Uncertain Formal Concept Analysis for the Study of a Text Corpus

Guillaume Petiot
CERES, Catholic Institute of Toulouse, 31 rue de la Fonderie, 31068, Toulouse, France

Keywords: Data Analysis, Formal Concept Analysis, Natural Language Processing, Possibility Theory, Uncertainties.

Abstract: The analysis of a corpus by an expert takes a relatively long time. The development of digital tools made it possible to generate instantly a summary of information contained in the corpus. In this paper, we will focus on the contribution of formal concept analysis (FCA) to the analysis of a corpus. FCA makes it possible to build a model also called the Hasse diagram which can be queried to find relevant formal concepts. Uncertainties can be present in all steps of the processing from the corpus processing to the visualization of the results. Indeed, if the words of the corpus are misspelled or additional quantitative variables are associated with the corpus, then uncertainties can appear. Uncertainties may also arise in queries when human knowledge is imprecise. Possibility theory allows us to represent and process these imperfections. The combination of textual analysis solutions and FCA allow us to present more relevant results that take into consideration uncertainties.

1 INTRODUCTION

A corpus is a collection of documents. These documents come from books, articles, transcripts of interviews, open questions in a questionnaire, websites, etc. The analysis of the lexicon of a corpus can be a time-consuming task for an expert. Indeed, when the corpus grows, it becomes more and more difficult to analyze the lexicon and accurately represent the relationships between words. The methods of Text Mining (Hotho et al., 2005) or lexicometry (Salem, 1986) make it possible to summarize a corpus more efficiently. Lexicometry (Salem, 1986) deals with the quantitative analysis of the lexicon using statistical methods. Many software tools have been proposed to summarize text corpora. Alceste and Iramuteq software, for example, are particularly interesting. In these software, a dictionary is previously built after a preprocessing of the corpus. The preprocessing can be a pipeline of operations leading to a corpus cleaning, followed by lemmatization to reduce the size of the dictionary. Then, a segmentation of the corpus is performed. These tools make it possible to calculate statistical summaries, to perform classification, factorial correspondence analysis, similarity analysis, etc. Finally, the latter offers graphical representations that highlight previous results.

The Iramuteq software performs a text segmentation of the corpus into segments. Then we can apply a factorial correspondence analysis and hierarchical clustering proposed by Reinert (Reinert, 1983). We obtain on the one hand a classification of terms and on the other hand a representation of terms on the first two principal components. Tables are presented in the tool and allow us to explore intermediate data and all results. For example, the result of the classification makes it possible to consult for each class the words associated with it as well as the χ^2 distance and the p-value. A concordancer allows you to consult the segments of text that contain the words selected by the user. The similarity analysis (SA) (Degenne and Vergès, 1973) proposed in this tool greatly contributes to the analysis of the link between terms by gradually representing the links in a graph.

The representation of a document-term matrix (DTM) is very close to the formal context of formal concept analysis (FCA). Indeed, it is possible to binarize the DTM or to define linguistic variables or classes concerning the frequencies of words. If we use possibility distributions of possibility theory to represent linguistic variables, we can compute degrees of necessity for each modality. We can also represent the uncertainty of words by a degree of necessity. Indeed, during lemmatization, a misspelled word can be associated with several words because the spelling of the words is very close. We can choose the word with the highest possibility. This, however, generates an uncertainty that must be propagated in the analysis.

The variables of different kinds – binary, qualita-

Petiot, G.
Uncertain Formal Concept Analysis for the Study of a Text Corpus.
DOI: 10.5220/0012316400003636
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence (ICAART 2024) - Volume 3*, pages 218-225
ISBN: 978-989-758-680-4; ISSN: 2184-433X
Proceedings Copyright © 2024 by SCITEPRESS – Science and Technology Publications, Lda.

tive (nominal or ordinal) or quantitative – can be associated with the texts of the corpus. The processing of variables is mandatory to represent them with the terms (the words of the corpus) in an uncertain context that gathers all information. Applications of FCA have already been proposed to analyze a corpus (Cimiano et al., 2005; Tovar et al., 2015), however, uncertainties and additional variables are rarely discussed in these studies.

In this research, we will focus our interest on processing uncertainties. We will extend the work already done in our previous research (Petiot, 2019). We will propose a new approach to analyzing a corpus by using uncertain formal concept analysis. We will combine traditional textual analysis approaches such as factorial correspondence analysis and similarity analysis with FCA.

To do this, we will in the first part recall the basis of the possibility theory and FCA. Then we will describe the steps of the corpus processing. We will distinguish the preliminary analysis of the context and the analysis of the formal concepts. We will present an example of a graphical query language that allows us to select pertinent formal concepts and to improve the visualization of information. Finally, we will show an example of the Hasse diagram leading to the computation of rules that highlight the dependence of terms.

