Using lattices for reconstructing stemma

Marc Le Pouliquen

1. Département LUSSI TAMCIC, UMR CNRS 2872
ENST Bretagne, BP 832, 29285 Brest Cedex
2. Université de Bretagne Occidentale IUP GMP
6 av. Le Gorgeu - CS93837 29238 Brest Cedex 3
{marc.lepouliquen@enst-bretagne.fr}

Abstract. The product of textual criticism is an edited text that the editor believes comes as close as possible to a lost original manuscript called the archetype. Usually, the editor compares different manuscripts of a single text, and represents it as an inverted tree showing all the steps in the transmission of a specific text, reconstructed by establishing relationships with other manuscripts. This tree is called the "stemma codicum"(cf. [7]). Because of the graphic proximity of the stemma with a semi-lattice, we propose to use two lattice construction techniques in order to reconstitute the filiation tree of manuscripts. First, we try the traditional methods to build the lattice of a binary relation (cf. [13]). Then a more specific solution to the problem is proposed. These techniques are finally tested on a real corpus of manuscripts by Rimbaud, "Les Effarés" (cf. [17]).

1 Introduction

In this paper, we use lattices as a pattern for the construction of the family tree of manuscripts within the framework of the critical edition.As far as possible, the editor must try to reconstitute , the original manuscript[1] as the author wrote it, starting from the various preserved manuscripts. The corpus is made up with many manuscripts which are copied from each other. To do so, it appears interesting to draw up a family tree of these manuscripts called the "stemma codicum".

As can be seen on Figure 1, the stemma is a kind of graph or a tree. We will extract our stemma from a lattice by pruning vertices and edges. The lattice is built starting from a binary relation between the manuscripts and their differences. This information is contained in the collation table[2]. Two methods are proposed to carry out the lattice pruning:

- An expert (in this case an editor) orders the most relevant concepts (in this case the "differences" between the manuscripts) according to his judgment.
- An algorithm helps the expert by removing the lattice vertices which have not enough "difference" on each level. After many iterations, the lattice becomes a tree or a graph representing a stemma.

This paper is organized as follows: In section 2, we persent philological methods for the establishment of stemma. In section 3, we describe visualization techniques for the building of the stemma which are tested in section 4 on a real corpus of poems.

[1] The **original manuscript** or archetype is the most recent common ancestor of all extant manuscripts in an textual tradition.

[2] **collation** is the comparison between a manuscript and the other manuscripts from the corpus for the sake of producing a list of the differences

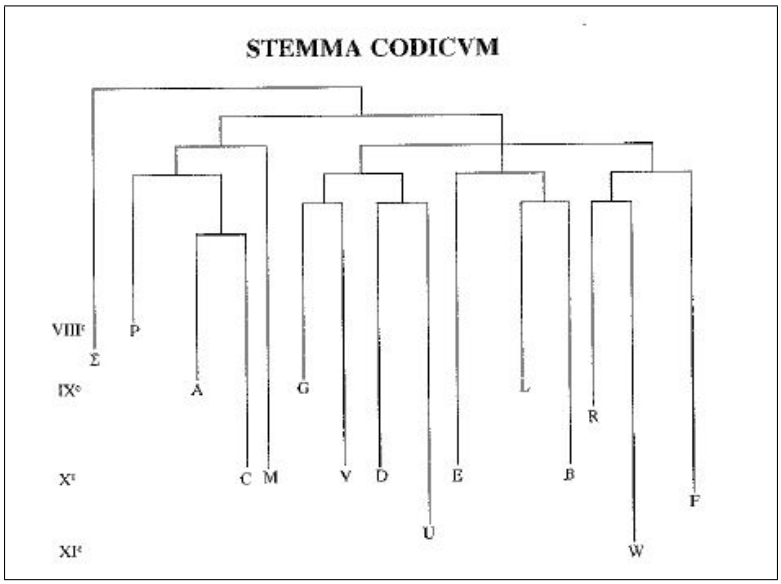

Fig. 1. Stemma codicum established by Reydellet in connection with Poems of Venance Fortunat [6]

2 Some philological methods to establish the stemma codicum

A text which was copied several times constitutes a "textual tradition" and all the specimens that have reached us are called the "witnesses". Usually, the editor compares different witnesses of a single text, and makes a selection of variants ("readings") taken from many sources to restore the original manuscript. The editors use a stemma codicum to evaluate readings, and vice versa.

