ALGORITHMS AND METHODS FOR MINING ASSOCIATION RULES IN IMAGE DATABASES

David ZEMAN, Master Degree Programme (5)
Dept. of Information Systems, FIT, VUT
E-mail: xzeman14@stud.fit.vutbr.cz

Supervised by: Ing. Vladislav Kubíček

ABSTRACT

Subject of this work is to show algorithms and new methods for mining content-based associations from multimedia databases. We focus on associations with recurrent items and with spatial relationship, describe weak points and find possibilities of further improvement. Three algorithms are introduced as a soution of problem with mining recurrent items, one of them also consider spatial relationship of objects in image.

1 ÚVOD

Dolování dat z databází se stává námětem mnoha prací a výzkumných projektů. Doposud bylo však vždy spojeno dolování dat s alfanumerickými databázemi. Rozvoj internetu a firemních sítí vede k stále masovějšímu použití grafických a zvukových databází a tomu odpovídající formy dat. Dostupnost a nové možnosti vedou k nebývalému zájmu a rozšíření použití multimediálních dat. S tím vyvstává potřeba jejich vhodného ukládání, vyhledávání a zpracování Objevují se rozsáhlé databáze obrázků a záznamů zvuků. Několik zajímavých studií a aplikací bylo představeno zejména pro řešení problémů v oblasti zpracování meteorologických a satelitních dat. Ovšem jejich objemnost a rozličná struktura informací nedovoluje použití klasických algoritmů a postupů pro zpracování a vyhledávání. Proto se hledají alternativní cesty jak rychle a efektivně tato data používat. Obecný postup práce s daty a jejich použití v procesu získávání znalostí lze zjednodušeně popsat následovně : čistění dat, sjednocení dat, transformace dat, dolování dat, vyhodnocení.

2 ASOCIAČNÍ PRAVIDLA

2.1 ÚVOD

Procesem uvedeným v (1) je možné získat celou řadu zajímavých informací jako jsou různé evoluční či deviační analýzy, charakteristická, klasifikační či asociační pravidla. Právě posledně zmíněná asociační pravidla a jejich dolování jsou námětem této práce. Umožňují nám získat obraz o vztazích mezi prvky v množině dat na základě statistických údajů.

Právě hledání nejrůznějších vztahů mezi jednotlivými objekty databáze, ať už jsou tyto objekty jakékoli, vede k stále častějšímu použití metod pro dolování asociačních pravidel. Naskýtá se tak možnost zpracovávat obrovské množství dat, kde jednotlivé položky samy o sobě mají takřka nulovou hodnotu, ovšem pokud jsou zpracovány ze statistického hlediska jako celý soubor dat, přináší velmi cenné informace, v dnešní době zejména pro bussiness sféru. Asociační pravidla nám tak umožňují získat obraz o vztazích mezi prvky v dané množině dat.

3 OBRAZ A JEHO POPIS

Klasická databáze obrázků je pro aplikaci zpracování nebo vyhledávání zcela nepoužitelná. Existuje řada přístupů jak se s tímto problémem vypořádat. Na jedné straně je to simulace obrazové databáze jako transakční, na druhé straně je to zavedení úplně nové struktury s obrazovými atributy. Takovýto popis obrazových dat nahrazující obrázek můžeme rozdělit na fyzický popis a logický popis.

3.1 FYZICKÝ A LOGICKÝ POPIS OBRAZU

Nejběžnějším formou fyzické reprezentace je rastrová forma, kdy se každý obrázek skládá z hlavičky a těla. Logický popis obsahuje řadu atributů charakterizujících obrázek. Daly by se rozčlenit do kategorii, jako jsou meta-atributy, sémantické atributy, atributy barev, atributy textur, atributy textur a prostorové atributy.

Různé typy atributů mohou být dále seskupovány a vytvářet tak **vektor rysů**. Hodnoty pak bývají uspořádány tak aby vytvářeli co nejlepší reprezentaci původního obrázku a co nejvíce snižovali kriteria jako je čas zpracování, obsáhlost popisu a paměťová náročnost.

3.2 MODELY OBRAZOVÝCH DAT

Model obrazových dat je abstrakce, která může být vhodnou náhradou za obrázek a může poskytovat jeho dostačující popis. Proces popisu obrázku se stává z extrakce globálních charakteristik obrázku, rozpoznání obrázku a přiřazení významu objektům. Přístupy k modelování obrazových dat mohou být rozděleny do skupin na základě pohledů, které jednotlivé modely podporují. Mezi hodnotné návrhy na modely obrazových dat patří : **VIMSYS** model obrazových dat, kde obrázek je uvažován jak čtyři vrstvy, **EMIR^2** rozšířený model pro obrazovou reprezentaci a vyhledávání, **AIR** adaptivní model pro práci s obrázky.