2 POSSIBILITY THEORY

Possibility theory (Zadeh, 1978) is an extension of the fuzzy sets theory proposed by L. A. Zadeh in 1965. It makes it possible to represent imprecise knowledge by distributions of possibility (noted π) and to compute degrees of certainty. It also offers a representation of ignorance. There are two important measures defined on the powerset of a universe Ω denoted $P(\Omega)$ in $[0,1]$:

- The measure of possibility Π

$$\forall A \in P(\Omega), \Pi(A) = \sup_{x \in A} \pi(x). \quad (1)$$

- The measure of necessity N

$$\forall A \in P(\Omega), N(A) = 1 - \Pi(\bar{A}). \quad (2)$$

The conditioning in possibility theory was discussed by researchers D. Dubois and H. Prade in (Dubois and Prade, 1988). They proposed the following solution for the conditioning:

$$\Pi(A|B) = \begin{cases} \Pi(A \cap B) & \text{if } \Pi(A \cap B) < \Pi(B), \\ 1 & \text{if } \Pi(A \cap B) = \Pi(B). \end{cases} \quad (3)$$

3 FORMAL CONCEPT ANALYSIS

Formal concept analysis is a method of data analysis proposed by R. Wille (Wille, 1982) which consists in describing the formal concepts present in a given context. Formal concepts encompass recurring features of the context. This method is an application of lattice theory that allows formal concepts to be represented by a Hasse diagram when a partial order relation is defined. Two solutions can be proposed to explore the formal concepts: a navigation in the Hasse diagram and the consideration of a formal concept and its neighbours. The second solution is to perform queries to search relevant formal concepts. Many applications exist concerning FCA (Poelmans et al., 2013; Poelmans et al., 2014; Snášel et al., 2008; Bělohlávek et al., 2007; Fernandez-Manjon and Fernandez-Valmayor, 1998) in text mining, linguistics, social media, education, bioinformatics, psychology, ontology engineering, etc.

A formal concept has two sets: the intension and the extension. The intension represents the set of common properties that the objects of the concept have, and the extension represents the set of objects to which they apply. Mathematically, a formal context is a triplet $(O,P,\Re)$ where $O = \{o_1,...,o_n\}$ is the set of objects, $P = \{p_1,...,p_m\}$ the set of properties, and $\Re$ a binary relation such that $\Re \subseteq O \times P$. If $(o,p) \in \Re$ then the object o has property p. A context is often represented by a table where the rows are objects and the columns are properties. The cells in the table represent the relation $\Re$ between the object and the property: 0 if $(o,p) \notin \Re$ or 1 if $(o,p) \in \Re$. One can define a function $\vartheta(o,p)$ that returns the value of the table for an object o and a property p. A formal concept of $(O,P,\Re)$ is a couple (O_i,P_i) such that $O_i \in O$ and $P_i \in P$ such that P_i is the set of properties shared by the set of objects of O_i. It can be noted $O_i^{\uparrow} = P_i$ or $P_i^{\downarrow} = O_i$. For example $(\{o_1,o_4,o_5\},\{p_1,p_2\})$ is a formal concept of the following binary context:

Table 1: Example of formal context.

Objects \ Properties	p_1	p_2	p_3
o_1	1	1	0
o_2	1	0	1
o_3	0	1	1
o_4	1	1	0
o_5	1	1	1

Definition 3.1. *The set of all formal concepts of* $(O,P,\Re)$ *is denoted* $\chi(\Re)$.

We have the following property for each formal concept (O_i,P_i) of $\chi(\Re)$:

$\left\{(O_i,P_i)|O_i^{\uparrow}=P_i,P_i^{\downarrow}=O_i\right\}$. It is possible to compare formal concepts with each other by defining a partial order:

Definition 3.2. *Let be two formal concepts* $(O_1,P_1),(O_2,P_2)$ *of* $\chi(\Re)$. *We define a partial order* $\leq$ *such that* $(O_1,P_1)\leq(O_2,P_2)$ *if and only if* $O_1\subseteq O_2$ *or* $P_2\subseteq P_1$.

