Textual Information Extraction in Document Images Guided by a Concept Lattice

Cynthia Pitou[1,2] and Jean Diatta[1]

[1] EA2525-LIM, Saint-Denis de La Réunion, F-97490, France
[2] Groupe Austral Assistance, 16 rue Albert Lougnon, 97490 Sainte Clotilde, France
cpitou@gaa.fr jean.diatta@univ-reunion.fr

Abstract. Text Information Extraction in images is concerned with extracting the relevant text data from a collection of document images. It consists in localizing (determining the location) and recognizing (transforming into plain text) text contained in document images. In this work we present a textual information extraction model consisting in a set of prototype regions along with pathways for browsing through these prototype regions. The proposed model is constructed in four steps: (1) produce synthetic invoice data containing the textual information of interest, along with their spatial positions; (2) partition the produced data; (3) derive the prototype regions from the obtained partition clusters; (4) build the concept lattice of a formal context derived from the prototype regions. Experimental results, on a corpus of 1000 real-world scanned invoices show that the proposed model improves significantly the extraction rate of an Optical Character Recognition (OCR) engine.

Keywords: textual information extraction, concept lattice, clustering

1 Introduction

Document processing is the transformation of a human understandable data in a computer system understandable format. Document analysis and understanding are the two phases of document processing. Considering a document containing lines, words and graphical objects such as logos, the analysis of such a document consists in extracting and isolating the words, lines and objects and then grouping them into blocks. The subsystem of document understanding builds relationships (to the right, left, above, below) between the blocks. A document processing system must be able to: locate textual information, identify if that information is relevant comparatively to other information contained in the document, extract that information in a computer system understandable format. For the realization of such a system, major difficulties arise from the variability of the documents characteristics, such as: the type (invoice, form, quotation, report, etc.), the layout (font, style, disposition), the language, the typography and the quality of scanning. In the literature, works in pattern recognition [16] and character recognition [28] provide solutions for textual information extraction in a computer system understandable format. Works in automatic natural language

ISBN 978-5-600-01454-1, National Research University Higher School of Economics, 2016.

processing [6] contribute to solving the problem about the detection of relevant information. This paper is concerned with scanned documents, also known as document images. We are particularly interested in locating textual information in invoice images. Invoices are largely used and well regulated documents, but not unified. They contain mandatory information (invoice number, unique identifier of the issuing company, VAT amount, net amount, etc.) which, depending on the issuer, can take various locations in the document. For instance, it seems difficult to identify a trend as to the position of the date and invoice number. However, similarities may occur locally for one or many information. To take an example, the amount is usually positioned at bottom-right in the French and English systems. Recent approaches such as those presented in [3, 4, 9] are specifically concerned with the extraction of information in administrative documents such as invoices. These works have in common the search, within a base, for a document similar to an input document. Each document of this base is assigned a template that lists some attributes (position, type, keywords) to be used in order to locate information contained in similar input documents. Bartoli et al. [3] propose a system of selecting, for an input document, the nearest wrapper based on a distance measure. A wrapper is an object containing information about geometric properties and textual content of elements to extract. Belaid et al. [4] propose a case-based-reasoning approach for invoice processing. Cesarini et al. [9] propose a system to process documents that can be grouped into classes. The system comprises three phases: (1) document analysis, (2) document classification, (3) document understanding.
The present paper is in the framework of region-based textual information localization and extraction [29, 30]. We present a textual information extraction model consisting in a set of prototype regions along with pathways for browsing through these prototype regions. The proposed model is constructed in four steps:

1. produce synthetic invoice data from real-world invoice images containing the textual information of interest, along with their spatial positions;
2. partition the produced data;
3. derive the prototype regions from the obtained partition clusters;
4. derive pathways for browsing through the prototype regions, from the concept lattice of a suitably defined formal context;

The paper is organized as follows. Section 2 is devoted to the construction of prototype regions. The formal context defined using the obtained prototype regions, and the determination of paths from the concept lattice of that formal context are described in Section 3. Section 4 presents our approach for textual information extraction, using the defined paths. Finally, some experimental results are presented in Section 5 and the paper is closed with a conclusion and perspectives.

