Dokumentografické informační systémy

D.Hůsek[a], V.Sklenář[b], V.Snášel[b]

[a]Institute of Computer Science,Acad of Sci. of Czech Republic, Pod vodárenskou věží 2, 180 00 Praha 8, Česká republika

[b]Computer Science Department, Palacky University of Olomouc, Tomkova 40, 779 00 Olomouc,Česká republika

Abstrakt

Současná doba je charakterizována prudkým nárůstem potřeby a vyhodnocení informací. Vývoj a zdokonalování dokumentografických informačních systémů je stále předmětem výzkumu. Mezi základní problémy patří řešení vyhledávacího problému, tj. k uživatelskému požadavku přiřadit odpovídající dokumenty. V tomto článku jsou diskutovány základní algoritmy, struktury a specializovaný hardware pro konstrukci dokumentografických informačních systémů.

1. Úvod

Motto

Proroctví o informačním průmyslu koncem devadesátých let jsou mnohá, ale nejistá, neboť není dosud známo, jak ovlivní pátá generace, porozumění textu, rozvoj znalostního inženýrství, technicko-ekonomické parametry hardware a služeb, vliv velkých projektů, sociální reakce a jaké budou přitom brzdné momenty. Nicméně se zdá, že Gutenbergovská éra papíru jako monopolního nosiče informací končí. Vyčkejme tedy - aniž složíme však ruce v klín - jistě se o tom dočteme v přemnohých textech.

M.Bloch, A.Scheber. Textové báze dát. Sofsem 86.

V současné době neustále stoupá potřeba zpracování velkého množství informací - novinových článků, odborné literatury, korespondence, agenturních zpráv, vyhlášek, zákonů, příspěvků z konferencí na počítačových sítích atd., které jsou přístupné v textovém tvaru (dále budeme používat pojem text). Jedná se většinou o stovky nebo tisíce stran obsahujících nejrůznější informace, z nichž však je pro konkrétního čtenáře v mnoha případech zajímavý pouze nepatrný zlomek. Ostatní stránky textu nejsou pro něj potřebné, protože obsahují informace, které již zná a nebo nepatří do oblasti jeho zájmu. Na první pohled je tedy všechno jednoduché. Během několika málo minut se čtenář seznámí s tím, co ho zajímá a zbytkem se vůbec nezabývá. Problém ale spočívá v tom, že většinou neví, kde přesně se pro něj zajímavé informace nacházejí nebo dokonce ani neví, zda se vůbec žádané informace v dostupných textech nacházejí.

Potřeba počítačového zpracování textů je tedy velmi aktuální. To lze dokumentovat i na tom, jak uváději některé statistiky, že vyhledávání informací v takovýchto dokumentech zabere určeným pracovníkům až 2 hodiny pracovního času denně.

Cílem článku je popsat některé současné používané modely textových databází. V úvodní kapitole je uvedena stručná historie této problematiky. Ve druhé kapitole je definován dokumentografický informační systém. Ve třetí kapitole je rozebráno obecné schéma DIS. Pátá kapitola se zabývá přesností a úplností DIS. V šesté kapitole jsou uvedeny některé možnosti hardwarového řešení předkládané problematiky. V sedmé kapitole je uvedeno stručné porovnání DIS a Hypertextu. Závěr hodnotí současný vývoj a perspektivy.

2. Historie

Motto

Jak se budeš Sokrate ptát na to , co neznáš...
A když nalezneš, co hledáš, jak budeš vědět, že je to ta věc,
kterou neznáš...
Člověk se nemůže ptát na to, co zná nebo na to, co nezná:
protože, když to zná, nemusí se ptát a když to nezná, pak se na
to nemůže ptát, protože nezná přesně subjekt, na který se má
ptát.

Meno, Plato

Historie základních idei v oblasti vyhledávání informací je těsně svázána s vývojem výpočetní techniky, a to v obou oblastech - jak programového vybavení, tak v oblasti hardware. Fairthorne uvažoval o možnosti použití Holleritových strojů na děrné štítky v oblasti vyhledávání informací již v r. 1953.

Nicméně největší vliv na vývoj textových informačních systémů měly před více než 30 lety (v roce 1962) experimenty se systémem Cranfield I, viz C.W.Cleverdon [3]. Tento systém obsahoval 1400 dokumentů a 255 dotazů. Indexování se v tomto systému provádělo ručně. V roce 1966 byl vyvinut systém Cranfield II s automatickým indexováním.

