ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE
FAKULTA ELEKTROTECHNICKÁ

CZECH TECHNICAL UNIVERSITY IN PRAGUE
FACULTY OF ELECTRICAL ENGINEERING

Ing. Zdeněk Kouba, CSc.

On-line analýza geografické informace

Geographical Information On-line Analysis

Summary

Data warehouses aggregate substantial information contained in data and filter out non-useful details from data sets. Data warehouse is an ideal means for storing vast history data files. Data warehouses provide very often inputs to data mining processes, which make advantage of having such condensed information at disposal to discover useful knowledge hidden in the data.

The main importance of data warehouses is their ability to support on-line analytical processing (OLAP), which relies on fast evaluation of sums, mean values, and similar aggregates of data stored in the data warehouse in the context of its dimensions.

One of the main objectives of the research presented in this text is to prove the feasibility of applying a data warehouse to the geographical information on-line processing. The result is a method of integrating a geographical information system (GIS) with a data warehouse. Such integration makes possible to replace the computationally expensive spatial query with a corresponding OLAP query passed to the data warehouse. The possibility of taking advantage of spatial indices for creating the aggregation hierarchy of the data warehouse's geographical dimension has been further explored in order to support geographical information on-line analytical processing.

The author and his colleagues call this approach GOLAP. Set of experiments has proven that it substantially speeds-up the evaluation of geographical information on-line analytical queries in practically interesting cases.

Souhrn

Datové sklady umožňují převzít z primárních dat podstatnou informaci a oprostit ji od nepodstatných detailů. Představují ideální prostředek pro archivaci historických souborů údajů. Metody odhalování znalostí v datech umožňují následně odhalit informace a znalosti obsažené v archivovaných datech.

Hlavní význam datových skladů spočívá v tom, že umožňují provádět on-line analýzy dat, které spočívají v rychlém vyhodnocování průměrů, součtů a dalších agregátů údajů uchovávaných v datovém skladu v kontextu jeho dimenzí.

Jedním z prvořadých cílů výzkumu, prezentovaného v této habilitační přednášce, bylo ověřit možnost využití datových skladů k provádění on-line analýzy geografických dat. Výsledkem je koncepce integrace geografického informačního systému (GIS) s datovým skladem, která umožňuje výpočetně náročné vyhodnocování prostorového dotazu systémem GIS nahradit vyhodnocením OLAP dotazu položeného datovému skladu. Pro podporu on-line analytického zpracování geografické informace byla dále zkoumána možnost využít prostorových indexů pro vytvoření agregační struktury geografické dimenze. Experimenty prokázaly, že tato koncepce, označovaná autorem a jeho kolegy zkratkou *GOLAP*, přináší v prakticky významných případech výrazné zrychlení vyhodnocení prostorových analytických dotazů.

Klíčová slova

geografická infomace, geografické informační systémy, GIS, prostorový index, R-strom, datové sklady, agregační hierarchie, geografická dimenze, on-line zpracování analýz, OLAP

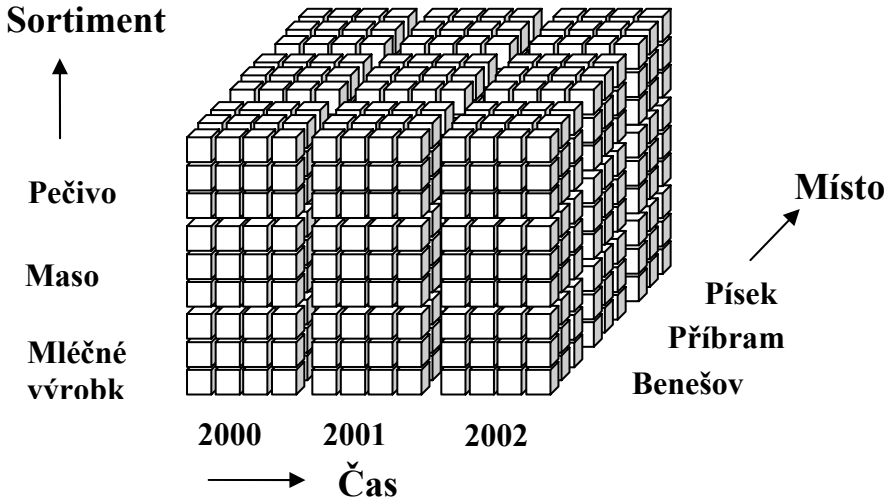
Key words

geographical information, geographical information systems, GIS, spatial index, R-tree, data warehouse, aggregation hierarchy, geographical dimension, on-line analytical processing, OLAP

Obsah

1 Úvod

Obchodní společnosti ve svých databázích zpravidla uchovávají údaje o obchodních transakcích uskutečněných v průběhu řady uplynulých let. Analýzou těchto údajů lze získat cenné znalosti o chování trhu, které jsou bezprostředně aplikovatelné při formulování obchodních strategií a přijímání manažerských rozhodnutí. Provádění analýz přímo nad primárními daty by však vzhledem k jejich obrovskému objemu bylo výpočetně velmi náročné a odezva na analytické dotazy by byla velmi pomalá. Z tohoto důvodu se předzpracovaná data ukládají do speciálního typu databází, tzv. datových skladů, které umožňují provádět jejich analýzu mnohem efektivněji. Hovoří se o systémech okamžitého zpracování analýz, označovaných jako OLAP (z anglického on-line analytical processing) systémy.



Obr. 1 Schématický pohled na datový sklad jako vícerozměrnou datovou strukturu

Datové sklady uchovávají agregovaná data v kontextu dalších údajů, jež má smysl při provádění analýz uvažovat. Smyslem OLAP systémů je poskytnout uživateli co nejrychleji požadované agregace dat na různých úrovních granularity, popřípadě výsledky analýz provedených nad těmito agregacemi. Cíli dosáhnout co nejrychlejší odezvy na analytické dotazy je podřízen i datový model datového skladu – viz např. [7].

Při návrhu běžných databází, často označovaných zkratkou OLTP (z anglického on-line transaction processing) pro odlišení od OLAP systémů, je naprosto nežádoucí uchovávání redundantních údajů. Naproti tomu datové sklady v zájmu dosažení co nejrychlejší odezvy redundantní údaje běžně

obsahují. Datový sklad tak běžně uchovává některé často vyhodnocované datové agregáty vyšších granularit, ačkoliv by bylo možné je získat agregací datových agregátů nižší granularity, které jsou rovněž v datovém skladu uchovávány.

Ke zpracování dat charakteru geografické informace, k jejichž analýze je zapotřebí vykonávat výpočetně náročné kombinatorické algoritmy či algoritmy výpočetní geometrie, se používají tzv. geografické informační systémy (GIS).

Předložený text rozšířené habilitační přednášky prezentuje možnost využití integrace geografického informačního systému s datovým skladem s cílem umožnit on-line analýzu geografické informace. Prezentovaných výsledků bylo dosaženo výzkumným týmem katedry kybernetiky ČVUT FEL pod vedením autora tohoto textu při řešení výzkumného projektu GOAL (Geographical Information On-line Analysis) programu INCO-COPERNICUS.

2 Geografické informační systémy

Data zpracovávaná systémy GIS mohou být klasifikována do dvou významných skupin:

- **strukturovaná data** bývají obvykle uložena v relační databázi, která může být součástí systému GIS. Častěji se však jedná o samostatný relační databázový server jiného dodavatele (typicky Oracle), vůči němuž se systém GIS chová jako klient.
- **nestrukturovaná data** vyjadřují geometrické vlastnosti jednotlivých GIS objektů a geometrické vztahy mezi nimi. Jsou obvykle reprezentována grafickými daty.

