Mining spatial association rules from document layout structures

Margherita Berardi Michelangelo Ceci Donato Malerba
Dipartimento di Informatica – Università degli Studi di Bari
via Orabona 4 - 70126 Bari
{berardi, ceci, malerba}@di.uniba.it

Abstract

In this paper we investigate the discovery of spatial association rules from a particular kind of images, namely document images. Document images are initially processed to extract both their layout structures and their logical structures. To take into account the inherent spatial nature of the layout structure, a spatial data mining algorithm is applied, which returns spatial association rules. We present possible applications of spatial association rules detected from document layout. We also illustrate and comment experimental results on a set of multi-page documents extracted by IEEE PAMI.

1. Introduction

The discovery of association rules has attracted a great deal of attention in data mining research. Association rules are a class of regularities introduced by [1] that can be expressed by an implication:

$$X \rightarrow Y$$

where X and Y are a sets of *items*, such that $X \cap Y = \emptyset$. The meaning of such rules is quite intuitive: Given a database D be of transactions, where each *transaction* $T \in D$ is a set of items, $X \rightarrow Y$ expresses that whenever a transaction T contains X than T probably contains Y also. The conjunction $X \wedge Y$ is called *pattern*.

Two parameters are usually reported for association rules, namely the *support*, which estimates the probability $p(X \in T \wedge Y \in T)$, and the *confidence*, which estimates the probability $p(Y \in T \mid X \in T)$. The goal of association rule mining is to find all the rules with support and confidence exceeding user specified thresholds, henceforth called *minsup* and *minconf* respectively. A pattern $X \wedge Y$ is *large* (or *frequent*) if its support is greater than or equal to *minsup*. An association rule $X \rightarrow Y$ is *strong* if it has a large support (i.e. $X \wedge Y$ is large) and high confidence.

The traditional application of association rules is in the business world, where they are used to take more precise marketing actions on the basis of what products are frequently bought together. In this application, the items are products and the transactions are customer purchases at the checkout. However, it is becoming clear that association rules are not restricted to market basket analysis, but can be successfully applied to a wide range of domains, such as web access patterns discovery [3] and building intrusion detection models [6].

An interesting application is faced in the work by Ordonez and Omiecinski [10] where a method for mining knowledge from images is proposed. The method is an association rule miner that automatically identifies similarities in images on the basis of their content. The content is expressed in terms of objects automatically recognized in a segmented image. The work shows that even without domain knowledge it is possible to automatically extract some reliable knowledge. Mined association rules refer to the presence/absence of an object in an image, since images are viewed as transactions while objects as items. No spatial relationship between objects in the same image is considered.

In this paper we investigate the discovery of association rules from a particular kind of images, namely document images. Document images are initially processed to extract their layout structures, which describe the geometrical arrangement of content portions on a page. Then the logical structures are extracted. The logical structure of a document image consists of a hierarchy of segments of the document, each of which corresponds to a visually distinguished semantic component of the document (e.g., title, paragraph, caption or heading) [12]. Some layout components with no visually distinguished semantic are labelled as *undefined*. The extraction of both layout and logical structures is performed by means of the system WISDOM++ [2], whose main characteristic is the application of machine learning algorithms in several document processing steps. In WISDOM++ documents are grouped into classes (e.g. paper published on IEEE Transactions of Pattern Analysis and Machine Intelligence), such that document images in the same class show approximately the same layout/logical structure.

The discovery of association rules follows the processes of layout structure extraction (layout analysis)

and logical structure extraction (document image understanding). We are interested in association rules expressing regularities among logical components of a set of document images belonging to the same class.

2. The approach

Differently from the work by Ordonez and Omiecinski [10], we also intend to take into account the inherent spatial nature of the layout structure, that is, we intend to discover, if any, spatial patterns between logical components. Therefore, association rule mining methods developed in the context of spatial data mining are considered [9].

There are three main peculiarities of the proposed approach. First, the spatial property of logical components is considered. Logical components are described in terms of their content type (e.g., text, graphics, etc.), their logical meaning (e.g., title, authors, etc.) and their geometrical features, which can be either relational or attributional. Relational features relate two logical components on the basis of their mutual position in the document (e.g. *on_top(A,B)*, *to_right(A,B)*). Attributional features refer to geometrical properties of layout components, such as width and height, as well as locational properties (position along x/y axis).

