenie poznámok zlyhalo.

---

<sup>21</sup>Adobe Acrobat Reader, <http://www.adobe.com/products/reader.html>

- Umiestňovanie musí byť založené na obsahu dokumentu. Pre dokumenty, ktoré sa nemenia alebo majú nemenné rozloženie (PDF dokumenty, obrázky) môže byť akceptovateľné aj absolútne umiestnenie na nejaké súradnice v zobrazenom dokumente.
- Musí pracovať bez nutnosti spolupráce so zdrojom dokumentu, ktorý by poskytol informácie o zmenách v dokumente.
- Musí pracovať s existujúcou infraštruktúrou (dokumenty, servery, klientské aplikácie).
- Musí vyžadovať len malé množstvo údajov potrebných na umiestnenie vzhľadom na veľkosť celého dokumentu.

Súčasne v tejto práci navrhujú niekoľko prístupov, ktoré tieto kritériá spĺňajú:

- Unikátny identifikátor pre jednotlivé časti dokumentu, ku ktorým sú pridané poznámky. Tento prístup poskytuje najlepšiu odolnosť voči zmenám v dokumente, vyžaduje však spoluprácu autora alebo zdroja dokumentov.
- Umiestnenie poznámok na základe pozície v strome dokumentu. Dokument rozdelili do stromu na základe hierarchie elementov v dokumente a poznámky sa pripájajú k niektorému uzlu v strome. Tento postup je do istej miery odolný voči zmenám obsahu, ale náchylný na chyby v prípade, ak sa mení štruktúra dokumentu.
- Umiestnenie na základe kontextu používa na umiestnenie poznámky v dokumente okolie tejto poznámky. Tento spôsob pripojenia poznámky do dokumentu je pomerne odolný voči zmenám v texte dokumentu a aj v texte, ku ktorému je pripojená poznámka v prípade, ak sa nehľadá presná zhoda medzi textom dokumentu a okolím poznámky.

Rôzne aplikácie používajú rôzne reprezentácie pre pripojenie poznámok k dokumentom. Napríklad nástroj Annotea (Kahan 2002) alebo MADCOW (Bottoni et al. 2005) umiestňuje poznámky do HTML dokumentu na základe XPointeru. V práci (Agosti and Ferro 2007) opisujú pripájanie poznámok pomocou XPath selektoru do XML dokumentov. Tento postup sa dá použiť okrem XML dokumentov aj pre bežné webové stránky. Je však spojený s problémami v prípade ak sa dokument zmení.

V práci (Ciccarese et al. 2011) je opísaná schéma na jednotnú reprezentáciu poznámok medzi rôznymi aplikáciami a doménami. Poznámky sú reprezentované pomocou RDF oddelene od dokumentov. Poznámky rôznych typov môžu byť pripojené ku konkrétnej časti dokumentu. Definovaných je niekoľko spôsobov ako je možné pridávať poznámku ku konkrétnemu miestu v texte:

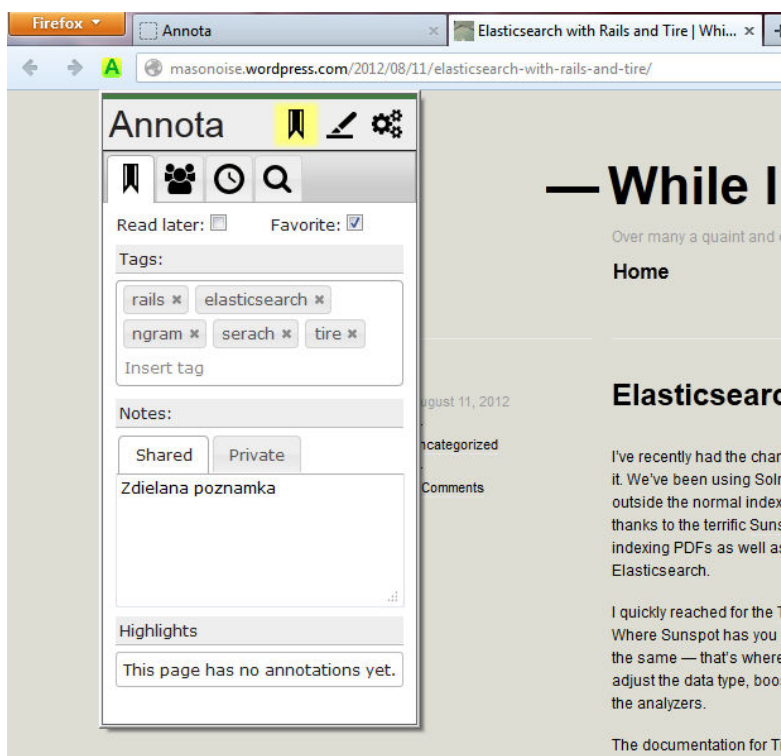
- Umiestnenie na základe okolia. Poznámka je pripojená k nejakej časti textu v dokumente. Tento text je špecifikovaný presným označením textom a jeho okolím.
- Umiestnenie na nejakú pozíciu v rade znakov. Text sa považuje za sériu znakov a poznámka je označená začiatočnou a koncovou pozíciou v tomto rade.
- Umiestnenie pomocou plochy špecifikuje plochu v zobrazenom dokumente, ku ktorej je pripojená poznámka. V tomto prípade sa neodkazuje na obsah dokumentu ale na geometrické súradnice plochy, ku ktorej sa má poznámka pripojiť.

Na umiestňovanie poznámok do dokumentov je možné používať niekoľko metód, ktoré sú vhodné pre rozdielne typy dokumentov a sú do rôznej miery odolné voči zmenám v dokumentoch. Pri pridávaní poznámok do dokumentov ako sú napríklad obrázky alebo PDF dokumenty, kde sa štruktúra dokumentu nemení, je možné pridávať poznámky na konkrétnu súradnicu v dokumente. Pri pridávaní poznámok do dokumentov ako sú napríklad webové stránky, kde sa mení obsah, ale aj zobrazenie podľa toho na akom zariadení stránku prezeráme, však takýto prístup nemôžeme použiť. Preto musíme umiestňovať poznámky vzhľadom na obsah dokumentu. Tu je možné použiť metódy založené na určení začiatočnej a koncovkej pozície textu, ku ktorému je poznámka priradená, metódy zohľadňujúce okolie textu a podobne. Pri voľbe metódy na prepájanie častí dokumentu a poznámok je vždy potrebné brať do úvahy možnosť, že dokument sa môže zmeniť.

## 5.2 Realizácia pomocou rozšírenia webového prehliadača

Pre potreby získania poznámok na experimentovanie a na skúmanie správania používateľov pri poznámkovaní sme implementovali rozšírenie prehliadača Firefox s názvom Annota, ktoré umožňuje pridávať poznámky do webových stránok, ale aj PDF dokumentov zobrazených v prehliadači. Neskôr sa projekt rozšíril a rozšíril sa aj o riešiteľský kolektív, a tak v súčasnosti Annota (Ševcech et al. 2012) zahŕňa okrem poznámkovania aj zdieľanie vytvorených záložiek a poznámok v skupinách, import údajov zo služby Mendley a podobne.

Pomocou rozšírenia prehliadača je možné do bežných webových stránok, ale aj do PDF dokumentov zobrazených v prehliadači pridávať rôzne typy poznámok. Podporované je vytváranie záložiek, zvýrazňovanie textu v dokumente, pripájanie komentárov k zvýraznenému textu, pripájanie poznámok k dokumentu ako celku, označovanie dokumentu na neskoršie prečítanie ako aj označovanie obľúbených dokumentov hviezdíčkou. Tieto funkcie sú sprístupnené pomocou vyskakovacieho



Obr. 5.1: Vyskakovacie okno, ktoré sprístupňuje rôzne funkcie rozšírenia prehliadača

okna, ktoré sa zobrazí po kliknutí na ikonu aplikácie v hornej lište prehliadača. Príklad vyskakovacieho okna je zobrazený na obrázku číslo 5.1

Základný prípad použitia, pre ktorý je Annota určená je spolupráca študenta a jeho mentora pri štúdiu vedeckých zdrojov v prostredí digitálnej knižnice. Pre potreby tohto scenáru sme implementovali podporu pre zdieľanie záložiek a poznámok. Študent má možnosť vytvárať záložky a pridávať poznámky do dokumentov v digitálnej knižnici. Vytvorené záložky a poznámky môže zdieľať so svojím mentorom prostredníctvom zdieľania v skupinách. Mentor takto môže sledovať postup študenta pri hľadaní dokumentov, môže mu odporúčať ďalšie dokumenty z ohľadom na tie, ktoré už prečítal, prípadne mu môže pridávať komentáre a ďalšie poznámky k záložkám. Zvýraznenia v texte, ktoré pridá k záložke, ktorú zdieľa v skupine sa zobrazia všetkým členom skupiny.

Rozšírenie prehliadača umožňuje vytvárať poznámky, ktorá sa viažu k celému dokumentu (tagy, poznámka vo forme voľného text), ale umožňuje pridávať aj poznámky ku konkrétnemu textu v tele dokumentu (zvýraznenie textu). Keďže pridávame poznámky do webových stránok, ktoré sa môžu často meniť, na umiestnenie poznámky do dokumentu sme museli použiť metódu, ktorá bude odolná voči týmto zmenám. Na pripájanie poznámok k dokumentom používame duplicitný spôsob, aby bolo možné v prípade potreby vymeniť alebo pozmeniť algoritmus na opätovné pripájanie poznámok do dokumentu. Na umiestnenie poznámky v texte sa uchováva text, ku ktorému je poznámka pripojená a poradie

výskytu tohto textu v dokumente. Ak je napríklad poznámka pripojená k tretiemu výskytu slova „auto“ v texte, tak sa uchová slovo „auto“ a číslo 3 ako poradie výskytu. Okrem toho sa uchováva okolie textu, ku ktorému je poznámka pripojená.

Umiestňovanie poznámky na základe textu, ku ktorému je poznámka pripojená a jeho poradie v dokumente je odolné voči zmenám v dokumente pokiaľ sa pred miesto kde má byť poznámka pripojená nepridá alebo nezmaže tento výskyt textu. V prípade, ak by sme použili približné porovnávanie reťazcov, tak táto metóda môže byť do istej miery odolná aj voči zmenám v samotnom označenom texte.

Umiestňovanie poznámok na základe okolia je odolné voči zmenám v dokumente a pri používaní približného porovnávania reťazcov je do istej miery odolné aj voči zmenám v samotnom označenom texte.

Pri vkladaní poznámok do webových stránok sme riešili problém stránok, ktoré menia svoj obsah počas práce so stránkou, napríklad pomocou technológie AJAX. V tomto prípade časť stránky, ku ktorej je pripojená poznámka, nemusí byť zobrazená bezprostredne po načítaní dokumentu, ale zobrazí sa až počas práce so stránkou. Pri vkladaní poznámok do dokumentu počas načítavania stránky teda môže proces vkladania poznámky zlyhať, pretože sa nenájde obsah, ku ktorému je ju potrebné pridať. Tento obsah sa môže objaviť až neskôr. Z tohto dôvodu sme implementovali funkciu na sledovanie zmien v obsahu stránky a v prípade zmeny spustíme funkciu na opätovné vykreslenie poznámok, ktoré sa nepodarilo vykresliť pri načítaní stránky. Sekvenčný diagram pre algoritmus vkladania poznámok do dokumentu je zobrazený na obrázku B.6 v technickej dokumentácii k nástroju Annota v prílohe B.



# Metóda na tvorbu dopytu pomocou poznámok

Pri čítaní dokumentov potrebujeme častokrát vyhľadať dokumenty, ktoré s týmto dokumentom súvisia. Obyčajne vtedy siahneme po vyhľadávači, manuálne vytvoríme dopyt a vyhľadáme ďalšie dokumenty, ktoré by nás mohli zaujímať. V tejto časti opisujeme metódu na vyhľadávanie súvisiacich dokumentov k práve študovanému dokumentu. Navrhnutá metóda vytvára dopyt na základe obsahu dokumentu a poznámok, ktoré k dokumentu pripája používateľ. Poznámky používame ako indikátory záujmu o konkrétne časti dokumentu.

V súčasnosti najpoužívanejšie vyhľadávače dokumentov alebo webových stránok umožňujú zadávať dopyty v podobe zoznamu kľúčových slov. Navrhli sme metódu na vytvorenie dopytu z textového dokumentu, kde vytvorený dopyt má formu zoznamu slov. Vytvorený dopyt je možné použiť v ľubovoľnom vyhľadávači, ktorý prijíma dopyty v podobe zoznamu slov. Pomocou navrhnutej metódy vytvárame dopyt, ktorého cieľom je získanie súvisiacich dokumentov. Pri výbere slov do dopytu sa používajú poznámky, ktoré k dokumentu pripojil jeho čitateľ ako indikátory jeho záujmu o konkrétne časti dokumentu a vyhľadávajú sa dokumenty súvisiace práve s tými časťami dokumentu, ktoré používateľa najviac zaujali.

Základný scenár pre použitie navrhnutej metódy je takýto:

1. Používateľ študuje dokument.
2. Používateľ k dokumentu pridáva tagy, zvýraznenia častí textu, pridáva komentáre k zvýraznenému textu a pridáva textové poznámky k dokumentu ako celku.
3. Na základe obsahu dokumentu a pripojených poznámok sa vytvára dopyt na získanie súvisiacich dokumentov a používateľovi sa popri štúdiu dokumentu zobrazujú vyhľadané súvisiace dokumenty.



4. Dopyt sa s pribúdajúcim počtom poznámok mení, a teda aj zobrazené výsledky vyhľadávania sa menia.

Vyhľadávanie súvisiacich dokumentov môže používať napríklad študent, ktorý sa snaží naštudovať si problematiku z nejakej oblasti. Našiel z tejto oblasti jeden dokument, ktorý ho zaujal. Píše si do neho poznámky, zvýrazňuje dôležité časti a podobne. Na základe dokumentu a k nemu pridaných poznámok sa mu vyhľadajú ďalšie súvisiace dokumenty, ktoré by si mohol prečítať.

Podobné použitie môže byť aj v prípade, ak študent študuje nejakú knihu alebo dlhší dokument, v ktorom našiel jednu tému, ktorá ho zaujala. K textu o tejto téme si pridáva poznámky. Počas čítania sa mu vyhľadávajú ďalšie dokumenty súvisiace práve s touto témou.

Bežne používaná metóda na tvorbu dopytu z obsahu dokumentu, ktorú používa napríklad nástroj ElasticSearch je založená na TF-IDF a zohľadňuje len frekvenciu výskytov slov v dokumente a v kolekcii dokumentov, v ktorej sa vyhľadáva. Veríme že nie len počet výskytov slov v dokumente, ale aj štruktúra dokumentu je veľmi dôležitá pri tvorbe dopytu. O to viac je štruktúra dokumentu dôležitá ak predpokladáme, že poznámky, ktoré do textu pridávajú používatelia a ktoré používame pri tvorbe dopytu, sú pripojené najmä k tým častiam dokumentu, ktoré používateľ a nejakým spôsobom zaujali. Navrhli sme metódu na tvorbu dopytov z obsahu dokumentu a pripojených poznámok, ktorá používa:

1. transformáciu textu na graf na zachovanie štruktúry textu a
2. šírenie aktivácie vo vytvorenom grafe na výber slov do dopytu.

Graf, ktorý vznikol z textu na základe susedností slov zachováva početnosť výskytov slova v dokumente pomocou stupňa vrchola, ako aj štruktúru textu pomocou štruktúry hrán, ktoré vznikli medzi uzlami. Na takto vzniknutý graf sa dajú aplikovať rôzne grafové algoritmy a pomocou nich sa z grafu dajú získavať informácie o jeho štruktúre (Paranyushkin 2011). Pomocou metriky betweenness centrality sa napríklad dajú ohodnotiť uzly podľa významnosti v texte a pomocou detekcie komúnit sa dajú extrahovať skupiny slov, ktoré spolu súvisia. My sme tento graf použili na nájdenie najvýznamnejších slov z pohľadu používateľa, ktorý do textu pridáva poznámky.

Pomocou šírenia aktivácie vo vzniknutom grafe vyhľadávame najvýznamnejšie slová pričom na vloženie počiatočnej aktivácie do grafu používame poznámky, ktoré do textu pridal používateľ. Zavedením aktivácie pomocou poznámok zohľadňujeme záujmy používateľa pri extrakcii slov do dopytu.

Navrhnutá metóda sa dá priamočiaro rozšíriť o používanie poznámok, ktoré pridal používateľ k ostatným dokumentom ako aj o poznámky vytvorené inými používateľmi.

## 6.1 Transformácia textu na graf

Prvým krokom pri transformácii textu na graf je predspracovanie textu v niekoľkých fázach: segmentácia, tokenizácia, odstránenie stop-slov a lematizácia alebo stemovanie. Po tomto prvom kroku z textu vznikol zoznam slov a môže začať proces transformácie tohto zoznamu na graf. Každé jedinečné slovo zo zoznamu prejde do grafu ako jeden uzol. Hrany v grafe vzniknú medzi dvoma uzlami v prípade, ak zodpovedajúce slová v texte sa nachádzajú vedľa seba alebo dostatočne malej vzdialenosti. Pseudokód pre tento algoritmus je nasledovný:

```
function createGraph(text, dist){
    words = text.removeStopwords.stem.split;
    nodes = words.uniq;
    edges = [];
    for (int i = 0; i < words.size; i++) {
        for (int j = i; i < min(i+dist, words.size-1); j++) {
            edges.add(words[i], words[j]);
        }
    }
    return Graph.new(nodes, edges);
}
```

Pre potreby experimentov sme skúšali niekoľko nastavení pre maximálnu vzdialenosť slov v texte, kde ešte môžeme vytvoriť hranu medzi zodpovedajúcimi uzlami grafu. Najlepšie výsledky sme dosiahli pre kombináciu dvoch prechodov cez slová v grafe opísanú v (Paranyushkin 2011). Algoritmus pre pridávanie hrán do grafu sme spustili dvakrát, raz s maximálnou vzdialenosťou slov nastavenou na 2 a raz nastavenou na 5. Pomocou týchto nastavení sme dosiahli, že boli previazané slová vo väčšej vzdialenosti, teda napríklad slová v susedných vetách. Zároveň medzi slovami, ktoré boli blízko pri sebe vzniklo viac hrán, a teda boli lepšie prepojené.

Pri vytváraní hrán sme nenechali váhu hrán, ale spoliehali sme sa na fakt, že medzi slovami, ktoré sa často objavujú blízko seba vznikne oveľa viac hrán ako medzi slovami, ktoré sa vedľa seba objavili len raz. Pre potreby skrátenia času behu algoritmu šírenia aktivácie v nasledujúcom kroku sme však po vytvorení grafu spojili násobné hrany do jednej a nastavili jej váhu rovnú počtu spojených hrán.

## 6.2 Výber slov do dopytu

V texte transformovanom do grafu používame algoritmus šírenia aktivácie na to, aby sme našli najvýznamnejšie uzly/slová. Tento algoritmus sa bežne používa napríklad na nájdenie uzlov v grafe, ktoré najviac súvisia s nejakým iným uzlom. Vtedy sa do tohto počiatočného uzlu zavedie aktivácia a nechá sa šíriť v grafe. Po tom ako sa už hodnoty aktivácie v jednotlivých uzloch menia len o menej ako stanovená hranica sa šírenie zastaví. Najviac súvisiace uzly majú po dokončení šírenia aktivácie najväčší podiel z pôvodnej aktivácie. Tento algoritmus sa však dá použiť aj na extrakciu kľúčových slov z textu (Palshikar 2007).

My používame algoritmus šírenia aktivácie na nájdenie najvýznamnejších slov v texte s ohľadom na poznámky, ktoré do textu pridal používateľ. Na vloženie počiatocnej aktivácie do grafu vytvoreného transformáciou textu na základe susednosti slov používame poznámky pripojené k dokumentu. Táto počiatocná aktivácia sa šíri v grafe a zhromažďuje sa v uzloch, ktoré sú najvýznamnejšie. Ak boli poznámky, a teda aj aktivácia, priradené k tým uzlom, ktoré používateľ a najviac zaujali, tak hľadanie najvýznamnejších uzlov bude zohľadňovať záujmy používateľ a.

Algoritmus na zavedenie aktivácie do grafu a na výber najdôležitejších slov je opísaný diagramom na obrázku číslo 6.1.

Pri zavádzaní aktivácie do grafu musíme rozlišovať dva typy poznámok:

- poznámky, ktoré zvyrazňujú časť textu dokumentu a
- poznámky, ktoré rozširujú dokument o ďalší obsah.