Zmíněná nestrukturovaná (grafická) data zpracovávaná systémy GIS jsou dvou základních typů — *rastrová* a *vektorová*.

V případě rastrových dat se jedná se o vysoce nestrukturovaná data, jejichž zpracování je obtížné. Z tohoto důvodu se používají obvykle jako ilustrativní podklad vektorové vrstvy při prezentaci map uživateli prostřednictvím grafického uživatelského rozhraní.

Zásadní význam mají v systémech GIS vektorová data, jež graficky interpretují jednotlivé GIS objekty v jejich geografickém kontextu a vzájemných geometrických souvislostech. Každý GIS objekt tak může být reprezentován jistou strukturovanou složkou v relační databázi a vektorovou grafickou složkou v jedné nebo více vektorových vrstvách grafických dat.

Každý GIS objekt je v příslušné vektorové vrstvě reprezentován jako jedno ze tří možných topologických primitiv *bod*, *linie* a *oblast*[1].

Důležité je, že daný GIS objekt nemusí mít přiřazeno jediné topologické primitivum, ale v závislosti na zvoleném měřítku může být tentýž GIS objekt reprezentován různými topologickými primitivy, nebo nemusí být reprezentován vůbec. Například objekt typu *město* velikosti menší než jistý práh nebude na mapě světa zobrazován vůbec, zatímco město, jehož velikost přesahuje jistou hranici, bude na mapě světa reprezentován jako bod. V podrobnějším měřítku může být první objekt reprezentován na mapě jako bod, zatímco druhý z výše uvedených objektů jako oblast.

Geografický informační systém typicky zpracovává dotazy typu: *"Nalezni všechny objekty, které leží v jisté oblasti dané pozice a daného tvaru a splňují jistou logickou podmínku."*

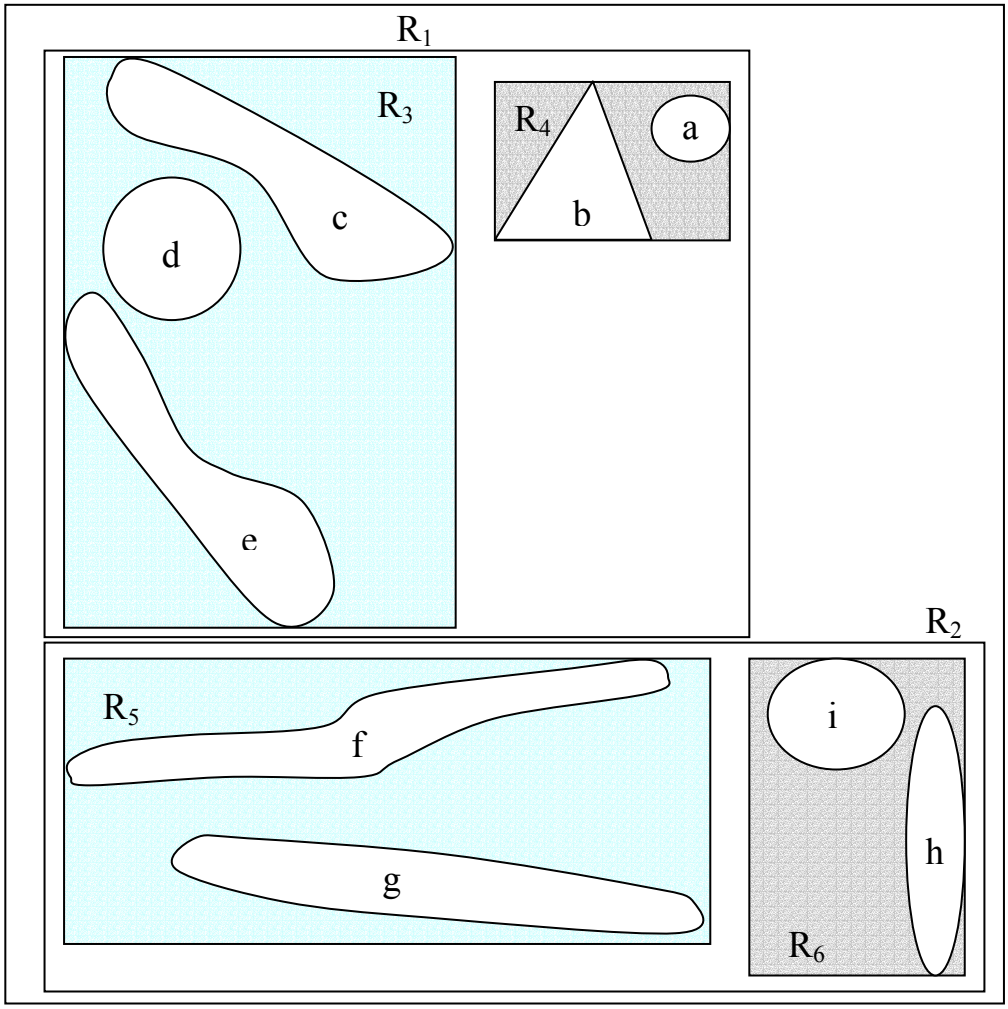
V tomto případě musí GIS testovat všechny v úvahu přicházející GIS objekty, zda leží v zadané oblasti. Test, zda daný GIS objekt leží uvnitř jiného GIS objektu, je výpočetně složitý i v případě, že se omezíme na jejich polygonální aproximace. Jde o to co nejvíce omezit množinu objektů, které je nutné testovat. Jinými slovy — snažíme se co nejvíce zúžit množinu kandidátů na pozitivní vyhodnocení. K tomu slouží vícerozměrné indexační techniky.

3 Vícerozměrné indexační techniky

Je známa celá řada vícerozměrných (nebo také prostorových) indexačních technik (přehled viz např. [3] nebo [8]) a implementace řady z nich jsou komerčně dostupné. Významní světoví producenti databázových systémů ke svým produktům nabízejí volitelné nadstavbové moduly implementující vícerozměrné indexy. Obdobně producenti systémů GIS dodávají moduly zefektivňující spolupráci jejich konkrétního systému GIS s libovolnou relační databází. Příkladem takového řešení je produkt *ArcSDE* firmy *ESRI*.

Z výše uvedeného odstavce tedy vyplývá, že vícerozměrné indexační techniky nejsou rozhodně objevem této habilitační práce. Jejím originálním přínosem ve vztahu k vícerozměrným indexačním technikám je způsob jejich využití pro zajištění efektivní spolupráce systému GIS s datovým skladem při provádění on-line analýz geografické informace.

[1] V případě trojrozměrných GIS by přibylo ještě topologické primitivum *prostor*. Naprostá většina systémů GIS je dvourozměrných, nebudeme se jím zde proto zabývat.



Obr. 2 - Princip tvorby *R-stromu.*

Stejně jako v případě indexů běžně používaných v relačních databázích (v naprosté většině implementovaných jako B-stromy) jde o to, aby byl hledaný záznam (zde reprezentující jistý prostorový objekt) lokalizován při minimálním počtu přístupů do databáze. Jednou z obvyklých technik používaných pro vícerozměrné indexování je technika využívající datových struktur označovaných jako tzv. *quad-tree* (někdy též *čtyřstrom*). Při použití této techniky je vždy prohledávaná oblast rozdělena na čtyři podoblasti: *severovýchod*, *jihovýchod*, *jihozápad* a *severozápad*. Každá z těchto oblastí je rekurzivně rozdělena opět na čtyři podoblasti a to až do jisté, předem dané, hloubky. Každá elementární oblast, t.j. taková, která se již dále nedělí na podoblasti, je spojena se seznamem geografických objektů, jež v dané oblasti leží. Vícerozměrný index je tedy tvořen stromem, jehož každý nelistový uzel má čtyři následníky. Při hledání objektu, který leží na jisté souřadnici, procházíme tímto stromem a v každém nelistovém uzlu, kterého jsme dosáhli, se rozhodneme, zda nás zajímá severovýchodní, jihovýchodní, jihozápadní nebo severozápadní část oblasti reprezentované tímto uzlem. Pak pokračujeme do

příslušného následnického uzlu, až dorazíme do listového uzlu. Ten definuje množinu objektů-kandidátů, které je pak nutno otestovat jeden po druhém.

