

# Obsah

<b>Předmluva k prvnímu vydání</b>	
<b>Předmluva k druhému přepracovanému vydání</b>	
<b>1. Úvod</b>	
1.1 Historie DIS	7
1.2 DIS v kontextu dalších informačních systémů	8
1.3 Model DIS	9
1.4 Vztah DIS k FIS: databázový pohled	10
1.5 Informační služby	12
1.6 Datové struktury	14
1.7 Architektura DIS	15
1.8 Hodnocení efektivity DIS	18
1.9 Specializovaný hardware	20
<b>2. Úvod do datových struktur a algoritmů</b>	
2.1 Regulární výrazy	21
2.1.1 Derivace regulárního výrazu	26
2.2 Konečné automaty	28
2.2.1 Deterministické konečné automaty	29
2.2.2 Nedeterministické konečné automaty	30
2.2.3 Převod nedeterministického konečného automatu na deterministický	30
2.2.4 Konstrukce deterministického konečného automatu k regulárnímu výrazu	31
2.3 Vyhledávací algoritmy pro přesné vyhledávání vzorků v textech	31
2.3.1 Vyhledávání hrubou silou	32
2.3.2 Knuth-Morris-Prattův algoritmus	33
2.3.3 Boyer-Mooreův algoritmus	34
2.3.4 Shift-Or algoritmus	35
2.3.5 Provnání efektivity algoritmů pro přesné vyhledávání jediného vzorku	36
2.3.6 Algoritmus Aho-Corasickové	37
2.3.7 Algoritmus Commentz-Walterové	39
2.3.8 Dvojcestný deterministický konečný automat se skokem	42
2.3.9 Funkce GREP()	44
2.4 Vyhledávací stromy	45
2.4.1 Trie, PAT-stromy a PAT-pole	47
2.4.2 B-stromy	50
2.5 Hašování	51
2.5.1 Otevřené adresování	51
2.5.2 Separátní řetězení	52
2.5.3 d-k funkce	56
2.6 Příklady	61
<b>3. Vyhledávání podle obsahu</b>	
3.1 Invertované soubory	61
3.1.1 Konjunktivní dotaz pomocí invertovaného souboru	61
3.1.2 Konstrukce invertovaného souboru	63
3.1.3 Zipfův zákon	63
3.1.4 Rozšíření invertovaných souborů	63
3.1.5 Implementace invertovaných souborů	65
3.2 Signaturové soubory	65
3.2.1 Metody tvorby signatur	66
3.2.2 Varianty signaturových metod	67
3.3 Příklady	69
<b>4. Lexikální analýza a slovník nevýznamových slov</b>	
4.1 Lexikální analýza při automatickém indexování a zpracování dotazu	70
4.2 Slovník nevýznamových slov	70
4.3 Přístupy k lemmatizaci	73
4.4 Realizace jednoduchého lemmatizátoru	74
4.5 Zpracování přirozeného jazyka	81

<b>5. Modely DIS</b>	89
5.1 <i>Boolovský model</i>	89
5.1.1 Použití tezauru v Boolovských modelech	91
5.1.2 ANSI Common Command Language	92
5.1.3 Kritika Boolovského modelu	93
5.2 <i>Vektorový model</i>	95
5.2.1 Výběr termů	98
5.2.2 Výpočet rozlišovací hodnoty termu	99
5.2.3 Určování vah	100
5.2.4 Implementace vektorového modelu	101
5.2.5 Kritika vektorového modelu	101
5.3 <i>Indexace latentní sémantiky</i>	102
5.3.1 Kritika indexování latentní sémantiky	104
5.4 <i>Rozšířený Boolovský model</i>	104
5.4.1 Fuzzy množiny	105
5.4.2 MMM model	109
5.4.3 Paiceův model	110
5.4.4 Model s měřítkem $p$	111
5.5 <i>Pravděpodobnostní model</i>	112
5.5.1 Binární nezávislostní model	112
5.6 <i>Model shluků dokumentů</i>	114
5.6.1 Metody generování shluků	115
5.6.2 Vyhledávání pomocí shluků	117
5.6.3 Generování shluků pomocí Kohonenových map	118
5.6.4 Generování shluků pomocí sférického K-mean algoritmu	119
5.6.5 Značkování shluků	120
5.7 <i>Uspořádání odpovědi a zpětná vazba</i>	121
5.7.1 Uspořádání odpovědi v indexové organizaci	122
5.7.2 Uspořádání odpovědi ve vektorovém modelu s použitím indexové organizace	124
5.7.3 Uspořádání odpovědi v signaturové organizaci	124
5.7.4 Zpětná vazba	125
5.7.5 Optimální vyhledávání s využitím zpětné vazby	127
5.8 <i>Příklady</i>	128
<b>6. Vyhledávání informací na webu</b>	131
6.1 <i>Dotazování nad webem</i>	132
6.1.1 Dotazování prostřednictvím dotazovacího jazyka	132
6.1.2 Listování strukturou	133
6.2 <i>Typy vyhledávacích strojů</i>	133
6.3 <i>Architektury vyhledávacích strojů</i>	134
6.3.1 Centralizovaná architektura	134
6.3.2 Metavyhledávání	136
6.3.3 Distribuované vyhledávání	137
6.3.4 Webové sklady	137
6.4 <i>Důležitost stránky a její použití pro vyhledávání</i>	137
6.4.1 Analýza odkazů pomocí algoritmu PageRank	137
6.4.2 Analýza odkazů pomocí algoritmu HITS	141
6.4.3 Podobnost obsahu	143
6.4.4 Text související s odkazy	144
6.5 <i>Problémy webových vyhledávacích strojů</i>	144
6.6 <i>K Sémantickému webu</i>	145
6.6.1 Jazyky RDF/S	147
6.6.2 Ontologie	151
6.6.3 Logika	152
6.6.4 Agenti	152
6.6.5 Inteligentní vyhledávací stroje	153
6.7 <i>Závěr</i>	154
6.8 <i>Příklady</i>	154
<b>7. Kompresce v DIS</b>	156
7.1 <i>Základy komprese</i>	156