Tyto práce pak inspirovaly takové veličiny jako jsou Salton, Lancester, Sparck Jones, C.J. van Rijsbergen a mnoho dalších. Cleverdon a jeho následníci zdůrazňovali experimentální ověřování teoretických poznatků týkajících vyhledávání informací v plných textech. Naopak teoretičtější přístup byl prosazován v USA nestory výpočetní techniky jako jsou Maron, Kuhns a Cooper. Později se tyto dva přístupy sešly na platformě pravděpodobnostního modelu.

První systémy byly založeny na sekvenčním prohledávání. Současný stav problematiky se nejlépe odráží na jednáních následujících konferencí.Konference, která je specializována na

teoretickou problematiku textových informačních systémů je ACM SIGIR Conference on Research and Development in Information Retrieval. Tato konference se soustavně zabývá problematikou která se v anglosaském světě nazývá information retrieval. Tato konference se střídavě koná v USA a v Evropě. Konference, která se specializuje na hodnocení efektivnosti textových informačních systémů, je konference TREC (Text Retrival Conference). Testovací soubor, který je základem pro hodnocení systémů pro zpracování plných textů obsahuje jeden milion ohodnocených dokumentů. Tzn., že na každý dotaz je známa správná množina dokumentů, která mu odpovídá [10]. Na konferencích [6], [7] a [8] bylo každoročně prezentováno několik desítek textových informačních systémů.

3. Pojem Dokumentografického informačního systému.

3.1 Dokument

Je zřejmé, že nás nebude zajímat pouze část textu vytržená z kontextu, ale ucelený text obsahující hledanou informaci. Takovýto text budeme dále nazývat dokumentem.

Tento pojem bude základem pro všechny činnosti související se zpracováním textových informací.

Na dokument se můžeme dívat ze dvou pohledů:
- Fyzický pohled určuje způsob uložení dokumentu na médiu.
- Logický pohled je dán informacemi, které jsou v dokumentu obsaženy.

Logický dokument nemusí odpovídat jednomu fyzickému dokumentu. Například kniha jako jeden logický celek může být rozdělena do více fyzických souborů.

3.2 Dokumentografické informační systémy

Dokumentografické informační systémy jsou třída programových nástrojů, určených pro zpracovávání, úschovu a výběr dokumentů.

Základním rozdílem mezi dokumentografickými a faktografickými informačními systémy je strukturovanost vkládaných dat. Faktografické systémy pracují s daty, majícími pevnou, předem danou strukturu, ve které každá položka má předem daný význam. Příkladem faktografických systémů jsou v současné době rozšířené relační databázové systémy.

Dokumentografické informační systémy (DIS) naproti tomu pracují s daty, která jsou ve své podstatě strukturována jen málo nebo vůbec ne. Základním prvkem dat v těchto systémech je text v přirozeném jazyce. Může se jednat o zákony, o knihy, o časopisy, o úřední spisy, o diplomové práce stejně jako o výzkumné zprávy, shrnující výsledky vědeckých pokusů. Například Institut pro standardizaci a technologii USA má k dispozici více než 2 miliony textových dokumentů, viz [7]. Po vzoru faktografických informačních systémů budeme na uložená textová data nahlížet jako na (textovou) databázi.

Nízká strukturovanost uchovávaných dat a použití přirozeného jazyka přináší nutnost vyřešit při tvorbě DIS i řadu jazykových problémů.

4. Obecné schéma DIS

Vstupní texty jsou po předzpracování zahrnuty do textové databáze, což je struktura obsahující informace o textech v podobě vhodné k aplikaci vyhledávacích technik. Uživatel klade dotaz, který je následně podroben analýze syntaktické a sémantické, pak je transformován a stává se vstupem pro algoritmus vyhodnocení dotazu. Výstupem tohoto algoritmu je nějaká množina záznamů, kterou je třeba dále zpracovat (seřazení, ohodnocení, dohledání vzorků atd. ...). Výsledné dokumenty jsou pak předloženy uživateli, který se rozhodne, zda se s dosavadní odpovědí spokojí, či zda bude pokračovat dalším, vylepšeným, dotazem, pomocí zpětné vazby. Vstupem algoritmu pro konstrukci zpětnovazebního dotazu je výsledek předchozího vyhledávání spolu s údajem uživatele, které vybrané dokumenty považuje za relevantní.