Jiná skupina často používaných prostorových datových struktur, jejímž představitelem jsou například dále podrobněji zmiňované *R-stromy*, je konstruována následovně. Oblast odpovídající jisté úrovni indexu pokryjeme několika (minimálně m_{min} , maximálně m_{max}) podoblastmi, každou z nich opět rekurzivně pokryjeme několika podoblastmi, až dospějeme k jisté elementární oblasti. Zřejmě musí být k dispozici předpis, podle něhož v každém kroku rozhodneme, zda je příslušná oblast oblastí elementární, či zda má algoritmus pokračovat rekurzivně dále. Současně musí být k dispozici (například heuristické) kritérium určující, které ze všech možných rozdělení na podoblasti má být použito. Výsledkem je m_{max}-ární strom, jehož listy reprezentují výše zmíněné elementární oblasti. Každý takový listový uzel pak obsahuje odkazy na prostorové objekty, které leží v danému listovému uzlu příslušející elementární oblasti.

Na tomto místě je třeba upozornit, že pokrývající oblasti mohou být konstruovány jako obecné polygony. Konstruují se přitom tak, aby bylo dosaženo co nejmenší tzv. *hluché plochy*, t.j. té části plochy pokrývající oblasti, jež není obsazena žádným prostorovým objektem. Z praktických důvodů se však nejčastěji používají obdélníky a to přesto, že při jejich použití je hluchá plocha ve srovnání s použitím obecného polygonu větší. Hovoří se o tzv. *minimálním ohraničujícím obdélníku*[2] - *MOO*.

Konstrukci *R-stromu* ilustrujme na příkladě uvedeném na obr. 2, který představuje mapu obsahující 9 prostorových objektů označených postupně *a*, *b*, … až *h*. Tyto prostorové objekty jsou na první úrovni pokryty *MOO* označenými R_1 a R_2. R_1 pokrývá prostorové objekty *a*, *b*, *c*, *d*, *e*, které jsou na další úrovni pokryty *MOO* označenými R_3 a R_4. Obdobně prostorové objekty pokryté *MOO* R_2 (t.j. *f*, *g*, *i*, *h*) jsou na další úrovni pokryty *MOO* R_5 *a* R_6. Výsledný *R-strom* je pak zobrazen na obr. 3.

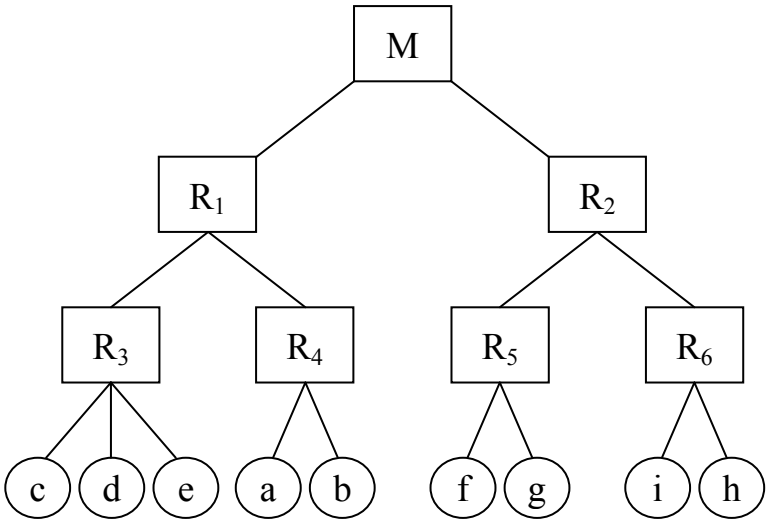
Při konstrukci *R-stromu* může dojít k tomu, že vzájemná konfigurace prostorových objektů nedovoluje pokrytí daným počtem vzájemně disjunktních *MOO*, aniž by některý objekt zasahoval do více než jednoho *MOO*. Z tohoto hlediska se jednotlivé prostorové indexační techniky dělí do dvou skupin:

- Indexační techniky, které **trvají na zachování disjunktnosti** *MOO*, výše naznačený konflikt řeší tak, že prostorový objekt zasahující do *n MOO* ($n > 1$) přiřadí každému z nich (*duplikace* objektů). Tato duplikace je někdy zamaskována tím, že se daný prostorový objekt rozdělí na *n* částí, z nichž

[2] Anglicky *minimal bounding rectangle* - *MBR*.

každá leží v právě jednom *MOO* (tzv. *střihání*[3] objektů). Výhodou tohoto přístupu je, že při vyhledávání stačí vyšetřit jedinou cestu stromu. Naopak operace vkládání nových geografických objektů je při tomto způsobu implementace složitější. Typickou datovou strukturou využívající tohoto přístupu jsou dále zmíněné *R^+-stromy*.

- Indexační techniky, které **připouštějí nedisjunktní** *MOO*, výše naznačený konflikt řeší tak, že referenci na prostorový objekt ležící uvnitř více *MOO* přiřadí jen jednomu z nich. Nevýhodou je, že v případě překrývajících se *MOO* musíme při lokalizaci objektů někdy vyšetřovat více než jednu cestu ve stromě. Operace vkládání nových prostorových objektů je naopak jednodušší. Zástupcem datových struktur využívajících tohoto přístupu jsou obecné *R-stromy* včetně jejich speciálního případu – *R*-stromů* (viz dále).



Obr. 3 - Vytvořený *R-strom.*

Zmíněné *R-stromy* jsou vícerozměrnou obdobou *B-stromů*, používaných běžně pro implementaci indexů záznamů klasických relačních databází. Datová struktura *B-stromů* je založena na skutečnosti, že máme k dispozici úplné (např. lexikografické) uspořádání klíčů záznamů databáze a není tedy problém sousední klíče uchovávat v téže stránce externího paměťového média. Naproti tomu v případě indexování prostorových objektů takové uspořádání neexistuje. Vícerozměrnost klíčů v tomto případě vždy poskytuje pouze částečné uspořádání, tj. může existovat dvojice klíčů, které jsou vzájemně neporovnatelné. Techniku *B-stromů* tedy nelze pro indexování prostorových objektů bezprostředně použít.

[3] Anglicky *clipping*.

Datová struktura označovaná jako *R-strom* byla prvně publikována v [5]. Podobně jako v případě *B-stromů* odpovídá každý uzel stromu jedné stránce externí paměti. Má však současně i geometrickou interpretaci — odpovídá totiž jistému *MOO* (v případě kořene je tímto *MOO* celá vyšetřovaná oblast — mapa). Každý nelistový uzel má tolik následovníků, na kolik *MOO* je jemu odpovídající oblast dále rozdělena. Listový uzel pak obsahuje odkazy na prostorové objekty ležící v *MOO* příslušné danému uzlu. *R-strom* má obvykle přiřazen tzv. řád, což je dvojice čísel ($m_{\min}$, $m_{\max}$), kde $m_{\min} < m_{\max} / 2$. Číslo $m_{\min}$ označuje minimální počet následníků, který musí mít každý nelistový uzel, resp. minimální počet odkazů na prostorové objekty, které musí mít listový uzel. Analogicky $m_{\max}$ je maximální počet takových následovníků/odkazů, které uzel daného *R-stromu* může mít.