The set $\chi(\Re)$ with the partial order $\leq$ is used to build a concept lattice that can be visualized by using a Hasse diagram. If the properties of the context are quantitative or multivalued, a transformation of the context must be performed to obtain a binary formal context. If it is not certain that an object has a property, it is necessary to adapt FCA. The study by (Bělohlávek, 2004) focused on the use of fuzzy sets to represent imprecise properties. The authors (Dubois et al., 2007; Dubois and Prades, 2015; Ait-Yakoub et al., 2016) propose to use possibility theory to take into account imprecision and uncertainties. The same authors also propose a solution to manage uncertainties and to provide a frame to represent ignorance which can be partial or total. A pair of necessity measures $(\alpha(o,p),\beta(o,p))$ with $\alpha(o,p)=N((o,p)\in\Re)$ and $\beta(o,p)=N((o,p)\notin\Re)$ is used to represent uncertainties. $N((o,p)\in\Re)$ is the necessity that the object o has the property p and $N((o,p)\notin\Re)$ is the necessity that the object o does not have the property p. The pair of necessity measures is required because of the equation 2. Each necessity measure is computed by using the possibility of the contrary event. The necessity measures α and β must satisfy the property $min(\alpha(o,p),\beta(o,p))=0$ of possibility theory. The $(1,0)$ and $(0,1)$ pairs denote a property or its lack. If $1>max(\alpha(o,p),\beta(o,p))>0$, the ignorance is partial. If one of the pair is $(0,0)$, then ignorance is total.

Definition 3.3. *An uncertain formal context can be defined as follows (Dubois and Prades, 2015):*

$$\Re'=\{(\alpha(o,p),\beta(o,p))|o\in O,p\in P\} \quad (4)$$

To compute formal concepts we can replace the $(\alpha(o,p),0)$ by 1 and $(0,\beta(o,p))$ by 0 to obtain a binary formal context. Then we can compute formal concepts by using an existing algorithm.

Definition 3.4. *The necessity measure (certainty) of a formal concept* $C=(O_i,P_i)$ *can be computed by using the following formula:*

$$N(C)=\min_{o\in O_i,p\in P_i} N((o,p)\in\Re) \quad (5)$$

To illustrate the certainty computation we propose the following example:

Table 2: Example of an uncertain formal context.

Objects \ Properties	p_1	p_2	p_3
o_1	(0,1)	(0,1)	(0.4,0)
o_2	(0,0.3)	(1,0)	(1,0)
o_3	(0,0.7)	(1,0)	(0,0.6)
o_4	(1,0)	(0.5,0)	(0.8,0)
o_5	(1,0)	(0,0.5)	(1,0)

By transforming this context we obtain a binary context:

Table 3: Uncertain formal context to binary formal context.

Objects \ Properties	p_1	p_2	p_3
o_1	0	0	1
o_2	0	1	1
o_3	0	1	0
o_4	1	1	1
o_5	1	0	1

In this example, we can see that $(\{o_4,o_5\},\{p_1,p_3\})$, $(\{o_4\},\{p_1,p_2,p_3\})$, $(\{o_2,o_4\},\{p_2,p_3\})$,$(\{o_2,o_3,o_4\},\{p_2\})$ and $(\{o_1,o_2,o_4,o_5\},\{p_3\})$ are the formal concepts of this formal context. We computed the certainty of these formal concepts:

Table 4: Computation of certainty.

Formal Concepts	Certainties
$(\{o_4,o_5\},\{p_1,p_3\})$	0.8
$(\{o_4\},\{p_1,p_2,p_3\})$	0.5
$(\{o_2,o_4\},\{p_2,p_3\})$	0.5
$(\{o_2,o_3,o_4\},\{p_2\})$	0.5
$(\{o_1,o_2,o_4,o_5\},\{p_3\})$	0.4

We propose another example to illustrate the processing of quantitative properties. We consider the following context:

Table 5: Multivalued context example.

Objects \ Properties	Age	Gender
o_1	5	Man
o_2	35	Woman
o_3	50	Man
o_4	19	Man
o_5	80	Woman

We propose to use possibility theory and a linguistic variable to transform the multivalued property concerning age. We define for example three distributions of possibility for age (young, adult and old):

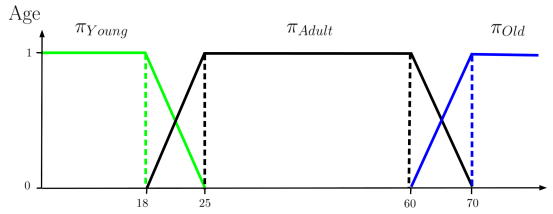

Figure 1: Linguistic variable.

We compute the membership degree of the three possibility distributions, then we perform a renormalization before computing the measure of necessity. When the property is qualitative (nominal or ordinal), we create as many properties as modalities. If we apply this to our example, we obtain:

Table 6: Transforming multivalued context into uncertain context.

