MINING ASSOCIATION RULES FROM TEXTURE DATABASES

Petr CHMELAŘ, Master Degree Programme (5)
Dept. of Information Systems, FIT, BUT
E-mail: xchmel04@stud.fit.vutbr.cz

Supervised by: Ing. Martin Heckel

ABSTRACT

This paper describes a new approach to database technology, automatic description of image content using data mining algorithms. Textures are characterized as a set of association rules, capturing co-occurrence of major data items, its statistical and structural information.

1 ÚVOD

Jsme schopni sbírat čím dál více dat až neuvěřitelně efektivně, ať už jde o ty, které potřebujeme a chceme, ale také ty, které nás nezajímají, což vyjadřuje obrat: „Topíme se v datech, ale žízníme po informacích." Potřebujeme najít spolehlivý přístup, kterým lze data automaticky zpracovávat, klasifikovat, objevovat zajímavé vlastnosti a vztahy či změny a z nich vybírat jen to podstavné, co nás zajímá. Tato práce se zaměřuje na multimediální data, která vytváří zvláště velké databáze, konkrétně na jejich obrazovou část.

Pro popis obsahu obrázku se využívá vizuálních vlastností obsahu – rysů. Textura je vedle barvy a tvarů důležitá vlastnost obrazových dat, je tvořena opakujícími se elementy, kterým se říká primitiva. Jeví se důležitá pro prohlížení velkých obrazových databází a má nespočet možných aplikací v praxi. Většina povrchů reálného světa má svoji texturu a každý úspěšný vizuální systém si musí poradit se zpracováním těchto informací. V této práci je popsán přístup identifikující textury pomocí asociačních pravidel.

2 ZÍSKÁVÁNÍ ZNALOSTÍ Z MULTIMEDIÁLNÍCH DATABÁZÍ

„Proč chceme získávat znalosti z databází?" Protože je potřebujeme, přirozený vývoj databázových technologií na popud průmyslu, obchodu i vědy dospěl do stadia, kdy je nutné efektivně popsat různorodé typy dat - temporální, dokumenty, prostorová a multimediální data, z nichž každý má svoje specifika, aplikace, ale i problémy.

Nabízejí se dva přístupy k identifikaci multimediálních dat - dle popisu dat nebo naopak z jejich obsahu, což se obecně jeví jako efektivnější a lze pro tento úkol použít několik metod pro dolování dat jako je diskriminace a charakterizace, shlukování nebo asociační analýzy.

2.1 ASOCIAČNÍ PRAVIDLA

„Co je to asociační analýza?“ Asociační analýza se snaží nalézt vztahy mezi analyzovanými daty. Ty se vyjadřují asociačními pravidly, které poté ukazují, které hodnoty atributů se vyskytují v dané množině dat společně. Typickým příkladem aplikace je analýza nákupního košíku, nicméně tato práce nabízí širší pohled.

Asociační pravidlo je tvaru $A \Rightarrow B$, kde A, B jsou množiny položek $\{i_1, i_2, \ldots, i_m\}$, vyskytující se v transakci T. Jejich kolekce tvoří transakční databázi D. Pravidlo se zapisuje:

$$i_1 \wedge i_2 \wedge \ldots \wedge i_k \Rightarrow i_{k+1} \wedge i_{k+2} \wedge \ldots \wedge i_{k+l} \quad (1)$$

Číslo $k+l$ označuje celkovou kardinalitu (mohutnost) K asociačního pravidla.

Asociační pravidla mohou být získávána z libovolného typu multimediálních dat a jejich skladů. Existují zde vztahy mezi obsahem média jako je barva, tvar, textura, tón nebo pohyb a jeho popisem. Lze získat i asociace obsahu dat s jejich lokalizací v prostoru a čase.

2.2 PROBLÉM ZAJÍMAVOSTI

„Jsou všechna pravidla zajímavá?“ Lze generovat obrovské množství vzorků, to co je dělá zajímavými je, že jsou pro lidi pochopitelné, použitelné, aktuální ale hlavně platné s jistou mírou pravděpodobnosti. Existuje několik metrik, které toto zajišťují:

Pro asociační pravidlo se používá podpora S, udává použitelnost – pravděpodobnost P se kterou se v transakcích T z testované kolekce D objevují současně položky A *i* B:

$$S(A \Rightarrow B) = P(A \cup B) \quad (2)$$

Další metrikou pravidla je jeho spolehlivost C, která určuje míru jeho kvality – s jakou (podmíněnou) pravděpodobností P se vyskytuje v každé transakci T obsahující A, také B:

$$C(A \Rightarrow B) = P(B \mid A) \quad (3)$$

Pravidla, která uspokojují minimální mez S_{min} jsou frekventovaná, uspokojující S_{min} i C_{min} nazýváme silná. Tyto pravidla, po možném doplnění korelací, jsou pro nás zajímavá.