Historically, several methods have been developed in order to try to visualize the genealogical relations between manuscripts. One of these methods, formalized by Lachmann[7] is now called the common error method. If an error is introduced into a manuscript, it is likely that the "descendant" of that text will show the same common error. So, a family of manuscripts is composed of the texts that have the same reading. Although this method has been largely criticized, both this method and its improvements have become indispensable to describe the history of the text

Another historical method is the method of Don Quentin[9]. He came up with the idea of reconstituting the sequence of the manuscripts by means of a three by three comparison. In fact, he assembled small chains of three manuscripts, one being between the other two and assembled these small chains in order to infer the complete tree.

After some counting, we notice that the number of different diagrams is exponentially dependent on the number of manuscripts. It is therefore impossible to consider all the stemmas, their construction and their comparison. This can explain why editors had difficulties formalizing both stemmas and their use. Thanks to the new visualization possibilities offered by lattices, the "stemmatization"methods can be modified and adapted to model the history of the text.

3 Lattice construction starting from binary relations

3.1 Algorithms and software

The last few years, Galois lattices of binary relations have been fields of important research in formal concept analysis (FCA) in particular for the visualization of many problems. FCA was basically inspired by the work of Birkhoff[3], Galois lattices were described by Barbut and Monjardet[1] and the whole approach was formalized by Ganter and Wille[11]. The lattice construction starting from binary relations and their visualization by the intermediary of the Hasse diagram allows greater comprehension of the binary table. Many algorithms have been developed for this construction:

- Those which build lattices in an incremental way i.e. they can update the lattice concept when a new object is added without re-computing the whole lattice
  - Godin[12]*
  - Carpineto and Romano[4]*
  - Norris[15]
- The other algorithms have to know the whole binary table before computing the lattice
  - Chein[5]
  - Ganter[10]
  - Bordat[2]*
  - Nourine and Raynaud[16]*

A detailed description of these algorithms and a comparison of lattice algorithms has been done in Guénoche[13] and Kuznetzov[14]. The methods which interest us are those followed by a * symbol because they can generate the Hasse diagram of Galois lattices.

In the experimentation, we need to visualize the lattices in a Hasse diagram to analyze them. With this purpose we opt for using two software for lattices representation:

- ConExp (Concept Explorer of Yevtushenko[19]) combines the creation and the visualization of the binary table in a simple tool. A view of the ConExp interface is in Figure 5. The diagrams can be exported to the JPEG or GIF format. With ConExp, it is possible to carry out many operations of Ganter and Wille[11].
- Galicia is the interactive lattice construction tool of Valtchev et al.[18]. Simple and valued contexts can be analyzed. The binary relations and the objects can also be described and stored. The lattices can be saved in JPEG, SVG or PDF formats (cf. Fig. 4).

3.2 Simple example

For the example, let us reduce the manuscripts to three sentences. Let there be the three following sentences, which correspond to the same sentence of manuscripts that were copied one from the other.

Mns1 =“Here is a sentence invented for the example”
Mns2 = “This is a sentence invented for the example”
Mns3 = “Here is a sentence built for the example”

There are two variant places here: (*Here/This*) and (*invented/built*), corresponding to four variants, as summarized in the collation table 1:

N° of variant place	Variants	Manuscripts	Variants	Manuscripts
1	Here	Mns1,Mns3	This	Mns2
2	invented	Mns1,Mns2	built	Mns3

Table 1. Collation table of our three manuscripts

Manuscripts/variants	Var1	Var1b	Var2	Var2b
Manuscript1	×		×	
Manuscript2		×	×	
Manuscript3	×			×

Table 2. Binary table

To obtain a binary table, we assign to the three manuscripts a boolean value according to the presence or absence of each variant (cf. Table 2).

