Discovering Knowledge through Multi-modal Association Rule Mining for Document Image Analysis

Corrado Loglisci, Michelangelo Ceci, Lynn Rudd, and Donato Malerba

Dipartimento di Informatica, Università degli Studi di Bari "Aldo Moro"
Via Orabona 4, 70125, Bari
{corrado.loglisci, michelangelo.ceci, lynnmargaret.rudd,
donato.malerba}@uniba.it

Abstract. The paper introduces a descriptive data mining method to discover knowledge for the task of automatic categorization in document image analysis. We argue that a document image is a multi-modal unit of analysis whose semantics is deduced from a combination of textual content, layout structure and logical structure. So, the method considers simultaneously different modalities of document representation, and, therefore different types of information: spatial information derived from a complex document image analysis process (layout analysis), information extracted from the logical structure of the document (by means of document image classification and understanding) and the textual information extracted by means of an OCR. The proposed method is based on a relational data mining approach to discover association rules, where the relational setting is justified, given its appropriateness to analyze data available in more than one modality. Experimental results on a real world dataset are reported.

1 Introduction

Document image analysis is the subfield of digital image processing that aims to convert document images to symbolic form for modification, storage, retrieval, reuse, and transmission. However, technologies and systems demand a large amount of domain-specific knowledge, in order to properly process document images, as well as to automatically catalog and organize them [20]. Hand-coding the necessary knowledge for document images with varied layout, such as those processed in real applications, is prohibitive. For this reason, there has been a growing interest in the application of data mining techniques in order to extract the required knowledge from document images [2],[5]. Text-mining is a technology of data mining which is attracting interest for its particular ability to analyze large collections of unstructured documents and extract interesting and non-trivial patterns or knowledge. Knowledge discovered from textual documents can be in various forms, including classification rules, which partition document collections into a given set of classes [23], clusters of similar documents or object composing documents [24], patterns describing trends, such as emerging topics

in a corpus of time-stamped documents [19], and ranking models used in the problems of order reading detection and document summarization [7].

Text mining and document image analysis have always been considered two complementary technologies: the former is appropriate for documents that are generated according to some textual format, while the latter is applicable to documents available on paper media. Document image mining aims to identify high-level spatial objects and relationships, while text mining is more concerned with patterns involving words, sentences and concepts. The possible interactions between spatial information extracted from document images and textual information, related to the content of some layout components, have never been considered in the data mining literature.

In this paper, we propose to extract the required knowledge in the form of association rules, which have been successfully applied both in the business and scientific realms. Some examples of applications to images have also been reported in the literature. In particular, Ordonez and Omiecinski [21] propose to mine association rules, in order to find similarities in images on the basis of their content. The content is expressed in terms of objects returned by a segmentation process. They show 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 are seen as items. No spatial relationship between objects in the same image is considered. Moreover, the proposed data mining method is *unimodal*, since it considers only one source of information - visual - conveyed by images. In order to deal properly with the multi-modal nature of documents we resort to relational data mining approaches [11], which permit the processing of data stored according to the relational data model. This model permits us to: *i)* naturally represent different types of data in different tables of a relational database; *ii)* relate collected data by means of foreign key constraints; *iii)* represent properly spatial relationships (e.g., topological or distance relationships) implicitly defined in the layout structure of a document. Relational data mining approaches permit us to directly analyze data stored in multiple database relations, thus preventing information loss, due to data transformation known as "propositionalization" [6].

In a generic library or archive, this kind of pattern can be used in a number of ways. First, discovered association rules can be used as constraints defining domain templates of documents both for classification tasks, such as in associative classification approaches [17] [15], and to support layout correction tasks. Second, the rules could also be used in a generative way. For instance, if a part of the document is hidden or missing, strong association rules can be used to predict the location of the missing layout/logical components [13]. 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 [22]. Association rules can be useful to define the layout specifications of such a system. Finally, this problem is also related to document reformatting [12].

The paper is organized as follows. In the next Section, background and motivations of this work are presented. The proposed solution is presented in Section 3. Finally, experiments are presented in Section 4.