Pro celkovou reprezentaci obrazu je klíčové zvolení také dalších atributů jako je barevný model, vybrání vhodné redukční metody, zvolení správné reprezentace rysů nebo výběr podobnostní funkce.

4 DOLOVÁNÍ ASOCIAČNÍCH PRAVIDEL

Asociační pravidla v obrazových databázích je možné dále členit podle typu položek na obou stranách asociačního pravidla. Lze získat asociace mezi obrazovým obsahem a neobrazovým obsahem, asociace mezi obrazovými obsahy bez prostorového uspořádání a asociace mezi obrazovými obsahy s prostorovým uspořádáním.

Tušíme, že samotné prostorové uspořádání bude jedním z problémů, který nelze při řešení opomenout. V transakčních databázích nic takového neexistovalo, zde jsou slova jako

nad, pod, před, mezi klíčová pro stanovení vzájemných vztahů. Zvláště pak v kombinaci s ostatními vlastnostmi jako je barva nebo tvar mohou vést k získávání zajímavých asociací.

4.1 SPECIALIZOVANÉ ALGORITMY

Mezi základní algoritmy pro práci s obrazovými databázemi patří:

- Apriori algoritmus, základní algoritmus pro práci s transakčními databázemi
- MaxOccur algoritmus zohledňující mnohačetné výskyty
- MM-Spatial algoritmus, který pracuje také s prostorovým uspořádáním

Ač se tyto algoritmy snaží vypořádat se specifickými problémy obrazových databází, stále můžeme najít řadu nevýhod jako je malá účinnost, kdy se k uživateli dostanou zbytečné a často nechtěné informace. Vyvstává s tím nutnost výsledky dále zpracovávat a ověřovat. Proto se objevují metody přidávání dalších restrikcí. Další nevýhodou je samotná náročnost procesu nalezení informací v obraze. Zde dochází k častému propojení s ostatními vědními obory a dochází k použití např. neuronových sítí atd. Pro efektivní zpracování je také důležité rozlišení obrázků. Zde se využívá přístup Progressive resolution refinement, který využívá malé časové složitosti při práci s hrubým rozlišením a při ověřování pomocí jemnějšího rozlišení je sice složitost větší, ovšem již nepracujeme s takovými objemy dat. Snahou je také upravit algoritmy tak aby výsledek obsahoval větší hodnotu informace, jde tedy o různé spojování podobných asociačních pravidel apod. Je třeba uvést, že oblast získávání asociačních pravidel z obrazových databází je velmi mladým odvětvím a hlavní rozvoj této oblasti teprve přichází se zpřístupněním multimediálních dat a internetu masové veřejnosti. Aplikace procesu vyhledávání asociačních pravidel v obrazových databázích je již dnes velmi důležitá pro předpověď počasí, magnetickou resonanci nebo v lékařství např. pro mamografii. Možnosti jsou však daleko větší a uplatnění je, s nadsázkou řečeno, omezeno naší fantazii. Nicméně vyvinout dobrý dolovací systém je nesmírně těžké právě proto, že jde o spojení několika vědních oborů a je třeba odborníků s celkovým přehledem.

LITERATURA

[1] Zeman, D.: Ročníkový projekt, Asociační pravidla v obrazových databázích, FIT VUT Brno 2001

[2] Zaiane, O. R., Han, J., Zhu, H., Mining Recurent Items in Multimedia with Progressive Resolution Refinement, Proc. 2000 Int. Conf. on Data Engineering (ICDE'OO), San Diego, CA, March 2000. Dostupne na adrese ftp://*ftp.fas*.sfu.ca/publ cs/han/pdflicdeOO.pdf

[3] Han, J., Kamber, M., Data Mining: Concepts and Techniques, Morgan Kaufrnann Publisher, 2000. Dostupné na ústavu, slajdy pro výuku dostupné též na adrese http://*www* .cs.sfu.ca/~han/DM-Book.html

[4] Kotásek, P., Článek Dolování dat z databází, FIT VUT Brno 2000

[5] Ondryas, Z.: Diplomová práce, FIT VUT Brno 2002

[6] Grosky, I. W., Stanchev, P. L.: An Image data model

[7] Stanchev, L. P.: General Image Retrieval model