DIS se skládá z několika spolupracujících komponent, viz [17]. Ne v každé implementaci se vyskytují všechny zde popsané části. Zdokonalování všech těchto komponent je stále ještě předmětem výzkumů. Kvalita jednotlivých prvků použitých v konkrétním systému potom určuje výslednou kvalitu, tedy míru uspokojení uživatele či uživatelů.

Jednou z komponent DIS může být vstupní textový filtr, který provádí lingvistickou analýzu čteného textu a převádí přečtené lexikální jednotky textu (slova) na základní tvar. Nazývá se často lematizátor.

V existujících jazycích (český jazyk je v tomto směru velmi bohatý) má většina slov mnoho tvarů lišících se podle rodu, pádu, jednotného či množného čísla. Jindy může mít jedno slovo několik odlišných významů (např. let = rok, létat).

Z našich testů vyplývá, že před indexováním nemá význam upravovat vstupní text filtrací. To proto, že filtrací dochází ke ztrátě informace a u velmi rozsáhlých textových databází nedojde ani k očekávané úspoře rozsahu indexů. Podstatně lepší výsledky lze dosáhnout kompresí textového dokumentu, viz [12] a kompresí indexů.

Další komponentou v klasických DIS je indexační jednotka. Tato komponenta má za úkol obohatit ukládané texty o doplňující informace, které umožní efektivní vyhledávání. V této fázi zpracování se ke každému textovému dokumentu doplní jeho počítačová reprezentace, která se nazývá záznam dokumentu.

Záznam obsahuje formální popis dokumentu, skládající se z hodnot vhodně specifikovaných atributů (položek), a z množiny termů, které ve stručné podobě vystihují obsah plného znění dokumentu. Vzhledem k nejednotné terminologii budeme pod termem chápat jistý vzorek textu (výraz), který může být víceslovný nebo také jednoslovný. Jednoslovným (ale mnohdy i víceslovným) termům se také někdy říká klíčová slova, místo o termech se také hovoří o deskriptorech.

Nalezení vhodné množiny termů je v obecnosti velmi náročná úloha, která v mezním případě vyžaduje porozumění sémantickému významu textu. Termy vybrané během indexace musí dostatečně přesně reprezentovat obsah dokumentu a také dát do souvislosti dokumenty týkající se podobného tématu. Přitom obecné termy, vyskytující se ve všech, resp. skoro ve všech dokumentech, nemají pro účel vyhledávání téměř žádný význam. Malý význam mají také ty termy, které se vyskytují v příliš malém počtu dokumentů. Tento problém se řeší pomocí StopListu. StopList je seznam slov, která se při indexování a dalším zpracování textu ignorují. Konstrukce stoplistu je uvedena v [5] a [9].

Proces přiřazení množiny termů dokumentu - indexace dokumentu - se proto v mnoha systémech provádí buď ručně nebo poloautomaticky. V prvním případě provede specialista v daném oboru - indexátor - sám výběr nejvhodnějších termů, v druhém případě systém poskytuje možnost upravit množinu termů, vytvořenou systémem na základě analýzy plného textu.

V [1] je popsán pokus, kdy 8 indexátorů mělo pro popis dokumentu vybírat ze 14 deskriptorů. Ukázalo se, že žádný deskriptor nebyl vybrán všemi, a žádní dva indexátoři nepoužili stejnou množinu deskriptorů.

4.1 Metody vyhledávání

Při vyhledávání v dokumentech se vychází ze slov, která charakterizují jeho obsah. Tato slova mohou být stanovena uměle, např. autorem dokumentu a mají pak podobu klíčových slov. To však přináší několik problémů. Například nemusí být jasné, kolik slov dostatečně charakterizuje obsah dokumentu. Volba klíčových slov je navíc značně závislá na pohledu jejich tvůrce.

