V klasických hypertextoch hrany a uzly určuje väčšinou ručne tvorca systému pri procese *autorizácie* (Shneidermann – Kearsley, 1989, s. 59) čiže pri návrhu typov uzlov, ich rozmiestnenia a spojenia.

3. Automatické generovanie hypertextu

Ideálnym by bol v tomto prípade taký stav, pri ktorom by autor napísal ľubovoľný text a bez akýchkoľvek iných úprav by bol stroj schopný vygenerovať hypertextovú verziu. To zatiaľ zrejme nie je možné, a to z viacerých dôvodov.

Prvým je miera vhodnosti textu na hypertextové spracovanie. Takýmto spôsobom sa spracúvajú predovšetkým vecné texty, ktoré sú pomerne dobre štruktúrované, úplné a logické (Mistrík, 1982, s. 17).

Druhým dôvodom je problém automatickej identifikácie uzlov. Ak sa obmedzíme na väzbu slova alebo slovného spojenia so začiatkom kapitoly, čo samo osebe je dosť výrazné zúženie možností, stojíme pred dvoma čiastkovými problémami: jednak býva niekedy problematické algoritmicky určiť začiatky jednotlivých kapitol pre nejednotnosť noriem a existenciu viacerých spôsobov členenia textu. Ak je kapitola identifikovaná, teda je určený nadpis a text je členený na odseky, potrebujeme priradiť tejto kapitole jedno alebo viac *kľúčových slov*, ktoré sa stanú hodnotami uzla tejto kapitoly. Tomuto problému je venovaná nasledujúca časť príspevku.

Text, v ktorom sú identifikované kapitoly a ich kľúčové slová, je pripravený na *indexáciu* (Shneidermann – Kearsley, 1989, s. 11).

3.1 Priradenie kľúčových slov kapitole

Minimálne nároky, ktoré by systém kládol na formu spracúvaného textu, a tým aj na autora, vyžadujú, aby bol automatický aj proces priradenia kľúčových slov ku každej kapitole. Tu sme však narazili na, zdá sa, najzávažnejší principiálny problém. Úloha by sa dala preformulovať na jednoduchú otázku: Povedzte, *o čom* je táto kapitola, veta, odsek, kniha...?

V prvej verzii systému sme sa pokúšali určiť kľúčové slová podľa počtu výskytov. Vychádzali sme z predpokladu, že kľúčové je to slovo, ktoré sa v texte vyskytuje najčastejšie a je plnovýznamové. Avšak tento predpoklad, a to aj ak zoberieme do úvahy synonymá, nemusí platiť, hoci vo vecných textoch je to obyčajne splnené. Problémom však bola neúnosná redundancia; tých „nepodstatných" – hoci plnovýznamových – slov bolo príliš veľa. Určité riešenie by sa ponúkalo, ak by existoval mechanizmus na generovanie parafráz (Páleš, 1994). Ten však vyžaduje mohutnú lingvistickú podporu a výsledok by zrejme v súčasnosti nezodpovedal potrebám. Preto sme sa rozhodli zaťažiť autora textu a vyžadovať kapitoly ohodnotené kľúčovými slovami.

Aj tak zostáva niekoľko problémov otvorených. V tejto verzii pracujeme iba s jednosmernými hranami bez váhy. Hranou je spojená kapitola so všetkými výskytmi kľúčových slov prislúchajúcich tejto kapitole. Je možné dostať sa od výskytu slova na začiatok príslušnej kapitoly, nie však naopak z kapitoly ku všetkým výskytom jej kľúčových slov v ostatnom texte. Túto funkciu do určitej miery plní register.

Ďalej bolo potrebné vyriešiť otázku, koľko kľúčových slov môže mať jedna kapitola, či to môžu byť viacslovné konštrukcie – a ak áno, upraviť spôsob indexácie; a nakoniec, či môže existovať viac kapitol ohodnotených tým istým kľúčovým slovom. Rozhodli sme sa v tejto verzii neobmedzovať počet kľúčových slov pri jednej kapitole, vyžaduje sa iba zachovanie formalizmu. Kľúčové slová môžu byť aj viacslovné konštrukcie, no v jednom texte sa nemôžu vyskytovať viaceré kapitoly s tým istým kľúčovým slovom. Tieto skutočnosti musí mať autor na zreteli; pri procese indexácie však systém upozorní na chybu, prípadne by mal ponúknuť možnosť opraviť ju.

3.2 Indexácia textu

Cieľom procesu indexácie je určenie hrán v hypertextovej sieti, čiže určenie krížových referencií v hypertextovej databáze. Ide o vyhľadávanie kľúčových slov jednotlivých kapitol vo zvyšnom texte a spojenie týchto slov s danou kapitolou. V tejto fáze algoritmu je možné, ako akýsi „postranný efekt", automaticky vytvoriť obsah textu a register, ktoré sú samozrejme tiež vybavené odkazmi na príslušnú kapitolu.

Ako všade, aj pri indexácii sa vyskytujú niektoré problémy. Azda najzávažnejším z nich je určenie, ktoré tvary slov sa majú vyhľadať k danému kľúčovému slovu. Slovenčina je flektívny jazyk, preto je nevyhnutné vyhľadávať všetky gramatické tvary. Avšak pri slovesách je týchto tvarov až 247 (Páleš, 1994). Niekedy je výhodné okrem gramatických tvarov brať do úvahy aj slová patriace do toho istého slovotvorného hniezda ako kľúčové slovo, a tým by počet vyhľadávaných tvarov prerástol únosné hranice. Použili sme heuristickú a tak trochu „neexaktnú" metódu prahovania, ktorej výsledky sú až prekvapujúco dobré. V kombinácii s klasickým gramatickým postupom by sa výsledok mohol približovať tej úspešnosti, s akou by to robil človek.

3.2.1 Metóda prahovania

Pomocou tejto metódy sme sa pokúsili efektívnejšie ako klasickým pridávaním pádových prípon a alternovaním koreňa zistiť, či sú dva tvary podobné natoľko, aby sa mohli byť považovať za jedno slovo (s jedným lexikálnym významom).