2 Background and Motivations

In tasks where the goal is to uncover structure in the data and where there is no target concept, the discovery of relatively simple but frequently occurring patterns is promising. Association rules are a basic example of this kind of setting. The problem of mining association rules was originally introduced in the work [1] that can be expressed by an implication:

$$X \rightarrow Y,$$

where X and Y are sets of items called *antecedent* and *consequent* respectively ($X \cap Y = \varnothing$). The meaning of such rules is quite intuitive: Given a database D of transactions, where each transaction $T \in D$ is a set of items, $X \rightarrow Y$ expresses that whenever a transaction T contains X, then T probably also contains Y. The conjunction $X \wedge Y$ is called a pattern. Two parameters are usually reported for association rules, namely the support, which estimates the probability $p(X \subseteq T \wedge Y \subseteq T)$, and the confidence, which estimates the probability $p(Y \subseteq T | X \subseteq T)$. 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 $X \rightarrow Y$ 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., $X \rightarrow Y$ is large) and high confidence. However, it is becoming clear that these rules can be successfully applied to a wide range of domains, such as web access pattern discovery [9] and mining data streams [14]. As previously before, an interesting application is discussed in the work by [21] where a method for mining knowledge from images is proposed.

Nevertheless, mining patterns from document images raises many different issues regarding document structure, storage, access and processing. Firstly, documents are typically unstructured or, at most, semi-structured data. In the case of structured data, the associated semantics or meaning is unambiguously and implicitly defined and encapsulated in the structure of data (i.e., relational databases), whereas unstructured information meaning is only loosely implied by its form and requires several interpretation steps in order to extract the intended meaning. Endowing documents with a structure that properly encodes their semantics adds a degree of complexity in the application of the mining process. This makes the data pre-processing step really crucial.

Secondly, documents are message conveyors whose meaning is deduced from the combination of the written text, the presentation style, the context, the reported pictures and the logical structure, at least. For instance, when the logical structure and the presentation style are quite well-defined (typically when some parts are pre-printed or documents are generated by following a predefined formatting style), the reader may easily identify the document type and

locate information of interest even before reading the descriptive text (e.g., the title of a paper in the case of scientific papers or newspapers, the sender in the case of faxes, the supplier or the amount in the case of invoices, etc.). Moreover, in many contexts, illustrative images fully complement the textual information, such as diagrams in socio-economic or marketing reports. By considering typeface information, it is also possible to immediately and clearly capture the notion the historical origin of documents (e.g., medieval), as well as the cultural origin (e.g., Arabic). The presence of spurious objects may inherently define classes of documents, such as revenue stamps in the case of legal documents. The idea of considering the multi-modal nature of documents falls into the novel research trend of the document understanding field, that encourages the development of hybrid strategies for knowledge capture, in order to exploit the different sources of knowledge (e.g., text, images, layout, type style, tabular and format information) that simultaneously define the semantics of a document [10].

However, data mining has evolved following a unimodal scheme instantiated according to the type of the underlying data (text, images, etc). Applications of data mining involving hybrid knowledge representation models are still to be explored. Indeed, several works have been proposed to mine association rules from the textual dimension [4] with the goal of finding rules that express regularities concerning the presence of particular words or particular sentences in text corpora. Conversely, mining the combination of structure and content dimensions of documents has not been investigated yet in the literature, even though some emerging real-world applications require mining processes able to exploit several forms of information, such as images and captions in addition to full text [25].

Thirdly, documents are a kind of data that do not match the classical attribute-value format. In the tabular model, data are represented as fixed-length vectors of variable values describing properties, where each variable can have only a single, primitive value. Conversely, the entities (e.g., the objects composing a document image) that are observed and about which information is collected may have different properties, which can be properly modeled by as many data tables (relational data model) as the number of object types. Moreover, relationships (e.g., topological or distance relationships that are implicitly defined by the location of objects spatially distributed in a document image or words distributed in a text) among observed objects forming the same semantic unit can also be explicitly modeled in a relational database by means of tables describing the relationship. Hence, the classical attribute-value representation seems too restrictive and advanced approaches are necessary to both represent and reason in the presence of multiple relations among data.