From this table, we assume that the history of the text is summarized in the following way:

- Either (a) manuscript 2 is the nearest manuscript to the original. So here, manuscript 1 is copied from manuscript 2. The scribe modifies *This* in *Here*. Manuscript 3 is copied from manuscript 1, and another scribe modifies *invented* in *built*. On the other hand, if the first scribe modifies *Here* in *This*, there is little chance that the following scribe could find again *Here* by modifying *This*. It is not a credible assumption and one can say that the manuscript 1 is an intermediate between the manuscripts 2 and 3.
- Or (b) manuscript 1 is the nearest manuscript to the original and manuscripts 2 and 3 are copied from it
- Or (c) manuscript 3 is the nearest manuscript to the original, then manuscript 1 is copied from it and manuscript 2 copied from manuscript 1.

According to the information contained in the sentences, it is not possible to choose between these three stemmas in Figure 2 without the help of external information as datation or codicologic studies[3].

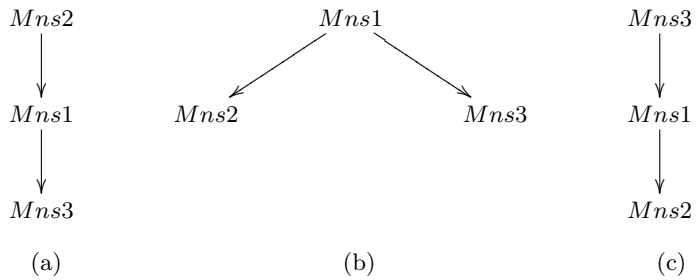

Fig. 2. Three possible stemmae

We now associate the binary table (cf. Table. 2) to the lattice (cf. Fig. 3) obtained by the previous algorithms (cf. 3.1). We note that we obtain a perfect representation of the manuscripts and their variants; indeed the lattice shows that manuscripts 1

[3] **Codicologic information** are for example: the color of the ink, the order of the page who can be modified over the time...

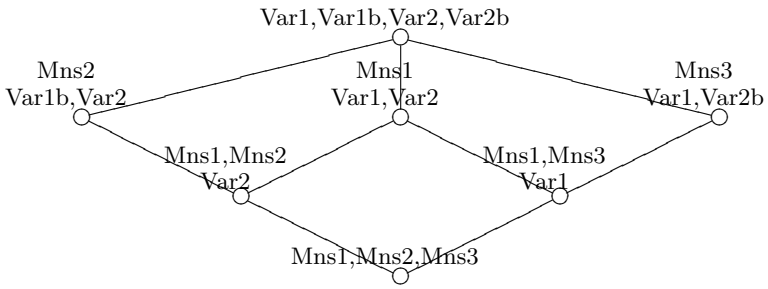

Fig. 3. Lattice

and 2 have Var2 (*invented*) as their common variant. To obtain the stemma starting from the Hasse diagram, two methods are proposed:

– the editor must remove the less significant variants until the lattice becomes a tree. In our example, if the editor assumes that Var1 is more judicious than Var1b, he obtains Figure 4 and this choice is called "emendation"[4]. After if the editor prefers Var2b to Var2, we obtain the Hasse diagram of Figure 5 which corresponds to the preceding stemma (c).

