Association Rule Mining in The Wider Context of Text, Images and Graphs

UNIVERSITY OF LIVERPOOL

Frans Coenen
Department of Computer Science
UKKDD'07, April 2007

PRESENTATION OVERVIEW

- Motivation.
- Association Rule Mining (quick overview).
- Challenges of wider ARM application.
- Text Mining
- Image Mining
- Graph Mining
- Image Graph Mining
- Conclusions

MOTIVATION

- Association Rule Mining (ARM) is a well established DM mechanism.
- The initial concept of ARM has been extended in a number of technical directions: Incremental ARM, Utility Mining, Unique pattern mining, Weighted ARM, Classification ARM, Distributed/Parallel ARM.
- We would also like to apply ARM technology to non-standard data sets such as document collections, image sets, graphs (i.e. non-tabular data sets).

WE LIKE ARM!

- Process is easy to understand (and therefore easy to explain to end users).
- Computationally efficient compared to many other DM techniques.
- Can cope with data sets that have very high dimensionality.
- Can cope with data sets that have many missing values.
- Results are expressed in an easy to understand rule format.

CHALLENGE OF NON-STANDARD DATA-SETS

- The challenge of applying ARM to non-standard data is to translate the data into a format that will allow the application of ARM, i.e. into a vector format.
- We can of course achieve text, image and graph mining using alternative techniques.
- But we like ARM and believe that it offers certain advantages!

TEXT MINING

BAG OF WORDS

- Each document can be represented as a subset of a global bag of words ⇒ Each word represents an attribute ⇒ Too many attributes.
- Limit number of words in bag using stemming and lists of stop words.
- Selecting only key words ⇒ How obtained?
  1) Statistically (e.g. TF-IDF).
  2) Using given vocabularies/dictionaries ⇒ How do we obtain these?
  3) Natural language parsing (NLP).

BAG OF PHRASES

- In the bag of words approach we lose information with regard to word ordering.
- The bag of phrases approach goes some way to addressing this ⇒ How do we I.D. phrases.
  1) Statistically.
  2) Word grams.
  3) NLP.

SOME IDEAS

- Identify *noise*, *stop* and *significant* words (all other words are *ordinary* words).

Delimiters	Contents
stop marks and noise words	sequence of one or more significant words and ordinary words
	sequence of one or more significant words and ordinary words replaced by "wild cards"
stop marks and ordinary words	sequence of one or more significant words and noise words
	sequence of one or more significant words and noise words replaced by "wild cards"

Coenen, Leng, Sanderson and Wang (2007)

IMAGE MINING

IMAGE REPRESENTATION FOR DATA MING

- Images are made up of pixel data where each pixel has two fields: (i) colour (RGB values), and (ii) relative location in a 2-D plane.
- We wish to represent images in some way that will support ARM.
- Too computationally intensive to work with all colours and all pixels.
- Need to adopt some approach to reduce the computational overhead by reducing the amount of data we are working with but without loosing too much image information.

IMAGE REPRESENTATION FOR DATA MING cont.

- Some ideas:

1. Convert to a 8 or 4 bit colour (256 or 16 colours respectively).
2. Convert to luminance (brightness/gray) scale and limit the number of luminance values.
3. Tesselation.
4. Quad trees.
5. Image segmentation.
4. Image primitives
5. Textual descriptions.
6. Conceptual hierarchies.

AN IDEA (IMAGE PRIMITIVES)

- Define images in terms of primitives ⇒ primitives= attributes.
- Primitives may be identified by first identifying *blobs* in the image using an appropriate segmentation technique.
- Match blobs to primitive shapes using a dictionary of primitives.

GRAPH MINING

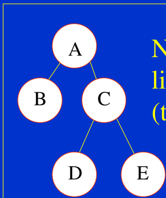
GRAPH MINING

- Graph mining, and especially mining for frequent patterns in graphs, can be catagorised as follows:

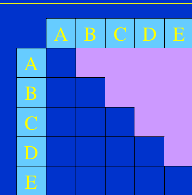
Transaction Graph Mining where the dataset comprises a collection of small graphs called examples. The goal is then to find frequent patterns that exist across the "transactions".

Single Graph Mining where the data set comprises a single large graph. The objective is then to discover frequently occurring patterns within this single graph.

ADJACENCY MATRIX



Node and link graph (tree).



Adjacency matrix

```
Attribute set = {AB, AC, AD, AE,
BC, BD, BE, CD, CE, DE}
```

IMAGE GRAPH MINING

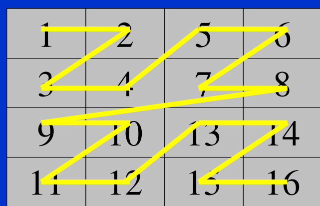
IMAGE CURVES

- Given an tiled (tesselated) image we can sequentially number the tiles so that the entire image can be unraveled into a “single strand”.
- Example:

1	2	5	6
3	4	7	8
9	10	13	14
11	12	15	16

IMAGE CURVES

- Given an tiled (tesselated) image we can sequentially number the tiles so that the entire image can be unraveled into a “single strand”.
- Example:

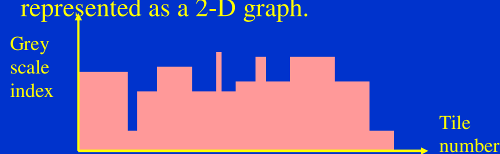


CONCEPT HIERARCHIES

- Given a set of image primitives these can be arranged in a concept hierarchy with the edges representing spatial relations (e.g. above, below, contains, etc).
- How we identify such relations may be problematic.
- But ideas behind adjacency matrix may be applicable.

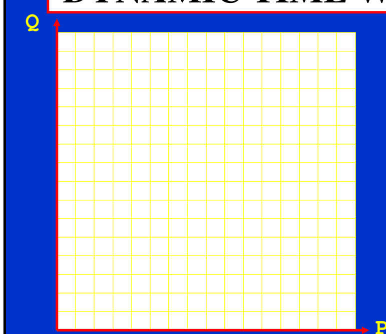
IMAGE CURVES

- If we convert the “tiled” images to gray scale (luminance) the image curve can be represented as a 2-D graph.

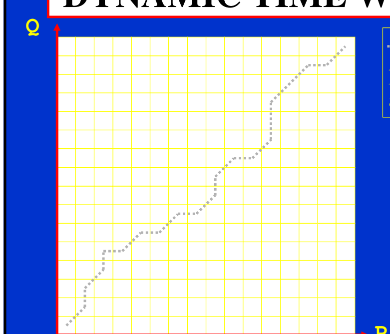


- An image curve of this form has similarities with time series curves therefore can be mined using time series analysis techniques.
- Fore example Dynamic Time Warping (DTW).

DYNAMIC TIME WARPING

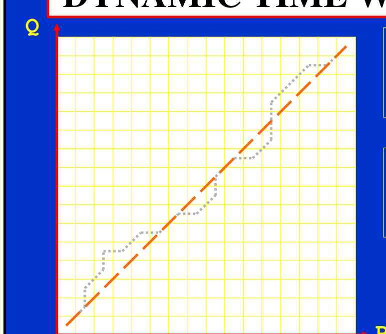


DYNAMIC TIME WARPING



.... = Best fit path between P and Q .

DYNAMIC TIME WARPING



.... = Best fit path between P and Q .

— — = Path if P and Q are the same.

SUMMARY AND CONCLUSUONS

- Motivation.
- Association Rule Mining (quick overview).
- Challenges of wider ARM application.
- Text Mining
- Image Mining
- Graph Mining
- Image Graph Mining
- Conclusions