Lastly, the data mining method should take into account external information, also called expert or domain knowledge, that can add semantics to the whole process and then obtain high-level decision support and user confidence. All these peculiarities make documents a kind of complex data that require methodological evolutions of data mining technologies, as well as the involvement of several document processing techniques. In our context, the extraction

of spatio-textual association rules requires the consideration of all these sources of complexity deriving from the inherent nature of processed documents.

3 Multi-modal Association Rule Mining

In our proposal, the system used for processing documents is WISDOM++[3]. WISDOM++ [1] is a document analysis system that can transform textual paper documents into XML format. This is performed in several steps. First, the image is segmented into basic layout components (non-overlapping rectangular blocks enclosing content portions). These layout components are classified according to their type of content (e.g., text, graphics, and horizontal/vertical lines). Second, a perceptual organization phase, called layout analysis, is performed to detect structures among blocks. The result is a tree-like structure, named layout structure, which represents the document layout at various levels of abstraction and associates the content of a document with a hierarchy of layout components, such as blocks, lines, and paragraphs. Third, the document image classification step aims at identifying the membership class (or type) of a document (e.g. censorship decision, newspaper article, etc.), and it is performed using some first-order rules which can be automatically learned from a set of training examples.

Document image understanding (or interpretation) creates a mapping of the layout structure into the logical structure, which associates the content with a hierarchy of logical components, such as title/authors of a scientific article, or the name of the censor in a censorship document. As previously pointed out, the logical and the layout structures are strongly related. For instance, the title of an article is usually located at the top of the first page of a document and it is written with the largest character set used in the document. Document image understanding also uses first-order rules [18]. Once the logical and layout structure have been mapped, OCR can be applied only to those textual components of interest for the application domain, and the content can be stored for future retrieval purposes. Thus, the system can not only determine the type of document, but is also able to identify interesting parts of a document and extract the information given in this part plus its meaning. The result of the document analysis is an XML document that makes the document image easily retrievable. Once the layout/logical structure as well as the textual content of a document have been extracted, association rules can be extracted.

The task of mining *relational* association rules is, in this work, solved by the SPADA system[16]. It represents relational data *à la* Datalog, a logic programming language with no function symbols, specifically designed to implement deductive databases. Moreover, SPADA takes into account background knowledge (BK) expressed in Prolog and handles background hierarchies over the objects to be mined. In document image understanding, hierarchies can be naturally defined, e.g., considering the organization of the logical components (Figure 1). So, association rules involving more abstract objects are better supported (although less precise), while association rules involving more specific objects have

[1] www.di.uniba.it/∼malerba/wisdom++/

higher confidence values (although lower support values). SPADA distinguishes between the set S of *reference* (or target) *objects*, which are the main subject of analysis, and the sets R_k, $1 \leq k \leq m$, of *task-relevant* (or non-target) objects, which are related to the former and can contribute to accounting for the variation. Each unit of analysis includes a distinct reference object and many related task-relevant objects. Therefore, the description of a unit of analysis consists of both properties of included reference and task-relevant objects as well as their relationships.

3.1 Document Description

In the logic framework adopted by SPADA, a relational database is boiled down into a deductive database. Properties of both reference and task-relevant objects are represented in the extensional part D_E, while the domain knowledge is expressed as a normal logic program which defines the intensional part D_I. As an example, we report a fragment of the extensional part of a deductive database D, which describes multi-modal information which can be extracted from any document image:

block(b1). block(b2). ... height(b2,[11..54]). width(b1,[7..82]). ...
on_top(b2,b1). ... on_top(b2,b3). ... part_of(b1,p1). part_of(b2,p1). page_first(p1).
... abstract(b1). title(b2). ... text_in_abstract(b1,'base'). text_in_title(b2,'model')....

In this example, $b1$ and $b2$ are two constants which denote many distinct layout components (reference objects) , while $p1$ denotes a document page (task-relevant object). Predicate *block* defines a layout component, $part\_of$ associates a block to a document page, *height* and *width* describe geometrical properties of layout components, *on_top* expresses a topological relationship between layout components, $page\_first(p1)$ refers to the position of the page in the document, *abstract* and *title* associates $b1$ and $b2$ with a logical label, $text\_in\_abstract$ and $text\_in\_title$ describe the textual content of the logical components.

