

Michal Laclavík
Martin Šeleng

Vyhľadávanie informácií

Slovenská technická univerzita
v Bratislave
2012

© RNDr. Michal Laclavík, PhD., Mgr. Martin Šeleng, PhD.

Lektori: Prof. Ing. Ján Paralič, PhD.
Mgr. Gabriela Kosková, PhD.

Vydala Slovenská technická univerzita v Bratislave
vo Vydavateľstve STU, Bratislava, Vazovova 5.

Text prešiel jazykovou úpravou vydavateľstva.

Schválilo vedenie Fakulty informatiky a informačných technológií STU v Bratislave
dňa 28.2.2012,
číslo rozhodnutia 2012.06.1 pre študijný program Softvérové inžinierstvo a Informačné
systémy

ANOTÁCIA

Učebný text sa venuje jednej z dôležitých oblastí súčasnej informatiky, ktorou je vyhľadávanie informácií (information retrieval). Zameriava sa na vyhľadávanie informácií na webe, ale tiež na oblasť extrakcie informácií. Diskutuje o základných pojmoch z oblasti vyhľadávania a získavania informácií z internetu a extrakcie informácií, ako aj o základných modeloch pre túto oblasť. Učebný text sa venuje jednotlivým témam v poradí cyklu spracovania informácií: získanie údajov, predspracovanie, indexovanie alebo extrakcia a následné vyhľadávanie. Dôležitou témou je hodnotenie úspešnosti systémov na vyhľadávanie informácií. Učebný text vysvetľuje aj existujúce architektúry a systémy, predovšetkým architektúry vytvorené firmou Google, ako aj voľne dostupné softvérové riešenia, ktoré umožňujú tvorbu takýchto systémov. Príloha obsahuje príklady úloh z vyhľadávania informácií aj s uvedením najčastejších chýb pri ich riešení.

Učebný text je určený študentom informatického zamerania na vysokých školách technického, prírodovedného a ekonomického smeru, najmä však študentom Fakulty informatiky a informačných technológií Slovenskej technickej univerzity v Bratislave, kde je predmet Vyhľadávanie informácií povinne voliteľným predmetom inžinierskeho štúdia pre študijné programy Softvérové inžinierstvo a Informačné systémy.

OBSAH

ANOTÁCIA	V
OBSAH	VII
Slovník	x
Zoznam obrázkov	xi
Zoznam tabuliek	xii
1 ÚVOD	1
1.1 Skrátená história vyhľadávania informácií	2
1.2 O knihe	3
1.2.1 Pre koho je kniha určená	3
1.2.2 Príklady dokumentov	4
1.2.3 Témy knihy	5
1.2.4 Ďalšie témy súvisiace s vyhľadávaním informácií	6
1.2.5 Príklady k jednotlivým témam	8
1.3 Základné pojmy a architektúra vyhľadávacieho systému	8
1.3.1 Informačný priestor, dopyt a výsledok	8
1.3.2 Architektúra	9
2 ZÍSKAVANIE DÁT ALEBO DOKUMENTOV	13
2.1 Techniky sťahovania dokumentov z webu	13
2.2 Sťahovače so zameraním	17
2.3 Spracovanie liniek	17
2.4 Príklady	18
3 ANALÝZA TEXTU, TEXTOVÉ OPERÁCIE	19
3.1 Konverzia dokumentov na text	20
3.2 Segmentácia textu	21

3.3	Tokenizácia textu.....	23
3.4	Konverzia značiek na termy	24
3.4.1	Základný tvar slov	25
3.4.2	Lematizácia a stemovanie v slovenčine.....	26
3.5	Analýza textu pomocou knižnice Lucene.....	27
3.6	Príklady	28
4	MODELY A INDEXOVANIE	29
4.1	Invertovaný index.....	32
4.2	Invertovaný index z web dokumentov.....	36
4.3	Frekvencia termov, váženie a normalizácia.....	38
4.4	Príklady	39
5	VYHĽADÁVANIE A USPORIADANIE	41
5.1	Vyhľadávanie	41
5.2	Usporiadanie.....	43
5.2.1	Podobnosť.....	44
5.2.2	Algoritmus PageRank.....	44
5.2.3	Algoritmus HITS	48
5.2.4	Algoritmus OPIC.....	52
5.2.5	Algoritmus SALSA	56
5.3	Príklady	59
6	REGULÁRNE VÝRAZY	61
6.1	Úvod do regulárnych výrazov	61
6.1.1	Zhrnutie vlastností regulárnych výrazov	65
6.2	Regulárne výrazy v úlohách vyhľadávania a extrakcie informácií.....	66
6.2.1	Sťahovanie dokumentov	66
6.2.2	Predspracovanie textu.....	67
6.2.3	Identifikácia entít	68
6.3	Príklady	71
7	EXTRAKCIA INFORMÁCIÍ.....	73
7.1	Metódy extrakcie informácií	75
7.1.1	Slovníkové metódy.....	75

