

Obsah

1 Automatické indexování	9
1.1 „Kanonický přístup“ k indexování a vyhledávání v textových databázích a plnotextových systémech	9
1.2 Problém rozpoznání (ohodnocení) významnosti jednotlivých výrazů v textu a možnosti jeho řešení	16
1.2.1 Možnosti jednoduchého dichotomického výběru	17
1.2.2 Otázka selektivní síly	18
1.2.3 Metody vážení selekčních znaků	19
1.3 Problém tvarosloví (morfologie) přirozeného jazyka a možnosti jeho řešení	27
1.3.1 Možnosti a meze použití operátoru pravostranného rozšíření	28
1.3.2 Základní pojmy morfologie	29
1.3.3 Derivátor slovních tvarů	31
1.3.4 Lematizátor slovních tvarů	38
1.3.5 Jiné možnosti řešení problému tvarosloví	46
1.4 MOZAIKA — příklad metody automatického indexování maximálně signifikantními termíny z textu	51
1.4.1 Základní charakteristika, cíl a způsob použití metody MOZAIKA	51
1.4.2 Základní myšlenky a způsob fungování metody MOZAIKA	52
1.4.3 Hodnocení, omezení, možnosti úprav	60
1.5 Automatické indexování pomocí tezauru	64
1.5.1 Základní představa tezauru a možnosti jeho počítačové reprezentace	64
1.5.2 Problémy formalizace indexování podle tezauru	69
1.5.3 Tezaurus jako pomůcka při formulaci dotazů	73
1.5.4 Specifická reprezentace některých prvků tezauru v databázovém systému CDS/ISIS	76
1.6 Automatické indexování pomocí sémantického analyzátoru SÉMAN	78
1.7 Shrnutí lingvistických problémů spojených s automatickým indexováním	81
1.7.1 Selekční významnost výrazů přirozeného jazyka v textu . .	81
1.7.2 Gramatická a obecně formální nejednotnost výskytů selekčních znaků v textech	82
1.7.3 Synonymie a podobné vztahy	84

1.7.4	Homonymie	84
2	Metody porovnávání dokumentů a dotazů založené na měření podobnosti	97
2.1	Booleovský model vyhledávání a jeho rozšíření	97
2.2	Základní pojmy a nástroje vektorového modelu vyhledávání	100
2.2.1	Informační vektor	100
2.2.2	Intuitivní vymezení podobnosti informačních vektorů a jejího využití	101
2.2.3	Matematické míry podobnosti informačních vektorů	101
2.2.4	Problém očíslování selekčních znaků	106
2.2.5	Výhoda vektorového modelu vyhledávání v méně rozdílné interpretaci dokumentů a dotazů	108
2.3	Měření podobnosti dokumentů a dotazu s využitím indexového souboru	110
2.4	Shlukování dokumentů	112
2.4.1	Idea shlukování a jeho smysl pro zodpovídání dotazů	112
2.4.2	Jednoduchý procedurální přístup: McQuittyho algoritmus shlukování	113
2.4.3	Definice shluku pomocí koheze: „klasifikační“ algoritmus	114
2.4.4	Idea centroidu	119
2.4.5	Mac Queenův algoritmus shlukování	121
2.4.6	Rocchiův algoritmus shlukování	123
2.4.7	Možnost kombinace booleovského modelu vyhledávání s vybranými prvky shlukování	126
3	Možnosti automatizace tvorby tezauru	131
3.1	Automatizace „ručního“ sestavování tezauru	131
3.2	Automatizovaný výběr lexika pro tezaurus z dané množiny textů	136
3.3	Automatizované vyhledávání tezaurových vztahů mezi termíny	137
3.3.1	Porovnávání termínů podle jejich společného výskytu v textech	137
3.3.2	Porovnávání termínů na základě metody SÉMAN (princip systému ATEZ)	138
4	Některé další algoritmizovatelné operace	145
4.1	Automatické referování (vytváření abstraktů, sumarizace)	145
4.1.1	Co to znamená „referovat“	145
4.1.2	Automatické referování založené na tezauru	147
4.1.3	Automatické referování založené na měření obsahových souvislostí mezi větami	149
4.1.4	Další možné přístupy, příklady a argumenty	154
4.2	Strojový překlad	159
4.3	Automatické získávání znalostí z textů	165