The complete list of predicates is reported in Table 1. The a-spatial feature $type\_of$ specifies the content type of a layout component (e.g. image, text, horizontal line). Logical features are used to associate a logical label to a layout object and depend on the specific domain. In the case of scientific papers (considered in this work), possible logical labels are: *affiliation*, *page_number*, *figure*, *caption*, *index_term*, *running_head*, *author*, *title*, *abstract*, *formulae*, *subsection_title*, *section_title*, *biography*, *references*, *paragraph*, *table*.

Textual content is represented by means of another class of predicates, which are true when the term reported as the second argument occurs in the layout component denoted by the first argument. Terms are automatically extracted by means of a text-processing module. All terms in the textual components are tokenized and the set of obtained tokens is filtered out, in order to remove punctuation marks, numbers and tokens of less than three characters. Standard text preprocessing methods are used to: *i)* remove stop-words, such as articles, adverbs, prepositions and other frequent words; *ii)* determine equivalent stems

Layout structure	Locational features	$x\_pos\_center/2$
		$y\_pos\_center/2$
	Geometrical features	$height/2$
		$width/2$
	Topological features	$on\_top/2$
		$to\_right/2$
	Aspatial feature	$type\_of/2$
Logical structure	Logical features	application dependent (e.g., *abstract/1*)
Text	Textual features	application dependent (e.g., *text_in_abstract/2*)

Table 1. Complete list of predicates used.

(stemming), such as 'topolog' in the words 'topology' and 'topological, by means of the well-known Porter's algorithm for English texts .

Only relevant tokens are used in textual predicates. They are selected by maximizing the product $maxTF \times DF^2 \times ICF$ [8] that scores high terms appearing (possibly frequently) in a single logical component c and penalizes terms common to other logical components. More formally, let c be a logical label associated to a textual component. Let d be the bag of tokens in a component labeled with c (after the tokenization, filtering and stemming steps), w a term in d and $TF_d(w)$ the relative frequency of w in d. Then, the following statistics can be computed:

1. the maximum value $TF_c(w)$ of $TF_d(w)$ on all logical components d labeled with c;
2. the document frequency $DF_c^2(w)$, i.e., the percentage of logical components labeled with c in which the term w occurs;
3. the category frequency $CF_c(w)$, i.e., the number of labels $c' \neq c$, such that w occurs in logical components labeled with c'.

Then, the score v_i associated to the term w_i belonging to at least one of the logical components labeled with c is:

$$v_i = TF_c(w_i) \times DF_c^2(w_i) \times 1/CF_c(w_i) \tag{1}$$

According to this function, it is possible to identify a ranked list of "discriminative" terms for each of the possible labels. From this list, we select the best n_{dict} terms in $Dict_c$, where n_{dict} is a user-defined parameter. The textual dimension of each logical component d labeled as c is represented in the document description as a set of ground facts that express the presence of a term $w \in Dict_c$ in the specified logical component.

3.2 The mining step

In the setting introduced so far, the problem of mining multi-modal association rules can be formalized as follows:

Given

– a set S of *reference objects* (layout components);

- some sets R_k, $1 \leq k \leq m$ of *task-relevant objects* (layout components related to those of the reference objects);
- a background knowledge BK which includes hierarchies H_k on objects in R_k, granularity levels M in the descriptions (1 is the highest, while M is the lowest) and a set of granularity assignments Ψ_k which associate each object in H_k with a granularity level (the hierarchical organization of the layout components included in the task-relevant objects);
- a couple of thresholds $minsup[l]$ and $minconf[l]$ for each granularity level;
- a language bias LB that constrains the search space.

Find strong multi-level association rules, i.e., association rules involving objects at different granularity levels.

