Plnotextová technologie a její aplikace

Richard Bébr

1 Úvodní poznámka

Pokud je čtenář s plnotextovou technologií obeznámen, neřeknou mu následující kapitoly nic nového (jsou určeny pro základní seznámení s problematikou). Znalému čtenáři doporučuji přejít rovnou ke kapitole 6.

Kapitoly 2 až 5 jsou míněny jako rámcový popis a nezabíhají s ohledem na vymezený rozsah příspěvku do hloubky odborných problémů.

2 Základní princip

Plnotextová technologie (fulltext) řeší ukládání a údržbu velkých objemů dat textového typu do počítače. Zavádí zajímavé a uživatelsky velmi efektivní způsoby vyhledávání v textových bázích (pro zajímavost: nalezení všech výskytů libovolného slova nebo sousloví v několika stech MBytech textu netrvá déle než 8 vteřin).

Principy plnotextových technologií byly známy již dříve, avšak teprve výrazné pokroky v oblasti hardware a software umožnily realizovat prakticky využitelné aplikace.

Patrně nejznámější implementaci plnotextové technologie provedla kanadská firma Fulcrum Technologies Inc. (50 % světového trhu v této oblasti); obecný systém plnotextové databanky této firmy má název Ful/Text. V ČR se plnotextovými aplikacemi zabývá např. firma EXPRIT, která na základě licence Fulcrum provedla lokalizaci Ful/Text do českého prostředí a vytvořila i systém TEXPRO, který usnadňuje programování konkrétních aplikací (jakýsi CASE pro fulltext).

Výklad v dalším textu budeme ilustrovat na příkladech, odvozených z Ful/Text a TEXPRO.

3 Srovnání s jinými systémy ukládání dat

Principy plnotextového zpracování pomůže objasnit porovnání klasických databankových systémů (DB) s fulltextem (FT):

Uložení dat:

DB používá přísně strukturovaná data (soubor – věta – položka nebo totéž nazvané jinak). FT ukládá texty nestrukturovaně.

Oživování dat:

DB provádí typické aktualizační transakce typu přírůstek – úbytek – změna. Tyto transakce se provádějí poměrně často, za plného uživatelského provozu a to interaktivně. FT zná v zásadě jen přírůstek a rušení. Báze textů se většinou pouze doplňuje, převážně mimo uživatelský provoz a to dávkově.

Vyhledávání:

V DB je vyhledávání spojeno se strukturalizací: hledáme věty, jejichž určité položky vyhovují jistým požadavkům. Ve FT hledáme spíše asociativně: zjišťujeme, kde a v jaké části textu se vyskytuje určité slovo nebo sousloví (a to i v různých gramatických tvarech).

Dále si ještě porovnejme FT s bibliografickými systémy (BB): BB pracují s popisy textů (klíčová slova, anotace), kdežto FT pracuje přímo s originálním textem.

A teď pozor: FT pouze doplňuje a rozšiřuje možnosti stávajících DB a BB systémů – v žádném případě je nechce nahradit! Aplikační oblasti tedy budou:

DB: Formalizovaně a strukturovaně popsatelné skutečnosti – personální, skladové, evidenční atd databáze. Vyhledávání podle formalizovaných podmínek.

BB: Evidence rozsáhlých textových komplexů (knih, článků atd), kde úplné texty nemůžeme nebo nemusíme uchovávat. Vyhledávání podle sémantiky popisu evidovaných písemností.

FT: Uložení kompletních textů: korespondence, spisy, zprávy, zákony, předpisy apod. Intuitivní vyhledávání podle slov a slovních spojení.

Upozornění: lokalizace do národního prostředí u DB není v řadě případů klíčová, u BB se doporučuje, u FT je zcela nezbytná!

4 Realizace plnotextového systému

V této kapitole popíšeme zásady ukládání a oživování dat. Vyhledávání bude rozebráno v kapitole 5.

Základní jednotkou v plnotextovém systému je **dokument**. Může obsahovat až 16 milionů znaků. **Kolekce** je sada dokumentů s určitou (např. tématickou) příbuzností, může však být tvořena i zcela různorodými dokumenty. Každý dokument může být zařazen do několika kolekcí. Několik kolekcí může být sjednoceno do **metakolekce**, např. pro účely vyhledávání.

V dokumentu mohou být vymezeny **zóny** (tj. souvislé části textu), pro které je možno stanovit určité podmínky (nezobrazovat, neindexovat, ...). Zóny se mohou překrývat. Je možné i vyhledávání ve vymezených zónách.