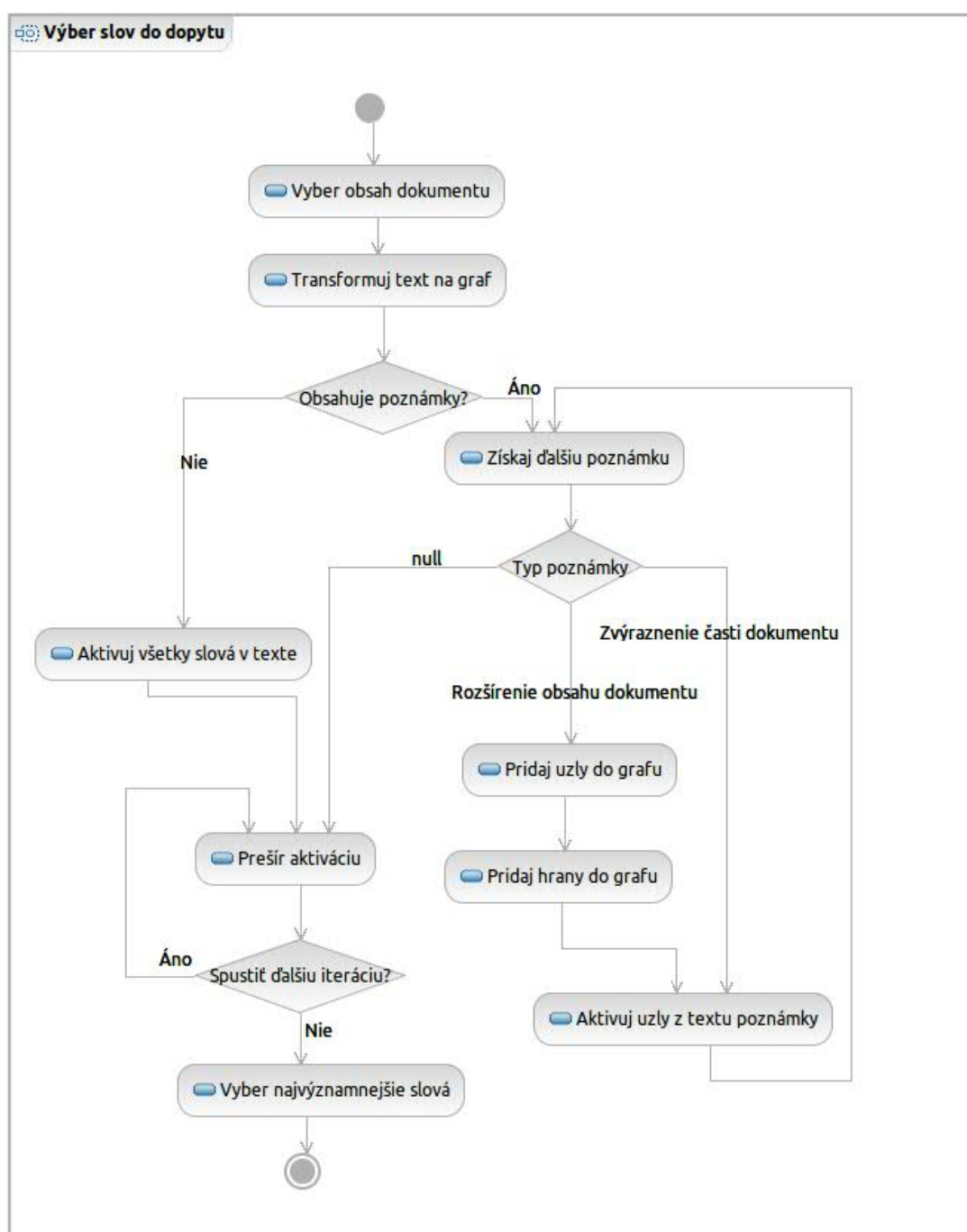
Navrhnutá metóda zohľadňuje oba typy poznámok. Poznámky, ktoré zvyrazňujú časť dokumentu zavádzajú počiatocnú aktiváciu do tých uzlov, ktoré reprezentujú slová zvyraznenej časti dokumentu.

Poznámky, ktoré rozširujú obsah dokumentu, rozširujú graf dokumentu o nové slová a hrany a zavádzajú aktiváciu do týchto nových uzlov. Pri zavádzaní aktivácie do rozširujúcej časti grafu predpokladáme, že používateľ pri písaní poznámok používa aspoň z časti spoločné slová ako sú slová v texte dokumentu. Prostredníctvom týchto spoločných slov sa môže aktivácia šíriť do zvyšku grafu. Tento predpoklad môže byť narušený napríklad v prípade, ak sú dokument a priradené poznámky v rozdielnych jazykoch. Pri experimentoch preto pracujeme s textami v angličtine a všetky priradené poznámky automaticky prekladáme do angličtiny.

Pri zavádzaní počiatocnej aktivácie do grafu používame poznámky, ktoré k dokumentu pridal používateľ. V prípade, ak k textu niesú priradené žiadne poznámky, tak sa aktivácia zavádza k všetkým slovám v texte a najvýznamnejšie slová získané šírením aktivácie budú slová významné pre text ako celok.

Po zavedení aktivácie do grafu sa počiatocná aktivácia šíri v grafe. Po ukončení šírenia aktivácie vyberieme uzly, ktoré majú najvyššiu aktiváciu a tieto používame ako slová do dopytu. Keďže používame poznámky ako indikátory záujmu používateľ a a počiatocná aktivácia sa šíri z uzlov, ku ktorým sú priradené poznámky, tak nájdené najvýznamnejšie slová odrážajú záujmy používateľ a.

Navrhnutá metóda umožňuje nájsť slová významné vzhľadom na opoznámkované časti dokumentu, ale dokáže nájsť aj globálne významné slová, ktoré sú významné pre celý dokument. Pomer slov významných pre opoznámkované časti dokumentu a globálne významných slov je možné kontrolovať počtom iterácií algoritmu. S rastúcim počtom iterácií sa počiatocná aktivácia šíri ku globálne významným slovám.



Obr. 6.1: Diagram činností, ktorý opisuje algoritmus na výber slov do dopytu s použitím obsahu dokumentu a pripojených poznámok

Pri použití tejto metódy je dôležité nájsť správny počet iterácií, po ktorých treba algoritmus zastaviť ako aj správne váhy pre rôzne typy poznámok.

### 6.3 Vlastnosti metódy na tvorbu dopytov

Navrhnutá metóda na tvorbu dopytu z obsahu textového dokumentu a pripojených poznámok vytvára dopyt v podobe zoznamu významných slov. Pri tvorbe dopytu sa používajú poznámky, ktoré zvyrazňujú časť dokumentu ako aj poznámky, ktoré pridávajú do dokumentu dodatočný obsah. Poznámky sa používajú ako indikátory záujmu používateľa o konkrétne časti dokumentu.

Pri práci s poznámkami predpokladáme, že používateľ pridáva poznámky k častiam dokumentu, ktoré ho nejakým spôsobom zaujali. Týmito poznámkami zvyrazňuje dôležité časti dokumentu, alebo označuje miesta, ktoré ho nejakým spôsobom zaujali. Pomocou obsahu poznámok používateľ vyjadruje svoje myšlienky, sumarizuje dokument alebo píše pripomienky k označenej časti textu. V prípade, ak k dokumentu niesú pripojené žiadne poznámky, navrhnutá metóda vytvorí dopyt len z obsahu dokumentu.

Pri používaní obsahu poznámok ako sú napríklad komentáre alebo voľný text pripojený k dokumentu predpokladáme, že pri písaní poznámok autor používa z časti spoločné slová ako sú slová v dokumente. V prípade, ak by boli v obsahu poznámok použité všetky slová rozdielne od slov v dokumente, pri pridávaní obsahu poznámok do grafu vytvoreného z textu by mohli vzniknúť dva podgrafy, ktoré by nespájali žiadne hrany a nemohla by sa medzi nimi šíriť aktivácia. V takom prípade by mohla nastat' situácia, že po šírení aktivácie v takomto grafe budú najvyššie ohodnotené len slová z používateľových poznámok a dopyt sa vytvorí len z nich. Toto by mohol byť problém napríklad v prípade ak by používateľ písal poznámky v inom jazyku ako je jazyk dokumentu. V tomto prípade by mohol vzniknúť dopyt, ktorý by obsahoval slová v inom jazyku ako je jazyk dokumentu. Z toho dôvodu pred vytváraním dopytu prekladáme automaticky všetok obsah poznámok do angličtiny.

Jednotlivé typy poznámok používané pri tvorbe dopytu je možné používať s rôznymi váhami. Metódu je možné obohatiť o ďalšie typy poznámok, ktoré zvyrazňujú časť dokumentu alebo pridávajú do dokumentu ďalší obsah. Metódu je možné priamočiaro rozšíriť o použitie poznámok, ktoré pridal používateľ k iným dokumentom. Tieto poznámky môžu zavádzať aktiváciu do grafu podobne ako poznámky z dokumentu, potrebné je len určiť váhu s akou sa majú zohľadňovať. Podobne je možné navrhnutú metódu rozšíriť aj o poznámky ostatných používateľov.

Šírením počiatkovej aktivácie v grafe dokumentu sa aktivácia hromadí v slovách významných z pohľadu používateľa. Postupným opakovaním šírenia aktivácie sa aktivita presúva zo slov významných pre používateľa do slov významných pre celý dokument. Pomocou nastavenia počtu iterácií algoritmu je možné kontrolovať

pomer slov v dopyte, ktoré sú významné pre používateľa a ktoré sú významné pre dokument ako celok.

Navrhnutá metóda vytvára dopyt zo študovaného dokumentu a poznámok, ktoré k nemu pridáva používateľ. Pri vytváraní dopytu sa používa len obsah študovaného dokumentu a proces získavania slov do dopytu nieje závislý od obsahu ostatných dokumentov v indexe vyhľadávača. Spracovanie dokumentu a vytvorenie dopytu je teda možné vykonať na strane klienta bez nutnosti spolupráce s vyhľadávačom. Výsledkom spracovania dokumentu je dopyt, ktorý sa dá použiť nezávisle od nástroja na vyhľadávanie a nezávisle od ďalších dokumentov.



# Vyhodnotenie

Navrhli sme metódu na tvorbu dopytu na vyhľadanie súvisiacich dokumentov, ktorá využíva poznámky priradené k dokumentu ako indikátory záujmu používateľa o konkrétnu časť dokumentu. Metódu sme overili prostredníctvom implementácie prototypu v podobe služby s názvom Annota, ktorá slúži na vytváranie záložiek a poznámok pri práci s digitálnou knižnicou. Súčasťou Annoty je rozšírenie prehliadača Firefox, opísané v kapitole 5, ktoré umožňuje pri čítaní webových stránok zobrazených v prehliadači vytvárať rôzne typy poznámok.

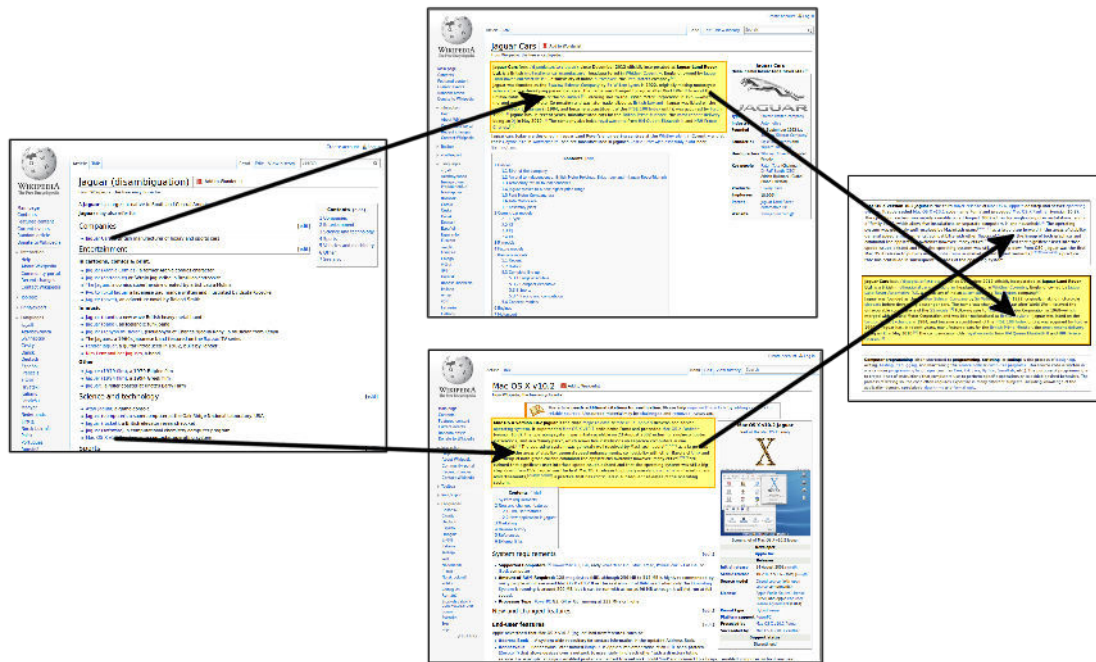
## 7.1 Stanovenie parametrov navrhnutej metódy pomocou simulácie

Pre používanie navrhnutej metódy na vyhľadávanie súvisiacich článkov je potrebné nastaviť váhy s akými sa majú zohľadňovať rôzne typy poznámok a počet iterácií algoritmu, po ktorých sa má zastaviť šírenie aktivácie. Na určenie týchto parametrov sme navrhli simuláciu, kde optimalizujeme parametre pre maximálnu presnosť vyhľadávania. Pri tejto simulácii generujeme poznámky do dokumentov z dátovej sady, vyhľadávame súvisiace dokumenty a meriame presnosť vyhľadávania pre rôzne kombinácie parametrov. Pre optimalizáciu parametrov používame hill climbing algoritmus.

### 7.1.1 Dátová sada

Ako dátovú sadu pri vykonávaní tejto simulácie používame sadu dokumentov získaných z Wikipédie. Pomocou tejto simulácie simulujeme pridávanie poznámok do dokumentov tak, ako keď používateľ pridáva poznámky len k tej časti dokumentu, ktorá ho najviac zaujala. Takýto dokument sme vytvorili





Obr. 7.1: Vytváranie zdrojového dokumentu pri simulácii na nájdenie parametrov metódy na vytváranie dopytu

pomocou takzvaných rozdeľovacích stránok (disambiguation page) na Wikipédii. Tieto stránky obsahujú odkazy na stránky pre rôzne významy viacvýznamových slov. Napríklad stránka Jaguar\_(disambiguation)<sup>22</sup> odkazuje na stránky pre rôzne významy slova jaguár. Z Wikipédie sme stiahli všetky rozdeľovacie stránky a vybrali z nich podmnožinu, s ktorou sme ďalej pracovali. K vybranej podmnožine rozdeľovacích stránok sme stiahli všetky stránky, na ktoré smerovali odkazy z rozdeľovacích stránok. Z týchto stránok sme vybrali abstrakty a spojili abstrakty pre každú rozdeľovaciu stránku v náhodnom poradí do jedného dokumentu. Tento dokument predstavoval zdrojový dokument, do ktorého sme pridávali poznámky a ku ktorému sme hľadali súvisiace dokumenty. Pri generovaní poznámok sme si vybrali jeden z abstraktov v zdrojovom dokumente a k nemu pridali poznámky. Obrázok 7.1 znázorňuje získanie stránok, na ktoré odkazuje rozdeľovacia stránka, výber abstraktov z týchto stránok, spojenie abstraktov do jedného dokumentu v náhodnom poradí a následné generovanie poznámok do jedného z abstraktov.

Na základe obsahu vytvoreného dokumentu a poznámok pripojených k jednému z abstraktov, z ktorých sa dokument skladá, sme vytvorili dopyt. Tento dopyt sme použili na vyhľadanie v celej sade stiahnutých dokumentov a zaznamenali sme si nájdené dokumenty a ich relevanciu k abstraktu, ku ktorému sme pridali poznámky. Za relevantné dokumenty sme považovali tie, ktoré patrili do rovnakej kategórie ako stránka, do ktorej abstraktu sme pridávali poznámky. Na zvýšenie počtu dokumentov v sade, v ktorej sme vyhľadávali sme stiahli všetky dokumenty

<sup>22</sup>Jaguar\_(disambiguation), [http://en.wikipedia.org/wiki/Jaguar\\_\(disambiguation\)](http://en.wikipedia.org/wiki/Jaguar_(disambiguation))

pre každú z kategórií, v ktorej sa nachádzal niektorý z článkov, na ktoré odkazovali rozdeľovacie stránky. Množstvo spracovaných rozdeľovacích stránok použitých pre vytvorenie simulácie a celkový počet spracovaných dokumentov je zosumarizované v tabuľke číslo 7.1.

| Atribút   | Počet  |
|---|--------|
| Počet stiahnutých rozdeľovacích stránok                         | 226363 |
| Počet vybraných rozdeľovacích stránok do testu                  | 86     |
| Počet stránok, na ktoré odkazovali vybrané rozdeľovacie stránky | 629    |
| Počet kategórií   | 2654   |
| Počet stiahnutých stránok                                       | 232642 |

Tabuľka 7.1: Veľkosť vytvorenej dátovej sady

### 7.1.2 Generovanie poznámok

Pri simulovaní pridávania poznámok do vytvorených dokumentov sme generovali poznámky na základe poznámok, ktoré vytvorili používatelia nástroja Annota medzi 8.10.2012 a 1.2.2013. Za tento čas 82 používateľov služby Annota spolu vytvorilo 1390 záložiek, 388 zvýraznení v texte, 81 komentárov pripojených k zvýraznenému textu a 198 záložiek s priradenou poznámkou vo forme voľného textu. Na vytvorených poznámkach sme sledovali niekoľko parametrov, ktoré sme potom použili pri generovaní poznámok. Sledovali sme:

- počet zvýraznení v texte pre používateľa a dokument,
- dĺžku zvýraznených úsekov textu,
- dĺžku poznámky priradenej k záložke,
- podiel zvýraznení textu, ku ktorým bol pripojený komentár a
- dĺžku komentáru pripojeného k zvýraznenému textu.

Na obrázku číslo 7.2 je zobrazené rozdelenie počtu zvýraznení v texte pre dokument a používateľa. Zvyšné sledované parametre sú opísané v prílohe E.

Pre jednotlivé parametre sme našli rozdelenia, podľa ktorých sa správajú. Podľa týchto rozdelení sme potom generovali zvýraznenia v texte, komentáre a poznámky vo forme voľného textu pripojené k dokumentu ako celku. Pri generovaní poznámok, ktoré do textu pridávajú ďalší obsah (komentáre a poznámky vo forme voľného textu) sme používali text zo stránky, ku ktorej abstraktu sme poznámky generovali.



Obr. 7.2: Rozdelenie počtu zvýraznení v texte pre používateľa a dokument

### 7.1.3 Simulácia

Pre optimalizáciu parametrov používame hill climbing algoritmus, ktorý mení parametre navrhutej metódy tak, aby sa maximalizoval prírastok k presnosti vyhľadávania. Problém uviaznutia v lokálnom maxime riešime pomocou nastavenia veľkého počiatočného kroku, s ktorým algoritmus upravuje parametre a postupným znižovaním tohto kroku.

Keďže sme pri simulácii pridávania poznámok do textu používali poznámky generované náhodne na základe nejakých rozdelení, každý krok simulácie sme zopakovali desať krát, aby sme znížili pravdepodobnosť, že dosiahnuté výsledky boli náhodné. Jeden krok simulácie opisuje nasledovný pseudokód:

```
def simulation_step(weights, iteration)
  for disambiguation in disambiguations do
    abstracts = disambiguation.pages.abstracts
    for abstract in abstracts do
      text = abstracts.shuffle.join(" ")
      graph = Graph.new(text)
      annotation = Annotation.generate(abstract)
      graph.activate(annotation, weights)
      iteration.times do
        graph.spread_activation
      end
      query = graph.top_nodes
      results = ElasticSearch(query)
```

```

    category = abstract.page.categories
    relevant = results.with_category(category)
  end
end
end

```

Počas jedného kroku simulácie algoritmus iteruje cez všetky rozdeľovacie stránky, vyberie abstrakty zo stránok, na ktoré táto rozdeľovacia stránka smeruje. Pre každý z vybraných abstraktov vytvorí jeden dokument, ktorý je spojením všetkých abstraktov v náhodnom poradí. Algoritmus vygeneruje poznámky do tohto jedného abstraktu, vytvorí graf a spustí zvolený počet iterácií šírenia aktivácie. Po skončení šírenia aktivácie vyberie z grafu najvýznamnejšie uzly, ktoré sa použijú na vyhľadanie súvisiacich článkov. Získame tak 10 článkov, pre ktoré overíme ich relevanciu. Relevantné sú tie dokumenty, ktoré patria do tej istej kategórie ako stránka abstraktu, ku ktorému sme generovali poznámky. Pre každé vykonané vyhľadanie súvisiacich článkov zaznamenáme presnosť tohto vyhľadávania. Ako výslednú presnosť metódy s nejakou kombináciou parametrov berieme priemer presností pre všetky vyhľadávania, ktoré sme vykonali s touto kombináciou parametrov.