Hierarchies H_k define *is-a* (i.e., taxonomic) relations on task-relevant objects. Both the frequency of patterns and the strength of association rules depend on the granularity level l at which patterns/rules describe data. Therefore, a pattern P with support s at level l is *frequent* if $s \geq minsup[l]$ and all ancestors of P with respect to H_k are frequent at their corresponding levels. An association rule $Q \rightarrow R$ with support s and confidence c at level l is *strong* if the pattern $Q \cup R$ is frequent and $c \geq minconf[l]$.

The expressive power of first-order logic is exploited to specify both the background knowledge BK, such as hierarchies and domain specific knowledge, and the language bias LB. The LB is important to allow the user to specify his/her bias for interesting solutions and then to exploit this bias to improve both the efficiency of the mining process and the quality of the discovered rules. In SPADA, the language bias is expressed as a set of constraint specifications for either patterns or association rules.

```
article
+ − − heading
|   + −− identification
|    | + −− (title, author, affiliation)
|   + −− synopsis
|      + −− (abstract, index_term)
+ − − content
|   + −− final components
|    | + −− (biography, references)
|   + −− body
|      + −− (section_title, subsect_title, paragraph, caption, figure, formulae, table)
+ − − page_component
|   + −− running_head
|   + −− page_number
+ − − undefined
```

Fig. 1. Hierarchy of logical components.

4 Experiments

We investigate the applicability of the proposed solution to real-world document images. In particular, we have considered a dataset composed of 3,603 logical components extracted from twenty-four multi-page documents, which are scientific 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 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 217 document images in all. The number of features to describe the twenty-four documents presented to SPADA is 38,948, about 179 features for each document page. The total number of logical components is partitioned as follows: 23 for affiliation, 191 for page_number, 357 for figure, 202 for caption, 26 for index_term, 231 for running_head, 28 for author, 26 for title, 25 for abstract, 333 for formulae, 65 for section_title, 21 for biography, 45 for references, 968 for paragraph, 48 for table and 1014 for undefined. About 150 descriptors for each document page have been extracted. To generate textual predicates we set $n_{dict} = 50$ and we considered the following logical components: title, abstract, index_term, references and running_head. Hence, the following textual predicates have been included in the document descriptions: *text_in_title*, *text_in_abstract*, *text_in_index_term*, *text_in_references*,*text_in_running_head*. The total number of extracted textual features is 1,681. The text was read with a commercial OCR.

In Table 2 we report a comprehensive view of the association rules mined for each logical component at different granularity levels. SPADA finds associations for all logical components. In particular, many spatial patterns involve logical components in the first page of an article, such as affiliation, title, author, abstract and index_term. Indeed, the first page generally presents a more regular layout structure and contains several distinct logical components. The situation is different for references where most of the rules involve textual predicates, because of the high frequency of discriminating terms (e.g., 'vol', 'ieee').

An example of an association rule discovered by SPADA at the second granularity level (L=2) is:

$$is\_a\_block(A) \rightarrow specialize(A, B), is\_a(B, heading),\ on\_top(B, C), C \neq B, is\_a(C, heading)$$
$$text\_in\_component(C, paper). \quad \text{[supp: 38.46\% conf: 38.46\%]},$$

This rule considers both spatial and textual properties. It is interpreted as follows: a portion equal to 38.46% of the heading blocks is above another heading block which contains the term 'paper'. Usually, this term occurs in the abstract (a typical sentence is "In this paper ..."), which means that the heading block C is an abstract, while B is another logical component that usually is above the abstract (e.g., author component or title component). The percentage value of 38.46% indicates that 10 out of 26 layout components of type heading (Figure 1), in the overall initial set of 3,603 layout components, match the association rule

reported above. Indeed, at a lower granularity level (L=4), SPADA discovers the following rule:

$is\_a\_block(A) \rightarrow specialize(A, B), is\_a(B, title),\ on\_top(B, C), C \neq B, is\_a(C, abstract),$
$text\_in\_component(C, paper).$ [supp: 38.46% conf: 38.46%],

The rule has the same confidence and support reported for the rule inferred at the first granularity level. This means that all heading components represented by B in the former rule (L=2) are titles. Another example of a discovered rule is:

$is\_a\_block(A) \rightarrow specialize(A, B), is\_a(B, references),\ type\_text(B), at\_page\_last(B).$
[supp: 46.66% conf: 46.66%],

which shows the use of the predicate at_page_last(B) introduced in the BK. This is an example of a pure spatial association rule. The rules reported above have only one predicate on the antecedent side. An example of rules with several predicates on the antecedent side is:

$is\_a\_block(A), specialize(A, B), is\_a(B, heading) \rightarrow specialize(A, C), is\_a(C, heading),$
$at\_page\_first(B).$ [supp: 100.0% conf: 100.0%],

which has been discovered at the granularity level L=2. It is characterized by the highest value of support (since it matches 23 out 23 affiliation components) and the highest value of confidence, which indicates a very strong implication between the set of predicates on the antecedent and the set of predicates on the consequent. This rule reports that whenever there is a heading block B (antecedent), then there is another heading block C and block B is collocated on the first page. The information associated to this rule is probably too general, since different types of components could satisfy it. To obtain more specific indications, we can explore the lower levels of granularity and use specialized patterns. The rule reported above is refined at the level L=4 with the following

$is\_a\_block(A), specialize(A, B), is\_a(B, affiliation) \rightarrow specialize(A, C), is\_a(C, affiliation),$
$at\_page\_first(B).$ [supp: 100.0% conf: 100.0%],

which keeps the same statistical parameters and states that whenever there is an affiliation block B, then it is collocated on the first page and there is another affiliation block C, which could be typically the affiliation details of a co-author. Finally, an example of pure a textual association rule discovered by SPADA is:

$is\_a\_block(A) \rightarrow specialize(A, B), is\_a(B, index\_term),\ text\_in\_component(B, index).$
[supp: 92.0% conf: 92.0%],

which simply states that a logical component index term contains the term "index".

5 Conclusions

Automated extraction of knowledge patterns from document images can boost the application of document analysis systems in various contexts. In this paper we investigate a particular form of knowledge pattern, namely association rules, which are useful for many knowledge-intensive tasks, such as document classification and indexing, document reformatting and document reconstruction. The proposed method is based on a multi-relational approach, in order to consider

No of Rules	Level 1	Level 2	Level 3	Level 4
min_conf	0.3	0.3	0.3	0.3
min_supp	0.3	0.3	0.3	0.3
Affiliation	18	18	18	18
Page_Number	62	62	61	0
Figure	26	26	26	23
Caption	33	33	33	33
Index_term	45	45	45	51
Running_head	184	184	184	0
Author	30	30	30	30
Title	27	27	32	32
Abstract	103	101	101	101
Formulae	26	26	25	28
Section_Title	23	23	23	23
Biografy	23	23	23	23
References	266	265	256	256
Table	30	30	30	18

Table 2. Number of extracted association rules per logical label.

the inherent multi-modality of documents, which convey layout, logical and textual information. Moreover, knowledge patterns are extracted at various levels of granularity. Experiments prove the viability of the proposed approach.

Acknowledgments

This work partially fulfills the project "PON020056-33489339 PUGLIA@SERVICE - Internet-based Service Engineering enabling Smart Territory structural development" funded by the Italian Ministry of University and Research (MIUR) and the project "MAESTRA - Learning from Massive, Incompletely annotated and Structured Data" (Grant number ICT-2013-612944) funded by the European Commission.

References

1. R. Agrawal, T. Imielinski, and A. N. Swami. Mining association rules between sets of items in large databases. In P. Buneman and S. Jajodia, editors, *International Conference on Management of Data*, pages 207–216, 1993.
2. M. Aiello, C. Monz, L. Todoran, and M. Worring. Document understanding for a broad class of documents. *International Journal on Document Analysis and Recognition-IJDAR*, 5(1):1–16, 2002.
3. O. Altamura, F. Esposito, and D. Malerba. Transforming paper documents into XML format with WISDOM++. *IJDAR*, 4(1):2–17, 2001.
4. A. Amir, Y. Aumann, R. Feldman, and M. Fresko. Maximal association rules: A tool for mining associations in text. *J. Intell. Inf. Syst.*, 25(3):333–345, 2005.
5. M. Berardi, M. Ceci, and D. Malerba. Mining spatial association rules from document layout structures. In *Proceedings of the 3rd Workshop on Document Layout Interpretation and its Application, DLIA03*, pages 9–13, 2003.
6. M. Ceci, A. Appice, and D. Malerba. Mr-SBC: a multi-relational naive bayes classifier. In *Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases*, volume 2838 of *LNAI*, pages 95–106. Springer-Verlag, 2003.