Při vkládání nového prostorového objektu může dojít k tomu, že nějaký uzel odpovídajícího *R-stromu*, jenž má právě $m_{\max}$ následovníků nebo odkazů, přesáhne svou kapacitu. V tom případě musí dojít k obdobě akce *štěpení* stránky externí paměti, známé z teorie *B-stromů*. *MOO* odpovídající danému uzlu se tedy nahradí dvěma *MOO*. To může vyvolat operaci štěpení na vyšší úrovni a ta se tak může rekurzivně šířit až do kořenového uzlu.

Analogicky rušení prostorového objektu může nastat situace, že jistému uzlu, který má právě $m_{\min}$ následovníků/odkazů, klesne počet následovníků/odkazů pod $m_{\min}$. V takovém případě musí následovat akce *slévání* stránek odpovídající slučování *MOO*. I tato akce se může rekurzivně šířit až do kořenového uzlu.

Pro konstrukci *R-stromu* je zásadním problémem rozdělení daného uzlu. Algoritmus štěpení uzlu *R-stromu* hledá optimální pokrytí dané oblasti minimálními ohraničujícími obdélníky. Kritérium optimality je přitom voleno tak, aby byl minimalizován součet jejich ploch (důvodem je minimalizace výše zmíněné hluché plochy). Pro hledání globálního minima výše naznačeného kritéria by bylo třeba vzít v úvahu všechny možnosti štěpení. Takový algoritmus by měl tudíž exponenciální složitost. Ve skutečnosti se proto používají heuristické algoritmy, které vyhledávají pouze suboptimální řešení. Nejčastěji se používají heuristické algoritmy s lineární nebo kvadratickou složitostí — viz [5].

Vylepšenou variantou *R-stromů* jsou tzv. *R*-stromy*, které používají kritéria, jež kromě minimalizace ploch *MOO* sleduje i minimalizaci jejich překryvu. Důvod je zřejmý — snížení pravděpodobnosti, že bude nutné prohledávat více než jednu cestu ve stromě (viz výše). Při konstrukci *R*-stromů* se uplatňuje heuristické kritérium, jež sleduje s klesající vahou:

- minimalizaci ploch *MOO*,

- minimalizaci obvodů *MOO*,
- minimalizaci ploch překryvů *MOO*.

V případě R^+*-stromů* se při štěpení uzlu konstruují oba minimální ohraničující obdélníky zásadně jako disjunktní (R^+*-stromy* patří mezi indexační techniky, které trvají na zachování disjunktnosti *MOO* – viz výše). Z toho vyplývá, že mohou existovat objekty, které leží jistou částí v jednom a jinou částí v druhém minimálním ohraničujícím obdélníku. Kritérium pro nalezení nejvhodnější dvojice minimálních ohraničujících obdélníků je přitom voleno tak, aby zohledňovalo požadavek na minimální počet objektů zasahujících do obou minimálních ohraničujících obdélníků. Pokud přesto nastane situace, že existuje objekt ležící na pomezí obou zkonstruovaných minimálních ohraničujících obdélníků, budou na tento objekt odkazovat oba dva. Při vkládání nového GIS objektu do indexu tak může nastat potřeba jej vložit do více než jednoho indexního záznamu (listu R^+*-stromu*).

4 Integrace datový sklad - GIS

Geografickou informací jsou míněny údaje o geografických objektech spolu s jejich vzájemnými topologickými a geografickými vztahy, jež umožňují vyhodnocovat prostorové dotazy kladené uživatelem. Současné aplikace geografických informačních systémů (GIS) udržují velké objemy geografické informace a to obvykle v relačních databázích.

Protože je vyhodnocování prostorových dotazů výpočetně velmi náročné, dodávají nejdůležitější výrobci ke svým databázovým systémům speciální moduly implementující některé prostorové indexační techniky, jež zvyšují efektivitu zpracování prostorových dotazů. Společnost Oracle nabízí kupříkladu produkt označovaný *Spatial Cartridge*, který je založen na dříve zmíněné prostorové datové struktuře *quad-trees*. Podobně společnost *Informix* vyvinula produkt označovaný *Spatial Data Blade*, založený na struktuře *R-stromů*. Takový přístup je výhodný pro prostorové dotazy, jejichž úkolem je nalézt (geografické) objekty ležící uvnitř jisté, dotazem definované, oblasti. V případě analytických dotazů však dosud používané přístupy nepřinášejí výrazné zefektivnění.

Na druhé straně vývoj v oblasti informačních technologií zaznamenal v posledních letech značný rozvoj zejména v oblasti výše uvedených OLAP systémů. Výzkum v oblasti OLAP systémů byl motivován snahou zrychlit proces analýzy velmi velkých objemů dat uložených v rozsáhlých databázích.

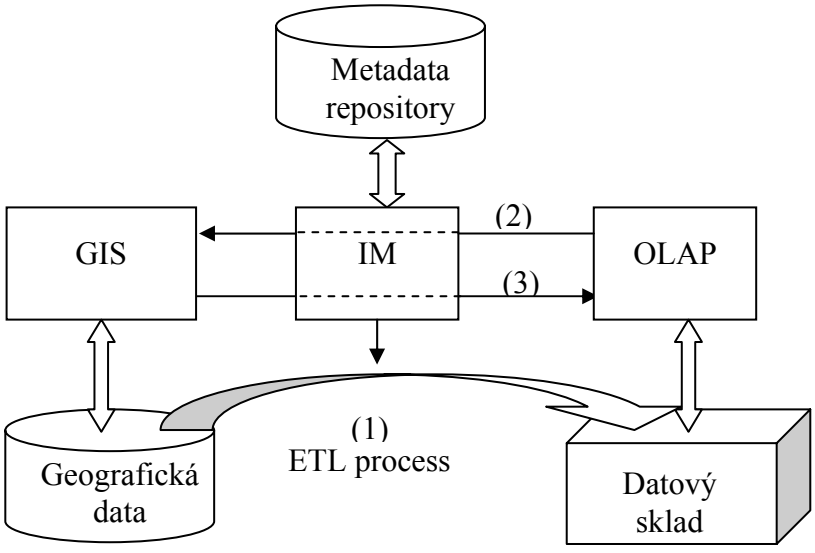
Myšlenka integrace datového skladu se systémem GIS, publikovaná v práci [6], nabízí řešení implementace systému pro on-line analýzu geografické informace přirozenými prostředky geografického informačního systému.

V tomto pojetí obsahuje datový sklad materializované pohledy[4] na geografická data. Namísto toho, aby se pokaždé, když je třeba vyhodnotit nějaký prostorově analytický dotaz, spouštělo složité a časově náročné vyhodnocování prostorového dotazu, provede se on-line analýza na předem připravených agregovaných datech. Tato cesta navíc umožňuje využít grafických možností systému GIS ke grafické prezentaci výsledku analytického dotazu v odpovídajícím geografickém kontextu.

Z předchozího odstavce vyplývá, že navrhovaná integrace nenabízí pouze jednosměrné propojení systému GIS s datovým skladem. GIS hraje v takovém integrovaném systému dvojí úlohu. Není pouze zdrojem dat, z něhož jsou data extrahována procesem ETL[5] a následně ukládána do datového skladu, ale je rovněž platformou pro prezentaci a interpretaci výsledků analýz.

Koncepce integrace systému GIS s datovým skladem byla rozpracována na katedře kybernetiky ČVUT FEL v rámci projektu GOAL programu INCO-COPERNICUS v letech 1998-2001. Vyvinutý prototyp byl ověřen na praktické úloze analýzy a predikce spotřeby v distribuční síti pitné vody společnosti VOSS, spol. s r.o.