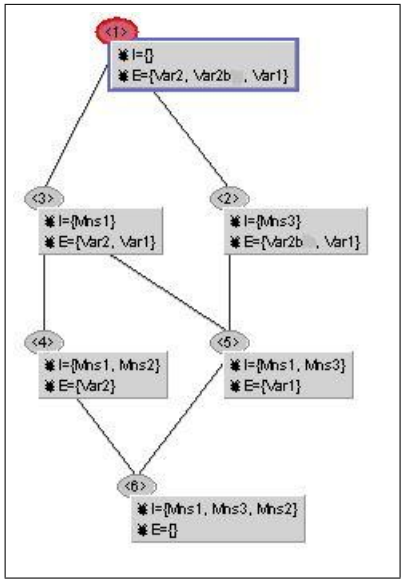

Fig. 4. Lattice using Galicia by removing Var1b

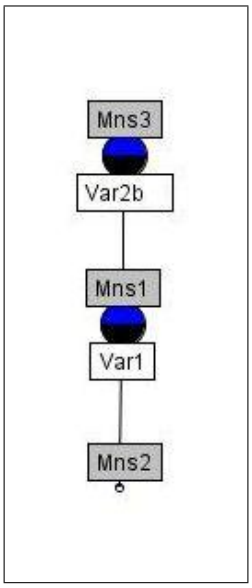

Fig. 5. Lattice using ConExp by removing Var1b and Var2

– An algorithm helps the editor by removing the lattice vertices witch have not enough relevance.

Algorithm 1

1. *We delete the vertice which have the least variants by level*

[4] **emendation**, a correction made to a text in the belief that the author's original wording has been wrongly altered

2. *We delete the variant of the same variant place which is contained in the previous deleted vertice.*
3. *We reiterate the process until the editor decides to stop it or as long as the graph remains connected.*
4. *REM: If the vertices have the same number of variants, we keep the variant of the same variant place which is contained in most manuscripts to begin the iteration.*

As in our example, there are only 2 variant places, so the algorithm keeps only Var1 and Var2. Actually, in the first variant place, Var1 is contained in manuscripts 1 and 3 whereas Var1bis is contained in manuscript 2 only, so the algorithm keeps Var1. Finally, we obtain the stemma (b) of the Figure 2. In this algorithm, the editor must constantly be able to impose his expert point of view on the interface which will be taken into account when realizing the stemma.

4 Application to a real corpus

We test this method on a real corpus of Rimbaud poems, "Les Effarés" or "Petit Pauvre" put together by Steve Murphy (cf. [17]). Here we consider five versions of this poem:

- GM reproduction of *Gentleman's Magazine* (1878)
- L Lutèce printed book (1883)
- JA Manuscripts' autograph[5] of Jean Aicard (1871)
- PD Manuscripts' autograph in Demeny's collection (1871)
- PV Copy of P. Verlaine (1872)

After collation (cf. Table 3) we use the same method described above, and investigate 14 different variants.

Variant place	Line	GM	L	JA	PD	PV
1	titre	Petits Pauvres	Les Effarés	Les Effarés	Les Effarés	Les Effarés
2	5	dos	culs	culs	culs	culs
3	7	cinq	Les	cinq	cinq	cinq
4	9	beaux	lourd	lourd	lourd	lourd
5	11-17		Ils voient	Ils voient	I ls voient	Ils voient
6	16		gros	gras	gras	gras
7	17		Chante	Chante	Chante	Grogne
8	23-25	boulanger	médianoche	médianoche	minuit sonne	médianoche
...	...	...	...	...	...	...

Table 3. Collation table of poems

With ConExp, we achieve a lattice that we will use to find the stemma. Initially, the visualization shows that common variants of L and PV (cf. Fig. 6) represent 10 variants out of 14. A high score of common variants means a close relation between these two manuscripts, for L is assumed to be a reconstruction from memory by Verlaine.

The second diagram (cf. Fig. 7) shows us that the two manuscripts JA and PD have the second hightest score of common variants. They are manuscripts that are the oldest and closest by date, which may explain their proximity.