Přesnější je tedy, když pracujeme s celým dokumentem a uvažujeme všechna slova, která jsou v něm uvedena. Ani to samozřejmě nezaručuje, zaznamenání všech informací obsažených v dokumentu. Jeden logický pojem můžeme vyjádřit pomocí různých slov. Hledáme-li například dokumenty obsahující informaci o kopané, tak nás budou zajímat i dokumenty, ve kterých se slovo kopaná vůbec nevyskytuje, např. dokument obsahující slovo fotbal. Je celkem pravděpodobné, že nás mohou zaujmout i dokumenty obsahující slova Sparta, Slavie, FIFA, penalta, ...Všechna tato slova totiž mohou (ale také nemusí) souviset s naším dotazem.

V DIS lze pak formulovat základní vyhledávací problém:
Nalézt k uživatelskému požadavku - dotazu - relevantní dokumenty.

Mezi problémy, souvisejícími s vyhledávacím problémem patří zejména:

- jak určit, co je relevantní a co ne,
- jak zajistit efektivnost zpracování,
- jak zajistit uspořádání výstupů podle relevance.

Je zřejmé, že řešení vyhledávacího problému vyžaduje další komponentu DIS - vyhledávací stroj. Tato komponenta využívá indexů a vybírá z textové databáze dokumenty, které vyhovují dotazu zadanému uživatelem.

Vyhodnocení dotazu spočívá většinou v porovnání termů uvedených uživatelem v dotazu s popisy, které specifikují jednotlivé dokumenty.

Indexace ovšem nemusí být vyjádřena pomocí termů. Obsah dokumentu lze vyjádřit i jinak, např. jistým zakódováním textu do podstatně kratšího řetězce znaků - signatury. Jinou možností může být ocenění termů popisujících dokument čísly (váhami) vyjadřujícími důležitost termu v dokumentu. Využití lingvistiky může znamenat např. využití jistých relací mezi termy. Takové relace tvoří další pomocné struktury dat (tezaury), které mohou zaručit výběr takových dokumentů, jejichž termy jsou jiné, než ty zadané v dotazu, a přesto je výsledek relevantní.

Zobecníme-li tuto úvahu, potřebujeme model textové databáze. Bude to soubor pojmů a nástrojů umožňujících popsat textovou databázi a formulovat základní vyhledávací algoritmy umožňující řešit vyhledávací problém.

Ze softwarově inženýrského hlediska nesmíme zapomenout na komponentu uživatelské rozhraní. Ta komunikuje s uživatelem a nabízí mu možnost pokládat dotazy informačnímu systému za pomoci dotazovacího jazyka. Zatímco v oblasti faktografických informačních systémů existují standardy pro komunikaci (nejznámějším dotazovacím jazykem je SQL), kterými se výrobci řídí, standardizace dotazovacích jazyků v textových databázích je teprve v začátcích a každý produkt obsahuje vlastní způsob formulace dotazů.

4.1.1 Booleovský model

Ve většině případů jsou dotazy formulovány pomocí přesně definovaného formálního jazyka (založeného většinou na Booleově algebře). Booleovská metoda dotazování je nejstarší a také nejrozšířenější způsob formulace dotazu uživatelem. V tomto modelu je každý dokument spojen s množinou termů, kterými je charakterizován a dotaz je booleovský výraz (složen z termů, logických operací AND, OR, NOT a uzávorkování). Z databáze jsou vybrány ty dokumenty, které obsahují hledané termy v kombinaci určené dotazem. Vyhodnocování těchto dotazů je založeno na vyhledávání v tzv. invertovaných seznamech (pro každý term máme uspořádaný seznam dokumentů, ve kterých se vyskytuje).

Uvedeme popis operací pro booleovský model.

X AND Y Výběr dokumentů, obsahujících jak term X, tak term Y.

X OR Y Výběr dokumentů, obsahujících buď term X, nebo term Y, nebo oba termy současně.

X XOR Y Výběr dokumentů, obsahujících buď term X, nebo term Y, ale ne oba současně.

X NOT Y Výběr dokumentů, obsahujících term X, ale ne term Y.

X ADJ Y (adjacent) - Výběr dokumentů, ve kterých se vyskytuje term X následovaný termem Y

X WORDS(n) Y Výběr dokumentů, ve kterých se vyskytuje term X následovaný termem Y nejdále ve vzdálenosti n slov.

Tento model se velice snadno implementuje a je velice efektivní z hlediska časové náročnosti na vyhodnocení dotazu.