Po spustení simulácie a optimalizácii parametrov pomocou hill climbing algoritmu sme najlepšie výsledky získali pri použití kombinácie parametrov zobrazenej v tabuľke číslo 7.2. Tieto parametre sme používali vo všetkých ďalších experimentoch.

| Parameter               | Váha |
|-------------------------|------|
| Počet iterácií          | 3    |
| Váha zvýraznení v texte | 7    |
| Váha komentárov         | 1    |
| Váha voľnej poznámky    | 13   |

Tabuľka 7.2: Typy poznámok používané rôznymi službami

## 7.2 Porovnanie voči metóde založenej na TF-IDF

Metódu na vytváranie dopytu z dokumentu obohateného o poznámky overujeme voči existujúcej metóde na vyhľadanie na základe dokumentu. Touto porovnanou metódou je MoreLikeThis dopyt v nástroji ElasticSearch. MoreLikeThis dopyt vnútorne používa transformáciu dokumentu na zoznam kľúčových slov pomocou TF-IDF metriky. Predpokladáme, že navrhovaná metóda, ktorá používa transformáciu dokumentu na kľúčové slová na základe analýzy grafu získaného z dokumentu, dosiahne lepšie výsledky ako metóda

založená na TF-IDF pri používaní poznámok vytváraných používateľmi v procese tvorby dopytu.

Budeme overovať niekoľko scenárov pre navrhnutú metódu ako aj pre metódu založenú na TF-IDF:

- Vyhľadávanie na základe dopytu vytvorenom bez použitia poznámok.
- Vyhľadávanie na základe dopytu vytvorenom z obsahu dokumentu a náhodne generovaných poznámok.
- Vyhľadávanie na základe dopytu vytvorenom s úplnou informáciou o záujmoch používateľa.

Na vykonávanie týchto experimentov sme použili podobnú simuláciu ako sme opísali v predchádzajúcej časti. Túto simuláciu sme spustili pre navrhnutú metódu ako aj pre MoreLikeThis dopyty. Pri vytváraní dopytu bez použitia poznámok sme použili na tvorbu dopytu celý text zdrojového dokumentu. Pri vytváraní dopytu s úplnou informáciou o záujmoch používateľa sme na tvorbu dopytu použili len text toho odseku, ktorý mal simulovať časť dokumentu, ktorá zaujala používateľa. Pre potreby porovnávania navrhnutej metódy s MoreLikeThis dopytom s použitím generovaných poznámok, sme rozširovali text pre MoreLikeThis dopyt pomocou zvýraznených častí textu a obsahu poznámok. Rôzne váhy pre rôzne typy poznámok sme zabezpečili opakovaným pridávaním zvýrazneného textu alebo obsahu poznámky. Počet opakovaní sme určili, podobne ako v predchádzajúcej časti, pomocou simulácie a optimalizácie parametrov pre najvyššiu presnosť vyhľadávania pomocou hill climbing algoritmu.

Priemery presností vyhľadávania získané pre porovnávané metódy a pre všetky tri scenáre sú zhrnuté v tabuľke číslo 7.3

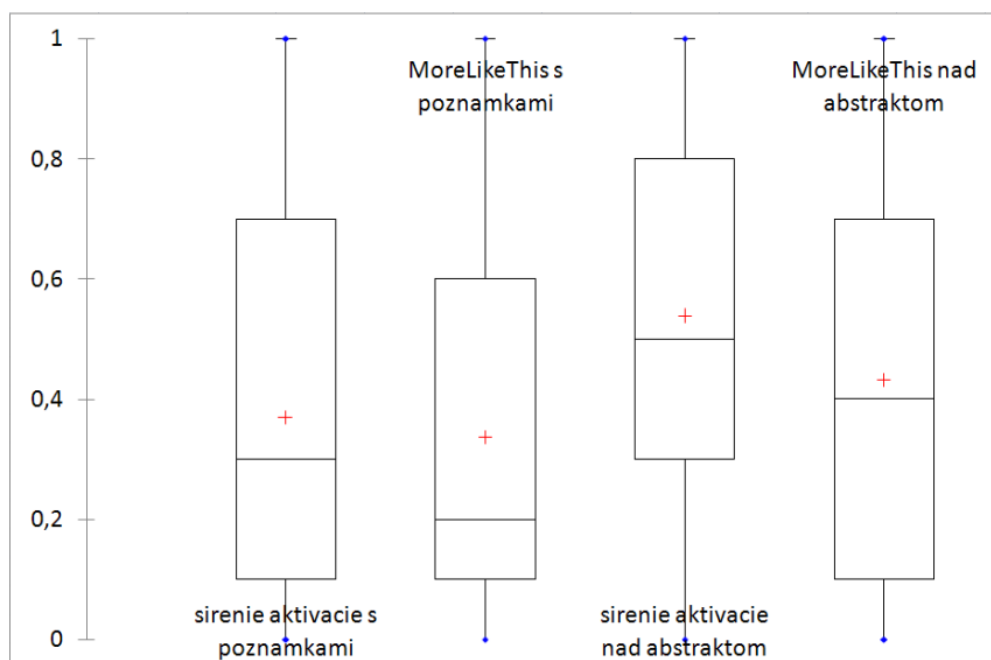
| Metóda                                     | Presnosť |
|--|----------|
| MoreLikeThis bez poznámok                  | 21,32%   |
| Navrhnutá metóda bez poznámok              | 21,96%   |
| MoreLikeThis s generovanými poznámkami     | 33,64%   |
| Navrhnutá metóda s generovanými poznámkami | 37,07%   |
| MoreLikeThis na základe abstraktu          | 43,20%   |
| Navrhnutá metóda na základe abstraktu      | 53,34%   |

Tabuľka 7.3: Typy poznámok používané rôznymi službami

Vykonalí sme studentove t-testy pre dvojice navrhnutej metódy a MoreLikeThis metódy pre jednotlivé scenáre. Nulová hypotéza bola: dve porovnávané metódy dosiahnu rovnakú priemernú presnosť. Testy sme vykonali na hladine

významnosti 0,05. Pre scenár, v ktorom sme nepoužívali poznámky pri tvorbe dopytu na základe vykonaného testu nemôžeme zamietnuť nulovú hypotézu. To znamená, že porovnávané metódy nedosiahli štatisticky významný rozdiel v presnosti vyhľadávania v scenári, kde sme na tvorbu dopytu použili len obsah dokumentu, a teda navrhnutá metóda dosahuje porovnateľné výsledky ako metóda založená na TF-IDF aj napriek tomu, že navrhnutá metóda na rozdiel od porovnáwanej metódy využíva len informácie zo zdrojového dokumentu a nevyužíva informácie o ostatných dokumentoch v kolekcii.

Pre zvyšné dva scenáre (s generovanými poznámkami a s použitím abstraktu pri tvorbe dopytu) na základe vykonaných testov môžeme zamietnuť nulovú hypotézu. To znamená, že rozdiel v priemernej presnosti porovnávaných metód je štatisticky významný. Z tabuľky číslo 7.3 vidíme, že v oboch týchto scenároch dosiahla navrhovaná metóda významne vyššiu presnosť v porovnaní s MoreLikeThis metódou. Na obrázku číslo 7.3 vidno rozdelenie nameraných hodnôt pre porovnávané metódy pre scenáre s použitím poznámok a s použitím abstraktu pri tvorbe dopytu.



Obr. 7.3: Rozdelenie presností získaných pomocou simulácie pre porovnávané metódy

## 7.3 Vyhodnotenie vyhľadávania súvisiacich dokumentov

Celkové vyhodnotenie implementovaného riešenia sme vykonali prostredníctvom riadenej používateľskej štúdie ako aj pomocou dlhodobého experimentu

s použitím nástroja Annota. Cieľom týchto experimentov bolo overiť úspešnosť navrhutej metódy na vyhľadávanie súvisiacich dokumentov počas čítania a poznámkovania dokumentu. Porovnávali sme úspešnosť vyhľadávania dokumentov pomocou vytvoreného dopytu s použitím poznámok a bez použitia poznámok.

### 7.3.1 Používateľská štúdia

Cieľom používateľskej štúdie bolo overiť či poznámky použité pri tvorbe dopytu na získanie súvisiacich článkov zlepšujú presnosť vyhľadávania. Popri tom sme chceli zistiť, či a ako používatelia pridávajú poznámky pri čítaní elektronických dokumentov. Pri týchto experimentoch sme mali osem dobrovoľníkov, ktorých sme sa najskôr spýtali na to, či a ako vytvárajú poznámky do dokumentov a následne mali za úlohu pridať poznámky do odborných článkov podľa vlastného výberu a ohodnotiť relevanciu dokumentov, ktoré sme im vyhľadali pomocou navrhutej metódy na tvorbu dopytu. Všetci oslovení dobrovoľníci si v nástroji Annota vopred vytvorili sadu záložiek alebo si do Annoty importovali záložky zo služby Mendeley.

#### Zvyky používateľov pri vytváraní poznámok

Predtým ako sme používateľov nechali vytvárať poznámky do dokumentov sme sa ich spýtali niekoľko otázok o tom, či a ako pridávajú otázky do dokumentov. Štúdie sa zúčastnilo 8 dobrovoľníkov, ktorých sme sa spýtali postupne tieto otázky:

1. Keď čítate tlačené dokumenty ako časopisy, knihy alebo vytlačené odborné články, píšete si do nich poznámky?
2. Pridávate poznámky do elektronických dokumentov? Ak áno, aký nástroj na to používate?
3. Vytvárate pri čítaní elektronických dokumentov poznámky ako sú napríklad záložky, tagy, zvýraznenia v texte, komentáre, voľní text k celému dokumentu, iné?
4. Keď zvýrazňujete text v dokumente, zvýrazňujete viac v niektorej časti dokumentu, napríklad v úvode, v závere, v hlavnej časti článku alebo väčšinou rozmiestňujete zvýraznenia rovnomerne v celom článku?
5. Čo vyjadrujete poznámkami: vlastné myšlienky, sumarizáciu dokumentu, vyberáte najzaujímavejšie časti, pridávate pripomienky k dokumentu, označenie na neskôr „todo“, iné?
6. Keď pridávate tagy k dokumentom, používate ich na opis dokumentu, na zaradenie dokumentu do nejakých kategórií alebo nejak inak?

Na základe týchto otázok sme získali niekoľko informácií o zvykoch používateľov pri písaní poznámok. Približne polovica z opýtaných dobrovoľníkov pridáva poznámky do tlačенých dokumentov, ale všetci okrem jedného pridávajú nejakú formu poznámok do elektronických dokumentov.

Na pridávanie poznámok do elektronických dokumentov používajú najmä nástroje Mendeley a Annota. Počet používateľov týchto dvoch nástrojov je však ovplyvnený tým, že sme oslovili práve takých používateľov, ktorí majú vytvorené záložky v nástroji Annota, alebo ktorí si nainportovali záložky z nástroja Mendeley do Annoty. Okrem týchto dvoch nástrojov však oslovení dobrovoľníci používajú aj ďalšie nástroje, ktoré používajú nie všeobecne na pridávanie poznámok do ľubovoľných dokumentov, ale majú pre ne špecifickejšie určenie. Napríklad služby Diigo alebo Delicious používajú dvaja dobrovoľníci na vytváranie záložiek a ich organizáciu pomocou tagov, službu Pocket používajú dvaja dobrovoľníci na odkladanie dokumentov na neskoršie prečítanie, službu Evernote používajú dvaja dobrovoľníci na písanie vlastných myšlienok a poznámok o nejakom dokumente, ale aj bez toho aby boli poznámky naviazané na nejaký študovaný dokument. Okrem týchto nástrojov viacerí dobrovoľníci používajú rôzne nástroje na vytváranie zoznamov úloh, takzvaných ToDo listov.

Opýtaní dobrovoľníci používajú všetky spomínané typy poznámok, najčastejšie však vytvárajú záložky a pridávajú k nim zvýraznenia v texte. Najmenej často pridávajú poznámky vo forme voľného textu pripojeného k dokumentu ako celku.

Všetci opýtaní dobrovoľníci pridávajú zvýraznenia v texte rovnomerne po celom dokumente, len jeden z nich pridáva zvýraznenia väčšinou rovnomerne, ale trochu viac v úvode dokumentu.

Dobrovoľníci pomocou pridaných poznámok vytvárajú sumarizáciu dokumentu a vyjadrujú pomocou nich vlastné myšlienky. Jeden z dobrovoľníkov používa zvýraznenia v texte a komentáre vložené priamo do textu na označenie zaujímavých častí dokumentu tak, aby sa k tomuto miestu vedel rýchlo vrátiť pri opätovnom čítaní a aby vedel rýchlo zistiť, čo ho napadlo pri tom ako študoval túto časť dokumentu. Opýtaní dobrovoľníci teda potvrdili náš predpoklad, že používatelia pomocou poznámok označujú práve tie časti dokumentu, ktoré ich najviac zaujímajú. Popri tom pridávajú k dokumentom sumarizácie a používajú poznámky na uchovanie vlastných nápadov a myšlienok.

Podľa toho ako používatelia pridávajú tagy do dokumentov sa dajú rozdeliť opýtaní dobrovoľníci do dvoch skupín: na tých, ktorí pomocou tagov dokumenty opisujú a tých, ktorí pomocou tagov zaraďujú dokumenty do kategórií. Pri pýtaní sa na spôsob pridávania tagov do dokumentov dobrovoľníkom častokrát ani nenapadlo, že tagy sa dajú pridávať aj iným spôsobom ako to robia práve oni a divili sa ako môže niekto pridávať tagy inak ako oni. Tagy pridané týmito dvoma skupinami používateľov sa odlišujú účelom ako aj počtom (používatelia, ktorí používajú tagy na opísanie dokumentov pridávajú väčší počet tagov na dokument a viac rôznych tagov (Körner et al. 2010)), a teda môžu mať rôzny význam pri používaní tagov pri organizácii dokumentov a navigácií medzi nimi.



## Úspešnosť navrhnutej metódy

V druhej časti štúdie dobrovoľníci študovali niekoľko odborných článkov podľa vlastného výberu a pridávali do nich poznámky pomocou nástroja Annota, tak ako bežne pridávajú poznámky do dokumentov. Po tom ako dočítali dokument sme vytvorili dva dopyty pomocou metódy navrhnutej v kapitole 6. Pri jednom dopyte sme použili poznámky ako indikátory záujmu používateľa a pri druhom dopyte sme nepoužívali poznámky, ale len obsah dokumentu. Pomocou týchto dopytov sme vyhľadali dve sady dokumentov spomedzi všetkých dokumentov nazbieraných v nástroji Annota. Následne sme požiadali dobrovoľníkov, aby vybrali z vytvorených zoznamov tie dokumenty, ktoré súvisia s dokumentom, ktorý práve čítali a aby určili, ktorý z dvoch porovnávaných zoznamov dokumentov obsahuje dokumenty, ktoré viac súvisia s prečítaným dokumentom. Získané zoznamy boli pri každom experimente náhodne vymenené, takže dobrovoľníci, ktorý ohodnocovali relevantnosť nájdených dokumentov nevedeli, ktorý zoznam bol získaný ktorou metódou.

Oslovení dobrovoľníci spolu prečítali a opoznámkovali 11 rôznych dokumentov. V deviatich prípadoch označili zoznam získaný pomocou metódy, ktorá používala poznámky za lepší. V jednom prípade metóda používajúca poznámky vytvorila dopyt v slovenčine kvôli tomu, že všetky poznámky pripojené k dokumentu boli v slovenčine. Keďže všetky dokumenty, v ktorých sme vyhľadávali boli v angličtine, pomocou tohto dopytu sme nenašli žiadne dokumenty. V jednom prípade metóda, ktorá nepoužívala poznámky dosiahla lepší výsledok ako metóda používajúca poznámky. Pomocou metódy bez použitia poznámok sme spolu našli 15 relevantných dokumentov a metódou s použitím poznámok 34 relevantných dokumentov.

Časť z dobrovoľníkov písala poznámky po slovensky, čo spôsobovalo menšie problémy, keďže metóda na vytváranie dopytu je navrhnutá pre prácu s anglickým textom. Aby ale pracovali v čo najpodobnejších podmienkach ako pri bežnom pridávaní poznámok, nechali sme ich písať poznámky tak ako sú zvyknutí.

Pri jednom dokumente, kde dobrovoľník pridával všetky poznámky v slovenskom jazyku sme požiadali dobrovoľníka, aby po vyhodnotení relevancie získaných dokumentov preložil všetky poznámky do angličtiny. Znovu sme vytvorili dopyty a vyhľadali dokumenty a požiadali dobrovoľníka, aby opätovne ohodnotil relevantnosť týchto nájdených dokumentov. V oboch prípadoch metóda používajúca poznámky našla rovnaký počet dokumentov, ktoré patrili do tej istej témy ako opoznámkovaný dokument. Pri používaní poznámok preložených do angličtiny však aj zvyšné nájdené dokumenty hovorili o súvisiacich témach k študovanému dokumentu.

Pri inom dokumente sme požiadali dobrovoľníka, aby po ohodnotení relevantnosti nájdených dokumentov pridal do dokumentu ďalšie poznámky a potom skúsil znovu ohodnotiť nové nájdené dokumenty. Dobrovoľník v tomto experimente približne zdvojnásobil počet pridaných poznámok. V prvom prípade

metóda používajúca poznámky našla 3 dokumenty s témou súvisiacou s pôvodným dokumentom. Podľa tohto dobrovoľníka sa však v dátovej sade, kde sme dokumenty vyhľadávali nachádzali 4 dokumenty, ktoré hovoria o rovnakej téme ako zdrojový dokument. Z týchto dokumentov metóda s pôvodným množstvom poznámok nenašla žiadny. Po zvýšení počtu poznámok metóda znovu našla dokumenty so súvisiacou témou a aj jeden z dokumentov s rovnakou témou. S rastúcim množstvom poznámok priradených k dokumentu sa teda zvyšuje presnosť vytvoreného dopytu pri vyhľadávaní súvisiacich dokumentov.

Pri používaní poznámok na tvorbu dopytu navrhnutá metóda dosiahla lepšie výsledky ako v prípade, ak sa pri tvorbe dopytu nepoužívali poznámky. Pri použití poznámok sme našli viac dokumentov, ktoré hovorili o rovnakej téme ako zdrojový dokument a viac dokumentov, ktoré opisovali súvisiace témy.

### 7.3.2 Dlhodobý experiment v nástroji Annota

Na kvantitatívne porovnanie zlepšenia presnosti vyhľadávania súvisiacich článkov pomocou navrhnutej metódy bez použitia poznámok a s použitím poznámok pri tvorbe dopytu sme vykonali dlhodobý experiment v nástroji Annota. Do rozšírenia prehliadača, pomocou ktorého je možné pridávať poznámky do dokumentov, sme pridali funkciu na vyhľadanie súvisiacich článkov k práve študovanému dokumentu. Pri návšteve stránky niektorej z vopred zvolených digitálnych knižníc alebo pri zobrazení PDF dokumentu v prehliadači sa po chvíli zobrazilo používateľovi v pravom dolnom rohu obrazovky tlačidlo na vyhľadanie súvisiacich článkov (obrázok 7.4). Po stlačení tohto tlačidla sa na strane používateľa vybral z dokumentu obsah, transformoval sa na graf, spustila sa implementácia metódy na tvorbu dopytu, ktorá použila poznámky vytvorené používateľom pri čítaní dokumentu a vytvorila dopyt na získanie súvisiacich článkov. Tento dopyt sa odoslal na server, kde sa pomocou neho vyhľadali dokumenty spomedzi všetkých dokumentov, ktoré boli nazbierané v nástroji Annota. Informácie o nájdených dokumentoch sa odoslali späť do prehliadača, kde sa zobrazili v zozname usporiadanom podľa relevancie (obrázok 7.5). Keď používateľ klikol na niektorý z dokumentov v zozname, tak sa mu na novej karte v prehliadači zobrazil tento dokument spolu s výzvou na ohodnotenie ako veľmi zobrazený dokument súvisí s dokumentom, na základe ktorého sme ho vyhľadali. Používatelia mohli hodnotiť relevanciu dokumentu na 4-stupňovej škále od rovnakej témy až po úplne nesúvisiacu tému. V prípade, ak sa používateľ nevedel rozhodnúť o relevancii dokumentu, mohol odoslať prázdnu odpoveď.

Pomocou AB-testovania sme náhodne volili jeden z dvoch spôsobov, ktoré sme používali pri tvorbe dopytu. Oba spôsoby používali navrhnutú metódu, jeden s použitím poznámok pri tvorbe dopytu a jeden bez použitia poznámok.