[5] An **autograph** is a document written entirely in the handwriting of its author

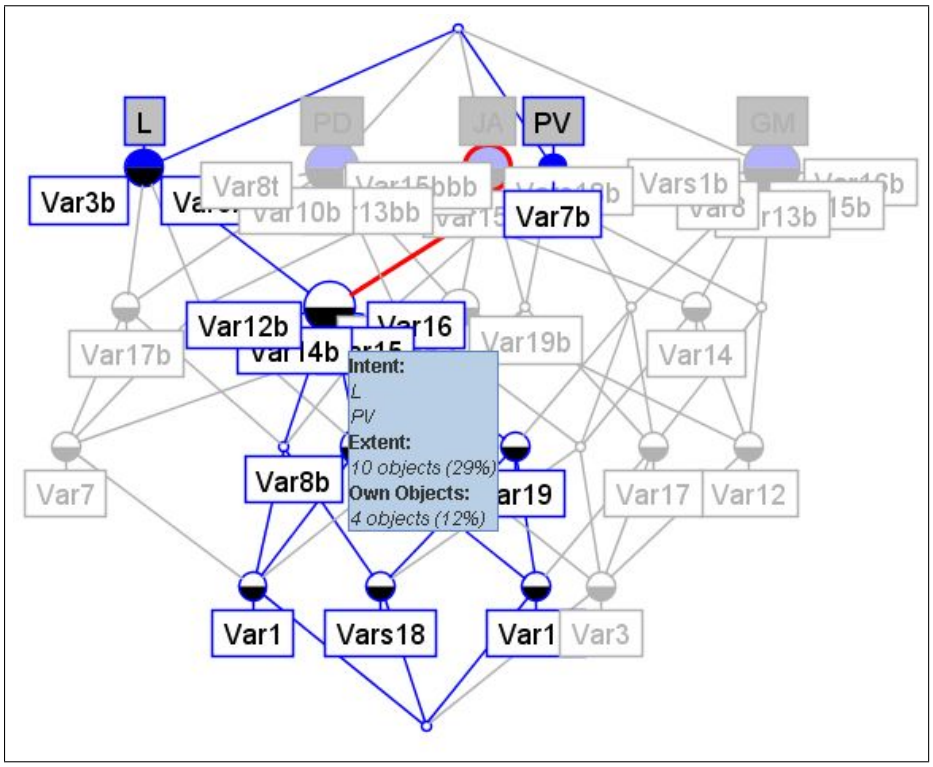

Fig. 6. Hasse diagram of Les Effarés. Relation between manuscripts PV and L

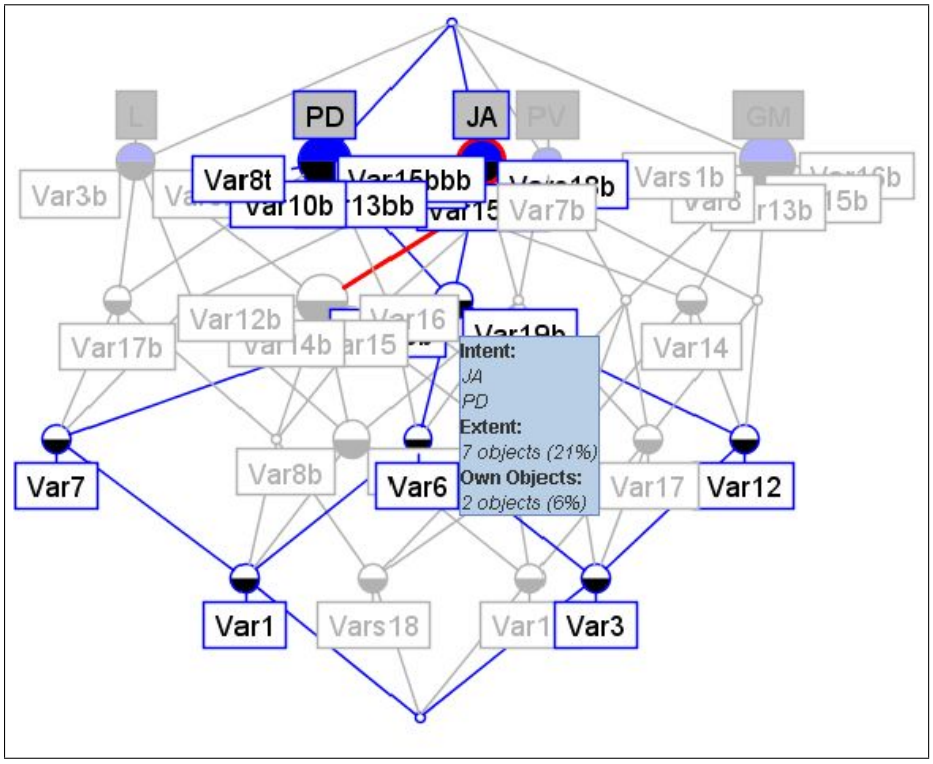

Fig. 7. Hasse diagram of Les Effarés. Relation between manuscripts JA and PD

To attempt to extract a family tree (stemma) from the Hasse diagram (cf. Fig. 8), we cannot extract the most significant variants, because they are almost all legitimate and undoubtedly by the author himself. We use the algorithm which involves in removing the least significant vertices i.e. those which have fewer variants on each level. Then we obtain Figure 9. If we continue the extraction as long as the "lattice" remains connected, we finish in removing vertice 4 and the dotted line edges. In this corpus, the stemma probably represents the proximity of the poems rather than a hypothetical filiation.

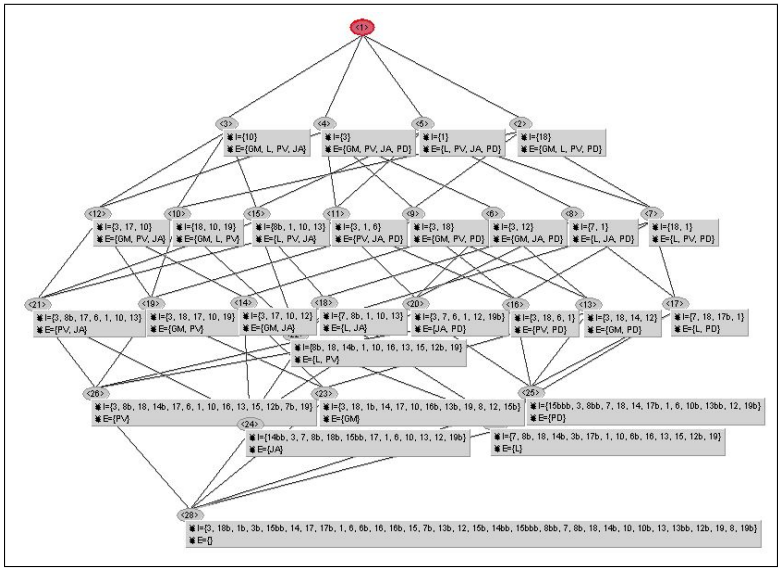

Fig. 8. Hasse diagram of Les Effarés using Galicia

5 Prospects and Conclusion

In future work, we therefore plan to write a program designed to be used as a stemma construction aid for the textual scholar. In all cases, the editor may interact with the program to improve the results using human insight. This interaction is necessary if we want to persuade editors of the usefulness and the interest of the system. The use of lattices is necessary to visualize the relations between manuscripts and their variants and providing the editor with the required interactions.

However, many tasks remain:

- Sometimes the corpus contains more than one hundred manuscripts and one thousand variants. Under these conditions, how can we optimize the visualization?
- Many statistical and probabilistic aspects must also be considered during the automatic lattice pruning.
- Methods based on phylogenetic trees are already used for drawing stemmas (cf. [8]); how can we combine these two methodologies ?