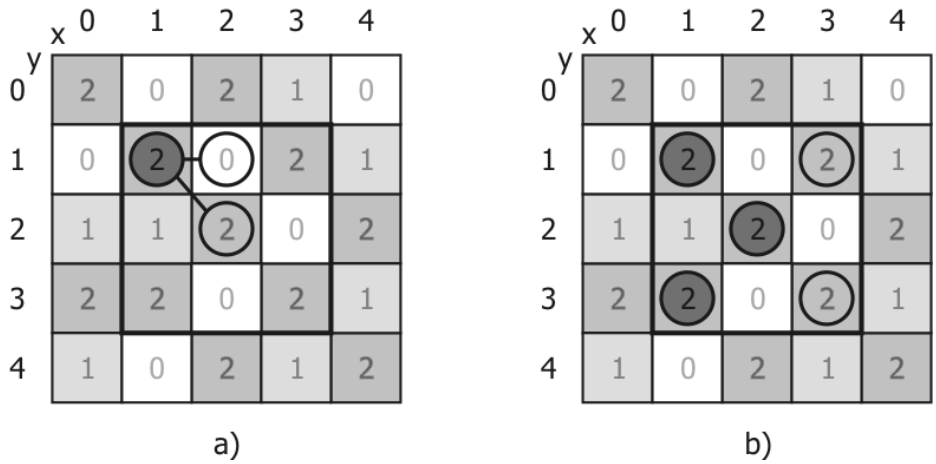
3 ANALÝZA TEXTURY

„Jak lze popsat texturu?“ Tímto problémem se zabývá texturní analýza, je to větev umělé inteligence, která čerpá i z jiných disciplín jako biologie a psychologie. Byly navrženy metody založené na statistických vlastnostech obrazu, zpracování signálů, na modelech textur a geometrické metody. Jedním z požadovaných cílů texturní analýzy je získání informace o rozmístění pixelů obrazu a jejich vzájemných vztahů. To spadá do oblasti asociační analýzy, potřebujeme ale specifikovat co je to položka a jak vypadá transakce:

- Kořenový pixel okolí $n \times n$ je ten, který je v jeho středu, jak je naznačeno v obrázku 1.
- Položka je trojice $i=(X, Y, G)$, kde X, Y je vzdálenost od kořenového pixelu, G je jeho intenzita (úroveň šedi). Položka tedy představuje libovolný pixel v daném $n \times n$ okolí.
- Transakci v texturní analýze představuje množina položek asociovaná s daným kořenovým pixelem, je to podmnožina položek v jeho okolí.

Asociační pravidlo je poté schopno zachytit často se vyskytující strukturu obrázku. V reálné aplikaci jsou intenzity pixelů kvůli absorpci mírných odchylek vhodně kvantovány,

velikost okolí například *32×32*, hodnoty S_{min} *5 %* a C_{min} *60 %*. Tvorba asociačních pravidel probíhá generováním množin frekventovaných položek, využívá vlastnosti Apriori [1].



Obr. 1: *Příklad obrázku a asociačního pravidla.*

Příklad na obrázku 1. je značně zjednodušen, jsou použity tři stupně šedi *G* v okolí velikosti *3×3* a ofset *X, Y* je omezen na *<-1, 1>*. V obrázku 1a) je pro kořenový pixel *[1, 1]* zobrazeno pravidlo $(0, 0, 2) \wedge (1, 1, 2) \Rightarrow (1, 0, 0)$ pro ilustraci data mining postupu.

{(0, 0, 2)}, S=5/9	*{(1, 0, 0)}, S=3/9*	*{(1, 1, 2)}, S=5/9*
{(0, 0, 2), (1, 0, 0)}, S=3/9	*{(0, 0, 2), (1, 1, 2)}, S=5/9*	*{(1, 0, 0), (1, 1, 2)}, S=3/9*
{(0, 0, 2), (1, 0, 0), (1, 1, 2)}, S =3/9		

Tab. 1: *Frekventované množiny použité při generování ilustračního pravidla.*

V obrázku 1 b) je vyznačeno devět kořenových pixelů, všechny tři položky se zde nalézají ve třech případech, jak je označeno tmavě, pravidlo má tedy podporu S = *3/9 = 33 %*. Levou polovinu pravidla lze najít v pěti zvýrazněných případech, společně s pravou ve třech, spolehlivost je pak C = *3/5 = 60 %*.

4 ZÁVĚR

Pro testování popsaného přístupu slouží úloha klasifikace textur z alba P. Brodatze [4], do databáze jsou uložena silná asociační pravidla a pro každou dvojici tříd textur je nalezeno pravidlo, které je nejlépe odlišuje. Metoda je implementována s optimalizacemi všech klíčových úloh a poskytuje kvalitní výsledky, srovnatelné s Gaborovými filtry [2]. Navíc zpřístupňuje další zajímavé informace z libovolné multimediální databáze.

REFERENCES

[1] Han, J., Kamber, M.: Data Mining: Concepts And Techniques, 2000

[2] Rushing, J. A., et al.: Using Association Rules as Texture Features, 2001

[3] Tuceryan, M, Jain, A.: The Handbook of Pattern Recognition and Computer Vision, 1998

[4] Brodatz, P.: Textures: A photographical album for artists and designers, 1966