V databázi je zřízen **katalog**, ve kterém je pro každý dokument jeden strukturovaný záznam, zvaný **profil**. Obsahuje jednak systémové údaje (status dokumentu) a jednak uživatelská data (autor, vydavatel, datum vydání apod); podle těchto posledně jmenovaných dat lze též vyhledávat, status dokumentu je možno zobrazit.

Počet dokumentů v kolekci je omezen pouze možnostmi hardware a operačního systému. Jako maximum se uvádí 32 milionů dokumentů v kolekci. Počet kolekcí není exaktně omezen.

Pořizování dokumentů a jejich ukládání do báze je odlišné od klasických databank. Plnotextové systémy neobsahují žádné prostředky pro přímé interaktivní vkládání a opravy dat. To vyplývá z principu a účelu plnotextových databází, do kterých se vkládají co možno bezchybné hotové texty, které se v průběhu svého uložení nemění. Text dokumentu se pořídí mimo systém některým textovým editorem. Provedou se **kontroly** dokumentu (minimalizace chyb). S výhodou se využívá spell–checkerů a někdy se vyplatí napsat i speciální účelové kontrolní programy. Zkontrolovaný text bývá ještě v některých případech třeba **verifikovat** (potvrdit jeho správnost). Dále se provádí **intelektuální předzpracování** dokumentu; v zásadě jde o vyplnění profilu pro daný dokument. Rozsah intelektuálního předzpracování se řídí účelem a řešením aplikace. Pro jednoduché úlohy je možné získat profilové údaje (při vhodné úpravě dokumentů) i automaticky; pro některé komplikované aplikace (viz např. kap. 6) vytvářejí profily celé týmy vysoce kvalifikovaných specialistů.

Dokument a jeho profil se pak zavedou do systému. Přitom se používá **filtr**, tj. programový prvek, který převede text z formátu editoru na jednotný vnitřní formát textové báze. Dodávané plnotextové systémy obsahují filtry pro všechny důležité editory, další filtry lze doplnit.

Po uložení dokumentů se provádí **indexace**. Systém má vytvořen **slovník**, ve kterém je (vždy jen jednou) uloženo každé slovo, které se vyskytuje v textu dokumentů. Slovník obsahuje odkazy od každého slova ke všem jeho výskytům. Při indexaci se doplňují odkazy na nový dokument, případně i slova, která dosud ve slovníku nebyla. Přitom v systému existuje tzv. **stoplist** = seznam **stop slov**; to jsou slova, podle kterých se nebude vyhledávat (spojky, předložky, zájmena, tvary slovesa „býti" apod) a která se tudíž do slovníku neukládají. Indexační program při své práci stoplist respektuje. Stoplist v základním tvaru je dodáván se systémem a za provozu může být doplňován (např. objevíme-li ve slovníku stopslova, s kterými autor systému nepočítal). Doporučuje se využívat takto vytvořený slovník i při kontrolách pořizovaných textů (mimo plnotextový systém). Zevrubnou kontrolou odhalíme překlepy, ale i nová stopslova.

Indexace je možná i za chodu systému, doporučuje se však provádět indexaci jako dávkový chod v době, kdy uživatelé nejsou připojeni. V dávkách se provádí:

- zavedení nových dokumentů a profilů
- indexace
- pořízení archivních kopií.

Pro tyto účely jsou v plnotextovém systému k dispozici příslušné utility. Texty je možno dle přání uchovávat i v komprimované formě.

Pro zvýšení uživatelského komfortu bývá v plnotextovém systému i **tezaurus synonym**, takže se vyhledávání provádí nejen podle zadaného slova, ale i podle jiných slov stejného významu. Někdy se vytváří slovník homonym (souzvučných slov); pak při zadání výběru podle slova s několika významy je uživatel dotázán, který význam má na mysli.

Fulltext může být doplněn i hierarchickými tezaury, známými z bibliografických systémů.

5 Uživatelská komunikace

Na rozdíl od klasických databank neprovádějí aktualizaci plnotextových systémů běžní uživatelé. Dialog s uživatelem slouží tedy hlavně k formulaci dotazu, prezentaci výsledků hledání a pro ovládání pomocných funkcí (tisk, udržování knihoven dotazů apod.).

Dotaz bývá sestaven ze dvou částí:

- jednak se mohou určit podmínky pro jednotlivé položky profilu – postupuje se jako u klasické DB, neboť jde o strukturovanou informaci
- jednak je možno určit slovo nebo sousloví, které má být v textu vyhledáno.