V období od 9.4.2013 do 3.5.2013 sme získali 61 hodnotení, z toho 37 hodnotení pre metódu používajúcu poznámky a 24 pre metódu, ktorá nepoužívala poznámky pri tvorbe dopytu. Každý zo štyroch možných odpovedí (dokumenty



Obr. 7.4: Tlačidlo na zobrazenie súvisiacich článkov



Obr. 7.5: Zoznam nájdených súvisiacich článkov

sú na: rovnakú tému, súvisiace témy, mierne súvisiace témy a na nesúvisiace témy) sme pridelili postupne bodovú hodnotu od 1 pre rovnaké témy po 4 pre nesúvisiace témy. Priemerné ohodnotenie pre metódu používajúcu poznámky bolo 2,486 a pre metódu bez použitia poznámok 3,083. Podľa studentovho t-testu vykonanom na získaných výsledkoch je rozdiel medzi dvoma porovnávanými hodnotami štatisticky významný.

# Zhodnotenie

V rámci projektu sme študovali možnosti použitia poznámok a značiek pripojených k dokumentom na zlepšenie navigácie medzi dokumentami. Analyzovali sme niekoľko existujúcich nástrojov, ktoré používajú poznámky na podporu navigácie a organizácie kolekcie dokumentov. Navrhli sme a implementovali sme metódu na tvorbu dopytu na základe obsahu dokumentu a pripojených poznámok. Dopyt slúži na vyhľadávanie súvisiacich dokumentov k práve študovanému dokumentu. Poznámky používame ako indikátory používateľovho záujmu o konkrétne časti dokumentu. Navrhnutá metóda vytvára dopyt vo forme zoznamu kľúčových slov, ktorý je najčastejšie používanou formou dopytu v bežne dostupných vyhľadávačoch. Dopyt sa vytvára len z obsahu dokumentu a pripojených poznámok nezávisle od sady dokumentov, v ktorej sa má vyhľadávať. Tvorba dopytu je preto nezávislá od vyhľadávača a vytvorený dopyt je možné používať v ľubovoľnom vyhľadávači, ktorý prijíma dopyty vo forme zoznamu kľúčových slov.

Implementovali sme rozšírenie prehliadača Firefox, ktoré umožňuje používateľom pridávať rôzne poznámky do bežných webových stránok alebo PDF dokumentov zobrazených v prehliadači. Toto rozšírenie je súčasťou služby Annota, na ktorej tvorbe autor tohto dokumentu spolupracoval.

Presnosť vyhľadávania pomocou dopytu vytvorenou navrhnutou metódou sme overili prostredníctvom simulácie na dátovej sade získanej z Wikipédie a porovnaním s bežne používanou metódou na vyhľadávanie na základe textu, ktorú sme pre potreby experimentov rozšírili o možnosť zohľadniť pri vyhľadávaní poznámky pripojené k dokumentu. Navrhnutú metódu sme porovnali voči MoreLikeThis dopytu používanom v nástrojoch ElasticSearch a Solr. navrhnutá metóda dosiahla porovnateľné výsledky ako porovnávaná metóda v prípade, ak sme nepoužívali poznámky pri tvorbe dopytu a dosiahla významne lepšie výsledky oproti porovnáwanej metóde v prípade, ak sme použili poznámky ako indikátory záujmu používateľa o časti dokumentu.

Vykonalí sme používateľskú štúdiu, pomocou ktorej sme študovali zvyky používateľov pri pridávaní poznámok do dokumentov a zisťovali sme či používanie poznámok ako indikátorov záujmu používateľa pomocou navrhutej metódy zlepšuje presnosť vyhľadávania súvisiacich dokumentov. Podľa výsledkov tejto štúdie pomocou dopytov vytvorených s použitím poznámok pripojených k dokumentu navrhnutá metóda vytvára dopyt, pomocou ktorého sa dá vyhľadať viac relevantných výsledkov ako v prípade dopytu vytvoreného bez použitia poznámok.

Podobné výsledky sme dosiahli aj pomocou dlhodobého experimentu v nástroji Annota, kde sme nechali používateľov tohto nástroja vyhľadávať súvisiace články k zobrazenému dokumentu a nechali sme ich hodnotiť relevanciu nájdených článkov, pričom sme pomocou AB-testovania porovnávali výsledky získané s použitím poznámok a bez použitia poznámok pri tvorbe dopytu.

# Literatúra

- Abbasi, Rabeeh (2011). “Query expansion in folksonomies”. In: vol. 6725. Springer, pp. 1–16.
- Abel, Fabian et al. (2009). “Context-based ranking in folksonomies”. In: *Proceedings of the 20th ACM conference on Hypertext and hypermedia*. ACM, pp. 209–218.
- Agosti, Maristella and Nicola Ferro (Nov. 2007). “A formal model of annotations of digital content”. In: *ACM Trans. Inf. Syst.* 26.1.
- Bao, Shenghua et al. (May 2007). “Optimizing web search using social annotations”. In: *Proceedings of the 16th international conference on World Wide Web*. WWW '07. New York, USA: ACM, p. 501.
- Biancalana, Claudio (2009). “Social tagging in query expansion: A new way for personalized web search”. In: *Computational Science and 4*, pp. 1060–1065.
- Billerbeck, Bodo et al. (2003). “Query Expansion using Associated Queries”. In: *Proceedings of the twelfth international conference on Information and knowledge management*, pp. 2–9.
- Bottoni, Paolo et al. (2005). “Storing and Retrieving Multimedia Web Notes”. In: *Databases in Networked Information Systems: 4th International Workshop, DNIS 2005, Aizu-Wakamatsu, Japan, March 28-30, 2005, Proceedings*. Vol. 3433. Springer, p. 119.
- Buchanan, George and Jennifer Pearson (2008). *Improving Placeholders in Digital Documents Research and Advanced Technology for Digital Libraries*. Ed. by Birte Christensen-Dalsgaard et al. Vol. 5173. Lecture Notes in Computer Science. Springer Berlin / Heidelberg. Chap. 1, pp. 1–12.
- Cai, Yi and Qing Li (2010). “Personalized search by tag-based user profile and resource profile in collaborative tagging systems”. In: *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, pp. 969–978.
- Carmel, David et al. (Dec. 2010). “Social bookmark weighting for search and recommendation”. In: *The VLDB Journal* 19.6, pp. 761–775.
- Cattuto, Ciro, Roma La, and I Roma (2007). “Network Properties of Folksonomies”. In: *AI Communications* 20.4, pp. 245–262.
- Chirita, Paul Alexandru, Claudiu S. Firan, and Wolfgang Nejdl (July 2007). “Personalized query expansion for the web”. In: *Proceedings of the 30th annual international ACM SIGIR'07*. New York, USA: ACM Press, p. 7.

- Ciccarese, Paolo et al. (2011). “An open annotation ontology for science on web 3.0”. In: *J Biomed Semantics* 2.Suppl 2, S4.
- Claypool, Mark et al. (Jan. 2001). “Implicit interest indicators”. In: *Proceedings of the 6th international conference on Intelligent user interfaces - IUI '01*. IUI '01. New York, New York, USA: ACM Press, pp. 33–40.
- Dasdan, Ali et al. (Nov. 2009). “Automatic retrieval of similar content using search engine query interface”. In: *Proceeding of the 18th ACM conference on Information and knowledge management - CIKM '09*. New York, New York, USA: ACM Press, p. 701.
- Golder, Scott a. and Bernardo A. Huberman (Apr. 2006). “Usage patterns of collaborative tagging systems”. In: *Journal of Information Science* 32.2, pp. 198–208.
- Golovchinsky, Gene, Morgan N. Price, and Bill N. Schilit (1999). “From reading to retrieval: freeform ink annotations as queries”. In: *SIGCHI Bulletin*. ACM Press, pp. 19–25.
- He, Qi et al. (Apr. 2010). “Context-aware citation recommendation”. In: *Proceedings of the 19th international conference on World wide web - WWW '10*. New York, USA: ACM Press, p. 421.
- Holub, Michal and Mária Bieliková (2011). “An Inquiry into the Utilization of Behavior of Users in Personalized Web”. In: *Journal of Universal Computer Science* 17.13, pp. 1830–1853.
- Hotho, Andreas et al. (2006). “Information Retrieval in Folksonomies : Search and Ranking”. In: *The Semantic Web: Research and Applications* 4011, pp. 411–426.
- Jiao, Shanghai et al. (July 2008). “Exploring folksonomy for personalized search”. In: *Proceedings of the 31st annual international ACM SIGIR '08*. New York, USA: ACM Press, p. 155.
- Joachims, Thorsten (2002). “Optimizing search engines using clickthrough data”. In: *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '02*. KDD '02. Edmonton, Alberta, Canada: ACM, pp. 133–142.
- Kahan, J (Aug. 2002). “Annotea: an open RDF infrastructure for shared Web annotations”. In: *Computer Networks* 39.5, pp. 589–608.
- Kleinberg, Jon M (1999). “Authoritative sources in a hyperlinked environment”. In: *Journal of the ACM (JACM)* 46.5, pp. 604–632.
- Korkontzelos, Ioannis, Ioannis P Klapaftis, and Suresh Manandhar (2008). “Reviewing and evaluating automatic term recognition techniques”. In: *Advances in Natural Language Processing*. Springer, pp. 248–259.
- Körner, Christian et al. (June 2010). “Of categorizers and describers”. In: *Proceedings of the 21st ACM conference on Hypertext and hypermedia - HT '10*. New York, USA: ACM Press, p. 157.
- Li, Rui et al. (2007). “Towards Effective Browsing of Large Scale Social Annotations”. In: *Human Factors*, pp. 943–952.
- Page, Lawrence et al. (1999). “The PageRank citation ranking: bringing order to the web.” In:

- Palshikar, Girish Keshav (2007). “Keyword extraction from a single document using centrality measures”. In: *Pattern Recognition and Machine Intelligence*. Springer, pp. 503–510.
- Paranyushkin, Dmitry (2011). “Visualization of Text’s Polysingularity Using Network Analysis”. In: *Prototype Letters* 2.3, pp. 256–278.
- Pereira, Álvaro R and Nivio Ziviani (2003). “Retrieving similar documents from the web”. In: *Journal of Web Engineering* 2.4, pp. 247–261.
- Phelps, Thomas A and Robert Wilensky (2000). “Robust intra-document locations”. In: *Computer Networks* 33.1, pp. 105–118.
- Schilit, Bill N., Morgan N. Price, and Gene Golovchinsky (May 1998). “Digital library information appliances”. In: *Proceedings of the third ACM conference on Digital libraries - DL '98*. New York, USA: ACM Press, pp. 217–226.
- Ševcech, Jakub et al. (Nov. 2012). “Logging activity of researchers in digital library enhanced by annotations”. In: *7th Workshop on Intelligent and Knowledge oriented Technologies*, pp. 197–200.
- Šimko, Marián et al. (2011). “Supporting Collaborative Web-based Education via Annotations”. In: *World Conference on Educational Multimedia, Hypermedia and Telecommunications 2011*. Vol. 2011. 1, pp. 2576 –2585.
- Wu, Xian, Lei Zhang, and Yong Yu (May 2006). “Exploring social annotations for the semantic web”. In: *Proceedings of the 15th international conference on World Wide Web - WWW '06*. WWW '06. New York, USA: ACM Press, p. 417.
- Yang, Yin et al. (2009). “Query by document”. In: *Proceedings of the Second ACM International Conference on Web Search and Data Mining - WSDM '09*, p. 34.
- Yin, Zhijun, Milad Shokouhi, and Nick Craswell (2009). “Query Expansion Using External Evidence”. In: *Advances in Information Retrieval* 2, pp. 362–374.
- Zhang, Xiaoxun et al. (2009). “sDoc : Exploring Social Wisdom for Document Enhancement in Web Mining”. In: *Proceedings of the 18th ACM conference on Information and knowledge management*.





# Annota - Používateľská príručka

*Autori: Michal Holub, Róbert Móro, Jakub Ševcech, Roman Burger,  
Juraj Kostolanský, Martin Lipták, Samuel Molnár*

Nástroj Annota je služba na vytváranie záložiek a na pridávanie poznámok do webových stránok. Podporuje pridávanie tagov, zvýraznenie v texte, pridávanie komentárov a ukladanie poznámok vo forme voľného textu k záložke ako celku. Tieto poznámky je možné pridávať do webových stránok ako aj do PDF dokumentov. Rozhranie na pridávanie poznámok je implementované prostredníctvom rozšírenia pre prehliadač Firefox. Webové rozhranie aplikácie umožňuje organizovanie vlastnej knižnice dokumentov ako aj vyhľadávanie medzi vlastnými, ale aj všetkými verejnými dokumentami. Pomocou webového rozhrania je možné vytvárať skupiny pre používateľov a zdieľať záložky v rámci skupiny.

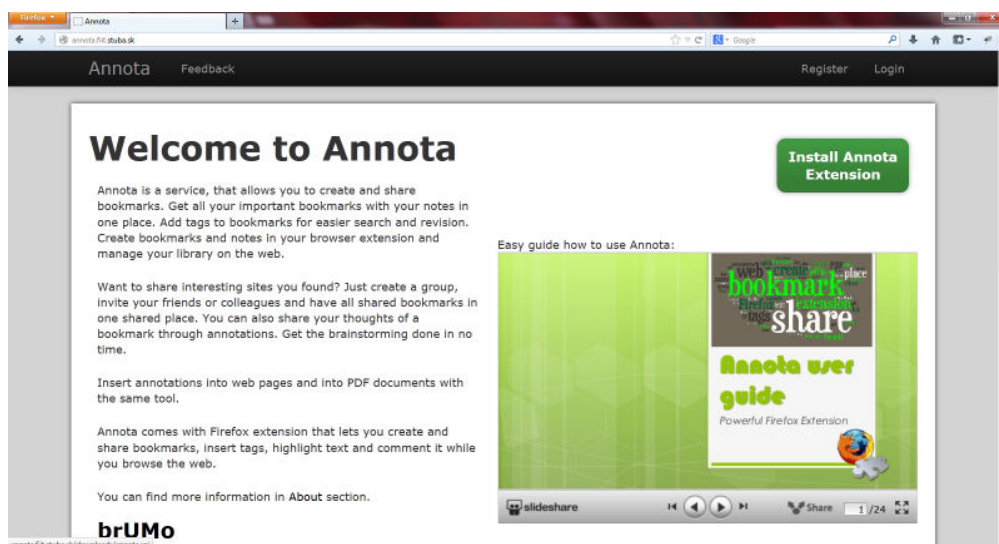
Webové rozhranie aplikácie ako aj rozšírenie pre prehliadač sú dostupné na adrese:

**<http://annota.fiit.stuba.sk/>**

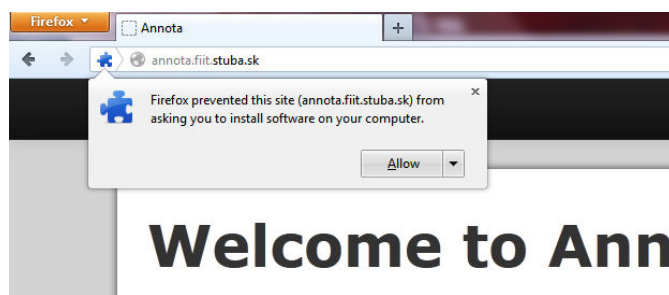
## Inštalácia

Po zobrazení adresy aplikácie v prehliadači Firefox sa zobrazí úvodná obrazovka, na ktorej sa nachádza základný opis nástroja a odkaz na inštaláciu rozšírenia (obrázok A.1).

Po kliknutí na odkaz na rozšírenie sa zobrazí výzva na povolenie inštalácie rozšírenia (obrázok A.2). Po potvrdení tejto výzvy prehliadač stiahne rozšírenie a zobrazí ďalšiu výzvu na potvrdenie inštalácie (obrázok A.3). Po potvrdení tejto

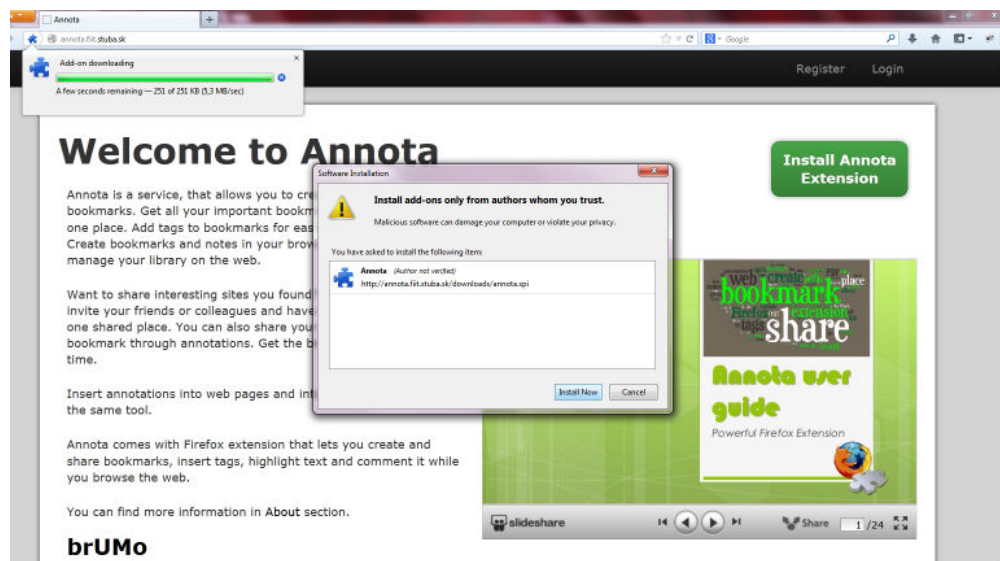


Obr. A.1: Úvodná stránka nástroja Annota.

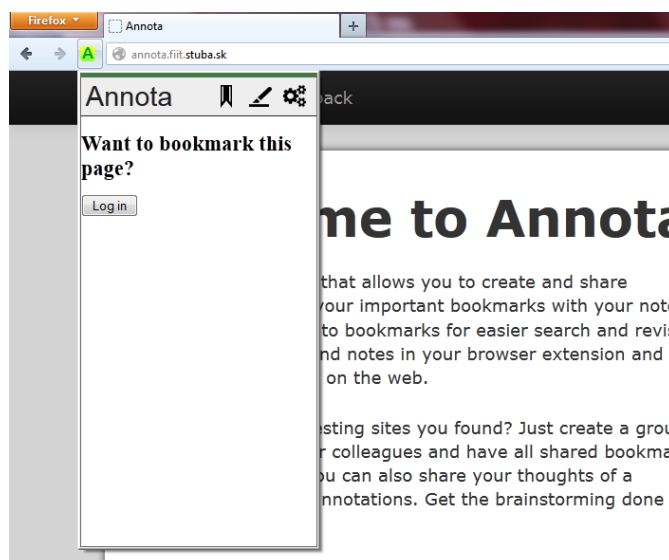


Obr. A.2: Výzva na potvrdenie inštalácie rozšírenia.

výzvy sa rozšírenie nainštaluje. Po reštartovaní prehliadača sa v hornej lište prehliadača zobrazí tlačidlo na zobrazenie vyskakovacieho okna (obrázok A.4). Vyskakovacie okno zobrazuje hlásenie o tom že používateľ nie je prihlásený do Annoty. Na plnohodnotné používanie rozšírenia je potrebné sa najskôr prihlásiť do Annoty prostredníctvom webového rozhrania.



Obr. A.3: Výzva na povolenie inštalácie rozšírenia.

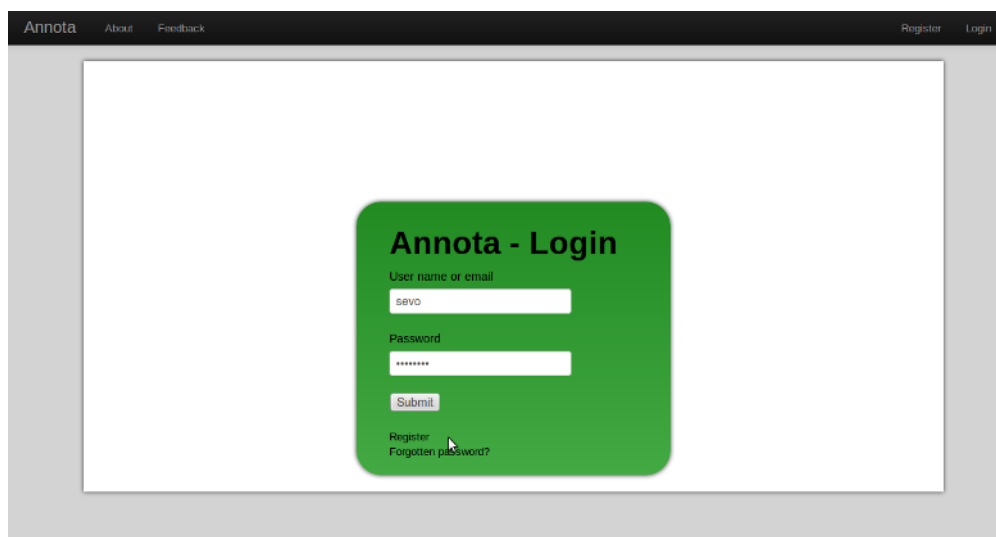


Obr. A.4: Vyskakovacie okno vyzývajúce používateľa k prihláseniu sa.