Objects \ Properties	Age_{Young}	Age_{Adult}	Age_{Old}	Man	Woman
o_1	(1,0)	(0,1)	(0,1)	(1,0)	(0,1)
o_2	(0,1)	(1,0)	(0,1)	(0,1)	(1,0)
o_3	(0,1)	(1,0)	(0,1)	(1,0)	(0,1)
o_4	(0.83,0)	(0,0.83)	(0,1)	(1,0)	(0,1)
o_5	(0,1)	(0,1)	(1,0)	(0,1)	(1,0)

Many algorithms can be used to compute all formal concepts (S. O. Kuznetsov, 2003). We chose for our experiment to use a parallel recursive algorithm (Krajča et al., 2008). This algorithm takes in input a binary context and computes all the formal concepts.

From the concept lattice, we can extract association rules that represent the dependencies between the properties.

Definition 3.5. *An association rule is a pair of itemset written $P_1 \rightarrow P_2$ where P_1 and P_2 are two sets of properties such as $P_1 \cap P_2 = \oslash$. P_1 is the condition of the rule and P_2 is the conclusion.*

Definition 3.6. *We define the support of the rule noted $\sigma(P_1 \rightarrow P_2)$ by using the following formula:*

$$\sigma(P_1 \rightarrow P_2) = \frac{\|(P_1 \cup P_2)^{\downarrow}\|}{\|O\|} \quad (6)$$

Definition 3.7. *The confidence of the rule $conf(P_1 \rightarrow P_2)$ can be computed as follows:*

$$conf(P_1 \rightarrow P_2) = \frac{\sigma(P_1 \rightarrow P_2)}{\sigma(P_1)} \quad (7)$$

With '$\sigma(P_1) = \frac{\|P_1^{\downarrow}\|}{\|O\|}$

We consider in this research only the properties that satisfy $N((o,p) \in \Re) > 0$ for the computation of the support and confidence. Usually, there are two thresholds θ_σ and θ_{conf} that allow us to select relevant association rules. If $conf(A \rightarrow B) = 1$ then the rule is exact, else the rule is approximate. If $\theta_\sigma = 1$ then the rule has at least one object with this profile. Finally, another measure can be proposed that corresponds to the necessity of the rule.

Definition 3.8. *The necessity degree of the rule noted $N(P_1 \rightarrow P_2)$ can be computed as follows:*

$$N(P_1 \rightarrow P_2) = N(\neg P_1 \cup P_2) = 1 - \Pi(P_1 \cap \neg P_2) \quad (8)$$

P_1 and P_2 are conjunctions of properties. This equation involves a discussion concerning $\Pi(P_1 \cap \neg P_2)$ and $\Pi(P_1 \cap P_2)$. The rule requires that the proposition $P_1 \cap P_2$ is more possible than the proposition $P_1 \cap \neg P_2$. If $\Pi(P_1 \cap P_2) > \Pi(P_1 \cap \neg P_2)$ then the rule $P_1 \rightarrow P_2$ is true. In fact we have $1 - \Pi(P_1 \cap P_2) < 1 - \Pi(P_1 \cap \neg P_2)$ so $N(\neg P_1 \cup \neg P_2) < N(\neg P_1 \cup P_2)$ and finally $N(P_1 \rightarrow \neg P_2) < N(P_1 \rightarrow P_2)$. This constraint means that the certainty of having P_2 if P_1 is true is higher than the certainty of having not P_2 if P_1 is true leading to the rule $P_1 \rightarrow P_2$. For example if $\Pi(P_1 \cap P_2) = 1$ and $\Pi(P_1 \cap \neg P_2) = \alpha$ then $N(P_1 \rightarrow P_2) = 1 - \alpha$ and $N(P_1 \rightarrow \neg P_2) = 0$.

If we consider the simple rule $p \rightarrow q$ where p and q are two different properties of a formal context, we compute the certainty of the rule as follows:

$$N(p \rightarrow q) = \min_{o \in O}[1 - \Pi((o,p) \in \Re \cap (o,q) \notin \Re)] \quad (9)$$

By using the property $\Pi(A \cap B) \leq \min(\Pi(A), \Pi(B))$ and if we consider that the minimum is the maximum value of the possibility $\Pi(A \cap B)$ then we propose:

$$N(p \rightarrow q) = \min_{o \in O}[1 - \min(\Pi((o,p) \in \Re), \Pi((o,q) \notin \Re))] \quad (10)$$

This formula can be easily generalized for several properties in the condition and conclusion of the rule. For example, we consider the following uncertain context and we want to compute the certainty of the rule $N(p_2 \rightarrow p_3)$.

Table 7: Example of uncertain formal context.