2 Construction of prototype regions

2.1 Construction of a synthetic data set

The present work is motivated by the request of a company interested in developing its own system enabling to automatically extract some textual information from scanned invoices. The company has provided us with a corpus of 1000 real-world scanned invoices, emitted by 18 service providers whose business is around car towing and auto repair. All the images are one page A4 documents. The whole set of information the company is interested in, comprises: invoice number, invoice date, net amount, VAT amount, customer reference, the type of service provided, the issuer identity. In our study, we consider only the following five information:

- I1: the key word of the service provided: towing or auto repair,
- I2: customer reference: a string of 9-13 characters,
- I3: the plate number of the assisted vehicle,
- I4: the invoice issuer unique identifier: a string of 14 digits,
- I5: the net amount of money requested for the provided service.

Each of the textual information is located in a region delimited by a rectangle defined by the coordinates (x, y) (in pixels) of its top left corner and the coordinates (z, q) of its bottom right corner. In the sequel, by the term region will be meant a rectangular area in an invoice image. Hence, a region may be represented by the four coordinates (x, y, z, q) of its top left and bottom right corners. As the information to be extracted are located in (rectangular) regions we adopt a region-based extraction approach. The regions which the proposed approach is based on are prototypes obtained from the more specific regions containing, each, a single information. Now, the coordinates of the regions containing the needed information are not available for the real-world scanned invoices at hand. To cope with this, we develop a JAVA program, with a graphical interface, enabling to create synthetic invoice data simulating the real-world scanned invoices along with the approximate coordinates of the specific rectangles containing the needed information. For instance, from a real-world scanned invoice an initial synthetic invoice is manually created. This synthetic invoice is a single black and white A4 page. This page will contain a string corresponding to a plate number approximately at the same location as the plate number information appears in the real-world invoice. Additionally, the string will be inserted approximately with the same size and the same font as in the real-world invoice in order to look like it. The string is inserted manually in the initial synthetic invoice as one can do with a text editor. However, many strings contained in the real-world invoice are not reproduced in the synthetic invoice. For instance, the information about the emitter and the receiver (address, phone number, ...) are not reproduced because they are not relevant for the study. Thus, the set of textual information I1 to I5 is placed manually on the synthetic invoice. Then, from the obtained initial synthetic invoice a fixed number of synthetic invoice images may be created automatically. In such synthetic invoice images,

the information locations are maintained identically to the initial synthetic invoice but the contained string may vary. Finally, for each distinct emitter of the real-world invoice images corpus, one initial synthetic invoice image is created manually and a fixed number of synthetic invoice images is created automatically from the initial synthetic invoice. A corpus of 1000 synthetic invoice images is thus produced and, for each synthetic invoice, both the textual information and the coordinates of the respective rectangles containing them, are stored in a database. The original distribution per emitter of the real-world invoices is preserved in the synthetic corpus of images. Synthetic data sets can then be generated from this database for closer insight. An example of such data sets is a set of (synthetic invoice) records described by 20 variables representing 5 blocks of 4 coordinates (x, y, z, q), each block being associated with one of the 5 considered information I1 to I5. It should be noted that this possibility to produce a synthetic representation of a real-world scanned invoice is an important step for updating the proposed model, namely when one has to extract information from previously unseen scanned invoice. This point will be discussed later in Section 6.