Second, the hierarchical structure of logical components is considered as well. It is possible to look at the set of logical components of a document image as a hierarchy where each single logical component is related to another one by a *is_a* or a *part_of* relation (e.g *title* is part_of *identification*, *page_number* is_a *page_component*). The levels in the hierarchy are called granularity levels.

Third, the logical components can play different roles. Indeed, in spatial data mining attributes of some spatial objects in the neighborhood of, or contained in, a unit of analysis [1] may affect the values taken by attributes of the unit of analysis. Therefore, it is necessary to distinguish units of analysis, which are the *reference* objects of an analysis, from other *task-relevant* spatial objects, and it is important to represent interactions between them. In our application, some logical components play the role of *reference objects* while other logical components play the role of *task relevant objects*.

To mine spatial association rules we use SPADA (Spatial Pattern Discovery Algorithm) [9] which is based on a multi-relational data mining approach and permits the extraction of multi-level spatial association rules, that is, association rules involving spatial objects at different granularity levels. An example of association rule discovered by SPADA is:

is_a(A,running_head) →
on_top(A,B) , *is_a(B,content)* ,*type_text(A)*
support: 90.9% confidence: 90.9%

This rule means that if a logical component (*A*) is a *running_head* then it is textual and it is on top of another layout component (*B*) which is a component of type *content*. This rule has a high support and a high confidence (both expressed as percentages).

3. Using association rules

The extracted association rules can be used in a number of ways. First, new documents can be recognized as satisfying the constraints that define the domain template (document classification and retrieval). Indeed, recent approaches propose to use discovered association rules for classification tasks [7].

Second, the rules can be profitably used in the automatic layout correction. Currently, in WISDOM++ the automatic correction of the layout is performed by means of a set of rules learned from the sequence of user actions [8]. By formulating the problem as a planning problem, it is necessary to define both a goal and a metric evaluating the distance between the current state (layout structure) and the goal. This metric can be based on the number of association rules supported by the extracted layout structure.

Third, the rules can be also used in a generative way. For instance, if a part of the document is hidden or missing, strong spatial association rules can be used to predict the location of missing layout/logical components [5]. Moreover, a desirable property of a system that automatically generates textual documents is to take into account the layout specification during the generation process, since layout and wording generally interact [11]. Spatial association rules can be useful to define the layout specifications of such a system. Finally, this problem is also related to document reformatting [4].

4. Mining spatial association rules

The problem of mining association rules by means of SPADA can be formally stated as follows:

Given

- a set of descriptions of the labelled documents (result of the document understanding)
- a set of reference objects S,
- some sets R_k, $1 \le k \le m$, of task-relevant objects,

[1] The unit of analysis is the basic entity or object about which generalizations are to be made based on an analysis and for which data are collected in the form of variables.

- a background knowledge *BK* including some spatial hierarchies H_k on objects in R_k and a domain specific knowledge,
- *M* granularity levels in the descriptions (1 is the highest while M is the lowest),
- a set of granularity assignments ψ_k which associate each object in H_k with a granularity level,
- a couple of thresholds *minsup*[*l*] and *minconf*[*l*] for each granularity level,
- a declarative bias *DB* that constrains the search space,

Find

strong multi-level spatial association rules.

The set of descriptions of the labelled documents is expressed in the form of first-order logic conditions. The use of the first order logic is necessary because the feature-vector representation, typically adopted in statistical approaches, cannot render the relational features.

Spatial features (relations and attributes) are used to describe the logical structure of a document image. In particular, we mention *locational* features such as the coordinates of the centroid of a logical component (*x_pos_center*, *y_pos_center*), *geometrical* features such as the dimensions of a logical component (*width*, *height*), and *topological* features such as relations between two components (*on_top*, *to_right*, *alignment*). We use the *non-spatial* feature *type_of* that specifies the content type of a logical component (e.g. *image*, *text*, *horizontal line*). In addition there are other non-spatial features, called *logical* features which define the label associated to the logical components. They are: *affiliation*, *page_number*, *figure*,*caption*, *index_term*, *running_head*, *author*, *title*, *abstract*, *formulae*, *subsection_title*, *section_title*, *biografy*, *references*, *paragraph*, *table*, *undefined*. In the following we present an example of document description on which run SPADA:

```
class(h,tpami),
is_a(a,running_head), is_a(b,title),...
page(h,first),
part_of(h,a),part_of(h,b)=true,...
width(a,390),width(b,490),...
height(a,7),height(b,54),...
type_text(a),type_text(b),...
x_pos_centre(a,215),x_pos_centre(b,288),...
y_pos_centre(a,26),y_pos_centre(b,83),...
on_top(a,b),on_top(b,c),on_top(b,d),...
to_right(e,f),...
only_left_col(a,l),only_left_col(b,e),
only_right_col(b,e),only_middle_col(b,e),
...
```