Objects \ Properties	p_1	p_2	p_3
o_1	(0,1)	(0,1)	(0.4,0)
o_2	(0,0.3)	(1,0)	(1,0)
o_3	(0,0.7)	(1,0)	(0.6,0)
o_4	(1,0)	(0.5,0)	(0.8,0)
o_5	(1,0)	(0,0.5)	(1,0)

By applying the previous formula we obtain:

$$N(p_2 \rightarrow p_3) = \min(1,1,0.6,0.8,1) = 0.6 \quad (11)$$

We can also compute the support of this rule:

$$\sigma(p_2 \rightarrow p_3) = \frac{\|(p_2 \cup p_3)^{\downarrow}\|}{\|O\|} = \frac{3}{5} = 0.6 \quad (12)$$

We can see that the support of the rule $p_2 \rightarrow p_3$ is the frequency of the rule in the context. It is also the probability that an object has the properties p_2 and p_3. Then we compute the confidence of the rule:

$$conf(p_2 \rightarrow p_3) = \frac{\sigma(P_2 \rightarrow P_3)}{\sigma(P_2)} = \frac{0.6}{0.6} = 1.0 \qquad (13)$$

We can deduce from the above formula that the confidence of the rule $p_2 \rightarrow p_3$ is the percentage of objects that have the property p_3 when they have the property p_2.

4 EXPERIMENTATION WITH A TEXT CORPUS IN FRENCH

4.1 Preprocessing and Context Generation

The goal of preprocessing is to gather variables that represent exogenous information and the texts of the corpus into an uncertain formal context. Variables can be quantitative and qualitative. We can represent knowledge about a quantitative variable by using a linguistic variable. Multivalued variables are also transformed. Segmentation can be performed to divide the initial texts into segments. Each segment inherits the values of the variables associated with the initial text. A processing is applied to clean the new corpus of texts made up of segments. This cleaning consists in changing the case and eliminating unwanted characters, punctuation, numbers, and unnecessary words. Below, we present a summary of the processing:

Figure 2: Corpus processing.

Then we apply a lemmatization that significantly reduces the size of the dictionary. To do this, we used an existing French lexicon that associates each word with its lemma. The synonyms were not considered in this study. When a word is not found in the lexicon, we look for the closest word. For each word of the lexicon we associate a degree of possibility computed by using the Jaro-Winkler (Winkler, 1999) distance between the current word and the word in the lexicon. The word of the lexicon with the highest degree of possibility is chosen. If the possibility of this word is less than a threshold, the original word is kept, leading to a pair of necessity measures $(1,0)$ in the context if the word is in a segment. Otherwise, the word is replaced by the lemma of the lexicon. Then we renormalise the degrees of possibility before computing a measure of necessity noted α. The pair of necessity degrees is $(\alpha,0)$ if the lemma is present in a segment of text. The uncertain context is then generated from the segments of texts and variables transformed into properties.

4.2 Context Information

The uncertain formal context can be first analysed by using usual data analysis tools before applying formal concept analysis. To provide some examples of results, we propose to analyse the computer science curriculum in French high schools in 2019. We do not use additional variables in this study. The corpus is split into segments and we consider only the 50 most frequent terms of the corpus to generate the uncertain context. The properties are the terms of the corpus and the object the segment identifier. First, it is possible to compute a co-occurrence matrix of properties from the uncertain context. Then, we perform a similarity analysis and generate a similarity graph. By computing the maximum spanning tree we obtain a much more readable graph. We compute the uncertainty $I(m_1,m_2)$ (noted I) of the co-occurrence between two properties m_1 and m_2 of the uncertain context $(O,P,\Re)$ by using the following formula:

$$I = \min_{o \in O}[\min(N((o,m_1) \in \Re), N((o,m_2) \in \Re))] \qquad (14)$$

A descending hierarchical classification (DHC) of the segments makes it possible to highlight the terms that are often found together in the segments. We propose a visualization of the classification on the first two principal components of the factorial correspondence analysis. For example we obtain:

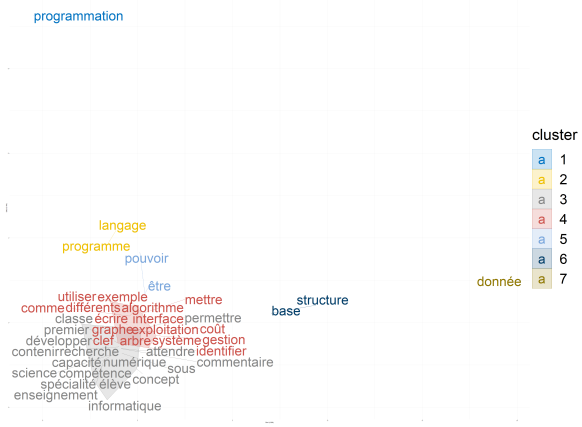

Figure 3: DHC on the first two principal components with 7 classes.