Nevýhody klasického booleovského modelu:

- Efektivně formulovat dotaz je značně obtížné, vyžaduje tourčité zkušenosti.
- Klasická booleovská metoda neumožňuje vyjádřit míru relevance jednotlivých vybraných dokumentů vzhledem k dotazu Považuje všechny nalezené dokumenty za stejně dobré.
- Není zde možnost řídit velikost výstupu. Snadno nastávají extrémní případy, kdy výstup je prázdný nebo naopak obsahuje veliké množství dokumentů.
- Standardní booleovská metoda chápe všechny termy v dotazu jako stejně důležité.
- Výsledky dotazu neodpovídají intuitivní představě uživatele:
  - v disjunktivních dotazech se na výstupu objeví dokumenty obsahující i jen jediný term vedle záznamů obsahujících všechny nebo většinu termů
  - konjunktivních dotazech jsou z výstupu vyloučeny dokumenty neobsahující i jen jediný term, stejně jako dokumenty neobsahující žádný z nich.

4.1.2 Vektorový model

Vektorová reprezentace dat je asi o dvacet let mladší než booleovský model. Dokumenty charakterizujeme pomocí vektorů typu $D = w_{i1}, w_{i2}, \ldots, w_{in}$ kde n je počet termů, kterými charakterizujeme dokumenty v databázi a w_{ij} jsou váhy těchto termů vztažené k dokumentu D.

Stejně tak dotaz lze vyjádřit jako vektor $Q = q_1, q_2, \ldots, q_n$ kde čísla q_i vyjadřují, do jaké míry je pro tazatele výskyt termu v dokumentu žádoucí.

Pak můžeme určit koeficient podobnosti dotazu a dokumentu, sim(D, Q), který vyjadřuje odhad míry relevance dokumentu. Vyhodnocení dotazu spočívá v seřazení dokumentů podle hodnot koeficientu podobnosti.

Tato metoda umožňuje triviálně řídit velikost výstupu. Její nevýhodou je ztráta možnosti strukturalizace dotazu. Ztrácí se zde rozdíl mezi AND a OR dotazy. Přesto se v praktickém provozu systémy založené na vektorovém modelu osvědčují lépe než booleovské.

4.1.3 Rozšířená booleovská logika

V těchto modelech předpokládáme, že dokument má přiřazenu nějakou váhu pro každý indexovaný term. Tato váha vyjadřuje míru, jakou je tento dokument termem charakterizován. Bez újmy na obecnosti můžeme předpokládat, že váhy leží v intervalu [0,1]. V klasickém booleovském modelu jsme ještě více omezeni a to pouze na dvě hodnoty 0 a 1. Vyhledání dokumentu podle dotazu, znamená vypočítat koeficient podobnosti dotazu a dokumentu. Uvedeme si tři metody výpočtu tohoto koeficientu.

4.1.3.1 MMM model

Tento model je založen na teorii fuzzy množin, viz [14]. V této teorii prvek má proměnlivou míru, d(A) příslušnosti k dané množině A, místo klasického je nebo není prvkem množiny.

V Mixed Min a Max (MMM) modelu je každý term spojen s jistou fuzzy množinou. Váha dokumentu vzhledem k termu A se bere jako míra příslušnosti dokumentu do fuzzy množiny termu A.

Podle teorie fuzzy množin dokument, který je vybrán dotazem A or B, musí náležet do sjednocení fuzzy množin A a B. Podobně dokument, který je vybrán dotazem A and B, musí náležet do průniku fuzzy množin A a B. Je proto možné definovat koeficient podobnosti dokumentu a dotazu or jako max(d(A),d(B)) a obdobně u dotazu and jako min(d(A),d(B)). MMM model počítá tento koeficient podobnosti jako lineární kombinaci min a max vah dokumentů.

Obecně lze říci, že výpočetní nároky tohoto modelu jsou malé a efektivnost vyhledávání dobrá, mnohem lepší než u klasického booleovského modelu.

4.1.3.2 Paice model

V roce 1984 uveřejnil Paice model založený také na teorii fuzzy množin. I v tomto modelu je každý term spojen s jistou fuzzy množinou. Váha dokumentu vzhledem k termu A se bere jako míra příslušnosti dokumentu do fuzzy množiny termu A. Na rozdíl od MMM modelu se do výpočtu koeficientu podobnosti berou v úvahu všechny váhy termů, nikoliv jen jejich maxima a minima.