4.1 Integrační modul



Obr. 3 - Schema integrace GIS - datový sklad.

[4] Materialized views
[5] Extraction-transformation-load

Základní komponentou navrženého přístupu je tzv. *integrační modul* (IM), jenž plní tři hlavní funkce:

- provedení akce ETL, t.j. naplnění datového skladu, resp. modifikace jeho obsahu daty pocházejícími ze systému GIS - viz datový proud označený symbolem (1) na obr. 3.
- synchronizace stavu systému GIS se současným stavem datového skladu — viz šipka (2) na obr. 3.
- účast systému GIS na formulování multidimensionálního OLAP dotazu — viz šipka (3) na obr. 3.

Integrační modul se opírá o úložiště (repository) metadat, jež obsahuje zejména následující typ metadat:

- datový model datového skladu,
- datový model datového zdroje GIS,
- skripty datových transformací,
- metadata popisující korespondenci mezi třídami/instancemi systému GIS a třídami/instancemi datového skladu.

4.1.1 Metadata repository

Úložiště metadat je vnitřně reprezentováno stromovou datovou strukturou a slouží jako prostředek pro společné uchovávání metadat všech výše uvedených typů na jediném místě. Součástí metadata repository je rovněž parser metadat, umožňující sestavit strom z textového zápisu metadat a metody pro přístup k metadatům a jejich údržbu.

Významnou součástí metadat jsou metadata popisující proces ETL, který představuje zásadní funkcionalitu integračního modulu. Tento proces se stal motivací vývoje systému *SumatraTT*[6] — obecného nástroje pro interaktivní návrh datových transformací — viz [1].

4.1.2 Proces ETL

Proces ETL je realizován pomocí jádra nástroje *SumatraTT*, který se tak zde stává součástí integračního modulu. ETL proces je definován pomocí dvou typů skriptů uložených v metadata repository:

- Skripty pro extrakci dat z datových zdrojů, jejich agregaci a uložení do datového skladu.

[6] *SumatraTT* je ochrannou známkou ČVUT zapsanou v rejstříku ochranných známek pod číslem 243786.

- Skripty pro validaci dat. Za určitých okolností a za předpokladu, že data obsahují redundantní údaje, umožňují tyto skripty data nejen validovat, ale i rekonstruovat chybné nebo chybějící údaje.

4.1.3 Korespondence tříd a instancí

Pro potřeby integrace obou systémů je nejvhodnější pohlížet na systém GIS i na datový sklad z pozic objektově-orientované analýzy. U systému GIS tak budeme rozlišovat dva základní pojmy:

- *GIS třídy* jsou abstraktní skupiny objektů téže sémantické i geografické interpretace (region, okres, budova, atd.).
- *GIS objekty* odpovídají instancím GIS tříd, t.j. jednotlivým konkrétním datovým elementům (konkrétní domy, silnice, kraje atd. reprezentované příslušným bodem, linií či oblastí) s přidruženými údaji.

Systém GIS je datovým zdrojem poskytujícím zbytku integrovaného systému geografická data. Spojujícím článkem mezi systémem GIS a datovým skladem jsou taxonomická hierarchie GIS tříd a agregační hierarchie geografické dimenze datového skladu. Integrace systému GIS s datovým skladem je založena na udržování následujících korespondencí:

- *Korespondence tříd* mapuje konkrétní agregační úrovně geografické dimenze datového skladu na odpovídající úrovně taxonomické hierarchie GIS tříd a naopak. Tato poměrně dlouhodobě stabilní informace je uložena v metadata repository.
- *Korespondence instancí* mapuje konkrétní instance agregačních úrovní na *GIS objekty*, t.j. instance *GIS tříd*, a naopak. Tato korespondence má větší dynamiku než předešlá. Integrační modul musí zaznamenávat změny v korespondenci instancí vyvolané akcemi na straně GIS i na straně datového skladu a propagovat ji vždy i tomu druhému z obou subsystémů.
- *Korespondence akcí* je nejdynamičtější ze všech třech analyzovaných korespondencí. Zajišťuje tzv. konzistenci navigace. Pokud například v prezentační vrstvě datového skladu přejdeme na jinou agregační úroveň, změní se informace o agregační úrovni v metadatech a systém GIS zareaguje přechodem na odpovídající úroveň taxonomické hierarchie GIS tříd. Jiným příkladem takové akce je výběr objektů na straně GIS. Prezentační vrstva datového skladu zareaguje výběrem odpovídající datové kostky.

4.2 GOLAP — on-line analytické zpracování geografické informace

Řešení integrace systému GIS s datovým skladem popsané v předchozím odstavci je omezeno na případy, kdy je struktura geografické dimenze předem dána a je relativně stabilní. V takovém případě je možné předem spočítat některé často používané datové agregáty (odpovídající například územním celkům daného územního členění) a uložit je v datovém skladu. Prostorový dotaz, který uživatel položí systému GIS, může být v takovém případě přeložen na odpovídající dotaz typu OLAP a ten je pak předložen ke zpracování datovému skladu. Představme si však, že v případě náhlého příchodu povodňové vlny potřebujeme rychle určit, pro kolik mužů, žen a dětí ze zatopené oblasti je třeba zajistit nouzové ubytování. Vzhledem k tomu, že hranice zatopené oblasti nebere ohled na administrativní členění státu, připravené datové agregáty při řešení takového dotazu příliš nepomohou.

Pro takové aplikace navrhujeme nový přístup popsaný v práci [6]. Hierarchická struktura geografické dimenze je vytvořena tak, že odpovídá prostorovému indexu zkonstruovanému pro dané geografické rozložení GIS objektů. Na základě tohoto přístupu byl vybudován prototyp systému nazvaného GOLAP[7]. Následující odstavce jej popisují podrobněji a prezentují rovněž výsledky experimentů, dokumentujících efektivitu takového řešení.

4.2.1 Architektura systému GOLAP

Architektura systému GOLAP vychází z přirozeného začlenění prostorového indexu do komerčního datového skladu. Prototypová implementace zavádí zmíněný prostorový index jako rozšíření služby *Analytical Services*, která je součástí databázového systému Microsoft SQL Server 2000. Prezentované řešení je však obecné a není omezeno na tento konkrétní produkt. Jak samotná idea, tak vyvinutý software může být snadno adaptován na libovolnou platformu datového skladu.

Systém GOLAP se sestává ze dvou základních součástí:

- komponenty pro vytváření prostorového indexu nazývané *spatial index builder* a
- komponenty pro vyhodnocování analytického dotazu nazývané *query evaluator*.

První část systému GOLAP — *spatial index builder* — je nástroj, který umožňuje pro danou mapu vytvořit prostorový index. Pro implementaci

[7] Geographical Information On-line Analysis Processing

prostorového indexu byla zvolena datová struktura *R*-stromů*. Jak již bylo uvedeno, prostorový index je strom, jehož listy odkazují na jednotlivé prostorové objekty. Každý nelistový uzel odpovídá jistému *MOO* a odkazuje na datový agregát odpovídající území pokrytému příslušným *MOO*.

Spatial index builder je spouštěn pouze v průběhu plnění/aktualizace datového skladu procesem *ETL*.

V naší, z hlediska současné teorie systémů GIS značně netypické úloze potřebujeme s každým *MOO* spojit příslušný datový agregát. Jistá komplikace vznikne, pokud nějaký GIS objekt leží v průsečíku několika *MOO*. Fakta takového GIS objektu zahrneme do datového agregátu odpovídajícímu jednomu (libovolně zvolenému) *MOO*, uvnitř něhož daný GIS objekt leží.