In this graph, proximity between the terms reveals their links in the corpus when the quality of projection is good. It is also possible to look for an interpretation of the factorial axes. We can visualize the dendrogram of the classification as follows:

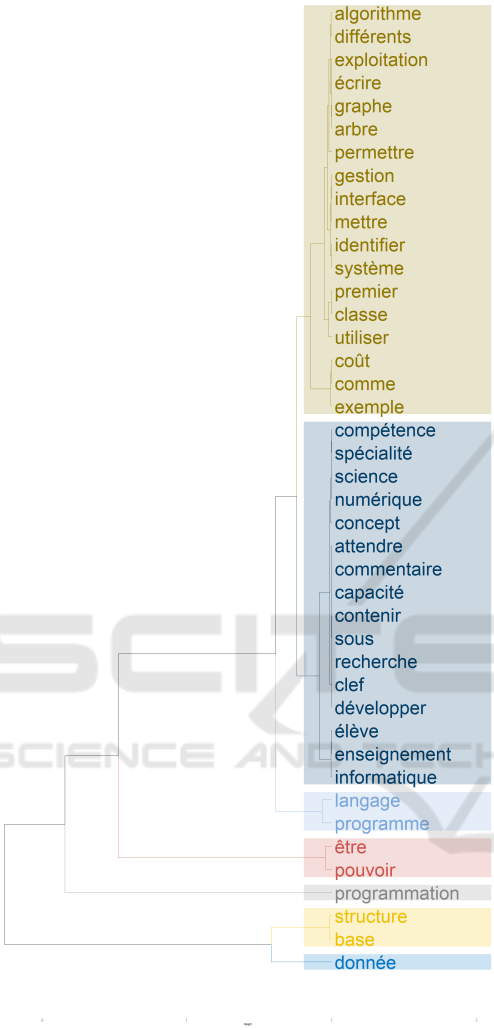

Figure 4: Dendrogram with 7 classes.

The dendrogram represents all words in classes and can be very useful for data analysis.

4.3 Formal Concepts Analysis

We consider now all terms to generate the uncertain context. Formal concepts can be used to represent terms that are present together in text segments. We have developed several visualization tools. First of all, we propose a visualization of the formal concepts in a table that can be sorted according to the number of objects, number of properties, certainty, or the relevance score computed from a query. Next, we propose a representation with a Hasse diagram of the concept lattice. We also propose for a formal concept the similarity graph of the properties of the formal concept to represent the links between the properties. Finally, we propose to visualize the properties of a formal concept in the first two principal components. When the number of formal concepts becomes very important, visualization tools do not necessarily allow us to see the information we are looking for. We therefore propose two solutions. The first is to consider only the words with a frequency above the threshold. This makes it possible to limit the size of the dictionary without losing the most important words. The second solution we propose is the use of queries in a relatively simple graphical language. Indeed, the graphical language makes it possible to take into account criteria of the presence or absence of a property or an object. It is possible to perform a multicriteria combination by using AND, OR and NOT operators. We also propose functions that calculate, for example, cardinality (the number of properties or objects in a formal concept), a score defined in (Petiot, 2019). Criteria can be applied to the results of these functions. Our solver can manage binary criteria or uncertain criteria. The solver generates a possibilistic network before evaluating the query. Combination operators are uncertain logical gates (Petiot, 2019) that can represent traditional logical combinations and uncertain logical combinations. We compile the query into a circuit to improve the computation time and we compute a relevance score for each formal concept. This solution allows us to manage the uncertainties of the query which is illustrated by the following example:

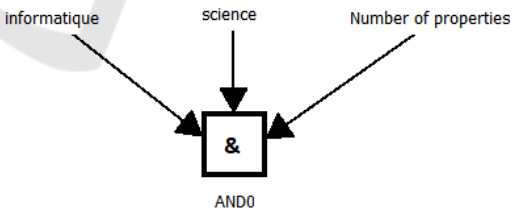

Figure 5: Graphical query to search formal concepts.

In this query, we want to retrieve the formal concepts that contain the French words *"informatique"* and *"science"*. We also want to keep only formal concepts with a limited number of properties because the visualization would be degraded. To translate this constraint we used two possibility distributions that represent the states of the variable *"Number of properties"*:

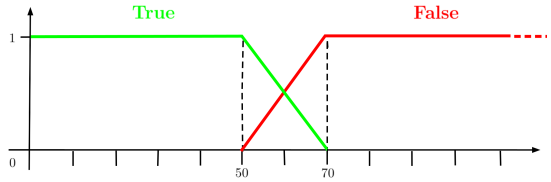

Figure 6: Possibility distributions of the variable *"Number of properties"*.