Složitost výpočtů je vyšší než u MMM modelu. Pro MMM model stačí zjistit pouze hodnoty min a max z množiny termů, což je možno provést v čase O(n). Paice model vyžaduje setřídit váhy termů. Toto vyžaduje čas nejméně O(nlogn). Také je zde větší množství výpočtů v reálné aritmetice.

4.1.3.3 P-norm model

P-norm model bere v úvahu mimo vah termů v dokumentu také váhy termů v dotazech. Četné experimenty prokázaly, že tento model je velice přesný. Nevýhodou však zůstává jeho vysoká výpočetní složitost.

5. Přesnost a úplnost

Efektivita vyhledávacích systémů je dána jednak rychlostí zpracování požadavku a uživatelským komfortem (projevujícím se zejména ve formě interakce se systémem, ve způsobu kladení dotazů a poskytování odpovědi), hlavně však schopností systému poskytnout informaci o relevantních dokumentech. Míra této schopnosti se vyjadřuje pomocí dvou ukazatelů, koeficientu přesnosti, P (z angl. precision), a koeficientu úplnosti, R (z angl. recall). Koeficient přesnosti bývá též nazýván poměr relevance. Označme:

R_Q počet vybraných relevantních dokumentů
R_V počet všech relevantních dokumentů v kolekci
A_Q počet všech vybraných dokumentů

Pak definujeme:
$R = R_Q / R_V$ a
$P = R_Q / A_Q$

Koeficient úplnosti lze chápat jako pravděpodobnost, že relevantní dokument byl vybrán, koeficient přesnosti jako pravděpodobnost, že vybraný dokument je relevantní.

Má-li být řešen kompromis mezi R a P, uživatelé volí obvykle jako cíl vyšší P [16]. Jsou tak zbaveni nepříjemné nutnosti prohledávání eventuálního velkého množství nerelevantního materiálu.

Koeficienty R a P zřejmě nejsou na sobě zcela nezávislé. Z definice R a P vyplývá, že při velmi úzce specifikovaném dotazu, kdy obdržíme málo nerelevantních dokumentů, ale také málo relevantních (relativně z celkového množství) dostaneme vysoké P, ale nízké R, a naopak, pokud je specifikace dotazu širší, obdržíme více relevantních dokumentů, tj. zvětší se R, ale také se zvýší počet vybraných nerelevantních, a tudíž klesne hodnota P.

Zkušenost ukazuje, že při výběru dokumentů v reálných systémech jsou tyto koeficienty ve vztahu nepřímé úměrnosti. Pokusy činěné na fungujících DIS ukázaly, že vyjdeme-li z dotazu, vyznačujícího se po vyhodnocení vyšším R a nízkým P, pak variováním dotazu s cílem zvýšit P, dosáhneme tohoto zvýšení pouze za cenu nižšího R (citováno podle [16]).

Existují metody jak zvyšovat hodnoty parametru R nebo P. Zvýšení R lze docílit rozšířením dotazu tak, že termy dotazu se nahradí obecnějšími výrazy, naopak zvýšení P vyžaduje upřesnění dotazu (viz [16], podrobně [17]). Tyto požadavky jdou však zřejmě proti sobě.

6. Specializovaný hardware

Hardware má vliv na návrh DIS, protože určuje alespoň částečně operační rychlost DIS, což je klíčový faktor interaktivních informačních systémů, a množství a typ informace, která může být prakticky uložena v informačním systému. Další podrobnosti o specializovaném hardware je možno najít v [15]. Většina DIS, které se v současné

době používají, je implementována na von Neumannovských počítačích, tzn. na universálních počítačích s jedním procesorem. Většina technik a algoritmů, které jsou v tomto článku diskutovány, implicitně předpokládá von Neumannovský stroj jako implementační platformu.

V posledních letech enormně vzrostla výpočetní rychlost těchto počítačů, ale stále existují aplikace z oblasti vyhledávání informací, pro které jsou tyto počítače stále příliš pomalé. Jako odpověď na tento problém někteří výzkumníci zkoumali alternativní architektury pro implementaci DIS. Existují dva základní přístupy - paralelní počítače a specializovaný hardware.

Zajímavá je implementace DIS na Connection machine, což je masivně paralelní počítač s 64000 procesory.

