

Obsah

Předmluva k prvnímu vydání

Předmluva k druhému přepracovanému vydání

1. Úvod

- 1.1 Historie DIS
- 1.2 DIS v kontextu dalších informačních systémů
- 1.3 Model DIS
- 1.4 Vztah DIS k FIS: databázový pohled
- 1.5 Informační služby
- 1.6 Datové struktury
- 1.7 Architektura DIS
- 1.8 Hodnocení efektivity DIS
- 1.9 Specializovaný hardware

2. Úvod do datových struktur a algoritmů

- 2.1 Regulární výrazy
 - 2.1.1 Derivace regulárního výrazu
- 2.2 Konečné automaty
 - 2.2.1 Deterministické konečné automaty
 - 2.2.2 Neterministické konečné automaty
 - 2.2.3 Převod neterministického konečného automatu na deterministický
 - 2.2.4 Konstrukce deterministického konečného automatu k regulárnímu výrazu
- 2.3 Vyhledávací algoritmy pro přesné vyhledávání vzorků v textech
 - 2.3.1 Vyhledávání hrubou silou
 - 2.3.2 Knuth-Morris-Prattův algoritmus
 - 2.3.3 Boyer-Mooreův algoritmus
 - 2.3.4 Shift-Or algoritmus
 - 2.3.5 Provnání efektivity algoritmů pro přesné vyhledávání jediného vzorku
 - 2.3.6 Algoritmus Aho-Corasickové
 - 2.3.7 Algoritmus Commentz-Walterové
 - 2.3.8 Dvojcestný deterministický konečný automat se skokem
 - 2.3.9 Funkce GREP()
- 2.4 Vyhledávací stromy
 - 2.4.1 Trie, PAT-stromy a PAT-pole
 - 2.4.2 B-stromy
- 2.5 Hašování
 - 2.5.1 Otevřené adresování
 - 2.5.2 Separátní řetězení
 - 2.5.3 d-k funkce
- 2.6 Příklady

3. Vyhledávání podle obsahu

- 3.1 Invertované soubory
 - 3.1.1 Konjunktivní dotaz pomocí invertovaného souboru
 - 3.1.2 Konstrukce invertovaného souboru
 - 3.1.3 Zipfův zákon
 - 3.1.4 Rozšíření invertovaných souborů
 - 3.1.5 Implementace invertovaných souborů
- 3.2 Signaturové soubory
 - 3.2.1 Metody tvorby signatur
 - 3.2.2 Varianty signaturových metod
- 3.3 Příklady

4. Lexikální analýza a slovník nevýznamových slov

- 4.1 Lexikální analýza při automatickém indexování a zpracování dotazu
- 4.2 Slovník nevýznamových slov
- 4.3 Přístupy k lemmatizaci
- 4.4 Realizace jednoduchého lemmatizátoru
- 4.5 Zpracování přirozeného jazyka

5. Modely DIS	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
6. Vyhledávání informací na webu	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
7. Kompresce v DIS	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 Kompresí bitových řetězců	172
7.6.1 Kódování Huffmanovým kódem	173
7.6.2 Kódování délek běhů	173
7.6.3 b-blokové kódování	173
Literatura	176
Rejstřík	180

V současné době nás stále stupně potřeba zpracování velkého množství informací – zejména článků, odborné literatury, korespondence, agenturních zpráv, vyhledávacích záznamů, přípravků z konferencí na počítačových síťech atd., která jsou přístupné prostřednictvím počítače. Pro uživatele je většinou problematické se naučit, kde přesně se nachází pro něj zajímavé informace, nebo jak sama ani neví, zda se k nim informace v dostupných textech vůbec nachází. Navíc učení o počtu napodobení vyhledávání podle katalogových linků v Internetu a následná hledání vytrácejí texty. O zpracování těchto datových souborů navíc běžně předpokládá jednotnou zadávací strukturu. V realitě život ani takto fungujeme toho, že se všechno spíše podřídí stanoveným pravidlům pro sestavení textu. Vzniká tedy potřeba nástrojů, umožňujících rychlé a snadné přehledy v datových souborech, o jejichž struktuře nevíme téměř nic. Převládá řešení, více používat, že se jedná o počítačové sítě.

Systémy pro úložku a vyhledávání textů budeme nazývat datobankografické informační systémy (DIS). DIS nejen jednoduše řeší pomocí klasické databázové technologie. Všechny ostatní možnosti on světě směřované dříve jsou totiž orientovány na práci s formátovanými daty. Čím dříve se začíná studium informatiky a tím, jak lze z prostředí DIS přistupovat, vyhledávat se k nim a metod, která informatika nabízí jako základní disciplína, t. j. např. teorie algoritmů, datové struktury, modelování apod. Jméno a význam, že jde v první řadě o slovo – vyhledávání a ukládání informací (information storage and retrieval), který je speciálním druhem technického – umělého zpracování s kmitočtovými. Náš pohled je orientován spíše softwarově analyticky, abychom zjistili, jak se DIS konstruují v prostředí počítačů, jaké metody algoritmy se k nim třeba používají.

Pro studium obou směrů je dnes k dispozici úlohy [Sa75], [Kr79], [SB83], z nečeských [K91], [FY92], [WB89], [Kor91], [PSV97], [Sch97], [HW97]. Základní a do dnes používaná učebnice [Kr79] je dokonce dostupná na Internetu. Z české literatury je možné využít soubor [M94], o souborových technikách pojednávají postrojnásle skripta [Po97].

V kapitole 1 začínáme základními informacemi o DIS, tj. o jejich architekturu, způsobech, vzniku a systémech faktografických. Závěry jsou i základní potřebné datové struktury v DIS, hodnocení jejich efektivity. Kapitola 2 poskytuje řadu formálních poznámek k jejich seřazení a záměru. Dvě rozsáhlé úlohy datových struktur – slovníkové vyhledávání záznamů inventurního souboru a souborový signatur. Jím je závěrečná samostatná kapitola 3. V kapitole 4 je stručně popsána lexikální analýza, konkrétně a konstrukce slovníku nevyznačených slov. Jde o první dvě kapitole 1 diskutujeme řadu důležitých modelů DIS, jako je Booleovský, vektorový, pro-íterpobíhací, model sítě a fuzzy přístup. Kapitola 6 uvádí do problematiky hypertextu, o němž se diskutujeme. Je to také velmi pro pochopení DIS v co největším počtu. Konkrétně kapitola 7 je věnována speciálnímu úskalí komprese dat v DIS. Bez komprese totiž nešlo rozsáhlé DIS uplatňovat prakticky. Text je doplněn stručným literaturou a rejstříkem. Materiál předpokládá základní znalosti programování. Ne náhodou, i když užitečné, je povědomí o řešení databázové technologie (viz např. v [PH98]).

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