The criterion *true* allows us to select only the formal concepts that satisfy the constraint. Moreover, we can see that formal concepts with less than 60 properties are accepted and others rejected. In the same way, we can define states for variables associated with the terms, for example here the terms *"informatique"* and *"science"*. The result of the query evaluation is as follows:

Concept	Certainty	Relevance	Number of properties	Number of objects
C3	1	1	27	1
C7	0.011	1	18	1
C0	1	1	14	1
C2	1	1	13	1
C1	1	1	12	1
C9	1	1	6	2
C8	1	1	6	2
C6	1	1	6	2
C5	1	1	6	2
C10	1	1	4	4
C4	1	1	3	5

Figure 7: Query result.

We can deduce the similarity graph for each formal concept of the query result. For the formal concept denoted C7 we obtain the similarity graph below:

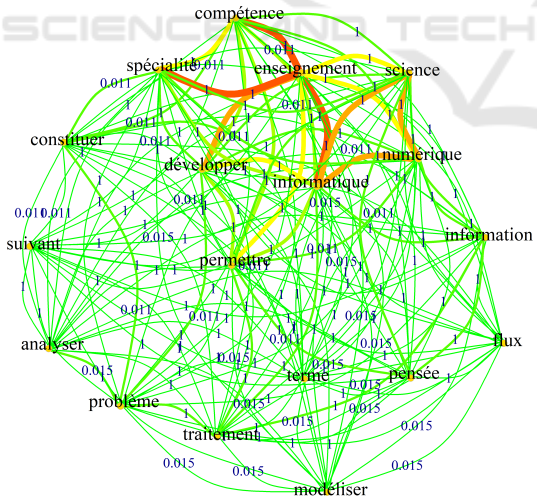

Figure 8: Similarity graph with uncertainties.

In the graph, the edges between the words are represented by a gradient of colours proportional to the co-occurrence index. A maximum tree can also be computed.

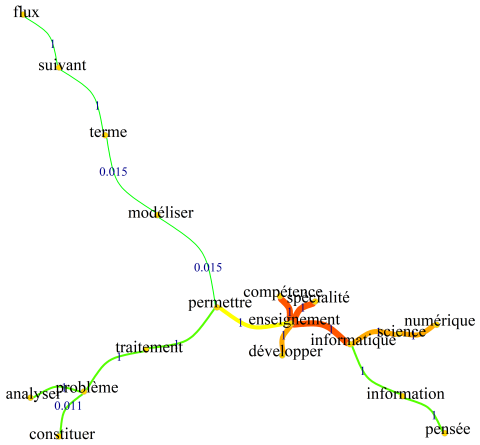

Figure 9: Maximum tree with uncertainties.

It is possible to generate a Hasse diagram with the certainty of the formal concepts.

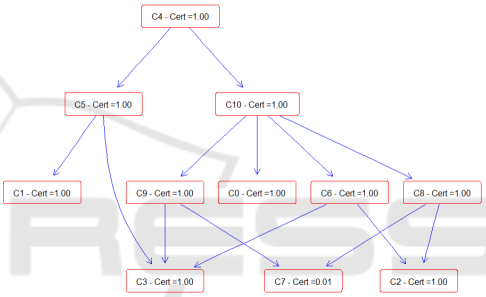

Figure 10: Hasse diagram with uncertainties.

Finally, we can deduce the rules from the Hasse diagram. The rules are presented with their quality measures: confidence, support and certainty.

	Rule	Confidence	Support	Certainty
0	compétence permettre -> enseignement	1	2.91262 %	1
1	cycle terminal -> enseignement	1	1.94175 %	1
2	développer spécialité -> enseignement	1	1.94175 %	1
3	classe collège commun dispenser école mathématique ...	1	0.970874 %	1
4	acquérir convenir culture maximum placer situation -> ...	1	0.970874 %	1
5	acquisition certification faire fin objet -> cycle terminal	1	0.970874 %	1

Figure 11: Example of the first five rules.

5 CONCLUSION

In this research, our goal was to combine text analysis with formal concept analysis to propose a new mixed data analysis solution. We associated each text of the corpus with a set of variables that can be qualitative or quantitative. Next, we performed a lemmatization of the corpus to reduce the vocabulary of the dictionary. Quantitative variables were replaced by linguistic variables. This made it possible to calculate cer-

tainty for each modality. Qualitative variables were transformed into binary variables. Finally, we obtained a representation of the corpus and variables in an uncertain context. We computed uncertain formal concepts and showed that it was possible to visualize the links between words in a formal concept by using similarity analysis. By projecting formal concepts on the first two principal components of factorial correspondence analysis we visualized the relationships between terms. Finally, the graphical queries made it possible to highlight the essential terms. Moreover, they improve computation time and they reduce exploration time for the user. Our perspective is to experiment and improve this approach. We will improve and compare the solutions for the preprocessing of the corpus. We plan to collaborate with researchers in the humanities to test our solution with practical applications.