Specializovaný hardware pro DIS znamená počítače speciálně navržené k vykonávání operací specifických pro DIS.

Jako příklad může sloužit specializovaný hardware efektivně realizující běžné operace jako je kombinace booleovských množin apod.

7. Hypertext a úplný text

Již v roce 1945 navrhl Vannevar Bush (viz [13]) základní koncepci nelineárního strukturování textu, které by korespondovalo asociativní podstatě lidské mysli. Tato nová metoda využívání a zpracování informace, která byla v roce 1965 Tedem Nelsonem nazvána hypertextem, však musela na své bohaté využití ještě několik desítek let počkat. Rozvoj a masové využívání osobních počítačů v polovině osmdesátých let vedly k explozi zájmů o tuto informační technologii. Současně s tím dochází k vývoji velkého počtu implementací hypertextu. Myšlenku využití hypertextu je vhodné objasnit na několika příkladech. Představme si člověka, který využívá textové vyjádření informací pro svoji práci. Kromě tužky a papíru může v dnešní éře osobních počítačů využívat s výhodou služeb textového editoru. Kromě sběru informací však často potřebuje vyhledávat části textů, ty případně dále klasifikovat či jinak separovat. Klasické textové editory v různé míře tyto funkce podporují. S přibývajícími informacemi se však bude stále obtížnější v textu orientovat. Lze sice vytvářet různé kartotéky (či databáze záznamů - v našem případě textů), nicméně pomalu se ztrácí přehled a mnoho úsilí se promarní listováním v takto lineárně organizované databázi a hledáním navazujících odkazů.

Jiným příkladem je studium rozsáhlého materiálu. Při procházení textu si například nezapamatujeme všechny definice, takže se k těmto definicím budeme chtít rychle vrátit. Můžeme si chtít vytvářet studijní plány a přehledy, které by již neobsahovaly partie, jež jsme zvládli, ale naopak partie, které potřebujeme zopakovat. I v tomto případě zřejmě potřebujeme inteligentnější nástroj než je textový editor či běžný databázový systém.

Dalším jednoduchým příkladem je potřeba rychlého nalezení informace v manuálu, v informační příručce, jízdním řádu či turistickém průvodci - tj. v kolekci plných textů.

Program k tomuto účelu sloužící se nazývá hypertextový systém. Data s nimiž tento program pracuje se nazývají hypertext, protože jsou informačně bohatší než obyčejný text.

Základní koncepty charakteristické pro hypertext:

- Asociativní vyhledávání pomocí vazeb mezi uzly reprezentovanými okny na obrazovce
- Prohledávání
- Počítačem podporovaná kooperativní práce
- Hypermedia rozšiřují principy hypertextu s komplexními objekty než jen s okny - na audio, video a další.

Problémy:

- Jak vytvářet vazby mezi jednotlivými uzly
- Navigace, jak zařídit aby se uživatel neztratil v hypertextové síti, nebo byla nadmíru rozptýlenajeho pozornost přemírou informace

- Neexistence standardu --- Zvítězí některý z existujících systémů např. systém Xanadu, nebo vzejde nový standard z World Wide Web-u (WWW) na Internetu.

Hypertext a DIS

Společné:

- Obecně
  - Oba koncepty navrhl Vannevar Bush již v r.1945
  - Oba koncepty mají pomoci prohledávat velké kolekce dat

- Společné techniky
  - Efektivní ukládání na vnější paměťová média
  - Sdílený přístup do databáze (k informacím)
  - Směs předzpracování (pre-analysis) a následného prohledávání (post-searching)
  - Používají: asociace, křížové odkazy a citace
  - U obou existuje potřeba automatického indexování a vytváření vazeb

Co by mělo být společného

Vyhledávání pomocí klíčových slov je potřeba zahrnout do hypertextových systémů

Prohlížení pomocí vazeb (odkazů) je třeba zahrnout do vyhledávacích systémů

Rozdílné:

Vyhledávání	x	Procházení
Statický	x	Uživatelsky modifikovatelný
Text	x	Multimedia
Indexování	x	Tvorba odkazů

8. Závěr

Strategickou sílu každé organizace dnes představuje informační systém. Počítačové zpracování informací, zvláště pak jejich využívání z různých zdrojů vede k vytváření uživatelských nástrojů - dotazovacích jazyků, pomocí kterých by měl uživatel být schopen formulovat své požadavky s cílem získat od IS relevantní odpovědi.