## Prvé použitie

Ako základný prostriedok na používanie rozšírenia slúži vyskakovacie okno, ktoré je možné zobrazit' kliknutím na logo aplikácie v navigačnej lište prehliadača (obrázok A.4). Keďže zatiaľ nie sme prihlásený do aplikácie, vyskakovacie okno zobrazuje výzvu na prihlásenie a odkaz na prihlasovací formulár.

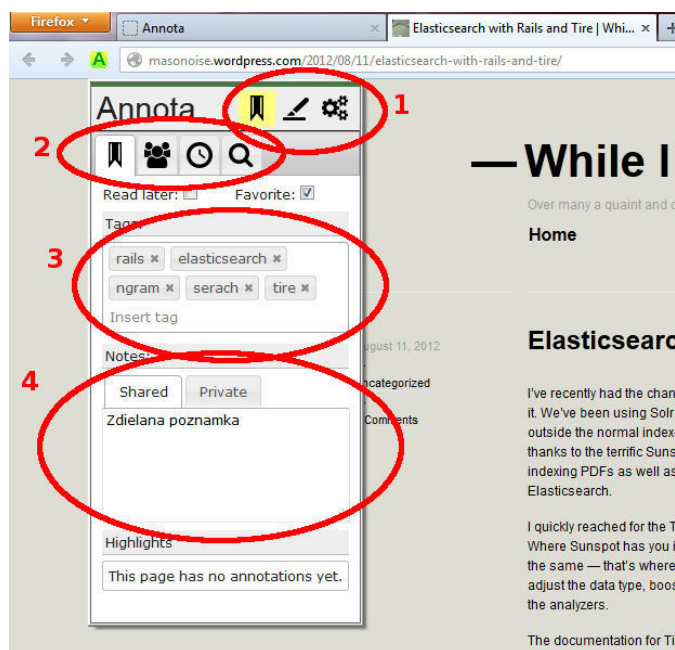
Pre používanie aplikácie je potrebné sa prihlásiť pomocou webového rozhrania. Prihlásenie do webového rozhrania slúži na prihlásenie tak do webového rozhrania ako aj na prihlásenie do rozšírenia. Stačí teda jedno prihlásenie na získanie prístupu do webového rozhrania ako aj na aktiváciu funkcií rozšírenia. Prihlasovací formulár (obrázok A.5) umožňuje zadať prihlasovacie údaje, ale tiež zobrazit' registračný formulár ako aj rozhranie na prihlásenie v prípade zabudnutého hesla. Registračný formulár je možné vyvolať aj pomocou odkazu „Register“ v hornej lište aplikácie.



Obr. A.5: Prihlasovací formulár.

Po prihlásení a zobrazení vyskakovacieho okna máme dostupné funkcie na pridávanie poznámok do webových stránok (obrázok A.6). Vyskakovacie okno poskytuje v hornej časti tlačidlo na vytvorenie záložky, ktoré podfarbením indikuje či je na danom dokumente vytvorená záložka, tlačidlo na zvýraznenie všetkých označených textov a tlačidlo na zobrazenie nastavení (1). Po stlačení tlačidla na zvýrazňovanie všetkých označených textov sa každý text, ktorý následne používateľ označí na stránke zvýrazní. Po opätovnom kliknutí na toto tlačidlo sa táto funkcia vypne a používateľ bude môcť znovu označovať text na stránke bez toho aby sa podfarboval. Vytvorit' zvýraznenie v texte je možné ešte druhým spôsobom, ktorý nevyžaduje spúšťanie tejto funkcie. V prípade ak používateľ označí ľubovoľný text na stránke, klikne naň pravým tlačidlom myši a zvolí možnosť „Highlight“ z kontextového menu. Pri tomto spôsobe zvýrazňovania hrozí menšie nebezpečenstvo podfarbenia nechceného textu.

Funkcie vo vyskakovacom okne sú organizované do niekoľkých záložiek (2). Prvá



Obr. A.6: Vyskakovacie okno na pridávanie poznámok do webových stránok.

záložka (zobrazená na obrázku A.6) poskytuje funkcie na pridávanie rôznych typov poznámok k dokumentu. Je možné pridávať tagy (3) alebo poznámky vo forme voľného textu (4). V dolnej časti tejto záložky sa zobrazuje zoznam vytvorených zvýraznení usporiadaný podľa času vytvorenia zvýraznenia. Po kliknutí na niektoré zo zvýraznení, ktoré sa tu môžu nachádzať, sa presunie pohľad obrazovky na dané miesto v dokumente.

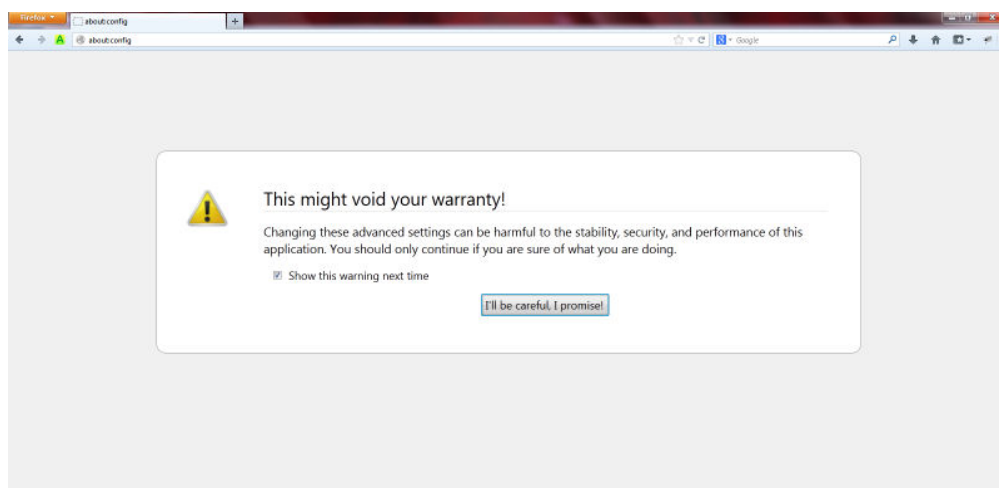
Ďalšie záložky vo vyskakovacom okne poskytujú funkcie na zdieľanie záložky so skupinami a na vytvorenie novej skupiny, zobrazenie záložiek označených na neskoršie prečítanie a na vyhľadanie v zozname vlastných záložiek (2).

K zvýrazneniu v texte je možné pridávať komentáre. Okno na pridanie komentáru sa zobrazí po prechode myšou nad zvýrazneným textom. V tomto okne je možné pridávať komentár, označovať poznámku za privátnu ako aj odstránenie zvýraznenia.

## Poznámky v PDF

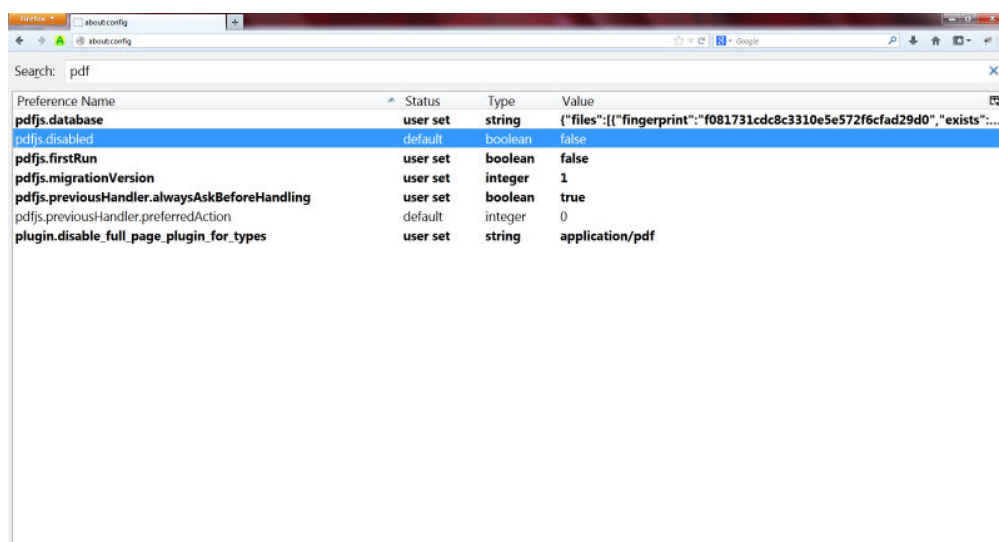
Prehliadač Firefox umožňuje od verzie 15 zobrazovať PDF dokumenty priamo v prehliadači<sup>23</sup>. Pre zobrazovanie PDF dokumentov pomocou vstavanej podpory, nesmie byť nainštalované v prehliadači žiadne iné rozšírenie na zobrazovanie PDF dokumentov. Pre verzie Firefoxu nižšie ako 20 je potrebné povoliť natívne zobrazovanie PDF dokumentov. Pre povolenie zobrazovania PDF je potrebné

<sup>23</sup>Staršie verzie podporujú zobrazovanie PDF dokumentov pomocou rozšírenia <https://addons.mozilla.org/en-US/firefox/addon/pdfjs/>



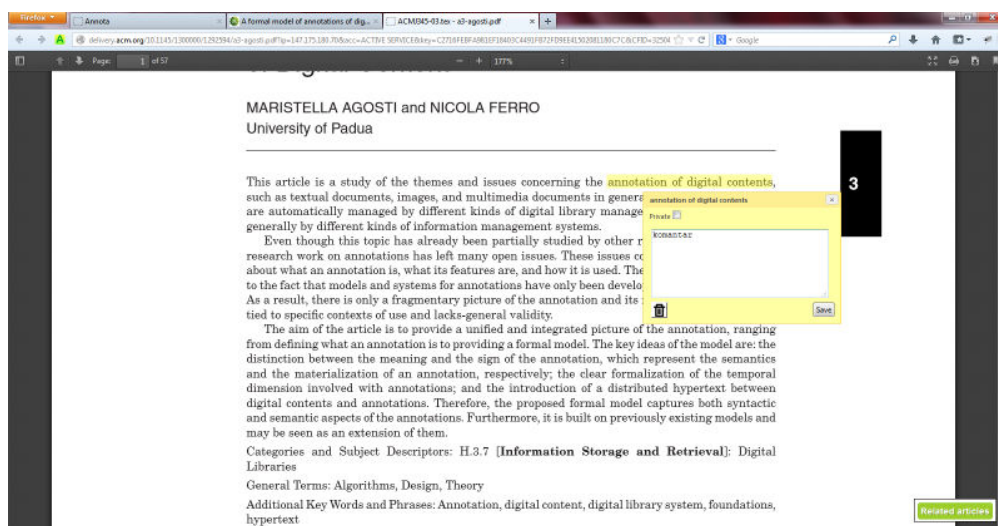
Obr. A.7: Potvrdenie výstrahy pri povoľovaní podpory zobrazovania PDF dokumentov.

navigovať sa na adresu „about:config“, potvrdiť upozornenie (obrázok A.7), vyhľadať záznam „pdfjs.disabled“ a nastaviť ho na hodnotu „true“ (obrázok A.8). Zmena sa prejaví po reštarte prehliadača.



Obr. A.8: Povolenie podpory zobrazovania PDF dokumentov.

Po povolení zobrazovania PDF dokumentov a po prihlásení sa do rozšírenia má používateľ možnosť pridávať všetky poznámky, ktoré mohol pridávať do webových stránok aj do PDF dokumentov (obrázok A.9).



Obr. A.9: Pridávanie poznámok do PDF dokumentov priamo v prehliadači.

## Webové rozhranie

Rozšírenie prehliadača slúži na vytváranie poznámok. Na organizáciu vytvorených záložiek slúži webové rozhranie. Toto rozhranie umožňuje organizovať vlastnú zbierku poznámok a zdieľať záložky v skupinách. Okrem toho, webové rozhranie umožňuje napríklad aj import záložiek zo služby Delicious alebo Mendeley.

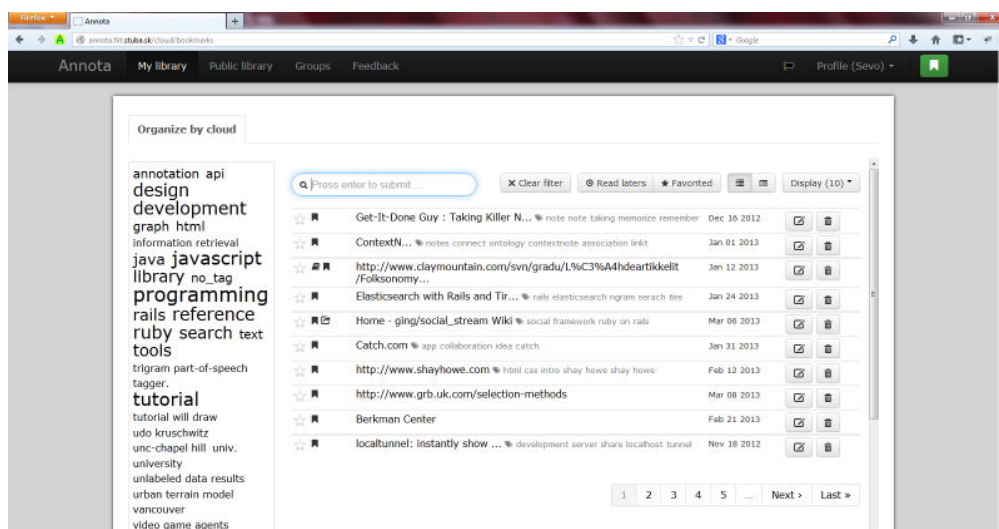
## Organizácia knižnice záložiek

Po kliknutí na voľbu „My library“ v navigačnej lište sa zobrazí zoznam záložiek, ktoré používateľ vytvoril. Ku každej záložke sú priradené tagy, ktoré k nim používateľ pridal a poznámky, ktoré k nim napísal (obrázok A.10). Záložky je možné označovať hviezdíčkou ako obľúbené. Na základe toho či je záložka obľúbená alebo či je označená na neskoršie prečítanie je možné záložky filtrovať. Hlavný prostriedok na na filtrovanie záložiek je pomocou oblaku kľúčových slov založeného na tagoch priradených k záložkám.

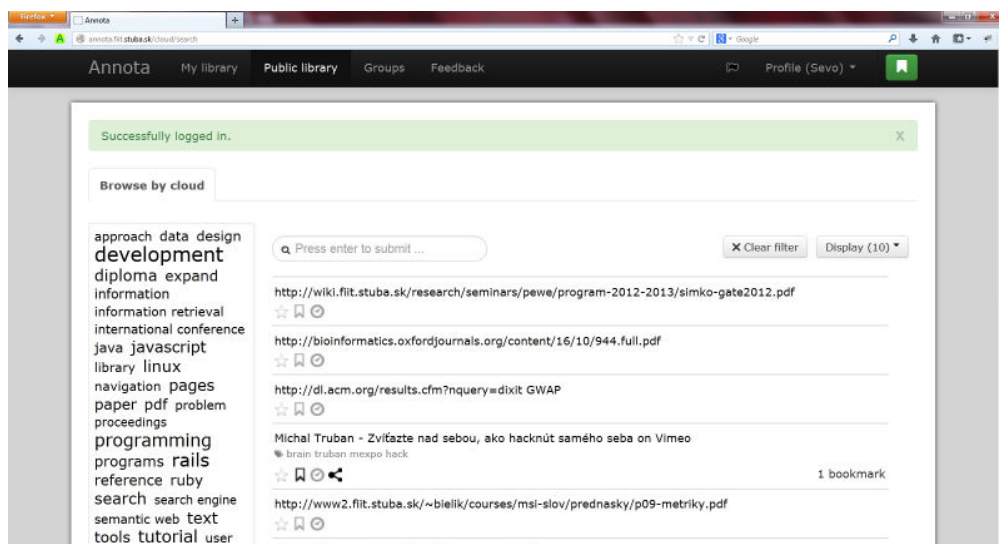
## Vyhľadávanie dokumentov

Po kliknutí na voľbu „Public library“ v navigačnej lište sa zobrazí zoznam všetkých verejných dokumentov, ktoré pridali medzi záložky používateľa Annoty (obrázok A.11). podobne ako pri organizácii vlastných záložiek aj tu sa dá navigovať pomocou oblaku kľúčových slov a pomocou fulltextového vyhľadávania.

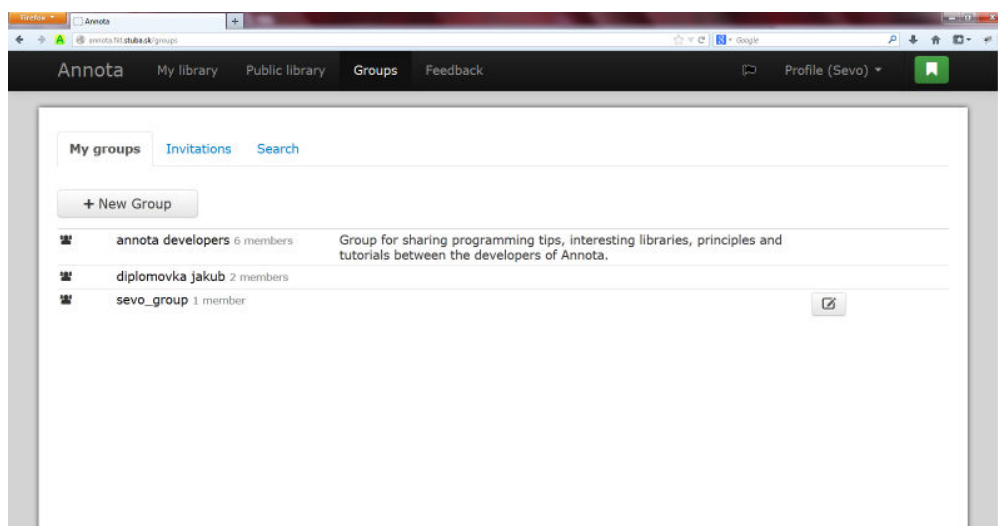




Obr. A.10: Rozhranie na organizáciu zbierky záložiek.



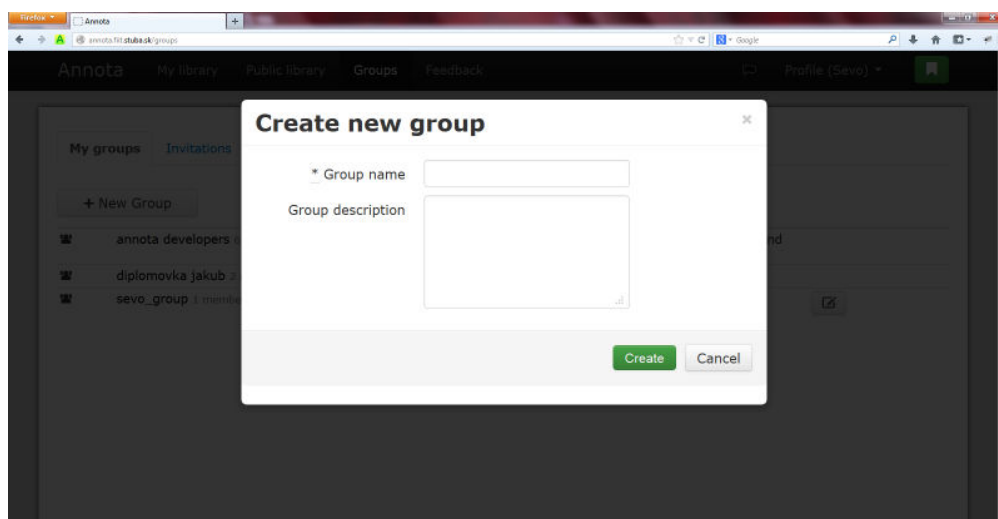
Obr. A.11: Rozhranie na vyhľadávanie medzi všetkými verejnými záložkami.