Dotaz může být položen v jazyku SQL nebo je vytvářen vyplňováním návodných obrazovek.

Pozastavme se nyní u dotazu na slovo nebo sousloví, který má řadu zajímavých možností:

- Uživatel nemusí brát ohled na stop slova. Je možno, ale není nutno je použít.
- Systém dotaz zpracuje gramaticky, určí a hledá i všechny tvary slova nebo sousloví (vytvoření množného čísla, skloňování apod.). Zde je vidět důležitost nejen běžného národního prostředí, ale i co nejlepší národní lingvistiky, která musí být integrální součástí systému.

– Lze použít „divokých" znaků „*" a „?" ve slovech dotazu.

– V dotazu je možno formulovat „vzdálenost" slov v sousloví. Systém pak hledá takové fráze, kde se zadaná slova vyskytují nejvýše v určené vzdálenosti (mezi nimi mohou být i jiná slova).

– Nabízejí se dva druhy hledání:

  — **neuspořádané**, kdy se žádaná slova objevují („uvnitř" zadané vzdálenosti) v libovolném pořadí

  — **uspořádané**, kdy se slova objeví v zadaném rozpětí vzdálenosti v zadaném pořadí.

– Slova a sousloví lze kombinovat pomocí spojek AND, OR, NOT s libovolným závorkováním.

– Číselné a datumové položky mohou být hledány dle přesné hodnoty nebo v zadaném rozpětí hodnot.

Výstupy po úspěšném prohledání jsou upraveny takto:

Neprve se objeví počet dokumentů, vyhovujících zadaným podmínkám. Pak může uživatel volit

– „HIT LIST" = seznam nalezených dokumentů

– **zobrazení profilu** určeného dokumentu

– **zobrazení plného textu** určeného dokumentu.

Hledaná slova nebo sousloví jsou vysvícena. Nabízí se možnosti listování textem, stránkování a dále tisku celých textů nebo zvolených částí (totéž platí i pro profil).

Každá aplikace má svou soustavu menu, pomocí kterých uživatel prochází systémem a volí jednotlivé funkce. Po každém menu může následovat buď menu nižší úrovně nebo návodná obrazovka nebo prezentace výsledků. V aplikaci může být zabudován obecný i kontextový HELP.

Uživatelské dotazy mohou být ukládány do knihoven; existují knihovny obecné, přístupné všem uživatelům a knihovny privátní pro jednotlivé uživatele. Dotazy, uložené ve své knihovně může uživatel vyvolávat, editovat, kombinovat atd. Privátní knihovna dotazů samozřejmě zůstává uživateli k dispozici pro další seance.

6 Příklad aplikace

Nejprve několik základních informací: Národní informační středisko České republiky NIS (pro starší znalce: dříve ÚVTEI) dostalo za úkol vybudovat celostátní systém právních informací (CSPI).

Pro takovéto centrální systémy je neobyčejně výhodný hardware typu mainframe (data jsou soustředěna v jediném místě a tak je velice usnadněna aktualizace; každá úprava báze se ihned projeví u všech uživatelů). Komunikační možnosti mainframe jsou rozsáhlé, stejně tak kapacity pamětí. Pro CSPI byl zvolen mainframe firmy DEC. Pro uložení právních textů je ideální plnotextová technologie a tak byla pro dodávku software vybrána firma EXPRIT.

Náběh zkušebního provozu CSPI se předpokládá od 1.7.1993. V dalším textu uvedeme pouze informativní popis jen těch nejzákladnějších prvků systému s cílem ilustrovat problematiku plnotextové technologie v praxi.

Báze dat CSPI bude obsahovat úplné texty Sbírky zákonů (předpisy publikované i registrované) z let 1945 až po současnost (výhledově budou zařazeny i zákony od r. 1918). Dále se připravují judikáty a perspektivně i relevantní právnická literatura. Uvažuje se také o textech právních dokumentů mezinárodních společenství s případnými překlady do češtiny.

Pro představu: např. kompletní ročník 1992 Sbírky zákonů má textový objem cca 20 MByte. Kapacita dodaných disků DEC je 4 GByte, předpokládá se rozšíření na 20 – 30 GByte. Vnitřní paměť mainframe DEC má nyní 64 MByte.