V současné době síli snahy o zahrnutí technik DIS do klasických databázových systémů. Příkladem může sloužit tezaurus implementovaný v produktu Oracle SQL*Text retrieval. Tento produkt obohacuje SQL jazyk o možnost dotazů do nad úplnými texty. Dotaz je zapsán ve formě:

SELECT <seznam_položek>
  FROM <seznam_tabulek>
  WHERE <položka> CONTAINS <textový výraz>

Textový výraz může být tvaru:

'text'	obyčejný term
'text'*	zprava rozšířený term
*'text'	zleva rozšířený term
'text'	oboustranně rozšířený term
't?xt'	term s libovolným znakem místo '?'
't%xt'	term s libovolným podřetězcem místo %
'text1' (m,n) 'text2'	text1 může být o m slov za text2 nebo text2 o n slov za text1

Jak bylo uvedeno, problematika DIS není zdaleka uzavřena, bouřlivý vývoj v této oblasti je možno sledovat zejména v [6], [7] a [8].

Rychlost růstu informačních zdrojů velmi výstižně zachycuje kniha [21].

Literatura

1. D.C.Blair. Indeterminacy in the Subject Access to Documents., Information Processing & Management, Vol.22, No.2, 1986
2. M.Bloch, A.Scheber. Textové bázy dát. Sofsem 86.
3. C.W.Cleverdon. Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems College of Aeronautics, Cranfield, England 1962
4. G.H.Gonnet, R.Beaza-Yates. Handbook of Algorithms and Data Structures. Addison-Wesley Publishing, 1991
5. W.F.Frakes, R.B.Yares Ed. Information Retrieval, Data Structures & Algorithms Prentice Hall 1992
6. D.K.Harman, Ed. The First REtrieval Conference (TREC-1) National Inst. of Standards and Technology, Gaithersburg, USA, 1993
7. D.K.Harman, Ed. The Second REtrieval Conference (TREC-2) National Inst. of Standards and Technology, Gaithersburg, USA, 1994
8. D.K.Harman, Ed. The Third REtrieval Conference (TREC-3) National Inst. of Standards and Technology, Gaithersburg, USA, 1995

9. S.Jones. Text and Context. Springer-Verlag, 1991.
10.K.S.Jones, Ed. Reflections on Trec (TREC-2). Information Processing & Management, vol.31(3) 1995.
11.J.Kostelanský. Použití rozšířené Boolské logiky a fuzzy logiky ve vyhledávání informací. DATASEM 93
12.B.Melichar. Textové informační systémy. Skriptum ČVUT, Praha 1994
13.J.Nilsen. Hypertext and Hypermedia. Academic Press, 1990.
14.V.Novák. Fuzzy množiny a jejich aplikace. SNTL, Praha, 1990.
15.E.Ozkaraham. Database Machines and Database Managment. Prentice Hall 1986
16.J.Pokorný. Vyhledávání ve velkých textových databázích. DATASEM 90.
17.J.Pokorný, M.Kopecký. Výzkumná zpráva UIVT AV CR V610. Vyhledávání v textových databázích, s využitím principů umělé inteligence a neuronových sítích. Grant č. 102/94/0728 Grantové agentury ČR.
18.G.Salton. The use of extended Boolean Logic on Information Retrieval. Proc. of ACM-SIGMOD, Int. Conf. on Management of Data, Boston, 1984.
19.V.Snášel, V.Sklenář, R.Nováková. Úplné texty a informační systémy. DATASEM 94.
20.V.Snášel, V.Sklenář, R.Nováková. Large Full Texts and Information Systems.Computer Based Learning, Opava 95.
21.J.Žbirka. Zpracování plných textů metodou třetí generace.DATASEM 93
22.I.H.Witten, A.Moffat, T.C.Bell. Managing Gigabytes: Compressing and Indexing Documents and Images. Van Nostrand Reinhold, 1994.

Poznámka recenzenta:

Obsahově jistě zajímavý příspěvek v obdržené verzi obsahoval řadu editačních a gramatických nedostatků. Přes veškeré úsilí následně provedené úpravy znamenají nápravu jen částečnou.