2.2 Clustering of the synthetic data

As we mentioned in Section 2.1, the regions which our proposed approach is based on are the so-called prototype regions, obtained from the specific regions that contain, each, a single information. More precisely, a prototype region associated with a given information should be a region containing a homogeneous set of specific regions related to various positions of this information in different invoice images. This makes cluster analysis methods, (such as the partitioning ones) good candidates for capturing such homogeneous sets of specific regions. Then, in the next section, the construction of prototype regions from such homogeneous sets of specific regions is explained.
The synthetic data obtained from the previous phase can be partitioned either: (a) in an overall view taking into account all of the 20 variables, or (b) in 5 independent views, each corresponding to one of the 5 information I1 to I5 and taking into account, for each view, the 4 associated variables. The approach in five independent views consists in creating five data sets: D1, D2, D3, D4 and D5. A record in Di is described by the four coordinates of regions containing information Ii. For both approaches we adopted the K-means [23] clustering with Euclidean distance. K-means is a popular, simple and efficient algorithm for cluster analysis. To determine the number of clusters, we conducted, on the one hand, an (agglomerative) ascending hierarchical clustering with Ward criterion (clustering method based on a classical sum-of-squares criterion, producing groups that minimize within-group dispersion) [24] and, on the other hand, executions of K-means for values of k between 2 and 18. Several validity criteria, such as within cluster sum of squares, silhouette and Calinkski-Harabasz [33], provided by the package clusterCrit of R software, were used to determine the optimal number of clusters for each data set. It turns out that considering five independent views leads to better clusters w.r.t. each of the considered validity

criteria. Therefore, we adopt the option consisting in partitioning each of the 5 views. The best values of k obtained for the respective five views are shown in Table 1.

Table 1. Number of clusters for data sets D1, D2, D3, D4 and D5.

	D1: Info I1	D2: Info I2	D3: Info I3	D4: Info I4	D5: Info I5
k	10	10	10	3	10

2.3 Determination of the prototype regions

As we mentioned in the previous section, a prototype region associated with an information should contain a homogeneous set of specific regions related to various positions of this information in different invoice images.
Recall that for each information Ii, the associated data set Di is partitioned into some number of clusters (see Table 1). Then, we associate to each of these clusters, say C, a *prototype region* defined as the smallest rectangle R containing each of the specific rectangles in C. Thus, we obtain 43 prototype regions R_1,...,R_{43}, with the first 10 related to information I1, the next 10 to I2, the next 10 to I3, the next 3 to I4 and the last 10 to I5. Figure 1 shows the prototype regions related to information I4. The next step of the construction of our proposed model is to set up pathways for efficiently browsing through the set of defined prototype regions. Such pathways will be obtained from the concept lattice of a suitably defined formal context.

3 Determination of pathways for browsing through the prototype regions

So far, we indicated how we determine prototype regions containing the textual information to be extracted. So we come to the fourth step in our approach, namely, define pathways for efficiently browsing through the set of these prototype regions. For this, Formal Concept Analysis (FCA) appears very appropriate. Indeed, the pathways we seek to determine may be obtained from the concept lattice of a suitably defined formal context.

3.1 Construction of the concept lattice

Recall that a *formal context* is a triple $\mathbb{K} = (O, A, \mathcal{R})$, where O is a set of objects, A a set of attributes and $\mathcal{R} \subseteq O \times A$ a binary relation from O to A. A *formal concept* of $\mathbb{K}$ is a pair (X, Y) such that $Y = X' = \{a \in A : x\mathcal{R}a \text{ for all } x \in X\}$ and $X = Y' = \{x \in O : x\mathcal{R}a \text{ for all } a \in Y\}$. Note that the double application of

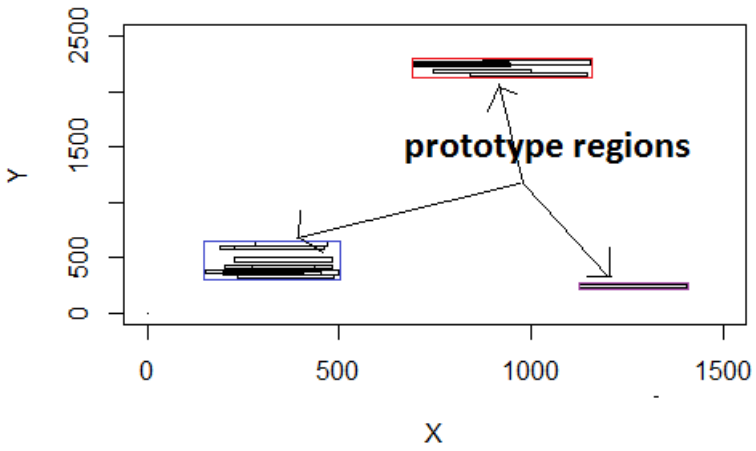

Fig. 1. Prototype regions related to information I4.