7. M. Ceci, C. Loglisci, and L. Macchia. Ranking sentences for keyphrase extraction: A relational data mining approach. In *Pushing the Boundaries of the Digital Libraries Field - 10th Italian Research Conference on Digital Libraries, IRCDL 2014, Padua, Italy, January 30-31, 2014*, pages 52–59, 2014.
8. M. Ceci and D. Malerba. Classifying web documents in a hierarchy of categories: a comprehensive study. *J. Intell. Inf. Syst.*, 28(1):37–78, 2007.
9. M. S. Chen, J. S. Park, and P. S. Yu. Data mining for path traversal patterns in a web environment. In *ICDCS '96: Proceedings of the 16th International Conference on Distributed Computing Systems (ICDCS '96)*, page 385, Washington, DC, USA, 1996. IEEE Computer Society.
10. A. Dengel. Making documents work: Challenges for document understanding. In *ICDAR*, pages 1026–. IEEE Computer Society, 2003.
11. S. Džeroski and N. Lavrač. *Relational Data Mining*. Springer-Verlag, 2001.
12. L. Hardman, L. Rutledge, and D. Bulterman. Automated generation of hypermedia presentation from pre-existing tagged media objects. In *In Proc. of the Second Workshop on Adaptive Hypertext and Hypermedia*, 1998.
13. K. Hiraki, J. H. Gennari, Y. Yamamoto, and Y. Anzai. Learning spatial relations from images. In *ML*, pages 407–411, 1991.
14. N. Jiang and L. Gruenwald. Research issues in data stream association rule mining. *SIGMOD Rec.*, 35(1):14–19, 2006.
15. B. Li, N. Sugandh, E. V. Garcia, and A. Ram. Adapting associative classification to text categorization. In P. R. King and S. J. Simske, editors, *ACM Symposium on Document Engineering*, pages 205–208. ACM, 2007.
16. F. A. Lisi and D. Malerba. Inducing multi-level association rules from multiple relations. *Machine Learning*, 55(2):175–210, 2004.
17. B. Liu, W. Hsu, and Y. Ma. Integrating classification and association rule mining. In *Knowledge Discovery and Data Mining*, pages 80–86, 1998.
18. D. Malerba. Learning recursive theories in the normal ilp setting. *Fundamenta Informaticae*, 57(1):39–77, 2003.
19. Q. Mei and C. Zhai. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *KDD '05: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 198–207, New York, NY, USA, 2005. ACM.
20. G. Nagy. Twenty years of document image analysis in pami. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(1):38–62, 2000.
21. C. Ordonez and E. Omiecinski. Discovering association rules based on image content. In *ADL '99: Proceedings of the IEEE Forum on Research and Technology Advances in Digital Libraries*, page 38, Washington, DC, USA, 1999. IEEE Computer Society.
22. K. Reichenberger, K. J. Rondhuis, J. Kleinz, and J. A. Bateman. Effective presentation of information through page layout: a linguistically-based approach. In *In Proc. of ACM Workshop on Effective Abstractions in Multimedia, Layout and Interaction, Association for Computing Machinery*, 1995.
23. F. Sebastiani. Introduction: Special issue on the 25th european conference on information retrieval research. *Inf. Retr.*, 7(3-4):235–237, 2004.
24. M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In *Proceedings of KDD-2000 Workshop on Text Mining*, 2000.
25. A. S. Yeh, L. Hirschman, and A. A. Morgan. Evaluation of text data mining for database curation: lessons learned from the kdd challenge cup. *CoRR*, cs.CL/0308032, 2003.