where *h* represents the page and *a*,…,*g* represent the logical components of the page. It is noteworthy that the features *class* and *page* are used to describe the class and the order page and the relation *part_of* is used to express the membership of a component to a page. Numerical features are automatically discretized before inferring spatial association rules.

The specification, by means of a set of rules, of the following domain specific knowledge:

```
at_page(X,first) <- part_of(Y,X),
page(Y,first)
at_page(X,intermediate) <-
        part_of(Y,X),page(Y,intermediate)
at_page(X,last_but_one) <-
       part_of(Y,X), page(Y,last_but_one)
at_page(X,last)       <-       part_of(Y,X),
page(Y,last)
```

permits to automatically associate information on page order to layout components. Since the presence of some logical components may depend on the page order (e.g. *author* is in the first page), the above rules allows SPADA to discover associations where this information is taken into account. The specification of the hierarchy (Figure 1) allows the system to extract spatial association rules at different granularity size.

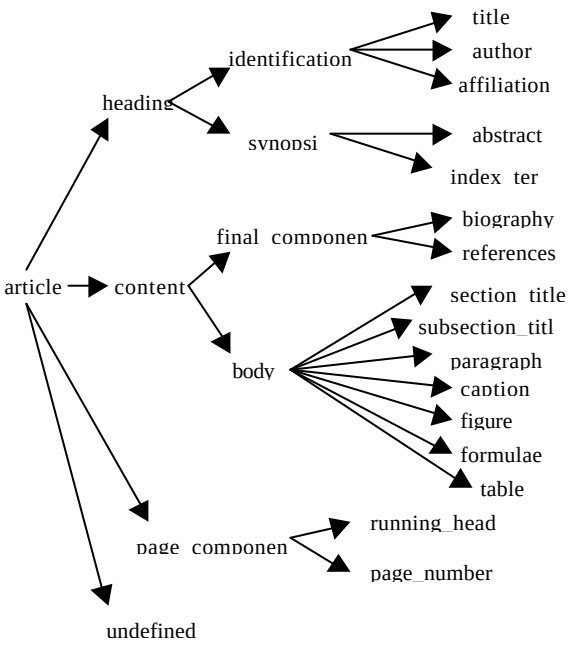

Figure 1. Hierarchy of logical components.

The declarative bias *DB* constrains the search space and aims at defining the *reference objects* (*ro*) and *task relevant objects* (*tro*). In our task, the *ro* are the logical components to which at least one satisfied logical feature, different from *undefined*, is associated. The *tro* are all the logical components.

5. Experimental results

To investigate the applicability of the proposed solution we considered six papers, published as either regular or short, in the IEEE Transactions on Pattern Analysis and Machine Intelligence, in the January and February 1996 issues. Each paper is a multi-page document and has a variable number of pages and layout components for page. A user of WISDOM++ labels some layout components of this set of documents according to their logical meaning. Those layout components with no clear logical meaning are labelled as *undefined*. All logical labels belong to the lowest level of the hierarchy reported in the previous section. We processed 54 document images in all.
In Table 1 logical components distribution on the processed documents is shown.

Table 1. Labels distribution.