REFERENCES

Ait-Yakoub, Z., Djouadi, Y., Dubois, D., and Prade, H. (2016). From a possibility theory view of formal concept analysis to the possibilistic handling of incomplete and uncertain contexts. In *5th International Workshop "What can FCA do for Artificial Intelligence?" (FCA4AI 2016) co-located with ECAI 2016*, pages 79–88.

Bělohlávek, R. (2004). Concept lattices and order in fuzzy logic. In *Annals of Pure and Applied Logic*, volume 128, pages 277–298.

Bělohlávek, R., Sklenar, V., Zacpal, J., and Sigmund, E. (2007). Evaluation of questionnaires by means of formal concept analysis. In *Int. Conference on Concept Lattices and Their Applications*, pages 100–111. J. Diatta, P. Eklund, M. Liquiere (Eds.): CLA 2007.

Cimiano, P., Hotho, A., and Staab, S. (2005). Learning concept hierarchies from text corpora using formal concept analysis. In *Journal of Artificial Intelligence Research*, volume 24, pages 305–339.

Degenne, A. and Vergès, P. (1973). Introduction à l'analyse de similitude. In *Revue française de sociologie [Sciences Po University Press, Association Revue Française de Sociologie]*, volume 14, pages 471–512.

Dubois, D., de Saint-Cyr, F. D., and Prade, H. (2007). A possibility-theoretic view of formal concept analysis. In *Fundamenta Informaticae, IOS Press*, volume 75, pages 195–213.

Dubois, D. and Prade, H. (1988). Possibility theory - an approach to computerized processing of uncertainty.

Dubois, D. and Prades, H. (2015). Formal concept analysis from the standpoint of possibility theory. In *J. Baixeries, C. Sacarea, M. Ojeda-Aciego (eds) Formal Concept Analysis. ICFCA 2015. Lecture Notes in Computer Science, Springer*, volume 9113, pages 21–38.

Fernandez-Manjon, B. and Fernandez-Valmayor, A. (1998). Building educational tools based on formal concept analysis. In *Journal of Education and Information Technologies*, volume 3, pages 187–201.

Hotho, A., Nürnberger, A., and Paass, G. (2005). A brief survey of text mining. In *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, volume 20, pages 19–62.

Krajča, P., Outrata, J., and Vychodil, V. (2008). Parallel recursive algorithm for FCA. In *In: Proc. CLA 2008, CEUR WS*, pages 71–82.

Petiot, G. (2019). Information retrieval in a concept lattice by using uncertain logical gates. In *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K 2019*, volume 1, pages 289–296.

Poelmans, J., Ignatov, D., Kuznetsov, S. O., and Dedene, G. (2013). Formal concept analysis in knowledge processing: A survey on applications. In *Expert Systems with Applications*, volume 40, pages 6538–6560.

Poelmans, J., Ignatov, D., Kuznetsov, S. O., and Dedene, G. (2014). Fuzzy and rough formal concept analysis: A survey. In *International Journal of General Systems*, volume 43.

Reinert, A. (1983). Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte. In *Cahiers de l'analyse des données 8.2*, pages 187–198.

S. O. Kuznetsov, S. A. O. (2003). Comparing performance of algorithms for generating concept lattices. In *J. Experimental & Theoretical Artificial Intelligence*, volume 14, pages 189–216.

Salem, A. (1986). Segments répétés et analyse statistique des données textuelles, etude quantitative à propos du père duchesne de hébert. In *Histoire & Mesure, Ed. du CNRS*, volume 1.

Snášel, V., Horák, Z., and Abraham, A. (2008). Understanding social networks using formal concept analysis. In *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 390–393.

Tovar, M., Pinto, D., Montes, A., Serna, G., and Vilariño, D. (2015). Patterns used to identify relations in corpus using formal concept analysis. In *Arrasco-Ochoa, J., Martinez-Trinidad, J., Sossa-Azuela, J., Olvera Lòpez, J., Famili, F. (eds) Pattern Recognition. MCPR 2015. Lecture Notes in Computer Science, Springer*, volume 9116, pages 236–245.

Wille, R. (1982). Restructuring lattice theory: an approach based on hierarchies of concepts. In *I. Rival, (ed.) Ordered Sets. Reidel, Dordrecht-Boston*, pages 445–470.

Winkler, W. E. (1999). The state of record linkage and current research problems. In *Statistical Research Division, U.S. Bureau of the Census*.

Zadeh, L. A. (1978). Fuzzy sets as a basis for a theory of possibility. In *Fuzzy Sets and Systems*, volume 1, pages 3–28.