

Obsah

1 Automatické indexování	7
1.1 „Kanonický přístup“ k indexování a vyhledávání v textových databázích a plnotextových systémech	7
1.2 Problém tvarosloví a možnosti jeho řešení	13
1.2.1 Možnosti a meze použití operátoru pravostranného rozšíření	14
1.2.2 Derivátor slovních tvarů	15
1.2.3 Lematizátor slovních tvarů	18
1.2.4 Jiné možnosti řešení problému tvarosloví	22
1.3 Automatické indexování pomocí tezauru	27
1.3.1 Základní představa počítačové reprezentace tezauru jako pomocného nástroje pro indexování	27
1.3.2 Problémy formalizace indexování podle tezauru	31
1.3.3 Tezaurus jako pomůcka při formulaci dotazů	35
1.3.4 Specifická reprezentace některých prvků tezauru v databázovém systému CDS/ISIS	37
1.4 Automatické indexování sémantickým analyzátozem SÉMAN	40
1.5 Vážení prvků selekčního jazyka v selekčních obrazech textů	44
1.6 Automatické indexování maximálně signifikantními termíny z textu — metoda MOZAIKA	51
1.6.1 Základní charakteristika, cíl a způsob použití metody MOZAIKA	51
1.6.2 Základní myšlenky a způsob fungování metody MOZAIKA	51
1.6.3 Omezení, možnosti aplikace a úprav	60
1.7 Shrnutí lingvistických problémů spojených s automatickým indexováním	64
1.7.1 Významnost slov v textu vzhledem k vyhledávání	64
1.7.2 Synonymie a podobné vztahy	65
1.7.3 Homonymie	66
2 Metody porovnávání dokumentů a dotazů založené na měření podobnosti	77
2.1 Základní pojmy a nástroje pro měření podobnosti objektů nesoucíh informací	78
2.1.1 Informační vektor	78

2.1.2	Intuitivní vymezení podobnosti informačních vektorů a jejího využití	79
2.1.3	Matematické míry podobnosti informačních vektorů	79
2.1.4	Problém očíslování selekčních prvků	85
2.2	Měření podobnosti dokumentů a dotazu s využitím indexového souboru	88
2.3	Shlukování dokumentů	90
2.3.1	Idea shlukování a jeho smysl pro zodpovídání dotazů	90
2.3.2	Jednoduchý procedurální přístup: McQuittyho algoritmus shlukování	91
2.3.3	Definice shluku pomocí koheze. „Klasifikační“ algoritmus	92
2.3.4	Idea centroidu	96
2.3.5	Mac Queenův algoritmus shlukování	98
2.3.6	Rocchiův algoritmus shlukování	100
3	Možnosti automatizace tvorby tezauru	105
3.1	Automatizace „ručního“ sestavování tezauru	105
3.2	Automatický výběr lexika pro tezaurus z dané množiny textů	109
3.3	Automatické vyhledávání tezaurových vztahů mezi termíny	112
3.3.1	Porovnávání termínů podle jejich společného výskytu v textech	112
3.3.2	Porovnávání termínů na základě metody SÉMAN (princip systému ATEZ)	114
4	Některé další algoritmizovatelné operace	119
4.1	Automatické referování	119
4.1.1	Automatické referování založené na tezauru	119
4.1.2	Automatické referování založené na měření obsahových souvislostí mezi větami	121
4.2	Strojový překlad	127
4.3	Automatické získávání znalostí z textů	133