the derivation operator $(.)'$ is a closure operator, i.e. $(.)''$ is extensive, idempotent and monotone. Sets $X \subseteq O, Y \subseteq A$, such that $X = X''$ and $Y = Y''$ are said to be closed. The subset $X \subseteq O$ is called the extent of the concept (X, Y) and Y its intent. The concept lattice of the formal context $\mathbb{K}$ [35], also known as the Galois lattice of the binary relation $\mathcal{R}$ [2], is the (complete) lattice $(\mathcal{L}(\mathbb{K}), \leq)$, where $\mathcal{L}(\mathbb{K})$ is the set of formal concepts of $\mathbb{K}$ and $\leq$ the subconcept/superconcept partial order. Thus, a concept lattice contains a minimum (resp. a maximum) element according to the relation $\leq$, called the bottom (resp. the top). In this work, we consider the formal context where the objects are the invoice images and the attributes the predicates $I_i = j$, where Ii, $i = 1, \ldots, 5$ denotes the five textual information mentioned in Section 2.1, and $j = 1, \ldots, 43$ denotes the ID of the 43 prototype regions R_1,...,R_{43}. An invoice o_n is in relation with a predicate $I_i = j$ if the textual information I_i is located at prototype region R_j in the invoice o_n. A summary of this formal context is shown in Table 2.

3.2 Determination of paths from the concept lattice

Recall that, given a formal context $\mathbb{K} = (O, A, \mathcal{R})$, an association rule is a pair (X, Y), denoted as $X \to Y$, where X and Y are disjoint subsets of A [1]. The set X is called the antecedent of the rule $X \to Y$ and Y its consequent. The support of an association rule $X \to Y$ is the proportion of objects that contain all the attributes in $X \cup Y$, i.e. $\frac{|(X \cup Y)'|}{|O|}$. The confidence of $X \to Y$ is the proportion of objects that contain Y, among those containing X. A (support,confidence)-valid

Table 2. Part of the Formal context of invoices data sets.

	I1=1	I1=2	I1=3	I1=4	I1=5	I1=6	I1=7	I1=8	I1=9	I1=10	...	I4=31	I4=32	I4=33	I5=34	I5=35	I5=36	I5=37	I5=38	I5=39	I5=40	I5=41	I5=42	I5=43
o_1	X										...	X			X									
...											...													
o_{895}										X	...	X				X								
...											...													
o_{1000}							X				...		X								X			

association rule is an association rule whose support and confidence are at least equal to a fixed minimum support threshold and a fixed minimum confidence threshold, respectively. An approximate association rule is an association rule whose confidence is less than 1. When the minimum support threshold is set to 0, the Luxenburger basis of approximate association rules is the set of rules of the form $X \rightarrow Y \setminus X$ where $X = X''$, $Y = Y''$, $X \subset Y$ and there is no Z such that $Z'' = Z$ and $X \subset Z \subset Y$ [21].

The Luxenburger basis can be visualized directly in the line diagram of a concept lattice. Each approximate rule in the Luxenburger basis corresponds exactly to one edge in the line diagram. The line diagram of a lattice contains paths by which one can move from the top concept to the bottom one. The pathways we adopt for browsing through the set of prototype regions are exactly those corresponding to sequences of association rules of the Luxenburger basis, i.e. top-down consecutive edges in the concept lattice. In other words, a *pathway* is a sequence $Y_0 \rightarrow Y_1 \rightarrow ... \rightarrow Y_n$, where Y_0 is the intent of the top formal concept and for all $0 \leq i < n$, $Y_i \rightarrow Y_{i+1}$ is an association association rule of the Luxenburger basis. Given a node of the concept lattice, there are as many approximate association rules of the Luxenburger basis whose antecedent is the intent of this node, as are the children nodes of this node in the concept lattice. Between two approximate association rules having the same antecedent, the one with highest support is considered first. For instance, let a pathway p_1: I5=42 $\rightarrow$ {I1=9, I3=25} holds with a support of 4% and a pathway p_2: I5=42 $\rightarrow$ {I2=19, I3=28} holds with a support of 6%. In the aim to extract information I1 to I5 from a candidate invoice image, and supposing that I5=42 is the lattice top node's direct child node which holds the highest support value, prototype region R_{42} is visited first in order to find information I5. Then, using pathway p_2, prototype regions R_{19} and R_{28} are visited for finding information I2 and I3 respectively. When, an information Ii is not found in a prototype region given by pathway p_2, so p_1 may be used to find it. Thus, all approximate association rules given by the Luxenburger basis are used for information I1 to I5 localization and extraction. In systems, such as, CREDO [8] and SearchSleuth [12] the browsing strategy consists in focusing on a concept and its neighbors.