7.1.2	Metódy založené na pravidlách a vzoroch.....	76
7.1.3	Metódy strojového učenia.....	76
7.2	Vyhodnotenie úspešnosti extrakcie informácií.....	77
7.2.1	Manuálna anotácia.....	77
7.3	Príklady	78
8	HODNOTENIE ÚSPEŠNOSTI.....	79
8.1	Príklady	86
9	SÚČASNÝ INTERNET	87
9.1	Internetový vyhľadávač Google	87
9.2	Škálovateľné riešenia na spracovanie veľkých objemov dát.....	90
9.2.1	Distribučované „súborové“ systémy	90
9.2.2	Distribučované spracovanie dát – Architektúra MapReduce	91
9.3	Príklady	94
10	SOFTVÉROVÉ KNIŽNICE A DÁTOVÉ ZDROJE.....	95
10.1	Softvérové knižnice a systémy	95
10.1.1	Textové operácie.....	95
10.1.2	Lucene a súvisiace projekty.....	96
10.1.3	Extrakcia informácií	98
10.2	Dostupné dátové zdroje v slovenskom jazyku.....	99
11	ZÁVER	101
12	POUŽITÁ LITERATÚRA	103
13	PRÍLOHA - SKÚŠKOVÉ TESTY.....	109
13.1	Skúškový text z roku 2008	109
13.2	Skúškový text z roku 2009	112
13.3	Skúškový text z roku 2010	116
13.4	Skúškový text z roku 2011	120
13.5	Skúškový text z roku 2012	124
INDEX		129

INDEX

Akurátnosť	81	Okapi	2, 31
Analýza textu	19	Ontea.....	99
Anchor text.....	17, 37, 38, 89	OpenNLP	99
BM25	2, 31	OPIC	52
Booleovský model.....	29	PageRank	44, 88
Business Intelligence.....	8	Podniková inteligencia.....	8
Cranfield experiment.....	2, 79	Pravdepodobnostný model.....	31
Crawler.....	13	Precision	79, 80
Data Mining	8, 77	Presnosť	79, 80
Dolovanie dát	8, 77	Recall.....	79, 80
Dopyt.....	9, 41, 80	Regexy	61
Euklidovská vzdialenosť	44	Regulárne výrazy.....	61
Extrakcia informácií.....	68, 73, 75, 98	Rozpoznávanie názvoslovných entít.....	73
Focused Crawler.....	17	SALSA	56
GATE	98	Segmentácia textu.....	21
Gazetteer	75	Sémantický web.....	6
Google.....	43, 87, 89	Solr	97
Hadoop.....	91, 97	Spracovanie prirodzeného jazyka	6
HITS.....	48	Sťahovač	13
Hodnotenie úspešnosti	79	Sťahovač so zameraním.....	17
Identifikácia jazyka	20	Stemovanie	25
Indexovanie	29, 31	Stop slová	24
Informačný priestor	9	Strojový preklad	7
Invertovaný index.....	32, 35, 37	Termy	24
Kontext používateľa.....	7	Text linky.....	17, 37, 38, 88
Kosínusová korelácia	30	Tf-idf.....	38
Lematizácia	25	Tika.....	98
Linky	17	Tokenizácia.....	23
Lucene.....	96	UIMA	98
MapReduce	90, 91, 92	Úplnosť.....	79, 81
Named Entity recognition	73	Usporiadanie.....	43
Natural Language Processing.....	6	Vektorový model	30
Normalizácia	38	Výsledok.....	9, 80
Nutch.....	96	Základný tvar slov	25