Druhá část systému GOLAP — *query evaluator* — dává uživateli k dispozici grafické uživatelské rozhraní pro definování oblasti ve tvaru polygonu, která má být analyzována. Pro jednoduchost výkladu předpokládejme, že agregační funkcí je **součet** faktů. *Query evaluator* pak vyhodnocuje danou oblast s využitím R^* stromu dle následujícího algoritmu:

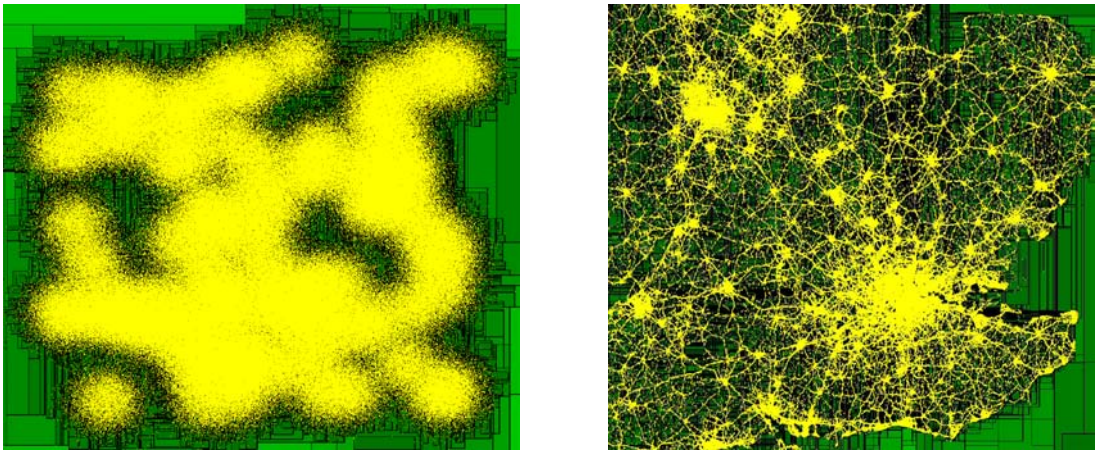
1. Do seznamu *Uzly* přiřaď kořen R^* stromu, proměnné *Výsledek* přiřaď hodnotu 0.
2. Je-li seznam *Uzly* prázdný, ukonči výpočet; proměnná *Výsledek* v takovém případě obsahuje hledaný součet faktů přes všechny GIS objekty ležící ve vyšetřované oblasti.
3. Vyjmi první prvek ze seznamu *Uzly* a ulož jej do proměnné *Uzel*.
4. Je-li *Uzel* listovým uzlem R^* stromu nebo reprezentuje-li MOO ležící celý uvnitř vyšetřované oblasti, přičti datový agregát odpovídající uzlu *Uzel* do proměnné *Výsledek* a jdi na bod 2.
5. Má-li MOO odpovídající uzlu *Uzel* s vyšetřovanou oblastí prázdný průnik, jdi na bod 2.
6. Na konec seznamu *Uzly* přidej všechny následníky uzlu *Uzel* v R^* stromě a jdi na bod 2.

4.2.2 Experimentální výsledky

Pro ověření, jak velké urychlení systém GOLAP přináší oproti určení výsledku na základě agregace údajů jednotlivých prostorových objektů ležících uvnitř vyšetřované oblasti konvenčním systémem GIS, byly provedeny následující experimenty. Pro jejich provedení bylo k dispozici následující prostředí:

- pracovní stanice Pentium IV, 1,8 GHz, 512 MB RAM,
- operační systém MS Windows XP,

- MS SQL Server 2000 se službou Analytical Services.



Obr. 4 - Ukázka geografického rozložení GIS objektů v souborech S_1 a S_2

Experimenty byly provedeny na mapě ve tvaru dvourozměrné mřížky rozměru 10 000 x 10 000 bodů se dvěma soubory dat - viz obr. 4:

- Soubor $\mathbf{S}_1$ byl vygenerován tak, že bylo v mapě náhodně vygenerováno 100 center s rovnoměrným rozložením souřadnic podél obou os. Kolem každého z těchto center bylo vygenerováno 10 000 GIS objektů s dvojrozměrným (Gaussovým) rozložením s rozptylem 1000 bodů. Celkem tak soubor $\mathbf{S}_1$ čítal 1 milion GIS objektů a simuloval tak 100 měst rozmístěných v mapě. Kolem každého z nichž je silniční síť přirozeně nejhustší v centru a směrem k okraji řídne.
- Soubor $\mathbf{S}_2$ vycházel z reálných údajů o rozložení dopravních nehod ve Velké Británii. Pouze počet GIS objektů byl zredukován náhodným výběrem s rovnoměrným rozložením tak, aby i tento soubor reprezentoval 1 milion GIS objektů.

Pro experiment byl navržen velmi jednoduchý model datového skladu, který sestával z jedné tabulky faktů a dvou dimenzí — geografické a časové. Tabulka faktů obsahovala jediný fakt reprezentující počet nehod v daném místě za jeden rok. Časová dimenze měla jedinou agregační úroveň *rok* a obsahovala její instance reprezentující 10 uplynulých let. Celkově tedy datový sklad odpovídal situaci, kdy je pro každý GIS objekt (např. nebezpečná křižovatka) evidován počet nehod za každý z uplynulých 10 roků.

Agregační hierarchie geografické dimenze byla vytvořena na základě odpovídajícího prostorového indexu implementovaného jako R^*-strom. Pro každou jednotlivou úroveň R^*-stromu obsahuje agregační hierarchie samostatnou agregační úroveň. Pro každou její instanci, která odpovídá právě

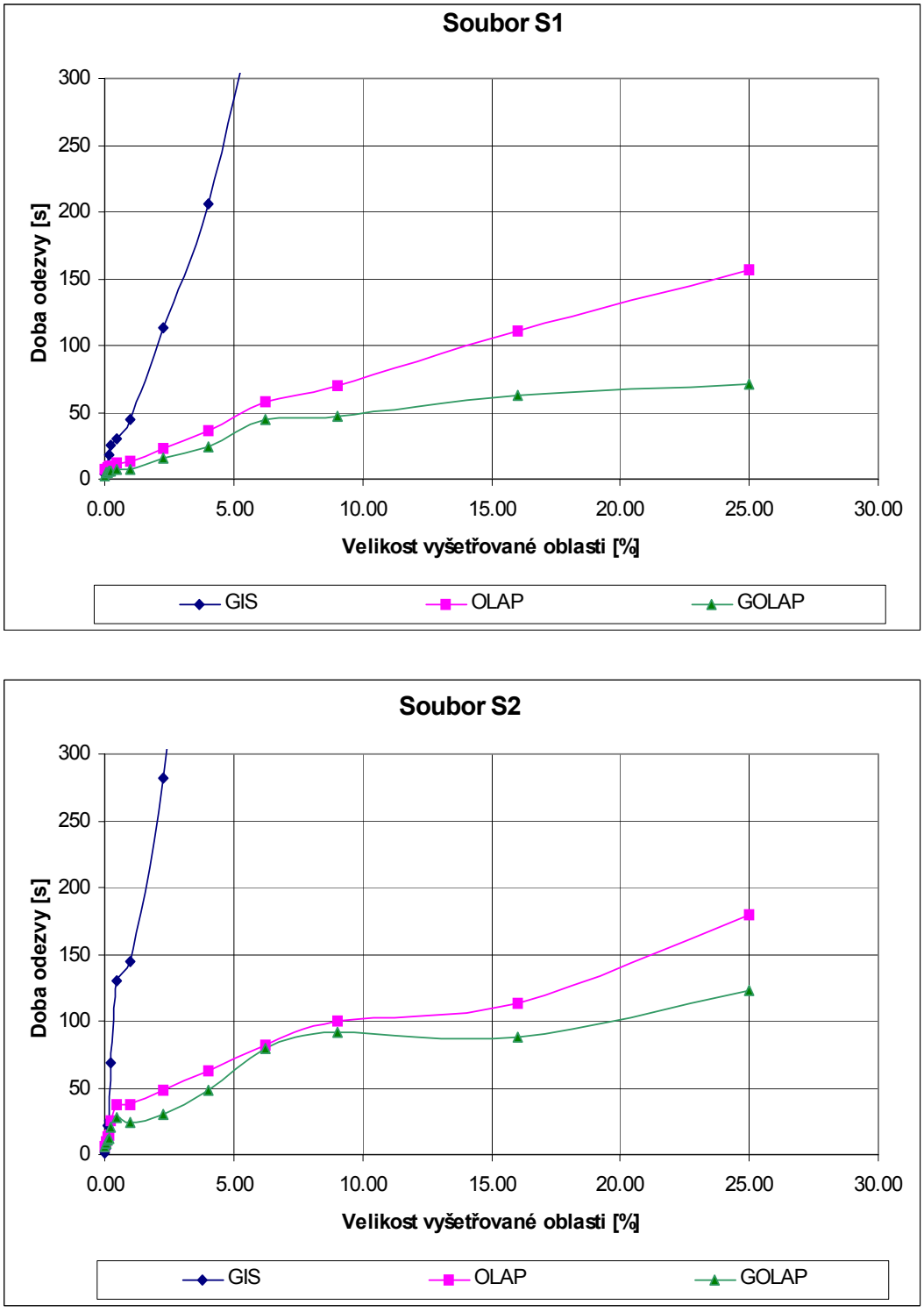
jednomu *MOO*, a každou instanci agregační úrovně *rok* časové dimenze obsahuje tabulka faktů spočítaný datový agregát.