The effectiveness and performance for using this type of strategy in Web search have been demonstrated in [8, 12].

4 Textual information extraction

To extract the textual information of interest, we perform an optical character recognition (OCR) engine on prototype regions, using the pathways determined in the previous step. Recall that a pathway is a sequence $Y_0 \rightarrow Y_1 \rightarrow ... \rightarrow Y_n$, where Y_0 is the intent of the top formal concept and for all $0 \leq i < n$, $Y_i \rightarrow Y_{i+1}$ is an approximate association rule of the Luxenburger basis. It should be noted that each node Y_α in such a sequence represents a set of predicates "Ii=j" indicating that information Ii belongs to prototype region R_j. First, the set of pathways is ordered by descending support value of the intents. Then, in each node given by a pathway, an OCR engine is performed on each prototype region in order to extract the corresponding information Ii. In formal language theory, a regular expression is a sequence of characters that defines a search pattern, mainly for use in pattern matching with strings. For each sought information Ii, a regular expression is built and then used to check whether the extracted string (by the OCR engine) matches with the given information.
In the literature, approaches such as in [34, 18, 19] are based on *concept lattice classifier* and use a concept lattice for a classification task. Such approaches aim to improve the task of character or symbol recognition in images. In [34], the authors developed a recognition system named Navigala and fitted to recognize noisy graphical objects and especially symbols images in technical documents such as architectural plans or electrical diagrams. The authors noted that Navigala is somewhat generic and can be successfully applied to other types of data. In [18], the authors proposed modifications of some classifiers (naive Bayes, nearest neighbor and random forest classifiers) in order to use the modified classifiers as a part of the ABBYY OCR Technologies recognition schema for its performance improving. The authors note that their approach based on random forest can be applied to combine results of concept lattice classifiers.
In this paper, the task of text extraction is done with a free OCR engine named Tesseract OCR *(https://github.com/tesseract-ocr)*. Tesseract OCR was chosen because it is a free software providing a JAVA API. In this paper, we focus on the textual information localization task in administrative document images. Indeed, OCR engine such as ABBYY OCR has a better recognition rate than free OCR such as Tesseract OCR, but both are not able to localize or pick out a given information such as the net amount in invoice images. Their task is just to transform, as efficiently as possible, the text contained in images into plain text. In this work, we propose to combine the proposed localization approach (based on clustering analysis and navigation in a concept lattice) with any OCR engine in order to extract a given information in document images without browsing and recognizing the entire images.

5 Experimental results

We achieved an experiment in order to test the proposed model for textual information extraction in real-world invoice images. The experiment consists in extracting information I1 to I5, in the set of 1000 real-word invoice images (Section2.1). Despite the fact that the approach was trained and tested with good results on the synthetic data, in this section we present test results of the approach on real-world invoice images. Indeed, the corpus of real-world invoice images contains some noise which is not present in the synthetic invoice images. The real-world invoice images may contain colored images such as a logo, shadow areas and handwritten text. Additionally, they may be scanned with poor quality and may present distortion. Thus, the corpus of real-world invoice images seems to us to be quite interesting for testing the proposed textual information localization and extraction model. We performed two types of extraction:

1. from full images: OCR is performed on the entire page images regardless to specific regions;
2. from prototype regions, using the pathways presented in Section 3: OCR is performed only on image sub-regions, using the pathways.

To perform OCR on images, the JAVA library of the free OCR engine named Tesseract in its 3.02 version is used. We considered two measures:

1. the rate of correct information among the total number of sought information (recall),
2. the rate of correct information among the total number of detected information (precision).