7.2 Konstrukce Shannon-Fanova a Huffmanova kódu	158
7.3 Algoritmus LZW	161
7.4 Algoritmus HuffWord	166
7.5 Algoritmus WLZW	171
7.6 Komprese bitových řetězců	172
7.6.1 Kódování Huffmanovým kódem	173
7.6.2 Kódování délek běhů	173
7.6.3 b-blokové kódování	173
<b>Literatura</b>	176
<b>Rejstřík</b>	180

V současné době narůstá potřeba zpracování velkého množství informací – novinových článků, odborné literatury, korespondence, agendových spisů, vyhledávacích příkazů z konferencí na počítačových sítích atd., které jsou přístupné prostřednictvím počítače. Pro uživatele je většinou problematické, že neví, kde přesně se nachází pro něj zájmové informace, nebo dokonce ani neví, zda se žadované informace v dostupných textech vůbec nacházejí. Mnozí uvítají o počtu napodoben vyhledávání podle katalogizačních linků v katalogové a následně filtrované vybrané texty. O zpracovávaných dokumentech navíc nelze předpokládat jednotnou sadu pravidel pro strukturu textu. Veniká tedy potřeba nástrojů, umožňujících rychlou a snadnou orientaci v dokumentech, o jejichž struktuře nevíme téměř nic. Přesnější řešení, více pouze to, že se jedná o postřehové slovo.

Systemy pro údržbu a vyhledávání textů budeme nazývat *documentografické informační systémy* (DIS). DIS nelze jednoduše řešit pomocí klasické databázové technologie. Všechny hlavní metody ve světě strukturovaných databází jsou totiž orientovány na práci s formátovanými daty. Cílem skript je seznámit studenty informatiky s tím, jak lze k problematice DIS přistupovat, využít-li se bohatství metod, která informatika nabízí jako základní disciplíny, tj. např. teorie algoritmů, datové struktury, modelování apod. Jméno si vědomi, že jde z jedné strany o otoc – vyhledávací a ukládací informace (Information storage and retrieval), který je speciálním oborem počítačové vědy a komunikace s knihovnictvím. Naš pohled je orientován spíše softwarově orientovaný, tj. jak se DIS konstruuje v prostředí počítačů, jaké metody, algoritmy se k tomu účelu používají.

Pro studium obou směrů je dnes k dispozici titulů (např. [Sa75], [vR79], [SM83], z nových [Jo91], [FY92], [WM1994], [Kor97], [PSV97], [Sch97], [SW97]). Základní a do dnes použitelná učebnice [vR79] je dokonce dostupná na internetu. Z české literatury je možno využít skript [Me94], o souborových technikách pojednávají podrobně skripta [Po97].

V kapitole 1 začínáme základními informacemi o DIS, tj. o jejich architektuře, úkolech, vzniku a systémech faktografickém. Zmíněny jsou i základní používané datové struktury v DIS, hodnocení jejich efektivity. Kapitola 2 poskytuje řadu významných pojmů týkajících se jazyků a sítě. Dvě rozsáhlé třídy datových struktur podporujících vyhledávání zahrnují kvantitativní soubory a soubory signatur. Jim je věnována samostatná kapitola 3. V kapitole 4 je stručně popsána lexikální analýza, lemmatizace a konstrukce slovníka nevýznamových slov. Jádro svého tvoří kapitola 5 diskutující řadu důležitých modelů DIS, jako je Booleanský, vektorový, pravděpodobnostní, model shluků a fuzzy přístup. Kapitola 6 uvádí do problematiky hypertextu, o němž se domníváme, že je také velmi pro pochopení DIS v co nejširším pohledu. Konečně kapitola 7 je věnována speciálnímu otázkám komprese dat v DIS. Bez komprese totiž nelze rozsáhlé DIS úspěšně realizovat. Text je doplněn seznamem literatury a rejstříkem. Materiál předpokládá základní znalosti programování. Na závěr, i když užitečné, je povídání o relační databázové technologii (viz např. v [PH98]).

Autoři děkují Mgr. J. Dvorskému, Mgr. M. Žemličkovi a Mgr. M. Kopeckému za pečlivé přečtení rukopisu textu a za řadu připomínek, které přispěly k jeho zkvalitnění.

V Praze dne 1.7.1998

autoři