Je zřejmé, že prostorový index přináší na jedné straně potenciální zrychlení dotazu, na druhé straně je s ním spojena jistá režie. Cílem experimentů bylo posoudit, zda zmíněná režie významně nedegraduje očekávaný přínos prostorového indexu.

Experimenty byly prováděny pro prostorový dotaz, jehož cílem bylo nalézt souhrnný počet dopravních nehod ve vyšetřované oblasti za posledních 10 let. Vyšetřovanou oblastí byl čtverec, jehož plocha měla 13 různých velikostí v rozsahu od 1% do 25% celkové plochy mapy. Výsledná doba odezvy na předmětný prostorový dotaz byla pro každou velikost vyšetřované oblasti určena jako průměrná hodnota ze 30 měření a to 3 měření v 10 různých, náhodně zvolených polohách vyšetřované oblasti.

Obr. 5 shrnuje výsledky experimentů prováděných na obou souborech dat $\mathbf{S}_1$ a $\mathbf{S}_2$. Křivka označená *GIS* odpovídá řešení prostorového dotazu bez podpory datového skladu. Křivka označovaná *OLAP* odpovídá situaci, kdy se využívá datového skladu k získání agregované hodnoty faktu pro každý GIS objekt, avšak každý GIS objekt se zpracovává samostatně. Konečně křivka označená *GOLAP* reprezentuje dobu odezvy dosaženou s využitím prostorového indexu.

Přestože koncepce systému GOLAP umožňuje vyšetřovat oblasti mající tvar libovolného polygonu, byly experimenty prováděny s vyšetřovanou oblastí ve tvaru pravoúhlého čtyřúhelníka (čtverce). Tento přístup byl zvolen proto, aby efektivitu prostorového dotazu nepříznivě neovlivnilo testování, zda daný prostorový objekt leží uvnitř obecného polygonu. Z tohoto důvodu je třeba považovat výsledné porovnání rychlosti odezvy systému GOLAP a prostého prostorového dotazu z pohledu systému GOLAP za pesimistické. Lze předpokládat, že v případě vyšetřované oblasti ve tvaru obecného polygonu bude zrychlení odezvy u systému GOLAP ještě výraznější.



Obr. 5 - Závislost doby odezvy prostorového dotazu na velikosti vyšetřované oblasti

Při pohledu na obr. 5 lze konstatovat, že díky integraci GIS s datovým skladem dochází ke značnému zkrácení doby odezvy na prostorový dotaz. Další zrychlení pak přináší využití prostorového indexu pro definici agregační hierarchie geografické dimenze v systému GOLAP.

5 Závěr

Datové sklady umožňují převzít z primárních dat podstatnou informaci a oprostit ji od nepodstatných detailů. Datové sklady tak představují ideální prostředek pro archivování historických souborů údajů. Ty mohou obsahovat informace a znalosti, následně vytěžitelné metodami odhalování znalostí v datech.

Jedním z prvořadých cílů prezentovaného výzkumu, prováděného pod vedením autora předložené rozšířené habilitační přednášky, bylo ověřit možnost využití datových skladů k provádění on-line analýzy geografických dat. Výsledkem je koncepce integrace geografického informačního systému GIS s datovým skladem, která umožňuje výpočetně náročné vyhodnocování prostorového dotazu systémem GIS nahradit vyhodnocením OLAP dotazu položeného datovému skladu. Pro podporu on-line analytického zpracování geografické informace byla dále zkoumána možnost využít prostorových indexů pro vytvoření agregační struktury geografické dimenze. Experimenty prokázaly, že tato koncepce, označovaná autorem zkratkou *GOLAP*, přináší v prakticky významných případech výrazné zrychlení vyhodnocení prostorového analytického dotazu.

6 Literatura

[1] Aubrecht P., Kouba Z.: *Metadata Driven Data Transformation*, In: World Multiconference on Systemics, Cybernetics and Informatics, International Institute of Informatics and Systemics, Orlando, 2001,vol. 1, p. 332-336. ISBN 980-07-7541-2.

[2] Beckmann N., Kriegel H.-P., Schneider R., Seeger B.: *The R*-tree: An Efficient and Robust Access Method for Points and Rectangles*, Proc. of ACM SIGMOD Int. Conference on Management of Data, Washington D.C., 1993

[3] Beneš R.: *Vícerozměrné datové struktury a jejich použití v GIS*, diplomová práce, MFF UK, Praha, 1999

[4] Finkel R., Bentley J.: *Quad Trees: A Data Structure for Retrieval on Multiple Keys*, Acta Informatica, Vol. 4, No.1, 1974

[5] Guttman A.: *R-trees: A Dynamic Index Structure for Spatial Indexing*, Proc. Of SIGMOD Int. Conference of Management of Data, 1984, pp. 47-54

[6] Mikšovský P., Kouba Z.: *GOLAP — Geographical Online Analytical Processing*, In: Database and Expert Systems Applications. Heidelberg : Springer, 2001, vol. 1, p. 442-449. ISBN 3-540-42527-6

[7] Kouba Z.: *Datové sklady a získávání znalostí*, kapitola v monografii Umělá inteligence 4, Mařík-Štěpánková-Lažanský (editoři), Academia, Praha, 2003, pp. 313-354, ISBN 80-200-1044-0.

[8] Pokorný J.: *Prostorové datové struktury a jejich použití k indexaci prostorových objektů*, GIS Ostrava 2000, Technická universita Ostrava, 2000

Ing. Zdeněk Kouba, CSc.

nastoupil po maturitě na gymnáziu v Benešově v roce 1978 ke studiu oboru technická kybernetika na Českém vysokém učení technickém v Praze, fakultě elektrotechnické, které ukočil s vyznamenáním v roce 1983. Po vykonání základní vojenské služby nastoupil v roce 1984 do interní vědecké aspirantury v tehdejším Středisku výpočetní techniky ČSAV (dnes ÚIVT Akademie věd České republiky).