On the one hand, a sought information is considered detected, if a string which matches the corresponding regular expression is found. On the other hand, a sought information is considered correctly extracted, if the extracted textual information corresponds exactly to the visual information that should be read in the image. For instance, let 'net amount' be a sought information and assume that the net amount is 107€ in the invoice image. During the process, if the retrieved information is "101€", the net amount will not be considered correctly extracted because the real net amount mentioned in the original invoice image is "107€". The results are presented in Table 3. On the one hand, despite the fact that according to [25], the Tesseract OCR engine has accuracy of 70% for text extraction in gray scale number plate images, we observe that the accuracy of the OCR engine is weak for text extraction in real-world invoice images. On the other hand, these results show that our proposed model improves significantly the performance of the OCR engine. Note that a p-value of 8.799e-05 was obtained for this experiment, which means that the results are significant.

6 Conclusion and perspectives

We presented a (prototype) region-based model for localizing and extracting textual information in document images. Experimental results show that the

Table 3. Results of experiments.

	recall	precision
OCR	29,84 %	37,36 %
OCR + pathways	35,88 %	50,46 %

proposed model improves significantly the correctness of a textual information extraction process based on an OCR engine. This model is constructed in four steps:

1. produce synthetic invoice data from real-world invoice images containing the textual information of interest, along with their spatial positions;
2. partition the produced data;
3. derive the prototype regions from the obtained partition clusters;
4. derive pathways for browsing through the prototype regions, from the concept lattice of a suitably defined formal context;

The Step 1 is important when one has to extract information from a previously unseen invoice image. Indeed, if some information of such an invoice image are not retrieved, then a synthetic representation of the considered invoice can be produced, and this triggers incremental updates of the synthetic data sets, the prototype regions, and the concept lattice. Our future work will focus on these update processes, and compare them with those proposed in [3, 4, 9]. We also plan to develop a classification model, as in [9], that will enable to predict the invoice emitter based on the five textual information I1 to I5 considered in the present paper. This will allow to easier retrieve the other textual information the company is interested in: invoice number, invoice date, tax rate, tax due.

Acknowledgment

This work was partially carried out within the project ClustOverlap supported by Reunion Island Region - grant DIRED 20140704. The authors are also very grateful to the 3 anonymous reviewers for their valuable comments.

References

1. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. ACM SIGMOD Record, 22.2, pp. 207–216 (1993).
2. Barbut, M., Monjardet, B.: Ordre et classification: algèbre et combinatoire. Hachette (1970).
3. Bartoli, A., Davanzo, G., Medvet, E., Sorio, E.: Semisupervised wrapper choice and generation for print-oriented documents. IEEE Transactions on Knowledge and Data Engineering 26.1, pp. 208–220 (2014).
4. Belaïd, A., D'Andecy, V. P., Hamza, H., Belaïd, Y.: Administrative document analysis and structure. Learning Structure and Schemas from Documents, pp. 51–71 (2011).