Document ID / Label	***1***	***3***	***4***	***6***	***7***	***9***	***Total***
affiliation	1	1	0	1	2	2	7
page_number	8	13	12	1	5	5	44
figure	19	13	12	3	12	12	71
caption	13	17	7	3	11	5	56
index_term	1	1	1	0	0	0	3
running_head	14	15	14	1	6	5	55
Author	1	1	2	1	1	1	7
Title	1	1	1	1	1	1	6
abstract	1	1	1	1	1	1	6
formulae	24	19	21	0	4	5	73
subsection_title	3	1	1	0	0	0	5
section_title	4	4	2	0	1	1	12
biografy	2	1	1	0	0	0	4
references	3	3	2	1	1	2	12
paragraph	54	55	50	3	19	21	202
table	0	9	1	0	2	1	13
undefined	21	26	27	10	14	16	114

The number of features to describe the six documents presented to SPADA is 17,880, about 331 features for each page document. The total number of logical components is 690 (114 of which are *undefined*) about 318 descriptors for each page document.
An example of association rule discovered by SPADA is:
is_a(A,author) $\rightarrow$ *only_middle_col(A,B)* ,
is_a(B,heading), height(B,[1..174]), type_text(A)
support: 85.71% confidence: 85.71%
The spatial pattern of this rule involves six out of seven (i.e. 85.71%) blocks labelled as authors. This means that six logical components which represent the author of some paper are textual components vertically centered with a logical component B at the heading of the paper, with height between 1 and 174.
At a lower granularity level, a similar rule is found where the logical component B is specialized as abstract:
is_a(A,author) $\rightarrow$ *only_middle_col (A,B),*
is_a(B,abstract), height(B,[1..174]), type_text(A)
support: 85.71% confidence: 85.71%
The rule has the same confidence and support reported for the rule inferred at the first granularity level.

Conclusions

This work presents an application of spatial data mining techniques to the problem of finding associations between logical components extracted from document images by means of document analysis and understanding methods. As future work, we intend to investigate the application of mined association rules in three different contexts, namely document classification and retrieval, automated layout correction, and automated generation of documents.

Acknowledgements

This work fulfills the research objectives set by the IST-1999-20882 project COLLATE (Collaboratory for Annotation, Indexing and Retrieval of Digitized Historical Archive Material) funded by the European Union.

References

1. R. Agrawal, and R. Srikant, Fast Algorithms for Mining Association Rules.*Proc.20th Int.Conf.Very Large Data Bases*, VLDB, 1994
2. O. Altamura, F. Esposito, and D. Malerba, Transforming paper documents into XML format with WISDOM++, *Int. Journal on Document Analysis and Recognition*, 4(1), 2-17, 2001.
3. M. S. Chen, J. S. Park, and P. S. Yu, Data Mining for Path Traversal Patterns in a Web Environment, *In Proceedings of the 16th International Conference on Distributed Computing Systems*, IEEE press, in 385-392, 1996.
4. L. Hardman, L. Rutledge, and D. Bulterman, Automated generation of hypermedia presentation from pre-existing tagged media objects, *Proc. Of the 2nd. Workshop on Adaptive Hypertext and Hypermedia*, 1998.

5. K. Hiraki, J.H. Gennari, Y. Yamamoto, and Y. Anzai, Learning Spatial Relations from Images, *Machine Learning Workshop*, Chicago, pages 407—411, 1991.
6. W. Lee, S. Stolfo, and K. Mok, Mining audit data to build intrusion detection models. *In KDD-98*, Agrawal, Stolorz, and PiatetskyShapiro, Eds., AAAI Press, pp. 66—72, 1998.
7. B. Liu, W. Hsu, and Y. Ma., Integrating classification and association rule mining. *In KDD'98*, New York, NY, 1998.
8. D. Malerba, F. Esposito, O. Altamura, M. Ceci, and M. Berardi, Correcting the Document Layout: A Machine Learning Approach, ICDAR 2003.
9. D. Malerba, and F.A. Lisi, Discovering Associations Between Spatial Objects: An ILP Application, in C. Rouveirol & M. Sebag (Eds.), *Inductive Logic Programming*, Lecture Notes in Artificial Intelligence, 2170, Springer, Berlin, 2001.
10. C. Ordonez, and E. Omiecinski, Discovering association rules based on image content, *Proceedings of the IEEE Advances in Digital Libraries Conference 99*.
11. K. Reichenberger, K. J. Rondhuis, J. Kleinz, and J. Bateman, Effective Presentation of Information Through Page Layout: a Linguistically-Based Approach. *Proceedings of the ACM Workshop on Effective Abstractions in Multimedia*. San Francisco, California, 1995.
12. K. Summers, Toward a taxonomy of logical document structures. *In Electronic Publishing and the Information Superhighway: Proceedings of the Dartmouth Institute for Advanced Graduate Studies (DAGS)*, pages 124--133, 1995.