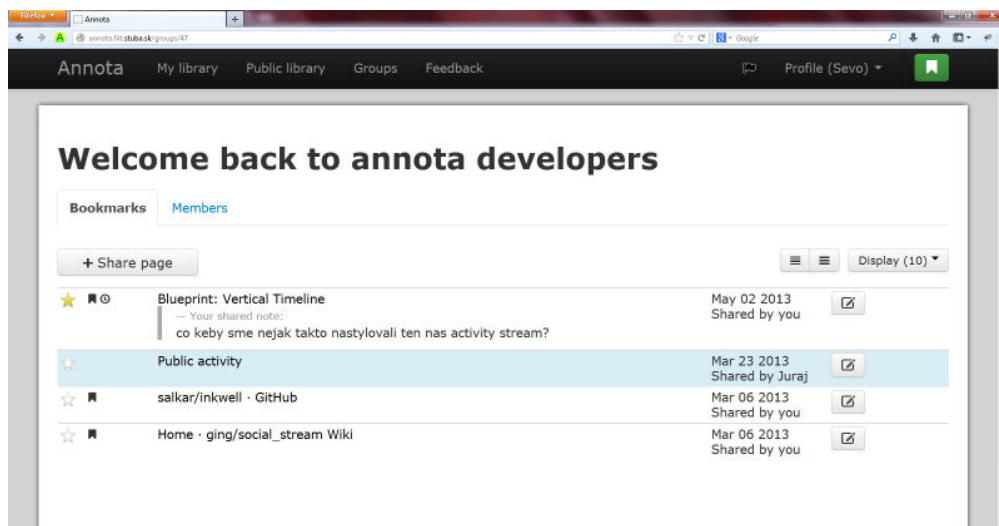
Obr. A.12: Zoznam skupín, v ktorých je používateľ členom.

## Zdieľanie záložiek

Vytvorené záložky je možné zdieľať so skupinou používateľov. Na zobrazenie zoznamu skupín, ktorých je používateľ členom slúži voľba „Groups“ v navigačnej lište. Po kliknutí na túto lištu sa zobrazí zoznam skupín spolu s tlačidlom na vytvorenie novej skupiny (obrázok A.12). Po kliknutí na toto tlačidlo sa zobrazí formulár na vyplnenie detailov o novej skupine a na vytvorenie tejto skupiny (obrázok A.13). Po kliknutí na niektorú skupinu v zozname skupín sa zobrazí zoznam záložiek, ktoré so skupinou zdieľali členovia skupiny (obrázok A.14). Zdieľanie záložky je možné zo zoznamu vlastných záložiek vo webovom rozhraní ako aj z rozšírenia prehliadača (obrázok A.6 časť 4)



Obr. A.13: Formulár na vytvorenie novej skupiny.

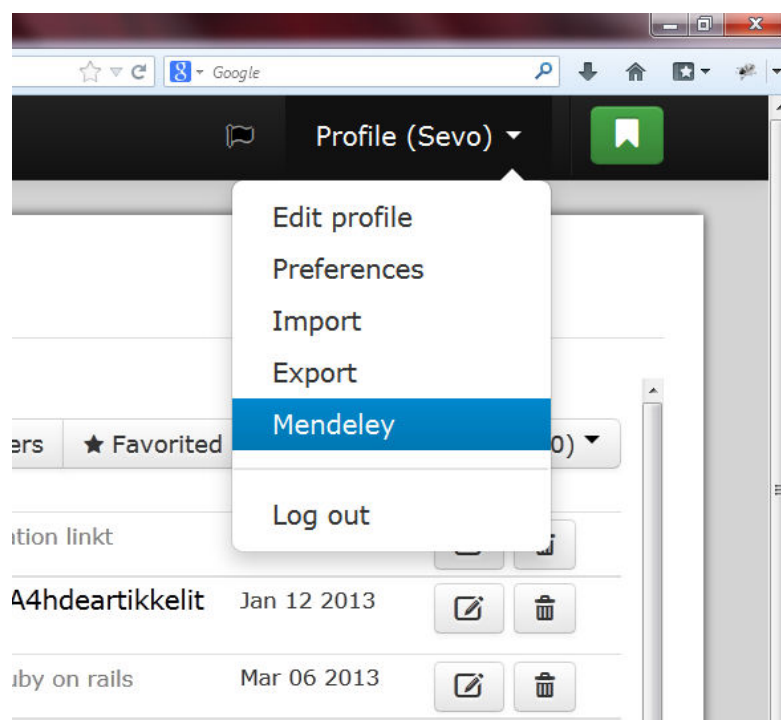


Obr. A.14: Zoznam záložiek zdieľaných v skupine.

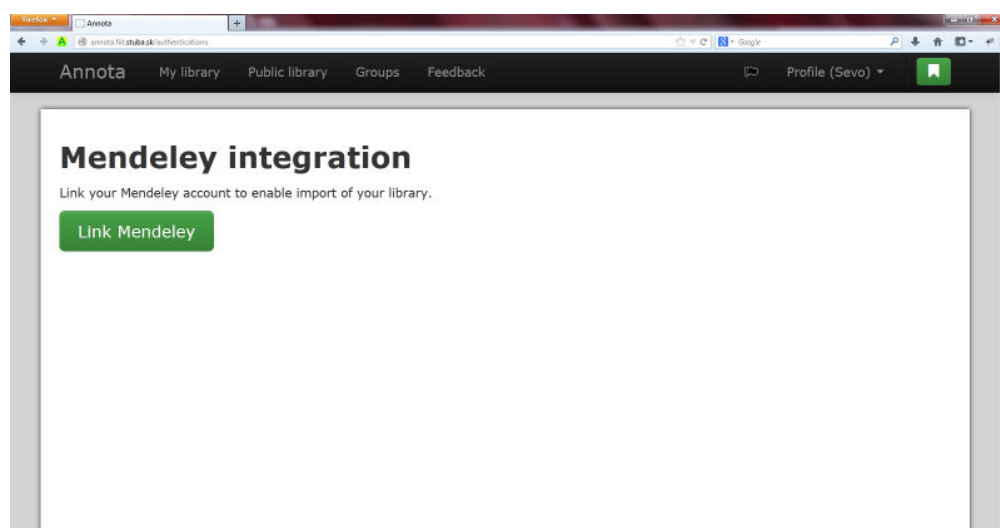
## Import údajov zo služby Mendeley

Annota podporuje možnosť importovať údaje nazbierané pomocou služby Mendeley. Importujú sa tu všetky dokumenty, z ktorých sa stanú záložky, priradené poznámky (poznámky k celému dokumentu, nie poznámky v texte), tagy priradené k dokumentom a zaradenie dokumentov do adresárov. Prístup k funkcionalite na import údajov je pomocou menu v hornej lište A.15. Menu sa zobrazí po kliknutí na meno prihláseného používateľa v pravej časti hornej lišty aplikácie. Po kliknutí na voľbu „Mendeley“ sa zobrazí výzva na prepojenie Annoty a Mendeleya A.16.

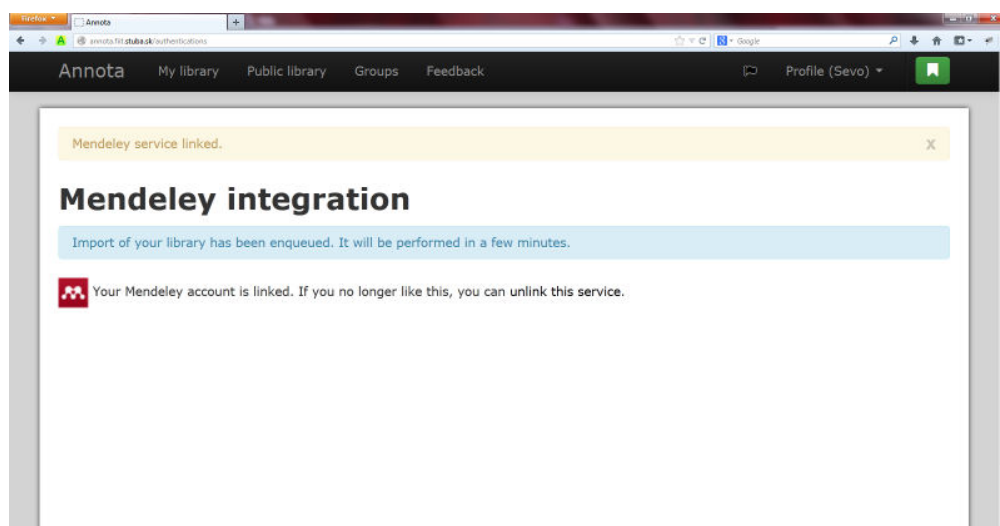
Po kliknutí na voľbu „Link Mendeley“ sa zobrazí výzva na potvrdenie prepojenia. V závislosti na tom či je používateľ práve prihlásený do webového rozhrania služby Mendeley sa môže najskôr zobrazit' výzva na prihlásenie. Po potvrdení prepojenia sa používateľovi opäť zobrazí webové rozhranie Annoty s hlásením že príkaz na import údajov bol zaradený na spracovanie a prebehne o chvíľu A.17. Po dobehnutí importu sa na tejto stránke zobrazí informácia o poslednom čase spustenia importu a možnosť import opätovne spustiť.



Obr. A.15: Menu na prístup k funkcii importovania údajov zo služby Mendeley.



Obr. A.16: Výzva na prepojenie služby Mendeley a Annoty.



Obr. A.17: Hlásenie o zaradení požiadavky na import na spracovanie.

# Annota - Technická dokumentácia

Táto časť dokumentu obsahuje technickú dokumentáciu k vytvorenej službe Annota, ktorá slúži na vytváranie záložiek a pridávanie poznámok do webových stránok. Aplikácia je zložená z dvoch častí: klient a server. Pri implementácii serverovej časti aplikácie sme použili programovací jazyk Ruby a vývojový rámec Ruby on Rails. Klientskú časť aplikácie tvorí rozšírenie pre prehliadač Firefox, ktoré komunikuje so serverovou časťou prostredníctvom REST webového rozhrania a na výmenu údajov používa posielanie správ vo formáte JSON. Klientská časť je implementovaná v jazyku JavaScript.

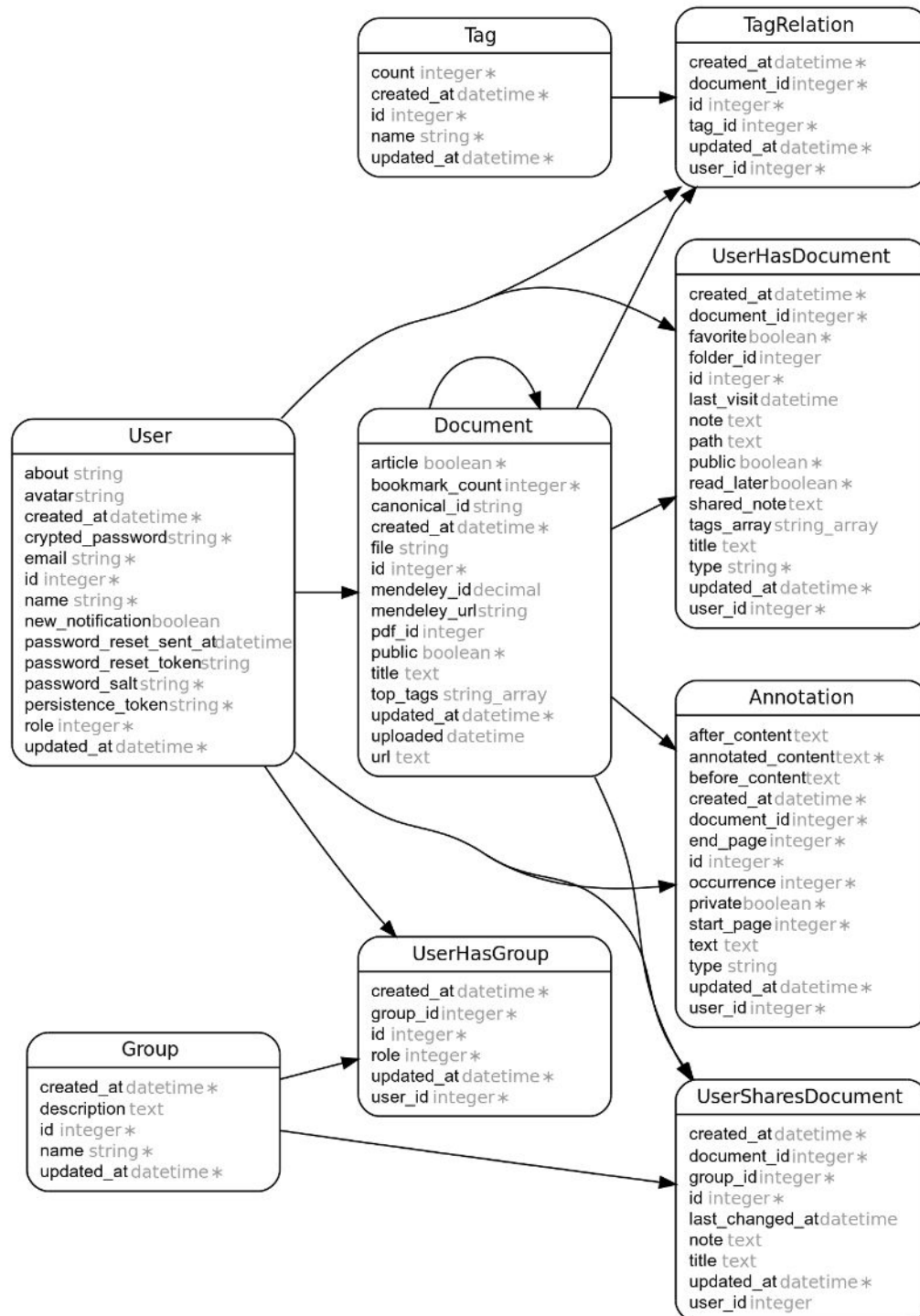
## Dátový model

Na diagrame na obrázku B.1 je znázornený fyzický dátový model serverovej časti aplikácie Annota. Ako úložisko údajov je použitá databáza PostgreSQL<sup>24</sup> a na vyhľadávanie medzi uloženými dokumentami používame vyhľadávač ElasticSearch<sup>25</sup>.

---

<sup>24</sup>PostgreSQL, <http://www.postgresql.org>

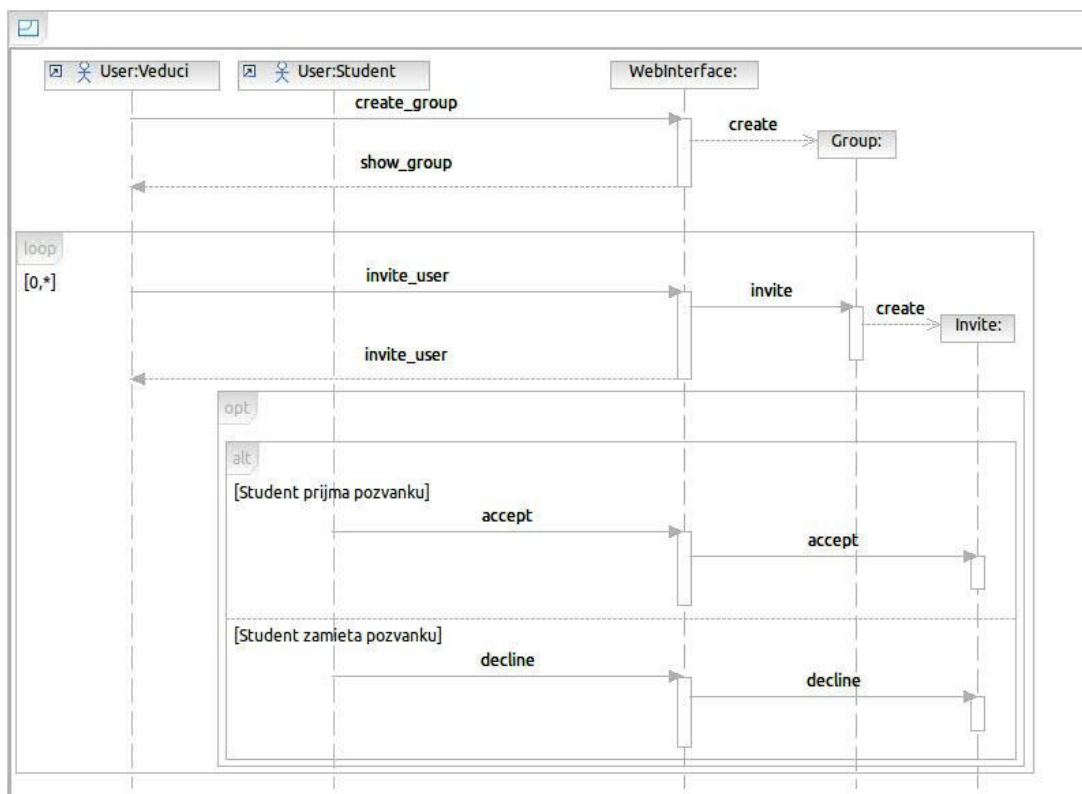
<sup>25</sup>ElasticSearch, <http://www.elasticsearch.org>



Obr. B.1: Najvýznamnejšie časti dátového modelu serverovej časti aplikácie Annota.

## Vytvorenie skupiny

Na diagrame na obrázku B.2 je znázornený sekvenčný diagram, ktorý opisuje postupnosť krokov potrebných na vytvorenie skupiny, pozvanie študenta do skupiny a prijatie resp. zamietnutie pozvánky študentom.

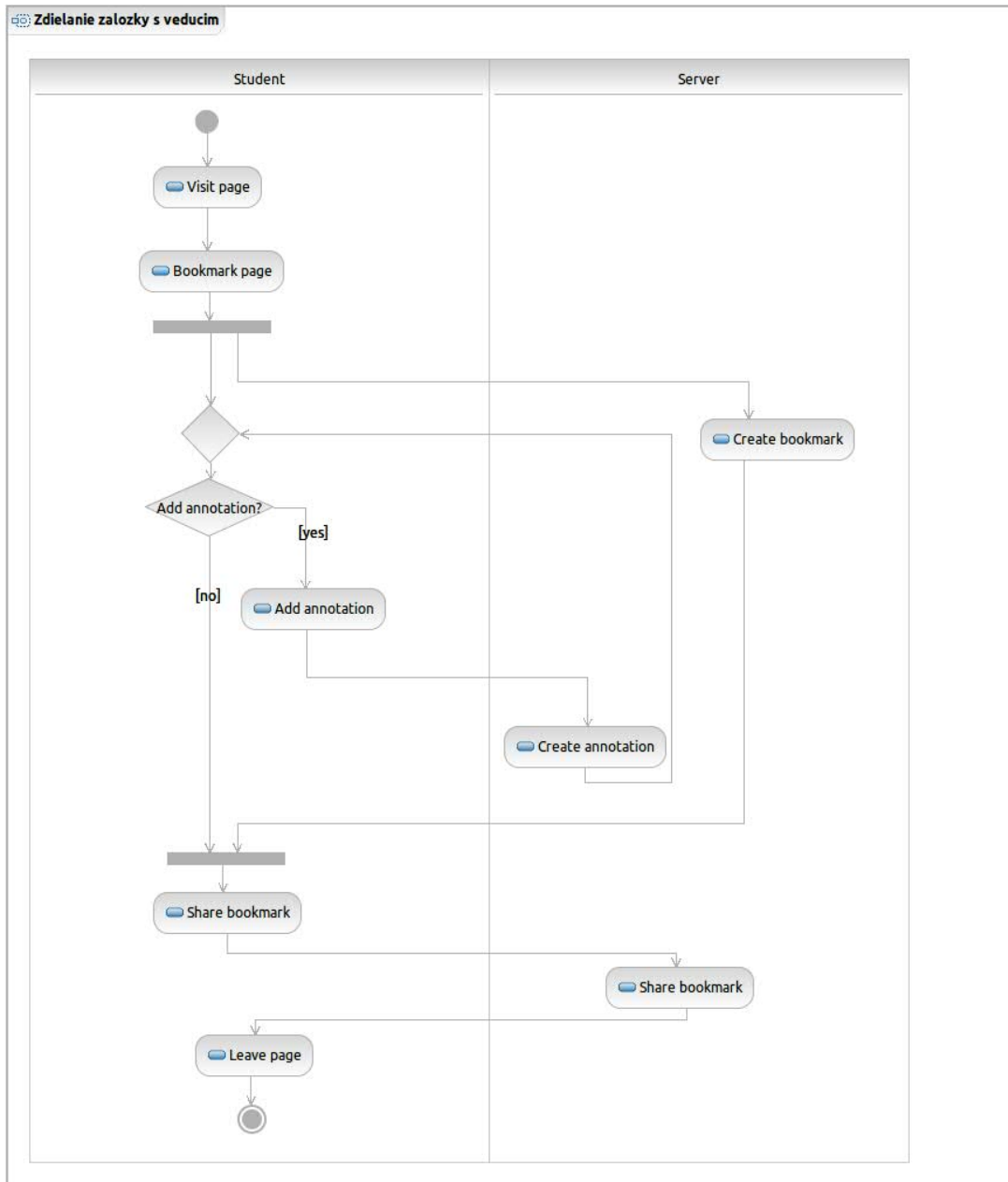


Obr. B.2: Sekvenčný diagram pre vytvorenie skupiny pre študentov.