5. Birkhoff, G.: Lattice Theory. American Mathematical Society 25.3 (1967).
6. Cambria, E., White, B.: Jumping NLP Curves: A Review of Natural Language Processing Research. IEEE Computational Intelligence Magazine 9.2, pp. 48–57 (2014).
7. Carpineto, C., Michini, C., Nicolussi, R.: A concept lattice-based kernel for SVM text classification. Formal Concept Analysis, pp. 237–250 (2009).
8. Carpineto, C., Romano G.: Expoiting the Potential of Concept Lattices for Information Retrieval with CREDO. Journal of Universal Computer Science, 10.8, pp. 985–1013 (2004).
9. Cesarini, F., Francesconi, E., Gori, M., Soda, G.: Analysis and understanding of multi-class invoices. International Journal on Document Analysis and Recognition 6.2, pp. 102–114 (2003).
10. Charikar, M., Chekuri, C., Feder, T., Rajeev Motwani: Incremental clustering and dynamic information retrieval. In: Proceedings of the twenty-ninth annual ACM symposium on Theory of computing. ACM, pp. 626–635 (1997).
11. Cho, W.C., Richards, D.: Improvement of Precision and Recall for Information Re- trieval in a Narrow Domain: Reuse of Concepts by Formal Concept Analysis. In Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence (WI '04), pp. 370–376 (2004).
12. Ducrou, J., Eklund, P. W.: SearchSleuth: The conceptual neighbourhood of an web query. In J. Diatta, P. Eklund, & M. Liquire (Eds.), Proc. CLA 2007, LIRMM & University of Montpellier II (2007).
13. Eisenbarth, T., Koschke, R., Simon, D.: Locating Features in Source Code. IEEE Transactions on software engineering 29.3, pp. 210–224 (2003).
14. Ganter, B., Wille, R.: Contextual attribute logic. Springer Berlin Heidelberg (1999).
15. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer Science & Business Media (2012).
16. Guyon, I., Cawley, G., Dror, G., Saffari, A.: Hands-on Pattern Recognition. Challenges in Machine Learning 1. Isabelle Guyon, Gavin Cawley, Gideon Dror, and Amir Saffari editors (2011).
17. Hyontai, S.: An Effective Sampling Method for Decision Trees Considering Comprehensibility and Accuracy. WSEAS Transactions on Computers 8.4, pp. 631–640 (2009).
18. Itskovich, L., Kuznetsov, S.O.: Machine Learning Methods in Character Recognition. Proc. 13th International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing (RSFDGrC 2011), Lecture Notes in Computer Science 6743, pp. 322–329 (2011).
19. Kaytoue, M., Kuznetsov, S.O., Napoli, A., Duplessis, S.: Mining gene expression data with pattern structures in Formal Concept Analysis. Inf. Sci. 181(10): 1989-2001 (2011).
20. Langley, R.: Practical statistics simply explained. Courier Corporation (1971).
21. Kuznetsov, S.O., Makhalova, T.P.: Concept interestingness measures: a comparative study. CLA 2015, pp. 59–72 (2015).
22. Luxenburger, M.: Implications partielles dans un contexte. Mathématiques et Sciences Humaines 113, pp. 35–55 (1991).
23. MacQueen, J. et al.: Some methods for classification and analysis of multivariate observations. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability 1.14, pp. 281–297 (1967).

24. Murtagh, F., Legendre, P.: Wardś hierarchical agglomerative clustering method: Which algorithms implement wardś criterion?. Journal of Classification 31.3, pp. 274–295 (2014).
25. Patel, C., Patel, A., Patel, D.: Optical character recognition by open source OCR tool tesseract: A case study. International Journal of Computer Applications, 55.10, pp. 50–56 (2012).
26. Poelmans, J., Ignatov, D.I., Kuznetsov, S.O., Dedene, G.: Formal Concept Analysis in knowledge processing: A survey on applications. Expert systems with applications 40.16, pp. 6538–6560 (2013).
27. R Core Team: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, (2013).
28. Singh, S.: Optical Character Recognition Techniques: A survey. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) 4.6, pp. 545–550 (2013).
29. Sumathi, C.P., Santhanam, T., Devi, G.G.: A survey on various approaches of text extraction in images 3.4, pp. 27–42 (2012).
30. Sumathi, C.P., Santhanam, T., Priya, N.: Techniques and challenges of automatic text extraction in complex images: A survey. Journal of Theoretical and Applied Information Technology, 35.2, pp. 225-235 (2012).
31. Taouil, R., Pasquier, N., Bastide, Y., Lakhal, L.: Mining bases for association rules using closed sets. ICDE2000 International Conference, pp. 307 (2000).
32. van De Vel, M.L.J: Theory of convex structures. Elsevier (1993).
33. Vendramin, L., Campello, R., Hruschka, E.: Relative Clustering Validity Criteria: A Comparative Overview. Statistical Analysis and Data Mining 3, pp. 209–235 (2010).
34. Visani, M., Bertet, K., Ogier, J.-M.: NAVIGALA: An original symbol classifier based on navigation through a galois lattice. International Journal of Pattern Recognition and Artificial Intelligence 25.4, pp. 449–473 (2011).
35. Wille, R.: Restructuring lattice theory: an approach based on hierarchies of concepts. Springer Netherlands (1982).