References

1. Barbut M. and Monjardet B. : Ordre et classification : Algèbre et combinatoire, *Collection Hachette Université*, (1970)

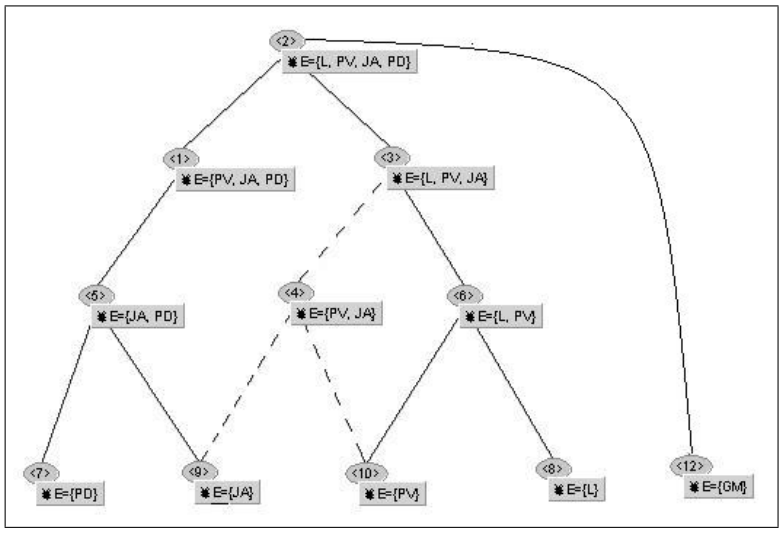

Fig. 9. Stemma of Les Effarés

2. Bordat J.P. : Calcul pratique du treillis de Galois dune correspondance, *Math. Sci. Humaines* No 96 (1986), 31–47
3. Birkhoff G. : Lattice Theory, *American Mathematical Society Publications*, (1940)
4. Carpineto C. and Romano G.: A lattice conceptual clustering system and its application to browsing retrieval, *Machine Learning* No 24 (1996), 95–122
5. Chein M. : Algorithme de recherche des sous-matrices premières d'une matrice, *Bull. Math. Soc. Sci. Math. R.S. Roumanie* No 13 (1969), 21–25
6. Fortunat V. and Reydellet M. : Poèmes Livres I-IV, *Belles lettres*, (1994)
7. Lachmann, K.K.F.W. : Kleinere Schriften, *Berlin*, (1876)
8. Le Pouliquen M., Barthélemy J.P. and Bertrand P. : Filiation de manuscrits sanskrits et arbres phylogénétiques, *soumise en vue d'une parution dans un numéro spécial de la revue RNTI (SFC06)* , (2006)
9. Quentin, H. : Essais de critique textuelle, *Picard*, (1926)
10. Ganter B. : Two Basic Algorithms in Concept Analysis, *Technische Hochschule Darmstadt*, Preprint No. 831 (1984)
11. Ganter B. and Wille R. : Formal Concept Analysis, Mathematical foundations, *Springer-Verlag*, (1999)
12. Godin R., Missaoui R., and Alaoui H.: Incremental concept formation algorithms based on galois lattices, *Computation Intelligence*, No. 11(2) (1995), 246–267
13. Guenoche A. : Construction du treillis de Galois d'une relation binaire, *Math. Inform. Sci. Humaines*, No. 109 (1990), 41–53
14. Kuznetsov S.O. and Obiedkov S.A. : Comparing performance of algorithms for generating concept lattices, *Exp. Theoret. Artificial Intelligence*, No 14 23 (2002), 189–216
15. Norris E.M. : An Algorithm for Computing the Maximal Rectangles in a Binary Relation, *Revue Roumaine de Mathématiques Pures et Appliquées*,No. 23(2) (1978), 243–250
16. Nourine L. and Raynaud O. : A fast algorithm for building lattice, *Information Processing Letters*, (1999), 199–204
17. Rimbaud, A. : Poésies, édition critique avec introduction et notes de Steve Murphy, *Honoré Champion*, (2000)
18. Valtchev P., Gosser D., Roume C. and Hacene M.: Galicia: an open platform for lattices, *In Aldo de Moor,Wilfried Lex, and Bernhard Ganter, editors, Contributions to the 11th Conference on Conceptual Structures*, (2003), 241–254
19. Yevtushenko A.S. : System of data analysis "Concept Explorer", *Proceedings of the 7th national conference on Artificial Intelligence KII-2000*, (2000), 127–134