V roce 1987 využil nabídky katedry řídicí techniky FEL ČVUT a nastoupil do pracovního poměru na FEL ČVUT. Nejprve byl zařazen do funkce vědecký pracovník, později do funkce odborný asistent na katedře řídicí techniky. Dnes působí v této funkci na katedře kybernetiky téže fakulty.

V roce 1991 obhájil kandidátskou disertační práci, zabývající se zpracováním neurčitosti v expertních systémech založeným na metodách vícerozměrné datové analýzy. Výsledky své práce v této oblasti aplikoval mimo jiné během svého tříměsíčního studijního pobytu v roce 1990 na universitě Tor Vergata v Římě.

Své pedagogické působení na FEL ČVUT zahájil v roce 1989, kdy vedl cvičení předmětu *Teorie automatického řízení*, později se podílel na přípravě výuky a vedl cvičení v předmětech *Kybernetické systémy*, *Systémy pro řízení počítači*, *Operační systémy pro řízení* a *Expertní a databázové systémy*.

Ve školním roce 1991/92 začal participovat na přednáškách předmětu *Znalostní systémy* a ve školním roce 1993/94 rovněž na přednáškách předmětu *Informační a databázové systémy*. V roce 1994 zahájil svou samostatnou přednáškovou činnost a to nejprve jako přednášející předmětů *Informační a databázové systémy* a *Znalostní systémy*, které byly v té době volitelnými předměty pro studenty 4. resp. 5. ročníku oboru technická kybernetika.

Předmět *Informační systémy*, který se soustředil zejména na metodiku návrhu informačních systémů, dal základ pro vybudování povinného předmětu bakalářského studia *Projektování informačních systémů*, který Ing. Kouba přednáší od školního roku 1998/99 dosud.

Předmět *Znalostní systémy* se zabýval obecnými principy znalostních systémů, zpracováním neurčitosti v systémech na podporu rozhodování a metodami extrakce znalostí z dat. Koncepce tohoto předmětu umožnila vybudovat povinný předmět inženýrského studia *Systémy na podporu rozhodování*, který Ing. Kouba přednáší od zimního semestru školního roku 1999/2000.

Ing. Kouba během svého pedagogického působení dovedl k úspěšné obhajobě diplomové práce více než 35 diplomantů.

Dne 24.1.2001 schválila vědecká rada ČVUT FEL jmenování Ing. Kouby do funkce školitele pro studijní obor doktorského studia *26-19-9 Umělá inteligence a biokybernetika*. V současné době je Ing. Kouba školitelem jednoho doktoranda třetího ročníku, jednoho doktoranda druhého ročníku a tří doktorandů prvního ročníku doktorského studia. Ing. Kouba je rovněž zván jako člen komise pro obhajobu doktorských disertačních prací na pracoviště školící doktorandy v příbuzných oborech, zejména na Vysokou školu ekonomickou v Praze a Fakultu aplikovaných věd Západočeské university v Plzni.

Během uplynulých let byl pozván k přednesení několika přednášek na zahraničních univerzitních pracovištích (1990 série tří přednášek o aplikovatelnosti metod vícerozměrné datové analýzy při zpracování neurčitosti během dvoutýdenního pobytu na universitě v Portu, Portugalsko; 1998 týdenní přednáškový pobyt na Milwaukee School of Engineering, USA; 2000 přednáška na téma metody integrace geografického informačního systému a datového skladu na Universitě Jana Keplera v Linci, Rakousko).

V roce 2000 na výzvu programového výboru mezinárodní konference Balanced Automated Systems in Manufacturing and Transportation zorganizoval workshop *Data Management and Data Warehousing*, který se uskutečnil jako samostatná sekce konference BASYS v Berlíně v září 2000.

V roce 2001 ho vyzval programový výbor mezinárodní konference DEXA (Database and Expert Systems Applications) k zorganizování mezinárodního workshopu *Presenting and Exploring Heritage on the Web* (PEH'02). První ročník tohoto workshopu se uskutečnil v září 2002 jako součást 13. mezinárodní konference DEXA 2002 v Aix-En-Provence ve Francii, druhý ročník byl součástí 14. mezinárodní konference DEXA 2003 v Praze. Ing. Kouba je předsedou programového výboru i 3. ročníku tohoto mezinárodního workshopu, který proběhne na přelomu srpna a září 2004 jako součást 15. mezinárodní konference DEXA 2004 v Zaragoze ve Španělsku.

Ing. Kouba je členem pedagogické komise FEL ČVUT. Dále je členem technické normalizační komise *TNK 122 Geografická informace / Geomatika* Českého normalizačního ústavu. Jako člen této komise se aktivně podílí na procesu začleňování řady norem ISO 19101 — ISO 19122 do soustavy norem ČSN.

Jeho odborným zájmem jsou metody formálního návrhu softwarových (zejména databázových) systémů, metody vícerozměrné datové analýzy a jejich aplikace v oblasti zpracování neurčité informace systémy na podporu rozhodování a v oblasti procesu data mining. Svou odbornou činnost rozvíjí v prostředí Gerstnerovy laboratoře pro inteligentní rozhodování a řízení na katedře kybernetiky, kde vybudoval a řídí pracovní skupinu *informačních a znalostních systémů*.

V minulosti vedl několik projektů FRVŠ a Interní grantové agentury ČVUT, podílel se na řešení několika projektů programu TEMPUS, PECO, COST, GAČR i řady průmyslových projektů. V této oblasti je nejvýznamnější jeho role technického vedoucího výzkumného týmu ČVUT FEL v projektech *EUROSAT* (1993-96, program PECO Evropské komise), *Evidence kulturních památek - počítačový systém* (1996-98, grant Ministerstva kultury ČR) a *GOAL - Geographical Information On-line Analysis* (1998-2001, program INCO-COPERNICUS Evropské komise).

Od dubna 2002 je Ing. Kouba zodpovědným zástupcem ČVUT FEL v mezinárodním řešitelském konsorciu projektu *Enabling Communities of Interest to Promote Heritage of European Regions (CIPHER)* podprogramu IST 5. rámcového programu Evropské komise, který je zaměřen na využití technik znalostního managementu při prezentaci kulturního dědictví.

Jako autor/spoluautor publikoval v posledních deseti letech své výsledky ve 3 kapitolách v mezinárodních monografiích, 1 kapitole v tuzemské monografii, 1 článku v mezinárodním časopise, ve 20 příspěvcích na mezinárodních a 2 příspěvcích na tuzemských konferencích a ve 29 oponovaných výzkumných zprávách, souvisejích s řešenými národními i mezinárodními projekty, jejichž byl (spolu)řešitelem. Publikované výsledky byly citovány v 7 pracích zahraničních a 10 českých autorů. Jeho nejvýznamnější publikace jsou:

[1] Kouba Z.: *Datové sklady a získávání znalostí*, in Umělá inteligence 4, Academia, Praha, 2003, pp. 313-354, ISBN 80-200-1044-0.

[2] Kouba Z., Vlček T.: *Man-Machine Interface for CIM*, In: LNCS 973, Springer-Verlag, Berlin-Heidelberg, 1995, p. 427-438, ISBN 3-540-60286-0.

[3] Kouba Z., Matoušek K., Mikšovský P.: *On-line Analysis of Utility Networks*, In: Knowledge and Technology Integration in Production and Services, Kluwer Academic / Plenum Publishers, 2002, p. 469-476, ISBN 1-4020-7211-2.

[4] Mikšovský P., Kouba Z.: *GOLAP - Geographical Online Analytical Processing*, In: Database and Expert Systems Applications, Springer, Heidelberg, 2001, vol. 1, p. 442-449. ISBN 3-540-42527-6.