## Zdieľanie záložky

Diagram na obrázku B.3 znázorňuje diagram činností, ktorý opisuje proces, v ktorom študent vytvára zvýraznenia v texte webovej stránky a opoznámkovanú záložku zdieľa v skupine so svojim vedúcim. Tento diagram popisuje jednu z hlavných úloh nástroja Annota, a to podporu pri spolupráci študenta a jeho vedúceho pri vyhľadávaní zdrojov.

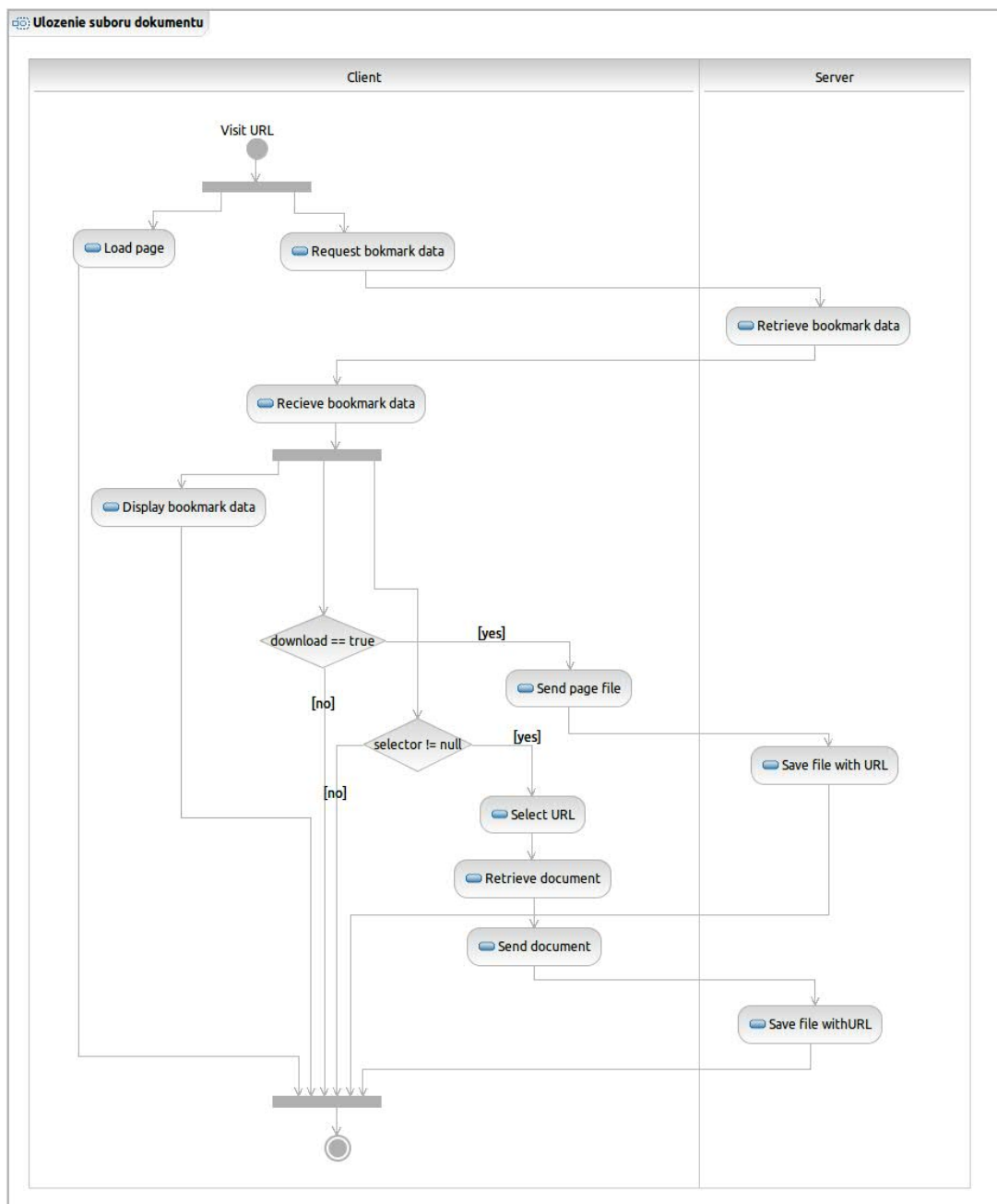




Obr. B.3: Diagram činností pre zdieľanie záložky s vedúcim.

## Ukladanie navštívených stránok

Obrázok B.4 znázorňuje pomocou diagramu činností, proces nahrávania súboru navštívenej stránky digitálnej knižnice a posielanie PDF dokumentu, na ktorý stránka odkazuje. Webové stránky sa posielajú len pri návšteve digitálnej knižnice a to len v prípade, ak tento dokument ešte nebol nahraný na server. Podobne aj PDF dokumenty sa posielajú len pri návšteve digitálnej knižnice a len ak sa tento dokument ešte nenachádza na servere.

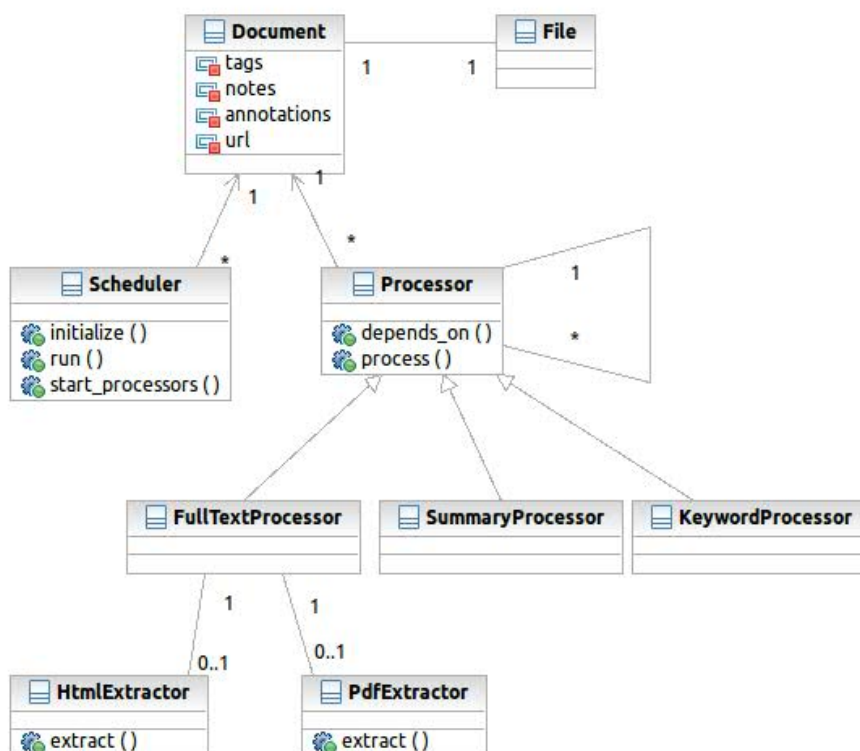


Obr. B.4: Diagram činností pre posielanie súborov webových stránok a PDF dokumentov na sever.

## Spracovanie dokumentu

Na obrázku číslo B.5 je znázornený diagram tried pre triedy, ktoré sa starajú o spracovanie súboru pripojeného k dokumentu na strane servera. Po nahraní dokumentu je potrebné z tohto extrahovať úplný text, kľúčové slová a vytvoriť sumariáciu, ktorá sa bude zobrazovať pri dokumente vo výsledkoch

vyhľadávania. Na získanie týchto informácií slúžia triedy, ktoré dedia od triedy *Processor*. Trieda *FullTextProcessor* slúži na extrakciu textu z dokumentu. V závislosti od typu dokumentu na to používa triedy *HtmlExtractor* a *PdfExtractor*. Trieda *SummaryProcessor* slúži na vytvorenie sumarizácie z textu dokumentu a trieda *KeywordProcessor* slúži na extrakciu kľúčových slov z textu dokumentu. Spúšťanie jednotlivých extrakcií nad dokumentom riadi trieda *Scheduler*. Spracovanie dokumentu sa vykonáva na pozadí v poradí vypočítanom na základe definícií závislostí v každej z tried na extrakciu údajov z dokumentu.



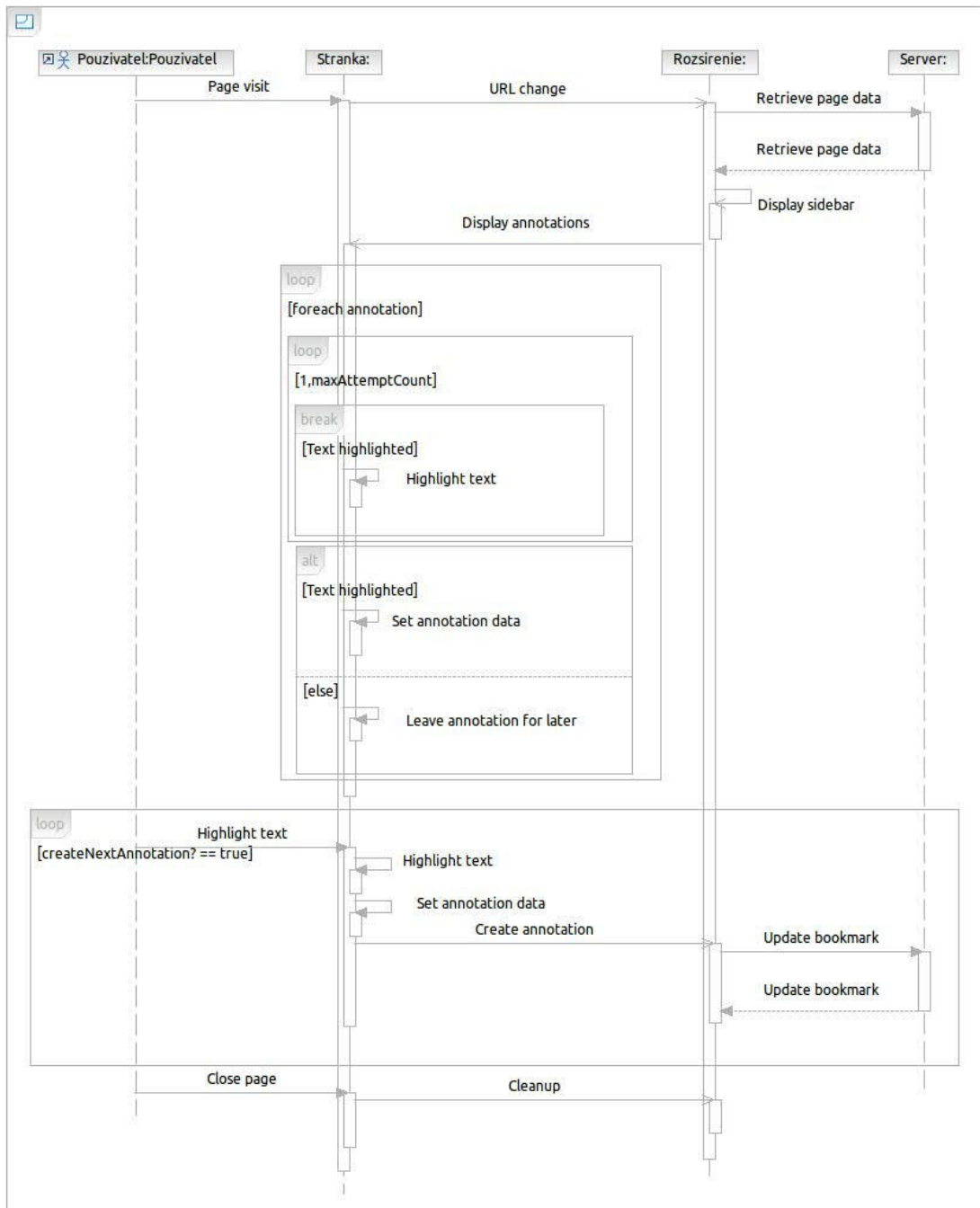
Obr. B.5: Diagram tried znázorňujúci spracovanie súboru pripojeného k dokumentu.

## Vykresľovanie zvýraznení vo webovej stránke

Sekvenčný diagram na obrázku číslo B.6 zobrazuje algoritmus na vykreslenie zvýraznení do textu webovej stránky.

Po zobrazení webovej stránky rozšírenie získa zo servera informácie o danej stránke a postupne vykreslí do webovej stránky všetky zvýraznenia. Proces vykresľovania zvýraznenia môže v prípade, ak sa označený text na stránke nenachádza zlyhať. Tento stav môže nastať v prípade, ak sa stránka zmenila, ak ide o stránku, kde sa obsah dynamicky mení alebo je to PDF dokument, ktorý sa vykresľuje po

jednotlivých stránkách. Z tohto dôvodu samotné vykresľovanie prebieha opakovane a zastaví sa po úspešnom vykreslení zvýraznenia alebo po maximálnom počte pokusov o vykreslenie. Ak sa nepodarí vykreslenie, tak sa zvýraznenie odloží na neskoršie vykreslenie, ktoré nastane napríklad po zachytení signálu o zmene obsahu stránky.



Obr. B.6: Sekvenčný diagram algoritmu na vykresľovanie zvýraznení v stránke.



# Annota - Inštalačná príručka

Pre nainštalovanie a spustenie služby Annota je potrebné mať nainštalované:

- interpretér jazyka Ruby verzie 1.9.3-p286
- PostgreSQL verzie 9.1 alebo vyššej
- Elasticsearch verzie 0.20.2 alebo vyššej

Z priloženého média stiahnite zdrojové súbory aplikácie Annota.

Vytvorte prázdnu PostgreSQL databázu a nastavte prístupové údaje k tejto databáze do súboru `config/database.yml`

Nainštalujte knižnice, ktoré aplikácia využíva. Najskôr knižnicu bundler na správu knižníc príkazom `gem bundler install` a potom všetky knižnice definované v súbore `Gemfile` príkazom `bundle install`

Spustite migrácie pre vytvorenie tabuliek v databáze príkazom `rake db:migrate`

Vytvorte indexy dokumentov v nástroji Elasticsearch príkazmi:

- `rake environment tire:import CLASS=Document FORCE=true`
- `rake environment tire:import CLASS=User FORCE=true`
- `rake environment tire:import CLASS=Group FORCE=true`
- `rake environment tire:import CLASS=UserNavigation FORCE=true`

Spustite aplikciu prkazom `rails s`

Po spustení serveru bude aplikcia dostupn na adrese `http://localhost:3000/` prostrednctvom webovho prehliadača.

V prpade nespechu je mozn si vyskšať aplikciu na adrese `http://annota.fiit.stuba.sk`

# Analýza údajov z používania nástroja Annota

V tejto časti dokumentu sa nachádza analýza vlastností údajov nazbieraných pomocou nástroja Annota k 9.5.2013

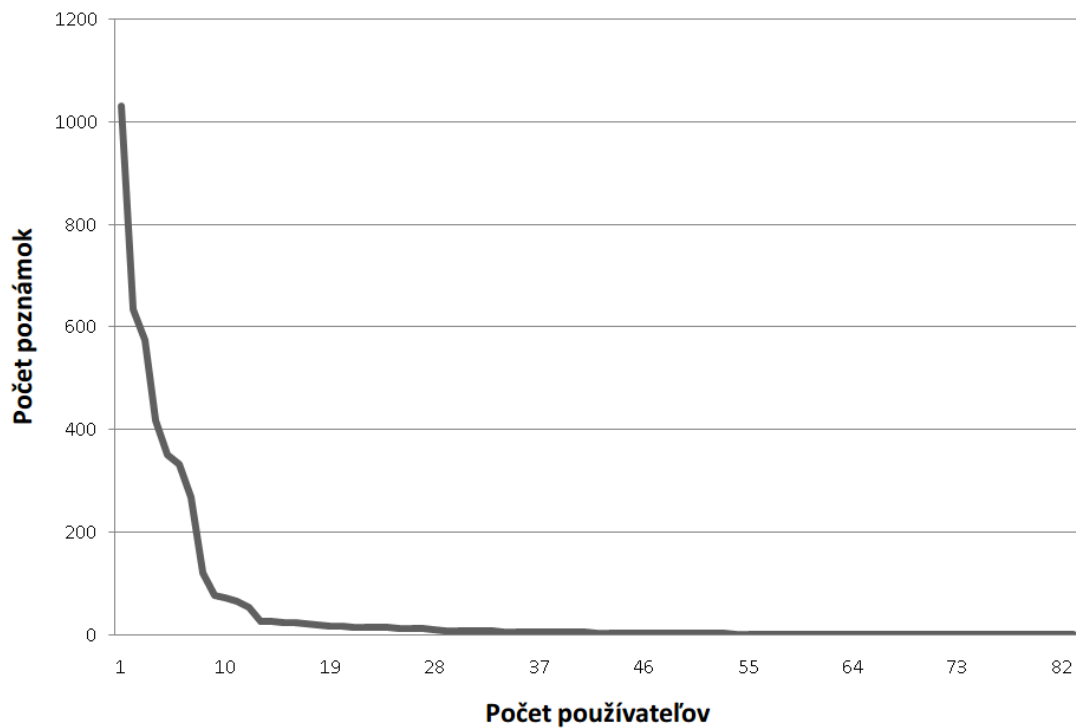
Nástroj Annota používa 113 registrovaných používateľov na vytáranie záložiek a pridávanie poznámok k webovým stránkam a PDF dokumentom. Spolu vytvorili 4444 záložiek, z ktorých 1294 sú záložky dokumentov v digitálnych knižniciach.

Pomocou ukladania súborov navštívených stránok digitálnej knižnice sa spolu nazbieralo 4151 súborov, z ktorých 2781 sú PDF dokumenty a zvyšok sú HTML súbory navštívených stránok.

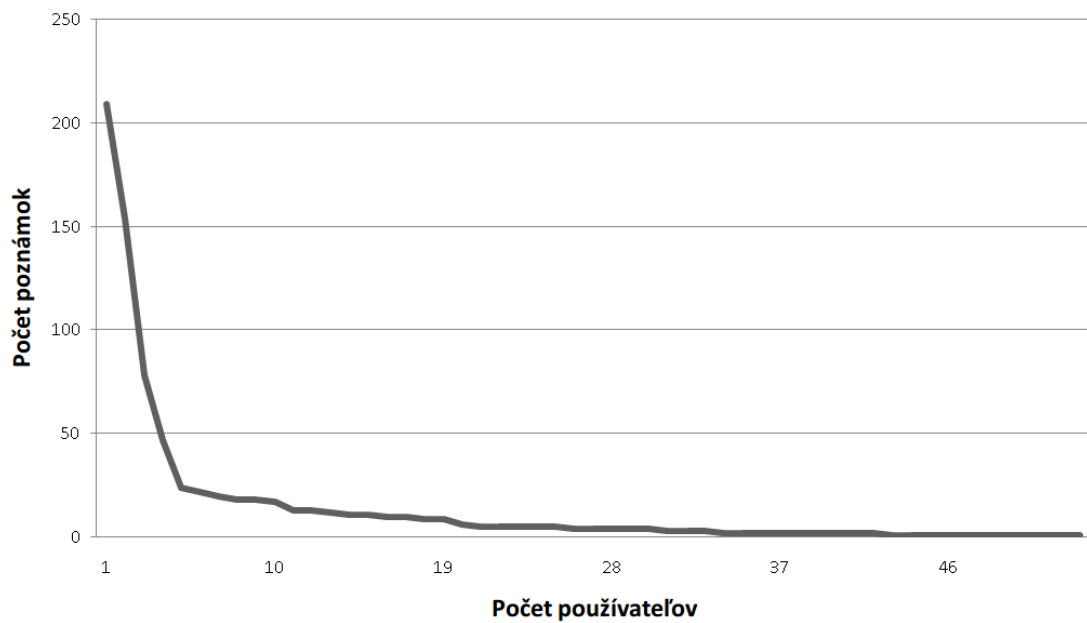
Používatelia vytvorili 53 skupín, do ktorých sa zaradilo 92 používateľov, pričom 54 rôznych používateľov je aspoň v jednej skupine.

Na grafoch na obrázkoch D.1, D.2 a D.3 môžeme vidieť postupne počet záložiek, poznámok a tagov, ktoré vytvorili používatelia. Na všetkých troch grafoch je vidieť že namerané hodnoty sa riadia podľa mocninového pravidla. To znamená že v systéme je niekoľko aktívnych používateľov, ktorí vytvoria väčšinu záložiek, poznámok a tagov. Ostatní používatelia prispievajú do celkového množstva nazbieraných údajov len veľmi malým podielom.