Právní texty vyžadují mimořádně náročné intelektuální předzpracování. Zejména je nutno v plném rozsahu zachytit derogace a novelizace a to jak aktivní (předpis mění jiný předpis) tak i pasivní (předpis je měněn jiným předpisem).

Protože existují určité cesty, jak získávat texty právních dokumentů v digitalizované formě ještě před vydáním ve Sbírce, bude systém aktualizován do několika dnů (případně i hodin) od vydání příslušné částky Sbírky. To je určitá výhoda proti stávajícím komerčním právním systémům, které většinou nabízejí čtvrtletní oživování.

U právních textů zvláště vyniká požadavek přesnosti, správnosti a naprosté shody s textem, publikovaným ve Sbírce. Pro CSPI jsou vyvíjeny metody, které budou tento požadavek v maximální míře garantovat.

CSPI bude poskytovat informace v několika úrovních – od prostého dotazu až k podkladům pro vědecké bádání v oblasti legislativy. Měla by být k dispozici i tzv. aktuální znění, tedy text právní normy s promítnutím novelizací k určitému datu. Uživatelský dialog je projektován velmi pečlivě s ohledem na nejrůznější typy a úrovně uživatelů.

Základní metodou získání informací z CSPI bude on–line dialog po telefonní nebo datové síti. Uživatelé budou vybaveni terminály nebo mohou použít PC s emulačním programem. Velkou výhodou on–line dialogu je přístup ke zcela aktuálním informacím a v neposlední řadě i platba pouze za dobu pobytu v databázi (nikoliv tedy – jak je zvykem u dodávaných komerčních systémů – nákladný nákup systému a pravidelné platby za aktualizaci, přičemž tyto finanční položky nejsou závislé na využívání systému).

Počítá se i s napojením na spojovou službu VIDEOTEX (takže každý účastník této služby by měl dostup i do systému CSPI); zatím však jsou příslušné spojové organizace velmi nekomunikativní a vyžadují za propojení na VIDEOTEX nehorázné finanční částky. NIS bude perspektivně vydávat i CD ROM s právními texty a obslužnými programy. Již v současné době prodává NIS zájemcům kompletní texty Sbírky zákonů (roč. 1990 – 1993, nabídka se průběžně rozšiřuje) na disketách ve formátu WP nebo T602. Tyto texty jsou pořizovány v předstihu pro vložení do CSPI.

7 Závěr

Plnotextová technologie se zařadila po bok databankovým a bibliografickým systémům. Je určena pro práci s rozsáhlými úplnými texty a má některé zvláštní prostředky pro vyhledávání. Plnotextová technologie musí být perfektně lokalizována do národního prostředí a musí mít co nejdokonalejší národní linguistiku.

V ČR již existuje řada aplikací (např. systém TEXPRO má realizováno více než 160 prodejů pro orgány státní správy, výzkumná a informační střediska apod.). Pro širší veřejnost může být zajímavá aplikace CSPI, popsaná výše.

Jak jsme již ukázali, závažný problém pro plnotextovou technologii představuje pořizování dat, neboť jde jednak o velké objemy a jednak je nutná co nejpřesnější shoda dat s originálním textem. S rozvojem výpočetní a organizační techniky a technologie se však nabízejí např. tato řešení:

- vlastní dopisy a spisy se dnes pořizují převážně na textových editorech, takže mohou být přímo přenášeny do plnotextových databází
- příchozí texty, přijímané počítačem s faxmodemovou kartou jsou digitalizované a připravené pro zápis do báze
- přímý zápis je rovněž možný u korespondence E-mail
- některé písemnosti je možno převést do digitální podoby scanováním; tato technika (zejména její software) se v poslední době prudce rozvíjí.

Jak je vidět, lze při vhodném vybavení a dobré organizaci práce závažně racionalizovat získávání dat pro textové databáze.

Plnotextová technologie se stále vyvíjí, její aplikace se množí a ve věku informací se stává důležitým pracovním nástrojem. Brzy se přímo či nepřímo dotkne každého z nás.

Literatura

O plnotextové technologii existuje bohatá literatura. Vážným zájemcům doporučuji využít služeb Národního informačního střediska, které má přímé propojení do mnoha světových bibliografických databank a může zajistit anotace literatury na libovolné téma v libovolném rozsahu.

Autor: Ing. Richard Bébr
Nádražní 80
370 01 České Budějovice

Pracoviště: Národní informační středisko
Havelkova 22
130 00 Praha 3
tel.: (02) 235 05 88 linka 324 (nebo 323)
fax : (02) 235 97 88