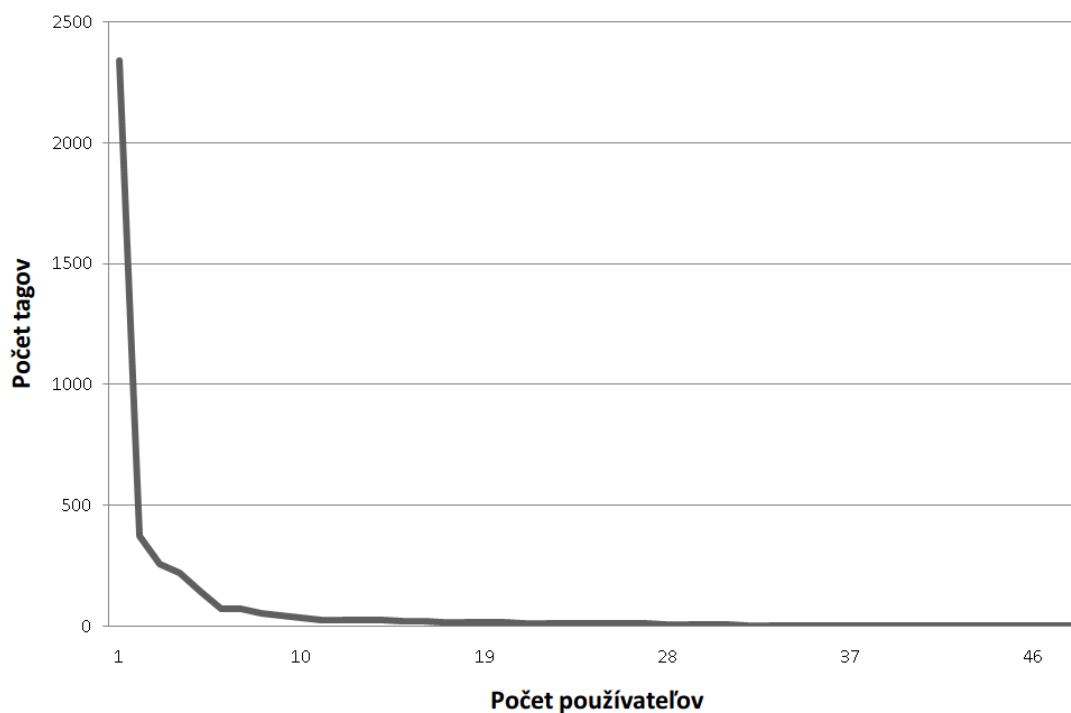




Obr. D.1: Príspevok používateľov k celkovému počtu záložiek.

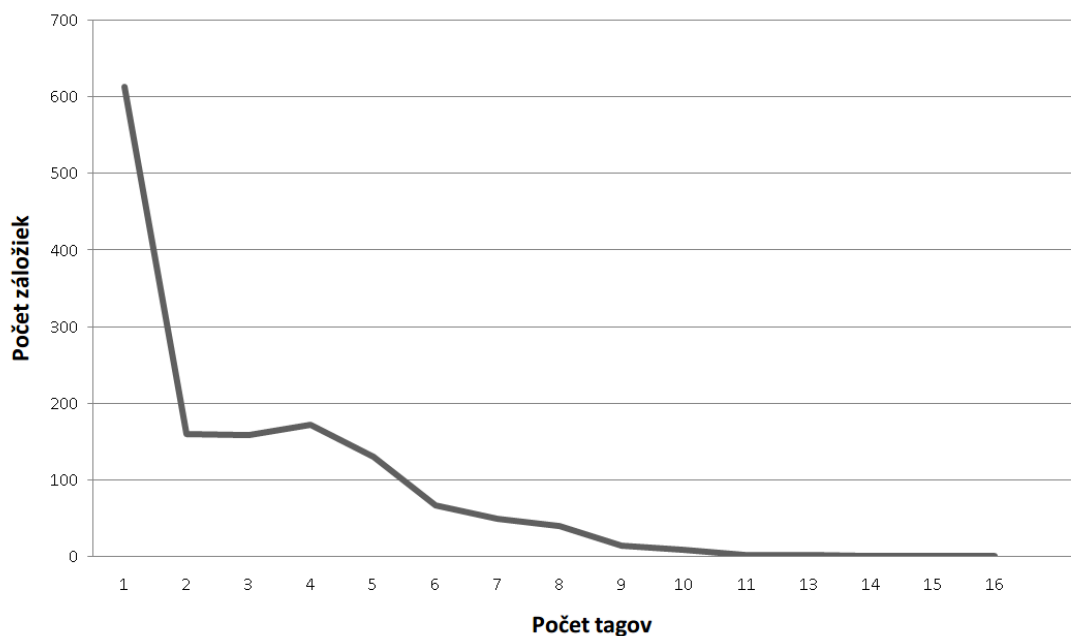


Obr. D.2: Príspevok používateľov k celkovému počtu poznámok.



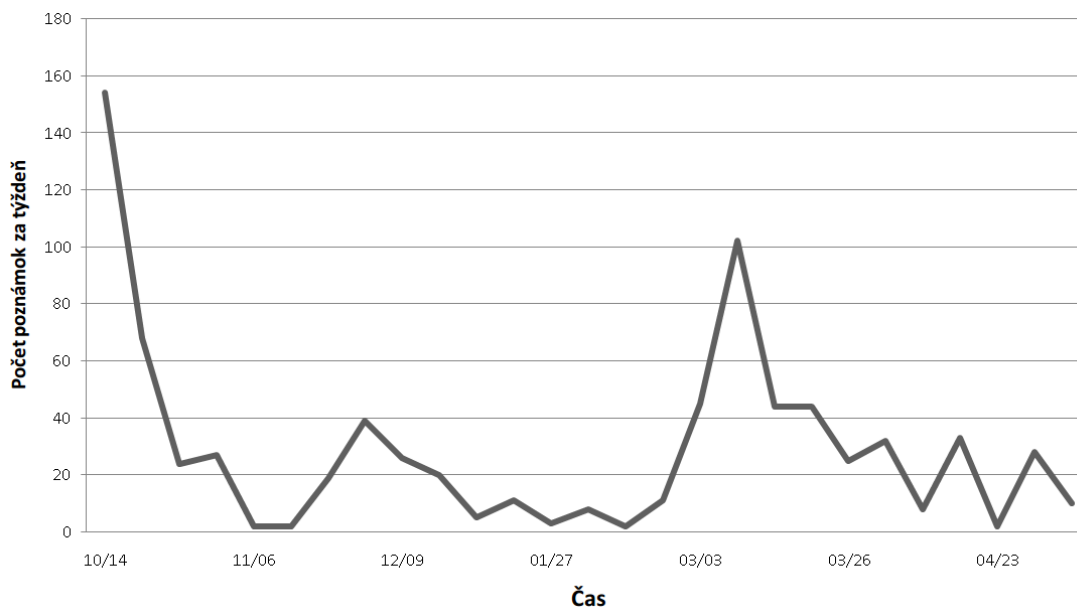
Obr. D.3: Príspevok používateľov k celkovému počtu tagov.

V grafe na obrázku D.4 je zobrazená početnosť záložiek podľa počtu priradených tagov. Na tomto grafe vidíme že počet tagov priradených k záložkám sa riadi mocninovým pravidlom. Väčšina otagovaných záložiek, má priradené jeden až štyri tagy, ale existujú záložky, ktoré majú aj viac ako desať priradených tagov.

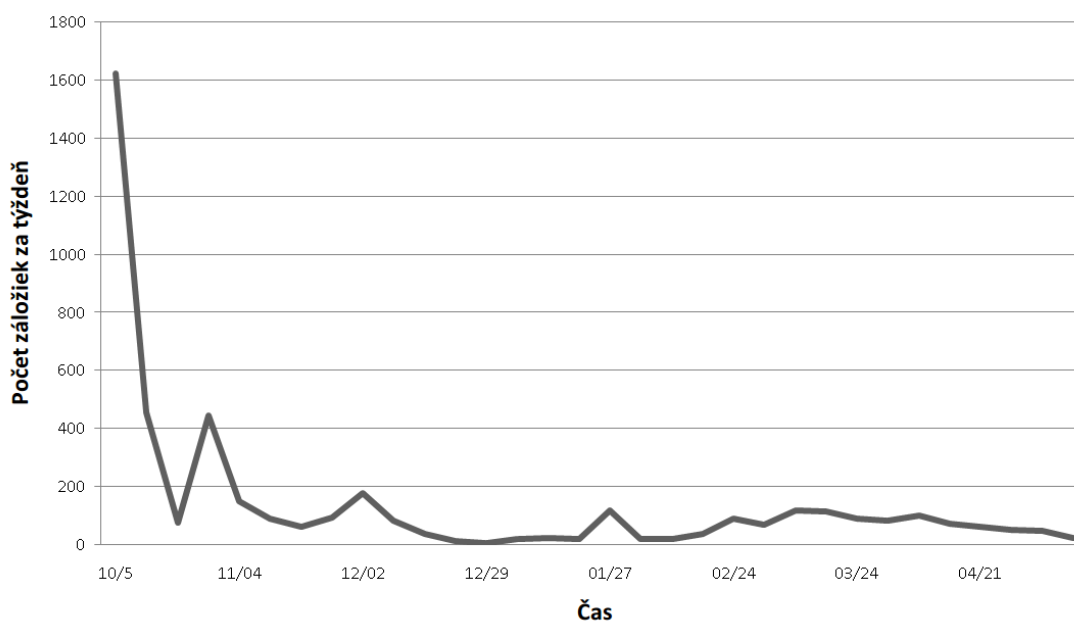


Obr. D.4: Početnosť záložiek podľa počtu priradených tagov.

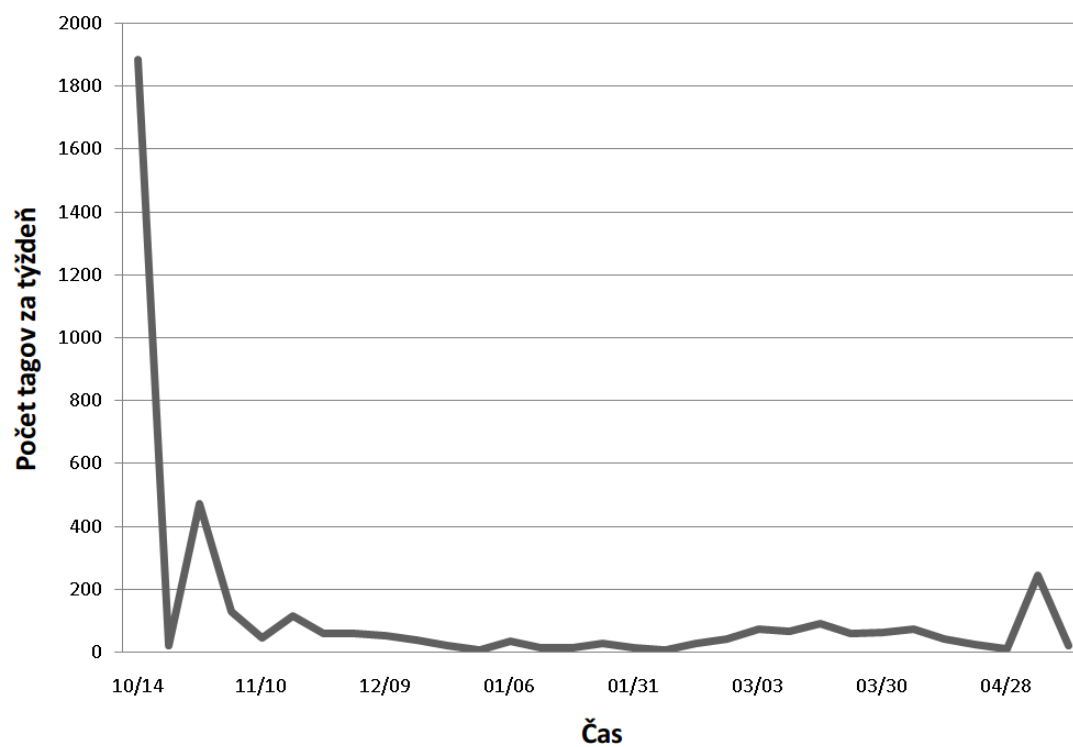
Pribúdanie poznámok, záložiek a tagov v čase je možné vidieť na obrázkoch D.5, D.6 a D.7. V týchto grafoch sú zahrnuté aj záložky a tagy vytvorené prostredníctvom importu údajov zo služieb Delicious a Mendeley. Výrazne vyššie hodnoty v prvých týždňoch používania Annoty sú spôsobené práve údajmi, ktoré si importovali viacerí používatelia. Množstvo vykonanej aktivity sa v čase pomerne výrazne mení, pravdepodobne v dôsledku rôznych udalostí počas roka, ako napríklad odovzdávanie projektov alebo študentská vedecká konferencia, ktoré motivovali používateľov študovať články a vytvárať poznámky.



Obr. D.5: Priebeh aktivity pri pridávaní poznámok v čase.



Obr. D.6: Priebeh aktivity pri vytváraní záložiek v čase.



Obr. D.7: Priebeh aktivity pri pridávaní tagov v čase.



# Parametre použité pri generovaní poznámok v simulácii

Pri spúšťaní simulácie na nájdenie optimálnych parametrov pre metódu na tvorbu dopytu z obsahu dokumentu a k nemu pripojených poznámok sme používali náhodné generovanie poznámok na základe rozdelení získaných z poznámok, ktoré pridávali používatelia v nástroji Annotate. V poznámkach získaných z Annotate sme sledovali niekoľko parametrov a získali sme pravdepodobnostné rozdelenia, s akými sa vyskytovali v reálnych dátach. Sledovali sme:

- Pomer medzi počtom všetkých záložiek a záložiek, ku ktorým bola pripojená poznámka vo forme voľného textu.
- Pomer medzi počtom všetkých záložiek a záložiek, v ktorých bol aspoň jeden kúsok textu zvýraznený.
- Pomer medzi počtom všetkých zvýraznení v texte a zvýraznení, ku ktorým bol pripojený komentár.
- Rozdelenie dĺžok zvýraznených textov.
- Rozdelenie dĺžok poznámok pripojených k záložkám.
- Rozdelenie dĺžok komentárov pripojených k zvýrazneniam v texte.
- Rozdelenie počtu zvýraznení v texte pre používateľa a dokument.

Z poznámok z Annotate sme získali nasledovné údaje:

- Pomer počtu všetkých záložiek a záložiek s poznámkou je 143/103

- Pomer počtu všetkých záložiek a záložiek so zvýraznením je 143/59
- Pomer počtu všetkých zvýraznení a zvýraznení s komentárom je 383/82

Parametre získaných rozdelení sú zhrnuté v tabuľke číslo E.1. Diagramy získaných rozdelení sú zobrazené na obrázkoch E.1, E.2, E.3 a E.4.

| Sledovaná vlastnosť                  | Rozdelenie   | Parametre          |
|--------------------------------------|--------------|--------------------|
| Rozdelenie dĺžok zvýraznených textov | Logaritmické | $\theta = 0.98821$ |
| Rozdelenie dĺžok poznámok            | Geometrické  | $p = 0.11182$      |
| Rozdelenie dĺžky komentárov          | Geometrické  | $p = 0.0671$       |
| Rozdelenie počtu zvýraznení          | Logaritmické | $\theta = 0.85613$ |

Tabuľka E.1: Parametre získaných rozdelení.



Obr. E.1: Rozdelenie dĺžok poznámok pripojených k záložkám.



Obr. E.2: Rozdelenie dĺžok zvýraznení v texte.



Obr. E.3: Rozdelenie dĺžok komentárov pripojených k zvýrazneniam v texte.





Obr. E.4: Rozdelenie počtu zvýraznení v texte podľa používateľa a dokumentu.

# Príspevok publikovaný na konferencii WIKT 2012

V tejto prílohe sa nachádza článok publikovaný na konferencii WIKT 2012<sup>26</sup>.

---

<sup>26</sup>WIKT, <http://wikt2012.fit.stuba.sk/>

# Zaznamenávanie aktivity výskumníka v digitálnej knižnici vedeckých zdrojov obohatené o poznámky

Jakub Ševcech, Mária Bieliková, Roman Burger, Michal Barla  
Fakulta informatiky a informačných technológií, Slovenská technická univerzita  
Ilkovičova 3, 842 16 Bratislava, Slovensko  
sevcech08@student.fiit.stuba.sk, bielik, barla@fiit.stuba.sk,  
roman.arnold.burger@gmail.com

**Abstrakt.** Pre efektívne poskytovanie informácií potrebujeme dobre poznať správanie sa konzumentov informácií. Zber informácií je však náročný aj preto, že základným predpokladom získania vhodnej sady údajov je dobrá motivácia používateľa použiť nástroj, ktorý zaznamenáva aktivitu. V tomto príspevku opisujeme prístup k zaznamenávaniu aktivity výskumníkov na webe, ktorý realizuje zbieranie informácií o navštívených webových stránkach a dopĺňa získané údaje o poznámky, ktoré predstavujú základnú motiváciu pre použitie nášho monitorovacieho nástroja. Pod poznámkami rozumieme záložky, tagy, zvýraznenia v texte a komentáre, ktoré k webovým stránkam (dokumentom) pridávajú samotní návštevníci webových stránok. Na zaznamenávanie aktivity sme vytvorili nástroj Annota, ktorý je realizovaný ako rozšírenie pre internetový prehliadač. Umožňuje pridávať poznámky do zobrazených webových stránok, ale aj PDF dokumentov zobrazených v prehliadači. Vytvorené záložky a poznámky je možné zdieľať so skupinami používateľov. Nástroj je realizovaný všeobecne pre ľubovoľné stránky, pričom však jeho motivačná časť je prispôbená pre digitálnu knižnicu a navigáciu v nej začínajúcim výskumníkom.

**Kľúčové slová.** poznámky, web, tagy

# Príspevok publikovaný na konferencii IIT.SRC 2013

V tejto prílohe sa nachádza článok publikovaný na študentskej vedeckej konferencii IIT.SRC, kde získal cenu dekana.

# Related Documents Search Using User Created Annotations

Jakub ŠEVCECH

Slovak University of Technology in Bratislava  
Faculty of Informatics and Information Technologies  
Ilkovičova 3, 842 16 Bratislava, Slovakia  
sevo\_jakub@yahoo.fr

**Abstract.** We often use various services for creating bookmarks, tags, highlights and other types of annotations while surfing the Internet or when reading electronic documents. These annotations can be used to support navigation, text summarization etc. We proposed a method for searching related documents to currently studied document. Proposed method uses annotations created by the document reader as indicators of user's interest in particular parts of the document. The method is based on spreading activation in text transformed into graph. For evaluation we created a service called Annota, which allows users to insert various types of annotations into web pages and PDF documents displayed in the web browser. We analyzed properties of various types of annotations inserted by users of Annota into documents. Based on these we evaluated our method by simulation and we compared it against commonly used TF-IDF based method.

# Príspevok do konferencie ASIR 2013

V tejto prílohe sa nachádza článok prijatý do konferencie ASIR 2013<sup>27</sup> s výstupom vydaným vydavateľom IEEE Computer Society. Tento článok získal cestový grant na prezentovanie výsledkov na konferencii ASIR.

---

<sup>27</sup>ASIR, <http://fedcsis.org/asir>

# Query Construction for Related Document Search Based on User Annotations

Jakub Ševcech, Mária Bieliková  
Faculty of Informatics and Information Technologies,  
Slovak University of Technology,  
Ilkovičova 3, 842 16 Bratislava, Slovakia  
Email: xsevcechj, maria.bielikova@stuba.sk

**Abstract.** We often use various services for creating bookmarks, tags, highlights and other types of annotations while surfing the Web or just reading electronic documents. These annotations represent additional information on particular information source. We proposed a method for query construction to search for related documents to currently studied document. We use the document content where we concentrate on user created annotations as indicators of user's interest in particular parts of the document. Our method for query construction is based on spreading activation in a graph created from the document content. We evaluated proposed method within a service called Annota, which allows users to insert various types of annotations into web pages and PDF documents displayed in the web browser. We analyzed properties of various types of annotations inserted by users of Annota into documents. Based on these properties, we also performed a simulation to determine optimal parameters and compare proposed method against commonly used tf-idf based method.

# Obsah elektronického média

- **Diplomová práca** - elektronická verzia dokumentu
- **Implementácia**
  - **Server** - zdrojové súbory serverovej časti aplikácie Annota
  - **Rozšírenie** - zdrojové súbory rozšírenia prehliadača
  - **Simulácia** - zdrojové súbory skriptu simulácie na nájdenie optimálnych parametrov metódy na tvorbu dopytu
- **Obraz databázy** - obraz testovacej databázy serverovej časti aplikácie (skutočné údaje niesú priložené s ohľadom na súkromie používateľov služby Annota)